GENOME ANNOTATION AND GENOTYPE-PHENOTYPE ASSOCATION IN TWO NON-MODEL PARASITES

by

Daniel D. Ence

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

December 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Daniel D. Ence**

has been approved by the following supervisory committee members:

Mark Yandell , Chair **10/20/2016**
Date Approved

Robert Weiss , Member **10/20/2016**
Date Approved

Charles Murtaugh , Member **10/20/2016**
Date Approved

Michael Shapiro , Member **10/20/2016**
Date Approved

Nels Elde , Member **10/20/2016**
Date Approved

and by **Lynn B. Jorde** , Chair/Dean of

the Department/College/School of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

With the rapid proliferation of high-throughput sequencing methods, both the number and variety of genome assemblies have increased and require rigorous and sophisticated methods for genome annotation. As more species' genomes are sequenced, the targets of genome annotation projects will more commonly be species with few closely related species that have been analyzed previously. This can be a serious challenge for genome annotation, here defined as the identification and demarcation of gene models in a genome assembly.

My PhD research has focused on the application and development of genome annotation methods for non-model organisms. In the first chapter of my thesis, I present a review of the field of genome annotation, which discusses current challenges and best-practice approaches. In the second chapter of my thesis, I present analyses of the important agronomic pest, *Cronartium quercuum* sp. fusiforme (CQF), which causes fusiform rust disease in loblolly pine trees. I annotated the genome and used genome-resequencing data to confirm results from a previous linkage mapping study that identified the location of virulence factor 1 (*Avr1*) and to identify candidate *Avr1* genes.

TABLE OF CONTENTS

Chapters

Appendices

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I'd like to thank my advisor Mark Yandell for allowing me the privilege of being his student. His passion for science and creativity are infectious to all around him and helped me keep focus and perspective. His mentorship, guidance, and encouragement have been invaluable as I've worked towards finishing my degree this year. I'd also like to thank current and former lab members, Carson Holt, Hao Hu, Zev Kronenberg, Michael Campbell, Barry Moore, Edwin Lin, Brett Kennedy, Shawn Rynearson, Steven Flygare, and Qing Li for their advice and friendship during our time together. I also want to acknowledge the contributions of the many collaborators who helped with the analysis of the genomes presented in this thesis: John M. Davis (University of Florida), Katherine E. Smith (USDA Southern Research Station), Ellen Pritham (University of Utah), C. Dana Nelson (USDA Southern Research Station), Lukas Schrader (University of Regensburg), Jürgen Gadau (Arizona State University), Jan Oettler (University of Regensburg), Claudia Marquez, Chris Elsik (The University of Missouri), Karen Kapheim (Utah State University), Hugh M. Robertson (University of Illinois at Urbana-champaign), Gene E. Robinson (University of Illinois at Urbana-Champaign), Brielle J. Fishcman, and Olav Ruepell (University of North Carolina).

CHAPTER 1

INTRODUCTION

A Brief Introduction to Genome Annotation

Genome annotation can be summarized as the process of identifying the location and structure of protein-coding genes within genomic sequence (1). In addition to identifying protein-coding genes, genome annotation can also be said to include the process of identification of repetitive sequence in a genome (2–4), as well as the process of assigning functional annotations to protein-coding genes (5–9). Genome annotation projects are usually undertaken on organisms whose genomes have recently been sequenced and assembled for the first time. Exceptions to this rule include the ongoing revisions and updates as reference genome assemblies are updated (as is currently done with the human reference genome) and whole-scale re-annotation projects necessitated by advances in the evidence available for the organism of interest (see Chapter 4 of this thesis for the re-annotation of the *Trichomonas vaginalis* (*T. vaginalis*) reference genome).

The first organisms to have their genomes sequenced, assembled, and annotated were model organisms, for which there were plenty of resources in the form of EST and cDNA libraries and amino-acid sequences of proteins, which could be used in the genome annotation process. The high-quality reference genomes produced in the first years of the “genomics-era” (10–13), in combination with the volume and quality of the evidence that had been produced by decades of prior research, allowed for confident analyses within and between species.

Genome Annotation of Non-Model Organisms

Following the release of the first annotated reference genomes, the advent of next-generation sequencing technologies (NGS) dramatically reduced the cost of genome sequencing (14, 15). Reference-based comparisons of the genomes of multiple individuals within a species allow for investigation of population genetics questions, while comparisons of *de novo* assemblies of genomes of multiple species can allow for studies of evolution at larger phylogenetic distances. Although NGS technologies allowed for the investigation of many more new and different kinds of genomic studies than before, the draft assemblies produced with these methods failed to reach the completeness and contiguity of the generation of reference genomes.

Additionally, genome projects are not equally distributed across the tree of life (16). As previously noted, multiple genome assemblies of related organisms enable comparative studies at varying phylogenetic distances. The comparisons can investigate questions such as the molecular evolution of orthologous genes in different species (17–20), the expansion or contraction of gene families (21), and the identification of so-called “orphan” genes that are present in only a single species within a clade (22). An additional use of those genomes is to leverage the annotated protein-coding genes from related species in the genome annotation of newly sequenced organisms. Early on in my thesis research, I contributed genome annotations for two species of hymenopteran insects (*Cardiocondyla obscurior* and *Megachile rotundata*), which benefited from many other hymenopteran genomes that have been annotated (23). These genome annotations formed the basis of published research which investigated evolutionary questions within and between species (23, 24, see Appendices A and B for full-text of publications). A study that includes the annotation of a newly sequenced and assembled genome as well as inter- and intraspecific genomic comparisons with multiple reference genomes is included in Chapter 3 of this thesis.

As a corollary to benefits of "annotation-rich" taxonomic neighborhoods noted above, genome projects in "annotation-poor" or understudied groups of organisms can encounter pitfalls in a lack of available resources from species closely related to the organisms of interest. These pitfalls can include a lack of annotated protein-coding genes that can be used for homology searches in the target genome, as well as novel transposable elements (TE) families in the target genome that will not be identified by repeat-masking software using data from previously published genomes. An example of these problems is the genome annotation of *T. vaginalis* in which 60,000 protein-coding genes were reported in a 160 Mbp genome (26) when an unknown number of those protein-coding genes are probably encoded by *Maverick* TE elements (27). The re-annotation of the *T. vaginalis* genome, along with inter- and intraspecific comparisons using a novel reference-free comparative genomics method, is reported in Chapter 4 of this thesis.

References

1. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13(5):329–342.

2. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.

3. Smit A, Hubley R, Green P RepeatMasker. Available at:

http://www.repeatmasker.org/.

4. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(SUPPL. 1):351–358.

5. Marchler-Bauer A, et al. (2005) CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res* 33(DATABASE ISS.):192–196.

6. Miranda-Saavedra D, Barton GJ (2007) Classification and functional annotation of eukaryotic protein kinases. *Proteins* 68(4):893–914.

7. Dutkowski J, et al. (2013) A gene ontology inferred from molecular networks. *Nat Biotech* 31(1):38–45.

8. Michael A, et al. (2000) Gene Ontology: tool for the unification of biology. *Nature* 25:25–29.

9. Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA (2013) Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol Ecol* 22(12):3216–41.

10. Celniker SE, et al. (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol* 3(12):1–14.

11. Adams MD, et al. (2000) The genome sequence of Drosophila melanogaster. *Science* 287(March):2185–2195.

12. Venter JC, et al. (2001) The sequence of the human genome. *Science (80-)* 291(February):1304–1351.

13. Hattori M (2005) Finishing the euchromatic sequence of the human genome. *Tanpakushitsu Kakusan Koso* 50(2):162–168.

14. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272.

15. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108(4):1513–1518.

16. Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. *J Hered* 100(5):648–55.

17. Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267(May):275–276.

18. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genet*

4(12):1–10.

19. Zhang J (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* 18(6):292–298.

20. Yang Z, Bielawski JR (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15(12):496–503.

21. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22(10):1269–71.

22. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E (2013) Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biol Evol* 5(2):439-455.

23. Waterhouse RM (2015) A maturing understanding of the composition of the insect gene repertoire. *Curr Opin Insect Sci* 7(January):15–23.

24. Schrader L, et al. (2014) Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat Commun* 16(5):1–10.

25. Kapheim KM, et al. (2015) Genomic signatures of evolutionary transitions from solitary to group living. *Science* (May):1–8.

26. Carlton JM, et al. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315(5809):207–12.

27. Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390(1–2):3–17.

CHAPTER 2

A BEGINNER’S GUIDE TO EUKARYOTIC GENOME ANNOTATION

The following chapter is a reprint of a review article coauthored by Mark Yandell and myself, and is presented here with permissions of the authors and kind permission of Springer Nature. This review article was first published in Ence D. and Yandell M. (2012) A beginner’s guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13(5):329-340. Available at: http://www.nature.com/nrg/journal/v13/n5/abs/nrg3174.html.

STUDY DESIGNS

A beginner's guide to eukaryotic genome annotation

Mark Yandell and Daniel Ence

Abstract | The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.

Genome annotation
A term used to describe two distinct processes. 'Structural' genome annotation is the process of identifying genes and their intron–exon structures. 'Functional' genome annotation is the process of attaching meta-data such as gene ontology terms to structural annotations. This Review focuses on structural annotation.

RNA-sequencing data
(RNA-seq data). Data sets derived from the shotgun sequencing of a whole transcriptome using next-generation sequencing (NGS) techniques. RNA-seq data are the NGS equivalent of expressed sequence tags generated by the Sanger sequencing method.

Department of Human Genetics, Eccles Institute of Human Genetics, School of Medicine, University of Utah, Salt Lake City, Utah 84112-5330, USA.
Correspondence to M.Y.
e-mail: myandell@genetics.utah.edu

doi:10.1038/nrg3174

Sequencing costs have fallen so dramatically that a single laboratory can now afford to sequence large, even human-sized, genomes. Ironically, although sequencing has become easy, in many ways, genome annotation has become more challenging. Several factors are responsible for this. First, the shorter read lengths of second-generation sequencing platforms mean that current genome assemblies rarely attain the contiguity of the classic shotgun assemblies of the *Drosophila melanogaster*[1,2] or human genomes[3,4]. Second, the exotic nature of many recently sequenced genomes also presents annotation challenges, especially for gene finding. Whereas the first generation of genome projects had recourse to large numbers of pre-existing gene models, the contents of today's genomes are often terra incognita. This makes it difficult to train, optimize and configure gene prediction and annotation tools.

A third new challenge is posed by the need to update and merge annotation data sets. RNA-sequencing data (RNA-seq data)[5–8] provide an obvious means for updating older annotation data sets; however, doing so is not trivial. It is also not straightforward to ascertain whether the result improves on the original annotation. Furthermore, it is not unusual today for multiple groups to annotate the same genome using different annotation procedures. Merging these to produce a consensus annotation data set is a complex task.

Finally, the demographics of genome annotation projects are changing as well. Unlike the massive genome projects of the past, today's genome annotation projects are usually smaller-scale affairs and often involve researchers who have little bioinformatics and computational biology expertise. Eukaryotic genome annotation is not a point-and-click process; however, with some basic UNIX skills, 'do-it-yourself' genome annotation projects are quite feasible using present-day tools. Here we provide an overview of the eukaryotic genome annotation process, describe the available toolsets and outline some best-practice approaches.

Assembly and annotation: an overview

Assembly. The first step towards the successful annotation of any genome is determining whether its assembly is ready for annotation. Several summary statistics are used to describe the completeness and contiguity of a genome assembly, and by far the most important is N50 (BOX 1). Other useful assembly statistics are the average gap size of a scaffold and the average number of gaps per scaffold (BOX 1). Most current genomes are 'standard draft' assemblies, meaning that they meet minimum standards for submission to public databases[9]. However, a 'high-quality draft' assembly[9] is a much better target for annotation, as it is at least 90% complete.

Although there are no strict rules, an assembly with an N50 scaffold length that is gene-sized is a decent target for annotation. The reason is simple: if the scaffold N50 is around the median gene length, then ~50% of the genes will be contained on a single scaffold; these complete genes, together with fragments from the rest of the genome, will provide a sizable resource for downstream analyses[10,11]. As can be seen in FIG. 1, median gene lengths are roughly proportional to genome size. Thus, if the size of the genome of interest is known, it is possible to use this figure to obtain a rough estimate of gene lengths and hence to obtain an estimate of the minimum N50 scaffold length for annotation. CEGMA[12] provides another,

N50
A basic statistic for describing the contiguity of a genome assembly. The longer the N50 is, the better the assembly is. See box 1 for details.

Long interspersed nuclear elements
(LINEs). Retrotransposons that encode reverse transcriptase and that make up a substantial fraction of many eukaryotic genomes.

Short interspersed nuclear elements
(SINEs). Retrotransposons that do no encode reverse transcriptase and that parasitize LINE elements. *ALU* elements, which are very common in the human genome, are one example of a SINE.

complementary means of estimating the completeness and contiguity of an assembly. This tool screens an assembly against a collection of more or less universal eukaryotic single-copy genes and also determines the percentage of each gene lying on a single scaffold.

Obtaining a high-quality draft assembly is an achievable goal for most genome projects. If an assembly is incomplete or if its N50 scaffold length is too short, we would recommend doing additional shotgun sequencing, as tools are available for the incremental improvement of draft assemblies[13–15].

Annotation. Although genome annotation pipelines differ in their details, they share a core set of features. Generally, genome-wide annotation of gene structures is divided into two distinct phases. In the first phase, the 'computation' phase, expressed sequence tags (ESTs), proteins, and so on, are aligned to the genome and *ab initio* and/or evidence-driven gene predictions are generated. In the second phase, the 'annotation' phase, these data are synthesized into gene annotations (BOX 2). Because this process is intrinsically complicated and involves so many different tools, the programs that assemble compute data (evidence) and use it to create genome annotations are generally referred to as annotation pipelines. Current pipelines are focused on the annotation of protein-coding genes, although Ensembl also has some capabilities for annotating non-coding RNAs (ncRNAs). Tools for annotation of ncRNAs are described in BOX 3.

Box 1 | Common statistics for describing genome assemblies

Genome assemblies are composed of scaffolds and contigs. Contigs are contiguous consensus sequences that are derived from collections of overlapping reads. Scaffolds are ordered and orientated sets of contigs that are linked to one another by mate pairs of sequencing reads.

Scaffold and contig N50s

By far the most widely used statistics for describing the quality of a genome assembly are its scaffold and contig N50s. A contig N50 is calculated by first ordering every contig by length from longest to shortest. Next, starting from the longest contig, the lengths of each contig are summed, until this running sum equals one-half of the total length of all contigs in the assembly. The contig N50 of the assembly is the length of the shortest contig in this list. The scaffold N50 is calculated in the same fashion but uses scaffolds rather than contigs. The longer the scaffold N50 is, the better the assembly is. However, it is important to keep in mind that a poor assembly that has forced unrelated reads and contigs into scaffolds can have an erroneously large N50. Note too that scaffolds and contigs that comprise only a single read or read pair — often termed 'singletons' — are frequently excluded from these calculations, as are contigs and scaffolds that are shorter than ~800 bp. The procedures used to calculate N50 may therefore vary between genome projects.

Percent gaps

Another important assembly statistic is its percent gaps. Unsequenced regions between mate pairs in contigs and between scaffolds are often represented as runs of 'N's in the final assembly. Thus two assemblies can have identical scaffold N50s but can still differ in their percent gaps: one has very few gaps, and the other is heavily peppered with them. Estimates of gap lengths are often made based on library insert sizes and read lengths; when these are available, the number of 'N's in these gaps usually, but not always, represents the most likely estimate of that gap's size; sometimes, all gaps are simply represented by a run of 50 'N's regardless of their size.

Percent coverage

Percent coverage is used in two senses: genome coverage and gene coverage. The first number, genome coverage, refers to the percentage of the genome that is contained in the assembly based on size estimates; these are usually based on cytological techniques[116,117]. Genome coverage of 90–95% is generally considered to be good, as most genomes contain a considerable fraction of repetitive regions that are difficult to sequence. So it is not a cause for concern if the genome coverage of an assembly is a bit less than 100%. Gene coverage is the percentage of the genes in the genome that are contained in the assembly. Gene and genome coverage can differ from one another, as hard-to-assemble repetitive regions are often gene-poor. As a result, the percentage gene coverage is often substantially larger than the percentage genome coverage for some difficult-to-assemble genomes.

Step one: the computation phase

Repeat identification. Repeat identification and masking is usually the first step in the computation phase of genome annotation. Somewhat confusingly, the term 'repeat' is used to describe two different types of sequences: 'low-complexity' sequences, such as homopolymeric runs of nucleotides, as well as transposable (mobile) elements, such as viruses, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)[16,17]. Eukaryotic genomes can be very repeat rich; for example, 47% of the human genome is thought to consist of repeats[18], and this number is likely to be the lower limit. Also, the borders of these repeats are usually ill-defined; repeats often insert within other repeats, and often only fragments within fragments are present — complete elements are found quite rarely. Repeats complicate genome annotation. They need to be identified and annotated, but the tools used to identify repeats are distinct from those used to identify the genes of the host genome.

Identifying repeats is complicated by the fact that repeats are often poorly conserved; thus, accurate repeat detection usually requires users to create a repeat library for their genome of interest. Available tools for doing so generally fall into two classes: homology-based tools[19–21] and *de novo* tools[22–25] (for an overview, see REFS 26,27). Note, however, that *de novo* tools identify repeated sequences — not just mobile elements — so their outputs can include highly conserved protein-coding genes, such as histones and tubulins, as well as transposon sequences. Users must therefore carefully post-process the outputs of these tools to remove protein-coding sequences. These same outputs probably also contain some novel repeat families. Repeats are interesting in and of themselves, and the life cycles and phylogenetic histories of these elements are growing areas of research[17,28,29]. Adequate repeat annotation should thus be a part of every genome annotation project.

After it has been created, a repeat library can be used in conjunction with a tool such as RepeatMasker[30], which uses BLAST[31–33] and Crossmatch[34] to identify stretches of sequence in a target genome that are homologous to known repeats. The term 'masking' simply means transforming every nucleotide identified as a repeat to an 'N' or, in some cases, to a lower case a, t, g or c — the latter process is known as 'soft masking'[32,35]. The masking step signals to downstream sequence alignment and gene prediction tools that these regions are repeats. Failure to mask genome

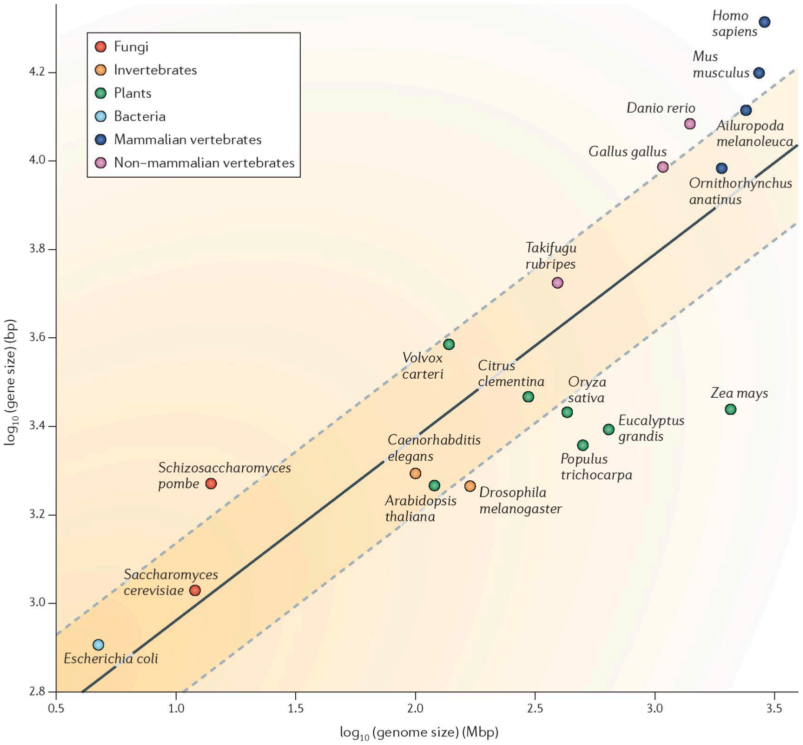


Figure 1 | **Genome and gene sizes for a representative set of genomes.** Gene size is plotted as a function of genome size for some representative bacteria, fungi, plants and animals. This figure illustrates a simple rule of thumb: in general, bigger genomes have bigger genes. Thus, accurate annotation of a larger genome requires a more contiguous genome assembly in order to avoid splitting genes across scaffolds. Note too that although the human and mouse genomes deviate from the simple linear model shown here, the trend still holds. Their unusually large genes are likely to be a consequence of the mature status of their annotations, which are much more complete as regards annotation of alternatively spliced transcripts and untranslated regions than those of most other genomes.

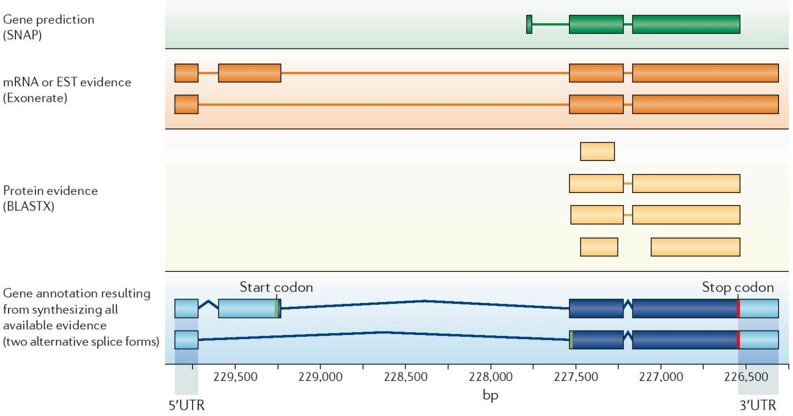
sequences can be catastrophic. Left unmasked, repeats can seed millions of spurious BLAST alignments[32], producing false evidence for gene annotations. Worse still, many transposon open reading frames (ORFs) look like true host genes to gene predictors, causing portions of transposon ORFs to be added as additional exons to gene predictions, completely corrupting the final gene annotations. Good repeat masking is thus crucial for the accurate annotation of protein-coding genes.

Evidence alignment. After repeat masking, most pipelines align proteins, ESTs and RNA-seq data to the genome assembly. These sequences include previously identified transcripts and proteins from the organism whose genome is being annotated. Sequences from other organisms are also included; generally, these are restricted to proteins, as these retain substantial sequence similarity over much greater spans of evolutionary time than nucleotide sequences do. In principle, TBLASTX[31,32,36] can be used to align ESTs and RNA-seq data from phylogenetically distant organisms but, owing to high computational costs, this is only done rarely.

UniProtKB/SwissProt[37–39] is an excellent core resource for protein sequences. As SwissProt is restricted to highly curated proteins, many users might want to supplement this database with the proteomes of related, previously annotated genomes. One easy way to assemble additional protein and EST data sets is to download sequences from related organisms using the NCBI taxonomy browser[40,41].

EST and protein sequence data sets are often aligned to the genome in a two-tiered process. Frequently, BLAST[31,32,36] and BLAT[42] are used to

Box 2 | Gene prediction versus gene annotation


Gene prediction (SNAP)
mRNA or EST evidence (Exonerate)
Protein evidence (BLASTX)
Gene annotation resulting from synthesizing all available evidence (two alternative splice forms)
Start codon
Stop codon
229,500
229,000
228,500
228,000
227,500
227,000
226,500
bp
5′UTR
3′UTR

Although the terms 'gene prediction' and 'gene annotation' are often used as if they are synonyms, they are not. With a few exceptions, gene predictors find the single most likely coding sequence (CDS) of a gene and do not report untranslated regions (UTRs) or alternatively spliced variants. Gene prediction is therefore a somewhat misleading term. A more accurate description might be 'canonical CDS prediction'.

Gene annotations, conversely, generally include UTRs, alternative splice isoforms and have attributes such as evidence trails. The figure shows a genome annotation and its associated evidence. Terms in parentheses are the names of commonly used software tools for assembling particular types of evidence. Note that the gene annotation (shown in blue) captures both alternatively spliced forms and the 5′ and 3′UTRs suggested by the evidence. By contrast, the gene prediction that is generated by SNAP (shown in green) is incorrect as regards the gene's 5′ exons and start-of-translation site and, like most gene-predictors, it predicts only a single transcript with no UTR.

Gene annotation is thus a more complex task than gene prediction. A pipeline for genome annotation must not only deal with heterogeneous types of evidence in the form of the expressed sequence tags (ESTs), RNA-seq data, protein homologies and gene predictions, but it must also synthesize all of these data into coherent gene models and produce an output that describes its results in sufficient detail for these outputs to become suitable inputs to genome browsers and annotation databases.

identify approximate regions of homology rapidly. These alignments are usually filtered to identify and to remove marginal alignments on the basis of metrics such as percent similarity or percent identity. After filtering, the remaining data are sometimes clustered to identify overlapping alignments and predictions. Clustering has two purposes. First, it groups diverse computational results into a single cluster of data, all supporting the same gene. Second, it identifies and purges redundant evidence; highly expressed genes, for example, may be supported by hundreds if not thousands of identical ESTs.

The term 'polishing' is sometimes used to describe the next phase of the alignment process. After clustering, highly similar sequences identified by BLAST and BLAT are realigned to the target genome in order to obtain greater precision at exon boundaries. BLAST, for example, although rapid, has no model for splice sites, and so the edges of its sequence alignments are only rough approximations of exon boundaries[43]. For this reason, splice-site-aware alignment algorithms, such as Splign[44], Spidey[45], sim4 (REF. 46) and Exonerate[43], are often used to realign matching and highly similar ESTs, mRNAs and proteins to the genomic input sequence. Although these programs take longer to run, they provide the annotation pipeline with much improved information about splice sites and exon boundaries.

Of all forms of evidence, RNA-seq data have the greatest potential to improve the accuracy of gene annotations, as these data provide copious evidence for better delimitation of exons, splice sites and alternatively spliced exons. However, these data can be difficult to use because of their large size and complexity. The use of RNA-seq data currently lies at the cutting edge of genome annotation, and the available toolset is evolving quickly[47]. Currently, RNA-seq reads are usually handled in two ways. They can be assembled *de novo* — that is, independently of the genome — using tools such as ABySS[48], SOAPdenovo[49] and Trinity[50]; the resulting transcripts are then realigned to the genome in the same way as ESTs. Alternatively, the RNA-seq data can be directly aligned to the genome using tools such as TopHat[51], GSNAP[52] or Scripture[53] followed by the assembly of alignments (rather than reads) into transcripts using tools such as Cufflinks[54]. See REF. 55 for guidance on the best way to use TopHat with Cufflinks.

Percent similarity
The percent similarity of a sequence alignment refers to the percentage of positive scoring aligned bases or amino acids in a nucleotide or protein alignment, respectively. The term positive scoring refers to the score assigned to the paired nucleotides or amino acids by the scoring matrix that is used to align the sequences.

Percent identity
The percent identity of a sequence alignment refers to the percentage of identical aligned bases or amino acids in a nucleotide or protein alignment, respectively.

Box 3 | Non-coding RNAs

Non-coding RNA (ncRNA) annotation is still in its infancy compared with protein-coding gene annotation, but it is advancing rapidly. The heterogeneity and poorly conserved nature of many ncRNA genes present major challenges for annotation pipelines. Unlike protein-encoding genes, ncRNAs are usually not well-conserved at the primary sequence level; even when they are, nucleotide homologies are not as easily detected as protein homologies, which limits the power of evidence-based approaches.

One common approach is to identify ncRNA genes using conserved secondary structures and motifs. Established examples of these types of tools include tRNAscan-SE[118] and Snoscan[119]. MicroRNA (miRNA) gene finders are also available[120]. A more general approach is first to align nucleotide sequences — genomic, RNA-seq and ESTs — from closely related organisms to the target genome and then search these for signs of conserved secondary structures. This is a complex process, however, and can require substantial computational resources; qRNA is one such tool[121], another is StemLoc[122]. Be aware that these tools have high false-positive rates. RNA sequencing is also greatly aiding ncRNA identification. For example, miRNAs can be directly identified using specialized RNA preps and sequencing protocols[123,124]. Even with such sophisticated tools and techniques, distinguishing between bona fide ncRNA genes, spurious transcription and poorly conserved protein-encoding genes that produce small peptides remains difficult, especially in the cases of long intergenic non-coding RNAs (lincRNAs)[125,126] and expressed pseudogenes[127,128].

Another approach is to annotate possible ncRNA genes liberally and then use Infernal[129] and Rfam[114] to triage and classify these genes based on primary and secondary sequence similarities. Even with these resources, however, many ncRNAs will remain unclassifiable. Currently, ncRNA annotation is cutting edge, and those using ncRNA annotations should bear in mind that ncRNA annotation accuracies are generally much lower than those of their protein-coding counterparts.

Opinions differ as to the best approach for using RNA-seq data, and the most promising avenue will probably heavily depend on both genome biology (for example, gene density) and the contiguity and completeness of the genome assembly. Gene density is an important consideration. If genes are closely spaced in the genome, then tools such as Cufflinks[54] sometimes erroneously merge RNA-seq reads from neighbouring genes. In such cases, *de novo* assembly of the RNA-seq data mitigates the problem; in fact, Trinity[50] is designed to deal with this issue. Several annotation pipelines are now compatible with RNA-seq data: these include PASA[56], which uses inchworm[50] outputs, and MAKER[10], which can operate directly from Cufflinks[54] outputs or can use preassembled RNA-seq data.

***Ab initio* gene prediction.** When gene predictors[57–60] first became available in the 1990s (see REF. 61 for an overview), they revolutionized genome analyses because they provided a fast and easy means to identify genes in assembled DNA sequences. These tools are often referred to as *ab initio* gene predictors because they use mathematical models rather than external evidence (such as EST and protein alignments) to identify genes and to determine their intron–exon structures.

The great advantage of *ab initio* gene predictors for annotation is that, in principle, they need no external evidence to identify a gene or to determine its intron–exon structure. However, these tools have practical limitations from an annotation perspective. For instance, most gene predictors find the single most likely coding sequence (CDS) and do not report untranslated regions (UTRs) or alternatively spliced transcripts (BOX 2). Training is also an issue; *ab initio* gene predictors use organism-specific genomic traits, such as codon frequencies and distributions of intron–exon lengths, to distinguish genes from intergenic regions and to determine intron–exon structures. Most gene predictors come with precalculated parameter files that contain such information for a few classic genomes, such as *Caenorhabditis elegans*, *D. melanogaster*, *Arabidopsis thaliana*, humans and mice. However, unless your genome is very closely related to an organism for which precompiled parameter files are available, the gene predictor needs to be trained on the genome that is under study, as even closely related organisms can differ with respect to intron lengths, codon usage and GC content[62].

Given enough training data, the gene-level sensitivity of *ab initio* tools can approach 100%[63,64] (BOX 4). However, the accuracy of the predicted intron–exon structures is usually much lower, ~60–70%. It is also important to understand that large numbers of pre-existing, high-quality gene models and near base-perfect genome assemblies are usually required to produce highly accurate gene predictions[63,65]; such data sets are rarely available for newly sequenced genomes.

In principle, alignments of ESTs, RNA-seq and protein sequences to a genome can be used to train gene predictors even in the absence of pre-existing reference gene models. Although many popular gene predictors can be trained in this way, doing so often requires the user to have some basic programming skills. The MAKER pipeline provides a simplified process for training the predictors Augustus[66,67] and SNAP[62] using the EST, protein and mRNA-seq alignments that MAKER has produced[10,56]. An alternative is to use GeneMark-ES[68,69]: a self-training, but sometimes less-accurate, algorithm[69,70].

Evidence-driven gene prediction. In recent years, the distinction between *ab initio* prediction and gene annotation has been blurred. Many *ab initio* tools, such as TwinScan[71], FGENESH[72], Augustus, Gnomon[73], GAZE[74] and SNAP, can use external evidence to improve the accuracy of their predictions. ESTs, for example, can be used to identify exon boundaries unambiguously. This process is often referred to as evidence-driven (in contrast to *ab initio*) gene prediction. Evidence-driven gene prediction has great potential to improve the quality of gene prediction in newly sequenced genomes, but in practice it can be difficult to use. ESTs and proteins must first be aligned to the genome; RNA-seq data must be aligned too, if they are available. Splice sites must then be identified, and the assembled evidence must be post-processed before a synopsis of these data can be passed to the gene finder. In practice, this is a lot of work, requiring a lot of specialized software. In fact, it is one of the main obstacles that genome annotation pipelines attempt to overcome.

Box 4 | How gene prediction and gene annotation accuracies are calculated

Three commonly used measures of gene-finder performance are sensitivity, specificity and accuracy[130]. Each is measured relative to some standard, usually a reference annotation. Sensitivity (SN) is the fraction of the reference feature that is predicted by the gene predictor. To be more precise, SN = TP / (TP + FN), where TP is true positives and FN is false negatives. By contrast, specificity (SP) is the fraction of the prediction overlapping the reference feature: for example, SP = TP / (TP + FP), where FP is false positives. Note that the definition of SP given here is the one that is commonly used by the gene-finding community[130] but, more correctly, this measure is positive predictive value (PPV) or precision.

Both measures can be calculated for any portion of a gene model, such as genes, transcripts or exons. At the nucleotide level, TP is the number of exonic nucleotides in the reference gene model, FN is the number of these that are not included in the prediction, and FP is the number of exonic nucleotides in the prediction that are not found in the reference gene model. At the exon level, SN is the number of correct exons in the prediction divided by the number of exons in the reference gene model, and SP is the number of correct exons in the prediction divided by the number of exons in the prediction[130]. So-called 'site measures' are also used: for example, the SN and SP for predicting features such as start codons or splice donors. SN and SP are often combined into a single measure called accuracy (AC): for example, AC = (SN + SP) / 2 (see REFS 130–132 for reviews of commonly used accuracy measures).

Panel **A** of the figure shows SN, SP and AC for two different gene models. The reference model is shown in blue, and the two different predictions at the same locus are shown in red. The table on the right gives the values of SN, SP and AC for the two predictions. For the purposes of calculation, exons 1, 2 and 3 of the reference gene model and of prediction 1 have identical start and end coordinates and are 100, 50 and 50 nucleotides long, respectively. In prediction 2, exons are 75 and 50 nucleotides long, respectively, and the start coordinate of its first exon is identical to that of the reference, but its end is not; its second exon is identical to the third exon in the reference.

