Université de Montréal

An approach to improved microbial eukaryotic genome annotation

par Matthew Sarrasin

Département de Biochimie et Médicine Moléculaire, Université de Montréal, Faculté de Médicine

Mémoire présenté en vue de l'obtention du grade de M.Sc.

en Bio-informatique

Décembre, 2017

© Matthew Sarrasin, 2017

Ce mémoire intitulé:

An approach to improved microbial eukaryotic genome annotation

présenté par Matthew Sarrasin

a été évalué par un jury composé des personnes suivantes :

Dre Sylvie Hamel

Présidente-rapporteure

Dr B. Franz Lang

Directeur de recherche

Dre Gertraud Burger

Co-directrice de recherche

Dr Sebastian Pechmann

Membre du jury

Résumé

Les nouvelles technologies de séquençage d'ADN ont accélérées la vitesse à laquelle les données génomiques sont générées. Par contre, une fois séquencées et assemblées, un défi continu est l'annotation structurelle précise de ces nouvelles séquences génomiques. Par le séquençage et l'assemblage du transcriptome (RNA-Seq) du même organisme, la précision de l'annotation génomique peut être améliorée, car les lectures de RNA-Seq et les transcrits assemblés fournissent des informations précises sur la structure des gènes. Plusieurs pipelines bio-informatiques actuelles incorporent des informations provenant du RNA-Seq ainsi que des données de similarité des séquences protéiques, pour automatiser l'annotation structurelle d'un génome de manière que la qualité se rapproche à celle de l'annotation par des experts. Les pipelines suivent généralement un flux de travail similaire. D'abord, les régions répétitives sont identifiées afin d'éviter de fausser les alignements de séquences et les prédictions de gènes. Deuxièmement, une base de données est construite contenant les données expérimentales telles que l'alignement des lectures de séquences, des transcrits et des protéines, ce qui informe les prédictions de gènes basées sur les Modèles de Markov Cachés généralisés. La dernière étape est de consolider les alignements de séquences et les prédictions de gènes dans un consensus de haute qualité. Or, les pipelines existants sont complexes et donc susceptibles aux biais et aux erreurs, ce qui peut empoisonner les prédictions de gènes et la construction de modèles consensus. Nous avons développé une approche améliorée pour l'annotation des génomes eucaryotes microbiens. Notre approche comprend deux aspects principaux. Le premier est axé sur la création d'un ensemble d'évidences extrinsèques le plus complet et diversifié afin de mieux informer les prédictions de gènes. Le deuxième porte sur la construction du consensus du modèle de gènes en utilisant les évidences extrinsèques et les prédictions par MMC, tel que l'influence de leurs biais potentiel soit réduite. La comparaison de notre nouvel outil avec trois pipelines populaires démontre des gains significatifs de sensibilité et de spécificité des modèles de gènes, de transcrits, d'exons et d'introns dans l'annotation structural de génomes d'eucaryotes microbiens.

Mots-clés : génome nucléaire, annotation structurale, eucaryote microbien, protistes, champignons, *Saccharomyces*, *Neurospora*, *Ustilago*, *Plasmodium*

Abstract

New sequencing technologies have considerably accelerated the rate at which genomic data is being generated. One ongoing challenge is the accurate structural annotation of those novel genomes once sequenced and assembled, in particular if the organism does not have close relatives with well-annotated genomes. Whole-transcriptome sequencing (RNA-Seq) and assembly—both of which share similarities to whole-genome sequencing and assembly, respectively—have been shown to dramatically increase the accuracy of gene annotation. Read coverage, inferred splice junctions and assembled transcripts can provide valuable information about gene structure. Several annotation pipelines have been developed to automate structural annotation by incorporating information from RNA-Seq, as well as protein sequence similarity data, with the goal of reaching the accuracy of an expert curator. Annotation pipelines follow a similar workflow. The first step is to identify repetitive regions to prevent misinformed sequence alignments and gene predictions. The next step is to construct a database of evidence from experimental data such as RNA-Seq mapping and assembly, and protein sequence alignments, which are used to inform the generalised Hidden Markov Models of gene prediction software. The final step is to consolidate sequence alignments and gene predictions into a high-confidence consensus set. Thus, automated pipelines are complex, and therefore susceptible to incomplete and erroneous use of information, which can poison gene predictions and consensus model building. Here, we present an improved approach to microbial eukaryotic genome annotation. Its conception was based on identifying and mitigating potential sources of error and bias that are present in available pipelines. Our approach has two main aspects. The first is to create a more complete and diverse set of extrinsic evidence to better inform gene predictions. The second is to use extrinsic evidence in tandem with predictions such that the influence of their respective biases in the consensus gene models is reduced. We benchmarked our new tool against three known pipelines, showing significant gains in gene, transcript, exon and intron sensitivity and specificity in the genome annotation of microbial eukaryotes.

Keywords : nuclear genome, structural annotation, microbial eukaryote, protists, fungi, *Saccharomyces*, *Neurospora*, *Ustilago*, *Plasmodium*

Table of contents

List of figures

List of tables

List of acronyms

A: Augustus

A#: Augustus round number

AC: Augustus with coverage information

AN: Augustus with no hints

AS: Augustus with splice junction and coverage information

BAM: Binary Alignment/Map

bp: Base pair

BUSCO: Benchmarking Universal Single-Copy Orthologs

CDS: Coding DNA Sequence

CM: Covariance Model

EST: Expressed Sequence Tag

DBN: Dynamic Bayes Network

DNA: Deoxyribonucleic Acid

dUTP: Deoxyuridine-triphosphatase

G: Genemark

GA: Genemark-Augustus consensus predictions

GFF: General Feature Format

GFF3: General Feature Format version 3

GSA: Gene-structure-aware

GTF: Gene Transfer Format

HMM: Hidden Markov Model

eSn: Exon Sensitivity

eSp: Exon Specificity

gSn: Gene Sensitivity

gSp: Gene Specificity

iSn: Intron Sensitivity

iSp: Intron Specificity

tSn: Transcript Sensitivity

tSp: Transcript Specificity

MPSA: Multiple Protein Sequence Alignment

ncRNA: non-coding Ribonucleic Acid

NGS: Next-generation Sequencing

NT: Nucleotide

ORF: Open Reading Frame

P: Pooled

PCR: Polymerase Chain Reaction

PPX: Protein Profile Extension

R: Representative

R#: Round number

RD: Relative Distance

RNA: Ribonucleic Acid

RNA-Seq: RNA Sequencing

S#: Snap round number

SAM: Sequence Alignment/Map

Sn: Sensitivity

SNP: Single-Nucleotide Polymorphism

Sp: Specificity

UTR: Untranslated Region

List of abbreviations

Contig: Contiguous Sequence Element

Indel: Insertion/Deletion

k-mer: Subsequence of length *k*

N. crassa: *Neurospora crassa*

P. falciparum: *Plasmodium falciparum*

S. cerevisiae: *Saccharomyces cerevisiae*

U. maydis: *Ustilago maydis*

To curiosity and discovery.

Acknowledgements

The work presented here would not have been possible without the help of at least a dozen people. First, I thank my research director, Franz Lang, for providing me the opportunity to ask my own questions, to be part of something bigger, but also for providing guidance to stay on track. I also thank my research co-director, Gertraud Burger, for invaluable lessons on how to tell a compelling scientific story.

Of course, I thank all my colleagues for their contributions to this project, and for enriching my experience as a graduate student. In no particular order: Lila Salhi, Lise Forget, Simon Laurin-Lemay, Sandrine Moreira, Matus Valach, Rachid Daoud, Mohamed Aoulad-Aissa, Mohammed Hafez, Nicholas Schweiger and Jean-François Théroux.

Thanks to family and friends for supporting me and believing that I would, one day, make it through.

1. Introduction

I. Genomes across the tree of life

The genome is an organism's hereditary basis. It is the genetic material contained in chromosomal DNA, the DNA of mitochondria and various plastids, and plasmid DNA. Viruses, albeit neither free-living nor cells, also contain genomes either in the form of DNA or RNA. As of December 5th, 2017, genomes of over 33,000 species have been deposited at the National Center for Biotechnology Information (NCBI) (ftp.ncbi.nlm.nih.gov/genomes)–not to mention the many more that are about to become publicly available. Of the 33,000 genome sequences, a quarter are from viruses, 60% from bacteria, 4% from archaea, and 8% (i.e., 2,589) from the eukaryotic nucleus. In addition, nearly 10,000 organelle genomes are available, of which 75% are mitochondrial and 25% plastidal. Though NCBI's repository of genomic information represents only a small fraction of the tree of life, the currently available data provide a glimpse into the striking diversity in genome architecture and content.

a. Genome architecture

The spectrum of genome sizes ranges from the 2 kilobase (kb) circovirus genomes, to purportedly hundreds of Gbp of some eukaryotic nuclear genomes (Pellicer, Fay et al. 2010). Genome sizes cluster into roughly three groups (Koonin 2011). Viruses tend to have the smallest genomes on average but can reach sizes greater than those of the second group, bacteria and archaea (prokaryotes) (Philippe, Legendre et al. 2013). In turn, certain bacteria, e.g. the myxobacterium *Sorangium cellulosum,* possess genomes that exceeds 13 Mbp, which is larger than the nuclear genome of *Saccharomyces cerevisiae* for example (Belyi, Levine et al. 2010). Thus, there is considerable overlap between genome-size clusters such that sharp boundaries cannot be drawn between the three groups. However, a distinctive feature of the respective groups is the arrangement of functional and non-functional genome regions (referred to as genome architecture), as specified in the following sections (b-e).

b. Viruses

Viral genomes consist of either DNA or RNA, are linear or circular, single- or double-stranded, and may be made up of multiple discrete fragments (Dimmock, Easton et al. 2016). Among the smallest known virus genomes are those of single-stranded DNA circoviruses, which contain

two protein-coding genes in roughly 2 kb-long molecule (Belyi, Levine et al. 2010). On the other hand, the largest virus genomes are those of double-stranded DNA pandoraviruses, at around 2.5 Mb containing about 2,500 genes (Philippe, Legendre et al. 2013). Viruses with genomes made of RNA are more abundant than DNA viruses. Their genome sizes average around 10 kb and are often multipartite (Dimmock, Easton et al. 2016). Despite a significant diversity in genome layout, a common feature to all known viral genomes is the extremely high gene density such that genes often overlap (Firth and Brown 2006).

c. Prokaryotes

Although bacteria and archaea are fundamentally different from each other, they have genome features in common that distinguish them from eukaryotes. Hence, bacteria and archaea are often referred to as 'prokaryotes'. Prokaryotic genomes are composed of double-stranded DNA, typically in circular conformation, yet cases of linear chromosomes exist. Additional circular and linear, self-replicating, double-stranded DNA molecules, known as a plasmid, are most often encountered in bacteria. These extrachromosomal elements carry genes that can impart a survival advantage under specific conditions, such as synthesis of an antibiotic or protection against an antibiotic. Prokaryotes have generally a high gene density—albeit not as high as viruses—with short intergenic regions and mostly uninterrupted genes ((Rogozin, Makarova et al. 2002); but see (Lambowitz and Belfort 1993) and references therein). That being said, the defining architectural feature of prokaryotes is the operon. Operons are modules of spatially close and functionally-linked genes that are co-regulated and co-transcribed (Jacob and Monod 1961). Such co-localisation usually includes two to four genes (Salgado, Moreno-Hagelsieb et al. 2000), and sometimes dozens more, as for example genes encoding subunits of the ribosomal supercomplex (Wolf, Rogozin et al. 2001). The composition of operons is more likely to be conserved than their synteny (Tatusov, Mushegian et al. 1996).

d. Eukaryotes

The range in eukaryotic nuclear genome sizes spans several orders of magnitude. One of the smallest genomes has been observed in the unicellular parasitic eukaryote *Encephalitozoon* at around 2.4 Mb (Katinka, Duprat et al. 2001), whereas the largest genomes have been documented in plants. For example, the nuclear genome size of the Japanese Canopy plant *Paris japonica* exceeds 120 Gb (Pellicer, Fay et al. 2010), i.e. it is more than 40 times as large as that of humans. The ~2,500 eukaryotic genome sequences currently available at NCBI are mostly of

fungi (~1,800), animals (~400) and plants (~200). The remaining ~200 nuclear genomes come from a taxonomically broad range of less well-known and less well-studied eukaryotic groups such as Archaeplastida, Stramenopiles, Alveolata, Rhizaria, Discoba, Amoebozoa, Hacrobia, Apusozoa and Opisthokonta (Adl, Simpson et al. 2005, Hampl, Hug et al. 2009, Okamoto, Chantangsi et al. 2009). These supergroups are often referred to collectively by the catch-all term 'protists', referring to any eukaryotic organism that neither belongs to plants, animals nor fungi (Adl, Leander et al. 2007).

Certain trends have been observed in nuclear genome organisation, but they are not universal. The majority of nuclear genes are apparently not organised as in prokaryotes. Clusters of co-transcribed genes have been documented in nematodes (Krause and Hirsh 1987), trypanosomes (Sutton and Boothroyd 1986), euglenozoans (Tessier, Keller et al. 1991), and others (Bitar, Boroni et al. 2013), though they do not function like operons. Non-random arrangement of functionally linked genes has been found to be correlated with tandem duplications, like the ANTP-like homeobox genes in animals (Ferrier and Holland 2001). Co-expression has also been correlated with co-localisation of genes. More than 25% of yeast genes co-expressed during the mitotic cell cycle are clustered on the same chromosome (Cho, Campbell et al. 1998). In animals, however, clustering of co-expressed genes only accounts for <5% of all coding genes (Sémon and Duret 2006).

The predominant architectural genome features that set eukaryotes apart from prokaryotes is the composition, abundance and location of non-coding DNA. One particular type almost uniquely found in eukaryotes are the non-coding segments (introns) that interrupt coding segments (exons). Other forms of non-coding DNA tend to be more abundant in eukaryotes compared to prokaryotes, including repetitive regions (simple and tandem, transposable elements, duplications, pseudogenes), non-coding genes (discussed later), and other intergenic DNA. The general trend is that genomes distended with numerous long introns and other types of non-coding DNA tend to be observed in higher eukaryotes such as humans (Mattick 2001).

The number, length and composition of introns varies substantially between taxa, and even within the same genus. Some nuclear genomes can contain relatively few (<200 total) introns that are often short (<100 bp median length) as seen, for example, in *S. cerevisiae*

(Spingola, Grate et al. 1999), the diplomonad *G. lamblia* (Nixon, Wang et al. 2002) and kinetoplastids in general (e.g. *T. brucei*) (Muhich and Boothroyd 1988). On the opposite end of the spectrum, protein-coding genes (averaging 2 kb in length, and some exceeding 10 kb) in vertebrates have between five to eight introns (Gibbs, Weinstock et al. 2004). Despite the variability in number and length, intron positions in coding genes tend to be conserved even across large evolutionary distances. For instance, as many as 30% of introns are inserted at the same positions in orthologous genes from vertebrates and plants (Fedorov, Merican et al. 2002).

The total amount of non-coding DNA, and the proportion of repetitive regions, non-coding genes or other types of intergenic DNA, also vary considerably. For example, around 52% of the *E. cuniculi* genome (Katinka, Duprat et al. 2001), and around 41% of the *Trypanosoma cruzi* genome (El-Sayed, Myler et al. 2005), is non-coding DNA. Higher eukaryotes are on the other end of the non-coding DNA abundance spectrum, where as much as 99% of the genome is non-coding DNA in humans (Lander, Linton et al. 2001) and plants (Bennetzen and Kellogg 1997). The nuclear genome of *E. cuniculi* contains as few as 2,000 densely packed genes, with short intergenic regions (~130bp) and with most of the non-coding DNA (~53%) found in telomeric and sub-telomeric regions (Katinka, Duprat et al. 2001). In contrast, the mean length of intergenic regions in *T. cruzi* is ~1,000, and a substantial amount of the repetitive regions are composed of pseudogenes (El-Sayed, Myler et al. 2005). Transposable elements make up 45% of the human genome, and even more than 50% of some plant genomes like maize ((Bennetzen and Kellogg 1997) and references therein).

e. Mitochondria and plastids

In addition to the nuclear genome, eukaryotes can also house organellar genomes. It is now widely accepted that mitochondria and plastids were once free-living bacteria (alpha-proteobacterial and cyanobacterial origins, respectively) that formed an endosymbiotic relationship with ancestral eukaryotes. The transition of the endosymbiont to an integral part of the eukaryotic cell has had a profound effect on the architecture of both the host's and the symbiont's genome—the latter having undergone massive genome reduction and rearrangement, even between closely related species ((Gray, Burger et al. 1999, Green 2011), and references therein). Genome conformations can be linear or circular, either in a single contiguous chromosome or spread between dozens of fragments (Gray, Burger et al. 1999, Stoebe and Kowallik 1999).

Mitochondria DNAs (mtDNAs) range from 5 kbp (Hikosaka, Watanabe et al. 2009) to 100 kbp (Burger, Gray et al. 2013) in size, with introns and intergenic regions representing anywhere from 1% to 99% of the genome ((Smith and Keeling 2015) and references therein). The highly reduced apicomplexan mitochondria (~5 kbp genomes) contain as few as three genes (*cox1, cox3, cob*) (Feagin, Mericle et al. 1997), whereas the mitochondria of *Andalucia,* the most alpha-proteobacterial-like, contains 100 genes. Of those 100 genes, at least two have been found to encode proteins (large subunit mitoribosomal protein, and a protein related to cytochrome oxidase assembly) that have, so far, only been detected in the nuclear genome of other eukaryotes (Burger, Gray et al. 2013). Gene fragmentation, manifesting in lineage-specific ways, is a peculiar feature documented in the mitochondria of green algae (Boer and Gray 1988), euglenozoans (Lukeš, Guilbride et al. 2002, Marande, Lukeš et al. 2005) and alveolates (Waller and Jackson 2009). Essentially, genes are present in multiple discrete pieces distributed across different DNA molecules. In *Chlamydomonas* mitochondria, for example, eight discrete fragments that encode the large subunit ribosomal RNA are present in different DNA molecules, and are post-transcriptionally spliced together (Boer and Gray 1988).

In contrast to mtDNAs, plastid DNAs (ptDNAs) can be as large as 1 Mbp in certain plants (Sloan, Alverson et al. 2012). The non-coding DNA content of ptDNAs ranges from 5-80% (Smith, Lee et al. 2010). On average, ptDNAs contain more genes than mitochondria (Barbrook, Howe et al. 2010), ranging from as few as ~20 genes in some dinoflagellates (Barbrook, Voolstra et al. 2014), up to ~250 in red algae (Janouškovec, Liu et al. 2013). Plastids can have a substantially different gene complement compared to mitochondria; the most apparent being photosynthesis-related genes in chloroplasts. Other types of plastid genomes can contain genes related to pigment synthesis and storage (chromoplasts), or monoterpene synthesis (leucoplasts) ((Wise 2007) and references therein).

