



UNIVERSITA' DEGLI STUDI DI VERONA

DIPARTIMENTO DI BIOTECNOLOGIE

SCUOLA DI DOTTORATO DI SCIENZE INGEGNERIA MEDICINA

DOTTORATO DI RICERCA IN BIOTECNOLOGIE APPLICATE

CICLO XXVIII

**Structural annotation of eukaryotic genomes
in 2nd generation sequencing era**

S.S.D. BIO/18

Coordinatore: Prof.ssa Paola Dominici

Tutor: Prof. Massimo Delledonne

Dottoranda: Dott.ssa Alessandra Dal Molin

UNIVERSITA' DEGLI STUDI DI VERONA

DIPARTIMENTO DI BIOTECNOLOGIE

SCUOLA DI DOTTORATO DI

SCIENZE INGEGNERIA MEDICINA

DOTTORATO DI RICERCA IN

BIOTECNOLOGIE APPLICATE

CICLO XXVIII

**Structural annotation of eukaryotic genomes
in 2nd generation sequencing era**

S.S.D. BIO/18

Coordinatore: Prof.ssa Paola Dominici

Tutor: Prof. Massimo Delledonne

Dottoranda: Dott.ssa Alessandra Dal Molin

Quest'opera è stata rilasciata con licenza Creative Commons Attribuzione – non commerciale

Non opere derivate 3.0 Italia . Per leggere una copia della licenza visita il sito web:

<http://creativecommons.org/licenses/by-nc-nd/3.0/it/>



Attribuzione Devi riconoscere una menzione di paternità adeguata, fornire un link alla licenza e indicare se sono state effettuate delle modifiche. Puoi fare ciò in qualsiasi maniera ragionevole possibile, ma non con modalità tali da suggerire che il licenziante avalli te o il tuo utilizzo del materiale.



NonCommerciale Non puoi usare il materiale per scopi commerciali.



Non opere derivate —Se remixi, trasformi il materiale o ti basi su di esso, non puoi distribuire il materiale così modificato.

Structural annotation of eukaryotic genomes in 2nd generation sequencing era
Alessandra Dal Molin
Tesi di Dottorato
Verona, 9 Maggio 2016
ISBN 978-88-69250-07-1

INDEX

ABSTRACT.....	5
INTRODUCTION.....	6
<i>Genome annotation in 2nd generation era.....</i>	<i>6</i>
<i>Structural annotation of eukaryotic genomes.....</i>	<i>7</i>
Main concepts.....	7
Computational Strategies.....	10
Repeat identification and masking.....	10
Non-coding RNAs annotation.....	13
Protein coding gene annotation.....	14
Extrinsic approach.....	16
Intrinsic approach.....	18
<i>Automated genome annotation and quality assessment.....</i>	<i>22</i>
AIM OF THESIS.....	27
EXPERIMENTAL CASES.....	28
<i>Annotation of a genome with a closely related reference.....</i>	<i>28</i>
Background.....	28
Materials.....	29
Methods.....	31
Results and discussion.....	35
<i>Annotation of a genome with a close but phylogenetically distinct reference...</i>	<i>43</i>
Background.....	43
Materials.....	44
Methods.....	46

Results and discussion.....	49
<i>Annotation of a genome with no reported reference.....</i>	<i>59</i>
Background.....	59
Materials.....	60
Methods.....	61
Results and discussion.....	65
CONCLUSIONS.....	76
BIBLIOGRAPHY.....	79

ABSTRACT

In the last decade the increase in efficiency and decrease in cost of new sequencing techniques led to a growing amount of genomic sequences in public databases. With this huge volume of sequences being generated from high-throughput sequencing projects, the requirement for providing accurate and detailed genome annotations has never been greater. Structural genome annotation is the process of identifying structural features in a DNA sequence and classifying them based on their biological role. Computer programs are increasingly used to perform structural annotation since they meet the high-throughput demands of genome sequencing projects even if they are less accurate than manual gene annotation which remains the ‘golden-standard’ for evaluating annotation confidence and quality.

The aim of this project is to meet the need of producing fast and accurate genome annotation by applying available computational means to different experimental cases, depending on the biological knowledge achieved so far and the quality of starting data. The contribution of different methods used to produce the final annotation has been analyzed along with the evaluation of results for the completeness of the study.

The results obtained showed that the complexity of eukaryotic genomes greatly affects the annotation process; a big fraction of the genes in a genome sequence can be found mostly by homology to other known genes or proteins and by the use of *ab initio* predictors and species-specific evidence. The integration of multiple sources of annotation greatly improved the accuracy of the final genome annotations, anyway being not error free. Quality assessment of results and filtering of low confidence sequences together with manual revision are always required to achieve higher accuracy.

INTRODUCTION

Genome annotation in 2nd generation sequencing era

In response to the increasing demand of complete genome and transcriptome sequences the last decade has seen the development and release of so called “Next Generation Sequencing” technologies, which offer a far higher throughput and lower sequencing costs compared to their prior ones¹⁻³. Indeed, with the release of Illumina Genome Analyzer in 2005, the use of short-read massively-parallel sequencing took sequencing runs from producing 84 kilobase (kb) per run to 1 Gigabase (Gb) per run, revolutionizing sequencing capabilities and launching the “next-generation” in genomic science. More recently, sequencing costs have fallen so dramatically that a single laboratory can afford to sequence large genomes in a relatively short time and researchers can analyze thousands to tens of thousands of samples in a single year. Indeed by 2014, the sequencing rate climbed to 1.8 Terabases (Tb) per run whereas, before NGS, the genome sequencing projects were massive in terms of time and costs, involving several research groups. As a matter of fact, the first human genome published in 2001 required 15 years to sequence and cost nearly 3 billion dollars in contrast to the recently released instruments, sequencing over 45 human genomes in a single day for approximately 1000 dollars each.

Although sequencing has become easy in many ways, genome annotation has become more challenging. Indeed, before NGS the genome sequencing projects were undertaken by consortia, such as the *C. elegans* Sequencing Consortium and the *Arabidopsis* Genome Initiative. In these cases the annotation process was performed on deeply-studied organisms, which genomes were reconstructed from the assembly of random cosmid clones with long inserts, YACs, fosmids⁴, large-insert Bacterial Artificial Chromosomes (BAC), phages, Transformation-competent Artificial Chromosome libraries (TAC) and Inverse Polymerase Chain

Reaction (IPCR) products derived from genomic DNA⁵. In the NGS era, the shorter read length of second-generation sequencing platforms (50–250 bp, depending on the platform) prevents current genome assemblies from fully attaining the contiguity of classic genome shotgun assemblies, while the exotic nature of many recently sequenced genomes complicates the already challenging gene annotation¹¹. Indeed, whereas the first genome projects could recur 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 tools¹⁷. Anyway, the exponential accumulation of genomic sequences in public databases requires fast, accurate and detailed annotations of an increasing amount of gene products. The availability of such resources of data, bioinformatics techniques as well as high throughput computing with limited manual annotation enhances the need of reducing the growing gap between the number of sequences and annotations⁷. Despite significant improvements, the accurate identification and structural elucidation of protein coding genes remains challenging. While automatically generated annotations are perfectible, manual gene annotation still remains the ‘golden-standard’ for evaluating annotation confidence and quality.

Structural annotation of eukaryotic genomes

Main concepts

By genome annotation we commonly refer to **the process of identifying structural features on the genome sequence and determining their biological function**. The annotation process is composed of two main steps: (1) structural annotation and (2) functional annotation. In the first step, known classes of elements encoded by the genome sequence are identified and properly labeled, namely ‘retrotransposon’, ‘protein-coding gene’, ‘ribosomal RNA’, and many others. The latter step describes the biological meaning of the identified elements as part of a certain process. In this thesis I will focus on the structural annotation of eukaryotic genomes, in particular of fungal and plant genomes.

Eukaryotic genomes vary from tens Megabases to several Gigabases involving a much more complex organization compared to prokaryotic genomes, which are small in size lacking introns and repetitive regions.

Fungal, plant and animal genomes contain similar numbers of protein coding genes and average coding sequence lengths, which range from 1.3 and 1.9 kb. In general, higher eukaryotes, as plants and animals, show lower gene densities with respect to their genome size and consecutively are characterized by longer intergenic regions, on average 3.9 kb in *Arabidopsis* and 3.3 kb in maize, as opposed to a range of 80-150 bp in many ascomycetes.

Lower eukaryotes, such as fungal genomes, display coding densities ranging from 37% to 61% with typically few and short introns. Intron densities in fungi range averagely from 5-6 introns per gene in basidiomycetes such as *Cryptococcus neoformans* (Loftus et al. 2005), to 1-2 introns per gene for many sequenced ascomycetes (e.g. *Neurospora crassa*, *Magnaporthe oryzae*) (Galagan et al. 2003; Borkovich et al. 2004; Dean et al. 2005), to <300 introns in total in the hemiascomycete yeast *Saccharomyces cerevisiae* (Goffeau et al. 1996). An exception to the rule is the basidiomycete *C. neoformans* possessing an unusual wide range of intron sizes, from 68 bp to 35 bp, where the shorter are the most represented⁸.

Fungal genomes are relatively densely populated with genes, which are characterized by a significant variation in exon-intron structure. For example, fungal introns contain several short sequences required to perform an efficient splicing, like acceptor and donor sites at either end of the intron and 'Branch Point sites' (BP sites)⁶.

Regarding plants, intron lengths are generally higher. In tomato *Solanum lycopersicum*, the median intron length is 264 bp with an intron density of 3 introns per gene, whereas in *Arabidopsis thaliana* and rice is shorter (100 bp and 145 bp, respectively).

Given the significant differences in the characteristics of exons and introns between lower and higher eukaryotes, the training of gene prediction tools on organism-specific data is paramount⁸. The relatively simple gene structures of

most fungi facilitate accurate gene prediction. Gene prediction in fungi has relied heavily on the *de novo* gene prediction as the majority of fungal species lack significant EST data. On the other hand, plants can usually rely on EST data or full-length cDNAs, a powerful resource in gene prediction. We described how the genome size affects the gene structure all alone, but several additional factors related to recombination rate, expression level and effective population size, are involved as well.

Repetitive, or interspersed, elements are an important feature of eukaryotic genomes, and indeed account for a large proportion of the variation in genome size. The repetitive DNA fraction may represent a high proportion of a particular genome and abundance of repetitive sequences correlates with genome size explaining the differences in genomic DNA contents of different species⁸.

The typical repeat content of fungal genomes ranges between 3% and 64%, often increasing the difficulties in achieving an highly contiguous assembly⁸. The majority of the repeat sequences are associated with mobile genetic elements, copies or remnants of retrotransposons or DNA transposons, likely concentrated in few chromosomes that are rich in genes related to pathogenicity. High levels of repetitiveness are also found in plant genomes, i.e. transposable elements cover >80% of the maize genome⁹. A major class, the retroelements, encode the proteins necessary for their own reverse transcription and integration, and sometimes represent the 50% of the genome⁹.

The process of genome annotation is focused mainly on the detection and annotation of repetitive regions and protein-coding genes, although recently there has been an increased interest in other functional elements, such as non-coding RNAs. The techniques used for annotating repeats, non-coding RNAs and protein-coding genes are distinct and will be described hereafter¹⁰.

Computational strategies

Repeat identification and masking

The identification and masking of repetitive regions is usually performed as the first step of genome annotation, in order to exclude repetitive regions during the gene annotation phase and try to avoid the introduction of biases in the following analyses¹¹.

Repeats occur in all shapes and sizes: they can be widely interspersed repeats, tandem repeats or nested repeats, they may comprise just two copies or millions of copies, and they can range in size from 1–2 bases (mono- and di-nucleotide repeats) to hundreds of thousands of bases⁹ (Table 1). Well-characterized repeats are sometimes separated into two classes:

- short tandem repeats, also called micro- and mini-satellites;
- interspersed repeats, called short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs).

Tandem DNA repeats (≥ 2 bp in length) are adjacent to each other and can involve as few as two copies or many thousands of copies. Centromeres and telomeres are largely comprised of tandem repeats.

Interspersed repeats are identical or nearly identical DNA sequences which occur in the genome every hundreds, thousands or even millions of nucleotides⁹. Repeats can be spread out through the genome by mechanisms such as transposition. Short interspersed nuclear elements (SINEs) are repetitive DNA elements typically of 100–300 bp in length, while long interspersed nuclear elements (LINEs) are typically larger of 300 bp; both SINEs and LINEs spread throughout the whole genome. Repeats can also take the form of large-scale segmental duplications, such as those found on some human chromosomes and even whole-genome duplication, such as in the *Arabidopsis thaliana* genome⁹.

Repeat class	Repeat type	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	2-100
SINE	Interspersed	100-300
DNA transposons	Interspersed	200-2,000
LTR retrotransposon	Interspersed	200-5,000
LINE	Interspersed	500-8,000
rDNA (16S,18S,5.8S and 28S)	Tandem	2,000-43,000
Segmental duplications and other classes	Tandem or Interspersed	1,000-100,000

Table 1. Classes of repeats in eukaryotic genomes.

The tools used to identify repeats are distinct from those used to identify protein coding genes. Available tools for repeat identification generally fall into two classes: homology-based tools and *de novo* tools. Repeats are often poorly conserved and their accurate detection is usually increased when users create a *de novo* repeat library for their own genome of interest. However, *de novo* tools identify repeated sequences — not just mobile elements — so their outputs can include members of highly conserved protein-coding gene families, such as histones and tubulins, in addition to transposon sequences. Users must, therefore, carefully post-process the outputs of these tools to remove protein-coding sequences. Moreover, a high level of fragmentation in the genome assembly may cause the absence of sequence contiguity, shortening the scaffold length and even losing information¹¹. Repeats are interesting in and of themselves, and the life cycles and phylogenetic histories of these elements are growing areas of research. Adequate repeat annotation should thus be a part of every genome annotation project¹¹.

Several repeats library annotation pipelines and protocols have been developed. As an example, REPET pipeline allows the identification and annotation of Transposable Elements (TEs) through two main phases: TEdenovo and TEannot. Briefly, the first one performs self-alignment of the genome and clustering to obtain an initial repeat consensus. In the second phase, the draft repeat consensus is used to mask the genome, followed by detection of SSRs and final repeats annotation export. In the repeat detection phase, REPET can be fed with custom

set of repeats and can check for the presence of potential host genes (potential species-specific genes).

Another software which includes a repeat detection tool is MAKER, which is bundled with a repeat library creation protocol, which is combination of structural-based and homology-based approach, used to maximize the opportunity for repeat collection:

1. MITEs (Miniature Inverted Repeat Transposable Elements);
2. LTR (Long Terminal Repeat) retrotransposons;
3. Collection of repetitive sequences by RepeatModeler;
4. Exclusion of gene fragments.

Firstly, MITEs are collected using MITE-Hunter⁸¹ with all default parameters; then LTR retrotransposons are collected using LTRharvest⁸² and filtered by LTRdigest⁸² (tool from the GenomeTools suite) and other custom programs. In plants, LTR retrotransposons represent the largest genomic percentage of all repeats, increasing the importance of collecting this type of elements with high confidence. Secondly, the TE sequences containing significant gaps (more than 50 Ns) are excluded from the analysis, since even after the above procedures, a considerable amounts of false positives could be generated.

Retrotransposons are frequently nested with each other or inserted by other elements. When unidentified, misclassification may occur along with other complications. To properly detect retrotransposones, LTR sequences from candidate elements are used to mask the putative internal regions. The detection of LTR sequences in the internal regions defines the case of elements nested with other insertions.

The last step implies the exclusion of potential gene fragments from the final library by searching against a plant protein database, which does not store proteins from transposons.

After it has been created, a repeat library can be used in conjunction with tools like RepeatMasker¹², which uses BLAST¹³ and crossmatch to identify stretches of

sequence in a target genome that are homologous to known repeats, and/or RepeatRunner, which integrates RepeatMasker with BLASTX¹³ to search a database of repeat encoded proteins (e.g., reverse transcriptases) providing a comprehensive way of identifying repetitive elements. The ‘masking’ action transforms 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 known as ‘soft masking’. Masking repeated regions helps the downstream sequence alignment and gene prediction tools to recognize these regions as repeats. The failure of properly masking repeats can be the reason of reduced accuracy in the final annotation. Unmasked repeats can seed millions of spurious BLAST alignments, producing false evidence for gene annotations¹¹. Furthermore, many transposons’ Open Reading Frames (ORFs) are mistaken for true host genes by gene predictors, causing portions of transposons’ ORFs to be considered as additional exons to gene predictions, extensively corrupting the final gene annotations. In conclusion, good repeat masking has been proven to be crucial for the accurate annotation of protein-coding genes¹¹.

Non-coding RNAs annotation

Non-coding RNAs (ncRNAs), also known as the secret regulators of the cells, have been discovered 20 years ago, but only recently the attention of the scientific community has been turned towards the structure and function of ncRNAs¹⁴.

NcRNAs are classified as (1) infrastructural and (2) regulatory ncRNAs. Infrastructural ncRNAs seem to have a housekeeping role in translation and splicing and include RNA components like ribosomal, transfer and small nuclear RNAs¹⁵. Regulatory ncRNAs are more interesting from an epigenetic point of view as they are involved in the modification of other RNAs. Non-coding RNA genes include RNAs, such as small nuclear (snRNAs), small nucleolar (snoRNAs) and telomere-associated RNAs (TERC, TERRA); while do not include small ncRNAs, such as microRNAs (miRNAs), endogenous small interfering (endo-siRNAs) that participate in RNA interference (RNAi), Piwi-associated (piRNAs) and long non-coding RNAs¹⁵.

Many of the newly identified ncRNAs have not been functionally characterized yet, raising the possibility those component are non-functional, the reason why they have been referred to as ‘junk RNA’ in the last few years, likely to be products of spurious transcription¹⁶.

The heterogeneity and poorly conserved nature of many ncRNA genes present a major challenge in annotation pipelines. Indeed, unlike protein-encoding genes, ncRNAs are usually not well-conserved at the primary sequence level and even when they are, nucleotide homologies are not as easily detected as protein homologies¹¹.

A common approach to identify ncRNA genes involves the detection of conserved secondary structures and motives, for example using Infernal and Rfam database, to triage and classify the genes depending on primary and secondary sequence similarities. The analysis of RNA sequencing (RNA-seq) greatly improves ncRNAs identification. In particular, miRNAs can be directly identified using specialized RNA preps and sequencing protocols. Despite such sophisticated tools and techniques, distinguishing between *bona fide* ncRNA genes, spurious transcription and poorly conserved protein-encoding genes producing small peptides is still difficult to accomplish, especially if long intergenic non-coding RNAs (lincRNAs) and expressed pseudogenes are involved¹¹.

Although advancing rapidly ncRNA annotation is cutting edge showing accuracies generally much lower than their protein-coding counterparts¹¹. Indeed, ncRNAs annotation is still in its infancy compared with protein-coding gene annotation, but it is. Current annotation pipelines in some cases also allow the integration of ncRNAs annotations.