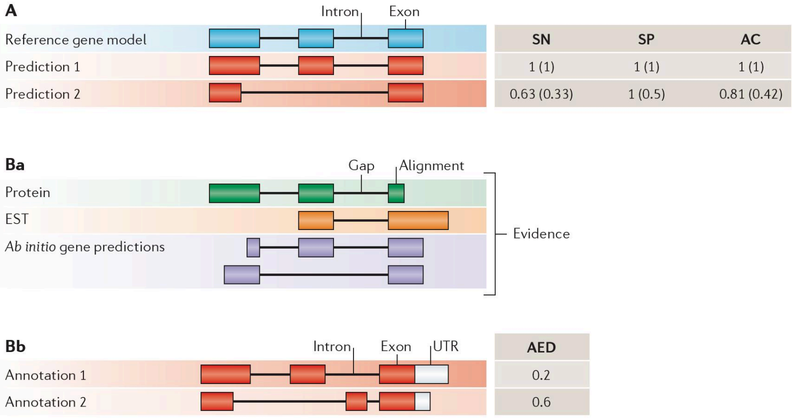
Numbers in parentheses are the values at the exon level; the others are nucleotide-level values. Note that the values for prediction 2 are lower at the exon level than they are at the nucleotide level. This is because exon-level calculations have an 'all-or-nothing' aspect to them: that is, a model in which the exons each differ by a single nucleotide from the reference will have nucleotide-level SN, SP and AC values near 1; its exon-level SN, SP and AC values, however, will all be 0.

With a few modifications, SN, SP and AC can also be used to compare two annotations to one another. This is the approach taken by the Sequence Ontology Project to calculate annotation edit distance (AED), which can be used to measure the congruence between an annotation and its supporting evidence[96]. AED is calculated in the same manner as SN and SP, but in place of a reference gene model, the coordinates of the union of the aligned evidence (see panel **Ba**) are used instead: AED = 1 – AC, where AC = (SN + SP) / 2.

An AED of 0 indicates that the annotation is in perfect agreement with its evidence, whereas an AED of 1 indicates a complete lack of evidence support for the annotation. More information regarding AED can be found in REF. 96.

Panel **B** illustrates how AED is used. Panel **Ba** shows the protein, expressed sequence tag (EST) and *ab initio* gene predictions that are produced during the computation phase of the annotation process. Panel **Bb** shows two hypothetical annotations based on this evidence. Solid portions of boxes in panel **Bb** delimit coding sequence; note that the two annotations differ at their 3′ untranslated regions (UTRs) as well as their coding-exon coordinates.

The table on the right in panel **Bb** shows how nucleotide-level AED values can be used to summarize the goodness of fit of an annotation to its overlapping evidence. Annotation 1 has the lower AED (of 0.2), meaning that it is a better fit to the evidence than annotation 2 (with an AED of 0.6) is; thus, bringing annotation 1 into perfect synchrony with the evidence would require fewer manual editing operations than would be required for annotation 2.



	SN	SP	AC
Prediction 1	1 (1)	1 (1)	1 (1)
Prediction 2	0.63 (0.33)	1 (0.5)	0.81 (0.42)

	AED
Annotation 1	0.2
Annotation 2	0.6

Step two: the annotation phase

The ultimate goal of annotation efforts is to obtain a synthesis of alignment-based evidence with *ab initio* gene predictions to obtain a final set of gene annotations. Traditionally, this was done manually; human genome annotators would review the evidence for each gene in order to decide on their intron–exon structures[75]. Although this results in high-quality annotation[76,77], it is so labour-intensive that, for budgetary reasons, smaller genome projects are increasingly being forced to rely on automated annotations.

There are almost as many strategies for creating automated annotations as there are annotation pipelines, but the common theme is to use evidence to improve the accuracy of gene models, usually through some combination of pre- and post-processing of the gene predictions. FIGURE 2 and TABLE 1 provide an overview of some of the more commonly used approaches.

Automated annotation. The simplest form of automated annotation is to run a battery of different gene finders on the genome and then to use a 'chooser algorithm' (also known as a 'combiner') to select the single prediction whose intron–exon structure best represents the consensus of the models from among the overlapping predictions that define each putative gene locus. This is the process used by JIGSAW[78]. EVidenceModeler (EVM)[79] and GLEAN[80] (and its successor, Evigan[81]) go one step further, attempting to choose the best possible set of exons automatically and to combine them to produce annotations. This is done by estimating the types and frequencies of errors that are made by each source of gene evidence and then choosing combinations of evidence that minimize such errors. Like *ab initio* gene predictors, JIGSAW must be retrained for each new genome, and so it requires a source of known gene models that were not already used to train the underlying *ab initio* gene predictors. EVM allows the user to set expected evidence error rates manually or to learn them from a training set. By contrast, GLEAN and Evigan use an unsupervised learning method to estimate a joint error model, and thus they require no additional training. In a recent gene prediction competition[64], the combiners nearly always improved on the underlying gene prediction models, and JIGSAW, EVM or Evigan performed similarly.

Another popular approach is to feed the alignment evidence to the gene predictors at run time (that is, evidence-driven prediction) to improve the accuracy of the prediction process — a chooser can then be used to identify the most representative prediction. The predictions can also be processed — before or after running the chooser — to attain still greater accuracies by having the annotation pipeline add UTRs as suggested by the RNA-seq and EST data. This is the process used by PASA[56,82], Gnomon[73] and MAKER[10]. The evidence can also be used to inform the choices made by the chooser algorithm — by picking the post-processed gene model that is most consistent with the protein, EST and RNA-seq alignments[83]; EVM, MAKER and PASA all provide methods for doing so (TABLE 1; FIG. 2).

So which approach should you use? Probably the best way to think about the problem is in terms of effort versus accuracy. Simply running a single *ab initio* gene finder over even a very large genome can be done in a few hours of central processing unit (CPU) time. By contrast, a full run by an annotation pipeline such as MAKER or PASA can take weeks, but because these pipelines align evidence to the genome, their outputs provide starting points for annotation curation and downstream analyses, such as differential expression analyses using RNA-seq data. Another factor to consider is the phylogenetic relationship of the study genome to other annotated genomes. If it is the first of its taxonomic order or family to be annotated, it would definitely be preferable to use a pipeline that can use the full repertory of external evidence, especially RNA-seq data, to inform its gene annotations; not doing so will almost certainly result in low-quality annotations[80].

Visualizing the annotation data

Output data: the importance of using a fully documented format. The outputs of a genome annotation pipeline will include the transcript and protein sequences of every annotation, which are almost always provided in FASTA format[84]. Although FASTA files are useful, they only enable a small subset of possible downstream analyses. Visualizing annotations in a genome browser and creating a genome database requires a more descriptive output file. At a bare minimum, output files need to describe the intron–exon structures of each annotation, their start and stop codons, UTRs and alternative transcripts. Ideally, these outputs should go one step further and should include information about the sequence alignments and gene predictions that support each gene model.

Four commonly used formats for describing annotations are the GenBank, GFF3, GTF and EMBL formats. Using a fully documented format is important for three reasons. First, doing so will remove the trouble of writing software to convert outputs into a format that other tools can use. Second, common formats, especially those such as GenBank and GFF3, which use controlled vocabularies and ontologies to define their descriptive terminologies, guarantee 'interoperability' between analysis tools. Third, unless a common vocabulary is used to describe gene models[85], comparative genomic analyses can be frustratingly difficult or downright impossible. In response to these needs, the Generic Model Organism Database (GMOD) project community has developed a series of standards and tools for description, analyses, visualization and redistribution of genome annotations, all of which use the GFF3 file format as inputs and outputs. Leveraging GMOD tools and GFF3 substantially simplifies curation, analysis, publication and management of genome annotations.

GMOD. The GMOD project is an umbrella organization that provides a large suite of tools for creating, managing and using genome annotations, including the analysis, visualization and redistribution of annotation data. Users who have browsed the

Unsupervised learning methods

Refers to methods that can be trained using unlabelled data. One example is a gene prediction algorithm that can be trained without a reference set of correct gene models; instead, the algorithm is trained using a collection of annotations, not all of which might be correct.

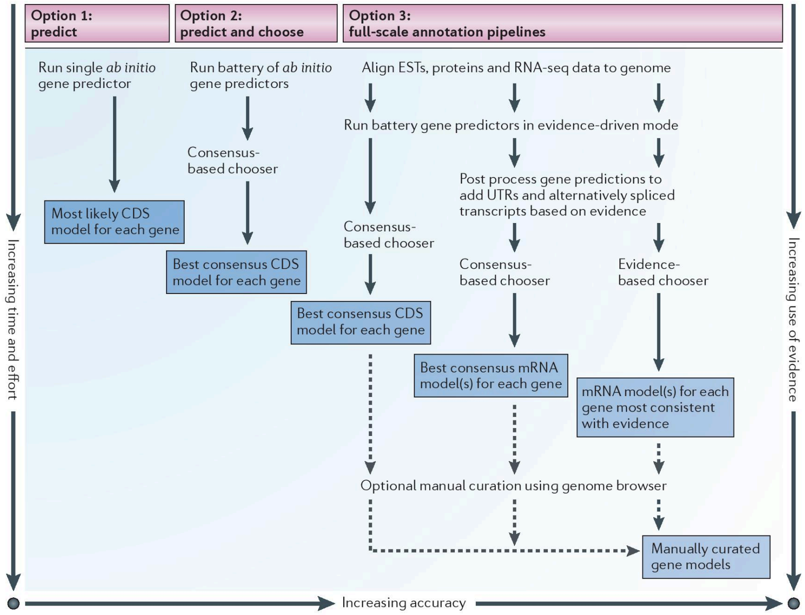

Option 1:
predict
Option 2:
predict and choose
Option 3:
full-scale annotation pipelines
Run single *ab initio* gene predictor
Most likely CDS model for each gene
Run battery of *ab initio* gene predictors
Consensus-based chooser
Best consensus CDS model for each gene
Align ESTs, proteins and RNA-seq data to genome
Run battery gene predictors in evidence-driven mode
Post process gene predictions to add UTRs and alternatively spliced transcripts based on evidence
Consensus-based chooser
Best consensus CDS model for each gene
Consensus-based chooser
Best consensus mRNA model(s) for each gene
Evidence-based chooser
mRNA model(s) for each gene most consistent with evidence
Optional manual curation using genome browser
Manually curated gene models
Increasing time and effort
Increasing use of evidence
Increasing accuracy

Figure 2 | **Three basic approaches to genome annotation and some common variations.** Approaches are compared on the basis of relative time, effort and the degree to which they rely on external evidence, as opposed to *ab initio* gene models. The y axis shows increasing time and effort; the x axis shows increasing use of external evidence and, consequently, increasing accuracy and completeness of the resulting gene models. The type of final product produced by each kind of pipeline is shown in the dark blue boxes. Relative positions in the figure are for summary purposes only and are not based on precisely computed values. See TABLE 1 for a list of commonly used software components. CDS, coding sequence; EST, expressed sequence tag; RNA-seq, RNA sequencing; UTR, untranslated region.

Saccharomyces Genome Database, WormBase, FlyBase, The *Arabidopsis* Information Resource (TAIR) or the University of California Santa Cruz (UCSC) Genome Browser will have used GMOD tools. GMOD tools also aid in creating an online genome database. The key is having annotations and their associated evidence in GFF3 format, which is useable by GMOD tools. Users can directly visualize these files using GBROWSE[86] and JBROWSE[87] to produce views of their data just like those offered at WormBase and UCSC. They can also directly edit the gene models using the Apollo genome browser and JBROWSE. BioPerl[88] also provides a set of database tools for loading GFF3 files into a ready-made Chado[89] database schema with which an online genome database can be rapidly created that contains a genome and its annotations in a 'browse-able' format.

Quality control

Incorrect annotations poison every experiment that makes use of them. Worse still, the poison spreads because incorrect annotations from one organism are often unknowingly used by other projects to help annotate their own genomes. Standard practices for genome annotation have been proposed for bacterial[90], viral[91] and eukaryotic genomes[92], but even when followed, quality control remains an issue. Even the best gene predictors and genome annotation pipelines rarely exceed accuracies of 80% at the exon level[63], meaning that most gene annotations contain at least one mis-annotated exon. Given these facts, assessing how accurately a genome is annotated is an important part of any project.

Over the years, there have been various contests aimed at assessing gene prediction accuracy[63,65] (BOX 4). These contests have played an important part in improving the power and accuracy of gene prediction. However, less progress has been made regarding genome annotations[64]. The heart of the problem is the absence of reference data sets with which to obtain accuracy estimates. The first generation of genome projects — *Saccharomyces cerevisae*, *C. elegans* and *D. melanogaster*, for instance — all had decades of work to draw on when training and measuring the accuracy of gene predictors and annotation pipelines. However, no such data set exists for most of the organisms being sequenced today. Moreover, just because a gene

Table 1 | **Five basic categories of annotation software and some selected examples**

Software	Description	Refs
Ab initio *and evidence-drivable gene predictors*		
Augustus	Accepts expressed sequence tag (EST)-based and protein-based evidence hints. Highly accurate	66,67
mGene	Support vector machine (SVM)-based discriminative gene predictor. Directly predicts 5′ and 3′ untranslated regions (UTRs) and poly(A) sites	133
SNAP	Accepts EST and protein-based evidence hints. Easily trained	62
FGENESH	Training files are constructed by SoftBerry and supplied to users	72
Geneid	First published in 1992 and revised in 2000. Accepts external hints from EST and protein-based evidence	134
Genemark	A self-training gene finder	69,70
Twinscan	Extension of the popular Genscan algorithm that can use homology between two genomes to guide gene prediction	71
GAZE	Highly configurable gene predictor	74
GenomeScan	Extension of the popular Genscan algorithm that can use BLASTX searches to guide gene prediction	135
Conrad	Discriminative gene predictor that uses conditional random fields (CRFs)	136
Contrast	Discriminative gene predictor that uses both SVMs and CRFs	137
CRAIG	Discriminative gene predictor that uses CRFs	138
Gnomon	Hidden Markov model (HMM) tool based on Genscan that uses EST and protein alignments to guide gene prediction	73
GeneSeqer	A tool for identifying potential exon–intron structure in precursor mRNAs (pre-mRNAs) by splice site prediction and spliced alignment	139
EST, protein and RNA-seq aligners and assemblers		
BLAST	Suite of rapid database search tools that uses Karlin–Altschul statistics	31–33
BLAT	Faster than BLAST but has fewer features	42
Splign	Splice-aware tool designed to align cDNA to genomic sequence	44
Spidey	mRNA-to-DNA alignment tool that is designed to account for possible paralogous alignments	45
Prosplign	Global alignment tool that uses BLAST hits to align in a splice-site- and paralogy-aware manner	140
sim4	Splice-aware cDNA-to-DNA alignment tool	46
Exonerate	Splice-site-aware alignment algorithm that can align both protein and EST sequences to a genome	43
Cufflinks	Extension to TopHat. Uses TopHat outputs to create transcript models	54
Trinity	High-quality *de novo* transcriptome assembler	50
MapSplice	Spliced aligner that does not use a model of canonical splice junction	141
TopHat	Transcriptome aligner that aligns RNA sequencing (RNA-seq) reads to a reference genome using Bowtie to identify splice sites	51
GSNAP	A fast short-read assembler	52
Choosers and combiners		
JIGSAW	Combines evidence from alignment and *ab initio* gene prediction tools to produce a consensus gene model	78
EVidenceModeler	Produces a consensus gene model by combining evidence from protein and transcript alignments together with *ab initio* predictions using weights for both abundance and the sources of the evidence	79
GLEAN	Tool for creating consensus gene lists by integrating gene evidence through latent class analysis	80
Evigan	Probabilistic evidence combiner that use a Bayeisan network to weigh and integrate evidence from *ab initio* predictors, alignments and expression data to produce a consensus gene model	81

Table 1 (cont.) | **Five basic categories of annotation software and some selected examples**

Software	Description	Refs
Genome annotation pipelines		
PASA	Annotation pipeline that aligns EST and protein sequences to the genome and produces evidence-driven consensus gene models	56,82
MAKER	Annotation pipeline that uses BLAST and exonerate to align protein and EST sequences. Also accepts features from RNA-seq alignment tools (such as TopHat). Massively parallel	10,83
NCBI	The genome annotation pipeline from the US National Center for Biotechnology Information (NCBI). Uses BLAST alignments together with predictions from Gnomon and GenomeScan to produce gene models	142
Ensembl	Ensembl's genome annotation pipeline. Uses species-specific and cross-species alignments to build gene models. Also annotates non-coding RNAs	107
Genome browsers for curation		
Artemis	Java-based genome browser for feature viewing and annotation. Can use binary alignment map (BAM) files as input	99
Apollo	Java-based genome browser that allows the user to create and edit gene models and write their edits to a remote database	97
JBROWSE	JavaScript- and HTML-based genome browser that can be embedded into wikis for community work. Excellent for Web-based use	87
IGV	Genome browser that supports BAM files and expression data	143

These tools are widely used both as standalone applications and as modular components of genome annotation pipelines. See FIG. 2 for a schematic of the roles of each class of tool in genome annotation.

predictor does well on one genome is no guarantee of a good performance on the next[83]. Assessing annotation quality in the absence of reference genome annotations is a difficult problem. Experimental verification is one solution, but few projects have the resources to carry this out on a large scale.

Approaches for assessing annotation quality. One simple approach for obtaining a rough indication of annotation quality is to quantify the percentage of annotations that encode proteins with known domains using tools such as InterProScan[93] and Pfam[94] or tools such as MAKER, which provides an automated means for carrying out such analyses[83]. Although relative numbers of domains vary between organisms and the expansion and contraction of particular gene families have a well-established role in organismal evolution, among the eukaryotes, the overall percentage of proteins that encode a domain of any sort is reasonably constant[83]. The domain content of the human, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae* proteomes varies between 57% and 75%[95]. Poorly trained gene finders do not perform nearly this well — 5% to 25% is typical. Thus, a eukaryotic proteome with a low percentage of domains is a warning sign that it could be poorly annotated[83].

Although domain content provides a rough estimate of overall annotation quality, it provides little guidance when trying to judge the accuracy of a given annotation. One approach towards solving this problem is to ask whether the protein, EST and RNA-seq evidence support or contradict the annotated intron–exon structure of the gene. This is fairly straightforward to assess by eye, but performing this task in an automated fashion requires a computable metric. In response, the Sequence Ontology Project[85] has developed several metrics for quality control of genome annotations[96]. Annotation edit distance (AED), for example, measures how congruent each annotation is with its overlapping evidence (BOX 4). AED thus provides a means to identify problematic annotations automatically and to prioritize them for manual curation. AED scores can also be used to measure changes to annotations between annotation runs. The MAKER2 genome annotation pipeline[83] provides some useful tools for automatically calculating AED.

Of course, identifying inaccurate annotations is only half of the problem; errors also need to be corrected. The most direct approach to fixing an erroneous annotation is to edit its intron–exon coordinates manually. The Apollo[97], Argo[98] and Artemis[99] browsers are widely used for this purpose. Gene models can be graphically revised using a series of 'drag-and-drops' and mouse clicks, and the resulting edits are written back to either files or to a remote database connection[89].

Annotation jamborees. Many genome projects choose to manually review and edit their annotation data sets. Although this process is time- and resource-intensive, it provides opportunities for community building, education and training.

Annotation jamborees (a term that was coined by the *D. melanogaster* community to describe the first such gathering[100]) provide a ready means for manual curation and analysis of the data and for putting together a genome paper. The key to hosting a successful jamboree is infrastructure. At a minimum, attendees must be able to search the annotated proteins and transcripts and to view the annotations in a genome browser. Searches can easily be handled by setting up a

BLAST database server coupled with a graphical user interface (GUI) such as a Web browser. The WWW BLAST server package[101] provides an easy means to do so. GBrowse[86,102] and JBrowse[87] can also easily be configured to allow remote users to view the annotated genome, as can the Apollo genome browser, which also provides a means to edit incorrect annotations. As all of these resources can be set up and configured remotely, it is now possible to support a distributed jamboree, in which the community collaborates via the Internet. This model recently proved to be successful for the ant genome community, which organized a distributed jamboree in which investigators and students collaborated to curate and analyse three different ant genomes quickly, all in a distributed manner[103–106].

Making data publicly available

Successful genome annotation projects do not just end with the publication of a paper; they also produce publicly available annotations. Genome annotations fuel the bench work and computational analyses that constitute the day-to-day operations of molecular biology and bioinformatics laboratories worldwide. They also provide an essential resource for other genome annotation projects; the transcripts and proteins produced by one annotation project will probably be used to help annotate other genomes. There are three basic routes to making annotations publicly available: you can build your own genome database and place it online; you can submit your annotations to GenBank and Ensembl; or you can submit them to any of a growing number of theme-based genome databases. We recommend taking all three routes.

Submitting annotations to public databases. One way to make annotations publicly available is to submit them to GenBank. Parties working on vertebrate genomes are also encouraged to contact Ensembl, which continues to incorporate new species at the rate of 5–10 per year in order to create a comprehensive annotation resource for vertebrate genomes, all of which are annotated by its gene build pipeline[107]. Both GenBank and Ensembl have much to offer to smaller genome projects, including powerful data-marts that allow users to browse and download data. Ensembl and GenBank also automatically handle the heavy lifting that is involved in relating gene models to those of other organisms and identifying homologues, paralogues and orthologues. They also provide an easy means to search and browse data; in short, they integrate a data set into the larger landscape of genomics and genome annotations. Best of all, the entire process is free, and submission to these sites in no way abridges the rights of the generators of the data to host and maintain their own genome database. For the research communities of most organisms, members will prefer to visit the specialized genome database for that organism, whereas the larger biological community will tend to access the data through GenBank and Ensembl. In addition to these large sites, intermediate-sized projects that host, manage and maintain sets of annotated genomes that are all related by a common theme are gaining in popularity. Examples include BeeBase[103], Gramene[108], PlantGDB[109], Phytozome[110] and VectorBase[111].

Data-mart
Provides users with online access to the contents of a data warehouse through user-configurable queries. A data-mart allows users to download data that meet their particular needs: for example, all transcripts from all annotated genes on human chromosome 3.

Updating annotations. Many genomes were annotated so long ago that the existing annotations could be dramatically improved using modern tools and data sets such as RNA-seq. In many cases, improved assemblies are possible as well. The question then becomes how to merge, update and improve the existing annotations and, at the same time, to document the process. Like annotation quality control, this is a thorny problem that until recently has garnered little attention, and few published tools yet exist to automate the process. Among existing tools, GLEAN and PASA can be used to report differences between pre-existing gene models and newly created ones. Ensembl has a procedure to merge annotation data sets to produce a consensus, and PASA has one for updating annotations with RNA-seq data. The MAKER annotation pipeline provides an automated toolkit with all of these functionalities and can revise, update and merge existing annotation data sets, as well as map them forwards to new assemblies[10,83].

GenBank provides two avenues for redistributing the results of updates and re-annotation of genomes. If the group that is updating the annotations includes the original authors, the update can simply be submitted; if not, there are two routes for submission. If the work involves substantial improvements to the original assembly, the parties producing them can submit the new annotations to GenBank as primary authors; if not — that is, if the revisions merely improve the original annotations — those producing them can submit their work through the third party submission channel. Ensembl also allows submission of such data, although the process is less formal, and interested parties should contact Ensembl directly.

Conclusions

In some ways, cheap sequencing has complicated genome annotation. As we have explained, the fragmented assemblies and exotic nature of many of the current genome-sequencing projects are part of the reason that this is so, but it is the ever-widening scope of annotation that is presenting the greatest challenges. Genome annotation has moved beyond merely identifying protein-coding genes to include an ever-greater emphasis on the annotation of transposons, regulatory regions, pseudogenes and ncRNA genes[112–115]. Annotation quality control and management are also increasingly becoming bottlenecks. As long as tools and sequencing technologies continue to develop, periodic updates to every genome's annotations will remain necessary. Those undertaking genome annotation projects need to reflect on this fact. Like parenthood, annotation responsibilities do not end with birth. Incorrect and incomplete annotations poison every experiment that makes use of them. In today's genomics-driven world, providing accurate and up-to-date annotations is simply a must.

1. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
2. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**, research0079 (2002).
3. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
4. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
5. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008).
6. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
7. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**, 621–628 (2008).
8. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
   This paper provides one of the most extensively documented surveys of alternatively spliced transcripts. It is a key publication for understanding how extensive alternative splicing is in human tissues, for understanding how powerful RNA-seq data are as a tool for discovering new transcripts and for quantifying their abundance and differential expression patterns.
9. Chain, P. S. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236–237 (2009).
10. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
11. Ye, L. *et al.* A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol.* **12**, R31 (2011).
12. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
13. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11**, R41 (2010).
14. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968–1969 (2009).
15. Husemann, P. & Stoye, J. r2cat: synteny plots and comparative assembly. *Bioinformatics* **26**, 570–571 (2010).
16. Kapitonov, V. V. & Jurka, J. A novel class of SINE elements derived from 5S rRNA. *Mol. Biol. Evol.* **20**, 694–702 (2003).
17. Kapitonov, V. V. & Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Rev. Genet.* **9**, 411–412; author reply 414 (2008).
18. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
19. Buisine, N., Quesneville, H. & Colot, V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics* **91**, 467–475 (2008).
20. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
21. McClure, M. A. *et al.* Automated characterization of potentially active retroid agents in the human genome. *Genomics* **85**, 512–523 (2005).
22. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
23. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (Suppl. 1), i351–i358 (2005).
24. Smit, A. & Hubley, R. RepeatModeler 1.05. *repeatmasker.org* [online], http://www.repeatmasker.org/RepeatModeler.html (2011).
25. Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
26. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Rev. Genet.* **13**, 36–46 (2012).
27. Bergman, C. M. & Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* **8**, 382–392 (2007).
28. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Rev. Genet.* **10**, 691–703 (2009).
29. Witherspoon, D. J. *et al.* Alu repeats increase local recombination rates. *BMC Genomics* **10**, 530 (2009).
30. Smit, A. F., Hubley, R. & Green, P. RepeatMasker 3.0 *repeatmasker.org* [online], http://www.repeatmasker.org/webrepeatmaskerhelp.html (1996–2010).
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
32. Korf, I., Yandell, M. & Bedell, J. *BLAST: an Essential Guide to the Basic Local Alignment Search Tool* 339 (O'Reilly & Associates, 2003).
   Everyone involved with a genome project should be familiar with BLAST. Reference 31 is the original paper describing this tool. Reference 32 is an entire book describing BLAST and how it is used.
33. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
34. Green, P. Crossmatch. A general purpose utility for comparing any two sets of DNA sequences. *PHRAP* [online], http://www.phrap.org/phredphrap/general.html (1993–1996).
35. Majoros, W. H. *Methods for Computational Gene Prediction* 2 (Cambridge Univ. Press, 2007).
36. Camacho, C. *et al.* BLAST + : architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
37. Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* **5**, 39–55 (2004).
38. Boeckmann, B. *et al.* Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C.R. Biol.* **328**, 882–899 (2005).
39. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39**, D214–D219 (2011).
40. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37**, D26–D31 (2009).
41. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15 (2009).
42. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
43. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
44. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3**, 20 (2008).
45. Wheelan, S. J., Church, D. M. & Ostell, J. M. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.* **11**, 1952–1957 (2001).
46. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974 (1998).
47. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**, 469–477 (2011).
48. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
49. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
50. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotech.* **29**, 644–652 (2011).
   This paper describes Trinity, a transcriptome assembler that was specifically designed for next-generation sequence data. It is required reading for anyone trying to use RNA-seq data for genome annotation.
51. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
52. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
53. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotech.* **28**, 503–510 (2010).
54. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech.* **28**, 511–515 (2010).
55. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protoc.* **7**, 562–578 (2012).
   This paper describes best practice approaches for combining TopHat and Cufflinks when using RNA-seq data for genome annotation.
56. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
57. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226**, 141–157 (1992).
58. Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 354–362 (1994).
59. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
   This study describes the *ab initio* gene predictor GenScan. It is a classic paper that is full of informative explanations of the problems associated with eukaryotic gene prediction.
60. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).
61. Brent, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15**, 1777–1786 (2005).
62. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
   This paper describes a gene predictor, SNAP, that is easy to use and to configure. It also clearly explains the pitfalls that are associated with using a poorly trained gene finder or one that has been trained on a different genome from the one that is being annotated.
63. Reese, M. G. & Guigo, R. EGASP: Introduction. *Genome Biol.* **7** (Suppl. 1), 1–3 (2006).
   This is the introduction to an entire issue of *Genome Biology* that is dedicated to benchmarking an entire host of eukaryotic gene finders and annotation pipelines. Anyone involved with a genome annotation project should have a look at every paper in this special supplement.
64. Coghlan, A. *et al.* nGASP—the nematode genome annotation assessment project. *BMC Bioinformatics* **9**, 549 (2008).
65. Guigo, R. & Reese, M. G. EGASP: collaboration through competition to find human genes. *Nature Methods* **2**, 575–577 (2005).
66. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** (Suppl. 2), ii215–ii225 (2003).
67. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
68. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
69. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
70. Zhu, W., Lomsadze, A. & Borodovsky, M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
71. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, S140–S148 (2001).
72. Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
73. Souvorov, A. *et al.* Gnomon — the NCBI eukaryotic gene prediction tool. *National Center for Biotechnology Information* [online], http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml (2010).

CHAPTER 3

A GENOMIC REGION IN THE FUSIFORM RUST PATHOGEN INTERACTS SPECIFICALLY WITH THE *PINUS TAEDA* L. FR1 RESISTANCE LOCUS

The following chapter is a manuscript coauthored by Katherine E. Smith, Amanda L. Pendleton, Thomas L. Kubisiak, Claire L. Anderson, Asaf Salamov, Andrea Aerts, Robert W. Riley, Alicia Clum, Erika A. Linquist, Michael Campbell, Zev Kronenberg, Nicolas Feau, Braham Dhillon, Richard C. Hamelin, Jason A. Smith, Mark Yandell, C. Dana Nelson, Igor V. Grigoriev, John M. Davis, and myself.

Katherine E. Smith and I contributed equally to this work. Michael Campbell and I prepared the genome annotation. Zev Kronenberg developed the tests for genomic selection. This manuscript will be submitted to *Fungal Genetics and Biology* in October or November 2016.

Abstract

Heteroecious rust pathogens colonize two or more plant hosts to complete their life cycles. Typically, interactions between rust pathogens and their telial (i.e., repeat) hosts have attracted much research attention given the economic importance of these hosts. Conversely, for oak-pine rusts, only minor symptoms are observed on telial hosts (*Quercus* spp.) whereas significant economic losses can occur on aecial hosts (*Pinus* spp.). We sequenced, assembled, and annotated the genome of the fusiform rust pathogen *Cronartium quercuum* f.sp. *fusiforme* (*Cqf*) which incites gall symptoms on the stems of commercially important species such as loblolly pine (*Pinus taeda* L.). To further refine and characterize the genomic region in *Cqf* that was previously shown by recombinational linkage mapping to specifically interact with the pine Fr1 gene, pycnial droplets were collected from *Fr1*/*fr1* (virulence selected) resistant hosts and *fr1*/*fr1* (non-selecting) susceptible hosts and were sequenced. Using this approach, we identified a selective sweep in the genomic region that was genetically mapped to *Avr1* in *Cqf*. To further our ongoing map-based cloning and genome finishing efforts, we aligned genetic marker sequences to scaffolds in order to place the *Cqf* genome within the context of the genetic map framework. These results suggest that bulk segregant sequencing should enable additional avirulence loci to be identified in *Cqf*, opening the door for the development of specific genetic marker assays to monitor *Cqf* virulence in the field, a technique that would allow for more informed predictions regarding the most resistant pine genotype to be planted across the range of loblolly pine.

Introduction

Fusiform rust disease is incited by the heteroecious, macrocyclic pathogen (*Cronartium quercuum* f.sp. *fusiforme* or *Cqf*; Fig. 3.1). In contrast to most rust-disease systems, disease symptoms on the oak (telial, repeating) host are quite minor and of little to no economic importance, and have therefore motivated little to no research to identify genetic resources of resistance in this host. However, genetic resistance on the pine (aecial, non-repeating) host is of great economic importance because the disease causes major damage to pine seedlings. Each year, ca. 1 billion seedlings of loblolly pine (*Pinus taeda* L.) are planted in the US, with the majority of those seedlings being genetically selected for improved resistance to fusiform rust disease via traditional breeding (1). Genetic resistance to fusiform rust in the aecial host is largely conditioned by major genes. This inference is based on a series of studies in which the no-gall phenotype was genetically mapped to single genetic loci in the host, after inoculation with single-aeciospore-derived spore population of *Cqf* (Fr1 (2); *Fr1-Fr9* (3)). Recently, a candidate gene for Fr1 was identified in the loblolly pine draft genome assembly (4). There is also evidence that each of the mapped Fr loci has been overcome, to varying degrees depending on the geographic origins of spore collections used to inoculate known *Fr1*/-resistant compared to *fr1*/*fr1* susceptible seedlings (5).

In the past decade, much progress has been made in discovering effectors, the gene products in rust fungi that enable resistance genes to be defeated, i.e., effect host manipulation to suppress host defenses, obtain nutrients, and generate disease symptoms (6–8). Despite remarkable progress made in understanding the structure and function of effectors in rusts, thus far experiment data for effectors on aecial hosts are lacking. Rust pathogen interactions with telial and aecial hosts are distinct with respect to the disease phenotype, the host organ in which the interaction occurs, the cellular structures formed, and length of coevolution (9, 10). This suggests that there may be distinct effectors and/or distinct targets for effectors in telial versus aecial hosts. Effectors often lack homologs in related species, potentially due to the rapid evolution of fungal pathogen genomes, which leads to taxon-specific gene models and thus limiting comparative analyses. In an analysis of gene family evolution in 15 basidiomycete fungi, ca. 20% of *Cqf* gene models with evidence of expression in the transcriptome were *Cqf*-specific (11). We chose to pursue genomic mapping of *Avr1*, since it was previously inferred to be a single locus based on linkage to markers on linkage group III in the genetic map of *Cqf* (12). Here we leveraged segregation ratio differences between aeciospore pools derived from resistant versus susceptible pines to compare the segregation of SNVs across the *Cqf* genome and identified candidate genes for *Avr1*.