II. From genome to transcriptome

The transcriptome is defined as the full range of molecules transcribed from a subset of the genetic repertoire under specific conditions in either a single cell or a (possibly heterogeneous) population of cells. Changes in conditions can rapidly induce specific readjustments in the transcriptome. Thus, the flow of information is more complex than the rather simple view of 'from genome to transcriptome'.

a. Transcriptome composition

The different types of RNA present in the transcriptome include messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), a host of mostly small non-coding RNAs (small interfering RNA (siRNA), micro RNA (miRNA), piwi-interacting RNA (piRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA)), long non-coding RNAs (lncRNA), and likely a range of yet-to-be characterized RNA. The respective length and abundance varies significantly between RNA classes. Small RNAs such as 5S rRNA, siRNA, miRNA, piRNA, snoRNA and tRNA, are shorter than 200 nucleotides (nt), whereas long RNAs such as mRNA, small subunit (SSU) and large subunit (LSU) rRNA and lncRNA can be as long as 17 kb (Brown, Hendrich et al. 1992). The abundance of the different RNA molecules varies by several orders of magnitude (Lodish, Berk et al. 2000). Only a small fraction of the RNA population is mRNA (~1%), while rRNA (~80%), and tRNA (15%) represent the majority, with the remainder being the various other ncRNA (Lodish, Berk et al. 2000).

b. Biological functions of transcripts

Each class of RNA molecules is linked to a specific biological function (Lodish, Berk et al. 2000). Protein synthesis involves the mRNA template to be translated, where tRNAs carry amino acids for the ribosome (composed of rRNAs and ribosomal proteins) to link together (Lodish, Berk et al. 2000). Ribosomal RNAs, in turn, are structural and enzymatically active components of ribosomes (ribozymes). Regulation in the cell can be mediated by siRNA which blocks gene expression (Hamilton and Baulcombe 1999), miRNA which blocks or accelerates degradation of mRNA (Lee, Feinbaum et al. 1993), or piRNA, which have been posited to be involved in retrotransposon silencing (Aravin, Gaidatzis et al. 2006, Girard, Sachidanandam et al. 2006).

c. Transcript processing

RNA molecules can also be modified post-transcriptionally through the processes of editing, or splicing in *cis* or in *trans*. The RNA editing machinery can substitute a nucleotide (e.g. A-to-I or C-to-U) for another, at a specific position (Benne, Van Den Burg et al. 1986). Another, but different, form of RNA processing commonly observed in eukaryotes is splicing (Lodish, Berk et al. 2000). Six snRNAs and hundreds of proteins form the small nuclear ribonucleoprotein (snRNP) complex known as the spliceosome that coordinates removal of introns from pre-mRNA (and pre-rRNA), and to join respective coding exons into a contiguous transcript. While

some interrupted protein-coding genes give rise to a single mRNA, some genes can give rise to multiple different mRNAs through the process of alternative splicing. Multiple different mRNA may be the result of exon skipping, intron retention, alternative splicing sites, alternative promoters, or a combination of any of the above (Lodish, Berk et al. 2000). The sequence of an mRNA can thus be different from an alternative mRNA originating from the same gene, which can confer an alternative function, or can change its localisation. While the most commonly observed form of splicing takes place on the same RNA molecule (in *cis*), *trans* splicing takes place between two discrete molecules.

III. Next-generation sequencing

DNA sequencing technologies have seen rapid advances along with falling costs since their inception 40 years ago—especially true for the last 15 years with the advent of next generation sequencing (NGS) (Mardis 2008, Shendure and Ji 2008, Mardis 2017). NGS offers several fundamental upgrades over previous sequencing technologies. The first major difference is in the way the library is prepared for NGS; input DNA (after fragmentation and adapter ligation) is amplified by polymerase chain reaction (PCR). Second, NGS technologies couple the sequencing and conversion of position-defined fragments to digital molecular information (Mardis 2017). The coupling of those two steps is colloquially described as "massively parallel". The advent of NGS has opened the door to genome-wide analyses (Mardis 2017) of methylation patterns (Cokus, Feng et al. 2008), transcription factor binding sites (Sanger, Air et al. 1977, Mikkelsen, Ku et al. 2007, Cokus, Feng et al. 2008, Mardis 2017) and variants (Korbel, Urban et al. 2007). The same technology has equally been applied to deep sequencing whole transcriptomes (Mortazavi, Williams et al. 2008, Wang, Gerstein et al. 2009).

IV. Genome assembly

The genome must be reassembled from the millions of reads that are generated from widely used platforms such as Illumina (<=300 bp reads) or PacBio (<=40 kb reads). Genome assembly is a complicated procedure that is constantly evolving in response to new technologies, particularly as novel chemistries are used to increase the length of reads (Baker 2012). Effectively, the bottleneck has shifted from biochemistry to bioinformatics (Yandell and Ence 2012).

The first genomes to be assembled (*Haemophilus influenzae* (Fleischmann, Adams et al. 1995), baker's yeast (Goffeau, Barrell et al. 1996) and human (Lander, Linton et al. 2001)) were done using the overlap layout consensus (OLC) method. Briefly, it can be likened to solving a jigsaw puzzle by exploiting overlap between pieces (Pop 2009). OLC assemblers (reviewed in (Miller, Koren et al. 2010)) first determine read overlaps in an all-against-all, pairwise alignment. The alignment algorithm is a heuristic search of read subsequences of length *k* (*k*-mer), otherwise known as a seed search. Next, a graph is constructed from the reads that fully or partially overlap (the jigsaw pieces that fit together) that approximates the read layout (Myers 1995). Finally, a multiple sequence alignment of the reads is computed to determine the precise layout and to resolve the consensus. An alternative was proposed by Pevzner *et al.* in 2001 (Pevzner, Tang et al. 2001) that can simplify assembly (particularly in repeat regions) to some extent using a De Bruijn graph. It starts by breaking reads into *k*-meres and building a graph of overlapping *k*-mers (by exactly *k*-1 nucleotides) and traversing the graph to reconstruct contigs, similarly to the OLC method. In both approaches, parameters are inferred from sequencing depth to help resolve repetitive regions that typically have much higher coverage than the global average (Treangen and Salzberg 2011).

a. Difficulties in genome assembly

The quality of a genome assembly is often difficult to measure given the lack of a gold standard to compare with (Salzberg, Phillippy et al. 2012). Even in long-standing assemblies such as that of the mouse nuclear genome, a number of segments are still unresolved. As much as 140 Mb (mostly from duplications) have recently been reintegrated in the mouse assembly (Church, Goodstadt et al. 2009). Issues with long repetitive regions and scaffold arrangement remain at the forefront of genome assembly, especially for libraries of shorter (20-400 bp) reads (Henson, Tischler et al. 2012, Schlebusch and Illing 2012). Difficulties can be exacerbated for certain protist genomes given that bacterial contamination—the source of food for some (Haas and Webb 1979)—is nearly impossible to avoid. No studies, to our knowledge, have yet been conducted to evaluate the extent of contamination in known protist genomes, nor to evaluate methods for removing bacterial contamination.

V. Whole transcriptome sequencing

RNA-Seq is a revolutionary procedure that, in a nutshell, provides a snapshot of all the genes expressed at a given time and condition (Mortazavi, Williams et al. 2008). It has a number of advantages over previous technologies like serial analysis of gene expression (Velculescu, Zhang et al. 1995) and tiling microarrays (Stoughton 2005). For instance, molecular information can be directly obtained from very lowly to highly expressed genes (i.e. higher dynamic range) without the need for specific restriction sites or *a priori* knowledge of transcripts (Parkhomchuk, Borodina et al. 2009). Thus, the full spectrum of RNAs and their absolute expression level—from mRNAs to the various non-coding RNAs (ncRNAs)—are readily available at the molecular level (Margulies, Egholm et al. 2005, Mortazavi, Williams et al. 2008, Parkhomchuk, Borodina et al. 2009). RNA-Seq has seen widespread adoption for characterising novel transcripts and quantifying expression in a population of cells (Cloonan, Forrest et al. 2008, Nagalakshmi, Wang et al. 2008, Wang, Baskerville et al. 2008, Marguerat and Bähler 2010) or, more recently, at the level of a single cell (Tang, Barbacioru et al. 2009, Islam, Kjallquist et al. 2011). Of course, depending on the type of study it is desirable to target a specific group of RNA. For example, the presence of highly expressed ribosomal RNAs could hinder signal from mRNA. Protocols have been developed to target poly-A enriched transcripts (since mRNAs typically carry an A-tail; (Cloonan, Forrest et al. 2008, Mortazavi, Williams et al. 2008), depletion kits for removing rRNA (He, Wurtzel et al. 2010), or size selection for enrichment of

small ncRNAs (reviewed in (Jacquier 2009)). Among the various experimental protocols, three are used most commonly (Figure 1).

All methods start with total RNA extraction and selection of the target class of RNA.

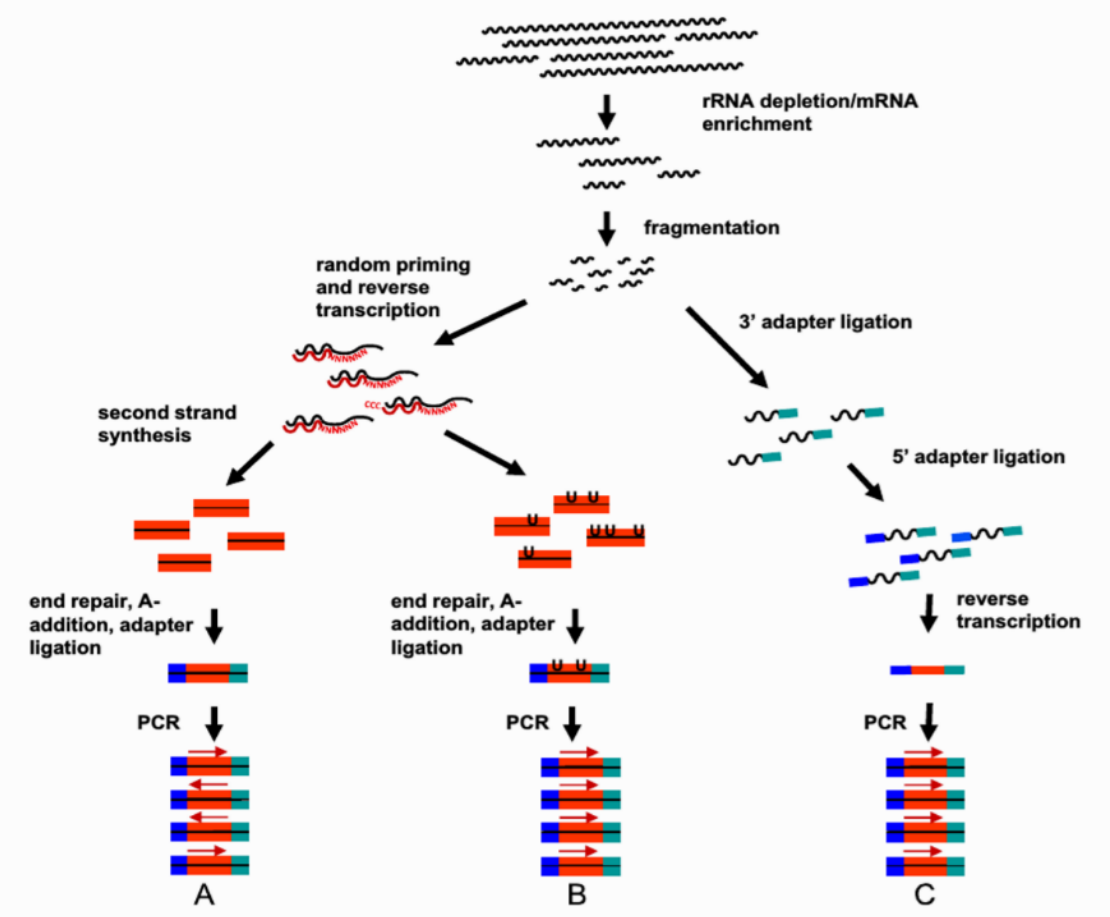

Figure 1: Overview of three commonly used protocols to prepare an RNA-Seq library. RNAs are selected and fragmented, then either synthesized to cDNAs by random priming and reverse transcription (either untagged (A) or dUTP tagged (B)), or (C) adapters are sequentially ligated followed by reverse transcription. In all cases, the final step is fragment amplification. Adapted from van Dijk et al (van Dijk, Jaszczyszyn et al. 2014).

Branches A and B of the flowchart in Figure 1 follow similar steps to generate fragments suited for sequencing. Double-stranded cDNAs are first synthesized by reverse transcription initiated by random priming, followed by technology-specific adapter ligation for PCR amplification (van Dijk, Jaszczyszyn et al. 2014). The difference is that dUTP tagging in method B prevents sequencing of the second strand, and thus preserving the orientation of the RNA template (Parkhomchuk, Borodina et al. 2009). Preserving strand information (strand-specificity) is useful in teasing apart overlapping genes encoded on opposite strands as well as sense transcripts and (regulatory) antisense transcripts of a given gene (Normark, Bergstrom et al. 1983, Guida,

Lindstädt et al. 2011, Wang, Jiang et al. 2014). Branch C of the flowchart describes the sequential ligation approach. It is typically employed in small RNA analyses (van Dijk, Jaszczyszyn et al. 2014).

VI. Whole-transcriptome assembly

In transcriptome assembly, similarly to genome assembly, either an OLC graph or De Bruijn graph can be applied (Florea and Salzberg 2013), but with certain adapted parametrisation. One major difference to genome assembly is that read count per gene is proportional to gene expression (Mortazavi, Williams et al. 2008). Therefore, genome assembly parametrisation is not applied to transcriptome assembly since it could lead to a loss of sensitivity in resolving transcripts whose coverage differs significantly from the average (Martin and Wang 2011). Secondly, graph parsing differs in that all transcript variants (e.g. arising from processing or alternative splicing) must be enumerated (Garber, Grabherr et al. 2011). A final consideration is whether the genome sequence is available or not; transcriptome assembly can be guided by

RNA-Seq reads mapped to the genome sequence, otherwise it is done *de novo* (Figure 2; (Martin and Wang 2011).

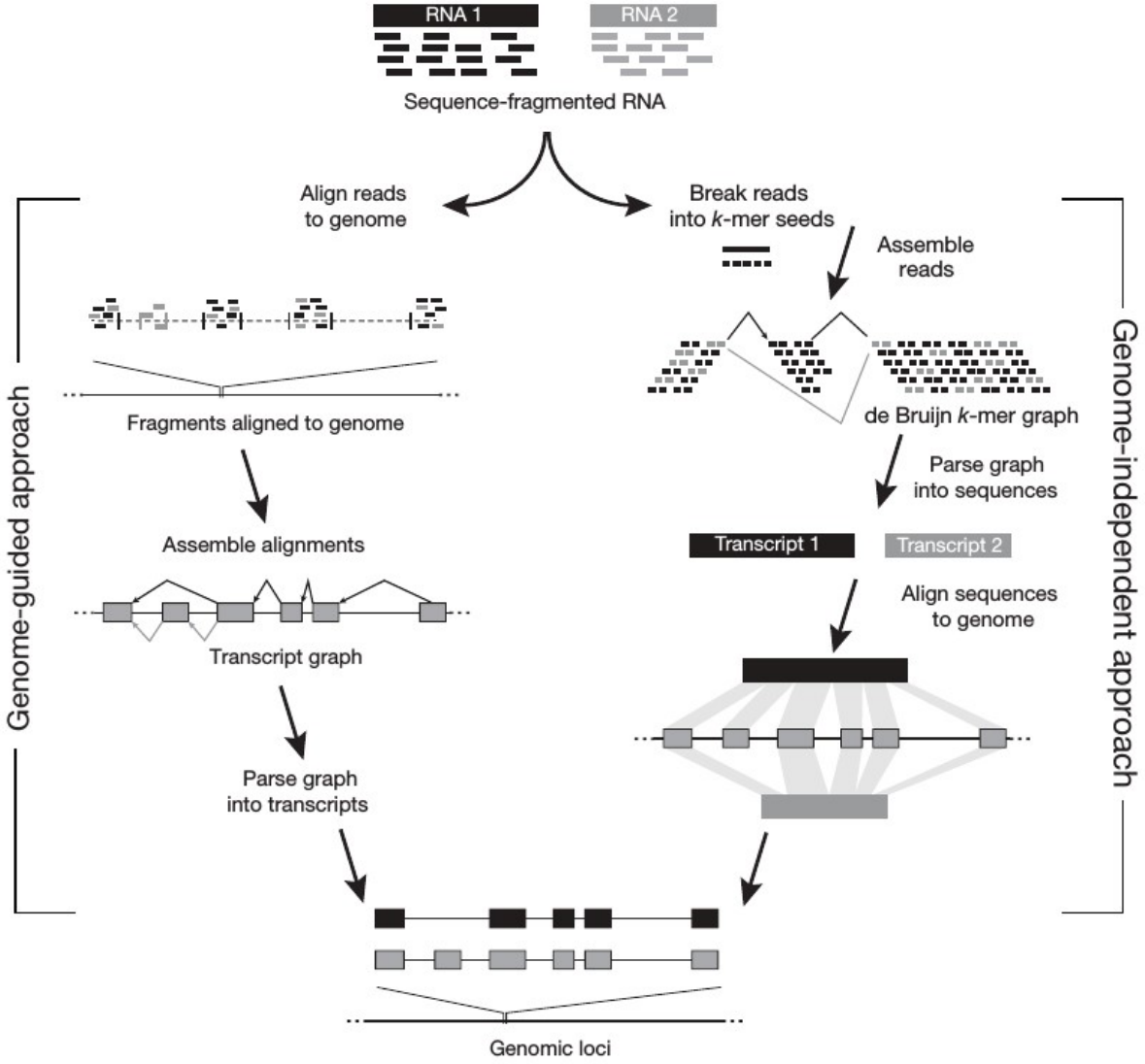

Figure 2: Overview of the genome-guided (left) and genome-independent (right) approaches to transcriptome assembly. Reads are mapped in the genome-guided approach. Reads are then clustered per genomic loci to infer feature boundaries, which are used to construct transcript graphs. Graphs are parsed to infer possible unique paths representing putative isoforms. On the other hand, the de novo approach builds De Bruijn k-mer graphs and traverses them to infer putative transcripts. Adapted from (Garber, Grabherr et al. 2011).

There are advantages and disadvantages to both methods. The genome-guided (reference-based) approach requires, evidently, a high quality genome assembly. By first mapping the RNA-Seq reads to the genome, it is possible to filter out lower-quality or erroneous reads, which can reduce the complexity of assembly, and also reduce the risk of artefactual fusions (Trapnell, Williams et al. 2010). The procedure involves clustering of mapped reads at their respective genomic loci to be incorporated into a transcript graph, which is parsed to reconstruct a maximal

set of transcripts that best explain each cluster of reads (Guttman, Garber et al. 2010). The approach can have a higher sensitivity in resolving transcripts since an assembler can fill in gaps using the genomic sequence where low coverage could make reconstructing a whole transcript difficult (Guttman, Garber et al. 2010, Trapnell, Williams et al. 2010). The obvious limitation is the requirement of a genome sequence on which to map reads. The second potential drawback is that the quality of the transcriptome assembly is dependent on the quality of the genome assembly. Insertions, deletions (indels), and un-joined contigs in the genome (Salzberg and Yorke 2005), as well as extensive contaminating contigs from other organisms in the genome assembly can reduce the effectiveness of RNA-Seq read mapping, thus reducing the completeness and accuracy of the transcript assembly (Martin and Wang 2011).

