A COMPUTATIONAL FRAMEWORK FOR TRANSCRIPTOME ASSEMBLY AND ANNOTATION IN NON-MODEL ORGANISMS: THE CASE OF *VENTURIA INAEQUALIS*

Name: Stanley Mbandi Kimbung

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor Philosophiae at the South African National Bioinformatics Institute, Department of Biotechnology, Faculty of Science, University of the Western Cape

Date submitted for examination: May 2014

Supervisor: **Professor Alan Christoffels**

Co-supervisor: **Professor D. Jasper G. Rees**

Keywords

RNA-Seq

Quality filtering

Transcriptome reconstruction

Transfrags

Coding potential

Comparative transcriptomics

Open reading frame

Host-pathogen interaction

Venturia inaequalis

Ortholog

Abstract

The genetic blueprint does not fully inform the range of possibilities that exists for a given species or the ways in which genes manifest in response to the environment. Transcriptome sequencing (RNA-seq) offers expression and nucleotide-level resolution of genes en masse without knowledge of the underlying genome. Recent advances in nucleotide sequencing technologies have reduced to a fraction of the cost and time required in procuring large amounts of sequence data. However, the characteristic diminutive sequence length and high base-calling errors rates of ultra-high throughput sequence data render exclusive assembly of RNA-seq computationally challenging. The paucity of genomic sequences for the apple scab pathogen (*Venturia inaequalis*) has limited the breeding of durable apple cultivars. Collaborative efforts over the past few years between the South African National Bioinformatics Institute and the South African Agricultural Research Council have culminated in sequencing the transcriptome and genome of *V. inaequalis*. This thesis is about the development of methodologies to interrogate RNA-Seq reads derived from non-model organisms. We implemented these methodologies to generate genomic information and describe a useful resource that will promote hypothesis-driven research into the biology of *V. inaequalis*.

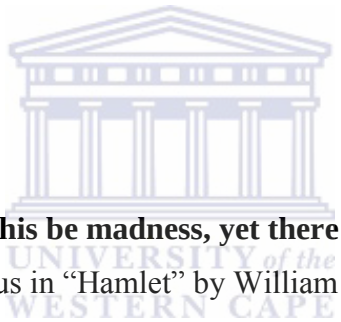
In this dissertation three computational approaches are presented that enable optimization of reference-free transcriptome reconstruction. The first addresses the selection of *bona fide* reconstructed transcribed fragments (transfrags) from *de novo* transcriptome assemblies and annotation with a multiple domain co-occurrence framework. We showed that selected transfrags are functionally relevant and represented over 94% of the information derived from annotation by transference. The second approach relates to quality score based RNA-seq sub-sampling and the description of a novel sequence similarity-derived metric for quality assessment of *de novo* transcriptome assemblies. A detail systematic analysis of the side effects induced by quality score based trimming and or filtering on artefact removal and transcriptome quality is describe. Aggressive trimming produced incomplete reconstructed and missing transfrags. This approach was applied in generating an

optimal transcriptome assembly for a South African isolate of *V. inaequalis*. The third approach deals with the computational partitioning of transfrags assembled from RNA-Seq of mixed host and pathogen reads. We used this strategy to correct a publicly available transcriptome assembly for *V. inaequalis* (Indian isolate). We binned 50% of the latter to Apple transfrags and identified putative immunity transcript models. Comparative transcriptomic analysis between fungi transfrags from the Indian and South African isolates reveal effectors or transcripts that may be expressed in planta upon morphogenic differentiation.

These studies have successfully identified *V. inaequalis* specific transfrags that can facilitate gene discovery. The unique access to an in-house draft genome assembly allowed us to provide preliminary description of genes that are implicated in pathogenesis. Gene prediction with *bona fide* transfrags produced 11,692 protein-coding genes. We identified two hydrophobin-like genes and six accessory genes of the melanin biosynthetic pathway that are implicated in the invasive action of the appressorium. The cazyome reveals an impressive repertoire of carbohydrate degrading enzymes and carbohydrate-binding modules amongst which are six polysaccharide lyases, and the largest number of carbohydrate esterases (twenty-eight) known in any fungus sequenced to date.

DECLARATION

I declare that "A computational framework for transcriptome assembly and annotation in non-model organisms: The case of *Venturia inaequalis*" is my own work, that it has not been submitted for any degree or examination in any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.



Though this be madness, yet there is method in't.

(Polonius in "Hamlet" by William Shakespeare)

Stanley Mbandi Kimbung May 2014

Acknowledgement

A considerable time has been put to the realization of this work which would not have come to fruition without the invaluable support of my mentors, colleagues and peers; directly or indirectly, whom I wish to acknowledge for making this thesis possible.

My utmost gratitude goes to my supervisor, Prof Alan Christoffels who has relentlessly invested valuable time to direct this work from conceptualization to implementation. Alan endowed me with enormous experience first as a workshop participant, to an intern and finally accepting me for PhD studies in his lab. I probably would not have done better if I did not consider your proposed area for my research. I look forward to engaging with you in many other projects in future.

I wish to thank Prof Jasper Rees and members of the Rees Laboratory (Dr Joseph Mafofo and PhD fellow, Lizex Hussslemann) for providing the raw data of *Venturia* on which this thesis is based.

Sincere gratitude also goes to Dr Uljana Hesse for the many interesting academic discourse. Your meticulousness and confidence in my work is greatly appreciated.

To the people whom I work with, thanks for putting up with me especially those in the Christoffels Lab.

I acknowledge the faculty members at SANBI for putting in place an enviable environment that nurtured me in various ways: Dr. Junaid Gamieldien, Dr. Gordon Harkins, Dr. Nicki Tiffin and Prof. Simon Travers.

Peter van Heusden (PVH), Mario Jonas, Long Yi and Dale Gibbs; for your superb technical support. PVH, I acknowledge in particular your implementation of my PERL substring matching algorithm in PYTHON with suffix tree and for replying to my emails after midnights and on holidays.

The bioinformatics community is also acknowledged especially those on SEQANSWERS, BIOSTARS and the assembly algorithm mailing lists for answering many of my queries. Thank you for your understanding especially when trivial questions came from me.

I wish to acknowledge the assistance rendered to me in various ways by the non-academic staff of SANBI: Junita Williams, Ferial Mullins, Maryam Salie, Fungiwe Mpithi and Samantha Alexander.

Lastly to my family and friends, here is your opportunity to understand what I have been up to during the last years in South Africa. Your incessant love, unwavering support and encouragement were the reason for me to remain attached to this academic venture.

This work was supported by the South African Research Chair Initiative of the Department of Science and Technology and National Research Foundation (NRF).

Dedication

To my parents who have graced my presence and my teachers for starting me on this crazy journey.

Table of contents

List of abbreviations

EST	Expressed Sequence Tag
DNA	Deoxyribonucleic acid
NGS	Next generation sequencing
RNA-Seq	Ribonucleic acid sequencing
TF	Tranfrags
NR	Non-redundant
E-value	Expectation value
BLAST	Basic local alignment search tool
ORF	Open Reading Frame
cDNA	Complementary Deoxy Ribonuceic Acid
CEG	Core eukaryotic gene
Asp	Aspartic acid
Cys	Cysteine
Glu	Glutamic acid
Ser	Serine
Thr	Threonine
WGS	Whole-genome shortgun

List of tables

Title **Page**

List of figures

Title **Page**

CHAPTER 1

Introduction: RNA-Seq in non-model organisms

1.1 Overview

The genetic blueprint provided by the genome sequence, does not fully inform the range of possibilities that exists for a given species or the ways in which these components will manifest in response to the environment. Whole transcriptome analysis by complimentary DNA (cDNA) or expressed sequence tag (EST) library sequencing, offers the possibility of interrogating genes and their expression en masse without knowledge of their underlying genomes. As such, the focus of genomic research has drifted towards the functional elements and transcribed regions (Harbers and Carninci, 2005). The latter represents the full set of transcripts (transcriptome): including large and small RNAs, splicing isoforms, gene-fusion transcripts and novel transcripts from unannotated genes (Martin and Wang, 2011). Transcriptome sequencing is largely a valid alternative for functional genomics due to the high functional information content (Miller et al., 2010). Transcriptomics aims at delineating the molecular phenotypes of transcripts (mRNA, rRNA, tRNA); transcriptional models of genes in relation to their translational start sites (UTRs, 5′ and 3′ ends, splicing patterns and PolyA signalling sites), the changing expression levels of each transcript under different physiological states and to understand the architecture of quantitative traits (Costa et al., 2010; Wang et al., 2009).

A number of technologies referred to as “next-generation” sequencing (NGS) that offer dramatic increases in cost-effective sequence throughput, albeit with shorter read lengths have emerged to curb the limitation faced with classical sequencing methods (Morozova and Marra, 2008; Mardis, 2009). Their application in sequencing cDNA has been termed 'RNA-Seq' and they do not require traditional cloning as with EST sequencing (Wilhelm and Landry, 2009). In contrast to microarray methods, RNA-seq achieves base-pair-level resolution over a higher dynamic range of expression levels without the requirement of *a priori* knowledge of transcribed regions (Wang et al., 2009). Despite these unprecedented benefits, the substantial expansions in NGS sequencing have led to an avalanche of data and spawn the demand of computational resources. The relatively short read length and the intrinsic error rate that increases with read length presents with new

algorithm challenges (Clifton and Mitreva, 2009). The plummeting cost per base in acquiring raw sequence data has somewhat 'democratize' large-scale sequencing. Transcriptome or genome sequencing is no longer a consortium-driven engagement or limited to model organisms (Hutchison, 2007; Ekblom and Galindo, 2011). For this reasons, the number of sequenced transcriptomes in non-model organism has grown by leaps and bounds (Ekblom and Galindo, 2011).

In agricultural genomic, NGS has lent itself very well to providing the potential for transcriptomic/genomic data to shape Rosaceae molecular bioscience research towards sustainable breedomics (Shulaev et al., 2008). Central to this are efforts taken by the Rosaceae community to develop genomics tools, identifying markers that segregate with resistance and improved fruit quality traits (Chagné et al., 2012; Gusberti et al., 2013). Very little has been done to establish genome resources for associated fungi pathogens of Rosaseae despite the inescapable economic realities encountered in producing saleable fruits. The genus *Venturia* is a member of the Dothideomycetes with many species that cause plant diseases. Apple scab, caused by *Venturia inaequalis* is arguably the most economically important pathogen of Rosaceae, infecting members of Maloideae worldwide and is the most studied pathosystem in woody plant species (Bowen et al., 2011). The apple genome is expected to facilitate the identification and manipulation of resistance genes present in the *Malus* germplasm that could potentially be marshalled to confer durable resistance against disease and selection of improved varieties (Velasco et al., 2010). However the availability of a genomic resource for *Venturia* is recourse for understanding the *Venturia-Malus* pathosystems and a prerequisite for effectively manipulating these host resistance factors towards rational design and implementation of control strategies.

In the following section, an abridged overview on the background of DNA sequencing technology with special focus on RNA-seq analysis and associated informatics challenges is highlighted. Insights are made to introduce novel analytical perspectives in RNA-Seq pre-assembly and post-assembly processing in non-model species. Their applicability is demonstrated in establishing a genomic resource for *V. inaequals* (South Africa isolate). The final section of this

chapter outlines the specific aims that are being addressed in order to achieve these goals.