Materials and Methods

The haploid genome of *Cronartium quercuum* f. sp. *fusiforme* was sequenced using the Illumina and 454 platforms. Genomic DNA was isolated from pycniospores collected from a single gall (G-11) on the Harrison Experimental Forest near Saucier, Mississippi. A panel of eight microsatellite markers (13) was used to confirm the haploid status (i.e., single genotype purity) of the sample prior to the library construction and sequencing. Two libraries, one with an insert size of 270 bp and one with a long mate-pair insert size of 27 kb, were sequenced using Illumina technology, generating 2 X 100 bp reads. A single 4kb paired-end library was sequenced using Roche 454 pyrosequencing technology. Reads were quality control filtered for artifacts and process contamination, then assembled with the short read genome assembler AllPathsLG release version R40582 (14). The genome was annotated using the JGI annotation pipeline (15), which combines several gene prediction and annotation methods (see below) and integrates the annotated genome into the web-based fungal resource MycoCosm (16) for comparative genomics.

Genome fragmentation was assessed with the Benchmarking Universal Single-Copy Orthologs (BUSCO) on the *Cqf* genome, as well as genome assemblies for two other rust fungi (*Melampsora tritic*, *Puccinia graminis*) obtained from MycoCosm (17).

Before gene prediction, assembly scaffolds were masked using RepeatMasker (18) with the current edition of the RepBase library (19), with the most frequent (observed >150 times) repeats recognized by RepeatScout (20). The following combination of gene predictors was run on the masked assembly: *ab initio* Fgenesh (21) and GeneMark (22) homology-based Fgenesh+ (21) and Genewise (23) seeded by blastx alignments against the NCBI non-redundant (NR) database; and transcriptome-based CombEST. In addition to protein coding genes, tRNAs were predicted using tRNAscan-SE (24). All predicted proteins were functionally annotated using SignalP (25, 26) for signal sequences, TMHMM (27) for transmembrane domains, InterProScan (28) for integrated collection of protein domains, and protein alignments to the NCBI NR database, SwissProt www.expasy.org/sprot/, KEGG (29) for metabolic pathways, and KOG (30) for eukaryotic clusters of orthologs. InterPro and SwissProt hits were used to map gene orthology (GO) terms (31). For each genomic locus, the best representative gene model was selected based on a combination of protein homology and EST support.

Secreted proteins (SPs) and small secreted proteins (SSPs) were annotated using a combination of signal peptide (SignalP 3.0 and 4.0) (25, 32) , trans-membrane domain (TMHMM 2.0) (33) and protein location (TargetP 1.1) (34) prediction algorithms. To compare closely related species, secreted protein predictions were carried out on the predicted proteomes of *Cqf*, *Melampsora larici-populina* and *Puccinia graminis* f.sp. *tritici*. SPs and SSPs were then assembled into clusters of homologs using the OrthoMCL algorithm (35) following an all-vs.-all blastp search (coverage and identity of at least 50%, e-value cutoff of 1e-05) with an inflation value set to 1.1.

Samples for RNA sequencing were collected from both the pine and the oak hosts of *Cqf*. Galls were collected from 5-year-old slash pine trees in October as pycniospores were forming and again in April as aeciospores were forming. Galls were freeze-dried for one week. Yellow-colored aeciospore (spring) or orange-colored pycniospores (fall) hymenial layers were chipped away from the outside of freeze-dried galls using a scalpel. Aeciospores were collected by knocking them off the surface of spring-collected pine galls. Oak associated tissues included infected oak leaves with attached telial columns, telial columns removed from oak leaves and basidiospores collected onto pH 2.0 water wetted filters (to prevent germination). Oak leaves and telial columns were stored at 20°C and basidiospores were stored in pH 2.0 water, at 4°C for up to 4 days. Pine associated tissues were collected in the field, at the University of Florida in Gainesville, Florida. Oak associated tissues were collected from greenhouse-inoculated, open-pollinated wild northern red oak (*Quercus rubra*) seedlings at the USDA Forest Service, Resistance Screening Center in Asheville, North Carolina.

RNA was extracted from the following tissues: aeciospores, basidiospores, telial columns, infected oak leaves, and the spring and fall hymenial layers of pine galls. Cetyltrimethylammonium bromide (CTAB) buffer was used, with different grinding procedures for each tissue. Infected oak leaves were frozen and ground in liquid nitrogen. The following tissues were ground in CTAB buffer pre-warmed to 65°C using a Geno/Grinder 2000 homogenizer (BT&C 24 Incorporated): 1) Aeciospores were ground in 4ml round bottom vials containing ~20mg of spores and a 1.0cm stainless steel ball; 2) Basidiospores were ground in 1.5ml Eppendorf tubes containing ~20mb of spores, 150mg zircon beads, and 12.5mb diatomaceous earth; 3) Telial columns were ground in 1.5ml Eppendorf tubes containing CTAB buffer and a 1.0cm steel ball. In all cases, extracted RNA was treated for 30 minutes at 37°C with RQ1 RNase-free DNase (Promega, M6101) and then purified using a Qiagen RNeasy Mini Spin Column.

RNA for library production and sequencing was analyzed on an Agilent 2100 Bioanalyzer and only RNA with a minimum RNA integrity score (RIN) of 6.3 was selected for sequencing. Libraries were constructed using RNA from the following five samples: 1) pycnial hymenial layer, 2) aecial hymenial layer, 3) aeciospores, 4) basidiospores, and 5) infected oak leaves with telial columns. Total RNA from these *Cqf*. samples was used to generate five individual, RNASeq libraries. Messenger RNA was purified from total RNA using the Absolutely mRNA™ purification kit (Stratagene). The isolation procedure was performed twice to ensure the sample was free of rRNA. Subsequently, the mRNA samples were chemically fragmented to the size range 200-250bp using 1x fragmentation solution for 5 minutes at 70°C (RNA Fragmentation Reagents, AM8740–Zn, Ambion). First strand cDNA was synthesized using Superscript

II Reverse Transcriptase (Invitrogen) and random hexamers. Complementary DNA (cDNA) was purified with Ampure SPRI beads. Then the second strand was synthesized using a dNTP mix (with dTTP replaced with dUTP), E.coli RNaseH, DNA Ligase, and DNA polymerase I for nick translation, resulting in double-stranded cDNA (dscDNA). The dscDNA were purified and selected for fragments in the range 200-300bp using a double Ampure SPRI bead selection. The dscDNA fragments were then blunt-ended, poly-A tailed, and ligated with Truseq adaptors using Illumina DNA Sample Prep Kit (Illumina). Adaptor-ligated DNA was purified using Ampure SPRI beads. Then the second strand was removed by AmpErase UNG (Applied Biosystems) similar to the method described previously (36). Digested cDNA was again cleaned with Ampure SPRI beads. Paired-end 76 bp reads were generated by sequencing using an Illumina HiSeq instrument.

Five lanes of Illumina HiSeq data from the individual RNA libraries were groomed using FASTQ Groomer v.1.0.4 and subsequently filtered using the Fastx Toolkit Quality Filter v1.0.0, first requiring 100% of each read to maintain a quality score of 20, then maintaining reads with at least 95% of the bases with quality scores greater than 30 (www.hannonlab.cshl.edu/fastx_toolkit/index.html). Paired-end reads were mapped to the final JGI assembly using Tophat (37). Assembly using the final JGI annotations and calculation of transcript abundance was completed using Cufflinks v.0.0.5 (38), allowing for no mismatches and intron specifications of 63-1457bp. Duplicate reads (exact matches) were removed by Picard v.1.56.0 (http://broadinstitute.github.io/picard/).

In order to identify variant sites that mapped to the *Avr1* genomic region, we employed bulk segregant sequencing, which involved a pooling strategy with Illumina sequencing. Briefly, we sequenced DNA extracted from pools of pycnial droplets produced on *Fr1*/*fr1* (resistant) pine hosts (i.e., ONLY virulent *avr1* alleles) and compared this to sequence data obtained from DNA extracted from pycnial droplets produced on *fr1*/*fr1* (susceptible) pine hosts (i.e., a mixture of *Avr1* and *avr1* alleles). Variant sites were identified with UnifiedGenotyper from GATK 3.6 and a genomic scan for markers associated with *Avr1* was performed with pFst from the vcflib suite (https://github.com/vcflib/vcflib).

Blastn (Blast2.3.0+) and GAP5 (39) were used to align and visually identify the most likely position for genetic markers within the *Cqf* assembly. For SSR markers, either the primer sequences or the cloned DNA sequence from which they were developed, were queried. Primers had to be found in the correct orientation, within the expected allele size range, and flanking the expected simple sequence repeat motif to be considered confidently placed. For RAPDs, 8-mers representing the 3’-most end of the original 10-mer primers were queried. Primer sequences had to be found in the correct orientation and within +/- 25 bp of agarose gel-esimated fragment sizes to be considered tentatively placed. For AFLPs, each primer-pair combination representing only the enzyme-specific nucleotides plus their corresponding selective amplification nucleotides were queried. Primer-pair sequences had to be found in the correct orientation and within +/- 3 bp of an adjusted sequencer-estimated fragment size (-24bp) to account for core AFLP primer nucleotides to be considered tentatively placed. For all marker types, in cases where multiple marker candidates were observed within the *Cqf* assembly, no conclusions could be drawn regarding placement.

The presence of telomere-like sequences was investigated using the simple sequence repeat $(TTAGGG)_5$ and BLAST. To be declared significant, hits had to have 100% identity to sequences within the *Cqf* assembly. Alignments were visually inspected and further characterized using GAP5.

Results and Discussion

The *Cqf* draft genome assembly (Croqu 1) totaled 76.6 Mb with 1,198 scaffolds (Table 3.1). The *Cqf* draft genome assembly contained 13,903 predicted genes supported by RNA sequence obtained from five developmental stages. This gene set was 84.7% complete according to the Benchmarking Universal Single-Copy Orthologs (BUSCO) approach and was the smallest among the sequenced rust genomes of *Puccinia graminis* f.sp. *tritici* (*Pgt*) and *Melampsora larici-populina* (*Mlp*) despite a significant fraction (17.6%) composed of transposable elements (Table 3.1) (40, 41).

To obtain expression evidence during multiple developmental stages of compatible interactions, RNA sequences were analyzed from infected leaves, mixed teliospores/basidiospores (both telial host collections), fall hymenial layer generating pycniospores, spring hymenial layer generating aeciospores, and aeciospores. Transcripts were detected for 77.4% of the predicted gene models from the final assembly (at >20RPKM) in at least one tissue source and from 54.3% of the gene models in all five tissue sources (Table 3S.1). These facts provide experimental support for expression of most *Cqf* genes at multiple stages of development. *Cqf* transcriptional activity appears similarly high to that of *Mlp*, where evidence of expression for 79.5% of gene models was obtained from at least one stage and 50.2% from all four stages collected from the telial host.

Several trends reported in other rust genomes were observed in *Cqf*; these include large families of small secreted proteins, protein kinases, Major Facilitator Superfamily (MFS) proteins, and proteins containing WD40 and zinc-finger domains (40). Similar to what was found in *Mlp*, the nitrate assimilation cluster in *Cqf* is partially complete. The nitrite reductase gene appears to be missing (protein ID 291348 in *Laccaria bicolor* (L. bicolor*)*). While the nitrate reductase gene is found on scaffold 1 of *Cqf* (Cqf649896), adjacent to this gene is a MFS transporter (Cqf649897) more simiar to a Git1p-related permease than to the nitrate/nitritie transporter (protein ID 723812 in *L. bicolor*) expected in the fungal nitrate assimilation cluster. Deficiency in nitrogen assimilation is presumably compensated by the capacity to directly assimilate peptides and amino acids, as has been suggested for *Mlp* and *Pgt* (40). Similar to the other rust fungal genomes, *Cqf* encodes 13 oligopeptide transporter (vs. 23 in *Mlp* and 21 in *Pgt*) and 10 amino acid permeases (vs. 15 in *Mlp* and 12 in *Pgt*). No sulfate reductase gene was detected in *Cqf*, suggesting that it, like other rust fungi, is similarly impaired in sulfur assimilation (40). *Cqf* contained 1140 genes encoding secreted proteins, with 666 of them encoding effector-like SSPs (small secreted proteins under 300 amino acids). We compared the predicted secretomes of the three sequenced rust pathogens to one another, and found the majority of secreted proteins and SSPs to be unique to individual rust species, possibly reflecting their rapid evolution (Fig. 2S.1).

Previously, we identified genetic markers flanking the *AVIRULENT TO FUSIFORM RUST RESISTANCE 1* (*Avr1*) locus (12), which is recognized in a gene-for-gene manner by the first well-characterized resistance locus in pine, *FUSIFORM RUST RESISTANCE 1* (*Fr1*; 1), and is located on *Cqf* Linkage Group III at 138 cM from the origin. Mapping the *Avr1* locus was feasible because the mapping population used to identify *Avr1* was derived from a heterozygous *Avr1*/*avr1* isolate (P2) and segregated in a Mendelian fashion. We cloned and sequenced single or low copy fragments containing genetic markers linked to *Avr1* in the *Cqf* genetic map. One marker was an amplified fragment length polymorphism marker; E13M6 (74 bp and 158 cM from origin), and one was a simple sequence repeat marker (DN_058; 80 bp and 113 cM). We searched the assembly for the sequence-based genetic markers and detected them in the order predicted by the genetic map, on scaffold #20 (Fig. 3.2, Table 3.2). The co-location of these three flanking markers, and in the correct order, suggested that the *Avr1* locus is located within a single scaffold in the draft reference genome.

Using our unique strategy, outlined above, it is possible to define a genomic interval for *Avr1* on the *Cqf* pine aecial host for two reasons. First, the haploid spore type present in pycnial drops allows genotyping of meiotic products arising from teliospores (produced on oak, the telial host) that recombine prior to infection of the aecial host. Second, resistant aecial (pine) hosts act as “filters” against avirulent spores, creating a selective sweep near the avirulence locus in the pools. Variable sites were identified from Illumina sequence of each bulk, and peak values or probabilistic F_{ST} (pFst) occurred over the identical genomic interval on scaffold #20 that was also identified using the cloned genetic markers (Fig 3.2). Although there was not strong evidence implicating a single gene model as *Avr1* we confirmed a genomic interval with 36 candidate genes, including

5 that contain a secretion signal peptide and do not share sequence similarity with a protein of known function (Table 3.3).

In order to gain some fundamental insight on the *Cqf* assembly, we focused on the localization of genetic markers previously mapped in *Cqf* isolate P2 which consisted of 421 markers distributed across 39 linkage groups and 9 pairs (12). The genetic map was composed primarily of dominant RAPD and AFLP markers (92%). However, 37 marker loci (33 SSRs, 20 cloned RAPDs, and 2 cloned AFLPs) provided an opportunity for "confident placement" given their longer query sequence lengths and/or requirement for the presence of an expected internal SSR motif. Based on blast, 36 markers were identified as a single locus within the *Cqf* assembly. Given their dominant inheritance, lack of intervening sequence information, and standard error associated with fragment size estimation, all RAPD and AFLP markers are considered "tentative".

Based on SSR and cloned marker sequence data, 19 LGs and one pair were confidently associated with 30 *Cqf* scaffolds (Table 3.3). Given the limited number of markers available, most joins were based on only a single marker. However, 3 scaffolds contained 2 or more markers and hence could be oriented with respect to their corresponding LGs. Considering the RAPD and AFLP sequence data, all 39 LGs and 9 pairs were tentatively associated with a total 87 scaffolds. A major goal moving forward will be to leverage the *Cqf* assembly and additional sequencing data to develop a much larger panel of genetic markers. High density mapping based on a prior information now made possible from the *Cqf* assembly will prove instrumental for both future map-based cloning and genome finishing efforts.

To gain additional insight on the *Cqf* assembly, we focused on the localization of eukaryotic telomere repeat sequences (42). Using the simple sequence repeat $(TTAGGG)_5$, we identified and characterized 17 distinct locations within the *Cqf* assembly. Information about these sequences including scaffold, length, scaffold end, scaffold location, and number of repeat motifs is reported in Table 3.4. Fifteen of the telomere-like sequences were localized to the ends of scaffolds as would be expected for true telomeres. One sequence was found to be centrally located on scaffold 130. Curiously, this sequence was found to be immediately adjacent to a long stretch of unknown sequence, i.e., N's. In addition, approximately 1200 bp of sequence was distally located to a telomere-like sequence observed on the high end of 230. Similar to scaffold 130, a stretch of unknown sequence (N's) was again observed to separate the telomere-like sequence from the distal sequence. It is possible that these anomalies are due to recent double-stranded breakpoints that have yet to be fully degraded/repaired (43), but they may also simply represent chimeric artifacts of cloning and library construction. Regardless, these 17 scaffolds should prove extremely useful as potential anchor points and at least 15 may tentatively be classified as chromosomal ends. Currently, there is no karyotype information specifically for *Cqf*. Assuming that *Cqf* has at least 17-18 chromosomes as has been observed for other closely related rust species in the order *Pucciniales* (44–46), additional telomere sequences likely exist, but similarly to other sequenced fungal species (47, 48), they simply were not captured in the current assembly. The identification and finishing of telomeric regions has often proven to be difficult and will require additional targeted mapping, cloning, and sequencing (49).

Conclusions

In this manuscript, we report the sequencing, assembly, and annotation of the *Cqf* genome, and present evidence for expression of gene models at multiple stages of the life cycle. We also demonstrate the utility of the *Cqf* genomic resource by identifying a single scaffold containing the avirulence (*Avr1*) in the pathogen that interacts with the corresponding *Fr1* resistance gene in the aecial host. Our analysis of the *Cqf* genome identifies candidate effectors that may be conditioning the specific interactions with its aecial host, and more generally provides useful genomic resources to explore coevolution between heteroecious rust fungi and their host plant species.

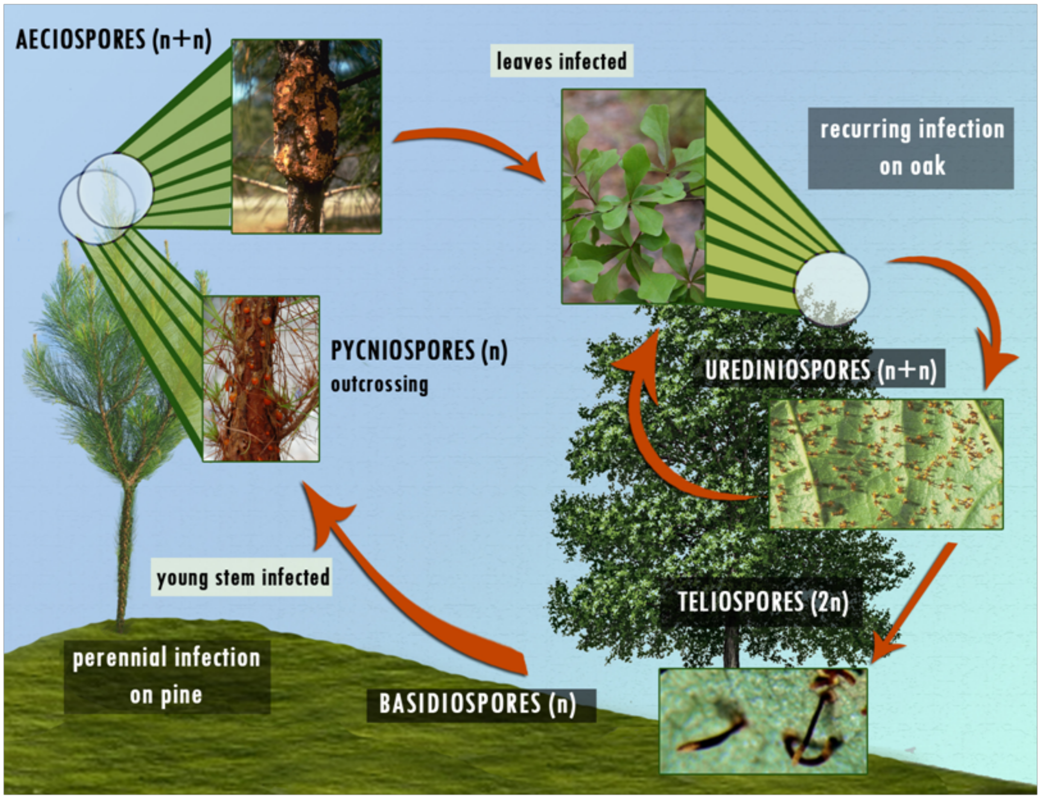


Figure 3.1 Life cycle of *Cronaritum quercuum* f.sp. *fusiforme* (*Cqf*)

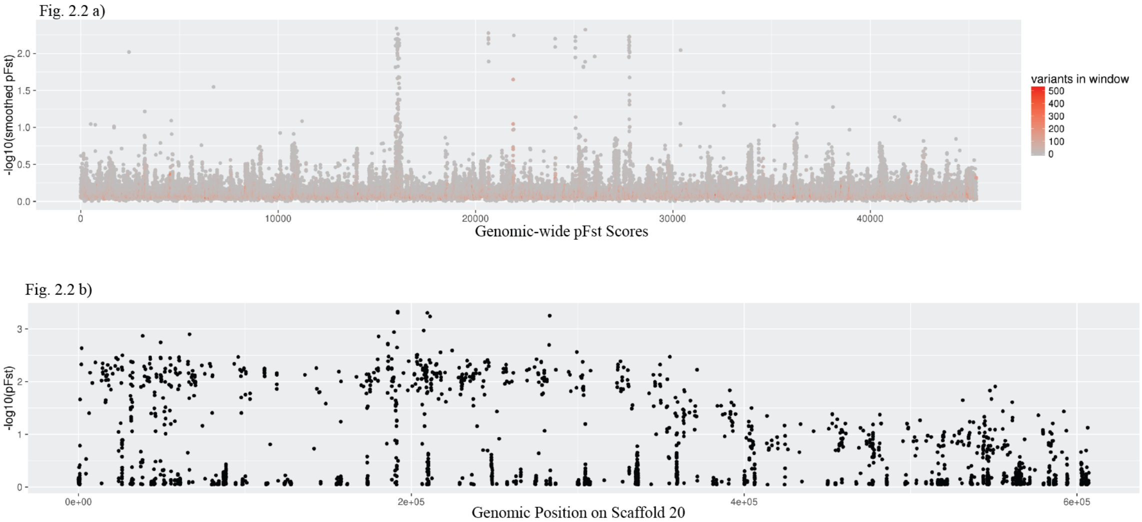


Figure 3.2 Annotation features of the *Cronartium quercuum* f.sp. *fusiforme* (*Cqf*) assembly (Croqu 1) a) Genome-wide pFst scores in the *Cqf* genome, b) pFst scores on the scaffold 20 genome.

Table 3.1 Assembly and annotation features of *Cqf* and two other rusts

Assembly Features	**Cqf**	**Mlp**	**Pgt**
Scaffold total, Mb	76.6	101.2	88.6
Scaffolds	1,198	462	392
Scaffold N50	70	27	30
Scaffold L50, bp	312,582	1,146,214	964,966
Contigs	10,431	3,254	4,557
Contig N50	1,691	265	556
Contig L50, bp	10,112	112,315	39,497
Repeats, %	17.6	44.0	6.7
Scaffold gaps, %	22.7	3.4	8.0
Annotation Features			
Protein coding genes	13,903	16,694	20,534
Protein length, median, aa	225	306	265
Exon length, median, bp	160	150	141
Gene length, median, bp	1,198	1,396	1,329
Transcript length, median, bp	868	1,002	860
Intron length, median, bp	87	81	94
Multi-exon genes, %	83.5	91.2	94.2
BUSCO (Complete or Fragment), % complete	84.7	89.2	86.4

Table 3.2 Association between the recombinational linkage map for *Cqf* isolate P2 and the Croqu1 assembly based on mapped genetic markers and BLAST.

AAC_05	I	208:213.5	47	332043-332643
Cqf-55	III	62.1:157.5	11	522780-522939
DN_058	III	113.3:157.5	20	461725-462250
BB07_750	III	136.7:157.5	20	252929-254246
E6M7-481	III	155.8:157.5	20	178998-179097
E13M6-92/93	III	157.5:157.5	20	4676-4749
AAC_38	V	126.8:149.4	163	115236-115824
AAC_57	V	58.5:149.9	56	135353-134530
GATA-46	VI	1.1:131.4	63	261774-261538
AAT_05	VI	102.1:131.4	124	40899-41312
AAC_28	VI	121:131.4	216	91277-90403
AAG_13	VIII	0.0:125.7	126	172154-171854
Cqf-84	VIII	20.1:125.7	502	3791-3955
AAG_08	VIII	50.1:125.7	55	335988-337599
AAC_84	VIII	67.2:125.7	49	118126-117153
GATA_07	VIII	69.5:125.7	49	167845-169145
Cqf-83	IX	53.1:125.2	251	37617-37787
AAC_30	XII	0.0:107.7	131	111067-111755
Cqf-78	XIII	61.4:93.4	22	518080-518223
DN_133	XIV	52.4:90.6	7	300486-300883
Cqf-151	XV	6.0:87.8	1149	253-234
Cqf-96	XVI	52.2:86.6	237	30465-30641
Cqf-65	XIX	52.4:78.2	111	41570-41997

Table 3.2 Continued

DN_079	XX	17.8:65.2	267	12286-12619

Table 3.3 Gene models on scaffold 20 with corresponding JGI IDs, located between *Avr1*-linked markers E13M6 and BB07. Functional descriptions are based on domain, KOG, and GO classification. Red font indicates secreted.

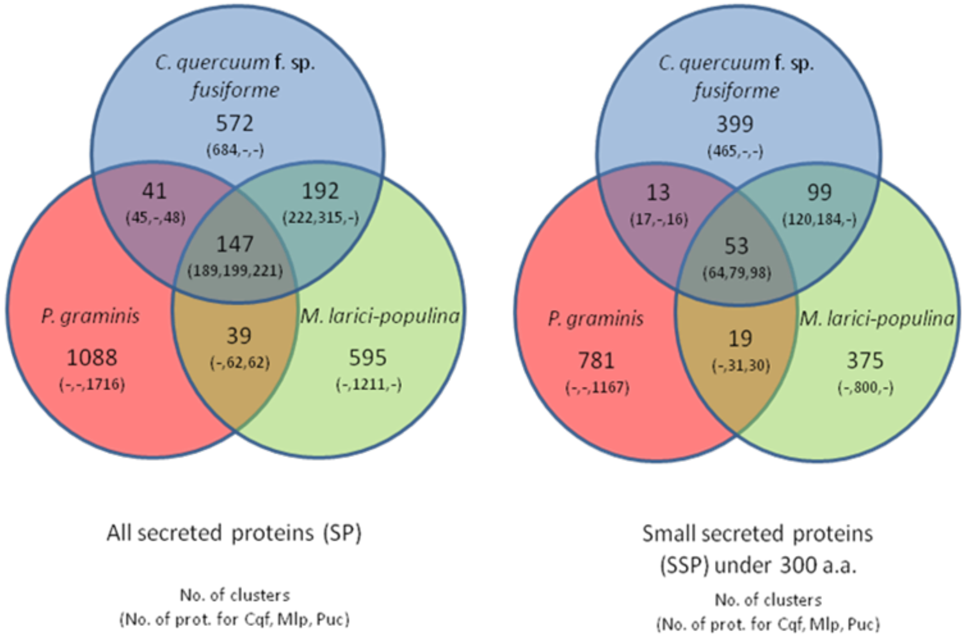
Protein/ Marker ID	Scaffold Position	Top Hit BLASTp	BLASTp Evalue	Functional Description	Length (aa)	Secreted Protein Cluster
E13M6	4676-4749					
131224	5318-9986	*Mlp*-hypothetical	0		1386	
318216	9951-10850	No hits	-		135	
39727	18739-22661	*Mlp*-hypothetical	0	Lipid transport and metabolism	1169	
39717	37567-38225	No Hits			144	
104995	41090-43578	*Mlp*-hypothetical	$8e^{-62}$		415	
318408	70702-73621	*Pgt*-hypothetical	$6e^{-75}$		408	
59217	77625-79309	*Mlp*-hypothetical	$6e^{-134}$	RNA processing and modification		
104998	79393-80078	No hits			123	No cluster
104999	85181-87979	*Pgt*-hypothetical	$8e^{-103}$		477	No cluster
318619	105080-105438	No hits			56	
720986	106082-109629	*Mlp*-hypothetical	$8e^{-120}$	Fungal transcription factor	778	
105001	115837-118941	*Mlp*-hypothetical	0	Aminopeptid-ase I zinc metalloprote-ase (M18)	494	
653529	119749-120131	No hits			69	
39766	133971-138495	*Mlp*-hypothetical	0	Nuclear pore, Nup160 component	1447	
653531	139251-143309	*Mlp*-hypothetical	0	Heat shock protein	939	cluster 1140

Table 3.3 Continued

Protein/ Marker ID	**Scaffold Position**	**Top Hit BLASTp**	**BLASTp Evalue**	**Functional Description**	**Length (aa)**	**Secreted Protein Cluster**
318901	144284-144831	No hits			92	No cluster
318956	146225-146629	No hits			73	
131233	146710-147654	*Mlp*-secreted protein	1e^{-22}		174	
89044	161734-162311	hypothetical protein (*Mycobacterium rhodesiae*)	9e^{-05}	Uncharacterized conserved protein	140	cluster 431 (4 members, 1 not secreted)
653534	164105-165020	hypothetical protein (*Puccinia striiformis*)	3e^{-55}	Peptidyl-prolyl cis-trans-isomerase	125	
89046	165106-166229	No hits			151	
319061	167417-168854	No hits			277	
653538	173065-173671	No hits			73	
59233	175817-177261	*Mlp*-secreted protein	2e^{-33}		223	cluster 292 (6 members, 2 not secreted)
687776	190223-190450	*Mlp*-hypothetical	9e^{-14}		76	
720994	194079-195314	*Mlp*-secreted protein	8e^{-90}		181	cluster 292(6 members, 2 not secreted)
39745	201048-202489	*Mlp*-hypothetical	1e^{-79}		370	
89056	203508-204096	No hits			152	
653547	204001-205224	*Pgt*-RING box protein	4e^{-71}	Unbiquitin ligase	116	

Table 3.3 Continued

Protein/ Marker ID	**Scaffold Position**	**Top Hit BLASTp**	**BLASTp Evalue**	**Functional Description**	**Length (aa)**	**Secreted Protein Cluster**
39695	213097-214941	*Mlp*- G-protein alpha subunit	3e^{-145}	Fungal G-protein alpha subunit	366	
653551	218731-225157	Myosin 5	0	Myosin head, motor region	1664	
39784	227153-228293	*Mlp*- G-protein alpha subunit	4e^{-50}	G-protein alpha subunit	237	
653554	240482-240834	No hits			68	
653557	241287-242771	*Mlp*-hypothetical	7e^{-114}	Signal transduction	200	
89063	243491-244153	*Mlp*-hypothetical	3e^{-09}	Protein binding (zinc finger, RING-type)	221	
59248	253753-255477	*Mlp*-hypothetical	8e^{-92}	Protein turnover (ubiquitin ligase, Skp1 component)	159	
BB07	252929-254246					



All secreted proteins (SP)

No. of clusters
(No. of prot. for Cqf, Mlp, Puc)

Small secreted proteins
(SSP) under 300 a.a.

No. of clusters
(No. of prot. for Cqf, Mlp, Puc)

Figure 3S.1 Comparison of secreted protein families between closely related rust fungal genomes.

Table 3S.1 Expressed gene transcript support for gene model predictions

No. of Tissue Sources with Expressed Gene Model	**Gene Count with >20 Reads/Kb Mapped**	**Percentage of Transcriptome Represented (%)**
5	7549	54.3
4	983	7.1
3	742	5.3
2	691	5.0
1	802	5.8
0	3136	22.6
Total	**13901**	**100**

References

1. McKeand SE, Amerson HV, Li B, Mullin TJ (2003) Families of loblolly pine that are the most stable for resistance to fusiform rust are the least predictable. *Can J For Res* 33(7):1335–1339.

2. Wilcox PL, Amerson HV, Kuhlmant EG, Liu B, Malley DMO (1996) Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. *Proc Natl Acad Sci* 93(April):3859–3864.

3. Amerson H, Nelson C, Kubisiak T, Kuhlman E, Garcia S (2015) Identification of nine pathotype-specific genes conferring resistance to fusiform rust inlLoblolly

pine (Pinus taeda L.). *Forests* 6(8):2739–2761.

4. Neale DB, et al. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* 15(3):R59.

5. Isik F, Amerson HV, Whetten RW, Garcia SA, McKeand SE (2012) Interactions of Fr genes and mixed-pathogen inocula in the loblolly pine-fusiform rust pathosystem. *Tree Genet Genomes* 8(1):15–25.

6. Catanzariti AM, Jones DA (2010) Effector proteins of extracellular fungal plant pathogens that trigger host resistance. *Funct Plant Biol* 37(10):901–906.

7. Dodds PN, Catanzariti AM, Lawrence GJ, Ellis JG (2007) Avirulence proteins of rust fungi: penetrating the host–haustorium barrier. *Aust J Agric Res* 58(6):512–517.

8. Kemen E, et al. (2005) Identification of a protein from rust fungi transferred from haustoria into Infected Plant Cells. *Mol Plant-Microbe Interact* 18(11):1130–1139.

9. Richardson BA, Zambino PJ, Klopfenstein NB, McDonald GI, Carris LM (2007) Assessing host specialization among aecial and telial hosts of the white pine blister rust fungus, Cronartium ribicola. *Can J Bot* 85(3):299–306.

10. Wingfield BD, Ericson L, Szaro T, Burdon JJ (2004) Phylogenetic patterns in the Uredinales. *Australas Plant Pathol* 33(3):327–335.

11. Pendleton AL, et al. (2014) Duplications and losses in gene families of rust pathogens highlight putative effectors. *Front Plant Sci* 5(June):299.

12. Kubisiak TL, et al. (2011) A genomic map enriched for markers linked to Avr1 in Cronartium quercuum f.sp. fusiforme. *Fungal Genet Biol* 48(3):266–74.

13. Burdine CS, Kubisiak TL, Johnson GN, Nelson CD (2007) Fifty-two polymorphic microsatellite loci in the rust fungus, Cronartium quercuum f.sp. fusiforme. *Mol Ecol Resour* 7(6):1005–1008.

14. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108(4):1513–1518.

15. Grigoriev I V., Martinez DA, Salamov AA (2006) Fungal genomic annotation. *Appl Mycol Biotechnol* 6:123–142.

16. Grigoriev I V, et al. (2011) The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* 40(November):1–7.

17. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM (2015)

BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.

18. Smit A, Hubley R, Green P RepeatMasker. Available at: http://www.repeatmasker.org/.

19. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.

20. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(SUPPL. 1):351–358.

21. Salamov AA, Solovyev V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res* 10(4):516–522.

22. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 18(12):1979–90.

23. Birney Clamp, M., Durbin, R E (2004) GeneWise and Genomewise. *Genome Res* 14(4):988–995.

24. Lowe TM, Eddy SR (1996) TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955–964.

25. Petersen TN, Brunak S, Von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786.

26. Nielsen H, Brunak S, Von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* 12(1):3–9.

27. Melén K, Krogh A, Von Heijne G (2003) Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol* 327(3):735–744.

28. Hunter S, et al. (2012) InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res* 40(D1):1–7.

29. Kanehisa M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34(Database issue):D354-7.

30. Tatusov RL, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

31. Michael A, et al. (2000) Gene Ontology: tool for the unification of biology. *Nature*

25:25–29.

32. Bendtsen JD, Nielsen H, Von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340(4):783–795.

33. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305(3):567–580.

34. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300(4):1005–1016.

35. Li L (2003) OrthoMCL: Identification of ortholog goups for eukaryotic genomes. *Genome Res* 13(9):2178–2189.

36. Parkhomchuk D, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37(18):e123.

37. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111.

38. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515.

39. Bonfield JK, Whitwham A (2010) Gap5-editing the billion fragment sequence assembly. *Bioinformatics* 26(14):1699–1703.

40. Duplessis S, et al. (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci U S A* 108(22):9166–71.

41. Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res* 37(1):289–297.

42. Casas-Vila N, Scheibe M, Freiwald A, Kappei D, Butter F (2015) Identification of TTAGGG-binding proteins in Neurospora crassa, a fungus with vertebrate-like telomere repeats. *BMC Genomics* 16(1):965.

43. Wellinger RJ, Zakian VA (2012) Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: beginning to end. *Genetics* 191(4):1073–1105.

44. Boehm EW a., McLaughlin DJ (1991) An ultrastructural karyotype for the fungus *Eocronartium muscicola* using epifluorescence preselection of pachytene nuclei. *Can J Bot* 69(6):1309–1320.

45. Boehm E, Bushnell W (1992) An ultrastructural pachytene karyotype for Melampsora lini. *Phytopathology* 82:1212–1218.

46. Boehm E, Wenstrom J, McLaughlin D (1992) An ultrastructural pachytene karyotype for Puccinia graminis f. sp. tritici. *Can J Bot* 70:401–413.

47. Rehmeyer C, et al. (2006) Organization of chromosome ends in the rice blast fungus, Magnaporthe oryzae. *Nucleic Acids Res* 34(17):4685–4701.

48. Wu C, et al. (2009) Characterization of chromosome ends in the filamentous fungus Neurospora crassa. *Genetics* 181(3):1129–1145.

49. Farman ML (2011) Targeted cloning of fungal telomeres. *Fungal Genomics: Methods and Protocols*, pp 11–31.

CHAPTER 4

REFERENCE-FREE COMPARATIVE GENOMICS AMONG TRICHOMONADS

This chapter is a manuscript written by Claudia Marquez, Dwight Kuo, Ellen Pritham, Mark Yandell, and myself. Claudia Marquez and Ellen Pritham constructed the transposable element library based on the reference genome for *Trichomonas vaginalis* strain G3. Dwight Kuo supplied the PacBio *de novo* assembly of *Trichomonas vaginalis* strain RP. Daniel Ence performed the genome annotation of both *Trichomonas vaginalis* genome assemblies. Daniel Ence implemented and applied the reference-free comparative genomics approach, under the guidance of Ellen Pritham and Mark Yandell. This manuscript will be submitted to the *Proceedings of the National Academy of Sciences* in November 2016.

Abstract

The trichomonads are a group of diverse, mostly parasitic eukaryotes, which include the agricultural and human pathogens, Tritrichomonas foetus (*Tritrich. foetus*) and *Trichomonas vaginalis* (*T. vaginalis*). *T. vaginalis* is notable for its extremely large number of annotated protein coding genes (>75,000) (1, 2). This has led to speculations as to its evolutionary significance, and its possible role in causing human disease and adaption to parasitic lifestyles. To date, absence of high-quality assemblies for multiple

trichomonads due to the difficulties inherent in assembling and annotating these highly repetitive genomes have prevented resolving these speculations (3). In response, we have re-annotated the *T. vaginalis* genome using a custom, highly curated repeat library in conjunction with the widely used, evidence-driven genome annotation pipeline MAKER (4). In annotations of both a previously published reference genome (1) and a *de novo* PacBio-based genome we describe here, this process reduces the number of annotated protein-coding genes in *T. vaginalis* by a third to ~23,000, while retaining the vast majority of core cell-metabolic genes. This process also identifies the majority of annotated *T. vaginalis* gene models as belonging to *Maverick* transposable elements, rather than the host genome. We also compare our gene models to WGS and RNASeq datasets from other trichomonads, using a reference-free analysis technique, that allowed us to carry out comparative analyses using only unassembled reads. These analyses demonstrate that the large gene families previously proposed to have arisen during *T. vaginalis*' adaption to its human host are shared by the seven trichomonad species examined here.

Significance Statement

Poorly sampled eukaryotic lineages present numerous challenges for genome analysis, including identifying novel transposable elements (TEs), which can hamper intra- and interspecific comparisons. One such lineage is the trichomonads, in which only one genome (*Trichomonas vaginalis* G3) has been annotated. Our results indicate that >50% of annotated protein-coding genes in *T. vaginalis* reside in TEs. We also present a method for reference-free comparative genomics to identify the core set of genes shared by trichomonads. Our method makes use of unassembled WGS and RNAseq data and bypasses the major requirement of comparative genomics, namely a reference genome for each species. Our core gene set identifies gene families that may have been expanded in the ancestral trichomonad relative to other crown eukaryotes.

Introduction

Genome annotation is problematic for many species that impact human health and agriculture due to a lack of genomic resources from themselves and/or closely related organisms. In addition, novel transposable elements (TEs) in the genome of interest and a lack of gene models for *ab initio* gene predictors and homology searches (5) also further complicate the annotation process.

The trichomonads stand out as an important but poorly sampled group of organisms. These eukaryotic unicellular parasites inhabit a variety of hosts (see Fig. 4.1) and are distantly related to other eukaryotes (6). The two best-known species of parabasalids are the human pathogen *Trichomonas vaginalis* (*T. vaginalis*) and the bovine pathogen, *Tritrichomonas foetus* (*Tritrich. foetus*), both of which cause significant disease and economic loss in humans and livestock (7–9).

The global human pathogen *T. vaginalis* was the first genome of the order trichomonadida to be annotated (1) and encountered the problems noted above. Although the original annotation identified one-third of the *T. vaginalis* G3 reference assembly as repetitive sequence, later reports suggested that an additional one-third of the genome was made up of a novel transposable element (TE) family, *Mavievicks* (10), thus casting doubt on the functional status of many of the original 60,000 protein coding genes reported in the original genome annotation. Even with recent publication of transcriptome data for several trichomonad species (11–14), comparative genomics in trichomonads are limited due to the large and repetitive genomes in this group (3, 15). Recent developments of long-read technology have been suggested as a way to overcome both the size and repetitiveness of genomes in this group (15).

We present a re-annotation of the *T. vaginalis* G3 reference assembly as well as an annotation of a genome assembly of *T. vaginalis* strain RP that was sequenced with PacBio long-read technology. Our re-annotation used a library of TE consensus sequences identified in the *T. vaginalis* G3 reference assembly in conjunction with the widely used genome annotation pipeline, MAKER (16–18).

We also present a method for reference-free comparative genomics and compare the genomes and transcriptomes of 7 trichomonad species. This method applies a kmer-based approach originally developed for metagenomic analysis, which allows for identification of core gene sets to be assembled and annotated without whole genome assembly (19). This approach works rapidly and accurately across large phylogenetic distances, and requires works directly from unassembled sets of reads. Thus it overcomes one of the major requirements of comparative genomics, namely a reference genome for each species to be studied.

Results

We used our custom TE library with the MAKER genome annotation pipeline (16–18) to annotate genome assemblies of two strains of *T. vaginalis*: the *T. vaginalis* G3 reference assembly (1) and a *de novo* assembly of *T. vaginalis* strain RP made with PacBio long-read assembly. This resulted in 25,965 (MAKER G3) and 23,958 (MAKER RP) protein coding genes, respectively (see Fig. 4S.1). These updated gene sets are less than one-third the size of the 75,000 protein-coding genes in the current trichDB 1.3 (20).

The percent of base pairs in the genome annotated to genes was similarly reduced from 72Mbp (56%) to 35Mbp (27%) (see Fig. 4.2b), while the percent of the genome annotated as repetitive elements increased from 39Mbp (30%) to 58Mbp (45%). Even more of the RP Pacbio-based assembly was identified as a repetitive element (52%, 84Mbp) than in the G3 Sanger-based assembly (see Fig. 4.2b).

Out of the 75,000 genes currently reported by trichDB, 21,194 had a matching gene annotated in the MAKER G3 set, as identified by reciprocal best hit (RBH). RBH identified 17,396 genes in common between the MAKER RP set and the trichDB gene set and similarly 16,592 genes in common between the MAKER G3 and MAKER RP sets ($p < 0.001$). Out of the 54,305 genes from trichDB without a matching MAKER G3 gene, 28,620 were entirely overlapped by a TE (see Fig. 4.2a). Even though the MAKER G3 set is one-third the size of prior estimates, virtually all core metabolic pathways present in the trichDB gene set are also present in the MAKER G3 set (see Figs. S4.2a,b).

The original annotation of the *T. vaginalis* G3 reference assembly reported 880 kinase genes (1). Annotation of the trichDB and MAKER G3 gene sets with the kinase annotation pipeline, Kinannote (21), identified 936 kinases in the trichDB gene set and 820 kinases in the MAKER G3 gene set (see Fig. 4S.3, details in Table 4S.1). Surprisingly, MAKER RP set identified 442 kinases, which is still expanded relative to other microbial kinomes. Thus all *T. vaginalis* kinome annotation datasets, including our own, are expanded relative to most other microbial kinomes (see Fig. 4S.3), and 88% of currently annotated kinases are included in the MAKER annotation of the G3 reference genome, despite its dramatically reduced gene count (see Fig. 4S.1, Table 4S.1).

The original annotation of the G3 reference assembly also reported 463 'degradome' genes. The vast majority of the degradome reported in the trichDB set of genes is present in the MAKER results (see Fig. 4S.4; 508 degradome genes in trichDB genes, 90% of these (456) are present in the MAKER G3 annotations and 82% (417) are present in MAKER RP set). Two families that were noted as expanded in the original annotation of *T. vaginalis* (cysteine peptidase clan CA, family C19 and cysteine peptidase clan CA, family C1 or papain-like cysteine peptidases) are also expanded in both the MAKER G3 and MAKER RP sets (see Fig. 4S.4).

Even though the re-annotation of the G3 reference assembly resulted in a gene set one-third the size of prior estimates, important classes of genes noted in prior publications are maintained. The re-annotation gene set includes a set of meiosis-specific genes noted in prior publications (22, 23) (see Table 4S.2). In 2014, it was reported that a 34kbp region of the *T. vaginalis* genome constituted a single, large lateral gene transfer event (24). This region was confirmed in three different strains of T. from three different labs, and thus is likely not a bacterial contamination. This region included 27 consecutive genes which had their origin in a close relative of the firmicute bacteria *Peptoniphilus harei*, which varied in their sequence conservation and coding potential. In this region, 14 out of the 27 genes are present in the MAKER genes. The 13 missing genes were not overlapped by any RNAseq or other expressed sequence evidence. Among the genes present in this region is the only example of an FtsH peptidase in the genome of *T. vaginalis* (TVAG_243780 in trichDB set, snap_masked-DS113827-processed-gene-0.43 in MAKER set). Although scaffold DS113827 is broken across several scaffolds in the

PacBio-based assembly of strain RP, seven of the genes from scaffold DS113827 are present in the MAKER RP set, including the FtsH peptidase.

Both the trichDB gene annotations and the subset of trichDB gene annotation without a matching MAKER gene and not overlapped by a TE contain a preponderance of genes with terms related to TE activity. By comparison, the MAKER annotations of both the G3 reference and strain RP PacBio assemblies have a much smaller proportion of genes with terms related to TE activity ("DNA-binding", see Fig. 4S.5).

A preliminary application of our reference-free comparative genomics method to data simulated from bacterial genomes indicated that our method can recover homologous genes at phylogenetic distances similar to those separating trichomonad species even with low coverage data (Table 4S.4, Fig. 4S.6).

Using the set of MAKER G3 genes that were supported by expression evidence, protein homology or possession of a functional domain as reference sequences, we classified WGS and RNAseq datasets from strains of *T. vaginalis* and other trichomonad species and recovered gene models from each sample, although the numbers of genes recovered and median percent coverage scores were less than for the bacterial simulated data. The percent coverage scores and size of the gene sets decreased as the phylogenetic distance from the reference species (*T. vaginalis* G3) increases (see Tables 4.1, 4.2).

We identified a core set of 21,895 genes present in either the RNAseq or WGS sample of each strain of *T. vaginalis* (see Fig. 4.3). The *T. vaginalis* core gene set obtained with Taxonomer shares 17,677 genes in common with a *T. vaginalis* core set of 18,133 genes obtained with a more standard Trinity+rbh method ($p < 0.001$). Out of the 820 kinases identified in the MAKER G3 annotation set, 782 were shared between the

Taxonomer core gene set and the Trinity+rbh core gene set. The distribution of GO terms between the Taxonomer and Trinity+rbh core gene sets is essentially identical (see Table 4S.5).

With the Taxonomer-based method, we identified a core set of 2,458 genes present in each species of trichomonad included in this study. In comparison, the Trinity+rbh method identified a core set of 5,058 genes. The intersection between the sets identified by the two methods is comprised of 1,557 genes. The GO terms associated with genes shared by both gene sets the trichomonad core gene set includes many basic biological functions such as translation, transport, cell cycle, and cell motility. The set of core trichomonad genes also included 181 kinase genes (Fig. 4.4, Table 4S.5).

Discussion

The results of our reannotation of the G3 reference assembly highlight the vital importance of accurate and thorough masking of transposable elements (TEs) from the genome before annotating protein-coding genes (25). They also highlight the difficulty of conducting genomic research in regions of the tree of life that lack genomic and transcriptomic resources. The fact that the number of annotated genes in this genome was reduced to one-third of prior estimates, without losing major metabolic pathways or most of the large and expanded kinome, indicates the majority of genes not present in the MAKER G3 set were not contributing to *T. vaginalis* biology in those functions. Conversely, the previously annotated genes that were not annotated by MAKER also highlight the difficulty of *T. vaginalis* as an object for a genome annotation. This genome harbors many copies of *Mavericks*, a family of particularly large TEs that can carry 6-22 open-reading frames (10, 26). The wide variety of GO terms associated with genes overlapping masked regions indicates that these TEs include genes associated with a variety of functions not normally associated with TE function.

The increased amount of sequence masked as TE in the RP PacBio assembly indicates that the majority of the increased contiguity in the genome came from sequencing and resolving TEs as has been previously suggested (15). Conversely, the disagreement in kinase annotation results between the Sanger-based G3 reference assembly and the RP PacBio assembly suggests that many kinases were fragmented across multiple contigs in the G3 reference assembly and are now collapsed and merged in the RP PacBio assembly, highlight the utility of long-read technologies for resolving duplicated genes.

One major issue frustrating trichomonad comparative genomics is the absence of high-quality assemblies for multiple trichomonads. Unfortunately, obtaining sufficient high-quality DNA, cost, and the difficulties of assembling these highly repetitive genomes continue to limit comparative approaches. Trichomonads are hardly alone in this regard. In response, we have used an ultrafast read classification engine called Taxonomer (19) to identify a core gene set for 7 taxa (*Tritrichomonas foetus*, *Trichomitus batrachorum*, *Tetratrichomons gallinarum, Pentatrichomonas hominis, Trichomonas gallinae, Trichomonas tenax, Trichomonas vaginalis*). Benchmarks and proof-of-concept results for the technique are provided in Flygare and Simmon et al. 2015. The close agreement between our method and a more standard Trinity+rbh method in identifying an intraspecies (*T. vaginalis*) core gene set indicates that this method is effective. However, the difference in efficacy of our method between the bacterial species used in the proof-of-concept study and our samples of trichomonad species beyond *T. vaginalis* suggests that the sensitivity of our method can be improved, and consequently, that the core trichomonad gene set identified here represents the lower bounds of the set of genes shared by the 7 species used in this study. These gene families (kinases and various cysteine peptidase families) have previously been suggested to be important for pathogenesis (3).

Summary

By re-annotating the G3 reference assembly of *T. vaginalis*, we reduced the number of protein-coding genes by two-thirds. With the reduced set of genes, we were able to approach genomic comparisons for the broader trichomonad group confident that we had excluded the vast majority of transposable elements.

Only about a third of genes are in common between the trichomonad core gene sets identified by our method and the Trinity+rbh. This may be due to the large number of duplicated genes in the *T. vaginalis* genome, which we used as our basis of comparison for identifying matches, or due to limited sensitivity of our method. That our method works directly from short-read sets, instead of requiring annotated draft genome or transcriptome assemblies, represents an advantage over existing approaches. Two large classes of genes (kinases and 'degradome') are highly represented in the trichomonad gene set, suggesting that regardless of host or lifestyle, all trichomonads possess expanded suites of these genes.

Materials and Methods

To determine the TE composition of the *T. vaginalis* G3, we employed four computational tools to thoroughly annotate the genome. To identify the repetitive sequence of the *T. vaginalis* G3 genome, we employed RepeatScout (27). This program identified repeats <50 nucleotides in length and >50% complexity. To reduce redundancy and overlaps, sequences were collapsed (assembled) using Sequencher v 5.1 (28) with default parameters (58% similarity over 20 nucleotides). As this program does not classify these regions as TEs, we then employed REPCLASS (29).

REPCLASS was used to classify the repeated sequences as possible TEs. Module 7 was selected to include three classification modules: homology, structure, and target site duplication.

Both the ExPASy translate tool (30) and ORF Finder (31) tools were utilized to identify open reading frames (ORFs) within the filtered RepeatScout and REPCLASS outputs. The longest ORF (starting with a methionine) was used as a query against the protein domain database curated by NCBI. Hits were considered significant if the e-value was greater than 0.01. The function of the hypothetical ORF was predicted by homology to proteins of known function and by the presence of conserved domains identified through a conserved domain database (CDD) search (32).

To identify the structural components of the repeats, blast searches (primarily blastn and tblastn) were conducted using the putative ORFs identified by ExPASy and ORF Finder as queries using default parameters and without filtering for simple and complex repeats to identify novel TEs. Searches were conducted against various GenBank databases including whole genome shotgun reads (WGS), nucleotide collection (NR) high throughput genomic sequences (HTGS), genome survey sequences (GSS), and expressed sequence tags (EST) databases. Hits were considered significant when the e-value was $<10^{-4}$ and were then used as seed queries at the DNA and protein level.

Sequences were mined and binned if sequence identity was greater than or equal to 95%.

Each significant hit was then examined for TE structures (i.e., long terminal repeats (LTRs), non-LTRs, terminal inverted repeats (TIRs), and target site duplications (TSDs). TIRs were identified by pairwise comparisons taking 3,000 nucleotides upstream and downstream of each significant hit using blast. TSDs were identified by aligning 100 nucleotides upstream and downstream from the TIRs of the elements. To maximize the probability of identifying all probable elements, newly identified elements and putative proteins were used as queries using blast against the WGS and NR databases. Two TE copies were defined as being members of the same family when they displayed 80% pairwise similarity over at least 80% of the nucleotide sequence (33). Majority rule consensus sequences were generated for each family with Clustal W2 (34) and MacVector 7.2.2 (35).

Autonomous elements were used as queries using blast to identify related non-autonomous elements. The non-autonomous elements share the same TIRs but do not encode a transposase.

To illustrate the mobility of TEs, paralogous (empty) sites devoid of the insertion were queried against the *T. vaginalis* G3 genome sequence. Empty sites were identified by homology searches using blastn (word size 7, expectancy 1000) (36) with a query constructed from the sequences directly flanking the insertion site containing the unduplicated target site. The chimeric query sequence (~100 nucleotides) was created by extracting the flanking sequence (~50 nucleotides) upstream from the element insertion containing the TSD and extracting ~50 nucleotides downstream from the element insertion (lacking the TSD). The empty site query then should represent the site before the TE insertion occurred. An empty site was annotated when an alignment was found that spanned the chimeric sequence.

Alignments of the putative ORFs were constructed using Muscle (37) with default parameters and phylogenetic trees that were generated with MEGA v. 4 (38) using neighbor joining with Poisson correction, allowing for multiple substitutions at sites. Consensus sequences for multi-element families can be found in supporting information. RepeatMasker (39) was then employed using our TE library to identify all copies of each element as well as the TE composition of the *T. vaginalis* G3 genome. Hits were considered significant when sequence identity was greater than 90% over 50% of the ORF.

To determine the age of the elements identified, we used two complementary computational analyses including both a molecular clock dependent and independent technique. First the divergence from the consensus sequences of each individual TE insertion was used as a proxy for age. This approach complemented the nested insertion strategy that relies on the simple logic that when TEs are inserted in each other, the newest insertion would be the youngest. These age estimates were completed using the following script parseRM_GetLandscape_AK.pl available at https://github.com/4ureliek.

The CAI was used to estimate the codon adaptation of the consensus sequence transposase genes encoded by the TE families identified as well as the *T. vaginalis* house keeping genes previously identified by Cornelius et al. 2012 (TVAG_299450, TVAG_258340, TVAG_343390, TVAG_054490) (Appendix D). CAI and estimated CAI (eCAI) values were generated using the CAICal Server http://genomes.urv.cat/CAIcal/E-CAI/ (40). Statistical support and correction for G+C composition bias was calculated using the eCAI. Normalized CAI values were obtained by taking the quotient of the CAI and eCAI. CAI values larger than the eCAI or expected values can be attributed to adaptation (40). Codon usage tables for *T. vaginalis* and other selected organisms were obtained from the Codon Usage Database (41). Organisms were selected based on previous literature identifying 152 possible cases of horizontal gene transfer from human pathogenic prokaryotes to *T. vaginalis* (23).

Trichomonas vaginalis strain RP was obtained from ATCC (30188), and cultivated using ATCC PRA-2154 LYI Entamoeba medium, as recommended for axenic cultivation of *Trichomonas*. The cultures were split and harvested at log phase with >95% of cells motile. Cells were counted with a Neubauer Chamber and harvested in eight aliquots of 3 x 107 cells.

The Qiagen genomic DNA preparation kit for cultured cells was modified for use with nuclease rich protozoans, and genomic DNA was purified with Qiagen Q100 anion exchange resin. Cells were lysed with sucrose and Triton X at low ionic strength. Subsequently, nuclei were then lysed with guanidine, and lysates were poured over the anion exchange resin. The negatively charged DNA backbone binds to the positively charged DEAE groups on the surface of the resin, at a low salt concentration. DNA remains tightly bound to the resin over a wide range of salt concentrations. Impurities such as degraded RNA, cellular proteins, and metabolites are removed by a medium-salt wash. Genomic DNA was eluted in a high-salt buffer, and concentrated and desalted by isopropanol precipitation at room temperature to minimize co-precipitation of salt. After centrifugation, the DNA pellet was washed with 70% ethanol to remove residual salt and to replace the isopropanol with ethanol, which is more volatile and easily removed.

Modifications to the extraction protocol included use of DEPC and EDTA, some adjusted volumes for the protozoan DNA content, and minimal vortexing and pipetting to preserve DNA integrity.

The DNA was quantified by Picogreen. Integrity and size distribution were identified using the Agilent Tapestation and pulsed-field gel electrophoresis.

SMRT sequencing on PacBio RSII with the P5C3 chemistry was performed. Loading titrations of 0.120 nM, 0.240 nM, 0.360 nM and 0.480 nM on-plate concentration were evaluated. An on-plate concentration of 0.240 nM was selected since significant improvements in total reads were not observed at the higher concentrations. This on-plate concentration is 2X higher than the standard loading for 20kb libraries and is likely due to the long libraries. Based on the PacBio binding calculator and selected on-plate concentration of 0.240 nM, 2.6 ug of SMRT bell library support a maximum of 100 SMRT cells. Sequencing of 30 SMRT cells resulted in a post-filtered read coverage of 82x at an assumed genome size of 170 MB.

de novo assembly was done using a hierarchical method. The seed read cutoff for pre-assembly was 8000 bp (21x coverage). Pre-assembly was performed using daligner followed by assembly using FALCON for diploid organisms. The final assembly was polished using Quiver (42). Assembly was iterative for a total of 24x and all settings converged on the genome size of 163.9 MB. Less stringent settings led to a significant amount of mis-assembly. Very little of the genome (2 Mb, ~1%) is in an alternative path. Because the level of heterozygosity is high, alleles were assembled as separate contigs.

We annotated both G3 reference assembly (23) and the *de novo* assembly of strain RP of *T. vaginalis* with MAKER(16–18). In order to match the N50 length (67.59 kbp) reported in the original paper, we selected the 17,290 longest scaffolds out of the G3 reference assembly. This subset of scaffolds is 129 Mbp out of the 176Mbp in the full reference assembly. The entire 164 Mbp of the *de novo* assembly of strain RP was used.

In order to update the genome annotation of the G3 reference assembly, we used the MAKER genome annotation pipeline. We masked both assemblies with the library of consensus sequences of transposable elements in the *T. vaginalis* G3 reference assembly.

After repeat masking the genome, MAKER used the blast suite of programs to align transcript and protein data to the reference assembly. For transcript data, I used *T. vaginalis* transcripts downloaded from dbEST, Trinity-assembled transcriptomes from two prior studies of *T. vaginalis* (13, 14, 43), as well as our own sample of *T. vaginalis* G3 RNA. Transcriptomes were assembled with Trinity (see RNAseq section below) (44). MAKER aligned the transcript evidence with blastn (45).

The protein sequence data consisted of the three proteomes used in producing the original genome annotation (*Dictyostelium discoideum*, *Giardia lamblia*, and *Entamoeba histolytica*) in addition to the UniRef90 (46) set of protein sequences. MAKER aligned protein sequence data with blastp (45).

Three *ab initio* predictors were used to generate gene predictions. Genemark was self-trained on the G3 reference assembly (47). Snap (48) and Augustus (49, 50) were both trained on the set of conserved proteins identified in the *T. vaginalis* G3 reference genome with CEGMA(51, 52). Given the paucity of introns previously reported in the *T.*

vaginalis genome (23, 53), we adjusted the default settings of MAKER to allow single-exon alignments of expressed sequence greater than 250bps to be considered as evidence.

We also evaluated how our set of evidence compared to the gene models publicly available in trichDB v1.3. *T. vaginalis* gene models were downloaded from trichDB.org (20), and provided to MAKER as models.

Functional annotations (Pfam domain and GO term annotations) for the MAKER genes and trichDB models were assigned with InterProScan v. 5.8-49.0 (54).

MAKER and trichDB genes were assigned to KO (KEGG Orthology) categories with the BBH method on the KAAS server (http://www.genome.jp/kaas-bin/kaas_main) (55) and visualized with the Interactive Pathways Explorer v2 (http://pathways.embl.de/iPath2.cgi) (56, 57).

We used the Kinannote pipeline (21) to identify kinases in our MAKER gene set. We obtained an expected number of kinases for each species from a literature search (see Fig. 4S.2, Table 4S.3).

Samples of genomic DNA for six strains of *Trich. vaginalis* (30235, B7RC2, G3, JRSTV41, and T1) were obtained from ATCC. Libraries were prepared for each library using the Illumina TruSeq DNA sample prep protocol with custom size selection for 200bps inserts.

Samples for an additional four strains of *Trich. vaginalis* (30238, 50143, and t016) as well as three other trichomonad species (*Trichomonas tenax*, *Pentatrichomonas hominis*, and *Tritrichomonas foetus*) were obtained from ATCC. DNA libraries were prepared from the samples with the Illumina TruSeq Nano DNA Sample Prep protocol.

RNA-Seq reads were obtained using Illumina sequencing on *Pentatrichomonas hominis* PhGII, *Tetratrichomonas gallinarum* M3, *Trichomitus batrachorum* BUB, *Trichomonas gallinae* GCB, and *Trichomonas tenax* HS-4 RNA, which were isolated as previously described for *T. vaginalis* (Woehle et al. 2014). The 100bp reads were trimmed for low-quality basepairs and adapter sequences with prinseq-lite version 0.20.4 (58) and fqtrim (http://ccb.jhu.edu/software/fqtrim/). Singleton reads were discarded.

RNAseq samples of *T. vaginalis* B7RC2 and G3 were prepared with the Illumina TruSeq Stranded prep. procedure, which selects poly-A transcripts out of total RNA and sequenced with Illumina HiSeq 125bp paired-end sequencing v4, and cleaned with seqyclean (59).

Previously published datasets of Illumina 100bp paired-end RNAseq reads for *Tritrichmonas foetus* from cat and cow (*Tritrich. foetus* feline, *Tritrich. foetus* bovine) were obtained from NCBI accession numbers SRX540117 and SRX540971, respectively, (12), and trimmed with prinseq-lite and fqtrim as described above.

RNAseq samples for *T. vaginalis* strain t016 were obtained from NCBI accession numbers SRP036029 and SRP015999 (13, 14), and trimmed with seqyclean (59).

Quality-cleaned and trimmed RNAseq samples were assembled individually with Trinity vs2.0.6 (44). Open-reading frame (ORF) gene models were annotated in the Trinity assembled transcripts with Transdecoder vs2.0.1 (60).

In repurposing Taxonomer (19) from its initial purpose of metagenomics analysis to comparative genomics, we first built a classification database from the amino acid sequences of the set of MAKER G3 gene sequences supported by either expressed transcripts, protein homology, or a Pfam domain for classification in protein space (comparable to blastx).

Short reads are then classified to the reference sequences in protein space and assigned to the reference sequence that maximizes the kmer-weight metric described in (19). Reads that were not classified to any reference sequence or were tied for classification between multiple references were discarded. The set of reads from a given sample that classified to a single reference sequence are then assembled with the velvet genome assembler (61).

The longest contig (e.g., recovered gene model) resulting from the velvet assembler was then aligned to the set of reference amino acid sequences with blastx. If the original reference sequence to which the reads were classified is among the top 100 blastx hits, then the “recovered” gene model was considered present in that sample. In order to evaluate the classification and assembly results, the percent of the original reference amino acid sequence covered by any blast HSP for a recovered gene model was recorded; this was termed the “percent coverage” score for a recovered gene model in a given sample. The percent coverage scores for an entire sample were summarized in cumulative cutoff tables.

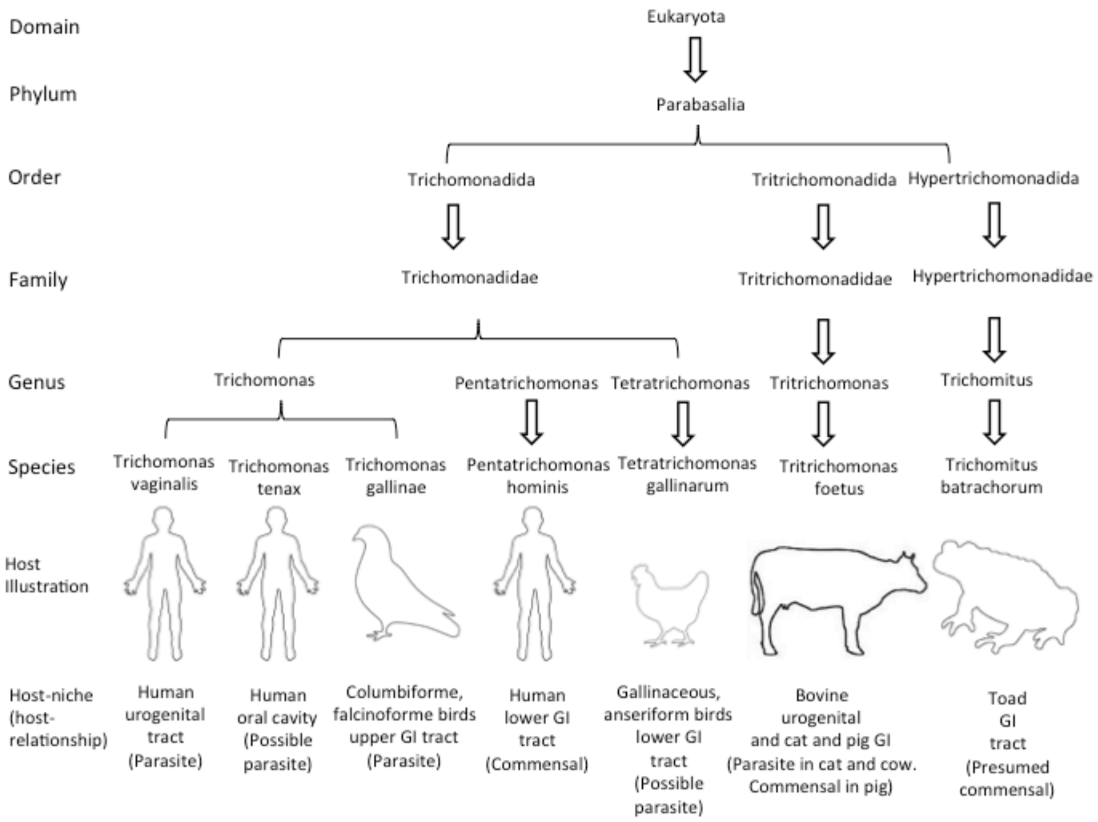

Domain
Eukaryota
Phylum
Parabasalia
Order
Trichomonadida
Tritrichomonadida
Hypertrichomonadida
Family
Trichomonadidae
Tritrichomonadidae
Hypertrichomonadidae
Genus
Trichomonas
Pentatrichomonas
Tetratrichomonas
Tritrichomonas
Trichomitus
Species
Trichomonas vaginalis
Trichomonas tenax
Trichomonas gallinae
Pentatrichomonas hominis
Tetratrichomonas gallinarum
Tritrichomonas foetus
Trichomitus batrachorum
Host Illustration
Host-niche (host-relationship)
Human urogenital tract (Parasite)
Human oral cavity (Possible parasite)
Columbiforme, falcinoforme birds upper GI tract (Parasite)
Human lower GI tract (Commensal)
Gallinaceous, anseriform birds lower GI tract (Possible parasite)
Bovine urogenital and cat and pig GI (Parasite in cat and cow. Commensal in pig)
Toad GI tract (Presumed commensal)

Figure 4.1 Diversity of trichomonad hosts and host-niches.
Taxonomical relationships of trichomonad species used in this study. Species are presented with an illustration of the best-known host, along with a description of the trichomonad’s habitat within the host and a statement of its parasitic status. Made according to the NCBI taxonomy. Host-parasite assignments made per the following sources: (1, 12, 62–66).