In the reference-free, or *de novo* approach (Figure 3), a *k*-mer library is generated from the reads to build the transcript graph, similar to the De Bruijn-graph based genome assembly

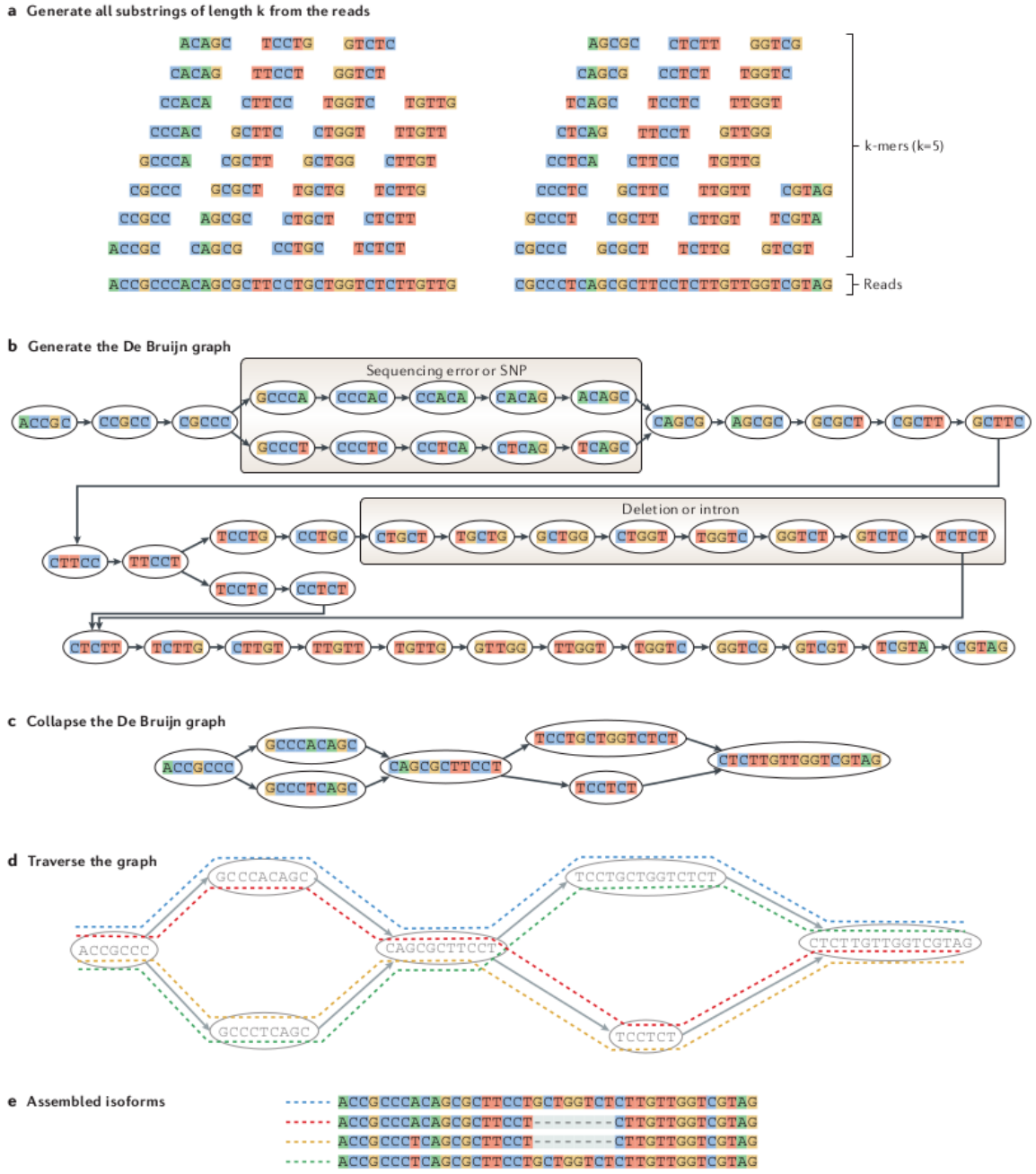

Figure 3: Breakdown of the De Bruijn graph approach to de novo transcriptome assembly. a) *k*-mers are enumerated from sequence reads; b) De Bruijn graph is constructed from k-mers; c) paths collapsed into plausible variants; d) the graph is traversed to enumerate all variants; e) transcripts are assembled from all plausible variants. Adapted from (Martin and Wang 2011).

(Grabherr, Haas et al. 2011). A unique *k*-mer, irrespective of coverage, represents a single node. The node is connected to another node if it overlaps by exactly *k*-1 nucleotides (Pevzner, Tang et al. 2001). Cases where *k*-mers overlap by *k*-2 nucleotides could indicate a single nucleotide

polymorphism (SNP), sequencing error, an exon-exon (or exon-intron) junction from a variant, at which point a new branch (of length *k*) is created on the graph (Pevzner, Tang et al. 2001, Zerbino and Birney 2008, Martin and Wang 2011). The process is repeated on every branch, creating downstream branches if there is a *k*-2 overlap, until all *k*-mers have been accounted for. The graph is either simplified and parsed (Trapnell, Williams et al. 2010), similarly to the genome-guided approach, or the graph is directly traversed to reconstruct transcripts (Grabherr, Haas et al. 2011). An advantage is that the assembler draws from the entire pool of reads (since none have been prefiltered by mapping), which allows discovery of novel transcripts (Mortazavi, Williams et al. 2008). Uneven or missing coverage can complicate the assembly process by leading to multiple contigs although their reads were originally derived from a single transcript (Martin, Bruno et al. 2010). Increasing sequencing depth may help resolve the issue but at the same time, very high coverage (>>1,000) tends to increase the number of false positive transcripts due to base-calling errors or chimeric reads (Tarazona, Garcia-Alcalde et al. 2011).

As discussed above, both the reference-based and *de novo* methods each have complementary strengths and weaknesses, and are each situation-dependent. Nonetheless, if a genome assembly is available, it is advised to adopt both approaches and combine their results into a non-redundant, comprehensive assembly (Haas and Zody 2010).

VII. Annotation

The ultimate goal of genome and transcriptome assembly is to establish the genetic repertoire of an organism, through a process known as genome annotation. Genome annotation can be broken down into two major steps: 1) structural annotation, followed by 2) functional annotation (Yandell and Ence 2012). Structural annotation involves modeling various genomic features and their precise locations, namely genes, regulatory and non-coding regions. The next step is to assign to each model a biological function based on sequence similarity to known genes (Yandell and Ence 2012). Structural annotation typically consists of a suite of steps glued together into a "pipeline". Several different pipelines exist, yet they all follow a similar workflow, which is described in the following sections.

a. Repeat masking

Generally, the first step aims at identifying repetitive regions, both tandem and dispersed repeats (Bao and Eddy 2002), and transposable elements such as long and short interspersed elements

(Bedell, Korf et al. 2000, Kapitonov and Jurka 2008). This is an important initial step since repeat regions left unmasked can misdirect Blast-like sequence alignment heuristics by anchoring seed searches in regions that yield sub-optimal alignments (Korf, Yandell et al. 2003). Unmasked repeats can also misinform gene prediction algorithms about exon-intron structure (Korf, Yandell et al. 2003, Yandell and Ence 2012).

b. Sequence alignment to generate evidence of coding gene structure

After repeat masking, it is common practice to construct a set of extrinsic evidence of gene locations and structure. Extrinsic evidence usually consists of organism-specific information (transcript assemblies) and of information from close neighbours (protein sequences), both of which can be aligned to the target organism's genome sequence (Haas 2003, Haas, Zeng et al. 2011, Yandell and Ence 2012). A widely used heuristic tool for sequence alignment is Blast (Altschul, Gish et al. 1990), which identifies potential coding regions by exact matches of small subsequences (seed searches) and expands on matched regions using a local alignment algorithm. One considerable disadvantage in the context of structural annotation is that Blast is not capable of precisely modelling exon-intron junctions (Korf, Yandell et al. 2003). An alternative is to use spliced-sequence alignment tools, such as Exonerate or Gmap, that apply similar seed search approach as Blast but employ a computationally demanding dynamic programming algorithm around alignment gaps to identify splice junctions (Slater and Birney 2005, Wu and Watanabe 2005). Some pipelines leverage both methods: Blast is used to quickly identify putative coding regions and then intervals are "polished" with a spliced alignment tool (e.g. exonerate (Slater and Birney 2005)) to more accurately resolve start and stop codons, and splice junctions (Cantarel, Korf et al. 2008).

c. Gene prediction

Gene prediction algorithms are based on Hidden Markov Models (HMMs), but they have been generalised to include statistical properties of the genome such as G+C content, codon usage preferences and intron structure (Korf 2004). Extrinsic information extracted from RNA-Seq and sequence alignments are being used to inform both the HMMs and their generalised statistical models (Stanke, Diekhans et al. 2008). HMMs are an entirely probabilistic framework to model the most likely state (e.g. constituting an exon, intron, or intergenic region) of a given interval of a primary sequence (Rabiner 1989). The state is the "hidden" part in the Hidden Markov Model, since we do not know the underlying state *a priori*. The HMM consists of a transition probability (from one state to another) and an emission probability that models the composition of a state (Durbin 1998). In the context of gene prediction, a simplified HMM (Figure 4) can consist of three states: exons, splice junctions and introns, each with their own probabilities of remaining in the same state or transitioning to another (Eddy 2004). Each state

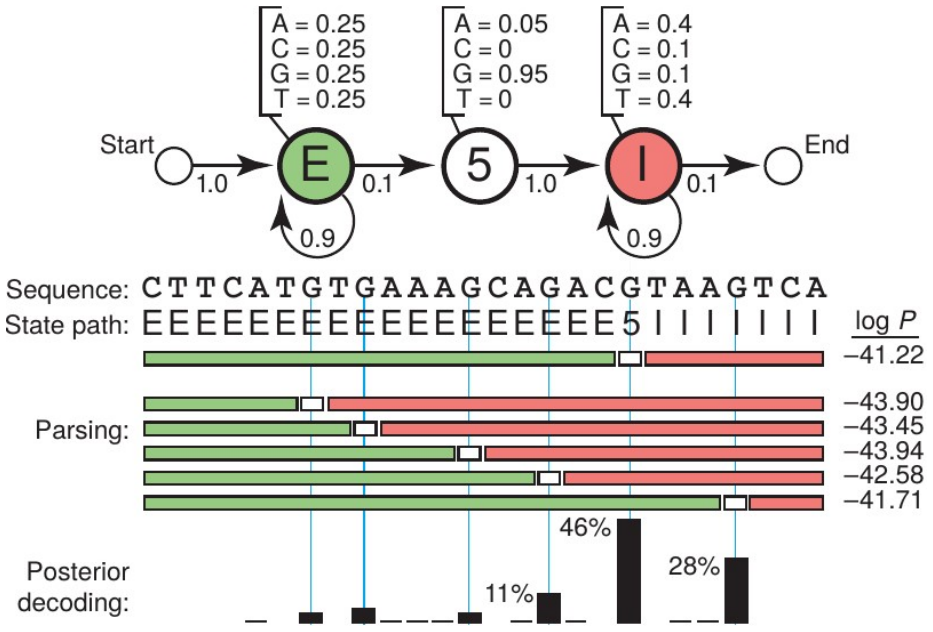

Figure 4: Cartoon of a HMM with three states with their respective transition and emission probabilities. For example, a simple HMM could model exons, splice junctions and introns. Adapted from (Eddy 2004).

contains its own emission probabilities that could model, for example, a high chance of observing a G at an exon-intron junction, a higher chance of observing As and Ts than Gs and Cs within an intronic sequence, and a similar chance of observing any of the four nucleotides within an exon. The algorithm walks along the sequence, inferring the underlying states at each position given the previously observed sequence (a Markov chain). Many different state paths

potentially exist in such a probabilistic framework. For instance, referring to Figure 4, it is assumed that the first state is an exon, and the task is to determine at which locations of the sequence switches the state. The splice junction is assumed to have 95% chance of being a G, so the most likely locations of that state can be reasonably reduced to six out of twenty nucleotide positions. The algorithm explores each of those state paths in a process known as posterior decoding (Durbin 1998). The Viterbi algorithm, commonly employed in gene predictors (Lukashin and Borodovsky 1998, Stanke and Waack 2003), chooses the state with the highest log-likelihood by dynamic programming (Durbin 1998), which turns out to be position 19 in Figure 4. The inclusion of splice junctions inferred from RNA-Seq, for example, extrinsically supports a particular state path, thus increasing the accuracy of a prediction. Coverage of RNA-Seq reads can also be included to inform transitions from intergenic to initial exon states, while aligned protein sequences could support a transition to an open reading-frame state.

d. Gene model consensus

With a wealth of information from sequence alignments and HMM-based predictions about functional regions, the final step is to choose a representative model (Yandell and Ence 2012). Transcript alignments (either organism specific or from a close relative) can provide information about untranslated regions (UTRs) and coding exons, but the open reading frames (ORFs) need to be inferred (Adamidi, Wang et al. 2011). Protein sequence alignments (typically from other organisms) can fill in the gap to resolving ORFs. However, protein sequences may sometimes be insufficient. Gene predictors can give a rough look at a genome's coding content, but are strongly dependent on a high-confidence and representative training set (Burset and Guigo 1996). Therefore, pipelines have been developed to automate the process of synthesizing gene-model creation from multiple different sources to reduce the amount of manual effort required (Cantarel, Korf et al. 2008, Reid, Nicholas et al. 2014, Hoff, Lange et al. 2016, Min, Grigoriev et al. 2017). They 'combine' evidence to create a consensus using either a supervised algorithm guided by a training set (Allen and Salzberg 2005), an *un*supervised algorithm using a Dynamic Bayes Network (DBN) (Liu, Mackey et al. 2008) or an algorithm based on user-defined weights per evidence source (Haas, Salzberg et al. 2008). Essentially, these pipelines each attempt to choose a model that best represents the evidence while also reducing the different types of

possible errors, such as in-frame stop codons, incorrect splice junctions, or frame shifts (Yandell and Ence 2012).

The weighted consensus process followed by EVidenceModeler (Haas, Salzberg et al. 2008) will serve as an example (Figure 5).

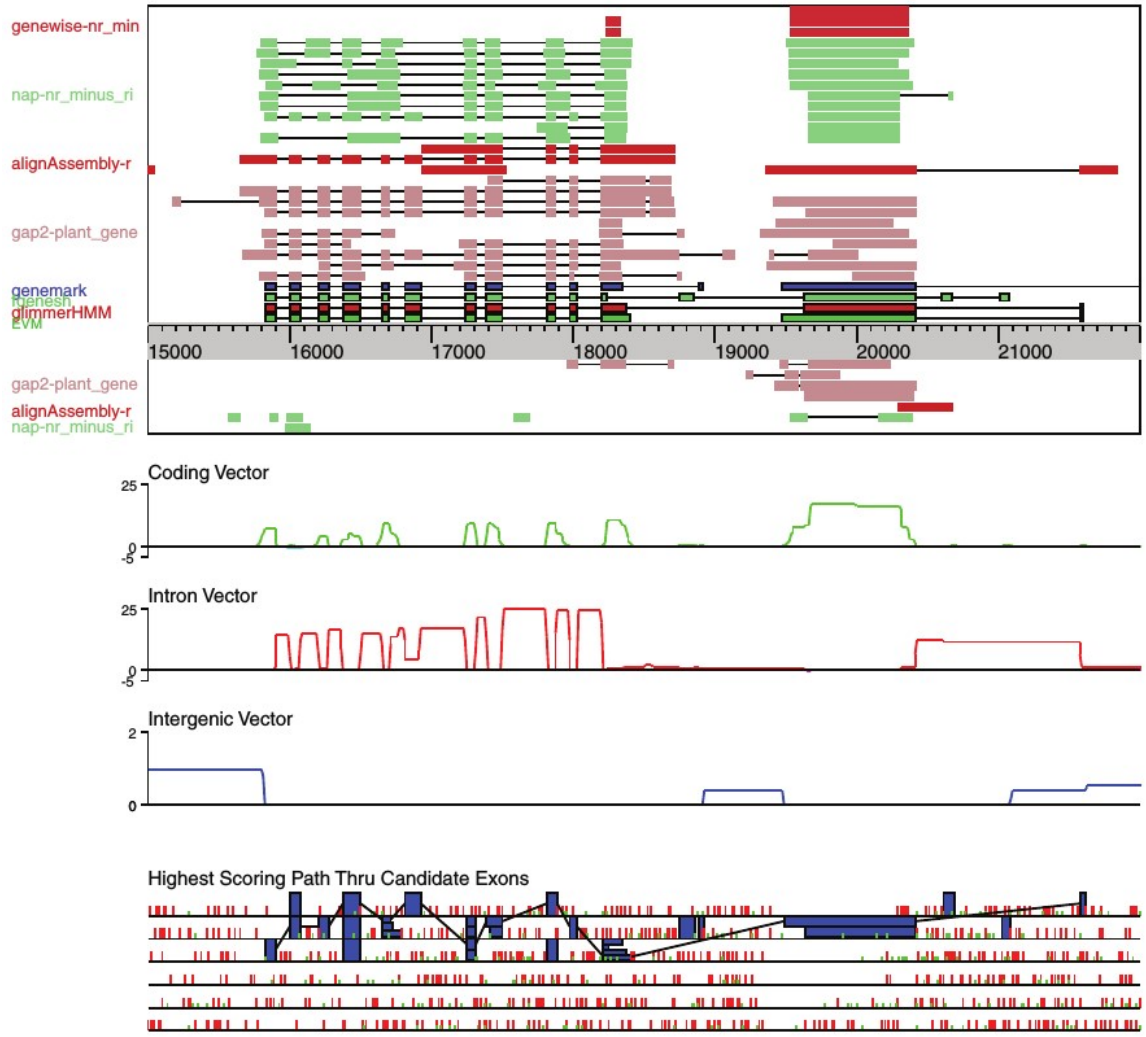

Figure 5: Overview of the weighted consensus algorithm implemented in EVM. The top window represents transcript alignments (Nap-nr_minus_ri, AlignAssembly-r), protein alignments (Gap2-plant_gene) and gene predictions (Genewise-nr_min, Genemark, Fgenesh, GlimmerHMM) used to build consensus gene models (EVM). The Coding, Intron and Intergenic vectors (middle window) are computed to evaluate the highest scoring path through candidate exons (bottom window). See text for a more detailed description. Adapted from (Haas, Salzberg et al. 2008).

The top box is a snapshot of a 7-kb window in the rice genome with putative protein-coding regions, and several different sources of evidence to suggest its structure. Tracks labelled 'Fgenesh', 'Genemark' and 'GlimmerHMM' contain intervals inferred by the corresponding gene predictors. The Genewise and 'Nap2' tracks come from rice and non-rice protein sequence

alignments, respectively, whereas the 'AlignAssembly' and 'Nap' tracks are derived from rice and rice neighbour EST alignments, respectively. The orientation of the intervals are separated by the genomic loci axis. Position-specific score vectors are calculated for genomic features, such as coding, intron and intergenic regions, based on the evidence intervals as seen in the middle graphs of Figure 5. All possible exonic and intronic regions in the six reading frames are computed from the feature vectors, as enumerated in the bottom part of Figure 5, where green and red ticks represent start and stop codons, respectively. The vertices connecting candidate exons trace the highest scoring path computed by a dynamic programming, yielding two distinct regions in the 'EVM' track. The two regions correspond to known genes encoding the peroxisomal membrane carrier protein and 50S ribosomal protein L4, chloroplast precursor, respectively (Haas, Salzberg et al. 2008).

In contrast to the manual assignment of weights as mentioned above, the supervised training approach taken by Jigsaw infers weights from a subset of curated models, and employs a dynamic programming algorithm to resolve the highest scoring path (Allen and Salzberg 2005). Conversely, instead of relying on a training set, the DBN algorithm (employed in the Evigan pipeline) learns the parameters that best explain the evidence and then generates a consensus by computing the maximum likelihood using the Viterbi algorithm (Liu, Mackey et al. 2008), similarly to posterior decoding computed by the generalised HMMs of gene predictors. Despite the differences between the three combiners, the nematode genome assessment project (nGASP) found their performance to be similar (Coghlan, Fiedler et al. 2008). Nevertheless, the quality of consensus models may differ significantly between organisms, depending on a plethora of factors such as the available experimental data, evidence quality, evolutionary distance, among others (Yandell and Ence 2012).

e. Functional annotation

Once a complete set of organism-specific gene models have been built, the interest shifts to inferring functional information for each model. Various levels of functional information can be assigned, including secretion signals, domain content and product name, as well as broader biological implications such as pathways and processes (Frishman 2007). Pairwise similarity searches and HMM-based searches are common methods employed to elucidate function (Haas,

Zeng et al. 2011), among other specialised methods for specific tasks (e.g. localisation prediction) (Horton, Park et al. 2006).

A basic routine is to run Blast to align a query database to the target gene models whereby information associated with sequences in the query database are transferred to their respective "best hit" (i.e. best score) target sequences. Swissprot is a widely used query database, which contains many sequences with known product names, gene ontology information (molecular function, cellular localisation, biological process (Ashburner, Ball et al. 2000)), and other types of information (Boeckmann, Bairoch et al. 2003). The main advantage with Blast is that it allows for rapid transfer of information from those databases to the target sequences. In contrast to similarity algorithms, profile HMMs are an inherently probabilistic framework to detect sequence relationships with higher sensitivity. Profile HMMs are position-specific statistical models to describe the consensus of a multiple sequence alignment (i.e. the profile). The model contains information, at every given positon, to compute the true frequency of a residue relative to the observed frequency (Durbin 1998). Commonly used profile HMM databases include Pfam to search for protein domains (Bateman, Birney et al. 2002), and TIGRfam to search for protein families (Haft, Selengut et al. 2003).

VIII. Pitfalls of automated annotation

Two important challenges in every automated (structural and functional) annotation pipeline are, depending on the information they accept, (i) the introduction of systematic error due to incomplete or misleading data, (ii) errors perpetuated from incorrect comparative data and (iii) the presence of inherent biases.

a. Pitfalls in automated structural annotation

Firstly, errors can be introduced from incomplete forms of evidence. For instance, lacking RNA-Seq coverage can reduce the number of detected introns and underrepresent expressed regions as well as limit the number of successfully assembled full-length transcripts (Sims, Sudbery et al. 2014). Such errors can increase the false negative rate of gene prediction algorithms and consensus model-building tools (Yandell and Ence 2012).