1.2 Transcriptome sequencing in context

Transcriptome sequencing is an efficient means to generate functional genomic level data for non-model organisms in an era where whole genome sequencing is still largely impractical for most eukaryotes (Parchman et al., 2010). Organisms for which we have accumulated knowledge about their biology or ecology are not often those with established genomic resources. This is a problem particular for research groups that would like to infer or explain phenotypic traits with underlying sequence variation (Fraser et al., 2011). Transcriptome sequencing is essentially DNA sequencing applied to sequencing RNA. In principle therefore, any sequencing technology ever developed, can readily be used in determining the primary sequence of RNA via reverse transcription of purified mRNA with oligo(dT) and random priming. In non-model organisms, *de novo* transcripome assembly is simplified by the fact that a large portion of a eukaryotic genome is non-coding DNA and transcribed genes contain fewer repetitive elements, no introns and intragenic regions that would otherwise render analysis more difficult (Bouck and Vision, 2007; Fraser et al., 2011). The landmark publication of Sanger and colleagues (Sanger et al., 1977) sparked decades long sequence-driven research that followed. Traditionally, EST sequencing has been the core method for revealing the complex transcriptional repertoire. Several variations of the chain terminator method (Sanger et al., 1977) have been applied to infer the primary sequence of nucleotides. Since then, large EST catalogues have been very useful in comparative and phylogenomics analysis (Nagaraj et al., 2006), reliable identification of large numbers of candidate SNPs (Buetow et al., 1999), peptide identification and proteome scale characterization (Lisacek et al., 2001), gene discovery and annotation (Lizotte-Waniewski et al., 2000). Imposed partly due to historical sequencing capacity, this method is low throughput, has inherent quantitative limitations (Mortazavi et al., 2008; Wang et al., 2009), is prohibitively expensive for deployment on a large scale, error prone (Aaronson et

al., 1996), and labour intensive and more crucially will only detect the more abundant transcripts. ESTs from Sanger sequencing often account for only 60% of an organism's genes (Burke et al., 1999; Bentley, 2006).

Sanger sequencing-driven quantitative techniques for gene expression profiling have largely been performed with serial analyses of gene expression (SAGE) in which multiple copies of known gene segments 'short sequence tag', are concatenated and sequenced (Velculescu et al., 1995). However, a large proportion of the SAGE tags is located in the 3'-UTR and would require a fully sequenced genome and reliable gene annotation (Palmieri and Schlötterer, 2009). Hitherto, mircroarrays have dominated the field of expression profiling (Tseng et al., 2012). Due to cross-hybridization and complex hybridization behaviour and kinetics, they are not well suited for measuring absolute expression levels (Palmieri and Schlötterer, 2009). In addition, knowledge of the genes being examined is required for probe design. With microarray based expression analyses and chain termination sequencing falling short with limitations, RNA-Seq is the first sequencing-based method that allows the entire transcriptome to be surveyed simultaneously at base pair resolution in a very high-throughput and quantitative manner (Wang et al., 2009). By all accounts, DNA sequencing is ascendant for transcriptome analysis, and microarrays are waning. A typical RNA-seq experiment may require several steps depending on the particular application. For simplicity, we organised them in these categories: data generation and analysis (Martin and Wang, 2011).

1.2.1 Data generation

1.2.1.1 Library preparation

Data generation often begins with a library preparation step where random fragments from each transcript are generated, repaired and end-polished to make them amendable for sequencing. It may be necessary to preserve the information about which strand was originally transcribed. However, the polarity of the

transcript lends its usefulness to genome annotation in resolving overlapping genes or in a situation where the aim of *de novo* assembly is to estimate the expression levels of assembled fragments so that antisense transcripts would be easy to resolve (Parkhomchuk et al., 2009). RNA-seq libraries can also be normalised and or perform PolyA selection to enrich mRNAs. Although some mRNAs that lack a polyA tail will be missed, both methods will increase the representation of lowly expressed transcripts (Ekblom et al., 2012). A few recently adopted sequencing methodologies such as Helicos and Pacific Biosciences obviate the need for library preparation but suffer from high error rates (Ozsolak et al., 2010; Coupland et al., 2012). The choice of library preparation methods are many and may be invariably dictated by the choice of sequencing methodology. Fuelled by constraints in high reagent cost, limited design flexibility, and protocol complexity; attempts have been made recently to design cross-platform sequencing library preparation methods (Nguyen-Dumont et al., 2013). Reads of longer lengths are generally preferred for *de novo* transcriptome experiments as they greatly reduce the complexity of transcript reconstruction. In many instances, the choice of sequencing technology largely depends on the technology to which a user has access and the budget constraints for sequencing where sequencing is outsourced (Wang et al., 2009).

1.2.1.2 Next/Ultra high-throughput DNA sequencing

There are currently five alternative strategies used to investigate the base-composition of the nucleic acid content of a sample, in a highly parallelized fashion that apparently successfully addressed the limitations in Sanger sequencing to varying degrees. The depth of coverage, through-put and large amounts of DNA sequenced at very low cost has earned them the names: Next Generation Sequencing (NGS) or Ultra high-throughput sequencing (Morozova and Marra, 2008; Shendure and Ji, 2008). Cyclic-array sequencing is the most in use and according to market share analysis; about two thirds of all NGS instruments in current operation have been manufactured by Illumina (http://www.genomeweb.com). The frequently encountered NGS commercial

products include 454 sequencing (used in the 454 Genome Sequencers, Roche Applied Science; Basel), Solexa technology (used in the Illumina Genome Analyzer, San Diego), the SOLiD platform (Applied Biosystems; Foster City, CA, USA) and Ion Torrent Systems Inc now owned by Thermo Fisher Scientific. The available sequence technologies are however not limited to list below. A closer inspection of sequence accumulation suggests a logarithmic growth rate with inflection points that appear to correspond to technical innovations, many of which are expected in the future (Hutchison, 2007).

1.2.1.2.1 454 sequencing: pyrosequencing in high-density pico-wells

This technology was parallelized by 454 Life Sciences, and was the first used in a proof-of-concept in *de novo* assembly (Margulies et al., 2005). Arguably the most successful non-Sanger method developed to date and is the first NGS technology to provide a cost-effective, and reliable method for development of functional genomic tools for non-model species (Cheung et al., 2006; Vera et al., 2008). Sequencing is performed by the pyrosequencing method which measures the release of inorganic pyrophosphate via a coupled reaction (Ronaghi et al., 1996). The sequencing reaction takes place in a microfabricated array of picoliter-scale wells. Each well contains a bead on which adapter ligated DNA fragments are attached. The well dimensions preclude it from accommodating more than one bead which renders it compatible with the biochemistry of array-based sequencing. This also allows hundreds of thousands of independent pyrosequencing to operate, massively increasing the sequencing throughput. Emulsion PCR (ePCR) is initially carried out on each bead to generate multiple copies of each attached fragment. Solutions of dNTPs are added one at a time; the release of inorganic pyrophosphate (PPi) whenever a compliment nucleotide is incorporated is converted into visible light by a series of enzymatic reactions (requiring the combined action of ATP sulfurylase and luciferase). At the unetched base of the slide, the emitted photo is relayed by a fiber-optic bundle for charge-coupled device-based signal detection. The intensity of the emitted photon is proportional to the amount of PPi released and hence the number of bases

incorporated (Margulies et al., 2005). However, the signal strength for homopolymer stretches is linear only up to eight nulceotides which is an inherent problem in the resolution of low complexity sequences (Gilles et al., 2011). As a consequence, the dominant error type is insertion-deletion, rather than substitution. At the time of writing, the GS FLX+ (Droege and Hill, 2008) is capable of producing reads of up to 1kb in length. There is often a trade-off between read length and quantity. Since they produce fewer reads, they are particularly not well suited for gene expression analysis as a sequencing method of choice where a reference is available and cost is a limiting factor. For certain applications where long read-lengths are critical (e.g., *de novo* assembly and metagenomics), it may be the method of choice. According to GenomeWeb Daily News (http://www.genomeweb.com), the 454 sequencing technology will be phased out by 2016. 454 sequencing has little scalability and has not met up with the rapid improvement offered by competing technologies. Roche has invested with Pacific Biosciences to develop diagnostic products based on PacBio's Single Molecule Real Time technology.

1.2.1.2.2 Illumina/Solexa: reversible terminators chemistry with bridge amplification

The Illumina sequencing platform was originally commercialised as 'Solexa'. It is also a sequencing-by-synthesis NGS platform (Bentley et al., 2008). The technology behind is based on cyclic reversible termination (CRT). CRT uses a protecting group attached to the nucleotide that terminates DNA synthesis after incorporation by polymerase. Upon removal of the reversible terminator, the natural nucleotide substrate is restored allowing subsequent addition of reversible terminating nucleotides (Metzker, 2005). In Illumina platform, flow cells are used instead of microtitre plates. There can be up to 8 flow cells which can allow fragments of 8 libraries to be processed simultaneously. Primer sequences that are complimentary to adapter sequences in library derived fragments are tethered on the flow cell through a flexible linker so that any amplicon that arise during the bridge PCR is clustered to a single physical location. Clonal amplification can

produce up to 40 million clusters each containing ~1,000 clonal amplicons. With this design, all four nucleotides (tagged with a fluorescent moiety, each with a unique colour to distinguish among the bases) are introduced at the same time through the flow cell. The sequence of the template is deduced by reading off the colour at each successive nucleotide addition step.

Although the Illumina approach is more effective at sequencing homopolymeric stretches due to the use of modified polymerase and dye terminator nucleotides reversible terminators, more base-substitution errors are observed (Hutchison, 2007). The use of reversible dye terminator nucleotides also places constraints on efficient base incorporation which leads to shorter read length. However, they are able to produce paired end reads of 300 bases (2 x 300). Incomplete cleavage of fluorescent labels or terminating moieties introduces signal decay and dephasing with the result that Illumina reads display a considerable drop in read quality towards the 3'-end (Shendure and Ji, 2008). The Illumina platform has the most prolific through-put of all NGS technologies. This stems from the very high density of clusters that can reportedly produce > 1 billion bases in a single run. This make the Illumina technology ideal for genome re-sequencing and expression profiling. By the time of writing, Illumina had just unveiled the HiSeqX Ten Sequencing System with a 10-fold increase in daily throughput over its earlier machines.

1.2.1.2.3 ABI/SOLiD: sequencing by ligation

The SOLiD sequencing platform from Applied Biosystems (now branded by Thermo Fisher Scientific) is the third commercially successful NGS technology. The sequencing biochemistries rely on the discriminatory capacities of polymerases (ePCR) and enzymatic ligation reaction of anchored primer to a population of degenerate octamers that are labelled with fluorescent dyes (Shendure et al., 2005). On the SOLiD instrument, clonal sequencing features are generated by ePCR of single-molecule on the surface of 1-μM paramagnetic primer tethered beads similar to that used in the 454 technique. The amplification

products are selectively immobilized to a solid planar glass surface after breaking the emulsion. The sequence interrogation is done through repeated cycles of hybridization of a mixture of sequencing primers and fluorescently labelled probes, rather than a polymerase. The universal primer is hybridized to the array of a complementary adaptor on amplicon-bearing beads. Each octamer is structured such that the identity of one of its two central positions which allow a di-base encoding system is correlated with the identity of the fluorophore. When octamers anneal, ligation follows during which time image capture is performed across four channels. Thereafter, the octamer is chemically excised between positions 5 and 6, removing the fluorescent label. As such, the next round of hybridization-ligation can only interrogate a 5th base downstream. The process is repeated, at the end of which the entire system is re-set by denaturation. The entire process is repeated with a new set of primesr complementary to the n-1 position for a second round of ligation cycles. Five rounds of primer reset are completed for each sequence tag. In this system, each position is effectively probed twice which can increase the precision when calling biological variations over sequencing errors.