Figure 4.2a.

Figure 4.2b.

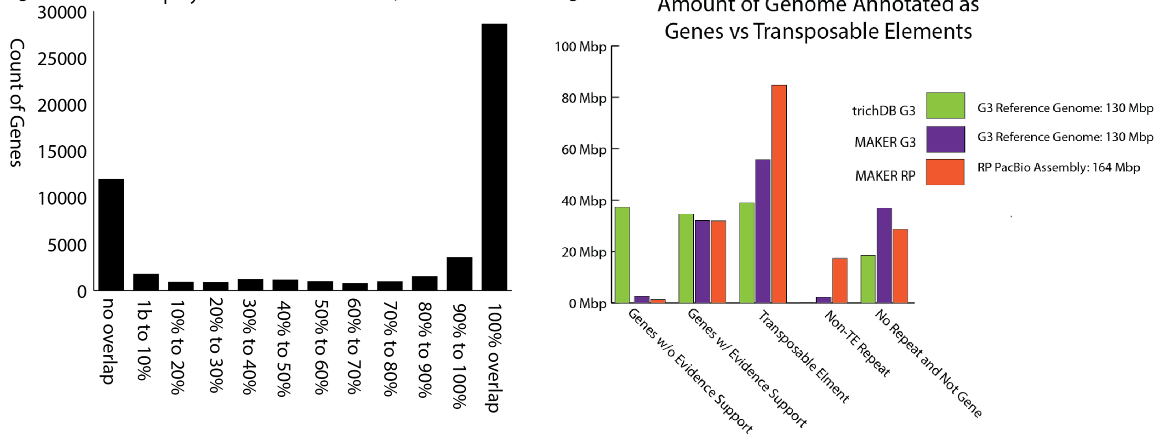
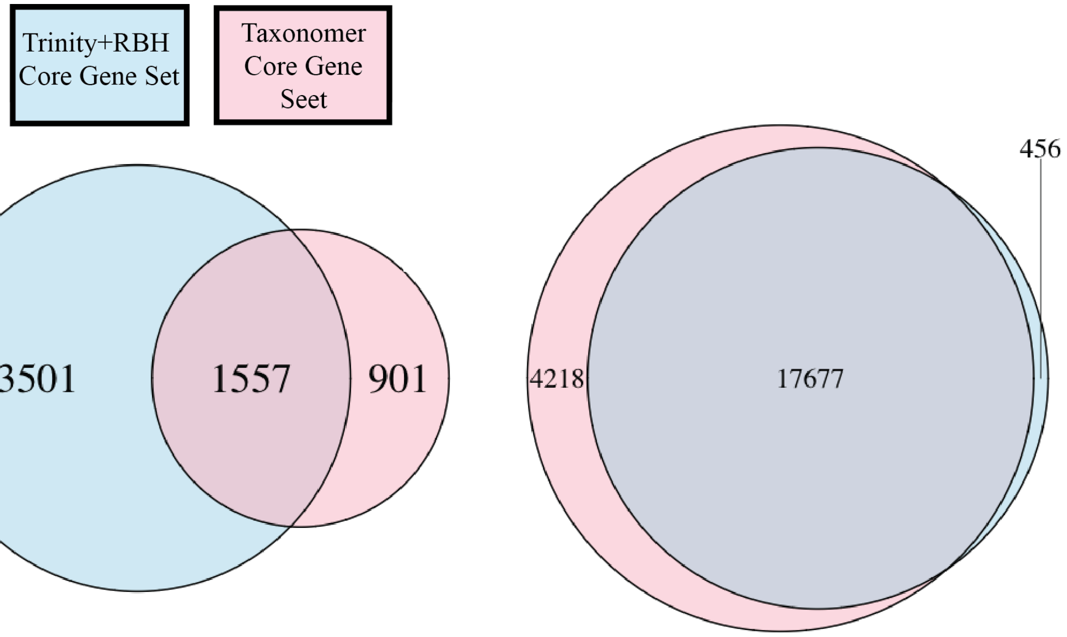


Figure 4.2 Impact of custom TE library on genome annotation of *T. vaginalis*. Fig. 4.2a) Overlap of trichDB genes by regions identified as TEs by the *T. vaginalis*-specific TE library. Fig. 4.2b) Amount of *T. vaginalis* Genome annotated as a gene without evidence support (EST or protein), a gene with evidence support, TE, non-TE repetitive sequence, and none of the above.



Core Gene Set of 7 Trichomonad Species

Core Gene Set of 8 *T. vaginalis* strains

Figure 4.3 Comparison of gene sets identified by Trinity+RBH or Taxonomer methods. Core gene sets within the *T. vaginalis* strains and within the 7 trichomonad species were identified by the Taxonomer-based method and the Trinity+RBH method. Overlap between the core gene sets were identified by RBH to MAKER annotated protein-coding genes.

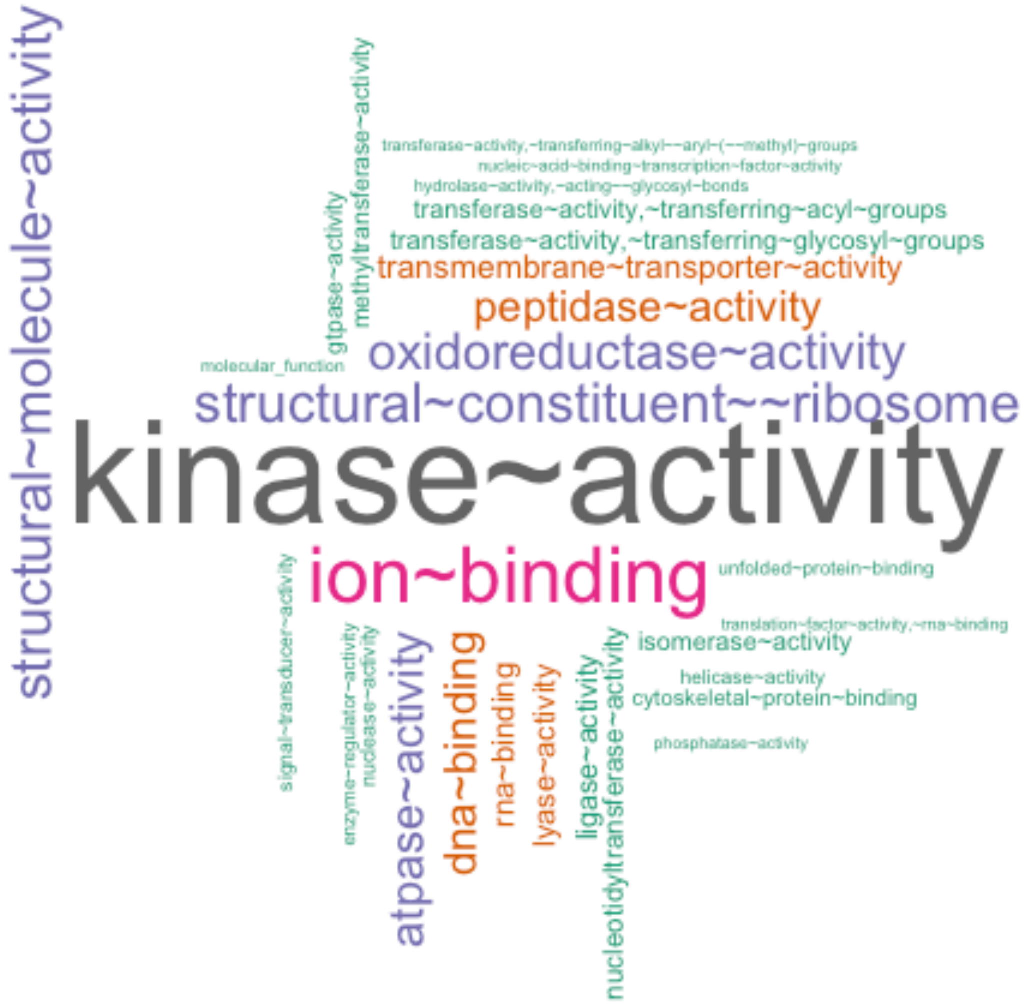


Figure 4.4 Word cloud of GO terms in Taxonomer trichomonad core gene set.

Table 4.1 WGS sample classification results

Sample	Percent of Reference Sequence Length Recovered at Greater Than or Equal to Cutoff					
	> 99%	> 90%	> 75%	> 50%	> 25%	> 0%
Tvag.G3						
Tvag.30235	0.56	0.65	0.66	0.67	0.68	0.69
Tvag.30238	0.32	0.35	0.37	0.4	0.43	0.45
Tvag.50143	0.31	0.35	0.37	0.4	0.43	0.45
Tvag.B7RC2	0.3	0.34	0.36	0.4	0.43	0.45
Tvag.JRSTV41	0.3	0.34	0.36	0.4	0.43	0.46
Tvag.t016	0.37	0.52	0.57	0.66	0.72	0.75
Tvag.T1	0.29	0.33	0.36	0.4	0.43	0.46
Trich. tenax	0	0	0.01	0.02	0.04	0.05
Pentatrich. hominis	0	0.01	0.02	0.04	0.07	0.09
Tritrich. foetus	0	0.01	0.01	0.03	0.06	0.07

Table 4.2 RNAseq sample classification results

Sample	Percent of Reference Sequence Recovered at Greater Than or Equal to Cutoff					
	> 99%	> 90%	> 75%	> 50%	> 25%	> 0%
Tvag.B7RC2	0.13	0.4	0.53	0.64	0.75	0.84
Trich.tenax	0.01	0.04	0.07	0.15	0.31	0.54
Trich. gallinae	0.01	0.03	0.06	0.13	0.28	0.5
Tetratrich. gallinar.	0.0	0.0	0.01	0.03	0.1	0.26
Pentatrich. hominis	0.0	0.0	0.01	0.03	0.11	0.27
Tritrich. foetus bovine	0	0	0	0.01	0.04	0.15
Tritrich. foetus feline	0	0	0	0.01	0.04	0.15
Trichomit. batrach.	0	0	0.01	0.03	0.1	0.26

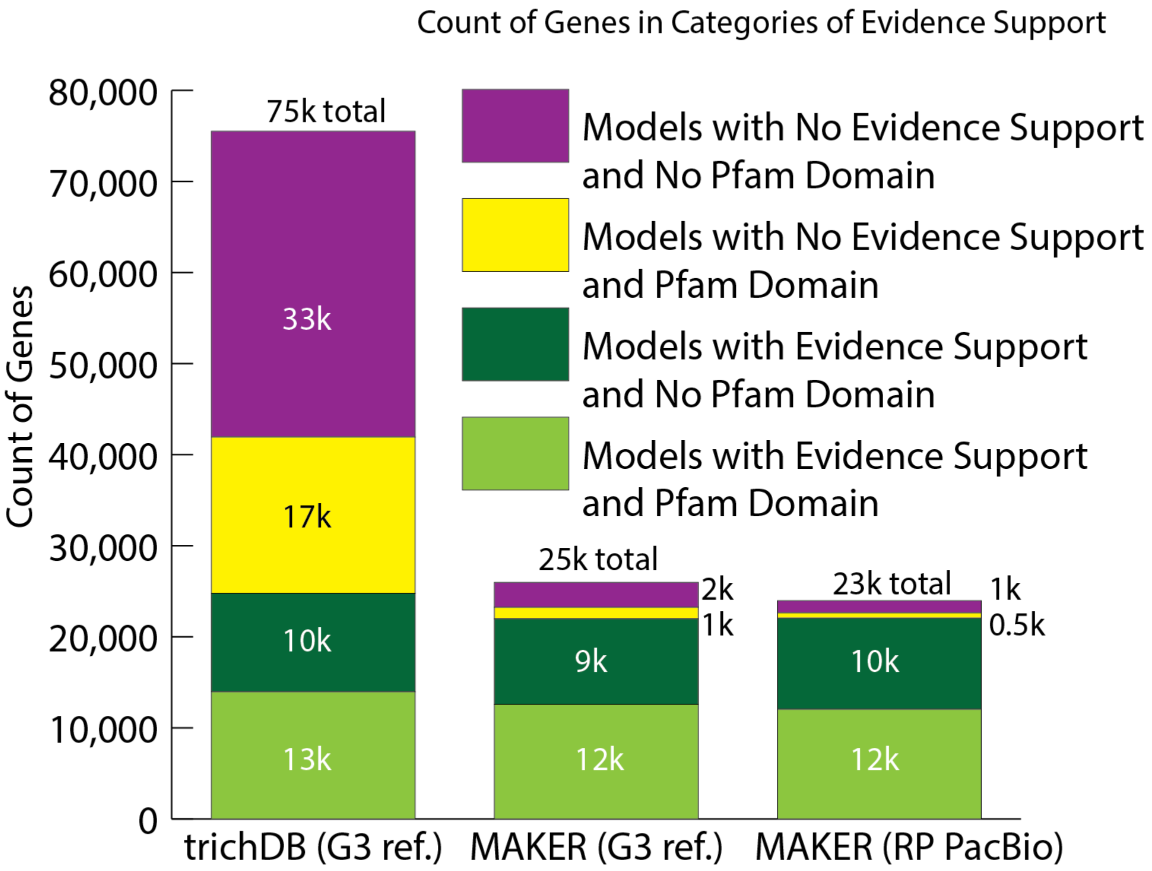


Figure 4S.1 Evidential support for MAKER and trichDB gene models.
Evidence support for the gene models of the annotated gene sets (trichDB on G3 reference assembly, MAKER on G3 reference assembly, and MAKER on RP PacBio-based assembly) was assessed by overlap of gene models by an expressed sequence alignment, protein alignment, or presence of a Pfam protein domain.

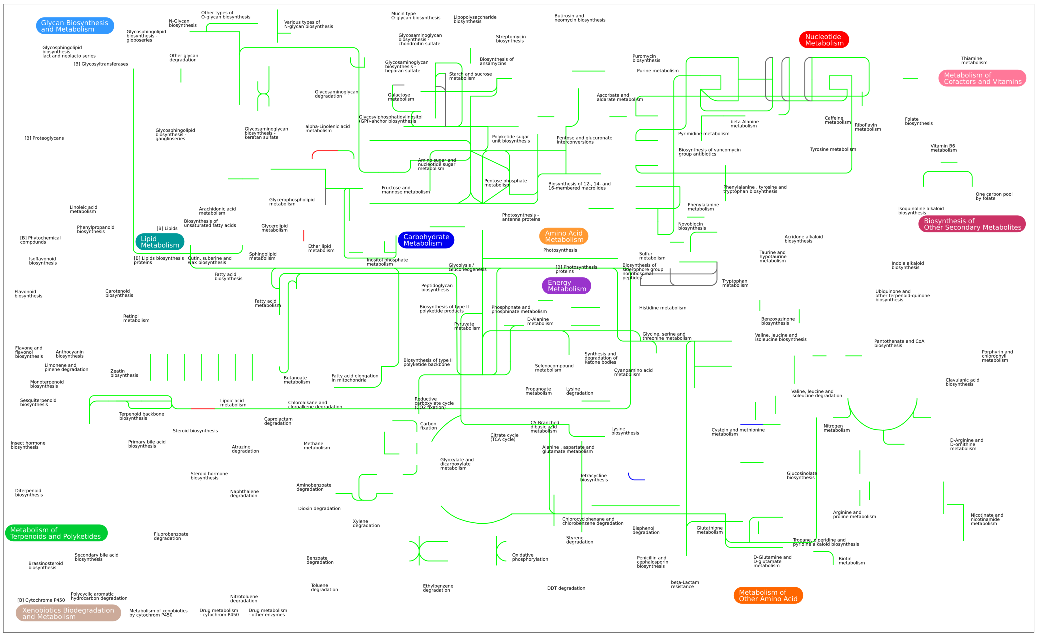

Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Biosynthesis of Other Secondary Metabolites
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid

Figure 4S.2 KEGG metabolic pathways in MAKER vs. trichDB models.
KEGG assignment of MAKER G3 and trichDB models was done with the KEGG Automated Annotation Server (KAAS) and visualized with the Pathways tool on the *T. vaginalis*-specific subset of metabolic pathways. Pathway segments in both MAKER and trichDB are colored green; segments in only MAKER are colored blue; segments only in trichDB are colored red.

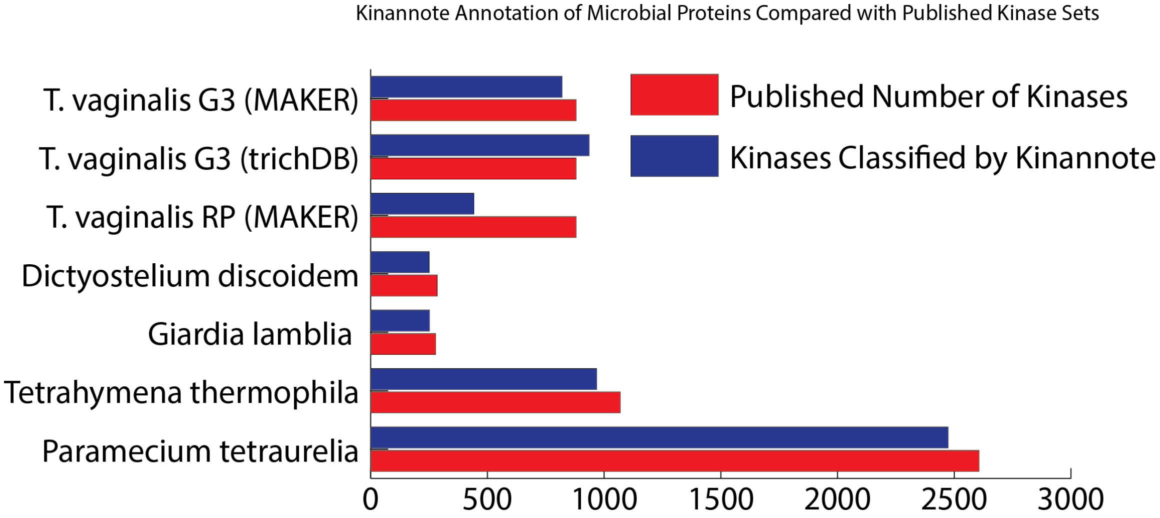


Figure 4S.3 Kinome annotation of annotated microbial genes.
Kinases genes in annotated gene sets were annotated with Kinannoate and compared to previously published estimates of kinases genes for each species.

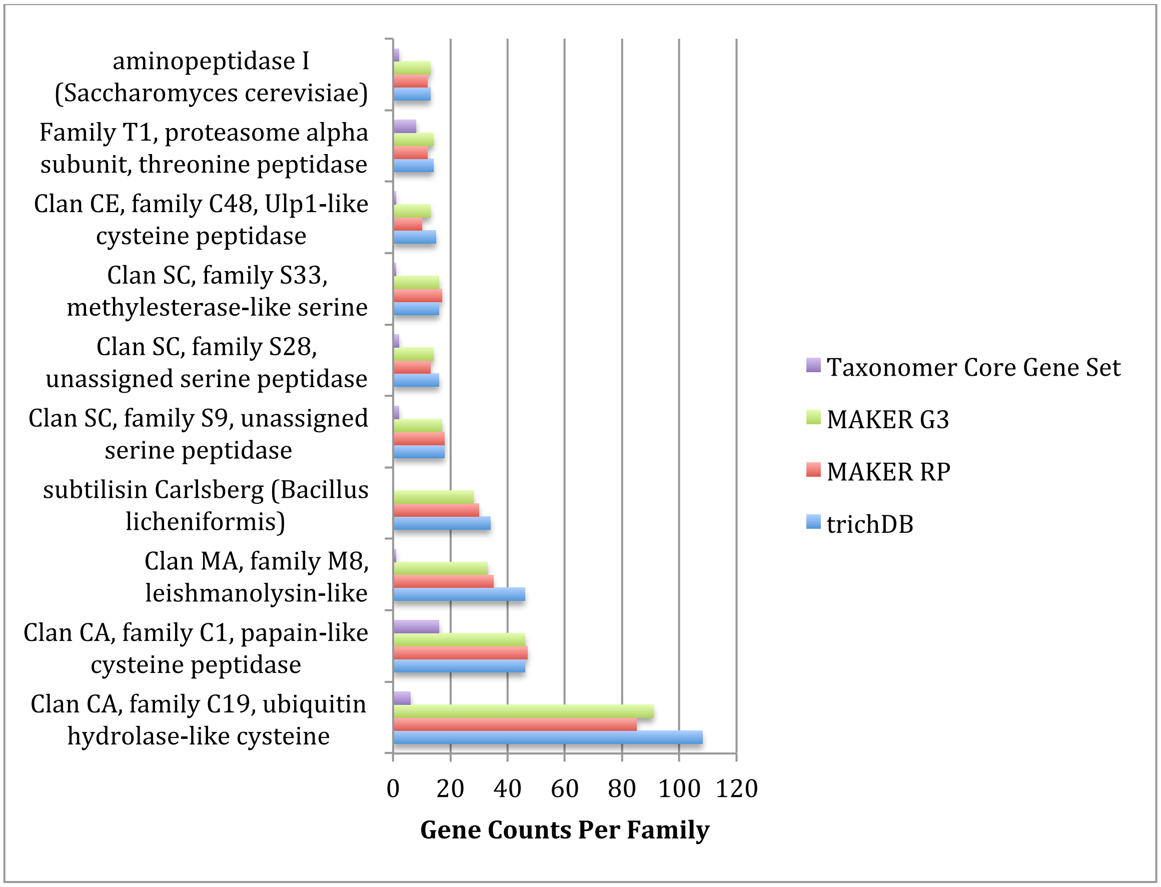


Figure 4S.4 Degradome gene family counts in annotated gene sets.
Degradome family assignment was done with blast searches using the MEROPS server.

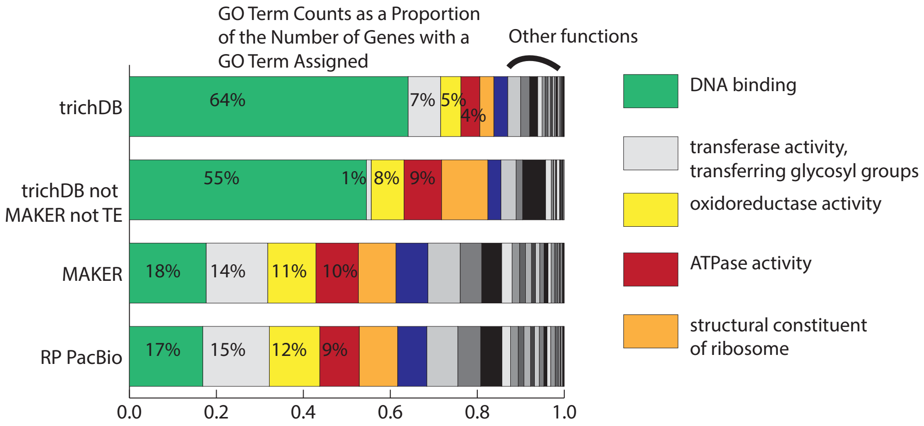


Figure 4S.5 Comparison of GO term counts between trichDB, MAKER G3, and MAKER RP gene sets.
GO Terms were mapped to protein domains using InterPro2GO, and then reduced with the GO generic slim file with AmiGO’s slimmer tool.

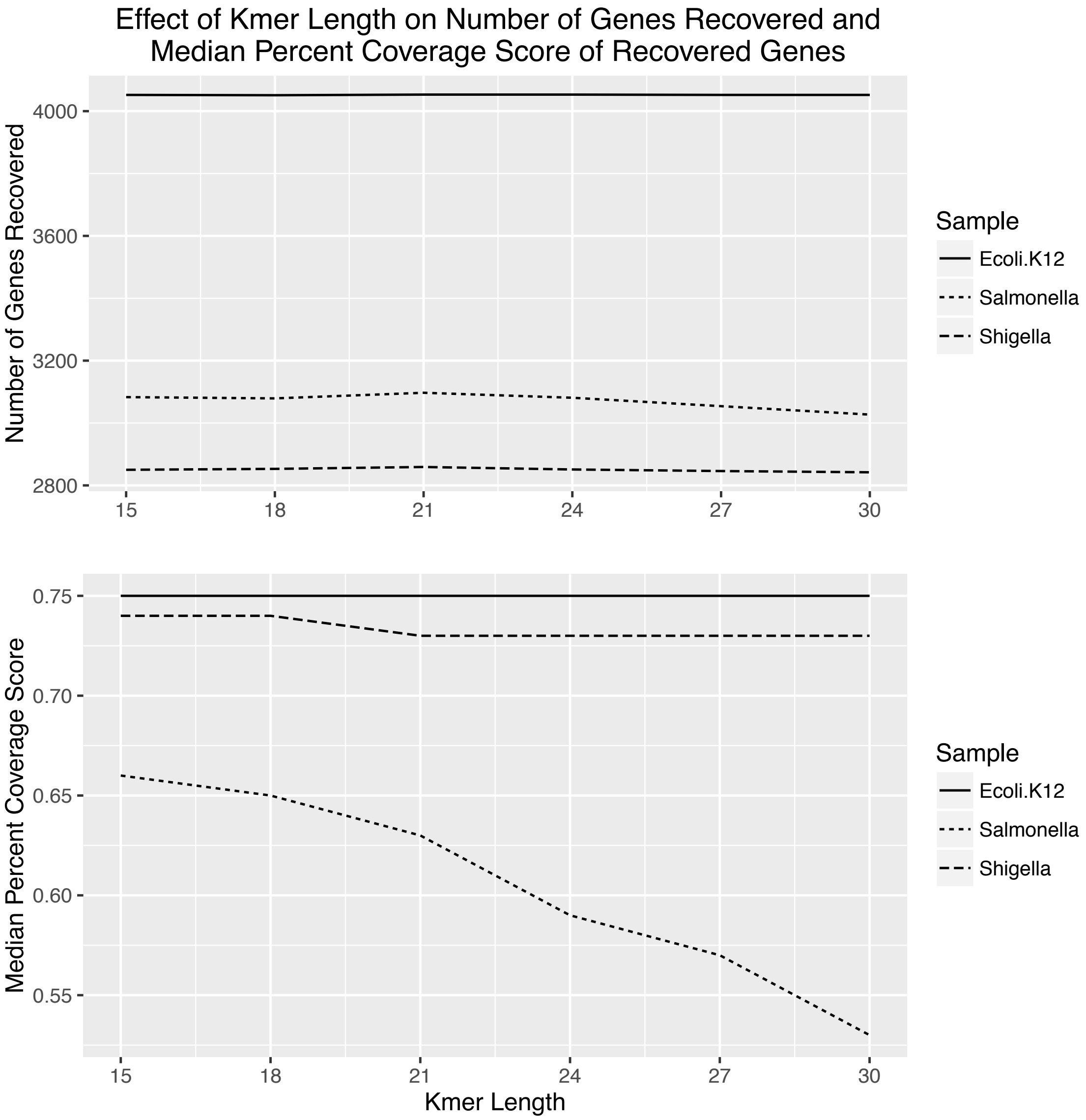


Figure 4S.6 Summary of proof-of-concept results with simulated 5X coverage on bacterial genes.

Simulated 5X coverage reads from three species were classified to E.coli.K12 genes using a range of kmer lengths. Number of genes recovered and the median percent coverage of the reference sequences were assessed with a blastx query of the recovered contigs against the Ecoli.K12 reference sequences.

Table 4S.1 Summary and sources for kinases

Dataset	#Proteins	#Kinases by Kinannote	#Kinases by Source	%Deviation from Source	Source
T.vaginalis (MAKER)	25,936	820	880	6.82%	Carlton et al. 2007
T.vaginalis (trichDB)	75,499	1,122	880	27.5%	Carlton et al. 2007
Dictyostelium discoideum	12,318	328	285	15.1%	http://kinase.com
Tetrahymena thermophila	26,996	1,193	1,069	5.53%	http://kinase.com
Paramecium tetraurelia	39,519	2,750	2,606	3.64%	http://kinase.com

Table 4S.2 MAKER equivalents of meiosis-specific genes in trichDB

(query) trichDB ID	(subject) MAKER ID	E-value
TVAG_455180	snap_masked-DS113894-processed-gene-0.39-mRNA-1	1.30E-114
TVAG_151700	augustus_masked-DS113425-processed-gene-0.15-mRNA-1	5.90E-119
TVAG_258950	augustus_masked-DS114140-processed-gene-0.1-mRNA-1	1.00E-206
TVAG_155030	augustus_masked-DS114127-processed-gene-0.4-mRNA-1	1.90E-174
TVAG_230730	augustus_masked-DS113364-processed-gene-0.18-mRNA-1	5.20E-174
TVAG_472000	augustus_masked-DS114293-processed-gene-0.0-mRNA-1	0
TVAG_058400	augustus_masked-DS113455-processed-gene-0.4-mRNA-1	3.30E-118
TVAG_292060	augustus_masked-DS113233-processed-gene-1.10-mRNA-1	0
TVAG_062830	genemark-DS113216-processed-gene-1.38-mRNA-1	1.40E-109

Table 4S.3 Reference-free comparative genomics results from bacteria

Sample	> 99%	> 90%	> 75%	> 50%	> 25%	> 0%
Ecoli.K12	0.88	0.98	0.98	0.98	0.98	0.98
Shigella	0.56	0.65	0.66	0.67	0.68	0.69
Salmonella	0.37	0.52	0.57	0.66	0.72	0.75

Table 4S.4 Top 7 ‘slimmed’ molecular function GO term counts in the *T. vaginalis* core gene sets

GO Term	Trinity+RBH	Taxonomer	Common to Both
Kinase Activity	946	946	946
Ion Binding	389	389	389
Peptidase Activity	315	315	315
Glycosyl Transferase Activity	198	198	198
Transmembrane Transporter Activity	139	139	139
Oxidoreductase Activity	110	110	110
Enzyme Regulator Activity	105	104	104

Table 4S.5 Top 7 ‘slimmed’ molecular function GO term counts in the trichomonad core gene sets

GO Term	Trinity+RBH	Taxonomer	Common to Both
Kinase Activity	286	316	181
Ion Binding	99	159	63
Peptidase Activity	64	124	41
Glycosyl Transferase Activity	57	61	30
Transmembrane Transporter Activity	49	48	22
Oxidoreductase Activity	48	41	20
Enzyme Regulator Activity	45	40	17

References

1. Carlton JM, et al. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315(5809):207–12.

2. Aurrecoechea C, et al. (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. *Nucleic Acids Res* 37(Database issue):D526-30.

3. Barratt J, Gough R, Stark D, Ellis J (2016) Bulky Trichomonad Genomes: Encoding a Swiss Army Knife. *Trends Parasitol* 32(10):783–797.

4. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database

management tool for second-generation genome projects. *BMC Bioinformatics* 12(1):491.

5. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13(5):329–342.

6. Hampl V, et al. (2005) Inference of the phylogenetic position of oxymonads based on nine genes: support for metamonada and excavata. *Mol Biol Evol* 22(12):2508–18.

7. Satterwhite CL, et al. (2013) Sexually transmitted infections among US women and men: prevalence and incidence estimates, 2008. *Sex Transm Dis* 40(3):187–93.

8. Rae DO, Crews JE, Greiner EC, Donovan GA (2004) Epidemiology of Tritrichomonas foetus in beef bull populations in Florida. *Theriogenology* 61(4):605–618.

9. Mendoza-Ibarra JA, et al. (2012) High prevalence of Tritrichomonas foetus infection in Asturiana de la Monta??a beef cattle kept in extensive conditions in Northern Spain. *Vet J* 193(1):146–151.

10. Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390(1–2):3–17.

11. Barratt JLN, Harkness J, Marriott D, Ellis JT, Stark D (2011) The ambiguous life of Dientamoeba fragilis: the need to investigate current hypotheses on transmission. *Parasitology* 138(5):557–572.

12. Morin-Adeline V, et al. (2014) Comparative transcriptomics reveals striking similarities between the bovine and feline isolates of Tritrichomonas foetus: consequences for in silico drug-target identification. *BMC Genomics* 15(1):955.

13. Woehle C, et al. (2014) The parasite Trichomonas vaginalis expresses thousands of pseudogenes and long non-coding RNAs independently from functional neighbouring genes. *BMC Genomics* 15(1):906.

14. Gould SB, et al. (2013) Deep sequencing of Trichomonas vaginalis during the early infection of vaginal epithelial cells and amoeboid transition. *Int J Parasitol* 43(9):707–719.

15. Conrad MD, Bradic M, Warring SD, Gorman AW, Carlton JM (2013) Getting trichy: Tools and approaches to interrogating Trichomonas vaginalis in a post-genome world. *Trends Parasitol* 29(1):17–25.

16. Cantarel BL, et al. (2008) MAKER: an easy-to-use annotation pipeline designed

for emerging model organism genomes. *Genome Res* 18(1):188–96.

17. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12(1):491.

18. Campbell MS, et al. (2014) MAKER-P : A Tool Kit for the Rapid Creation , Management , and Quality Control of Plant. 164(February):513–524.

19. Flygare S, et al. (2016) Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol* 17(1):111.

20. Aurrecoechea C, et al. (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. *Nucleic Acids Res* 37(Database issue):D526-30.

21. Goldberg JM, et al. (2013) Kinannote, a computer program to identify and classify members of the eukaryotic protein kinase superfamily. *Bioinformatics* 29(19):2387–2394.

22. Malik S-B, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM (2008) An expanded inventory of conserved meiotic genes provides evidence for sex in Trichomonas vaginalis. *PLoS One* 3(8):e2879.

23. Carlton JM, et al. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315(5809):207–12.

24. Strese Å, Backlund A, Alsmark C (2014) A recently transferred cluster of bacterial genes in Trichomonas vaginalis - lateral gene transfer and the fate of acquired genes. 14(1):1–13.

25. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13(5):329–342.

26. Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. *J Hered* 100(5):648–55.

27. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(SUPPL. 1):351–358.

28. Gene Codes Corporation Sequencher. Available at: http://www.genecodes.com/.

29. Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol Evol*

1:205–20.

