Université de Montréal

Genomic variation in recombination patterns: implications for disease and cancer

par

Julie Hussin

département de biochimie

Faculté de médecine

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Ph. D.

en Bio-informatique

Février 2013

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Genomic variation in recombination patterns: implications for disease and cancer

présentée par :

Julie Hussin

évaluée par un jury composé des personnes suivantes:

Sylvie Mader

président-rapporteur

Philip Awadalla

directeur de recherche

Guy Rouleau

membre du jury

Gil McVean

examinateur externe

Guillaume Lettre

représentant du doyen de la FES

Abstract

The intergenerational mixing of DNA through meiotic recombination of homologous chromosomes is, along with mutation, a major mechanism generating diversity and driving the evolution of genomes. In this thesis, I use bioinformatics and statistical approaches to analyse modern genomic data in order to study the implication of meiotic recombination in human disease. First, using high-density genotyping data from French-Canadian families, we studied sex- and age-specific effects on recombination patterns. These analyses lead to the first observation of a significant decrease in recombination rates with advancing maternal age in humans, with potential implications for understanding trisomic conceptions. Second, using next-generation sequencing of exomes from families of children with leukemia, we discovered unusual distributions of recombination breakpoints in some leukemia patients, which implicates PRDM9, a protein involved in defining the location of recombination breakpoints, in leukemogenesis. Third, using single nucleotide polymorphisms (SNPs) called from RNA sequencing data, we present a detailed comparison of the mutational burden between high and low recombining regions in the human genome. We further show that the mutational load in regions of low recombination at the individual level varies among human populations. In analysing genomic data to study recombination in population and disease cohorts, this work improves our understanding of how recombination impacts human health. Furthermore, these results provide insights on how variation in recombination modulates the expression of phenotypes in humans.

Keywords : genetic recombination, sequencing, PRDM9, population genetics, leukemia

Résumé

Durant la méiose, il se produit des échanges réciproques entre fragments de chromosomes homologues par recombinaison génétique. Les chromosomes parentaux ainsi modifiés donnent naissance à des gamètes uniques. En redistribuant les mutations génétiques pour générer de nouvelles combinaisons, ce processus est à l'origine de la diversité haplotypique dans la population. Dans cette thèse, je présente des résultats décrivant l'implication de la recombinaison méiotique dans les maladies chez l'humain. Premièrement, l'analyse statistique de données de génotypage de familles québécoises démontre une importante hétérogénéité individuelle et sexe-spécifique des taux de recombinaisons. Pour la première fois chez l'humain, nous avons observé que le taux de recombinaison maternel diminue avec l'âge de la mère, un phénomène potentiellement impliqué dans la régulation du taux d'aneuploïdie associé à l'âge maternel. Ensuite, grâce à l'analyse de données de séquençage d'exomes de patients atteints de leucémie et de ceux de leurs parents, nous avons découvert une localisation anormale des évènements de recombinaison chez les enfants leucémiques. Le gène *PRDM9*, principal déterminant de la localisation des recombinaisons chez l'humain, présente des formes alléliques rares dans ces familles. Finalement, en utilisant un large spectre de variants génétiques identifiés dans les transcriptomes d'individus Canadiens Français, nous avons étudié et comparé le fardeau génétique présent dans les régions génomiques à haut et à faible taux de recombinaison. Le fardeau génétique est substantiellement plus élevé dans les régions à faible taux de recombinaison et nous démontrons qu'au niveau individuel, ce fardeau varie selon la population humaine. Grâce à l'utilisation de données génomiques de pointe pour étudier la recombinaison dans des cohortes populationnelles et médicales, ce travail démontre de quelle façon la recombinaison peut affecter la santé des individus.

Mots clés: recombinaison génétique, séquençage, PRDM9, génétique des populations, leucémie

Table of Content

List of Figures

CHAPTER I

CHAPITRE II

CHAPITRE III

List of Tables

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

Abbreviations and Acronyms

ACA : Acadians of Quebec

ALL : Acute Lymphoblastic leukemia

B-ALL : B-cell precursor ALL

BCR : Breakpoin Cluster Region

ANOVA : Analysis of variance

CaG : CARTaGENE

CEU : HapMap population from Utah residents with Northern and Western European ancestry

CEPH: Centre d'Etude du Polymorphisme Humain

CI : Confidence Intervals

CVD : Cardio-Vascular Disease

CNO : North Shore Region

cM : centiMorgan

CMT1A : Charcot-Marie Tooth Disease type 1A

DGS : DiGeorge syndrome

d_N : rate of nonsynonymous substitution

DNA : Deoxyribonucleic acid

d_S : rate of synonymous substitution

DSB : Double-Strand Break

FAB-L1 : French American British subtype L1

FC : French-Canadian

FoSTeS : Fork Stalling and Template Switching

FRS : Framingham Risk Score

GAS : Gaspesia Region

Gb : Gigabase

GC : Guanine-Cytosine

gBGC : GC-Biaised Gene Conversion

H3K4me3 : Histone H3 lysine 4 trimethylation

hg18 : March 2006 Human Genome Version

hg19 : February 2009 Human Genome Version

HGDP : Human Genome Diversity Project

HJ : Holliday Junction

HLA : human leukocyte antigen

HNPP : hereditary neuropathy with liability to pressure palsies

HRR : High Recombination Regions

HS: Hotspot

HWE : Hardy-Weinberg equilibrium

IgH : Immunoglobulin Heavy chain

indOR : Odds Ratio per individual

Kb : Kilobase

L1 : Long Interspersed Element 1

LCR : Low Copy Repeats

LD : Linkage Disequilibrium

LS-CHD : Left-Sided Congenital Heart Disease

LOH : Loss of heterozigosity

LOY : Loyalists of Quebec

MAF : Minor Allele Frequency

Mb : Megabase

MMBIR : Microhomology-Mediated Break-Induced Replication

MRCA : Most Recent Common Ancestor

MTL : Montreal Area

NA : non-applicable

NAHR : Non-Allelic Homologous Recombination

NCBI : National Center for Biotechnology Information

N_e : Effective population size

NF1 : Neurofibromatosis type 1

NGS : Next-Generation Sequencing

NHEJ : Non-homologous End-Joining

ns : non-significant

OR : Odds Ratio

PCA : Principal Component Analysis

PCR : Polymerase Chain Reaction

PHD : Plant Homeodomain

Ph+ : Philadelphia chromosome positif

PWS : Prader-Willi syndrome

QCC : Quebec City

REML : Restricted maximum likelihood

RNA : Ribonucleic acid

RNAseq : RNA sequencing

RPQ : Reference Panel of Quebec

RSS : Recombination Signal Sequence

SAG : Sagnenay Region

SC : Synaptonemal Complex

SCID : Severe Combined Immunodeficiency

SDSA : Synthesis-Dependent Strand-Annealing

SJ : St. Jude Children's Research Hospital

SMA : Spinal Muscular Atrophy

SMS : Smith-Magenis syndrome

SNP : single-nucleotide polymorphism

T-ALL : T-cell precursor ALL

TCR : T-cell receptor

TSS : Transcription Starting Site

UCSC : University of California, Santa Cruz

WBS : Williams-Beuren syndrome

YRI : HapMap Yoruba population from Ibadan, Nigeria

ZnF : Zinc Fingers

Acknowledgments

It is my great pleasure to acknowledge all of those who have contributed in many ways to make this thesis possible.

First of all, I wish to express my endless thanks to my advisor, Philip Awadalla. I am grateful for his expert and valuable guidance and constant encouragement, for his enthusiasm and inspiring vision. I particularly want to thank him for giving me the opportunity to join his laboratory and work on diverse and exciting projects, as it was him who convinced me to pursue doctoral studies.

I would also like to thank all present and past members of the Awadalla lab 'family', for their help and friendship. In particular, I would like to thank Jacklyn Quinlan and Youssef Idaghdour, with whom I not only shared an office, but also times of laughters, panic and philosophical reflections. I also thank Ferran Casals and Elias Gbeha, for their valuable help with experimental work, and Vanessa Bruat, Thibault de Malliard and Jean-Christophe Grenier, for their great work and technical assistance.

I deeply thank my great friend and colleague, Claude Bherer, for always being available to talk endlessly about life and science, for spending hours questioning me about my ideas and for generously sharing hers.

I would like to thank all my professors and colleagues from the Bioinformatics program at University of Montreal. Particularly, I warmly thank Marie Pier Scott-Boyer : we went through so much together during all these years as undergraduate and graduate students. I am also glad to acknowledge the help I received from the administrative staff at Ste-Justine and at the biochemistry department. Elaine Meunier, Alida Hounyovi, Sandy Lalonde and Dominika Kozubska deserve special mention for their kind assistance.

I am very grateful to Sylvie Mader, Gil McVean and Guy Rouleau for taking the time to review this thesis. I also wish to acknowledge Miklos Csuros, François Major, Luis Barreiro, Mark Samuels and Marie-Helene Roy-Gagnon for being on my thesis commitees and for providing useful suggestions and comments.

My sincere thanks go to Gregor Andelfinger, Daniel Sinnett and Charles Mullighan and their research teams, for allowing access to their disease cohorts from Ste-Justine and St. Jude children's hospitals. In addition, I share the credit of my work with all my co-authors on the manuscripts presented in this thesis, and I would like to thank all of the researchers on whose work I was able to build.

I take this opportunity to sincerely acknowledge all patients and their families that participated in these studies, for their generosity in sharing their informations and biologicals to improve health research. I also thank the volunteers in Quebec that participated in the Cartagene project and all members of the Cartagene team, who make this great project a reality.

It is with immense gratitude that I thank my fiancé, Olivier Gandouet, for his unceasing encouragements and patience, and for providing brillant solutions to the many problems I encountered. This thesis would not have been possible without his love and support. I would also like to thank all my family in Belgium, for believing in me since I was a little girl. I'm also very grateful to my dearest friends, Elina Betman, Emy Behmoaram, Daphné Bui, Jade Damdjee and Nadine El Kurdi, for always being understanding and supportive in all my excitements, struggles and frustrations.

Finally, I would like to express my eternal appreciation towards the two most wonderful researchers I know : my parents. Towards my father, Jules Beckers, who left much too early, but who continues to be my source of inspiration in my research and in my life. Towards my mother, Véronique Hussin, who has always been my role model and who has instilled in me confidence and the love of science. Thank you for all your precious advice.

I ackowledge the financial support from biT excellence scholarships, Natural Sciences and Engineering Research Council of Canada (NSERC), the Foundation of Stars, le Réseau de médecine génétique appliquée (RMGA), la Faculté des études supérieures et postdoctorale de l'Université de Montréal (FESP).

I dedicate this work to the memory of my father, Jules Beckers

Only through my future work will I be able to properly thank him

Je t’aime Papa

CHAPTER I: Introduction

THE IMPACT OF RECOMBINATION

IN HUMAN DISEASE

INTRODUCTION

« Ce qui donne à un individu sa valeur génétique, ce n'est pas la qualité propre de ses gènes. C'est qu'il n'a pas la même collection de gènes que les autres. »

Sciences de la vie et société – 1980

François Jacob (1920-2013)

Although mutation is the ultimate source of genetic variation, meiotic recombination is a contributing mechanism playing a powerful role in driving the evolution of genome structure by generating haplotypic diversity. The recombinational process creates new combinations of parental genetic material and thus enables every child to receive a unique mosaic of parental pairs of chromosomes, drawn from an extremely large number of possibilities. Furthermore, mutations involving large insertions and deletions, transpositions and rearrangements mainly arise through recombination mechanisms. It follows that, in breaking down parental associations between variants, recombination allows phenotypic variability across generations. Natural selection acting on novel haplotypes likely favors recombination among lineages, depending on variable environmental factors or life histories. However, tight regulation of the mutational and recombinational processes is needed for the maintenance of genome stability in somatic and germinal cells, as multicellular species experience a very large number of cell divisions per generation resulting in their genomes being particularly prone to errors.

Substantial variation in the rate and distribution of crossovers has been found between and among species, genders, ages, populations and individuals, within and among chromosomes, at a megabase and kilobase scales. In mammals, the distribution of crossovers along the genome is known to vary, and substantial regions of DNA with unusually low recombination are observed, known as coldspots,

while highly localized peaks of recombination, known as hotspots, are also seen. The histone modifier PRDM9 has been identified as a major determinant of the location of recombination hotspots in mammals (Baudat et al. 2010). However, comparing recombination landscapes between primate species shows that hotspots are not conserved and fine-scale recombination rates are markedly different, suggesting a rapid turnover of the recombination landscape on an evolutionary scale. Among humans, variation in fine-scale rates of recombination was observed between populations and individuals and important sex-differences in recombination rates exist. Therefore, many different selective forces may be acting on recombination rates over time. Since natural selection will not act independently on mutations at linked loci, the rate of adaptation will depend on the local mutation and recombination rates.

What causes variation in recombination rates to be so widespread, and what are the consequences for the human genome, remain open questions with important implications for understanding human disease and adaptation. The goal of this chapter is to review work that considers how recombination processes in humans influence disease. Before addressing the potential molecular roles of recombination in human disease, I will address the question of why is recombination adaptative and what benefits and consequences it has on genome evolution. I will then review the approaches available for estimating recombination rates in natural populations and detail the various levels of variation observed within and between genomes. I will further describe the molecular players involved in meiotic and somatic processes of recombination, influencing genome stability. These processes altogether influence human diseases, and pediatric disorders in particular, and the present state of knowledge linking disease and recombination is reviewed.

I. GENOME EVOLUTION AND VARIATION IN RECOMBINATION

Meiotic recombination is the process by which two genomes fuse to produce cells that contain a mixture of genes, before separating as a result of chromosome segregation. Errors in recombination affect genomic integrity in a critical manner, the frequency of crossover should therefore be tighly regulated. However, extensive variation in recombination rates has been observed at all scales, within and between genomes. In this chapter section, I discuss the effects of this variation on genome evolution and how it modulates the efficacy of selection. The existing methods to detect recombination will be presented, followed by a description of the variation of recombination that has been observed in humans.

1. The Evolutionary Advantage of Recombination

How recombination evolved initially and why it has been maintained are critical to understanding why sexual reproduction is so widespread in our world. Producing offspring sexually has a two-fold cost in fitness, because only females are capable of bearing youngs, such that male and female individuals do not contribute resources to the offspring equally (Lehtonen et al. 2012). Furthermore, only half the genes in the offspring are from each parent, who survived to reproductive age. As these mates have genomes that proved themselves successful in the current environment, it is puzzling that they would shuffle their genotypes by recombination and risk creating genomes with lower fitness. In line with this idea, there is theoritical evidence that at loci that control recombination, alleles that decrease it are favored (Kimura 1956; Lewontin 1971). Selection should therefore operate to eliminate sexual recombination in a population. However, and despite the additional cost of securing a mate to reproduce, the vast majority of species reproduce sexually (Engelstadter 2008), with varying levels of genetic mixing (Awadalla 2003). This is known as the paradox of sex (Otto and Lenormand 2002).

A direct effect of genetic recombination, that might explain its origin, is the transfer of genetic elements from one genomic background to another, allowing for the spread of genetic elements such as phage, bacterial plasmids or transposable elements (Otto and Lenormand 2002). Another molecular phenomenon that could account for the origin of genetic exchange is that it is associated with the repair of DNA damage (Bernstein et al. 1981). These associations might explain the origin of sex and recombination, but are unlikely to explain why recombination was not eliminated by natural selection, despite its costs.

Evolutionary explanations for recombination, stating that recombination exists because mixing genetic material from two individuals is by itself beneficial, were first presented in the classical work of Fisher (1930) and Muller (1932). These authors described that favorable mutants which arise in different genomes can be combined into the same genome by recombination, which favors the fixation of beneficial alleles in populations with recombination. In a population with no recombination, two favorable mutations will succeed in fixing only if they appeared on the same genetic background, i.e. if the second mutation occurs in a descendant of the individual in which the first mutation arose. Otherwise, only one can ultimately be fixed and most of the new favorable mutations will be lost, slowing down the rate of evolution. These concepts were further extended by Hill and Roberston to take into account the effect of genetic drift, linkage and selection (Hill and Robertson 1966). They showed that, with no recombination, beneficial mutations at linked loci and occurring on different genetic background will interfere with one another's fixation, thereby decreasing the effectiveness of selection. This is known as "Hill-Robertson interference" (Figure 1a). H.J. Muller introduced a similar effect, known as "Muller's ratchet", to describe how genomes of an asexual population accumulate deleterious mutations in an irreversible manner (Muller 1964). Muller considered the case of deleterious alleles instead of advantageous ones, and pointed out that in the absence of recombination, natural selection could never reduce the number of linked deleterious alleles. This is because repeated losses of chromosomes with the

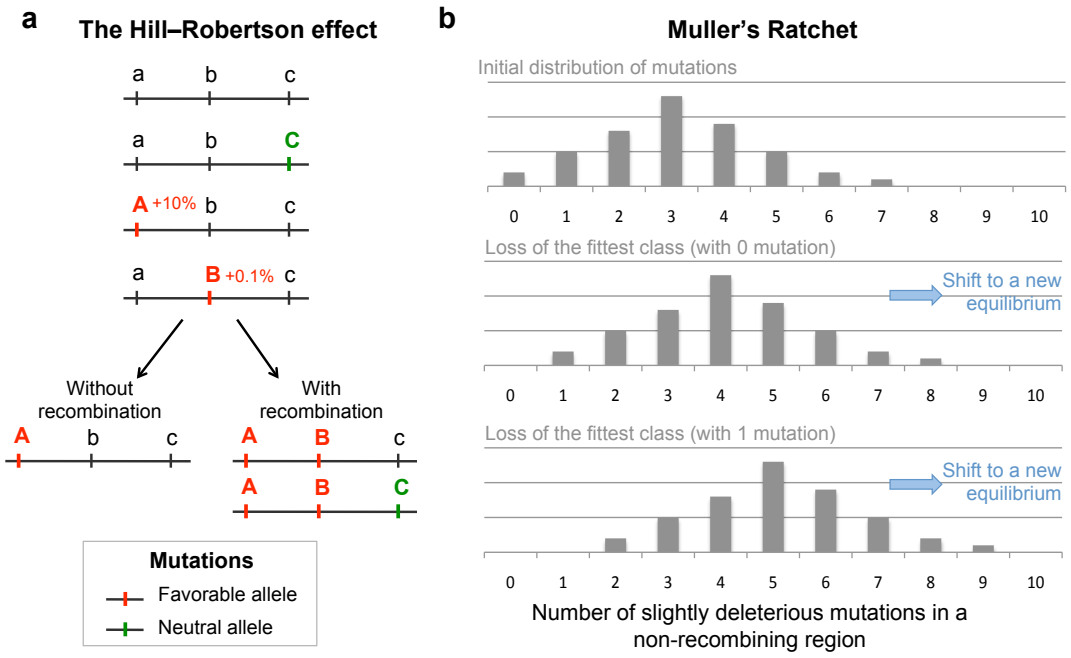


Figure 1. Effect of recombination on accumulation of favorable and deleterious alleles.

(a) The Hill–Robertson interference. Three linked sites are considered with nucleotide a, b and c. The new alleles A and B are advantageous and provide fitness increases of 10% and 0.1%, respectively, whereas C is neutral. Without recombination, A is rapidly fixed in the population because of its large fitness advantage, along with b and c due to genetic hitchhiking. With recombination, the fittest combinations ABc and ABC can be generated and become predominant in the population [redrawn from (Marais and Charlesworth 2003)]. (b) Muller's ratchet. The top diagram shows the initial distribution of individuals with 0, 1, 2, … slightly deleterious alleles in a population without recombination at equilibrium. The middle diagram shows the same population when the fittest class is lost by chance. With no recombination and no back mutation, this class can never be recovered. The whole distribution shift to the right after the ratchet "has clicked round one notch" (Muller 1964). The process is then repeated (lower diagram), leading to an increase in mutational load and decrease in fitness.

smallest number of mutations will lead to a gradual accumulation of deleterious mutations (Figure 1b). Genomic regions with recombination are far less subject to this effect, since mutations at different sites can disentangle themselves from initial chance associations. Muller's ratchet mechanism thus predicts an increase in the average number of deleterious alleles in genomic regions with no recombination. Following these seminal articles, several authors presented computer simulations to verify the theoritical models (Felsenstein 1974; Birky and Walsh 1988; McVean and Charlesworth 2000; Gordo et al. 2002), confirming that recombination should speed up the response to selection in a population, that slightly deleterious mutations should accumulate in non-recombining regions and that the chance of fixation of a positively selected new mutation is greatly reduced. J. Felsenstein stated that this impact on mutational load may be the most quantitatively important evolutionary effect of recombination (Felsenstein 1974), since there must be far more deleterious alleles occurring than beneficial ones.

There are, however, conditions under which uncoupling variants at different sites is not advantageous. In the absence of linkage, genes are already well mixed in the population and shuffling genomes further by recombination will have no effect on the population fitness. Therefore, the advantageous effect of recombination will depend on the patterns of linkage between loci that have been built over time. The actual linkage disequilibrium patterns may therefore be of major importance in determining if and where natural selection will act. Furthermore, recombination will be advantageous only if the associations between loci are negative, when favorable and deleterious alleles are linked to each other. Linkage disequilibrium may also be positive, when favorable (or deleterious) alleles are linked to one another, a scenario where recombination is not advantageous, because it slows down the fixation of beneficial mutations (or the removal of detrimental mutations).

2. Intragenomic Variation in Recombination and Natural Selection

Within nuclear genomes of eukaryotes, considerable variation in the local rate of recombination has been observed from genetic data. At a genomic-scale, at least one crossover event per autosome is necessary and the average recombination rate is negatively correlated with genome size (Awadalla 2003). Sex chromosomes harbor large regions with increased recombination, such as pseudoautosomal regions, and regions depleted of recombination, such as for the mammalian Y chromosome. Broad-scale variation at the chromosomal level has been identified in mammals: centromeric regions show consistently low recombination rates whereas telomeric regions tend to have particularly high rates (Choo 1998; Jensen-Seaman et al. 2004). Furthermore, factors such as DNA methylation, genetic imprinting and germline transcription correlate with large scale recombination rates (Lercher and Hurst 2003; Sigurdsson et al. 2009; McVicker and Green 2010). At a finer scale, it is now well described that recombination rates are not uniformly distributed throughout a chromosome, and occur largely in recombination hotspots of 1-2Kb that exhibit a rate of recombination up to 100-fold higher than the genome average.

Empirical estimation of the local recombination rate have prompted researchers to study the consequences of recombination on genetic diversity since the late 1980s, after the introduction of the concept of hitchhiking of beneficial mutations by positive selection (Smith and Haigh 1974). A positive correlation between local recombination rate and level of silent variability was found, originally described in *Drosophila* (Aguade et al. 1989; Begun and Aquadro 1992). A decade later, it was established that genomic regions with essentially no recombination deteriorate faster than recombining regions (Charlesworth and Charlesworth 2000), forming partial evidence that suppression of recombination *per se* reduces the efficacy of selection. However, there is little empirical evidence that variation in the recombination rate leads to significant variation in the efficacy of selection across the genome (Webster and Hurst 2012).

Studies have investigated this question by evaluating the relationship of local recombination rates and measures of the rate of adaptation, such as as d_N/d_S and codon usage bias (Yang and Nielsen 1998; Hey and Kliman 2002). Codon usage bias measures the proportion of "preferred" codons, coded by the most abundant tRNAs, and is used as an estimate of selection in species with large population size. It was found to correlate positively with recombination rate in *Drosophila* and *Caenorhabditis elegans*. However, recombination rates correlate with content in guanine and cytosine (GC-content) in many taxa, potentially explaining the effect of recombination on codon usage bias (Marais et al. 2001). The ratio of the rate of nonsynonymous substitution on the rate of synonymous substitution, d_N/d_S, reflects the selective forces acting on non-silent sites relative to silent ones that are much less constrained. When selection is acting, d_N/d_S departs from 1: it is higher if most nonsynonymous sites are beneficial, or reduced if most sites are deleterious. Because both adaptation and purifying selection may occur at many genes, it is difficult to know *a priori* whether a positive or negative correlation between d_N/d_S and recombination rate is expected. Conflicting results about the effect of recombination rates on d_N/d_S have been found in *Drosophila* (Larracuente et al. 2008) and, in humans, d_N/d_S is constant across crossover rate categories (Bullaughey et al. 2008).

However, many studies have shown that recombination rates are different between species, even closely related ones, questioning whether the use of d_N/d_S for examining the impact of recombination on the efficacy of selection is a satisfactory approach. Although the average number of recombination events per chromosome is likely conserved across a wide array of species, there is substantial variation in fine-scale recombination rates among multicellular organisms (Awadalla 2003; Wilfert et al. 2007). Furthermore, although humans and chimpanzees share around 99% of their genome, the precise location of recombination hotspots is not conserved between them, suggesting a rapid turnover of the recombination landscape between closely related species (Ptak et al. 2005; Auton et al. 2012). If the

local recombination rate is fast evolving, currently observed recombination rates are unlikely to reflect the recombinational history that may have diverged since the reproductive isolation of the two species.

3. Detecting Recombination in Human Data

Recombination rates can be detected directly, by cytological approaches and by sperm typing methods, or indirectly, by building genetic maps using genetic data from pedigrees or unrelated individuals from a population (Figure 2). The recombination rate, r, is usually measured in terms of the expected number of recombination events per generation between two loci, expressed in centimorgan (cM) units, which is defined as a 1% chance that two loci will be separated by a recombination event in one meiosis. Genetic maps define the linear order of markers along a chromosome, with distances between pairs of markers measured in cM. Commonly used genetic markers in human genetics include microsatellite markers and single-nucleotide polymorphisms (SNPs). SNPs became the markers of choice for constructing genetic maps, their greater density – more than four per Kb (The 1000 Genomes Project Consortium 2010), compared to the highly polymorphic microsatellite markers – around one every 15Kb (Subramanian et al. 2003), compensate for the smaller amount of information per SNP by creating local haplotypes that have greater linkage information content.

Direct approaches : Cytogenetics and Sperm Typing

The direct approches to study recombination examine the direct result of meiosis, at the individual level (Figure 2d). Cytological approaches aim at analysing cells at specific meiotic time points to examine chiasma, the physical manifestation of recombination, and crossover-associated proteins (Figure 2e). These approaches allowed for detailed comparisons of crossover frequency among individuals in a chromosome-specific manner, leading to primary observations that each chromosome has at least one crossover, that females have more crossovers than males and that the presence of one crossover inhibits the formation of a second one

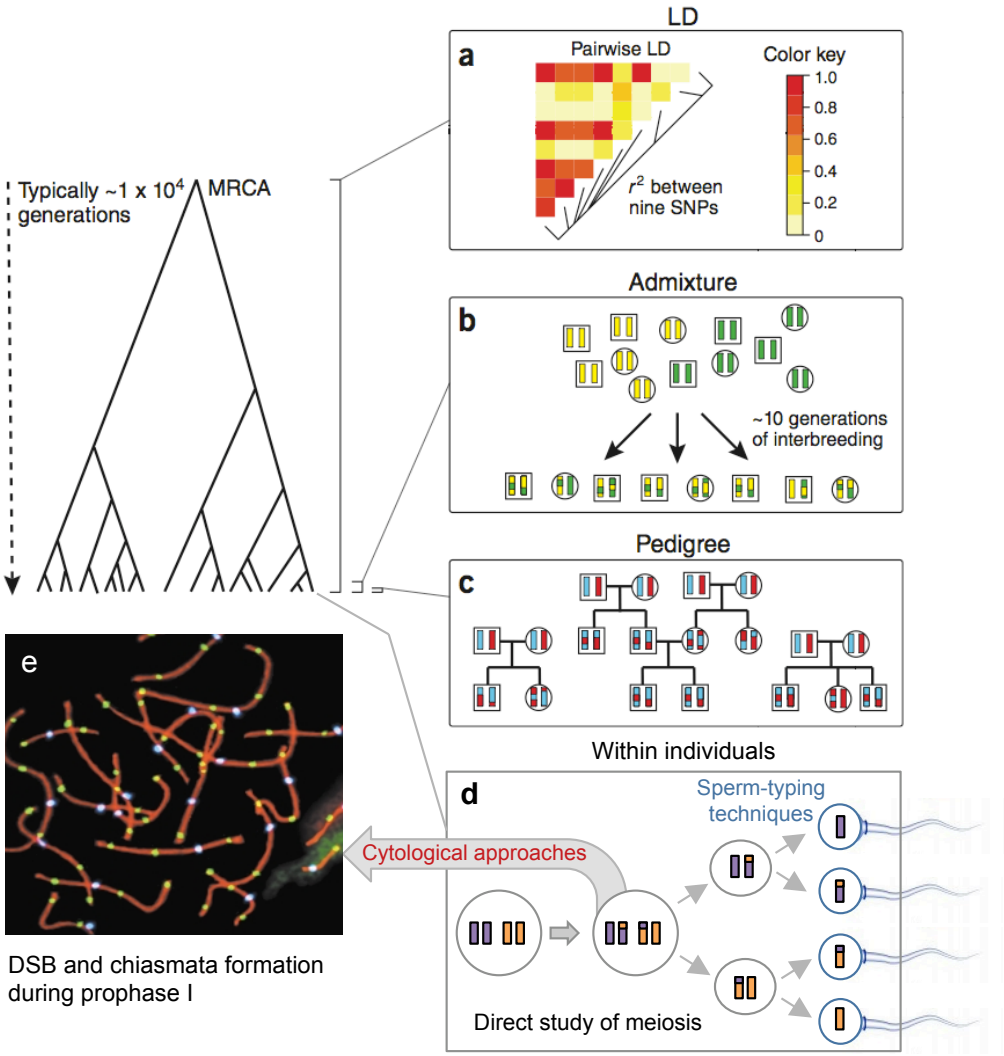


Figure 2. Methods to infer recombination events occuring within various time frame captured.

(a) Linkage disequilibrium (LD) based methods calculate historical recombination rates from events inferred since the most recent common ancestor (MRCA) of a sample at each locus. (b) Admixture-based method infer recombination events that occur since the admixture of the green and yellow ancestral populations. (c) Pedigree-based recombination maps capture sites of recombination in parent-child transmissions (d) Direct approaches such as sperm typing and (e) cytological techniques locate recombination breakpoints for several meioses within individuals, completely free of the effect of selection. [Modified from (O'Reilly and Balding 2011); Cytogenetics image from (McDougall et al. 2005)]

in the same region (Lynn et al. 2004). However, acquiring material is a major obstacle since these techniques require fetal ovarian tissue or testicular biopsies. Furthermore, preparations are sometimes difficult to analyse and these methods cannot be used for high-resolution analyses.

For mapping crossovers at higher resolution, geneticists have developed sperm genotyping assays (Figure 2d), that rely on PCR amplification of DNA from single-sperm and pooled-sperm. Using this approach, it is possible to observe the outcome of thousands of meioses from a single individual at resolutions of less than 0.5 Kb (Kauppi et al. 2004). These techniques presented the first direct evidence that the human genome harbor recombination hotspots (Hubert et al. 1994). Sperm typing has been used to characterize more than thirty human hotspots so far, including those from the major histocompatibility complex region (Jeffreys et al. 2001). This approach captures all recombination events occuring in the germline of an individual, which may differ slightly from recombination events segregating in the population. A serious drawback of this approach is that it is limited to studying male recombination, as typing a population of mature oocytes from women is impractical. Finally, and perhaps more importantly, it is unable to create large scale recombination maps, as it is technically challenging to study genomic regions larger than 300 Kb at high resolution.

Building Genetic Maps

There are two main types of approaches to construct genome-scale genetic maps from polymorphism data : methods that use genetic information from pedigree data and, more recently, methods using SNP data from a population sample (Figure 2a-c). These maps became important technology to understand the molecular basis of human disease. Indeed, genetic maps are at the basis of fine genetic mapping, a strinkingly successful approach to identify new genes linked to disease. They are also important ressources for imputation of missing genotypes (Marchini et al. 2007).

The construction of genetic maps using pedigree data in humans (Figure 2c) predates the sequencing of the human genome (Murray et al. 1994; Dib et al. 1996; Broman et al. 1998). It was enabled by the characterisation of large reference families (e.g. CEPH extended pedigrees) and the development of computational methods for multilocus linkage analysis (Lander and Green 1987; Kruglyak et al. 1996). Analyses of genetic information from the two parents and at least two siblings enable the detection of recombination events, mapped at high resolution in the offspring using informative parental markers and sibling information. Alternatively, if the parental phase is known or inferred with high accuracy, recombination events can be located by considering parent-offspring pairs. The highest resolution map in humans today is the deCODE map (Kong et al. 2010), contructed based on 15,257 parent–offspring pairs, which revealed many local differences between individuals and genders. The main advantage of pedigree-based approaches is that inferred recombination events can be assigned to specific individuals. The main drawback of pedigree-based maps is the cost in collecting and genotyping thousands of individuals, as it requires a very large number of meioses to be sampled in order to ascertain events at a high resolution. Furthermore, they give information on meiotic recombination events occuring in recent generations, but do not help answering the question of how much evolutionary recombination has occurred in different genomic regions over time, a key question with major implications for disease-mapping studies.

To understand variation in historical recombination rates, methods were developed to indirectly infer population recombination rates in a population, by sampling a much larger number of meiotic events. They are based on the observation that adjacent SNPs tend to form clusters (or blocks), showing high levels of linkage disequilibrium (LD). LD is the preferential association of allelic combinations on chromosomal segments and can be measured by statistics such as D’ and r^2 (Lewontin 1964; Hill and Robertson 1968). These associations are broken down by recombination, with greater recombination resulting in more rapid decay of LD. The

relationship between the historical recombination rates and patterns of LD has been widely studied in population genetics (Weir 1979; Hudson and Kaplan 1985; Myers and Griffiths 2003; Song and Hein 2005). The parameter estimated is the population recombination rate $\rho = 4N_e r$ (Griffiths and Marjoram 1996), where N_e is the effective population size – a measure of the number of independent breeding individuals assuming a panmictic population, estimated to be in the range of 10,000 in humans, and r is the rate of recombination per meiosis between two loci. The coalescent (Kingman 1982; Wakeley 2009) provides a statistical description of the genealogical history of sequences sampled in an ideal population with precise equations linking decay of LD, historical recombination rates and effective population size. To infer a local estimate of ρ, the idea is to combine the SNPs in a region to calculate a likelihood of pairwise inferences of recombination between markers. This may be achieved using composite-likelihood methods (Hudson 2001; Fearnhead and Donnelly 2002; McVean et al. 2002) and true-likelihood methods (Kuhner et al. 2000; Fearnhead and Donnelly 2001; Li and Stephens 2003). Because ρ is a compound parameter confounded by N_e, the resulting LD-based map may vary due to the many factors that can affect variation in N_e across populations or along the genome. The model generally assumes no changes in population size between generations and homogeneous rates of recombination through time and between individuals. Furthermore, the effects of natural selection and migration, which will distort the inference of local recombination rates, are not modeled. Therefore, these methods cannot be applied rigourously to recently admixed populations.

More recently, novel methods have been developed to compute genetic maps based on ancestry switch points in recently admixed individuals (Hinch et al. 2011; Wegmann et al. 2011). These approaches take advantage of the fact that admixed individuals carry chromosomes that are mosaics of segments from two or more diverged populations. They use an explicit population genetic model to perform local ancestry inference based on fine-scale variation data (Price et al. 2009) using a hidden Markov model with ancestry as the hidden state. Switch points between

ancestral segments are seen at recombination sites that occurred since the admixture event. The recombination rates can then be infered via sophisticated statistical methods such as empirical Bayes (Wegmann et al. 2011) or Markov chain Monte Carlo (Hinch et al. 2011) approaches. These approaches have so far been applied to African-American and African-Caribbean individuals. Contrary to LD-based maps, admixture-based maps will reflect contemporary recombination rates and are unlikely to be sensitive to selection. However, in order to be built, admixture-based maps require the ancestral populations to be sufficiently diverged from each other and well characterized as modern populations. Therefore, although admixed populations exist in different parts of the world, it is unclear to which extent this approach will be useful in most human populations.

Recombination Hotspots

Thanks to these approaches, the existence of recombination hotspots in humans is now widely acknowledged. LD-based approaches (McVean et al. 2004; Myers et al. 2005) and more recently, pedigree-based (Kong et al. 2010) and admixture-based methods (Hinch et al. 2011) allowed for large-scale inference of recombination hotspots throughout the human genome. The direct scoring of recombinant gametes helped confirm the presence of hotspots at the individual level (Jeffreys et al. 2005). However, the correspondance between sperm-typing, pedigree-based, and LD-based hotspots is not perfect. A hotspot may be absent from LD-based maps due to the action of natural selection or drift on the patterns of LD. Alternatively, it may have evolved recently, and not have had sufficient time to leave a mark on haplotype diversity (Jeffreys et al. 2005). Conversly, a hotspot that has been inactivated recently may still show up in LD-based analyses but will not be observed as a pedigree-based or sperm-typing hotspot. Many hotspots must therefore be transient features of the genome, suggesting that hotspots may frequently be polymorphic in the population (Coop and Myers 2007). This prediction was confirmed by the recent discovery of the mechanism of initiation of recombination

in hotspots, consisting of the binding of the histone methyltransferase PR domain containing 9 (PRDM9) to specific DNA motifs (Baudat et al. 2010; Myers et al. 2010; Parvanov et al. 2010). Following the analyses of approximately 25,000 hotspots found using the HapMap LD-based map, a 13-pb degenerate motif, CCNCCNTNNCCNC, was identified, associated with the activity of 40% of these hotspots (Myers et al. 2008). The threefold periodicity in the motif was found to reflect a motif bound by 3-bp binding unit of individuals fingers in zinc-finger (ZnF) proteins, which led to the discovery of PRDM9 (Figure 3).

4. Patterns of Variation in Human Recombination

Despite the important differences between the many existing methods to detect recombination, the results obtained from them have been concordant and complementary. The number of crossovers per meiosis and the relative locations of hotspots along chromosomes concord remarkably well among methodologies. These approaches have helped reveal extensive natural variation in human recombination. In this section, I review how recombination rates vary between chromosomes, genders, populations and individuals.

Chromosomal effects

Studies of recombination using cytological approaches suggest that variation in recombination along chromosomes likely depends on factors specific to individual chromosomes (Lynn et al. 2004). Chromosome structure and centromere position may be factors influencing recombination rates. For example, although chiasmata are generally uncommon near the centromere, early studies of human sperm observed that large acrocentric chromosomes (chr. 13, 14, and 15) typically have two chiasmata, with one close to centromere (Laurie and Hulten 1985). Their results also suggest that these chromosomes have higher frequencies of crossovers than similar-sized non-acrocentric chromosomes and that their crossover distribution differs from small acrocentric chromosomes (chr. 21 and 22).

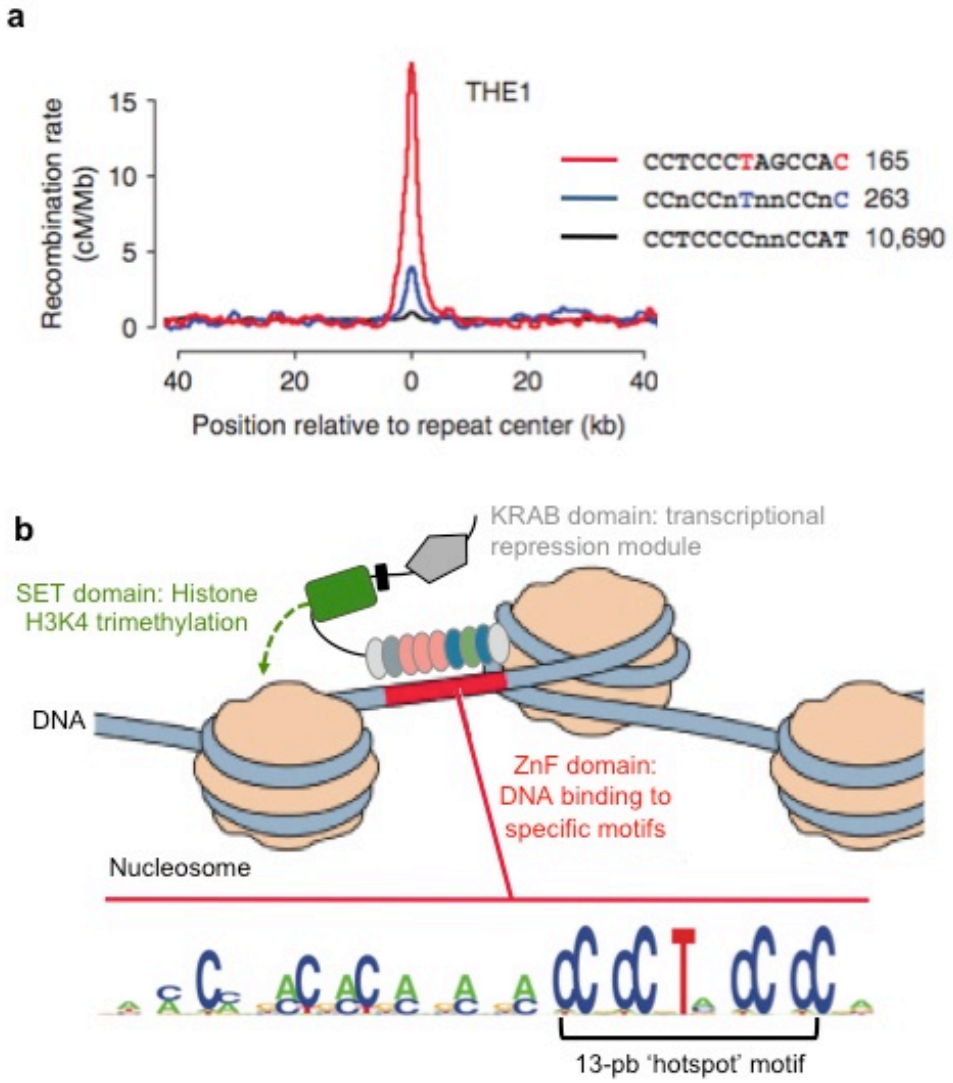


Figure 3. Recombination hotspot motif and PRDM9.

(a) The 13-bp hotspot motif within THE1 repeat family. Average recombination rates around sequences with repeat-specific hotspot motifs (red), degenerate hotspot motifs (blue) and THE1 repeat consensus motifs (black) are plotted. The hotspot motif sequences show increased recombination activity. [Modified from (Myers et al. 2008)] (b) PRDM9 zinc-finger protein binds to the 13-bp hotspot motif (Baudat et al. 2010). PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation (H3K4me3) catalyzed by its SET domain, a process that initiates meiotic recombination in mammals.

More recently, LD-based hotspots inferred from HapMap data (Myers et al. 2005) revealed that chromosome 19 has a much lower density and intensity of hotspots than other chromosomes. This chromosome also has the highest gene density and proportion of open chromatin. These findings suggest that chromosomal size is not the sole factor mediating rates of recombination, and that patterns of human recombination vary among different chromosomes. However, a recent pedigree-based study observed that beyond one crossover per chromosome, additional crossovers occur in rough proportion to physical length, which led the authors to conclude that there are likely few chromosome-specific factors affecting the total recombination rate per chromosome in humans (Fledel-Alon et al. 2011).

Sex-specific effects

Sex-specific differences in recombination rate and distribution have been reported for many organisms, from *Drosophila*, where recombination between homologous chromosomes occurs only in female meiosis (Morgan 1914), to mammals, where several sex-specific genetic maps have been generated. However, no single pattern has emerged, as in some mammalian species males have larger genetic maps than females (Maddox et al. 2001) and in others, such as humans, females recombination rates are higher. Genome-wide, human female rates are approximately 1.6-fold greater than in human males (Kong et al. 2002), with an average of 27 and 42 recombination events per paternal and maternal meiosis, respectively (Coop et al. 2008). One possible explanation is that the chromosomes are less compacted in female meiosis than in male meiosis, but the distribution of the events is also different between genders. Indeed, the ratio between female and male recombination rate varies along the chromosome and is consistently <1.0 near telomeres and >1.0 in centromeric regions (Kong et al. 2002). Male recombination is preferentially localized in telomeric and subtelomeric regions whereas females display a more uniform distribution of crossovers along chromosomes. The basis for differences in telomeric recombination between genders is not known, but may be

due to differential distribution of initiation sites in spermatocytes and oocytes (Lynn et al. 2004). Furthermore, the correlation between GC-content and recombination rates is stronger in males than in females (Duret and Arndt 2008). However this effect is likely a consequence of the sex-specific distribution of recombination events along chromosomes, since GC-content also correlates with distance to telomere (Popa et al. 2012).

Sex-specific differences in recombination rates are also found in localized genomic regions, such as imprinted chromosomal regions. Imprinting is a mechanism of gene expression regulation, whereby certain genes are expressed in a parent-of-origin-specific manner. Imprinted genes tend to have an excess of recombination hotspots yielding recombination rates higher than average (Lercher and Hurst 2003; Sandovici et al. 2006) and several imprinted regions recombine more frequently in one sex than in the other (Paldi et al. 1995). Additionnaly, the latest deCODE map (Kong et al. 2010) revealed that approximately 15% of human hotspots appear to be sex-specific. Although the mechanism of sex-specificity in hotspot formation is currently unknown, it has been hypothesized that differential DNA accessibility during male and female meiosis may lead to differential epigenetic marking affecting both recombination and imprinting control (Paigen and Petkov 2010).

Finally, potential age-related effects on recombination also appear to differ between genders. Direct analyses of spermatocytes indicate no relationship between recombination rate and paternal age in normal meioses. Although these direct approaches are impractical in human female, genome-wide analyses in pedigrees have produced some evidence that the variation in number of recombination events per maternal meiosis correlates with maternal age, but no correlation with paternal age has been found (Kong et al. 2004; Coop et al. 2008). Reduction of recombination associated with maternal age has been observed in other mammals (Henderson and Edwards 1968; Sugawara and Mikamo 1983), but surprisingly, the correlation reported in humans is in the opposite direction.

Inter-individual effects

Extensive variation has been observed at the individual level in recombination patterns in humans. Phenotypic variation has been observed in the total number of recombination per meiosis and in hotspot usage, with genetic factors found to be implicated in such variability. A third level of inter-individual variation, the strength of crossover interference, has been reported to differ among individuals (Broman and Weber 2000; Sun et al. 2006). Crossover interference is a phenomenon that distributes meiotic crossovers such that adjacent crossovers tend to occur further apart than expected by chance. Evidence for inter-individual variation in crossover interference is based on a small number of observations and is therefore tentative, and no genetic factors have so far been associated with variation in interference distances between individuals in humans.

Important heterogeneity in the total number of recombination has been observed in humans. Significant variation in recombination rates was first detected among human females using pedigree-based methods (Broman et al. 1998) and the heritablity of this trait is estimated to be around 30% (Kong et al. 2004). Cytological methods further demonstrated extreme variation in the number of recombination events among oocytes of the same female (Lenzi et al. 2005). Among human males, variation has also been reported (Hassold et al. 2004; Coop et al. 2008) but is less pronounced than in females. In the Icelandic population, the mean recombination rate in females was found to be associated with a 900-kb inversion polymorphism at 17q21.31 (Stefansson et al. 2005) and both male and female recombination rates correlate with polymorphism in RNF212 (Kong et al. 2008). Intriguingly, SNPs at this latter locus influence genome-wide recombination rates of males and females in opposite directions, as the haplotype associated with increased recombination rates in males is associated with decreased rates in females, suggesting either antagonistic effects or distinct causative SNPs between sexes. These associations have been replicated in independent european cohorts (Chowdhury et al. 2009;

Fledel-Alon et al. 2011), and additional putative genetic associations in either males or females have been discovered. It is therefore clear that genetic factors contribute to variation in the mean recombination rate and in the sex-specific regulation of the recombination process.

Most hotspot locations in humans, if not all, are determined by PRDM9. Interestingly, *PRDM9* is highly polymorphic and is considered one of the most rapidly evolving genes in the human genome. *PRDM9* variation is concentrated in its ZnF array and consists of the type and number of fingers, and the order in which they appear in the array (Figure 4). Within humans, *PRDM9* ZnF array presents substantial variation, with one major allele A and more than 30 rare alleles identified worldwide (Baudat et al. 2010; Berg et al. 2010; Parvanov et al. 2010; Borel et al. 2012). It is likely that many more alleles, harbouring different combinations of fingers, remain to be uncovered in human populations. The major allele A binds to the 13-pb motif identified by Myers and colleagues (2008) (Figure 3) whereas other alleles show different DNA-binding characteristics. This translates into individual differences in hotspot usage and population specific hotspot activity (Baudat et al. 2010; Berg et al. 2010; Berg et al. 2011; Hinch et al. 2011). Furthermore, PRDM9 is the only loci found to be associated with hotspot usage genome-wide (Kong et al. 2010; Fledel-Alon et al. 2011; Hinch et al. 2011). Finally, hotspot usage was found to be heritable (Coop et al. 2008), with variation in PRDM9 alone explaining most of the estimated narrow sense heritability in this trait (Fledel-Alon et al. 2011).

There is little evidence that mean recombination rates and hotspot usage are correlated as phenotypes (Kong et al. 2010; Fledel-Alon et al. 2011), which suggest that genetic map length and fine-scale positioning of events are separately determined. However, the three loci associated with variation in recombination phenotypes, namely *PRDM9*, *RNF212* and the inversion at 17q21.31, all show unusual degree of divergence between African and non-African populations. For PRDM9, diversity in Africans is much higher than in non-African populations (Figure

3) and nearly half of the african-specific variants are predicted to result in impaired binding at the 13-bp motif relative to the common A allele. In particular, PRDM9 C-type variants (Figure 4) do not activate hotspots presenting the 13-bp motif (Berg et al. 2011). The analysis of these variants, common in Africans but rare in Europeans, and the construction of a high-resolution African-American genetic map (Hinch et al. 2011) led to the characterization of African-enhanced hotspots, specifically activated by other sequence motifs. Africans therefore use a much broader spectrum of recombination hotspots, which translate into important differences in hotspot usage and fine-scale rates between populations. Furthermore, both RNF212 and the 17q21.31 inversion appear to show unusually high differentiation among populations (Stefansson et al. 2005; Kong et al. 2008) but it remains unclear if divergence in patterns of diversity translate into differences in genome-wide recombination rates between populations. Evidence for population differences in overall map length exist (Jorgenson et al. 2005; Ju et al. 2008; He et al. 2011), however map length estimates may vary due to differences in sample sizes, pedigree structure, and marker heterozygosity as well.

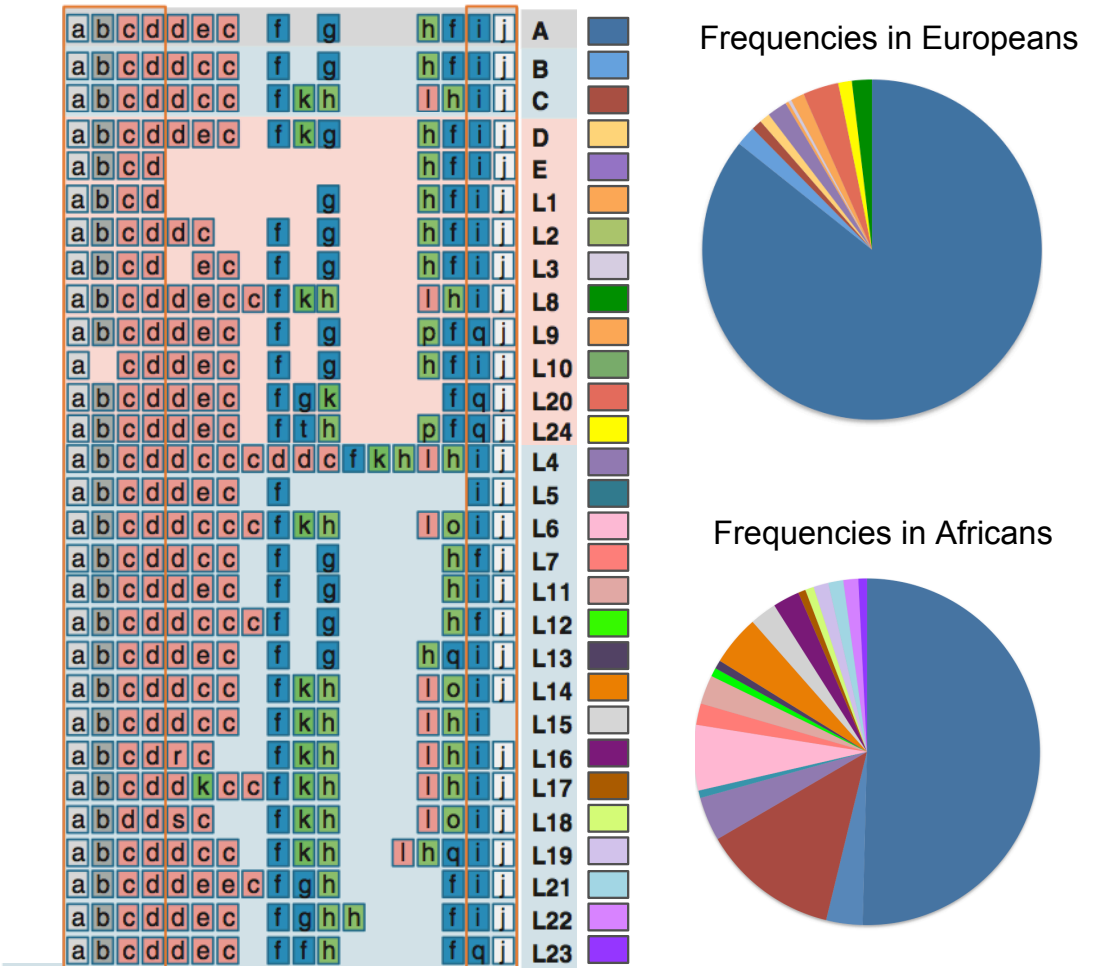


Figure 4. Structure and frequencies of human PRDM9 ZnF alleles.

Each coloured box corresponds to a 84-bp ZnF repeat. The letter within each box correponds to the type of repeat, that differ by a few nucleotides (repeat types differing by only 1 nucleotide are indicated with boxes of the same color). The 5' four repeats and the 3' two repeats are generally conserved between alleles. The pie charts on the right-hand side show allele frequencies in the European and African populations (Berg et al. 2010). Alleles A, B and C are shared between Africans and Europeans, alleles D to L24 have only been found in individuals from European descent, and alleles L4 to L23 were observed in individuals from African descent. Africans have a higher frequency of C-type alleles, containing the *k* and *l* ZnF repeat types [Modified from (Ponting 2011)].

II. MOLECULAR PLAYERS AND GENOMIC DISORDERS

Although one of the main effect of meiotic recombination is the shuffling of alleles, recombination is also essential for accurate chromosomal disjunction and maintenance of genomic stability during meiosis in most eukaryote species. Variation in recombination rates, highligted in the above section, generally have harmless effects in humans, but it may also be a source of harmful mutations with important clinical consequences.

The non-random distribution of crossovers along chromosomes is likely caused by two meiotic molecular phenomena. First, DNA double-stranded breaks (DSBs) are not formed at random and, second, not all DSBs are resolved crossovers, with crossover choice governed by interference (Housworth and Stahl 2003). The successful repair of DSBs in crossovers depends upon many molecular processes and ensures proper segregation of homologous chromosomes. In humans, altered meiotic recombination is the first molecular correlate associated with non-disjunction of chromosomes, leading to aneuploid gametes. Furthermore, DSBs may aberrantly pair with non-homologous loci, leading to structural rearrangements via non-allelic homologous recombination (NAHR). Many of theses rearrangements are highly deleterious and can be detrimental to the survival of the organism. These aberrations may also occur somatically, with somatic acquired genomic instabilities being one hallmark of cancer (Hanahan and Weinberg 2011).

In this section, I review the molecular players involved in the biology of meiotic recombination and its possible resulting outcomes : crossovers, gene conversion and NAHR. Next, the main molecular determinants and processes involved in mitotic recombination are presented. These concepts are reviewed in the context of human diseases, including infertility, aneuploidies, congenital genetic defects, immunodeficiencies and cancers.

1. Meiosis, Recombination and Fertility in Humans

Meiosis is a complex and strongly conserved developmental process that comprises two cell divisions, in order to go from one diploid cell to four haploid cells. It is characterized by an extended prophase, which includes special steps governing the movement and organization of meiotic chromosomes, such as homologous chromosome pairing, synapse and recombination. Most of our understanding of the genetic control of meiosis comes from studies in model organisms such as yeast and mouse (Handel and Schimenti 2010), however regulators of the mammalian meiotic program remain largely unidentified. Furthermore, major distinctions exist between mammalian male and female meiosis. In females, the meiotic process is initiated in fetal ovaries. Before birth, the oocytes undergo meiotic arrest at the end of prophase I. Meiosis is then resumed after puberty for a subset of the oocyte population. In males, meiosis is continuously initiated from spermatagonia stem cells, producing sperm throughout the reproductive lifespan. As pre-meiotic cells enter prophase I, homologous chromosomes condense (leptonema), start synapsis (zygonema), complete synapsis (pachynema), and form chiasmata (diplonema), before segregating to the opposite poles in metaphase I (Zickler and Kleckner 1999) (Figure 5).

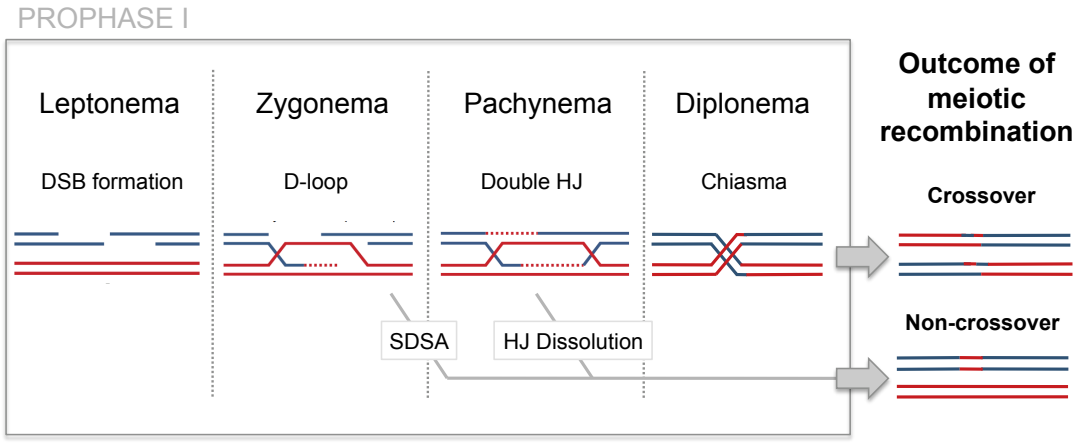


Figure 5. Schematic representation of the meiotic recombination process.

The major steps of crossover formation are illustrated. See details in the text.

Synapsis is the close pairing of homologous chromosomes and is mediated by the synaptonemal complex (SC) that spans the gap between paired chromosomes. Briefly, a chromosomal scaffold begins to form during leptonema through the assembly of REC8 and SMC1B cohesin proteins with SC-specific proteins, such as SYCP3 and SYCP2. DSBs are then generated by the topoisomerase-like enzyme SPO11. Their location is influenced by chromatin structure and by the action of PRDM9, placing histone H3 lysine 4 trimethylation marks (hereafter termed H3K4me3). DSBs are then resected to generate 3' single-strand DNA overhangs that are recognized by homologous recombination repair machinery (Bannister and Schimenti 2004). It includes DMC1 and RAD51 which colocalize to recombination nodules (RNs) to form nucleoprotein filaments and catalyze the search for a complementary sequence on the homologous chromosome to use for repair. This leads to one of the two ends of the DSB invading its homologous chromatid, generating a D-loop intermediate. Maturation of a subset of meiotic recombination sites (<10%) into crossovers occurs during pachynema. These sites are marked by the mismatch repair proteins MLH1 and MLH3, which also colocalize to RNs, to complete recombination. In the final diplonema substage, the homologues are physically held together by the crossovers, a structure called chiasmata, during the end of prophase when cohesins are removed (Page and Hawley 2003). Furthermore, chiasmata are necessary to stabilize the homologs on the metaphase I plate and they promote normal segregation at anaphase I.

Severe defects in chromosome synapsis and recombination will likely result in infertility. In model organisms, it has been shown that defects during prophase I, that prevent pairing and leave breaks unrepaired, trigger abnormal meiotic arrest in the pachytene stage. In mammals, mutations in genes thought to be involved in meiotic control during chromosome synapsis and DNA repair also lead to meiotic arrest and infertility (Roeder and Bailis 2000). Furthermore, successful formation of the SC is required for fertility in mammals, as null alleles of *SMC1*, *REC8*, *SYCP2* and *SYCP3* causes infertility or highly reduced fertility with elevated aneuploidy rates in

mice. In humans, infertility is a relatively common problem and involves numerous pathways from gamete formation to implantation. Although many meiotic genes in infertility mouse models have been identified, infertility-causing mutations have remained largely elusive in humans, with the exception of SPO11 and SYCP3 mutations (Miyamoto et al. 2003; Christensen et al. 2005).

Another important meiotic gene that has been implicated in infertility in human males is PRDM9. As described above (section I.4-5), PRDM9 has recently been identified as a major determinant of the location of recombination hotspots in mammals, although we still have little understanding of the mechanism through which PRDM9 helps to initiate recombination. What is known is that PRDM9 terminal Cys_2His_2-type ZnF array recognizes small DNA motifs and that its histone methyltransferase activity catalyzes the placement of H3K4me3 near these sites (Figure 3) (Grey et al. 2011), which directs SPO11 and additional proteins to initiate DSBs. A recent study in mice demonstrated that PRDM9 is not required for DSBs to occur, but rather determines where these events take place, moving them away from H3K4me3-marked functional elements, such as gene promoters and enhancers (Brick et al. 2012). Prior to discovering its primary function in recombination, this gene had been identified as a 'speciation' and 'fertility' gene (Oliver et al. 2009; Thomas et al. 2009). Indeed, PRDM9 is the first (and only) hybrid sterility gene to be described in vertebrates, as allelic incompatibility can cause hybrid male sterility in mice, consistent with a role in speciation (Mihola et al. 2009). Furthermore, PRDM9 knockout in mice causes sterility in both sexes (Hayashi et al. 2005). In humans, rare dominant nonsynonymous mutations in the ZnF domain of PRDM9 were significantly enriched among infertile men and may thus cause azoospermia (Irie et al. 2009).