1.2.1.2.4 Pacific biosciences: polymerase active-site monitoring

Pacific biosciences, Inc., (PacBio) have developed a procedure that combines nanotechnology with molecular biology and highly sensitive fluorescence detection to achieve single-molecule DNA sequencing for sequencing based on single molecule real time (SMRT^{TM}). DNA polymerase molecules, bound to a DNA template at the bottom of a silicon wafer well as a regular array that may contain tens to thousands of nanophotonic structures known as zero-mode waveguides (ZMWs) (Eid et al., 2009). Here, the fluorophore is linked to the terminal phosphate moiety (phospholinked) and not base-linked with the advantage that the release of the fluorophore generated natural, unmodified DNA. The width of the ZMW is opaque to light, but energy can penetrate through a short distance and excite the fluorophores attached to those nucleotides in the vicinity of the polymerase. This format allows a single-molecule detection of

fluorescently labelled nucleotide incorporation events in real time (Quail et al., 2012; Mardis, 2013). Recently studies have revealed a difference in incorporation kinetics with DNA modified bases. This makes it a promising technology for methylome studies (Murray et al., 2012; Song et al., 2012). The read lengths obtained can be quite long (up to 10,000 nucleotides at maximum). SMRT sequencing has an overall higher error rate in single-molecule sequencing due to a small fraction of nucleotides that escape fluorescent labelling or dwell longer. Insertion and deletion errors predominate but with a few substitution errors and error rate can reach approximately 15% (Mardis, 2013).

1.2.1.2.5 Heliscope Single Molecule Sequencer (Helicos Biosciences)

The Helicos sequencer is another commercially available technology that relies on cyclic interrogation of single-molecule. It differs from PacBio in that a highly sensitive fluorescence detection system is used to directly interrogate single DNA molecules via sequencing by synthesis without any prior clonal amplification step (Harris et al., 2008). An advantage of this is that uncontrolled bias in template representation from amplification efficiencies is eliminated. A disorder array of surface-tethered poly-T oligomers are hybridised to template DNA from randomly generated library fragments. Floursecently labelled nucleotides are added one after another resulting in template dependent extension. Chemical cleavage and release of the fluorescent label permits subsequent cycles of extension and imaging (Shendure and Ji, 2008). In the absence of a termination moiety homopolymeric template DNA is an important issue.

1.2.1.2.6 Ion torrent: pH change monitoring

This sequencing approach is amongst the latest cyclic array-based sequencing instrument that was commercialized in 2010 by Ion Torrent and later acquired by Life Technologies™ Corp (now Thermo Fisher Scientific). A major difference between this platform and others is that the release of hydrogen ions serves as the cue for deciphering the incorporation of a complimentary base in a polymerase

reaction on a semiconductor-sensing device or ion chip (Rothberg et al., 2011). The format is akin to the microtitre plate of 454 sequencing but unlike it, no flours are used which obviates the needs for advance optical photon detection. Clonal sequencing features are generated by ePCR of adaptor-flanked fragments, attached to beads that are encapsulated in a 1:1 ratio in oils micelles containing the reactants. During sequencing, one side of the silicon chip serves as a microfluidic conduit to deliver the reactants needed for the sequencing reaction, whereas the other side is bonded to a hydrogen ion detector that translates a pulse of hydrogen ions from each well into a quantitative readout of nucleotide bases. Unlike reversible terminator sequencing, a wide range of read lengths are obtained from any sequencing run due to the use of native nucleotides with beads containing different DNA fragments (Mardis, 2013). The error model of Ion Torrent sequencing is defined largely by insertion or deletion errors that come from asynchronous nucleotide incorporation in regions of homopolymers.

1.2.2 NGS data analysis in non-model organism

The emergence of NGS has created a flurry of data that prohibits manual analysis. It is immediately invaluable to recognise the usefulness in storing, analysing and annotating huge data in a highly parallelized fashion, with auto/semi-automated or structured manual workflows (Nagaraj et al., 2007); while optimizing throughput and efficiency. Extensive computational strategies or work-flows have been developed to organize and analyse NGS data for gene discovery, transcript and single nucleotide polymorphism analysis as well as functional annotation of putative gene products. Work-flows allow easy experimentation and reproducibility of findings by third parties. To this end, Bioinformaticists have had to deal modestly and quickly by investing, developing and deploying effective data compression and processing algorithms as well as in cyberinfrastructure (Jones et al., 2012). The proliferation of tools and analysis pipelines for both commercial and/or non-commercial use, available for Linux, Windows and Machintosh has often created an embarrassing choice for optimal selection of tools for NGS data analysis. This is further compounded by the fact that some

proprietary tools have limited cross-applicability. Tools such a CLC-Bio (http://www.clcbio.com) offer better operability but at prohibitively procurable costs, while integrated platforms such as Blast2go (Conesa et al., 2005; Götz et al., 2008) and Galaxy (Giardine et al., 2005) are open source but are limited by their dependence on internet connectivity to interrogate remote databases and data transferability respectively. The urge to develop novel pipelines continues to prevail as existing ones do not fully capture the semantics and parameters (Hoon et al., 2003) which depend on the highly changing field of NGS analysis. Tools that obviate the technical requirements on programming competences and big computational infrastructure are likely to be of interest to smaller research groups (Cantacessi et al., 2010).

Until recently, NGS reads were considered too short and restricted to resequencing the genome and elucidating the genetic variation between cells or individuals of a species or organisms with closely related genomes. This unavoidably necessitates the availability of a reference which is not always available or incomplete at best (Paszkiewicz and Studholme, 2010). The suitability of NGS in *de novo* assembly of genomes has been addressed with recent algorithmic (Butler et al., 2008) and experimental advances (mate-pair, short-read paired-ends libraries, moderate read length increment), especially in combination with long-read technologies (Maher et al., 2009). A plethora of *de novo* assembly tools have been developed for NGS which are based on graph theory and a number of attempts have been made to extend their usage in assembling transcriptomes (Paszkiewicz and Studholme, 2010). De bruijn graph based assemblers have remained the most popular and provide tractable representation of each read as a *k*-mer (Compeau et al., 2011). *De novo* genome assemblers implicitly assumed the target has been chosen uniformly at random such that the coverage is even and explore sequence depth to resolve low complexity regions as well as to compute optimal set of parameters for sequence linearity, which will probably treat highly expressed transcripts as repetitive sequences and favour the assembly of only a subset of transcripts (Martin and Wang, 2011).

Computational analysis of NGS data involves one or more of the following tasks: (i) alignment of sequence reads to a reference; (ii) *de novo* assembly from paired or unpaired reads which may be followed by alignment; (iii) base-calling and/or polymorphism detection; and (iv) genome browsing and/or annotation (Shendure and Ji, 2008). While transcriptome assembly at first glance may appear less cumbersome due to lack of low complexity regions (Paszkiewicz and Studholme, 2010), exclusive assembly of short reads into transfrags (assembly-derived transcribed fragments), possesses enormous computational and algorithmic challenges. A typical transcriptome assembly work-flow can be divided into three main steps: pre-assembly, assembly and post-assembly (El-Metwally et al., 2013). Each step offers the opportunity for development of novel algorithms that increase or improve the wider applicability of NGS technologies.

1.2.2.1 Pre-assembly (preprocessing)

NGS platforms have a perceived limitation in higher base-call error rates and novel platform-specific artefacts which affects downstream application (Cox et al., 2010). Preprocessing algorithms have emerged in response to variability in raw sequence type (paired and unpaired reads), output (base or color space) and data quality. As a result many assembly projects include a pre-processing step where NGS data can either undergo trimming and or filtering or error correction. NGS technologies all rely on a complex interplay of chemical reaction, hardware and optical or ion gradient sensors and depending on the technology, these contribute unequally to the overall noise that must be considered when transforming base incorporation signal into a sequence of bases (Ewing and Green, 1998). NGS platforms differ in mechanistic details which cause stochastic failures in nucleotide incorporation or block removal, or incorporation of more than one nucleotide in a particular cycle that corresponds to a random base call (Erlich et al., 2008). The FASTQ format (**Figure 1.1**) has emerged as the *de facto* format for data exchange between tools.

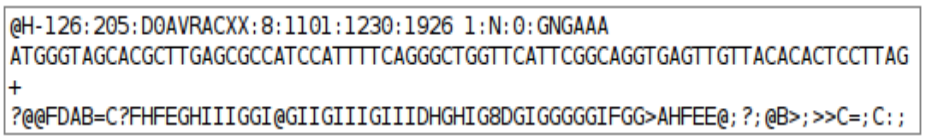


Figure 1.1 FASTQ read from DNA sequencing

There are four line types in the FASTQ format in the order: identifier and optional description, sequence line, a plus (+) or an optional repeat of title line, and ASCII quality line. Each base is associated with a ASCII-encoded quality character corresponding to a PHRED score (Cock et al., 2010). Each score represent the probability *p* that the corresponding base was called incorrectly by the following equation:

$$q = -10 \times \log_{10}(p)$$

Thus a base-call quality score of 30, has a 1/100 (0.001) chance of being incorrect.

Data acquisition is typically followed by: (i) parsing sequencer output; calculating and visualizing summary statistics on quality scores and nucleotide distributions; (ii) trimming and filtering reads if necessary by quality score and other various manipulations (Blankenberg et al., 2010).

1.2.2.1.1 General quality assessment

Although NGS platforms perform quality filtering, several sequence artefacts still remain in the dataset. Before data analysis, an overall assessment of quality is required which essentially provides a diagnosis of whether the data set is of high quality and whether it accurately represents the underlying biological sample. Deviation from what is expected indicates the presence of biases or artefacts in the sequence read data. The preprocessing strategy to embark on is decided on after quality assessment. The following metrics can be highly informative in identifying potential problems: composition per cycle, composition per read, duplication level, length distribution, and over-represented *k*-mer, over-

represented sequence, quality per cycle, quality per read, and quality per tile and adapter contamination (Zhou and Rokas, 2014). Many tools have been developed to perform quality assessment and important features to consider when selecting a tool include: Graphical interface, parallelization and support compressed input. A popular example with a plethora of features is FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and can be used in conjunction with quality filtering or trimming tools. Other tools such as NGS QC TOOLKIT (Patel and Jain, 2012) and Qtrim (Shrestha et al., 2014) perform both quality checks and filtering.

1.2.2.1.2 Quality score dependent preprocessing

Read trimming and or filtering algorithms rely on base quality scores to retain or discard an entire read or a portion of it. As such, they target errors that are likely represented by low quality scores and ambiguous bases (e.g N). The choice of a quality filtering tool depends on the NGS data type, portability and the sequence transformation features. Important features to consider include: visualization, simultaneous barcode and adapter processing, restoration of read pairings, alignment logging and demultiplexing and color-space support (Dodt et al., 2012). Reads can be trimmed in varying modes: ConDeTri (Smeds and Künstner, 2011) trims the reads from the 5'-end, 3'-end or both ends over a defined number of bases (window) or per base and tools such as FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) will retain or discard an entire read after assessing quality over a fraction of bases in the read. 5'-end and 3'-end trimming can be iterative over each base. Here, the ends of each read are interrogated and bases are excised until a base with a score above threshold is encountered. This is suitable when there are no bases upfront with a lower score after a high quality score base. The sliding window approach is particularly suitable for Roche 454 reads with fluctuations of high and low score bases, where tools such as Qtrim (Shrestha et al., 2014) maximizes read length with low prevalence of low quality bases. The majority of these tools perform blunt-end trimming. This approach is inefficient because it relies on selecting an arbitrary number of bases and treats all

reads as if they all have the same quality scores at their end. The advantage of having multiple read processing features in one algorithm eliminates the need for time-consuming file operations since independent steps can be carried out concurrently.