30. ExPASy - Translate tool Available at: http://web.expasy.org/translate/.

31. ORF Finder.

32. Marchler-Bauer A, et al. (2005) CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res* 33(DATABASE ISS.):192–196.

33. Wicker T, et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8(12):973–982.

34. ClustalW2.

35. Rastogi PA (1999) MacVector. *Bioinformatics Methods and Protocols*, eds Misener S, Krawetz SA (Humana Press, Totowa, NJ), pp 47–69.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.

37. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 5(32):1792–1797.

38. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24(8):1596–1599.

39. Smit A, Hubley R, Green P RepeatMasker. Available at: http://www.repeatmasker.org/.

40. Puigbò P, Bravo IG, Garcia-Vallve S (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* 3:38.

41. Nakamura Y, Gojobori T, Ikemura T (1999) Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic Acids Res* 27(1):292.

42. Chin C-S, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10(6):563–569.

43. Fang Y-K, et al. (2014) Gene-expression analysis of cold-stress response in the sexually transmitted protist Trichomonas vaginalis. *J Microbiol Immunol Infect* 48(6):662–675.

44. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–52.

45. NCBI Blast+ Available at: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.

46. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH (2015) UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6):926–932.

47. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 18(12):1979–90.

48. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.

49. Stanke M, Tzvetkova A, Morgenstern B (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* 7 Suppl 1(May 2005):S11.1-8.

50. Hoff KJ, Stanke M (2013) WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res* 41(Web Server issue):W123-8.

51. Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.

52. Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res* 37(1):289–297.

53. Vanácová S, Yan W, Carlton JM, Johnson PJ (2005) Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. *Proc Natl Acad Sci U S A* 102(12):4430–4435.

54. Hunter S, et al. (2012) InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res* 40(D1):1–7.

55. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35(SUPPL.2):182–185.

56. Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci* 33(3):101–103.

57. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) IPath2.0: Interactive pathway explorer. *Nucleic Acids Res* 39(SUPPL. 2):412–415.

58. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6):863–864.

59. seqyclean Available at: https://github.com/ibest/seqyclean.

60. TransDecoder.

61. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.

62. Tachezy J, et al. (2002) Cattle pathogen tritrichomonas foetus (Riedmüller, 1928) and pig commensal Tritrichomonas suis (Gruby & Delafond, 1843) belong to the same species. *J Eukaryot Microbiol* 49(2):154–63.

63. Carlos L, Santos C, Benchimol M (2015) Is Trichomonas tenax a Parasite or a Commensal ? *Ann Anat* 166(2):196–210.

64. Stabler RM (1954) Trichomonas gallinae: a review. *Exp Parasitol* 3(4):368–402.

65. Cepicka I, Hampl V, Kulda J, Flegr J (2006) New evolutionary lineages, unexpected diversity, and host specificity in the parabasalid genus Tetratrichomonas. *Mol Phylogenet Evol* 39(2):542–551.

66. Zubácová Z, Cimbůrek Z, Tachezy J (2008) Comparative analysis of trichomonad genome sizes and karyotypes. *Mol Biochem Parasitol* 161(1):49–54.

CHAPTER 5

CONCLUSIONS AND PERSPECTIVES

Genome Projects with Non-Model Organisms

Genome annotation is a process that in large part depends on leveraging resources generated from prior research on the organism of interest or closely related species (1). This presents problems when the organism is a member of a lineage poorly represented in public repositories or is highly diverged from the organisms that are available or both. The problems encountered include a lack of knowledge of novel transposable elements (TEs) in the genome of interest and a lack of gene models that can be leveraged for *ab initio* gene predictors and homology searches.

In spite of these problems, many species that impact human health and agriculture fall into these categories. For example, the genome of the global human pathogen (2) *Trichomonas vaginalis* (*T. vaginalis*) was the first genome of its order (trichomonadida) to be published (3) and lacked many of the resources available to better sampled lineages. *Cronartium quercuum* f. sp. *fusiforme* (*Cqf*) and other congenic species (*Cronartium rubicola*) have a large economic impact, but *Cqf* is the only genome of its genus to be sequenced (4–6). In contrast, the annotation of a new insect genome will benefit from the 80 other insect genomes and 1.2M genes from those genomes that have already been annotated (7).

Novel abundant TEs are always a potential pitfall for genome annotation projects. Although tools to automate the process of identifying repetitive sequence in a new genome do exist (i.e., RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html; also see Yandell and Ence, 2012 for a list of other tools), a thorough but time-intensive approach, such as was undertaken for our *T. vaginalis*-specific TE library, will achieve a precision and accuracy not possible with automated methods. The 2007 genome annotation of *T. vaginalis* presents an extreme example of this problem. The *Mavericks* family of TEs was not identified until after the genome annotation had been published, but made an extensive impact on the genome annotation results. Approximately 30,000 annotated protein-coding gene models were entirely overlapped by elements from our TE library. In contrast, an automated tool such as RepeatModeler appears to have sufficed for the annotation of the small and TE-rich genome of *Cqf*, since gene counts for *Cqf* are within the expected range based on genome annotations of other rust fungi (8, 9).

The genome annotation projects presented in this thesis provide a clear contrast between a project that has data resources to leverage and a project that lacks those resources. The annotation of *Cqf* benefited from a thorough RNAseq dataset that sampled several lifecycle stages on both its hosts, as well as from leveraging the genome annotations of several other rust fungi (8, 9). In contrast, the 2007 genome annotation of *T. vaginalis* did not have complete genomes or transcriptomes from any other parabasalid, and had to use the then available resources of the distantly related *Dictyostelium discoideum*, *Giardia lamblia*, and *Entamoeba histolytica*. Our annotation of the *T. vaginalis* genome with the updated TE library made use of the same genome annotations used in the 2007 annotations, but also used thorough RNAseq datasets of *T. vaginalis* strain t016 at several time points under different environmental conditions (10–12). In fact, these recent studies already suggested that many of the protein-coding genes annotated in *T. vaginalis* are not expressed, which our re-annotation study demonstrated.

The Role of Improvements in Sequencing Technology and Analysis Methods

The genome studies in this thesis highlight the role that new sequencing technologies like PacBio SMRT sequencing and other long-read technologies can play in genomics for non-model organisms (13–16). The *de novo* assembly of *T. vaginalis* strain RP with PacBio reads is approximately the same size as the Sanger-based G3 reference assembly, but has an N50 length two times longer (3). The fact that the long-read based assembly appears to have both resolved the long TEs in the *T. vaginalis* genome and merged the kinases in this genome to half their previous number shows how this technology has changed views of the biology of this organism. Likewise, future publications on the *Cqf* genome will make use of a PacBio-based assembly of *Cqf* made in order to resolve gaps in the assembly located near the *Avr1* identified by our selective sweep searchers.

The development of Illumina short-read sequencing also contributed to both of these projects. The RNASeq datasets used in both studies are a relatively recent development in genomics. The whole genome sequence (WGS) samples of several different *T. vaginalis* strains are also a relatively recent development over past efforts, which relied on PCR assays of a set of validated single-copy genes in the G3 reference genome (17).

In addition to new sequencing technologies, advances in sequence analysis allow for new and exciting questions to be asked and answered. The MAKER genome annotation pipeline (18–20) used for annotating both of the genomes in this thesis integrates all the steps of genome annotation (masking of repeats, *ab initio* gene prediction, and nucleotide and protein homology searches). It outputs files that are compatible with downstream tools to facilitate visualization and manual curation of gene models (1). It also provides annotation edit distance, AED, along with other quality control metrics to help human curators of gene models to prioritize their tasks (18, 21). All these features allow genome annotation projects to be conducted by scientists in their own fields, but with minimal bioinformatics or computation expertise.

In addition to improvements in genome annotation, the emergence of kmer-based methods like Taxonomer and others promise to allow fast and accurate exploration of large gene spaces, not only in the metagenomics field often targeted initially, but also in comparative genomics studies as was presented for *T. vaginalis* here (22–27). Although each of these methods differ in implementation and performance, in general, kmer-based searches in "protein space" allow scientists to complete the equivalent of ten of millions of blastx or tblastx searches in a small fraction of the time that the equivalent blast searches would take.

Summary and Future Directions

In summary, advances in genome annotation and the emergence of kmer-based sequence analysis techniques were both essential to the success of the genome annotations projects included in this thesis. The challenges inherent to genome annotation projects in phylogenetically distant and understudied organisms were overcome with a combination of expert analysis to identify a new family of TEs and quantify their abundance in the *T. vaginalis* genome and application of new methods and data sources. The results from those analyses resolved questions regarding both the TE content of the *T. vaginalis* genome and the number of protein-coding genes in the genome. The resolution of those questions allowed for the investigation of comparative genomics questions among seven different trichomonad species through the application of an ultrafast and accurate kmer-based read classifier tool, Taxonomer (23).

Future directions to extend on this work include further analysis of candidate *Avr1* genes as well as candidate *Fr1* genes in the loblolly pine genome. Future directions for the trichomonad core gene set work include improved parameters to increase sensitivity at further phylogenetic distances, as well as examination of genes specific to – or absent from – species that inhabit certain taxa, niches, or that are parasitic vs. commensal in their hosts.

References

1. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13(5):329–342.

2. Satterwhite CL, et al. (2013) Sexually transmitted infections among US women and men: prevalence and incidence estimates, 2008. *Sex Transm Dis* 40(3):187–93.

3. Carlton JM, et al. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315(5809):207–12.

4. Knight HA, Dutrow GF (1970) Incidence and financial impact of fusiform rust in the South. *J For* 72(7):398–401.

5. Cubbage FW, Pye JM, Holmes TP, Wagner JE (2000) An economic evaluation of Fusiform rust protection research. *South J Appl For* 24(2):77–85.

6. Richardson BA, Zambino PJ, Klopfenstein NB, McDonald GI, Carris LM (2007) Assessing host specialization among aecial and telial hosts of the white pine blister

rust fungus, Cronartium ribicola. *Can J Bot* 85(3):299–306.

7. Waterhouse RM (2015) ScienceDirect A maturing understanding of the composition of the insect gene repertoire. *Curr Opin Insect Sci* 7(January):15–23.

8. Duplessis S, et al. (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci U S A* 108(22):9166–71.

9. Nemri A, et al. (2014) The genome sequence and effector complement of the flax rust pathogen Melampsora lini. *Front Plant Sci* 5(March):98.

10. Woehle C, et al. (2014) The parasite Trichomonas vaginalis expresses thousands of pseudogenes and long non-coding RNAs independently from functional neighbouring genes. *BMC Genomics* 15(1):906.

11. Gould SB, et al. (2013) Deep sequencing of Trichomonas vaginalis during the early infection of vaginal epithelial cells and amoeboid transition. *Int J Parasitol* 43(9):707–719.

12. Fang Y-K, et al. (2014) Gene-expression analysis of cold-stress response in the sexually transmitted protist Trichomonas vaginalis. *J Microbiol Immunol Infect* 48(6):662–675.

13. Zhang W, Ciclitira P, Messing J (2014) Pacbio sequencing of gene families - A case study with wheat gluten genes. *Gene* 533(2):541–546.

14. Lee H, Gurtowski J, Yoo S (2014) Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv*:1–17.

15. English AC, et al. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 7(11):1–12.

16. Goodwin S, et al. (2015) Oxford Nanopore sequencing and de novo assembly of a eukaryotic genome. *bioRxiv*:13490.

17. Conrad MD, et al. (2012) Extensive genetic diversity, unique population structure and evidence of genetic exchange in the sexually transmitted parasite Trichomonas vaginalis. *PLoS Negl Trop Dis* 6(3):e1573.

18. Campbell MS, et al. (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol* 164(2):513–24.

19. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12(1):491.

20. Cantarel BL, et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18(1):188–96.

21. Eilbeck K, Moore B, Holt C, Yandell M (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 15:1–15.

22. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* 32(5):462–4.

23. Flygare S, et al. (2016) Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol* 17(1):111.

24. Pérez N, Gutierrez M, Vera N (2016) Computational performance assessment of k-mer counting algorithms. *J Comput Biol* 23(4):cmb.2015.0199.

25. Chor B, Horn D, Goldman N, Levy Y, Massingham T (2009) Genomic DNA k-mer spectra: models and modalities. *Genome Biol* 10(10):R108.

26. Rizk G, Lavenier D, Chikhi R (2013) DSK: K-mer counting with very low memory usage. *Bioinformatics* 29(5):652–653.

27. Audano P, Vannberg F (2014) KAnalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics* 30(14):2070–2072.

APPENDIX A

TRANSPOSABLE ELEMENT ISLANDS FACILITATE ADAPTATION TO NOVEL ENVIRONMENTS IN AN INVASIVE SPECIES

This appendix is a reprint of a research article coauthored by Lukas Schrader, Jay W. Kim, myself, Aleksey Zimin, Antonia Klein, Katharina Wyschetzki, Tobias Weicheselgartner, Carsten Kemena, Johannes Stökl, Eva Schultner, Yannick Wurm, Christopher D. Smith, Mark Yandell, Jürgen Heinze, Jürgen Gadau, and Jan Oettler and is presented here with permissions of the authors and kind permission of Springer Nature.

I contributed the genome annotations of *Cardiocondyla obscurior* that formed the basis of the genomic research presented in this article. This research article was first published in Schrader et al. (2014) Transposable Element Islands Facilitate Adaptation to Novel Environments in an Invasive Species. *Nature Communications* 16(5):1-10. Available at: http://www.nature.com/nrg/journal/v13/n5/abs/nrg3174.html.

nature COMMUNICATIONS

ARTICLE

Received 14 Apr 2014 | Accepted 6 Oct 2014 | Published 16 Dec 2014

DOI: 10.1038/ncomms6495

OPEN

Transposable element islands facilitate adaptation to novel environments in an invasive species

Lukas Schrader[1,*], Jay W. Kim[2], Daniel Ence[3], Aleksey Zimin[4], Antonia Klein[1], Katharina Wyschetzki[1], Tobias Weichselgartner[1], Carsten Kemena[5], Johannes Stökl[1], Eva Schultner[6], Yannick Wurm[7], Christopher D. Smith[8], Mark Yandell[3,9], Jürgen Heinze[1], Jürgen Gadau[10] & Jan Oettler[1,*]

Adaptation requires genetic variation, but founder populations are generally genetically depleted. Here we sequence two populations of an inbred ant that diverge in phenotype to determine how variability is generated. *Cardiocondyla obscurior* has the smallest of the sequenced ant genomes and its structure suggests a fundamental role of transposable elements (TEs) in adaptive evolution. Accumulations of TEs (TE islands) comprising 7.18% of the genome evolve faster than other regions with regard to single-nucleotide variants, gene/exon duplications and deletions and gene homology. A non-random distribution of gene families, larvae/adult specific gene expression and signs of differential methylation in TE islands indicate intragenomic differences in regulation, evolutionary rates and coalescent effective population size. Our study reveals a tripartite interplay between TEs, life history and adaptation in an invasive species.

[1] Institut für Zoologie, Universität Regensburg, 93053 Regensburg, Germany. [2] Department of Biomolecular Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA. [3] Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA. [4] Institute for Physical Sciences and Technology, University of Maryland, College Park, Maryland 20742, USA. [5] Institute for Evolution and Biodiversity, Westfälische Wilhelms-Universität, 48149 Münster, Germany. [6] Department of Biosciences, University of Helsinki, 00014 Helsinki, Finland. [7] School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK. [8] Department of Biology, San Francisco State University, San Francisco, California 94132, USA. [9] Utah Center for Genetic Discovery, University of Utah, Salt Lake City 84112, USA. [10] School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.O. (email: joettler@gmail.com).

Depletion of genetic variation is detrimental to species evolution and adaptation[1]. Low genetic and phenotypic variation is common in founder populations, where only one or a few genotypes are isolated from a source population. Under such conditions, reduced effective population size (N_{e}) should decrease selection efficiency and increase genetic drift, resulting in only weak selection against mildly deleterious alleles which can thus accumulate[2]. These effects should be even stronger in inbreeding species[3] and taxa with generally low N_{e} such as social insects[4]. Despite these constraints on adaptive evolution, many inbred or selfing species thrive and are able to invade novel habitats. This raises the question of how genetic variation as the raw material for adaptation is generated in such systems.

Single-nucleotide substitutions are an important factor in adaptation[5] and species diversification[6,7]. However, other structural and regulatory units, such as transposable elements (TEs) and epigenetic modifications, may act as drivers in adaptation and evolution[8]. TEs play a particularly vital role in genome evolution[9] and recurringly generate adaptive phenotypes[10–13] primarily through (retro-)transposition[14], and secondarily through ectopic recombination and aberrant transposition[15].

The invasive, inbreeding ant *Cardiocondyla obscurior* (Fig. 1) provides a suitable model to study how species adapt to novel habitats in spite of constraints imposed by invasion history, life history or both. Originally from Southeast Asia, *C. obscurior* has established populations in warm climates around the globe from founder populations that presumably consisted of only one or a few inbred colonies, each with a few reproductive queens and several dozen sterile workers. In this species, related wingless males and females (queens) mate within the colony, after which queens leave the colony with a group of workers to find a new nest nearby. While greatly reducing the extent of gene flow between colonies, this behaviour enables sexual reproduction within the same colony and allows single founder colonies to rapidly colonize novel habitats. At the same time, the combination of prolonged inbreeding with severe genetic bottlenecks strongly reduces N_{e} in this species. Under such conditions, genetic drift is predicted to drastically deplete genetic variation, thus leaving little for selection to act on.

Figure 1 | Two workers of *C. obscurior* and the remains of a fly. Hidden in small cavities of plants, the inconspicuous colonies of this species are frequently introduced to new habitats by global commerce. In spite of strong genetic bottlenecks, even single colonies with few reproductive individuals suffice to establish stable populations.

Here we explore the genomes of *C. obscurior* from two invasive populations (Brazil BR and Japan JP) to identify signatures of divergence on a genomic level and to determine how the species can rapidly adapt to different habitats. We find clear phenotypic differences between the populations and strong correlation between accumulations of TEs ('TE islands') and genetic variation. Our results suggest that TE islands might function as spring wells for genetic diversification in founder populations of this invasive species. The distinct organization of TE islands, their gene composition and their regulation by the genome adds compelling evidence for the role of TEs as players in differentiation, adaptation and speciation.

Results

Phenotypic differences between BR and JP lineages. Colonies from the two populations contained similar numbers of workers (Mann–Whitney U-test = 778.5, $Z = -0.634$, $P = 0.526$; BR: median = 28, quartiles 21.75 and 51.25, $n = 27$ colonies; JP: median = 29, quartiles 16 and 47, $n = 64$), but queen number was higher in Japan (Mann–Whitney U-test = 501, $Z = -3.084$, $P < 0.003$; BR: 5 queens, quartiles 3, 8; JP: median = 10, quartiles 4 and 19). Body sizes of queens and workers from BR were significantly smaller than in JP individuals, yet wingless males did not differ in any of the measured characters (see Supplementary Information).

In ants, cuticular chemical compounds play a particular prominent role in kin recognition, which is crucial for species integrity but on a deeper level also a requirement for the maintenance of altruism[16]. Analysis of cuticular compound extracts from BR and JP workers showed that compound composition differed significantly between the two lineages (multivariate analysis of variance: df = 2, F = 10.33, $R^2 = 0.39$, $P < 0.001$) and samples were classified correctly according to population of origin in 83.3% of cases (Supplementary Table 1; Supplementary Fig. 1).

The lineages also differed in behaviour, with BR colonies being significantly more aggressive towards both workers and queens from their own lineage, while JP colonies more readily accepted JP workers and queens (P_{Workers} JPxJP versus BR × BR = 0.000296, P_{Queens} JP × JP versus BR × BR = 7.98e − 07, Supplementary Fig. 2). Confronted with individuals from the other lineage, BR colonies were as aggressive as in within-population encounters (P_{Workers} BR × JP versus BR × BR = 0.39, P_{Queens} BR × JP versus BR × BR = 0.94), while JP colonies were again significantly less aggressive (P_{Workers} JP × BR versus BR × BR = 0.000131, P_{Queens} BR × JP versus BR × BR = 1.23e − 07). Testing discrimination against workers of another ant species, *Wasmannia auropunctata,* evoked similarly high aggressive responses in both lineages, suggesting that the BR and JP populations do not generally differ in their aggressive potential.

The *C. obscurior* genome is compact and rich in class I TEs. Using MSR-CA version 1.4, we produced a 187.5-Mb draft reference genome based on paired-end sequencing of several hundred diploid females (454 Titanium FLX sequencing) and a 200-bp library made from five haploid males (Illumina HiSeq2000; Supplementary Table 2), all coming from a single Brazilian colony. Automatic gene annotation using MAKER version 2.20 (ref. 17) was supported by 454 RNAseq data of a normalized library made from a pool of all castes and developmental stages. We filtered the assembly for prokaryotic scaffolds and reduced the initial 11,084 scaffolds to 1,854 scaffolds, containing all gene models and a total of 94.8% (177.9 Mb) of the assembled sequence. The genome can be accessed under antgenomes.org/ and hymenopteragenome.org.

The final gene set contains 17,552 genes, of which 9,552 genes have a known protein domain as detected by IPRScan (www.ebi.ac.uk/interpro/), and falls within the range of recent estimates for eight other sequenced ant species[18–26]. Of all genes, 72.5% have an annotation edit distance of less than 0.5, which is consistent with a well-annotated genome[27] (Supplementary Table 3).

The *C. obscurior* genome is the smallest so far sequenced ant genome[18–26]. Although there is no physical genome size estimate for *C. obscurior*, assembled sequences and physical estimates are tightly correlated in seven ant genomes (LM in R: $R^2 = 0.73$, $F_{1,\,5} = 13.7$, $P = 0.014$, from ref. 28), suggesting that *C. obscurior* has the smallest genome reported so far for an ant species [29]. Overall, the draft genome size of the analysed sequenced ants is negatively correlated to relative exon content (GLM in R: df = 6, F = 150.55, $P < 0.001$) but not to relative intron content (df = 5, F = 0.65, $P = 0.460$; Fig. 2), indicative of stabilizing selection on coding sequence. In contrast, intron size distribution is diverse between ant genomes and is not correlated with genome size (Supplementary Fig. 3; Supplementary Table 4).

We used a custom pipeline (see Supplementary Information) to identify simple repeats, class I retrotransposons and class II DNA transposons in *C. obscurior*, seven ant genomes (*Acromyrmex echinatior* (*Aech*), *Atta cephalotes* (*Acep*), *Solenopsis invicta* (*Sinv*), *Linepithema humile* (*Lhum*), *Pogonomyrmex barbatus* (*Pbar*), *Harpegnathos saltator* (*Hsal*), *Camponotus floridanus* (*Cflo*)), the parasitic wasp *Nasonia vitripennis* (*Nvit*) and the honeybee *Apis mellifera* (*Amel*). Across the analysed ants, genome size is significantly correlated with relative simple repeat content (lm, $R^2 = 0.66$, F = 11.83, $P = 0.014$; Fig. 2) but not with class I and class II TE content. However, it appears that the larger genomes contain more relative class II sequence. Relative class I retrotransposon content was highest in *C. obscurior* (7.6 Mb, 4.31%, Supplementary Fig. 4) and in particular, many class I non-LTR retrotransposons (for example, 14 types of LINEs) and several types of LTR transposons (Ngaro, Gypsy, DIRS and ERV2), TIR elements (for example, hAT, MuDR, P) and Helitrons are more abundant in *C. obscurior* (Supplementary Table 5).

Genomic signatures of an inbred lifestyle. On the basis of TE content calculations for 1 and 200 kb sliding windows, we identified 18 isolated 'TE islands' located in 'LDR' (low-density regions) in the *C. obscurior* genome. These TE islands were defined as containing TE accumulations in the 95–100% quantile within scaffolds over 200 kb (87 scaffolds, representing 96.02% or 170.8 Mb of the assembly). In total, TE islands cover 12.78 Mb of sequence (7.18% of total sequence) and range between 0.19 and 1.46 Mb in size. The TE islands contain 27.54% (4.92 Mb) of the assembly-wide TE sequence (17.87 Mb), 6.6% of all genes (1,160), and have reduced exon content (TE islands 87.0 exon bp kb^{-1}, LDRs 124.5 exon bp kb^{-1}). Note that some larger scaffolds contain more than one TE island.

Retroelements of the superfamilies BEL/Pao, DIRS, LOA/Loa, Ngaro, R1/R2 and RTE as well as DNA transposons of the superfamilies Academ, Kolobok-Hydra, Maverick, Merlin, on and TcMar-Mariner/-Tc1 populate TE islands with significantly higher copy numbers than other elements (Fisher's exact test, false discovery rate <0.05, Fig. 3, Supplementary Table 6). Furthermore, both class I and class II elements show a length polymorphism, with elements in TE islands being significantly longer compared with elements in LDRs (*U*-tests, $W = 109089018$, $P < 2e-16$ for class I and $W = 152340067$, $P < 2e-16$ for class II, Fig. 4a, Supplementary Fig. 5).

We also assessed the genome-wide TE distributions for seven published ant genomes, *Amel* v4.5 and *Nvit* v2.0 (Fig. 5). The smaller ant genomes (*Pbar*, *Lhum* and *Cflo*) and *Amel* are similar in TE sequence distribution. In contrast, the larger genomes (*Aech*, *Acep*, *Sinv* and *Hsal*) are more variable, have higher median TE content and a much broader and tailed TE frequency distribution with longer stretches of high or low TE content. The genome of *C. obscurior* is distinct from the other ant genomes, with low TE content in LDRs but exceptional clustering with high TE densities in TE islands. The genome of the inbred wasp *N. vitripennis* contains regions with up to 60% TE content that are surrounded by LDRs containing much less TE sequence (~10%), resembling the pattern observed in *C. obscurior*.

TE islands diverge faster than LDRs in the two populations. We mapped ~140 Gb of genomic DNA Illumina reads (~60 × coverage for each population) from pools of 30 (BR) and 26 (JP) male pupae, respectively, against the reference genome (BWA; bio-bwa.sourceforge.net) and analysed the local coverage ratio to detect genetic divergence. Deviations from the mean coverage ratio (Fig. 6) are in part caused by sequence deletions, insertions and duplications[30]. Such variations are particularly frequent in TE islands (Figs 4b and 6), suggesting accelerated divergence within islands (median deviation from mean coverage ratio: 0.288 in TE Islands, 0.163 in LDRs; *U*-test, $W = 640300902$; $P < 2e-16$).

We retrieved SNV (single-nucleotide variants) calls using consensus calls from samtools (samtools.sourceforge.net) and the GATK (broadinstitute.org/gatk/). Although TE islands only comprise 7.18% of the genome, they combine 15.59% (86,236 of 553,052) of all SNV calls. Given that we sequenced haploid males from highly inbred lineages, heterozygous SNVs should be rare. A large fraction of heterozygous SNVs in both lineages are within TE islands (62.95% of 62,879 in BR, 50.52% of 98,353 in

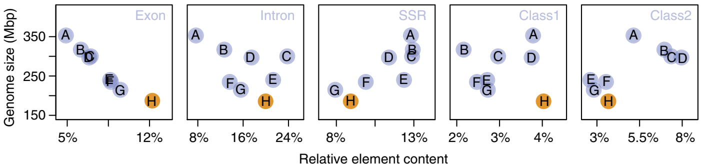


Figure 2 | Assembly size in Mbp plotted against the relative proportion of exons, introns and different repetitive elements. The analysed genomes show a negative correlation between relative exon but not intron content. Genome size is positively correlated with relative short simple repeat but not class I and II TE content. A, *S. invicta*; B, *A. cephalotes*; C, *A. echinatior*; D, *H. saltator*; E, *C. floridanus*; F, *P. barbatus*; G, *L. humile*; H, *C. obscurior*.

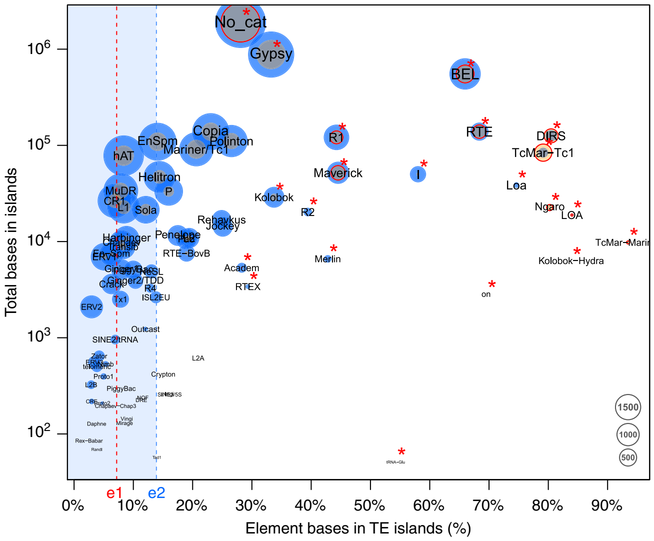


Figure 3 | The proportion of bases annotated in TE islands in *C. obscurior* against the log-scaled total base count in TE islands for each TE superfamily. Point size is relative to the copy number of the respective element found in TE islands (orange) and in LDRs (blue). Red circles indicate superfamilies with significantly higher frequency in TE islands than other superfamilies. Superfamilies with a significantly higher base count in TE islands are denoted by a red asterisk. e1: Percentage of the genome contained in TE islands (7.18%), e2: median across all types of TEs (13.89%).

JP), while rates of homozygous calls (Fig. 6) are not increased (11.88% of 16,277 in BR, 6.91% of 445,316 in JP). High numbers of false positive heterozygous SNVs calls can arise in duplicated regions that collapsed into a single locus due to misassemblies[31]. Accordingly, such SNVs can be identified by a twofold increase in coverage and in fact mark diverging duplicated loci within the same lineage (Fig. 4c).

Genes in TE islands should also show signatures of accelerated divergence from orthologues if overall sequence evolution is increased in these regions. Indeed, BLASTp searches against seven ant proteomes produced significantly lower bit scores for genes within TE islands when compared with genes in LDRs (Fig. 4d, U-test, $W = 120460260$, $P < 2e-16$). In accordance, SNV annotation revealed higher rates of non-synonymous substitutions between the BR and JP lineage in TE island genes (Fig. 4e, U-test, $W = 923754$, $P < 2e-16$). Surprisingly, however, on average, TE island genes contained less synonymous SNVs than LDR genes (LDR 0.67 kb^{-1}, TE island 0.42 kb^{-1}, U-test, $W = 10743397$, $P < 2e-16$).

Copy number variation within and between TE islands. We inspected 512 candidate loci (155 in TE islands) of 1 kb length by plotting the coverage of each lineage relative to SNVs, genes, and TEs at the respective position, to find genes potentially affected by deletion or copy number variation events and compiled a list of 89 candidate genes (Supplementary Table 7). Experimental proof-of-principle was conducted by PCR and Sanger sequencing for two deletion candidates (*Cobs_13563* and *Cobs_01070*) and by real-time quantitative PCR for four duplication candidates (*Cobs_13806*, *Cobs_17872*, *Cobs_13486*, and *Cobs_16853*) (Supplementary Fig. 7). A majority of these genes are located in TE islands (61.8%) and 34 genes show at least weak expression in BR individuals in RNAseq data (see below). The affected genes play roles in processes that may be crucial during invasion of novel habitats, such as chemical perception, learning and insecticide resistance. In particular, four different odorant/gustatory receptor genes show signs of either multiple exon (*Cobs_05921*, *Cobs_13418*, *Cobs_14265*) or whole-gene duplication (*Cobs_17892*). A gene likely involved in olfactory learning, *Cobs_13711*, a homologue to *pst*[32], also shows signs of duplication. Three genes homologous to fatty acid synthase (FAS) genes, a key step in cuticular odour production, contain partial deletions (*Cobs_16510*, *Cobs_14262*) or duplications (*Cobs_15866*). Furthermore, we found differences in genes associated with insecticide response (*Cobs_00487*, a homologue of *nAChRα6* (FBgn0032151) (ref. 33) and *Cobs_17834*, coding for a homologue to Cyp4c1 (EFN70878.1) (ref. 34). Other key genes affected are associated with circadian rhythm (*Cobs_17789*, homologue to *per* (FBgn0003068)), caste determination (*Cobs_01070*, with homology to *Mrjp1* (gi406090) (ref. 35), development (*Cobs_17755*, coding for a homologue of VgR (Q6X0I2.1) (ref. 36) and aging (*Cobs_14758*, with homology to *Mth2* (FBgn0045637) (ref. 37).

De novo assembly of ∼23M Illumina paired-end reads from the JP lineage that could not be mapped to the BR reference genome resulted in 17 contigs after filtering with highly significant BLASTx hits against proteins of other ants, suggesting that these conserved sequences were lost in the BR lineage instead of being gained in the JP lineage. According to functional annotation, among others these contigs code for homologues involved in development (Vitellogenin-like (XP_003689693))[38], cellular trafficking (Sorting nexin-25 (EGI65030))[39], immune response (Protein Toll (EGI66069))[38] and neuronal organization (Peripheral-type benzodiazepine receptor-associated protein 1 (EFN68490))[40] (Supplementary Table 8).