Protein coding gene annotation

The protein-coding gene annotation, the important step of genome annotation, usually integrates various resources to compute consensus gene structures. The typical gene structure of eukaryotic genes consist of exonic regions alternated by intronic regions. Generally, all exons can be separated into four classes: 5' exons, internal exons, 3' exons and intronless exons (as known as monoexonic genes)¹⁷.

In vertebrate organisms, genes are typically characterized by several exons and the precise identification of internal coding exons represent the most delicate step in gene-prediction algorithms.

The terms ‘gene prediction’ and ‘gene annotation’ are often wrongly used as synonyms. With a few exceptions, gene predictors identify for each gene the most likely Coding Sequence (CDS) with no mentioning of Untranslated Regions (UTRs) or any alternative splicing¹⁸. Gene annotations, conversely, might include UTRs, alternative splice isoforms and have attributes such as evidence trails. Gene annotation is, thus, a far 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 databases¹¹.

The information used to annotate genes comes generally from three types of analysis: (i) *ab initio* gene finding programs, which are runs on the DNA sequence to predict protein-coding genes; (ii) alignments of cDNAs and expressed sequence tags (ESTs), if available, from the same or related species; and (iii) alignments of the translated DNA sequence to known proteins.

The abundance of the different types of evidence depends on the organism, but for less well-studied species cDNA and ESTs evidences are often missing¹⁹.

Depending on the available type of data, the computational gene structural annotation is usually carried out by using an extrinsic and/or intrinsic approach. The extrinsic approach is homology-based, meaning that the genome annotation occurs using information coming from proteins of related species, or species-specific data, like ESTs and RNA-seq data. The intrinsic approach refers to *ab initio* gene prediction, which recognizes coding sequences using Hidden Markov Model (HMM) profiles and other functional elements based on intrinsic properties of the genome. The latter approach is generally used for *de novo* genome annotation, while the combination of extrinsic and intrinsic approach can solve

borderline cases, where the previous available information is partial and too incomplete to totally rely on it.

Extrinsic approach

After the repeat masking step, most of the genome annotation pipelines perform protein, ESTs and RNA-seq data alignment to the genome assembly. These sequences generally include previously identified transcripts or ESTs from the organism whose genome is being annotated and/or sequences from other organisms; generally, these are restricted to proteins, as these retain substantial sequence similarity over much greater periods of evolutionary time than nucleotide sequences do¹¹.

UniProtKB SwissProt (<http://www.uniprot.org/>) is an excellent core resource for protein sequences. As SwissProt is restricted to few highly curated proteins, it is advisable to supplement this database with the proteomes of related, previously annotated genomes. Frequently, BLAST and BLAT are used to 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 identity or coverage. After filtering of the protein alignments, highly similar sequences identified by BLAST and BLAT are realigned to the target genome using more sophisticated tools in order to obtain greater precision at exon boundaries¹¹.

Indeed BLAST, although being rapid, doesn't perform spliced alignment, therefore the edges of its sequence alignments are only approximations of exon boundaries. For this reason splice-site-aware alignment algorithms, such as Splign²⁰, sim4cc²¹ and Exonerate²², are often used to realign matching and highly similar ESTs, mRNAs and proteins to the genomic input sequence. Although these programs take more time to run, they provide the annotation pipeline with much improved information about splice sites and exon boundaries.

The alignment of a reference species' gene models onto the genome in study could be also used to transfer the annotation from the first to the latter one, based on the hypothesis to have high levels of homology between the two organisms.

The transfer of annotation can be used between any closely related species, either to transfer annotations between successive versions of a draft genome, or also to annotate new strains or species.

To face with the diminishing annotation resources available for each new sequenced genome and the need for more rapid annotation of new sequences, various annotation transfer strategies have been developed.

They are based simply on the sequence homology between closely related species or even to the synteny information which could aid the transfer of gene coordinates. As an example, The *map2assembly* script bundled with MAKER²⁴ uses BLASTN and Exonerate to map transcripts from a reference genome onto the new genome and refine the alignment to reliably transfer the structural annotation from one genome to another. This approach has been used to map maize reference transcripts onto the genome and to re-annotate a 22 Mb region of the *Zea mays* (maize).

On the other hand, RATT²³ (Rapid Annotation Transfer Tool) transfers annotations from a high-quality reference to a new genome on the basis of conserved synteny. In RATT, positional data based on conserved synteny and similarity between a reference and query are used to infer orthology, and hence function, more accurately. Furthermore, as genes differ in their underlying sequence between strains, the program refines all genes features in a correction step, to take into account changes to start and stop codons, length or the presence of internal stop codons²³.

Of all forms of evidence that could be used to be aligned on the new genome, cDNAs and RNA-seq data provide copious evidence for better delimitation of exons, splice sites and alternatively spliced exons. However, these data could be difficult to use because of their large size and complexity¹¹.

Currently, RNA-seq reads are usually handled in two ways. They can be assembled *de novo* — that is, genome-independent approach — using tools such as ABySS²⁵, SOAPdenovo-Trans²⁶ and Trinity²⁷; 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 TopHat2²⁸,

GSNAP²⁹ followed by the assembly of alignments into transcripts using tools such as Cufflinks³¹ or Scripture³⁰.

Several annotation pipelines are now compatible with RNA-seq data: these include PASA³², which uses inchworm outputs, EVidence Modeler³³, and MAKER³⁴, which can operate directly from Cufflinks outputs or can use preassembled RNA-seq data. Another way to use RNA-seq data is to ‘drive’ *ab initio* predictors and will be discussed in next paragraph.

Intrinsic approach

When gene predictors first became available in the 1990s they revolutionized genome analyses because they provided a fast and easy means to identify genes in assembled DNA sequences¹¹.

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 but the genome itself to identify a gene. However, these tools have some practical limitations from an annotation perspective.

For instance, most gene predictors find the single most likely CDS and do not report UTRs or alternatively spliced transcripts. 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 pre-calculated parameter files that contain such information for a few classic genomes, such as *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, humans and mice¹¹.

However, unless the genome is very closely related to an organism for which pre-compiled 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 content¹¹.

Gene finding in smaller eukaryotes tends to be more accurate because of their smaller introns and greater gene density, and gene finders for *bacteria*, *archaea*

and viruses are very accurate, predicting >99% of protein-coding genes correctly for most genomes. All of these methods assume that the DNA sequence is (mostly) correct, and certain types of errors will lead to erroneous gene predictions. In particular, any sequencing error that introduces an in-frame stop codon is likely to result in a mistaken gene prediction, because *ab initio* methods organize their searches around open reading frames.

Given enough training data, the gene-level sensitivity of *ab initio* tools could approach 100%; however, the accuracy of the predicted intron–exon structures is usually much lower, ~60–70%. It is also important to understand that the use of large numbers of pre-existing, high-quality gene models and near base-perfect genome assemblies is preferable to produce highly accurate gene predictions; but such data sets are rarely available for newly sequenced genomes¹¹.

In case of the absence of pre-existing reference gene models, the alignments of ESTs, RNA-seq and protein sequences to a genome can be used to train gene predictors. This process is often referred to as ‘evidence-driven’ 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, RNA-seq or protein data can be used to identify exon boundaries unambiguously: first they must be aligned to the genome, splice sites must be identified, and the assembled evidence must be post-processed before a synopsis of these data can be passed to the gene finder. As an example, the MAKER pipeline³⁴ provides a simplified process for training the predictors AUGUSTUS and SNAP using the EST, protein and mRNA-seq alignments that MAKER has produced.

In practice, this work requires a lot of specialized software and it is one of the main obstacles that genome annotation pipelines attempt to overcome¹¹.

Software	Description
<i>Ab initio and evidence-drivable gene predictors</i>	
Augustus	Accepts expressed sequence tag (EST)-based and protein-based evidence hints. Highly accurate
mGene	Support vector machine (SVM)-based discriminative gene predictor. Directly predicts 5' and 3' untranslated regions (UTRs) and poly(A) sites
SNAP	Accepts EST and protein-based evidence hints. Easily trained
FGENESH	Training files are constructed by SoftBerry and supplied to users
Geneid	First published in 1992 and revised in 2000. Accepts external hints from EST and protein-based evidence
Genemark	A self-training gene finder
Twinscan	Extension of the popular Genscan algorithm that can use homology between two genomes to guide gene prediction
GAZE	Highly configurable gene predictor
GenomeScan	Extension of the popular Genscan algorithm that can use BLASTX searches to guide gene prediction
Conrad	Discriminative gene predictor that uses conditional random fields (CRFs)
Contrast	Discriminative gene predictor that uses both SVMs and CRFs
CRAIG	Discriminative gene predictor that uses CRFs
Gnomon	Hidden Markov model (HMM) tool based on Genscan that uses EST and protein alignments to guide gene prediction
GeneSeqer	A tool for identifying potential exon–intron structure in precursor mRNAs (pre-mRNAs) by splice site prediction and spliced alignment

Figure 1. Most popular *ab initio* and evidence-drivable gene predictors.

In recent years various tools able to manage evidence-driven prediction have been developed (Figure 1). AUGUSTUS⁵¹ and SNAP⁵⁰ (Semi-HMM-based Nucleic Acid Parser) gene prediction tools are based on generalized Hidden Markov Models (gHMMs). AUGUSTUS also integrates together an accurate method for modeling the intron length distribution and includes a training procedure to first create the parameters for the species and a windowed weight array matrix (WWAM). It is able to perform also an optimization step to increase the prediction accuracy by a few percent points⁵¹.

SNAP is one of the simplest and most lightweight *ab initio* predictors currently available and it is also provided with a training module that makes it easily adaptable to different organisms so that the gHMM parameters are adjusted in a species-specific manner.

Twinscan⁸⁶ is also a system based on gHMMs for predicting gene-structure in eukaryotic genomic sequences and combines the information from predicted coding regions and splice sites with conservation measurements between the

target sequence and sequences from a closely related genome. Twinscan is bundled with a training utility, which makes use of BLAST and BLAT alignments to derive conservation sequences.

GeneMark is a family of gene prediction programs as the eukaryotic gene predictor GeneMark-ES⁵³ and GeneMark-ET⁵², the semi-supervised version of GeneMark-ES. GeneMark-ES⁵³ is an *ab initio* gene finding tool based on HMMs which performs unsupervised self-training. It doesn't need curated training sets but the genomic sequence to be trained. GeneMark-ET instead uses mapped RNA-seq reads or transcripts evidence to improve training. GeneMark-ET has an additional parameter "--fungus" specifically designed for fungal genomes. This design enables the algorithm to work equally well for species with the kinds of variations in splicing mechanisms seen in the fungal phyla *Ascomycota*, *Basidiomycota*, and *Zygomycota*⁵⁴. In this case the HMM model underlying the *ab initio* algorithm takes also into account for introns possessing conserved BP sites, as there is existing evidence of a significant role of the fungal BP sites in splicing⁶. The model consists in an enhanced intron sub-model that accommodates intron sequences with and without BP sites.

Geneid⁵⁵ is one of the first *de novo* predictors that became available (1992), and differently from the previous ones, it is based on the recognition of signals on a genomic sequence using Position Weight Arrays (PWAs). The program is distributed with an official training guide that permits an accurate modeling of coding and non-coding regions, as well as CDS and splice sites, starting from a training set of gene models.

In short, the program first calculates each possible k-mer probability to occur in the CDS region compared with a non-coding region (the intron). A different procedure is performed for the Start, Acceptor and Donor signals. The genomic sequence is scanned for instances of canonical sites (i.e., 'ATG', 'AG', 'GT') and the program will try to determine whether the surrounding sequence is more likely to be found in the presence of a real signal rather than being a random occurrence. Geneid determines this by comparing the surrounding region with the available PWAs, which is calculated during the training.

The PWM is calculated by first comparing the sequences of real signals with all the sequences of false signals and then is used to create an initial version (un-optimized) of geneid parameter file. Geneid has included, as for AUGUSTUS, an optimization step. This step is performed by taking the training sequences and tuning particular values called the exon weight factor “ewf” and the oligo weight factor “owf” inside the template, iteratively. At the end geneid predictions done on the training data are compared with the training annotations themselves in order to finally choose the best combination of weight factors to optimize prediction accuracy.

Automated genome annotation and quality assessment

Manual annotation is an expensive and time-consuming process. To make this process faster, several automated annotation pipelines have been developed with the purpose to integrate existing software tools into one package that produces database-ready genome annotations in a relatively short time.

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' 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 locus¹¹. This is the process used for example by Jigsaw³⁵, EVidenceModeler³³ and GLEAN³⁶.

Other popular and much refined 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. This is the process used by PASA³², Gnomon³⁷ and MAKER³⁴.

Classification and prioritization of annotations for later manual review is a crucial step of the genome annotation process. A classification scheme requires that each annotation be tagged with information describing the type of evidence that supports each gene model. The pipeline of MAKER⁴ in particular does so together with the integration of associated evidence and quality control statistics. It's able to identify and mask repetitive elements in the genome, to align ESTs and protein

evidence, and to produce *ab initio* gene predictions inferring five and three prime UTRs (Figure 2).

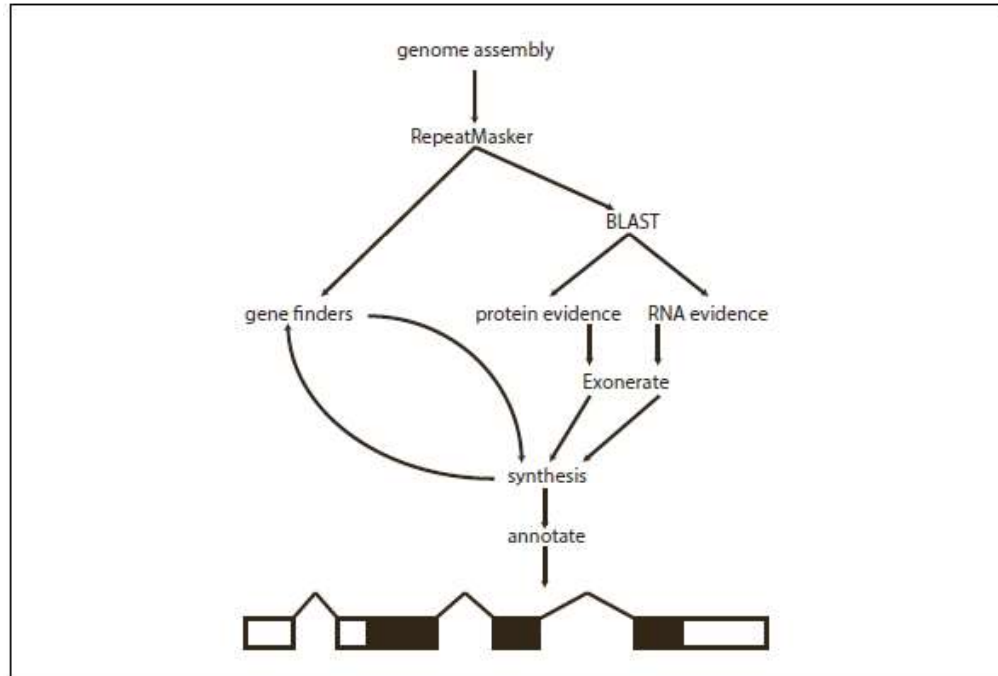


Figure 2. The MAKER annotation pipeline workflow.

MAKER uses evidence alignments to provide ‘hints’ to the prediction programs as to the location of probable introns, exons, and coding regions. It actively modifies the resulting predictions to include features like UTRs that can be inferred from EST or mRNA alignments. In this way, it guides the behavior of *ab initio* prediction programs using experimental evidence to produce improved models. MAKER then takes the entire pool of *ab initio* and evidence informed gene predictions, updates features such as 5' and 3' UTRs based on EST evidence, tries to determine alternative splice forms where EST data permits, produces quality control metrics for each gene model (these are included in the output), and then ‘chooses’ from among all the gene model possibilities the one that best matches the evidence. This is done using a modified sensitivity/specificity distance metric³⁴. At the end the pipeline integrates all the evidences to produce final gene annotations with quality control statistics that help prioritize genes for downstream review and manual curation²⁴.

The assessment of the quality of an annotation is fundamental in an automated genome annotation process. The gene predictions have to be possibly checked to avoid the creation of artifacts. Indeed, even the best gene predictors and genome annotation pipelines rarely exceed accuracies of 80% at the exon level¹¹, meaning that most gene annotations contain at least one mis-annotated exon.

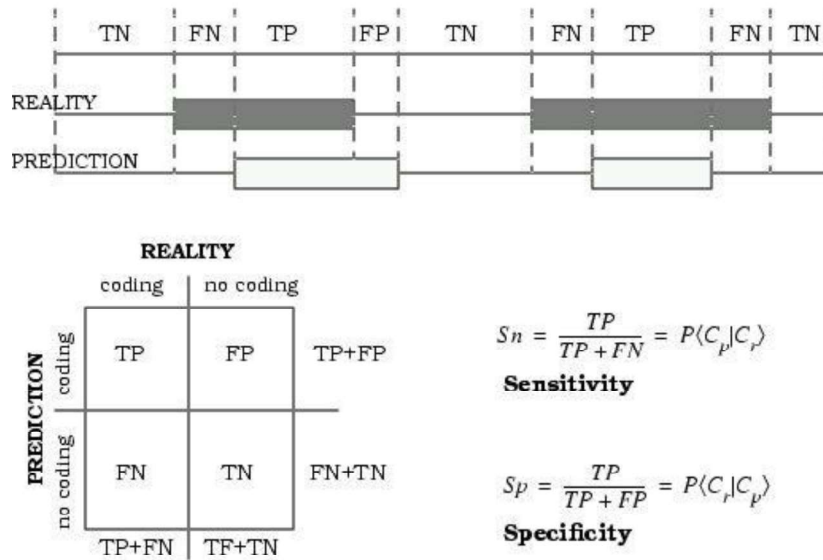


Figure 3. Statistical metrics used to evaluate the best matches between evidence and prediction.

The commonly used metrics for measuring the accuracy of gene prediction are sensitivity and specificity (Figure 3). Sensitivity (SN) is the fraction of the test set 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, $SP = TN / (TN + FP)$, where FP is false positives¹¹. 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 prediction⁴⁹.

Eval⁴⁹ is a flexible tool for analyzing the performance of gene-structure prediction programs as it provides summaries and graphical distributions for many statistics describing any set of annotations, regardless of their source. It compares sets of predictions to standard annotations and to one another using standard statistical measures as previously described; it calculates sensitivity and specificity for any portion of a gene model, such as genes, transcripts or exons each time relative to a reference annotation.

In addition to the quality control of gene prediction also the quality of the final gene models should be checked. In this sense, MAKER has introduced a measure of the congruence between a gene annotation and its supporting evidence, called the Annotation Edit Distance (AED)³⁴. AED is calculated as $AED = 1 - AC$ where $AC = (SN + SP) / 2$, where SN and SP are calculated respect to the union of the aligned evidences (Figure 4).