The inversion on chromosome 17 found to be associated with high recombination in human females, correlates with increased female fertility (Stefansson et al. 2005). This may be due to the fact that the inversion also contains duplications of *KANSL1*, a gene putatively involved in germ cell differentiation (Boettger et al. 2012). These

duplications produce novel transcripts and have rapidly reached high frequencies in European populations, suggesting they are under positive selection (Boettger et al. 2012; Steinberg et al. 2012). However, individuals carrying the inversion are also predisposed to the 17q21.31 microdeletion syndrome. This newly discovered syndrome is characterized by developmental delay, congenital malformations and intellectual disability (Lupski 2006).

2. The Crossover Pathways, the Holliday Junction and Aneuploidies

Studies in model organisms revealed that crossover and non-crossover pathways are distinct (Allers and Lichten 2001; Guillon et al. 2005). These pathways are specified in late leptonema, although subsequent processing of crossover recombination likely appears in early-mid pachynema (Figure 5) (Strong and Schimenti 2010). Crossover formation is well described by the Szostak model (Szostak et al. 1983). This model predicts that the central intermediate of crossover formation is a four-way DNA junction structure, known as Holliday junction (HJ), that physically connects the two recombining DNA molecules. HJ resolution is catalysed by specialized nucleases, such as newly identified GEN1 and SLX1-SLX4 in humans. They are key proteins, required for HJ resolution and for maintaining genome stability. Mutations of human *SLX4* have been reported in a subtype of Fanconi anemia patients (Kim et al. 2011). Fanconi anemia is an autosomal recessive genetic disease with an incidence of 1 per 350,000 births, characterized by developmental anomalies and breakdown of the hematopoietic system (Moldovan and D'Andrea 2009). Mutations in any of the 15 FANC genes that cooperate to repair DSBs by homologous recombination cause the disease. These genes are involved in DNA repair processes in somatic cells but many of them are, like *SLX4*, involved in meiotic recombination as well (Crismani et al. 2012). For instance, *FANCD2* (also known as *BRCA2*) binds to meiotic chromosomes in SC and is required for normal progression of spermatocytes in mice (Garcia-Higuera et al. 2001).

Proper resolution of the HJ and maintenance of physical connections between chromosomes until anaphase I are primordial for adequate segregation of chromosomes and avoidance of gametic aneuploidy. Abnormal chromosome segregation occurs when there is non-disjunction of homologues or premature separation of sister chromatids. This happens in at least 5% of clinically recognized human pregnancies, making aneuploidy the leading cause of pregnancy loss. Indeed, chromosome segregation in human meiosis is surprisingly error-prone, especially in females, and the reasons for this are still unclear. Altered genetic recombination is the first characterized molecular correlate of non-disjunction : failure to resolve chiasmata between homologous chromosomes in anaphase I results in non-disjoint chromosomes segregating together, while failure to establish chiasmata results in the independent segregation of chromosomes to the same pole.

In all model organisms studied so far, disturbances in crossover pathways are associated with abnormal chromosome segregation in meiosis, with both the number and the location of the exchanges at fault (Hassold and Hunt 2001). In fact, significant reduction in the number of recombination events is a feature of all human trisomies studied so far. This has been demonstrated by comparing the frequency and distribution of crossovers occurring in trisomy-generating meioses with those from normal meioses. Along with reduced map lengths, suboptimally positioned chiasmata have been observed. For example, the presence of distally placed recombination events (i.e. far from the centromere) is likely a risk factor for trisomy 16 and 21 (Hassold et al. 1995; Lamb et al. 1997). Moreover, maternally derived cases of sex chromosome trisomies sometimes involve recombination very close to the centromere (Thomas et al. 2001), suggesting that exchanges occurring too close to the centromere, as well as too far, are risk factors for nondisjunction. However, it is important to note that these recombinational anomalies are not present for all trisomies, with little evidence for a reduction in recombination in trisomy 16 and with abnormal recombination in trisomies 15 and 18 restricted to chromosomes with reduced number of crossovers (Hassold et al. 1995; Bugge et al.

1998; Robinson et al. 1998). Therefore, even if recombination is the main molecular process known to be associated with non-disjunction, it is unlikely that all chromosomes are affected by abnormal recombination in the same way.

Interestingly, although human males have lower recombination rates, most aneuploidies result from female meioses, with 20-25% of human oocytes being aneuploid, against only 2% of spermatocytes (Hassold and Hunt 2001). Furthermore, the rate of aneuploidy increases with age in females. This 'maternal age effect' is particularly pronounced : under the age of 25, a woman has a 2% chance of having a trisomic pregnancy, but over the age of 40, this chance rises to 35%. This effect is thought to be due to age-related insults to the meiotic system at each stages of the oocyte developement (Hassold and Hunt 2009).

3. The Non-Crossover Pathway and Gene Conversion

As stated before, only a small proportion of DSBs (<10%) are resolved in crossovers. The vast majority of DSBs are resolved in non-crossovers, also contributing to homologous chromosome pairing. Non-crossovers are often accompanied by gene conversion events, which involve an unidirectional transfer of short segments of DNA without the exchange of flanking markers. These events appear as a double crossover event within a small interval. They can either result from HJ dissolution or from the synthesis-dependent strand-annealing (SDSA) model, when the invading strand is displaced from the template before the formation of the HJ (Chen et al. 2007).

Although its contribution to haplotype structure in humans remains unclear, gene conversion has been proposed as a recombination-driven process that can considerably influence genomic diversity and evolution via meiotic drive. Meiotic drive is a type of intragenomic conflict, arising when alleles do not have equal probabilities of being present in a gamete (Zimmering et al. 1970). GC-biased gene conversion (gBGC) accelerates the fixation of G and C nucleotides, because

heterozygotes with one AT and one GC allele will produce slightly more gametes containing the GC allele. This is due to a biased incorporation of G or C during repair of mismatches in homologous recombination (Duret and Galtier 2009) and, in yeast, experimental evidence found gBGC to be associated with meiotic non-crossover events (Mancera et al. 2008).

This process is widespread in eukaryotes (Pessia et al. 2012) and is likely responsible for the correlation between GC-content and recombination rates. It can also be responsible for an increase of the rate of substitutions in localized genomic regions (Kostka et al. 2012). gBGC may also result in the biaised transmission of the recombinogenic motifs recognized by PRDM9 (Myers et al. 2010), as molecular genetics experiments showed that there is a bias in favour of transmitting non-recombinogenic alleles at human recombination hotspots (Jeffreys and Neumann 2002). Therefore, bias gene conversion may be partly responsible for the rapid changes in the location of recombination hotspots between species (Coop and Myers 2007), along with the fast molecular evolution of PRDM9 ZnF array (Ponting 2011). Finally, studies suggested that gBGC can contribute to the spreading of mutations within a population and may promote the fixation of weakly deleterious mutations (Galtier et al. 2009). More recently, gBGC has been proposed as a factor responsible for the presence of high frequency disease-causing mutations (Necsulea et al. 2011), making this process highly relevant to human health and disease.

4. Non-Allelic Homologous Recombination

Aberrant gametogenesis leading to recurrent structural genetic abnormalities is a major cause of congenital birth defects. DSBs will sometimes aberrantly pair with non-homologous loci, in a process called non-allelic homologous recombination (NAHR), which results in structural rearrangements that are, in general, deleterious. Indeed, a group of human diseases, termed genomic disorders, arise due to NAHR. Here are some examples : Charcot-Marie Tooth Disease type 1A (CMT1A), neurofibromatosis type 1 (NF1), Williams-Beuren syndrome (WBS), Smith-Magenis

syndrome (SMS), hereditary neuropathy with liability to pressure palsies (HNPP), DiGeorge syndrome (DGS), Prader-Willi syndrome (PWS), childhood spinal muscular atrophy (SMA) and the 17q21.31 microdeletion syndrome. Many of them result from megabase-scale duplications, as in CMT1A (Lupski et al. 1991), or deletions, as in WBS, DGS, PWS and SMS. In most cases, the rearrangements are flanked by low copy repeats (LCRs) that typically share homology greater than 98%. Generally, repeated DNA sequences play an important role in mediating disease-causing recombination errors. Pairing and homologous recombination between misaligned repetitive elements has been observed at rearrangement breakpoints related to disease and is thought to be the main mechanism of NAHR (Purandare and Patel 1997). If repeats are in opposite orientation on the same chromosome, NAHR will result in inversions, while NAHR between repeats present on different chromosomes will lead to chromosomal translocations. These rearrangements are likely to dramatically disrupt genes, possibly creating fusion genes (Figure 6a-b).

NAHR breakpoints are not distributed evenly along the LCRs and cluster in narrow hotspots (Lupski 2004), that are often found at strikingly similar positions to those of hotspots resulting from allelic recombination (Lindsay et al. 2006; Raedt et al. 2006). Furthermore, NAHR hotspots and recombination hotspots share similar properties of distribution of strand exchange (Lindsay et al. 2006), suggesting that these two types of hotspots are functionally related. Many lines of evidence also suggested that *PRDM9* variation correlates with instability in minisatellite repeats and with recurrent pathological rearrangements, such as 17p11.2 deletions/duplication events (Berg et al. 2010) and 7q11.23 microdeletions (Borel et al. 2012). Recurrent duplications or deletions at 17p11.2 are implicated in CMT1A and HNPP, whereas 7q11.23 microdeletions cause WBS. PRDM9 thus appears to be involved in meiotic instabilities leading to genomic disorders.

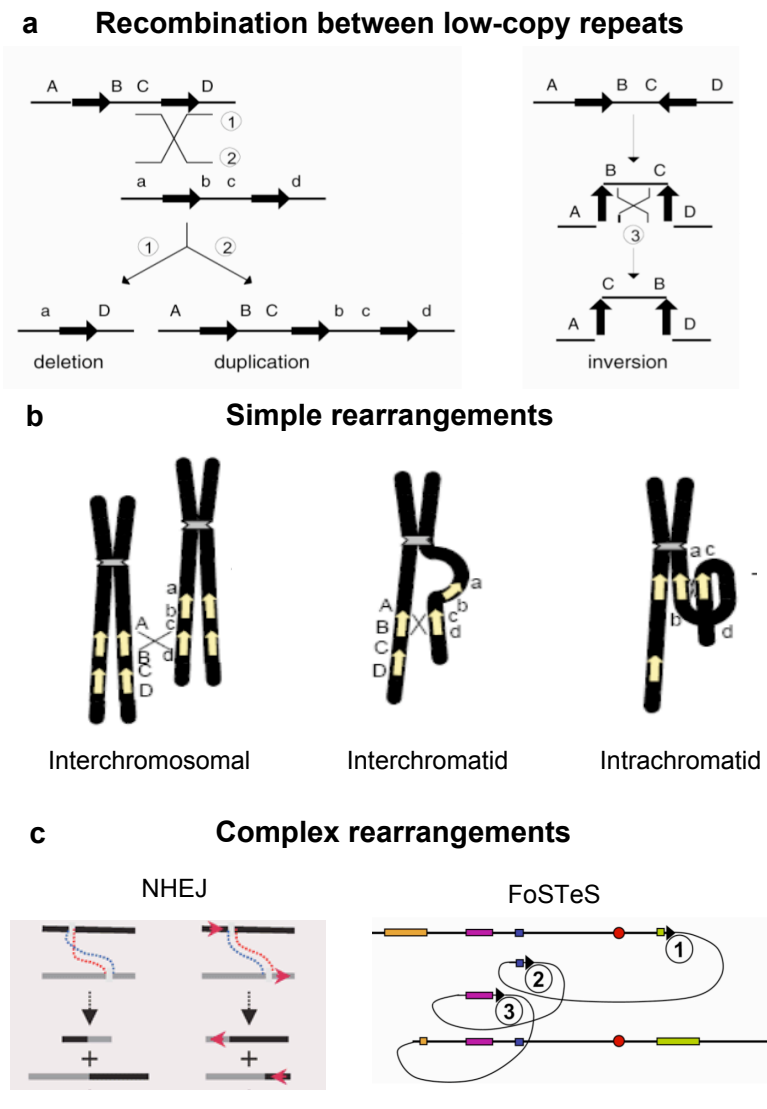


Figure 6. Genomic rearrangements mediated by recombination processes

(a) Genomic rearrangements resulting from recombination low-copy repeats (black arrows). Thin diagonal lines represent recombination events, resulting in deletion (1) and duplication (2) for direct repeats, or inversion for inverted repeats (3). (b) Interchromosomal, interchromatid and intrachromatid non-allelic homologous recombination (NAHR) between LCR pairs results in reciprocal translocations, duplications and deletions. (c) Example of large-scale rearrangements resulting from non-homologous end-joining (NHEJ) and multiple FoSTeS events (see section II.5). [Modified from (Gu et al. 2008; Chen et al. 2010)].

There seems to be a sex-dependent component to some rearrangements, which do not arise at the same frequencies in paternal and maternal meioses. For example, the duplication or deletion at 17p11.2, associated with CMT1A or HNPP, respectively, arise from two distinct sex-dependent mechanisms (Lopes et al. 1997). Most *de novo* rearrangements are from paternal origin and arise by NAHR between the two chromosome 17 homologues. The rare rearrangements that are of maternal origin result from an intra-chromosomal process, such as unequal sister chromatid exchange (Lopes et al. 1998) (Figure 6). Interestingly, this region of chromosome 17 appears to have higher recombination rates in females than in males. This suggests that oogenesis may afford greater protection from misalignment during synapsis, or that male-specific factors may operate during spermatogenesis to help stabilize the rearrangements. Alternatively, these sex-specific differences might reflect different selection bias against the rearranged alleles in male and female germ lines. Differences in NAHR frequency between male and female were also found for other loci, with childhood SMA deletions originating mainly in spermatogenesis (Wirth et al. 1997) whereas 80% of *de novo* NF1 deletions are of maternal origin (Lazaro et al. 1996).

Even when they are not pathological, large-scale rearrangements may have phenotypic impacts. For instance, as in *Drosophila* and many other species, recombination is suppressed in individuals heterozygous for inversions (Roberts 1976; Ishii and Charlesworth 1977). Indeed, a crossover that occurs within a pericentric inversion would produce recombinant chromatids that have duplications and deletions. These recombinant gametes will likely not lead to viable progeny. Furthermore, if the inversion is paracentric (i.e. spamming the centromere), it would produce an acentric recombinant chromatid, that will be lost, and a dicentric fragment that is likely to break and result in chromatids with large deletions. These gametes will likely be nonviable as well. However, it has been proposed that suppressed recombination in large inversions contributes in the formation of new species (Rieseberg 2001).

5. Somatic Recombination, Combined Immunodeficiencies and Cancers

Despite its central role in meiosis, recombination is also a universally important mechanism for the repair of DSBs due to DNA damage, such as replication-fork breakage. During mitosis, DSBs can be induced by specialized endonucleases, by oxidative free radicals or by natural ionizing radiation and need to be properly repaired to maintain genome integrity. There are two main types of somatic recombination identified : mitotic recombination between homologous chromosomes, which is a rare process on a per cell division basis, and V(D)J recombination, which takes place only in the primary lymphoid tissue.

Mechanisms of somatic recombination

Despite the importance of mitotic recombination in DNA repair damage and maintenance of genome stability during cell division, many aspects of this process remain poorly understood. Evidence suggests that spontaneous mitotic recombination happens during interphase and that hotspots for spontaneous mitotic recombination exist (Lee et al. 2009). In general, mitotic DSBs are repaired by homologous recombination and resolved by the formation of HJ or by SDSA, as in meiotic recombination. This is done only during the S and G2 phase of the cell cycle, using preferentially the sister chromatid as a template to repair DSBs. When the homologous chromosomes involved have heterozygotes markers, this process will lead to loss of heterozygosity (LOH) distal to the recombination site, in half of the daugther cells.

NAHR may also occur in mitosis and can generate sub-populations of somatic cells carrying genomic rearrangements that can cause genomic disorders with mosaic manifestations (Dempsey et al. 2007; Steinmann et al. 2007). Repeats involved in meiotic NAHR events can mediate mitotic NAHR (Barbouti et al. 2004; Carvalho and Lupski 2008), although the exact positions of hotspots used may differ (Turner et al. 2008). Interindividual variation in mitotic recombination has been observed (Holt et

al. 1999) and, surprisingly, patterns of mitotic NAHR can differ between females and males, with female having significantly higher mitotic recombination rates (Holt et al. 1999; van der Maarel et al. 1999; Steinmann et al. 2007). As discussed before, the differences between males and females in meiotic recombination patterns may be explained by the fundamental differences in gametes formation, but in mitosis, the reasons underlying this sex-bias are not known and little data on this phenomenon are available at the present time.

Although NAHR was the first major DNA rearrangement mechanism identified, non-recurrent rearrangements are mainly thought to arise by non-homologous end-joining (NHEJ) or Fork Stalling and Template Switching (FoSTeS) mechanisms (Figure 6c). They may occur in germ cells, although they have mainly been observed in somatic cells and can cause errors leading to cancer and immunodeficiencies. NHEJ is used in human cells to repair normal or pathological DSBs at any time during the cell cycle. After the detection of DSBs, NHEJ proceeds by bridging the two DNA ends and ligating them after modification to make the ends compatible (Lieber 2008). This modification includes cleavage and addition of nucleotides at these ends and thus leaves an 'information scar'. NEHJ is therefore an error-prone process. These breakpoints do not need to show any homology with other genomic regions, although they often appear close to DNA breaking elements, such as repetitive (e.g. LINE, Alu) or transposon elements, by which they may be stimulated (Stankiewicz et al. 2003; Shaw and Lupski 2005). The replication-based model FoSTeS (Lee et al. 2007) has been proposed for complex genomic rearrangements (i.e. with many breakpoints, such as inverted duplications interrupted by deleted fragments). It is thought to occur when the DNA replication fork stalls, leading to disengagement of the lagging strand that then anneals to another replication fork in physical proximity and restarts DNA synthesis. This can occur many times in a row, leading to very complex rearrangements (Figure 6c). This process may be initiated by a single-end DSB induced by a microhomology during replication (MMBIR) (Hastings et al. 2009).

V(D)J recombination and Severe Combined Immunodeficiencies

During lymphocyte development, immunoglobulin and T-cell receptor genes undergo somatic DNA rearrangements through a mechanism called V(D)J recombination. V(D)J recombination randomly combines variable (V), diverse (D), and joining (J) gene segments and generates a diverse repertoire of antigene receptors to match antigens from various pathogens and dysfunctional cells. The gene segments are flanked by recombination signal sequences (RSSs), composed of an heptamer (7 bp) and a nonamer (9 bp) element separated by a spacer containing either 12 or 23 bp (Schatz and Swanson 2011). RSSs are bridged by a protein-DNA complex, that includes RAG1 and RAG2, that initiates the rearrangement process by introducing site-specific DSBs, subsequently repaired by NHEJ. Briefly, RAG proteins first bind to a 12RSS or a 23RSS and then capture the second RSS to form a paired complex, within which DNA is cleaved. The cleavage step generates a covalently sealed hairpin at the end of the gene segments. The RAG proteins then cooperates with NHEJ DNA repair factors to rejoin the DNA ends (Schatz and Ji 2011).

RAG proteins are highly conserved among vertebrates. The core domain of RAG1 mediates RSSs contact and interacts with RAG2 through a C_2H_2 ZnF domain. RAG2 contains a core domain that interacts with RAG1 and a C-terminal noncanonical plant homeodomain (PHD) finger that specifically recognizes H3K4me3 marks, necessary for efficient V(D)J recombination (Matthews et al. 2007). RAG proteins are selectively recruited to small regions of highly active chromatin, called recombination centres, where RSSs sequences are made accessible. RAG1 recognize the RSSs in a sequence-specific manner, whereas RAG2 specifically binds to H3K4me3 and increases the affinity and specificity of the DNA-protein interaction (Swanson 2004).

RAG proteins are essential for lymphocyte differentiation, as inactivation of RAG1 or RAG2 arrests both T- and B-lymphocyte development. Furthermore, mutations in humans that inactivate the recombination capacity of RAG1 and RAG2 lead to severe

combined immunodeficiency (SCID), the most serious inherited immunological deficit, manifested by the absence of both T and B cells. Moreover, mutations in either RAG1 or RAG2 that cause partial loss of V(D)J recombination result in Omenn syndrome (Villa et al. 1998). Omenn syndrome is a rare autosomal recessive form of SCID occuring in infants within weeks of birth, making them extremely vulnerable to infectious diseases. Interestingly, Omenn syndrome has been associated with mutations of critical residues within the RAG2 PHD finger that interact with H3K4me3, which considerably alters V(D)J recombination (Gomez et al. 2000). Additionally, dysfunction of many other players of V(D)J recombination are implicated in human SCID. In fact, defects in NHEJ account for an important fraction of SCID (Schwarz et al. 2003). For example, mutations in the *Artemis* gene, involved in the hairpin-opening step, are known to cause radiosensitive SCID (Moshous et al. 2001).

Recombination, Genomic instabilities and Cancer

Genomic instability is one of the hallmarks of cancer and a major driving force of tumorigenesis. It corresponds to the failure of parental cells to accurately replicate their genome and correctly distribute the genomic material among daughter cells. Accumulation of genomic alterations from parental cells to daughter cells will cause deregulation of cell division, imbalance between cell growth and cell death, and ultimately, cancer.

Maintenance of genomic integrity during cell division depends, among other mechanisms, upon high-fidelity of DNA replication and error-free repair of DNA damage that may occur sporadically throughout the cell cycle. Recombination processes are involved in these two mechanisms and error-free repair is ensured by homologous recombination or NHEJ. Abnormalities in homologous recombination may cause genetic defects associated with cancer initiation and progression by different mechanisms. First, because mitotic crossovers by homologous recombination lead to LOH in half of the daugther cells, the inactivation of an allele

of a tumor suppressor gene that already harbor a mutated allele may cause tumorigenesis initiation. Second, detrimental mutations in key players of the DSB repair pathway by homologous recombination may lead to the accumulation of damaged DNA within cells and greatly increase the risk of tumor initiation. For example, *BRCA1* and *BRCA2*, the two major breast cancer susceptibility genes, participate in pathways that facilitate homologous recombination DNA repair (Roy et al. 2012). Third, unresolved HJs during mitosis leads to missegregation of chromosomes and aneuploid somatic cells (Rodrigue et al. 2012). Aneuploidy is a common abnormality in cancer and often correlates with poor prognostic (Beerman et al. 1990; Sciallero et al. 1993). How aneuploidy contributes to development and progression of tumors remain however unclear, but recent advances provided a putative mechanistic link between aneuploidy and genomic instability : aneuploidy could enhance mitotic recombination and defective DNA damage repair (Sheltzer et al. 2011).

Several forms of cancer, mainly leukemias and lymphomas, are caused by acquired chromosomal translocations that disrupt gene function and regulation. They may lead to activation of oncogenes, or inactivation of tumor suppresor genes. NHEJ is currently considered to be the major mechanism rejoining translocated chromosomes in cancer (Lieber et al. 2008). When two independent breaks occur simultaneously in a cell, NHEJ will cause the genomic instability if the incorrect DNA ends get rejoined (Figure 6c). Somatic NAHR is also responsible for generating some of the genomic rearrangements present in cancer cells. In fact, several studies suggested a correspondence between evolutionary and cancer-related breakpoints (Kost-Alimova et al. 2003; Murphy et al. 2005). Particularly, segmental duplication breakpoints show signs of instability in human carcinoma cells (Darai-Ramqvist et al. 2008) and osteosarcoma cells (Martin et al. 2010). Tumor-break prone segmental duplications are structurally unstable in evolution and in malignancies, and may have selective value at both scales (Darai-Ramqvist et al. 2008). Lastly, recent analyses of cancer cell genomes have revealed extraordinary complexity, with many

cancer cells harbouring tens to hundreds of clustered genome rearrangements. This has been termed chromothripsis (Stephens et al. 2011) and could arise through mechanisms involved in the restoration of collapsed replication forks (as described in FoSTeS and MMBIR) leading to chromosome shattering within cancer cells.

Childhood diseases caused by defects in recombination processes are often associated with higher risks of cancer. Many of these associations have been described, and here are some representative examples. Bloom syndrome is a rare autosomal recessive disorder characterized by severe genomic instabilities and excessive homologous recombination. Affected individuals have a high risk of developing early onset cancer, including rare tumors of early childhood (German 1997). Another example is Fanconi Anemia, which results from genetic defects in DNA repair proteins, as described above (section II.2). The majority of patients develop cancer, most often acute myelogenous leukemia and squamous cell carcinomas. Finally, constitutional aneuploidies, known to result from abnormal meiotic recombination patterns, are also associated with increased risk of cancer (Ganmore et al. 2009). In most cases, the type of cancer observed reflect the embryonic development abnormalities caused by the type of trisomy.

Genomic alterations in Childhood Leukemia

Chromosomal translocations and aneuploidies are among the most common changes in many neoplasms of the hematopoietic system. In particular, the most frequent childhood cancer, acute lymphoblastic leukemia (ALL), is characterized by malignant immature white blood cells with abnormal karyotypes overproduced in the bone marrow. Ploidy is a highly significant prognostic factor in childhood ALL : a favorable outcome is likely in patients with a hyperdiploid karyotype (with more than 50 chromosomes, particularly those with trisomies for 4, 10, 17, 18) whereas hypodiploidy (fewer than 46 chromosomes) is associated with a poor outcome (Nachman et al. 2007). Non-random chromosomal translocations are frequently observed in childhood leukemias and also correlates with prognosis. The most

common is t(12;21) producing the *ETV6-RUNX1* fusion gene and occurs in 20-25% of cases, but other translocations are recurrently found such as t(1;19) and t(9;22), with *EA2-PBX1* and *BCR-ABL* fusion genes, respectively. Fusions involving the immunoglobulin heavy chain (IgH) and the T-cell receptor (TCR) loci are also observed. Infants (<12 months) generally present a distinct type of ALL, with chromosome rearrangements involving the *MLL* gene, associated with poor prognosis (Greaves and Wiemels 2003). Oncogenic transformation likely result from the action of these fusion proteins, with capabilities beyond those of the original constituent proteins. For a given translocation, the genomic regions within which recombination occurs are generally localized within breakpoint cluster regions (BCRs). Most of these genomic abnormalities arise pre-natally, but are not sufficient for disease progression (Greaves 2003). However, the consistent association of specific alterations with specific leukemic phenotypes supports the hypothesis that these genetic abnormalities are causal events during leukemic transformation.

Different mechanisms have been proposed to explain how the recurrent translocations associated with childhood leukemias arise (Lieber et al. 2010). These include : (1) illegitimate V(D)J recombination; (2) homologous recombination mediated by repetitive sequences, such as Alu elements; (3) NHEJ in regions that show increased susceptibility to DSBs. *Alu*-mediated oncogenic rearrangements in leukemic cells involve *BCR* and *ABL*, or rearrangements of *MLL* (Zhang et al. 1995; Strout et al. 1998). Recently, it has become increasingly clear that the occasional errors during V(D)J recombination contribute to the development of leukemias. Indeed, there are locations in the genome that can look similar to an RSS, other than the ones located at antigen receptor loci. These are called pseudo RSSs and are aberrant target sites of RAG complex (Lewis et al. 1997). In many subtypes of T-cell ALL, these pseudo RSSs are cut and rejoined to the TCR locus. In other cases, two DSBs can occur at pseudo RSSs simultaneously, resulting in translocations independent of the antigen receptor loci. Some evidence suggests that the *EA2-PBX1* fusion gene may result from such a process (Wiemels et al. 2002). To function

as targets for V(D)J recombination, pseudo RSSs must however behave like recombination centres, and the functional requirements and site specificity of this process remain to be fully elucidated.

Finally, a series of inherited syndromes associated with genetic instablity predispose individuals to leukemias. For instance, patients with Ataxia Telangiectasia and Nijmegen's breakage syndrome, which are due to mutations in genes that have an important role in recognition of DSBs, are prone to develop chromosomal translocations and lymphoid malignancies (Rotman and Shiloh 1998; Digweed and Sperling 2004). Mosaic Variegated Aneuploidy is a rare condition that presents mosaic aneuploidy, with an increased risk of developping childhood leukemia (Jacquemont et al. 2002). Furthermore, constitutional trisomy 21, also known as Down Syndrome, is associated with markedly increased risk for childhood ALL (Hasle 2001). These associations suggest that there is a link between cancer and genetic disorders, however it remains unknown whether these syndromes directly cause cancer or whether some genetic defects within an individual can lead to the two types of diseases arising through common mechanisms.

Studying the genetic architecture of cancer at the subclonal and single-cell level is now feasible using next generation sequencing (NGS) technologies. NGS approaches (see our review (Casals et al. 2012)) and new methodological developments demonstrate that great genetic diversity occurs in leukemia-initiating cells. Malignant cell populations are genetically variegated and the cell expansion fits a branching multi-clonal evolution model better than a linear process (Anderson et al. 2011; Notta et al. 2011). Therefore, the genetic heterogeneity and complexity of leukemia is only now becoming appreciated through genomic analyses, and "evokes a remarkably Darwinian perspective of the evolution of leukaemia-initiating cells" (Burgess 2011).

RESEARCH QUESTIONS AND THESIS OUTLINE

From the diverse studies described in this chapter, it is clear that recombination is a fundamental biological process that has many important roles in the maintenance of genome integrity, within individuals and from one generation to the next. It is also a key evolutionary force that likely modulates the effect of natural selection across the genome as well as between individuals. Major defects in recombination lead to disruption of genome stability, which underlies many human conditions, from infertility to cancer.

Congenital aneuploidies in humans mainly arise from female meiosis and are associated with two phenomena : altered recombination and increasing maternal age. Are the sex-specific differences in recombination patterns implicated in this effect? How does recombination vary with age in humans? In Chapter II, I present a study evaluating age-dependant and sex-specific variation in recombination rates in human pedigrees using high density genotyping data from French Canadian families.

Defects in meiotic recombination are implicated in many inherited syndromes that predispose individuals to cancer. Also, many cancer-related genes are involved in both DNA repair and chromosomal recombination. Are there links between meiotic and mitotic defects in recombination? Does the variation in meiotic recombination, observed between and within individuals, impact susceptibility to cancer? In Chapter III, I present a study of meiotic recombination patterns observed in a cohort of families of patients affected by childhood ALL.

Research exploring the advantages of recombination and its impact on the efficacy of selection have mainly been discussed in theoretical work and demonstrated by simulations. Do patterns of variation in recombination rates along chromosomes influence the removal of new deleterious mutations occuring in human populations? How does mutational load in low recombination regions influences human diseases? In Chapter IV, I present a study that compares the mutational load between regions

of high and low recombination rates using mutations observed in next-generation RNA sequencing data from a cohort of 521 phenotyped French-Canadian individuals.

The main goal in this work is to better understand the costs for human health associated with the recombinational process and the impact of variation in recombination patterns on the susceptility to disease.

CHAPTER II:

Age-dependent recombination rates in human pedigrees

Julie Hussin, Marie-Helene Roy-Gagnon, Roxanne Gendron, Gregor Andelfinger and Philip Awadalla

Reference: Hussin J., Roy-Gagnon M-H., Gendron R., Andelfinger G. and Awadalla P. 2011. Age-dependent recombination rates in human pedigrees. PLoS Genet 7(9): e1002251. doi:10.1371/journal.pgen.1002251

AUTHORS' CONTRIBUTION

In this paper, my contribution is:

- Design of the study with PA;
- Quality control of the pedigree genotyping data;
- Computational package for infering recombination events in pedigrees;
- Statistical analyses;
- Writing of the paper.

Contributions of other authors are: PA M-HR-G and GA contributed reagents, patient materials and samples. M-HR-G provided statistical support. GA recruited the families. RG performed the experimental work and was in charge of the individuals database. PA revised the manuscript.

ACKNOWLEDGMENTS

I am thanksful to the families that participated in this study. I thank J. F. Angers, C. Bherer, F. Casals, P. Donnelly, O. Gandouet, Y. Idaghdour, E. Stone, J. Quinlan, and M. Zilversmit for useful discussions and comments and C. Ober, G. Coop, and M. Przeworski for providing data from the Hutterite cohort and for commenting on the results. I thank three anonymous reviewers and the editors for their valuable comments on the analyses, which helped improve this manuscript. I acknowledge the Genome Quebec Innovation Centre at McGill University and A. Montpetit for genotyping services.

ABSTRACT

In humans, chromosome-number abnormalities have been associated with altered recombination and increased maternal age. Therefore, age-related effects on recombination are of major importance, especially in relation to the mechanisms involved in human trisomies. Here, we examine the relationship between maternal age and recombination rate in humans. We localized crossovers at high resolution by using over 600 thousand markers genotyped in a panel of 69 French-Canadian pedigrees, revealing recombination events in 195 maternal meioses. Overall, we observed the general patterns of variation in fine-scale recombination rates previousely reported in humans. However, we report for the first time a significant decrease in recombination rate with advancing maternal age in humans, likely driven by chromosome-specific effects. The effect appears to be localized in the middle section of chromosomal arms and near subtelomeric regions. We postulate that, for some chromosomes, protection against non-disjunction provided by recombination becomes less efficient with advancing maternal age, which can be partly responsible for the higher rates of aneuploidy in older women. Our model reconciles our findings with reported associations between maternal age and recombination in cases of trisomies.

AUTHOR SUMMARY

Aging is a genetically and environmentally modulated process. One particular manifestation of aging in humans is the age-related changes that affect the female reproductive system. It is well established that chromosome-number abnormalities in offspring occur more frequently as maternal age advances, but the meiotic mechanisms involved remain unclear. Meiotic recombination has been associated with maternal age in different species but contrasting effects of maternal age on recombination rates have been reported among mammals. In this study, we found a decrease of recombination rates with increasing maternal age in a French-Canadian cohort, with the most pronounced decline possibly occurring before 32 years of age. We observed chromosome-specific age effects and, in older women, recombination frequencies are notably reduced in the middle portion of chromosomal arms and near subtelomeric regions. No paternal age effect on recombination was found, highlighting differences in patterns of variation among sexes. Many studies have shown significant inter-individual variation in genome-wide recombination rates, and our results points to an additional, intra-individual, source of variation in recombination rates among transmissions from the same mother.

INTRODUCTION

Meiotic recombination is crucial in both driving the evolution of genomes and ensuring faithful segregation of pairs of homologous chromosomes during gametogenesis. The initiation of genetic recombination during the first meiotic prophase enables homologous chromosomes to orient properly on the spindle and helps form physical connections between chromosomes (Smith and Nicolas 1998). This process results in strand crossovers and further leads to zygotes harboring new combinations of parental genetic material. Every descendant is therefore provided with a unique mosaic of both pairs of parental chromosomes.

In most mammals, including humans, there are important sex-differences in recombination rates and patterns (Coop and Przeworski 2007). First, the distribution of crossovers along the genome differs between sexes, tending to be lower at the telomeres in females relative to males (Broman et al. 1998). Second, the average size of the genetic map for females is 1.6 times longer than that for males (Donis-Keller et al. 1987; Kong et al. 2002). Third, 15% of female and male hotspots are sex-specific (Kong et al. 2010). Evidence indicates that these differences result from sexual dimorphism in the regulation of the meiotic process (Hassold and Hunt 2001; Cohen et al. 2006; Chowdhury et al. 2009), but high levels of heterogeneity in recombination rate is also observed within the same sex.

Pedigree studies have identified extensive variation in rates among females (Broman et al. 1998; Kong et al. 2002) and more recent studies reported significant variation in both female and male crossover rates (Cheung et al. 2007; Coop et al. 2008; Chowdhury et al. 2009; Kong et al. 2010). In addition to interindividual variation, the number of crossovers among different gametes of an individual has been reported to vary (Hassold et al. 2004; Lenzi et al. 2005). However, variation in gamete recombination does not necessarily translate into variation in offspring recombination, since only a small subset of the gamete variation may be consistent

with live-born offspring. For example, more than 20% of oocytes exhibit an abnormal number of chromosomes, and yet very few aneuploid embryos are viable (Hassold and Hunt 2001). Since reduction or failure of meiotic recombination is associated with improper disjunction of chromosomes, leading to genetically unbalanced gametes, high rates of recombination protect oocytes from non-disjunction events (Roeder 1997; Smith and Nicolas 1998), and oocytes with many crossovers are likely to result in a live embryo. Conversely, oocytes exhibiting too few crossovers are particularly prone to aneuploidy.

The most important factor linked to chromosomal aneuploidy in women is advancing maternal age (Hassold and Hunt 2001). Since very little is known about age-related causes of non-disjunction, it remains important to establish associations between patterns of recombination and maternal age in normal meioses. Although recombination is initiated during fetal development in mammal females, age can still influence recombination. In mice, it has been demonstrated that oocytes do not exit the mitotic phase of oogenesis all at once, but rather in successive waves (Polani and Crolla 1991). Furthermore, oocytes ovulate in the same order in which they entered meiosis (Polani and Crolla 1991). This 'production line' model thus suggests that eggs ovulated late in life are the result of more premeiotic mitotic divisions.

Contradictory observations for relationships between maternal age and recombination rates have been reported in mammals, with studies reporting a weak increase of recombination count estimates in humans (Kong et al. 2004; Coop et al. 2008) whereas decreases in frequency of crossovers in mice and hamsters were reported (Henderson and Edwards 1968; Sugawara and Mikamo 1983). To further investigate the maternal age effect on recombination in humans, we densely genotyped individuals from 68 French-Canadian families in Quebec and localized recombination events at high resolution using a previously described method (Coop et al. 2008). We report, for the first time to our knowledge, a significant genome-wide decrease in recombination rate with advancing maternal age in humans and

we compare our results with observations from similar studies. Chromosome-specific effects likely drive the observed reduction in recombination with age. Our observations are consistent with a proposed model in which protection against non-disjunction through recombination becomes less efficient with advancing maternal age in some chrosomosomes.

RESULTS

Significant Variation in Fine-Scale Recombination Patterns

Maternal recombination rates in meioses can be examined by inferring recombination events in viable offspring using dense genome-wide genotyping of pedigrees. To capture crossovers occurring during parental gametogenesis, a total of 478 individuals from 68 French-Canadian pedigrees were genotyped using the Affymetrix 6.0 1M Chip. Over 650,000 SNPs were retained after stringent quality control, providing information on 195 maternal and paternal meioses. Following the procedure described by Coop and colleagues (Coop et al. 2008), we localized crossovers at high-resolution in 68 nuclear families with at least two children and examined variation in fine-scale recombination patterns among individuals. We observed an average of 41.7 (40.2 - 43.3 95%CI) and 27.7 (26.9 - 28.4 95%CI) recombination events among maternal and paternal transmissions, respectively, in close agreement with published estimates (Kong et al. 2002; Cheung et al. 2007; Coop et al. 2008; Kong et al. 2008; Chowdhury et al. 2009).

We confirmed the presence of significant variation for fine-scale patterns of recombination (Coop et al. 2008; Chowdhury et al. 2009), suggesting that we have sufficient power to detect fine-scale variation patterns among individuals in this cohort. In particular, we observed significant variation in recombination rates among males and females for individual chromosomes (Table S1), including chromosome 19 in males. Although recombination is positively correlated with gene density, chromosome 19 has been previously reported to be an outlier, as this chromosome has the lowest density of recombination hotspots (Myers et al. 2005) but the highest gene density (Lander et al. 2001). It also carries the highest proportion of open chromatin (Gilbert et al. 2004). We also evaluated the overlap between the recombination events inferred in our cohort and known population recombination hotspots inferred from HapMap3 CEPH haplotypes. To do so, we considered a

subset of recombination events inferred to be less than 30 Kb apart. We found that 70% and 68% of maternal and paternal events, respectively, overlapped described recombination hotspots (Myers et al. 2005), whereas less than 35% overlap is expected if recombination events are randomly distributed across genomes. Overall, these results demonstrate that there is substantial heterogeneity in recombination counts among families, sexes and individuals.

Genome-Wide Negative Maternal Age Effect

The number of observed crossovers in children of our cohort is negatively correlated with maternal age at time of birth (β = -0.49 crossovers/year, Pearson r = -0.28, p = 0.0017). This negative maternal age effect was determined using a linear mixed model that account for the effects of the mother on recombination rates. This effect remained significant after including the number of children of a mother as a covariate in the model (β_{age} = -0.44 crossovers/year, p = 0.007). We also used family-adjusted recombination counts and ages to evaluate if the age trend detected exists 'within family' (see Materials and Methods). Maternal age remained negatively correlated to the number of recombination events across transmissions within families (β = -0.42 crossovers/year, Pearson r = -0.25, p = 0.0047), ruling out the possibility that this pattern is due to variation in recombination rates among mothers. To determine the period of reproductive life in which the maternal age effect is strongest, we used a linear spline smoothing while specifying a random effects structure to account for the within-family correlations (Gurrin et al. 2005). The fitted spline regression is displayed in Figure 1 along with the fit of the linear regression. The spline fit suggests that recombination counts decrease for all ages, with the greatest decline found among children born from mothers that are 32 years of age or younger. We note, however, that the spline fit is not a significant improvement relative to the linear fit (p =0.0695).

Most double recombination events called within less than 1Mb were a result of genotyping errors (see Materials and Methods), nevertheless, including these events

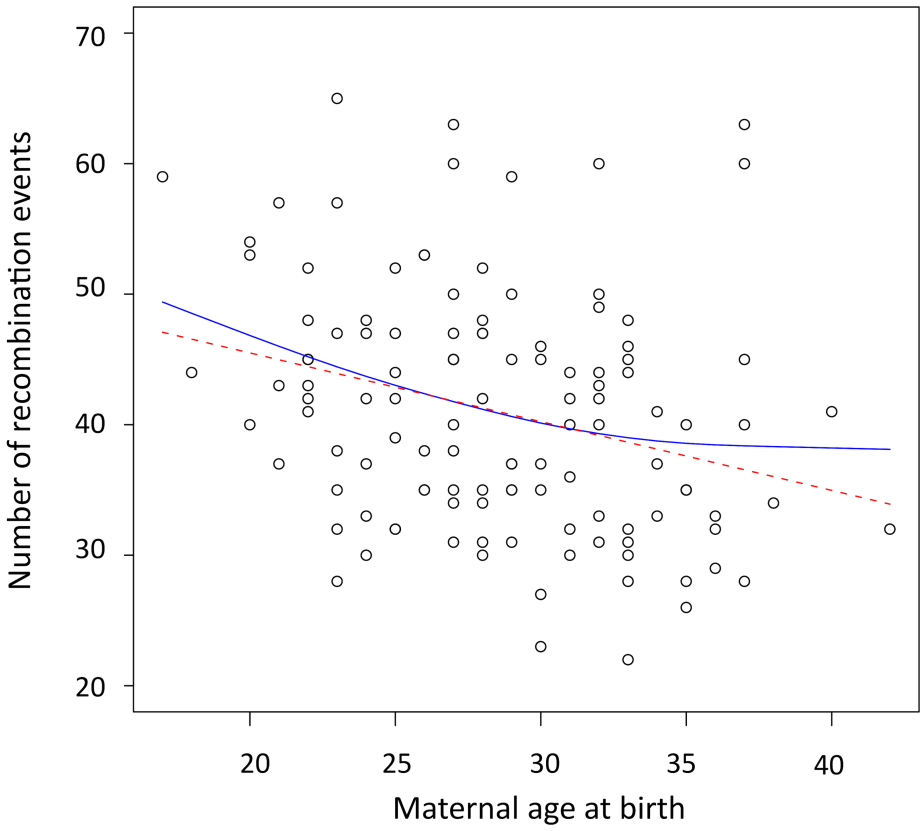


Figure 1. Negative correlation between the maternal age at birth and the number of recombination events

Scatterplot and fitted regression functions showing negative correlation between the maternal age at birth and the number of recombination events in offspring. The red dashed line is the linear regression (β = -0.49, p = 0.0017, r^2 = 0.081). The solid blue line represents the result of the linear spline regression with knots at each distinct value of maternal age at birth and the smoothing parameter λ estimated by restricted maximum likelihood (REML) (λ = 22.29, p = 0.005, r^2 = 0.091).

did not change the direction of the negative trend with age observed in females. Even when called double recombination events occurring within 2, 5, 10 and 20 Mb were excluded from the analyses, the significant negative correlation between recombination counts and maternal age remained (Table S2). All analyses were also performed for males, and no significant correlation was observed between paternal age and the number of paternal crossovers inferred (with family-adjusted values: β = -0.18, Pearson r = -0.15, p = 0.12) as previously reported (Coop et al. 2008).

Evaluating Maternal Age Effects along Chromosomal Arms

We investigated whether the observed genome-wide negative correlation between maternal age and crossover counts is specific to certain chromosomes or genomic regions. For all chromosomes, the mean number of crossovers observed in mothers older than 30 years of age was less than for younger mothers, and significantly so (p<0.05) for chromosomes 5 to 10, 15, 18, and 20 (Figure 2 and Table S3). Putting aside these nine significant chromosomes, the observation that the remainging chromosomes all show reduced mean recombination rates in older mothers (p = $1.22 \cdot 10^{-4}$) is a robust signal for a systematic negative effect. However, the negative correlation is no longer significant for these individually non-significant chromosomes grouped together (with family-adjusted values: β = -0.13, r = -0.14, p = 0.101), suggesting that the genome-wide effect detected is mainly driven by effects present on specific chromosomes. Also, the shift in mean between younger and older mothers seen for the above significant chromosomes is significantly greater than that for the remaining chromosomes (one-tailed p = $1.2 \cdot 10^{-3}$). We further found, based on simulations, that no more than seven chromosomes would be expected to be significant if the genome-wide effect is shared uniformly across all chromosomes (see Materials and Methods). Our result of nine statistically significant chromosomes therefore appears as an outlier (one-tailed p < 0.0002) where submetacentric chromosomes are overrepresented (7 out of 9, one tailed p = 0.043), suggesting that chromosomal arm size or structure are potential determinants.

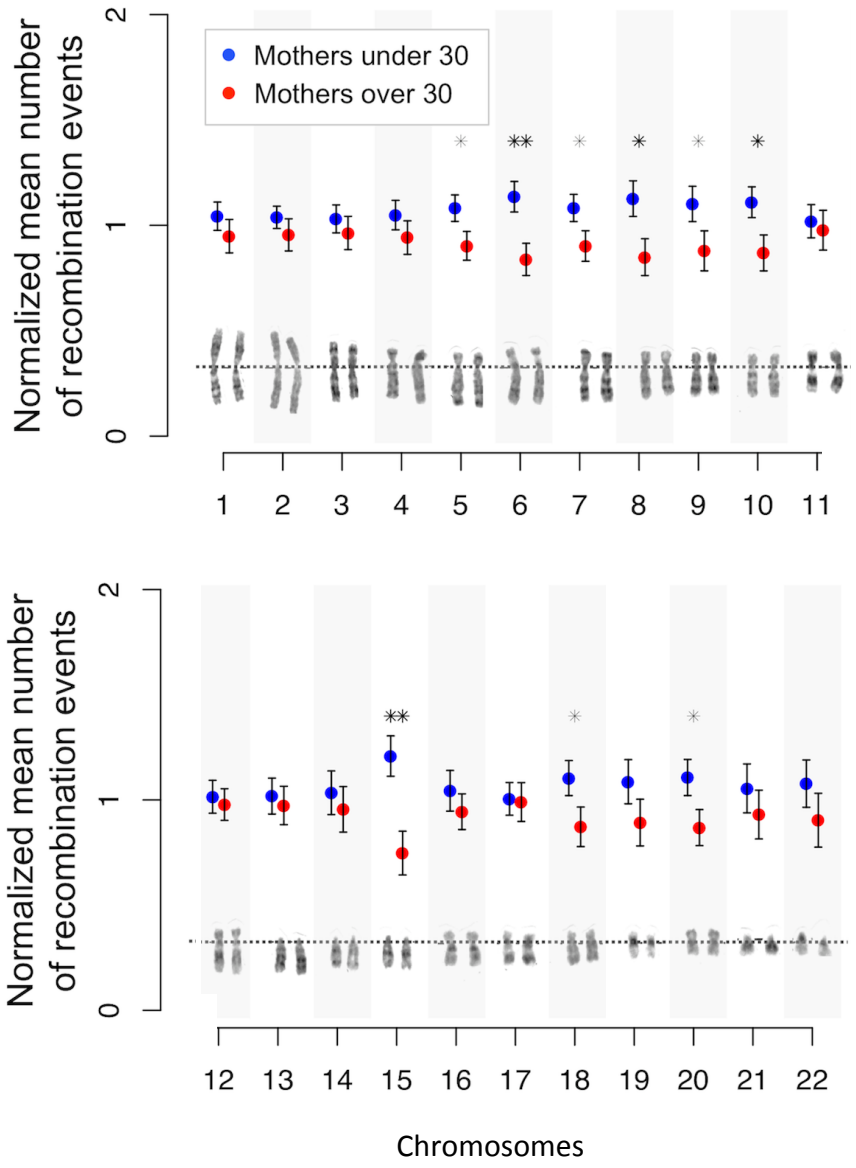


Figure 2. Chromosome-specific effects

Chromosome-specific shifts in normalized means (and standard errors) of the number of maternal crossovers for mothers under and over 30 years of age. Position of centromere is shown for each chromosome (dotted line). Significance of the shift at the 5% (*) and 1% (**) levels is assessed by permutations.

We computed the distribution of maternal and paternal recombination events along chromosomal arms for parents under and over 30 years of age, independently (Figure 3). Male recombination rates increase as we approach telomeric ends of chromosomes, as seen in other studies (Rouyer et al. 1990; Blouin et al. 1995; Broman et al. 1998; Badge et al. 2000; Kong et al. 2002), whereas female rates drop substantially at subtelomeric regions. The difference in recombination counts between mothers from the two age groups is clearly visible and no such pattern is seen in males. The statistical correlation between recombination counts and maternal age was evaluated with respect to relative location on chromosomal arms (Table S3). The decay in crossovers with maternal age appears to be localized in specific portions of chromosomal arms. More precisely, the reduction in recombination rates observed for older mothers in the middle section of chromosomal arms and near the subtelomeric regions is significantly greater than those in the other bins (one-tailed $p = 0.0464$).

Finally, we compared recombination hotspots locations between mothers younger and older than 30 years of age. A large proportion of events (70%) overlapped with previously identified population recombination hotspots in both age groups. Furthermore, no significant differences were found among younger and older mothers in the distribution of hotspots along chromosomal arms (Figure S1).

Phenotypes Show No Association with Maternal Age and Recombination

Among our study cohort, 40 children have left-sided congenital heart disease (LS-CHD), a cardiac malformation where there is substantial evidence for a genetic component (Cripe et al. 2004; McBride et al. 2005; Hinton et al. 2007). We therefore tested for possible associations between the disease phenotype and both maternal age at birth and recombination rates, and found none. Moreover, the negative correlation between family-adjusted crossovers and maternal age remained significant when only unaffected children were considered ($p = 0.0023$). Five mothers had LS-CHD and were involved in 21 transmissions. Again, a significant

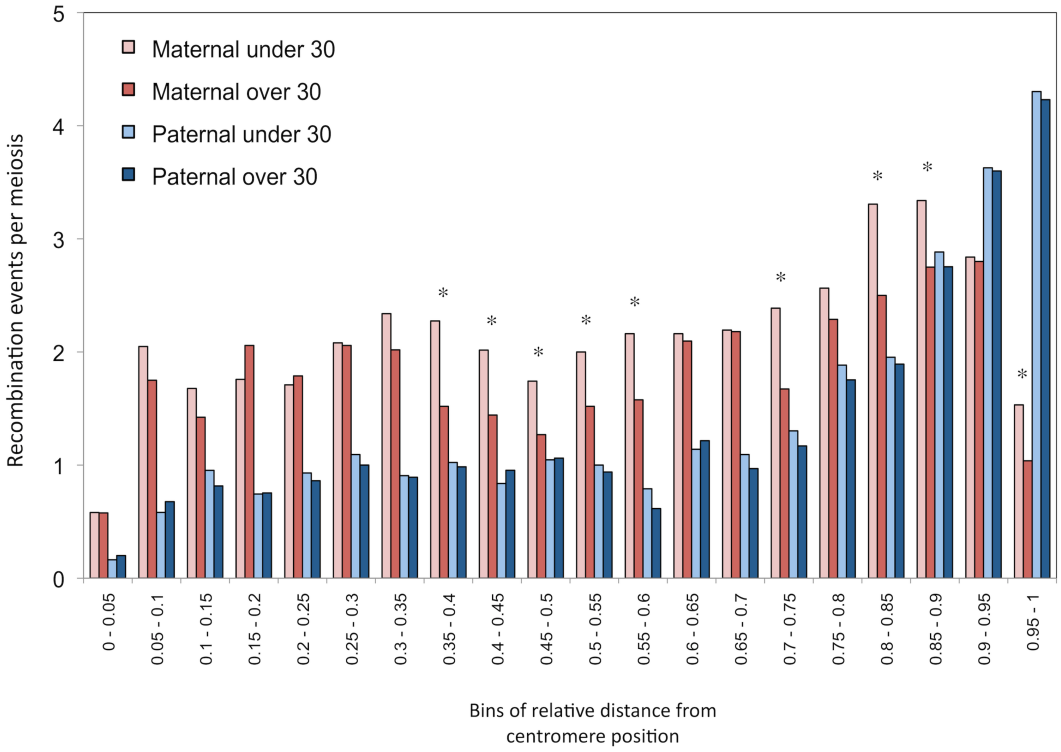


Figure 3. Distribution of recombination events along chromosomal arms.

Histograms of mean number of events per transmission, grouped in 20 bins of relative distances from centromere (increments of 0.05 units). Paternal and maternal events are shown separately and transmissions are partitioned according to the age of the parent at birth. Parents of 30 years old are part of the over-30 groups. Significance of the shift at the 5% level (*) is assessed by permutations. All autosome arms are included.

negative correlation between family-adjusted values is observed when these transmissions were removed from the analysis (p = 0.0059). These results indicate that clinical phenotypes in a subset of our study cohort have little to no effect on our findings.

Comparisons with previous studies in humans

Our main finding that the maternal age effect is negatively correlated with recombination rate is in sharp contrast with a previous finding in an Icelandic cohort (Kong et al. 2004) where a positive correlation between maternal age and recombination rates was observed. There are three main differences in design between the two studies. First, the Icelandic study has a much larger sample size, allowing the detection of what is a very weak positive effect (β = 0.043 recombination events per year) that could not have been detected in our study. Second, approximately 1000 microsatellite markers were used to map recombination events. Third, maternal age at birth was approximated by rounding ages up to the nearest five years. Through simulations, we showed that the discrepancy in results between studies is unlikely to be due to sample size effects (Supporting Text S1). To evaluate to what extent the number of sampled markers and age approximations affect the power to detect an effect among the French-Canadians, we recreated these conditions with our dataset. When we approximate maternal age the same way as in the Icelandic study and used 1000 randomly selected informative markers per mother, the trend remained but the correlation was no longer significant at the 5% level. The mean number of crossovers across transmissions for younger mothers drops from 43.07 to 35.13. For older mothers, the mean drops from 38.04 to 31.62 crossovers per transmission, shifting the difference in means between younger and older mothers from 5.03 to 3.51 crossovers. The use of a relatively large number of markers is particularly important to have enough power to detect changes in recombination counts in specific chromosomal regions, highlighting the need for high marker density. However, our

correlation remained significantly negative when we used only 100 000 SNPs, corresponding to 6000-7000 informative markers in our analysis.

Coop and colleagues also reported a positive effect observed among related Hutterites (Coop et al. 2008) and kindly provided us with recombination rates and parental age at birth in 52 nuclear families. Marker density and the methodology used to infer recombination rates are both similar to those in our study. Using these data, we reproduced their finding and observed a significant positive correlation between recombination counts and maternal age using a linear mixed model (Figure S2a) and family-adjusted values ($\beta = 0.22$, Pearson r = 0.13, $p = 0.034$), however with an explained variance in recombination rate of less than 2%. Moreover, non-parametric tests showed no significant correlation (Spearman $\rho = 0.10$, $p = 0.11$). All results remain unchanged when only recombination events seen once in a family are kept in the analyses.

The distribution of maternal and paternal recombination events along chromosomal arms is very similar to those observed in the French-Canadian cohort (Figure S2b), except that the age effect is barely visible. The positive effect does not seem to be specific to particular chromosomal regions, since the correlations were not significantly different between regions. When examining chromosome-specific age effects among the Hutterites, no significant increase was observed on any chromosome (Table S4). However, two chromosomes showed a significant reduction in the mean number of crossovers for mothers over 30 years of age: chromosomes 20 ($p = 0.0354$) and 22 ($p = 0.0321$) with a one-tailed probability of 0.0455 that at least two chromosomes exhibit such p-values by chance alone (see Materials and Methods).

In order to compare to data in the Icelandic study (Kong et al. 2004), where age data was binned in age categories of five years, we binned the French-Canadian and Hutterite data into similar age category, for each cohort separately (Figure S3). We observed significant differences in recombination rates among categories ($p = 5 \cdot 10^{-4}$)

in the French-Canadians, but not in the Hutterites (p = 0.091). It is worth noting that the average number of crossovers per transmission decreases between mothers aged 25 to 29 and those aged 30 to 34 at time of birth in the Hutterites and in the Icelanders (see Figure 1 in (Kong et al. 2004)), although the differences are not significant.

DISCUSSION

In this study, we examined age-related effects on recombination and observed a negative correlation between the number of maternal crossovers and the mother's age at the time of birth. The proportion of the total variance explained by the genome-wide correlation is significant, yet relatively small (8.1%). This observation is striking considering no strong effect is expected, because considerably reduced levels of recombination are associated with non-viable offspring. The maternal-age effect is pronounced in the middle and distal portions of chromosomal arms. The decrease in recombination might be more pronounced for mothers younger then 32 years of age, after which the rate of maternal non-disjunction is reported to accelerate (Hassold and Hunt 2001).

The possibility that age might influence recombination rates has been examined in several organisms. An age-related decline in recombination has been demonstrated in plants and Drosophila (Griffing and Landridge 1963; Ashburner 1989). In the latter, however, an increase at older age (>16 days) has consistently been reported (Redfield 1966; Ashburner 1989). In mammals, while maternal age has been associated with recombination rate in several studies, paternal age effects on recombination have not been demonstrated. This asymmetry may be explained by important differences in the time of entry, duration and outcome of meiotic processes between sexes (Hassold and Hunt 2001; Cohen et al. 2006). While male germ cells are produced continuously and progress from prophase I to the second meiotic division in several days, the life cycle of oocytes is longer and more complex, beginning during early fetal life (Hunt and Hassold 2002). After a period of mitotic proliferation, oocytes progress through prophase I and initiate genetic recombination, before entering an arrest phase. In humans, meiotic arrest can be maintained for decades, until the oocyte resumes the first meiotic division and proceeds to metaphase II, prior to ovulation. In each of these meiotic stages, errors

affecting chromosome segregation may occur and become more frequent as women age (Hassold and Hunt 2009). Particularly, the physical manifestation of recombination has a critical role in tethering homologous chromosomes together during meiosis (Smith and Nicolas 1998), and a significant reduction in recombination has been identified as a causal mechanism underlying non-disjunction of pairs of chromosomes (Hassold et al. 2004). If the association we observe in this study reflects reduced recombination in oocytes ovulated later in life, one might consider this reduction to be partially responsible for higher level of aneuploidies in older women.

While our results are in agreement with the effects reported in mammals, they directly contradict previous studies demonstrating recombination rates increase with maternal age in humans (Kong et al. 2004; Coop and Przeworski 2007; Coop et al. 2008). All other analyses we performed studying recombinational patterns among families, sexes and individuals corroborate results found in other cohorts. Because of a lack of resolution in recombination estimates and possible misspecification of maternal ages at birth, we are left to wonder whether the Icelandic study had sufficient resolution to properly estimate the maternal age effect. Furthermore, the effect of maternal age on recombination rate reported by Kong and colleagues is very small (Kong et al. 2004) (0.043 crossovers per year). We showed that our effect, which is almost 10-times stronger in the opposite (negative) direction, would not have been detectable in our sample using a marker density and maternal age accuracy similar to that of Kong and colleagues (Kong et al. 2004). It is not possible to assess whether the positive effect is real or spurious, but further analyses of the Icelandic cohort, using recently published data (Kong et al. 2010), may be informative.

The positive correlation in the Hutterite cohort is also very weak, albeit significant (Coop et al. 2008). Our data and methods are very similar and are unlikey to be the cause of the discrepancy. Therefore, only two likely explanations remain : either the

populations are intrinsically different or the studies are capturing different aspects of variation in female recombination rates. It is possible that the trends observed in the different studies reflect that the relationship between maternal age and recombination is a variable phenotype in females, similar to other “low” or “high” recombination phenotypes (Kong et al. 2008; Chowdhury et al. 2009). In humans, genetic determinants could have evolved to counter the “age-dependent reduction” recombination phenotype present in rodents, leading to the absence of maternal age effect, or to a weak increase of recombination rates due to selection against low-recombinant oocytes as age increases (Kong et al. 2004). If these phenotypes coexist in human populations, one may observe increasing, decreasing or even U-shaped trends in any given population. Moreover, the selective pressure acting in human is unlikely to act in rodents, who rarely exhibit age-related meiotic dysfunctions, allowing the negative maternal age effect to be observed consistently.

Because of the association between aneuploidy and recombination rates, women that do not harbor the “age-dependent reduction” recombination phenotype will tend to have more children later in life than the ones subjected to the negative maternal age effect. This would lead to a slight increase in the mean number of recombinations for mothers with more children (Kong et al. 2004; Coop et al. 2008). We note that, in comparison with our cohort, there are a greater proportion of larger families in the Icelandic and the Hutterites cohort. If risk of non-disjunction or other chromosomal anomalies increases with age, then only oocytes with more recombination will survive and be observed in larger families.

The age-effect observed among French-Canadians may also be a consequence, and not a cause, of the higher frequencies of non-disjunction with advancing maternal age. The patterns we observed in viable offspring do not necessarily reflect a decrease of recombination among oocytes in the female ovary, where all eggs might be recombining at the same rate. Rather, the number of crossovers sufficient for proper homologous segregation in young women may not be protective against

non-disjunction in older women. More oocytes would give rise to aneuploid zygotes when ovulated later in life, a model consistent with the observations that increasing age of women increases the likelihood of trisomy. Under this hypothesis, one can show that the mean number of crossovers observed is expected to increase with maternal age in aneuploid conceptions and to decrease in normal, properly disjoined, fertilized eggs (Figure 4 and Supporting Text S1).