Systematic biases can be introduced when inferring putative coding regions from protein sequence alignments and from gene prediction training. Since protein sequence alignment

algorithms are based on similarity searches, evolutionary distance can have varying negative effects on the diversity and completeness of information that can be leveraged. The further the distance the less useful protein sequence alignments become, which can be problematic for organisms with few characterised neighbours (Slater and Birney 2005), as is the case for most protists.

Gene prediction algorithms are also dependent on the diversity and completeness of information. A preliminary set of gene models are used to parameterise and tune generalised HMMs. Thus, a lack of preliminary models that adequately and accurately represent the full gene repertoire can bias predictions towards a particular gene structure. In turn, biased training models decrease the likelihood of detecting coding intervals (Burset and Guigo 1996).

b. Pitfalls in automated functional annotation

Systematic errors can be subtly introduced in functional annotation procedures that make use of Blast. The reason is that Blast ultimately computes a similarity score that is difficult to tie with biological relevance, especially for distantly-related sequences. As such, judging an adequate cutoff score to transfer functional information is more or less arbitrary and error prone (Galperin and Koonin 1998).

A continual challenge for databases such as Swissprot is curating data to maintain a high level of accuracy. The increasing amount of genomic data becoming available is increasingly being annotated automatically. Those models might then be used to infer functions in future projects, which could perpetuate erroneous information (Bork and Bairoch 1996). Conversely, inherent biases introduced in the structural annotation itself can have a similar effect of missing relevant information.

IX. Goals

Automated structural annotation pipelines have seen widespread adoption, yet they still cannot attain the level of accuracy of a dedicated human curator. Thus, the first goal of this project was to assess the state of current annotation pipelines by benchmarking them with “gold standard” structural genome annotations (i.e. those with considerable manual curation). The second goal was to propose a method of structural annotation to improve gene model confidence/quality, thereby reducing expert intervention.

X. Objectives

This Master research project had two objectives. The first was to identify the plausible sources of error and bias in three freely distributed, and commonly used annotation pipelines. The second objective was to develop an approach—borrowing from the same (and partially modernised) tool set of current pipelines—to specifically mitigate, if not prevent, the propagation of biases and errors. The work performed is described in the following manuscript that will be submitted for publication to Genome Biology.

2. Manuscript:

An approach to improved microbial eukaryotic genome annotation

Matthew Sarrasin, Gertraud Burger and B. Franz Lang

Target journal: Genome Biology

Contribution of authors

BFL and GB contributed to conceptual design of the approach, orchestrating the project, and writing the manuscript. MS contributed to conceptual design, implementation and validation of the study, and manuscript writing.

Keywords

Annotation, genome, eukaryote, protist, fungi, *Saccharomyces*, *Neurospora*, *Ustilago*, *Plasmodium*

I. Abstract

Challenges in automating structural annotation of eukaryotic genomes are ever-present, particularly for eukaryote nuclear genomes without a well-annotated, closely neighbouring species. The main shortcoming of freely distributed pipelines for structural annotation come from (1) errors in incomplete input data, (2) biases in gene prediction training, and (3) inaccurate gene model consensus construction procedures. Here, we propose an improved approach that mitigates the impact of those three shortcomings. Our approach capitalizes on two main aspects; first, it leverages a more complete and diverse set of extrinsic evidence, derived from RNA-Seq and homology data, to better inform gene predictions. Second, gene models are constructed from the extrinsic evidence and gene predictions using a weighted consensus approach such that the impact of potential errors and biases is reduced. Comparative benchmarking against three widely-used pipelines shows that our approach has higher sensitivity and specificity in detecting genes, transcripts, exons and introns.

II. Background

Recent development in high-throughput whole-genome and whole-transcriptome sequencing (RNA-Seq) technology is both a boon and a burden to novel eukaryotic genome projects (Shendure and Ji 2008, Ozsolak, Platt et al. 2009, Wang, Gerstein et al. 2009, Marguerat and Bähler 2010). Although the cost of sequencing has substantially decreased, the process of finding genes and determining their exon-intron structure in a genome assembly (structural genome annotation) continues to be a challenge (Yandell and Ence 2012). Structural genome annotation is a complex multi-step process, referred to as a pipeline, to build gene models by leveraging experimental data (RNA-Seq, homology data) and predictive modeling algorithms (generalised Hidden Markov Model (HMM) gene predictions) (Yandell and Ence 2012). Several different automated pipelines have been developed to facilitate structural genome annotation by modelling the decision-making process an expert curator would follow to consolidate multiple sources of information. Among the most commonly used are Snowyowl (Reid, Nicholas et al. 2014), Maker (Cantarel, Korf et al. 2008), and Braker (Hoff, Lange et al. 2016). Snowyowl, which was specifically developed for fungal genome annotation, builds consensus gene models from concordance between gene predictions, RNA-Seq coverage and protein sequence similarity (Reid, Nicholas et al. 2014). Likewise, Maker, a generic eukaryotic annotation pipeline, extracts information about gene structure from transcript and protein sequence alignments to inform predictions and consensus gene models (Cantarel, Korf et al. 2008). In contrast, Braker was developed to only use splice junctions inferred from RNA-Seq read mapping. Though the above pipelines greatly facilitate the annotation process, some important considerations remain that can affect the quality of gene models.

One consideration is that lacking or incomplete experimental evidence can hinder gene predictions (Mathé, Sagot et al. 2002). Further, such systematic errors propagate and bias the construction of consensus gene models informed by those gene predictions and experimental evidence. Genome annotation becomes increasingly difficult for eukaryotes where relatively few of its neighbours have been characterised, and even more so for isolated, divergent taxa, as the usefulness of inferences based on sequence similarity quickly degrades (Korf 2004). Such issues can be exacerbated in genome assemblies inconspicuously contaminated with DNA from other species, e.g. bacteria being the food source of numerous unicellular eukaryotes. It becomes difficult to discern between a truly eukaryotic gene with no introns, a laterally transferred or

mitochondrial gene, or a contaminant—particularly in intron-poor eukaryotes. Leveraging RNA-Seq data can alleviate the difficulty, to an extent, by providing organism-specific information about genes (expressed under a given condition) and their structure (Denoeud, Aury et al. 2008, Ozsolak, Platt et al. 2009). Even then, building complete and reliable gene models for such organisms remains an important bottleneck.

Protist genomes continue to be largely uncharacterised due to the above-mentioned issues. In addition, protists are biologically more diverse than, for instance, animals or fungi. Therefore, it is not surprising that the nuclear genomes of certain protist groups is even more gene-dense than yeast (Goffeau, Barrell et al. 1996, Galagan, Calvo et al. 2003), while others are as inflated as in plants (Hou and Lin 2009). In line with this variability, some protists have a genome that is intron-poor with a total count below 100 (e.g., *Giardia* (Morrison, McArthur et al. 2007), *Trypanosoma* (Hall 2003)), whereas others possess close to four per gene on average, as seen in the ciliate *Tetrahymena* (Coyne, Hannick et al. 2011). Further, some protists add a canonical 5'UTR to pre-mRNAs by trans-spicing, such as trypanosomes (Sutton and Boothroyd 1986), euglenids (Tessier, Keller et al. 1991), rotifers (Pouchkina-Stantcheva and Tunnacliffe 2005) and dinoflagellates (Lidie and Van Dolah 2007)—a phenomenon also observed in cnidarians (Stover and Steele 2001), certain chordates (Vandenberghe, Meedel et al. 2001), nematodes (Krause and Hirsh 1987), and Platyhelminthes ((Rajkovic, Davis et al. 1990)(reviewed in (Bitar, Boroni et al. 2013)). Such features can severely hinder the genome annotation process where less abundant similarity-based data exists, and where RNA data is not fully exploited.

Given the above considerations, current generic pipelines may not equally generate high-confidence gene models for all eukaryotic genomes (or all fungal genomes, in the case of Snowyowl). Such factors may exacerbate the shortcomings related to gene model creation, such as incorrect start or stop sites, splitting of models into multiple discrete pieces, artificial fusions between close neighbours (Mathé, Sagot et al. 2002), or simply missing models where one should exist (Burset and Guigo 1996). Thus, in this study, the three pipelines mentioned above were examined regarding their performance in creating gene models for four model genomes: *Neurospora crassa*, *Saccharomyces cerevisiae*, *Ustilago maydis* and *Plasmodium falciparum*. On this basis, we formulated a new annotation approach that specifically aimed at mitigating, if not rectifying, the shortcomings of those pipelines.

III. Results & Discussion

i. Defining the major steps of each pipeline

Each pipeline leverages different forms of experimental evidence, tools and methods in a suite of steps to generate a consensus gene model set. Each step that outputs gene models is considered a major step.

Maker employs Blast and Exonerate (Slater and Birney 2005) for transcript (from an RNA-Seq reads assembly) and protein sequence alignments (supplied by the user), and has the capacity to internally run the output from the semi-HMM-based gene predictor Snap (Korf 2004), the generalised HMM-based Augustus (Stanke, Diekhans et al. 2008), and the self-training predictor Genemark (Ter-Hovhannisyan, Lomsadze et al. 2008). Maker employs a novel algorithm to generate gene models depending on the step and the input data. First, Maker employs Blast (Altschul, Madden et al. 1997) to rapidly align transcript sequences and protein sequences to the genome. Sequences with hits are passed to Exonerate to better resolve their exon-intron structure. Maker extracts the exon-intron boundaries from those alignments (extrinsic evidence) for the user to train (outside of the pipeline) the gene predictors Snap (step S1) and Augustus (step A1). The output from those gene predictors, along with Genemark output (step G; done outside of the pipeline), are passed to Maker to run its consensus building algorithm. Those consolidated models serve as a basis to retrain (bootstrap training) Snap (step S2) and Augustus (step A2). Their output is given to Maker so that a final consensus set can be built (step F). The final models are based on concordance between transcript alignments, protein alignments and gene predictions.

Snowyowl, on the other hand, differs from Maker in several ways. In the first step, Genemark is run (step G) to generate preliminary gene models. Next, Augustus is trained with the transcriptome assembly as input to generate an independent set of initial gene models (step A). Models common between those of Genemark and Augustus are selected to retrain Augustus (step GA). The next phase, termed by the authors as the 'proliferation stage', which runs three independent rounds of Augustus with relaxed parameters to infer multiple plausible models at a given genomic region. The first round of Augustus predictions are made without any hints (step AN) after retraining in step GA. The second round leverages the locations of RNA-Seq reads mapped to the genome (step AC) using Blat (Kent 2002), whereas the last round leverages

both mapped reads (mapped using Blat) and splice junctions inferred from the mapping (step AS). Finally, common gene models between the three rounds of prediction in the proliferation stage are pooled together (stage P). The last two steps involve a model-scoring algorithm (step R), followed by model selection based on a defined threshold (step F). Each model is scored based on similarity search results against the user-defined sequence database, concordance with splice junctions, and level of RNA-Seq coverage (Reid, Nicholas et al. 2014). The selection algorithm finds models that have similarity hits above a defined score, completely agree with inferred splice junctions and have coverage above a defined amount, and labels them as high-confidence models. Models that meet all the same criteria but fall below the defined RNA-Seq coverage threshold are also accepted, but flagged as 'lowly expressed'. All others are labelled as 'imperfect' and not accepted in the final consensus set.

In contrast to both Maker and Snowyowl, Braker is considerably less complex. Braker accepts as input the mapped RNA-Seq reads. Splice junctions are extracted from the read mapping and serve as the only source of extrinsic evidence. In the first step, splice junction locations are passed to Genemark to help generate preliminary models (step G). Of the preliminary models, only those whose predicted introns are concordant with the inferred RNA-Seq splice junctions are selected to train Augustus. After training, predictions from Augustus are generated with the help of the same inferred splice junction information used in Genemark (step F). Those models constitute the consensus set.

ii. Sources of error and bias

a. In RNA sequencing and transcriptome assembly

Several considerations pertain to the use of RNA-Seq data that can affect the output of annotation pipelines that make use of it. Biases in library preparation and sequencing can cause uneven coverage (e.g. lower coverage over GC-rich regions) that makes it difficult to adequately resolve expressed regions ((Sims, Sudbery et al. 2014) and references therein). Increasing the sequencing depth can help resolve the issue to an extent, as well as increase the statistical power to differentiate between signal and noise (Haas and Zody 2010), but some genes are inherently lowly expressed that remain challenging to obtain sufficient coverage (Mortazavi, Williams et al. 2008). On the other hand, a higher number of errors may occur in highly expressed regions, owing to the fact that more reads are generated and that sequencing technologies can introduce

a number of errors in base-calling (~1-3% in Illumina sequencing (Nakamura, Oshima et al. 2011), as much as 17% in PacBio sequencing (Chin, Sorenson et al. 2011)).

The above issues can have a negative impact on 1) tools that exploit RNA-Seq read mapping, and 2) transcriptome assembly routines. Firstly, low or uneven coverage in a read mapping file may convince gene predictors (Augustus and Genemark) of an incorrect start/stop site, or the lack of an intron where one exists, or even that a single gene model should erroneously be split into multiple (Mathé, Sagot et al. 2002). Secondly, low or uneven coverage and errors in reads can negatively impact transcriptome assembly. Low and uneven coverage could lead to reconstruction of multiple discrete transcripts in place of a single, contiguous transcript, whereas errors in base-calling could lead to multiple spurious transcripts ((Martin and Wang 2011) and references therein). In turn, expressed genes that are closely located in the genome run the risk of having their transcripts artificially fused by the assembly algorithm (Grabherr, Haas et al. 2011), particularly in cases where genes are co-transcribed (Hoffmann, Otto et al. 2014).

b. In protein similarity searches

The use of homology-based data comes with a few important drawbacks, namely, 1) the quality of protein sequences is not uniform in a given database and, 2) the level of sequence divergence strongly determines the quality of an alignment.

Publicly available protein sequence databases, such as Swissprot (Boeckmann, Bairoch et al. 2003), contain a wealth of both predicted and expertly-curated information for use in genome annotation. Assuring the quality of those protein sequences is, however, an ongoing challenge. Protein sequences that may be used to detect coding regions in an uncharacterized genome may be incomplete or contain defects, depending on the level of expert curation. Recently, as much as 60% of the cytochrome P450 genes in 47 different plant genomes were found to have some defect (Gotoh, Morita et al. 2014), which suggests that leveraging protein sequence information, particularly where few well-characterised neighbours exist, can potentially introduce biases.

Secondly, current protein-to-genome sequence alignment algorithms (as implemented in, e.g., Exonerate, Genewise (Birney, Clamp et al. 2004)) are based on similarity, which can be derailed by homologous proteins under different selective pressure such that their sequences sufficiently diverge (reviewed in (Batzoglou 2005)). Thus, an important caveat is that increasing

sequence divergence can yield partial models, or none at all (Korf, Yandell et al. 2003). In other cases, the algorithm can potentially introduce false introns to optimise the alignment (Wu and Watanabe 2005, Gotoh 2008).

c. In gene prediction

Ongoing challenges related to gene prediction include 1) creating adequate and accurate training sets (Burset and Guigo 1996, Guigo, Agarwal et al. 2000), and 2) inherent limitations in predictive algorithms (Mathé, Sagot et al. 2002, Ter-Hovhannisyan, Lomsadze et al. 2008). The quality of gene predictions is dependent on the quality and diversity of organism-specific gene models used to parameterize (train) HMM-based predictors, since universal parameters cannot be applied to all eukaryotes (Burset and Guigo 1996). Thus, parameters inferred from incorrect gene models reduces the predictive power, which can lead to a higher rate of incorrect or false models (Mathé, Sagot et al. 2002). On the other hand, a small number of highly accurate training models can fail to capture the full gene structure diversity, which can lead to fewer predicted models (Burset and Guigo 1996).

Aside from training sets, gene prediction algorithms have inherent limitations that can introduce errors. For instance, an ongoing challenge is the capacity to model very long introns (>10kb as in humans (Gibbs, Weinstock et al. 2004)) and very short exons (<25bp, see (Volfovsky, Haas et al. 2003)). A related pitfall to predicting short features such as exons is the potential failure to recognize short intergenic regions between two closely-neighbouring (or overlapping, as in, e.g., *Arabidopsis* (Quesada, Ponce et al. 1999)) genes on the same strand (Mathé, Sagot et al. 2002). Such small intergenic regions may be misinterpreted as an intron, which could lead to artificially fused models (Pavy, Rombauts et al. 1999). Conversely, genes can be falsely predicted inside long intergenic regions (Ter-Hovhannisyan, Lomsadze et al. 2008).

iii. Benchmarking the major steps to identify errors and biases

The gene models output at the major steps (defined above) of Maker, Braker and Snowyowl (fungi only), as run on *N. crassa, U. maydis* and *P. falciparum,* were benchmarked against the gene models of the corresponding reference models. This was done to determine the steps at which models may have been affected by errors or biases for each pipeline. The benchmarking metrics employed for comparison were sensitivity and specificity. Sensitivity measures the

amount a given feature (e.g. transcript, exon) in the reference models overlaps with the same feature in the predicted models, whereas specificity is the amount a given feature in the predicted models overlaps with the same feature in the reference models (Burset and Guigo 1996)). Sensitivity and specificity were computed (in %) at the gene, transcript, exon and intron levels using the Eval package (Keibler and Brent 2003) (Figure 6).

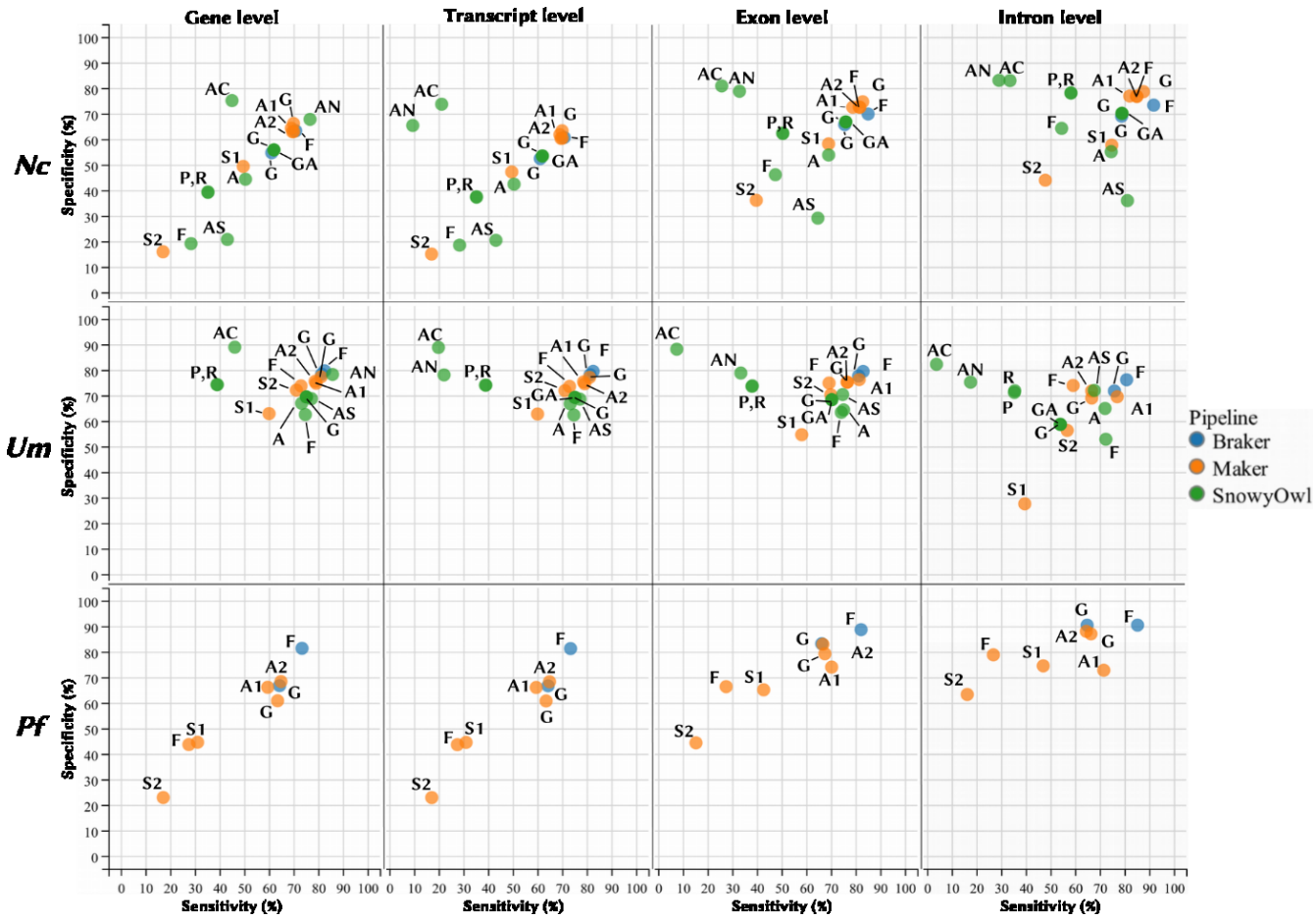

Figure 6: Specificity vs. sensitivity plots of the major steps of each pipeline at the level of genes, transcripts, exons and introns for *N. crassa (Nc), U. maydis (Um) and P. falciparum (Pf)*. The major steps of Braker, shown in blue, include Genemark (G) and final models (F) for Maker are shown in orange. In green, the Snowyowl steps include (in order) initial Genemark predictions (G), the first round of Augustus (A), consensus predictions from Genemark and Augustus (GA), Augustus predictions with no hints (AN), coverage only (AC), coverage with splice junctions (AS), pooled models (P), representative models (R) and final models (F).

a. Maker

The extrinsic evidence extracted from protein sequence alignments are not full gene models, however it is an important step done to train gene predictors. It is of interest to gauge how

representative the extrinsic evidence is of the reference models. Thus, the relative distance (RD) metric was used to measure how much the protein sequence alignments correspond with the reference gene (denoted as Maker-Ref), and *vice versa* (denoted as Ref-Maker, Table 1). Lower RD corresponds to a shorter distance between the alignments and the gene models (or *vice versa*), thus implying more congruence.