Consensus between independent reads improves the accuracy of transcriptome analyses. However, use of machine-assigned quality scores on next generation platforms does not necessarily correlate with accuracy (Eren et al., 2013). It is difficult to differentiate between biological variations and sequencing error using quality score without a reference, since error becomes dominant with volume of sequence data (Conway and Bromage, 2011). The literature is replete with published *de novo* transcriptome assemblies that have typically used a quality score threshold of 20 (Zhao et al., 2011; Li et al., 2012; Zhang et al., 2013). Other studies have used quality score below 20 (Q20) such as: Q13 (Duan et al., 2012; Haznedaroglu et al., 2012), and Q18 (Parchman et al., 2010). Different quality score thresholds can be applied on the 3'-end (Q25) and 5'-end (Q10 for 5 bases and Q20 for 10 bases) simultaneously (Chiara et al., 2013). The lesson drawn from these studies is that the implementation of quality score threshold is entirely subjective. Quality score based trimming leads to significant loss of data (Le et al., 2013). For *de novo* transcriptome assembly, there is an optimal number of reads balancing coverage and errors (Francis et al., 2013). The trimming or filtering stringency will determine the quantity of reads available for assembly without guaranteeing an accurate basecall. For each expression level there is a spectrum of parameters (typically *k*-mer) for optimal transcript assembly (Schulz et al., 2012). The effects of quality score based trimming have not been systematically addressed in the context of transcript assembly. Prior to analyses in this thesis, the only published attempt made use of a genome assembler which assumes uniform coverage and is not suitably optimized for *de novo* transcriptome assembly (Garg et al., 2011). There is no consensus on quality filtering/trimming thresholds since the quality score distribution is non-uniform across samples and the technologies for sequencing are constantly evolving. This is addressed in detail in chapter two and can be used in conjunction with an appropriate metric for

transcriptome quality assessment in defining optimal quality score filtering conditions.

1.2.2.1.3 Error correction

A very attractive alternative to quality score dependent filtering that mitigates bulk loss of data is to revert the mistakes made by a sequencing platform. Error correction algorithms have largely been applied on genomic reads. Since each base is usually sampled many times, reads with the correct value will prevail. However, reads from repetitive sequences or non-uniform sampling of the genome appear too often leading to multiple equal correction choices (El-Metwally et al., 2013). The following approaches have been explored in error correction: suffix tree, multiple sequence alignment and *k*-mer spectrum.

Suffix tree/array-based approach: Variable *k*-mers (substrings) are derived from each read representing various suffixes. All suffixes from the reads and their frequencies are represented as branch points (nodes) in a tree-like data structure known as a suffix tree/array. By traversing the tree, nodes with frequencies below a specified threshold are considered erroneous. The suffix tree or array–based filters attempts to find the most similar neighbouring node that will serve as the target candidate for updating erroneous bases. The HiTEC algorithm uses a thorough statistical analysis of the suffix array built on the string of all reads and their reverse complements (Ilie et al., 2011).

Multiple sequence alignment-based approach (MSA): All *k*-mers that occur in the reads or their reverse complements are indexed in a hash table which associates a *k*-mer with a list of reads where it occurs. The index structure is kept in main memory and is repeatedly accessed to answer queries like "given a *k*-mer, get the reads containing this *k*-mer". The idea here is to take reads that share common *k*-mers and build a multiple sequence alignment. The first read is set as the initial consensus of the alignment and reads are then aligned against the consensus using a variant of the Needleman–Wunsch algorithm (Needleman and

Wunsch, 1970). As such, gaps are inserted to accommodate subsequent reads which are updated with the most prevalent base in the consensus. An implementation of MSA as in the case of CORAL (Salmela and Schröder, 2011) is particularly useful when the reads are very similar. For very divergent reads, the return multiple alignment might not be very good.

***K*-spectrum–based** error correction: Algorithms in this category generate a catalogue of all possible *k*-mers from each read. Each catalogue of *k*-mer, known as *k*-mer spectrum is compared to one another. Those with small differences are likely from reads that originate from the same genomic locus. The similarity between *k*-mers is estimated from a weighted value assigned to each *k*-mer depending on its frequency and the quality scores of the bases in the *k*-mers. Reads with *k*-mers below a threshold (cutoff point) are considered untrusted *k*-mers and are edited against a target candidate with the smallest hamming distance as in REPTILE (Yang et al., 2010).

All the above-mentioned algorithms apply the same general assumption that reads with errors are less frequent and random and can be detected by counting the reads. In a transcriptome assembly scenario, these multiple options could represent genuine transcript variants with unequal expression levels which may be disproportionately collapsed. Error correction maximizes the quantity of reads for downstream analyses but may reinforce errors and eliminate genuine reads with low frequency *k*-mer (Martin and Wang, 2011). This limits the wider applicability of error correction algorithms in preprocessing reads prior to *de novo* RNA-Seq assembly. Recently, MacManes and Eisen, (2013) implemented error correction algorithms on RNA-Seq data where REPTILE (Yang et al., 2010) performed best with *de novo* transcriptome assembly of error corrected reads. However, the effects of coverage on error correction and that of error correction in enumerating splice variants remains an outstanding question.

The only known error correction algorithm specifically design for RNA-Seq data is SEECER, SEquencing Error CorrEction in Rna-seq (Le et al., 2013). Like those

design for error correction of genomic reads, it also makes use of *k*-mers to establish a hash table which it rapidly queries *k*-mers to build contigs by selecting (without replacement) a random read. It uses spectral clustering and relaxation of *k*-means to find related clusters with coherent subsets of reads. These reads are then used as the initial set for training hundreds of thousands of HMMs and uses these to correct sequencing errors. Despite its power to address and handle overlapping effects of non-uniform abundance, polymorphisms and alternative splicing, it performed poorly when compared to REPTILE in the study of MacManes and Eisen, (2013). However, the authors suggested that a lower coverage in their study could have resulted in this difference in performance.

A very promising algorithm worthy of mention performs filtering independent of quality scores but would not correct the reads. Diginorm examines *k*-mer abundance distribution within individual reads and builds many small de Bruijn graphs from them. By setting a coverage threshold past which reads can no longer be collected, it discards both data and errors. The net effect is digital normalization (Brown et al., 2012).

1.2.2.2 *De novo* assembly: graph construction process and simplification

De novo transcriptome assembly is a process whereby individual sequence reads are pieced together to form long contiguous sequences. The reconstructed sequences are technically referred to as transfrags and they share the same nucleotide sequence as the original template mRNA. Due to the perceived large volume of data, short read length, and different error rate than capillary sequencing, tools have been specifically design for analysing NGS type data. As with Sanger sequencing, the assembly task is relegated to computer algorithms (Staden, 1979). Transcriptome assembly offers a different type of informatics challenge which makes genome assemblers unsuitable for assembly of RNA-Seq data. Firstly, sequencing depth is used as a proxy by genome assemblers to distinguish low complexity regions, a feature that would flag highly expressed transcripts as repetitive. In addition, coverage across the genome is expected to be

uniform and because the expression levels of transcripts are uneven, assembly of certain transcripts will be favoured. Reads are generally short and may not represent entire exons, making it difficult to track which splice variant or isoform they emerge from (Martin and Wang, 2011).

The transcriptome assembly problem has a mathematical formalism typically requiring a graph data structure. In the field of graph theory, an assembly is a graph reduction problem that is represented as a hierarchical data structure which maps the sequence data to a putative reconstruction of the target (Miller et al., 2010). Transcriptome assemblers differ in their initial graph construction process, configuration, traversing, and simplification (El-Metwally et al., 2013). In this thesis, I have grouped transcriptome assemblers depending on their approach to graph construction: overlap graphs and de Bruijn graphs. It is important to note that prior to the introduction of *de novo* transcriptome assemblers, genome assemblers were used to perform RNA-Seq assembly (Strickler et al., 2012). The examples in the subsequent section are restricted to algorithms specifically designed to address *de novo* fragment assembly of transcriptome data.

1.2.2.2.1 Overlap-layout consensus approach

Assemblers in this category take a common approach known as overlap-layout-consensus (OLC). The initial step to overlap-layout graph construction is to perform a computationally expensive all-versus-all comparison of each sequence read to the other. The *k*-mer content is used to select possible overlapping candidate reads (Miller et al., 2010). This objective is to produce a composite string that contains all the *k*-mers as substrings. The *k*-mer length, minimum overlap and percent identity in the overlapping region are important parameters that affect sensitivity in alignment discovery. The next step is to construct an overlap graph where nodes correspond to reads and edges encode the overlap is generated from all satisfactory pairwise comparisons. Manipulating the graph produces an approximate read layout. Paths through the graph are the potential transfrags that can be converted to sequence information.

Finding all pairwise overlap is computationally expensive, limiting the application of OLC on Illumina reads. Longer reads have sufficient characters to detect overlaps such as 454 technologies. As a result, tools for EST assembly of Sanger reads are used for assembly of NGS EST-like reads from 454 technologies. EST assembly algorithms such as CAP3 (Huang and Madan, 1999), MIRA (Chevreux et al., 2004) and its wrapper EST2ASSEMBLY (Papanicolaou et al., 2009) that apply OLC have performed particular well (Kumar and Blaxter, 2010). A recent review suggests that the NEWBLER (Margulies et al., 2005) algorithm which makes use of OLC has been customized for *de novo* transcriptome assembly.

1.2.2.2.2 De Bruijn graph approach

The most appealing and frequently used approach to transcriptome assembly is based on the de Bruijn graph. A de Bruijn graph does not necessarily store individual reads or their overlap. The graph is built with *k*-mers derived from the reads, which makes it attractive for handling the large number of short reads from NGS platforms. This formalism was inspired by the works of Leonhard Euler and thus has earned the name Eulerain approach (Compeau et al., 2011). Here, a *k*-mer is represented by a node while each edge represents k–1 overlap between the nodes. The graph can have a reverse formulation where edges correspond to *k*-mers and the nodes are represented by a *k*-mer suffix or prefix of size *k*-1. The graph itself stores only one occurrence of any given *k*-mer irrespective of its frequency. This has an advantage of reduced memory usage over using the raw data. This simplification allows linear time construction or constant time hash table look-up for the existing *k*-mer despite the vast majority of available *k*-mers. However, when the traversal count of each edge is known, the Eulerian path that visits each edge in the graph exactly once requires polynomial time (Pevzner et al., 2001).

Reducing reads to *k*-mers has a major drawback especially when handling long reads in that information about the reads is lost. This is further compounded by

sequencing errors which induce false positive and false negative overlaps in the graph. False overlaps increase the branching nodes in the graph (Miller et al., 2010). As such, space consumption is a pressing practical problem for assembly with de Bruijn graph-based algorithms. Various modifications of the de Bruijn graphs are actively being researched mostly for genome assemblers. One such approach is to use a probabilistic data structure (known as Bloom filter) that stores all the observed *k*-mers implicitly in memory with 4 bits per *k*-mer (Melsted and Pritchard, 2011). The probability of false positives increases with the number of elements inserted in the Bloom filter. Edges are implicitly deduced by querying the Bloom filter. This may return a false positive answer for an arbitrary node. To avoid bogus branching, a subset of false positives can be stored in a separate data structure that guides the bloom filter querying process (Chikhi and Rizk, 2013). Another memory-efficient approach is to use a small fraction of the observed *k*-mers (sparse *k*-mers) as nodes and the links between these nodes (Ye et al., 2012). An alternative to subset representation is to perform additional path navigation on each compressed node/edge pair using an extension of Burrows-Wheeler transform (Bowe et al., 2012).

As mentioned earlier, the process of assembling a transcriptome violates many of the assumptions on which assemblers developed for application on genomic DNA data rely. Since there is no single parameter set that can give the best results for all genes, a clever approach that mainly addresses the uneven expression levels has been to merge transcriptome assemblies generated by a genome assembler with varying *k*-mer sizes (Martin et al., 2010; Surget-Groba and Montoya-Burgos, 2010). This approach can also be considered as a post-assembly processing procedure which I discuss further in the next section. The majority of transcriptome assemblers that are based on the de Bruijn graph paradigm are an extension from genome assembly models (Paszkiewicz and Studholme, 2010). They include OASES (Schulz et al., 2012), Trans-ABySS (Robertson et al., 2010) and SOAPdenovo-Trans (Xie et al., 2014).