Trisomy studies provide evidence that recombination rates may increase with maternal age in aneuploid conceptions. Robinson and colleagues (Robinson et al. 1998) studied non-disjunctions of chromosome 15 and reported that the mean maternal age at birth for cases harboring more than two crossovers is substantially higher than for cases with zero, one or two crossovers. This suggests a positive association between maternal age and recombination rate in aneuploid conceptions involving chromosome 15, and similar associations have been reported for chromosomes 18 and X (Bugge et al. 1998; Thomas et al. 2001). In trisomy 13, 16 and 21, however, no age-effect was reported (Hassold et al. 1995; Lamb et al. 2005; Bugge et al. 2007; Oliver et al. 2008) except for one study in trisomy 21 which found such an effect (Sherman et al. 1994). Interestingly, in the normal conceptions studied here, chromosomes 15 and 18 had a significant decrease in recombination with maternal age, whereas no significant effects were found for chromosomes 13, 16 and 21. Therefore, our model is consistent with significant association found for some chromosomes but not for others (Lamb et al. 1996; Hassold and Hunt 2001). Not all trisomies are affected by increasing maternal age equally and it seems unlikely that the same mechanisms apply to all aneuploid conceptions (Morton et al. 1988).

Altogether, these data highlight the fact that different chromosomes are subjected to distinct selective, mechanistic or structural constraints influencing recombination patterns over successive generations. This points to chromosome-specific effects that might be critical determinants of the complex relationship between maternal

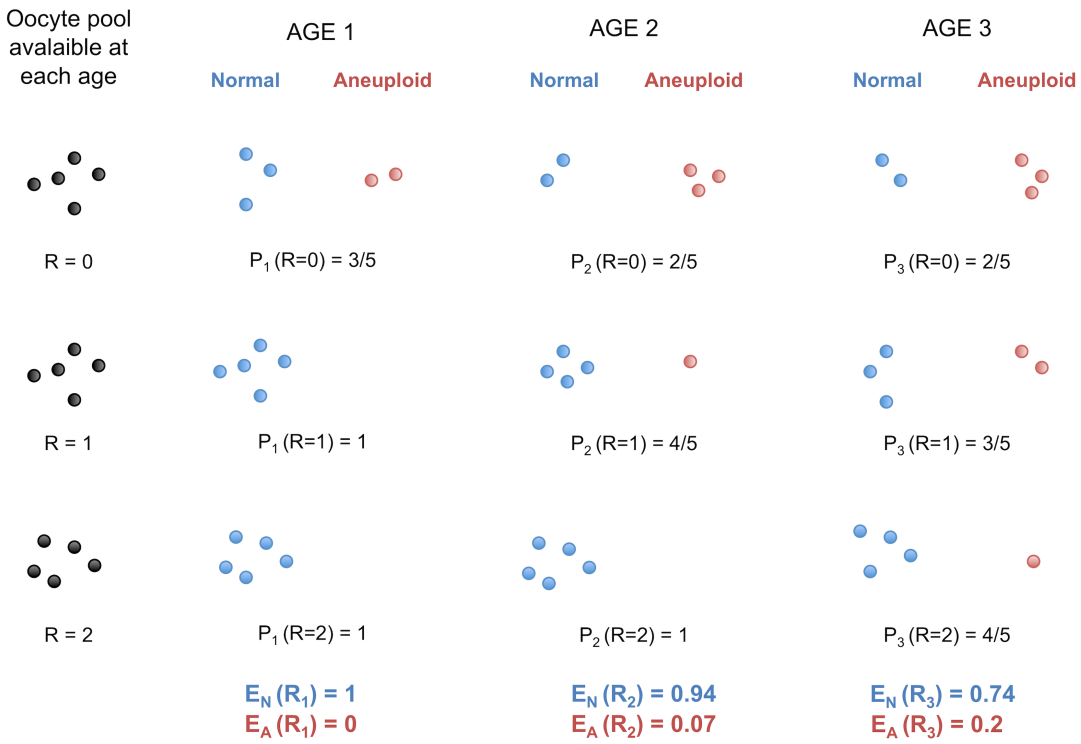


Figure 4. Protection against non-disjunction may be reduced as women ages.

We propose that protection given by high recombination becomes less efficient with increasing maternal age. Here, we depict oocytes containing only one chromosome, with R recombination events. We suppose that, at each of the three arbitrary age periods (k = 1,2,3), the proportion of oocytes having R recombinations stays the same (i.e. it does not decrease or increase with age). During each age period, several oocytes enter their final stage of maturation and give properly disjoined gametes with one chromosome (Normal) or non-disjoined gametes with zero or two chromosomes (Aneuploid). P_k (R = *r*) is the probability of proper disjunction in an oocyte with *r* crossovers for the age period k. E_N and E_A are the mean numbers of recombination in properly disjoined and non-disjoined oocytes, respectively. Under this model, E_N is expected to decrease with k whereas E_A is expected to increase with k (see Text S1).

age and recombination in humans. Chromosome-specific effects may vary among populations depending on genetic differences in factors regulating the recombination machinery. Our results support this hypothesis, as a significant decay was found on nine chromosomes in our French-Canadian cohort and on two chromosomes in the Hutterites. The result for chromosome 20 was significant in both cohorts, but the overlap could be explained by chance alone.

Many factors could reduce the protection provided by recombination from meiotic breakdown, such as factors acting when meiosis resumes after arrest during the final stages of oocyte growth and maturation (Revenkova et al. 2004; Hodges et al. 2005; Kan et al. 2008). Furthermore, factors related to the functional significance of telomeres in meiotic recombination might be implicated. According to the telomere theory of reproductive aging in women (Keefe et al. 2006; de La Roche Saint-Andre 2008) shorter telomeres could be detrimental to segregation of chromosomes, especially for those with recombination event near subtelomeric regions (Lee et al. 1998; Liu et al. 2004). Moreover, the telomere length and rate of erosion might be associated with sex- and chromosome-specific genetic factors (Graakjaer et al. 2006; Mayer et al. 2006) that vary among human populations or cohorts.

In conclusion, high-density genotyping of nuclear families enabled us to capture individual heterogeneity in recombination rates. The results described here are in favor of adaptative theories of sex-specific recombination rates (Coop and Przeworski 2007) suggesting that increased rates in females may have evolved to compensate for improper chiasma formations later in life. The biological causes that underlie recombinational variation and sex-differences have been under investigation (Stefansson et al. 2005; Kong et al. 2008; Chowdhury et al. 2009; Baudat et al. 2010), but the implications of variable rates for population genetic inferences and disease mapping remain unknown.

MATERIAL AND METHODS

Ethics Statement

The ethics committee of Sainte-Justine Hospital Research Center, University of Montreal, approved the study protocol and all participants gave their informed consent. The study was in accordance with the principles of the current version of the Declaration of Helsinki.

Cohort Description and Genomic Data

A French-Canadian cohort was recruited to discover genomic variants contributing to left-sided congenital heart disease (LS-CHD). The cohort is composed of 68 three-generational French Canadian pedigrees, together consisting of more than 700 individuals, including 242 individuals affected with LS-CHD. All participants underwent physical exams, ECG and echocardiography. A total of 478 individuals from 89 overlapping nuclear families were genotyped using the Affymetrix 6.0 platform. Further analysis of this cohort will be presented elsewhere.

We applied standard quality control SNP filters such as call rates (< 95%), departures from Hardy-Weinberg ($p < 0.01$), replicate concordance and Mendelian errors, resulting in a data set of 657,823 autosomal polymorphic SNPs. Genotypic data are available (Dataset S1).

Algorithm to Call Recombination Events

To localize crossover events in autosomes, we only considered the 69 nuclear families in the French–Canadian cohort that had at least two children. We used a previously described heuristic algorithm (Coop et al. 2008) that identifies parental informative markers and phases each child using sibling information. Three modifications to the procedure reported by Coop et al. (Coop et al. 2008) were

made. First, in order to compare recombination rates among families, we evaluated the same SNPs in all families, removing 209,816 SNPs with missing data in at least one family (Dataset S1). Second, to filter out potential remaining genotyping errors, we discarded double recombinants over short intervals. We used a pre-treatment strategy to remove SNPs that result in an observed double recombinant, inferred within 1 Mb (~1 cM, with genomic average of 1 cM/Mb) rather than discarding double recombinants occurring within five informative markers (Coop et al. 2008). The majority of double recombinants removed were found in many individuals at the same positions and are therefore unlikely to be real double-crossover events. Third, the Coop et al. algorithm counts as recombination events the crossovers that are not unique in large families (with four children or more). This means that two offspring can have the same recombination event occurring between the same markers. For smaller families however, only events classified as unique would be captured. This leads to a downward bias in the total number of events detected in small families, relative to larger families (Coop et al. 2008). Thus, in our analyses, we chose to only consider crossovers that are unique in both small and large families. Because this can lead to a downward bias in the number of crossovers for larger families, we partitioned large families into all possible combinations of families of three children (reduced families). For every child, we inferred the recombination counts for the reduced families that include this child, and computed the unbiased recombination counts, averaged over all reduced families. All the results presented in this study remained statistically significant when unbiased recombination counts were used.

To ensure that variation in call rate did not lead to miscalling of recombination events, we examined the correlation between genotype call rates and inferred recombination rates. The number of recombination events observed in a child is uncorrelated with the genotype call rate in this child (Spearman $\rho = 0.098$, $p = 0.29$ for maternal transmissions). The mean number of recombination events per mother

is not correlated with the genotype call rate in the mother (Spearman ρ = -0.035, p = 0.71).

Fine-scale Recombination Patterns Among Individuals

On average, 23,165 informative markers per transmission were used to infer recombination events in our cohort. To verify whether we had sufficient power to detect variation in fine-scale recombination patterns among individuals, we computed the average number of recombination events inferred among maternal and paternal transmissions. Confidence intervals were estimated by bootstrap. Following (Coop et al. 2008), we confirmed the presence of significant variation in the mean number of events genome-wide among females (p = 0.0032) and males (p = 0.0065) using ANOVA. We detected significant variation among individual chromosomes using a linear mixed model that corrects for genome-wide variation in recombination rates (Coop et al. 2008). Significance was determined using a randomization procedure whereby children were randomly reassigned to parents without modifying family sizes. We assessed the congruence of Phase II Hapmap recombination hotspots (Myers et al. 2005) with events localized between informative markers less than 30 kb apart, because the location of these events is considered to be more accurate. The expected proportion of events overlapping a hotspot by chance has been computed as detailed by Coop and colleagues (Coop et al. 2008).

Correlation between recombination and maternal age across transmissions

To study the correlation between recombination in offspring and maternal age at birth, we considered the 34 nuclear families with more than two genotyped children, because with only two children the number of events in each child cannot be determined. Following Kong and colleagues (Kong et al. 2004), we used a linear regression to assess the association between family-adjusted recombination counts and family-adjusted age of mothers at birth and computed the Pearson correlation

coefficient, *r*. The family-adjusted value is the difference between the value for a child and the value averaging over all children from a given mother. The family-adjusted values are used to evaluate the effect of age on recombination across transmissions within families, so that detected effects are not confounded by differences among mothers. To examine whether maternal age is the critical variable, as opposed to time between births, we used non-adjusted values to evaluate the maternal age effect across all transmissions with a linear mixed model that allows for correlated recombination rates by including random effects shared within each family. The number of children was also added as a covariate in the model, to adjust for this potential confounder. The results were confirmed by a non-parametric test: we found a significant Spearman correlation for adjusted counts and ages ($\rho = -0.25$, $p = 0.0078$) and for the non-adjusted values ($\rho = -0.31$, $p = 6 \times 10^{-4}$). To describe the local structure of the relationship between recombination and maternal age, we used a semi-parametric regression model that achieves smoothing using splines and provides a good fit to the data as we move across the range of maternal ages (see Supporting Text S1). The R packages lmeSplines and nlme were used to implement our model. The knots were specified at each distinct value of maternal age at birth ($k = 23$) and the smoothing parameter λ was estimated by REML ($\lambda = 22.29$). *P*-values were determined based on 10 000 randomized data sets, generated by permuting the maternal age across transmissions. For analyses involving family-adjusted values, permutations were performed within families.

Chromosome-Specific Effects

To evaluate chromosome-specific effects, we grouped the transmissions into two categories according to the age of the mother at birth: under 30 years old and 30 years old and above. We tested whether the shift in mean between the two age groups was significant in individual chromosomes (Table S4). Putting aside the significant chromosomes, we used a sign test to evaluate if a systematic effect remained among the non-significant chromosomes, with a standard binomial test

used to assess significance. To determine the number of chromosomes expected to show a significant shift given the genome-wide correlation, we performed simulations to redistribute crossovers of each mother randomly across chromosomes, while taking into account the mean number of recombination occuring on each chromosome (see Supporting Text 1). We also assessed by simulations whether the shift found in significant chromosomes was significantly different from the shift found in other chromosomes using normalized recombination counts (see Supporting Text 1). Normalized values are obtained by dividing the recombination counts by the mean number of recombinations observed on each chromosome across the cohort. *P*-values were obtained using the randomization scheme as described in the previous section.

Distance from centromere

Centromere positions were extracted from the UCSC Table Browser http://genome.ucsc.edu/cgi-bin/hgGateway (assembly Mar. 2006). Genomic positions of recombination events were converted to relative positions with respect to centromere location, i.e. a value of 0 for an event at the centromere and 1.0 for an event at chromosomal edges (telomeric regions). Recombination events were grouped in distance bins of 0.05 (Figure 3 and Figure S2b) and 0.1 (Table S3) and were separated according to parental origin and age group (under or over 30 years old). We evaluated the correlation between distances and the number of recombinations inferred in 0.05-bins. The distances were positively correlated with recombination, resulting in a Pearson $r = 0.86$ ($p < 10^{-4}$) when both paternal and maternal recombinations were considered. The positive correlation remained significant when paternal events (Pearson $r = 0.79$, $p < 10^{-4}$) and maternal events (Pearson $r = 0.58$, $p = 0.0047$) were considered separately, even though the correlation was weaker in females. *P*-values were determined based on 10 000 permutations of the recombination counts within bins. For each separate bin of size 0.1 and 0.05, we tested whether the shift in mean between mothers younger and

older than 30 was significant in individual bins (Figure 3, Figure S2b and Table S3). We assessed by simulations whether the shift found in significant bins was significantly different from the shift found in other bins using normalized recombination counts (see Supporting Text 1). We also evaluated the correlation between maternal age and recombination rates using a linear regression model with family-adjusted values for distance bins of 0.1 (Table S3). Similar effects and distributions of events were observed when chromosomal arms shorter and longer than 85 Mb were considered separately.

Maternal age effect and clinical phenotype

We tested for associations between the LS-CHD phenotype (affected vs. unaffected) and maternal age at birth by an analysis of variance using ANOVA and Kruskal-Wallis rank sum test. The same analyses were performed to test for a relationshp between the clinical phenotype and the number of recombination events found in every child. No significant differences in either recombination rates or maternal age at birth were observed between unaffected and affected individuals.

Factors influencing power to detect the maternal age effect

To evaluate the effect of sampling on the correlation between recombination rates and maternal age, we used resampling methods. We performed boostrap analyses over families within both the French-Canadian and Hutterite datasets. We also used a jackknife approach to generate samples similar to the French-Canadian dataset, using subsets of available and simulated data (Text S1).

Power to detect variation among transmissions can be affected by low SNP density. To evaluate the impact of different SNP density on our results, we used the --thin option of PLINK toolset (Purcell et al. 2007) to keep only a random 80%, 40%, 30%, 20% and 5% of SNPs. Five percent of SNPs corresponds to analysis with an average of 1000 informative markers per mother, which is the marker density used in the Icelandic study (Kong et al. 2004). Four reduced datasets were created per SNP

density. Recombinations were inferred for the 20 reduced datasets and the maternal age effect was evaluated on family-adjusted values.

Using approximations for the ages of individuals can lower the power to detect a correlation between maternal age at birth and recombination. Ages of all individuals (children and parents) were rounded up to the nearest five years and maternal age at birth was calculated by substracting the new child's age from the new mother's age. Since linear relationship between recombination rate and maternal age is no longer consistent with this data, the maternal age effect was evaluated by ANOVA, categorising estimates based on approximate ages.

Analyses of the maternal age effect on recombination found in Hutterites

We were provided access to the list of recombination events inferred in the Hutterite study and parental age at birth for individuals in 52 nuclear families, providing information for 282 female meiosis out of 364 analysed by Coop and colleagues (Coop et al. 2008). We evaluated the genome-wide correlation between family-adjusted recombination counts and maternal age using Pearson and Spearman correlation coefficients. In large families (>3 children), events that are not unique within a family, for example, seen in at least two children, were called by Coop et al. (Coop et al. 2008). All analyses were performed with the unique events only (947 events were removed) and the results remained unchanged.

Effects specific to chromosomal regions and chromosomes were evaluated as previously described. To obtain the probability, by chance alone, of at least two chromosomes showing a significant decay, we assume that the shift has the same probability to be either positive or negative. Using a binomial distribution, we computed the probability of having at least $k = 2$ chromosomes out of $n = 22$ at $p = 0.0355$, $p_{chr} = (1- P(k = 0) - P(k = 1)) \cdot 0.5^2 \sim 0.182 \cdot 0.25 \sim 0.0455$. We also performed simulations where we redistributed, for each transmission, the recombination events uniformly across chromosomes (Text S1) and computed the shift and

significance by chromosome. We find that, among 5000 simulations, only 201 had at least 2 chromosomes exhibiting a negative shift with a *p*-value lower than 0.036 (p_{chr} = 0.0402).

To compare the results observed in the French-Canadian and Hutterite cohorts with those obtained by Kong and colleagues (Kong et al. 2004), we treated maternal age as a categorical variable. For each cohort, transmissions were grouped into 4 categories according to age of the mother at birth: under 25 years old, between 25-29 years old, between 30 to 34 years old, 35 years old and above. We tested differences among categories by ANOVA (French-Canadians: $p = 5 \cdot 10^{-4}$, Hutterites: $p = 0.091$) and using Kruskal-Wallis rank sum test (French-Canadians: $\chi^2 = 10.77$ $p = 0.013$, Hutterites: $\chi^2 = 6.32$ $p = 0.098$). All *p*-values were obtained using the randomization scheme described above.

SUPPLEMENTARY METHODS

1. Spline smoothing with a linear mixed model

We used a linear mixed model to implement a semi-parametric regression model (Gurrin et al. 2005) of the relationship between recombination counts and maternal age at birth for nuclear families.

We seek to fit the following model :

$$\texttt{rec.counts} = m(\texttt{age.at.birth}) + h + \varepsilon$$

where *m* is the smoothing function of ages at birth, with the random effects *h*, with variance σ_h^2, capturing the within family correlation structure and ε representing an uncorrelated random error term, with variance σ_ε^2.

We used the R packages `lmeSplines` and `nlme` to implement our model. The package `lmeSplines` adds smoothing spline modelling capability to mixed linear models. We located a knot at each distinct value of age, although choosing a smaller number of knots did not change the results. The smoothing parameter $\lambda = \sigma_\varepsilon^2 / \sigma_h^2$ is estimated by maximizing the restricted log-likelihood.

This model was used to analyse the correlation between recombination counts and maternal age at birth in the French-Canadian cohort and in the Hutterite cohort (Coop et al. 2008). In the French-Canadians, the spline fit revealed new features of the data (Figure 1) although the improvement in goodness of fit provided by the linear spline model, evaluated by ANOVA, is not significant. In the Hutterites, the spline fit is very close to the linear fit (Figure S2a).

2. Evaluation of chromosome-specific effects

To evaluate if the significant correlations found on 9 out of 22 chromosomes are due to chromosome-specific effects and are not simply due to lack of power to detect

the correlation on all the chromosomes, we used two simulation-based approaches. The first approach was also used with the data from the Hutterite cohort (Coop et al. 2008) to evaluate the probability of having 2 chromosomes out of 22 exhibiting a significant negative correlation.

2.1 Evaluation of the expected number of significant chromosomes

We performed 5000 simulations where maternal recombination events inferred in each child were redistributed uniformly across chromosomes, while taking into account the mean number of recombination events expected to occur on each chromosome. Specifically, for each transmission, the number of crossovers for each chromosome was drawn from a multinomial distribution with probabilities c and sample size n, where n is the total number of recombination events for that transmission and $c = (c_1,c_2,\ldots,c_{22})$ with c_i the proportion of events found on chromosome i across all transmissions. We grouped the transmissions into two categories according to the age of the mother at birth: under 30 years old and 30 years old and above. For each simulation, we tested whether the shift in mean between the two age groups was significant and evaluated the number of chromosomes exhibiting a significant shift.

For the French-Canadian cohort, the maximum number of significant chromosomes found among the 5000 simulations was 7 and only 28/5000 had more than 4 significant chromosomes. For the Hutterite cohort, only 201 simulations among 5000 showed at least 2 chromosomes exhibiting a negative shift and a p-value lower than 0.036 (one tailed $p = 0.0402$).

2.2 Differences between significant and non-significant chromosomes

We separated the 9 significant chromosomes (SC) from the 13 non-significant chromosomes (NSC) found in the French-Canadian cohort. No significant correlation with maternal age was found for NSC ($\beta = -0.13$, $r = -0.14$, $p = 0.101$). We normalized the values by dividing the recombination counts by the average number of

recombinations observed for each chromosome across the cohort, we grouped the SC and NSC together and computed the shifts between older and younger mothers for the two groups separately. The normalized shift estimates are 0.251 for SC and 0.082 for NSC. We then permuted the normalized values among chromosomes, and grouped the nine chromosomes having the largest shifts separately from the remaining ones. The shifts between older and younger mothers for the two groups were then calculated. The difference observed in the shifts between SC and NSC, 0.169, is significantly different from that expected based on simulations (p = 0.0012). Even though we expect significant chromosomes to show a higher average effect than the non-significant ones, such a significant difference between the two groups was not seen in the simulations above (section 2.1) where the effect is distributed uniformly across chromosomes.

3. Evaluation of effects in specific chromosomal regions

Maternal recombination events were divided in bins of relative distance from centromere of 0.1 and were separated according to age group of the mother at time of birth (under or over 30 years old). For each bin, we tested whether the shift in mean between the two groups was significant. The five significant bins are 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.7-0.8 and 0.8-0.9 (Table S3). For each bin, the recombination counts were normalized by dividing each count by the average number of recombination counts observed in the bin across the cohort, and the normalized shifts were computed. The normalized shift estimates are 1.08 for the significant bins and 0.13 for the remaining ones, leading to a difference in shift of 0.95. To assess whether the normalized shift for the significant bins was significantly higher than the shift found in non-significant bins, we permuted the normalized values between bins. We then grouped the five bins showing the largest shifts separately from the remaining ones and computed the difference observed in the shifts between these two groups. The difference in shift observed in the data is significantly higher (p = 0.0464) than what is expected based on these simulations.

4. Evaluation of sample size effects

To evaluate the impact of sample size effects on the results, we used bootstrap and jackknife procedures :

4.1 Bootstrapping over families

We randomly draw families with replacement from the set of 34 French-Canadian families, to create 5000 simulated datasets and we examined the correlation between family-adjusted recombination counts and family-adjusted maternal ages at birth. For 4051/5000 bootstrap repetitions (81%), we obtain a significant negative correlation. Only 2/5000 (0.04%) exhibit a positive β, but the correlations were not significant. We performed the same experiment with the 52 families from the Hutterite cohort. For 3200/5000 bootstrap repetitions (64%), we obtain a significant positive correlation and 107/5000 (2.1%) exhibit a negative β, but none of them were significant. These results show that the confidence intervals for the correlation coefficients observed in the French-Canadian (95% CI -0.84 to -0.18) and Hutterite cohort (95% CI -0.10 to 0.43) do not deviate considerably from the negative and positive point estimates, respectively.

4.2 Jackknifing over Hutterites families to match the French-Canadian cohort

We used the 52 families from the Hutterite cohort to create samples of 34 families with the same number of children as in the French-Canadian families. Only 5/5000 simulations showed a significant negative correlation between family-adjusted values. This analysis suggests that it is unlikely (one-tailed p = 0.001) to observe a negative maternal age effect in a subset of 34 out of 52 families exhibiting a positive maternal age effect of ~0.22 recombinations per year.

4.3 Simulation of the Icelandic cohort and jackknifing.

We used the informations provided by Kong and colleagues (Kong et al. 2004) to simulate a dataset, with the same characteristics:

- the total number of nuclear families with at least 3 children (2177) ;
- the number of sampled children per family (Table 1 in (Kong et al. 2004)) ;
- the distribution of mother's age at birth (Table 2 in (Kong et al. 2004)) ;
- the average number of recombination events per maternal meiosis of 44.6, reported in (Kong et al. 2002), with standard deviation of σ = 8 recombinations ($\sigma_{Hutterites}$ = 7.90 and $\sigma_{French\text{-}Canadians}$ = 8.28) ;
- the estimated maternal age effect of 0.043 recombinations per year.

We then used this simulated dataset to create samples of 34 families with the same number of children as in our French-Canadian families. Only 94/5000 simulations showed a significant negative correlation between family-adjusted values. This analysis suggests that it is unlikely (one-tailed p = 0.0188) to observe a negative maternal age effect in a subset of 34 out of 2177 families exhibiting a positive maternal age effect of ~0.045 recombinations per year.

5. MODEL: Recombination provides a reduced protection against non-disjunction as women age.

We propose that a high recombination rate protects an oocyte ovulated earlier in life from non-disjunction, but that this protection becomes less efficient with increasing maternal age.

Let $P_k(r)$ be the probability of proper disjunction in an oocyte with r crossovers after k years. Under our hypothesis, $P_k(r)$ decreases with k for $r \geq \varepsilon$ (with $\varepsilon \geq 0$). Let R_k be the number of crossovers observed in a properly disjoined oocyte after k years. The mean number of crossovers in these normal oocyte (E_N) is :

$$E_N(R_k) = \sum_{r \geq 0} r \cdot P_k(r) \cdot P(R = r)$$

with $P(R = r)$ the probability of finding *r* recombinations in an oocytes, assuming here that this probability does not depends on *k*. For $x < x' : P_x(r) > P_{x'}(r)$ therefore $E_N(R_x) > E_N(R_{x'})$ (see Figure 4).

Under this model, the mean number of crossovers in properly disjoined oocytes (E_N) is expected to decrease with maternal age, whereas the mean number of recombination in non-disjoined oocytes (E_A) is expected to increase with maternal age.

SUPPLEMENTARY FIGURES AND TABLES

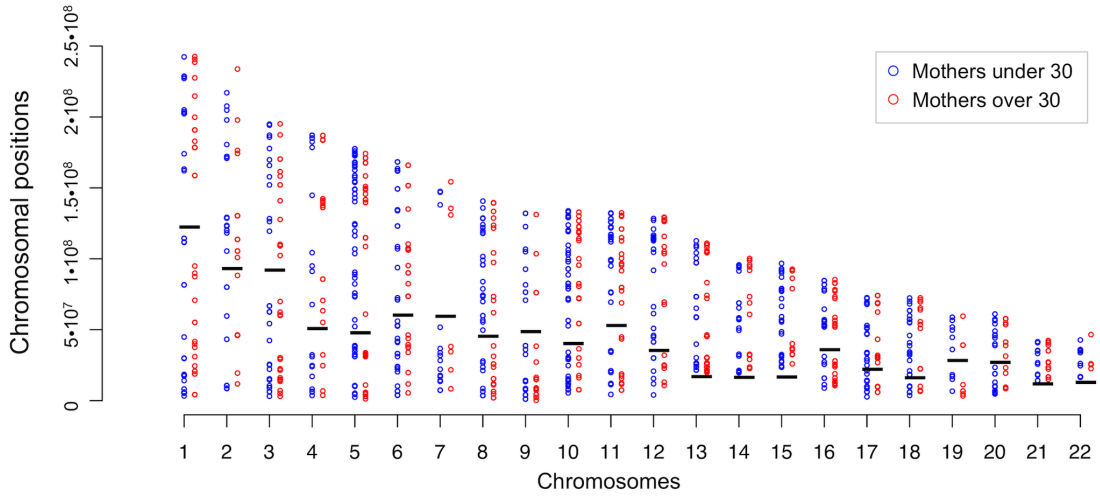


Figure S1. Recombination hotspots and maternal age

Congruence of Phase II Hapmap recombination hotspots with events localized between markers less than 30 KB apart. Positions of active hotspots on each autosome in mothers under 30 years old (blue) and 30 years old and over (red) are plotted. Black lines represent positions of centromeres.

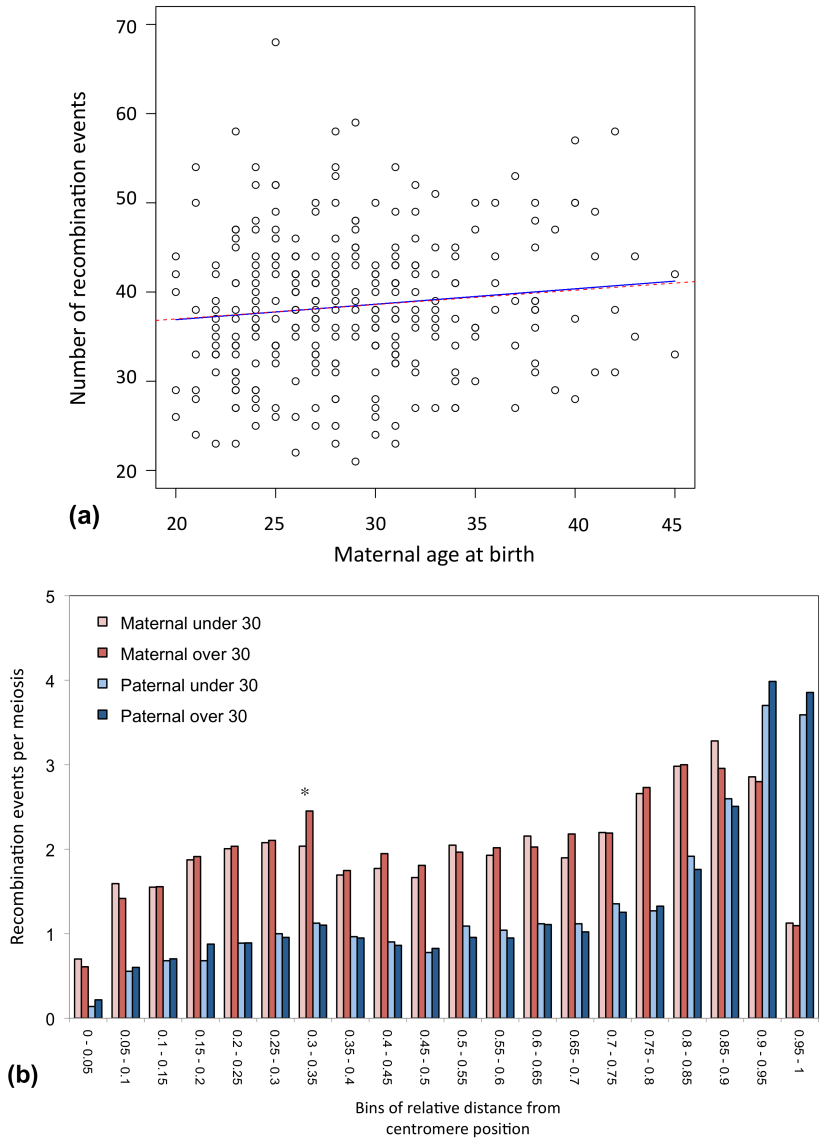


Figure S2. Maternal age effect in the Hutterite study.

(a) Scatterplot and fitted regression functions showing negative correlation between the maternal age at birth and the number of recombination events in offspring. The red dashed line represents the linear regression ($\beta = 0.17$, $p = 0.031$, $r^2 = 0.012$) and the solid blue line represents the result of the linear spline regression with knots at each distinct value of maternal age at birth ($\lambda = 17.05$, $p = 0.0354$, $r^2 = 0.012$). (b) Distribution of recombination events along chromosomal arms (see Figure 3 for detailed description). Significance of the shift at the 5% level (*) is assessed by permutations.

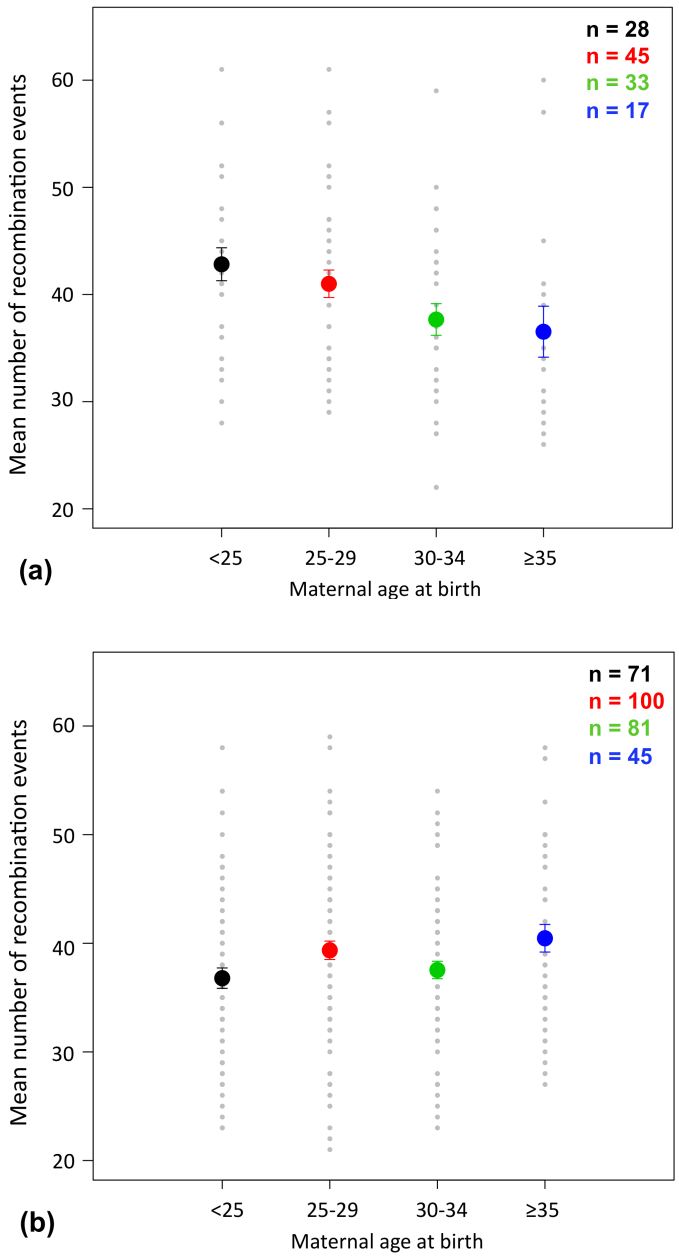


Figure S3. Maternal age effect with age categories

Relationship between maternal age and recombination in autosomes using categorised data for (a) the French-Canadian cohort and (b) the Hutterite cohort. The number of recombinations for all transmissions are plotted (smaller dots), sample means and standard errors for each age group are shown. The numbers of transmissions (n) in each category are reported.

Table S1. Significant variation among autosomes in number of recombination events among male and female transmissions.

The mixed and adjusted models are described in (Coop et al. 2008) and significance was assessed based on permutations (see Materials and Methods) using a likelihood-ratio test. Significant p-values ($p < 0.05$) are reported, otherwise, they are not significant (ns) is indicated. Values for chromosomes that were significant among the Hutterites (see Table S1 in (Coop et al. 2008)) are presented in red.

Chr	Males p-value		Females p-value	
	Mixed model	Adjusted model	Mixed model	Adjusted model
1	ns	ns	ns	ns
2	ns	ns	ns	ns
3	ns	ns	** 0.005648	ns
4	ns	ns	ns	ns
5	* 0.0476	ns	** 0.006067	ns
6	ns	ns	** 0.002188	ns
7	ns	ns	ns	ns
8	ns	ns	** 0.000868	ns
9	ns	ns	ns	ns
10	ns	ns	ns	ns
11	ns	ns	ns	ns
12	* 0.0452	* 0.0400	** 0.000738	ns
13	** 0.00149	** 0.00582	ns	ns
14	ns	ns	ns	ns
15	ns	ns	** 0.00114	ns
16	** 0.00872	ns	** 0.00391	* 0.0262
17	* 0.0189	ns	ns	ns
18	ns	ns	ns	ns
19	** 0.00750	* 0.0294	* 0.0251	ns
20	ns	ns	* 0.04965	ns
21	ns	ns	* 0.02196	* 0.0272
22	ns	ns	ns	ns

Table S2. Exclusion of double recombinants.

Correlation between family-adjusted age of mothers at birth and family-adjusted recombination counts was evaluated without double recombinants occurring between 2, 5, 10 or 20 Mb intervals. Permutations were used to assess significance.

Exclusion interval	β	*P*-value	Pearson *r*
< 2 Mb	-0.42	0.0012	-0.225
< 5 Mb	-0.44	0.0013	-0.214
< 10 Mb	-0.42	0.0018	-0.207
< 20 Mb	-0.42	0.0056	-0.239

Table S3. Correlations between recombination counts and maternal age along chromosomal arms.

Shifts of the mean number of maternal crossovers between mothers under and over 30 years of age are presented at different distances relative to centromere position. Linear correlations are evaluated using family-adjusted values grouped in 10 bins of distance relative to centromere location. Permutations were used to assess significance ($p < 0.05$) and significant results are reported in bold.

Bins		**Shift in recombination counts between mothers under and over 30**			**Linear correlation using adjusted values**		
min	max	Shift	Direction	Shift *p*-value	β	Pearson r	*p*-value
0	0.1	0.30	-	0.325	0.03	0.068	0.512
0.1	0.2	0.04	+	0.911	0.02	0.046	0.652
0.2	0.3	0.06	+	0.871	- 0.04	- 0.093	0.366
0.3	0.4	1.07	-	**0.006**	- 0.12	- 0.269	**0.007**
0.4	0.5	1.05	-	**0.007**	- 0.06	- 0.139	0.18
0.5	0.6	1.06	-	**0.005**	- 0.04	- 0.091	0.375
0.6	0.7	0.01	+	0.984	0.004	0.009	0.927
0.7	0.8	0.99	-	**0.01**	- 0.13	- 0.268	**0.007**
0.8	0.9	1.39	-	**0.003**	- 0.16	- 0.247	**0.014**
0.9	0.0	0.37	-	0.479	- 0.09	- 0.149	0.146

Table S4. Mean number of recombination events among maternal transmissions for each autosome in the French-Canadian and Hutterite studies.

For each study, transmissions are partitioned according to the age of the mother at birth (mothers of 30 years-old are part of the over-30 group). Permutations were used to test whether the shift was significant and significant results are reported in bold.

Chromosomes	**Means in the French-Canadian cohort**			Permutation test *p*-values	**Means in the Hutterite cohort**			Permutation test *p*-values
	Mother under 30	Mother over 30	Sign		Mother under 30	Mother over 30	Sign	
Chr 1	3.19	2.90	-	0.1843	3.29	3.15	-	0.4335
Chr 2	3.18	2.92	-	0.1829	2.88	3.12	+	0.1869
Chr 3	2.47	2.31	-	0.2641	2.46	2.71	+	0.1072
Chr 4	2.42	2.17	-	0.1585	2.39	2.60	+	0.1975
Chr 5	2.58	2.15	-	**0.0302**	2.33	2.30	-	0.8127
Chr 6	2.71	2.00	-	**0.0027**	2.32	2.37	+	0.7512
Chr 7	2.35	1.96	-	**0.0302**	2.09	2.25	+	0.2542
Chr 8	2.32	1.75	-	**0.0131**	2.07	2.07	+	0.9792
Chr 9	2.00	1.60	-	**0.0421**	1.90	1.99	+	0.5435
Chr 10	2.21	1.73	-	**0.0163**	2.26	2.02	-	0.0735
Chr 11	1.81	1.73	-	0.3621	1.81	1.76	-	0.6742
Chr 12	1.94	1.87	-	0.3746	1.92	2.00	+	0.5317
Chr 13	1.55	1.48	-	0.3574	1.45	1.42	-	0.7540
Chr 14	1.29	1.19	-	0.2981	1.34	1.49	+	0.1628
Chr 15	1.65	1.02	-	**0.0005**	1.39	1.36	-	0.7823
Chr 16	1.74	1.58	-	0.2267	1.60	1.54	-	0.5923
Chr 17	1.48	1.46	-	0.4469	1.58	1.43	-	0.2075
Chr 18	1.58	1.25	-	**0.0338**	1.39	1.47	+	0.5086
Chr 19	1.19	0.98	-	0.1007	1.07	1.02	-	0.5686
Chr 20	1.37	1.08	-	**0.0267**	1.37	1.14	-	**0.0354**
Chr 21	0.81	0.71	-	0.2303	0.66	0.59	-	0.4006
Chr 22	0.87	0.73	-	0.1556	0.77	0.60	-	**0.0321**
Autosomes	43.07	38.04	-	**0.0011**	40.10	40.63	+	0.5801

CHAPTER III:

Rare allelic forms of PRDM9 associated with childhood leukemogenesis

Julie Hussin, Daniel Sinnett, Ferran Casals, Youssef Idaghdour, Vanessa Bruat, Virginie Saillour, Jasmine Healy, Jean-Christophe Grenier, Thibault de Malliard, Stephan Busche, Jean-François Spinella, Mathieu Larivière, Greg Gibson, Anna Andersson, Linda Holmfeldt, Jing Ma, Lei Wei, Jinghui Zhang, Gregor Andelfinger, James R. Downing, Charles G. Mullighan, Philip Awadalla

Reference:

Hussin J., Sinnett D, Casals F, Idaghdour Y, Bruat V, Saillour V. et al 2012. Rare allelic forms of PRDM9 associated with childhood leukemogenesis. Genome Research (in press)

AUTHORS' CONTRIBUTION

In this paper, my contribution is:

- Design of the study with PA;
- Sample preparation and amplification for Sanger sequencing;
- Bioinformatics pipelines for SNP calling from exome sequencing data;
- Downstream computational/statistical analyses of genomic data;
- Development of motif search/annotation procedures and analyses;
- Writing of the manuscript.

Contributions of other authors are: DS YI JH GG AA LH GA JRD CGM and PA contributed reagents, patient materials and samples. FC JFS ML SB performed experimental work. VB VS JCG and TdM performed bioinformatics analyses of the Sainte-Justine cohort data and JM, LW, JZ performed bioinformatics analyses of the St. Jude cohort data. JCG implemented the motif search algorithm. PA designed the study and revised the manuscript.

ACKNOWLEDGMENTS

I would like to thank all patients and their parents from Sainte-Justine University Hospital and the St. Jude Children's Research Hospital for participating in this study. I acknowledge P. Legendre and G. Bourret from Genome Quebec Innovation Center, J. Langdon and M. Hurles for providing *PRDM9* sequencing primer sequences, B. Ge and T. Pastinen for the genotyping data, T. Sontag, S. Leclerc and D. Arafat for preparing DNA samples, B. Li and G. Abecasis for the polymutt executable. I thank C. Bherer, M. Capredon, P. Donnelly, E. Kritikou and J. Quinlan for helpful discussions. This work was supported by MDEIE of Quebec, Canadian Foundation for Innovation, NSERC, Terry Fox Foundation CIHR, PCGP, the American Lebanese and Syrian Associated Charities of St. Jude Children's Research Hospital.

ABSTRACT

One of the most rapidly evolving genes in humans, *PRDM9*, is a key determinant of the distribution of meiotic recombination events. Mutations in this meiotic-specific gene have previously been associated with male infertility in humans (Irie et al. 2009) and recent studies suggest that PRDM9 may be involved in pathological genomic rearrangements. In studying genomes from families with children affected by B-cell precursor acute lymphoblastic leukemia (B-ALL), we characterized meiotic recombination patterns within a family with two siblings having hyperdiploid childhood B-ALL and observed unusual localization of maternal recombination events. The mother of the family carries a rare *PRDM9* allele, potentially explaining the unusual patterns found. From exomes sequenced in 44 additional parents of children affected with B-ALL, we discovered a substantial and significant excess of rare allelic forms of *PRDM9*. The rare *PRDM9* alleles are transmitted to the affected children in half the cases, nonetheless there remains a significant excess of rare alleles among patients relative to controls. We successfully replicated this latter observation in an independent cohort of 50 children with B-ALL, where we found an excess of rare *PRDM9* alleles in aneuploid and infant B-ALL patients. *PRDM9* variability in humans is thought to influence genomic instability, and these data support a potential role for PRDM9 variation in risk of acquiring aneuploidies or genomic rearrangements associated with childhood leukemogenesis.

INTRODUCTION

Most effort in cancer genomics has focused on capturing somatic mutations from the screening of tumor and normal somatic tissue genomes, to identify factors mutated somatically during tumor progression. Genetic mapping approaches aim to find genomic regions predisposing individuals to cancer, to capture inherited predisposing mutations segregating in the population by using genetic linkage or association studies. For late-onset cancers, such as breast and colorectal cancers (Turnbull et al. 2010; Peters et al. 2012), many predisposing allelic variants have been described, supporting a polygenic model of susceptibility (Easton and Eeles 2008) but only few genetic risk factors for pediatric cancer have been established (Healy et al. 2007; Sherborne et al. 2010). Dominant mutations causing cancer early in life are likely to be rapidly eliminated from the population, and as a result, it is unlikely that affected children will share inherited mutations. Parental germline events may play a role in pediatric cancer development, with early evidence for epigenetically marking of imprinted genes during meiosis (Joyce and Schofield 1998), that may be involved directly in tumorigenesis for cancers of embryonal origin, such as Wilms' tumours, rhabdomyosarcoma, adrenocortical carcinoma and hepatoblastoma. Besides this, little is known about the contribution of meiotic events to the genetic instability driving the early onset of childhood cancer. In particular, novel genomic changes that occur during meiosis will not be detectable using standard genetic mapping approaches. However, interrogating normal and tumor genomes from families of patients provides an ideal framework to study *de novo* genomic events potentially linked to childhood malignancies.

Recent genomic studies using family data have shown that many early onset diseases arise from defects caused by *de novo* genetic aberrations, be they point mutations (Awadalla et al. 2010), copy number variants (Greenway et al. 2009), structural rearrangements (Kloosterman et al. 2011) or aneuploidies (Hassold et al. 2007). Recombination rates in children correlate with maternal age at birth (Hussin

et al. 2011), which may have implications for understanding aneuploid conceptions. Intriguingly, children born with constitutional aneuploidies and rearrangements are at an increased risk for various malignancies (Ganmore et al. 2009). For example, children with Down syndrome have nearly a 20-fold increased risk for acute leukemia (Ross et al. 2005), suggesting that carcinogenesis and congenital anomalies may have a common basis for some pediatric cancers (Bjorge et al. 2008). Known recombination associated factors, such as DNA repair and histone modifications, are associated with genomic instabilities and cancers (Fernandez-Capetillo et al. 2004; Helleday 2010), and congenital genomic rearrangements and aneuploidies have been associated with errors in meiotic recombination (Hassold and Hunt 2001; Sasaki et al. 2010). Such gross genomic events are frequent in pediatric cancers.

Cancer is the leading cause of death by disease among children in western countries, and the overall incidence rate continues to rise steadily. The most common pediatric cancer, acute lymphoblastic leukemia (ALL), is a hematological malignancy resulting from chromosomal alterations and mutations affecting molecular pathways that disrupt lymphoid progenitor cell differentiation (Greaves 1999). Childhood ALL is likely explained by a combination of genetic predisposition and environmental exposure during early development, in fetal life and in infancy. However, genetic association studies for childhood ALL have been hampered by insufficient sample sizes (Healy et al. 2007; Sherborne et al. 2010). Furthermore, ALL is a heterogenous disease presenting many molecular subtypes, with different populations having different incidence rates, such that the power of stratified analyses will be limited due to small number of cases in each subgroup. Finally, there is well-established evidence for prenatal initiation of the leukemogenesis process in children (Wiemels et al. 1999; Greaves 2006), and focusing exclusively on child genetic material in ALL association studies may be insufficient for understanding disease etiology.

To characterize the importance of parental germline events in susceptibility to childhood ALL, we first set out to determine whether meiotic recombination

patterns can lead to factors associated with the development of childhood ALL. From exome sequencing and genotyping data, we characterized meiotic recombination patterns in a unique family (referred herein as the ALL quartet) with two siblings having hyperdiploid B-cell precursor ALL (B-ALL). We observed unusual localization of maternal meiotic recombination events, with a small number of crossovers taking place in previously well-characterized population recombination hotspots. Such hotspots are short segments (1-2Kb) identified to be highly recombinogenic in the human genome (Myers et al. 2005). The mother of the family carries a rare *PRDM9* allele, potentially explaining the unusual placement of recombination events observed (Berg et al. 2011). PRDM9 is a meiosis-specific histone H3 methyltransferase that controls the activation of recombination hotspots via it's zinc finger (ZnF) DNA binding domain recognizing a short sequence motif, which then triggers hotspot activity through modification of the chromatin state (Grey et al. 2011). Analyses of next-generation sequencing and Sanger re-sequencing read data from a cohort of parents with B-ALL affected children of French Canadian descent, revealed a substantial excess of rare allelic forms of *PRDM9*. This association was successfully replicated in an independent cohort of children with B-ALL diagnosed in Tennessee, USA, where the effect was found particularly in aneuploid and infant B-ALL patients. Not only has *PRDM9* variability been associated with with hotspot activation, but in humans it has been suggested to influence genomic instability (Berg et al. 2010; McVean and Myers 2010). The results presented here point to rare *PRDM9* allelic forms involvment in the development of preleukemic clones in B-ALL patients and we propose that PRDM9 histone H3K4 methyltransferase activity in the parental germline could lead to the genomic instability associated with childhood ALL, a plausible mechanism consistent with the current understanding of molecular pathways of leukemogenesis and the disease.

RESULTS

The ALL family quartet

To study germline processes such as recombination, data from families with at least two siblings is required. Within the Quebec Childhood ALL cohort, we identified a family with two siblings having hyperdiploid B-ALL diagnosed at Sainte-Justine University Hospital. Families with two cases of childhood ALL in a sibship are rare and it is not clear whether siblings of children with ALL have an increased risk of developing ALL themselves (Draper et al. 1996; Winther et al. 2001). From studies published between 1951 and 2009 and registry-based childhood ALL data, an international collaboration only identified three sibships that were concordant for hyperdiploid ALL (Schmiegelow et al. 2011). However, the high concordance rate in ALL subtype within sibships is somewhat incompatible with a scenario where all cases in sibships occur randomly through independent events.

Sampling from the family included six biological samples from four family members: the mother and father, sampled once, and their two sons, patients 383 and 610, sampled at diagnosis and in remission (Figure 1). The brothers were both diagnosed with B-ALL with FAB-L1 morphology, at the age of 2 for patient 383 and 3 years later, at 14 years of age, for patient 610. At diagnosis, both siblings showed hyperdiploid leukemia clones (>50 chromosomes, Supplementary Results), a childhood B-ALL subtype that is very likely to be prenatally initiated (Gruhn et al. 2008). However, chromosomal instabilities found in preleukemic clones are generated prenatally in the normal population at approximately 100-times the rate of overt ALL (Mori et al. 2002), thus a second hit is required to trigger ALL during childhood, and results from other genetic and/or environmental factors. Because the patients were both diagnosed within a 3 year time period, it is likely that the second hit is due to environmental exposure (Greaves 2006).

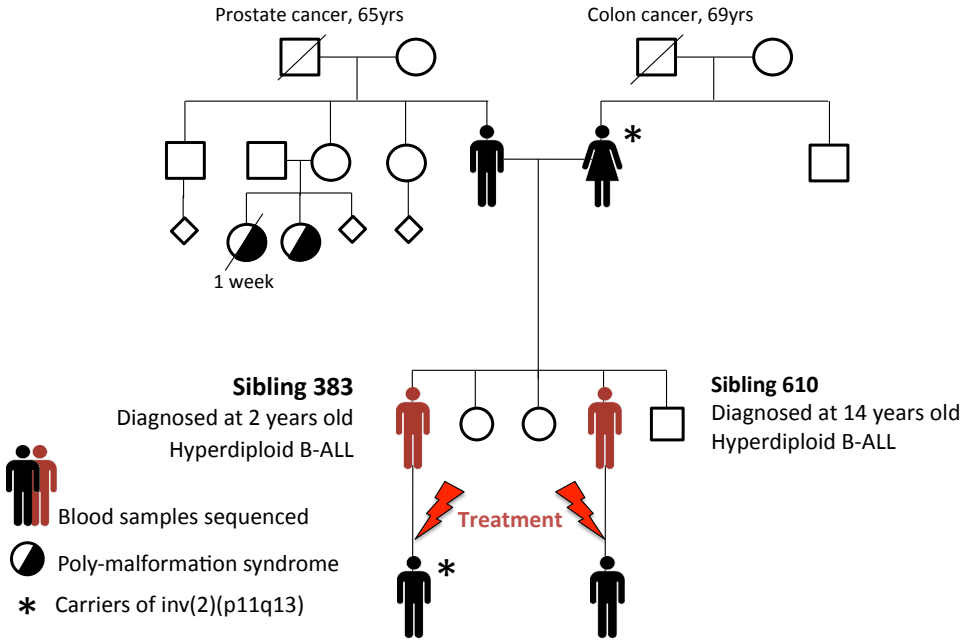


Figure 1. The ALL quartet family pedigree

The ALL Quartet is composed of the two parents and two brothers (patients 383 and 610) affected by hyperdiploid B-cell precursor childhood ALL, sampled prior to and after chemotherapy treatment. The brothers were diagnosed within a 3-year time period. The parents report Moroccan origins. Both maternal and paternal grandfathers are deceased from cancer. One of the father's sisters had children with poly-malformation syndromes, likely due to the high degree of consanguinity reported. Age at death is shown for deceased individuals.

De novo mutation and recombination events

Entire exomes were sequenced at high coverage using the SOLiD platform (Table S1A) to allow the full interrogation of the mutational and recombinational landscape occurring in coding regions (Material and Methods). Among variable positions discovered in the patients' exome, we looked for *de novo* mutations (Supplementary Methods). Given the human mutation rate, no more than one *de novo* point mutation is expected in a normal exome (Conrad et al. 2011). We only identified a single coding *de novo* point mutation in one of the patients (patient 383, Figure S1), which suggests that the parental germline mutation rate in this family is not higher than expected. Nevertheless, the putative *de novo* mutation identified is predicted to affect the structure and function of SMAD6. SMAD6 functions as an inhibitor of TGF-beta family signaling and was found to be a ligand-specific inhibitor of growth arrest and apoptosis in mouse B-cells (Ishisaki et al. 1999). Furthermore, SMAD6 is required for HL-60 myeloid leukemia cell line differentiation (Glesne and Huberman 2006) and is a key determinant of hematopoietic stem cell development (Pimanda et al. 2007).

Combining the exome sequencing with genotyping data obtained from Illumina Omni2.5 arrays, we identified over 816,000 high-confidence variable genomic positions within the ALL quartet exomes that showed no aberration in allele inheritance (Table S1B). We performed fine-scale dissection of meiotic recombination events on autosomes and located a total of 102 and 47 crossovers in maternal and paternal meioses, respectively (Figure 2A, Supplementary Methods). We also identified nine short tracts flanked by crossover events, possibly indicating gene conversion events (Supplementary Results, Table S2). The maternal and paternal mean number of recombination per meiosis for the ALL quartet were compared to the distribution of maternal and paternal means in a control cohort of French-Canadian families (FC family cohort, Material and Methods). The mother of the ALL quartet exhibits a high recombination rate with respect to the FC family

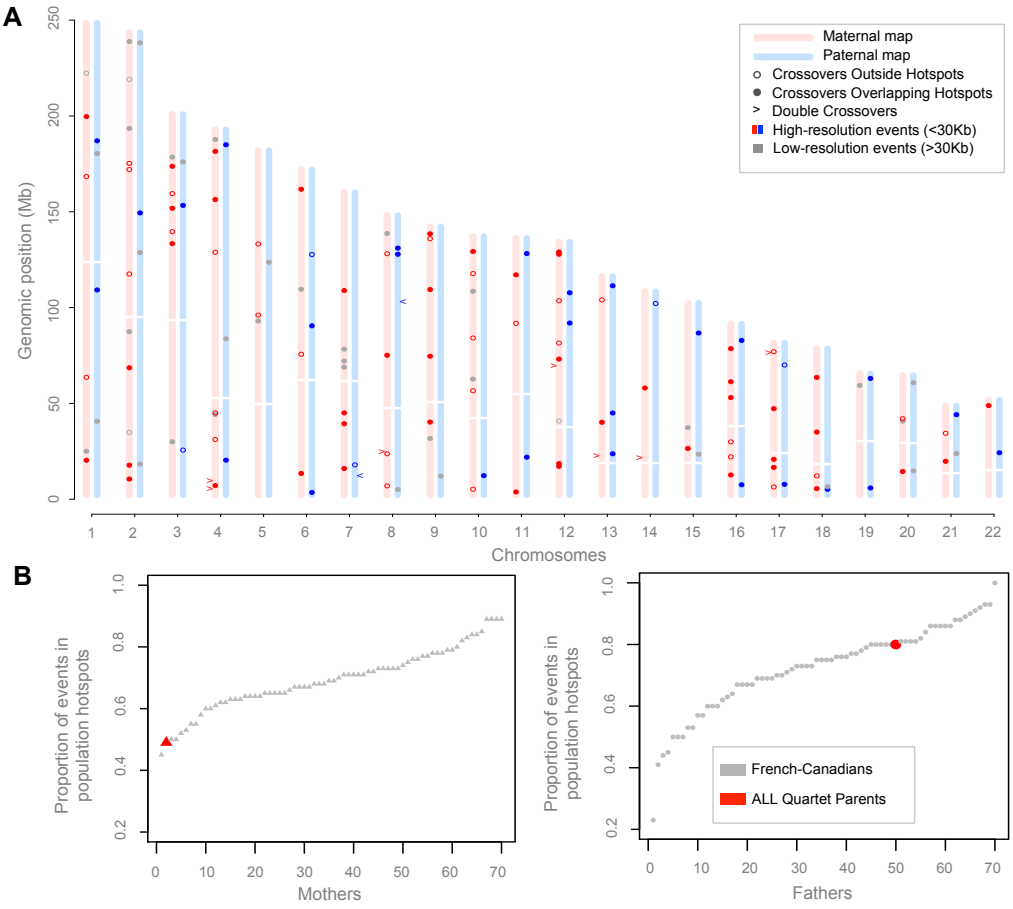


Figure 2. Map of recombinations events and hotspot usage in the ALL quartet

(A) Single and double crossovers in the two meioses that give rise to the patients, determined from analyses of SNPs from exome sequencing and genotyping data. Analyses were performed using pre- and post-treatment samples and only kept crossovers inferred in both. Using two somatic tissues allowed us to remove genotyping errors and double recombination events resulting from errors. All crossovers displayed are supported by at least 3 informative markers and high-resolution events are localized between informative markers less than 30 Kb apart (B) Fraction of high-resolution crossover intervals overlapping population hotspots in the FC family cohort and in the ALL quartet. Mothers (triangles) and fathers (circles) are ordered according to their proportion of overlap. We estimate that 11.78% (10.56-13.24 CI 95%) of these crossover intervals are expected to overlap population hotpots by chance.

cohort, with her mean number of crossovers per gamete found to be at the end of the spectrum (Figure S2). She carries two copies of the haplotype at the RNF212 locus associated with high recombination rates in females (haplotype [T, C] at SNP rs3796619 and rs1670533) (Kong et al. 2008). The father carries one copy of the rs3796619 T allele, which was estimated to decrease male genome-wide recombination rate by 2.62% per copy. This is consistent with the lower recombination rate seen in the father of the ALL quartet (Figure S2).

We next evaluated hotspot usage by computing the proportion of high-resolution recombination events (localized between informative markers less than 30 Kb apart) overlapping known population hotspots inferred from HapMap2 data (Myers et al. 2005). This measure does not directly evaluate hotspot usage, since a fraction of crossovers is expected to map in population hotspots by chance alone, but it is a good proxy to compare relative hotspot usage. Among recombination events identified in the ALL quartet, 85 maternal events and 34 paternal events were localized between informative markers less than 30 Kb apart. Fifty-four percent (46/85) of maternal events and 79% (27/34) of paternal events overlapped with HapMap2 recombination hotspots. These proportions are significantly different between the parents (p = 0.0124, Fisher's exact test). Since it was reported that, on average, 70% of events are expected to overlap population hotspots (Coop et al. 2008; Hussin et al. 2011), this result suggests that the mother has a lack of recombination in population hotspots. To validate this result, we further derived a null distribution of the proportion of recombination events expected to overlap with HapMap2 recombination hotspots in the FC family cohort (Material and Methods). The paternal crossovers show the expected enrichment in population hotspots inferred from HapMap2 data, while the maternal recombination landscape is unusual with a particularly low proportion of meiotic recombination events occurring in HapMap2 hotspots (Figure 2B).

Characterising PRDM9 in the ALL family quartet

Variability in recombination hotspot usage correlates with variation in *PRDM9*, a gene identified as a major hotspot determinant in mammals (Baudat et al. 2010; Berg et al. 2010; Myers et al. 2010; Parvanov et al. 2010) and the only locus known to be involved in hotspot specification in humans (Kong et al. 2010; Hinch et al. 2011). Allelic variation at the *PRDM9* locus consists of variable repeating units, encoded by a minisatellite formed by tandem-repeat C2H2 zinc finger (ZnF), and has a strong effect on recombination hotspots positioning and activity (Berg et al. 2010; Berg et al. 2011). The reduced proportion of recombination events overlapping population hotspots in the mother of the ALL quartet lead us to investigate genetic variation at the *PRDM9* gene. The sequencing data revealed that the father of the two affected brothers is homozygote for the most common 13-repeat allele (allele A) whereas the mother carries an allele A and a 14-repeat allele C, inherited by one of the two brothers (Supplementary Results). We further validated the presence of the C allele, which encodes major changes in the PRDM9 ZnF array (Baudat et al. 2010), by Sanger re-sequencing (Material and Methods).

Although this allele is rare in populations of European ancestry (~1%), it is more frequent in individuals of African descent (~13%) (Berg et al. 2010). Because the parents have Moroccan Arab ancestry (Supplementary Results, Figure S3), we studied *PRDM9* diversity in 27 Moroccan individuals to establish whether the presence of the C allele in the mother reflects a different distribution of Moroccan *PRDM9* alleles relative to populations of European descent. Among 54 Moroccan *PRDM9* alleles sequenced, no C allele was found (Figure S4), suggesting that the frequency of the C allele in the Moroccan population is similar to the observed frequency in populations of European descent (Supplementary Results).