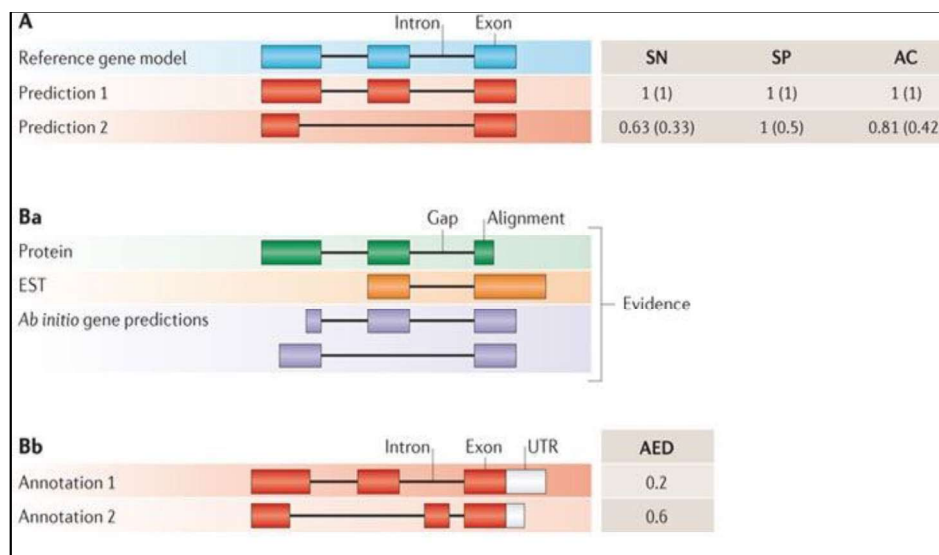


Figure 4. Example of calculation of AED measure by MAKER.

An AED of 0 indicates the perfect agreement between the annotation and its supporting evidence, whereas an AED of 1 indicates a complete lack of evidence support for the annotation¹¹.

MAKER also generates the Quality Index (QI) metric for each annotated gene model. Quality index is a nine-dimensional summary of a transcript's key features and how they are supported by the data gathered by MAKER's compute pipeline⁵⁷.

A typical QI might look as follows: “QI:0|0.77|0.68|1|0.77|0.78|19|462|824”. Table 2 provides a key for the QI data fields.

Position	Definition
1	Length of the 5' UTR
2	Fraction of splice sites confirmed by an EST/mRNA-seq alignment
3	Fraction of exons that match an EST/mRNA-seq alignment
4	Fraction of exons that overlap EST/mRNA-seq or protein alignments
5	Fraction of splice sites confirmed by ab initio gene prediction
6	Fraction of exons that overlap an ab initio gene prediction
7	Number of exons in the mRNA
8	Length of the 3' UTR
9	Length of the protein sequence produced by the mRNA

Table 2. MAKER quality index summary (adapted from Cantarel et al., 2008).

Over the years, there have been various contests aimed at assessing gene annotation accuracy¹¹. 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.

Indeed no ‘quality control’ gene sets exists for most of the organisms being sequenced today. One of the most used way to assess the quality of obtained annotation is to check the homology with other sequence present in public databases or to check if the CEGs are represented in the gene space. Moreover, just because a gene predictor does well on one genome is no guarantee of a good performance on the next³⁴.

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.

AIM OF THESIS

The purpose of this thesis was to meet the need to produce fast and accurate annotations on complex genomes such as the eukaryotic ones. Working on different experimental cases, different annotation strategies have been implemented tailored to each specific case, by integrating available software tools and evidence with quality control statistics.

The analyses have been conducted on three different cases:

- The annotation of a genome with closely related and well-characterized reference,
- The annotation of a genome with a close but phylogenetically distinct reference,
- The annotation of a genome with no reported reference.

The contribution of the different methods used was analyzed along with the evaluation of results obtained, also by the comparison with published data, focusing particularly on protein coding genes. The relevance of available data and computational means in the process of genome annotation will be discussed as well as the open issues in genome annotation process.

EXPERIMENTAL CASES

Annotation of a genome with a closely related reference

Background

The genus *Fusarium* represents the most important group of fungal plant pathogens, causing various diseases on nearly every economically important plant species. Members of the *Fusarium oxysporum* species complex exhibit extraordinary genetic plasticity and cause some of the most destructive and intractable diseases across a diverse spectrum of hosts, including many economically important crops, such as bananas, cotton, canola, melons, and tomatoes³⁸.

Fusarium comparative genomics has revealed that horizontal chromosome transfer introduces host-specific pathogenicity among members of this species complex and is responsible for the broad host range and the strong host specificity revealed by the members within the *F. oxysporum* species complex as well as for some of the most destructive and intractable plant diseases³⁸.

F. oxysporum f. sp. *melonis* (FOM) is a fungal pathogen that causes *Fusarium* wilt disease on melon (*Cucumis melo*). Risser et al. (1976) divided FOM into four races (races 0, 1, 2 and 1,2) based on the reaction to inoculation of three melon different cultivars. Recent studies show that *F. oxysporum* sp. *melonis* and the reference *F. oxysporum* sp. *lycopersici* (FOL) are closely related at evolutionary level at the point of creating a separated *clade* on a *Fusarium* strains phylogenetic tree³⁹ (Figure 4).

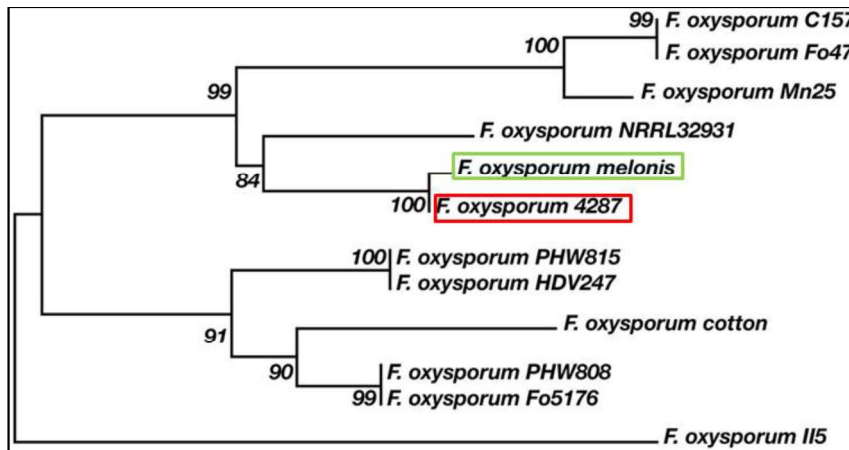


Figure 4. Phylogenetic tree of *Fusarium* genus by Kwiatos et al. Front. Microbiol. 2015.

An important finding is that the genus *Fusarium* has a core genome that is shared among all strains and covers approximately 60% of the entire genomic sequence. The *Fusarium* genomes consist of a core region with approximately 9,000 genes considered to be orthologous due to high sequence similarity and conserved gene order⁴⁰.

On the other hand, the unique sequences, designated as lineage-specific (LS) regions, are a substantial fraction (40%) of the genome assembly. The LS regions include four entire chromosomes (chromosomes 3, 6, 14 and 15) and small parts of chromosome 1 and 2⁴¹.

These regions contain more than 74% of the identifiable transposable elements (TEs) in the FOL genome, including 95% of all DNA transposons and about 28% of the entire FOL reference genome was identified as repetitive including many retro-elements, LINEs and SINEs and DNA transposons as well as several large segmental duplications⁴¹.

Materials

The genome of isolate FOM1018 has been previously sequenced with Illumina GAIIX and Illumina HiSeq 1000 (Illumina Inc, San Diego, CA) generating one standard 100 paired-end and three mate-pair libraries with different fragment sizes (5, 8, 10 kb) resulting in an average of 140X coverage. Scaffolding has been done using mate-pair reads and genome assembly performed with SOAPdenovo2⁴²

obtaining a 52.9 Mb genome assembly divided in 4,658 scaffolds with N50 length of 3.59 Mb.

Due to the inability to resolve part of repeated sequences, the FOM1018 genome is ~8 Mb shorter than the published reference for the species *Fusarium oxysporum* sp. *lycopersici* (FOL) which has a genome of ~61 Mb and 15 chromosomes⁴¹.

Comparison of FOM1018 genome with that of the FOL 4287 reference sequence (assembly ASM14995v2) using MUMmer⁴³ showed that the core genome's chromosomes are mostly co-linear and syntenic between the two *formae specialis*. As an example, Figure 5 shows an example of the perfect 1-to-1 correspondence between FOM1018 scaffold 4655 and *F. oxysporum* sp. *lycopersici* chromosome 8.

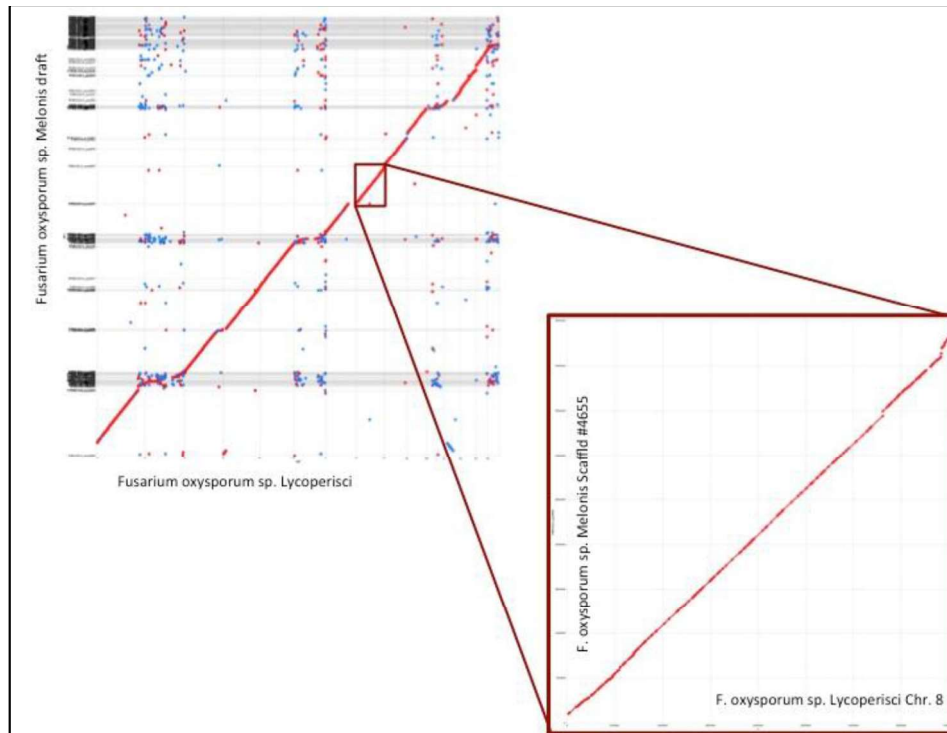


Figure 5. MUMmer plot of the alignment of FOM1018 assembly on FOL chromosomes.

On the other hand, many scaffolds were short and highly-fragmented (rows reach in blue and red dots in Figure 5). Based on the MUMmer alignment of the genomes, the average percent identity of the nucleotides was ~ 95%. Among a total of 4,658 scaffolds, 4,095 had scaffold length between 200bp and 1kb.

Approximately half of the shorter scaffolds (1,860), amounting to ~ 0.94 Mb (1.8% of FOM1018 genome), didn't align onto FOL reference genome.

Methods

The procedure adopted for the generation of the final annotation was the following:

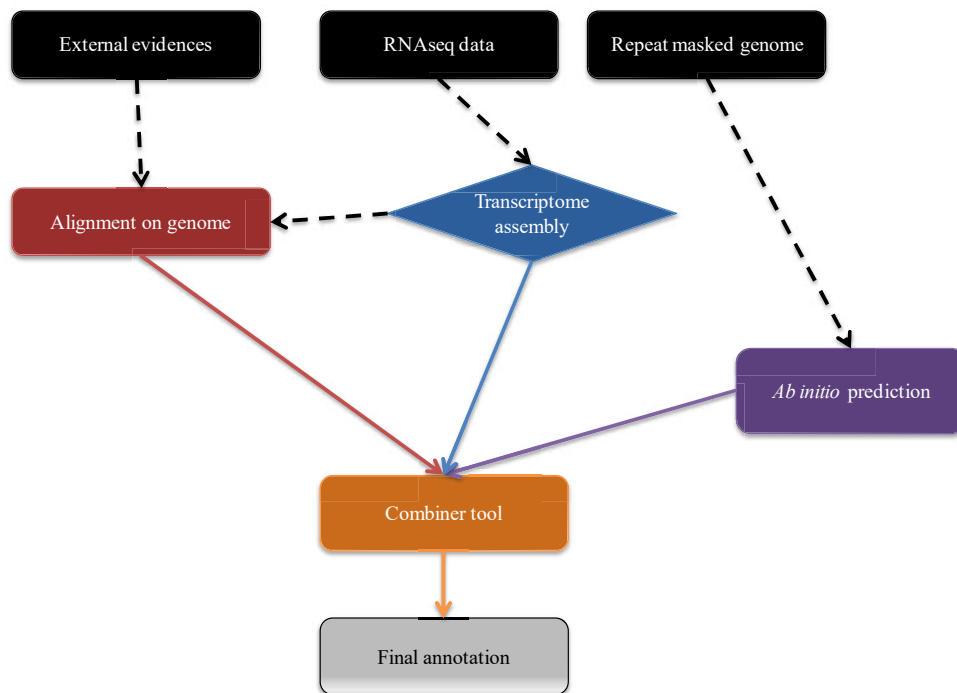


Figure 6. Diagram of procedure adopted for FOM1018 genome annotation.

- **Repeat masking of the genome and detection of putative non-coding RNAs**

The repeat masking of the genome has been performed using REPET software ver. 2.2⁴⁴. Simultaneously with the masking of the genome, REPET was fed also with the set of repeats (both nucleotides and aminoacidic sequences) of RepBase database (ver. 18.08)⁴⁵. During the repeat detection phase REPET also checks for the presence of potential host genes (potential FOM genes). These bases were not masked in the genome for being potentially related to protein coding genes.

In order to detect non-coding RNA sequences in the genome, sequence similarities were inferred using the reference-based method Infernal⁴⁶ ver. 1.1 using the covariance model of Rfam database⁴⁷ of ncRNA families release 12.0.

In Rfam database each RNA family is represented by a multiple sequence alignment (called a ‘seed alignment’) and a Covariance Model (CM) built from that alignment. Infernal builds a profile from a structurally annotated multiple sequence alignment of an RNA family with a position-specific scoring system for substitutions, insertions, and deletions. Positions in the profile that are base-paired in the consensus secondary structure of the alignment are modeled as dependent on one another, allowing Infernal’s scoring system to consider the secondary structure, in addition to the primary sequence, of the family being modeled⁴⁶.

The output of Infernal search was then converted to Generic Feature Format 3 (GFF3) file using a custom script. GFF files are plain text, 9 column, tab-delimited files which are frequently used as standard schema to for data exchange and representation of genomic data.

The repeat library was then fed to MAKER ver. 2.31.7, which first run RepeatMasker to identify all classes of repeats that match entries both in the RepBase repeat library and in the REPET library. Next MAKER uses RepeatRunner to identify transposable elements and viral proteins using the RepeatRunner protein database.

Complex repeats are hard-masked to remove this sequences from any further consideration at any later point of the annotation process, whereas simple repeats are soft-masked to prevent alignment programs such as BLAST from seedling any new alignments in the soft-masked region.

Anyway alignments that begin in a nearby (non-masked) region of the genome may extend into the soft-masked region since low-complexity regions are found within many real genes.

- **Transfer of FOL cDNAs on FOM genome**

The information obtained with the synteny analysis confirmed the strong similarity between the two genomes of FOM1018 and FOL, so it was decided to transfer the gene annotation of FOL to FOM1018. To transfer FOL cDNAs to

FOM1018 genome was used the *map2assembly* script bundled with MAKER²⁴. The script performs a BLASTN search onto the genome and only alignments of cDNA nucleotide sequences with default values 70% of identity, 70% coverage, e-value of 1E-10 and bit score greater than 40 are retained. The script then refines the alignment using Exonerate with 20% maximal score threshold, cleans the hits and clusters them into transcripts for annotations choosing the best ones. This step produced a standardized GFF3 annotation file of gene models for FOM1018.

- **Training dataset selection and *ab initio* prediction**

The transferred gene models were re-aligned on FOM genome and those gene models having 100% identity and coverage based on GMAP⁴⁸ alignment were selected. The resulting gene models were filtered for having a predicted protein starting with a methionine and for the presence full ORF and canonical splice sites using the VALIDATE_GTF.pl script of software Eval⁴⁹.

VALIDATE_GTF.pl is a flexible Perl script that checks if the annotation file of a gene model contains errors. It can detect most common syntactic errors, such as including the stop codon within the CDS annotation. It can also detect semantic errors, such as annotated coding sequence that contains stop codons spanning splice sites. The gene models passing this filters were selected and used as dataset for the training of *ab initio* predictors.

To improve the sensitivity and specificity of the annotation, multiple gene predictors were used, in particular gHMMs-based predictors SNAP and AUGUSTUS, GeneMark-ET and geneid.

For all predictors but GeneMark, the training procedure was undertaken by randomly selecting 90% of dataset for the training and the remaining 10% as test-set; this procedure was repeated ten times to perform 10-fold cross-validation. Eval was used at the end of each run to compare the predictions with the test-set and evaluating the performance of *ab initio* predictors.

The parameter `-flank` passed to the “fathom” program included in SNAP was left to the default value of 1000, to indicate that 1000 flanking bases for each gene model will be considered for training and testing.

GeneMark-ET was trained by using as input the file of intron coordinates of the high quality gene models and the soft masked genome FASTA file. The training was performed using with the flag ‘—fungus’ specifically designed for fungal genomes.

SNAP, AUGUSTUS and GeneMark prediction was run ‘inside’ MAKER whereas Geneid prediction was run ‘outside’ MAKER, since it doesn’t support it, and the file of geneid predictions was given in input to MAKER in order to be used together with other evidences to produce final consensus gene annotation.

- **External evidences selection and alignment**

In order to produce an evidence support to *ab initio* prediction, a database of 135,770 proteins comprising Swissprot Fungi curated database and proteins of other 7 fungi as external evidences to align onto FOM1018 genome was provided to MAKER (see Results). The choice of such organisms was inspired by the paper of Ma et al. (2010)⁴¹, which outlines the phylogenetic relationships of *Fusarium* species in relation to other ascomycete fungi. Only protein alignments with default minimum of 70% coverage, 50% identity, e-value of 1E-06 were retained. Exonerate similarity score (ESS) was instead set as $\geq 70\%$ ⁵⁶.

- **Final annotation generation and quality assessment**

To produce the final annotation, MAKER was supplemented with the masked genome assembly and evidence alignments, the ncRNAs annotation, the SNAP and GeneMark-ET HMMs, the FOL transferred gene models, the AUGUSTUS model for the species and the geneid predictions.

MAKER parameter for extending evidence clusters to gene predictors was set as 200 bases, as gene finders require flanking sequence on either side of a gene to correctly find start and stop locations. This option also affects how close evidence islands must be before clustering together, this means that increasing this value can capture exons missing from the evidence; while decreasing this value can help decrease gene mergers in organisms with high gene density. The parameter has been set to a maximum of 200 in relation to the short intergenic distances in fungi.

The maximum length for the splitting of hits was set to 100 (that is the expected maximum intron size for evidence alignments), because of the similar value for intergenic distance and intron length in the 7 Fungi proteomes given as external evidence support. Also the identification of monoexonic genes was allowed.