The only known implementation of a de Bruijn graph out of the realms of genome assembly is in TRINITY (Robertson et al., 2010). TRINITY deserved a separate attention because it implements a hybrid approach to transcript reconstruction that incorporates a greedy graph. TRINITY applies a scoring scheme to the graph structure based on the original read sequences that discard nonsensical paths (wrongly assembled transfrags). TRINITY is highly effective compared with alternative methods (Duan et al., 2012; Xu et al., 2012; Zhao et al., 2012; Pang et al., 2013). TRINITY is currently the top assembler in Alternative Splicing Challenge on the Dialogue for Reverse Engineering Assessments and Methods project (http://www.the-dream-project.org/result/alternative-splicing). As such, I have used it in all the transcriptome assembly endeavours in this thesis. TRINITY begins by generating transfrags using a greedy extension based on (*k*-1)-mer overlaps. Sufficient information about the most dominant isoform can be retrieved at this stage albeit only the unique portion. The next stage is to group the transfrags using the information from the raw reads. This process clusters together regions that have probably originated from alternatively spliced transcripts. Each cluster is used to generate independent de Bruijn graphs (ideally one graph per expressed gene). The final step traces the RNA-Seq reads through the graph and supported graph paths are used to enumerate sequences in a manner that reflects the original cDNA molecules (Grabherr et al., 2011; Haas et al., 2013).

1.2.2.3 Post-assembly: post-processing filtering

Studies on comparative assessment of transcriptome assemblers have strongly suggested that there is no best performing assembler (Duan et al., 2012; Clarke et al., 2013). As such, the choice of an assembler is subjective and very much dependent on a number of technical considerations such as computational speed and memory efficiency (Zhao et al., 2011). Furthermore, there is no clear consensus of what sequencing depth is adequate, vis-a-vis genome complexity that would contribute to optimal reconstruction (Francis et al., 2013). The challenge is more of how to approach an optimal solution based on available information on data. For example, while TRINITY was able to reconstruct more

authentic wheat transcripts than Trans-ABySS in one analysis where both were outperformed by OASES (Oono et al., 2013), the reverse was true with a change of parameters in another wheat transcriptome study (Duan et al., 2012). The general consensus from these studies is that multiple parameters should be attempted (particularly the *k*-mer size) during the initial assembly phase in order to reconstruct a broad range of sequences from the original transcriptome (transcript diversity). Intuitively, different transcriptome assemblers can be used and based on the performance measurement under consideration, one can be chosen (Sadamoto et al., 2012; Oono et al., 2013). However, these procedures are lengthy and will be limited by the aforementioned technical considerations. Alternatively, multiple assemblies from different assemblers can be combined to take advantage of the specific benefits each assembler has on a particular dataset (Tao et al., 2012; Thakur et al., 2013; Nakasugi et al., 2014). These approaches have their downside in that they often result in transcripts that do not represent authentic gene models or not represented in the genome (Strickler et al., 2012). Underpinning all of these approaches is redundancy between assemblies and between *k*-mers. Post-assembly strategies have aimed at reducing redundancy and producing longer transfrags. The following are typical strategies for post-assembly processing: *de novo* clustering, meta-assembly, dissimilar sequence clustering, locus specific clustering. *De novo* clustering in its simplest form will eliminate shorter transfrags which can be duplicates, substrings or based on a percentage of shared similarity between any pair of sequences without changing the sequence of the larger transfrags (Li and Godzik, 2006). This is suitabe for low complexity transcriptome assemblies particularly when generated with a larger *k*-mer (Surget-Groba and Montoya-Burgos, 2010). Meta-assembly on the other-hand merges assemblies from different *de novo* assembly platforms with another round of *de novo* assemblies (Feldmeyer et al., 2011). The initial longer transfrags (with EST-like lengths) generated from each *k*-mer are reassembled with an OLC assembler that can assemble EST sequences (Huang and Madan, 1999). In a slightly different scenario, the initial set of transfrags could be from different *k*-mers generated with OASES as in the Oases-Merge pipeline (Schulz et al., 2012).

1.2.2.3.1 Locus-specific (unigene-like) **clustering** clusters transfrags into secondary loci using CD-HIT-EST (Li and Godzik, 2006) which will eliminate duplicates and exact substrings. The longest representative of each locus is searched with BLASTx (Altschul et al., 1997) at high stringency against an appropriate section of a protein database. Query sequences with identical cross species profiles of best hit annotation are merged into "tertiary loci". Secondary loci without a BLAST hit are clustered using BLASTn searches of ESTs and cDNAs from the same combination of species. By aligning all tertiary transfrags per locus, the minimum identity and overlap thresholds for a subsequent final clustering step is estimated. Locus-specific transcript clusters (LSTC) are generated by an all-versus-all BLAST. A representative with the highest scoring BLASTx match against the original peptide database is selected. For unannotated LSTCs, the longest transcript model is chosen (Chiara et al., 2013).

1.2.2.3.2 Dissimilar sequence (DS) **clustering** also makes use of a protein database to scan the transfrag using BLASTx. However, the sequences are not merged on bases of similarity amongst themselves. It assumes that sequences that have the same blast hit belong to the different regions of a single gene (Gahlan et al., 2012). For each such cluster, the longest sequence with highest bit score is taken as the representative sequence. This would be particularly useful when a homology based approach cannot find similar regions between transfrags (Bhardwaj et al., 2013; Thakur et al., 2013).

Transcriptome assembly continues to be a significant challenge that would require novel heuristic approaches. All existing *de novo* assemblers tend to corrupt with increasing alternative splicing events (Chang et al., 2014). The success of post-assembly processing depends entirely on the quality of the initial transcript assemblies which in turn is affected by a wide range of assembly parameters. While post-assembly clustering or reassembly has proven very useful in recovering unique transfrags, they can lead to loss of unique functional annotation (Haznedaroglu et al., 2012). *De novo* clustering is based on common word

heuristics (Hazelhurst et al., 2008), that ignore the biological nature of assembled transcripts which can propagate chimeras (Sharov et al., 2005).

1.3 The scope and purpose of this thesis

The plummeting cost per base in procuring raw sequence data has dramatically changed our approach in studies of non-model organisms. Sequencing can be applied virtually to any biological phenomena. Access to genomic resources is no longer limited to organisms for which we have the best ecological knowledge (Fraser et al., 2011). In addition, the power of contemporary sequencing technologies can be harnessed by individual research groups to generate reliable genomic knowledgebases (Gibbons et al., 2009). These genomic resources are expected to promote hypothesis-driven research that fuels the bench work of our day-to-day operations in molecular biology and bioinformatics laboratories worldwide (Yandell and Ence, 2012).

Despite the immense potential offered by NGS technologies, the challenges of fragment assembly are enormous and further approaches that allow for an unbiased analysis of the transcriptome are required. A prerequisite to most of, if not all, downstream transcriptomic analyses in non-model organisms, such as differential expression and functional genomics require accurately assembled full length transcript models. The vast majority of computational analysis of NGS data has been thwarted towards assembly methods. In the absence of suitably longer reads, transcriptome assembly will continue to rely on approximate computation. For example, genes commonly use multiple start sites (Harbers and Carninci, 2005) and reads originate from both matured and incompletely spliced precursor RNA (Garber et al., 2011) that confers an inherent complexity on the transcriptome. Such complexity cannot be easily modelled and existing *de novo* assemblers tend to corrupt as a result of this (Chang et al., 2014).

Interestingly, emerging knowledge on *de novo* transcriptome assembly suggests that the pre-assembly and post-assembly stages are equally important avenues for

algorithm design and implementation. Recently, Francis et al., (2013) demonstrated the marginal gain in full-length transcript reconstruction beyond 60 million reads, suggesting that the quantity of read fed to assemblers is crucial. In a separate study, error correction which used to be relegated to updating genomic reads has recently been shown to be equally useful in error correction of RNA-Seq reads prior to assembly (MacManes and Eisen, 2013). *De novo* transcriptome assembly has been dramatically improved over the last few years, but existing challenges continue to suggest that additional heuristic approaches are required to harness the full potential of RNA-Seq in non-model organisms.

Access to genomic data is an important recourse to expedite efforts in the control of parasitic diseases. An attractive avenue for the application of NGS technologies is in apple scab. Apple scab is a disease of apples caused by *Ventuira inaequalis* (MacHardy et al., 2001). The pathogen has invaded almost every apple growing nation leading to significant losses in yields (Gladieux et al., 2008). Despite the widespread pernicious agronomic impact, genomic datasets are sparse, limiting the molecular biological research of this parasite to a few genes (Kucheryava et al., 2008; Bowen et al., 2009). In combination with the apple genome sequence released 4 years ago (Velasco et al., 2010) and the recently elaborated gene model (Bai et al., 2014), it is more invaluable than ever before that a similar resource for *V. inaequalis* will illuminate our understanding of the *Venturia-Malus* pathosystem and facilitate rational design and implementation of control strategies.

The profile of the South African apple market value chain is not free from the scourge of apple scab. South Africa was the fourth largest producer (16.3% in 2010) of apples in the southern hemisphere after Brazil, Chile and Argentina. It's a relatively small apple grower in terms of global hectares, but is a major volume exporter in global terms. The cost from fungicidal control, greatly impacts the economic profit of apple producers and the export-driven South African apple industry (http://www.daff.gov.za, 2011-12). Significant efforts are being made with funds from the Deciduous Fruit Producers Trust, Technology and Human

Resources for Industry Programme (THRIP) of the South African Department of Trade and Industry and the Research Chair in Health genomics (http://www.sanbi.ac.za/agri_genomics) for research in *V. inaequalis* (Celton et al., 2010). Since 2008, an isolate of *V. inaequalis* from the Experimental Farm of Bien Donné, Simondium, South Africa had been successfully maintained in culture and subsequently sampled for both genomic and RNA-seq reads using the Illumina platform. In collaboration with the South African National Bioinformatics Institute, the objective has been to establish a genomic resource that will be publicly accessible. Such datasets will be useful for comparative genomics of geographically distinct isolates worldwide and provide the basis for a comprehensive annotation of a newly sequenced *V. inequalis* genome.

In this thesis, we formulated the hypothesis that the generated short reads can be used to reconstruct transcript models sufficient enough for biological inferences on pathogen biology to be made. To achieve this, we start by developing a computational workflow for *de novo* transcriptome reconstruction in non-model organisms. In addition, we provided a novel perspective on the preprocessing analysis of Illumina RNA-Seq data. Both approaches are then applied to transcriptome reconstruction in *V. inaequalis*. We develop an iterative screening workflow for untangling transfrags assembled from mixed host and pathogen RNA-Seq reads. We demonstrated the usefulness of these methods in establishing a transcriptome assembly in context of the publicly available Apple genomic resources. Finally we demonstrated the usefulness of the *V. inaequalis* transfrags in genome annotation and provided preliminary insight in to pathogenesis.

1.4 Specific research aims

i. Implementation of coding potential assessment in post-assembly processing of assembly-derived transfrags.

ii. Explore quality score based filtering or trimming for pre-processing optimization of *de novo* transcriptome assembly.

iii. Implement optimal protein local alignments as a metric for evaluating transcriptome.

iv. Comparative transcriptomics of two geographical isolates of *Venturia inaequalis.*

v. Preliminary analysis of *V. inaequalis* draft genome assembly and identification of candidate genes involved in pathogenesis.