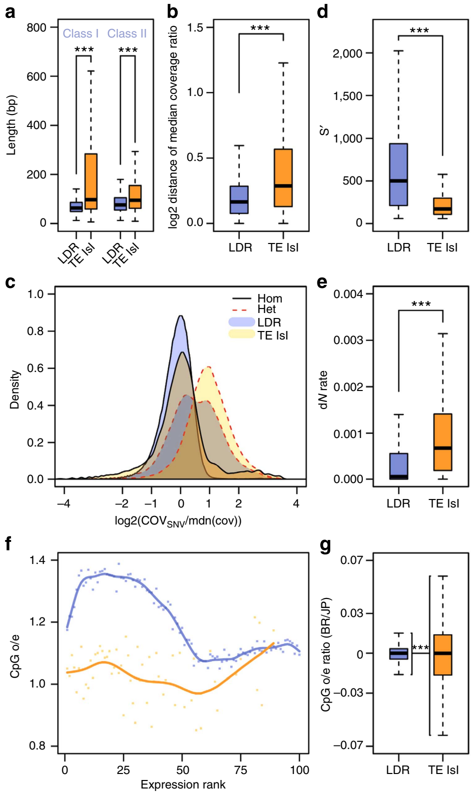


Figure 4 | Quantitative measures on the divergence of TE islands and LDRs. (**a**) Length polymorphism for Class I and Class II TEs in LDRs (blue) and TE islands (orange). *U*-tests, $n_{LDR} = 54{,}950$, $n_{TE} = 6{,}466$ for class I and $n_{LDR} = 59{,}054$, $n_{TE} = 6{,}813$ for class II. (**b**) Deviations from the median coverage ratio calculated for 1 kb windows in LDRs (blue) and TE islands (orange). *U*-test, $n_{LDR} = 157{,}296$; $n_{TE} = 12{,}165$. (**c**) Log2-scaled density plots of the coverage for all homozygous (solid black lines) and heterozygous SNV (dotted red lines) calls divided by the median coverage (orange, calls within TE islands; blue, calls in LDRs). Coverage at homozygous calls is not different from the median overall coverage, neither in TE islands nor in LDRs. The shift for heterozygous SNV calls within TE islands shows that most calls result from diverging duplicated loci. The bimodal distribution for heterozygous calls in other genomic regions suggests two distinct populations of SNV calls, that is, true heterozygous loci (first peak) and diverging sequence in duplicated loci (second peak). (**d**) Bit scores for genes in LDRs (blue) and TE islands (orange) retrieved by BLASTx against annotated proteins from seven ant genomes. *U*-test, $n_{LDR} = 12{,}065$; $n_{TE} = 902$. (**e**) Rates of non-synonymous substitutions (calculated as dN/(dN + dS)) in LDR (blue) and TE island genes (orange). *U*-test, $n_{LDR} = 6{,}806$; $n_{TE} = 423$. (**f**) Exon-wide CpG o/e values were plotted against the expression rank from 0 (least expressed) to 100 (most expressed) genes for LDRs (blue) and TE islands (orange). (**g**) Calculated ratios (BR/JP) for exon CpG o/e values in LDRs (blue) and TE islands (orange). F-test, $n_{LDR} = 16{,}379$; $n_{TE} = 1{,}159$. (***$P < 0.0001$, boxplots show the median, interquartile ranges (IQR) and 1.5 IQR.).

Gene composition and regulation of TE islands. Increased TE activity may incur costs to fitness by disrupting gene function. A two-tailed Gene Ontology (GO) enrichment analysis revealed that 59 GO terms associated with conserved processes (for example, cytoskeleton organization, ATP binding, organ morphogenesis) are under-represented in TE islands, while 18 GO terms are enriched (Supplementary Tables 9 and 10). Four of the over-represented terms relate to olfactory receptors (ORs; GO:0004984, GO:0005549, GO:0050911, GO:0007187) and two terms relate to FAS genes (GO:0005835, GO:0016297). The remaining 12 terms most likely relate to TE-derived genes.

Gene body CpG depletion as a result of increased CpG to TpG conversion due to cytosine methylation is a measure for germline methylation (that is, epigenetic regulation) in past generations. In TE island genes, the exon-wide median observed/expected (o/e) CpG ratio is significantly lower than in other genes (*t*-test, TE island genes: 1.05, LDR genes: 1.20, $P < 1e-16$). However, both sets of genes show strikingly different correlations of expression and o/e CpG values (Fig. 4f). For LDR genes, o/e CpG values are high in moderately expressed genes and low in highly expressed genes. In contrast, in TE islands, weakly to moderately expressed genes contain less CpG dinucleotides, while highly expressed genes have higher o/e CpG values. To further identify traces of differential regulation of TE islands, we compared the exon o/e CpG values between the lineages by calculating BR/JP ratios for each exon's o/e CpG values and found higher variance in BR/JP ratios in TE islands than in LDRs (Fig. 4g, F-test, F = 0.136, $P < 2e-16$, ratio of variances = 0.136).

Finally, to assess whether gene expression levels differed between LDRs and TE islands, we generated ~14 and ~17 Gb transcriptomic RNAseq data of seven queens and seven queen-destined larvae (third larval stage), respectively, from the BR lineage. We estimated mean normalized expression values for each gene using DESeq2 (bioconductor.org/packages/release/bioc/html/DESeq2.html), revealing that expression in TE islands was much lower than in LDRs (median expression of all LDR genes = 25.45; in TE islands: 0.49; *U*-test, $W = 14461310$, $P < 2e-16$). While larvae and adult queens did not differ in the expression of LDR genes (median expression in queens = 21.16; in larvae = 23, 72; *U*-test, $W = 133301709$, $P = 0.221$), TE island genes were more expressed in adult queens (median expression in queens = 0.84; in larvae = 0; $W = 1031038$, $P < 2e-16$; Fig. 7, see Supplementary Fig. 6 for details on differential expression between queen and larvae).

Discussion

C. obscurior is a textbook example for successful biological invasion. Its small size allows for interspecific avoidance, it can rapidly establish colonies in disturbed habitats, and multiple generations per year allow for fast adaptation. While variation in CHCs and body size between the populations point to adaptations to different environments, higher queen number in the JP lineage is likely correlated with reduced intraspecific aggression.

The small genome of *C. obscurior* differs markedly from the other analysed ant genomes in TE distribution and over-abundance of several class I subclasses. Importantly, the genome contains low frequencies of TEs in LDRs but well-defined islands with high densities of TEs. In these islands, TEs are on average longer than in LDRs, suggesting overall higher TE activity[41]. Differences in mutation rates and sequence divergence between LDRs and TE islands reveal distinct evolutionary dynamics acting within the *C. obscurior* genome. Moreover, in TE islands, key genes are removed and the majority of genes is less expressed in

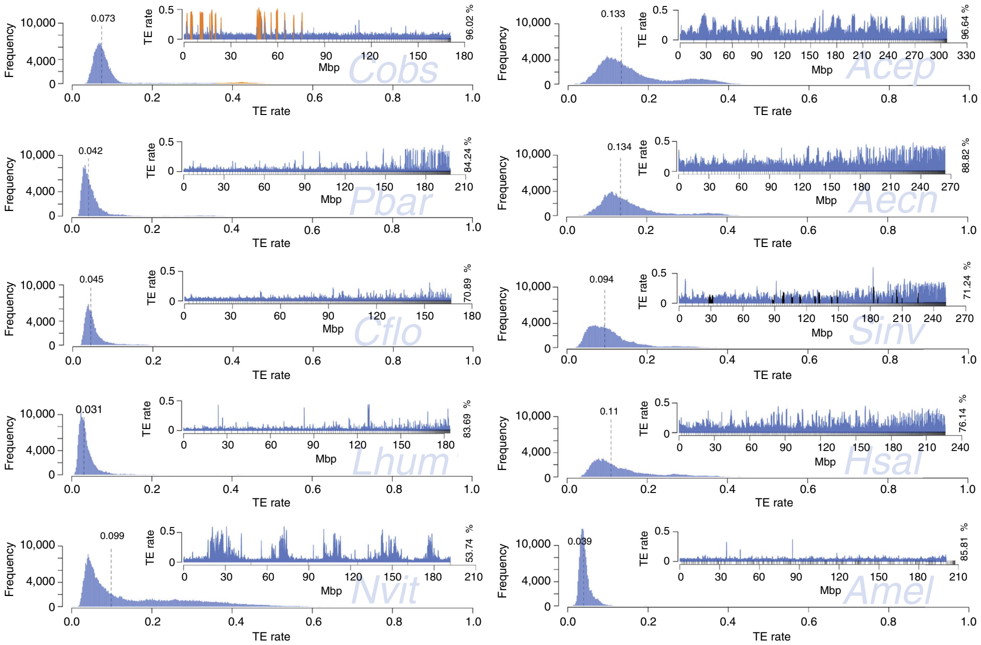


Figure 5 | Frequency and distribution (insert plots) of TE content in 200 kb windows. Frequency plots: dashed lines denote median TE content. Distribution plots: different proportions of total draft genome sequence were analysed (in %), depending on assembly quality. Scaffolds are sorted by size, small upward tick marks indicate scaffold boundaries. For *C. obscurior,* regions defined as TE islands are coloured in orange. For *S. invicta,* scaffolds mapping to a non-recombining chromosomal inversion[73] are shown in black. For *A. mellifera,* scaffolds were sorted according to linkage group.

larvae than adult queens. The non-random distribution of TEs suggests that intragenomic differences in selection efficiency against TEs may have further supported the formation of such locally confined TE accumulations.

Inbreeding can facilitate the accumulation of TEs[3] and repeated exposure to stress induced by novel environmental conditions can further amplify TE proliferation[42]. Small N_e is expected to increase the effects of genetic drift and in turn reduce selection efficiency against mildly deleterious mutations[2]. Under such conditions, local accumulations of TEs might have formed in genomic regions under relaxed selection. Similarly, a reduction in N_e in inbred *Drosophila* leads to a shift in the equilibrium between TE proliferation and purifying selection against TEs, thus allowing TEs to accumulate[43].

How can we explain extensive proliferation and diversification of TEs within islands, but purifying selection against TEs in LDRs? Coalescent effective population size of a genomic region is positively correlated with its recombination frequency and thus the local efficiency of selection and mutation rate[11]. The initial foundation of TE islands could hence be facilitated in genomic regions with low recombination frequency, providing a refugium of relaxed selection for TE insertions. Indeed, elevated rates of non-synonymous substitutions suggest relaxed selection on TE island genes. Increased frequency of DNA repair processes as a consequence of higher DNA transposition frequencies in TE islands should lead to more errors in DNA replication and double strand break repair[44] in comparison with LDRs. Large-scale mutations on the other hand, such as exon or gene duplications/deletions or gene shuffling, can directly be introduced during TE transposition[45]. TE islands may frequently produce genetic novelty and eventually, by chance, but despite high stochastic drift, adaptive phenotypes, corroborating the view of TEs as genetic innovators.

The list of genes affected by duplications or deletions contains a number of candidates that might be key to the divergence of the lineages. For example, differences in homologues to genes involved in larval development (for example, *Mrjp1*) might explain body-size differences. Two other candidates, *Cobs_00487* and *Cobs_17834,* show homology to genes that are involved in pesticide resistance against Chlorpyrifos and Imidacloprid (*nAChRα6*) and Deltamethrin (*Cyp4c*) in different invertebrate species[46–49]. Imidacloprid treatment of gall wasp infested *Erythrina variegate* coral trees of the Japan habitat occurred at least once the year before collection of the colonies in 2010 (personal communication S. Mikheyev). In the Brazil habitat, Chlorpyrifos, Deltamethrin and the organophosphate Monocrotophos have routinely been used over the last 10 years (personal communication J.H.C. Delabie).

Furthermore, several within-island genes involved in the production (FAS[50]) and perception (ORs) of chemical cues contained deletions or duplications in one of the lineages. These results suggest that variation in FAS genes may be responsible for diverging CHC profiles in *C. obscurior*[51], while variation in OR genes affects olfactory perception. Chemosensory neurons express highly sensitive ORs[52], which are particularly diverse[53] and under strong selection in ants[54]. Gene loss and duplication in the OR

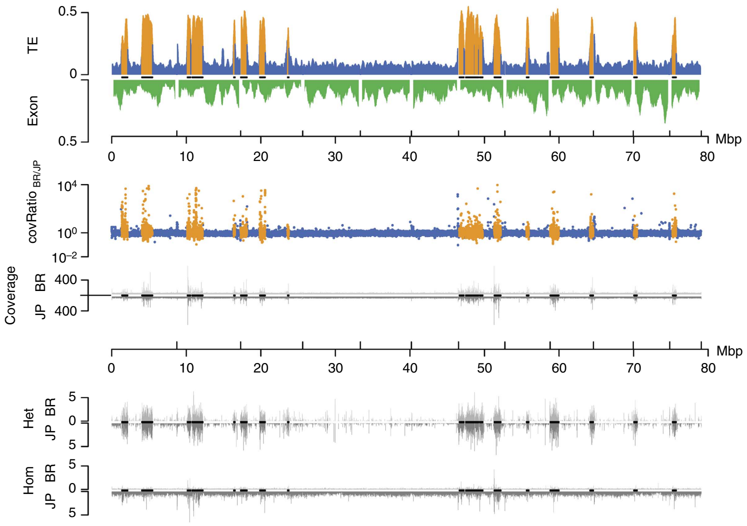


Figure 6 | Genomic divergence and subgenomic structure of the 12 largest *C. obscurior* genome scaffolds (including all 18 TE islands). High TE content in TE islands correlates with deviations from the average coverage ratio, very high absolute coverage in both lineages and high numbers of SNV calls. First track: relative TE (blue and orange within TE islands) and exon content (green) per 200 kb. Second track: coverage ratio BR/JP (blue and orange within TE islands). Third track: absolute coverage for BR (top) and JP (bottom). Fourth track: heterozygous SNV calls per kb in BR (top) and JP (bottom) relative to the reference genome. Fifth track: homozygous SNV calls per kb in BR (top) and JP (bottom) relative to the reference genome. Black lines on *x* axes indicate localization of TE islands.

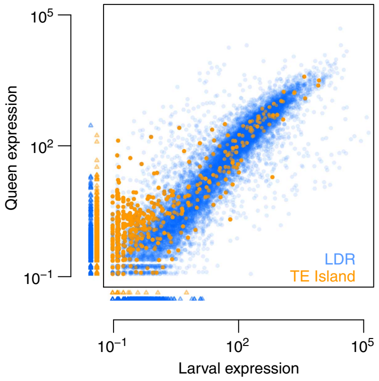


Figure 7 | Mean normalized expression in third instar queen larvae and mated adult queens for all Cobs1.4 genes. Small triangles indicate genes with no expression in queens (plotted below the *x* axis) or larvae (plotted left to the *y* axis). Ninety-five TE island genes and 1,382 LDR genes were not expressed at all (orange, TE island genes; blue, LDR genes).

gene family has been significantly frequent[55] and differences are assumed to be shaped by adaptive processes in response to a species' ecological niche[56,57]. Intriguingly, the diversification of OR genes is thought to be largely caused by gene duplications and interchromosomal transposition[58], two mechanisms known to be by-products of TE activity. While the distinct patterns of kin recognition and aggressive behaviour in the two lineages of *C. obscurior* may in part be explained by TE-mediated variation in these genes, they also suggest lineage-specific dynamics of the interaction of phenotype and genome evolution. Reduced aggression between colonies in the JP lineage should promote gene flow by exchange of reproductives and thus increase Ne, heterozygosity, and the efficiency of sexual recombination, facilitating the spread of novel arising genotypes. Our findings contrast the view of reduced aggression between colonies of invasive ants[59], but so far it is unclear whether lineage-specific differences are caused by variation in perception or downstream neuronal processes.

Mechanisms controlling TEs are as old as prokaryotes[9] and in fact most TEs are epigenetically silenced[45,60], through either methylation, histone modifications[61] or RNAi[62]. Even though many genes in TE islands are expressed, the overall expression is significantly lower than in LDRs. In line with previous correlations on methylation and expression in eusocial insects[63,64], o/e CpG ratios in *C. obscurior* LDR genes are negatively correlated with expression. However, TE island genes do not follow this trend, in that they are weakly expressed while having low o/e CpG rates. Proximity to TEs can increase gene body methylation[65], which could explain stronger methylation of TE island genes and thus CpG depletion. Also, relaxed selection in island genes should in general increase fixation frequency of base mutations, including CpG to TpG conversions thus depleting CpG content. Gene expression differences in TE island genes between larvae and adult queens suggest stronger regulation of these potentially disruptive genes during the sensitive developmental phase. Finally, key regulatory genes are under-represented in TE islands. These gene set differences between TE islands and LDRs can either be explained by selection processes, removing vital genes from linkage to TE islands or by selective restriction of TE accumulations to genomic regions devoid of such genes.

The current understanding of TE activity dynamics in genomes is that periods of relative dormancy are followed by bursts of activity, often induced by biotic and abiotic stress, such as exposure to novel habitats. Frequent TE transposition during bursts leads to genomic rearrangements, thus producing new genetic variants and eventually even promoting speciation[66–69]. TE dynamics can also be strongly affected by mating system[3,70–72], and the life history of *C. obscurior* likely challenges the genomic integrity resulting in genomic regions with over 50% TE content. In conclusion, TE dynamics in *C. obscurior* seem to have shifted from a serial to a parallel mode, where a fraction of the genome is reshaped repeatedly in a continuous burst of TE activity. Strikingly, the inbred parasitoid wasp *N. vitripennis* has similar TE frequency patterns suggesting that similar life history strategies and their consequences on N_e and drift can lead to convergent genomic organization. TEs represent a major force in evolution, contributing to the generation of genetic variation especially in species confronted with hurdles like inbreeding or repeated bottlenecks. They furthermore seem to play an important role in the rapid adaption of invasive species to novel environments, making it particularly crucial to understand their origin, function and regulation.

Methods

Detailed methods and accompanying Supplementary Tables 11 to 16 are available as Supplementary Information online.

Organisms. Live colonies of *C. obscurior* were collected from aborted fruits on coconut trees (*Cocos nucifera*) in Brazil (collected in 2009) and from bark cavities in coral trees (*Erythrina* sp.) in Japan (collected in 2010). The colonies were transferred to Regensburg and placed in plastered petri dishes. Food (honey-soaked shreds of paper; *Drosophila* or small chunks of *Periplaneta americana*) and water were provided every 3 days and colonies were kept in incubators under constant conditions (12 h 28 °C light/12 h 24 °C dark). All animal treatment guidelines applicable to ants under international and German law have been followed. Collecting the colonies that form the basis of the laboratory population used in this study was permitted by the Brazilian Ministry of Science and Technology (RMX 004/02). No other permits were required for this study.

***De novo* genome assembly.** The reference genome is based on one colony that was kept under strict inbreeding in the lab for four generations before extractions. Whole DNA was extracted with CTAB. We extracted DNA from ~900 ants, which were pooled to be sequenced with 454 technology. Extracts of 5, 10 and 30 Brazilian males and 26 Japanese males, respectively, were used for Illumina libraries.

We generated 200 and 500 bp insert libraries with Illumina's TruSeq DNA sample preparation kits from 5 μg of total DNA. Quality control and library preparation were carried out by the KFB sequencing centre of the University Regensburg, sequencing runs were performed by Illumina (Hayward, USA) on a HiSeq2000. Quality control, library preparation and sequencing of 8 and 20 kb long paired end libraries (454, Roche) were carried out by Eurofins MWG Operon (Ebersberg, Germany). Extracted DNA was fragmented into the appropriate fragment sizes (8 and 20 kb) using the HydroShear DNA Shearing Device (GeneMachine). Further library preparation was performed according to 'GS FLX Titanium Paired End Library Prep 20 + 8 kb Span Method Manual' before sequencing on a GS FLX Titanium (Roche).

The *de novo* genome assembly was created with MSR-CA version 1.4 open source assembler (University of Maryland genome assembly group at ftp://ftp.genome.umd.edu/pub/MSR-CA/). The MSR-CA assembler combines a deBruijn graph strategy with the traditional Overlap-Layout-Consensus employed by various assembly programmes for Sanger-based projects (Arachne, PCAP, CABOG). The MSR-CA uses a modified version of CABOG version 6.1 for contiging and scaffolding. The combined strategy allowed us to natively combine the short 100 bp Illumina reads and longer 454 reads in a single assembly without resorting to an approach that would require one to assemble each type of data separately and then creating a combined assembly.

Mapping. For each lineage, we randomly sampled 140 M 100 bp reads from libraries generated from 26 (JP) and 30 (BR) male pupae. Raw reads were parsed through quality filtration and adapter trimming (Trimmomatic v0.22 (usadellab.org/cms/?page=trimmomatic), options: HEADCROP:7 LEADING:28 TRAILING:28 SLIDINGWINDOW:10:10) and mapped against the BR reference genome with BWA (bio-bwa.sourceforge.net) and Stampy v1.0.21 (www.well.ox.ac.uk/project-stampy).

Variant calling. SNV calling was carried out combining samtools (samtools.sourceforge.net) and the GATK (www.broadinstitute.org/gatk/) retaining only those variants called consistently by both tools. The final variant set of 553 052 SNVs and 67,987 InDels was stored in a single VCF file. SNVs were annotated with SNPeff (snpeff.sourceforge.net) to identify non-synonymous and synonymous substitutions.

Calculation of sliding windows. One kb windows of different stats (TEs, exons, SNPs, coverage) were calculated for all scaffolds based on GFF, VCF and SAM files. For GFF and VCF files, custom bash and perl scripts were used to calculated TE and exon bases per 1 kb, and variant calls per 1 kb. Coverage per 1 kb was calculated from SAM files, using samtools' depth algorithm and custom bash and perl scripts. Subsequent processing, calculating of 200 kb sliding windows and plotting of the data was performed with R v3.0.0 (r-project.org).

Gene expression analysis with RNAseq. We extracted whole RNA with the RNeasy Plus Micro kit (Qiagen). Single end Illumina libraries from amplified RNA (Ovation RNAseq system V2) were generated following the manufacturers protocol (Ovation Rapid Multiplexsystem, NuGEN). Sequencing on an Illumina HiSeq1000 at the in-house sequencing centre (KFB, Regensburg, Germany) generated ~20M 100 bp reads per sample (Supplementary Table 16). Raw reads were filtered for adapter contamination (cutadapt, code.google.com/p/cutadapt/), parsed through quality filtration (Trimmomatic v0.27, options: LEADING:10 TRAILING:10 SLIDING:4:10 MINLEN:15), and mapped against the reference genome using the tophat2 (v2.0.8, ccb.jhu.edu/software/tophat/index.shtml) and bowtie2 (v2.1.0, bowtie-bio.sourceforge.net/bowtie2/index.shtml) package (--b2-sensitive mode, mapping rate ~50%). Gene expression analysis was carried out with DESeq2 (bioconductor.org/packages/release/bioc/html/DESeq2.html), based on count tables produced with HTSeq (www-huber.embl.de/users/anders/HTSeq/doc/overview.html) against the Cobs1.4 MAKER annotation (Supplementary Table 16). Genes were considered to be differentially expressed at a false discovery rate <0.05 and expression values are reported as untransformed base means of read counts per treatment group, after correcting for library size differences ('size factor normalization').

References

1. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* **18,** 237–268 (1987).
2. Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates Inc, 2007).
3. Charlesworth, D. & Wright, S. I. Breeding systems and genome evolution. *Curr. Opin. Genet. Dev.* **11,** 685–690 (2001).
4. Romiguier, J. *et al.* Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *J. Evol. Biol.* **27,** 593–603 (2014).
5. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
6. Lanfear, R., Ho, S. Y. W., Love, D. & Bromham, L. Mutation rate is linked to diversification in birds. *Proc. Natl Acad. Sci. USA* **107,** 20423–20428 (2010).
7. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26,** 345–352 (2010).
8. Fontdevila, A. *The Dynamic Genome* (Oxford Univ. Press, 2011).
9. Fedoroff, N. V. *Plant Transposons and Genome Dynamics in Evolution* (John Wiley & Sons, 2013).
10. González, J., Karasov, T. L., Messer, P. W. & Petrov, D. A. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*. *PLoS Genet.* **6,** e1000905 (2010).
11. Casacuberta, E. & González, J. The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **22,** 1503–1517 (2013).
12. Madlung, A. & Comai, L. The effect of stress on genome regulation and structure. *Ann. Bot. Lond.* **94,** 481–495 (2004).
13. Rostant, W. G., Wedell, N. & Hosken, D. J. Transposable elements and insecticide resistance. *Adv. Genet.* **78,** 169–201 (2012).
14. Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science* **303,** 1626–1632 (2004).
15. Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J. & Capy, P. The struggle for life of the genome's selfish architects. *Biol. Direct* **6,** 19 (2011).
16. van Zweden, J. S. & D'Ettorre, P. in *Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology* (Cambridge Univ. Press, 2010).
17. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).

18. Nygaard, S. *et al.* The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res.* **21,** 1339–1348 (2011).
19. Suen, G. *et al.* The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.* **7,** e1002007 (2011).
20. Wurm, Y. *et al.* The genome of the fire ant *Solenopsis invicta*. *Proc. Natl Acad. Sci. USA* **108,** 5679–5684 (2011).
21. Smith, C. R. *et al.* Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc. Natl Acad. Sci. USA* **108,** 5667–5672 (2011).
22. Smith, C. D. *et al.* Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc. Natl Acad. Sci. USA* **108,** 5673–5678 (2011).
23. Bonasio, R. *et al.* Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329,** 1068–1071 (2010).
24. Weinstock, G. M. *et al.* Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443,** 931–949 (2006).
25. Werren, J. H. *et al.* Functional and evolutionary Insights from the genomes of three parasitoid *Nasonia* species. *Science* **327,** 343–348 (2010).
26. Oxley, P. R. *et al.* The genome of the clonal raider ant *Cerapachys biroi*. *Curr. Biol.* **24,** 451–458 (2014).
27. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13,** 329–342 (2012).
28. Gadau, J. *et al.* The genomic impact of 100 million years of social evolution in seven ant species. *Trends Genet.* **28,** 14–21 (2012).
29. Tsutsui, N. D., Suarez, A. V., Spagna, J. C. & Johnston, J. S. The evolution of genome size in ants. *BMC Evol. Biol.* **8,** 64 (2008).
30. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6,** S13–S20 (2009).
31. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13,** 36–46 (2012).
32. Dubnau, J. *et al.* The staufen/pumilio pathway is involved in *Drosophila* long-term memory. *Curr. Biol.* **13,** 286–296 (2003).
33. Millar, N. S. & Denholm, I. Nicotinic acetylcholine receptors: targets for commercially important insecticides. *Invert. Neurosci.* **7,** 53–66 (2007).
34. Hemingway, J. & Ranson, H. Insecticide resistance in insect vectors of human disease. *Annu. Rev. Entomol.* **45,** 371–391 (2000).
35. Drapeau, M. D., Albert, S., Kucharski, R., Prusko, C. & Maleszka, R. Evolution of the yellow/major royal jelly protein family and the emergence of social behavior in honey bees. *Genome Res.* **16,** 1385–1394 (2006).
36. Chen, M.-E., Lewis, D. K., Keeley, L. L. & Pietrantonio, P. V. cDNA cloning and transcriptional regulation of the vitellogenin receptor from the imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae). *Insect Mol. Biol.* **13,** 195–204 (2004).
37. Duvernell, D. D., Schmidt, P. S. & Eanes, W. F. Clines and adaptive evolution in the methuselah gene region in *Drosophila melanogaster*. *Mol. Ecol.* **12,** 1277–1285 (2003).
38. Gilbert, L. I. *Insect Molecular Biology and Biochemistry* (Academic Press, 2010).
39. Worby, C. A. & Dixon, J. E. Sorting out the cellular functions of sorting nexins. *Nat. Rev. Mol. Cell Biol.* **3,** 919–931 (2002).
40. Galiegue, S. *et al.* Cloning and characterization of PRAX-1. A new protein that specifically interacts with the peripheral benzodiazepine receptor. *J. Biol. Chem.* **274,** 2938–2952 (1999).
41. Kaminker, J. S. *et al.* The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* **3** research0084 (2002).
42. Capy, P., Gasperi, G., Biémont, C. & Bazin, C. Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85,** 101–106 (2000).
43. Nuzhdin, S. V., Pasyukova, E. G. & Mackay, T. F. Accumulation of transposable elements in laboratory lines of *Drosophila melanogaster*. *Genetica* **100,** 167–175 (1997).
44. Shee, C., Gibson, J. L. & Rosenberg, S. M. Two mechanisms produce mutation hotspots at DNA breaks in *Escherichia coli*. *Cell Rep.* **2,** 714–721 (2012).
45. Fedoroff, N. V. Transposable elements, epigenetics, and genome evolution. *Science* **338,** 758–767 (2012).
46. Casida, J. E. & Durkin, K. A. Neuroactive insecticides: targets, selectivity, resistance, and secondary effects. *Annu. Rev. Entomol.* **58,** 99–117 (2013).
47. Xu, L., Wu, M. & Han, Z. Overexpression of multiple detoxification genes in deltamethrin resistant *Laodelphax striatellus* (Hemiptera: Delphacidae) in China. *PLoS ONE* **8,** e79443 (2013).
48. Slotkin, T. & Seidler, F. Transcriptional profiles reveal similarities and differences in the effects of developmental neurotoxicants on differentiation into neurotransmitter phenotypes in PC12 cells. *Brain Res. Bull.* **78,** 211–225 (2009).
49. Bergé, J. B., Feyereisen, R. & Amichot, M. Cytochrome P450 monooxygenases and insecticide resistance in insects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **353,** 1701–1705 (1998).
50. Blomquist, G. J., Nelson, D. R. & Derenobales, M. Chemistry, biochemistry, and physiology of insect cuticular lipids. *Arch. Insect Biochem. Physiol.* **6,** 227–265 (1987).
51. Foley, B., Chenoweth, S. F., Nuzhdin, S. V. & Blows, M. W. Natural genetic variation in cuticular hydrocarbon expression in male and female *Drosophila melanogaster*. *Genetics* **175,** 1465–1477 (2007).
52. Vosshall, L. B., Wong, A. M. & Axel, R. An olfactory sensory map in the fly brain. *Cell* **102,** 147–159 (2000).
53. Zhou, X. *et al.* Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. *PLoS Genet.* **8,** e1002930 (2012).
54. Kulmuni, J., Wurm, Y. & Pamilo, P. Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. *Heredity* **110,** 538–547 (2013).
55. Guo, S. & Kim, J. Molecular evolution of *Drosophila* odorant receptor genes. *Mol. Biol. Evol.* **24,** 1198–1207 (2007).
56. Hill, C. A. *et al.* G protein-coupled receptors in *Anopheles gambiae*. *Science* **298,** 176–178 (2002).
57. Bohbot, J. *et al.* Molecular characterization of the *Aedes aegypti* odorant receptor gene family. *Insect Mol. Biol.* **16,** 525–537 (2007).
58. Conceição, I. C. & Aguade, M. High incidence of interchromosomal transpositions in the evolutionary history of a subset of or genes in *Drosophila*. *J. Mol. Evol.* **66,** 325–332 (2008).
59. Tsutsui, N. D., Suarez, A. V. & Grosberg, R. K. Genetic diversity, asymmetrical aggression, and recognition in a widespread invasive species. *Proc. Natl Acad. Sci. USA* **100,** 1078–1083 (2003).
60. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9,** 397–405 (2008).
61. Rebollo, R. *et al.* A snapshot of histone modifications within transposable elements in *Drosophila* wild type strains. *PLoS ONE* **7,** e44253 (2012).
62. Buchon, N. & Vaury, C. RNAi: a defensive RNA-silencing against viruses and transposable elements. *Heredity* **96,** 195–202 (2006).
63. Bonasio, R. *et al.* Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Curr. Biol.* **22,** 1755–1764 (2012).
64. Hunt, B. G., Glastad, K. M., Yi, S. V. & Goodisman, M. A. D. The function of intragenic DNA methylation: insights from insect epigenomes. *Integr. Comp. Biol.* **53,** 319–328 (2013).
65. Veluchamy, A. *et al.* Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nat. Commun.* **4,** 2091 (2013).
66. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73,** 823–834 (2003).
67. de Boer, J. G., Yazawa, R., Davidson, W. S. & Koop, B. F. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* **8,** 422 (2007).
68. Ungerer, M. C., Strakosh, S. C. & Zhen, Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr. Biol.* **16,** R872–R873 (2006).
69. Hurst, G. D. & Werren, J. H. The role of selfish genetic elements in eukaryotic evolution. *Nat. Rev. Genet.* **2,** 597–606 (2001).
70. Charlesworth, D. & Charlesworth, B. Transposable elements in inbreeding and outbreeding populations. *Genetics* **140,** 415–417 (1995).
71. Boutin, T. S., Le Rouzic, A. & Capy, P. How does selfing affect the dynamics of selfish transposable elements? *Mobile DNA* **3,** 5 (2012).
72. Wright, S. I., Ness, R. W., Foxe, J. P. & Barrett, S. C. H. Genomic consequences of outcrossing and selfing in plants. *Int. J. Plant Sci.* **169,** 105–118 (2008).
73. Wang, J. *et al.* A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493,** 664–668 (2013).

Acknowledgements

We would like to thank John Wang for helpful discussions during the initial phase of the project and Andrés Moya for encouraging comments at an earlier stage of writing. We are most grateful to Masaki Suefuji and Jacques H.C. Delabie for collecting the colonies. We thank Christine Elsik, Deepak Unni and Chris Childers, who provide great support in hosting Cobs1.4 on hymenopteragenome.org. We thank Axel Strittmatter (MWG) and Thomas Stempfl (KFB) for advising on sequencing and Nicole Lehmann for advice on RNA amplification. We thank Helena Lowack for PCR validation of small-scale structural variants, Erich Bornberg-Bauer who assisted with CpG o/e calculation, Anurag Priyam for help with BLAST analyses and Robert Waterhouse for comparing the gene sets with OrthoDB. Y.W. is supported by BBSRC (BB/K004204/1) and NERC (NE/L00626X/1), A.Z. by NIH (R01 HG002945) and C.K. by DFG (BO2544/4-1). This study was funded by a DFG grant He1623/31 to J.O., J.G. and J.H.

Author contributions

J.O. and L.S. designed the study; J.O., L.S., J.G. J.H. wrote the manuscript; L.S. and J.O. analysed the data; A.Z. was responsible for genome assembly; J.W.K. and C.D.S. were responsible for repeat annotation; D.E., M.Y. and L.S. were responsible for gene prediction; C.K. was responsible for CpG o/e calculation; Y.W. was responsible for data

logistics; J.O., L.S., T.W., K.W. and A.K. were responsible for analyzing phenotypic differences; J.S., E.S. and J.O. performed the chemical analyses; all authors read and commented on the manuscript.

Additional information

Accession codes: Raw sequencing data from this study have been deposited in GenBank/EMBL/DDBJ under the BioProject accession code PRJNA237579.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Schrader, L. *et al.* Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.* 5:5495 doi: 10.1038/ncomms6495 (2014).