Species	Mapping	Count	Total	Fraction (Count/Total, %)
N crassa	Ref-Maker	16044	28587	55.9
N crassa	Maker-Ref	96575	131110	73.6
U maydis	Ref-Maker	6241	9637	64.9
U maydis	Maker-Ref	28230	40077	70.6
P falciparum	Ref-Maker	2288	13067	17.6
P falciparum	Maker-Ref	46896	60210	77.9

Table 1: Summary of the number (count) of protein sequence alignment falling within a relative distance of 0-0.10 (out of 0.50) out of the total number of alignments. Relative distance was computed as a function of the reference models, relative to the protein alignments, and conversely as a function of protein alignments relative to the reference models.

Protein sequence alignments tend to correspond with the reference gene models, and *vice versa*, for *N. crassa* and *U. maydis*. For *P. falciparum,* the percentage of alignment RDs with respect to the reference models was close to that seen in the fungi. However, there was a larger discrepancy in the RD between the reference models and the alignments. The low fraction of congruent gene models with respect to the alignments in conjunction with the high fraction of alignments congruent with the models suggests that most protein sequences aligned to a smaller number of genomic loci. This interpretation is corroborated by the fact that a benchmarking of single-copy orthologous genes (BUSCO, (Simão, Waterhouse et al. 2015)) on the protein alignments reflect the same trend (Supplemental Table 7). Fewer complete putative orthologous sequences were identified in the Maker alignments, and 80 out of the 94 hits were duplicates. It is likely that the proteomes given as input to Maker to annotate *P. falciparum* were more divergent compared to the proteomes used as input to annotate the two fungi.

Transcriptome assemblies were also computed for *N. crassa, S. cerevisiae, U. maydis* and *P. falciparum*. Thus, we sought to measure potential errors in those assemblies. A lower sequencing depth would yield fewer full-length mRNAs. The number of full-length mRNAs was measured by: 1) running Blast was run on the reconstructed transcripts, for each species, against the Swiss-prot database (v2017-07-15), and 2) by running BUSCO on each of the transcriptome assemblies.

The best hit was chosen for each transcript (if present), and its alignment length (with respect to the best hit in the Swiss-prot sequences) was computed as a percentage of the total transcript length (Figure 7). The results indicate that there are significantly fewer full-length or near-full-length hits in *P. falciparum* relative to the fungi.

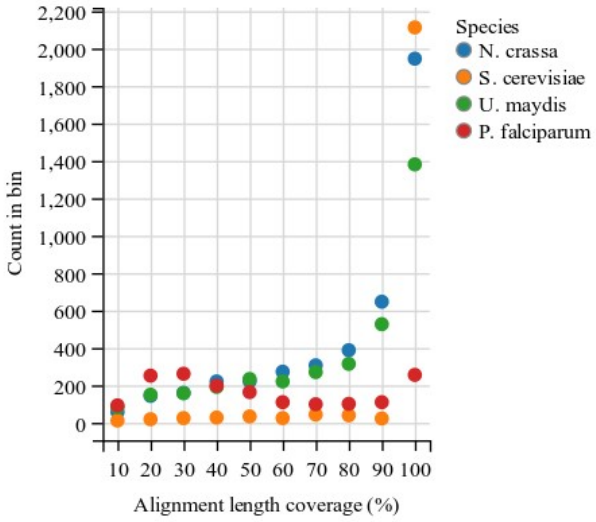

Figure 7: The number of Blast hits as a function of alignment length coverage of protein sequences against the transcriptome for the four tested species. On average, the fungi have a higher number of Blast hits, whereas *P. falciparum* has significantly fewer full-length hits in comparison.

The BUSCO results indicate a similar trend to the Blast results (Table 2). Over 200 orthologous genes are missing in the *P. falciparum* transcriptome assembly, whereas less than 10 are missing in the fungal transcriptomes.

BUSCO predictions	*N. crassa*	*S. cerevisiae*	*U. maydis*	*P. falciparum*
Complete	281	278	286	94
Complete and single-copy	6	8	24	14
Complete and duplicated	275	270	262	80
Fragmented	1	3	3	7
Missing	8	9	1	202
Total	290	290	290	303

Table 2: Breakdown of the BUSCO results. BUSCO's fungal and protozoa databases were searched against the fungal and protist transcriptomes, respectively. More than 200 orthologous genes are missing in the *P. falciparum* transcriptome assembly, whereas relatively few are missing in the fungal transcriptomes.

Despite the less complete protein sequence alignments and transcriptome assembly in *P. falciparum,* the gene models output by Augustus and Genemark are, on average, a few sensitivity and specificity percentage points lower than those of Braker. Snap, however, tended to underperform compared to Genemark and Augustus, despite being trained on the same data as Augustus. Nonetheless, a similar trend has been observed by the authors of Maker (Cantarel,

Korf et al. 2008). On the other hand, the bootstrap retraining method imparts an overall gain in sensitivity and specificity of the final models, at each level—particularly in the case of *N. crassa*. The final *P. falciparum* models, however, are significantly less sensitive and specific relative to the Augustus and Genemark models. Thus, the sensitivity and specificity of Maker's final consensus model set can vary significantly depending on the completeness of transcript and protein sequence alignments, despite potentially high-confidence gene predictions.

Here, we proposed a three-fold solution to improve gene model accuracy over Maker, particularly in the case of *P. falciparum*. First, the spliced aligner Spaln could replace Exonerate, given that Spaln has been shown to produce higher quality alignments compared to several competing aligners (Gotoh 2008). It does so in part by modelling organism-specific intron length distributions to better capture natural splice junctions (Iwata and Gotoh 2012). Second, a bootstrap retraining method, as implemented in Maker, need not be implemented. The sensitivity and specificity of models after retraining were either about the same for Augustus, or lower as for Snap. Lastly, the substantially lower sensitivity and specificity in the final models is likely due to the dependence Maker's consensus-building algorithm has on transcript and protein sequence alignments. Thus, we propose implementing an alternative that can distribute the weight of extrinsic evidence and gene predictions (as in the tool Evidencemodeler (Haas, Salzberg et al. 2008)) such that any type of evidence has less impact on the consensus.

b. Snowyowl

The gene predictions by Augustus after the second round of training show an increase in sensitivity and specificity across all levels, but it is less clear that there is an advantage to the proliferation and scoring phases. The significantly lower specificity relative to the sensitivity, as seen in the pooled *U. maydis* and unhinted *N. crassa* predictions, is expected given that Snowyowl deliberately over-predicts at those stages to evaluate multiple models per locus posteriorly. In some cases there is an increase in sensitivity and specificity, as seen in the splice-hinted models of *U. maydis*. However, in every case the final models suffer either a marginal or substantial degradation in sensitivity and specificity compared to the initial predictions. This suggests that there may yet be a benefit to exploring the multiple possible models at a given locus, but it would require a different scoring scheme to determine the most biologically relevant model.

The implication of Snowyowl's scoring scheme being primarily based on similarity searches is that plausible models can be falsely discarded. An additional consequence of the scoring scheme is that gene models with an uneven distribution of RNA-Seq coverage—despite having an average coverage (over the length of the model) above the defined threshold—are equally discarded. Such a cut-off may discard valid models, particularly since it is known that coverage across a transcript can vary, as described above. Despite the reported benefit in exploring alternatives and choosing models that best represent the evidence (Reid, Nicholas et al. 2014), the results shown in Figure 6 suggest that the second round of Augustus models could potentially yield a higher sensitivity and specificity consensus than Snowyowl's final models.

The solution we proposed here is centered on generating a more complete training set from a consensus between evidence types (transcriptome assembly alignments, protein sequence alignments and a high-confidence set of gene models), and to avoid the complexities of generating and selecting models as done in in the proliferation and scoring stages of Snowyowl.

c. Braker

The implication of building a training set from concordance between initial gene predictions and inferred splice junctions is that models with no introns—in other words, no evidence support—are equally selected for the training set (since they are technically consistent with splice junctions). Interestingly, the exon sensitivity and exon specificity shown in Figure 6 suggest that the impact of this bias may not be detrimental to the consensus models. Nonetheless, we proposed to include information from transcript and protein sequence alignments, RNA-Seq read coverage and splice junction locations, since it would permit the selection of initial models—whose features are all fully congruent with at least two sources of evidence—to construct a more robust training set. More specifically, initial predictions from Genemark that match either a transcript or a protein sequence alignment are selected for the training set. In addtion, transcript alignments that match protein alignments are also selected. This procedure could resolve the problem of validating intronless models while enforcing a wider set of criteria for training model selection. Secondly, implementing a weighted consensus between multiple different sources of evidence could mitigate the biases and errors that may otherwise proliferate in a pipeline that only employs gene predictors (Mathé, Sagot et al. 2002).

iv. Assembling an improved pipeline

Our approach combines aspects of each pipeline, and includes an additional aspect related to RNA-Seq read processing and extrinsic evidence creation (Figure 8).

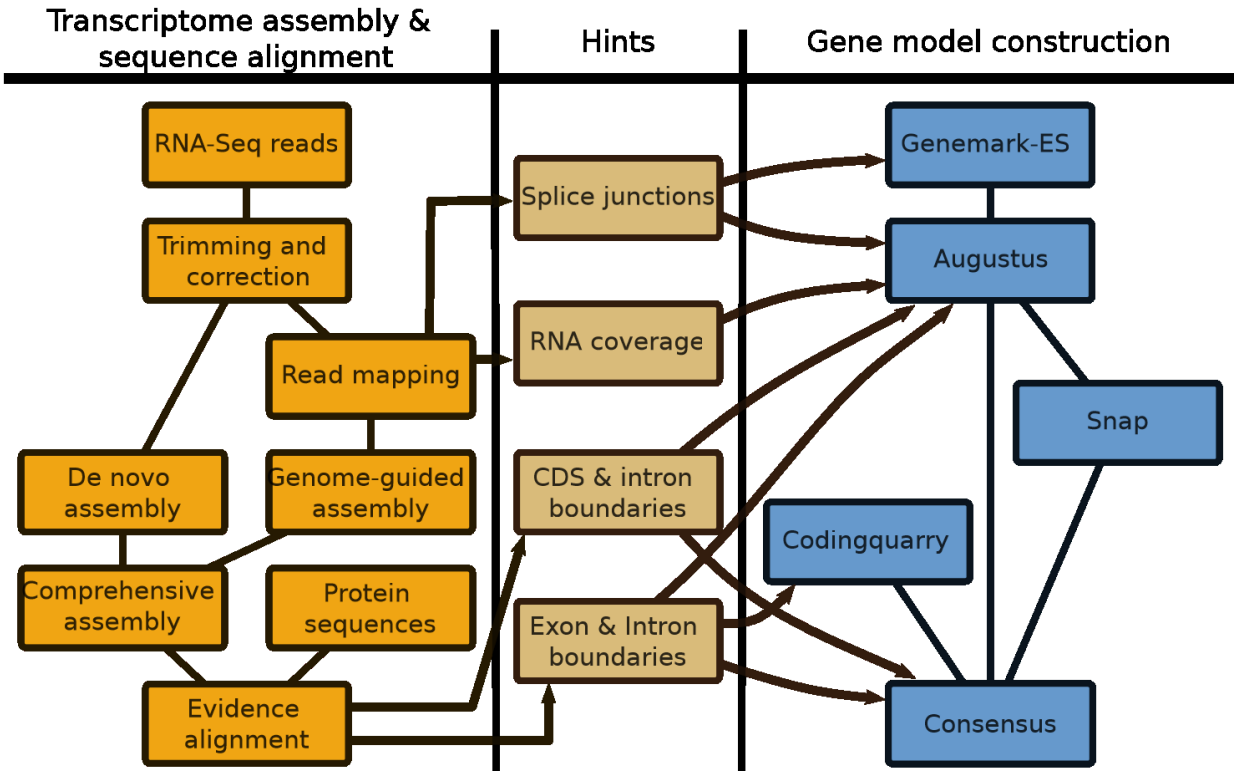

Figure 8: Proposed annotation approach. RNA-Seq reads a cleaned, corrected and mapped to the genome. Mapped reads are used to guide the transcriptome assembly and to build the de novo assembly, which are combined into a non-redundant assembly. Exon-intron and CDS hints are extracted from transcript and protein sequences, respectively. Preliminary models are constructed using Genemark-ES with splice junction hints. A subset of those predictions consistent with evidence are chosen as a training set for Augustus. Next, Augustus predictions are informed by all the extrinsic information from the assembly and alignment phase. Snap is trained on the output from Augustus, while Codingquarry is independently trained using the transcriptome assembly. Finally, a consensus set is built from the gene predictions, the transcriptome assembly and and protein alignments.

Splice junction and coverage evidence from corrected (and spliced leader sequences removed, if present) RNA-Seq reads, exon boundary evidence extracted from a non-redundant, combined *de novo* and genome-guided transcriptome assembly, and CDS hints from protein sequence alignments (even if they are sparse) are collected for the gene prediction phase. Coverage and splice junctions provide direct sources of evidence for expressed regions and their introns (Mortazavi, Williams et al. 2008). The RNA-Seq reads (coverage) are assembled into a non-redundant transcriptome, using both a *de novo* and genome-guided approach, which

provides a mostly high-confidence set of reconstructed transcripts (Grabherr, Haas et al. 2011, Haas, Papanicolaou et al. 2013) (though the benefit to combining both approaches has not yet been definitively shown, to our knowledge). The transcripts are then aligned to the genome, using PASA (Haas, Salzberg et al. 2008), to extract exon-intron boundaries, while protein sequences are aligned using Spaln (Gotoh 2008) with organism-specific parameters (Iwata and Gotoh 2012) to extract hints about coding regions. Genemark (Lomsadze, Burns et al. 2014) is trained with splice junctions, Augustus (Stanke, Diekhans et al. 2008) is trained with all experimental evidence, Snap (Korf 2004) is trained on the Augustus models, and Codingquarry (Testa, Hane et al. 2015) is trained with the assembled transcripts. Though Codingquarry was developed for fungi, it can also be used to annotate protists given gene structure similarities, but the authors state that it cannot be extended to higher eukaryotes at the moment. A diversity of predictors were included given that they each implement generalised HMMs differently which could increase the robustness of the model pool compared to generating alternatives from the same predictor (as in Snowyowl). The output from each predictor, along with the aligned transcriptome assembly and protein sequences, are consolidated into a consensus set through the weighted framework implemented in Evidencemodeler (Haas, Salzberg et al. 2008). We followed the authors' guideline on assigning weights intuitively (transcript assembly > protein alignments >= gene predictions), and empirically derived a weight combination of 50%, 33% and 17%, respectively, that consistently generated high-confidence models for the tested organisms (data not shown). The 3-2-1 ratio is also more consistent with the goal of reducing the impact of any one source of evidence.

v. Performance of the pipeline

Improvements in annotation quality were partly achieved by leveraging a more diverse set of extrinsic evidence to train gene predictors and to inform the construction of gene models. For instance, error-corrected RNA-Seq reads were used as evidence of expressed genome regions and of splice junctions, as well as for reconstructing and aligning transcript sequences to infer exons. Protein sequences were aligned, using an improved tool that also models intron length distributions, to infer putative coding regions. Reducing biases was further achieved by including a diversity of gene prediction software, each trained on different sources of evidence. Finally, a consensus building step was introduced to distribute the weight of evidence and predictions in a way that can prevent one element from dictating the structure of a model.

However, organism-specific RNA data is slightly favoured over cross-species protein sequence alignments, which is slightly favoured over gene predictions.

Table 3 indicates that the proposed approach can achieve better sensitivity and specificity with respect to the reference compared to other pipelines, with the largest gains observable in intron sensitivity and specificity (more than 2% in sensitivity and specificity in some cases, e.g. *P. falciparum*). These results demonstrate that reduction of errors and biases at the level of RNA-Seq reads, transcript reconstruction, protein sequence alignment, gene predictions, and consensus building yields improved gene models.

Species	Predictions	Braker	IH	Maker	Snowyowl
N crassa	gSn (%)	63.28	**71.06**	64.3	19.1
N crassa	gSp (%)	70.87	**74.05**	69.1	28.45
N crassa	tSn (%)	60.68	**68.05**	61.78	18.49
N crassa	tSp (%)	70.87	**74.05**	69.1	28.45
N crassa	eSn (%)	69.88	**77.28**	72.52	46.08
N crassa	eSp (%)	85.07	**86.13**	78.84	47.49
N crassa	iSp (%)	91.83	**92.92**	82.1	54.54
N crassa	iSn (%)	73.34	**81.94**	76.91	64.24
U. maydis	gSn (%)	79.67	**80.95**	73.75	62.51
U. maydis	gSp (%)	82.59	**82.89**	73.02	74.77
U. maydis	tSn (%)	79.46	**80.73**	73.55	62.35
U. maydis	tSp (%)	82.59	**82.89**	73.02	74.77
U. maydis	eSn (%)	79.42	**80.08**	74.87	63.38
U. maydis	eSp (%)	83.05	**84.25**	69.25	74.19
U. maydis	iSp (%)	80.84	**85.33**	59.18	72.12
U. maydis	iSn (%)	76.16	**76.4**	73.92	64.92
P. falciparum	gSn (%)	81.33	**82.64**	43.64	-
P. falciparum	gSp (%)	73.42	**75.29**	27.57	-
P. falciparum	tSn (%)	81.23	**82.55**	43.59	-
P. falciparum	tSp (%)	73.42	**75.29**	27.57	-
P. falciparum	eSn (%)	88.66	**90.33**	66.33	-
P. falciparum	eSp (%)	82.21	**85.1**	27.54	-
P. falciparum	iSp (%)	85.28	**89.78**	26.86	-
P. falciparum	iSn (%)	90.42	**93.29**	78.79	-

Table 3: Gene (gSn, gSp), transcript (tSn, tSp), exon (eSn, eSp) and intron (iSn, iSp) sensitivity and specificity of Braker, Maker, Snowyowl, and our in-house approach (IH), with respect to the reference annotations of *U. maydis, N. crassa, P. falciparum.* Snowyowl was not run on *P. falciparum*, given that it is not a fungus (-). The highest sensitivity and specificity achieved between the pipelines, at each level, are highlighted in bold.