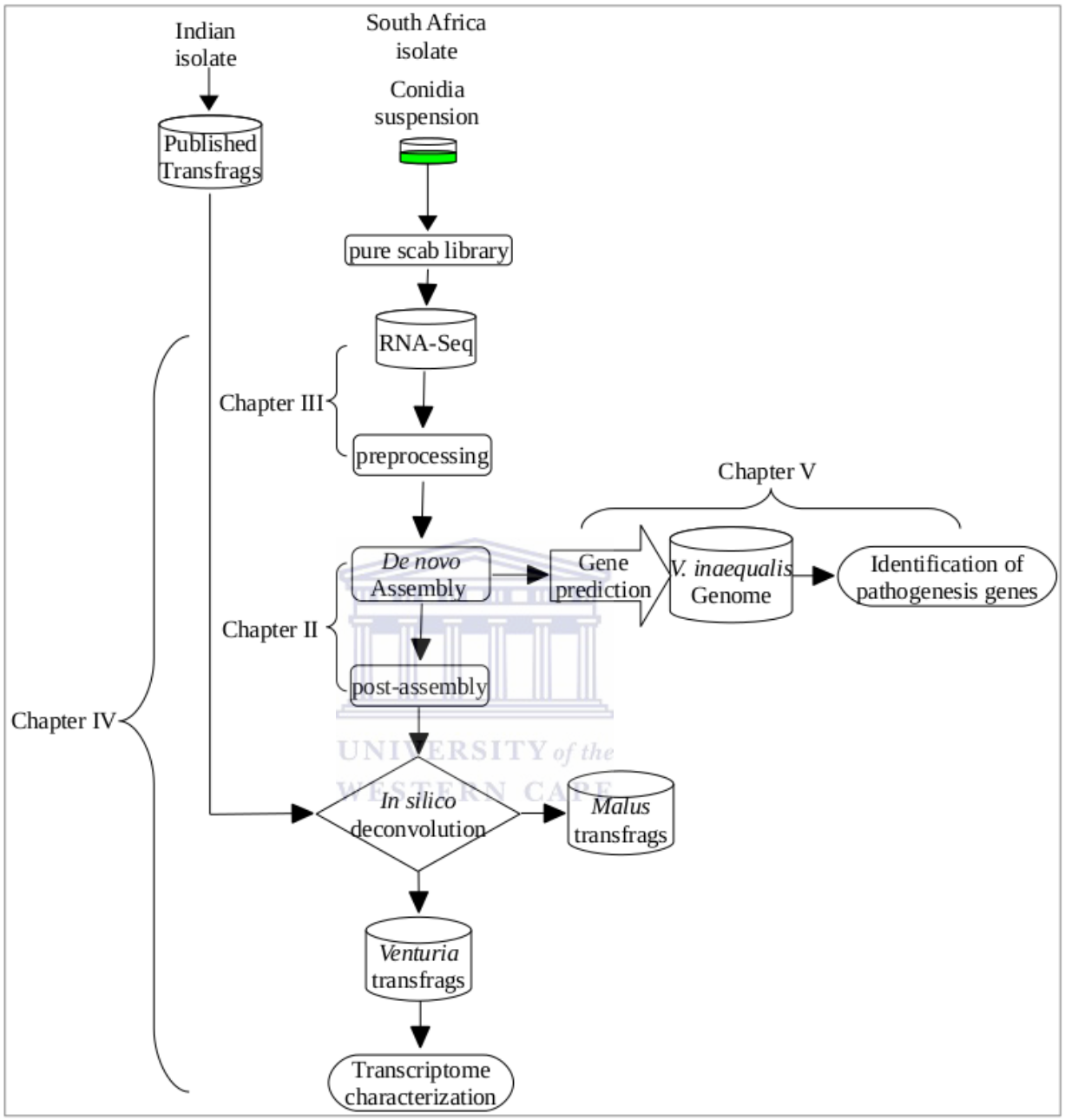


Figure 1.2 Schematic overview of the methods developed and used in this thesis. Sections are labled according to the thesis chapters.

1.5 List of publications

This thesis is organized into a series of journal manuscripts, published, in review or in preparation that are presented as independent chapters. Each chapter includes an introduction, results and or discussion and a reference section. The introduction and the conclusions follow the standard thesis/journal requirements.

Chapters and associated manuscripts

I. Mbandi SK, Hesse U, van Heusden P, Christoffels A. (2014). Inferring *bona fide* transcripts in RNA-Seq derived transcriptomes of non-model organisms. (*submitted* Front. Genet.)

II. Mbandi SK, Hesse U, Rees DJG and Christoffels A. (2014). A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. Front. Genet. 5:17. doi:10.3389/fgene.2014.00017

III. Mbandi SK, Husselmann L, Hesse U, Mafofo J, Rees DJG and Christoffels A. (2014). Identification of Venturia inaequalis putative effector candidates and apple resistance genes: a comparison of two trancriptomes. (*submitted* BMC genomics)

IV. Hesse U, Mafofo J, Mbandi SK, Oreetseng M, van Heusden P, Husselmann L, Rees DJG and Christoffels A. (2014). Genome of *Venturia inaequalis* – the causal agent of apple scab (*In preparation*)

Co-authorship statement for related manuscripts and publications

I. SKM design, performed all the experiments and analysed the data and wrote the manuscript. SKM and UH interpreted the results. UH provided intellectual assistance. PVH implemented the 'PERL redundancy removal' in PYTHON with suffix array. SMK and AC conceived the project. AC critically evaluated the manuscript, provided reagents, materials and supervision for the entire project.

II. SKM conceived and designed the experiments, performed all computational experiments, analysed the data and wrote the manuscript. UH facilitated transcript visualization. DJGR provided RNA-Seq data for *V. inaequalis*. AC provided reagents, materials and supervision.

III. SKM performed all computational experiments, analysed the data and wrote draft manuscript. SKM, UH and AC design the computational experiments. LH performed fungal culture establishment and RNA isolation. JM supervised the RNA isolation and RNA-seq. JR conceived and secured funding for *Venturia* RNA isolation experiments. AC critically evaluated the manuscript and coordinated this research.

IV. The analyses of chapter 5 form an integral part of the manuscript describing the genome of *V. inaequalis* in preparation.

CHAPTER 2

IFRAT: Inferring Functionally Relevant Assembly-derived Transcripts

2.0 Abstract

De novo transcriptome assembly of short transcribed fragments produced from sequencing-by-synthesis technologies often results in redundant datasets with differing levels of unassembled, partially assembled or miss-assembled transcripts. Post-assembly processing intended to reduce redundancy typically involves reassembly or clustering of assembled sequences. However, these methods are mostly based on common word heuristics that create clusters of biologically unrelated sequences, resulting in loss of unique transfrag annotations and propagation of miss-assemblies. Here, we propose a structured framework that consists of a few steps in pipeline architecture for Inferring Functionally Relevant Assembly-derived Transcripts (IFRAT). IFRAT combines 1) removal of identical subsequences, 2) error tolerant CDS prediction, 3) identification of coding potential, and 4) complements BLAST with a multiple domain architecture annotation that reduces non-specific domain annotation. We demonstrate, that independent of the assembler, IFRAT selects *bona fide* transfrags (with CDS and coding potential) for model organisms and non-model organisms (with no reference genome), without relying on post-assembly clustering or reassembly. We show that unselected transfrags mostly represent truncated sequences from intronic and untranslated (5' and 3') regions and non-coding gene loci. Therefore, IFRAT simplifies post-assembly processing providing a reference transcriptome enriched with functionally relevant assembly-derived transcripts.

2.1 Background

Whole transcriptome analysis using next generation sequencing (NGS) or sequencing-by-synthesis (SBS) technologies offers the possibility of interrogating genes and their expression en masse without knowledge of their underlying genomes. Transcriptome sequencing is often preferred over genome sequencing because of the reduced size of the sequence target space and the high functional information content (Pettersson et al., 2009; Martin and Wang, 2011). However, sequences generated from NGS platforms are often too short to represent entire protein-coding transcripts, and genomes for reference-guided transcriptome reconstruction are rare. De Bruijn graph assemblers allow *de novo* assembly of transcripts but represent only approximate computational solutions (Martin and Wang, 2011). The final assembly is one of many possibilities for which there is no universally accepted heuristic verification method; it is often highly redundant and contains miss-assemblies that are difficult to identify (Duan et al., 2012). Post-assembly processing intended to reduce redundancy typically involves reassembly or clustering of assembled sequences. This however may lead to propagation of miss-assemblies (Sharov et al., 2005) and assignment of sequences to unrelated gene clusters, resulting in loss of unique annotations (Haznedaroglu et al., 2012).

The main objective of transcriptome sequencing by synthesis is to ascribe functional labels to assembled transcribed fragments (transfrags). This is usually done via significant sequence similarity (Jones et al., 2005) or domain signature annotations (Quevillon et al., 2005). Similarity based approaches predominantly rely on transfer of functional labels of the best BLAST hits to the sequence in question (Conesa et al., 2005; Jones et al., 2005). However, low BLAST annotation coverage is often observed, in particular in transcriptomes of non-model organisms (Miller et al., 2012; Sun et al., 2010). The implementation of a significant BLAST hit as a proxy for functional annotation has further limitations: sequences that produce significant similarity may be functionally unrelated due to divergence (Koestler et al., 2010), low complexity sequences may produce high scoring hits but have no biological relationships (Mount, 2007), and functional homologs may lack sequence similarity (Galperin et al., 1998). Consequently, a

first large-scale assessment of protein function shows that BLAST is often ineffective at predicting functional labels (Radivojac et al., 2013). Domain based annotation methods (e.g. InterProScan) appreciate only presence/absence of domains. Given that domains seldom function in isolation (Vogel et al., 2004), a reliable approach should involve a method that recognises the overall domain co-occurrence architecture of the sequences under examination. Prerequisite for domain-based annotation is a reliable protein prediction method that tolerates sequencing errors and frame shifts.

Here, we introduce IFRAT, which allows for selection and annotation of functionally relevant transfrags (*bona fide*) without clustering. This is achieved through 1) removal of identical subsequences, 2) error tolerant CDS prediction, 3) identification of coding potential, and 4) complementation of BLAST with a multiple domain architecture annotation. The effectiveness and versatility of this approach is shown on published datasets from non-model organisms.

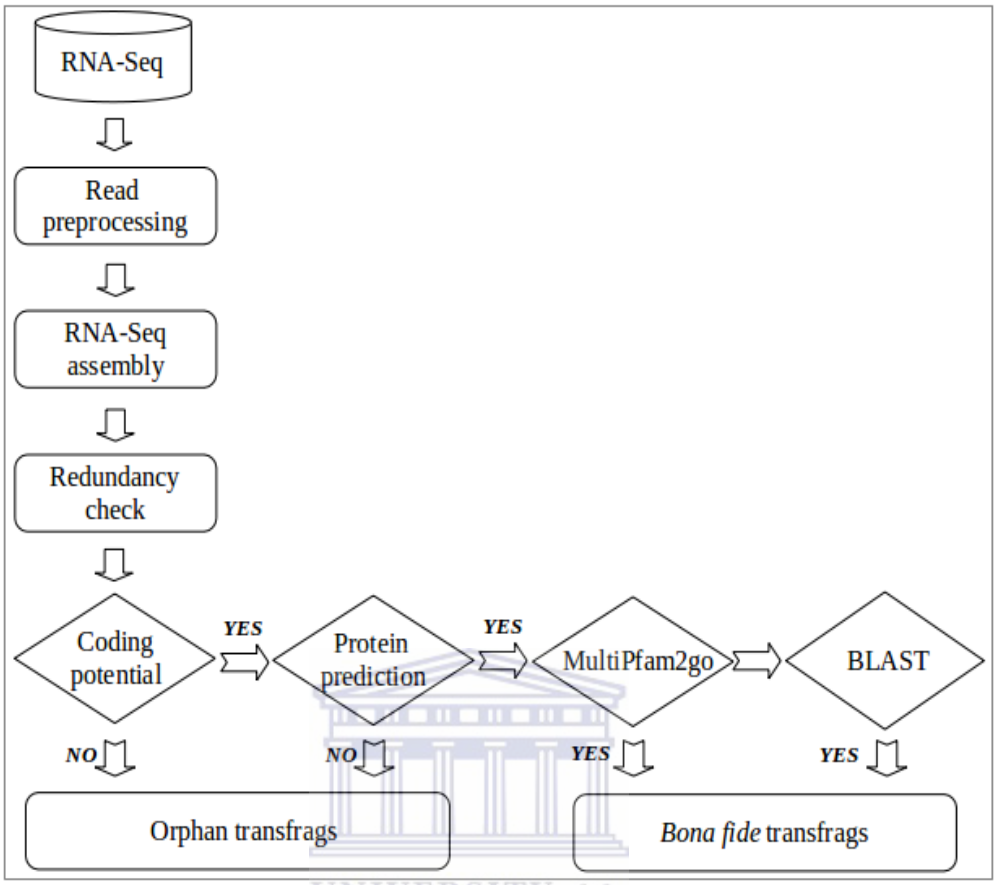


Figure 2.1 The basic portfolio of IFRAT pipeline.