Motifs overrepresented in recombination hotspots in European and African American individuals have been inferred from population studies (Myers et al. 2008; Hinch et al. 2011): a 13-mer motif is enriched in linkage disequilibrium-based

hotspots inferred from HapMap2 data, whereas a 17-mer motif is overrepresented in African-enriched hotpots. The A allele has been shown to bind to the 13-mer CCNCCNTNNCCNC motif, whereas the C allele has demonstrated inability to activate recombination hotspots presenting this motif (Berg et al. 2010). The C allele is predicted to specifically bind to a 17-mer motif CCNCNNTNNNCNTNNCC (Berg et al. 2011), very close to the motif found to occur at an increased frequency in African-enriched hotspots (Hinch et al. 2011). Compared to the distribution of these motifs seen at recombination events in the FC family cohort, we observed that the 17-mer motif is highly represented in the ALL mother's recombination events whereas the 13-mer motif is under-represented (Figure S5, Material and Methods). These observations confirm that the presence of the C allele in the mother is likely to have caused the genome-wide shift from HapMap2 hotspots observed in the maternal recombination landscape (Berg et al. 2011; Hinch et al. 2011).

Association between PRDM9 and ALL in parents

Recent studies suggest that PRDM9 is implicated in genomic rearrangements leading to congenital diseases (Myers et al. 2008; Berg et al. 2010; Borel et al. 2012). Because large-scale genomic rearrangements are common events in childhood leukemia, we sought to type *PRDM9* alleles in the parents of the B-ALL cohort. We assayed *PRDM9* ZnF alleles in a panel of 44 additional parents from 22 French-Canadian families with children affected by B-ALL (FCALL cohort) by analysing reads aligning to the PRDM9 ZnF array, in the parental trios where exome sequencing was previously performed (Material and Methods). The read data allows for the detection of all zinc finger repeats present in the common allele A, along with two rarer repeats, *k* and *l,* present in the C allele (Figure S6). Around one fourth of the parents (12/46, Table S3 and S4) have alleles with *k* finger repeats (*k*-finger alleles), usually rare in populations of European descent (Berg et al. 2010), suggesting that *k*-finger alleles are in excess in the FCALL cohort (p = 0.0181, Supplementary Results). To validate this result, Sanger re-sequencing of the *PRDM9* ZnF alleles was further

performed in 13 pairs of parents from the FCALL cohort and in 76 parents from the ethnically matched FC family cohort (Figure S7, Supplementary Results). Among the 26 B-ALL parents re-sequenced, evidence for the presence of rare alleles in the read data had been found in 9 parents from 8 families, and all of them were confirmed. Through re-sequencing, we discovered rare alleles in two additional families, which were not originally detected in the exome sequencing reads. In contrast, only 5 parents from the FC family cohort carried *k*-finger alleles (Table S5). This confirms the excess of rare alleles in the FCALL cohort with respect to controls ($p = 3.22 \times 10^{-3}$, Figure 3), with 76.9% of B-ALL families with at least one parent carrying a rare *PRDM9* allele. The rare alleles were preferentially carried by the mother ($p = 0.0235$, Fisher's exact test), although this maternal effect is not observed for *k*-finger alleles alone. Furthermore, the observation is not restricted to families with children having hyperdiploid B-ALL, since alternative *PRDM9* ZnF alleles are also found in parents of children presenting translocations and as yet uncharacterized genetic defects (Table S6). Finally, we found no significant evidence for transmission distortion, as *k*-finger parental alleles were transmitted to the affected child in 6 out of 11 cases with carrier parents (Table S3 and S4), resulting in 25% of the children of the FCALL cohort (6/24) carrying a *k*-finger allele.

Replication in a B-ALL patient cohort

The association was also detectable in the patients themselves ($p = 0.0123$). We replicated this latter association in an independent cohort of 50 children, sequenced whole genome, from St. Jude Children's Research Hospital. The children were affected by B-ALL from four subtypes: ETV6-rearranged, Phildelphia chromosome-positive, hypodiploid and infant B-ALL. We observed an excess of B-ALL patients with rare alleles in the St. Jude ALL cohort, with read data showing evidence for the *k* and/or *l* fingers in 10 children (Table S7). This excess is significant with respect to the French-Canadian controls ($p = 0.0143$, Table 1) and to 1000 Genomes Project controls from the CEU population ($p = 0.0353$, Supplementary Results). No *k*-finger

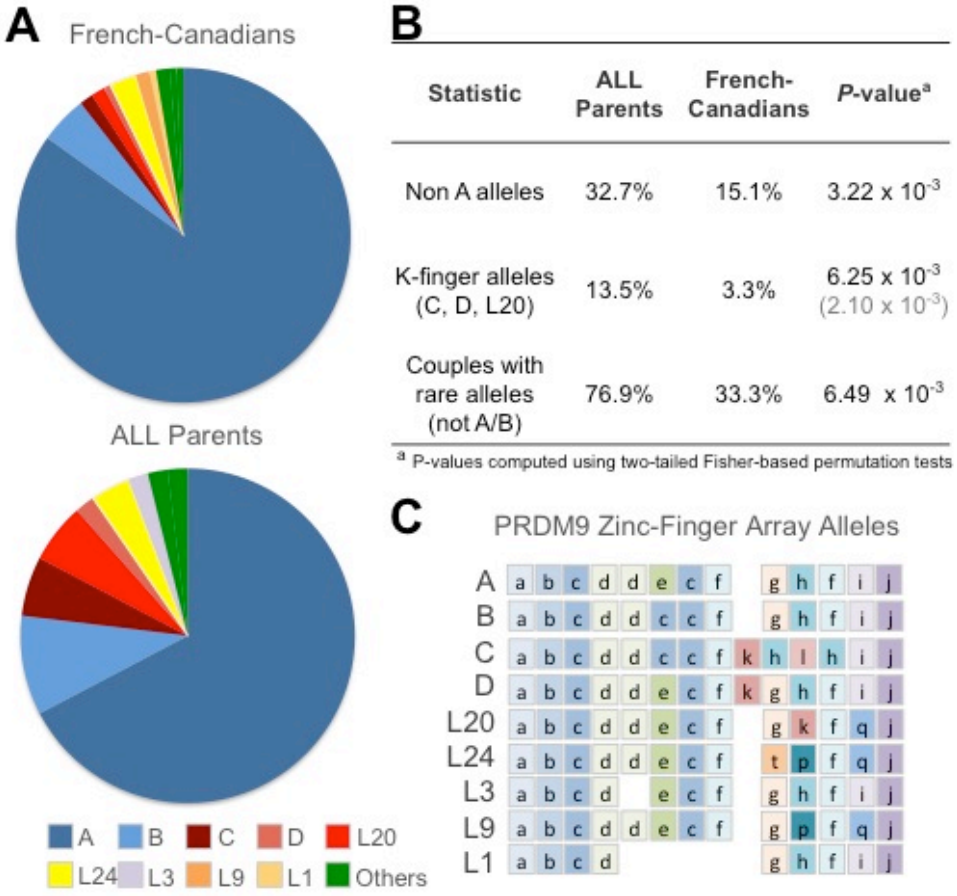


Figure 3. Excess of rare *PRDM9* alleles in parents from the FCALL cohort

(A) Pie charts showing frequencies of *PRDM9* zinc-finger (ZnF) alleles obtained through Sanger sequencing of 26 parents of patients with B-ALL and 76 parents from the FC family cohort (controls). Alleles labeled as 'Others' are population-specific alleles. Individuals' alleles are detailed in Table S3 and S5. (B) Differences in allele frequencies between parents of patients and controls. The *p*-value in parentheses was calculated by including alleles inferred from exome sequencing reads for the 20 ALL parents for which *PRDM9* ZnF arrays were not re-sequenced by Sanger (Table S4). Applying a Bonferroni correction for performing the same test in two subsets of alleles (non-A and *k*-fingers alleles) would yield $\alpha = 0.025$, although this correction is particularly conservative since subsets are correlated. (C) Allelic structure of *PRDM9* ZnF array for alleles found in these cohorts (population-specific alleles not shown).

Table 1. Replication of the association between *PRDM9 k*-finger alleles and in patients from St. Jude ALL cohort

ALL subtypes	All Patients			Patients of European Ancestry		
	Individuals	*k*-finger alleles	*P*-value[a]	Individuals	*k*-finger alleles	*P*-value[a]
B-ALL	50	10	0.0143	39[b]	7	0.0396
Hypodiploid B-ALL	16	5	7.50×10^{-3}	12	4	8.85×10^{-3}
Infant B-ALL	18	5	0.0122	14	3	0.0630
ETV6 B-ALL	9	0	-	9	0	-
Ph+ B-ALL	7	0	-	4	0	-

[a] P-values from permutation tests based on one-tailed Fisher's exact tests on counts between cases and controls

[b] The 39 patients represent a subgroup of the cohort of 50 patients

Patient's ethnicities were verified by principal component analyses on genetic variation using Eigensoft package (Figure S8, Supplementary Results. Patients do not have African ancestry. Controls consist of 76 French-Canadians individuals sequenced at the *PRDM9* locus and showing a total of 5 *k*-finger alleles (Table S5). Association testing between cases and controls was performed for subgroups with sample size greater than 10 individuals. Applying a conservative Bonferroni correction to correct for testing in two independent B-ALL subgroups would yield $\alpha = 0.025$.

alleles were detected in B-ALL patients from Philadelphia chromosome-positive and ETV6-rearranged subtypes. In hypodiploid and infant B-ALL subtypes, the excess of *k*-finger alleles is significant (Table 1). The children have no African ancestry, and 39 of them ethnically cluster with the controls of European ancestry, whereas the other children have different levels of Hispanic, Asian and Native American ancestry (Figure S8, Supplementary Results). Although the frequencies of PRDM9 *k*-finger alleles in Chinese and Mexican individuals are likely similar to the ones observed in Europeans (Parvanov et al. 2010), to be conservative we also tested for the association between *k*-finger alleles and B-ALL when only non-admixed white patients were included. The association remains significant overall and in the hypodiploid subtype, although the effect becomes marginal in the infant B-ALL subtype (Table 1).

PRDM9 binding motifs and ALL translocations

Chromosome-number abnormalities and chromosomal rearrangements have been associated with altered recombination (Hassold and Hunt 2001; Sasaki et al. 2010). Given that PRDM9 may be responsible for causing recombination-associated pathological genomic rearrangements (Myers et al. 2008; Berg et al. 2010; Borel et al. 2012), the high frequency of translocations and aneuploidies in leukemia raises the question of whether PRDM9 is implicated in the generation of preleukemic cells early in development. In particular, the C allele accounts for an important fraction of the rare alleles detected in the ALL cohorts, and its binding motif as been predicted and validated (Baudat et al. 2010; Berg et al. 2011). To assess whether the C *PRDM9* allele was more likely than the common A allele to bind to genes known to be involved in ALL translocations (ALL gene list, Table S8), we performed a motif search to identify putative A and C binding sequences in the human reference genome. We found an enrichment of sequences potentially recognized by allele C relative to allele A in the ALL gene list in comparison to the rest of the reference genome (OR=1.53 [1.15;2.04]) and to other coding regions (OR=1.61 [1.21;2.15]) (Table 2).

Table 2. PRDM9 alleles binding motifs in the Human Reference Genome.

Reference genome	PRDM9 allele	Motif	Number of Motifs		
			Genome	Genes	ALL gene list
Standard	A	*A:* G..C.....CC.CC....C.CC	12319	6953	34
	C	*C:* G..C.....CC.....C....CC	176826	95241	748
Degenerate	A	*A:* G..C.....CC.CC....C.CC	13335	7504	42
	C	*C:* G..C.....CC.....C....CC	186374	100041	789

Motif comparisons	ALL gene list vs Genome OR [CI]	ALL gene list vs Genes OR [CI]	Random gene lists
C vs A in Standard	1.53 [1.15;2.04] **	1.61 [1.21;2.15] **	43/1000
C vs A in Degenerate	1.53 [1.16;2.01] **	1.60 [1.22;2.12] **	48/1000

** significant based on 95% CI (one-tailed $p < 0.025$)

The motif search was performed on the Human Reference Genome (hg18). Motifs consist in DNA binding sequences predicted by Berg and colleagues (Berg et al. 2010) for each allele. The number of motifs is reported for whole genome, coding regions and the ALL gene list. The ALL gene list was built based on the most frequent ALL translocations reported in databases (Table S8). We compared counts using odds ratios (OR), to measure the association between motifs and their occurrence in the ALL gene list. For a given motif comparison, OR values were computed for 1000 random gene lists (*13*) and we computed the number of times significant ORs (two-tailed) are seen for both whole genome and genic regions. Results for unique and repetitive DNA are shown in Table S9.

The excess of C binding motif relative to A binding motif within the ALL gene list was found to be significant in unique DNA (OR = 2.39, Table S9) but not in repetitive DNA, except for segmental duplications. The C motif is highly overrepresented relative to the A motif in segmental duplications in the ALL gene list compared to segmental duplication in other genes. Indeed, after correcting for the higher proportion of sequences matching the C motif than the A motif in the genome, the C binding motif is more than 4 times more likely to be found in a segmental duplication within the ALL genes than the A motif (OR = 4.78, Table S9).

Infant B-ALL patients show an excess of C *PRDM9* alleles in the St. Jude ALL cohort and harbor leukemic clones with translocations involving the *MLL* gene on chromosome band 11q23. We therefore scanned the nucleotide sequence of *MLL* and found a motif matching the C putative binding sequence occurring within the breakpoint cluster region (Figure 4A) whereas no A binding motif was present in *MLL*. However, 75–90% of genomic DNA sequences are packaged into nucleosome particles, blocking the DNA from interacting with DNA binding proteins (Segal et al. 2006). Studies suggest that the sequence itself is highly predictive of nucleosome positioning, we thus used an *in silico* approach (Xi et al. 2010) to predict nucleosome positioning within the *MLL* breakpoint cluster region (Material and Methods). The tool predicts a potential starting position of the nucleosome at the end of the motif identified (Figure 4B), suggesting that the motif might be accessible in a stretch of unwrapped linker DNA. It follows that *PRDM9* C allele can potentially bind the MLL breakpoint cluster region, although this needs to be demonstrated *in vitro* and *in vivo*. Additionally, we re-analysed translocation data from sperm cells in men with known *PRDM9* alleles (Berg et al. 2010). The t(11;22)(q23;q11) translocations, often resulting in a MLL-rearrangements, occur at significantly higher frequencies in European males with *k*-finger alleles compared to those without *k*-finger alleles (p = 0.0436, Kruskal-Wallis test). However, no significant difference in translocation frequencies was observed between individuals of African descent with and without *k*-finger allele (p = 0. 7998, Kruskal-Wallis test).

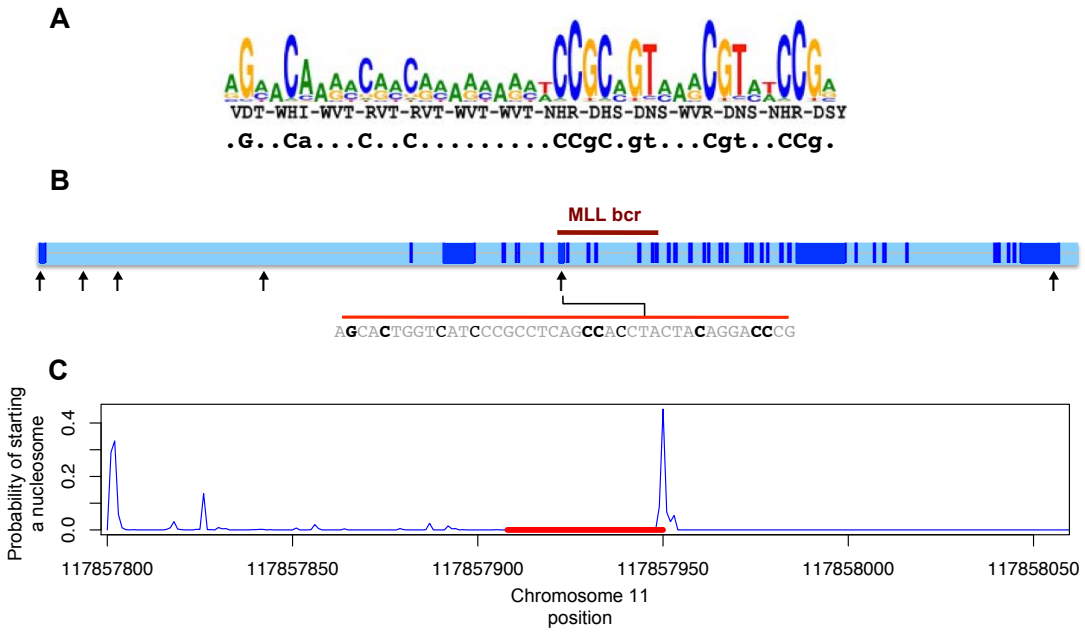


Figure 4. PRDM9 C binding motif in the MLL breakpoint cluster region.

(A) Logo plot of the C allele binding motif (Baudat et al. 2010), predicted based on the 3 indicated residues forming the binding unit of the ZnF repeats (positions -1, 3 and 6 of the ZnF alpha helices) and the consensus sequence motif simplified showing the most strongly predicted bases (in lowercase for >80% consensus for a specific base and in uppercase for >95% consensus (Berg et al. 2010) (B) Presence of a motif at chr11:117857908-117857950 (hg18), within the breakpoint cluster region of *MLL*, matching the predicted *PRDM9* C allele binding motif for the 7 strongly predicted bases shown in uppercase in the consensus sequence presented in (A), and 3 predicted bases shown in lowercase. Intronic regions are displayed in light blue. Black arrows show the positions of all occurrence of sequences matching the motif at uppercase characters. No *PRDM9* A allele binding motif was found in *MLL*. (C) Nucleosome starting positions predicted by NuPoP (Xi et al. 2010). The red line shows the position of the predicted C binding motif.

DISCUSSION

In this study, we examined germline processes in families with children having childhood pre-B acute lymphoblastic leukemia (B-ALL). With exome sequencing and dense genotyping data, we were able to capture parental germline recombination events in a unique family with two affected siblings. We identified PRDM9 as being associated with unusual recombination patterns and discovered a substantial excess of rare allelic forms of *PRDM9* in two independent ALL cohorts. In both the initial and replication ALL cohorts, care has been taken in controlling for population structure (Supplementary Results), the cause of many false-positive genetic associations. These data support the hypothesis that PRDM9 rare allelic variants are associated with ALL in children, but represent a relatively small dataset and the findings require further support in independent cohorts. The association should also be investigated in other types of childhood leukemias, such as T-lineage ALL and acute myeloid leukemia, as well as in parents of children with constitutional aneuploidies (Ganmore et al. 2009) or in woman experiencing molar pregnancies (Roman et al. 2006). The minisatellite alleles of *PRDM9* have to be carefully typed from sequencing read data and rare alleles should be validated through re-sequencing. These alleles are known to cause functional biological variation, as variants in the *PRDM9* gene influence recombination locations, although they have little effect on the total genome-wide recombination rate (Kong et al. 2010). If confirmed, this novel association suggests additional biological function for *PRDM9* allelic variation that might impact other processes than meiotic recombination.

These findings raise many important questions about both leukemogenesis and PRDM9 function. First of all, PRDM9 activity is likely exclusive to parental germ cells but it remains unclear if it acts in the patient somatic cells. The parents carrying rare *PRDM9* alleles only transmit the susceptibility allele to half of their affected children in the FCALL cohort, indicating that these alleles may act during meiosis, giving rise to gametes that predispose the offspring to B-ALL. Furthermore, PRDM9 specific

expression and function at early stages of meiosis (Hayashi et al. 2005) support the parental model and point to a germline mechanism of ALL development. However, *PRDM9* expression has been observed, albeit at low levels, in several hematopoetic tissues including leukemia cell lines (Johnson et al. 2003). Genomic data from additional family cohorts is needed in order to definitely resolve this question. Additional family data will also be informative with respect to the sex-specificity of the effect. Indeed, the higher frequency of maternal rather than paternal carriers of rare alleles among parents in the FCALL cohort suggests a strong sex-specific effect, at least for some alleles. Maternal-specific effects on recombination with implications for child health have been demonstrated in humans, such as maternal age effect on recombination (Hussin et al. 2011) and a maternal origin of most division errors leading to trisomies (Hassold and Hunt 2009). The sex-specific effect of PRDM9 alleles on B-ALL risk could result from the differences in recombination patterns along chromosomes and in hotspot usage between males and females mammals (Paigen et al. 2008; Kong et al. 2010) or from sexual dimorphism in the regulation of the meiotic process (Cohen et al. 2006).

The mechanism underlying this novel association is not known, yet previous evidence suggests that PRDM9 may be responsible for chromosomal translocations due to its ability to determine sites of genetic crossing over (Myers et al. 2008; Berg et al. 2010; Borel et al. 2012). Therefore, PRDM9 could be implicated in the generation of some chromosomal rearrangements found in leukemia, a hypothesis supported by analysis of putative binding motifs of *PRDM9* alleles occurring in a subset of genes involved in chromosomal rearrangements frequently found in ALL (Table S8). In these analyses, we used the *in silico* predicted binding motifs for the *PRDM9* C and A alleles as predictors for the binding activity of PRDM9 ZnF array to DNA sequences. PRDM9 binding properties are still mysterious (Segurel et al. 2011), and PRDM9 *in silico*-derived binding sites are not necessarily reliable for predicting PRDM9 binding activity. This is because *in silico* predictions using zinc finger databases and the algorithm developed by Persikov and colleagues (Persikov et al.

2009) give a vast excess of sites compared to those actually bound by PRDM9. Nevertheless, the binding predictions for *PRDM9* common allele A led to the discovery of the role of this gene in human recombination (Myers et al. 2010) and the predicted C binding motif matches almost perfectly the DNA motif found in excess in African-enriched hotspots (Hinch et al. 2011). Therefore, it appears that human alleles A and C are able to bind to at least a subset of these genomic motifs. It follows that the significant excess of sequences matching the C motif compared to A motif in the ALL translocated genes, although not a demonstration that C binds these sequences, suggests that the C allele is more likely than the A allele to bind in these fragile regions. In particular, we identified a motif matching the C allele binding motif occurring within the breakpoint cluster region of *MLL* (Figure 4), a gene translocated in infant B-ALL patients, however *in vitro* and *in vivo* binding of PRDM9 C allele in this region remains to be demonstrated.

Importantly, translocations in ALL patients generally occur somatically. In humans, perturbation of the H3K4me3 dynamics at early stage of development specifically leads to inappropriate differentiation of haematopoietic progenitor cells (Chi et al. 2010). Furthermore, the H3K4me3 mark, specifically placed by PRDM9 (Hayashi et al. 2005), is a histone methylation event mis-regulated in many pediatric cancers (Schwartzentruber et al. 2012; Wu et al. 2012; Zhang et al. 2012) and has been linked to leukemia initiation (Chi et al. 2010). In particular, deregulation of factors that mediate H3K4me3 interferes with RAG-mediated V(D)J recombination, crucial for B cell maturation, and affects hematopoietic cell populations. For example, local accumulation of H3K4me3 has recently been shown to occur within the breakpoint cluster region of BTG1, a driver gene affected by deletions that result from aberrant somatic recombination events in B-ALL (Waanders et al. 2012). Aberrant histone methylation in germline may therefore help establish tumor-initiating cell populations in early leukemogenesis. However, this hypothesis implies that H3K4me3 marks would be passed on to the child and maintained until early development. Transgenerational epigenetic inheritance has been recently reported

for the H3K4me3 mark in *C. elegans* (Greer et al. 2011) and depends on chromatin modifiers but also on the H3K4me3 demethylase RBR-2 acting in germline, suggesting that other contributors, acting in concert with PRDM9, would be required to disrupt normal H3K4me3 patterns.

As these results potentially link allelic variation at *PRDM9* with childhood ALL risk, it is reasonable to expect that higher frequencies of alternative alleles in individuals of African descent (Berg et al. 2010; Hinch et al. 2011) would indicate a potentially higher incidence of childhood leukemia in African populations. Incidence of childhood leukemia in sub-saharian Africa is not well documented, however the incidence rate among admixed African-American children is approximately half the rate in children of European descent (Gurney et al. 1995). Therefore, PRDM9 potential role in ALL will likely involve multi-locus interactions arising in specific genomic backgrounds. The significant difference in frequency of t(11;22)(q23;q11) translocations between men with and without *PRDM9 k*-finger alleles found in Europeans but not in Africans suggests that different alleles, or combination of alleles in a heterozygote, may not have the same impact on different genetic backgrounds. These differences could arise due to the existence of variation between European and African individuals in factors implicated in regulating H3K4me3 in meiosis after DSB repair. Moreover, since PRDM9 interacts with specific binding motifs to regulate histone methylation and recombination, these loci, if mutated, could modify downstream PRDM9 deleterious functions. Two studies (Jeffreys and Neumann 2002; Myers et al. 2010) have shown that self-destructive drive due to biased gene conversion disrupts the common-allele binding motifs and, in the African population, the same process appears to be eliminating binding motifs for alleles at higher frequencies (Berg et al. 2011). Furthermore, a recent study argued that genetic ancestry is critical to understand ALL risk and failure to go into remission, an indicator of relapse risk in ALL (Yang et al. 2011). In this context, it makes sense to consider that not only are *PRDM9* alleles critical, but also that the

ancestral background of the patients is a key factor for the role of PRDM9 in leukemogenesis.

While PRDM9 is known to be involved in sterility in human (Irie et al. 2009) and mice (Hayashi et al. 2005), this is the first study to specifically implicate PRDM9 in human disease and this novel association will hopefully inspire further investigation. PRDM9 clearly interacts with multiple factors to facilitate proper histone methylation and recruit recombination, but its molecular partners remain largely unidentified (Segurel et al. 2011) and the biological importance of PRDM9 is not fully understood. Finally, if PRDM9 is implicated in a germline mechanism of ALL development, it would mean that risk factors for ALL could be established as soon as during meiosis in the parental germline. Therefore, the results reported here raise the intriguing possibility that germline events and recombination processes play a role in the susceptibility to pediatric cancer, which have considerable implications for mapping strategies as well as prognosis and treatment of childhood leukemia.

MATERIAL AND METHODS

Datasets

The initial French-Canadian B-cell precursor acute lymphoblastic leukemia cohort (FCALL cohort) includes blood samples from 23 French-Canadian nuclear families, which include 22 parental trios (both parents and an affected child) and one family composed of both parents and two affected brothers, referred herein as the ALL quartet. DNA from each sample was extracted from peripheral blood cells or bone marrow as previously described (Baccichet et al. 1997). Exome sequencing using the Applied Biosystems SOLiD 4.0 System, read mapping and variant calling were performed as outlined below. Study subjects are from the established Quebec Childhood ALL cohort (Healy et al. 2010) diagnosed in the Hematology-Oncology Unit of Sainte-Justine University Hospital, Montreal, Canada, between October 1985 and November 2006.

The replication cohort (St Jude ALL cohort) is composed of 50 unrelated children affected with B-ALL treated at St-Jude Children's Research Hospital, Memphis, Tennessee, USA. Whole-genome DNA sequencing was performed using Illumina HiSeq paired-end sequencing at coverage of 30X, aligned using BWA (0.5.5) aligner to the human NCBI Build 36 reference sequence. The Institutional Review Board of the respective hospitals approved the research protocol and informed consent was obtained from all participants and/or their parents.

Exome sequencing data from two control cohorts were used in this study. The first one is a French-Canadian exome dataset (FCEXOME) consisting of 68 French-Canadian controls for whom exome sequencing reads aligning to PRDM9 ZnF array were available. The second exome data control cohort comprises 99 individuals of European descent from the CEU population sequenced in the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). The read data from these control cohorts are further described in Supplementary Results. *PRDM9* re-sequencing data

and genotyping data from two additional control cohorts were included: a cohort of families of self-declared French-Canadian origin (FC family cohort) (Hussin et al. 2011) and a cohort of unrelated Moroccan individuals (Moroccan cohort) (Idaghdour et al. 2010). Table S10 summarizes cohorts' composition, data generated and analyses performed on each dataset. We also used SNP data from the publicly available HGDP (Cann et al. 2002) and HapMap3 (Consortium 2005) datasets as controls in some ancestry and association analyses.

Exome Sequencing in the FCALL cohort

Exome capture was performed with the *SureSelect Target Enrichment System* from Agilent Technologies using the protocol optimized for Applied Biosystems' SOLiD sequencing. For each sample of the FCALL, single-end exome sequencing cohort was performed and generated approximately 5 Gb of mappable sequence data per sample. Color space reads were mapped to the NCBI Build 36 reference sequence with BioScope v1.2 (Ondov et al. 2008). Base quality recalibration was performed using GATK and BAQ (Li et al. 2009; McKenna et al. 2010). PCR duplicates were removed using Picard implemented in Samtools (Li et al. 2009). In the ALL quartet, SNP calling from exome data was performed as described in Supplementary Methods. *De novo* mutation discovery was performed using the DND software (Cartwright et al. 2012). In the 22 trios, we called SNPs in parents using Samtools and only SNPs within targeted regions with Phred score above 30 were kept.

Genome-wide SNP Arrays

Genotyping data was available for all individual included in this study except for the 22 trios from the FCALL cohort. The ALL quartet samples were genotyped using the Illumina HumanOmni 2.5-quad BeadChips and the Affymetrix 6.0 arrays (Supplementary Methods). Normal samples from children in the St Jude ALL cohort were genotyped on Affymetrix SNP 6.0 (48 individuals) or 250k Nsp and 250k Sty (2 individuals). The Moroccan cohort comprises 163 unrelated individuals genotyped

on the Illumina's Human 610-Quad SNP Beadchip (Idaghdour et al. 2010). The FC family cohort comprises 69 nuclear families with at least two offsprings, genotyped on the Affymetrix 6.0 array (Hussin et al. 2011).

Recombination Analyses

Recombination events were called in two datasets using the algorithm described previously (Hussin et al. 2011) available at www.iro.umontreal.ca/~hussinju/NucFamTools.html, on two datasets :

(i) the ALL quartet recombination SNP dataset, which is obtained by combining the exome and the Illumina SNPs (Supplementary Methods). The recombination events were called separately for the pre-treatment and post-treatment samples. All double recombination events separated by 1 and 2 informative markers were ignored. The pre-treatment and post-treatment sets of recombination events were subsequently compared and only events seen in both were kept. All events excluded were double recombinants separated by less than 5 informative markers. Full description of the recombination landscape in the parents of the ALL quartet is provided in Supplementary Results.

(ii) the Affymetrix SNP dataset for the 69 families from the FC family cohort and the ALL quartet. These samples were genotyped on the same chip, allowing direct comparison of the ALL quartet recombination landscape with recombination patterns observed in the FC family cohort. The 355,271 SNPs used in (Hussin et al. 2011) were selected and 20,118 SNPs were removed because of missing genotypes or mendelian errors in the ALL quartet data. Recombination events were located using two children in each family : for families with three offspring or more in the FC family cohort, two of the latter were chosen at random to match the ALL quartet family structure. We removed all double recombination events separated by less than 5 informative SNPs.

The congruence of HapMap2 recombination hotspots (Consortium 2005) was assessed using events localized between informative markers less than 30 kb apart. We estimated the proportion of recombination events that overlap a HapMap2 hotspot by chance using the approach described in (Coop et al. 2008). We also investigated the overlap between the inferred maternal and paternal recombination events and the putative *PRDM9* alleles A and C binding motif. Again, we only considered events localized between informative markers less than 30 kb apart. We also excluded recombination events overlapping a DNA sequence matching both motifs and, for each individual, we computed the proportion of events overlapping sequences matching exclusively A or C predicted binding motifs.

PRDM9 Zinc Fingers Typing in Short Read Data

To identify PRDM9 zinc fingers (ZnF) repeat types present in an individual, we analysed sequencing reads that aligned to the *PRDM9* ZnF array (exon 11) of the human NCBI Build 36 reference sequence (hg18), corresponding to the region chr5:23562605-23563612. The reads were extracted from BAM files before removing PCR duplicates i.e. multiple reads starting at the same reference coordinate. This is because applying this filter would cause the removal of reads sampling additional ZnF repeats, absent from the reference genome, that will align to the repeats present in the reference genome. Each read was aligned to the known human PRDM9 ZnF types identified in previous studies (zinc finger repeat types *a* to *t*, Figure S6). We computed the proportion of reads that aligned uniquely and without mismatch to each ZnF type. Given the proportion of reads aligning to types *b*, *c*, *d* and *f*, included in all PRDM9 ZnF alleles reported so far (Berg et al. 2010), we determine an inclusion criteria of 1% to infer the presence of a ZnF in a sample. Validation experiments by Sanger sequencing allowed us to confirm the accuracy of the 1% empirical criteria, since ZnF types predicted using this approach were present in the re-sequenced *PRDM9* alleles.

Sanger Sequencing of PRDM9 ZnF Alleles

We sequenced the ZnF array of *PRDM9* in 26 parents from the ALL cohort (including the ALL quartet parents), 76 parents from the FC family cohort and 27 unrelated individuals from the Moroccan cohort, resulting in a total of 258 alleles sequenced. *PRDM9* ZnF alleles were amplified from 5-20 ng of genomic DNA using the primers HsPrdm9-F3 and HsPrdm9-R1 (Baudat et al. 2010), designed to discriminate *PRDM9* from his paralogous copy *PRDM7*. Alleles were sequenced from diploid PCR products with primers 214F, 731F, 1742R and 1992R (Figure S9). Nonmixed sequence traces matching the A allele of *PRDM9*, indicating A/A homozygosity, were identified. We subsequently used the web-based tool Mutliple SeqDoc (Crowe 2005) to compare mixed traces with nonmixed A/A traces. This algorithm produces aligned images of a reference and a test chromatogram together with a subtracted trace showing differences between chromatograms. These difference profiles allow rapid visual identification of base substitutions, insertions and deletions in the test sequence. The differences highlighted by the algorithm are then visually checked and interpreted to avoid potential artifact calls often introduced by automatic base-calling software. This procedure allowed us to determine allele status for all individuals (Table S3 and S5, Figure S4). Most of the individuals were homozygotes A/A (64%), and all remaining individuals were heterozygotes. We identified 10 previousely-characterized alleles (B, C, D, E, L1, L3, L9, L20, L24, L14) (Baudat et al. 2010; Berg et al. 2010) and 7 novel alleles found only in this study (L32-38). The novel alleles are described in Supplementary Results.

Association Testing and Ancestry Analyses

Association between *PRDM9* alleles and ALL in the FCALL cohort was evaluated using randomization inference based on two-tailed Fisher's exact test with 10,000 permutations. For replication in the SJDALL cohort, one-tailed Fisher-based permutation tests were performed. To test whether rare alleles of *PRDM9* were over-represented in ALL subtypes, we performed a permutation test where we

permuted the 50 children from the St Jude cohort across subtypes 10,000 times and computed how many times patients with *k*-finger alleles were only seen in hypodiploid and infant B-ALL subtypes. The FC family cohort was used as control cohort for association between PRDM9 ZnF alleles and disease and HapMap3 CEU individuals were considered as controls for association between SNP rs12153202 and disease (Supplementary Results). Ancestry was studied by Principal Component Analyses of genotyped genetic variation of subjects and controls using the smartpca module from the Eigensoft package (Price et al. 2006). Detailed ancestry analyses can be found in Supplementary Results.

Genomic Motif Search

A motif search was performed to identify PRDM9 ZnF allele A and C binding sequences in the human genome. Motif search was performed on the human reference genome (hg18) and on a custom-made degenerate reference genome, constructed using biallelic SNPs in dbSNP v134 that were validated (VLD flag, set if the SNP has at least 2 minor allele counts). At each position where a SNP was reported, the nucleotide in the reference genome was replaced by the IUPAC code corresponding to allele variation. We then scan degenerate or non-degenerate genomes to search for specific degenerate motifs, denoted *A* and *C* (Table 2), representing DNA binding motifs predicted by Berg and colleagues (Berg et al. 2010) for PRDM9 ZnF allele A and C, containing only nucleotides predicted with >95% accuracy, based on the algorithm by Persikov and colleagues (Persikov et al. 2009). The sequences found were then annotated based on their starting position using a list of coordinates for regions of interest. We counted the number of non-overlapping motifs occurring whole-genome, in genes and in segmental duplications within genes. Coordinates of all annotated human genes and segmental duplications were obtained from UCSC tables.

Mapping PRDM9 binding motifs within the ALL gene list

We built an ALL gene list using an unbiased strategy based on Mitelman and dbCRID databases (Kong et al. 2011; Mitelman et al. 2011) as of July 2011. Translocations were selected only if they were reported in more than 10 entries for ALL in these databases (Table S8A). We next retreived fusion genes involved in these translocations from the databases and a final total of 38 genes were kept, following a litterature search performed to verify that they have been implicated in ALL in peer-reviewed publications (Table S8B). We computed the number of sequences matching the *A* and *C* motifs occuring within and outside of selected ALL genes. Chi-square tests are sensitive to sample size and will tend to reject the null when the sample becomes sufficiently large. Because we have huge numbers for the motif counts, we used Odds Ratios (OR) to compare the frequencies between motifs:

$$OR = \frac{p_{m=A} / (1 - p_{m=A})}{p_{m=C} / (1 - p_{m=C})}$$

with p_m the ratio between the number of motifs *m* in the ALL gene list and the total number of motifs *m* in a particular genomic dataset, such as the whole genome, genic regions (repetitive and unique DNA, Table S9) and in segmental duplication occuring in genes. This provides a measure of the strength of non-independence between the motifs and their occurrence in the ALL gene list. Confidence intervals were calculated following the procedure described in (Morris and Gardner 1988). We generated 1000 lists of randomly chosen genes, with the inclusion criteria being a gene length of 3Kb or more. The experiments described above were repeated with each random gene list in place of the ALL gene list and we computed the number of time significant OR were seen to obtain a two-tailed *p*-value. We used the program NuPoP (Xi et al. 2010) to predict nucleosome positioning within the *MLL* breakpoint cluster region (chr11:117857800-117858100, hg18).

Translocation Data

We used t(11;22)(q23;q11) translocation frequencies previously published by Berg and colleagues (Berg et al. 2010), based on *de novo* detection of translocations in sperm from men with different *PRDM9* alleles. We separated African from European men and performed a Kruskal-Wallis test to compare men carrying *k*-finger alleles and men carrying other alleles in each population. For Africans, the *k*-finger alleles (C, L4, L6, L14-19) showed no significant influence on translocation frequency (χ^2 =0.0643, p = 0.7998). However, in Europeans, men with *k*-finger alleles (C, L20) showed significantly increased translocation frequencies (χ^2 = 4.0735, p = 0.04356).

Data Access

Sequences of the *PRDM9* novel alleles have been deposited in GenBank under the accession number JQ044371-JQ044377. Genomic reads aligning to *PRDM9* exon 12 from all individuals in the FCALL and control cohorts are deposited in NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA060797. Whole-exome sequencing reads and genotyping data for the ALL quartet will be made available through the Quebec childhood leukemia web portal (http://childhoodleukemiagenomics.org). Genomic sequence data from the SJALL cohort is available through the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/) under the accession number EGAC00001000044 and can be accessed by application to the Pediatric Cancer Genome Project (PCGP) Data Access Committee. More information can be found at http://explore.pediatriccancergenomeproject.org.

SUPPLEMENTARY METHODS

1. SNPs in the ALL Quartet

In order to distinguish germline events from somatic events, each affected child from the ALL quartet was sampled twice, pre- and post-treatment. The pre-treatment samples had less than 20% tumor cells making it difficult to identify tumor-specific mutations in these ALL cases.

1.1 Exome SNP Dataset

Exome sequencing was performed as described in Material and Methods. Basic statistics are presented in Table S1A. To have the most accurate set of SNPs possible, we used two approaches that make use of inheritance patterns to call SNPs in the six samples:

Samtools dataset: Mismatches according to the hg18 human reference genome were called using Samtools and were considered as SNPs when the quality score (Phred score) is $\geq$ 20 and when the position is covered in all samples. Extra SNPs with a quality score below 20 in the offspring (pre and post-treatment) were rescued when the position is covered in each sample and the variant allele is confirmed in 30% of the reads in one parent.

Polymutt dataset: The program Polymutt implements a likelihood-based method for calling SNPs in trio data (Li and Abecasis 2011). Each of the four trios was processed independently and produced a list of genotype calls for SNPs in each individual and a mean quality score for the trio as a whole. We selected the SNPs having a mean quality score $\geq$ 20 and covered in all samples. We removed 1343 SNPs for which the genotype calls were inconsistent for the parents between the different trio analyses.

A total of 30,360 and 47,243 SNPs were inferred by the Samtools and Polymutt approaches, respectively. A total of 28,747 SNPs, called by both methods, were retained for further analysis. Genotypes called by Polymutt at these positions were

selected for the six samples, because the genotype mendelian error rate for the Samtools dataset (9115 errors, mendelian error rate of 7.93%) was higher than for the Polymutt dataset (511 errors, mendelian error rate of 0.44%). Finally, we removed 4799 SNPs, homozygous with non-reference alleles in all six samples, and obtained a final exome SNP dataset of 23,437 polymorphic positions.

1.2 Genotyping SNP Datasets

The ALL quartet samples were genotyped on two different platforms: the Illumina HumanOmni 2.5-quad BeadChips and the Affymetrix 6.0 SNP arrays.

For the Illumina Omni 2.5 array, 2,383,178 SNPs passed standard quality control filters. The parents were genotyped twice, with a concordance rate of 99,987%. We removed 617 discordant SNPs between duplicated parents, 1339 SNPs with Mendelian errors, 21,532 SNPs with missing data and 1,552,225 SNPs homozygous in all samples. We obtained a final Illumina SNP dataset of 808,082 polymorphic markers.

For the Affymetrix 6.0 array, we obtained 909,515 SNPs after applying standard quality control filters. We subsequently removed 2447 duplicated SNPs, 4403 SNPs with Mendelian errors and 110,423 SNPs with missing data. The final Affymetrix SNP dataset contained a total of 792,242 markers.

1.3 Recombination SNP Dataset

To obtain a complete recombination SNP dataset for the ALL quartet, that allows a finer detection of recombination events in coding regions, we merged the exome SNP dataset and the Illumina SNP dataset. Among Illumina Omni2.5 SNPs that passed standard quality control filters, 16,839 SNPs are positioned in Agilent SureSelect targeted regions used in the exome sequencing experiment. Among these, we removed 161 positions detected as polymorphic in the exome sequencing data but homozygous for all samples in the genotyping data. For 3347 polymorphic Illumina SNPs that were not detected in the exome sequencing data, most of which

went undetected because the positions were not covered in all samples in the exome data, genotype calls from the Illumina SNP dataset were kept. For the remaining 13,331 coincident positions, the genotype concordance rate was 0.976. We removed 3 SNPs that had different alleles in the two datasets and 1309 SNPs that had the same alleles but at least one sample with discordant genotypes. Combining the concordant SNPs with the remaining SNPs from both datasets, we obtained a final recombination SNP dataset of 816,715 for the ALL quartet (Table S1B).

2. *De novo* mutation discovery in the ALL quartet

2.1 DND Software (Conrad et al. 2011; Cartwright et al. 2012)

To discover *de novo* point mutation in sequencing data, we used a probabilistic approach DND developed by our group (Conrad et al. 2011; Cartwright et al. 2012). The method uses the relatedness between individuals in a pedigree to produce the posterior probability of *de novo* mutation at each genomic site, given the error rate of the sequencing technology, the somatic mutation rate and the population mutation rate for the population from which the samples were drawn.

DND was run on four trios: [M, F, 383R1]; [M, F, 383R2]; [M, F, 610R1]; [M, F, 610R2], where M is the mother, F is the father, R1 denotes pre-treatment samples and R2 denotes post-treatment samples for the two brothers, patients 383 and 610. We used a sequencing error rate of $\varepsilon = 0.005$, a somatic mutation rate of $\mu = 2\times10^{-7}$ and a population mutation rate of $\theta = 0.001$, and examined all sites within targeted regions covered by at least 15 reads in samples from each trios. Simulations revealed that highest confident calls require a read depth of at least 15 in both parents and the non-mutant allele (Cartwright et al. 2012). To infer a germline *de novo* mutation in a child, we considered all positions that had a probability of *de novo* > 0.90 in at least one of the two samples (R1 or R2), with the same alternative allele present in at least one read in the other sample. We found 27 such candidates in 383 and 15 in 610. We removed candidates for which at least two reads in the

parents showed the alternative allele: previous validation experiments showed that such *de novo* candidates are very likely to be false positives. Similarly, we removed candidates where at least two reads with the same alternative allele was seen in the sibling. This filter was applied if the children shared at least one parental chromosome at that locus (Figure S10). Although these candidate de novo mutations could reflect mutations that happened in premeiotic divisions in germinal stem cells, it is more likely that the allele was not sampled in the sequencing experiment for that parent or resulted from similar sequencing or mapping errors in both children. Unfortunately, although 7 passed the parental filter, none of the *de novo* candidates passed the sibling filter.

2.2 Sample-independent approach

Using the DND software with stringent parameters, we only interrogated regions covered at more than 15X in the parents, which in this case corresponds to ~12Mb of the human exome. To investigate other positions targeted by the sequencing experiments, we used SNPs from the *samtools dataset* called via a sample-independent strategy. We selected positions with at least 8X coverage in the parents (25,286,190 bp) and considered all Mendelian errors. To be called a *de novo* candidate, the variant allele must be seen at least 5 times in one of the child samples, twice in the other sample from the same sibling and must not be seen at all in the parental samples or in the other sibling. We identified one such mutation in patient 383 (Figure S1), located on chromosome 5 at position 64,860,544 (hg18). At this locus, the two brothers copied the same paternal chromosome and different maternal chromosomes (Figure S10). If we assumed a binomial distribution with a probability of 0.5 of sequencing the variant allele at a heterozygous position, the probability that this variant was transmitted from the father and was not sampled in neither the father nor patient 610 (total coverage of 27X) is $p = 7.45\times10^{-9}$. On the other hand, the probability that this variant was transmitted from the mother to 383 (coverage of 10X) is $p = 9.77\times10^{-3}$.

SUPPLEMENTARY RESULTS

1. Patients' Karyotype

The brothers from the ALL quartet family were both diagnosed, within a 3 years period of time, with B-cell precursor childhood ALL with FAB-L1 morphology at Sainte-Justine Hospital, Montreal, Canada, at the age of 2 for patient 383 and subsequently, at 14 years of age for patient 610. At diagnosis, both siblings showed hyperdiploid leukemia clones. Cytogenetic analyses were performed using standard procedures. G-banded chromosomal analysis for patient 610 revealed clones with the following karyotypic features: 55XY,+X,?del(2q),+5,+8,+10,+14,+17,+18,+21, +21[5]/46XY[21]. By employing the Illumina Omni 2.5 genotyping data, we also detected the following additional chromosomes in patient 610: +4,+?Y. For patient 383, clones were detected with karyotype: 5153,XY,inv(2)(p11q13),i(17)(q10),+4,+6, +?12,+15,+17,+18,+21[9]/46,XY,inv(2)(p11q13)[23]. The additional chromosomes 17, 18 and 21, shared between siblings' leukemic clones, are frequently gained in ALL hyperdiploidy. Chromosome 2 pericentric inversion in patient 383 is also carried by the mother and occurs at a higher frequency in African Americans compared to individuals of European descent (Phillips 1978). This aberration is not associated with a specific syndrome and no abnormal phenotype has been described (Srebniak et al. 2004). However, such inversions may lead to recombinants gametes with abnormal karyotes, through crossing over between the normal and inverted homologues.

2. Fine-scale Dissection of Recombination Events

Recombination analyses were performed separately for the pre-treatment and post-treatment samples, forming two quartets that will be referred to herein as quartet Q1 (pre-treatment) and quartet Q2 (post-treatment). The algorithm used to call recombination events (Hussin et al. 2011) first looks for errors, i.e markers that create double recombinants (Material and Methods). This procedure identified 1214

errors (433 in Q1, 456 in Q2 and 325 in Q1 and Q2). A total of 222 errors came from SNPs from the Illumina SNP dataset and 992 came from the exome SNP dataset. The highest density of sequencing errors detected by the recombination algorithm is located in chromosome 6, between 29 and 34 Mb, in the the complex region of the human leukocyte antigen (HLA) system, very likely caused by mapping errors, given that assembly of this locus using single-end short reads data is difficult.

The algorithm identified a total of 236 switches in the quartet Q1 and 224 in the quartet Q2. Switches not shared between quartets Q1 and Q2 were separated from their closest neighbouring event by at most 5 markers and are very likely to be caused by calling or alignment errors. All shared switches are presented in Figure S10. From this list, switches were called as crossovers only if they were separated from their closest neighbour by more than 2 informative markers. They were divided into two categories: single crossovers, when the nearest crossover is at more than 50Kb and double crossovers, when two crossovers were found within 50Kb of each other (Table S2, Figure 1).

3. Putative Gene Conversion Events

A region on chromosome 16 was intriguing because both parents showed double crossovers at exactly the same locus (separated by 4 and 8 markers in the father and mother, respectively). Genotyping and sequencing markers both supported the double crossovers. Although these double crossovers could reflect gene conversion events, it is unlikely that such events happened in both parents in the same region. A most likely explanation is that this region is not unique in the genome and that the detected genotypes tag polymorphisms positioned somewhere else in the genome. This is probably the case since these double recombinants occur in gene *HYDIN*, a gene that has been duplicated very recently, with a nearly identical 360-kb paralogous segment inserted on chromosome 1q21.1 (Doggett et al. 2006). We therefore removed these double crossovers from our final list of paternal and maternal recombination events.

We identified nine other short regions (<50Kb) flanked by recombination events (Table S2) having no known paralogous segment. All regions comprise genotyping SNPs, they are not resulting from errors in mapping of sequencing reads. These double crossovers could reflect meiotic gene conversions in the same meiosis or recombination events occuring in the same genomic region in the two brothers, since we can not distinguish to which child each event belongs. However, four maternal double crossovers on chromosome 4, 8 and 17 (as well as four additional double crossovers supported by only 2 markers on chromosomes 7, 8, 10 and 18 – see Figure S10) occurred within 1 Mb of another clearly defined recombination event. This would necessarily mean that, in one of the two children, at least two recombination events happened closeby. Under the model of crossover interference, however, the presence of one crossover event in a region reduces significantly the possibility of a second event nearby in the same individual. Therefore, these small double crossovers may reflect unique patterns of recombination or gene conversion, not yet identified in humans.

4. *PRDM9* alleles in the ALL Quartet

Because of the reduced proportion of recombination events overlapping with population hotspots observed in the mother of the ALL quartet, we investigated whether variants in the PRDM9 coding region were identifiable based on the read data. We detected two SNPs in the ZnF array domain of exon 11 of *PRDM9*, both present in dbSNP v134 (rs74710141 and rs77287813). SNP rs74710141 corresponds to the known C/G substitution in the sixth ZnF repeat, the only difference between PRDM9 allele A and B. It appeared that the hg18 reference allele is the B allele. SNP rs77287813 corresponds to a A/C substitution in the tenth finger, corresponding to the only difference existing between ZnF type *h* and *k*. Due to the structure of the PRDM9 ZnF array, the presence of this SNP is likely to reflect a supplementary repeat instead of a point mutation in the H ZnF repeat.

To infer which *PRDM9* ZnF allele was carried by the mother, we identified the ZnF types present in the read data for the mother (Material and Methods). Out of 1477 reads that mapped to the *PRDM9* ZnF array, 26 (1.76%) aligned specifically to the *k* ZnF type and 8 reads aligned specifically to the *l* ZnF type, which corresponds to 0.54% of the read data mapped in the array, a value below our inference criteria of 1% (Material and Methods). However, the *l* ZnF type, which is one mismatch away from the *e* ZnF type, is two mismatches away from its closest match in the reference: the ZnF type *c* of the *b* allele. This is expected to hamper the mapping of reads sampling the *l* ZnF repeat (for example in cases where there is an error at the end of the read). The two brothers did not copy the same maternal chromosome in this region of chromosome 5. Child 610 had 4417 reads mapping to the PRDM9 ZnF array and very few (a total of 6) aligning to either the *k* or *l* ZnF type. On the other hand, child 383 had 59/2625 reads (2.25%) and 18/2625 reads (0.69%) aligning to the *k* and *l* ZnF type, respectively. The unusual patterns of recombination in the mother and the occurrence of 85 and 26 reads sampling the *k* and *l* ZnF type on one maternal chromosome suggested the presence of the C allele in the mother.

Because SOLiD single-end short reads (50 bp) will not overlap a full ZnF repeat, which is 84 bp long, our data does not allow to determine the order of repeats or insertions in the ZnF array. The A/C genotype in the mother was thus validated by Sanger re-sequencing of the ZnF array (Material and Methods).

5. Description of *PRDM9* alleles and Novel ZnF Types

Labeling of the *PRDM9* zinc finger (ZnF) alleles and repeat types follows that of Berg and colleagues (Berg et al. 2010), but to differentiate “allele” from “finger” nomenclature, alleles are in uppercase and fingers are written in lowercase italic characters (for example, allele A has 13 repeats type and is coded: *abcddecfghfij*). The ZnF repeat types *a* to *x* are presented in Figure S6.

In the 22 parental trios of the FCALL cohort, we detected 11 parents for which read data show evidence for the *k* and/or *l* fingers. For 2 additional parents, the *p* and *t*

fingers were detected (data not shown), suggesting the presence of ZnF allele L24. We performed Sanger sequencing of the ZnF array for 12 families, which included 9 of the 11 parents with *k, l* or *p* and *t* fingers. Sanger sequencing experiments confirmed these 9 *PRDM9* alternative alleles and revealed the presence of undetected rare alleles in other families: the allele L3 and two alleles not reported in previous studies, L30 and L31 (Table S3). Repeat structure for L30 is *abcddecfghfqj* and for L31, *abcddecughfqj*, with repeat types described in Figure S6. Sanger sequencing of PRDM9 ZnF array was also performed for 76 French-Canadian parents from the FC family cohort and 27 Moroccans (Table S5, Fig S4). We discovered 5 novel alleles in this data denoted L32 to L36. The newly discovered allele L30 was seen three times in Moroccans. The repeat structure for the novel alleles are: L32=*abcvdecfghfij*, L33=*abcdddecfghfij*, L34=*abcddecfghfwj*, L35=*abcddecxghfij* and L36=*abrddecfghfij*, with repeat types described in Figure S6.

In total, we sequenced 258 *PRDM9* ZnF alleles and identified four novel ZnF types: *u*, *v*, *w* and *x* (Figure S6):

The *u* repeat type is a mutated *f* repeat type, encoding a missense change at a well-conserved position of the repeat sequence: GAG (E) → AAG (K). Only one synonymous mutation was previously observed at this codon (repeat type O, GAG (E) → GAC (E)). This change is not located within the binding unit of the ZnF repeat but it is predicted by SIFT (Ng and Henikoff 2003) to have a damaging effect on the resulting ZnF array (SIFT score = 0). This repeat type was found in the PRDM9 array from one ALL parent.

The *v* repeat type is a mutated *d* repeat type, encoding a missense change, TAT (Y) → TTT (F), at a position where only synonymous changes were previously observed (TAT (Y) → TAC (Y) in repeat types *a*, *i*, *j* and *m*). This change is not located within the binding unit of the ZnF repeat and is predicted as tolerated (SIFT score = 0.06). This repeat type was found in the PRDM9 array of two individual of the FC cohort.

The *w* repeat type is a mutated *i* repeat type, encoding a missense change, CGC (R) → CTC (L), at a position not particularly conserved between repeat types. This change is not near the binding unit of the ZnF repeat and is predicted as tolerated (SIFT score = 0.14). This repeat type was found in the PRDM9 array of one individual of the FC cohort.

The *x* repeat type is a mutated *f* repeat type, encoding a synonymous change (AAG (K) → AAA (K)). This repeat type was found in the *PRDM9* array of one individual in the Moroccan cohort.

6. Ancestry Analyses

6.1 ALL quartet

The parents of the ALL quartet are reported to be distant cousins of Moroccan ancestry. To verify their ethnicity and select the appropriate set of controls to use in our analyses, we performed principal components analyses (PCA) of the genetic variation using the smartpca and twstats modules from the Eigensoft package (Price et al. 2006). We first performed a PCA using genotyping data from the parents of the ALL quartet and 163 unrelated individuals from a Moroccan cohort (Idaghdour et al. 2010). We selected 27 Moroccan individuals among these with the closest eigenvalues to the 2 first significant PCs (Figure S3A), where DNA were available. We performed a second PCA including 28 unrelated French-Canadians individuals from the FC family cohort (Figure S3B). The results suggest that the parents of the ALL quartet have some Arab ancestry but are genetically closer to the French-Canadians than the Moroccans, although they also are outlier individuals relative to the French-Canadian cluster.

6.2 FCALL cohort and FC family cohort

We performed a PCA of genetic variation to study genetic ancestry of the parents from the FCALL cohort (parents of patients). We include parents from the FC family cohort (controls) and European and African individuals from the HGDP dataset using

positions of SNPs in common between exome sequencing variant found in the FCALL parents, genotyped SNPs in HGDP populations and the Affymetrix 6.0 array used to genotype individuals in the FC family cohort. The analysis demonstrate that French-Canadian parents of patients and controls cluster together with the French HGDP individuals, although the FCALL parents do not exactly overlap with the other two groups on the plot, likely because of the small differences in allele frequencies driven by the different technologies used to type genetic variation (exome SNP calling vs genotyping). In any case, the FCALL parents do not show a higher contribution of African genetic background than individuals from the FC family cohort. Therefore, differences in frequencies of PRDM9 ZnF allele C cannot be explained by a higher African ancestry in individuals from the FCALL cohort.

6.3 St Jude ALL cohort

The entire B-ALL St Jude cohort includs 61 B-ALL patients for which reported ethnicities were available. We verified these ethnicities by performing a PCA of the patients' genotyping variation genome-wide and removed from subsequent analyses children with an African genetic ancestry, for which we do not have controls. The PCA confirmed the reported ethnicities for most patients, however five individuals reported as "White" are potentially admixed (Figure S8). From the individuals that were reported as "Other", two individuals are likely mixed (black/white), one is likely Asian and one is likely Hispanic or Native American. Fifty patients showed no African component and were included in our analyses, with 39 children clustering with the French-Canadian controls. Therefore, association testing was performed with and without the 11 children with Hispanic, Asian or Native American ancestry.

SUPPLEMENTARY FIGURES AND TABLES

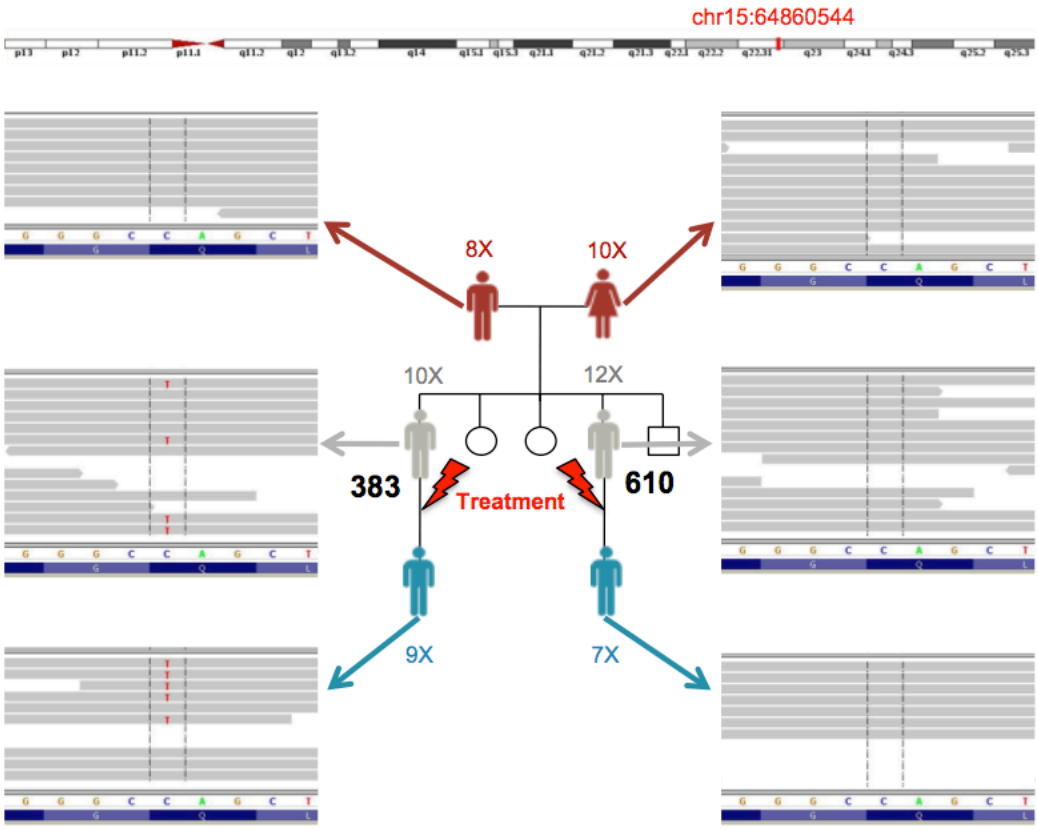


Figure S1. Identification of a *de novo* mutation in *SMAD6* on chromosome 15.

The variant allele is evenly sampled in patient 383 samples but is not seen in any of the other samples. Given the data, the probability that this mutation was inherited is $p = 9.77 \times 10^{-3}$ (Supplementary Methods).