We also performed pairwise comparisons between gene models predicted by each pipeline, and extracted only those models that are also in common with the reference (Supplemental Table 8, Table 9, Table 10, Table 11), providing a direct comparison of True Positive accuracy. Most notable is the performance of our tool compared to Braker (Supplemental Table 10). Only 10% of the pairwise comparisons had Braker modelled a genomic feature either equally or better than our approach. For instance, the intron sensitivity

in *U. maydis*, and the gene and transcript specificity in *P. falciparum*. A similar trend was observed in the pairwise comparison of models from our approach to those of the other pipelines (Supplemental Table 11). The gene, transcript and exon specificity of Braker at the loci predicted by our approach was slightly higher in *P. falciparum*. Nonetheless, on average, the pairwise analysis further supports the suggestion that the proposed approach is an improvement on the known pipelines.

vi. Annotation is less challenging for intron-poor eukaryotes

Baker's yeast was chosen as a test for annotation performance on an extreme case where gene density is high, with few total introns (280 total) among genes that also have short exons (~117 exons <25bp), and is well annotated. Table 4 indicates that sensitivity and specificity achieved across all the pipelines were similar. Snowyowl achieved the highest specificity, but also the lowest sensitivity, suggesting that it generated highly representative models of the reference at the cost of an increased False Negative rate—similarly to its models of *U. maydis*. On the other hand, our approach achieved the highest overlap with the reference at every level while still obtaining a relatively high Sp. This implies that there may still be a greater incentive to employ the proposed approach given its increased capacity to discover novel genes.

Species	Predictions	Braker	IH	Maker	Snowyowl
S. cerevisiae	gSn (%)	85.27	**88.36**	78.42	81.02
S. cerevisiae	gSp (%)	90.41	89.38	82.88	**92.06**
S. cerevisiae	tSn (%)	85.27	**88.36**	78.42	81.02
S. cerevisiae	tSp (%)	90.41	89.38	82.88	**92.06**
S. cerevisiae	eSn (%)	85.48	**88.32**	77.99	77.99
S. cerevisiae	eSp (%)	89.6	88.93	76.37	**92.09**
S. cerevisiae	iSp (%)	72.67	79.67	19.64	**92.86**
S. cerevisiae	iSn (%)	86.43	**86.79**	51.07	13.93

Table 4: Gene, transcript and exon sensitivity and specificity of Braker, Maker, our in-house approach, and SnowyOwl (fungi-only) gene models with respect to the *S. cerevisiae* reference. The highest sensitivity and specificity achieved between the pipelines, at each level, are highlighted in bold. The pipelines perform similarly for an intron-poor organism such as baker's yeast.

i. Annotating organisms that are either evolutionarily divergent, or have few characterised neighbours, or both

Arguably most challenging is the annotation of newly sequenced nuclear genomes that are divergent, i.e that have no close neighbours to extract meaningful information about coding regions. Obviously, Maker and Snowyowl are not optimal choices for such cases given their dependence on sequences from homologous proteins (protein-to-genome alignments in Maker, and similarity searches of predicted genes in Snowyowl). An alternative is to forgo such information, as in Braker. However, protein sequence alignments may still provide valuable

insight into gene structure, even if from phylogenetically distant relatives. In such cases, the question was, if the consensus building process can still exploit information from protein sequence alignments, even if incomplete, without reducing sensitivity and specificity.

The same weighted consensus-building step applied to *N. crassa, S. cerevisiae, P. falciparum* and *U. maydis* was applied to annotating *N. crassa* using the *Aspergillus nidulans* proteome as sole protein sequence input to simulate a phylogenetically distant and/or derived organism. Surprisingly, a larger fraction of the reference gene models was relatively closer to the Maker protein sequence alignments (Ref-Maker, Maker-Ref) than the alignments in our approach (Ref-IH, IH-Ref, Table 5).

Mapping	Count	Total	Fraction
Ref-IH	8531	28596	0.298
Ref-Maker	10183	28550	0.358
IH-Ref	3288	5163	0.637
Maker-Ref	5578	7672	0.726

Table 5: Summary of protein sequence alignment intervals falling within a relative distance of 00.10 (out of 0.50) for Maker and our approach tested in *N. crassa* with a single proteome. Relative distance was computed as a function of the reference relative to the protein alignments, and conversely as a function of protein alignments relative to the reference.

Yet, Table 6 indicates that our proposed pipeline is still capable of generating higher accuracy models compared to Maker and Snowyowl, as well as compared to the RNA-Seq-only Braker (except in exon specificity and intron sensitivity).

Predictions	IH	Maker	Snowyowl
gSn (%)	**65.3***	63.76	13.53
gSp (%)	**72.19***	62.84	8.74
tSn (%)	**62.61***	61.28	13.12
tSp (%)	**72.19***	63.13	8.74
eSn (%)	**72.24***	69.11	44.05
eSp (%)	**85.23***	77.46	22.83
iSp (%)	**90.8**	86.31	32.46
iSn (%)	**75.74***	71.87	71.83

Table 6: Sensitivity and specificity of the IH, Maker and Snowyowl models with respect to the *N. crassa* reference annotation when the proteome of a single, evolutionarily distant relative is used as protein sequence input. The values marked with an asterisk indicate better performance with respect to Braker results of **Table 3**, values in bold indicate the highest specificity and sensitivity between the three listed pipelines.

This suggests that, despite fewer protein sequence alignments, our approach is better capable of modelling the reference. The results further support the hypothesis that lacking protein sequence alignments can have a non-trivial impact on Maker's final models given that the sensitivity and specificity listed in Table 6 are even lower than those of Figure 6. More importantly, it supports the hypothesis that protein sequence data can still be leveraged without a significant impact on the consensus models.

I. Conclusion

By evaluating the sensitivity and specificity of gene models at major steps Maker, Braker and Snowyowl, it was possible to determine the points where errors or biases could have potentially been introduced. By targeting those sources of error, we developed a novel pipeline that mitigates the impact those errors thereby improving the consensus gene models. Future improvements could be the development of a protein sequence alignment tool that employs a (perhaps probabilistic) model that better captures evolutionary signal. Advances in artificial intelligence algorithms could also provide alternatives to the popular generalised-HMM framework to, for example, also model the phylogenetic component.

IV. Materials & Methods

a. Data

The data used in this study are all publicly available at https://www.ncbi.nlm.nih.gov/. *Neurospora crassa* GenBank assembly accession: GCA_000182925.2 (and corresponding annotations); RNA-Seq read accessions: SRR5000486, SRR5000484, SRR5000482. *Saccharomyces cerevisiae* GenBank assembly accession: GCA_000146045.2; RNA-Seq read accessions: SRR3396393, SRR3396392, SRR3396391, SRR3396389, SRR3396388, SRR3396387, SRR3396386, SRR3396385, SRR3396384, SRR3396382, SRR3396381. *Ustilago maydis* GenBank assembly accession: GCA_000328475.2; RNA-Seq read accessions: SRR5235721. *Plasmodium falciparum* GenBank assembly accession: GCA_000002765.1; RNA-Seq read accessions: SRR638980, SRR638979.

b. RNA-Seq read cleaning

The RNA-Seq read adapters are trimmed, followed by a Phred quality score trimming using the tool Trimmomatic v. 0.35 (Bolger, Lohse et al. 2014). A separate fasta file containing the known adapter sequences was passed to the software. Sequences were trimmed if they fell below the seed mismatch rate of 3bp and palindrome clipping threshold of 30bp. Secondly, the leading and trailing bp were trimmed if they fell below a Phred score of 3. A minimum score of 5 over a 4bp sliding window was kept for the internal bp calls. The resulting read pairs were kept if each of their sequences were greater than 20b. If one of the read pairs were dropped because it did not meet the minimum criteria, then its orphaned pair was labelled as a single-end read. If

neither of the reads in a pair met minimum criteria, then they were rejected. The tool Rcorrector (Song and Florea 2015), version 1.0.2, was run on the adapter- and quality-trimmed reads to correct both random sequencing error inherent to the sequencing technology as well as potential residues of adapters left over from the trimming process. A k-mer window of 32 with a maximum of one correction per window was used. Potential PhiX spike-ins were removed by mapping reads to the PhiX genome with bowtie2 (Langmead, Trapnell et al. 2009), then extracting hits with an in-house perl script (extract-reads).

c. RNA-Seq read mapping

The adapter- and quality-trimmed reads were mapped to the genome using STAR v. 2.5.2b (Dobin, Davis et al. 2013). A genome assembly fasta file index was built by running STAR in '--genomeGenerate' mode with '—runThreadN 8' and '—limitGenomeGenerateRAM 31'. Next, the paired-end reads were locally mapped to the genome, with a minimum and maximum intron length of 20 and 1000, respectively. The '--outSJFilterIntronMaxVsReadN' switch was invoked with parameters 100, 200, 500 and 1000 to limit the maximum splice junction size supported by 1, 2, 3 and >=4 reads, respectively. The output BAM file was sorted by coordinate with '--outSAMstrandField intronMotif' to append a field containing the intron motif, as well as the formatting parameters '--outSAMattributes Standard' and '--outSAMattrIHstart 0' for downstream compatibility. Coverage information from the BAM file was extracted, using the bam2wig program packaged with the Augustus gene prediction software and converted to gff format.

d. *De novo* and genome-guided transcriptome assembly

The Trinity transcriptome assembly software version 2.1.0 (Grabherr, Haas et al. 2011) was run in both *de novo* mode and genome-guided mode. The resulting BAM file output from the STAR (Dobin, Davis et al. 2013) read mapping was passed to the genome-guided mode of Trinity, with the same maximum intron limit as STAR and a minimum contig length of 90 to potentially capture small proteins, and all other parameters set to default. The *de novo* procedure was run with the same default parameters, except that the cleaned paired-end reads in fastq format are used instead of the mapped reads.

e. Repeat region masking

Repeat masking is done with RepeatMasker version 4.0.6 (Smit, Hubley et al.) using RepBase release 2015-08-07 (Jurka, Kapitonov et al. 2005), which is a library of known simple and complex eukaryotic repeats. Second, a *de novo* simple repeat library was constructed of 10-mers appearing >150 times in the genome was constructed using RepeatScout v. 1.0.5 (Price, Jones et al. 2005) and subsequently masked by running a second round of RepeatMasker. Third, transposable elements were identified using TransposonPSI v. 08222010 (unpublished, http://transposonpsi.sourceforge.net/) which employs a PSI-Blast (Altschul, Madden et al. 1997) identification of transposable element profiles shipped with the software. The output from the three repeat-finding tools was clustered and combined at >= 70% sequence identity using Usearch version 7.0.1090 (Edgar 2010), yielding an organism-specific repeat library. The final library served as input to RepeatMasker to convert upper case nucleotides in repetitive sequences to lower case (soft-masked) in the genome fasta file.

f. Spliced alignment of transcript and/or protein sequences

Transcripts, assembled from the genome-guided and *de novo* approaches, are first trimmed of poly-adenylated regions using seqclean (unpublished, https://sourceforge.net/projects/seqclean/) and aligned to the genome with the Program to Assemble Spliced Alignments (PASA, (Haas 2003). PASA uses Blat (Kent 2002) and Gmap (Wu and Watanabe 2005) to align transcripts, and then employs a dynamic programming algorithm to refine aligned intervals at exon/intron junctions. Only transcripts whose sequence identity to a genomic region was >=95% over >=95% of its sequence length were kept. A comprehensive transcriptome in gff format was created by clustering aligned transcripts from both assemblies that overlapped >=30% of their sequence. Protein sequences from the ten close neighbours were aligned to the genome with Spaln version 2.2.2e (Gotoh 2008) using the default parameters for aligning crossspecies proteins to an indexed genome file as described in the vignette (http://www.genome.ist.i.kyotou.ac.jp/~aln_user/spaln/index.html#Seq).

g. Evidence-based gene prediction

First, splice junctions inferred from the STAR mapping were supplied to GeneMark v. 4.32 (Lomsadze, Burns et al. 2014) for intron-aware, organism-specific self-training. The GeneMark models whose coding regions are fully supported by either a predicted mRNA (using

TransDecoder https://github.com/TransDecoder/TransDecoder/wiki) or protein sequence alignment were selected to train Augustus v. 3.2.1 (Stanke, Diekhans et al. 2008). The splice junction gff from STAR, the gff containing coverage information and the spliced alignments of transcripts were used as evidence to guide Augustus predictions. The Augustus gene models were used to train the Snap (Korf 2004) gene predictor (release 2013-11-29), as outlined in the manual shipped with the software (http://korflab.ucdavis.edu/Software/snap-2013-11-29.tar.gz). CodingQuarry v2.0 (Testa, Hane et al. 2015) was trained on the assembled and aligned transcripts, and run on the genome assembly.

h. Gene model construction by consolidating evidence and predictions

Evidencemodeler (Haas, Salzberg et al. 2008) was incorporated into the pipeline to consolidate the alignment information from transcripts, proteins and gene predictions into consensus gene models. The procedure for Evidencemodeler was performed as suggested by the developers (http://evidencemodeler.github.io/), except that a weighting scheme file was created that specified 2 for spliced protein alignments, 1 for all gene predictions and 3 for the transcriptome assembly.

V. Supplemental data

Species	Pipeline	Count	Prediction
N. crassa	Maker	281	Complete BUSCOs
N. crassa	Maker	6	Complete and single-copy
N. crassa	Maker	275	Complete and duplicated
N. crassa	Maker	1	Fragmented BUSCOs
N. crassa	Maker	8	Missing BUSCOs
N. crassa	Maker	290	Total BUSCOs
N. crassa	IH	114	Complete BUSCOs
N. crassa	IH	19	Complete and single-copy
N. crassa	IH	95	Complete and duplicated
N. crassa	IH	121	Fragmented BUSCOs
N. crassa	IH	55	Missing BUSCOs
N. crassa	IH	290	Total BUSCOs
S. cerevisiae	Maker	278	Complete BUSCOs
S. cerevisiae	Maker	8	Complete and single-copy
S. cerevisiae	Maker	270	Complete and duplicated
S. cerevisiae	Maker	3	Fragmented BUSCOs
S. cerevisiae	Maker	9	Missing BUSCOs
S. cerevisiae	Maker	290	Total BUSCOs
S. cerevisiae	IH	265	Complete BUSCOs
S. cerevisiae	IH	28	Complete and single-copy
S. cerevisiae	IH	237	Complete and duplicated
S. cerevisiae	IH	5	Fragmented BUSCOs
S. cerevisiae	IH	20	Missing BUSCOs
S. cerevisiae	IH	290	Total BUSCOs

U. maydis	Maker	286	Complete BUSCOs
U. maydis	Maker	24	Complete and single-copy
U. maydis	Maker	262	Complete and duplicated
U. maydis	Maker	3	Fragmented BUSCOs
U. maydis	Maker	1	Missing BUSCOs
U. maydis	Maker	290	Total BUSCOs
U. maydis	IH	234	Complete BUSCOs
U. maydis	IH	38	Complete and single-copy
U. maydis	IH	196	Complete and duplicated
U. maydis	IH	35	Fragmented BUSCOs
U. maydis	IH	21	Missing BUSCOs
U. maydis	IH	290	Total BUSCOs
P. falciparum	Maker	94	Complete BUSCOs
P. falciparum	Maker	14	Complete and single-copy
P. falciparum	Maker	80	Complete and duplicated
P. falciparum	Maker	7	Fragmented BUSCOs
P. falciparum	Maker	202	Missing BUSCOs
P. falciparum	Maker	303	Total BUSCOs
P. falciparum	IH	218	Complete BUSCOs
P. falciparum	IH	4	Complete and single-copy
P. falciparum	IH	214	Complete and duplicated
P. falciparum	IH	10	Fragmented BUSCOs
P. falciparum	IH	75	Missing BUSCOs
P. falciparum	IH	303	Total BUSCOs

Table 7: Summary of the BUSCO results on the protein sequence alignments of Maker and the proposed approach.

Species	Pipeline	Predictions	Braker	IH	Maker	Snowyowl
N crassa	Maker	gSn (%)	69.2	**77.07**	70.38	20.56
N crassa	Maker	gSp (%)	71.45	**76.54**	71.11	32.82
N crassa	Maker	tSn (%)	65.82	**73.21**	67.07	19.76
N crassa	Maker	tSp (%)	71.45	**76.54**	71.11	32.82
N crassa	Maker	eSn (%)	75.62	**83.3**	78.67	49.24
N crassa	Maker	eSp (%)	85.47	**88.05**	81.36	51.72
N crassa	Maker	iSp (%)	92.01	**94.11**	84.68	58.4
N crassa	Maker	iSn (%)	79	**88.02**	82.92	68.51
S. cerevisiae	Maker	gSn (%)	89.27	**91.88**	83.13	85.19
S. cerevisiae	Maker	gSp (%)	91.82	92.43	87.08	**93.06**
S. cerevisiae	Maker	tSn (%)	89.27	**91.88**	83.13	85.19
S. cerevisiae	Maker	tSp (%)	91.82	92.43	87.08	**93.06**
S. cerevisiae	Maker	eSn (%)	89.36	**91.89**	82.75	82.12
S. cerevisiae	Maker	eSp (%)	91.06	92.21	81.59	**93.27**
S. cerevisiae	Maker	iSp (%)	74.76	**83.51**	23.83	97.5
S. cerevisiae	Maker	iSn (%)	**87.97**	87.59	53.76	14.66
U. maydis	Maker	gSn (%)	82.41	**83.67**	76.39	64.41
U. maydis	Maker	gSp (%)	84.03	**84.75**	79.41	80.14
U. maydis	Maker	tSn (%)	82.18	**83.44**	76.18	64.23
U. maydis	Maker	tSp (%)	84.03	**84.75**	79.41	80.14
U. maydis	Maker	eSn (%)	82.63	**83.29**	78.07	65.7
U. maydis	Maker	eSp (%)	84.68	**86.14**	76.57	81.66
U. maydis	Maker	iSp (%)	82.9	**87.09**	67.44	84.18
U. maydis	Maker	iSn (%)	80.37	**80.48**	78.19	68.23
P. falciparum	Maker	gSn (%)	81.64	**82.8**	44.13	-
P. falciparum	Maker	gSp (%)	**81.69**	81.65	41.27	-
P. falciparum	Maker	tSn (%)	81.55	**82.71**	44.08	-
P. falciparum	Maker	tSp (%)	**81.69**	81.65	41.27	-
P. falciparum	Maker	eSn (%)	88.86	**90.47**	66.84	-
P. falciparum	Maker	eSp (%)	**90.28**	89.76	47.18	-
P. falciparum	Maker	iSp (%)	92.83	**92.94**	48.4	-
P. falciparum	Maker	iSn (%)	90.54	**93.36**	79.2	-

Table 8: Sensitivity and specificity of the Maker models that overlap with the reference, compared to the models of the other pipelines at the same genomic loci. The highest sensitivity and specificity achieved between the pipelines, at each level, are highlighted in bold.