To assess the quality of final annotation the MAKER AED and QI measures were used. Another approach that was followed was to search similarity against Pfam database⁵⁸ with Hmmer ver. 3.1b1⁵⁹ with an e-value cut-off of 1E-07 and to annotate GO-terms, Enzyme codes and relative pathways using Blast2GO suite⁶⁰ with default parameters. Apollo⁶¹ ver. 2.0 was used to examine the correctness and perform a visual inspection of the annotations.

Results and discussion

Due to the high fragmentation of part of the assembly, REPET was able to identify only 0.54 Mb of repeats (1% of total repeat content) in regions corresponding to *Fusarium* non-syntenic chromosomes, in contrast to the identified 12.91 Mb (20% of repetitive content) on the published FOL reference genome⁴¹ (Table 3). Indeed it was expected that the presence of such high fragmentation could have caused problems for repeat annotation and loss of information.

The fraction of FOM1018 repeats corresponding to the core genome⁴¹ was instead comparable with the one of the reference FOL, that is 3.99 Mb (7.5 % of total repeat content) compared to the 4.30 Mb of FOL (7% of total repeat content).

The biggest fraction of repeats belongs to DNA transposons, comprising a good quantity of TIR elements (2.1 Mb) but also to retrotransposons such as LTRs (2.6 Mb) and LINEs (0.35 Mb), in line with results obtained with FOL⁴¹ (Table 3).

	FOL	FOM1018
assembly length (Mb)	61.47	52.93
Repeats in core regions (Mb)	4.30 (7%)	3.99 (7.5 %)
Repeats in LS-regions (Mb)	12.91 (21%)	0.54 (1 %)
total masked (Mb) (%)	17.21 (28%)	4.53 (8,5%)

Table 3. Summary statistics of repetitive content of FOM1018 genome compared with FOL.

Non-coding RNAs annotation was done using the Infernal homology-based approach since no small RNA-seq data sets were available to eventually provide evidence to support *de novo* ncRNA annotations.

The analysis resulted in the identification of 314 tRNAs in FOM1018, slightly higher but in line with the number of tRNAs identified in FOL (Table 4). Also the number of small non coding RNAs is similar whereas in FOM1018 have been also identified 77 5S_rRNAs, which have not been previously identified in FOL. Based on the literature, 61 rRNA were identified also in *Fusarium oxysporum sp. melonis* isolate NRRL 26404³⁸ and this data confirmed the results of FOM1018.

	FOL	FOM1018
tRNA	268	314
5S_rRNA	-	77
snRNA	27	25

Table 4. Results of the analysis in Rfam database and comparison with FOL.

The soft masked genome of FOM1018 was used to transfer FOL gene models in order not to annotate genes in repetitive regions.

The *map2assembly* script bundled with MAKER was able to reliably transfer a total of 14,469 gene models upon 17,977 FOL cDNAs onto FOM genome (Table 5).

Total FOL cDNAs	19,777
Transferred on FOM1018 with 1 match	13,598
Transferred on FOM1018 with > 1 match	527
Non-transferred genes	3,852

Table 5. Statistics of FOL transferred gene models on FOM1018 genome.

A percentage of ~ 35% of transferred gene models had an AED = 0.00 that means that such gene models are identical on FOM1018 respect to FOL.

Among the genes reliably transferred, 527 had more than one *locus* on FOM1018 genome. This could be due to the presence of fragmentation and redundancy of some regions. Some real proteins contain low-complexity regions and if the program is left to align to a low-complexity region, spurious alignments would be produced.

Even if given a soft-masked genome, BLAST sometimes could allow ‘seedling’ to extend through low-complexity regions. This could have produced some spurious alignments but it is just a small proportion of overall alignments and the majority of gene models are of a good quality.

Based on results obtained with the transfer of annotation, a more reliable dataset for the training of *ab initio* predictors was produced. The 13,598 transcripts representing uniquely transferred gene models, obtained with MAKER, were re-aligned on FOM genome using GMAP.

A total of 11,462 gene models mapped with GMAP 100% coverage and identity back to FOM1018 genome. After filtering the quality of the transcripts with a custom script (see Methods) a final high quality dataset of 4,573 gene models was obtained. A total of 6,889 models didn’t pass the filters due mostly to problems in the annotation file of the gene model but also the FASTA file for predicted protein encoded (see EVALUATE_GTF.pl specifications in Introduction). *Ab initio* predictors parameter files were obtained with the 10-fold cross-validation using the high quality dataset of 4,573 gene models and their performance was evaluated (Table 6).

	AUGUSTUS	SNAP	Geneid
Gene Sensitivity	49.05%	38.69%	41.79%
Gene Specificity	56.01%	31.26%	40.66%
Exon Sensitivity	62.57%	53.39%	60.11%
Exon Specificity	71.34%	50.82%	60.26%
Nucleotide Sensitivity	93.51%	95.28%	97.18%
Nucleotide Specificity	100.00%	88.70%	90.53%

Table 6. Average sensitivity and specificity measures obtained with the training of gene predictors.

Both AUGUSTUS and geneid show good performance in terms of nucleotide and exon sensitivity and specificity. Gene sensitivity has in general lower values but in line with those obtained in other publications³⁴.

The proteins and of 7 fungi and SwissProt fungi were aligned on FOM1018 genome using MAKER⁴ (Table 7). Among all these external evidences, about ~ 60% aligned onto the FOM1018 genome with the parameters described in Methods.

Source	# proteins	#aligned on FOM1018
SWISSPROT	31,315	9,446
FOL 4827	17,696	15,224
FOM 26406	26,719	24,132
<i>F.Verticilloides</i>	14,185	12,261
<i>F.Graminearum</i>	13,313	9,797
<i>P.tritici-repentis</i>	12,169	3,086
<i>A.Nidulans</i>	10,534	3,368
<i>N.Crassa</i>	9,839	3,895
TOTAL	135,770	81,209

Table 7. External evidences aligned on FOM1018 genome.

After generation of all source of evidence, MAKER was run to produce the final annotation, which contained 18,689 gene models and the same number of

transcripts. Compared with the number of FOL genes, the number of FOM1018 gene models is slightly higher (Table 8). The main structural parameters such as the mean intergenic distance and the mean gene length are generally in line with those of FOL, even if the median intergenic distance is a bit lower. This could be due to the higher number of genes in a smaller assembly respect to the reference FOL or to potential gene fusions in the dataset that could be checked with the functional annotation and the manual curation; although the mean gene length is 1,323 bp and this fact could possibly exclude the second hypothesis. Mean exon and intron length are similar but the proportion of monoexonic genes is 10% higher. This could be a case of multi-exonic gene not splitted by the gene predictor and retained as monoexonic due to the absence of an evidence support such as a fungal protein or ESTs or RNA-seq data or maybe FOM1018 could have an higher proportion of monoexonic genes respect to its reference. Among 6,428 monoexonic genes, 2,229 (~ 35%) have an AED=1, which means no evidence support. This also should be investigated in manual curation and functional annotation phase. Among all gene models, 1,803 have at least one annotated UTR.

STRUCTURAL FEATURES	FOL	FOM1018
Number of genes:	17,696	18,689
Mean intergenic distance:	-201	-108.50
Median intergenic distance:	1,054	802
Mean gene length:	1,345	1,323.54
Mean Exon length:	497	487.15
Mean intron length:	100	116.85
Mean number of exons:	2.70	2.71
Monoexonic:	4,392	6,428
% monoexonic:	24.81%	34.40%
Monoexonic mean length:	1,094	854.98

Table 8. Summary structural statistics of FOM1018 final annotation and comparison with FOL.

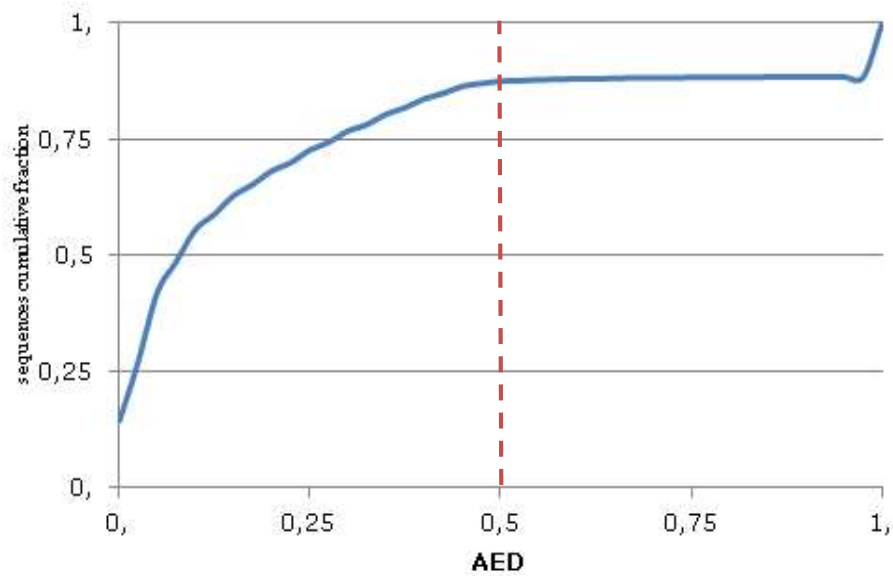


Figure 7. AED graph of FOM1018 gene models.

This final dataset of 18,689 then, comprises both low- and high-quality gene models, which can be first - but not absolutely - distinguished using the AED quality measures provided by MAKER itself.

More than 85% of gene models have an AED ≤ 0.5 (Figure 7) which is a standard cut-off to evaluate the quality of the gene model³⁴. Moreover, more than 15,900 gene models (85% of total) result to be supported by any type of evidence (AED < 1), meaning that almost every gene model with an AED < 1 has an AED ≤ 0.5 (Table 9).

Source	ALL	AED <1
AUGUSTUS	1,524	1,236
SNAP	3,943	3,266
GeneMark-ET	9,159	9,159
Geneid	2,929	1,420
FOL gene models	1,134	849
TOTAL	18,689	15,930

Table 9. Statistics of gene models and AED values in MAKER annotation.

Analysis of MAKER QI indicates that 13,416 gene models are fully overlapping an external evidence and 16,090 gene models overlap completely an *ab initio* gene prediction.

Indeed most of genes are supported by all type of evidence, such as predictions, external evidences and FOL genes. In Figure 8, 13,312 (71.2%) genes are supported either by transcript evidence and/or protein evidence, while 13,029 (69.7%) genes are supported by all three kinds of evidence. Most of genes, 18,406 (98.5%) are supported by *de novo* predictions and the 85.5% of them (15,973) are supported by FOL genes and protein evidence.

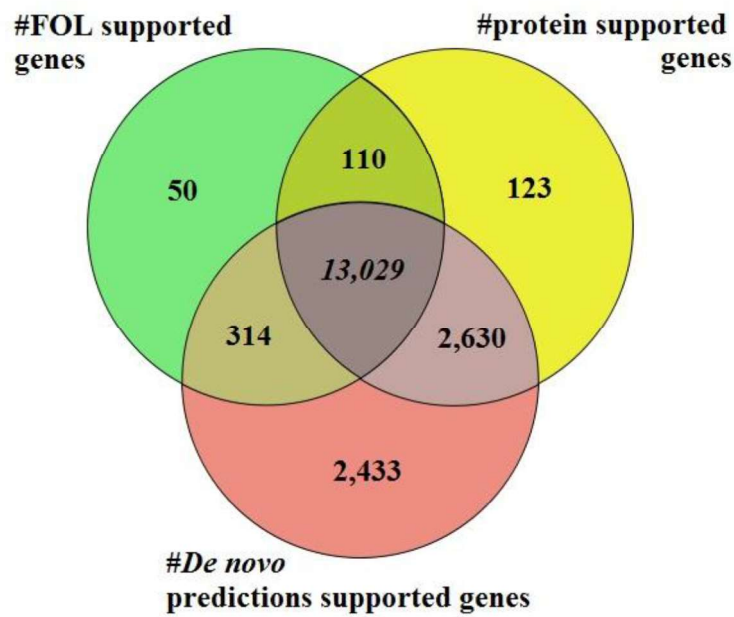


Figure 8. Venn diagram for sources of evidence for FOM1018 gene models. The different colors indicate various sources of evidence, and the numbers are the number of gene models supported by each kinds of evidence.

A good fraction of genes (18,168 - 97% of total) has at least one hit in NCBI non-redundant database, and 10,484 genes have at least one InterPro domain annotated.

Among all gene models, 2,907 have an AED=1 – and among them 2,229 are monoexonic, as said before – meaning that MAKER annotation has no experimental support for them. These low quality gene models may be artifacts or orphan genes and need to be further analyzed in detail; to get a first idea the BLAST results were checked, observing that 2,564 genes (83.5% of genes with AED=1) have at least one hit in NCBI non-redundant database even if the majority (2,205) are annotated as “hypothetical protein”.

More in-depth analysis to evaluate the ‘goodness’ of these genes are still required even if, based on this results, the quality of the overall annotation can be considered as good.

Annotation of a genome with a close but phylogenetically distinct reference

Background

The genus *Solanum* ranks among the largest of plant *genera* and includes several cultivated crops of regional or worldwide significance including potato (*Solanum tuberosum*) which is the most important non-cereal food crop worldwide. Abiotic stress factors such as cold, heat, drought, and salinity have a significant effect on cultivated potato, affecting yield, tuber quality, and market value (Wang-Pruski and Schofield, 2012). To improve resistance to these adverse environmental factors, potato breeders can exploit the ~200 tuber-bearing *Solanum* species native to South, Central, and North America⁶³.

Solanum commersonii is a tuber-bearing wild potato species native to Central and South America. Analyses of chloroplast genome restriction sites and nitrate reductase gene sequence confirmed that *S. commersonii* is phylogenetically distinct from cultivated potato (Rodriguez and Spooner 2009) and this distinction is confirmed also by the phylogenetic analysis done by Fajardo and Spooner in 2011 when they divided 29 diploid *Solanum* species in 4 sister clades by using used conserved orthologous set (COSII) nuclear loci (Figure 9).

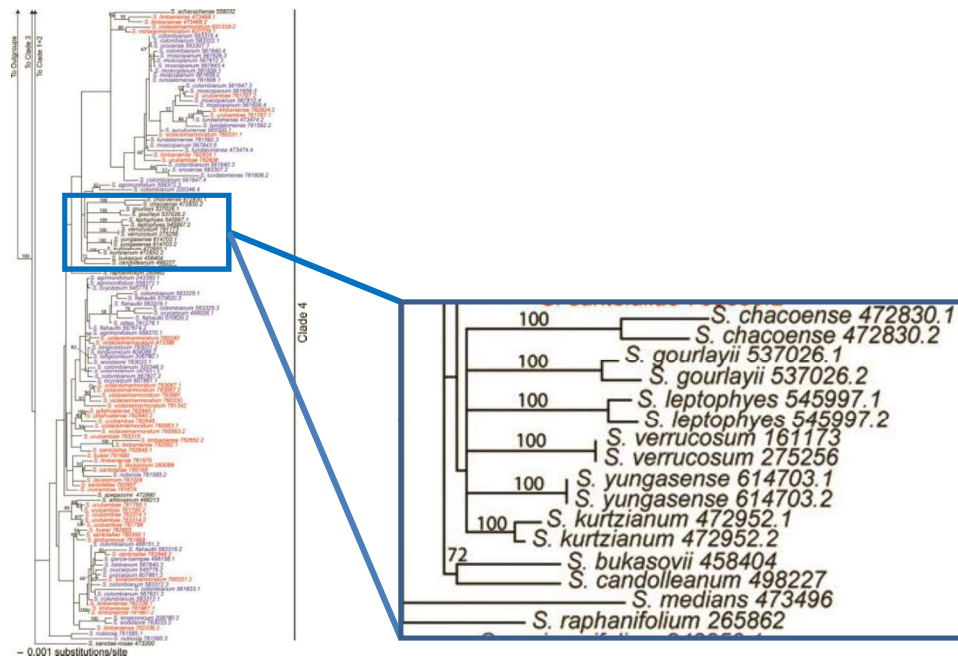


Figure 9. Extract of maximum likelihood phylogram of *Solanum* series Conicibaccata (Fajardo and Spooner, Systematic Botany 2011). *S. chacoense* alias *S. commersonii*.

Consistent with these analyses, *S. commersonii* and *S. tuberosum* are sexually incompatible and have been assigned different endosperm balance numbers (EBNs) (Johnston et al., 1980), with *S. commersonii* reported as 1 EBN and *S. tuberosum* reported as 4 EBN. Despite being genetically isolated from cultivated potato, *S. commersonii* has garnered significant research interest since it possesses resistance traits and particularly attractive is its freezing tolerance and capacity to cold acclimate (i.e., ability to increase cold tolerance after exposure to low, non-freezing temperatures). By contrast, the cultivated potato is classified as sensitive to low temperatures and is unable to cold acclimate (Palta and Simon, 1993).

Materials

The draft genome sequence of the wild potato species *S. commersonii* PI 243503 has been sequenced with Illumina HiSeq 1000 (Illumina Inc, San Diego, CA) from six 100 paired-end libraries with different fragment sizes (400, 450, 550 and 700 bp) and three mate-pair libraries with different fragment sizes (3,5,10 kb) obtaining roughly 105x coverage. The genome assembly was performed using the *S. tuberosum* genome sequence published in 2011 as a reference⁶³, obtaining a total assembly length of 830 Mb with an N50 scsffold length of 44,3 kb.

Preliminary analyses on genome assembly reported *S. commersonii* has a total of 9,894,571 reliable single-nucleotide polymorphisms (SNPs) among 662,040,919 reliable genome bases, yielding a SNP frequency of 1.49%.

The generation of RNA-seq dataset has been done starting from 100 paired-end sequencing libraries obtained from RNA of four different tissues, and sequencing with Illumina HiSeq1000 (Illumina Inc, San Diego, CA), obtaining a total of ~16,8 Gb (Table 10).

Tissue	Sequenced fragments (100x2)
flower	36,323,578
root	46,372,066
tuber	49,885,340
leaf	35,855,731

Table 10. Summary statistics of *S. commersonii* RNA-seq data.