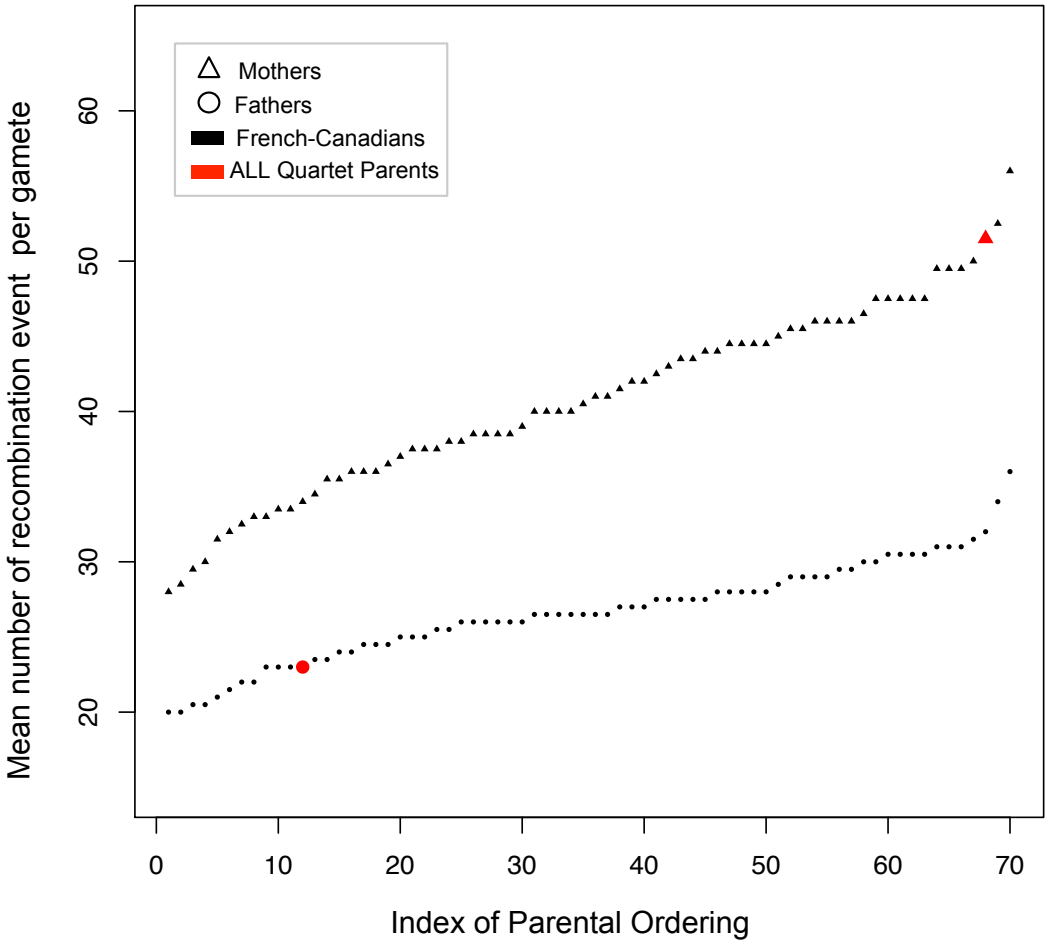


Figure S2. Mean recombination rate in the parents from the ALL quartet and the FC cohort.

Recombination events were called using genotyping data from the Affymetrix 6.0 array. The FC cohort is composed of 69 French-Canadian quartets. For the ALL quartet, only children's post-treatment samples were used to infer recombination. Mothers (triangles) and fathers (circles) are ordered according to their recombination rate. Genetic markers from Affymetrix 6.0 platform were used for all individuals, including the ALL quartet parents.

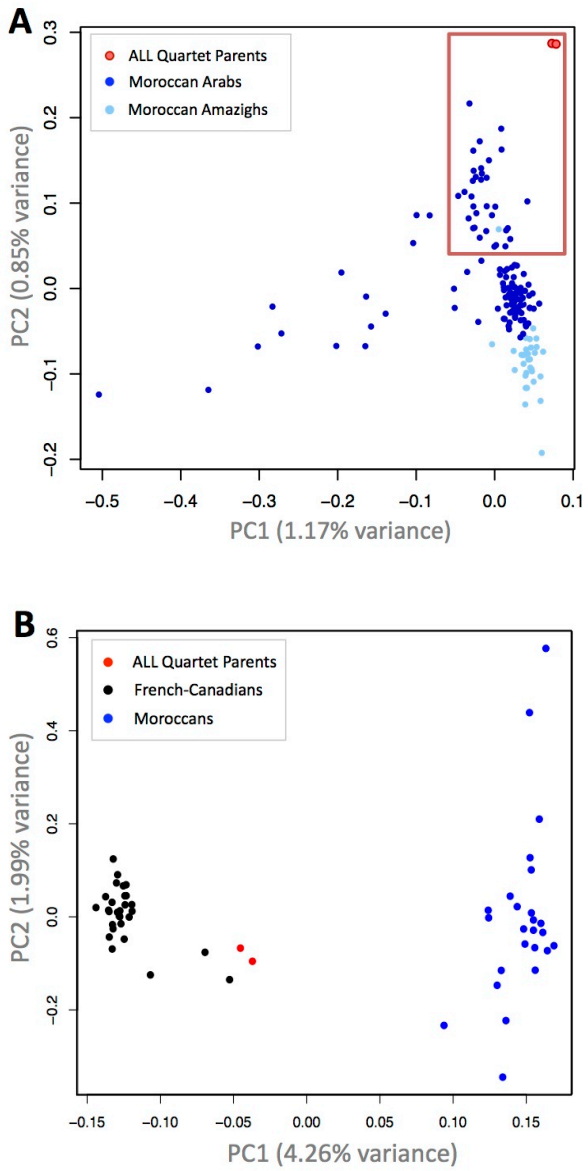


Figure S3. Genetic Ancestry of the ALL Quartet Parents.

We preformed a Principal Component Analysis of genetic variation using 61,454 SNPs from (A) the ALL quartet parents and 163 unrelated Moroccans; (B) the ALL quartet parents, the 27 unrelated Moroccans, chosen to be the closest to the ALL quartet parents (falling in the red rectangle in panel A), and 28 unrelated French-Canadians (Supplementary Results). The ALL quartet parents are closer to Moroccans of Arab ancestry then to Moroccan Berbers (Amazighs).

Individual ID	Gender	Alleles
A55	Female	A/A
A51	Female	A/A
A41	Female	A/L35
A109	Female	A/A
B04	Female	L36/L14
B08	Female	A/L20
I06	Female	A/A
I54	Female	A/A
I59	Female	A/A
I71	Female	A/A
I62	Female	A/A
I66	Female	B/L20
A60	Female	A/E
I09	Female	A/A
I28	Female	A/A
A68	Female	A/A
A65	Female	A/L37
I52	Female	A/A
A46	Male	A/A
A32	Male	A/A
A145	Male	A/A
A147	Male	A/A
A150	Male	A/A
A117	Male	L37/L20
A138	Male	A/L37
I19	Male	A/L37
A112	Male	A/A

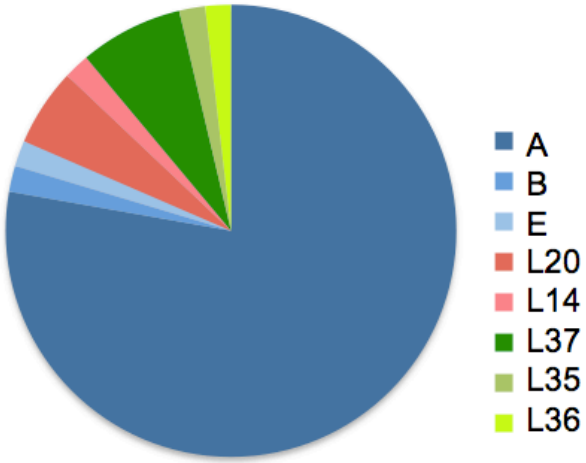


Figure S4. PRDM9 ZnF alleles in 27 unrelated Moroccan individuals.

Individuals were selected from the Moroccan cohort (Idaghdour et al. 2010) based on ancestry, chosen to be the closest to the ALL quartet parents (Figure S3). PRDM9 ZnF alleles were assayed by Sanger sequencing. Alleles A, B, E, L14 and L20 are described previously (Berg et al. 2010). Alleles L30, L35 and L36 are novel alleles, described in Supplementary Results.

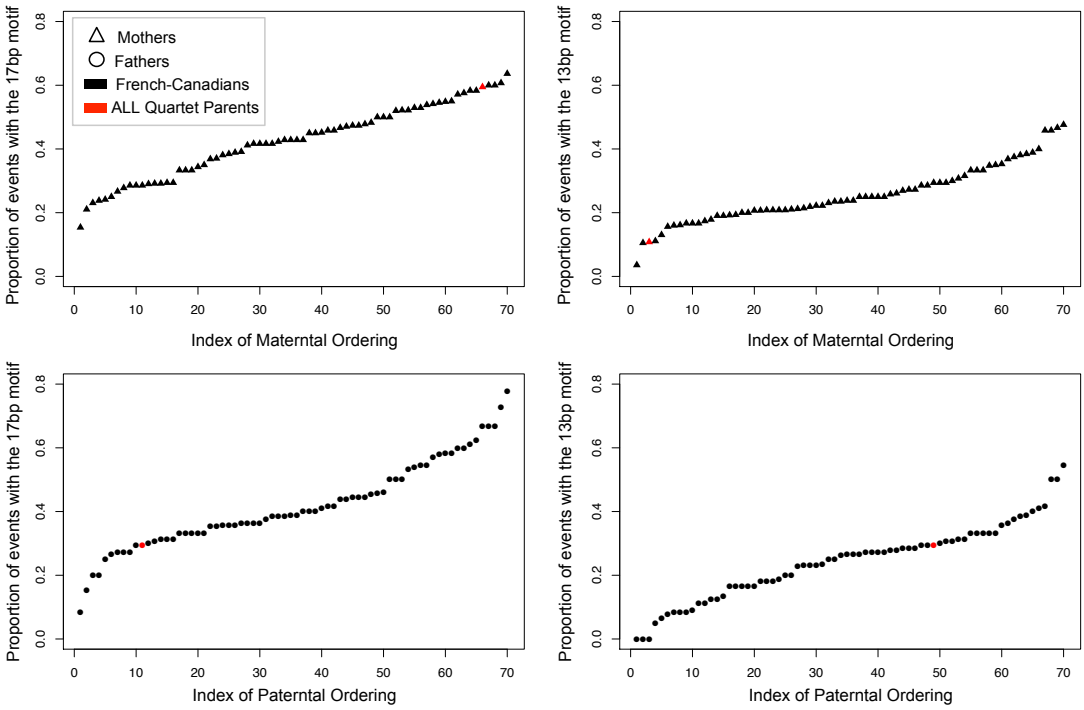


Figure S5. Proportion of recombination events called near *PRDM9* binding motifs.

We computed the proportion of recombination events overlapping the 17-bp and 13-bp predicted to be recognized by the C and A alleles of *PRDM9*, respectively, for parents of the ALL quartet, compared to parents from 69 French-Canadian quartets. The 13-bp motif, CCNCCNTNNCCNC is enriched in linkage disequilibrium-based hotspots inferred from HapMap data, whereas a 17-bp motif, CCNCNNTNNNCNNNNCC, is associated with African-enriched hotspots. Recombination events were called using genetic markers from Affymetrix 6.0 platform for all individuals, including the ALL quartet parents for which only children's post-treatment samples were used to call recombination events. Mothers (triangles) and fathers (circles) are ordered according to the proportion of motifs found near their recombination events.

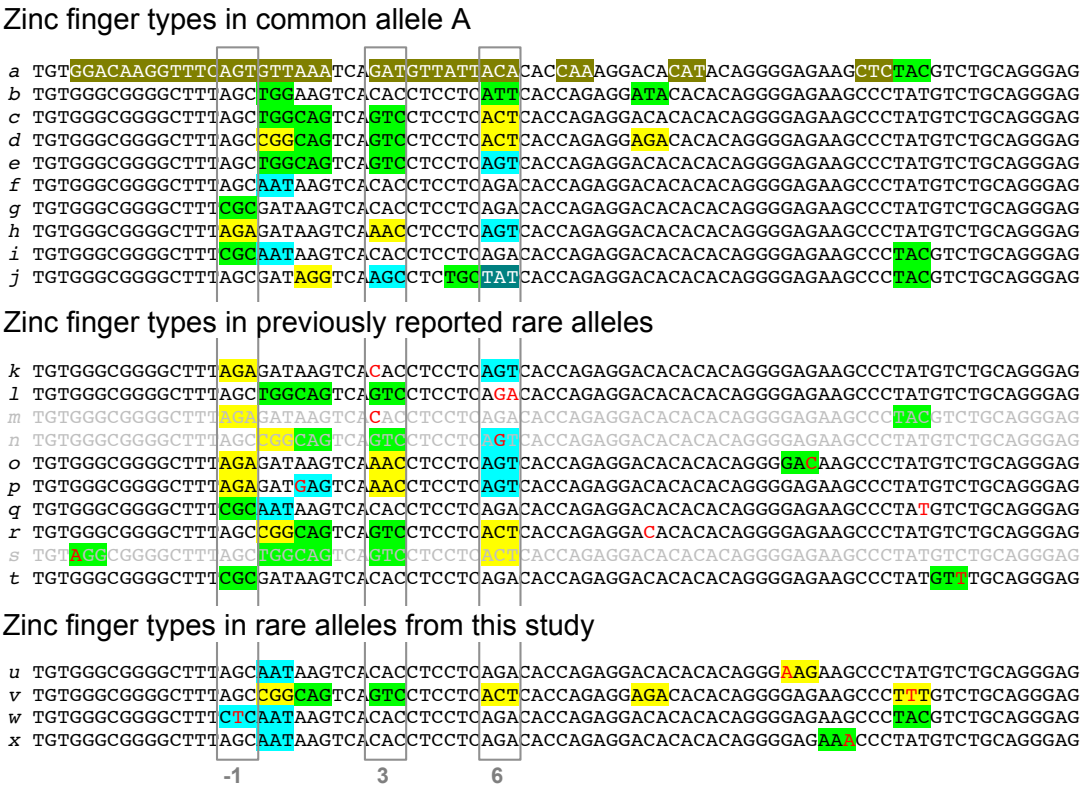


Figure S6. ZnF repeat types of PRDM9 alleles.

Modified from (Berg et al. 2010). Variable codons are colored. Red bases in ZnF repeat types from K to X are differences with respect to the closest repeat type present in the hg18 reference genome (note that the hg18 reference genome does not encode the E finger and therefore, the L finger is two differences away from the C repeat type). Repeat types M, N and S (Baudat et al. 2010; Berg et al. 2010) were not found in this study. The 3 codons coding for the binding unit of each repeat (positions -1, 3 and 6 of the ZnF alpha helices) are indicated.

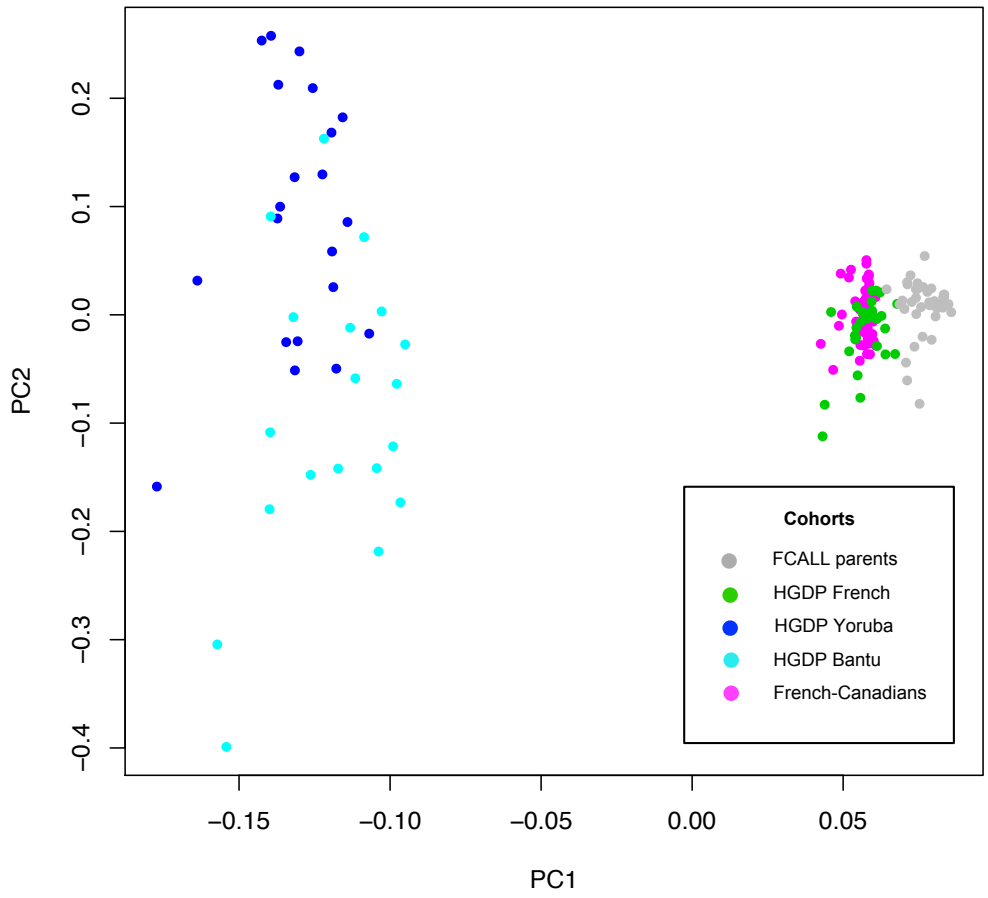


Figure S7. Genetic ancestry of parents from the FCALL cohort.

We performed a Principal Component Analysis of genetic variation using 358 SNPs in common between exome sequencing SNPs from the FCALL parents and genotyped SNPs from HGDP populations and the FC family cohort.

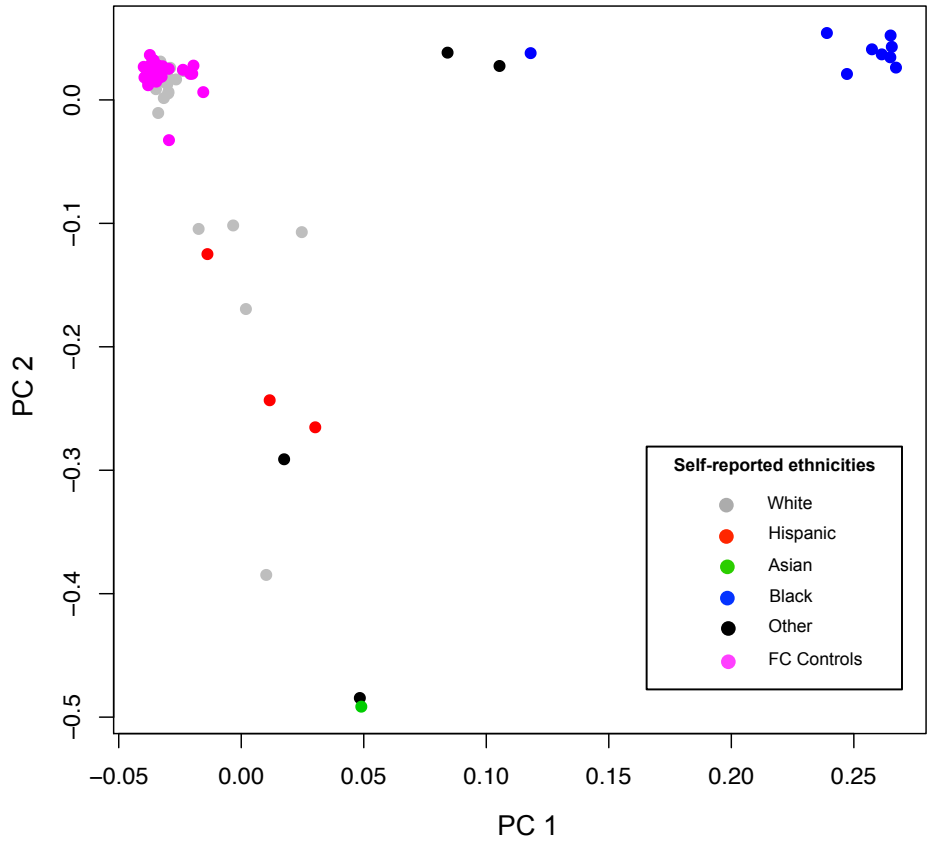


Figure S8. Genetic ancestry of SJDALL patients.

We preformed a Principal Component Analysis of genetic variation using 201 474 SNPs from 61 St Jude patients and 76 French-Canadians controls. The first Principal Component (PC1) separates individuals of European descent from individuals from African descent: the 11 individuals showing African ancestry (PC1> 0.05) were removed from analyses.

Primer name	Sequence
HsPrdm9-F3	TGTAAGGAATGACACTGCCCTGA
HsPrdm9-R1	ATGTCCCCCGAACACTTACAGAA
214F	TGATTGTTTCTTCATTTGATCTTCA
731F	TGGAGAGTGTGGACAAGGTTT
1742R	AGCAGAGGCTTGACCTATCG
1992R	GTCATGAAAGTGGCGGATTT

Figure S9. PCR primers used for amplifying and sequencing *PRDM9* ZnF alleles.

The ZnF array in exon 11 of *PRDM9* was amplified as described in (Baudat et al. 2010). Sanger sequencing was performed with primers 214F, 731F, 1742R and 1992R.

Figure S10. Chromosomal crossover breakpoints and shared haplotypes in the ALL quartet.

Graphical view of all paternal (blue) and maternal (red) crossover breakpoints that occurred in affected children inferred based on exome and genotyping SNPs (Supplementary Methods). When the lines are a part, the two brothers copied different parental chromosomes and when they are close together, they copied the same parental chromosome and share the same haplotype. White spaces represent regions were no informative markers were available. Small dots between parental tracks represent single markers that caused double crossover events and are likely to be SNP calling errors (Hussin et al. 2011). For small double crossovers, occurring within ≤50Kb and resulting from more than one marker, we indicated the number of informative markers separating them. Small double recombinants are likely to be false positive or reflect gene conversion events (see Supplementary Methods for details on checks performed in order to control for false positive breakpoints). The y-axis shows the position on the chromosome in tens of Mb. (*next page*)

Figure S10 (*continued*)

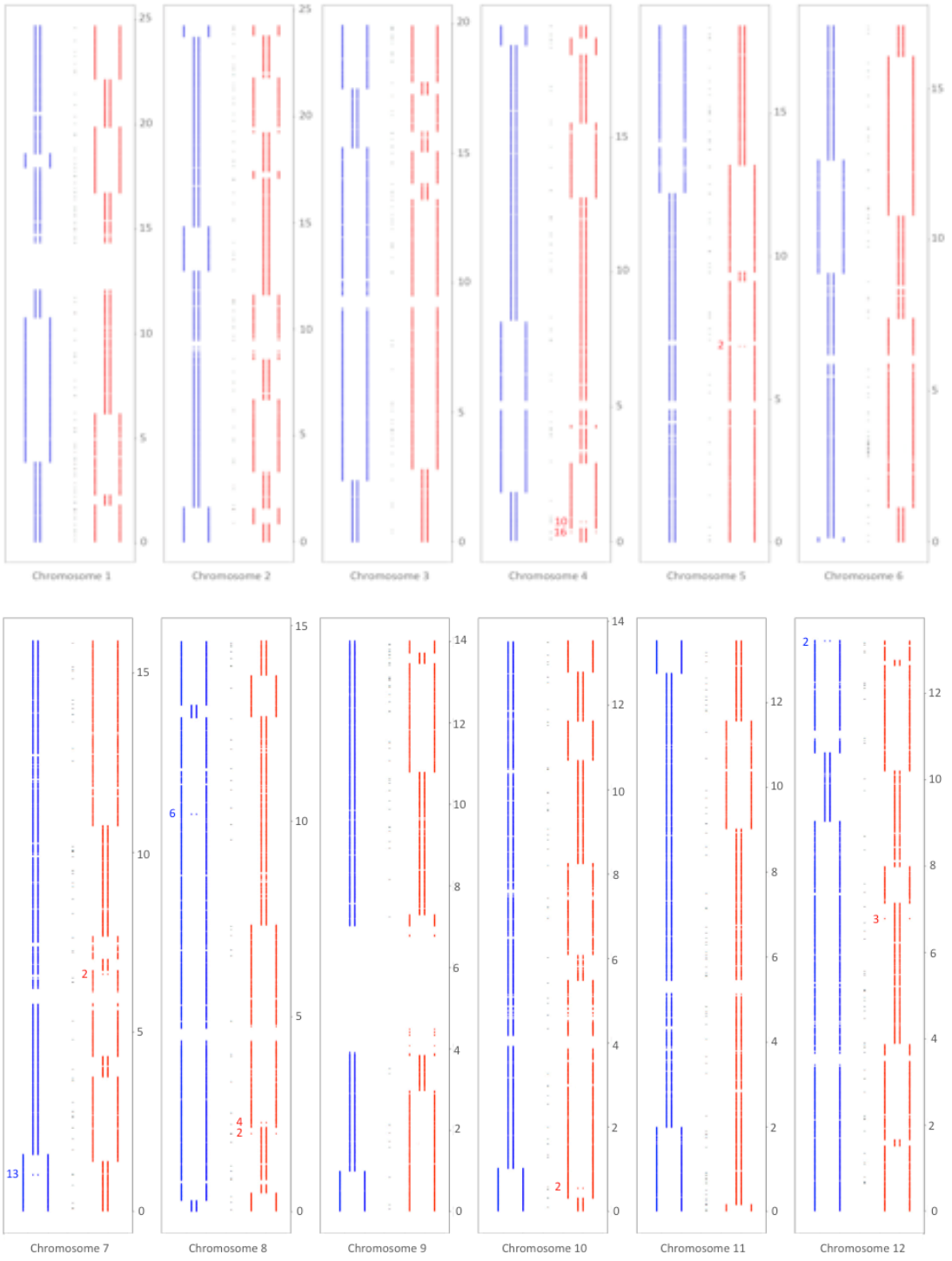


Figure S10 (*continued*)

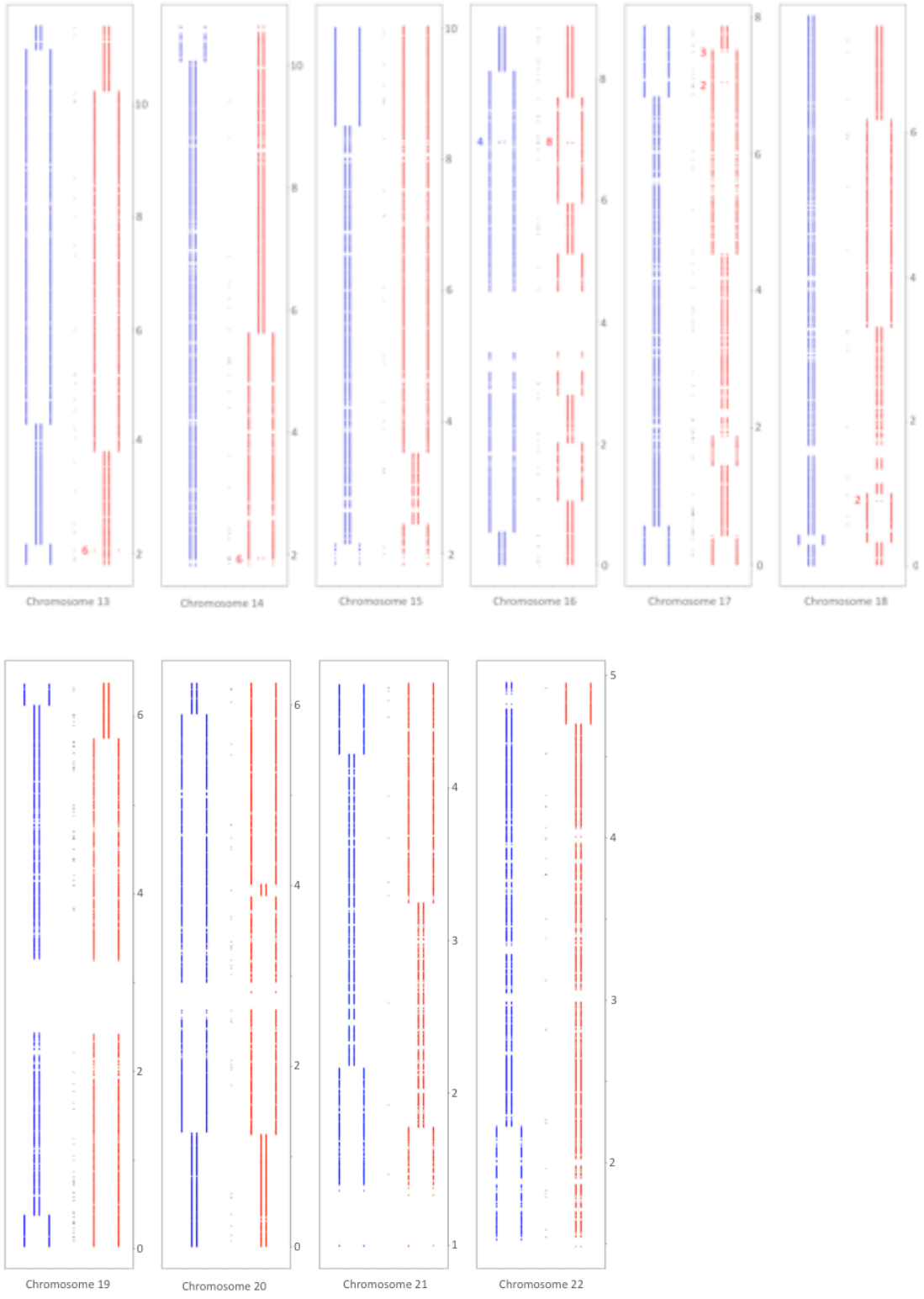


Table S1. Coverage and SNPs statistics in the ALL quartet.

A

Sample	Bioscope assembly		Exome statistic on coverage				
	Total number of mappable reads	% of reads aligned	% of reads aligned exome	% Exome coverage	Mean coverage	Average Base Quality	Average Mapping Quality
383 R1	52 802 074	77,49	62,27%	94,63%	**45,96**	28,63	76,81
383 R2	57 835 258	82,11	62,96%	96,46%	**49,93**	29,31	78,61
610 R1	52 729 534	83,07	63,83%	96,99%	**45,89**	29,40	78,82
610 R2	48 659 235	76,98	60,40%	96,84%	**40,14**	28,88	75,05
M	65 453 411	85,72	57,90%	96,73%	**51,81**	28,94	79,36
F	59 254 618	81,74	61,21%	96,02%	**49,95**	28,95	76,46

B

Number of SNPs	All samples	383 R1		383 R2		610 R1		610 R2	
		Hom	Het	Hom	Het	Hom	Het	Hom	Het
Total	816 715	366 234	450 481	366 137	450 578	370 994	445 721	371 094	445 621
Exome Sequencing	9 945	4 038	5 907	3 929	6 016	3 841	6 104	3 949	5 996
Genotyping	794 751	356 905	437 846	356 917	437 834	362 002	432 749	361 994	432 757
Overlap	12 019	5 291	6 728	5 291	6 728	5 151	6 868	5 151	6 868

Number of SNPs	All samples	Mother		Father	
		Hom	Het	Hom	Het
Total	816 715	363 805	452 910	368 133	448 582
Exome Sequencing	9 945	3 962	5 983	4 204	5 741
Genotyping Only	794 751	354 363	440 388	358 297	436 454
Overlap	12 019	5 480	6 539	5 632	6 387

R1 and R2 are the two somatic tissues sampled from both brothers. The exome is defined by the regions targeted by *Agilent SureSelect All Exon kit* covering 37,806,033 bp (1,22% of the human genome).

Table S2. Number of maternal and paternal recombination events per chromosome.

Chr	Total number of markers	Maternal events							Paternal events						
		Informative markers	Quartet Q1		Quartet Q2		Shared k>2		Informative markers	Quartet Q1		Quartet Q2		Shared k>2	
			k>1	k>2	k>1	k>2	Single	Double		k>1	k>2	k>1	k>2	Single	Double
1	61451	22517	6	6	8	6	**6**	**0**	20368	4	4	10	6	**4**	**0**
2	66882	23826	11	11	13	11	**11**	**0**	22667	6	4	4	4	**4**	**0**
3	55477	19146	7	7	9	7	**7**	**0**	18703	5	3	3	3	**3**	**0**
4	52750	18497	14	14	14	14	**8**	**2**	18151	3	3	3	3	**3**	**0**
5	50401	17975	5	3	5	3	**3**	**0**	16910	1	1	1	1	**1**	**0**
6	51056	18760	10	6	8	4	**4**	**0**	17441	7	3	3	3	**3**	**0**
7	45728	16101	9	7	9	7	**7**	**0**	15891	5	3	5	3	**1**	**1**
8	43517	14836	9	7	9	7	**5**	**1**	15993	7	5	9	5	**3**	**1**
9	37351	12838	8	6	8	6	**6**	**0**	13168	3	1	1	1	**1**	**0**
10	43005	14528	11	7	11	7	**7**	**0**	15546	1	1	3	1	**1**	**0**
11	40778	13952	3	3	6	3	**3**	**0**	13738	2	2	4	2	**2**	**0**
12	39814	13785	10	10	12	10	**8**	**1**	13674	6	2	4	2	**2**	**0**
13	30103	9605	4	4	4	4	**2**	**1**	11324	3	3	3	3	**3**	**0**
14	28292	10643	3	3	3	3	**1**	**1**	9107	3	1	1	1	**1**	**0**
15	26049	8928	2	2	2	2	**2**	**0**	8984	2	2	2	2	**2**	**0**
16	28506	9827	10	8	8	8	**6**	**(1)**	9602	6	4	8	4	**2**	**(1)**
17	24570	8149	11	7	9	7	**5**	**1**	8804	4	2	10	2	**2**	**0**
18	23846	8080	6	4	6	4	**4**	**0**	8173	2	2	2	2	**2**	**0**
19	19424	6459	1	1	1	1	**1**	**0**	6665	2	2	4	2	**2**	**0**
20	21582	7930	3	3	3	3	**3**	**0**	7241	2	2	2	2	**2**	**0**
21	12137	3917	3	2	2	2	**2**	**0**	4379	2	2	2	2	**2**	**0**
22	13970	5164	1	1	1	1	**1**	**0**	4632	1	1	1	1	**1**	**0**
Total	816689	285463	147	122	151	120	**102**	**7**	281161	77	53	85	55	**47**	**2**

Quartet Q1 (parents + children's post-treatment samples) and Q2 (parents + children's pre-treatment samples) were analysed separately. Parameter k is the number of informative markers that separate any two consecutive recombination events. Crossovers shared between quartets are considered to be real recombination events, separated into two categories: single crossovers, if the nearest neighbour is within >50Kb, and double crossovers, if the nearest neighbour is within ≤50Kb (Supplementary Methods). Double crossovers found in chromosome 16 in both parents were likely to be artefacts resulting from the HYDIN duplicated gene and were ignored.

Table S3. *PRDM9* alleles in the ALL quartet and 12 ALL trios based on read data and re-sequencing.

Family	Individual	Coverage	Proportion of Exome Sequencing Reads Aligning to PRDM9 ZnF Repeats												Validation[a]
			a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	
Quartet	F	67X	0.0103	0.0304	0.0556	0.0950	0.0274	0.0697	0.0215	0.0438	0.0222	0.0244	0.0014	0.0007	A/A
	M	73X	0.0121	0.0345	0.0697	0.0941	0.0128	0.0724	0.0270	0.0433	0.0291	0.0264	0.0176	0.0054	A/C
	610N	66X	0.0273	0.0287	0.0378	0.0570	0.0178	0.0516	0.0255	0.0278	0.0208	0.0144	0.0011	0.0002	-
	383N	71X	0.0502	0.0499	0.0632	0.096	0.0121	0.0780	0.0300	0.0491	0.0331	0.0236	0.0224	0.0069	-
375	F	11X	0.025	0.0607	0.0142	0.0392	0.0071	0.0821	0.0357	0.025	0.0107	0.0107	0	0	A/A
	M	14X	0.0088	0.0616	0.0352	0.0572	0.0088	0.0616	0.0176	0.0264	0.0220	0.0044	0	0	B/L24
	375N	74X	0.0472	0.0340	0.0546	0.0874	0.0185	0.0941	0.0340	0.0394	0.0374	0.0148	0.0057	0.0010	
380	F	17X	0.0138	0.0635	0.0220	0.0580	0.0165	0.0165	0.0083	0.0110	0.0110	0.0027	0	0.0055	A/A
	M	18X	0.0401	0.0401	0.0321	0.0455	0.0053	0.0642	0.0080	0.0214	0.0160	0.0160	0.0053	0	A/B
	380N	59X	0.0440	0.0343	0.0478	0.0994	0.0292	0.0816	0.0364	0.0575	0.0364	0.0296	0.0017	0.0008	-
390	F	22X	0.0221	0.0287	0.0265	0.0486	0.0110	0.0464	0.0221	0.0221	0.0110	0.0066	0.0022	0	A/A
	M	17X	0.0239	0.0418	0.0269	0.0239	0.0060	0.0537	0.0149	0.0447	0.0269	0.0149	0	0	A/L31
	390N	73X	0.0349	0.0322	0.0728	0.0856	0.0311	0.0802	0.0325	0.0450	0.0349	0.0244	0.0024	0.0003	-
420	F	30X	0.0101	0.0268	0.0385	0.0469	0.0168	0.0519	0.0084	0.0117	0.0101	0.0101	0.0117	0	A/L20
	M	27X	0.0127	0.0362	0.0416	0.0398	0.0036	0.0307	0.0126	0.0325	0.0181	0.0145	0.0036	0.0108	A/C
	420N	11X	0.0963	0.0229	0.0367	0.0505	0.0046	0.0826	0.0229	0.0092	0.0229	0.0092	0.0183	0	-
443	F	32X	0.0296	0.0172	0.0250	0.0499	0.0094	0.0374	0.0109	0.0296	0.0125	0.0047	0	0	A/A
	M	41X	0.0142	0.0303	0.0242	0.0763	0.0085	0.0303	0.0097	0.0097	0.0109	0.0109	0	0.0012	A/L24
	443N	8X	0.0613	0.0123	0.0307	0.0675	0.0184	0.0429	0.0184	0.0308	0.0245	0.0429	0	0	-
579	F	25X	0.0160	0.0321	0.0481	0.0561	0.0200	0.0401	0.0361	0.0341	0.0180	0.0140	0	0	A/A
	M	22X	0.0158	0.0271	0.0249	0.0633	0.0090	0.0520	0.0136	0.0113	0.0158	0.0181	0.0113	0.0023	A/L20
	579N	10X	0.0653	0.0201	0.0402	0.0553	0.0151	0.0503	0.0201	0.0302	0.0050	0.0050	0	0	-
580	F	21X	0.0238	0.0285	0.0404	0.0451	0.0024	0.0618	0.0071	0.0380	0.0166	0.0071	0	0.0071	A/L3
	M	25X	0.0121	0.0382	0.0221	0.0543	0.0101	0.0423	0.0141	0.0201	0.0121	0.0060	0.0020	0	A/L30
	580N	20X	0.0483	0.0266	0.0193	0.0459	0.0097	0.0700	0.0266	0.0242	0.0193	0.0048	0.0024	0	
595	F	37X	0.0134	0.0255	0.0282	0.0523	0.0094	0.0282	0.0121	0.0255	0.0255	0.0188	0.0013	0	A/A
	M	17X	0.0060	0.0150	0.0300	0.0390	0.0030	0.0390	0.0120	0.0270	0.0150	0.0060	0	0	A/A
	595N	26X	0.0594	0.0249	0.0287	0.1054	0.0096	0.0421	0.0115	0.0326	0.0192	0.0230	0	0.0038	-
728	F	18X	0.0088	0.0352	0.0235	0.0323	0.0029	0.0557	0.0176	0.0029	0.0293	0.0088	0.0264	0	A/L20
	M	17X	0.0254	0.0254	0.0226	0.0452	0.0085	0.0565	0.0056	0.0367	0.0226	0.0085	0	0.0028	A/B
	728N	111X	0.0646	0.0704	0.0267	0.0602	0.0120	0.0508	0.0152	0.0218	0.0290	0.0120	0.0201	0.0027	-
752	F	15X	0.0373	0.0271	0.0407	0.0237	0.0102	0.0305	0.0136	0.0237	0.0339	0.0034	0	0	A/B
	M	20X	0.0170	0.0414	0.0365	0.0852	0.0024	0.0341	0.0024	0.0292	0.0073	0.0049	0.0097	0.0146	A/C
	752N	127X	0.1768	0.0613	0.0223	0.0539	0.0055	0.0320	0.0152	0.0133	0.0156	0.0047	0.0141	0.0031	
764	F	13X	0.0197	0.0551	0.0433	0.0354	0.0079	0.0590	0.0079	0.0433	0.0315	0.0157	0	0	A/A
	M	23X	0.0191	0.0404	0.0297	0.0489	0	0.0319	0.0064	0.0063	0.0234	0.0042	0.0021	0.0043	A/A
	764N	92X	0.1136	0.0302	0.0210	0.0668	0.0005	0.0248	0.0081	0.0118	0.0065	0.0172	0.0011	0	-
794	F	15X	0.0130	0.0519	0.0357	0.0519	0.0065	0.0390	0.0130	0.0422	0.0292	0.0195	0.0065	0.0032	A/A
	M	14X	0.0141	0.0141	0.0424	0.0212	0	0.0530	0.0141	0.0247	0.0247	0.0212	0.0212	0	B/D
	794N	40X	0.1038	0.04	0.0375	0.0513	0.0175	0.0425	0.0188	0.0175	0.0075	0.0113	0.0088	0.0013	-

[a]Alleles A, B, C, D, L3, L20 and L24 are described in Berg et al. 2010 [8]. Alleles L30 and L31 are novel alleles, described in Supplementary Results.

For fathers (F), mothers (M) and patients (N) in each family, the ZnF repeat types from *PRDM9* alleles were first inferred from SOLiD sequencing read data. Repeat types (*a* to *l*) are described in Figure S6. Repeat types with a proportion above 0.01 (highlighted) are inferred to be present in the individuals. Sanger sequencing was subsequently performed in the parents and genotypes with rare alleles are highlighted. All *k* and *l* fingers inferred were validated (alleles C, D, L20). *(next page)*

.

Table S4. *PRDM9* alleles in an additional 10 ALL trios with B-ALL children based on read data.

Family	Individual	Coverage	Proportion of Exome Sequencing Reads Aligning to PRDM9 ZnF Repeats											
			a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
392	F	30X	0.101	0.053	0.0215	0.0381	0.0033	0.0712	0.0083	0.043	0.0248	0.0182	0	0
	M	27X	0.0731	0.0567	0.0347	0.0475	0.0037	0.0969	0.0183	0.0457	0.0475	0.0201	0	0.0018
	392N	78X	0.0829	0.0383	0.0625	0.074	0.0261	0.0778	0.0274	0.0657	0.0338	0.03	0.0019	0
424	F	33X	0.1161	0.0268	0.0357	0.067	0.0149	0.0818	0.0402	0.0685	0.0327	0.0298	0.006	0
	M	26X	0.0854	0.0645	0.0361	0.0323	0.0114	0.0721	0.0209	0.0361	0.0304	0.019	0.0038	0
	424N	47X	0.1876	0.0486	0.0423	0.0455	0.0095	0.0708	0.0381	0.0476	0.0275	0.019	0.0011	0
614	F	33X	0.1271	0.0651	0.0182	0.0272	0.0076	0.0227	0.0045	0.0605	0.0121	0.0136	0.0182	0.0106
	M	12X	0.1548	0.0502	0.0418	0.0167	0.0084	0.0502	0.0209	0.0209	0.0251	0.0293	0	0
	614N	76X	0.162	0.0567	0.0423	0.0476	0.0078	0.0495	0.015	0.0436	0.0267	0.0137	0.0007	0
617	F	13X	0.1231	0.0423	0.0346	0.0615	0.0346	0.0462	0.0423	0.0385	0.0231	0.0154	0	0
	M	11X	0.129	0.0599	0.0276	0.0276	0	0.0323	0.0184	0.0323	0.0323	0.0046	0.0046	0
	617N	13X	0.1088	0.083	0.0226	0.1132	0.0113	0.0415	0.0226	0.0491	0.0302	0.0113	0	0
657	F	9X	0.1703	0.033	0.022	0.0385	0.0055	0.0604	0.011	0.0604	0.033	0.0165	0.0055	0
	M	12X	0.1331	0.0403	0.0282	0.0524	0.004	0.0605	0.0161	0.0323	0.0403	0.0242	0	0
	657N	5X	0.0612	0.011	0.0659	0.1099	0.0549	0.0659	0	0.0769	0.011	0.022	0.009	0
685	F	24X	0.0763	0.0495	0.0268	0.0392	0.0062	0.0763	0.033	0.033	0.0371	0.0186	0.0144	0.0051
	M	20X	0.2545	0.0375	0.03	0.04	0.0075	0.05	0.025	0.05	0.035	0.02	0.0025	0
	685N	75X	0.1651	0.0617	0.0504	0.0338	0.0179	0.067	0.0239	0.0498	0.0319	0.01	0.002	0.0007
761	F	38X	0.0652	0.0417	0.0248	0.0495	0.0091	0.0717	0.0313	0.0456	0.0183	0.0209	0.0013	0.0013
	M	40X	0.0891	0.054	0.0151	0.0276	0.0025	0.0552	0.0113	0.0452	0.0213	0.0364	0.0025	0
	761N	92X	0.1351	0.0346	0.0362	0.0324	0.0086	0.0789	0.0243	0.0546	0.0286	0.0178	0.0005	0
762	F	25X	0.097	0.0257	0.0436	0.0535	0.0119	0.0614	0.0158	0.0416	0.0238	0.0139	0	0
	M	20X	0.0973	0.0389	0.0414	0.0389	0.0049	0.0681	0.0292	0.0316	0.0292	0.0292	0	0
	762N	81X	0.1193	0.0555	0.0262	0.0366	0.0085	0.0634	0.0128	0.0469	0.0189	0.014	0.0018	0.0006
767	F	21X	0.1253	0.0394	0.0487	0.0557	0.0046	0.0626	0.0116	0.0232	0.0255	0.0093	0.0116	0
	M	32X	0.1144	0.058	0.0329	0.0502	0.0094	0.0533	0.0078	0.0094	0.011	0.0157	0.0204	0.0031
	767N	80X	0.1277	0.0646	0.0292	0.0385	0.0062	0.0752	0.0143	0.0242	0.0311	0.0168	0.0242	0
777	F	18X	0.102	0.0737	0.0283	0.0255	0.0113	0.0567	0.0227	0.0482	0.0198	0.0368	0	0
	M	22X	0.1615	0.0664	0.0221	0.0442	0.0066	0.0708	0.0221	0.0243	0.0243	0.0243	0.0111	0
	777N	59X	0.2017	0.0524	0.0304	0.0355	0.0127	0.0609	0.0144	0.0262	0.0262	0.0144	0.0203	0

ZnF repeat types present in *PRDM9* were inferred based on SOLiD read data, for fathers (F), mothers (M) and patients (N) in each family. Repeat types (a to l) are described in Figure S6. Repeat types with a proportion above 0.01 (highlighted) are inferred to be present in the individuals.

Table S5. *PRDM9* alleles in 76 French-Canadian individuals.

A

Individual ID	Parent	Alleles
8	M	A/A
118	M	A/L24
183	M	A/A
385	M	A/A
210	M	A/A
190	M	A/L9
229	M	A/A
743	M	A/L1
51	M	A/L24
772	M	A/A
608	M	A/B
245	M	A/A
20	F	A/A
257	F	A/L20
270	F	A/A
596	F	A/A
64	F	A/A
713	F	A/A
717	F	A/A
90	F	A/A
944	F	A/A
146	F	A/A

B

Couple ID	Alleles	
	M	F
15_11	A/A	A/B
223_222	A/A	A/A
304_303	A/L32	A/A
348_347	A/A	A/A
38_39	A/A	L20/L24
393_392	A/C	A/A
413_412	A/A	A/A
424_423	A/A	A/A
428_427	A/A	A/A
43_48	A/A	A/A
460_459	A/L9	A/C
584_585	A/A	A/L33
626_625	A/A	A/A
647_651	A/A	A/L34
656_657	A/A	A/B
66_68	A/B	A/L24
692_691	A/L32	A/B
728_311	A/A	A/A
740_739	A/A	A/B
748_749	A/A	A/A
755_756	A/A	A/A
75_74	A/A	A/A
812_815	A/A	A/A
818_817	A/A	A/A
830_823	A/A	A/A
854_853	A/D	A/A
96_95	A/A	A/B

Individuals are mothers (M) and fathers (F) of at least 2 children from 49 families **(A)** one parent was sampled per family and with **(B)** both parents sampled per families. *PRDM9* ZnF alleles were assayed by Sanger sequencing. Genotypes with rare alleles are highlighted. Alleles A, B, C, D, L1, L9, L20 and L24 are described in Berg et al. 2010 (8). Alleles L32, L33 and L34 are novel alleles, described in Supplementary Results.

Table S6. B-ALL molecular subtypes for the 24 patients included in this study.

Child	Sex	Molecular Group	Detected translocations	Leukemic clone ploidy	*k*-finger in Family	Parent Carrier
610	Male	H	None	51-53	Yes	M
383	Male	H	None	55	Yes	M
375	Female	T	t(12;21)	46	No	-
380	Female	O	n/d	46	No	-
390	n/d	H	None	54	No	-
420	n/d	T	t(12;21)	n/d	Yes	M and F
443	Female	T	t(12;21)	46	No	-
579	n/d	O	None	46	Yes	M
580	n/d	T	t(12;21)	n/d	No	-
595	Male	O	None	47	No	-
728	n/d	O	None	n/d	Yes	F
752	Male	T	t(9;12)	45	Yes	M
764	Male	H	None	56	No	-
794	Male	O	None	46	Yes	M
392	Male	T	t(12;21)	46	No	-
424	Male	T	t(1;19)	46	No	-
614	Female	T	t(12;21)	46	Yes	F
617	n/d	O	None	n/d	No	-
657	n/d	T	t(12;21)	n/d	No	-
685	n/d	O	None	n/d	Yes	F
761	Male	H	None	56	No	-
762	n/d	O	None	n/d	No	-
767	Female	O	None	n/d	Yes	M and F
777	Female	H	None	49-54	Yes	M

The patients present different subtypes of B-ALL: high hyperdiploid clones (H), clones with translocation (T) and other uncharacterized translocations or genetic defects (O). There is no significant difference between subtypes for the presence of *k*-finger alleles in a family (Freeman-Halton test with 3 categories (Freeman and Halton 1951), $p = 0.268$). There is no significant difference between maternal (M) and paternal (F) origin of the *k*-finger alleles ($p = 0.369$, Fisher's exact test).

Table S7. *PRDM9* alleles in 50 children from SJDALL cohort based on read data.

Patients are separated in 4 B-ALL subtypes: ETV6 translocation (SJETV), hypodiploid (SJHYPO), infant (SJINF) and Philadelphia chromosome-positive (SJPHALL). Reported ethnicities were verified by PCA of genotyped data (Figure S8). Illumina read data from tumor and normal sample sequencing were used. Repeat types found in the read data with a proportion above 0.01 (highlighted) are inferred to be present in the individuals. Repeat types (*a* to *l*) are described in Figure S6. *(next page)*

Table S7 (*continued*)

Sample	Reported ethnicity (PCA)	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
SJETV010	White	0,065	0,052	0,1194	0,1525	0,0437	0,169	0,0686	0,0792	0,0697	0,0922	0	0,0012
SJETV022	White	0,0956	0,0566	0,127	0,1597	0,0314	0,1597	0,0616	0,0541	0,0717	0,073	0,0025	0,0025
SJETV024	White	0,1004	0,0588	0,1136	0,1714	0,0385	0,1572	0,0619	0,0598	0,0619	0,0822	0,003	0
SJETV027	White	0,1155	0,0599	0,1484	0,1284	0,0485	0,1469	0,0728	0,0613	0,0542	0,0942	0,0014	0,0043
SJETV028	White	0,0906	0,0523	0,1115	0,151	0,0441	0,1521	0,0743	0,0476	0,0662	0,0952	0,0023	0,0012
SJETV073	White	0,0871	0,0389	0,1191	0,1294	0,047	0,1569	0,0664	0,0825	0,0573	0,0653	0,0011	0,0034
SJETV085	White	0,0947	0,0579	0,107	0,1719	0,0526	0,1377	0,0588	0,0675	0,0781	0,0895	0,0018	0
SJETV089	White	0,0952	0,0476	0,1125	0,1743	0,0547	0,1358	0,0689	0,0598	0,0912	0,075	0	0,001
SJETV194	White	0,1035	0,0472	0,1068	0,1631	0,0422	0,1548	0,072	0,0522	0,0927	0,0844	0	0,0025
SJHYPO004	White	1,4588	0,0547	0,1258	0,1569	0,0588	0,1422	0,0662	0,0596	0,0743	0,0825	0	0,0016
SJHYPO006	White (adx)	0,0778	0,0488	0,1394	0,1092	0,0174	0,0871	0,0163	0,1254	0,0778	0,1045	0,0476	0,0395
SJHYPO013	Other(Asian)	0,1008	0,0584	0,142	0,1386	0,0309	0,1581	0,0653	0,0561	0,0687	0,0882	0	0,0046
SJHYPO021	White	0,1084	0,0515	0,1041	0,1468	0,0438	0,1391	0,0635	0,0624	0,0756	0,069	0,0011	0,0044
SJHYPO022	White	0,0929	0,0506	0,0988	0,1247	0,0565	0,1412	0,0824	0,0788	0,0624	0,0765	0	0,0024
SJHYPO040	White	0,0978	0,0422	0,0989	0,1578	0,05	0,13	0,0711	0,0222	0,08	0,0844	0,0322	0
SJHYPO042	White	0,0964	0,0321	0,1358	0,1378	0,0394	0,1492	0,0591	0,0808	0,0881	0,0725	0,0259	0,0114
SJHYPO044	White	0,0912	0,0592	0,1208	0,1739	0,0567	0,1406	0,0629	0,0838	0,0826	0,0703	0,0012	0,0012
SJHYPO046	White	0,0922	0,0454	0,121	0,121	0,055	0,1692	0,0646	0,0523	0,0688	0,088	0	0,0028
SJHYPO051	White (adx)	0,1194	0,0525	0,1089	0,1325	0,0617	0,1483	0,0499	0,0604	0,0722	0,0682	0	0,0026
SJHYPO052	White	0,0859	0,0452	0,0914	0,1333	0,0562	0,1773	0,0617	0,0738	0,0804	0,0881	0	0
SJHYPO055	White	0,1009	0,0482	0,1377	0,136	0,0684	0,1675	0,0667	0,0272	0,0614	0,0728	0,0342	0,0018
SJHYPO056	White	0,0883	0,0331	0,1145	0,1283	0,0483	0,2	0,1034	0,0676	0,0717	0,0952	0,0014	0
SJHYPO119	White	0,1138	0,0588	0,1077	0,1457	0,0465	0,12	0,06	0,0575	0,0612	0,0942	0,0037	0,0024
SJHYPO120	Other (Hisp)	0,1115	0,0632	0,0979	0,1586	0,0483	0,1561	0,0595	0,062	0,0595	0,0768	0,0012	0
SJHYPO123	White	0,0903	0,0593	0,1442	0,1226	0,0013	0,1119	0,0472	0,0822	0,0916	0,0836	0,0202	0,0162
SJINF001	White	0,0989	0,053	0,0901	0,1325	0,0336	0,1237	0,0636	0,0601	0,0583	0,0901	0,0018	0
SJINF002	White	0,8073	0,0477	0,1134	0,1314	0,0322	0,1198	0,0619	0,058	0,0464	0,0851	0,0026	0
SJINF003	White	0,1079	0,0539	0,0991	0,1181	0,0466	0,1254	0,0729	0,0466	0,051	0,0802	0,0102	0,0029
SJINF004	White (adx)	0,0971	0,0659	0,0989	0,1099	0,0403	0,1337	0,0659	0,0513	0,0769	0,0842	0	0,0018
SJINF005	White	0,1058	0,0276	0,0982	0,135	0,0537	0,138	0,0613	0,0675	0,0537	0,0813	0,0031	0,0015
SJINF006	White	0,1079	0,0468	0,1079	0,1571	0,0528	0,1439	0,0576	0,0635	0,0743	0,0743	0,0012	0,0012
SJINF007	White	0,081	0,0444	0,1059	0,1752	0,0471	0,1529	0,1007	0,0458	0,0745	0,0706	0,0013	0,0013
SJINF009	White	0,1008	0,0485	0,1263	0,1505	0,0306	0,148	0,0842	0,0268	0,051	0,0663	0,0472	0
SJINF011	White	0,0914	0,054	0,1167	0,1189	0,0474	0,1355	0,0584	0,0727	0,0793	0,0771	0,0011	0,0011
SJINF012	White	0,0885	0,0369	0,1118	0,1339	0,0344	0,1413	0,0725	0,0762	0,0676	0,0909	0	0,0025
SJINF013	White	0,091	0,0556	0,1327	0,0999	0,0164	0,1466	0,0721	0,0708	0,0582	0,0999	0,0025	0
SJINF014	Hispanic	0,098	0,059	0,1359	0,1258	0,0223	0,0991	0,029	0,1102	0,0724	0,0835	0,0379	0,0212
SJINF015	White	0,129	0,0496	0,1042	0,1191	0,0323	0,1377	0,0658	0,0484	0,098	0,0744	0,005	0,0025
SJINF016	Hispanic	0,0943	0,0432	0,1489	0,1	0,017	0,1034	0,0295	0,0693	0,067	0,0886	0,0625	0,0216
SJINF017	Hispanic	0,1032	0,0523	0,0921	0,1004	0,0418	0,1402	0,0635	0,06	0,0676	0,0781	0,0021	0,0014
SJINF019	White	0,118	0,0504	0,0878	0,1468	0,036	0,1583	0,0647	0,0576	0,0619	0,0835	0,0014	0,0014
SJINF020	White	0,1169	0,0438	0,1036	0,1474	0,0372	0,1421	0,073	0,0664	0,0558	0,0704	0,008	0
SJINF022	White	0,9667	0,0516	0,1371	0,117	0,0189	0,1006	0,0239	0,0994	0,0516	0,073	0,0365	0,0352
SJPHALL001	White	0,0713	0,0526	0,132	0,1507	0,0654	0,1507	0,0841	0,0678	0,0759	0,0654	0	0
SJPHALL003	White	0,0906	0,0482	0,1189	0,148	0,0532	0,1654	0,0657	0,0673	0,059	0,0798	0,0017	0,0025
SJPHALL004	Asian	0,0902	0,0486	0,1266	0,1449	0,0466	0,1459	0,0811	0,0669	0,0719	0,0973	0,001	0,0041
SJPHALL005	White (adx)	0,0931	0,0486	0,1275	0,1306	0,0547	0,1528	0,0739	0,0628	0,0648	0,0921	0	0,002
SJPHALL006	White	0,085	0,0385	0,1023	0,1859	0,0332	0,1554	0,0637	0,0531	0,0491	0,085	0,0027	0
SJPHALL007	White (adx)	0,0847	0,0367	0,1195	0,159	0,0555	0,1515	0,0593	0,0724	0,0931	0,0753	0,0028	0
SJPHALL008	White	0,089	0,0503	0,1285	0,1533	0,055	0,154	0,0789	0,0557	0,0875	0,072	0,0008	0,0015

Table S8: Most frequent translocations and fusion genes in ALL.

(A) The most frequent translocations involved in ALL found in the dbCRID and Mitelman database (Kong et al. 2011; Mitelman et al. 2011) (B) The ALL gene list is composed of fusion genes reported for the translocations from databases interrogated in (A) and found to be implicated in ALL in peer-reviewed publications.

A

Translocations	Cytogenetic bands		Entries in Databases	
			dbCRID	Mitelman
t(1,11)	1p32/1q23	11q23	0	11
t(1,14)	1p32(p33)	14q11	5	6
	1q21	14q32	8	9
t(1,19)	1q23	19p13.3	7	49
t(4,11)	4q21	11q23.3	20	88
t(5,14)	5q34(q35)	14q11/14q32	3	9
t(6,11)	6q27	11q23	1	10
t(7,9)	7q34/q11	9q34/p13	1	16
t(8,14)	8p24	14q11/q32	3	24
t(9,11)	9p21/q34	11q23.3	5	21
t(9,22)	9q34	22q11.2	19	136
t(10,11)	10p12	11q14/q23	2	13
t(10,14)	10p24.31	14q11.2/q32	8	8
t(11,19)	11q23	19p13.3	5	30
t(17,19)	17q22	19p13	2	11
t(12,21)	12p13.2	21q22.12	29	58

Table S8 (*continued*)

B

Gene	Chr	Start	End	Translocations	Nb of *C* motifs	*C* motifs by Kb
TAL1	1	47454550	47469974	t(1;14)(p32;q11)	5	0.314
EPS15	1	51592522	51757583	t(1;11)(p32;q23)	3	0.018
BCL9	1	145479805	145564639	t(1;14)(q21;q32)	2	0.024
MLLT11	1	149298774	149307597	t(1;11)(q21;q23)	0	-
PBX1	1	162795560	163082934	t(1;19)(q23;p13)	5	0.017
SEPT11	4	78089918	78178792	t(4;11)(q21;q23)	5	0.056
AFF1	4	88075186	88232005	t(4;11)(q21;q23)	9	0.044
RANBP17	5	170221599	170659624	t(5;14)(q34;q11)	16	0.037
TLX3	5	170668892	170671743	t(5;14)(q34;q11/q32)	4	1.403
NKX2-5	5	172591743	172594868	t(5;14)(q34;q32)	0	-
MLLT4	6	167970519	168115552	t(6;11)(q27;q23)	8	0.055
AUTS2	7	68701840	69895821	t(7;9)(q11;p13)	39	0.033
POM121	7	71987871	72059915	t(7;9)(q11;p13)	7	0.097
ELN	7	73080362	73122172	t(7;9)(q11;p13)	12	0.287
TCRB	7	141674678	141987064	t(7;9)(q34;q34)	9	0.046
MYC	8	128816946	128820200	t(8;14)(q24;q11/q32)	0	-
PVT1	8	128875960	129182681	t(8;14)(q24;q11/q32)	16	0.052
MLLT3	9	20334967	20612514	t(9;11)(q21;q23)	9	0.032
PAX5	9	36828530	37024476	t(7;9)(q11;p13)	23	0.117
ABL1	9	132579088	132752883	t(9;22)(q34;q11)	4	0.023
NOTCH1	9	138508716	138560059	t(7;9)(q34;q34)	21	0.409
MLLT10	10	21863107	22072560	t(10;11)(p12;q14/q23)	8	0.038
TLX1	10	102880251	102887526	t(10;14)(p24;q11)	3	0.412
LMO2	11	33836698	33870412	t(7;11)(q34;p13)	2	0.059
PICALM	11	85346132	85457756	t(10;11)(p12;q14)	2	0.018
MLL	11	117812414	117901146	t(1;11)(p32;q23),t(4;11)(q21;q23),t(6;11)(q27;q23), t(9;11)(q21;q23),t(10;11)(p12;q23),t(11;19)(q23;p13)	6	0.066
CCND2	12	4253198	4284777	t(7;12)(q34;p13)	2	0.063
ETV6	12	11694054	11939592	t(12;21)(p13;q22),t(7;12)(q34;p13),t(9;12)(q34;p13)	14	0.052
TRA@	14	21432409	21604421	t(1;14)(p32;q11), t(5;14)(q34;q11), t(8;14)(q24;q11),	17	0.030
TRD@		21987946	21995540	t(10;14)(p24;q11), t(11;14)(p13/q23;q11)	0	-
BCL11B	14	98705377	98807575	t(5;14)(q34;q32)	22	0.215
IGH@	14	105124270	105401515	t(1;14)(q21;q32), t(5;14)(q34;q32), t(8;14)(q24;q32)	245	0.884
HLF	17	50697320	50757425	t(17;19)(q22;p13)	0	-
DAZAP1	19	1358583	1386682	t(1;19)(q23;p13)	13	0.463
TCF3	19	1560294	1603328	t(1;19)(q23;p13)	20	0.465
MLLT1	19	6161391	6230959	t(11;19)(q23;p13)	19	0.273
RUNX1	21	35081967	35343511	t(12;21)(p13;q22)	64	0.053
BCR	22	21852551	21990224	t(9;22)(q34;q11)	26	0.189

Table S9. PRDM9 alleles binding motifs in unique and repetitive DNA

	Unique DNA		Repetitive DNA		Segmental Duplications	
	Genes	ALL gene list	Genes	ALL gene list	Genes	ALL gene list
A	3227	13	3726	21	371	3
C	30313	390	64928	458	5489	206
C vs *A* OR [CI]	2.39 [1.50;3.81]**		1.81 [0.69;4.76]		4.78 [1.84;12.46] **	

** significant based on 95% CI (one-tailed $p < 0.025$)

Number of motifs *A* and *C,* as presented in Table 2, in coding regions (similar results are obtained for the whole genome) and in the ALL gene list (Table S8). We compared counts using odds ratios (OR), to measure the association between motifs and their occurrence in the ALL gene list. The motif search was performed on the non-degenerate version of the Human Reference Genome (hg18). Repetitive regions were obtained from UCSC tables, with regions found by RepeatMasker (Smit 1996-2012) and Tandem Repeat Finder (Benson 1999) programs considered as repetitive DNA. Segmental duplications coordinates were also obtained from UCSC tables.