APPENDIX B

GENOMIC SIGNATURES OF EVOLUTIONARY TRANSITIONS FROM SOLITARY TO GROUP LIVING

This appendix is a reprint of a research article coauthored by Karen M. Kaphheim, Hailin Pan, Cai Li, Steven L. Salzberg, Daniele Puiu, Tanja Magoc, Hugh M. Robertson, Matthew E. Hudson, Aarti Venkat, Brielle J. Fischman, Alvaro Hernandez, Mark Yandell, myself, Carson Holt, George D. Yocum, William P. Kemp, Jordi Bosch, Robert M. Waterhouse, Evgeny M. Zdobnov, Eckart Stolle, F. Bernard Kraus, Sophie Helbing, Robin F.A. Moritz, Karl M. Glastad, Brendan G. Hunt, Michael A. D. Goodisman, Frank Hauser, Cornelis J. P. Grimmelikhuijzen, Daniel Guariz Pinheiro, Francis Morais Franco Nunes, Michelle Prioli Miranda Soares, Érica Donato Tanaka, Zilá Luz Paulino Simões, Klaus Hartfelder, Jay D. Evans, Seth M. Barribeau, Reed M. Johnson, Jonathan H. Massey, Bruce R. Southey, Martin Hasselmann, Daniel Hamacher, Mattias Biewer, Clement F. Kent, Amro Zayed, Charles Blatti III, Saurabh Sinha, J. Spencer Johnston, Shawn J. Hanrahan, Sarah D. Kocher, Jun Wang, Gene E. Robinson, and Guojie Zhang and is presented here with permissions of the authors and kind permission of The American Association for the Advancement of Science.

I contributed the genome annotations of *Megachile rotundata* that was one of 10 insect genome annotations that formed the basis of the genomic research presented in this article. This research article was first published in Kapheim KM, et al. (2015) Genomic signatures of evolutionary transitions from solitary to group living. *Science* (May1):1-8. Available at: http://science.sciencemag.org/content/348/6239/1139.

SOCIAL EVOLUTION

Genomic signatures of evolutionary transitions from solitary to group living

Karen M. Kapheim,[1,2,3]*† Hailin Pan,[4]* Cai Li,[4,5] Steven L. Salzberg,[6,7] Daniela Puiu,[7] Tanja Magoc,[7] Hugh M. Robertson,[1,2] Matthew E. Hudson,[1,8] Aarti Venkat,[1,8,9] Brielle J. Fischman,[1,10,11] Alvaro Hernandez,[12] Mark Yandell,[13,14] Daniel Ence,[13] Carson Holt,[13,14] George D. Yocum,[15] William P. Kemp,[15] Jordi Bosch,[16] Robert M. Waterhouse,[17,18,19,20] Evgeny M. Zdobnov,[17,18] Eckart Stolle,[21,22] F. Bernhard Kraus,[21,23] Sophie Helbing,[21] Robin F. A. Moritz,[21,24] Karl M. Glastad,[25] Brendan G. Hunt,[26] Michael A. D. Goodisman,[25] Frank Hauser,[27] Cornelis J. P. Grimmelikhuijzen,[27] Daniel Guariz Pinheiro,[28,29] Francis Morais Franco Nunes,[30] Michelle Prioli Miranda Soares,[28] Érica Donato Tanaka,[31] Zilá Luz Paulino Simões,[28] Klaus Hartfelder,[32] Jay D. Evans,[33] Seth M. Barribeau,[34] Reed M. Johnson,[35] Jonathan H. Massey,[2,36] Bruce R. Southey,[37] Martin Hasselmann,[38] Daniel Hamacher,[38] Matthias Biewer,[38] Clement F. Kent,[39,40] Amro Zayed,[39] Charles Blatti III,[1,41] Saurabh Sinha,[1,41] J. Spencer Johnston,[42] Shawn J. Hanrahan,[42] Sarah D. Kocher,[43] Jun Wang,[4,44,45,46,47]† Gene E. Robinson,[1,48]† Guojie Zhang[4,49]†

The evolution of eusociality is one of the major transitions in evolution, but the underlying genomic changes are unknown. We compared the genomes of 10 bee species that vary in social complexity, representing multiple independent transitions in social evolution, and report three major findings. First, many important genes show evidence of neutral evolution as a consequence of relaxed selection with increasing social complexity. Second, there is no single road map to eusociality; independent evolutionary transitions in sociality have independent genetic underpinnings. Third, though clearly independent in detail, these transitions do have similar general features, including an increase in constrained protein evolution accompanied by increases in the potential for gene regulation and decreases in diversity and abundance of transposable elements. Eusociality may arise through different mechanisms each time, but would likely always involve an increase in the complexity of gene networks.

The evolution of eusociality involves changes in the unit of natural selection, from the individual to a group (*1*). Bees evolved eusociality multiple times and are extremely socially diverse (*2*) (Fig. 1), but all pollinate angiosperms, including many crops essential to the human diet (*3*). Simple eusociality may be facultative or obligate, and both forms are characterized by small colonies with a reproductive queen and one or more workers that, due to social and nutritional cues, forego reproduction to cooperatively care for their siblings (*2*). Further evolutionary elaborations have led to complex eusociality, "superorganisms" with colonies of several thousand individuals, sophisticated modes of communication, and morphological specializations for division of labor (*4*).

Theory predicts that the evolution of simple eusociality involves increased regulatory flexibility of ancestral gene networks to create specialized reproductive and nonreproductive individuals, and the evolution of complex eusociality requires genetic novelty to coordinate emergent properties of group dynamics (*5*). To test these predictions, we analyzed five de novo and five publicly available draft genome sequences of 10 bee species from three families, representing two independent origins of eusociality in Apidae and Halictidae and two independent elaborations of simple to complex eusociality in two apid tribes [Apini (honeybees) and Meliponini (stingless bees); Fig. 1]. The draft genomes were of comparable, high quality (supplementary materials).

We found that the transition from solitary to group life is associated with an increased capacity for gene regulation. We scanned the promoter regions of 5865 single-copy orthologs among the 10 species to calculate a motif score [representing the number and binding strength of experimentally characterized transcription factor binding sites (TFBSs)] for 188 *Drosophila melanogaster* TFs (*6*) with at least one ortholog in each of the 10 bees, and correlated motif score with social complexity, using phylogenetically independent contrasts (*7*). Of 2101 significantly correlated motif-gene pairs, 89% were positive and 11% negative, showing that TFs tend to have increased capacity to regulate genes in eusocial species of bees, relative to solitary species (Fig. 2A, supplementary materials).

Further evidence for increased capacity for gene regulation throughout social evolution is a positive ranked correlation between social complexity and the number of genes predicted to be methylated (*7*) (Spearman's rho = 0.76, $P = 0.01$; phylogenetically corrected Spearman's rho = 0.64, $P = 0.06$; Fig. 2B; bioinformatics predictions validated with bisulfite sequencing data for three invertebrate species; supplementary materials). DNA methylation affects gene expression in a variety of ways (*8*). Thus, this result suggests an expansion in regulatory capacity with increasingly sophisticated sociality.

The potential for increased regulatory capacity was further revealed at the protein-coding level. Increased social complexity also is associated with rapid evolution of genes involved in coordinating gene regulation. A Bayesian phylogenetic covariance analysis (*9*) of 5865 single-copy orthologs identified 162 genes with accelerated evolution in species with increased social complexity (*7*) (additional data table S3). These rapidly evolving genes were significantly enriched ($P < 0.05$) for Gene Ontology (GO) terms related to regulation of transcription, RNA splicing, ribosomal structure, and regulation of translation (supplementary text and tables S11 and S12). Similar results have been reported for bee and ant species (*10*–*13*); our findings reveal the underlying causes. Approximately two-thirds of these genes are under stronger directional selection in species with increasingly complex eusociality, but we also detected nonadaptive evolution. One-third of the rapidly evolving genes are under relaxed purifying selection in species with complex eusociality, possibly due to reduced effective population sizes (*14*).

We also found an additional 109 genes, significantly enriched ($P < 0.05$) for functions related to protein transport and neurogenesis, which evolve slower with increased social complexity (supplementary text, table S13, and additional data table S3). This includes orthologs of *derailed 2* and *frizzled*, which function as Wnt signaling receptors in *Drosophila* synaptogenesis (*15*), and *rigor mortis*, a nuclear receptor involved in hormone signaling (*16*). A similar pattern of reduced evolutionary rate has been described for genes expressed in human and honey bee brains, potentially due to increasing pleiotropic constraint in complex gene networks (*17*, *18*). Constrained protein evolution of neural and endocrine-related genes seems at odds with the evolution of complexity, but this constraint appears to be compensated for, or perhaps driven by, increased capacity for gene regulation.

We next investigated whether these molecular evolution patterns involve similar sets of genes and cis-regulatory elements among the early (facultative and obligate simple eusociality) and advanced (complex eusociality) stages of independent social transitions. We identified lineage-specific differences in coding sequences and promoter regions of 1526 "social genes" for which evolutionary rate (d*N*/d*S*) is faster or slower with increased social complexity in two independent origins and two independent elaborations of eusociality (*7*)

(Fig. 1). Among these lineage-specific social genes, we found common patterns of cis-regulatory evolution: gains of TFBSs in the promoters of genes that evolve slower with increasing social complexity (Fig. 2C and supplementary text). This suggests that a shared feature of both independent origins and elaborations of eusociality is increasingly constrained protein evolution with increasing potential for novel gene expression patterns. The TFs responsible for this pattern were different for each social transition, even though our analysis was limited to highly conserved TFs (Table 1). Several function in neurogenesis or neural plasticity, or are prominent regulators of endocrine-mediated brain gene expression in honeybees (*19*, *20*).

We found further lineage-specific differences among the rapidly evolving "social genes" themselves. Genes undergoing accelerated evolution at the origins of eusociality were significantly enriched for GO terms related to signal transduction in both Apidae and Halictidae, but they shared only six genes (6 out of 354 and 167 genes, respectively; hypergeometric test, $P = 0.82$; Fig. 2D and additional data tables S5 and S6). Rapid evolution of signal transduction pathways may be a necessary step in all origins of eusociality to mediate intracellular responses to novel social and environmental stimuli (*10*), but selection appears to have targeted different parts of these pathways in each independent transition. Caste-specific expression and other analyses of these genes are needed to determine their function in eusociality.

Genes showing signatures of rapid evolution with the elaborations of complex eusociality were also highly disparate between honeybees and stingless bees, with only 43 shared genes and no shared enriched GO terms (43 out of 625 and 512 genes, respectively; hypergeometric test, $P = 0.70$; Fig. 2D and additional data tables S5 and S6). In addition, only 2 out of 5865 single-copy orthologs showed a signature of convergent evolution by fitting a dendrogram based on social complexity significantly better than the accepted molecular phylogeny (*7*) (supplementary text and fig. S21). Similarly, families of major royal jelly protein genes, sex-determining genes, odorant receptors, and genes involved in lipid metabolism expanded in some, but not all, lineages of complex eusocial bees (*7*) (Table 2 and supplementary text). These results suggest that gene family expansion is associated with complex eusociality as predicted (*5*), but involves different genes in each case. Despite striking convergence of social traits among the superorganisms (*4*), the final stages of transformation to this level of biological organization do not necessarily involve common molecular pathways.

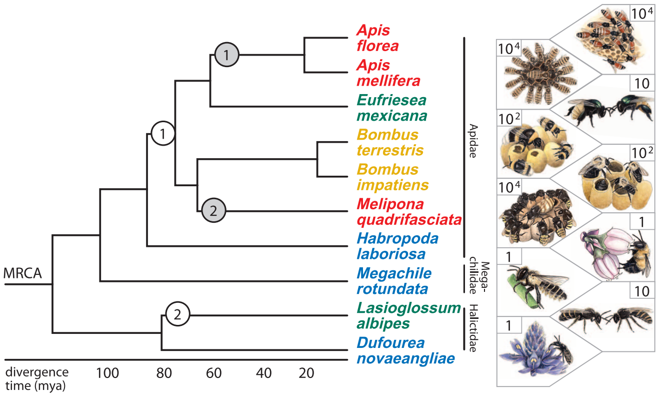


Fig. 1. Phylogeny and divergence times (*28*) of bees selected for genome analysis. We analyzed two independent origins of simple eusociality from a solitary ancestor, one each in Apidae (white circle 1) and Halictidae (white circle 2), and two independent elaborations of complex eusociality in honeybees (gray circle 1) and stingless bees (gray circle 2). Most bees mate once, but honeybees mate with multiple males. All bees eat pollen and nectar from flowering plants. Species names are colored according to degree of social complexity: blue: ancestrally solitary; green: facultative simple eusociality; orange: obligate simple eusociality; red: obligate complex eusociality. The social biology of *E. mexicana* is unknown, but is representative of the facultative simple eusocial life history (*29*). Numbers in each box are approximate colony size on a log scale. MRCA, most recent common ancestor; mya, millions of years ago.

[1]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [2]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]Department of Biology, Utah State University, Logan, UT 84322, USA. [4]China National GeneBank, BGI-Shenzhen, Shenzhen, 518083, China. [5]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, 1350, Denmark. [6]Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, MD 21218, USA. [7]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [8]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [9]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA. [10]Program in Ecology and Evolutionary Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [11]Department of Biology, Hobart and William Smith Colleges, Geneva, NY 14456, USA. [12]Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [13]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA. [14]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA. [15]U.S. Department of Agriculture–Agricultural Research Service (USDA-ARS) Red River Valley Agricultural Research Center, Biosciences Research Laboratory, Fargo, ND 58102, USA. [16]Center for Ecological Research and Forestry Applications (CREAF), Universitat Autonoma de Barcelona, 08193 Bellaterra, Spain. [17]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [18]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. [19]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [20]The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [21]Institute of Biology, Department Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, D-06099 Halle (Saale), Germany. [22]Queen Mary University of London, School of Biological and Chemical Sciences Organismal Biology Research Group, London E1 4NS, UK. [23]Department of Laboratory Medicine, University Hospital Halle, Ernst Grube Strasse 40, D-06120 Halle (Saale), Germany. [24]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, 04103 Leipzig, Germany. [25]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA. [26]Department of Entomology, University of Georgia, Griffin, GA 30223, USA. [27]Center for Functional and Comparative Insect Genomics, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [28]Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 14040-901 Ribeirão Preto, SP, Brazil. [29]Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista (UNESP), 14884-900 Jaboticabal, SP, Brazil. [30]Departamento de Genética e Evolução, Centro de Ciências Biológicas e da Saúde, Universidade Federal de São Carlos, 13565-905 São Carlos, SP, Brazil. [31]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900 Ribeirão Preto, SP, Brazil. [32]Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900 Ribeirão Preto, SP, Brazil. [33]USDA-ARS Bee Research Lab, Beltsville, MD 20705 USA. [34]Department of Biology, East Carolina University, Greenville, NC 27858, USA. [35]Department of Entomology, Ohio Agricultural Research and Development Center, Ohio State University, Wooster, OH 44691, USA. [36]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA. [37]Department of Animal Sciences, University of Illinois, Urbana, IL 61801, USA. [38]Department of Population Genomics, Institute of Animal Husbandry and Animal Breeding, University of Hohenheim, Germany. [39]Department of Biology, York University, Toronto, ON M3J 1P3, Canada. [40]Janelia Farm Research Campus, Howard Hughes Medical Institue, Ashburn, VA 20147, USA. [41]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [42]Department of Entomology, Texas A&M University, College Station, TX 77843, USA. [43]Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA. [44]Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark. [45]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [46]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. [47]Department of Medicine, University of Hong Kong, Hong Kong. [48]Center for Advanced Study Professor in Entomology and Neuroscience, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [49]Centre for Social Evolution, Department of Biology, Universitetsparken 15, University of Copenhagen, DK-2100 Copenhagen, Denmark.
*These authors contributed equally to this work. **†Corresponding author. E-mail: karen.kapheim@usu.edu (K.M.K.); wangj@genomics.org.cn (J.W.); generobi@illinois.edu (G.E.R.); zhanggj@genomics.org.cn (G.Z.)**

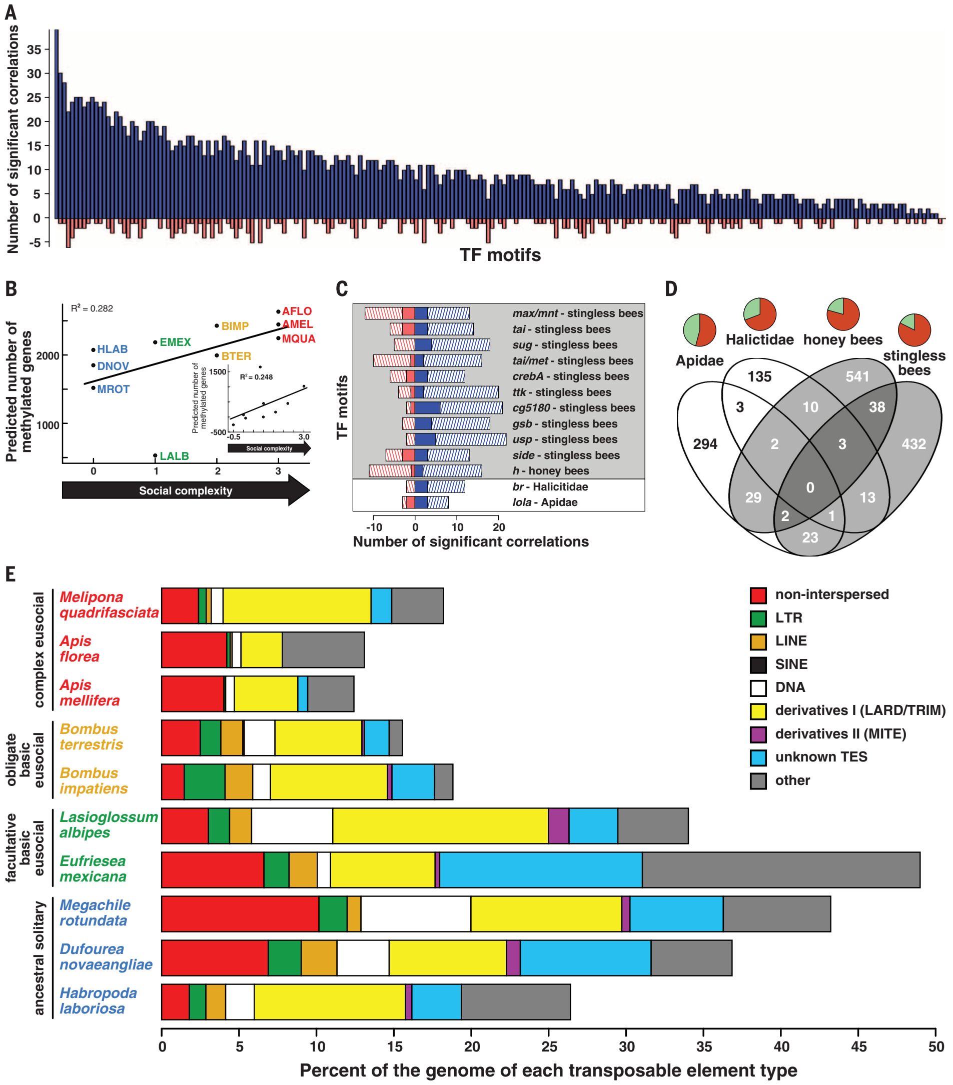


Fig. 2. Genomic signatures of evolutionary transitions from solitary to group life. (**A**) Increasing social complexity is associated with increasing presence of cis-regulatory TFBSs in promoter regions. Each bar represents a TFBS for which presence correlates significantly with social complexity (blue: positive; red: negative). (**B**) Relationship between predicted number of methylated genes and social complexity before and after (inset) phylogenetic correction (see text for statistics). (**C**) TFBS motifs showing a relationship between social complexity and evolutionary rate of coding and noncoding sequences in different lineages. Bar length indicates the number of significant correlations (blue: positive; red: negative) between each motif score and social complexity (from Table 1) among genes evolving faster (solid) or slower (hatched) in lineages with different levels of social complexity [from (D)]. Background shading follows circle shading in Fig. 1. (**D**) Number of genes for which evolutionary rate is faster or slower in lineages with higher compared to lower social complexity. Pie charts represent the proportion of genes evolving slower (light green) or faster (dark orange) with increased social complexity. Venn diagram shading follows circle shading in Fig. 1. (**E**) Complex eusocial species have a reduced proportion of repetitive DNA compared to other bees (see text for statistics). LTR, long terminal repeat; LINE, long interspersed element; SINE, short interspersed element; DNA, DNA transposon; LARD, large retrotransposon derivative; TRIM, terminal repeat retrotransposon in miniature; MITE, miniature inverted-repeat transposable element; TES, transposable elements.

The major transitions in evolution involve a reduction in conflict as the level of natural selection rises from the individual to the group (*1*). Extending this to intragenomic conflict may explain our finding of decreased diversity and abundance of transposable elements (TEs) with increasing social complexity (*7*) (regression after phylogenetic correction, F = 8.99, adjusted R^2 = 0.47, P = 0.017; Fig. 2E, figs. S42 to S44, and supplementary text). This may be a consequence of increased recombination rates among highly eusocial insects (*21*, *22*) or because key features of complex eusociality lead to decreased exposure to parasites and pathogens that horizontally transmit TEs (*4*, *23*). Eusociality in bees may thus provide natural immunity against certain types of intragenomic conflict.

Our results and those in (*10*–*13*) support the prediction that changes in gene regulation are key features of evolutionary transitions in biological organization (*5*). Our results further reveal the convergent adaptive and nonadaptive evolutionary processes common to both the early and advanced stages of multiple independent transitions from solitary to group living. It is now clear that there are lineage-specific genetic changes associated with independent origins of eusociality in bees, and independent elaborations of eusociality in both bees and ants. This includes different sets of genes showing caste-biased expression across species (*24*–*26*) and, as we have shown, evolutionary modifications of TEs, gene methylation, and cis-regulatory patterns associated with the suite of life-history traits that define eusociality. This suggests that if it were possible to "replay life's tape" (*27*), eusociality may arise through different mechanisms each time, but would likely always involve an increase in the complexity of gene networks.

Table 1. Transcription factors (TFs) and corresponding motifs associated with origins and elaborations of eusociality in bees. [Motif names: Fly Factor Survey (*6*); supplementary text.]

Motif	*D. melanogaster* TFs	Hypergeometric test *P*-value
	Solitary to simple eusociality–Apidae	
lola_PQ_SOLEXA	*Lola*	0.0047
	Solitary to simple eusociality–Halictidae	
br_PL_SOLEXA_5	*Br*	0.0016
	Simple eusociality to complex eusociality–honeybees	
h_SOLEXA_5	*dpn,h*	0.0027
	Simple eusociality to complex eusociality–stingless bees	
Side_SOLEXA_5	*E_spl, HLHm3, HLHm5, HLHm7, HLHmbeta, HLHmdelta, HLHmgamma, Side*	0.0008
usp_SOLEXA	*EcR,svp,usp*	0.0013
CrebA_SOLEXA	*CrebA*	0.0040
CG5180_SOLEXA	CG5180	0.0044
tai_Met_SOLEXA_5	*Mio_bigmax,tai_Met*	0.0045
ttk_PA_SOLEXA_5	*Ttk*	0.0078
gsb_SOLEXA	*gsb,Poxn,prd*	0.0083
tai_SOLEXA_5	*Tai*	0.0100

Table 2. Relative size of select gene families as related to social complexity in bees.

Family	Function	Eusocial bees compared to solitary bees
	Differences among bees	
Major royal jelly	Brood feeding	Expanded only in *Apis*
Sex determination pathway genes	Sex-specific development	Expanded in some eusocial lineages
Odorant receptors	Olfaction	Expanded in complex eusocial lineages
Lipid metabolism genes	Metabolic processing of lipids	Expanded in complex eusocial lineages
	Similarities across bees	
Biogenic amines receptors, neuropeptides, GPCRs*	Neural plasticity	Similar
Insulin-signaling and ecdysone pathway genes	Insect development, caste determination in honeybees, behavioral plasticity as adults	Similar
Immunity	Infectious disease protection	Similar
Cytochrome P450 monooxygenase genes	Detoxification	Similar

*GPCRs, G protein–coupled receptors.

REFERENCES AND NOTES

1. J. Maynard Smith, E. Szathmáry, *The Major Transitions in Evolution* (Oxford Univ. Press, Oxford, UK, 1995).
2. C. D. Michener, *The Social Behavior of the Bees* (Harvard Univ. Press, Cambridge, MA, 1974).
3. A.-M. Klein *et al.*, *Proc. Biol. Sci. B* **274**, 303–313 (2007).
4. H. Hölldobler, E. O. Wilson, *The Superorganism: The Beauty, Elegance and Strangeness of Insect Societies* (Norton, New York, 2009).
5. B. R. Johnson, T. A. Linksvayer, *Q. Rev. Biol.* **85**, 57–79 (2010).
6. L. J. Zhu *et al.*, *Nucleic Acids Res.* **39**, D111–D117 (2011).
7. Materials and methods are available as supplementary materials on *Science* Online.
8. H. Yan *et al.*, *Annu. Rev. Entomol.* **60**, 435–452 (2015).
9. N. Lartillot, R. Poujol, *Mol. Biol. Evol.* **28**, 729–744 (2011).
10. S. H. Woodard *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7472–7477 (2011).
11. B. A. Harpur *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2614–2619 (2014).
12. J. Roux *et al.*, *Mol. Biol. Evol.* **31**, 1661–1685 (2014).
13. D. F. Simola *et al.*, *Genome Res.* **23**, 1235–1247 (2013).
14. J. Romiguier *et al.*, *J. Evol. Biol.* **27**, 593–603 (2014).
15. M. Park, K. Shen, *EMBO J.* **31**, 2697–2704 (2012).
16. J. Gates, G. Lam, J. A. Ortiz, R. Losson, C. S. Thummel, *Development* **131**, 25–36 (2004).
17. D. Brawand *et al.*, *Nature* **478**, 343–348 (2011).
18. D. Molodtsova, B. A. Harpur, C. F. Kent, K. Seevananthan, A. Zayed, *Front. Genet.* **5**, 431 (2014).
19. D. W. Pfaff, A. P. Arnold, A. M. Etgen, R. T. Rubin, S. E. Fahrbach, Eds., *Hormones, Brain and Behavior* (Elsevier, New York, 2009).
20. S. Chandrasekaran *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18020–18025 (2011).
21. L. Wilfert, J. Gadau, P. Schmid-Hempel, *Heredity* **98**, 189–197 (2007).
22. E. S. Dolgin, B. Charlesworth, *Genetics* **178**, 2169–2177 (2008).
23. S. Schaack, C. Gilbert, C. Feschotte, *Trends Ecol. Evol.* **25**, 537–546 (2010).
24. B. Feldmeyer, D. Elsner, S. Foitzik, *Mol. Ecol.* **23**, 151–161 (2014).
25. P. G. Ferreira *et al.*, *Genome Biol.* **14**, R20 (2013).
26. B. G. Hunt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15936–15941 (2011).
27. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (Norton, New York, 1989).
28. S. Cardinal, B. N. Danforth, *Proc. Biol. Sci.* **280**, 20122686 (2013).
29. S. Cardinal, B. N. Danforth, *PLOS ONE* **6**, e21086 (2011).

ACKNOWLEDGMENTS

Data deposition at National Center for Biotechnology Information: *H. laboriosa*, *D. novaeangliae*, *E. mexicana*, *M. quadrifasciata*, and *M. rotundata* genome assemblies accession numbers PRJNA279436, PRJNA279825, PRJNA279814, PRJNA279820, and PRJNA66515. Funding for genome sequencing and analysis: BGI, a U.S. National Institutes of Health Pioneer Award (DP1 OD006416) to G.E.R., and European Union Marie Curie International Incoming Fellowship (300837) to G.Z. Additional funding: U.S. National Science Foundation grant DEB-0640690 (M.A.D.G.) and DEB-0743154 (G.E.R. and M.E.H); Danish Council for Independent Research grants 10-081390 (C.L.) and 0602-01170B (C.J.P.G.); Lundbeck Foundation (C.J.P.G); Georgia Tech–Elizabeth Smithgall Watts endowment (M.A.D.G.); Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312 (R.M.W.); Swiss National Science Foundation award 31003A-125350, Commission Informatique of the University of Geneva, Schmidheiny Foundation, and Swiss Institute of Bioinformatics (E.M.Z.); and Natural Sciences and Engineering Research Council of Canada Discovery Grant and Early Research Award (A.Z.). This project was conducted under the auspices of the i5K Initiative. We thank the Roy J. Carver Biotechnology Center (sequencing services); N. Lartillot (advice on Coevol); T. Newman (assistance with DNA extractions); and E. Hadley (assistance with figures). Computational support: D. Davidson, N. Band, D. Slater (University of Illinois), J. H. Kidner and H. Scharpenberg (Martin-Luther-University Halle-Wittenberg); Compute Canada; and Center for High Performance Computing at the University of Utah. J. Himes created the illustrations in Fig. 1. We are grateful to T. Pitts-Singer, H. G. Hall, B. N. Danforth, J. Gibbs, and S. Cardinal for providing bee specimens; R. Ayala for identification of *E. mexicana*; J. Vega and the staff at Estación de Biología Chamela, Universidad Nacional Autónoma de México (UNAM), for support during collecting trips; and S. A. Cameron for insightful discussion.

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/348/6239/1139/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S44
Tables S1 to S32
Additional Data Tables S1 to S12
References (*30*–*151*)
Author Contributions

19 December 2014; accepted 6 May 2015
Published online 14 May 2015;
10.1126/science.aaa4788

HUMAN OOCYTES

Error-prone chromosome-mediated spindle assembly favors chromosome segregation defects in human oocytes

Zuzana Holubcová,[1] Martyn Blayney,[2] Kay Elder,[2] Melina Schuh[1]*

Aneuploidy in human eggs is the leading cause of pregnancy loss and several genetic disorders such as Down syndrome. Most aneuploidy results from chromosome segregation errors during the meiotic divisions of an oocyte, the egg's progenitor cell. The basis for particularly error-prone chromosome segregation in human oocytes is not known. We analyzed meiosis in more than 100 live human oocytes and identified an error-prone chromosome-mediated spindle assembly mechanism as a major contributor to chromosome segregation defects. Human oocytes assembled a meiotic spindle independently of either centrosomes or other microtubule organizing centers. Instead, spindle assembly was mediated by chromosomes and the small guanosine triphosphatase Ran in a process requiring ~16 hours. This unusually long spindle assembly period was marked by intrinsic spindle instability and abnormal kinetochore-microtubule attachments, which favor chromosome segregation errors and provide a possible explanation for high rates of aneuploidy in human eggs.

Meiosis in human oocytes is more prone to chromosome segregation errors than mitosis (*1*, *2*), meiosis during spermatogenesis (*3*, *4*), and female meiosis in other organisms (*3*, *5*). Despite its importance for fertility and human development, meiosis in human eggs has hardly been studied. Human oocytes are only available in small numbers, warranting single-cell assays capable of extracting maximal information. Although high-resolution live-cell microscopy is an ideal method, oocyte development in the ovary poses challenges to direct imaging. We therefore established an experimental system (*6*) for ex vivo high-resolution fluorescence microscopy of human oocytes freshly harvested from women undergoing gonadotropin-stimulated in vitro fertilization cycles. To establish the major stages of meiosis in this system, we simultaneously monitored microtubules and chromosomes for ~24 to 48 hours (Fig. 1 and movie S1). Similar to the situation in situ (*7*), human oocytes matured into fertilizable eggs over this time course, as judged by the formation of a polar body. The morphologically identifiable stages (Fig. 1A) at characteristic times after nuclear envelope breakdown [(NEBD), set to 0 hours] provided a time-resolved framework for human oocyte meiosis (Fig. 1B). This reference timeline post-NEBD is used throughout this paper.

Before NEBD, chromosomes were highly condensed and clustered around the nucleolus. Instead of rapidly nucleating microtubules upon NEBD, human oocytes first formed a chromosome aggregate that was largely devoid of microtubules (Fig. 1A; movie S1; and fig. S1, A and B). Microtubules were first observed at ~5 hours, when they started to form a small aster within the chromosome aggregate. As the microtubule aster grew, the chromosomes became individualized and oriented on the surface of the aster with their kinetochores facing inwards. The microtubule aster then extended into an early bipolar spindle that carried the chromosomes on its surface (Fig. 1A; movie S1; and fig. S1, C to E). The chromosomes then entered the spindle but remained distributed throughout the entire spindle volume. Chromosomes first congressed in the spindle center at ~13 hours but continued to oscillate around the spindle equator. Stable chromosome alignment was typically only achieved close to anaphase onset (Fig. 1, A and B, and movie S1). Unexpectedly, the spindle volume increased over the entire course of meiosis, up until anaphase onset (Fig. 1, C and D). The barrel-shaped spindle formed in this process consisted of loosely clustered bundles of microtubules and lacked astral microtubules (movie S2 and fig. S2). At ~17 hours, the oocytes progressed into anaphase and eliminated half of the homologous chromosomes in a polar body. Nearly a day after NEBD, the oocytes had formed a bipolar metaphase II spindle and matured into a fertilizable egg. The stages and timing of meiosis were highly reproducible among oocytes (Fig. 1, A and B) and could also be observed in fixed oocytes (fig. S1, A to I). Importantly, 79.0% of imaged human oocytes extruded a polar body. This indicates that the imaging assays, as well as the methods by which the oocytes were obtained and processed, did not have a prominent effect on meiotic progression.

The surprisingly slow and gradual build-up of the spindle over 16 hours (Fig. 1, C and D) is in stark contrast to mitosis, where spindle assembly takes only ~30 min (*8*), or meiosis in mouse oocytes, where it takes 3 to 5 hours (*9*–*11*). During mitosis, two centrosomes ensure the rapid assembly of a spindle. In oocytes of many species, centrosomes are absent but functionally replaced by microtubule organizing centers (MTOCs) that lack centrioles (*9*, *12*). Human oocytes also lack centrosomes (*13*–*15*), but whether acentriolar MTOCs participate in spindle assembly is unclear (*16*–*19*). We consistently detected pericentrin- and γ-tubulin–positive MTOCs at the spindle poles of mitotic cells and metaphase I and II (MI and MII) mouse oocytes, but never at MI or MII spindles in human oocytes (Fig. 2, A and B, and fig. S3). Thus, our data suggest that meiotic spindles in human oocytes lack detectable MTOCs.

In *Xenopus* egg extracts, chromosomes can serve as sites of microtubule nucleation if centrosomes are absent (*20*). The human oocytes we imaged also initiated microtubule nucleation in the region of the chromosome aggregate (78 of 78 live human oocytes). High-resolution imaging of fixed human oocytes confirmed that microtubules were first nucleated on chromosomes, emanating primarily from kinetochores (Fig. 2C, movie S3, and fig. S4). MTOC-nucleated cytoplasmic asters, such as those seen in chromosomal proximity upon NEBD in mouse oocytes (*9*), could not be detected. Thus, chromosomes, not MTOCs, serve as major sites of microtubule nucleation in human oocytes.

[1] Medical Research Council, Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK. [2] Bourn Hall Clinic, Bourn, Cambridge CB23 2TN, UK.
*Corresponding author. E-mail: mschuh@mrc-lmb.cam.ac.uk