Methods

The procedure adopted for the generation of the final annotation was the following:

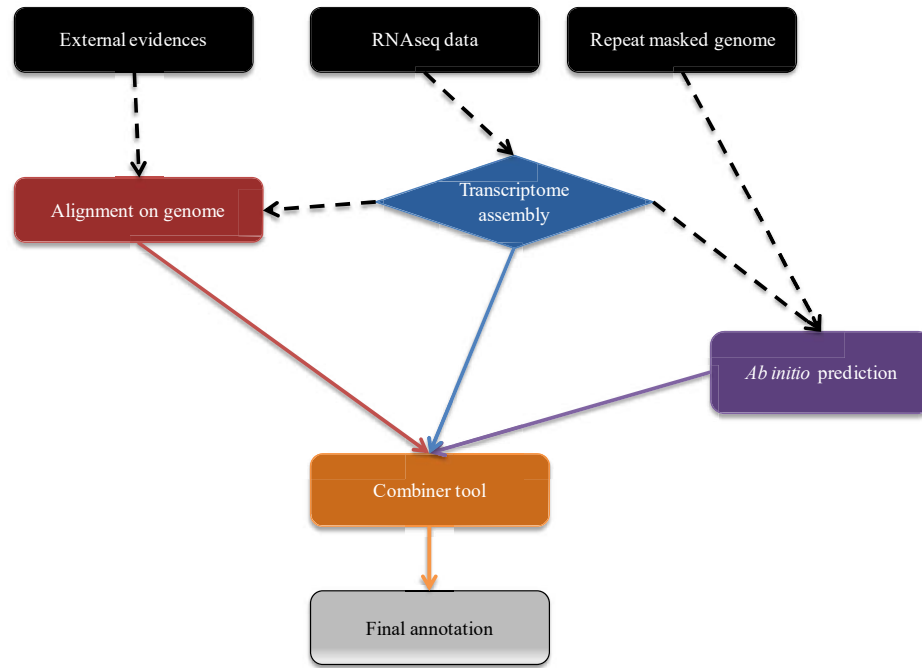


Figure 10. Diagram of strategy adopted for *S. commersonii* genome annotation.

- **Repeat masking of the genome and detection of putative non-coding RNAs**

Annotation of repeats and protein coding genes was performed using the MAKER³⁴ pipeline ver. 2.27. In particular, repeats were annotated and masked on *S. commersonii* genome assembly using RepeatMasker ver. 3.2.8 with the *Solanaceae* repeats database and RepeatRunner with the database of TE-encoded proteins, both included into MAKER installation³⁴. For the genome annotation procedure only scaffolds ≥ 1 kb were used.

The repeated fraction was also evaluated by graph-based clustering of repetitive elements in unassembled reads using the RepeatExplorer Web server⁶⁴ and by analysis of k-mer content using Jellyfish and GCE software (Liu et al., 2012). Putative SINEs were identified using the SINE-Finder tool⁶⁵ and were used to search against published SINE sequences of *S. tuberosum* and other *Solanaceae* using FASTA (E-value $\leq 1e-10$)⁶⁵.

Different E-value thresholds at increasing stringency were tested without significant differences. Members of each family detected in *S. commersonii* were aligned with MUSCLE⁶⁶, and consensus sequences were calculated.

NcRNAs were identified using cmscan (E-value $\leq 1e-10$) from Infernal ver. 1.1 against the database of covariate models of Rfam 11.0 and lncRNAs were identified using the approach described by Boerner and McGinnis (2012). Non-coding transcripts were BLAST searched as well against a database of plant mature miRNA sequences in miRBase (<http://www.mirbase.org/>) to identify homologous miRNAs.

- **RNA-seq data analysis**

For the creation of a more reliable dataset, raw sequencing reads from RNA-seq experiments performed on root, flower, tuber, and leaf samples were checked for quality using FastQC⁶⁷ ver. 0.10.1. Trimming and removal of adapters were performed with CutAdapt⁶⁸ ver. 1.5.2 and FASTX Toolkit⁶⁹ ver. 0.0.13.2.

Trimmed reads were then mapped against the *S. commersonii* genome sequence with TopHat¹³ ver. 2.0.11. Duplicated reads were removed with Picard Tools ver. 1.110 (<http://picard.sourceforge.net>) and the resulting files were used to annotate new transcripts with Cufflinks³¹ ver. 2.2.0 removing the isoforms contained in other isoforms and creating a new annotation file comprising those isoforms belonging to the class 's' as reported by Cuffmerge³¹.

Filtered RNA-seq reads were also *de novo* assembled into contigs using Trinity²⁷ release 2013/02/25 setting minimum contig length parameter equal to 200 bp and requiring at least two independent reads covering each contig.

- **External evidences alignment**

Proteins of length ≥ 33 amino acids and ESTs ≥ 100 bp were selected as external evidences to align onto *S. commersonii* genome together with the curated SwissProt plants protein database (see Results).

Assembled contigs and selected external evidences have been then aligned on *S. commersonii* genome using MAKER with BLAST and EXONERATE and default parameters defined by the authors.

- ***Ab initio* gene prediction**

Ab initio prediction of protein coding genes was performed using AUGUSTUS⁵¹, SNAP⁵⁰ and GeneMark-ES⁵³. SNAP⁵⁰ training procedure undertaken was indeed a formalization of the steps described in the MAKER tutorial. Briefly, a first iteration of MAKER was run using repeats, previously established evidence and default parameters of BLAST and Exonerate. MAKER GFF of derived gene models then has been converted to ZFF format with the script bundled with MAKER ‘maker2zff’. This scripts generates the ZFF format file and a FASTA file with the coordinates of gene models that can be referenced against. These will be used to train SNAP.

The parameter ‘–flank’ passed to the fathom program (included in SNAP) was left to the default value of 1000 flanking bases. The final HMM file has been given in input to the second iteration of MAKER for the gene prediction.

GeneMark-ES⁵³ was chosen as it performs self-training and doesn’t need curated training sets but the genome FASTA file. Indeed the training of *S. commersonii* was performed using randomly selected scaffolds covering about 40 Mbps of *S. commersonii* genome, in accordance with author's instructions⁵³.

- **Final annotation generation and quality assessment**

The final annotation was generated using MAKER pipeline integrating the RNA-seq data and external evidences. In total, two MAKER annotation iterations were carried out, the first one to build a dataset of genes for *ab initio* prediction and the second one to build final annotation.

Gene models with an AED higher than 0.5 were discarded from the final annotation (for further details about AED measure see Introduction).

Predicted ORFs were aligned against the NCBI Non Redundant (NR) database retrieved on 06/2012 with BLAST¹³ (BlastP, e-value < 10-5) and functionally annotated by automatic annotations performed with Blast2GO⁶⁰.

Results and discussion

Roughly 383 Mb of repetitive sequences were identified, accounting for 44.5% of the assembly of the *S. commersonii* genome. Analysis of k-mer distribution in unassembled reads estimated 51.3% of the genome as non-repetitive and graph-based clustering with RepeatExplorer detected a fraction of repeated sequences equal to 36% of the total genome (data not shown). Although these data are not conclusive, they suggest that, compared with potato (Potato Genome Sequencing Consortium, 2011), *S. commersonii* might have a lower amount of repetitive DNA (44.5% versus 55%), which might predict different genome dynamics in these two species since their separation from a common ancestor.

The repetitive fraction of *S. commersonii* genome is dominated by long terminal repeat-retrotransposons (LTR-RTs) (34%) with lower levels of several other repeat types (Figure 11). Characterization of SINE families allowed annotating 1,925 SINEs with significant similarity to families previously described in *S. tuberosum* and in other *Solanaceae* (Wenke et al., 2011; Seibt et al., 2012).

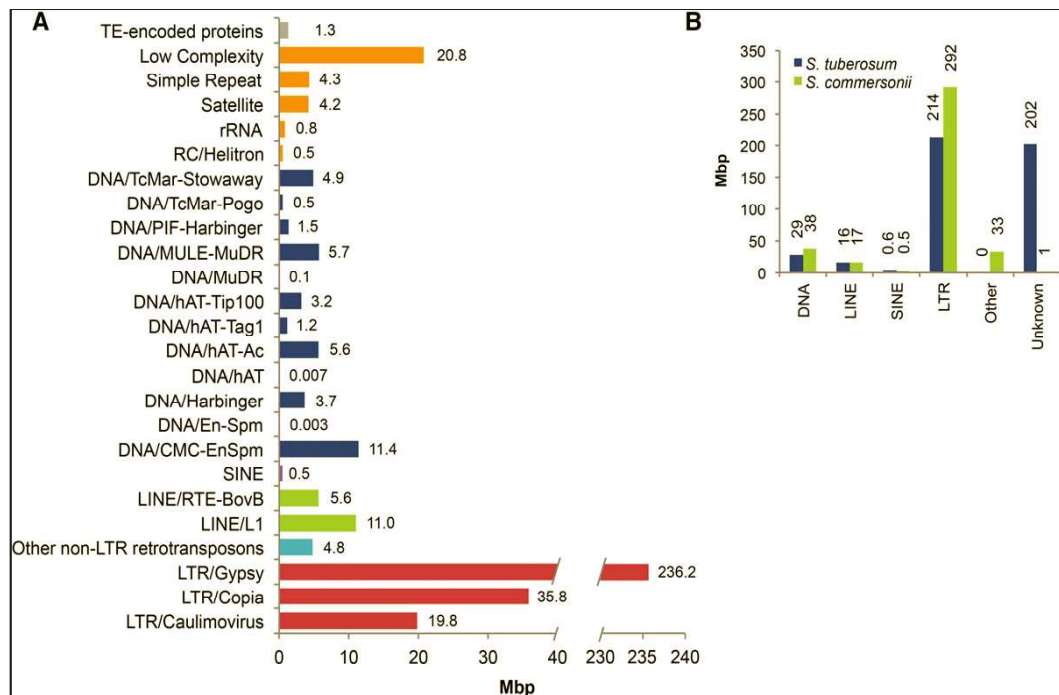


Figure 11. Repetitive Sequence Annotation in the Draft Genome of *S. commersonii*. (A) Classification of repetitive sequences in *S. commersonii*. (B) Comparison of transposable element lengths between *S. commersonii* and *S. tuberosum*⁶³.

The comparison between *S. commersonii* and other Solanaceous species showed differences in terms of repetitive sequences. In a comprehensive review of the first 50 sequenced plant genomes, Michael and Jackson (2013) reported that genome repetitive content ranged from 3% (*Utricularia gibba*) to 85% (*Zea mays*). Compared with potato (55%) and tomato (63%), *S. commersonii* showed a lower amount of repetitive DNA (44.5% of the assembly). As in other *Solanaceae* species, there were many more Ty3-gypsy type than Ty1-copia type LTR-RTs identified in *S. commersonii*, suggesting that the former elements have been somewhat more successful in colonizing and persisting in *Solanaceae* genomes.

LTR-RTs play a substantial role in genome size variation, and the lower frequency of TEs in *S. commersonii* may contribute to its smaller assembly size as well as underline the occurrence of different evolutionary dynamics⁶³.

Also ~21,000 *S. commersonii* ncRNAs were identified. Emerging evidence has revealed that ncRNAs are major products of the plant transcriptome (Rymarquis et al., 2008) and that they may have significant regulatory importance, especially during stress situations (Matsui et al., 2013).

A large number of transcripts (20,994) with no apparent coding capacity were predicted in *S. commersonii*. These ncRNAs comprised a diverse group of transcripts, including 40 among tRNAs and rRNAs, 18,882 long noncoding RNAs (lncRNAs), and 1703 putative microRNA (miRNA) precursors (Figure 12).

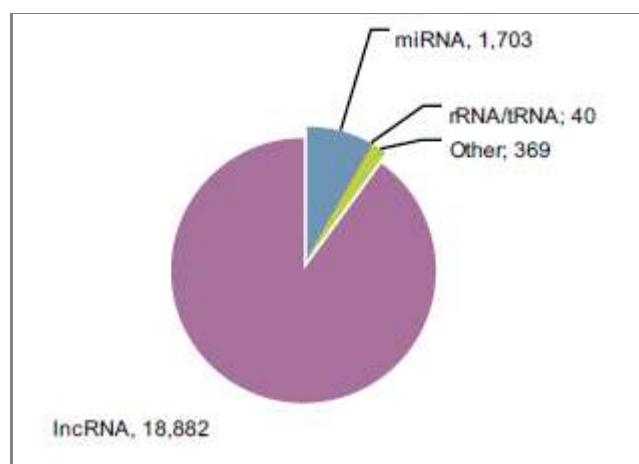


Figure 12. ncRNAs classes in *S. commersonii*.⁶³

A key step toward understanding the biological functions of the predicted miRNAs was achieved through the identification of 4,437 target sites. According to GO term classification, 22% (976) of the target genes are involved in cold response and 10 are potential regulators of transcripts annotated as responsive to cold.

The RNA-seq data filtering procedure resulted in a total 11,7 Gb of data that will be used to create support for *ab initio* predictions (Table 11).

Tissue	Sequenced fragments (100x2)	Filtered fragments (100x2)
flower	36,323,578	26,058,003
root	46,372,066	35,519,519
tuber	49,885,340	22,704,705
leaf	35,855,731	32,935,569

Table 11. Statistics of filtering of *S. commersonii* RNA-seq data.

The reference-based assembly of the transcriptome from leaf, flower, root and tuber samples produced 68,208 transcripts (Table 12) with an N50 length of 2,100, slightly higher than *S. tuberosum* (N50 length of 1,862 bp).

Number of contigs	68,208
Total assembly length (bp)	114,251,410
Average length of contigs (bp)	1,675.04
Minimum length of contigs (bp)	39
Maximum length of contigs (bp)	16,995
N50 length (bp)	2,100
Number of contigs >= 100bp	68,196
Average length of contigs >= 100bp	1,675.32
N50 length of contigs >= 100bp	2,100

Table 12. Statistics of *S. commersonii* reference-based transcriptome assembly.

The *de novo* assembly of RNA-seq filtered reads resulted in a total of 117,816 contigs with more than 96% mapping on *S. commersonii* genome (Table 13).

Assembled sequences. number	117,816
Maximum length. bp	53,539
Average length. bp	1,369.13
Minimum length. bp	301
Median	1,026
N50	1,887
# mapping against assembly	113,559
% mapping against assembly	96.39%

Table 13. Summary statistics of *de novo* assembled *S. commersonii* contigs.

The N50 length of the contigs was 1,887 bp, in line with the one of *S. tuberosum* PGSC v3.4 annotated transcripts (1,862 bp). The first MAKER iteration was run using selected ESTs and protein evidence (Table 14) producing a starting dataset of 79,627 gene models that have subsequently used for the training of SNAP and AUGUSTUS.

Selected external evidences aligned to *S. commersonii* genome with an average rate of ~62% and an overall alignment rate of ~58% (Table 14).

Species	Source	# sequences	# aligned sequences
<i>A. Thaliana</i>	TAIR10 Proteins	35,386	28,492
<i>S. Tuberosum</i>	PGSC v.3.4 Proteins	56,218	52,990
<i>S. Lycopersicum</i>	ITAG 2.3 Proteins	34,727	31,263
<i>Swiss-Prot Plants</i>	13/04/2013 Proteins	36,104	31,531
<i>S. commersonii</i>	NCBI ESTs	548,500	124,272
<i>S. Tuberosum</i>	NCBI ESTs	250,127	59,611
<i>S. Lycopersicum</i>	NCBI ESTs	298,306	74,280
<i>S. commersonii</i>	RNA-seq contigs	117,816	83,567
TOTAL		1,377,184	486,006

Table 14. External evidences used as support to *S. commersonii* gene predictions.

The GFF3 annotation derived from reference-based assembly and the FASTA file of *de novo* assembled contigs were aligned on *S. commersonii* genome together with other external evidences and trained *ab initio* predictors in a second iteration of MAKER.

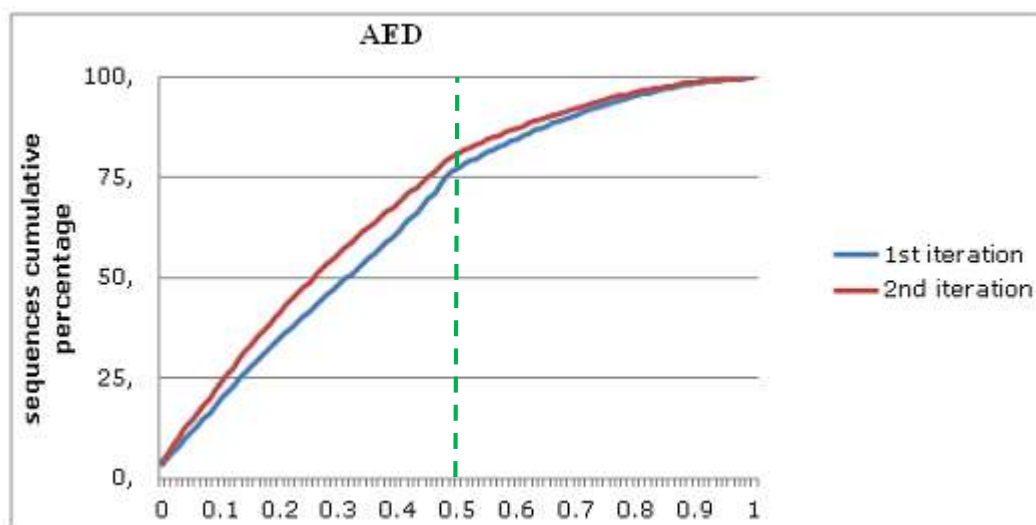


Figure 13. AED cumulative curve of *S. commersonii* 1st and 2nd MAKER iterations.

Genes with an AED ≤ 0.5 , protein length $> 100\text{bp}$ and scaffold length on which the gene was annotated $\geq 1\text{Kb}$ were filtered producing the final *S. commersonii* annotation (Table 15).

	<i>S.commersonii</i> complete	<i>S.commersonii</i> filtered
# genes	70,097	37,662
#mRNAs	72,088	39,493
mRNA mean length	933.6	1,346.3
Exon mean length	221.0	239.4
Intron mean length	588.6	603.0
Mean n. of exon	4.22	5.62
N. of monoexonic	14,022	5,782
% monoexonic	19.45%	14.64%

Table 15. Summary statistics of *S. commersonii* complete and filtered gene annotation.

This filtering procedure has been applied to obtain a final catalogue of high quality gene models and to exclude from the final annotation putative artifacts, pseudogenes and/or orphan genes.

The final gene annotation was composed by 37,662 protein-coding genes. Considering this numbers, fewer genes were predicted in *S. commersonii* than in potato (~39,000) (Figure 14), but the wild species has more predicted genes than tomato (34,727) and other *Solanaceae* (Figure 15).

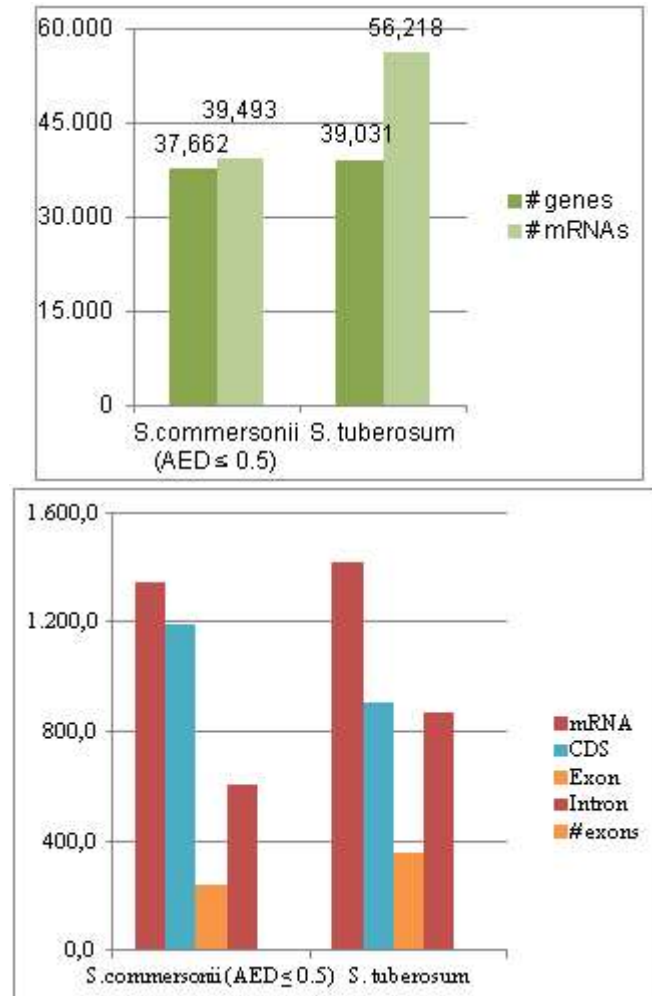


Figure 14. Summary statistics of *S. commersonii* genes and comparison with *S. tuberosum*. On the top are compared the number of genes and mRNAs, below structural features are compared⁶³.