Table S10. Data and analyses performed in this study.

Dataset	Families	Individuals per family	Data	Analyses
ALL quartet	1	2 parents + 2 offsprings	Genotyping on Illumina 2.5M	Ancestry analyses
			Genotyping on Affymetrix 6.0	Recombination Analyses
			Exome Sequencing on SOLiD 4.0	*De novo* Mutation Discovery
			Sanger sequencing of PRDM9 ZnF alleles of parents	PRDM9 ZnF alleles determination
FCALL cohort (Total of 22 parental trios)	22	2 parents + 1 offspring	Exome Sequencing on SOLiD 4.0	Ancestry analyses
	12*	2 parents + 1 offspring	Sanger sequencing of PRDM9 ZnF alleles of parents	PRDM9 ZnF alleles determination
SJDALL cohort (Total of 61 children)	61	1 individual	Genotyping on Affymetrix 6.0	Ancestry analyses
	50*	1 individual	Paired-end WGS on Illumina HiSeq	PRDM9 ZnF alleles determination
FC cohort (Total of 69 families)	69	2 parents + 2 offsprings	Genotyping on Affymetrix 6.0	Recombination Analyses
				Ancestry analyses
	27*	2 parents	Sanger sequencing of PRDM9 ZnF alleles of parents	PRDM9 ZnF alleles determination
	22*	1 parent		
Moroccan cohort (Total of 163 indiv)	163	1 individual	Genotyping on Illumina Human 610-Quad	Ancestry analyses
	27*	1 individual	Sanger sequencing of PRDM9 ZnF alleles	PRDM9 ZnF alleles determination

* A subset of the complete cohort

This study uses genetic information from a total of 639 individuals. The ALL quartet is part of the FCALL cohort but is displayed seperately since many analyses were performed on this family only.

CHAPTER IV:

Impact of variable recombination on human mutation load

Julie Hussin, Youssef Idaghdour, Alan Hodgkinson, Melanie Capredon, Jean-Christophe Grenier, Jean-Philippe Goulet, Thibault de Malliard, Elias Gbeha, Elodie Hip-Ki, Yves Payette, Catherine Boileau, Philip Awadalla

Reference:

Hussin J., Idaghdour Y., Hodgkinson A., Capredon M., Grenier J-C., *et al*. 2013. Recombination and Efficiency of Selection in a founding population. In preparation

AUTHOR'S CONTRIBUTION

In this paper, my contribution is:

- The idea to test the study's hypothesis;
- Contribution to bioinformatics pipelines for SNP calling from RNAseq data;
- Annotation of SNPs and exons from RNAseq data;
- Development of haplotype phasing procedure;
- Downstream computational and statistical analyses;
- Writing of the manuscript.

Contributions of other authors are: PA CB and YP contributed reagents, materials and samples. YI EG and EH performed the experimental work. YI AH MC JCG JPG and TdM performed bioinformatics analyses on the data. JCG implemented the haplotype phasing procedure. YI and PA conceived the cardiometabolic RNAseq project. PA revised the manuscript.

ACKNOWLEDGMENTS

I would like to thank all participants from the CARTaGENE project and all people from the CARTaGENE team that contributed to the recruitment. I thank C. Bherer, L. Excoffier and J. Novembre for helpful discussions and comments. I acknowledge the Genome Quebec Innovation Centre at McGill University for genotyping and sequencing services.

ABSTRACT

A major prediction in evolutionary biology is that linkage between sites that are simultaneously under selection will reduce the overall efficacy of natural selection in finite populations. However, evidence that variation in recombination rate across the genome leads to variation in the efficacy of selection is lacking in humans. In this study, we use genomic data to investigate the differences in mutational load between regions of high (>5cM/Mb) and low (<0.5cM/Mb) recombination in the human genome. We built the genetic map of the Quebec population from genotyping data and called SNPs from RNA sequencing (RNAseq) data in 521 French-Canadian individuals recruited by the CARTaGENE Project. We calculated the differential mutational load between high and low recombination regions by comparing the amount of variants that are likely to impact fitness, characterized based on functional annotations, conservation scores and frequency-based statistics. We find that SNPs in low recombination regions are significantly enriched for highly-constrained, low frequency nonsynonymous variants, relative to SNPs in high recombination regions, both at a population-level and at the individual-level. These variants are observed in linkage with each other more often in low recombination regions than in highly recombining parts of the genome, which indicates that they accumulate on the same haplotypes. Finally, we replicated our finding in the 1000 Genomes Project populations and observed that the differential mutational load per individual varies among human populations. These results strongly suggest that weakly deleterious mutations are less efficiently removed by natural selection in regions of low recombination rate in the human genome and this phenomenon might impact disease susceptibility at the individual level.

INTRODUCTION

In sexually reproducing species, it is well documented that non-recombining genomic regions, such as the Y chromosome (Charlesworth and Charlesworth 2000), tend to accumulate deleterious mutations and lose functional genes faster than the rest of the genome. This is because selection at one site will interfere with the action of selection at linked sites, thereby reducing the efficiency of selection in non-recombining regions (Hill and Robertson 1966; Felsenstein 1974). According to theoritical expectations, variation in recombination rate along the chromosomes should modulate the strength of interference between selected alleles. However, little evidence supports the hypothesis that the observed variation of crossover rate across the recombining chromosomes leads to variation in the efficiency of selection across the human genome.

Several studies have specifically investigated this question in humans and other species, mainly through an evaluation of between-species nucleotide divergence across recombination environments (Pal et al. 2001; Haddrill et al. 2007; Bullaughey et al. 2008). These studies assume that the local recombination rates will affect both within-species nucleotide diversity and between-species nucleotide divergence in a similar fashion. However, the pertinence of this prediction is questionable given the existence of pervasive differences in recombination rates between species and the distribution of genes between high and low recombination regions within species. In fact, recombination rates are often not conserved between species, and vary at fine scales between closely related species, such as humans and chimpanzees (Ptak et al. 2005; Auton et al. 2012). Furthermore, in humans, highly conserved genes implicated in essential cellular processes are particularly enriched in regions of high linkage disequilibrium (Smith et al. 2005) .

Another approach is to use patterns of diversity within populations to test whether the efficiency of selection is reduced in autosomal regions with low recombination. In

low recombination regions, favorable mutations arising in different individuals of a population will rarely be combined onto the same haplotype. Different alleles under selection will, on average, interfere with one another's fixation thus reducing the efficiency of selection, a phenomenon known as Hill-Robertson interference (Hill and Robertson 1966). Conversly, low recombination rates will cause disadvantageous mutations to accumulate on the same haplotype, which will lead to an increase in the average number of unfavorable mutations per genomic region. This is known as Muller's ratchet mechanism (Muller 1964; Felsenstein 1974). At the population level, this increased mutational load is expected to be quantitatively more important than interference between favorable mutations, because weakly deleterious mutations are occurring in larger numbers than advantageous ones.

The population mutational load, defined as the cumulative effect of deleterious mutations at the population level, results in a reduction in fitness for the population compared to a case where all individuals would have the most favored genotype at each site (Muller 1950; Agrawal and Whitlock 2012). Mutational load in a given genomic region can be estimated from genomic data as the number of deleterious variants present in that region. To assess whether a variant is deleterious, genic sequences are annotated based on the protein sequence they code for as synonymous or nonsynonymous (missense and nonsense). Coupled with prediction tools such as Polyphen (Adzhubei et al. 2010) and SIFT (Kumar et al. 2009) that classify missense variants as being either damaging or benign, this approach is widely used in medical genetics to characterize the putative impact of a variant on protein function. Furthermore, it has been shown that the degree of conservation at the variant site is one of the most reliable methods for assessing its pathogenicity (Flanagan et al. 2010). Conservation scores such as GERP (Davydov et al. 2010) and PhyloP (Siepel et al. 2006) are therefore powerful measures to identify mutations that are likely deleterious. Finally, measures based on allele frequencies are generally good indicators of the functional importance of a variant. Indeed, low frequency variants are enriched for mutations affecting protein function (Marth et al. 2011; Nelson et al.

2012) and the minor allele frequency (MAF) of a mutation correlates negatively with the level of evolutionary constraint at the mutated site (Cooper et al. 2010; Goode et al. 2010; Hodgkinson et al. 2013).

In this work, we use genomic data to investigate whether patterns of variation in recombination rates in the human genome lead to differential efficacy of natural selection in removing deleterious alleles. Fine-scale variation in human recombination rates consists of large coldspots with very low recombination rates, punctuated by short hotspots of recombination, where most of the crossover events occur. Here, we quantified the differences in mutational load between coldspots and regions with high density of hotspots. SNPs were called from RNA sequencing (RNAseq) data and population recombination rates were estimated from genotyping data in French-Canadian individuals recruited by the CARTaGENE Project (Awadalla et al. 2012). Mutational load is evaluated separately in low and high recombination regions, by characterizing variants based on functional annotations, conservation scores and frequency-based statistics. We find that SNPs in low recombination regions are significantly enriched in low frequency and nonsynonymous variants and are found at constrained positions moreso than SNPs in high recombination regions. These variants are likely mildly deleterious, hence impacting fitness. This enrichment is seen at both the population-level and the individual-level and deleterious variants are observed in linkage with each other more often in low recombination regions than in highly recombining parts of the genome. This finding indicates that they accumulate on the same haplotypes in coldspots, in line with the predicted outcome of Muller's ratchet mechanism. Finally, we observed a positive correlation between the differential mutational load per individual and the risk of cardio-vascular diseases. These results strongly suggest that deleterious mutations are less efficiently removed by negative selection in regions of low recombination rate and that this phenomenon likely impacts human health at the individual level.

RESULTS

The French-Canadian Study Population

The six million French-Canadians in Quebec are descendants of about 8,500 settlers, mostly of French origin, who colonized the province between 1608 and 1759 (Scriver 2001). The CARTaGENE project (CaG) collected biologicals and data from 20,000 participants recruited throughout the province of Quebec (Awadalla et al. 2012), and high-density genotyping and RNA sequencing data was generated for 521 French-Canadians participants (Material and Methods). Sampling includes individuals from three distinct metropolitan regions of Quebec: the Montreal area (MTL), Quebec City (QCC) and the Saguenay region (SAG) (Figure 1). Regional origins of the individuals were validated with a principal component analysis (PCA) of genetic diversity using genotypic data and including individuals from the Reference Panel of Quebec (RPQ) (Roy-Gagnon et al. 2011). The Saguenay population stands out in this graphical analysis whereas MTL and QCC cluster with the CEU population from HapMap3. Pairwise F_{st} between regional populations was computed using genotyping SNPs and shows little differentiation between CEU, MTL and SAG, although SAG is more differentiated than CEU and MTL (Figure S1). This likely results from the very recent regional founder effect that occured in the Saguenay region. This territory was colonized during the 19th century by a reduced number of settlers, who contributed massively to the genetic pool of individuals living in this region today (Bherer et al. 2011).

Recombination Rates, Coldspots and High Recombination Regions

We examined patterns of linkage disequilibrium (LD) and inferred recombination rates for the regional populations of Quebec. We compared these patterns with those found in CEU and YRI populations from HapMap3 (Material and Methods).

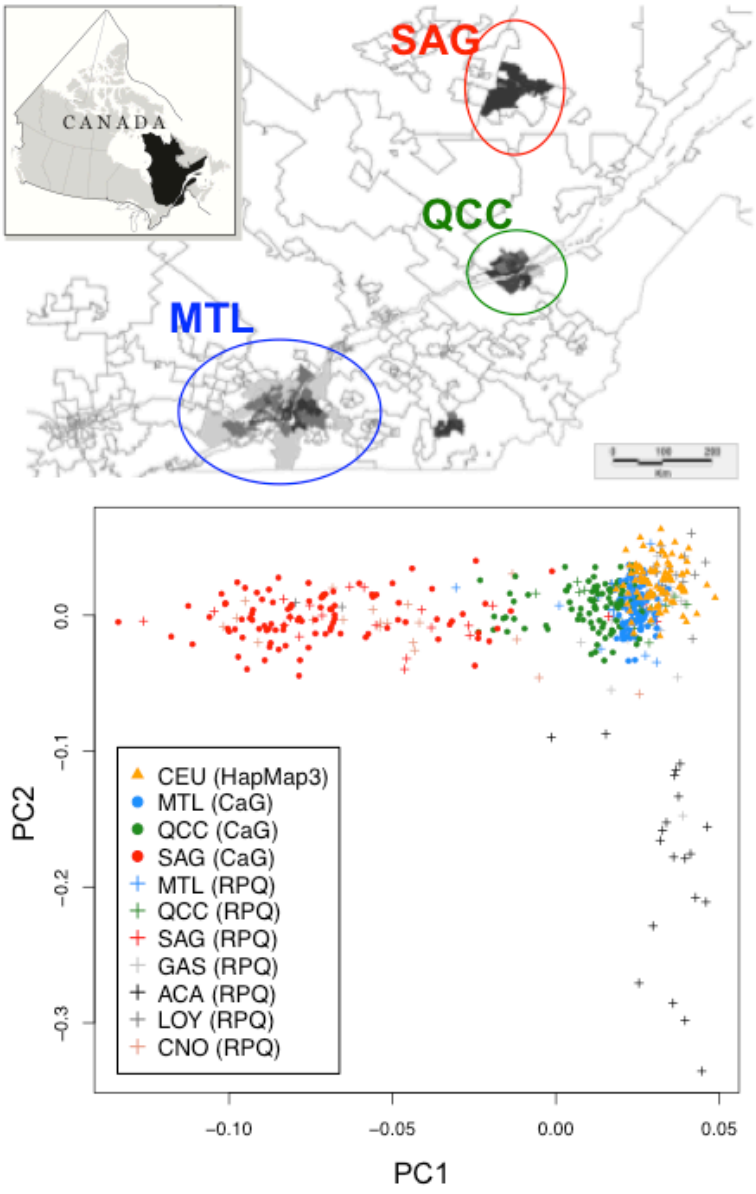


Figure 1. CARTaGENE sampling in three regional populations of Quebec

Sampling includes individuals from the Montreal area (MTL), Quebec City (QCC) and the Saguenay region (SAG). Regional origin of individuals was confirmed by a principal component analysis of genetic diversity in CARTaGENE (CaG) individuals compared with genetic diversity within the Reference Panel of Quebec (RPQ) and in the CEU population from HapMap3. Other populations included in the RPQ are : GAS : Gaspesia Region, ACA : Acadians, LOY : Loyalists, CNO : North Shore Region. [Canada map from http://atlas.gc.ca]

For patterns of LD, no notable differences were observed between MTL and CEU whereas SAG shows slower LD decay, more linkage between distant SNPs and a smaller number of small LD blocks (<5Kb) (Figure 2A and 2B). These patterns are likely due to the recent demographic history of the region. Genetic maps for all chromosomes were computed based on population recombination rate estimates, converted into centiMorgan (cM) per Mb using the deCODE genetic map (Kong et al. 2010) (Material and Methods). As expected, the concordance of genetic maps between populations is extremely high (Figure 2C), although they differ slightly from the HapMap genetic map (HapMap Consortium et al. 2007), computed with the 2002 version of the deCODE map.

We used these genetic maps to locate coldspots and hotspots of recombination. Coldspots are defined as non-centromeric regions of more than 50 Kb with recombination rates consistently lower than 0.5 cM/Mb in CaG, CEU and YRI populations (Material and Methods) and are likely shared among human populations. We obtained a list of 7,851 autosomal coldspots, with a mean size of 133.4 Kb, spanning about a third of the human genome, for a total of 1.049 Gb (Table S1). A hotspot is defined as a short segment (<15Kb) with recombination rates falling in the 90th percentile (> 5 cM/Mb). The vast majority of hotposts are shared, but occasional differences exist between CEU and CaG populations (Figure 2D) and between SAG and MTL/QCC (Supplementary results). Finally, we define high recombination regions (HRRs) as regions with a high density of hotspots, such that the distance separating neighbouring hotspots (>5 cM/Mb) is smaller than 50 Kb. We identified 12,500 HRRs genome wide shared between CaG, CEU and YRI populations, with a mean size of 50.74 Kb, covering a total of 634.2 Mb (Table S1). The definition of coldspot, hotspot and HRR are illustrated in Figure S2.

Increased Diversity at Nonsynonymous Positions in Coldspots

Regions of low crossing-over have reduced levels of neutral genetic variation in humans and other species (Begun and Aquadro 1992; Nachman 2001). This is likely

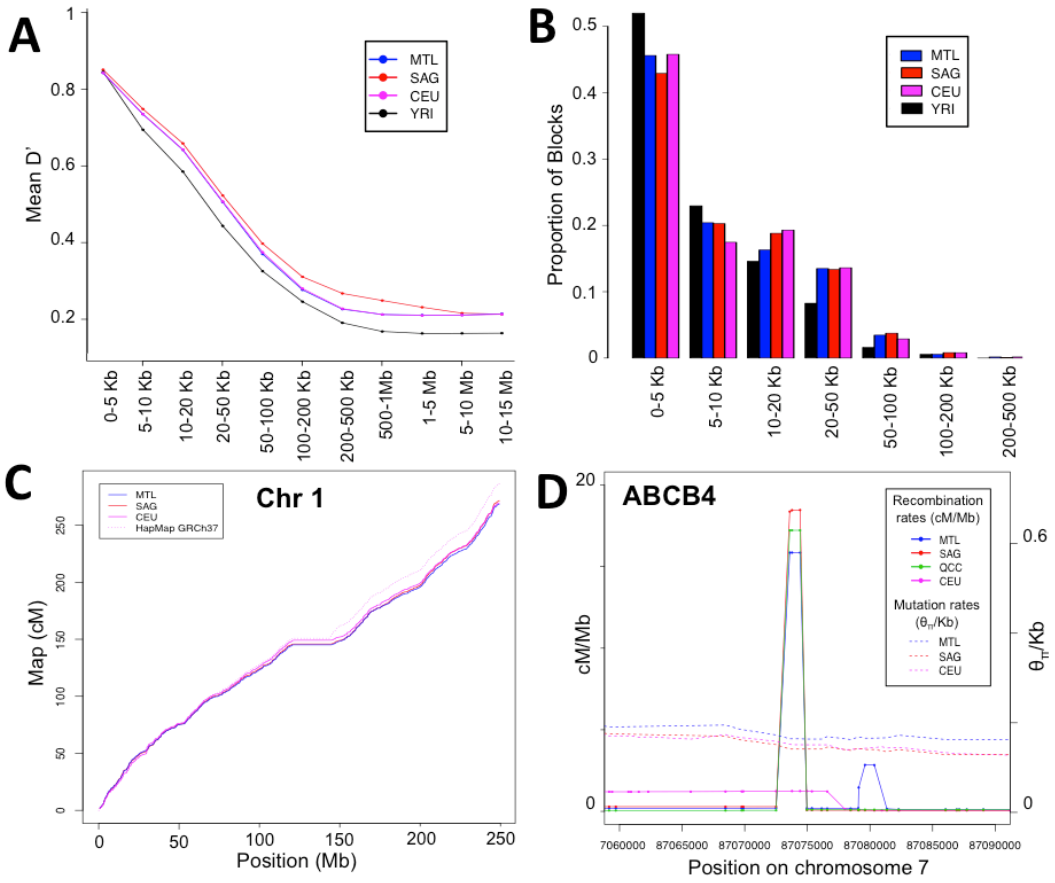


Figure 2. Patterns of linkage disequilibrium and recombination in French-Canadians.

(A) Average LD over the genome for CaG populations (MTL : Montreal area, QCC : Quebec city, SAG : Saguenay region) and CEU and YRI populations from HapMap3. (B) Distribution of linkage disequilibrium (LD) block length across populations. LD blocks and mean D' were computed using 96 individuals from each population with Haploview. (C) Graphical representation of the Chromosome 1 genetic map for MTL, SAG and CEU populations computed using recombination rates inferred by LDhat and the deCODE 2010 map, compared to the currently available HapMap genetic map. (D) Hotspot of recombination detected in CaG populations and not observed in CEU population. The mutation rate estimated by θ_π is not markedly different between populations.

due to stronger linkage between neutral variation and strongly selected mutations in low recombination regions. Using RNA sequencing SNPs, called in exons with high coverage at all positions (Material and Methods), we observed a SNP density in coldspots of 4.46 SNP/kb, which is significantly lower than the SNP density outside coldspots (5.08 SNP/kb, Odds Ratio (OR) = 0.87 [0.864;0.889 95%CI]) and within HRR (7.30 SNP/kb, OR = 0.61 [0.596;0.620 95%CI]). This result holds for variants at all minor allele frequency (MAF) classes, although the effect is smaller for low frequency variants (Figure 3A). We further observed a significant positive correlation between SNP density and mean recombination rate per exon (β=0.073, $p<2\times10^{-16}$, Table S2), after accounting for GC-content, average gene expression and exon size, indicating that these potential confounding factors do not fully account for the observed correlation between recombination and SNP density.

The consequence of a mutation on the protein sequence is a measure of the deleterious impact of a variant. Many studies have reported that the majority of nonsynonymous mutations are weakly deleterious and are expected to reduce fitness (Eyre-Walker et al. 2006; Boyko et al. 2008; Li et al. 2010). We annotated all SNPs in highly covered exons from the RNA sequencing data using public databases (Material and Methods) to identify sites that are putatively functional and classified them either as synonymous or nonsynonymous, and for nonsynonymous, as missense or nonsense. Next, for each given functional category, we compared the fraction of mutations between coldspots and HRRs using Odds Ratios (ORs) (Material and Methods). We observed an excess of nonsynonymous mutations in coldspots relative to HRR (Figure 3B). The corollary is an excess of synonymous mutations in HRRs relative to coldspots. This effect is mainly due to missense mutations, as they represent the majority of nonsynonymous mutations. On the other hand, the observed nonsense mutations were not significantly overrepresented in coldspots, although a relatively low number of nonsense were called (n = 363), which may affect the power to detect a significant enrichment. Overall, these results indicate that, despite the overall decreased diversity in low recombination regions, SNPs in

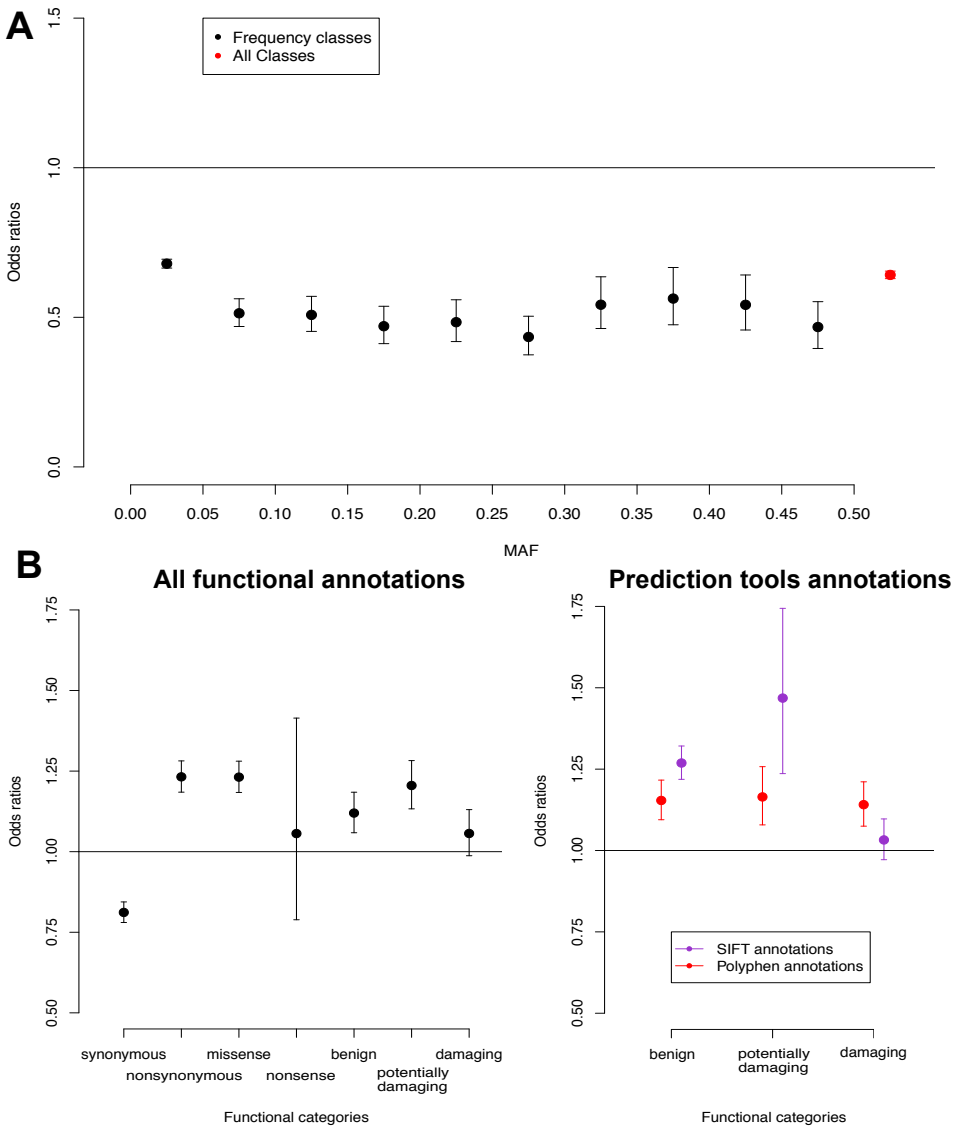


Figure 3. Comparison of diversity and functional classes of mutations between coldspots and HRRs.

Differential load is computed using Odds Ratios (ORs). OR < 1 corresponds to an enrichment in HRRs relative to coldspots, OR > 1 corresponds to an enrichment in coldspots relative to HRRs (A) OR comparing SNP density between coldspots and HRRs. (B) OR comparing the proportion of SNPs in each functional category between coldspots and HRRs. SIFT and Polyphen were used to predict the impact of missense. OR for consensus predictions (Material and Methods) are reported in the left panel. OR for SIFT and Polyphen predictions separately are reported in the right panel.

coldspots are more likely to impact protein function than SNPs in regions of high recombination.

The functional impact of missense mutations are either benign, when they do not affect the protein function, or damaging when they lead to appreciable protein changes. Polyphen (Adzhubei et al. 2010) and SIFT (Kumar et al. 2009) are widely used computational approaches to predict amino acid changes affecting protein function and activity. We used a combination of the two to reduce the number of false positives (Material and Methods) and annotated all missense variants as "Benign" (when both methods predict the mutation as being benign/tolerated), "Damaging" (when both methods predict the mutation as being damaging) or "Potentially Damaging" (when only one method predicts the mutation as being damaging). We compared these groups of mutations in our dataset and observed that benign and potentially damaging mutations are seen in excess in coldspots relative to HRRs, whereas damaging mutations are not enriched in coldspots (Figure 3B).

These results suggest that damaging mutations are likely removed from sequences with the same efficiency in low and high recombination regions and that the strength of purifying selection acting on highly deleterious mutations is similar between coldspots and HRR. However, the overrepresentation in coldspots of other mutations causing a change in the amino acid, which are likely to be weakly deleterious, indicates that these variants are less efficiently removed from sequences in regions of low recombination.

Enrichment of Highly Conserved Mutations in Coldspots

Genomic regions that remain conserved across multiple species are likely to be involved in essential cellular functions and have previously been used to identify putative functional sequences in the human genome (Pennacchio and Rubin 2001; Boffelli et al. 2003). Therefore, the level of conservation across multiple species at a given nucleotide is another measure of the functional impact of a mutation arising at

this position. We used conservations scores caculated by PhyloP (Siepel et al. 2006) and GERP (Davydov et al. 2010) to estimate the level of constraint at all positions and at all SNPs in the sequenced exons (Material and Methods). Positions with high GERP and PhyloP values represent sites that have accumulated fewer substitutions than expected under a neutral rate of evolution.

In the human genome, regions of strong linkage disequilibrium are more conserved than regions exhibiting high recombination rates, likely because low recombination regions are enriched with highly conserved genes with essential cellular functions (Smith et al. 2005). Indeed, coldspots show an excess of positions with high values of both GERP and PhyloP relative to positions in HRRs (Figure 4A). To evaluate if a mutated position is more likely to be found at a conserved position in coldspots, we thus need to take into account the baseline excess of conserved positions in coldspots compare to HRRs (Material and Methods, Figure S3). Interestingly, the excess is significant at mutated positions, as there is an enrichment of SNPs with GERP scores above 3 in coldspots compare to HRR after correcting for the baseline effect (Figure 4B).

SNPs in coldspots also tend to have higher PhyloP scores than SNPs in HRRs, however the enrichment is marginally significant for extreme PhyloP scores (>5). Because damaging mutations were not observed in excess in coldspots (Figure 3C), this result is likely due to an enrichment of damaging mutations in the extreme PhyloP class (Supplementary results). However, we note that the majority of damaging mutations (63%) are found to have PhyloP scores ranging from 1 to 5 (Table S3). These results indicate that positions under strong evolutionary constraint are more likely to be polymorphic in coldspots than in regions of high recombination. Therefore, mutations at conserved sites, which are likely deleterious or weakly deleterious, appear to be less efficiently removed from sequences by negative selection in coldspots than in HRRs.

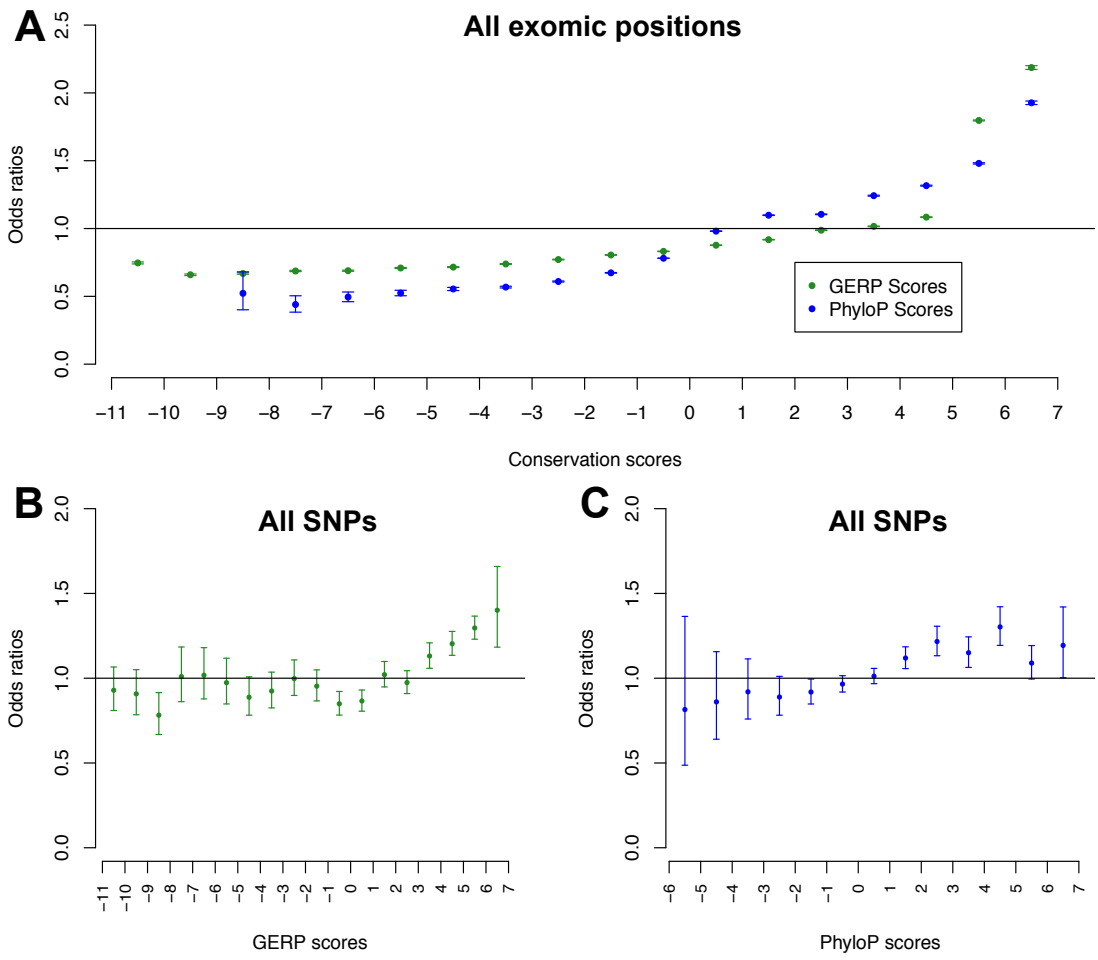


Figure 4. Comparison of conservation scores between coldspots and HRRs exomic positions and SNPs.

Differential load is computed using Odds Ratios (ORs). OR < 1 corresponds to an enrichment in HRRs relative to coldspots, OR > 1 corresponds to an enrichment in coldspots relative to HRRs (A) OR comparing the proportion of positions in each category of GERP and PhyloP scores between coldspots and HRRs. (B-C) OR comparing the proportion of SNPs in each category of GERP (B) and PhyloP (C) scores between coldspots and HRRs. OR were computed for categories including more than 30 SNPs.

Frequency-based Measures of Mutational Load

Low frequency variants represent the majority of genetic variants in human populations (Keinan and Clark 2012) and are enriched with functional mutations that affect protein function (Marth et al. 2011; Nelson et al. 2012). For example, variants segregating at nonsynonymous sites tend to be at lower frequency than those at synonymous sites, regardless of the function of the gene (Blekhman et al. 2008). Furthermore, a strong inverse relationship has been described between evolutionary constraint and allele frequencies of mutations segregating in human populations (Cooper et al. 2010; Goode et al. 2010; Hodgkinson et al. 2013). Therefore, the frequency of an allele segregating at a site is generally a good indicator of its functional importance.

The enrichment of variants at nonsynonymous positions and at conserved sites in coldspots is driven almost entirely by alleles with minor allele frequency (MAF) below 0.01 (Figure 5). Furthermore, we confirm the previously reported positive correlation between average MAF per exon and recombination rates (Lohmueller et al. 2011), which is significant after correcting for GC-content, average gene expression and exon size (Table S2). We also examined the impact of recombination rates on the accumulation of novel variants and singletons (Figure 5B). Novel variants are new variants that have neither been reported in dbSNP (Sherry et al. 2001) nor been seen in genomic data from the 1000 genome project (The 1000 Genomes Project Consortium 2010). These variants are at very low frequency and likely arose as private mutations in founders or originated *de novo* since the founding of the French-Canadian population. Novel variants are enriched in coldspots, suggesting that the reduction in the efficiency of negative selection in coldspots may be observed over very brief evolutionary time scale, as short as fifteen to twenty generations (Roy-Gagnon et al. 2011). Singletons are novel variants seen only once in the population and are not enriched in coldspots of recombination, consistent with the idea that these mutations correspond to the

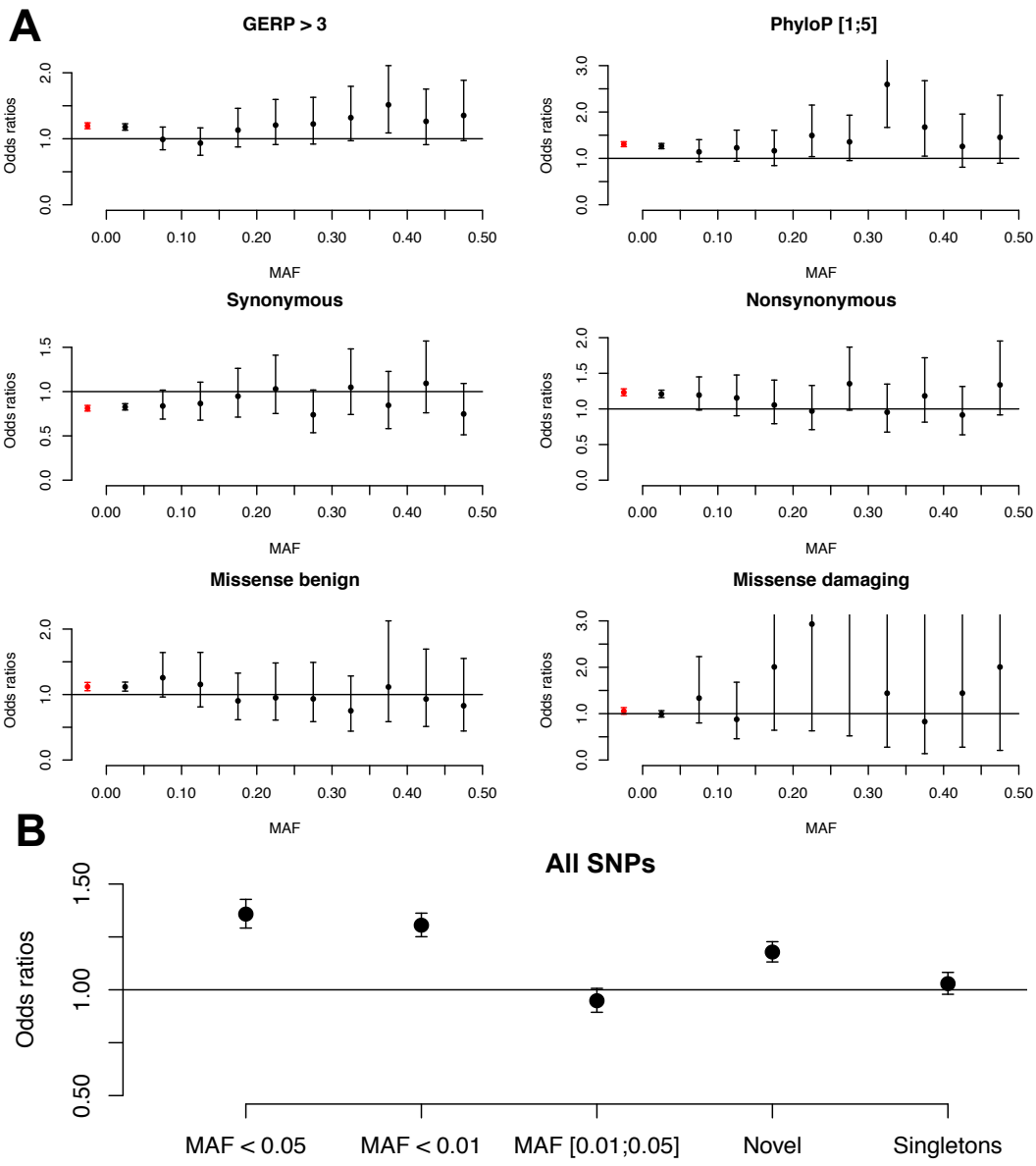


Figure 5. Differential mutational load based on frequency-based statistics

Differential mutational load is computed using Odds Ratios (ORs). OR < 1 corresponds to an enrichment of SNPs in that frequency classes in HRRs relative to coldspots, OR > 1 corresponds to an enrichment in coldspots relative to HRRs. (A) OR in each minor allele frequency class for SNPs in functional and conservation categories. (B) OR in frequency-based categories for all SNPs. Novel SNPs are SNPs not observed in dbSNP and 1000 Genome Project. Singletons are novel SNPs seen only once in the CaG sample.

most recent mutations in the population, on which natural selection is unlikely to have had enough time to act, even in HRR. In line with this observation, a recent study looking at mutations in 14002 individuals revealed that little difference between the proportions of variants in intronic, untranslated exonic, synonymous and nonsynonymous classes are seen when minor allele count is low (Nelson et al. 2012). The most recent class is formed by *de novo* mutations, which are not expected to occur preferentially in coldspots if they are randomly distributed in the genome. Also, singletons are likely enriched with sequencing errors that occur at random in the genome and are unlikely to cluster in regions of low or high recombination. Alternatively, mutations that are highly damaging in the homozygous state will be maintained at very low frequencies by purifying selection and may be observed as singletons in our sample.

Robustness to Potential Confounding Factors

Altogether, our analyses indicate that SNPs in coldspots are enriched in low frequency and nonsynonymous variants and are found at constraint positions moreso than SNPs in HRRs. This result implies that the efficiency at which selection removes slightly deleterious variants depends on recombination rate variation in the human genome. However, other genomic features that correlate with either recombination or efficiency of selection may influence this effect as well. We evaluated the linear correlation between recombination rates and SNP density per exon for nonsynonymous, damaging, novel, singletons and high conservation score variants. We accounted for potential confounding factors such as GC-content, average gene expression levels, exon size and total SNP density (Table S2). Recombination rates correlates negatively with the density of nonsynonymous, novel and high conservation score variants at the exon level, which validates our previous observations. Importantly, we verified that our results are robust to a wide array of recombination rate thresholds for defining coldspots and HRRs (Table S4).

We further considered two main confounding factors: GC-content and gene expression. GC-content is known to correlate with recombination rate in many taxa, including humans (Meunier and Duret 2004). Exons were ranked based on their GC-content and stratified into four categories of equal size and analyses were performed separately for each category, so that coldspots and HRRs contain similar proportion of GC (Table S5). The higher mutational load observed in coldspots remains significant for all GC categories, with the exception of high conservation score and missense mutations that are not significantly enriched in coldspots for the highest GC-content exons. This result may reflect a lack of power in this category, given that coldspots with high GC-content are likely to be quite rare.

Similarly, we separated exons in four categories according to their average gene expression levels across the entire population sample (Material and Methods). Expression level of a gene is one of the best predictors of its evolutionary rate (Pal et al. 2006) with the efficiency of selection being weaker in lowly expressed genes. Furthermore, within-gene recombination rates appear to correlate with transcription patterns, such as expression breadth and allelic expression (Necsulea et al. 2009). Indeed, we observed a very weak but significant negative correlation between recombination rates and mean gene expression in our data (Table S2), providing further support for a negative association between recombination and transcription in humans (Necsulea et al. 2009). Despite this, the increased mutational load seen in coldspots remains significant for all gene expression categories (Table S5). Singletons, that were found not to be in excess in coldspots in previous analyses, are significantly enriched in coldspots for regions of low gene expression and high GC-content. This result may reflect the heterogeneity of this class of mutations that likely contains *de novo* mutations and sequencing errors as well as deleterious mutations maintained at very low frequencies.

Coldspot Mutational Load in Regional Populations and Per Individual

Despite their short divergence time and relatively similar geographic and environmental range, the regional populations of Quebec are stratified and differences in disease prevalence exist (Scriver 2001). In particular, the Saguenay population is famous for the discovery of genes underlying Mendelian disorders because of the high carrier rate for many recessive mutations. We computed the differential mutational load in the regional populations separately. Overall, all three populations show an increased mutational load in coldspots, based on functional annotations, conservation scores and frequency-based measures (Figure 6A), however no significant differences are seen between populations. One interesting difference, however, is that the SAG and QCC population show a significant excess of mutations predicted as damaging in coldspots whereas the MTL population does not.

To obtain a better picture of the differences between regions, we computed the odds ratios between coldspots and HRR per individual (indOR) for all mutational load measures (Table 1). Most individuals show an enrichment of missense, low frequency and high conservation score variants, showing that SNPs in coldspots are likely to affect fitness at the individual level. Interestingly, a large proportion of individuals show an excess of damaging mutations in coldspots, which contrasts with the general result at the population level, where damaging mutations were not significantly enriched in coldspots overall (Figure 3). This observation suggests that the accumulation of damaging mutations in coldpots may be limited to a subset of individuals in the population.

Next, we tested whether regional populations have different distributions of indOR using a non-parametric Mann-Whitney U-test. Significant differences are observed between SAG and the other populations for indOR computed for frequency-based measures. To understand the nature of these differences, we plotted the indOR distributions for low frequency (MAF<0.01), novel and private variants in the three

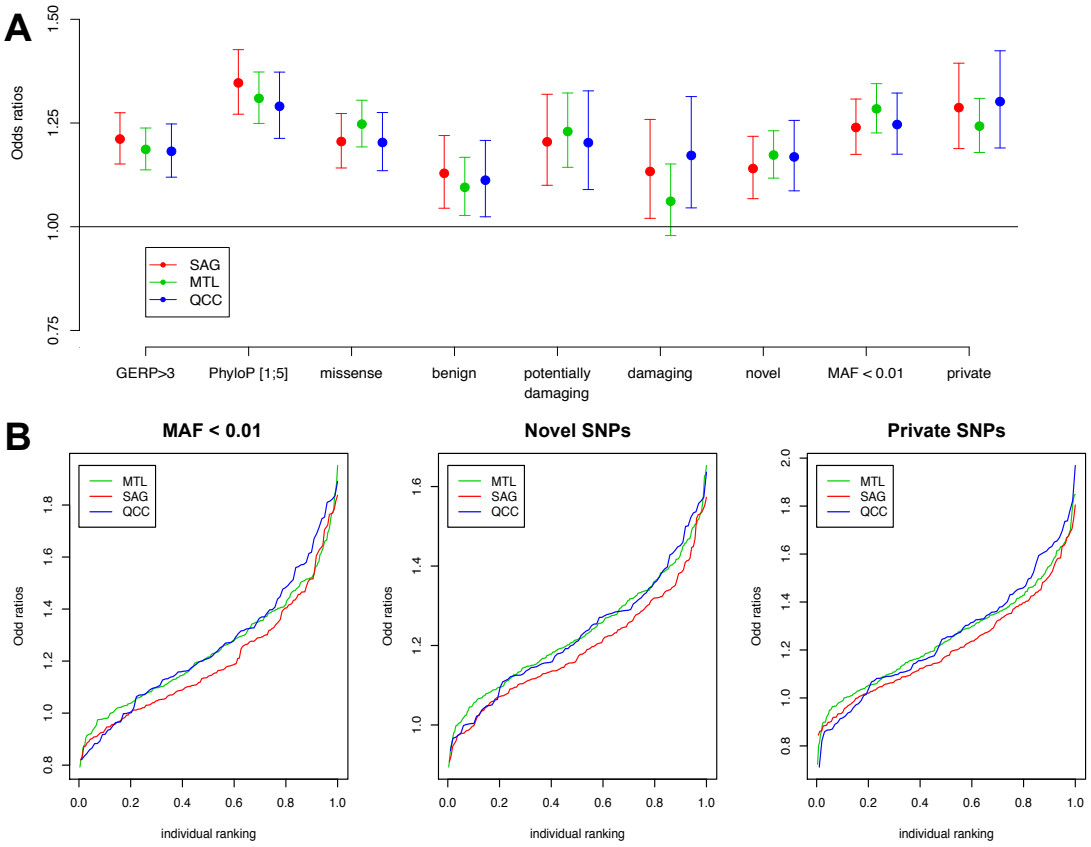


Figure 6. Differential mutational load in CaG regional populations

Populations : the Montreal area (MTL), Quebec city (QCC) and Saguenay region (SAG). Differential mutational load is computed using Odds Ratios (ORs). OR < 1 corresponds to an enrichment of SNPs in that frequency classes in HRRs relative to coldspots, OR > 1 corresponds to an enrichment in coldspots relative to HRRs. (A) OR based on conservation, functional and frequency-based annotations. (B) Distribution of OR per individuals (indOR) in the three regional populations.

Table 1. Per-individual differential mutational loads

Differential mutation load is computed using Odds Ratios (ORs) based on conservation, functional and frequency-based annotations, in the CaG populations together (All) and separately. OR < 1 corresponds to a significant enrichment of SNPs in that category in HRRs relative to coldspots, OR > 1 corresponds to a significant enrichment in coldspots relative to HRRs. OR = 1 corresponds to no significant differences between coldspots and HRRs. For each mutational category, we tested if the distributions of OR per individuals are different between CaG regional populations.

Mutation load	All			MTL			SAG			QCC			P-values for differences between regions					
measures	OR>1	OR=1	OR<1	OR>1	OR=1	OR<1	OR>1	OR=1	OR<1	OR>1	OR=1	OR<1	MTL vs SAG		QCC vs SAG		MTL vs QCC	
GERP >3	460	61	0	236	28	0	141	17	0	83	16	0	0.273		0.605		0.821	
PhyloP [1;5]	483	38	0	249	15	0	144	14	0	90	9	0	0.215		0.474		0.934	
Missense	484	37	0	245	19	0	148	10	0	91	8	0	0.324		0.535		0.783	
Nonsense	4	513	4	3	258	3	0	158	0	1	97	1	0.086		0.224		0.616	
Damaging	263	258	0	151	125	0	69	89	0	57	42	0	0.312		0.305		0.37	
MAF < 0.01	365	130	26	190	66	8	99	50	9	76	14	9	0.018	*	0.043	*	0.827	
Novel	401	118	2	208	55	1	115	42	1	78	21	0	0.006	**	0.018	*	0.796	
Private	330	181	10	175	85	4	91	65	2	64	31	4	0.047	*	0.12		0.976	
Singletons	72	445	4	36	226	2	20	137	1	16	82	1	0.417		0.71		0.7	

MTL : Montreal area, QCC : Quebec city, SAG : Saguenay region.
P-values computed using a Mann-Whiney U test. ** p<0.01; * p<0.05

regional populations (Figure 6B). The curves from the SAG population are consistently below the ones from populations in Metropolitan Quebec (MTL and QCC), suggesting a difference in the amplitude of the effect observed. This result may be explained by the difference in effective population size (N_e) between regions of Quebec, due to the different demographic histories. In particular, the population in the Saguenay region originated from a smaller number of founders (Bherer et al. 2011). We computed N_e by regional population from recombination rate estimates (Supplementary Results, Table S6), and found N_e to be significantly reduced in SAG relative to MTL and QCC. A smaller N_e implies a reduction in the efficiency of selection, impacting both low and high recombination regions. This observation may therefore explain the slightly smaller indOR in SAG, since the efficiency of selection is also reduced in high recombination regions.

Increased linkage of rare and deleterious variants in coldspots

The results presented in Table 1 show that slightly deleterious mutations accumulate preferentially in coldspots at the individual level. Because recombination is limited and does not redistribute variants among haplotypes in these regions, the theory predicts that many deleterious mutations will be found linked to each other on the same chromosome (Felsenstein 1974). To test this hypothesis, we obtained “mini-haplotypes” by extracting pairs of SNPs found on the same sequencing read or read pair, which are hence phased (Material and Methods, Figure S4). Separately for variants in coldspots and HRR, we compared the number of mini-haplotypes where two minor alleles are linked to each other with the number of mini-haplotypes where a minor (Min) and a major (Maj) allele are coupled together. Because intermediate frequency alleles at neighbouring SNPs are more likely to be found linked to one another than low frequency alleles, we looked at haplotypes with pairs of rare variants, with MAF ≤ 0.01.

Table 2. Differential load of mini-haplotypes in coldspots (CS) and HRRs.

Rare Variant Type	OR CAG	95%CI	Within CS		Within HRR	
			Min/Min	Min/Maj	Min/Min	Min/Maj
all rare	**2.64**	[2.38;2.93]	1782	71026	453	47706
conserved	**8.04**	[6.25;10.35]	1433	22187	64	7970
nonsyn	1.01	[0.89;1.15]	750	27526	398	14762
damaging	**13.50**	[1.79;101.99]	16	1709	1	1443

Number of mini-haplotypes with two minor alleles linked to each other (Min/Min) and with a minor allele linked to a major allele (Min/Maj) for rare variants (with MAF<0.01) in coldspots (CS) and HRRs. Odds Ratios (ORs) with confidence intervals were computed to compare the proportions of Min/Min between CS and HRRs. OR < 1 corresponds to an enrichment of Min/Min mini-haplotypes in HRRs relative to coldspots, OR > 1 corresponds to an enrichment in coldspots relative to HRRs.

Mini-haplotypes are 2.64 times more likely to carry two paired rare alleles rather than a rare allele paired with common allele in coldspots compared to HRRs (Table 2). This excess of Min/Min haplotypes for low frequency variants in coldspots indicates that selection is unable to efficiently remove these haplotypes in coldspots. Furthermore, coldspots are strongly enriched for mini-haplotypes with two rare alleles at conserved positions (OR = 8.04) compared to HRRs, where there is a severe lack of haplotypes with two rare mutations at conserved sites. A larger proportion of mini-haplotypes with two rare damaging mutations were found in coldspots compared to HRR (OR = 13.50), although these mini-haplotypes are very unfrequent. These results directly show that rare and weakly deleterious mutations arising on the same haplotype accumulate in coldspots and are not removed as efficiently as in HRRs.

Replication in the 1000 Genomes Project Populations

In order to replicate our findings, we performed all analyses using SNPs called from exome data in the African, Asian and European populations of the 1000 Genomes

Project (Material and Methods) in our selected exons. Overall, the results remain the same as the one seen in the CaG population, with categories of mutations enriched in slightly deleterious mutations overrepresented in coldspots (Figure 7). However, differences in the amplitude of the effect are observed between populations, with African exomes showing the weakest effects and the French-Canadian data showing the more pronounced effect, suggesting that demography is affecting this effect. We also compared the distribution of indOR between populations (Figure 8). For indOR computed for rare variants, the distribution differ significantly between populations, with the mean indOR for the 'out of Africa' populations being shifted to the right. We also observe an increased variance in the CAG population relative to the 1000 Genomes populations. For indOR, the distribution are similar among populations from the 1000 Genomes Project but the CAG dataset shows a shift in mean to the right, suggesting that the French-Canadian population has an increased mutational burden in coldspots relative to other populations worldwide. This effect may be due to greater variance in French-Canadian differential mutation load due to demographic events or relaxation of selection following the founding of the population, but it might also be caused by the differences in DNA vs. RNA mutational load. Finally, we replicated the mini-haplotype analyses in populations from the 1000 Genomes Project (Material and Methods) and very similar results were obtained, with mini-haplotypes with two rare variants linked together enriched in coldspots (Table S5).

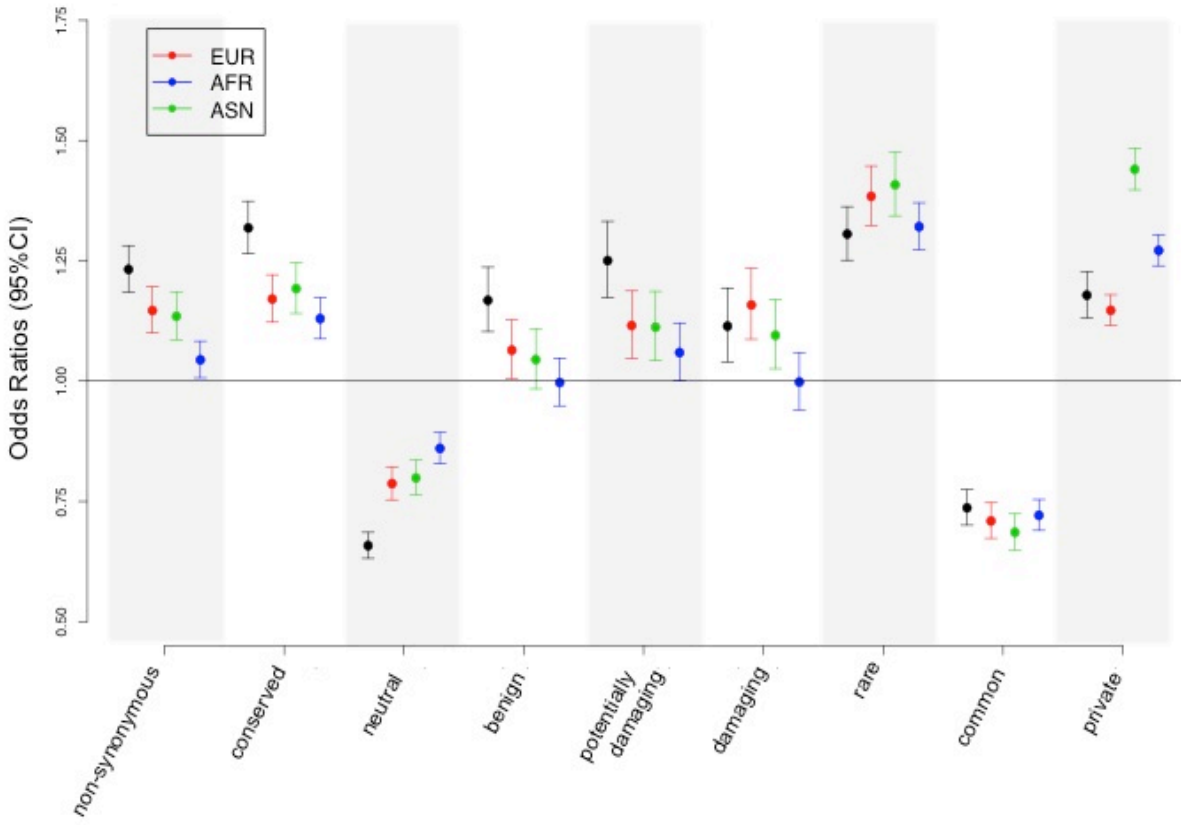


Figure 7. Differential mutational load in populations from The 1000 Genomes Project.

Neutral SNPs are synonymous variants at non-conserved positions. Common variants have MAF>0.05 and rare variants have MAF<0.01.

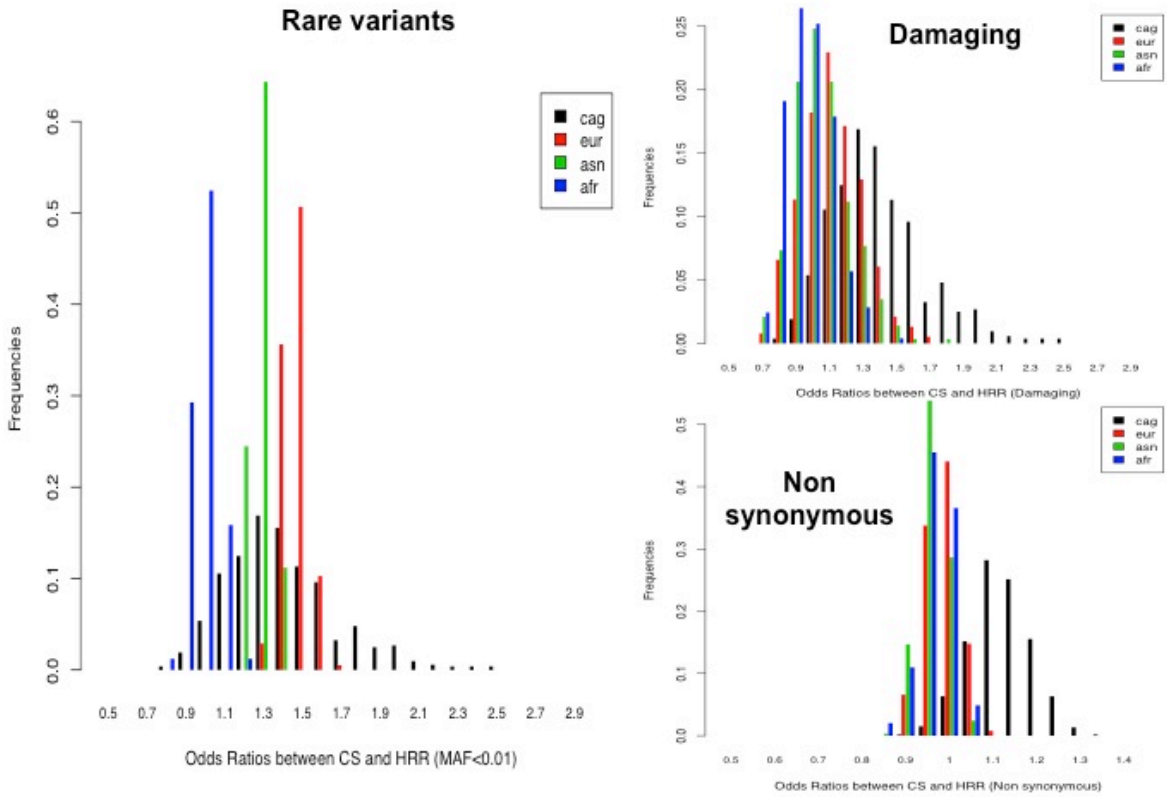


Figure 8. Distribution of individuals Odds Ratios (indOR).

In each population, indOR are computed for rare, damaging and non-synonymous variants in the CaG population and the 1000 Genomes population from Europe, Africa and Asia.

DISCUSSION

The genomics era has greatly improved our ability to estimate the abundance of deleterious mutations in a population and within individuals. In this study, using SNPs called from RNAseq data in 521 French-Canadians and several mutational load measures, we demonstrated empirically that, although diversity is reduced in coldspots of recombination, mutations in these low recombining regions are likely to have a stronger negative impact on fitness than mutations segregating in high recombination regions in the human genome. This indicates that negative selection works less efficiently at removing weakly deleterious mutations in regions of low recombination. To our knowledge, this is the first empirical demonstation that the efficiency of selection varies across the human genome owing to the variation in the local recombination rate. Generally, these results show that patterns of recombination rates in humans will likely alter the rate at which functional mutations accumulate in individuals.