Species	Pipeline	Predictions	Braker	IH	Maker	Snowyowl
N crassa	Snowyowl	gSn (%)	62.5	**71.59**	63.78	27.05
N crassa	Snowyowl	gSp (%)	67.22	**72.87**	67.71	32.26
N crassa	Snowyowl	tSn (%)	59.38	**67.81**	60.7	25.73
N crassa	Snowyowl	tSp (%)	67.22	**72.87**	67.71	32.26
N crassa	Snowyowl	eSn (%)	71.09	**79.62**	74.37	60.47
N crassa	Snowyowl	eSp (%)	84.1	**86.81**	80.3	51.21
N crassa	Snowyowl	iSp (%)	91.42	**93.61**	84.03	57.76
N crassa	Snowyowl	iSn (%)	75.23	**84.9**	79.24	80.75
S. cerevisiae	Snowyowl	gSn (%)	90.57	91.84	83.4	**91.96**
S. cerevisiae	Snowyowl	gSp (%)	91.59	92.04	88.04	**92.41**
S. cerevisiae	Snowyowl	tSn (%)	90.57	91.84	83.4	**91.96**
S. cerevisiae	Snowyowl	tSp (%)	91.59	92.04	88.04	**92.41**
S. cerevisiae	Snowyowl	eSn (%)	90.5	**92.24**	83.21	90.58
S. cerevisiae	Snowyowl	eSp (%)	90.81	92.33	82.21	**92.87**
S. cerevisiae	Snowyowl	iSp (%)	66.84	80.92	17.81	**97.5**
S. cerevisiae	Snowyowl	iSn (%)	**85.62**	84.25	58.9	26.71
U. maydis	Snowyowl	gSn (%)	85.17	**85.68**	78.07	79.96
U. maydis	Snowyowl	gSp (%)	85.85	**86.2**	80.82	79.93
U. maydis	Snowyowl	tSn (%)	84.97	**85.48**	77.89	79.78
U. maydis	Snowyowl	tSp (%)	85.85	**86.2**	80.82	79.93
U. maydis	Snowyowl	eSn (%)	85.22	**85.34**	79.71	80.38
U. maydis	Snowyowl	eSp (%)	86.67	**87.73**	77.97	81.47
U. maydis	Snowyowl	iSp (%)	85.98	**89.33**	69.42	83.93
U. maydis	Snowyowl	iSn (%)	**83.24**	83.03	79.92	80.55

Table 9: Sensitivity and specificity of the Snowyowl models that overlap with the reference, compared to the models of the other pipelines at the same genomic loci. The highest sensitivity and specificity achieved between the pipelines, at each level, are highlighted in bold.

Species	Pipeline	Predictions	Braker	IH	Maker	Snowyowl
N crassa	Braker	gSn (%)	71.84	**78.66**	71.31	21.01
N crassa	Braker	gSp (%)	71.27	**77.44**	72.54	33.22
N crassa	Braker	tSn (%)	68.21	**74.59**	67.87	20.17
N crassa	Braker	tSp (%)	71.27	**77.44**	72.54	33.22
N crassa	Braker	eSn (%)	78.57	**84.86**	79.52	49.95
N crassa	Braker	eSp (%)	85.33	**88.63**	82.38	51.81
N crassa	Braker	iSp (%)	91.96	**94.53**	85.59	58.37
N crassa	Braker	iSn (%)	82.19	**89.7**	84.01	69.52
S. cerevisiae	Braker	gSn (%)	90.56	**90.89**	81.81	85.06
S. cerevisiae	Braker	gSp (%)	91.09	91.13	87.48	**92.38**
S. cerevisiae	Braker	tSn (%)	90.56	**90.89**	81.81	85.06
S. cerevisiae	Braker	tSp (%)	91.09	91.13	87.48	**92.38**
S. cerevisiae	Braker	eSn (%)	90.71	**91.64**	81.39	82.17
S. cerevisiae	Braker	eSp (%)	90.36	91.71	82.11	**92.84**
S. cerevisiae	Braker	iSp (%)	74.69	84.23	24.78	**97.5**
S. cerevisiae	Braker	iSn (%)	**89.96**	87.36	53.16	14.5
U. maydis	Braker	gSn (%)	84.15	**84.96**	76.96	65.59
U. maydis	Braker	gSp (%)	83.97	**84.85**	80.26	80.34
U. maydis	Braker	tSn (%)	83.92	**84.73**	76.74	65.41
U. maydis	Braker	tSp (%)	83.97	**84.85**	80.26	80.34
U. maydis	Braker	eSn (%)	84.92	**84.94**	78.83	67.24
U. maydis	Braker	eSp (%)	84.56	**86.18**	77.23	82.09
U. maydis	Braker	iSp (%)	82.62	**87.02**	68.03	85.22
U. maydis	Braker	iSn (%)	**83.82**	82.93	79.87	70.7
P. falciparum	Braker	gSn (%)	82.02	**82.95**	43.91	-
P. falciparum	Braker	gSp (%)	**81.63**	**81.63**	41.36	-
P. falciparum	Braker	tSn (%)	81.92	**82.85**	43.86	-
P. falciparum	Braker	tSp (%)	**81.63**	**81.63**	41.36	-
P. falciparum	Braker	eSn (%)	89.29	**90.6**	66.55	-
P. falciparum	Braker	eSp (%)	**90.22**	89.8	47.26	-
P. falciparum	Braker	iSp (%)	92.79	**93.05**	48.54	-
P. falciparum	Braker	iSn (%)	91	**93.51**	79.05	-

Table 10: Sensitivity and specificity of the Braker models that overlap with the reference, compared to the models of the other pipelines at the same genomic loci. The highest Sensitivity and specificity achieved between the pipelines, at each level, are highlighted in bold.

Species	Pipeline	Predictions	Braker	IH	Maker	Snowyowl
N crassa	IH	gSn (%)	68.93	**77.29**	69.5	20.44
N crassa	IH	gSp (%)	71.35	**75.86**	70.76	32.74
N crassa	IH	tSn (%)	65.61	**73.45**	66.31	19.67
N crassa	IH	tSp (%)	71.35	**75.86**	70.76	32.74
N crassa	IH	eSn (%)	75.46	**83.55**	77.96	49.15
N crassa	IH	eSp (%)	85.38	**87.63**	80.9	51.66
N crassa	IH	iSp (%)	91.99	**93.94**	84.26	58.27
N crassa	IH	iSn (%)	78.93	**88.27**	84.44	68.44
S. cerevisiae	IH	gSn (%)	88.4	**90.76**	81.1	83.55
S. cerevisiae	IH	gSp (%)	91.12	90.91	86.93	**92.4**
S. cerevisiae	IH	tSn (%)	88.4	**90.76**	81.1	83.55
S. cerevisiae	IH	tSp (%)	90.42	90.91	86.93	**92.4**
S. cerevisiae	IH	eSn (%)	75.08	**91.5**	80.66	80.68
S. cerevisiae	IH	eSp (%)	86.69	**91.42**	81.45	**92.86**
S. cerevisiae	IH	iSp (%)	83.86	83.22	23.83	**97.5**
S. cerevisiae	IH	iSn (%)	84.28	**87.41**	51.44	14.03
U. maydis	IH	gSn (%)	83.86	**85.58**	77.19	65.59
U. maydis	IH	gSp (%)	84.28	**84.68**	78.6	80.46
U. maydis	IH	tSn (%)	83.62	**85.34**	76.98	65.4
U. maydis	IH	tSp (%)	84.28	**84.68**	78.6	80.46
U. maydis	IH	eSn (%)	84.19	**85.42**	78.88	66.97
U. maydis	IH	eSp (%)	84.81	**86.07**	75.62	82.25
U. maydis	IH	iSp (%)	82.76	**87**	66.57	85.48
U. maydis	IH	iSn (%)	82.18	**83.06**	79.34	69.81
P. falciparum	IH	gSn (%)	81.72	**83.1**	43.86	-
P. falciparum	IH	gSp (%)	**81.69**	81.62	41.36	-
P. falciparum	IH	tSn (%)	81.63	**83.01**	43.81	-
P. falciparum	IH	tSp (%)	**81.69**	81.62	41.36	-
P. falciparum	IH	eSn (%)	88.95	**90.7**	66.53	-
P. falciparum	IH	eSp (%)	**90.3**	89.71	47.33	-
P. falciparum	IH	iSp (%)	92.88	**92.9**	48.59	-
P. falciparum	IH	iSn (%)	90.65	**93.55**	78.95	-

Table 11: Sensitivity and specificity of the IH models that overlap with the reference, compared to the models of the other pipelines at the same genomic loci. The highest sensitivity and specificity achieved between the pipelines, at each level, are highlighted in bold.

3. General discussion

As outlined in the previous manuscript section, eukaryotic genome annotation continues to be a challenging task—exacerbated for species that are phylogenetically distant from model organisms. Despite the substantial improvements of our pipeline, which builds on three of the currently most popular published procedures, is has not completely reached the level of an experienced human curator (e.g., when comparing with the expert annotation by our collaborator Dr. M.W. Gray, of the set of mitochondrial genes in the jakobid nuclear genome, *Andalucia godoyi* (unpublished results; see also below)). Possible improvements for our new procedure became evident from comparative benchmarking results of the three current pipelines, determining the stages at which respective shortcomings are introduced. Our new approach encompasses RNA-Seq data processing and pre-assembly, and its integration into a genome annotation pipeline that makes use of a similar toolset and workflow as the benchmarked pipelines, but instead introduces a weighting of the various sources of evidence. Novel approaches, such as gene-transcript alignment methods based on the more sensitive HMM-method are still under development, i.e. the replacement of Exonerate that uses Blast-like methodology. The integration into our current pipeline is expected to increase the validity of gene models by a large margin.

Intron positions tend to be conserved over large evolutionary distances (Irimia and Roy 2008), yet such information is underexploited in current annotation pipelines (Hoff and Stanke 2015). Recently, a gene-structure-aware (GSA) multiple protein sequence alignment (MPSA) approach has been developed for post-annotation validation of gene models (Gotoh, Morita et al. 2014). A MPSA is computed from translated sequences of genes in the same family, and then supplemented with information on the genomic position and phase of introns of each translated sequence. Thus, a query gene model (belonging to the same family) can be assessed by including it in the GSA-MPSA and also computing its intron distribution and gap sizes relative to the other sequences. Assuming that long insertions and deletions are uncommon in closely related protein sequences (Rogozin, Carmel et al. 2012), and assuming the majority of intron positions are concordant in the MPSA, then the method can provide automated assessments and corrections of gene models. The method has been tested on the cytochrome P450 and ribosomal proteins from 47 plant genomes and has revealed that almost 50% of those gene models are either fragmented or present in unfinished areas of their respective genome

assemblies (Gotoh, Morita et al. 2014). In addition, any of the current intron prediction procedures have another major flaw – the use of a single intron splice site consensus (extended versions of the predominant GT – AG consensus), despite the recurrent presence of two distinct motifs (GT – AG plus a 'minor' AT – AC motif). This is currently handled by inferring a splice site consensus based on a mixture of both motifs, apparently biasing against recognition of introns with the less frequent minor motif. A software development that recognizes more than one set of consensus splice junctions would be highly desirable, although difficult to implement.

An approach to further strengthening extrinsic evidence from generalised HMMs is to include protein signature information derived from MPSAs of protein families, as implemented in the protein profile extension (PPX) of Augustus (Keller, Kollmar et al. 2011). While the method relies on correct *a priori* annotations, similarly to above, it can provide a complementary source of evidence that can model conserved regions in protein sequences of neighbouring organisms. The PPX models protein families from ungapped and highly conserved sections of a MPSA in a position-specific frequency matrix. The PPX computes similarity scores of all gene models (emitted from a normal run of Augustus) to protein profiles. A bonus is applied to gene models that share similarity with the profile. The authors reported >20% improvement in annotating the dynein heavy chain family in six vertebrate genomes compared to predictions from the same software without the profiles (Keller, Kollmar et al. 2011). However, a substantial challenge is that such an approach would require genome-wide protein family profile construction and validation.

Finally, a large-scale and rapid comparative approach to simultaneous structural genome annotation of multiple closely related organisms has recently been developed in light of the rapidly growing number of available genome sequences. The approach exploits negative selection and sequence conservation in multiple genome sequence alignments to construct gene models (using Augustus) (König, Romoth et al. 2016). Concordant gene structures among the input sequences tend to be preferred, but structures with missing introns or exons can also be accepted depending on the species tree. The authors report improved accuracy in gene models for 12 different species of *Drosophila* simultaneously compared to the gene models independently built for each genome (König, Romoth et al. 2016). Its use case may be somewhat limited given that it requires multiple closely related genomes.

4. Future directions

The annotation pipeline developed in this project was based on mitigating errors in current pipelines. The benchmarking results suggest that some improvements have been achieved in gene, transcript, exon and intron modelling compared to the other three current pipelines. Nonetheless, the results also suggest that further (although more difficult to reach) improvements can still be made, as automatic procedures are not yet at the level of manual curation. One avenue lies in further exploiting the conservation of introns in multiple protein sequence alignments to assess and correct gene models. In combination with the above-mentioned possible improvement by using more than one set of splice junction motifs, this may turn out to be a major step forward. A second avenue involves integrating protein profiles from multiple sequence alignments into generalised HMM prediction algorithms. Thirdly, a comparative gene-finding approach can be implemented from multiple genome alignments of closely neighbouring organisms whereby sequence conservation can be exploited. Ultimately, the improvements previously discussed all help build a more solid foundation for downstream analyses.

References

Adamidi, C., Y. Wang, D. Gruen, G. Mastrobuoni, X. You, D. Tolle, M. Dodt, S. D. Mackowiak, A. Gogol-Doering, P. Oenal, A. Rybak, E. Ross, A. S. Alvarado, S. Kempa, C. Dieterich, N. Rajewsky and W. Chen (2011). "De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics." Genome Research **21**: 1193-1200.

Adl, S. M., B. S. Leander, A. G. Simpson, J. M. Archibald, O. R. Anderson, D. Bass, S. S. Bowser, G. Brugerolle, M. A. Farmer and S. Karpov (2007). "Diversity, nomenclature, and taxonomy of protists." Systematic Biology **56**: 684–689.

Adl, S. M., A. G. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. U. Y. Brugerolle, R. A. Fensome and S. Fredericq (2005). "The new higher level classification of eukaryotes with emphasis on the taxonomy of protists." Journal of Eukaryotic Microbiology **52**: 399–451.

Allen, J. E. and S. L. Salzberg (2005). "JIGSAW: integration of multiple sources of evidence for gene prediction." Bioinformatics **21**: 3596-3603.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." Journal of molecular biology **215**: 403-410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research **25**: 3389-3402.

Aravin, A., D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M. J. Brownstein, S. Kuramochi-Miyagawa and T. Nakano (2006). "A novel class of small RNAs bind to MILI protein in mouse testes." Nature **442**: 203–207.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig and others (2000). "Gene Ontology: tool for the unification of biology." Nature genetics **25**: 25–29.

Baker, M. (2012). "De novo genome assembly: what every biologist should know." Nature methods **9**: 333-337.

Bao, Z. and S. R. Eddy (2002). "Automated de novo identification of repeat sequence families in sequenced genomes." Genome Research **12**: 1269–1276.

Barbrook, A. C., C. J. Howe, D. P. Kurniawan and S. J. Tarr (2010). "Organization and expression of organellar genomes." Philosophical Transactions of the Royal Society B: Biological Sciences **365**: 785.

Barbrook, A. C., C. R. Voolstra and C. J. Howe (2014). "The chloroplast genome of a Symbiodinium sp. clade C3 isolate." Protist **165**: 1–13.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. Sonnhammer (2002). "The Pfam protein families database." Nucleic acids research **30**: 276-280.

Batzoglou, S. (2005). "The many faces of sequence alignment." Briefings in bioinformatics **6**: 6–22.

Bedell, J. A., I. Korf and W. Gish (2000). "MaskerAid: a performance enhancement to RepeatMasker." Bioinformatics **16**: 1040-1041.

Belyi, V. A., A. J. Levine and A. M. Skalka (2010). "Sequences from Ancestral Single-Stranded DNA Viruses in Vertebrate Genomes: the Parvoviridae and Circoviridae Are More than 40 to 50 Million Years Old." Journal of Virology **84**: 12458-12462.

Benne, R., J. Van Den Burg, J. P. Brakenhoff, P. Sloof, J. H. Van Boom and M. C. Tromp (1986). "Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA." Cell **46**: 819–826.
Bennetzen, J. L. and E. A. Kellogg (1997). "Do Plants Have a One-Way Ticket to Genomic Obesity?" The Plant Cell **9**: 1509-1514.
Birney, E., M. Clamp and R. Durbin (2004). "GeneWise and genomewise." Genome research **14**: 988–995.
Bitar, M., M. Boroni, A. M. Macedo, C. R. Machado and G. R. Franco (2013). "The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles." Frontiers in Genetics **4**.
Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'donovan and I. Phan (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic acids research **31**: 365–370.
Boer, P. H. and M. W. Gray (1988). "Scrambled ribosomal RNA gene pieces in Chlamydomonas reinhardtii mitochondrial DNA." Cell **55**: 399–411.
Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**: 2114-2120.
Bork, P. and A. Bairoch (1996). "Go hunting in sequence databases but watch out for the traps." Trends in Genetics **12**: 425–427.
Brown, C. J., B. D. Hendrich, J. L. Rupert, R. G. Lafrenière, Y. Xing, J. Lawrence and H. F. Willard (1992). "The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus." Cell **71**: 527–542.
Burger, G., M. W. Gray, L. Forget and B. F. Lang (2013). "Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists." Genome Biology and Evolution **5**: 418-438.
Burset, M. and R. Guigo (1996). "Evaluation of gene structure prediction programs." Genomics **34**: 353–367.
Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado and M. Yandell (2008). "MAKER : An easy-to-use annotation pipeline designed for emerging model organism genomes." Genome Research: 188–196.
Chin, C.-S., J. Sorenson, J. B. Harris, W. P. Robins, R. C. Charles, R. R. Jean-Charles, J. Bullard, D. R. Webster, A. Kasarskis and P. Peluso (2011). "The origin of the Haitian cholera outbreak strain." New England Journal of Medicine **364**: 33–42.
Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman and D. J. Lockhart (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Molecular cell **2**: 65–73.
Church, D. M., L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z. Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M. Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, C. P. Ponting and T. M. G. S. Consortium (2009). "Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse." PLoS Biology **7**: e1000112.
Cloonan, N., A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan and S. M. Grimmond (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing." Nature Methods **5**: 613-619.

Coghlan, A., T. J. Fiedler, S. J. McKay, P. Flicek, T. W. Harris, D. Blasiar, T. n. Consortium and L. D. Stein (2008). "nGASP - the nematode genome annotation assessment project." BMC Bioinformatics **9**: 549.
Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). "Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**: 215.
Coyne, R. S., L. Hannick, D. Shanmugam, J. B. Hostetler, D. Brami, V. S. Joardar, J. Johnson, D. Radune, I. Singh and J. H. Badger (2011). "Comparative genomics of the pathogenic ciliate Ichthyophthirius multifiliis, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control." Genome biology **12**: R100.
Denoeud, F., J.-M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli and others (2008). "Annotating genomes with massive-scale RNA sequencing." Genome Biol **9**: R175.
Dimmock, N. J., A. J. Easton and K. N. Leppard (2016). Introduction to modern virology, John Wiley & Sons.
Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**: 15-21.
Durbin, R. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge university press.
Eddy, S. R. (2004). "What is a hidden Markov model?" Nature biotechnology **22**: 1315-1316.
Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**: 2460-2461.
El-Sayed, N. M., P. J. Myler, D. C. Bartholomeu, D. Nilsson, G. Aggarwal, A.-N. Tran, E. Ghedin, E. A. Worthey, A. L. Delcher and G. Blandin (2005). "The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease." Science **309**: 409-415.
Feagin, J. E., B. L. Mericle, E. Werner and M. Morris (1997). "Identification of additional rRNA fragments encoded by the Plasmodium falciparum 6 kb element." Nucleic acids research **25**: 438–446.
Fedorov, A., A. F. Merican and W. Gilbert (2002). "Large-scale comparison of intron positions among animal, plant, and fungal genes." Proceedings of the National Academy of Sciences **99**: 16128–16133.
Ferrier, D. E. and P. W. Holland (2001). "Ancient origin of the Hox gene cluster." Nature Reviews Genetics **2**: 33–38.
Firth, A. E. and C. M. Brown (2006). "Detecting overlapping coding sequences in virus genomes." BMC bioinformatics **7**: 75.
Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty and J. M. Merrick (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." science: 496-512.
Florea, L. D. and S. L. Salzberg (2013). "Genome-Guided Transcriptome Assembly in the Age of Next-Generation Sequencing." IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**: 1234-1240.
Frishman, D. (2007). "Protein annotation at genomic scale: the current status." Chemical reviews **107**: 3448–3466.
Galagan, J. E., S. E. Calvo, K. A. Borkovich, E. U. Selker, N. D. Read, D. Jaffe, W. FitzHugh, L.-J. Ma, S. Smirnov, S. Purcell and others (2003). "The genome sequence of the filamentous fungus Neurospora crassa." Nature **422**: 859–868.