Even though the number of genes found in *S. commersonii* was lower to that reported for *S. tuberosum*, the number of transcripts differed a lot between the two species. Indeed *S. commersonii* had 39,493 alternative isoforms, which was lower respect to the 56,218 isoforms annotated in potato (Figure 13). This might highlight the presence of more prominent alternative splicing activities in potato than in *S. commersonii*⁶³. This is consistent with observations by the Potato Genome Sequencing Consortium (2011) that 25% of potato genes encoded two or more isoforms, indicative of more functional variation than is represented by the gene set alone.

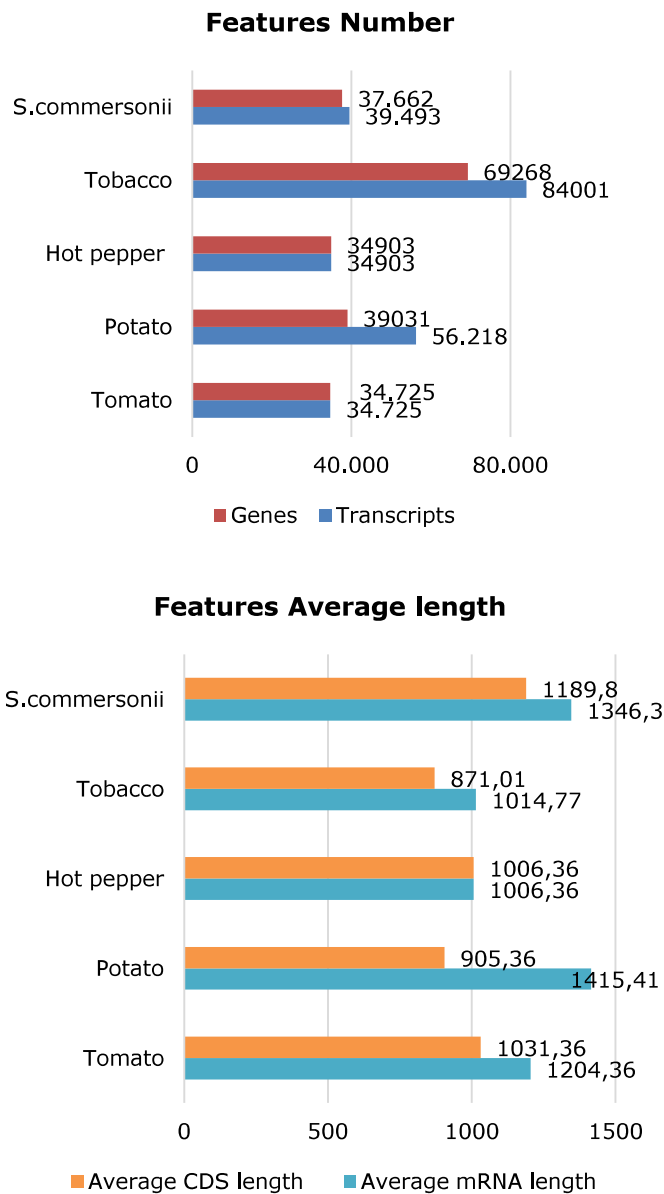


Figure 15. Summary statistics of *S. commersonii* gene features and comparison with other *Solanaceae*. On the top are compared the number of genes and the number of transcripts. Below CDS and mRNA lengths are compared⁶³.

A number of 13,996 genes (~37% of total) are supported by all sources of evidence (Figure 16). A number of ~22,000 genes (~58% of total) are supported either by transcript evidence, protein evidence and *ab initio* predictors while 7,670 (20.4% of total) are supported solely by *ab initio* predictors and proteins. No genes are supported solely by ESTs or RNA-seq or *ab initio* predictors. In general, few genes are supported by solely 2 out the 4 sources of evidence except from the pairing of *de novo* and protein supported gene models (Figure 16).

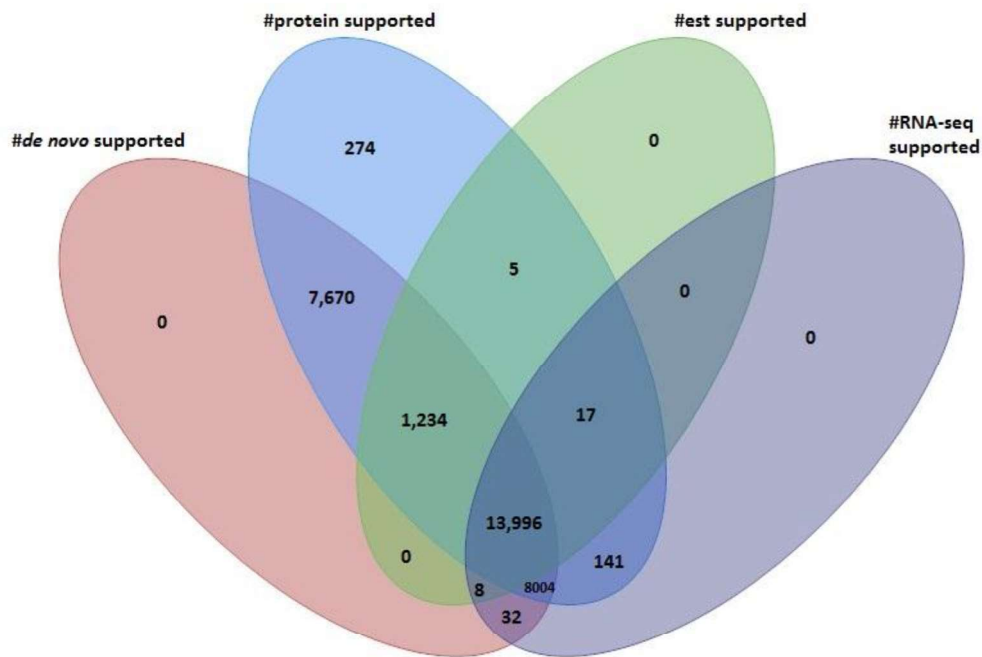


Figure 16. Venn diagram for sources of evidence for *S. commersonii* gene annotation.

Of predicted *S. commersonii* genes, 30,477 (~85.5%) predicted protein-coding genes had significant BLAST similarity to protein-coding genes from other organisms in the nonredundant NCBI database. Nearly 20,500 *S. commersonii* genes were assigned to Gene Ontology (GO) terms, and more than 4,900 proteins were annotated with a four-digit EC number (Figure 17).

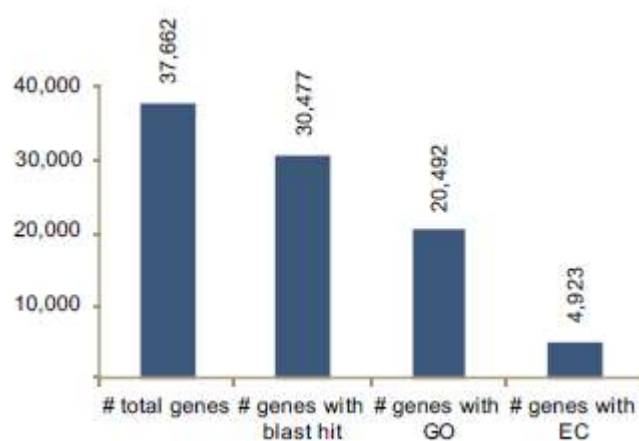


Figure 17. Functional annotation of *S. commersonii* transcriptome.⁶³

These results imply that more than 85% of *S. commersonii* gene models have a putative corresponding gene of another species annotated in NCBI database and that more than 24% of the predicted proteome of *S. commersonii* may have an enzymatic function, thus increasing the reliability of the annotation produced automatically.

Annotation of a genome with no reported reference

Background

Eggplant (*Solanum melongena* L. $2n = 2x = 24$, projected genome size 1.1 Gbp) belongs to the economically important family of the *Solanaceae*, which also includes a number of other important crops like tomato, potato, pepper and tobacco. Genomic studies in the *Solanaceae* family resulted firstly in the availability of the genome sequence of tomato⁷⁰ (Tomato Genome Consortium) and potato⁷¹, followed in 2014 by chili pepper⁷².

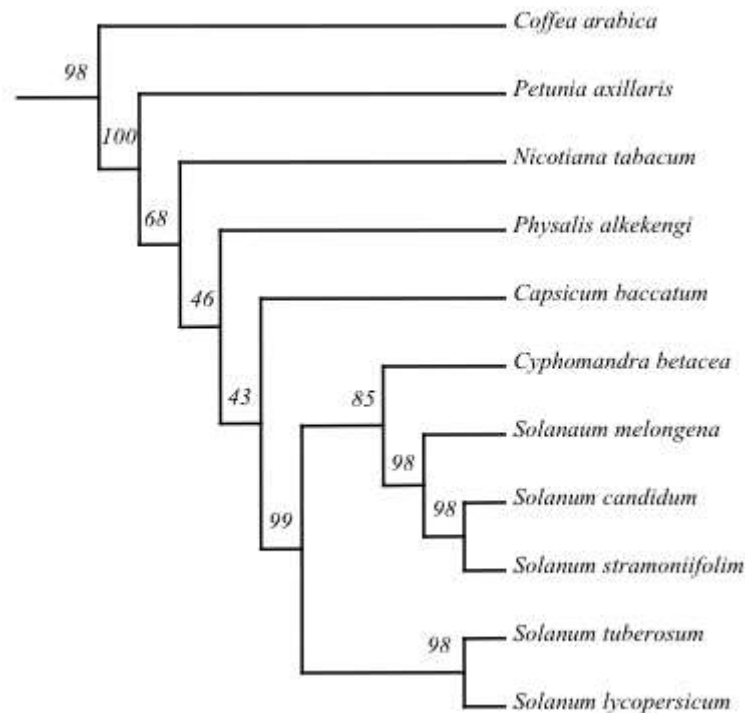


Figure 18. Phylogeny of the Solanaceae (solgenomics.net).

Genomic studies on eggplant so far were mainly focused on the development of intra-specific⁷³⁻⁷⁴ as well as inter-specific⁷⁵⁻⁷⁶ maps. Fukuoka et al. (2010) developed a set of 16,245 unigenes while Barchi et al. (2011) made available a very first set of SNP markers for the species, obtained by combining the so-called “Restriction-site Associated DNA”⁷⁷ method with Illumina DNA sequencing. Finally, a first draft of the genome sequence has been recently released, although

only 12% of its sequence was anchored to the genetic map⁷⁸. An high quality reference genome for *S. melongena* is not yet available.

Materials

The genome annotation procedure started with the production of an high quality *S. melongena* genome assembly obtained from sequencing at >150x coverage of 3 Illumina standard libraries and 5 mate pair libraries with different insert sizes (respectively 400, 500, 600 bp and 3.4, 5, 6.5, 7.8, 10 kb). The assembly obtained was filtered by masking the sequences potentially deriving from contaminants and removing the sequences shorter than 1Kbp, leading to a final genome of 1.28Gb with N50 length of 641Kb.

In order to annotate the assembled genome, based on the fact that no or few available data of *S. melongena* are present in public databases, also stranded RNA-Seq samples from 19 tissues were sequenced.

Also a dataset of 111 genes (thereafter called ‘control genes’), composed either by manually curated *S. melongena* CDS or from flcDNA alignments, were used as high quality gene set in the genome annotation process for the evaluation of the quality of annotation since it was too small as dataset to be used as evidence for gene annotation.

Methods

The genome annotation strategy for *S. melongena* assembly is the following:

Figure 19. Diagram of strategy adopted for *S. melongena* genome annotation.

- **Repeat masking of the genome and detection of putative non-coding sequences**

The annotation strategy was based on first identification and masking of repetitive sequences and other non coding sequences on genome.

A repeat library for *S. melongena* was created following the custom advanced protocol suggested for MAKER⁸⁰ ver. 2.31.3, a combination of structural-based and homology-based approach is used to maximize the opportunity for repeat collection (see Introduction).

The final repeat library has been fed to MAKER which run RepeatMasker together with the *Solanaceae* repeat library and RepeatRunner.

In order to detect non-coding RNA sequences in the genome, sequence similarities were inferred using the reference-based method Infernal⁴⁶ ver. 1.1 using the covariance model of Rfam database⁴⁷ release 12.0. The analysis of ribosomal regions was performed with RNAmmer⁸³.

- **RNA-seq data *de novo* assembly and training dataset selection**

Raw sequencing reads from RNA-seq experiments were checked for quality using FastQC⁶⁷ ver. 0.10.1, trimming and removal of adapters were performed with CutAdapt⁶⁸ ver. 1.5.2 and FASTX Toolkit⁶⁹ ver. 0.0.13.2.

RNA-Seq reads from the different 19 samples analyzed stages then were aligned against the reconstructed genome using the Tuxedo suite¹³. The TopHat program was used with the options "--b2-very-fast" and for each library was specified the average insert size and its standard deviation as estimated during library preparation, while the GFF of repetitive sequences identified by RepeatMasker was provided with the "-M" option. The alignments in BAM format were then analyzed by the Cuffmerge program with default parameters.

RNA-Seq assemblies were constructed separately for each of all the samples using the *de novo* pipeline of Velvet-Oases⁸⁴ ver. 0.2.08 and subsequently merged using EvidentialGene⁸⁵ "tr2aacds.pl" ver. 2013.07.27 pipeline. EvidentialGene allowed removing redundancy among samples and selecting high quality transcripts. To get a first impression about the completeness of the assembled transcriptome, the dataset was analyzed with CEGMA⁷⁹.

To define a first set of transcript assemblies that could be used to train gene predictors reliably, the *de novo* assembled transcriptome was compared with the proteomes of other four species – *N. benthamiana*, *S. lycopersicum*, *S. tuberosum* and *A. thaliana* - retaining only sequences with over 50% identity and 99% reciprocal coverage. The strict parameters ensured that only conserved sequences were retained and minimized the risk of including gene fusions in our assemblies.

In following analyses, these parameters have been used to define high-quality conserved transcripts independently of genome-related quality measures (such as AED, cDNA length, or similar).

The resulting sequences were aligned against the masked genome using MAKER, increasing its default parameters to include only alignments with 95% identity, 95% coverage and 70% of the Exonerate score threshold. In this first iteration, MAKER was used to assign a CDS to multiexonic aligned assemblies. The

alignments have been subsequently filtered to remove any BLAST alignment with an intron longer than 10 kbps, to minimize the chance of gene fusions.

The predicted models were compared again with the four original proteomes, and only those that matched again the original criteria – 50% identity and 99% reciprocal coverage – were selected, in order to eliminate any potential artifact introduced by the alignment.

Gene matching the control gene set were further removed from the training dataset to avoid bias in the annotation quality evaluation.

- **External evidences alignment**

In order to annotate the coding genes of *S. melongena*, publicly available evidences for *S. melongena* with other ESTs and protein resources available for related species, such as *S. torvum*, *lycopersicum* and *tuberosum*, were pooled together (see Results).

Also eggplant ESTs and RNA-Seq assemblies were aligned to the assembled genome and those with 95% of identity, 95% coverage, and 70% of Exonerate maximal score threshold were retained⁵⁶. On the other hand ESTs and proteins of other species have been aligned with a minimum of 70% coverage, 50% identity, and 70% of Exonerate maximal score threshold. The alignments have been subsequently filtered to remove any BLAST alignment with an intron longer than 10 kbps, to minimize the chance of gene fusions.

- ***Ab initio* predictors training and gene prediction**

The obtained dataset of refined gene models were used for the first round of training of five *ab initio* predictors: SNAP, AUGUSTUS, GeneMark-ET, geneid and Twinscan.

SNAP training has been performed setting the parameter “–flank” as the default value of 1000 flanking bases for training and testing. AUGUSTUS has been trained using the *S. melongena* genome to create the final model for the species.

GeneMark-ET was trained using as input the file of intron coordinates from Tophat and Cufflinks of the training dataset and the soft masked genome FASTA file.

Geneid training has been optimized using the 10-fold cross-validation procedure to estimate the accuracy of each choice of exon weight factor (ewf) and oligo weight factor (owf). The best combination of exon weight factor and exon factor was then selected as final.

TwinScan⁸⁶ was trained using the utilities provided together with the tool, which make use of BLAST and BLAT alignments to derive conservation sequences. Sequence conservation was calculated with MEGABLAST using as reference the tomato genome sequence release ITAG2.4. Also the *S. melongena* masked genome sequence has been provided to the programme. The parameter estimation for TwinScan was made using iParameterEstimation⁸⁷, a configurable maximum likelihood parameter estimation package for gHMMs. It points to an annotation set (the training examples), a gHMM file (the parameter definitions), and one or more feature map files. At the end “parameters.zhmm” and “parameters.xml” parameters files were created for the training. These are all either probabilities, log scores, or log-odds ratios of probabilities, depending on what is expected by *iscan*, and indicated by the gHMM file.

For AUGUSTUS, geneid and Twinscan, the training has been optimized manually by using a 10-fold cross validation to assess the sensitivity and specificity at the gene, exon and nucleotide levels. The 111 control genes dataset has been used as test-set to estimate the performance of the gene predictors.

Geneid and Twinscan gene predictions were run outside MAKER, since the pipeline doesn't support such predictors, and the GFF file of gene predictions was given to MAKER together with the other evidences to be used to produce final consensus gene annotation.

- **Final annotation generation and quality assessment**

Every form of evidence at the end was integrated in MAKER to produce the final annotation. The quality of annotation has been evaluated using MAKER AED quality measure, comparison with other published annotations and the 111 *S. melongena* control genes dataset.

Results and discussion

The repeat content analysis identified 56.12% of the genome as repetitive (Table 16) with LTR Gypsy/DIRS1 being the most abundant repeat class as these transposons alone covered about 314 Mb of the genome.

Type	Number	Bases	% of genome
Simple repeats	7,444	2,990,912	0.23
LINE/SINE	173,292	46,064,916	3.57
DNA transposons	368,644	77,027,848	5.98
Unclassified	188,245	33,213,758	2.58
LTR	1,047,883	573,711,377	44.55
TOTAL		722,670,286	56.12

Table 16. Summary statistics of *S. melongena* genome repetitive content.

This percentage is slightly lower but in line to that present in other *Solanaceae* like tomato (63.28%) and potato (62.20%).