In our analyses, we used three different types of measure to estimate the amount of deleterious mutations: functional annotations, conservation scores and frequency-based statistics. For each of these categories, we computed the differential mutational load between high and low recombination regions. We observed that low-frequency variants and nonsynonymous alleles are found significantly enriched in low recombination regions. Furthermore, SNPs in these genomic regions are found at constraint positions moreso than SNPs in high recombining regions, even after correcting for the overall higher conservation of coldspots in the human genome. Importantly, these observations remain valid when controlling for potential confounding, such as gene expression levels and GC-content. The effect observed is also robust to the recombination parameters used to define low and high recombining regions. Because evidence for the reduction in the efficiency of selection mainly come from non-recombining chromosomes (Charlesworth and Charlesworth

2000) and organelle genome (Lynch and Blanchard 1998), we favored an approach that contrasts coldspots of recombination and highly recombinogenic regions in the human genome. However, we detected a weak but significant linear correlation between recombination rates and densities of deleterious alleles after correcting for GC-content and gene expression levels, although the reduction in the efficiency of selection is unlikely to be linear in the recombination rate.

We used prediction tools, widely used in medical genetics, to assess the potential impact of a mutation on the function of a protein. Mutations that were predicted as damaging were found to be differentially distributed between coldspots and HRRs in in two out of three French-Canadian populations. Furthermore, at the individual-level, damaging mutations are enriched in coldspots half of the individuals (263/521, Table 1). The reduction in the strengh of selection may also affect the amount of damaging mutations segregating in coldspots. Mutations with harmful consequences accumulate preferentially in these genomic regions and per-individual differential mutational loads for damaging mutations are significantly higher in individuals in the French-Canadian population compared to other populations world-wide (Figure 8).

Singletons are not significantly enriched in coldspots in our data. These are novel variants seen only once in our sample. They therefore represent the most recent class of mutations and include *de novo* mutations, on which selection is unlikely to have had enough time to act. It has previously been suggested that double-strand break may generate mutations, but if this were the case, an enrichment of singletons in high recombination regions would have been expected. This observation adds further support to the view that recombination is not mutagenic and that the correlation between recombination and diversity is driven by natural selection (Lohmueller et al. 2011).

Muller pointed out that mutations will tend to accumulate on non-recombining genomes in finite populations, and described this mechanism as a ‘ratchet’ (Muller 1964). This accumulation is due to the effect of genetic drift that can cause the loss of

haplotypes with the lowest number of weakly deleterious mutations despite their higher fitness. The 'least-loaded' class may only be reconstituted by recombination. In the absence of recombination, there will be a gradual and irreversible buildup of deleterious mutation, and the loss of each least-loaded class can be seen as a turn of the ratchet. This argument also applies to any non-recombining genomic region. In regions where recombination rates are extremely low, it is hard to predict how pronounced this process would be. Here, we showed not only that human coldspots are enriched in deleterious variants overall, but also that the variants accumulate linked to each other on the same haplotype, in line with predictions from Muller's ratchet mechanism.

Because the buildup of negative linkage disequilibrium (when deleterious mutations are linked to beneficial ones) is inevitable in this situation, Muller's ratchet also predicts, in the long run, an increased rate of fixation of deleterious alleles under particular circumstances. This is because every deleterious mutation linked to a beneficial one driven to fixation by positive selection, will be swept to fixation with it. Under this model, both the mutational load and the rate of fixation of deleterious variants should be increased and the rate of fixation of advantageous mutation should be decreased. However, regions of strong linkage disequilibrium in the human genome have been found to be more conserved than regions exhibiting high recombination rates (Smith et al. 2005). Our results confirmed this finding, as positions in coldspots are significantly more conserved across multiple species than positions in HRR (Figure 4A). Furthermore, Bullaughey and colleagues found no correlation between recombination rates and the ratio of the rates of nonsynonymous and synonymous substitution (d_N/d_S) (Bullaughey et al. 2008). These observations suggest that the accumulation of deleterious mutations in coldspots does not result in a higher rate of fixation of nonsynonymous substitutions.

Several factors may explain why the increased mutational load in coldspots does not result in an increased rate of fixation. First, these genomic regions still have a small

amount of recombination. In the CaG population, the mean recombination rate across coldspots is 0.1348 cM/Mb [0.1269-0.1359 95%CI]. Although this value is very low, it is possible that it is high enough to prevent deleterious alleles to increase in frequency up to fixation. Likewise, it is possible that the fine-scale recombination rates evolve rapidly enough to break down negative associations of alleles before deleterious alleles are fixed. Both these hypotheses are testable by simulations. Finally, it has been demonstrated that when mutations are close enough to recessivity (i.e. they have low dominance coefficients) in diploid individuals, the decoupling of fixation and turns of the ratchet is likely to occur (Charlesworth and Charlesworth 1997). Furthermore, if the reduction in homozygous fitness is sufficiently severe, the probability of fixation will be decreased to a negligible level, since the selective advantage of favorable alleles will not be able to counter balance the harmful effects of these deleterious mutations.

Interestingly, most of the genes in low recombination regions are implicated in essential cellular processes, and are responsible for many genetic diseases in humans (Smith et al. 2005). The accumulation of slightly deleterious mutations in genes that are primordial for response to DNA damage or cell cycle progression, for example, can ultimately be detrimental to the health of individuals and are expected to explain an important fraction of the genetic etiology of human disease. Therefore, the effects observed probably result from different evolutive pressures. On one side, purifying and background selection will dominate in genes with essential cellular functions. These essential genes have been under selection for few billions of years to improve the performance of the biological machinery and may not need any more improvement through generation of new haplotypes. These genes show low recombination rates and are therefore less exposed to events of unequal crossing-over that would cause major mutations (such as large insertions, deletions, inversions or translocations) likely deleterious to the survival of the individual. Recombination thus needs to be redirected away from these coding sequences. The down side of this effect is the irreversible accumulation of weakly deleterious mutations. Each

individual will thus receive a burden of unfavorable mutants, which continually increases over successive generations in the absence of recombination. If selection against these deleterious mutations is very weak, these variants may even drift to intermediate frequencies. Higher recombination rates, on the other hand, will break down random associations of alleles, will reduce the average number of unfavorable mutations segregating in the population and will prevent weakly deleterious variants from accumulating on the same haplotypes. These results further imply that selective sweeps in humans are more likely to occur in high recombination regions, where advantageous alleles are more efficiently decoupled from deleterious alleles, and arise on haplotypes that did not accumulate high amount of deleterious alleles.

MATERIAL AND METHODS

Ethics Statement

The ethics committee of CARTaGENE approved the study and all participants gave their informed consent. The study was in accordance with the principles of the current version of the Declaration of Helsinki.

Cohort Description

Participants in this study were recruited as part as the CARTaGENE (CaG) population-based health study in Quebec, Canada. CaG randomly targeted 20,000 participants of 40–69 years of age, the segment of the Quebec population that is most at risk of developing chronic disorders, from three areas in Quebec : the Montreal (MTL) Area, the Quebec city (QCC) Area and the Saguenay Lac-St-Jean (SAG) region. The participants from the CaG project were not recruited for a particular disease but represent a random selection among the population. A total of 6500 participants were earmarked for RNA work by sampling whole blood in Tempus tubes. Based on data from these participants, 521 individuals were chosen to represent different regional populations of Quebec and were selected based on cardio-vascular phenotypes. Our selection is enriched with individuals representing high and low risk of developing cardiovascular diseases, based on their Framingham Risk Score (FRS) (D'Agostino et al. 2008). Specifically, we stratified participants by region and gender and ranked them based on their FRS, while excluding individuals treated for hypertension. We then selected participants based on their ranking from high to low and from low to high scores ensuring similar samples sizes for men and women. Regional origin of participants was verified by principal component analysis (PCA) of genetic diversity using the EIGENSTRAT method (Price et al. 2006). Specifically, the PCA was performed using SNPs genotyped in the 521 CaG individuals and in 140 individuals from the Reference Population Panel of Quebec (RPQ) sampled in seven

sub-populations of Quebec characterized by different demographic histories (Roy-Gagnon et al. 2011). The final selection consist in 264 participants from MTL, 99 from QCC and 158 from SAG.

Genomic data: RNA-sequencing and Genotyping

Approximately 3 mL of blood was collected for RNA work in Tempus Blood RNA Tubes (Life Technologies). Total RNA was extracted by using a Tempus Spin RNA Isolation kit (Life Technologies) followed by globin mRNA depletion by using a GLOBINclear-Human kit (Life Technologies). RNAseq 100bp pair-ends indexed libraries were constructed using the TruSeq RNASeq library kit (Illumina). Sequencing was done on HiSeq machines (Illumina), multiplexing three samples per lane. After initial filtering based on sequencing read quality, paired-end reads were aligned using TopHat (V1.4.0) (Trapnell et al. 2009) to the hg19 European Major Allele Reference Genome (Dewey et al. 2011). PCR removal was performed using Picard (picard_tools/1.56, http://picard.sourceforge.net).

Raw gene-level count data was generated using htseq 0.5.3p3 (Anders and Huber 2010). These counts were then normalized using EDASeq v1.4.0 and a procedure that adjust for GC-content as well as for distributional differences between and within sequencing lanes (Bullard et al. 2010; Risso et al. 2011). Average normalized gene expression levels per gene were determined by averaging expression levels of each gene across all individuals (Idaghdour et al. 2013. *In preparation*). Every exon of a gene was attributed the gene-level value.

SNP were called from RNAseq data using a procedure similar to SNP calling in exon sequencing data. However, prior to SNP calling, bowtie2 (0.12.7)(Langmead et al. 2009) was used to removed abundant sequences (polyA, polyT, tRNA). Only reads that were properly paired and uniquely mapped were kept. Mapping quality score were recalibrated using GATK (McKenna et al. 2010) and SNP calling was performed with samtools (0.1.18) (Li et al. 2009). Filtering of SNPs was done using vcftools v0.1.7

(Danecek et al. 2011). We kept SNPs with variant quality of 30 and genotype quality of 20 (Phred scores). Minor allele frequencies (MAF), the proportion of individuals with non-missing genotypes and Hardy-Weinberg equilibrium (HWE) p-values were computed using plink v1.07. SNPs showing departures from HWE at $p > 0.001$ were excluded. We obtained a total of 178,486 polymorphic SNPs (MAF > 0) in the 521 French-Canadians individuals. All 521 individuals were also genotyped on the Illumina Omni2.5M array. A total of 1,554,440 autosomal SNPs were obtained after filtering (Quality control HWE p<0.001, Missingness < 0.05, MAF>0).

Estimation of Recombination Rates and Genetic Map Construction

We used the genotyping SNPs and the *interval* program from the LDhat package (McVean et al. 2002) to estimate population recombination rates ρ on the autosomes, in units of $4N_e r$ per Kb, where N_e is the effective population size and r is the recombination rate per meiosis in cM. The largest chromosomes (1 to 12) were broken into two segments (p and q arms) and all genomic segments were phased with Shape-IT (Delaneau et al. 2008). Because the pre-computed likelihood tables for the *interval* program accept a maximum number of 192 haplotypes, we selected 96 unrelated individuals from each CaG subpopulations. For each population, we ran the *interval* program on each genomic segment for 30,300,000 iterations with a burn-in of 300,000 iterations and the estimate of the recombination rate ρ between each pair of adjacent SNPs was computed by taking the average rate across iterations of the rjMCMC procedure implemented in *interval* (McVean et al. 2002). To convert the population recombination rate estimates in units of $4N_e r$ per Kb into units of centiMorgan per Megabase (cM/Mb), we inferred N_e for each population using the estimates of r computed for the 2010 deCODE map in cM units (Kong et al. 2010). Specifically, we identified chromosomal segments where both CaG data and deCODE SNP positions allowed estimates of rates and we summed rates across these genomic regions to obtain the total estimated distance ($4N_eR$) and the total genetic distance (R in cM units) from the deCODE map. To calculate 95% confidence intervals, we

sampled the rjMCMC every 15,000 iterations and computed N_e using these 2000 samples of ρ values. Using the same methods, we estimated population recombination rates and constructed genetic maps using 96 unrelated individuals from the CEU and YRI populations from HapMap3. To allow comparison between populations, we further computed recombination rates in the five populations using the subset of SNPs genotyped in the CaG and HapMap3 populations.

Coldspots of recombination are defined as regions of more than 50Kb with recombination rates between adjacent SNPs below 0.5 cM/Mb in CaG, CEU and YRI populations. We excluded centromeric regions and required that at least 5 SNPs support the coldspot, to avoid regions with dramatically reduced power to estimate recombination rates. A hotspot is defined as a short segment (<15Kb) with recombination rates falling in the 90th percentile (> 5 cM/Mb). We define high recombination regions (HRRs) as regions with a high density of hotspots of any length, such that the distance separating neighbouring hotspots (>5 cM/Mb) is smaller than 50 Kb. See Figure S2 for an illustration of these definitions. The recombination rate thresholds used to define coldspots and hotspot were chosen to maximize the overall number of SNPs included in the analyses while minimizing the difference between the number of SNPs in coldpots and in HRRs. Changing these threshold values do not change the conclusions of the study (Table S4).

RNA-seq SNPs annotation and exon inclusion

For each of the 178,486 RNAseq SNPs called, we retrieved functional annotations and conservation scores from publicly available databases and ressources. Functional annotations were obtained from four different sources (as of october 2012) : Seattleseq (http://snp.gs.washington.edu/SeattleSeqAnnotation137/), dbSNP (dbSNP135), SIFT (http://sift.jcvi.org/) and wAnnovar (http://wannovar.usc.edu/) (Chang and Wang 2012). A SNP is annotated as ‘nonsense’ if at least one annotation tool annotated it as ‘stop gain’ or ‘stop loss’. Similarly, a SNP is annotated as ‘missense’ if at least one annotation tool characterized it as ‘missense’. The other

coding SNPs were annotated as 'synonymous' mutations. The remaining SNPs reported as intronic, in untranslated regions (UTR) or in non-coding RNA were labeled as 'other'. The functional impact of missense mutations was obtained with widely used prediction tools : PolyPhen (Adzhubei et al. 2010) predicts functional effects using sequence conservation, physiochemical properties, proximity to functional domains and protein structure. SIFT (Kumar et al. 2009) relies on the alignment of highly similar orthologous and paralogous protein sequences and predicts functional effects based on conservation in protein families. These methods have high sensitivity but low specificity (Flanagan et al. 2010), we therefore used a combination of the two to annotate RNAseq missense mutations and reduced false positives. A SNP was annotated as 'benign' if it is predicted as "tolerated" by SIFT and "benign" by Polyphen. It is annotated as "damaging" if it is predicted as "damaging" (low and high confidence calls) by SIFT and "possibly damaging" or "probably damaging" by Polyphen. When SIFT and Polyphen prediction do not agree, the SNP is annotated as "potentially damaging".

To estimate the level of contraint at nucleotides in DNA sequences, we retrieved GERP and PhyloP conservation scores for all positions within the human exome. These score were used in order to compare conservation levels between coldspots and HRRs and to annotate RNAseq SNPs. GERP scores were obtained from the Sidow lab website (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html) and PhyloP scores were obtained from UCSC Genome Bioinformatics website (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP46way/). Scores from both methods are obtained based on an alignment and a model of neutral evolution. Sites predicted to be conserved are assigned positive scores, while sites predicted to be fast-evolving are assigned negative scores. GERP quantifies position-specific constraint in terms of "rejected substitutions", the difference between the neutral rate of substitution and the observed rate at individual alignment positions, estimated by maximum likelihood. The PhyloP method is based on a phylogenetic hidden Markov model and computes *p*-values based on prior and posterior distributions of

the number of substitutions that have occurred. The absolute value of PhyloP scores represent -log *p*-values under a null hypothesis of neutral evolution.

Frequency-based statistics were computed for all RNAseq SNPs called. We obtained minor allele frequencies overall and within each regional population using the --freq option in PLINK (Purcell et al. 2007). We obtained SNP frequencies in the 1000 genomes project (as of february 2012) from wAnnovar annotation and obtained the dbSNP id of each SNP present in dbSNP135 from Seattleseq annotation. This allowed us to identify novel SNPs, not reported in dbSNP or as part as the 1000 genomes project. Among novel SNPs, we further defined private SNPs, seen only in one regional population of CaG, and singletons, seen only once in the sample.

To insure that sequencing SNPs are called throughout the length of exons, and to reduce the possible biases due to read depth, we selected exons with all positions of their sequence covered at a minimum of 20× in more than 50% of the sequenced individuals (i.e. at least 261 individuals). We used BAMStats-1.25 to obtain the minimum coverage per exon per individual for 208,226 autosomal exons. A total of 89,390 exons (hereafter termed Min20x exons) passed this stringent filter containing a total of 73,630 SNPs. For each exon, we tabulated the number of SNPs called, average MAF, exon size, GC-content, average recombination rates in CaG, CEU and YRI populations, average gene expression, average GERP and PhyloP scores. We also computed, for each exon, the density (SNP/kb) of synonymous, missense, damaging and novel SNPs, singletons, and SNPs with high GERP (>3) and PhyloP scores (>1). To evaluate the effect of confounders, exons were ranked based on their proportion of GC or based on their expression level. In each case, they were stratified in four categories of equal sizes : low (25% lowest values), low median (ranked between 25 and 50%), high median (ranked between 50 and 75%) and high (25% highest values).

Mutational Load and Odds Ratios

We computed the differential mutational load between coldspots and HRRs using Odd Ratios (OR). The OR is computed as :

$$OR = \frac{n_{CS} \cdot x_{HRR}}{n_{HRR} \cdot x_{CS}}$$

with n_{CS} and n_{HRR} are the number of mutations of a given type in coldspots and HRRs respectively, and x_{CS} and x_{HRR} are the number of other mutations in coldspots and HRRs, respectively. Confidence intervals were calculated following the procedure described in (Morris and Gardner 1988). If the OR value is significantly larger than 1 for a given type of mutation, this type of mutation is enriched in coldspots. Conversly, if the OR value is significantly smaller than 1, the given type of mutation is enriched in HRRs. OR_{Div} was calculated for evaluating differential diversity between coldspots and HRRs. In that case n_{CS} and n_{HRR} correspond to the total number of SNPs, and x_{CS} and x_{HRR} correspond to the total number of non-mutated positions in coldspots and HRRs, respectively. Similarly, OR_{Cons} was computed for evaluating differential level of constraint between coldspots and HRRs, with n_{CS} and n_{HRR} corresponding to the total number of positions in a given conservation score category and x_{CS} and x_{HRR} corresponding to the total number of positions outside this category, in coldspots and HRRs, respectively.

Differential mutational load was evaluated considering all SNPs found in the CaG cohort, per regional population and per individual. OR were computed for all categories of functional annotations (synonymous, nonsynonymous, missense, nonsense, benign, potentially damaging, damaging) and frequency-based measures (low MAF, novel, private, singletons). For conservation scores, SNPs were stratified in bins of GERP and PhyloP scores of size 1, from -11 to 7 for GERP and from -9 to 7 for PhyloP. When less than 30 SNPs were found in a category, the OR was not computed for that category. Because the conservation score distribution at the sequence level

differ significantly between coldspots and HRRs, we computed OR_{Cons} for each category and used this value to correct counts and adjust for the baseline differential constraints observed. For results with uncorrected OR, see Figure S3. For most analyses, we used GERP > 3 and PhyloP between 1 and 5 to assess the differential mutational load based on conservation scores (Supplementary Results). Linear regression models were also built to evaluate the correlation between recombination rates and mutational load per exon, while adjusting for counfounders such as expression level, GC-content, exon size and SNP density.

The ORs were also computed for each individual (indOR) based on the variants he/she carries. We used a non-parametric Mann-Whitney U-test to test whether distributions of indOR between groups of individuals are significantly different. Individuals were grouped according to their regional populations or based on phenotypic information and disease status.

Extracting Mini-Haplotypes from Mapped Paired-end Reads

We took advantage of sequencing paired-end reads to create mini-haplotypes consisting of pairs of SNPs that are found on the same read of read pair. We first considered the 178,486 polymorphic SNPs detected in CaG individuals by our SNP calling procedure (minimum variant quality of 30, minimum genotype quality of 20). We used bedtools (Quinlan and Hall 2010) to extract all mapped read pairs that overlap with these positions, with the exception of reads for which the position of the SNP falls in the two first or two last nucleotides of the read, since these bases are generally of lower quality. We then kept only read pairs overlapping with at least two SNP positions and where the read base qualities at the SNP positions were above 13. We found that 33.63% of extracted reads covered at least one other SNP position.

For each individual separately, we called mini-haplotypes for all pairs of SNPs covered at least 20 times (Figure S4). In order to keep the data independent, we excluded mini-haplotypes for which the two SNPs in the pair were already part of another pair

considered in this individual. Each mini-haplotype had to be seen at least 5 times and in more than 10% of the total number of read pairs covering the pair of SNPs in the individual. For each pair of SNPs, a diploid individual is expected to carry at most two mini-haplotypes, however, we observed pairs of SNPs harboring more than two highly-covered haplotypes (Figure S4, Case 4). This is likely due to paralogy or mapping errors, we thus excluded these pairs of SNPs from further analyses. Based on the allele frequencies of SNPs in the sample, each mini-haplotype was classified as Min/Min, if two minor alleles are linked with each other, as Maj/Maj, if two major alleles are linked to each other, or as Min/Maj, if the minor allele at one position is linked to the major allele at the other position. Mini-haplotypes are classified either as coldspots mini-haplotypes, if both SNPs in the pair are found in the same coldspot region, or as HRR mini-haplotypes, if both SNPs are found in the same HRR region. Mini-haplotypes with SNPs found in different regions are rare, and were ignored in our analyses.

SNP data in Populations from The 1000 Genomes Project

Phase 1 SNP data from the 1000 Genomes Project were downloaded from the 1000 genomes ftp site and consists of 1092 individuals from 14 populations. For these analyses, we used SNPs called from the high coverage exon-targeted data, that is sequenced to an average coverage of 50-100x, and only SNPs falling within targetted exons were extracted from exome vcf files. The false discovery rate of exome SNPs is 1.6%. Details on 1000 Genomes populations, sequencing protocol, snp calling, and validation can be found in the 1000 Genomes phase 1 publication (Abecasis et al. 2012).

SUPPLEMENTARY RESULTS

1. Comparison of Hotpots Locations in CaG populations and CEU population

For these analyses, recombination hotspots were determined using LDhat rates (ρ) computed for the subset of SNPs shared between HapMap3 and CaG data (Material and Methods). A hotspot is inferred as a segment (<15Kb) with ρ values in the 95th percentile per chromosome per population. A hotspot is shared if at least one ρ value in the segment is in the 90th percentile in at least one other population. Conversly, a hotspot is population-specific if no other population has ρ values in the segment in the 90th percentile. The vast majority of hotposts are shared which is expected given the short time since the population diverged. However occasional differences exist between CEU and CaG populations and SAG population and all others. For example, a strong hotspot in the *ABCB4* gene, not seen in the CEU population, was inferred in all Quebec populations (Figure 2D). This result was validated by the LDHOT approach (Auton et al. 2012). Although the allele frequencies in this region are very similar between populations, the haplotype diversity is larger in CaG populations. Similar differences are found within the SAG population (Table S1 and Figure S2). For example, some hotspots consistently found in humans are not detected in genes such *SHANK2*, *COL9A1* and *FOXP1* in SAG population. These genes have important biological functions (Nikopoulos et al. 2011; Bacon and Rappold 2012; Tse 2012), and the unique patterns of linkage disequilibrium found in this population suggest that using population-specific linkage map may be important when performing disease mapping, at least in some genomic regions.

2. Comparison between functional annotations and conservation scores

Conservation scores, such as GERP and PhyloP, are widely used to identify mutations that impact protein function. In general, mutations at sites that are conserved across multiple species are highly deleterious. Other approaches have been developed to predict the functional impact of missense SNPs, such as Polyphen and SIFT. When

looking at annotations from theses two prediction tools separately, we found that Polyphen finds all categories of missense mutations equally enriched in coldspots, whereas differences between categories are seen for SIFT annotations (Figure 3B), with the ones predicted as being damaging not preferentially found in coldspots.

We compared the annotations obtained using GERP and PhyloP conservations scores with Polyphen and SIFT predictions (Table S3). We consider a mutation to be conserved if it has a GERP score higher than 3 and/or a PhyloP score higher than 1. Sites with PhyloP score higher than 1 are the top 10% of conserved sites in the human genome. The GERP threshold was chosen to be comparable to PhyloP, so that the overall proportion of RNAseq SNPs and missense SNPs found to be conserved is approximately the same for both methods (Table S3). Overall, more than half missense mutations and a large proportion of damaging mutations (78.2%) are conserved according to both conservation scores. Polyphen predictions for damaging mutations concord slightly better with conservation annotations than SIFT predictions, as 73.5% of SIFT damaging mutations vs. 78.2% of Polyphen damaging mutations are predicted to be conserved. In our analyses, sites with extreme PhyloP conservation scores (>5) behave differently than sites having PhyloP score between 1 and 5 (Figure 4). This is not seen for GERP annotation. Although only 20% of damaging mutations are seen in this category, 38.7% of sites with extreme PhyloP scores are damaging whereas only 23.9% of sites with extreme GERP scores are damaging. This difference is highly significant (Chi-square test, $p<0.0001$), indicating that there is an enrichment of damaging mutations in the class of mutations having extreme PhyloP values.

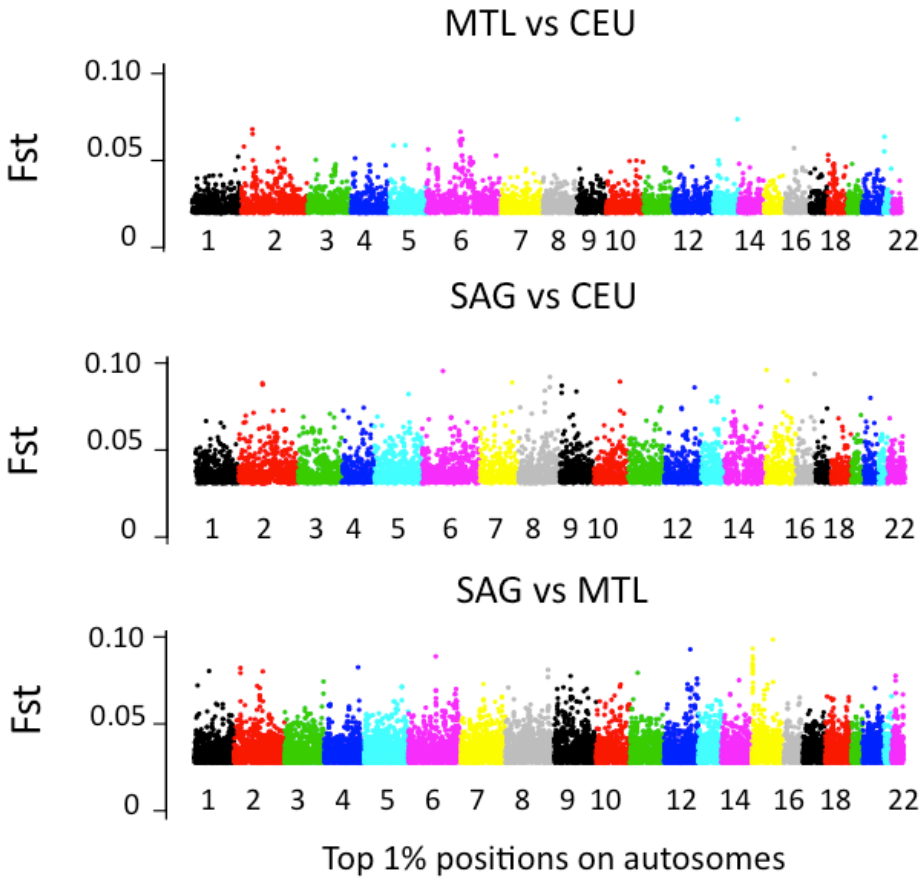
3. Power and confounding for phenotypic analysis

Different results were obtained across regions and they may be due to true signal, or to lack of power to capture the effect. We investigated three factors that can be affecting power : the distribution of FRS across populations, the sample size and the proportion of individuals with high enough coverage per SNP.

The distribution of FRS is different between CAG populations (Figure S5). SAG has the most individuals in average FRS values (between 6 and 12) compared to MTL and QCC. We expect the power to detect an effect to be increased when the majority of individuals have extreme values, such as in MTL population. The distribution of FRS for QCC and MTL are similar, however the size of the QCC sample is smaller and may be too low to allow the detection of an effect. To test this, we resampled 100 times MTL individuals to create datasets matching QCC sample size and number of individuals with cardiometabolic disorders. In the majority of samples (64/100), the effect seen in MTL with all individuals was not detected for ORs computed for rare variants (MAF<0.01). Furthermore, only 11 datasets showed significant results for all five measures of mutational load where the effect was initially detected (rare, non-synonymous, damaging, conserved, novel).

The observation of a significant correlation between FRS and indOR across mutational load measures led us to test several linear models to evaluate the correlation of FRS with other factors that could explain the observed association. We found that the total number of SNP is also correlated with FRS. We therefore corrected for the effect of the total number of SNPs when evaluating the correlation between indOR and FRS. In our analyses, we considered all exons that are entirely covered at 20x in more than 50% of individuals. When we consider all SNPs covered in more than 80% of individuals, we detected the same effect between FRS and indOR and between cardio-vascular disease status and indOR (Table S7) even though the power decreases as a result of the smaller number of SNPs considered to compute indOR. However, the linear correlation between FRS and the total number of SNPs remained significant, which suggest that this effect is not only due to differences in expression patterns between individuals.

SUPPLEMENTARY FIGURES AND TABLES



Comparison	1st Qu.	Median	Mean	3rd Qu.
MTL vs CEU	2.91×10^{-4}	1.35×10^{-3}	2.98×10^{-3}	4×10^{-3}
SAG vs CEU	4.82×10^{-4}	2.21×10^{-3}	4.76×10^{-3}	6.34×10^{-3}
SAG vs MTL	4.54×10^{-4}	2.05×10^{-3}	4.31×10^{-3}	5.71×10^{-3}

Figure S1. Genetic differentiation between MTL, SAG and CEU

F_{ST} computed to compare the genetic differentiation between MTL, SAG and CEU populations using SNPs in common in CEU, MTL and SAG. The top 1% values are plotted for each chromosome and mean and quartile values for each comparison are presented.

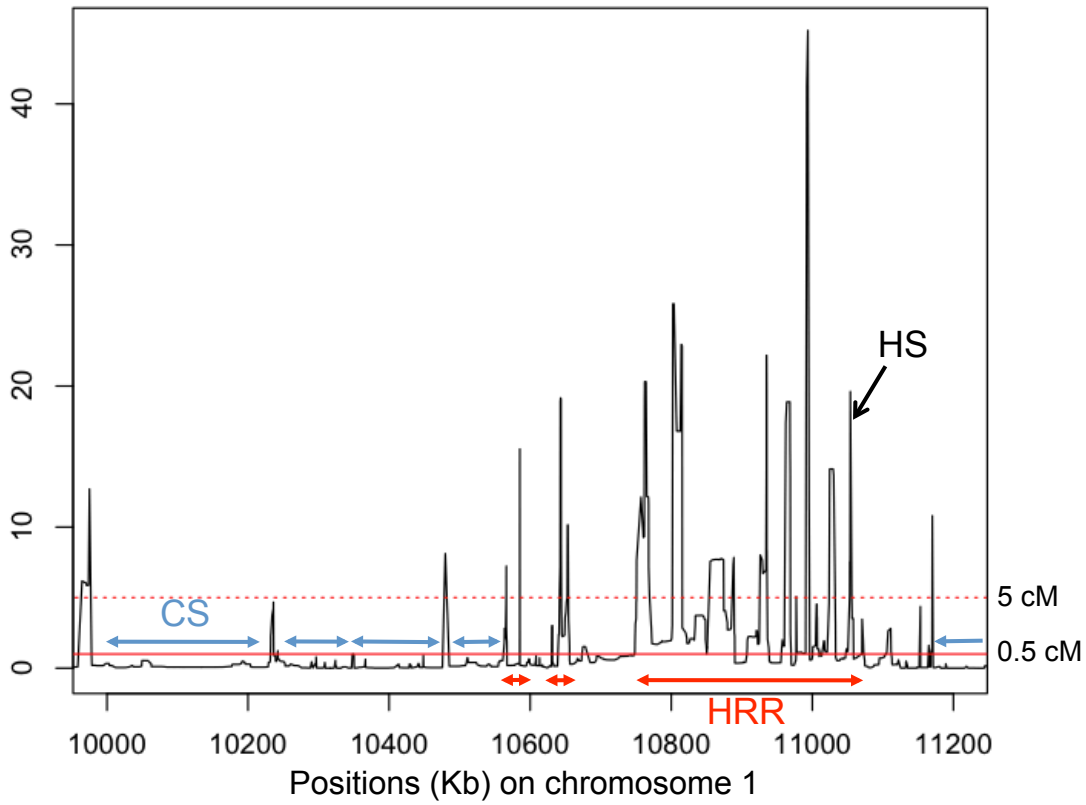


Figure S2. Illustration of the definition of coldspot (CS), hotspot (HS) and high recombination regions (HRRs).

CS are defined as regions of more than 50Kb with recombination rates between adjacent SNPs below 0.5 cM/Mb. HS are defined as a short segment (<15Kb) with recombination rates above 5 cM/Mb. HRRs are regions with a high density of hotspots of any length, such that the distance separating neighbouring hotspots (>5 cM/Mb) is smaller than 50Kb. CS and HRRs have to be conserved across CaG, CEU and YRI populations to be kept in our study.

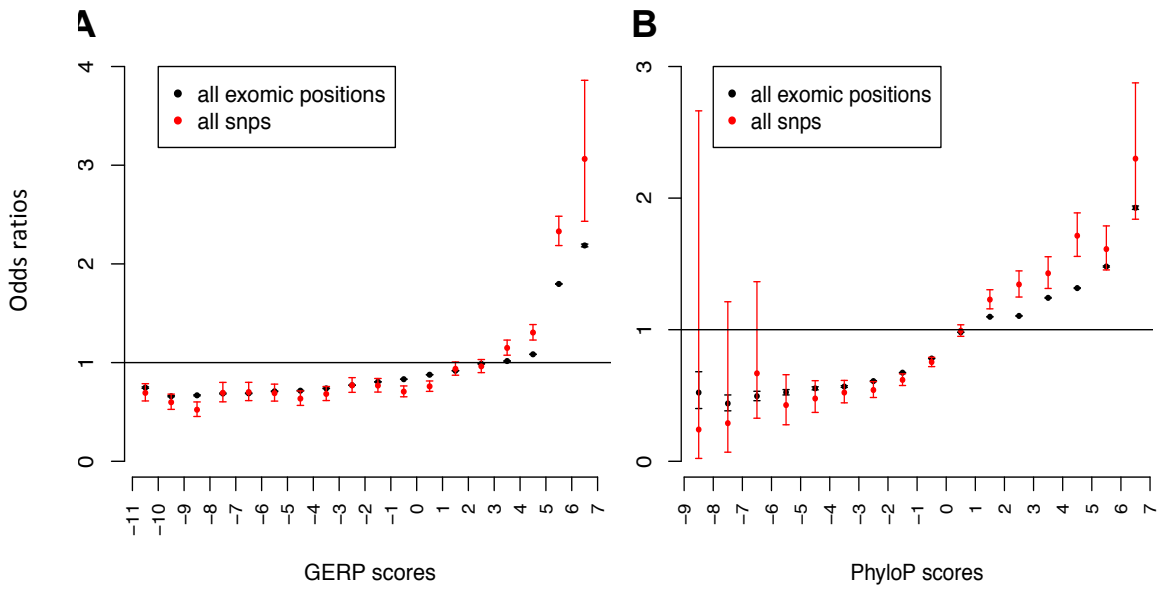
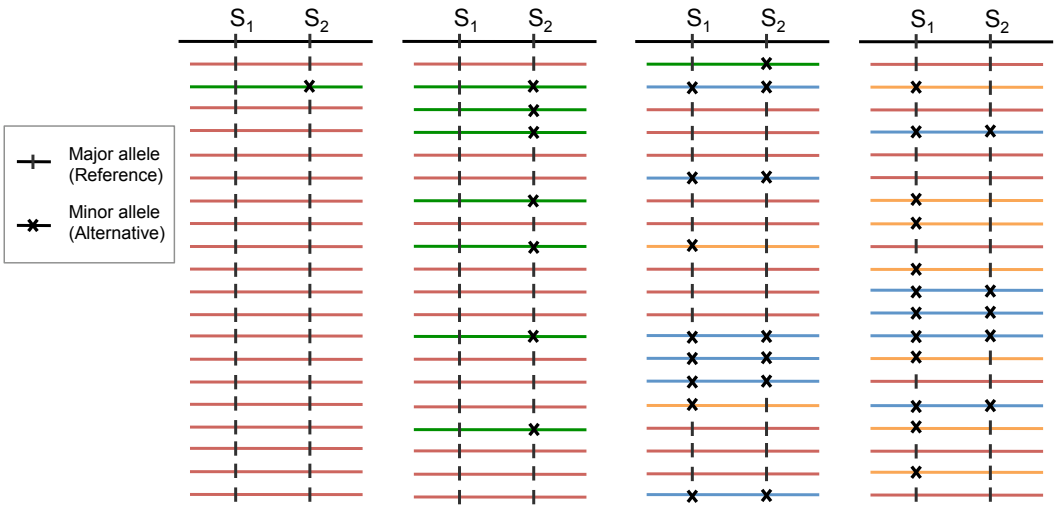


Figure S3. Uncorrected odds ratios (ORs) for SNPs compared to ORs for exomic positions for conservation categories.

As described with corrected ORs (Figure 4B-C), SNPs with GERP score >3 and for PhyloP score between 1 and 5 are significantly enriched in coldspots and this enrichement is significantly larger than the enrichement observed at the position level.



Haplotypes	Case 1	Case 2	Case 3	Case 4
S_1maj / S_2maj	19/20	13/20	11/20	7/20
S_1min / S_2maj	0/20	0/20	2/20	7/20
S_1maj / S_2min	1/20	7/20	1/20	0/20
S_1min / S_2min	0/20	0/20	6/20	6/20
Call	S_1maj / S_2maj	S_1maj / S_2maj S_1maj / S_2min	S_1maj / S_2maj S_1min / S_2min	No call

Figure S4. Calling "mini-haplotypes" from sequencing data.

Sequencing reads are aligned to the hg19 European Major Allele Reference Genome (Dewey et al. 2011). SNPs are called from sequencing read data and SNPs that pass quality control filters in at least one individual are kept. Each read (or read pair) covering two SNP positions are retreived. For each individual, the mini-haplotypes are called for each pair of SNPs covered by at least 20 reads (e.g. S_1 and S_2). In case 1, all reads except for one show the reference (major) alleles at both SNPs. The S_1maj/S_2min haplotypes is likely cause by a sequencing error at S_2 and is ignored. In case 2, the same haplotype is covered by more than 5 reads and is called. In case 3, the four haplotypes are seen, but only two are called because they are covered by more than 5 reads. In case 4, three haplotypes are detected and are covered with more than 5 reads. No call is made because no more than 2 haplotypes are possible in diploid individuals.

Table S1. Distribution of Coldspots (CS) and High Recombination Regions (HRR) genome-wide and in analysed exons.

Regions	CS	HRR
Whole genome	1,048,937,114	634,243,758
Whole exome	25,302,008	17,768,017
Min20x exons	6,906,137	2,036,963

The whole-exome includes 66,617,267 bp of sequences and the Min20x exons include 15,333,786 bp of sequences.

Table S2. Summary of linear models evaluating the correlation between recombination rates per exon and various variables.

Correlations are evaluated using per-exon values, considering only exons covered with minimum coverage of 20x at all positions in more than 50% of individuals.

Variable	Rec rates[a]	GC content	Expression	Exon Size	SNPs/Kb	R^2
SNP density	+ ***	+ ***	+ ***	ns	NA	0.038
MAF	+ ***	+ ***	+ ***	+ ***	NA	0.046
Expression[b]	- ***	+ ***	NA	ns	NA	0.016
Density of :						
GERP > 3	- ***	ns	ns	- ***	+ ***	0.267
PhyloP [1;5]	- ***	- ***	- **	- ***	+ ***	0.323
Nonsynonymous	-*	- ***	- ***	- ***	+ ***	0.538
Damaging	ns	ns	- ***	- ***	+ ***	0.151
Novel	- ***	+ ***	ns	ns	+ ***	0.387
Singletons	ns	+ ***	+ ***	+ *	+ ***	0.292

*** p<0.001; ** p<0.01; * p<0.05; ns=non-significant; +/- = positive/negative correlation; NA : non-applicable. P-values for the linear correlation are based on a Student's t-test that tests whether the coefficient for the term differs significantly from zero.

[a] Average recombination rates per exon in cM/Mb are computed based on the genetic map calculated based on 96 individuals from MTL and scaled using the 2010 deCODE map.

[b] This correlation is evaluated considering all exons covered with minimum coverage of 20x in at least ten individuals.

Table S3. Comparison between conservation scores and functional annotations

for 178,486 polymorphic SNPs called from RNAseq data.

Types of SNPs		GERP			PhyloP		
		> 3	**[3;5]**	**[5;7]**	**> 1**	**[1;5]**	**[5;7]**
	All SNPs	0.269	0.179	0.09	0.272	0.245	0.027
	Missense	0.584	0.33	0.254	0.593	0.497	0.096
	Consensus Damaging	0.813	0.401	0.413	0.834	0.631	0.203
Polyphen	Benign	0.455	0.292	0.163	0.455	0.413	0.042
Polyphen	Possibly Damag.	0.7	0.383	0.317	0.708	0.593	0.115
Polyphen	Probably Damag.	0.812	0.393	0.419	0.83	0.631	0.199
SIFT	Tolerated	0.343	0.226	0.117	0.341	0.307	0.034
SIFT	Damaging (LC)	0.585	0.347	0.238	0.61	0.53	0.08
SIFT	Damaging	0.77	0.392	0.378	0.797	0.614	0.183

LC : low confidence

The proportions of conserved SNPs from different functional categories are reported. Funtional impact was predicted by Polyphen and SIFT. SNPs classified in the Consensus Damaging class are mutations predicted as damaging by both methods.

Table S4. Robustness of the effect to recombination parameters.

Parameters used in this study are presented in red and were chosen to maximize the overall number of SNPs included in the analyses while minimizing the difference between the number of SNPs in coldpots and in HRRs.

Rec. rates (cM/Mb)		Odds ratios Coldspots vs HRRs				Number of SNPs		
L	**H**	**PhyloP [1;5]**	**Missense**	**MAF<0.01**	**Novel**	**in CS**	**in HRR**	**in between**
	5	1.55	1.29	1.34	1.2	9263	15208	49159
0.1	10	1.57	1.25	1.33	1.19	9263	10406	53961
	20	1.49	1.2	1.26	1.11	9263	7517	56850
	5	1.38	1.24	1.32	1.18	20482	15054	38094
0.25	10	1.39	1.21	1.31	1.17	20482	10210	42938
	20	1.32	1.16	1.24	1.08	20482	7272	45876
	5	**1.31**	**1.23**	**1.31**	**1.18**	**30789**	**14902**	**27939**
0.5	10	1.33	1.21	1.31	1.18	30789	9903	32938
	20	1.27	1.17	1.25	1.1	30789	6857	35984
	5	1.23	1.19	1.27	1.15	41034	14770	17826
1	10	1.24	1.17	1.27	1.16	41034	9620	22976
	20	1.17	1.13	1.2	1.07	41034	6352	26244
	5	1.2	1.15	1.24	1.13	49583	14915	9132
2	10	1.22	1.15	1.25	1.14	49583	9389	14658
	20	1.17	1.12	1.19	1.06	49583	5892	18155

Coldspots Regions : SNPs within a 50Kb region with no recombination rate higher than L

High Recombination Regions (HRRs) : SNPs within 50Kb of at least two hotspots with rate higher than H

Table S5. Robustness of the effect to GC-content and gene expression levels.

Feature	Category	PhyloP [1;5]	Missense	Novel	MAF< 0.01	Sing.
GC-content	lower	1,467	1,543	1,185	1,407	ns
	low median	1,503	1,510	1,268	1,584	ns
	high median	1,302	1,273	1,269	1,412	ns
	upper	ns	ns	1,192	1,230	1,177
Average Expression	lower	1,153	1,226	1,338	1,317	1,587
	low median	1,230	1,182	1,184	1,257	ns
	high median	1,280	1,085	1,198	1,272	ns
	upper	1,424	1,328	1,174	1,435	ns
All categories		1,305	1,232	1,179	1,305	ns

ns: non-significant

Odds Ratios values comparing numbers of variants in a SNP category between coldspots and HRR are reported. Gene expression levels and GC-content were computed per exon (Material and Methods). Exons were ranked according to their values, and categorised as lower, low median, high median and upper depending on their ranking (first 25%, 25-50%,50-75% and last 25%, respectively).

Table S6. Effective population size estimation based on inferred recombination rates.

Population	Estimate of N_e	CI 95%
MTL	11401	[10718 ; 11724]
QCC	11202	[10585 ; 11512]
SAG	9123	[8598 ; 9439]
CEU	10427	[9821 ; 10792]

Ancestral N_e is computed based on population recombination rates estimates calculated with LDhat (McVean et al. 2002) and the deCODE 2010 pedigree map (see Material and Methods).

Table S7. Mini-haplotype analysis in the 1000 Genomes Project Populations.

Rare Variant Type	OR AFR	95%CI	Within CS		Within HRR	
			Min/Min	Min/Maj	Min/Min	Min/Maj
all rare	**3.13**	[2.22;3.14]	409	25134	113	21785
conserved	**6.62**	[4.34;10.08]	251	6168	24	3904
non-synonymous	**9.66**	[6.72;13.87]	371	11809	32	9840
damaging	0.75	[0.20;7.33]	2	964	2	724

Rare Variant Type	OR CEU	95%CI	Within CS		Within HRR	
			Min/Min	Min/Maj	Min/Min	Min/Maj
all rare	**2.65**	[2.22;3.14]	535	30070	171	25439
conserved	**7.29**	[3.93;13.50]	136	11	5624	3316
non-synonymous	**11.22**	[7.53;16.71]	363	12570	26	10101
damaging	0.71	[0.10;5.05]	2	683	2	485

CHAPTER V: Discussion

THE COSTS OF RECOMBINATION

IN HUMANS

DISCUSSION AND PERSPECTIVES

Recombination is a decisive phenomenon driving the individual's genetics and phenotype, this is what makes us so different from each other. In this thesis, I highlighted some of the main costs of recombination for human populations and individuals. First, the cost of requiring recombination for proper chromosome segregation, in meiosis, predisposes humans to aneuploidies. Second, mutations arising from recombination processes are frequently associated with pediatric cancers and may not be independent of past meiotic events and mechanisms. Third, the non-random distribution of recombination rates in the human genome leads to an accumulation of deleterious mutations in essential cellular genes, found in low recombination regions. These costs, associated with patterns of variation in recombination rates, impact fitness at the population and individual levels.

Disease, Maternal Effects and Parental Genetics

In our first study, we looked at variation in recombination patterns in a cohort of French-Canadians families, and replicated findings from other studies, except for one result: the maternal age effect on recombination. Two studies had reported a positive maternal age effect (Kong et al. 2004; Coop et al. 2008) whereas we observed a negative correlation between maternal age and recombination. After our study was published, another group found a negative maternal age effect on recombination in an asian pedigree cohort (Bleazard et al. 2013). The fact that the direction of the correlation varies between studies either reflects methodological differences or real populational differences. Nevertheless, all studies have found a significant correlation between maternal age and recombination suggesting that, contrary to males, one or more factors are influencing the number of recombination within eggs along the life of human females. The same factors may be associated with age-dependent aneuploidies.

The source of the maternal age effect on aneuploidy remains unknown. To explain this effect, it has been proposed that insults to the meiotic process as women age become more frequent, or that meiotic checkpoints become less stringent (Hassold and Hunt 2009). Under such circumstances, the number of crossovers that generally protects against aneuploidies when no insult has been made may not be protective anymore. Under this "protection" model, aneuploid conceptions will tend to have more crossovers at older age than at a younger age of the mother, and as a result, a decrease of recombination with maternal age could be seen in the normal population (Chapter II, Figure 5). Presumably, the number of crossovers in properly disjoint oocytes remains higher than in aneuploid conceptions, regardless of maternal age.

If there is a negative correlation between maternal age and recombination in normal conceptions, the maternal age effect on aneuploidy should be less pronounced in woman with low recombination rates than in woman with high recombination rates. In this situation, the number of low recombining viable oocytes will be decreased less so than the number of high recombining viable oocytes. Alternatively, if women with higher rates of recombination are less prone to age-related aneuploidies than women with low recombination rates, then the correlation between recombination and maternal age would be positive. These considerations imply that age-related aneuploidies should be studied in the context of factors associated with maternal recombination rates. For instance, it will be important to establish whether women experiencing age-related aneuploid pregnancies are likely to carry particular alleles at loci associated with variation in mean recombination rates, such as the RNF212 gene and the inversion at 17q21.31. Because not only the number, but also the placement of recombination events on some chromosomes is linked with aneuploidies, typing PRDM9 alleles in women of all ages experiencing aneuploid pregnancies may also be informative.

Although aneuploidies mainly occur sporadically, recent studies have suggested genetic predispositions to non-disjunctions in humans (Gair et al. 2005; Kovaleva

2010; Silva-Grecco et al. 2012). So far, most studies of trisomy 21 have only used parental genetic information to identify the parental origin of the meiotic errors and study recombination patterns. However, because it is clear that the causal event for most human trisomies occur during meiosis, genetic variation in the parental meiotic genes should be further investigated in studies that aim to find genetic variants associated with aneuploidies in humans.

In our second study, we once again observed maternal effects. We discovered a large number of rare allelic forms of *PRDM9* in families of children affected with childhood acute lymphoblastic leukemia. In almost 77% of the families where PRDM9 was fully resequenced (Chapter III, Figure 3), we observed at least one parent with a rare allele of PRDM9, preferentially carried by the mother. However, other tests for subsets of alleles did not yield a significant maternal enrichement. Since the publication of our study, we analysed more families of affected children and further confirmed this maternal bias (p=0.0077, Table 1).

Table 1. Supplementary data on PRDM9 association with pre-B and pre-T ALL.

Study	Number of Families	With PRDM9 *k*-finger alleles		
		Total (transmitted)	Maternal	Paternal
B-ALL[a]				
Hussin et al. 2012	23	11 (6)	8	5
Unpublished data	30	12 (6)	9	3
Total	53	23 (12)	17[b]	8
T-ALL				
St. Jude cohort[c]	12	NA (4)	NA	NA
Ste-Justine cohort	12	6 (2)	4	2

[a] Families from Ste-Justine cohort

[b] The effect is preferentially maternal (Chi-square test, $p = 7.708 \times 10^{-3}$)

[c] Patient samples only. Total number of families with *k*-finger alleles and parental origin not available.

The higher frequency of maternal rather than paternal carriers of rare PRDM9 alleles suggests a sex-specific effect, possibly resulting from differences in hotspot usage in the two sexes. The mechanisms of sex-specificity in hotspot formation are currently unknown, but one possibility is that differential DNA accessibility during male and female meiosis leads to differential epigenetic marking. Human chromatids are less compacted in females than in males (Tease and Hulten 2004), different PRDM9 alleles may thus show differences in sex-specificity, with some alleles showing a very strong gender bias and others less so. The maternal bias we observe may suggest that the capacity of different unusual alleles to induce chromosomal instabilities depends on whether these alleles act in male or in female meiosis.

PRDM9 Function and Beyond

The association we found between PRDM9 and childhood leukemia is puzzling in one major aspect: if PRDM9 acts in meiosis and is not preferentially transmitted to the patient, how can it be associated with a carcinogenic process appearing during hematopoetic differentiation? My hypothesis is that abnormal PRDM9 trimethylation activity in germ-line predisposes the resulting gametes to genomic instabilities, that may be triggered during later stages of development and differentiation, and particularly during hematopoetic differentiation. This hypothesis is based on three major observations. First, the PRDM protein family is formed by 17 methyltransferases, with at least 11 involved in cancer. Second, V(D)J recombination centres resemble hotspot environments, as they both display enrichment of H3K4me3 and are regulated by specific DNA motifs. Third, incomplete erasure of histone marks can lead to the propagation of meiotic events to the new generation. Transgenerational inheritance of H3K4me3 has been observed in *C. elegans* and depends on a demethylase activity.

Many other PRDM proteins exist and the majority are involved in human diseases and cancer. They all contain a SET domain and Zinc Finger (ZnF) domains (Fog et al. 2011). Except for PRDM9, human genetic diversity of these ZnF domains has not

been characterized yet. Next-generation sequencing data, *de novo* assembly computational tools and sperm-typing techniques will be useful to characterize the variation within these ZnF arrays in population and disease cohorts. The main mechanism by which they trigger carcinogenesis is likely by promoting aberrant histone methylation in a tissue-specific manner (Fog et al. 2011). Therefore, the critical function of PRDM9 likely associated with leukemogenesis is its methyltransferase activity, which subsequently triggers DSBs. Whether PRDM9 plays a direct or indirect role in recruiting SPO11 to these sites also remains unknown. Recently, Brick and colleagues observed that PRDM9 is not essential for DSB formation, but rather drives DSB activity away from transcription starting sites (Brick et al. 2012). Why are PRDM9 marks preferentially recognized by the recombination machinery? The molecular role of H3K4me3 in DSB formation is not fully understood.

Although meiotic and V(D)J recombination result from different molecular pathways, the DSB initiation process has three important similarities. First, DSBs cluster in narrow regions, either called meiotic recombination hotspots or V(D)J recombination centres. Second, in both cases, the initiation involves in both cases the recognition of specific DNA motifs. Such motifs are found in thousands of locations in the human genome and only a subset of these are functional targets. Third, these highly active regions display enrichement in H3K4me3 marks. To initiate V(D)J recombination, RAG2 recognizes H3K4me3 and acts as an anchor for RAG1, that will catalyze DSB formation. The recognition of a pseudo recombination signal sequence (RSS) in environments that behave like recombination centres can result in genomic instabilities leading to leukemia initiation. Some alleles of PRDM9 may by chance put H3K4me3 marks in close vicinity of pseudo RSS, which is more likely in heterozygotes as their number of potential targets is expected to be largely increased compared to homozygotes.

The factors that regulate H3K4me3 demethylation at DSB sites remain unidentified in mammals. H3K4me3 at TSS is abundant in leptonema and zygonema and is

decreased in pachynema (Godmann et al. 2007). Are the H3K4me3 marks near recombination hotspots removed by the same mechanisms as TSS marks? H3K4me3 is a mark of active chromatin viewed as "long-lived", as it is involved in memory during cell state inheritance in metazoans and has been suggested to play a role in epigenetic inheritance in *Drosophila* and Budding Yeast (Ringrose and Paro 2004; Radman-Livaja et al. 2010). In *C. elegans*, perturbation of H3K4me3 during meiosis leads to transgenerational inheritance of longevity (Greer et al. 2011). RBR-2 is a H3K4me3 demethylase important to achieve incomplete erasure of histone marks and transgenerational inheritance. In human, little is known about the demethylation processes occuring in meiosis and early developement, however the potential involvement of histone demethylases in human diseases, including cancer, has been highlighted (Cloos et al. 2008). Our work support a new hypothesis to explain how congenital abnormalities predispose children to cancer, they possibly both result from the same unstable epigenomic patterns.

There are other hypotheses that might explain the association between PRDM9 unusual alleles and childhood leukemia. When transmitted to the affected child, the rare PRDM9 alleles may be aberrantly expressed in hematopoetic differentiating cells and directly creates defects. However, under such a model, transmission of the rare allele to the affected child would have been expected. This model therefore fails to explain why the excess of rare alleles is seen in parents more so than in the affected children. Another hypothesis is that unusual PRDM9 alleles activate retrotransposon elements during meiosis, that can be integrated in the genome later during developement, causing somatic mosaicism that may have a role in initiation of the leukemogenesis process. For instance, it has been described that activation of the Long Interspersed Element 1 (L1) in germ cells can be carried over through fertilization and can integrate during embryogenesis, thus creating somatic mosaicism during mammalian development (Kano et al. 2009). Furthermore, L1 activity has been proposed as one cause for genomic instability observed during the progression of leukemia (Kirilyuk et al. 2008). PRDM9 binding sites have been found

within specific retrotransposons, however the link between PRDM9 and activation of retrotransposons remains to be established.

We are now following up on these results in several ways. First, we are exploring the possible association between PRDM9 C-like alleles and chromosomal aberrations, through a collaboration with the Camerini-Otero group at the NIH, who developed DSB hotspot maps for human PRDM9 A homozygotes and AC heterozygotes (Kevin Brick, personal communication). Second, preliminary data suggest that the association is also present in T-ALL patients (Table 1). We are now obtaining more data for different types of leukemias and for other types of pediatric cancers, such as pediatric brain tumors (Schwartzentruber et al. 2012). Finally, genome-wide H3K4me3 ChIP-Seq profiling of samples of the ALL cohort will be performed, including the quartet family analysed in our study.

PRDM9 and Allelic Incompatibilities

An additional question regarding PRDM9 concerns the impact of heterozygosity at this locus in human populations. It might not be the effect of a particular PRDM9 allele that is important to understand it's link to leukemia, but rather the effect of two PRDM9 alleles interacting in the same cell during meiosis.

Prior to finding its role in meiotic recombination, PRDM9 had been identified as the first gene involved in speciation in mammals (Mihola et al. 2009). Most of the speciation genes were identified in *Drosophila* and many of them appear to be DNA binding proteins functioning in epigenetic modifications pathways (Bayes and Malik 2009; Ferree and Barbash 2009; Phadnis and Orr 2009). Genetic conflict between a DNA binding protein and its target sequence can contribute to genome instabilities through the Dobzhansky-Muller model (Brown and O'Neill 2010), which proposes that epistatic interactions between loci are incompatible in heterozygous hybrids (Dobzhansky 1937; Muller 1942). In the case of PRDM9, hybrid infertility in mouse is thought to result from aberrant H3K4me3 distribution in germ cells and altered gene

expression (Hayashi et al. 2005; Mihola et al. 2009). This means that conflict between PRDM9 alleles exist in nature and may have very serious consequences.

Within humans, competition between PRDM9 alleles has been reported. In the Hutterite population, the allele I, specific to that population, out-competes the allele A in determining hotspot usage within heterozygotes (Baudat et al. 2010). Some alleles have recently been qualified as “destabilizers”, which are low frequency PRDM9 variants potentially capable of destabilizing their own coding sequence (Jeffreys et al. 2013). Interestingly, complex defects are seen with the C and D alleles, both carrying a *k*-finger repeat and found associated with childhood leukemia in our study.

Furthermore, heterozygotes may be differentially tolerated on different genetic backgrounds, which could explain the differences in allele frequencies between European and African populations. Variation in Europeans is considerably low, which could be attributed to drift alone, or may result from selection against some incompatible alleles. As our results link allelic variation at *PRDM9* with the risk of childhood ALL, one might expect that the increased diversity in individuals of African descent, and in particular the higher frequencies of C-type alleles (Chapter I, Figure 4), would indicate a higher susceptibility to childhood leukemia in African populations. However, incidence of childhood leukemia in sub-saharian Africa is reported to be lower (Stiller and Parkin 1996), although it is hard to tell whether this is due to a genuinely lower risk or to under-ascertainment or under-diagnosis. Furthermore, information on the distribution of ALL subtypes among African affected children is lacking. Interestingly, these populations do not show evidence for the early childhood peak of incidence between 2 and 5 years old, seen in populations of European descent. In admixed African-American children, the incidence rate is approximately half that among children of European descent (Gurney et al. 1995), largely because of a much reduced early childhood peak (Stiller and Parkin 1996). This suggests that children from African descent may have a reduced chance of acquiring a second hit, critical for the development of leukemia.

Alternatively, the role of allelic forms of *PRDM9* in the development of ALL may involve multi-locus interactions arising only in specific genomic backgrounds. Since *PRDM9* interacts with specific binding motifs to regulate histone methylation and recombination, these loci, if mutated, could modify downstream *PRDM9* deleterious functions. To test this, we performed supplementary analyses to compare the rate at which mutation accumulate within *PRDM9* binding targets (Table 2). Genome-wide, we observed a higher proportion of SNPs within sequences matching the common allele binding motif than within those matching the C-type allele binding motif ($p < 0.0001$, χ^2 test). Conversely, for motifs found within known ALL genes (Chapter III, Table S8B), the sequences matching the C-type allele motif are significantly more mutated than the common allele motifs ($p = 0.0337$, χ^2 test). Additionally, two studies (Jeffreys and Neumann 2002; Myers et al. 2010) have shown that self-destructive drive due to biased gene conversion disrupts the common-allele binding motifs. In the African population, the same process appears to be eliminating binding motifs for alleles at higher frequencies (Berg et al. 2011). The finding that C-type binding targets in genes implicated in ALL are more frequently mutated in humans suggests that drive against these motifs may act in African genomes, which may implicate genomic background as a critical factor for the role of PRDM9 in leukemogenesis.

Table 2. SNPs within sequences matching the common allele binding motif (*A*) and the C-type allele binding motif (*C*).

Genomic Region	Motif	Motif Counts	Positions with SNP[a]	Positions without SNP[a]	SNP proportion P_{SNP}	*A* vs *C*
Genome-wide	*A*	13293	9015	457710	0.0193	$P_{SNP}(Core) > P_{SNP}(C)$ OR=1.23 [1.20; 1.25]
	C	185585	111896	6970316	0.0158	
Within ALL gene list	*A*	42	18	1452	0.0122	$P_{SNP}(Core) < P_{SNP}(C)$ OR=0.60 [0.38; 0.97]
	C	789	604	29378	0.0201	

[a] SNP positions are from dbSNP v134 with validated (VLD) status.

Taken together, these new observations raise the intriguing possibility that PRDM9 ZnF incompatibilities contribute to disease in humans in some populations but not necessarily in others, which may translate into reproductive incompatibilities in the population.

A Model of Heredity for Childhood Leukemia

This work raises the possibility that the presence of abnormal meiotic events, happening during gamete formation and triggered by key meiotic genetic players such as PRDM9, can predispose the resulting children to leukemia. The susceptibility to develop childhood leukemia may therefore be partially heritable, because a significant number of children with ALL inherit the rare PRDM9 gene variant. I thus propose a possible heredity mechanism that predisposes children to ALL and puts their own children at risk of having ALL-predisposed offspring. In this model, the abnormal genetic variant associated with predisposition to leukemia does not need to be passed from parent to child for offspring to develop the disease. Instead, it is the result of the genetic abnormality within the gamete that gives rise to the children that would predispose them to leukemia. However, children who inherit the genetic variant run the risk of transmitting ALL predisposition to their offspring. As all gametes of a parent may not be affected by the genetic abnormality, the predisposing factor is not necessarily segregating along with the predisposition itself. Furthermore, while an abnormal gamete may lead to ALL development, this condition alone is not enough. Triggering the process of cancer cell proliferation inevitably requires a second hit, such as other mutations and/or environmental factors. A number of different outcomes are thus possible, as described in Figure 1. In line with this model, pre-leukemic clones are present in the blood of children at birth in the normal population, and only 1 on 100 children will subsequently develop ALL (Mori et al. 2002).