Galperin, M. Y. and E. V. Koonin (1998). "Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption." In silico biology **1**: 55–67.
Garber, M., M. G. Grabherr, M. Guttman and C. Trapnell (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." Nature methods **8**: 469–477.
Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley and P. E. Burch (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**: 493–521.
Girard, A., R. Sachidanandam, G. J. Hannon and M. A. Carmell (2006). "A germline-specific class of small RNAs binds mammalian Piwi proteins." Nature **442**: 199–202.
Goffeau, A., B. G. Barrell, H. Bussey and R. W. Davis (1996). "Life with 6000 genes." Science **274**: 546.
Gotoh, O. (2008). "A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence." Nucleic Acids Research **36**: 2630-2638.
Gotoh, O., M. Morita and D. R. Nelson (2014). "Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment." BMC bioinformatics **15**: 1.
Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. a. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature biotechnology **29**: 644–652.
Gray, M. W., G. Burger and B. F. Lang (1999). "Mitochondrial evolution." Science **283**: 1476–1481.
Green, B. R. (2011). "Chloroplast genomes of photosynthetic eukaryotes." The plant journal **66**: 34–44.
Guida, A., C. Lindstädt, S. L. Maguire, C. Ding, D. G. Higgins, N. J. Corton, M. Berriman and G. Butler (2011). "Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis." BMC genomics **12**: 628.
Guigo, R., P. Agarwal, J. F. Abril, M. Burset and J. W. Fickett (2000). "An assessment of gene prediction accuracy in large DNA sequences." Genome Research **10**: 1631–1642.
Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander and A. Regev (2010). "Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nature Biotechnology **28**: 503-510.
Haas, B. J. (2003). "Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies." Nucleic Acids Research **31**: 5654-5666.
Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman and A. Regev (2013). "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." Nature Protocols **8**: 1494-1512.
Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell and J. R. Wortman (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments." Genome biology **9**: R7.
Haas, B. J., Q. Zeng, M. D. Pearson, C. A. Cuomo and J. R. Wortman (2011). "Approaches to fungal genome annotation." Mycology **2**: 118–141.

Haas, B. J. and M. C. Zody (2010). "Advancing RNA-seq analysis." Nature biotechnology **28**: 421–423.
Haas, L. W. and K. L. Webb (1979). "Nutritional mode of several non-pigmented microflagellates from the York River Estuary, Virginia." Journal of Experimental Marine Biology and Ecology **39**: 125-134.
Haft, D. H., J. D. Selengut and O. White (2003). "The TIGRFAMs database of protein families." Nucleic acids research **31**: 371–373.
Hall, N. (2003). "The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism." Nucleic Acids Research **31**: 4864-4873.
Hamilton, A. J. and D. C. Baulcombe (1999). "A species of small antisense RNA in posttranscriptional gene silencing in plants." Science **286**: 950–952.
Hampl, V., L. Hug, J. W. Leigh, J. B. Dacks, B. F. Lang, A. G. Simpson and A. J. Roger (2009). "Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups"." Proceedings of the National Academy of Sciences **106**: 3859–3864.
He, S., O. Wurtzel, K. Singh, J. L. Froula, S. Yilmaz, S. G. Tringe, Z. Wang, F. Chen, E. A. Lindquist, R. Sorek and P. Hugenholtz (2010). "Validation of two ribosomal RNA removal methods for microbial metatranscriptomics." Nature Methods **7**: 807-812.
Henson, J., G. Tischler and Z. Ning (2012). "Next-generation sequencing and large genome assemblies." Pharmacogenomics **13**: 901-915.
Hikosaka, K., Y.-i. Watanabe, N. Tsuji, K. Kita, H. Kishine, N. Arisue, N. M. Q. Palacpac, S.-i. Kawazu, H. Sawai and T. Horii (2009). "Divergence of the mitochondrial genome structure in the apicomplexan parasites, Babesia and Theileria." Molecular biology and evolution **27**: 1107–1116.
Hoff, K. and M. Stanke (2015). "Current methods for automated annotation of protein-coding genes." Current Opinion in Insect Science **7**: 8-14.
Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky and M. Stanke (2016). "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1." Bioinformatics **32**: 767-769.
Hoffmann, S., C. Otto, G. Doose, A. Tanzer, D. Langenberger, S. Christ, M. Kunz, L. M. Holdt, D. Teupser, J. Hackermüller and P. F. Stadler (2014). "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection." Genome Biology **15**: R34.
Horton, P., K.-J. Park, T. Obayashi and K. Nakai (2006). Protein Subcellular Localisation Prediction with WoLF PSORT., World Scientific.
Hou, Y. and S. Lin (2009). "Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes." PLoS One **4**: e6978.
Irimia, M. and S. W. Roy (2008). "Spliceosomal introns as tools for genomic and evolutionary analysis." Nucleic Acids Research **36**: 1703-1712.
Islam, S., U. Kjallquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lonnerberg and S. Linnarsson (2011). "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq." Genome Research **21**: 1160-1167.
Iwata, H. and O. Gotoh (2012). "Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features." Nucleic Acids Research **40**: e161-e161.
Jacob, F. and J. Monod (1961). On the regulation of gene activity, Cold Spring Harbor Laboratory Press.
Jacquier, A. (2009). "The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs." Nature Reviews Genetics **10**: 833-844.

Janouškovec, J., S.-L. Liu, P. T. Martone, W. Carré, C. Leblanc, J. Collén and P. J. Keeling (2013). "Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers." PLoS One **8**: e59001.
Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and genome research **110**: 462-467.
Kapitonov, V. V. and J. Jurka (2008). "A universal classification of eukaryotic transposable elements implemented in Repbase." Nature Reviews Genetics **9**: 411–412.
Katinka, M. D., S. Duprat, E. Cornillot, G. Méténier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier and P. Wincker (2001). "Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi." Nature **414**: 450–453.
Keibler, E. and M. R. Brent (2003). "Eval: a software package for analysis of genome annotations." BMC bioinformatics **4**: 50.
Keller, O., M. Kollmar, M. Stanke and S. Waack (2011). "A novel hybrid gene prediction method employing protein multiple sequence alignments." Bioinformatics **27**: 757–763.
Kent, W. J. (2002). "BLAT—the BLAST-like alignment tool." Genome research **12**: 656–664.
König, S., L. W. Romoth, L. Gerischer and M. Stanke (2016). "Simultaneous gene finding in multiple genomes." Bioinformatics **32**: 3388-3395.
Koonin, E. V. (2011). The logic of chance: the nature and origin of biological evolution, FT press.
Korbel, J. O., A. E. Urban, F. Grubert, J. Du, T. E. Royce, P. Starr, G. Zhong, B. S. Emanuel, S. M. Weissman, M. Snyder and M. B. Gerstein (2007). "Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome." Proceedings of the National Academy of Sciences **104**: 10110-10115.
Korf, I. (2004). "Gene finding in novel genomes." BMC bioinformatics **5**: 59.
Korf, I., M. Yandell and J. Bedell (2003). Blast, " O'Reilly Media, Inc.".
Krause, M. and D. Hirsh (1987). "A trans-spliced leader sequence on actin mRNA in C. elegans." Cell **49**: 753-761.
Lambowitz, A. M. and M. Belfort (1993). "Introns as mobile genetic elements." Annual review of biochemistry **62**: 587–622.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle and W. FitzHugh (2001). "Initial sequencing and analysis of the human genome."
Langmead, B., C. Trapnell, M. Pop, S. L. Salzberg and others (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome biol **10**: R25.
Lee, R. C., R. L. Feinbaum and V. Ambros (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**: 843–854.
Lidie, K. B. and F. M. Van Dolah (2007). "Spliced Leader RNA-Mediated trans-Splicing in a Dinoflagellate, Karenia brevis." Journal of Eukaryotic Microbiology **54**: 427-435.
Liu, Q., A. J. Mackey, D. S. Roos and F. C. N. Pereira (2008). "Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction." Bioinformatics **24**: 597-605.
Lodish, H., A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell (2000). "Molecular cell biology 4th edition." National Center for Biotechnology InformationÕs Bookshelf.
Lomsadze, A., P. D. Burns and M. Borodovsky (2014). "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm." Nucleic Acids Research **42**: e119-e119.

Lukashin, A. V. and M. Borodovsky (1998). "GeneMark. hmm: new solutions for gene finding." Nucleic acids research **26**: 1107-1115.
Lukeš, J., D. L. Guilbride, J. Votỳpka, A. Zíková, R. Benne and P. T. Englund (2002). "Kinetoplast DNA network: evolution of an improbable structure." Eukaryotic Cell **1**: 495–502.
Marande, W., J. Lukeš and G. Burger (2005). "Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids." Eukaryotic cell **4**: 1137–1146.
Mardis, E. R. (2008). "Next-Generation DNA Sequencing Methods." Annual Review of Genomics and Human Genetics **9**: 387-402.
Mardis, E. R. (2017). "DNA sequencing technologies: 2006–2016." Nature Protocols **12**: 213-218.
Marguerat, S. and J. Bähler (2010). "RNA-seq: from technology to biology." Cellular and Molecular Life Sciences **67**: 569-579.
Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature.
Martin, J., V. M. Bruno, Z. Fang, X. Meng, M. Blow, T. Zhang, G. Sherlock, M. Snyder and Z. Wang (2010). "Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads." BMC genomics **11**: 663.
Martin, J. A. and Z. Wang (2011). "Next-generation transcriptome assembly." Nature Reviews Genetics **12**: 671-682.
Mathé, C., M.-F. Sagot, T. Schiex and P. Rouzé (2002). "Current methods of gene prediction, their strengths and weaknesses." Nucleic acids research **30**: 4103–4117.
Mattick, J. S. (2001). "Non-coding RNAs: the architects of eukaryotic complexity." EMBO reports **2**: 986–991.
Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander and B. E. Bernstein (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**: 553-560.
Miller, J. R., S. Koren and G. Sutton (2010). "Assembly algorithms for next-generation sequencing data." Genomics **95**: 315–327.
Min, B., I. V. Grigoriev and I.-G. Choi (2017). "FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation." Bioinformatics **33**: 2936-2937.
Morrison, H. G., A. G. McArthur, F. D. Gillin, S. B. Aley, R. D. Adam, G. J. Olsen, A. A. Best, W. Z. Cande, F. Chen and M. J. Cipriano (2007). "Genomic minimalism in the early diverging intestinal parasite Giardia lamblia." Science **317**: 1921-1926.
Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature Methods **5**: 621-628.
Muhich, M. L. and J. C. Boothroyd (1988). "Polycistronic transcripts in trypanosomes and their accumulation during heat shock: evidence for a precursor role in mRNA synthesis." Molecular and cellular biology **8**: 3837–3846.

Myers, E. W. (1995). "Toward simplifying and accurately formulating fragment assembly." Journal of Computational Biology **2**: 275–290.
Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyder (2008). "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." Science **320**: 1344-1349.
Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai and H. Takahashi (2011). "Sequence-specific error profile of Illumina sequencers." Nucleic acids research **39**: e90–e90.
Nixon, J. E., A. Wang, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus and J. Samuelson (2002). "A spliceosomal intron in Giardia lamblia." Proceedings of the National Academy of Sciences **99**: 3701–3705.
Normark, S., S. Bergstrom, T. Edlund, T. Grundstrom, B. Jaurin, F. P. Lindberg and O. Olsson (1983). "Overlapping genes." Annual review of genetics **17**: 499–525.
Okamoto, N., C. Chantangsi, A. Horák, B. S. Leander and P. J. Keeling (2009). "Molecular phylogeny and description of the novel katablepharid Roombia truncata gen. et sp. nov., and establishment of the Hacrobia taxon nov." PloS one **4**: e7080.
Ozsolak, F., A. R. Platt, D. R. Jones, J. G. Reifenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz and P. M. Milos (2009). "Direct RNA sequencing." Nature **461**: 814-818.
Parkhomchuk, D., T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach and A. Soldatov (2009). "Transcriptome analysis by strand-specific sequencing of complementary DNA." Nucleic Acids Research **37**: e123-e123.
Pavy, N., S. Rombauts, P. Dehais, C. Mathe, D. VV Ramana, P. Leroy and P. Rouze (1999). "Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences." Bioinformatics **15**: 887–899.
Pellicer, J., M. F. Fay and I. J. Leitch (2010). "The largest eukaryotic genome of them all?" Botanical Journal of the Linnean Society **164**: 10–15.
Pevzner, P. a., H. Tang and M. S. Waterman (2001). "An Eulerian path approach to DNA fragment assembly." Proceedings of the National Academy of Sciences of the United States of America **98**: 9748–9753.
Philippe, N., M. Legendre, G. Doutre, Y. Couté, O. Poirot, M. Lescot, D. Arslan, V. Seltzer, L. Bertaux and C. Bruley (2013). "Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes." Science **341**: 281–286.
Pop, M. (2009). "Genome assembly reborn: recent computational challenges." Briefings in Bioinformatics **10**: 354-366.
Pouchkina-Stantcheva, N. N. and A. Tunnacliffe (2005). "Spliced Leader RNA–Mediated trans-Splicing in Phylum Rotifera." Molecular Biology and Evolution **22**: 1482-1489.
Price, A. L., N. C. Jones and P. A. Pevzner (2005). "De novo identification of repeat families in large genomes." Bioinformatics **21**: i351-i358.
Quesada, V. c., M. a. R. Ponce and J. L. Micol (1999). "OTC and AUL1, two convergent and overlapping genes in the nuclear genome of Arabidopsis thaliana." FEBS letters **461**: 101–106.
Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE **77**: 257–286.
Rajkovic, A., R. E. Davis, J. N. Simonsen and F. M. RoTTMAN (1990). "A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni." Proceedings of the National Academy of Sciences **87**: 8879-8883.
Reid, I., O. Nicholas, O. Zabaneh, R. Nourzadeh, M. Dahdouli, M. Abdellateef, P. M. Gordon, J. Soh, G. Butler, C. W. Sensen and others (2014). "SnowyOwl: accurate prediction of fungal

genes by using RNA-Seq and homology information to select among ab initio models." BMC bioinformatics **15**: 229.
Rogozin, I. B., L. Carmel, M. Csuros and E. V. Koonin (2012). "Origin and evolution of spliceosomal introns." Biology direct **7**: 1.
Rogozin, I. B., K. S. Makarova, J. Murvai, E. Czabarka, Y. I. Wolf, R. L. Tatusov, L. A. Szekely and E. V. Koonin (2002). "Connected gene neighborhoods in prokaryotic genomes." Nucleic acids research **30**: 2212–2223.
Salgado, H., G. Moreno-Hagelsieb, T. F. Smith and J. Collado-Vides (2000). "Operons in Escherichia coli: genomic analyses and predictions." Proceedings of the National Academy of Sciences **97**: 6652–6657.
Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop and J. A. Yorke (2012). "GAGE: A critical evaluation of genome assemblies and assembly algorithms." Genome Research **22**: 557-567.
Salzberg, S. L. and J. A. Yorke (2005). "Beware of mis-assembled genomes." Bioinformatics **21**: 4320-4321.
Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith (1977). "Nucleotide sequence of bacteriophage φX174 DNA." nature **265**: 687-695.
Schlebusch, S. and N. Illing (2012). "Next generation shotgun sequencing and the challenges of de novo genome assembly." South African Journal of Science **108**: 62–70.
Sémon, M. and L. Duret (2006). "Evolutionary origin and maintenance of coexpressed gene clusters in mammals." Molecular biology and evolution **23**: 1715–1723.
Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nature Biotechnology **26**: 1135-1145.
Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics **31**: 3210-3212.
Sims, D., I. Sudbery, N. E. Ilott, A. Heger and C. P. Ponting (2014). "Sequencing depth and coverage: key considerations in genomic analyses." Nature Reviews Genetics **15**: 121-132.
Slater, G. S. C. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC bioinformatics **6**: 31.
Sloan, D. B., A. J. Alverson, J. P. Chuckalovcak, M. Wu, D. E. McCauley, J. D. Palmer and D. R. Taylor (2012). "Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates." PLoS biology **10**: e1001241.
Smit, A. F. A., R. Hubley and P. Green 1996–2010. RepeatMasker Open-3.0.
Smith, D. R. and P. J. Keeling (2015). "Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes." Proceedings of the National Academy of Sciences **112**: 10177–10184.
Smith, D. R., R. W. Lee, J. C. Cushman, J. K. Magnuson, D. Tran and J. E. Polle (2010). "The Dunaliella salina organelle genomes: large sequences, inflated with intronic and intergenic DNA." BMC plant biology **10**: 83.
Song, L. and L. Florea (2015). "Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads." GigaScience **4**.
Spingola, m., L. Grate, D. Haussler and A. Manuel (1999). "Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae." Rna **5**: 221–234.
Stanke, M., M. Diekhans, R. Baertsch and D. Haussler (2008). "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics **24**: 637-644.

Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel." Bioinformatics **19**: ii215-ii225.
Stoebe, B. and K. V. Kowallik (1999). "Gene-cluster analysis in chloroplast genomics." Trends in genetics **15**: 344–347.
Stoughton, R. B. (2005). "Applications of DNA microarrays in biology." Annu. Rev. Biochem. **74**: 53-82.
Stover, N. A. and R. E. Steele (2001). "Trans-spliced leader addition to mRNAs in a cnidarian." Proceedings of the National Academy of Sciences **98**: 5693–5698.
Sutton, R. E. and J. C. Boothroyd (1986). "Evidence for trans splicing in trypanosomes." Cell **47**: 527-535.
Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao and M. A. Surani (2009). "mRNA-Seq whole-transcriptome analysis of a single cell." Nature Methods **6**: 377-382.
Tarazona, S., F. Garcia-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). "Differential expression in RNA-seq: A matter of depth." Genome Research **21**: 2213-2223.
Tatusov, R. L., A. R. Mushegian, P. Bork, N. P. Brown, W. S. Hayes, M. Borodovsky, K. E. Rudd and E. V. Koonin (1996). "Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli." Current biology **6**: 279–291.
Ter-Hovhannisyan, V., A. Lomsadze, Y. O. Chernoff and M. Borodovsky (2008). "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training." Genome Research **18**: 1979-1990.
Tessier, L. H., M. Keller, R. L. Chan, R. Fournier, J. H. Weil and P. Imbault (1991). "Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena." The EMBO Journal **10**: 2621-2625.
Testa, A. C., J. K. Hane, S. R. Ellwood and R. P. Oliver (2015). "CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts." BMC Genomics **16**.
Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, J. V. Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms." Nature biotechnology **28**: 511–515.
Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature Biotechnology **28**: 511-515.
Treangen, T. J. and S. L. Salzberg (2011). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nature Reviews Genetics.
van Dijk, E. L., Y. Jaszczyszyn and C. Thermes (2014). "Library preparation methods for next-generation sequencing: Tone down the bias." Experimental Cell Research **322**: 12-20.
Vandenberghe, A. E., T. H. Meedel and K. E. Hastings (2001). "mRNA 5′-leader trans-splicing in the chordates." Genes & development **15**: 294-303.
Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "Serial analysis of gene expression." Science **270**: 484.
Volfovsky, N., B. J. Haas and S. L. Salzberg (2003). "Computational Discovery of Internal Micro-Exons." Genome Research **13**: 1216-1221.
Waller, R. F. and C. J. Jackson (2009). "Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology." Bioessays **31**: 237–245.

Wang, L., N. Jiang, L. Wang, O. Fang, L. J. Leach, X. Hu and Z. Luo (2014). "3′ Untranslated Regions Mediate Transcriptional Interference between Convergent Genes Both Locally and Ectopically in Saccharomyces cerevisiae." PLoS Genetics **10**: e1004021.
Wang, Y., S. Baskerville, A. Shenoy, J. E. Babiarz, L. Baehner and R. Blelloch (2008). "Embryonic stem cell–specific microRNAs regulate the G1-S transition and promote rapid proliferation." Nature Genetics **40**: 1478-1483.
Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews. Genetics **10**: 57–63.
Wise, R. R. (2007). The diversity of plastid form and function. The structure and function of plastids, Springer**:** 3–26.
Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov and E. V. Koonin (2001). "Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context." Genome research **11**: 356–372.
Wu, T. D. and C. K. Watanabe (2005). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." Bioinformatics **21**: 1859-1875.
Yandell, M. and D. Ence (2012). "A beginner's guide to eukaryotic genome annotation." Nature reviews. Genetics **13**: 329–342.
Zerbino, D. R. and E. Birney (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." Genome Research **18**: 821-829.