Based on the homology to known miRNAs of 13 species with different evolutionary relationships, 168 different *Solanum melongena* miRNAs from 45 families were predicted from the search against miRBase.

A total of 905 ribosomal regions on 557 different scaffolds were detected with RNAmmer: 91 of 18S rRNA, 56 of 28S rRNA, and 758 of 8S rRNA (Table 17).

Family	#
<i>tRNAs</i>	2,856
<i>rRNA</i>	905
<i>Small nucleolar RNAs</i>	283
<i>miRNAs</i>	168
<i>Other</i>	2,269
	6,481

Table 17. Summary statistics of ncRNAs annotation.

The search in Rfam non-coding RNA families database resulted in the identification of 2,856 tRNAs in line with other sequenced plants⁷¹ but much more for example respect to tomato ITAG 2.4 annotation (886).

RNA-Seq samples from 19 tissues were sequenced obtaining on average ~19,4 millions of reads per sample (Figure 20).

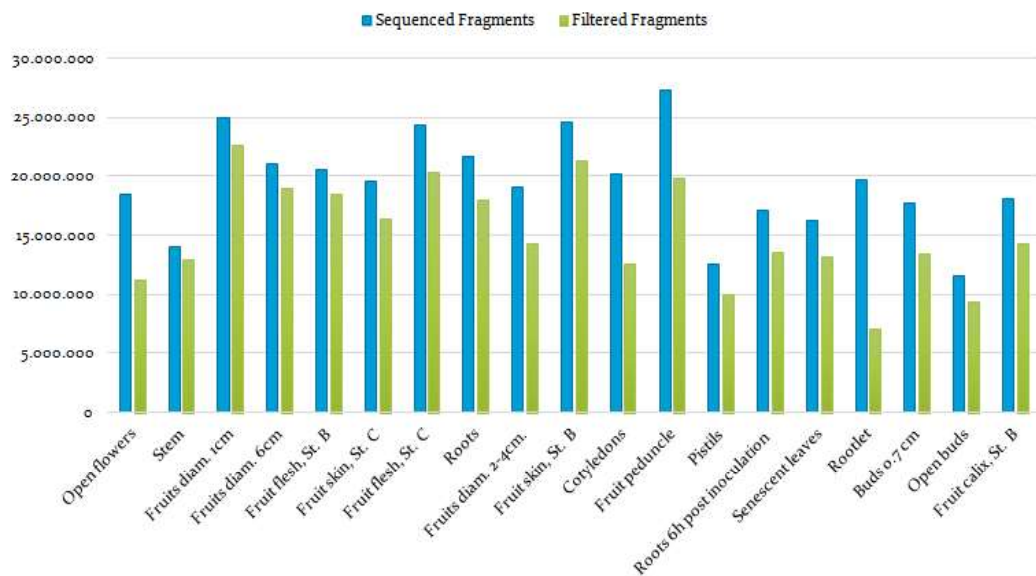


Figure 20. Summary statistics of the sequencing of *S. melongena* RNA-seq samples.

In order to reconstruct transcripts and both use them as evidence for annotation as well as for the raining of *ab initio* predictors, the filtered RNA-seq reads have been *de novo* assembled, producing 39,408 main isoforms (each corresponding to a gene) and 87,836 alternative isoforms. The *de novo* assembled transcripts obtained for each RNA-seq sample have been pooled together and clustered with EvidentialGene in order to reduce the redundancy of the dataset (Table 18).

	«Main isoform» set	«Alternative isoform» set	Filtered reliable sequences
<i>Number of sequences</i>	39,408	87,836	14,353
<i>Maximum length</i>	4,997	3,761	12,179
<i>Minimum length</i>	40	40	188
<i>Average</i>	316.73	278.2	1,658.58
<i>Median</i>	242	226	1,978
<i>Sequences with complete ORFs</i>	26,082	38,524	-

Table 18. Summary of clustered *de novo* assembled transcripts. Statistics of both main isoforms and alternative isoforms clustering sets are reported.

The assembled transcriptome of *S. melongena* and other *Solanaceae* was checked at first for Core Eukaryotic Genes (CEGs) to get a final list of those CEGs actually present in the transcriptome, their degree of completeness and the paralogy. The analysis shows that almost the entire dataset could be traced on the eggplant transcriptome (Figure 21).

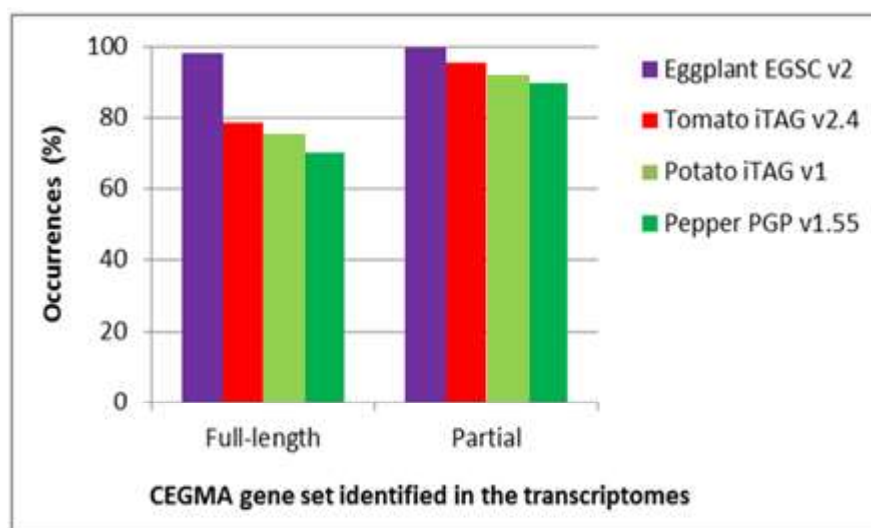


Figure 21. CEGMA analysis on *Solanaceae* transcriptomes.

The predicted aminoacidic sequences from the *S. melongena de novo* assembled transcriptome have been aligned to the four proteomes of *Solanaceae* and filtered

based on a 50% cut-off on identity and 99% reciprocal coverage with the proteomes, obtaining therefore a final set of 14,353 sequences (Table 18).

The resulting 14,353 transcripts have been then aligned to the genome assembly of *S. melongena* using MAKER software without any gene predictor, just setting identity and coverage cut-offs of 95%. In addition, the repeats and the ncRNAs annotation files have been included in the program run. MAKER used all these evidences to infer gene models directly and at the end predicted 8,751 gene models that are located in 1,217 scaffolds (Table 19).

Program	# gene models	# scaffolds
Blastn	12,035	1,425
Exonerate	12,529	1,403
MAKER	8,751	1,217

Table 19. Results of first run of MAKER with *S. melongena* transcriptome.

In order to check the 8,751 gene models predicted by MAKER, also 104 manually curated control gene models and 49 full-length cDNA sequences have been aligned to the *S.melongena* genome assembly using MAKER with the same parameter used for the RNA-seq dataset. Out of the 153 control genes, MAKER at the end of its run predicted 110 gene models on 83 scaffold (Table 20).

Program	# gene models	# scaffolds
Blastn	133	104
Exonerate	133	104
MAKER	110	83

Table 20. Results of MAKER run with *S. melongena* control genes.

Among this 111 gene models, 102 are derived from manually curated control genes and the remaining 9 from the full-length cDNAs. These genes will be used for the evaluation of the training of *ab initio* predictors. Of the 8,751 gene models, 70 matched the control gene set and were removed from the training dataset to avoid bias in the annotation quality evaluation.

MAKER was used in conjunction with predictions from geneid and TwinScan and the parameters files resulting from the training of SNAP, AUGUSTUS and GeneMark-ET (see Methods) to obtain a first iteration of the genome annotation (Table 21). The selected 8,681 gene models that passed the quality filters (see Methods) were used to perform a second round of training for *ab initio* gene predictors. While a marked improvement for geneid was observed, the other two gene predictors maintained a similar accuracy in the second round.

		<i>AUGUSTUS</i>		<i>TwinScan</i>		<i>Geneid</i>	
		<i>First iteration</i>	<i>Second iteration</i>	<i>First iteration</i>	<i>Second iteration</i>	<i>First iteration</i>	<i>Second iteration</i>
Gene	Sensitivity	55,56%	57,41%	54,05%	55,86%	32,43%	42,34%
	Specificity	38,96%	40,00%	49,59%	47,33%	21,43%	24,10%
Exon	Sensitivity	91,33%	92,16%	84,26%	86,44%	76,26%	76,39%
	Specificity	79,62%	79,57%	85,30%	85,04%	71,87%	69,85%
Nucleotide	Sensitivity	98,22%	98,30%	91,35%	97,47%	93,56%	94,44%
	Specificity	84,62%	84,83%	90,63%	82,64%	85,17%	88,91%

Table 21. Summary of estimated accuracy of the gene predictors, measured against the 111 control genes. For each iteration of the prediction software are reported sensitivity, specificity and F-Score at Gene, Exon and nucleotide levels.

Gene prediction for geneid and TwinScan resulted in 263,859 non-unique gene models with an N50 length respectively of 3,270 bp and 1,347 bp that will entirely be fed to MAKER and integrated with the other gene predictions to produce the final *consensus* annotation (Table 22).

	Geneid	TwinScan
<i>Number of gene models</i>	93,935	169,924
<i>Maximum transcript length (bp)</i>	79,281	58,764
<i>Minimum transcript length (bp)</i>	3	3
<i>Mean transcript length (bp)</i>	1,443.80	799
<i>N50 transcript length (bp)</i>	3,270.0	1,347

Table 22. Summary statistics of geneid and TwinScan gene predictions.

Selected external evidences have been aligned to *S. melongena* genome obtaining an average alignment rate of ~53% and an overall alignment rate of ~37% (Table 23).

Species	Evidence type	Source and date of retrieval	#total sequences	# aligned sequences
<i>Solanum melongena</i>	RNA-Seq <i>de novo</i>	Internal data	127,117	87,652
<i>Solanum melongena</i>	TopHat junctions	Internal data	307,611	34,270
<i>Solanum melongena</i>	Sanger ESTs	NCBI (2014/06/03)	98,087	90,892
<i>Solanum torvum</i>	Sanger ESTs	NCBI (2014/06/03)	28,743	21,148
<i>Solanum lycopersicum</i>	Sanger ESTs	NCBI (2014/06/03)	300,359	116,850
<i>Solanum lycopersicum</i>	Predicted proteins	ITAG 2.4	34,725	25,649
<i>Solanum tuberosum</i>	Sanger ESTs	NCBI (2014/06/03)	250,128	92,448
<i>Solanum tuberosum</i>	Predicted proteins	ITAG 1	35,004	28,870
<i>Nicotiana benthamiana</i>	Sanger ESTs	NCBI (2014/06/03)	21,749	5,156
<i>Nicotiana benthamiana</i>	Predicted proteins	Niben genome v. 0.4.4	76,379	44,577
<i>Capsicum annuum</i>	Sanger ESTs	NCBI (2014/06/03)	118,651	32,835
<i>Capsicum annuum</i>	Predicted proteins	Pepper annotation v. 1.55	34,899	29,184
SwissProt Plants	Proteins	Uniprot (2014/05/30)	37,494	25,493
<i>Ricinus communis</i>	Predicted proteins	JCVI	31,452	14,811
<i>Arabidopsis thaliana</i>	Predicted proteins	TAIR10	35,386	17,145

Table 23. Table of external evidences used to create support to *ab initio* gene predictions for *S. melongena*.

The final MAKER annotation produced 48,412 transcripts in 44,618 gene *loci*. This dataset comprises both low- and high-quality gene models, which can be distinguished using the AED quality measures provided by MAKER itself. Since the AED score indicates its congruence with available evidence at its genomic locus, the feasibility of using such a measure to flag reliable gene models was investigated.

Therefore the whole *S. melongena* proteome was compared with the four reference solanaceous proteomes to identify highly conserved sequences.

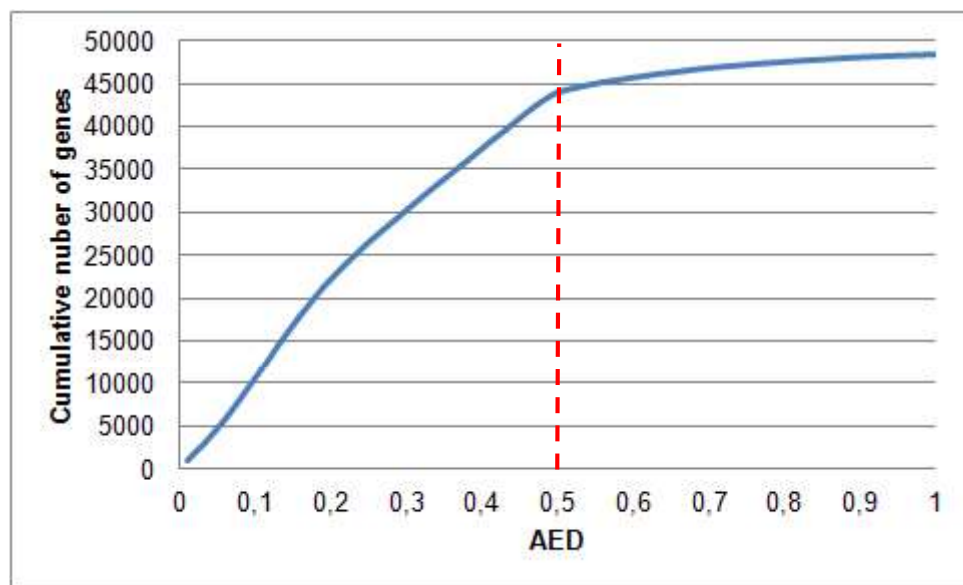


Figure 22. AED curve of *S. melongena* gene models.

For the final annotation, gene models with an AED of 0.48 or lower were retained, as 99% of the transcripts with a clear ortholog in another proteome (50% identity or higher, 99% or higher reciprocal BLASTP coverage, length ratio between 99% and 101%) had an AED equal or lower than this threshold (Figure 22).

The final annotation contained 43,579 transcripts in 39,921 loci, comparable with the tomato and potato annotations, but greatly better than the one recently published (Table 24, Figure 23).

The gene annotation was performed on the previously repeat masked 23,039 scaffolds with length > 1 kbp. Shorter (non informative) sequences have been excluded as suggested by the authors of MAKER^{4,34,80}.

	Eggplant	Tomato ITAG2.4	Potato ITAG1
<i>Number of genes</i>	39.921	34.725	35.004
<i>Number of mRNAs</i>	43.579	34.725	35.004
<i>Mean intergenic distance</i>	6.972,87	8.760,20	5.092,15
<i>Mean mRNA length</i>	1.393,33	1.204,36	1.070,27
<i>Mean CDS length</i>	1.203,30	1.031,36	1.070,27
<i>Mean Exon length</i>	287,8	261,38	246,86
<i>Mean intron length</i>	647,79	539,41	598,21
<i>Average number of exons</i>	4,84	4,61	4,34
<i>Average number of CDS exons</i>	4,64	4,53	4,34
<i>Monoexonic</i>	10.943	8.508	10.344
<i>Monoexonic Percentage</i>	25,11%	24%	29%

Table 24. Summary *S. melongena* annotation compared to other solanaceous crops (*S. lycopersicum* and *S. tuberosum*).

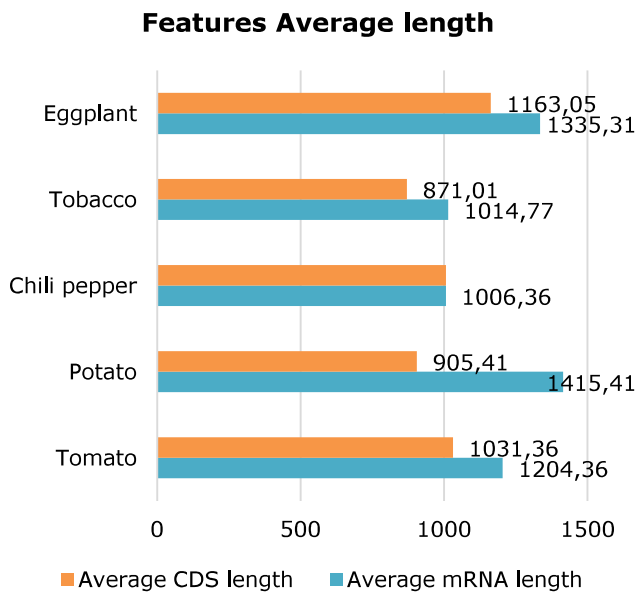
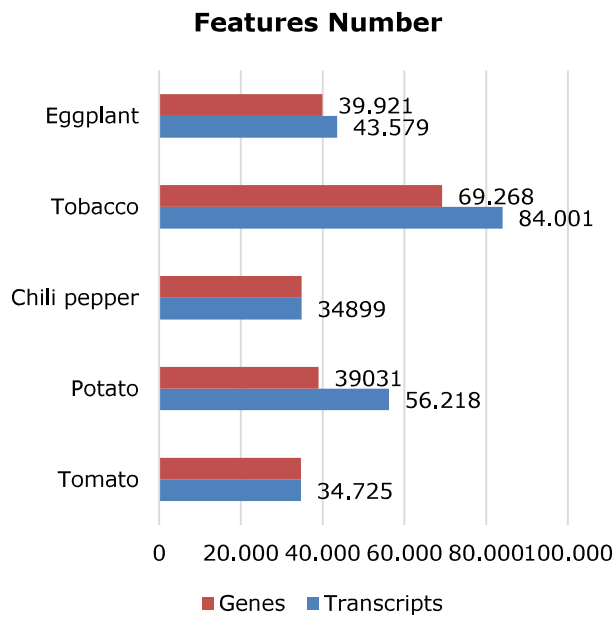


Figure 23. Comparison of the Eggplant MAKER annotation with the annotations of other *Solanaceae*. On the top the number of genes and transcripts are compared. Below average CDS and mRNA length are compared.

The comparison of structural features between *S. melongena* and other *Solanaceae* showed that eggplant had some annotated alternative isoforms respect to the other plants and that it had the longest average mRNA length, but in means of number of genes and lengths the final results are in line with other published plant annotations (Figure 23).

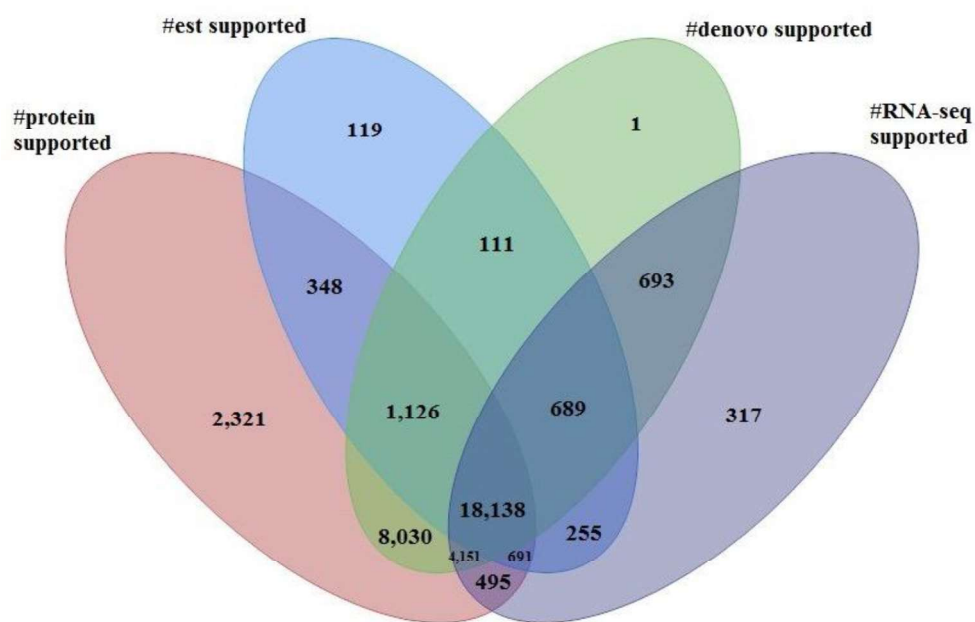


Figure 24. Venn diagram for sources of evidence for *S. melongena* gene annotation.