Flow diagram to illustrate the method of integrating protein-coding potential and multiple domain functional annotation to infer bona-fide assembly derived-transcripts.

2.2 Material and methods

2.2.1 Datesets

To establish a robust work-flow for prioritizing and selecting functionally relevant (*bona fide*) transfrags, we selected the fungal plant pathogen *Neurospora crassa* (Galagan et al., 2003) as a species with a reference genome. Publicly available non-strand specific RNA-Seq data (SRR100067) from wild type *N. crassa* 74-OR23-1VA was obtained from the NCBI Sequence Read Archive (SRA,

http://www.ncbi.nlm.nih.gov/Traces/sra). The associated genomic and predicted coding sequences were obtained from the whole genome shotgun project (accession AABX00000000, http://www.ncbi.nlm.nih.gov/Traces/wgs/) and the FungiDB database (Stajich et al., 2011), respectively. We verified the pipeline in post-assembly processing of recently published transcriptomes of non-model organisms: buckwheat (*Fagopyrum esculentum*) (Logacheva et al., 2011); hydra (*Hydra vulgaris*) (Wenger and Galliot, 2013); fresh water snail (*Radix balthica*) (Feldmeyer et al., 2011); centipede (*Alipes grandidieri*), marine worm (*Cerebratulus marginatus*), sea cradle (*Chiton olivaceus*), mediterranean sponge (*Crella elegans*), and earthworm (*Hormogaster samnitica*) (Riesgo et al., 2012). The datasets and scripts can be accessed via ftp://ftp.sanbi.ac.za.

2.2.2 Core analyses steps in the work-flow

2.2.2 Step 1: Preprocessing reads

Quality scores of ILLUMINA reads generally depreciate towards the 3'-end. Prior to assembly, low quality bases were trimmed from the 3'-end of each sequence if above an error probability of 0.01 (PHRED base quality score of 20) using custom PERL scripts. Reads shorter than 36 bp were discarded. The quality based filtering and trimming process ensured that orphan reads whose partner failed the quality threshold, were retained in a separate file and used for *de novo* transcriptome assembly. The quality of the sequencing output was determined using FastQC V0.7.0 (www.bioinformatics.babraham.ac.uk/projects/fastqc/).

2.2.2 Step 2: De novo assembly of putative transcripts

Reference free transcriptome reconstruction was performed separately using either TRINITY (release 2012-06-08; Grabherr et al., 2011), or VELVET (version 1.2.03; Zerbino and Birney, 2008) in combination with OASES (version 0.2.06; Schulz et al., 2012). TRINITY implements greedy algorithmic traversal of the *k*-mer graph prior to building a de Bruijn graph from clusters of pre-assembled

sequences. As a result, assembled transfrags are represented by actual reads. OASES on the other-hand, interrogates a pre-assembly from velvet to address alternative splicing and coverage variation across transcripts. TRINITY was specifically designed for transcriptome assembly using a single, fixed *k*-mer size (*k*-25). Therefore we tested OASES *k*-25 and two variations of multiple *k*-mer assemblies: an additive assembly by pooling (Oases-P) as described by Surget-Groba and Montoya-Burgos (2010), and a merged assembl using the Oases-merge pipeline (Oases-M). Only transfrags above 100 bp were kept for downstream analysis.

Final assemblies for OASES where generated after a series of insert size optimization steps. To achieve this we investigated the implication of inferred optimal library size on reconstruction efficiency for OASES assembler with and without insert size information. OASES performs an automatic estimate of the insert size from reads that map to a common node when the insert size is unspecified. Where the insert size information was used, the nominal (gel estimate) and the *in silico* estimated insert sizes where evaluated.

We estimated an *in silico* insert size for each draft assembly generated without specifying an insert size to OASES as follows: forward and reverse duplicated, and forward substrings where removed using PERL scripts; paired-end reads were mapped on to the unique transfrags from each single *k*-mer assembly and to the *N. crassa* reference CDS from FungiDB database (Stajich et al., 2011) using bowtie (Langmead et al., 2009) '-p 20 -m 1 -n 0 -X 1000 -I 0 -l 28 --chunkmbs 300 –f'. The bowtie parameters ensure that only uniquely mapped read pairs were reported. The distance between uniquely mapped paired reads was extracted using a PERL script.

In OASES, the basic assumption for graph topology constraint is that the distribution of distances between read pairs is normal. The insert size provides a reliable estimate of the likelihood that a read pair is at their observed location on the transfrags. We compared the insert sizes, estimated from each single *k*-mer

assembly and the reference *N. crassa* CDS using R for Statistical Computing (http://www.r-project.org/). We assume a homogeneous density distribution of extracted distances between the test (draft assemblies) and reference (CDS) observation (paired distances). Non-parametric analysis was applied to insert sizes across each *k*-mer assembly and CDS collection and the differences between each estimate was assessed *post hoc* using Agricolae package version 1.1-1 (Mendiburu, 2012).

We compared each draft assembly at the protein level using tBLASTn. *N. crasss* predicted proteins were searched against each customizable database of TFs from each draft assembly. By selecting the best BLAST hit at E-value threshold of 1e-3 and a minimum reference coverage of 50%, High-scoring segment pairs (HPSs) were analysed according to equation E1 (Wasmuth and Blaxter, 2004) using custom PERL scripts. In a situation where more than one top scoring hit had an equal E-value, their HPSs were sorted and ranked by bit-score. We computed the fractional coverage, referred to as HSP ratio, described in Chapter 3 or (Mbandi et al., 2014) in the context of scoring transcriptome assemblies in protein space.
TransfragBitScore is the bit score for the alignment of the highest scoring six frame translation of a reconstructed transcript; ReferenceBitScore is the bit score for the alignment between the reference protein and itself; ReferenceProteinLength is the length of the reference protein.

Eqn 1

$$\text{NormBitscore} = \frac{\text{TransfragBitScore}}{\text{ReferenceBitScore}} \times \frac{3 \times \text{ReferenceProteinLength}}{\text{TransfragLength}}$$

2.2.2 Step 3: Removing redundant sequencing

To avoid inflation in assembly statistics, a custom PERL script, was used to remove conventional duplicated transfrags and reverse compliment duplicates. Only one copy of each duplicate was retained. Short transfrags (reverse and forward) with 100% identity (substring or subsequence) to other sequences were completely removed. Some variants of reconstructed transcripts are different only for small variations, such as small insertions or deletions and SNPs. We consider them as distinct since we do not know to what extent these small variations are from sequencing errors or miss-assembly which could introduce premature termination of translation. To compare our filtering approach with a typically applied post-assembly clustering step, we used CD-HIT-EST (Li and Godzik, 2006) with the following parameters: -M 0 -T 20 -g 0 -c 1.0 -b 1 -aL 1.0 -aS 1.0 -n 10 -d 0 -p 1 (duplicated removal, +\-) and -M 0 -T 20 -g 0 -c 1.0 -b 1 -aS 1.0 -n 10 -d 0 -p 1 (substring removal, +/+). In addition, we evaluated the redundancy in each assembly using CD-HIT-EST as described by Haznedaroglu et al. (2012).

2.2.2 Step 4: Coding potential assessment and conceptual translation

Transfrags were evaluated for protein coding attributes using PORTRAIT (Arrial et al., 2009; version 1.1 with personal modifications)**.** Non-stranded RNA-Seq data are used to create a de Bruijn graph such that paths along the graph obey the semantics of double-stranded DNA (Miller et al., 2010). We corrected PORTRAIT to run ANGLE (Shimizu et a., 2006) in 6 frames, since the biological orientation of transfrags from non-strand specific libraries cannot be readily ascertained.

ANGLE exploits a hidden Markov model and a frame-shift detector to predict CDS in incompletely assembled or truncated transfrags. Below is a PERL pseudocode implementation that was modified to account for transfrag orientation.

```
foreach transfrag in ASSEMBLY{
        $FDP_score = angle($transfrag);
        $transfrag = reverse(transfrag);
        transfrag =~ tr/A,T,C,G/T,A,G,C/;
        $RDP_score = ANGLE($transfrag);
        $ORF = (frame with higher DP_score)
}
```

The ORF for a particular transfrag is chosen as that with the higher dynamic programming score (DP_score), whether forward (FDP_score) or reverse (RDP_score). A modification was made to agree with the grammar of PERL programming for the comparison operator in selecting the frame with higher DP_score like so:

```
if ($DP_score1 >= $DP_score2){
        print OUT "$header\n";
        print OUT "$predicted_protein1\n";
        print WITHORF "$header\n";
        print WITHORF "$sequencia\n";
}
```


The high scoring ORF is conceptually translated into protein using a standard codon usage table. Transfrags without putative predicted proteins are evaluated for coding capability through a protein-independent model. In both models, only intrinsic features are extracted. However, only the protein dependent model is used to assess protein coding potential in IFRAT.

2.2.2 Step 5: Functional annotation

We assigned protein domains to the predicted protein sequences using HMMER version 3.0 (Eddy, 2011) with the manually curated protein profile Hidden

Markov Models from Pfam (release 26.0, ftp://ftp.sanger.ac.uk). We then applied MultiPfam2go to explore co-occurrence relationships between the domains of each protein and assigned functional labels (gene ontology terms) if the underlying domain architectures predicted protein function (Forslund and Sonnhammer, 2008).

To mimic annotation of non-model organisms, we generated a BLAST-able database of UniProt Knowledgebase (FUNGI) release 2013_02 (The UniProt Consortium: http://www.uniprot.org/), excluding *N. crassa* sequences. We screened for highly significant BLASTx hits (max E-value $1e^{-10}$) using the NCBI BLAST package (version 2.2.25) and identified the top hit (lowest E-value, best scoring HSP covers minimum 25% of the hit) using custom PERL scripts.

2.3 Validating *bona fide* transcripts by mapping to reference genome and predicted CDS

The *bona fide* transfrags were aligned to the reference CDS with BLAT v. 34 (Kent, 2002) to assess the integrity of assembly-derived transcripts. BLAT alignment in sim4 format was generated under intron restriction (-fastMap) and post-alignment processing was performed through a series of custom PERL scripts.

Genome based clustering was performed to assess gene coverage by aligning *bona fide* transfrags to *N. crassa* reference genome with GMAP 2013-03-31.v5 (Wu and Watanabe, 2005). Using an PERL API, introns for *N. crassa* were obtained from EnsemblFungi (http://fungi.ensembl.org) to compute the maximum total length of intron per gene. Information about intron length statistics in fungi were obtain as described by (Kupfer et al., 2004) to parameterize transfrag and CDS alignment to the genome: min-intron length = 20, max-intron length = 2000, total length = 5904. A pictorial representation for finding the transfrag and CDS that overlap is shown in **Figure 2.2**. Known gene loci are compared to transfrag loci in a pair-wise manner using in-house PERL scripts to avoid building cluster

chains. Transfrags that do not overlap with CDS are clustered using Bedtools (Quinlan and Hall, 2010).