The early onset of ALL, with a peak incidence between 2 and 5 years old, is thought to result from inherited genetic predisposition interacting with environmental

factors during early development, in fetal life and infancy. While several somatic genetic defects have been identify to drive pediatric ALL (Zhang et al. 2012; Holmfeldt et al. 2013), few established genetic risk factors for ALL have been identified through genome-wide association studies (GWAS). The model proposed here may explain why GWAS fail to find convincing hits explaining suceptibility to childhood leukemia. Researchers generally focus on studying the children, their tumors and their environment, with genetic data from parents rarely being taken into account. Our findings demonstrate the importance of including parents' genetic information for the understanding of childhood leukemia, as well as other early childhood diseases.

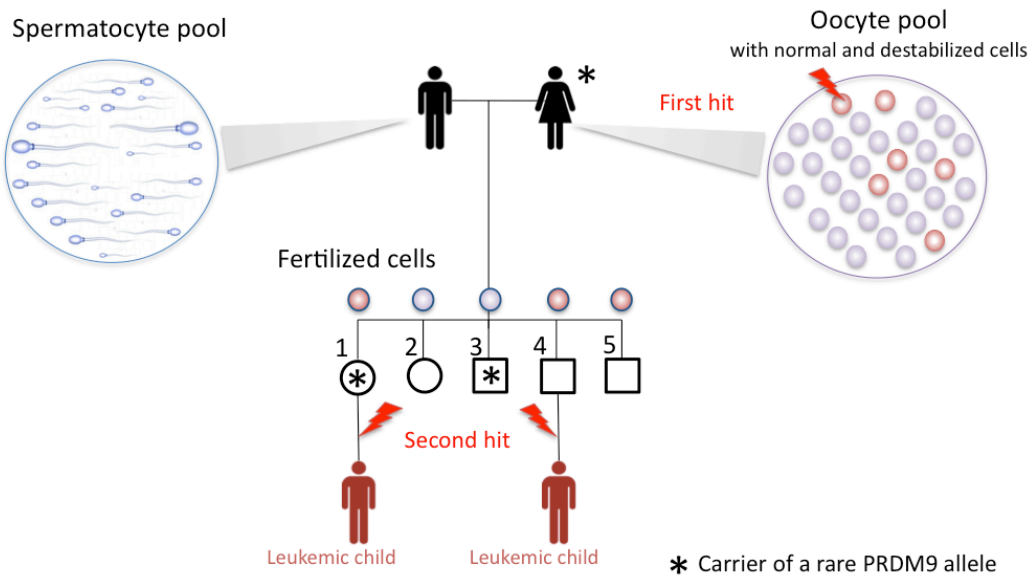


Figure 1. Proposed model of heredity for childhood acute lymphoblastic leukemia.

The predisposition comes from meiotic events in germline, triggered by the action of genetic factors, that affect genome stability (destabilized germ cells). Children who result from a destabilized germ cell (children 1 ,4 and 5) have an increased risk of developing leukemia. The onset of cancer depends on the occurrence of a second hit (children 1 and 4). Only children that received the predisposition factor risk transmitting the predisposition to their descendants (children 1 and 3). The predisposition segregates with onset of leukemia only in child 1.

Adaptative Patterns of Linkage

The third study included in this thesis (Chapter IV) presents unpublished results. Supplementary analyses are ongoing to replicate this effect in other populations and genomic datasets, including data from the 1000 genomes project (The 1000 Genomes Project Consortium 2010). It is, to my knowledge, the first study in humans establishing the reduced efficacy of selection on weakly deleterious mutations in regions of low recombination, in accordance with theoritical predictions made decades ago by Fisher, Muller and Hill and Robertson (Fisher 1930; Muller 1932; Muller 1964; Hill and Robertson 1966). A similar demonstration has been made recently in *Drosophila* (McGaugh et al. 2012). These results indicate that recombination rates have an important impact on how selection shapes diversity across the genome.

The advantageous effect of recombination comes from its capacity to create novel positive associations on which natural selection acts efficiently. Therefore, the action of natural selection in a species depends on the patterns of linkage between loci that have been built over time. These patterns themselves result from historical recombination rates and natural selection acting on genetic diversity. High recombination rates preserve variation within species that would otherwise be eliminated by natural selection. These regions are enriched with common variants, that drifted to intermediate and high frequencies mainly by genetic drift or, in some cases, by positive selection (Bersaglieri et al. 2004; Yi et al. 2010).

The ideal form of recombination would therefore be one that breaks down unfit combinations with a higher probability than fitter combinations of alleles. Coldspots are enriched for essential or 'housekeeping' genes that are relatively conserved across species (Smith et al. 2005), suggesting that in these regions, any mutation is very likely to be harmful and very unlikely to be favorable. Positive associations between negatively selected alleles are expected to be prevalent, which is supported by our demonstration that deleterious alleles are more frequently linked

to each other in coldspots than in high recombination regions (see Chapter IV, Table 2). Because a lack of recombination decreases the chance of *de novo* rearrangements within these regions, this distribution of recombination is likely beneficial for the fitness of the species. However, the cost associated is that individuals accumulate weakly deleterious mutations in conserved and essential genes, likely involved in diseases such as cardio-metabolic diseases, or late-onset cancers. It is therefore possible that the local rate of recombination is adaptative and the different selective forces acting among lineages may partly explain the extensive variation in recombination found among and within species.

During periods of rapid evolutionary change, selective pressures may act on local modifiers of recombination rates, such as PRDM9. PRDM9 ZnF array is evolving rapidly, with compelling evidence of positive selection acting at the DNA-binding determinant residues (Thomas et al. 2009; Ponting 2011). Strong sites that influence DSB activity are expected to be replaced by weaker ones by gene conversion, and erosion of PRDM9 binding sites has been observed in humans (Myers et al. 2010), providing a possible mechanism explaining why hotspots are short-lived. This erosion is likely compensated by the fast evolving nature of the minisatellite that form the PRDM9 ZnF array (Jeffreys et al. 2013). Therefore, the rapid turnover of recombination hotspots is consistent with the rapid evolution of PRDM9 DNA sequence (Ponting 2011). According to simulations, mutation and genetic drift alone could account for current PRDM9 diversity (Jeffreys et al. 2013), although this model does not explain the clear signature of strong and sustained positive selection at DNA contact residues within individual ZnF. Novel PRDM9 alleles would be positively selected if they increase recombination rates between linked polymorphisms under selection and disrupt linkage between deleterious alleles. Modifier alleles that increase recombination in a given genomic region also increase the fixation probability of beneficial alleles and subsequently hitchhike along with these variants to high frequencies. Conversly, novel PRDM9 alleles that increase crossover rates within essential ‘housekeeping’ genes may be selected against, as they can cause

deleterious effects by increasing the probability of unequal crossing-over causing rearrangements. Therefore, PRDM9 may be viewed as a component of a system that searches through the space of recombination landscapes, with the fittest PRDM9 alleles being the ones that reduce the mutational load at the population level while redirecting crossovers away from genomic regions associated with conditions that impact survival and fertility.

Mutation and Recombination

Indirect evidence suggests that mutation and recombination rates are associated in humans, as nucleotide diversity is weakly but significantly correlated with local recombination rates (Hellmann et al. 2003; Spencer et al. 2006). It remains unclear whether these factors are causally linked at the molecular level or if this correlation is driven by the action of natural selection. Recombination mechanisms, such as DSB formation and mismatch repair, might be mutagenic or indirectly affecting mutation rates. For example, recombination-associated mismatch repair in mammals is GC-biased while mutation rate is greatly influenced by base composition (Nachman and Crowell 2000). Another possibility is that both mutation and recombination rates covary with processes such as replication timing or chromatin organization (Stamatoyannopoulos et al. 2009; Schuster-Bockler and Lehner 2012).

Although the main goal of the third study presented in this thesis was to test for reduced selection in coldspots of recombination, we explored some aspect of the correlation between mutation and recombination rates as well. First, we found increased diversity in high recombination regions relative to coldspots at all minor allele frequencies. The correlation between SNP density and recombination rate remains significant after accounting for GC-content, suggesting that, within coding regions, the association between recombination and diversity is not fully explained through covariance of both factors with base composition (Spencer et al. 2006). Second, we found a significant correlation between minor allele frequency and recombination rate at the exon level. Along with the observation that these

correlations are stronger in genic regions than in non-genic regions (Lohmueller et al. 2011), these results suggest that the correlations observed are partly driven by natural selection. Finally, we found that the most recent mutations that arised in the sample, singletons, are observed in the same proportions across recombination environments. If the recombination process directly resulted in higher mutation rates, a slight enrichment of singletons in highly recombinogenic regions relative to coldpots would have been expected.

However, the mutational class formed by singletons is likely heterogeneous and may include sequencing errors and highly deleterious mutations kept at very low frequencies. Hence, our results do not directly show that the recombination process is not mutagenic. Furthermore, recombination has been shown to influence genetic diversity at the hotspot level (Spencer et al. 2006), therefore this effect may need to be evaluated at a finer scale than at the exon level. Biological samples analysed using NGS technologies now provide direct estimates of the number of new mutations occuring during meiosis in humans. In comparing DNA sequences from members of a nuclear family, *de novo* mutations can be directly observed, providing the most accurate estimates of the human mutation rate to date (Awadalla et al. 2010; Conrad et al. 2011; Cartwright et al. 2012).

Using powerful computational tools and genomic data from families, it is now possible to locate both recombination events and *de novo* mutations that occur during one meiosis and to determine whether they are directly associated. However, a large number of families are required to answer this question and more than two offsprings are needed to accurately assign recombination events to individuals. Alternatively, one may compile all *de novo* mutations identified in population and medical genomic projects so far, and assess whether they more likely occur near LD-based hotspots (or PRDM9 binding targets) than expected by chance. This should be done separately for males and females, as gender and age effects not only modulate recombination rates, but also mutation rates. In humans, males have a higher point mutation rate than females, possibly because men have more germ-line cell

divisions than women. Furthermore, the human mutation rate is known to increase with paternal age (Crow 1997; Kong et al. 2012) while no effect of maternal age on mutation rate has been observed. Age and gender impact mutation and recombination rates in opposite ways, suggesting that these effects result from different factors, and need to be accounted for in the analyses proposed here.

CONCLUSION

Sexual reproduction in eukaryotes arose about 850 million years ago and have been maintained ever since. Meiosis is a major evolutionary innovation and allows for recombination between homologous chromosome. In this thesis, I analysed genomic data generated by new biotechnologies, that have transformed the field of medical and population genetics in recent years. I developed and used bioinformatic and statistical approaches to explore the recombination process and the variability it creates. The results presented here, combined with results from other studies, highlight the fact that variation in genetic recombination has costs associated with human diseases. In mammals, recombination is required for the proper segregation of chromosomes. A functioning PRDM9, which redirects recombination within hotspots, is also essential for completion of meiosis. In humans, weakly deleterious mutations accumulate in essential genes within low recombination regions, possibly modulating susceptibility to late onset diseases. Assessing the mutational load within these genes at the individual level may be an interesting pursuit for improved personalized medicine. In high-recombination regions, some sequences become fragile breakpoints enriched for unequal crossover in germline and somatic cells, causing major developmental disorders and, possibly, childhood cancer. It will be important to develop family-based designs to further understand these genetic conditions in children. Samples from parents of patients, along with accurate computational methods to analyse family data, should be routinely included in medical and cancer genomics projects.

REFERENCES

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**(4): 248-249.

Agrawal AF, Whitlock MC. 2012. Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. *Annu Rev Ecol Evol Syst*(43): 115–135.

Aguade M, Miyashita N, Langley CH. 1989. Reduced variation in the yellow-achaete-scute region in natural populations of Drosophila melanogaster. *Genetics* **122**(3): 607-615.

Allers T, Lichten M. 2001. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* **106**(1): 47-57.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**(10): R106.

Anderson K, Lutz C, van Delft FW, Bateman CM, Guo Y, Colman SM, Kempski H, Moorman AV, Titley I, Swansbury J et al. 2011. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**(7330): 356-361.

Ashburner M. 1989. *Drosophila : a laboratory handbook*. Cold Spring Harbor Laboratory Press.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**(6078): 193-198.

Awadalla P. 2003. The evolutionary genomics of pathogen recombination. *Nat Rev Genet* **4**(1): 50-60.

Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, Hamet P, Laberge C. 2012. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int J Epidemiol*.

Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR, Cote M, Henrion E, Spiegelman D, Tarabeux J et al. 2010. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet* **87**(3): 316-324.

Baccichet A, Qualman SK, Sinnett D. 1997. Allelic loss in childhood acute lymphoblastic leukemia. *Leuk Res* **21**(9): 817-823.

Bacon C, Rappold GA. 2012. The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. *Hum Genet* **131**(11): 1687-1698.

Badge RM, Yardley J, Jeffreys AJ, Armour JA. 2000. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum Mol Genet* **9**(8): 1239-1244.

Bannister LA, Schimenti JC. 2004. Homologous recombinational repair proteins in mouse meiosis. *Cytogenet Genome Res* **107**(3-4): 191-200.

Barbouti A, Stankiewicz P, Nusbaum C, Cuomo C, Cook A, Hoglund M, Johansson B, Hagemeijer A, Park SS, Mitelman F et al. 2004. The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats. *Am J Hum Genet* **74**(1): 1-10.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**(5967): 836-840.

Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. *Science* **326**(5959): 1538-1541.

Beerman H, Kluin PM, Hermans J, van de Velde CJ, Cornelisse CJ. 1990. Prognostic significance of DNA-ploidy in a series of 690 primary breast cancer patients. *Int J Cancer* **45**(1): 34-39.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**(6369): 519-520.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**(2): 573-580.

Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* **42**(10): 859-863.

Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ. 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc Natl Acad Sci U S A* **108**(30): 12378-12383.

Bernstein H, Byers GS, Michod RE. 1981. Evolution of Sexual Reproduction: Importance of DNA Repair, Complementation, and Variation. *The American Naturalist* **117**(4): 537-549.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**(6): 1111-1120.

Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vezina H. 2011. Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* **144**(3): 432-441.

Birky CW, Jr., Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci U S A* **85**(17): 6414-6418.

Bjorge T, Cnattingius S, Lie RT, Tretli S, Engeland A. 2008. Cancer risk in children with birth defects and in their families: a population based cohort study of 5.2 million children from Norway and Sweden. *Cancer Epidemiol Biomarkers Prev* **17**(3): 500-506.

Bleazard T, Ju YS, Sung J, Seo JS. 2013. Fine-scale mapping of meiotic recombination in Asians. *BMC Genet* **14**: 19.

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M. 2008. Natural selection on genes that underlie human disease susceptibility. *Current biology : CB* **18**(12): 883-889.

Blouin JL, Christie DH, Gos A, Lynn A, Morris MA, Ledbetter DH, Chakravarti A, Antonarakis SE. 1995. A new dinucleotide repeat polymorphism at the telomere of chromosome 21q reveals a significant difference between male and female rates of recombination. *Am J Hum Genet* **57**(2): 388-394.

Boettger LM, Handsaker RE, Zody MC, McCarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**(8): 881-885.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**(5611): 1391-1394.

Borel C, Cheung F, Stewart H, Koolen DA, Phillips C, Thomas NS, Jacobs PA, Eliez S, Sharp AJ. 2012. Evaluation of PRDM9 variation as a risk factor for recurrent genomic disorders and chromosomal non-disjunction. *Hum Genet* **131**(9): 1519-1524.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**(5): e1000083.

Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485**(7400): 642-645.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* **63**(3): 861-869.

Broman KW, Weber JL. 2000. Characterization of human crossover interference. *Am J Hum Genet* **66**(6): 1911-1926.

Brown JD, O'Neill RJ. 2010. Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annu Rev Genomics Hum Genet* **11**: 291-316.

Bugge M, Collins A, Hertz JM, Eiberg H, Lundsteen C, Brandt CA, Bak M, Hansen C, Delozier CD, Lespinasse J et al. 2007. Non-disjunction of chromosome 13. *Hum Mol Genet* **16**(16): 2004-2010.

Bugge M, Collins A, Petersen MB, Fisher J, Brandt C, Hertz JM, Tranebjaerg L, de Lozier-Blanchet C, Nicolaides P, Brondum-Nielsen K et al. 1998. Non-disjunction of chromosome 18. *Hum Mol Genet* **7**(4): 661-669.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**: 94.

Bullaughey K, Przeworski M, Coop G. 2008. No effect of recombination on the efficacy of natural selection in primates. *Genome Res* **18**(4): 544-554.

Burgess DJ. 2011. Cancer genetics: Initially complex, always heterogeneous. *Nat Rev Genet* **12**(3): 154.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A et al. 2002. A human genome diversity cell line panel. *Science* **296**(5566): 261-262.

Cartwright RA, Hussin J, Keebler JE, Stone EA, Awadalla P. 2012. A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. *Stat Appl Genet Mol Biol* **11**(2).

Carvalho CM, Lupski JR. 2008. Copy number variation at the breakpoint region of isochromosome 17q. *Genome Res* **18**(11): 1724-1732.

Casals F, Idaghdour Y, Hussin J, Awadalla P. 2012. Next-generation sequencing approaches for genetic mapping of complex diseases. *J Neuroimmunol* **248**(1-2): 10-22.

Chang X, Wang K. 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* **49**(7): 433-436.

Charlesworth B, Charlesworth D. 1997. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genet Res* **70**(1): 63-73.

-. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**(1403): 1563-1572.

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* **8**(10): 762-775.

Chen JM, Cooper DN, Ferec C, Kehrer-Sawatzki H, Patrinos GP. 2010. Genomic rearrangements in inherited disease and cancer. *Semin Cancer Biol* **20**(4): 222-233.

Cheung VG, Burdick JT, Hirschmann D, Morley M. 2007. Polymorphic variation in human meiotic recombination. *Am J Hum Genet* **80**(3): 526-530.

Chi P, Allis CD, Wang GG. 2010. Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nat Rev Cancer* **10**(7): 457-469.

Choo KH. 1998. Why is the centromere so cold? *Genome Res* **8**(2): 81-82.

Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG. 2009. Genetic analysis of variation in human meiotic recombination. *PLoS Genet* **5**(9): e1000648.

Christensen GL, Ivanov IP, Atkins JF, Mielnik A, Schlegel PN, Carrell DT. 2005. Screening the SPO11 and EIF5A2 genes in a population of infertile men. *Fertil Steril* **84**(3): 758-760.

Cloos PA, Christensen J, Agger K, Helin K. 2008. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes Dev* **22**(9): 1115-1140.

Cohen PE, Pollack SE, Pollard JW. 2006. Genetic analysis of chromosome pairing, recombination, and cell cycle control during first meiotic prophase in mammals. *Endocr Rev* **27**(4): 398-426.

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**(7): 712-714.

Consortium TIH. 2005. A haplotype map of the human genome. *Nature* **437**(7063): 1299-1320.

Coop G, Myers SR. 2007. Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet* **3**(3): e35.

Coop G, Przeworski M. 2007. An evolutionary view of human recombination. *Nat Rev Genet* **8**(1): 23-34.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**(5868): 1395-1398.

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. 2010. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* **7**(4): 250-251.

Cripe L, Andelfinger G, Martin LJ, Shooner K, Benson DW. 2004. Bicuspid aortic valve is heritable. *J Am Coll Cardiol* **44**(1): 138-143.

Crismani W, Girard C, Froger N, Pradillo M, Santos JL, Chelysheva L, Copenhaver GP, Horlow C, Mercier R. 2012. FANCM limits meiotic crossovers. *Science* **336**(6088): 1588-1590.

Crow JF. 1997. The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci U S A* **94**(16): 8380-8386.

Crowe ML. 2005. SeqDoC: rapid SNP and mutation detection by direct comparison of DNA sequence chromatograms. *BMC Bioinformatics* **6**: 133.

D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**(6): 743-753.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15): 2156-2158.

Darai-Ramqvist E, Sandlund A, Muller S, Klein G, Imreh S, Kost-Alimova M. 2008. Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome Res* **18**(3): 370-379.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**(12): e1001025.

de La Roche Saint-Andre C. 2008. Alternative ends: telomeres and meiosis. *Biochimie* **90**(1): 181-189.

Delaneau O, Coulonges C, Zagury JF. 2008. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**: 540.

Dempsey MA, Schwartz S, Waggoner DJ. 2007. Mosaicism del(22)(q11.2q11.2)/dup(22)(q11.2q11.2) in a patient with features of 22q11.2 deletion syndrome. *Am J Med Genet A* **143A**(10): 1082-1086.

Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, Whirl-Carrillo M, Wheeler MT, Dudley JT, Byrnes JK et al. 2011. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* **7**(9): e1002280.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**(6570): 152-154.

Digweed M, Sperling K. 2004. Nijmegen breakage syndrome: clinical manifestation of defective response to DNA double-strand breaks. *DNA Repair (Amst)* **3**(8-9): 1207-1217.

Dobzhansky T. 1937. Genetic nature of species differences. . *Am Nat* **71**: 404–420.

Doggett NA, Xie G, Meincke LJ, Sutherland RD, Mundt MO, Berbari NS, Davy BE, Robinson ML, Rudd MK, Weber JL et al. 2006. A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics* **88**(6): 762-771.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES et al. 1987. A genetic linkage map of the human genome. *Cell* **51**(2): 319-337.

Draper GJ, Sanders BM, Lennox EL, Brownbill PA. 1996. Patterns of childhood cancer among siblings. *Br J Cancer* **74**(1): 152-158.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4**(5): e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285-311.

Easton DF, Eeles RA. 2008. Genome-wide association studies in cancer. *Hum Mol Genet* **17**(R2): R109-115.

Engelstadter J. 2008. Constraints on the evolution of asexual reproduction. *Bioessays* **30**(11-12): 1138-1150.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**(2): 891-900.

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. *Genetics* **159**(3): 1299-1318.

-. 2002. Approximate likelihood methods for estimating local recombination rates. *J R Stat Soc Lond* **B**(64): 657–680.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* **78**(2): 737-756.

Fernandez-Capetillo O, Lee A, Nussenzweig M, Nussenzweig A. 2004. H2AX: the histone guardian of the genome. *DNA Repair (Amst)* **3**(8-9): 959-967.

Ferree PM, Barbash DA. 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS Biol* **7**(10): e1000234.

Fisher RA. 1930. The Genetical Theory of Natural Selection. *Clarendon Press, Oxford*.

Flanagan SE, Patch AM, Ellard S. 2010. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers* **14**(4): 533-537.

Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, Przeworski M. 2011. Variation in human recombination rates and its genetic determinants. *PLoS One* **6**(6): e20321.

Fog CK, Galli GG, Lund AH. 2011. PRDM proteins: Important players in differentiation and disease. *Bioessays*.

Freeman GH, Halton JH. 1951. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* **38**(1-2): 141-149.

Gair JL, Arbour L, Rupps R, Jiang R, Bruyere H, Robinson WP. 2005. Recurrent trisomy 21: four cases in three generations. *Clin Genet* **68**(5): 430-435.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet* **25**(1): 1-5.

Ganmore I, Smooha G, Izraeli S. 2009. Constitutional aneuploidy and cancer predisposition. *Hum Mol Genet* **18**(R1): R84-93.

Garcia-Higuera I, Taniguchi T, Ganesan S, Meyn MS, Timmers C, Hejna J, Grompe M, D'Andrea AD. 2001. Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. *Mol Cell* **7**(2): 249-262.

German J. 1997. Bloom's syndrome. XX. The first 100 cancers. *Cancer Genet Cytogenet* **93**(1): 100-106.

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**(5): 555-566.

Glesne D, Huberman E. 2006. Smad6 is a protein kinase X phosphorylation substrate and is required for HL-60 cell differentiation. *Oncogene* **25**(29): 4086-4098.

Godmann M, Auger V, Ferraroni-Aguiar V, Di Sauro A, Sette C, Behr R, Kimmins S. 2007. Dynamic regulation of histone H3 methylation at lysine 4 in mammalian spermatogenesis. *Biol Reprod* **77**(5): 754-764.

Gomez CA, Ptaszek LM, Villa A, Bozzi F, Sobacchi C, Brooks EG, Notarangelo LD, Spanopoulou E, Pan ZQ, Vezzoni P et al. 2000. Mutations in conserved regions of the predicted RAG2 kelch repeats block initiation of V(D)J recombination and result in primary immunodeficiencies. *Mol Cell Biol* **20**(15): 5653-5664.

Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, Karra K, Davydov E, Batzoglou S, Myers RM et al. 2010. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res* **20**(3): 301-310.

Gordo I, Navarro A, Charlesworth B. 2002. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161**(2): 835-848.

Graakjaer J, Londono-Vallejo JA, Christensen K, Kolvraa S. 2006. The pattern of chromosome-specific variations in telomere length in humans shows signs of heritability and is maintained through life. *Ann N Y Acad Sci* **1067**: 311-316.

Greaves M. 1999. Molecular genetics, natural history and the demise of childhood leukaemia. *Eur J Cancer* **35**(14): 1941-1953.

-. 2003. Pre-natal origins of childhood leukemia. *Rev Clin Exp Hematol* **7**(3): 233-245.

-. 2006. Infection, immune responses and the aetiology of childhood leukaemia. *Nat Rev Cancer* **6**(3): 193-203.

Greaves MF, Wiemels J. 2003. Origins of chromosome translocations in childhood leukaemia. *Nat Rev Cancer* **3**(9): 639-649.

Greenway SC, Pereira AC, Lin JC, DePalma SR, Israel SJ, Mesquita SM, Ergul E, Conta JH, Korn JM, McCarroll SA et al. 2009. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat Genet* **41**(8): 931-935.

Greer EL, Maures TJ, Ucar D, Hauswirth AG, Mancini E, Lim JP, Benayoun BA, Shi Y, Brunet A. 2011. Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. *Nature* **479**(7373): 365-371.

Grey C, Barthes P, Chauveau-Le Friec G, Langa F, Baudat F, de Massy B. 2011. Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol* **9**(10): e1001176.

Griffing B, Landridge J. 1963. Factors affecting crossing over in the tomato. *Aust J Biol Sci* **16**: 826-837.

Griffiths RC, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol* **3**(4): 479-502.

Gruhn B, Taub JW, Ge Y, Beck JF, Zell R, Hafer R, Hermann FH, Debatin KM, Steinbach D. 2008. Prenatal origin of childhood acute lymphoblastic leukemia, association with birth weight and hyperdiploidy. *Leukemia* **22**(9): 1692-1697.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**(1): 4.

Guillon H, Baudat F, Grey C, Liskay RM, de Massy B. 2005. Crossover and noncrossover pathways in mouse meiosis. *Mol Cell* **20**(4): 563-573.

Gurney JG, Severson RK, Davis S, Robison LL. 1995. Incidence of cancer in children in the United States. Sex-, race-, and 1-year age-specific rates by histologic type. *Cancer* **75**(8): 2186-2195.

Gurrin LC, Scurrah KJ, Hazelton ML. 2005. Tutorial in biostatistics: spline smoothing with linear mixed models. *Stat Med* **24**(21): 3361-3381.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biol* **8**(2): R18.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144**(5): 646-674.

Handel MA, Schimenti JC. 2010. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat Rev Genet* **11**(2): 124-136.

HapMap Consortium Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Hasle H. 2001. Pattern of malignant disorders in individuals with Down's syndrome. *Lancet Oncol* **2**(7): 429-436.

Hassold T, Hall H, Hunt P. 2007. The origin of human aneuploidy: where we have been, where we are going. *Hum Mol Genet* **16 Spec No. 2**: R203-208.

Hassold T, Hunt P. 2001. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet* **2**(4): 280-291.

-. 2009. Maternal age and chromosomally abnormal pregnancies: what we know and what we wish we knew. *Curr Opin Pediatr* **21**(6): 703-708.

Hassold T, Judis L, Chan ER, Schwartz S, Seftel A, Lynn A. 2004. Cytological studies of meiotic recombination in human males. *Cytogenet Genome Res* **107**(3-4): 249-255.

Hassold T, Merrill M, Adkins K, Freeman S, Sherman S. 1995. Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *Am J Hum Genet* **57**(4): 867-874.

Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5**(1): e1000327.

Hayashi K, Yoshida K, Matsui Y. 2005. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* **438**(7066): 374-378.

He C, Weeks DE, Buyske S, Abecasis GR, Stewart WC, Matise TC. 2011. Enhanced genetic maps from family-based disease studies: population-specific comparisons. *BMC Med Genet* **12**: 15.

Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. 2007. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood* **109**(2): 683-692.

Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. 2010. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. *Haematologica* **95**(9): 1608-1611.

Helleday T. 2010. Homologous recombination in cancer development, treatment and development of drug resistance. *Carcinogenesis* **31**(6): 955-960.

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**(6): 1527-1535.

Henderson SA, Edwards RG. 1968. Chiasma frequency and maternal age in mammals. *Nature* **218**(5136): 22-28.

Hey J, Kliman RM. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics* **160**(2): 595-608.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**(3): 269-294.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics* **38**(6): 226–231.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL et al. 2011. The landscape of recombination in African Americans. *Nature* **476**(7359): 170-175.

Hinton RB, Jr., Martin LJ, Tabangin ME, Mazwi ML, Cripe LH, Benson DW. 2007. Hypoplastic left heart syndrome is heritable. *J Am Coll Cardiol* **50**(16): 1590-1595.

Hodges CA, Revenkova E, Jessberger R, Hassold TJ, Hunt PA. 2005. SMC1beta-deficient female mice provide evidence that cohesins are a missing link in age-related nondisjunction. *Nat Genet* **37**(12): 1351-1355.

Hodgkinson A, Casals F, Idaghdour Y, Grenier J-C, Le Deist F, Haddad E, Awadalla P. 2013. Evolutionary constraint reveals individual mutation load variability among humans. *BMC Genomics* **(submitted)**.

Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, Payne-Turner D, Churchman M, Andersson A, Chen SC et al. 2013. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet*.

Holt D, Dreimanis M, Pfeiffer M, Firgaira F, Morley A, Turner D. 1999. Interindividual variation in mitotic recombination. *Am J Hum Genet* **65**(5): 1423-1427.

Housworth EA, Stahl FW. 2003. Crossover interference in humans. *Am J Hum Genet* **73**(1): 188-197.

Hubert R, MacDonald M, Gusella J, Arnheim N. 1994. High resolution localization of recombination hot spots using sperm typing. *Nat Genet* **7**(3): 420-424.

Hudson RR. 2001. Two-locus sampling distributions and their application. *Genetics* **159**(4): 1805-1817.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**(1): 147-164.

Hunt PA, Hassold TJ. 2002. Sex matters in meiosis. *Science* **296**(5576): 2181-2183.

Hussin J, Roy-Gagnon MH, Gendron R, Andelfinger G, Awadalla P. 2011. Age-dependent recombination rates in human pedigrees. *PLoS Genet* **7**(9): e1002251.

Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, Martin HC, Miclaus K, Jadallah SJ, Goldstein DB, Wolfinger RD et al. 2010. Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat Genet* **42**(1): 62-67.

Irie S, Tsujimura A, Miyagawa Y, Ueda T, Matsuoka Y, Matsui Y, Okuyama A, Nishimune Y, Tanaka H. 2009. Single-nucleotide polymorphisms of the PRDM9 (MEISETZ) gene in patients with nonobstructive azoospermia. *J Androl* **30**(4): 426-431.

Ishii K, Charlesworth B. 1977. Associations between allozyme loci and gene arrangements due to hitch-hiking effects of new inversions. . *Genet Res* **30**: 93-106.

Ishisaki A, Yamato K, Hashimoto S, Nakao A, Tamaki K, Nonaka K, ten Dijke P, Sugino H, Nishihara T. 1999. Differential inhibition of Smad6 and Smad7 on bone morphogenetic protein- and activin-mediated growth arrest and apoptosis in B cells. *J Biol Chem* **274**(19): 13637-13642.

Jacquemont S, Boceno M, Rival JM, Mechinaud F, David A. 2002. High risk of malignancy in mosaic variegated aneuploidy syndrome. *Am J Med Genet* **109**(1): 17-21; discussion 16.

Jeffreys AJ, Cotton VE, Neumann R, Lam KW. 2013. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. *Proc Natl Acad Sci U S A* **110**(2): 600-605.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**(2): 217-222.

Jeffreys AJ, Neumann R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* **31**(3): 267-271.

Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nat Genet* **37**(6): 601-606.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* **14**(4): 528-538.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**(5653): 2141-2144.

Jorgenson E, Tang H, Gadde M, Province M, Leppert M, Kardia S, Schork N, Cooper R, Rao DC, Boerwinkle E et al. 2005. Ethnicity and human genetic linkage maps. *Am J Hum Genet* **76**(2): 276-290.

Joyce JA, Schofield PN. 1998. Genomic imprinting and cancer. *Mol Pathol* **51**(4): 185-190.

Ju YS, Park H, Lee MK, Kim JI, Sung J, Cho SI, Seo JS. 2008. A genome-wide Asian genetic map and ethnic comparison: the GENDISCAN study. *BMC Genomics* **9**: 554.

Kan R, Sun X, Kolas NK, Avdievich E, Kneitz B, Edelmann W, Cohen PE. 2008. Comparative analysis of meiotic progression in female mice bearing mutations in genes of the DNA mismatch repair pathway. *Biol Reprod* **78**(3): 462-471.

Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH, Jr. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev* **23**(11): 1303-1312.

Kauppi L, Jeffreys AJ, Keeney S. 2004. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet* **5**(6): 413-424.

Keefe DL, Marquard K, Liu L. 2006. The telomere theory of reproductive senescence in women. *Curr Opin Obstet Gynecol* **18**(3): 280-285.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**(6082): 740-743.

Kim Y, Lach FP, Desetty R, Hanenberg H, Auerbach AD, Smogorzewska A. 2011. Mutations of the SLX4 gene in Fanconi anemia. *Nat Genet* **43**(2): 142-146.

Kimura M. 1956. A model of a genetic system which leads to closer linkage by natural selection. *Evolution* **10**: 278-287.

Kingman JF. 1982. The coalescent. *Stoch Process Appl* **13**: 235-248.

Kirilyuk A, Tolstonog GV, Damert A, Held U, Hahn S, Lower R, Buschmann C, Horn AV, Traub P, Schumann GG. 2008. Functional endogenous LINE-1 retrotransposons are expressed and mobilized in rat chloroleukemia cells. *Nucleic Acids Res* **36**(2): 648-665.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* **20**(10): 1916-1924.

Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, Sigurdardottir S, Richardsson B, Jonsdottir J, Thorgeirsson T, Frigge ML et al. 2004. Recombination rate and reproductive success in humans. *Nat Genet* **36**(11): 1203-1206.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Wong WS et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**(7412): 471-475.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31**(3): 241-247.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Gylfason A, Kristinsson KT, Gudjonsson SA et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**(7319): 1099-1103.

Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, Jonsdottir GM, Gudjonsson SA, Sverrisson S, Thorlacius T et al. 2008. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science* **319**(5868): 1398-1401.

Kong F, Zhu J, Wu J, Peng J, Wang Y, Wang Q, Fu S, Yuan LL, Li T. 2011. dbCRID: a database of chromosomal rearrangements in human diseases. *Nucleic Acids Res* **39**(Database issue): D895-900.

Kost-Alimova M, Kiss H, Fedorova L, Yang Y, Dumanski JP, Klein G, Imreh S. 2003. Coincidence of synteny breakpoints with malignancy-related deletions on human chromosome 3. *Proc Natl Acad Sci U S A* **100**(11): 6622-6627.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol* **29**(3): 1047-1057.

Kovaleva NV. 2010. Germ-line transmission of trisomy 21: Data from 80 families suggest an implication of grandmaternal age and a high frequency of female-specific trisomy rescue. *Mol Cytogenet* **3**: 7.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**(6): 1347-1363.

Kuhner MK, Yamato J, Felsenstein J. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**(3): 1393-1401.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**(7): 1073-1081.

Lamb NE, Feingold E, Savage A, Avramopoulos D, Freeman S, Gu Y, Hallberg A, Hersey J, Karadima G, Pettay D et al. 1997. Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjunction of chromosome 21. *Hum Mol Genet* **6**(9): 1391-1399.

Lamb NE, Freeman SB, Savage-Austin A, Pettay D, Taft L, Hersey J, Gu Y, Shen J, Saker D, May KM et al. 1996. Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II. *Nat Genet* **14**(4): 400-405.

Lamb NE, Yu K, Shaffer J, Feingold E, Sherman SL. 2005. Association between maternal age and meiotic recombination for trisomy 21. *Am J Hum Genet* **76**(1): 91-99.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**(8): 2363-2367.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet* **24**(3): 114-123.

Laurie DA, Hulten MA. 1985. Further studies on chiasma distribution and interference in the human male. *Ann Hum Genet* **49**(Pt 3): 203-214.

Lazaro C, Gaona A, Ainsworth P, Tenconi R, Vidaud D, Kruyer H, Ars E, Volpini V, Estivill X. 1996. Sex differences in mutational rate and mutational mechanism in the NF1 gene in neurofibromatosis type 1 patients. *Hum Genet* **98**(6): 696-699.

Lee HW, Blasco MA, Gottlieb GJ, Horner JW, 2nd, Greider CW, DePinho RA. 1998. Essential role of mouse telomerase in highly proliferative organs. *Nature* **392**(6676): 569-574.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**(7): 1235-1247.

Lee PS, Greenwell PW, Dominska M, Gawel M, Hamilton M, Petes TD. 2009. A fine-structure map of spontaneous mitotic crossovers in the yeast Saccharomyces cerevisiae. *PLoS Genet* **5**(3): e1000410.

Lehtonen J, Jennions MD, Kokko H. 2012. The many costs of sex. *Trends Ecol Evol* **27**(3): 172-178.

Lenzi ML, Smith J, Snowden T, Kim M, Fishel R, Poulos BK, Cohen PE. 2005. Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis i in human oocytes. *Am J Hum Genet* **76**(1): 112-127.

Lercher MJ, Hurst LD. 2003. Imprinted chromosomal regions of the human genome have unusually high recombination rates. *Genetics* **165**(3): 1629-1632.

Lewis SM, Agard E, Suh S, Czyzyk L. 1997. Cryptic signals and the fidelity of V(D)J joining. *Mol Cell Biol* **17**(6): 3125-3136.

Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**(1): 49-67.

-. 1971. The effect of genetic linkage on the mean fitness of a population. *Proc Natl Acad Sci U S A* **68**(5): 984-986.

Li B, Abecasis G. 2011. Polymutt: a tool for calling polymorphism and de novo mutations. http://genome.sph.umich.edu/wiki/Polymutt.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**(4): 2213-2233.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T et al. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42**(11): 969-972.

Lieber MR. 2008. The mechanism of human nonhomologous DNA end joining. *J Biol Chem* **283**(1): 1-5.

Lieber MR, Gu J, Lu H, Shimazaki N, Tsai AG. 2010. Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans. *Subcell Biochem* **50**: 279-296.

Lieber MR, Lu H, Gu J, Schwarz K. 2008. Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. *Cell Res* **18**(1): 125-133.

Lindsay SJ, Khajavi M, Lupski JR, Hurles ME. 2006. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am J Hum Genet* **79**(5): 890-902.

Liu L, Franco S, Spyropoulos B, Moens PB, Blasco MA, Keefe DL. 2004. Irregular telomeres impair meiotic synapsis and recombination in mice. *Proc Natl Acad Sci U S A* **101**(17): 6496-6501.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**(10): e1002326.

Lopes J, Ravise N, Vandenberghe A, Palau F, Ionasescu V, Mayer M, Levy N, Wood N, Tachi N, Bouche P et al. 1998. Fine mapping of de novo CMT1A and HNPP rearrangements within CMT1A-REPs evidences two distinct sex-dependent mechanisms and candidate sequences involved in recombination. *Hum Mol Genet* **7**(1): 141-148.

Lopes J, Vandenberghe A, Tardieu S, Ionasescu V, Levy N, Wood N, Tachi N, Bouche P, Latour P, Brice A et al. 1997. Sex-dependent rearrangements resulting in CMT1A and HNPP. *Nat Genet* **17**(2): 136-137.

Lupski JR. 2004. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biol* **5**(10): 242.

-. 2006. Genome structural variation and sporadic disease traits. *Nat Genet* **38**(9): 974-976.

Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA et al. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66**(2): 219-232.

Lynch M, Blanchard JL. 1998. Deleterious mutation accumulation in organelle genomes. *Genetica* **102-103**(1-6): 29-39.

Lynn A, Ashley T, Hassold T. 2004. Variation in human meiotic recombination. *Annu Rev Genomics Hum Genet* **5**: 317-349.

Maddox JF, Davies KP, Crawford AM, Hulme DJ, Vaiman D, Cribiu EP, Freking BA, Beh KJ, Cockett NE, Kang N et al. 2001. An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Res* **11**(7): 1275-1289.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454**(7203): 479-485.

Marais G, Charlesworth B. 2003. Genome evolution: recombination speeds up adaptive evolution. *Curr Biol* **13**(2): R68-70.

Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci U S A* **98**(10): 5688-5692.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**(7): 906-913.

Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X et al. 2011. The functional spectrum of low-frequency coding variation. *Genome Biol* **12**(9): R84.

Martin JW, Yoshimoto M, Ludkovski O, Thorner PS, Zielenska M, Squire JA, Nuin PA. 2010. Analysis of segmental duplications, mouse genome synteny and recurrent cancer-associated amplicons in human chromosome 6p21-p12. *Cytogenet Genome Res* **128**(4): 199-213.

Matthews AG, Kuo AJ, Ramon-Maiques S, Han S, Champagne KS, Ivanov D, Gallardo M, Carney D, Cheung P, Ciccone DN et al. 2007. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature* **450**(7172): 1106-1110.

Mayer S, Bruderlein S, Perner S, Waibel I, Holdenried A, Ciloglu N, Hasel C, Mattfeldt T, Nielsen KV, Moller P. 2006. Sex-specific telomere length profiles and age-dependent erosion dynamics of individual chromosome arms in humans. *Cytogenet Genome Res* **112**(3-4): 194-201.

McBride KL, Pignatelli R, Lewin M, Ho T, Fernbach S, Menesses A, Lam W, Leal SM, Kaplan N, Schliekelman P et al. 2005. Inheritance analysis of congenital left ventricular outflow tract obstruction malformations: Segregation, multiplex relative risk, and heritability. *Am J Med Genet A* **134A**(2): 180-186.

McDougall A, Elliott DJ, Hunter N. 2005. Pairing, connecting, exchanging, pausing and pulling chromosomes. *EMBO Rep* **6**(2): 120-125.

McGaugh SE, Heil CS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MA. 2012. Recombination modulates how selection affects linked sites in Drosophila. *PLoS Biol* **10**(11): e1001422.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**(9): 1297-1303.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**(3): 1231-1241.

McVean G, Myers S. 2010. PRDM9 marks the spot. *Nat Genet* **42**(10): 821-822.

McVean GA, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**(2): 929-944.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**(5670): 581-584.

McVicker G, Green P. 2010. Genomic signatures of germline gene expression. *Genome Res* **20**(11): 1503-1511.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* **21**(6): 984-990.

Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. 2009. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* **323**(5912): 373-375.

Mitelman F, Johansson B, Mertens F. 2011. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Miyamoto T, Hasuike S, Yogev L, Maduro MR, Ishikawa M, Westphal H, Lamb DJ. 2003. Azoospermia in patients heterozygous for a mutation in SYCP3. *Lancet* **362**(9397): 1714-1719.

Moldovan GL, D'Andrea AD. 2009. How the fanconi anemia pathway guards the genome. *Annu Rev Genet* **43**: 223-249.

Morgan T. 1914. No crossing over in the male of drosophila of genes in the second and third pairs of chromosomes. *Biol Bull* **26**: 195–204.

Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, Hows JM, Navarrete C, Greaves M. 2002. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci U S A* **99**(12): 8242-8247.

Morris JA, Gardner MJ. 1988. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed)* **296**(6632): 1313-1316.

Morton NE, Jacobs PA, Hassold T, Wu D. 1988. Maternal age in trisomy. *Ann Hum Genet* **52**(Pt 3): 227-235.

Moshous D, Callebaut I, de Chasseval R, Corneo B, Cavazzana-Calvo M, Le Deist F, Tezcan I, Sanal O, Bertrand Y, Philippe N et al. 2001. Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. *Cell* **105**(2): 177-186.

Muller H. 1932. Some genetics aspect of sex. *Am Nat* **66**: 118-138.

Muller H. 1942. Isolating mechanisms, evolution and temperature. . *Biol Symp* **6**: 71–125.

Muller H. 1950. Our load of mutations. *Am J Hum Genet* **2**(2): 111-176.

-. 1964. The Relation of Recombination to Mutational Advance. *Mutat Res* **106**: 2-9.

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**(5734): 613-617.

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillen J, Sheffield VC, Sunden S, Duyk GM et al. 1994. A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* **265**(5181): 2049-2054.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746): 321-324.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**(5967): 876-879.

Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40**(9): 1124-1129.

Myers SR, Griffiths RC. 2003. Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**(1): 375-394.

Nachman JB, Heerema NA, Sather H, Camitta B, Forestier E, Harrison CJ, Dastugue N, Schrappe M, Pui CH, Basso G et al. 2007. Outcome of treatment in children with hypodiploid acute lymphoblastic leukemia. *Blood* **110**(4): 1112-1115.

Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* **17**(9): 481-485.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1): 297-304.

Necsulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat* **32**(2): 198-206.

Necsulea A, Semon M, Duret L, Hurst LD. 2009. Monoallelic expression and tissue specificity are associated with high crossover rates. *Trends Genet* **25**(12): 519-522.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**(6090): 100-104.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**(13): 3812-3814.

Nikopoulos K, Schrauwen I, Simon M, Collin RW, Veckeneer M, Keymolen K, Van Camp G, Cremers FP, van den Born LI. 2011. Autosomal recessive Stickler syndrome in two families is caused by mutations in the COL9A1 gene. *Invest Ophthalmol Vis Sci* **52**(7): 4774-4779.

Notta F, Mullighan CG, Wang JC, Poeppl A, Doulatov S, Phillips LA, Ma J, Minden MD, Downing JR, Dick JE. 2011. Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. *Nature* **469**(7330): 362-367.

O'Reilly PF, Balding DJ. 2011. Admixture provides new insights into recombination. *Nat Genet* **43**(9): 819-820.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet* **5**(12): e1000753.

Oliver TR, Feingold E, Yu K, Cheung V, Tinker S, Yadav-Shah M, Masse N, Sherman SL. 2008. New insights into human nondisjunction of chromosome 21 in oocytes. *PLoS Genet* **4**(3): e1000033.

Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH. 2008. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* **24**(23): 2776-2777.

Otto SP, Lenormand T. 2002. Resolving the paradox of sex and recombination. *Nat Rev Genet* **3**(4): 252-261.

Page SL, Hawley RS. 2003. Chromosome choreography: the meiotic ballet. *Science* **301**(5634): 785-789.

Paigen K, Petkov P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet* **11**(3): 221-233.

Paigen K, Szatkiewicz JP, Sawyer K, Leahy N, Parvanov ED, Ng SH, Graber JH, Broman KW, Petkov PM. 2008. The recombinational anatomy of a mouse chromosome. *PLoS Genet* **4**(7): e1000119.

Pal C, Papp B, Hurst LD. 2001. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol Biol Evol* **18**(12): 2323-2326.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* **7**(5): 337-348.

Paldi A, Gyapay G, Jami J. 1995. Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies. *Curr Biol* **5**(9): 1030-1035.

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of mammalian recombination hotspots. *Science* **327**(5967): 835.

Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**(2): 100-109.

Persikov AV, Osada R, Singh M. 2009. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **25**(1): 22-29.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* **4**(7): 675-682.

Peters U, Hutter CM, Hsu L, Schumacher FR, Conti DV, Carlson CS, Edlund CK, Haile RW, Gallinger S, Zanke BW et al. 2012. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* **131**(2): 217-234.

Phadnis N, Orr HA. 2009. A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *Science* **323**(5912): 376-379.

Phillips RB. 1978. Pericentric inversions inv(2)(p11q13) and inv(2)(p13q11) in 2 unrelated families. *J Med Genet* **15**(5): 388-390.

Pimanda JE, Donaldson IJ, de Bruijn MF, Kinston S, Knezevic K, Huckle L, Piltz S, Landry JR, Green AR, Tannahill D et al. 2007. The SCL transcriptional network and BMP signaling pathway interact to regulate RUNX1 activity. *Proc Natl Acad Sci U S A* **104**(3): 840-845.

Polani PE, Crolla JA. 1991. A test of the production line hypothesis of mammalian oogenesis. *Hum Genet* **88**(1): 64-70.

Ponting CP. 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Trends Genet* **27**(5): 165-171.

Popa A, Samollow P, Gautier C, Mouchiroud D. 2012. The sex-specific impact of meiotic recombination on nucleotide composition. *Genome Biol Evol* **4**(3): 412-422.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8): 904-909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**(6): e1000519.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* **37**(4): 429-434.

Purandare SM, Patel PI. 1997. Recombination hot spots and human disease. *Genome Res* **7**(8): 773-786.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Radman-Livaja M, Liu CL, Friedman N, Schreiber SL, Rando OJ. 2010. Replication and active demethylation represent partially overlapping mechanisms for erasure of H3K4me3 in budding yeast. *PLoS Genet* **6**(2): e1000837.

Raedt TD, Stephens M, Heyns I, Brems H, Thijs D, Messiaen L, Stephens K, Lazaro C, Wimmer K, Kehrer-Sawatzki H et al. 2006. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. *Nat Genet* **38**(12): 1419-1423.

Redfield H. 1966. Delayed mating and the relationship of recombination to maternal age in Drosophila melanogaster. *Genetics* **53**(3): 593-607.

Revenkova E, Eijpe M, Heyting C, Hodges CA, Hunt PA, Liebe B, Scherthan H, Jessberger R. 2004. Cohesin SMC1 beta is required for meiotic chromosome dynamics, sister chromatid cohesion and DNA recombination. *Nat Cell Biol* **6**(6): 555-562.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol* **16**(7): 351-358.

Ringrose L, Paro R. 2004. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* **38**: 413-443.

Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**: 480.

Roberts PA. 1976. The genetics of chromosome aberration. In *The genetics and biology of Drosophila* (ed. M Ashburner, E Novitski, TRF Wright), pp. 67-184. Academic Press, London.

Robinson WP, Kuchinka BD, Bernasconi F, Petersen MB, Schulze A, Brondum-Nielsen K, Christian SL, Ledbetter DH, Schinzel AA, Horsthemke B et al. 1998. Maternal meiosis I non-disjunction of chromosome 15: dependence of the maternal age effect on level of recombination. *Hum Mol Genet* **7**(6): 1011-1019.

Rodrigue A, Coulombe Y, Jacquet K, Gagne JP, Roques C, Gobeil S, Poirier G, Masson JY. 2012. The RAD51 paralogs ensure cellular protection against mitotic defects and aneuploidy. *J Cell Sci*.

Roeder GS. 1997. Meiotic chromosomes: it takes two to tango. *Genes Dev* **11**(20): 2600-2621.

Roeder GS, Bailis JM. 2000. The pachytene checkpoint. *Trends Genet* **16**(9): 395-403.

Roman E, Doyle P, Lightfoot T, Ansell P, Simpson J, Allan JM, Kinsey S, Eden TO. 2006. Molar pregnancy, childhood cancer and genomic imprinting - is there a link? *Hum Fertil (Camb)* **9**(3): 171-174.

Ross JA, Spector LG, Robison LL, Olshan AF. 2005. Epidemiology of leukemia in children with Down syndrome. *Pediatr Blood Cancer* **44**(1): 8-12.

Rotman G, Shiloh Y. 1998. ATM: from gene to function. *Hum Mol Genet* **7**(10): 1555-1563.

Rouyer F, de la Chapelle A, Andersson M, Weissenbach J. 1990. An interspersed repeated sequence specific for human subtelomeric regions. *EMBO J* **9**(2): 505-514.

Roy R, Chun J, Powell SN. 2012. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer* **12**(1): 68-78.

Roy-Gagnon MH, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, Vezina H, Labuda D. 2011. Genomic and genealogical investigation of the French Canadian founder population structure. *Hum Genet* **129**(5): 521-531.

Sandovici I, Kassovska-Bratinova S, Vaughan JE, Stewart R, Leppert M, Sapienza C. 2006. Human imprinted chromosomal regions are historical hot-spots of recombination. *PLoS Genet* **2**(7): e101.

Sasaki M, Lange J, Keeney S. 2010. Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol* **11**(3): 182-195.

Schatz DG, Ji Y. 2011. Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* **11**(4): 251-263.

Schatz DG, Swanson PC. 2011. V(D)J recombination: mechanisms of initiation. *Annu Rev Genet* **45**: 167-202.

Schmiegelow K, Lausten Thomsen U, Baruchel A, Pacheco CE, Pieters R, Pombo-de-Oliveira MS, Andersen EW, Rostgaard K, Hjalgrim H, Pui CH. 2011. High concordance of subtypes of childhood

acute lymphoblastic leukemia within families: lessons from sibships with multiple cases of leukemia. *Leukemia*.

Schuster-Bockler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**(7412): 504-507.

Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, Jacob K, Sturm D, Fontebasso AM, Quang DA, Tonjes M et al. 2012. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**(7384): 226-231.

Schwarz K, Ma Y, Pannicke U, Lieber MR. 2003. Human severe combined immune deficiency and DNA repair. *Bioessays* **25**(11): 1061-1070.

Sciallero S, Giaretti W, Geido E, Bonelli L, Zhankui L, Saccomanno S, Zeraschi E, Pugliese V. 1993. DNA aneuploidy is an independent factor of poor prognosis in pancreatic and peripancreatic cancer. *Int J Pancreatol* **14**(1): 21-28.

Scriver CR. 2001. Human genetics: lessons from Quebec populations. *Annu Rev Genomics Hum Genet* **2**: 69-101.

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**(7104): 772-778.

Segurel L, Leffler EM, Przeworski M. 2011. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biol* **9**(12): e1001211.

Shaw CJ, Lupski JR. 2005. Non-recurrent 17p11.2 deletions are generated by homologous and non-homologous mechanisms. *Hum Genet* **116**(1-2): 1-7.

Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, Humpton TJ, Brito IL, Hiraoka Y, Niwa O, Amon A. 2011. Aneuploidy drives genomic instability in yeast. *Science* **333**(6045): 1026-1030.

Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, Papaemmanuil E, Bartram CR, Stanulla M, Schrappe M et al. 2010. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat Genet* **42**(6): 492-494.

Sherman SL, Petersen MB, Freeman SB, Hersey J, Pettay D, Taft L, Frantzen M, Mikkelsen M, Hassold TJ. 1994. Non-disjunction of chromosome 21 in maternal meiosis I: evidence for a maternal age-dependent mechanism involving reduced recombination. *Hum Mol Genet* **3**(9): 1529-1535.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1): 308-311.

Siepel A, Pollard K, Haussler D. 2006. New methods for detecting lineage-specific selection. *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006: April 2–5, 2006, Venice Lido, Italy)*: pp190–205.

Sigurdsson MI, Smith AV, Bjornsson HT, Jonsson JJ. 2009. HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res* **19**(4): 581-589.

Silva-Grecco RL, Navarro GC, Cruz RM, Balarin MA. 2012. Micronucleated lymphocytes in parents of Down syndrome children. *Braz J Med Biol Res* **45**(7): 573-577.

Smit A, Hubley, R & Green, P. . 1996-2012. RepeatMasker Open-3.0 http://www.repeat.masker.org.

Smith AV, Thomas DJ, Munro HM, Abecasis GR. 2005. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res* **15**(11): 1519-1534.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23**(1): 23-35.

Smith KN, Nicolas A. 1998. Recombination at work for meiosis. *Curr Opin Genet Dev* **8**(2): 200-211.

Song YS, Hein J. 2005. Constructing minimal ancestral recombination graphs. *J Comput Biol* **12**(2): 147-169.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet* **2**(9): e148.

Srebniak M, Wawrzkiewicz A, Wiczkowski A, Kazmierczak W, Olejek A. 2004. Subfertile couple with inv(2),inv(9) and 16qh+. *J Appl Genet* **45**(4): 477-479.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**(4): 393-395.

Stankiewicz P, Shaw CJ, Dapper JD, Wakui K, Shaffer LG, Withers M, Elizondo L, Park SS, Lupski JR. 2003. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am J Hum Genet* **72**(5): 1101-1116.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37**(2): 129-137.

Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M et al. 2012. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet* **44**(8): 872-880.

Steinmann K, Cooper DN, Kluwe L, Chuzhanova NA, Senger C, Serra E, Lazaro C, Gilaberte M, Wimmer K, Mautner VF et al. 2007. Type 2 NF1 deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. *Am J Hum Genet* **81**(6): 1201-1220.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1): 27-40.

Stiller CA, Parkin DM. 1996. Geographic and ethnic variations in the incidence of childhood cancer. *Br Med Bull* **52**(4): 682-703.

Strong ER, Schimenti JC. 2010. Evidence Implicating CCNB1IP1, a RING Domain-Containing Protein Required for Meiotic Crossing Over in Mice, as an E3 SUMO Ligase. *Genes (Basel)* **1**(3): 440-451.

Strout MP, Marcucci G, Bloomfield CD, Caligiuri MA. 1998. The partial tandem duplication of ALL1 (MLL) is consistently generated by Alu-mediated homologous recombination in acute myeloid leukemia. *Proc Natl Acad Sci U S A* **95**(5): 2390-2395.

Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**(2): R13.

Sugawara S, Mikamo K. 1983. Absence of correlation between univalent formation and meiotic nondisjunction in aged female Chinese hamsters. *Cytogenet Cell Genet* **35**(1): 34-40.

Sun F, Oliver-Bonet M, Liehr T, Starke H, Turek P, Ko E, Rademaker A, Martin RH. 2006. Variation in MLH1 distribution in recombination maps for individual chromosomes from human males. *Hum Mol Genet* **15**(15): 2376-2391.

Swanson PC. 2004. The bounty of RAGs: recombination signal complexes and reaction outcomes. *Immunol Rev* **200**: 90-114.

Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW. 1983. The double-strand-break repair model for recombination. *Cell* **33**(1): 25-35.

Tease C, Hulten MA. 2004. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet Genome Res* **107**(3-4): 208-215.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

Thomas JH, Emerson RO, Shendure J. 2009. Extraordinary molecular evolution in the PRDM9 fertility gene. *PLoS One* **4**(12): e8505.

Thomas NS, Ennis S, Sharp AJ, Durkie M, Hassold TJ, Collins AR, Jacobs PA. 2001. Maternal sex chromosome non-disjunction: evidence for X chromosome-specific risk factors. *Hum Mol Genet* **10**(3): 243-250.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.

Tse MT. 2012. Neurodevelopmental disorders: exploring the links between SHANK2 and autism. *Nat Rev Drug Discov* **11**(7): 518.

Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS et al. 2010. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**(6): 504-507.

Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**(1): 90-95.

van der Maarel SM, Deidda G, Lemmers RJ, Bakker E, van der Wielen MJ, Sandkuijl L, Hewitt JE, Padberg GW, Frants RR. 1999. A new dosage test for subtelomeric 4;10 translocations improves conventional diagnosis of facioscapulohumeral muscular dystrophy (FSHD). *J Med Genet* **36**(11): 823-828.

Villa A, Santagata S, Bozzi F, Giliani S, Frattini A, Imberti L, Gatta LB, Ochs HD, Schwarz K, Notarangelo LD et al. 1998. Partial V(D)J recombination activity leads to Omenn syndrome. *Cell* **93**(5): 885-896.

Waanders E, Scheijen B, van der Meer LT, van Reijmersdal SV, van Emst L, Kroeze Y, Sonneveld E, Hoogerbrugge PM, Geurts van Kessel A, van Leeuwen FN et al. 2012. The Origin and Nature of Tightly Clustered BTG1 Deletions in Precursor B-Cell Acute Lymphoblastic Leukemia Support a Model of Multiclonal Evolution. *PLoS Genet* **8**(2): e1002533.

Wakeley J. 2009. *Coalescent theory : an introduction*. Roberts & Co. Publishers, Greenwood Village, Colo.

Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet* **28**(3): 101-109.

Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, Sun YV, Torgerson DG, Rafaels N, Mosley T et al. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet* **43**(9): 847-853.

Weir BS. 1979. Inferences about linkage disequilibrium. *Biometrics* **35**(1): 235-254.

Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, Saha V, Biondi A, Greaves MF. 1999. Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet* **354**(9189): 1499-1503.

Wiemels JL, Leonard BC, Wang Y, Segal MR, Hunger SP, Smith MT, Crouse V, Ma X, Buffler PA, Pine SR. 2002. Site-specific translocation and evidence of postnatal origin of the t(1;19) E2A-PBX1 fusion in childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **99**(23): 15101-15106.

Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity (Edinb)* **98**(4): 189-197.

Winther JF, Sankila R, Boice JD, Tulinius H, Bautz A, Barlow L, Glattre E, Langmark F, Moller TR, Mulvihill JJ et al. 2001. Cancer in siblings of children with cancer in the Nordic countries: a population-based cohort study. *Lancet* **358**(9283): 711-717.

Wirth B, Schmidt T, Hahnen E, Rudnik-Schoneborn S, Krawczak M, Muller-Myhsok B, Schonling J, Zerres K. 1997. De novo rearrangements found in 2% of index patients with spinal muscular

atrophy: mutational mechanisms, parental origin, mutation rate, and implications for genetic counseling. *Am J Hum Genet* **61**(5): 1102-1111.

Wu G, Broniscer A, McEachron TA, Lu C, Paugh BS, Becksfort J, Qu C, Ding L, Huether R, Parker M et al. 2012. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet* **44**(3): 251-253.

Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP. 2010. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **11**: 346.

Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, Yang W, Neale G, Cox NJ, Scheet P et al. 2011. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat Genet* **43**(3): 237-241.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* **46**(4): 409-418.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**(5987): 75-78.

Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, Easton J, Chen X, Wang J, Rusch M et al. 2012. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**(7380): 157-163.

Zhang JG, Goldman JM, Cross NC. 1995. Characterization of genomic BCR-ABL breakpoints in chronic myeloid leukaemia by PCR. *Br J Haematol* **90**(1): 138-146.

Zickler D, Kleckner N. 1999. Meiotic chromosomes: integrating structure and function. *Annu Rev Genet* **33**: 603-754.

Zimmering S, Sandler L, Nicoletti B. 1970. Mechanisms of meiotic drive. *Annu Rev Genet* **4**: 409-436.