Most of annotated genes (93,8 %) are supported by reliable evidence, such as transcripts or proteins, RNA-seq data or *ab initio* predictors. A number of 18,138 genes (45.4% of total) are supported by all sources of evidence (Figure 24), while 3,133 (7.8% of total) are supported either by transcript evidence or protein evidence. In general, most of genes are supported by other plants' proteins evidence or *ab initio* predictors.

Finally, a quality assessment based on the 111 manually curated control genes also supported the general reliability of the final annotation: the majority of control genes (79, or 71.2%) are reconstructed correctly or exhibit minor discrepancies (14 transcripts, or 12.61%).

	<i>Description</i>	<i>Count</i>	<i>Percentage</i>	<i>Avg AED against reference</i>
Perfect	Perfect concordance between reference and prediction.	72	64,86%	0
Supported	Reference and prediction disagree, but the available evidence indicates that the predicted model might be correct.	7	6,31%	0,11
Unclear	Evidence in the region does not allow discerning whether the reference or the prediction is correct.	11	9,91%	0,1
Wrong (minor)	The predicted model is incorrect, but the difference is small (AED ≤ 0.05)	14	12,61%	0,025
Wrong	The predicted model is incorrect, with significant differences	7	6,31%	0,1662
Total		111	100%	0,03

Table 25. Annotation assessment with control genes. The 111 control gene were compared to the respective Maker gene models and the concordance of the structures was used to assess the quality of the annotation pipeline results.

Only in 21 cases, the prediction has been confirmed as incorrect by evidence alignment, although in most cases the discrepancy is very limited, as 14 of such gene models have an AED score lower than 0.05 (Table 25).



Figure 25. Manual revision of AROA_SOLLC control gene. Discrepancies are indicated by red arrows; red bars indicate congruence on exon boundaries.

For example, in one case the discrepancy arises from the alignment of a *N. benthamiana* protein with an enlarged exon, which prompted MAKER to try to reconcile this evidence with the available gene predictions (Figure 25).

The origin of the discrepancies seems to be the locus-specific and cannot be solved by using automatic filters, therefore, even if results show a good overall annotation quality, manual review would be always preferable.

CONCLUSIONS

Second Generation Sequencing technologies have revolutionized completely genome projects, reducing sequencing costs and diminishing the time necessary to obtain a complete genome, thus giving the opportunity to access whole genome sequencing data easily but on the other hand increasing the demand for rapid and accurate genome annotation.

In this PhD project the automated structural annotation of three different eukaryotic genomes has been performed in order to first annotate a species with a published reference genome annotation; the second experimental case regarded a species with a close but phylogenetically distinct reference and the third case regarded a species with any close reference genome annotation.

In the first case it has been experimented the transfer of a closely related genome annotation to a new sequenced genome in order to reliably annotate a good portion of protein coding genes in a relatively short time, followed by *ab initio* prediction of the remaining portion of protein coding genes driven by a refined high quality dataset of gene models derived from transferred reference gene models and supported by other organisms evidence.

When experimenting the case of a close but phylogenetically distinct reference genome, the transfer of the reference annotation has not been considered a reliable choice. So it has been decided to perform *de novo* genome annotation and to produce RNA-seq data in order to support the *ab initio* predictions together with the external evidences.

The results produced by the two experimental cases strongly influenced the strategy defined for the third case, the one about the annotation of a genome with no reported reference. Since no reference annotation can be transferred, also here the genome annotation strategy was based completely on a *de novo* approach.

In this case RNA-seq data has been produced not only to support *ab initio* prediction but also have been ‘polished’ to obtain an high quality dataset to drive the gene prediction. To this purpose, several external evidences have been selected not only to support gene predictions but also to filter the RNA-seq data to produce an high quality gene set and to improve the accuracy of final annotation.

The results obtained in this work show that, as expected, the complexity of eukaryotic genomes greatly influences the annotation process (e.g., the presence of repetitive regions, overlaps, duplications, etc.) but demonstrated that also in case of fragmented assemblies - although being aware of losing part of the information - it is possible to perform the automated genome annotation.

After the identification of repeats and other non-coding sequences, a big fraction of the genes in a genome sequence can be found by homology to other known genes or proteins and refined using RNA-seq evidence and/or *ab initio* predictors. This fraction is supposed to increase as more genomes get sequenced and annotated.

Furthermore, the accuracy of *de novo* predictions is expected to increase since *de novo* gene finders for annotating protein-coding genes in complex genomes are constantly improving¹⁰.

On the other hand, RNA-seq data could be helpful since a major limitation of using *ab initio* gene finders is the inability to investigate the biological phenomenon of alternative splicing. Even if the exons within a novel gene region are precisely predicted, there is currently no precise computational method to determine which exons should be included in the transcript.

Choosing a weight for *ab initio* prediction and RNA-seq data is difficult⁸⁹. A particular *de novo* gene prediction might be especially strong, on the basis of its conservation pattern and splice sites, whereas a cDNA alignment in the same locus might be weak owing to unusual splice sites or multiple mismatches near the splice sites or even heterozygosity.

Gene finders *per se* are hugely inaccurate and finding novel genes from a purely *ab initio* approach is still a major challenge, while species-specific data may help with this issue.

RNA-seq technology allows a more accurate annotation of exon-intron boundaries, the detection of UTR regions – not possible for *de novo* gene predictors – and helps the detection of alternative splicing isoforms.

RNA-seq is a powerful tool, even if covering all transcriptome landscape for a species would be very expensive and time-consuming task since it is impossible to do this with one sequencing run.

In general, there is no ‘best’ approach for genome annotation but results obtained suggest that the integration of multiple sources of annotation greatly improves the accuracy of the final annotation.

Each method has its own advantages, and it is clear that there are genes that can be found by each method may not be found by the other. For example, recently evolved genes are difficult to find using extrinsic, homology-based methods, and genes that are missing common features or that do not fit common profiles of genes are difficult to find using *ab initio* methods⁸.

In conclusion, in some ways cheap sequencing has complicated genome annotation. The fragmented assemblies and complex 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 challenges⁵.

Genome annotation has moved beyond merely identifying protein-coding genes to include an ever-greater emphasis on the annotation of transposons, regulatory regions, pseudo-genes and ncRNAs⁵.

Annotation quality control and management are also increasingly becoming bottlenecks. The process of genome annotation is not error free; manual curation is still required as is the periodic update to old genome annotations, since incorrect and incomplete annotations poison every experiment that makes use of them.

In today's genomics-driven world, as long as tools and sequencing technologies continue to develop, providing fast and accurate and up-to-date annotations is simply a must.

BIBLIOGRAPHY

1. Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, 11(1), 31–46. doi:10.1038/nrg2626.
2. Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)*, 6, 287–303. doi:10.1146/annurev-anchem-062012-092628.
3. Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*. doi:10.1016/j.jgg.2011.02.003.
4. The C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396), 2012–2018. doi:10.1126/science.282.5396.2012.
5. Arabidopsis, T., & Initiative, G. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), 796–815. doi:10.1038/35048692.
6. Barbazuk, W. B., Fu, Y., & McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. *Genome Research*. doi:10.1101/gr.053678.106.
7. Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews. Genetics*, 5(4), 276–287. doi:10.1038/nrg1315.
8. Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A., & Birren, B. (2005). Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Research*. doi:10.1101/gr.3767105.
9. Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1), 36–46. doi:10.1038/nrg3117.
10. Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews. Genetics*, 9(1), 62–73. doi:10.1038/nrg2220.

11. Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5), 329–342. doi:10.1038/nrg3174.
12. Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*. doi:10.1002/0471250953.bi0410s25.
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–10. doi:10.1016/S0022-2836(05)80360-2.
14. Arita, K., Kanno, T., Yoshikawa, M., & Habu, Y. (2011). Non Coding RNAs in Plants. *RNA Technologies*, 237–249. doi:10.1007/978-3-642-19454-2.
15. Costa, F. F. (2008). Non-coding RNAs, epigenetics and complexity. *Gene*. doi:10.1016/j.gene.2007.12.008.
16. Kung, J. T. Y., Colognori, D., & Lee, J. T. (2013). Long noncoding RNAs: Past, present, and future. *Genetics*. doi:10.1534/genetics.112.146704.
17. Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9), 698–709. doi:10.1038/nrg890.
18. Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*. doi:10.1016/j.tig.2007.12.007.
19. Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*. doi:10.1016/j.tig.2007.12.006.
20. Kapustin, Y., Souvorov, A., Tatusova, T., & Lipman, D. (2008). Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct*, 3, 20. doi:10.1186/1745-6150-3-20.
21. Zhou, L., Pertea, M., Delcher, A. L., & Florea, L. (2009). Sim4cc: A cross-species spliced alignment program. *Nucleic Acids Research*, 37(11). doi:10.1093/nar/gkp319.
22. Slater, G. S. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1), 31+. doi:10.1186/1471-2105-6-31.
23. Otto, T. D., Dillon, G. P., Degraeve, W. S., & Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 39(9). doi:10.1093/nar/gkq1268.

24. Holt, C., & Yandell, M. (2011). MAKER: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1), 491. doi:10.1186/1471-2105-12-491.
25. Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., ... Jones, S. J. M. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21), 2872–2877. doi:10.1093/bioinformatics/btp367.
26. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., ... Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12), 1660–1666. doi:10.1093/bioinformatics/btu077.
27. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–512. doi:10.1038/nprot.2013.084.
28. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), R36. doi:10.1186/gb-2013-14-4-r36.
29. Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7), 873–881. doi:10.1093/bioinformatics/btq057
30. Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., ..., Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5), 503–510. doi:10.1038/nbt0710-756b.
31. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3), 562–78. doi:10.1038/nprot.2012.016.
32. Haas, B. J., Delcher, A. L., Mount S.M., S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., ... White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31(19), 5654–5666. doi:10.1093/nar/gkg770.

33. Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., ... Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, 9(1), R7. doi:10.1186/gb-2008-9-1-r7.
34. Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., ... Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1), 188–196. doi:10.1101/gr.6743907.
35. Allen, J. E., & Salzberg, S. L. (2005). JIGSAW: Integration of multiple sources of evidence for gene prediction. *Bioinformatics*, 21(18), 3596–3603. doi:10.1093/bioinformatics/bti609.
36. Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V, Roos, D. S., & Weinstock, G. M. (2007). Creating a honey bee consensus gene set. *Genome Biology*, 8(1), R13. doi:10.1186/gb-2007-8-1-r13.
37. Souvorov, A. et al. (2010) Gnomon — the NCBI eukaryotic gene prediction tool. *National Center for Biotechnology Information*.
38. Ma, L.J., Shea, T., Young, S., Zeng, Q., Kistler, H.C. (2014). Genome Sequence of *Fusarium oxysporum* f. sp. *melonis* Strain NRRL 26406, a Fungus Causing Wilt Disease on Melon. *Genome Announc.* Jul-Aug; 2(4): e00730-14.
39. Kwiatos, N., Ryngajllo, M., & Bielecki, S. (2015). Diversity of laccase-coding genes in *Fusarium oxysporum* genomes. *Frontiers in Microbiology*, 6(SEP). doi:10.3389/fmicb.2015.00933.
40. Rep, M., & Kistler, H. C. (2010). The genomic organization of plant pathogenicity in *Fusarium* species. *Current Opinion in Plant Biology*. doi:10.1016/j.pbi.2010.04.004.
41. Ma, L.-J., van der Does, H. C., Borkovich, K. a, Coleman, J. J., Daboussi, M.-J., Di Pietro, A., ... Rep, M. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*, 464(7287), 367–373. doi:10.1038/nature08850.
42. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., ... Wang, J. J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), 18. doi:10.1186/2047-217X-1-18.

43. Delcher, A. L., Salzberg, S. L., & Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.], Chapter 10*, Unit 10.3.doi:10.1002/0471250953.bi1003s00.
44. Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, 6(1). doi:10.1371/journal.pone.0016526.
45. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4), 462–467. doi:10.1159/000084979.
46. Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933–2935. doi:10.1093/bioinformatics/btt509.
47. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. (2003). Rfam: An RNA family database. *Nucleic Acids Research*. doi:10.1093/nar/gkg006.
48. Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9), 1859–1875. doi:10.1093/bioinformatics/bti310.
49. Keibler, E., & Brent, M. R. (2003). Eval: a software package for analysis of genome annotations. *BMC Bioinformatics*, 4(1), 50. doi:10.1186/1471-2105-4-50.
50. Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5, 59. doi:10.1186/1471-2105-5-59.
51. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, 34(WEB. SERV. ISS.). doi:10.1093/nar/gkl200.
52. Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, 42(15), e119. doi:10.1093/nar/gku557.

53. Borodovsky, M., & Lomsadze, A. (2011). Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinformatics, Chapter 4*, Unit 4 6 1–10. doi:10.1002/0471250953.bi0406s35.
54. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research*, 18(12), 1979–1990. doi:10.1101/gr.081612.108.
55. Guigò, R., Knudsen, S., Drake, N., & Smith, T. (1992). Prediction of gene structure. *Journal of Molecular Biology*, 226(1), 141–157. doi:10.1016/0022-2836(92)90130-C.
56. Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., ... Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, 196(3), 891–909. doi:10.1534/genetics.113.159996.
57. Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics, 2014*, 4.11.1–4.11.39. doi:10.1002/0471250953.bi0411s48.
58. Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., ... Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*. doi:10.1093/nar/gkt1223.
59. Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(SUPPL. 2). doi:10.1093/nar/gkr367.
60. Conesa, A., Gotz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18), 3674–3676. doi:10.1093/bioinformatics/bti610.
61. Lewis, S. E., Searle, S. M. J., Harris, N., Gibson, M., Lyer, V., Richter, J., ... Clamp, M. E. (2002). Apollo: a sequence annotation editor. *Genome Biology*, 3(12), RESEARCH0082. doi:10.1186/gb-2002-3-12-research0082.
62. Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., & Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics*. doi:10.1186/1471-2164-12-540.

63. Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., ... Carputo, D. (2015). The *Solanum commersonii* Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. *The Plant Cell*, 27(4), 954–68. doi:10.1105/tpc.114.135954.
64. Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & MacAs, J. (2013). RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29(6), 792–793. doi:10.1093/bioinformatics/btt054.
65. Wenke, T., Döbel, T., Sörensen, T. R., Junghans, H., Weisshaar, B., & Schmidt, T. (2011). Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *The Plant Cell*, 23(9), 3117–28. doi:10.1105/tpc.111.088682.
66. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with improved accuracy and speed. *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, (Csb), 3–4. doi:10.1109/CSB.2004.1332560.
67. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. doi:citeulike-article-id:11583827.
68. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 10. doi:10.14806/ej.17.1.200.
69. Gordon, A., & Hannon, G. J. (2010). Fastx-toolkit. FASTQ/A short-reads pre-processing tools. *Unpublished Http://hannonlab. Cshl. Edu/fastx_Toolkit*.
70. The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400), 635–41. doi:10.1038/nature11119.
71. Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., ... Visser, R. G. F. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355), 189–195. doi:10.1038/nature10158.
72. Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., ... Choi, D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics*, 46(3), 270–278. doi:10.1038/ng.2877.

73. Barchi, L., Lanteri, S., Portis, E., Stàgel, A., Valè, G., Toppino, L., & Rotino, G. L. (2010). Segregation distortion and linkage analysis in eggplant (*Solanum melongena* L.). *Genome / National Research Council Canada = G  nome / Conseil National de Recherches Canada*, 53(10), 805–15. doi:10.1139/g10-073.
74. Fukuoka, H., Miyatake, K., Nunome, T., Negoro, S., Shirasawa, K., Isobe, S., ... Ohyama, A. (2012). Development of gene-based markers and construction of an integrated linkage map in eggplant by using *Solanum* orthologous (SOL) gene sets. *Theoretical and Applied Genetics*, 125(1), 47–56. doi:10.1007/s00122-012-1815-9.
75. Doganlar, S., Frary, A., Daunay, M. C., Lester, R. N., & Tanksley, S. D. (2002). Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. *Genetics*, 161(4), 1713–1726. doi:12196413.
76. Wu, F., Eannetta, N. T., Xu, Y., & Tanksley, S. D. (2009). A detailed synteny map of the eggplant genome based on conserved ortholog set II (COSII) markers. *Theoretical and Applied Genetics*, 118(5), 927–935. doi:10.1007/s00122-008-0950-9.
77. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2), 240–248. doi:10.1101/gr.5681207.
78. Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., ... Fukuoka, H. (2014). Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the old world. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 21(6), 649–60. doi:10.1093/dnares/dsu027.
79. Parra, G., Bradnam, K., & Korf, I. (2007). : A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9), 1061–1067. doi:10.1093/bioinformatics/btm071.
80. Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., ... Yandell, M. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, 164(2), 513–24. doi:10.1104/pp.113.230144.

81. Han, Y., & Wessler, S. R. (2010). MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, 38(22). doi:10.1093/nar/gkq862.
82. Gremme, G., Steinbiss, S., & Kurtz, S. (2013). Genome tools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3), 645–656. doi:10.1109/TCBB.2013.68.
83. Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9), 3100–3108. doi:10.1093/nar/gkm160.
84. Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8), 1086–1092. doi:10.1093/bioinformatics/bts094.
85. Nakasugi, K., Crowhurst, R., Bally, J., & Waterhouse, P. (2014). Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLoS ONE*, 9(3). doi:10.1371/journal.pone.0091776.
86. van Baren, M. J., Koebbe, B. C., & Brent, M. R. (2007). Using N-SCAN or TWINSKAN to predict gene structures in genomic DNA sequences. *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]*, Chapter 4(December), Unit 4.8. doi:10.1002/0471250953.bi0408s20.
87. Chaochun Wei. (2006) Using Expressed Sequence Tags to Improve Gene Structure Prediction. PhD thesis, Washington University in St. Louis.
88. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. doi:10.1093/bioinformatics/btv351.
89. Goodswen, S. J., Kennedy, P. J., & Ellis, J. T. (2012). Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. *PLoS ONE*, 7(11). doi:10.1371/journal.pone.0050609.

90. Campbell, M. S., & Yandell, M. (2015). An Introduction to Genome Annotation. *Current Protocols in Bioinformatics*. Editorial Board, Andreas D. Baxevanis [et Al.], 52, 4.1.1–4.1.17. doi:10.1002/0471250953.bi0401s52.