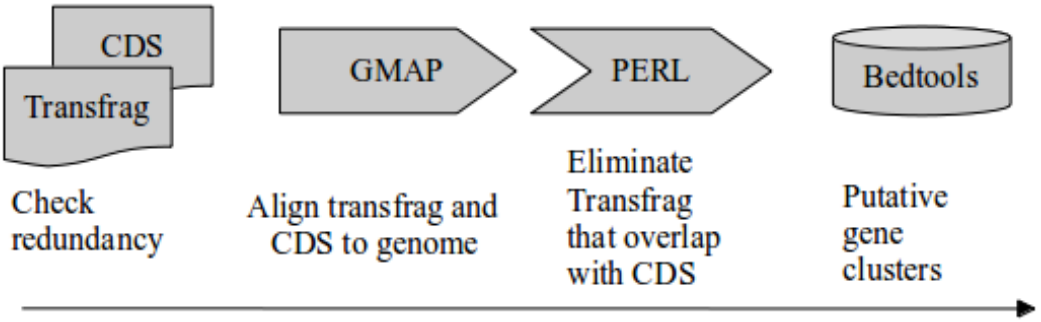


Figure 2.2 Validating *bona fide* transfrags by mapping to reference genome. Transfrags that overlap with CDS were removed to prevent formation of chained clusters.

2.4 Evaluating IFRAT

We teased out the robustness of IFRAT in selecting *bona fide* transfrags in reference-free assembly derived-transfrags. Published, publicly available transcriptome assemblies for selected non-model species were procured. Each assembly was checked for redundancy and only transfrags with a minimum length of 100 bp were used for downstream analysis. For each species, we prepared a customised database of UniProt corresponding to the taxon and performed BLASTx with E-value $\leq$ 1e-10. For each hit, the highest scoring HSP with a coverage $\geq$ 25% was selected as a reliable hit.

2.5 Results and Discussion

2.5.1 Quality assessment and Preprocessing

There exist no consensuses on the quality score threshold for filtering/trimming short-reads from NGS sequence platforms. Although NGS platforms are equipped with quality filtering tools, they are not usually accessible to researchers who

outsource sequencing services (Cox et al., 2010). A generalised guideline is to compute summary statistics with some visual inspection of the quality variation before embarking on a filtering/trimming strategy. Empirically, a minimum quality score of twenty, representing a 1 in 100 chance of uncertainty is considered a comfortable compromise for retaining a good proportion high quality reads for *de novo* assembly. The distribution of per base quality score for trimmed and untrimmed reads is shown for *N. crassa* reads ≥ 36 bp in **Figure 2.3**. The quality score distribution for trimmed reads indicates that every base in each read has a quality score > 20. The distribution of read length after trimming, suggest that the majority of reads are ≥ 75 bp as shown in **Figure 2.4**. Read trimming is known to improve downstream analysis but can lead to significant loss of data (Le et al., 2013). However, about 82.47% of reads were retained as a result of quality trimming indicating that the *N. crassa* data is of high quality (**Table 2.1**). This number is greater than 30 million reads, suggested to be optimal in studies of *de novo* transcriptome assembly in non-model species (Francis et al., 2013) and within the range of 15 – 50 million reads sufficient to detect the majority of genes in human tissue (Hou et al., 2013).

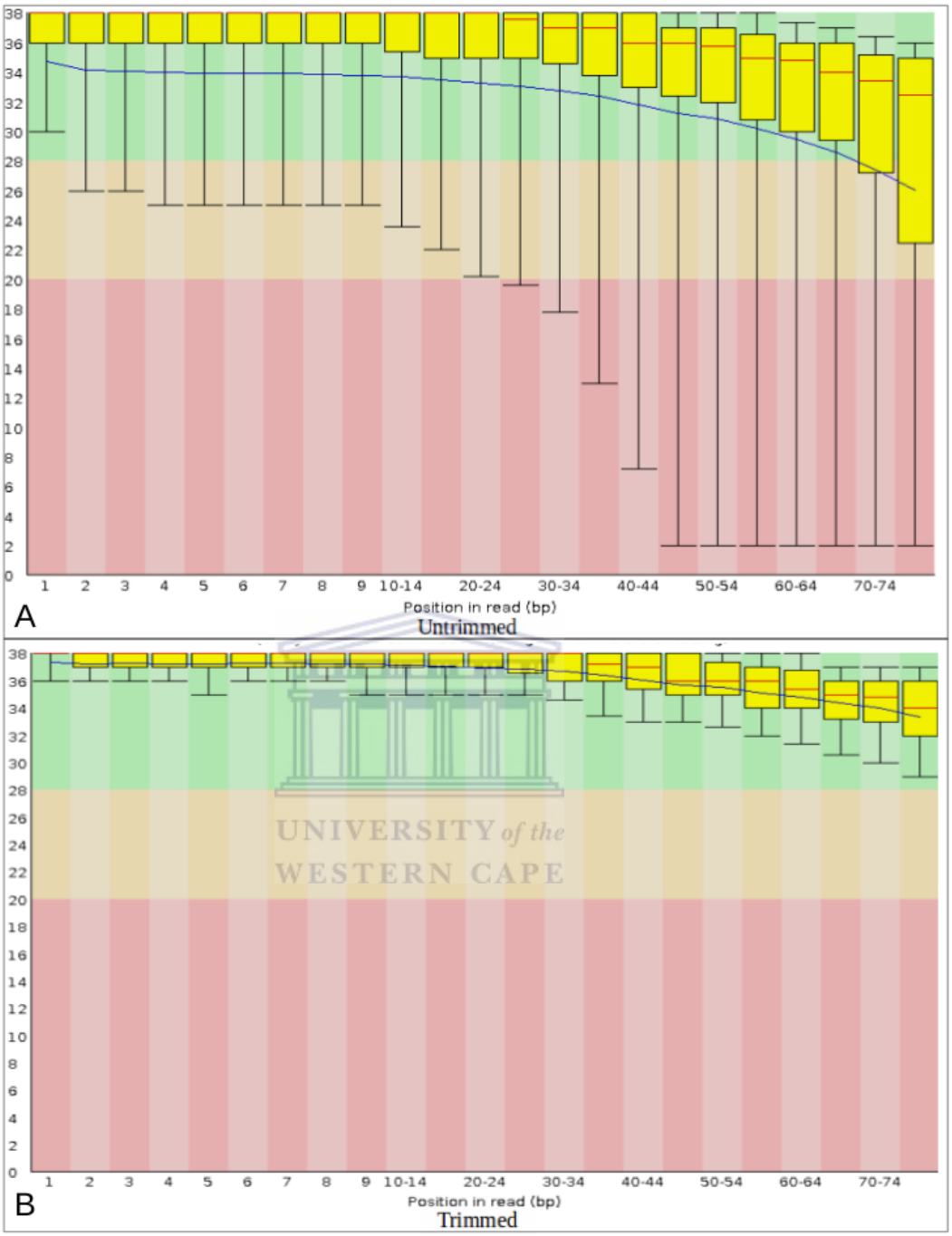


Figure 2.3 Per-base Quality Score Distributions for *N. crassa* reads.

The median and mean base quality scores were calculated for each base position of each read. The red and blues lines depict the fluctuation in median quality score and average quality scores respectively. Upper and lower whiskers refer to 90% and 10% quantiles, while the yellow boxes indicate interquartile range (25-75%). Panel A: distribution of quality scores of untrimmed reads from short-read achieve. Panel B: the narrow spread of whiskers for trimmed reads indicates that low quality regions near the 3' end of reads have been removed.

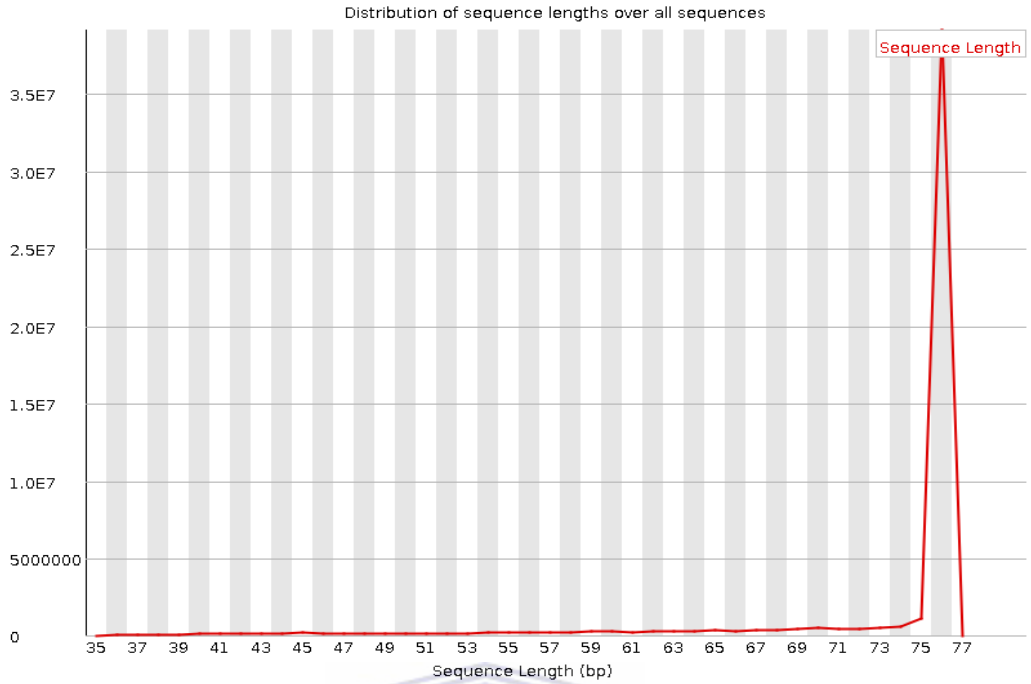


Figure 2.4 Read length distribution for *N. crassa* trimmed reads. Progressive removel of low quality bases from the 3'-end results in reads of varying length

Table 2.1 Summary statistics on sequence trimmed and unitrimmed *N. crassa* reads.

Reads	№ of bases	№ of reads		Basic length statistics		
		Paired	Single	Max	Min	Mean
Untrimmed	4,757,759,296	31,301,048 (100%)	nil	76	76	76
Trimmed	3,736,456,933	24,390,689 (78%)	2,849,486 (4.6%)	76	36	72

2.5.2 Reconstructing putative transcripts

The choice of a non-commercial assembler for reference free reconstruction of transcripts depends on a number of practical considerations such as memory usage, ease of installation, runtime, as well as community usage; the detail of

which is beyond the scope of this study (see Clarke et al., 2013; Lu et al., 2013; Zhao et al., 2011 for a review). We examine the robustness of IFRAT on various *de novo* assembly strategies and the implication of these assembly strategies on gene coverage coverage and *ab initio* annotation. The assembly strategies accomplished were: a non-adjustable single *k*-mer assembly (TRINITY), re-assembly (Oases-merge) or clustering (CD-HIT-EST) of single *k*-mer assemblies and additive multiple-*k* assembly approaches.

2.5.2.1 Optimal insert size estimate

OASES interrogates a pre-assembly generated by Velvet (Zerbino and Birney, 2008), employing a set of dynamic and static filters to enumerate full length isoforms (Schulz et al., 2012). When library size is not specified, Velvet attempts an estimate between reads sharing on common node. This however depends on an initial good draft assembly per *k*-mer. However, nominal insert size is applied across all *k*-mer values. The quality of reconstructed transcripts is sensitive to the size of the *k*-mer (Surget-Groba and Montoya-Burgos, 2010).

Systematic estimation of the *in silico* insert size was carried out to inform the parametrization of paired end assembly with VELVET/OASES. The distribution of insert sizes estimated from uniquely mapped read pairs to the *N. crassa* draft assemblies (generated without insert size specified) and *N. crassa* CDS are shown in **Figure 2.5.** Although the frequency distribution of insert sizes is similar across all assemblies, a cursory look at the box-plot distribution of insert sizes, suggest that they exhibit subtle differences. Differences in insert sizes between draft assemblies were inferred by statistical testing.

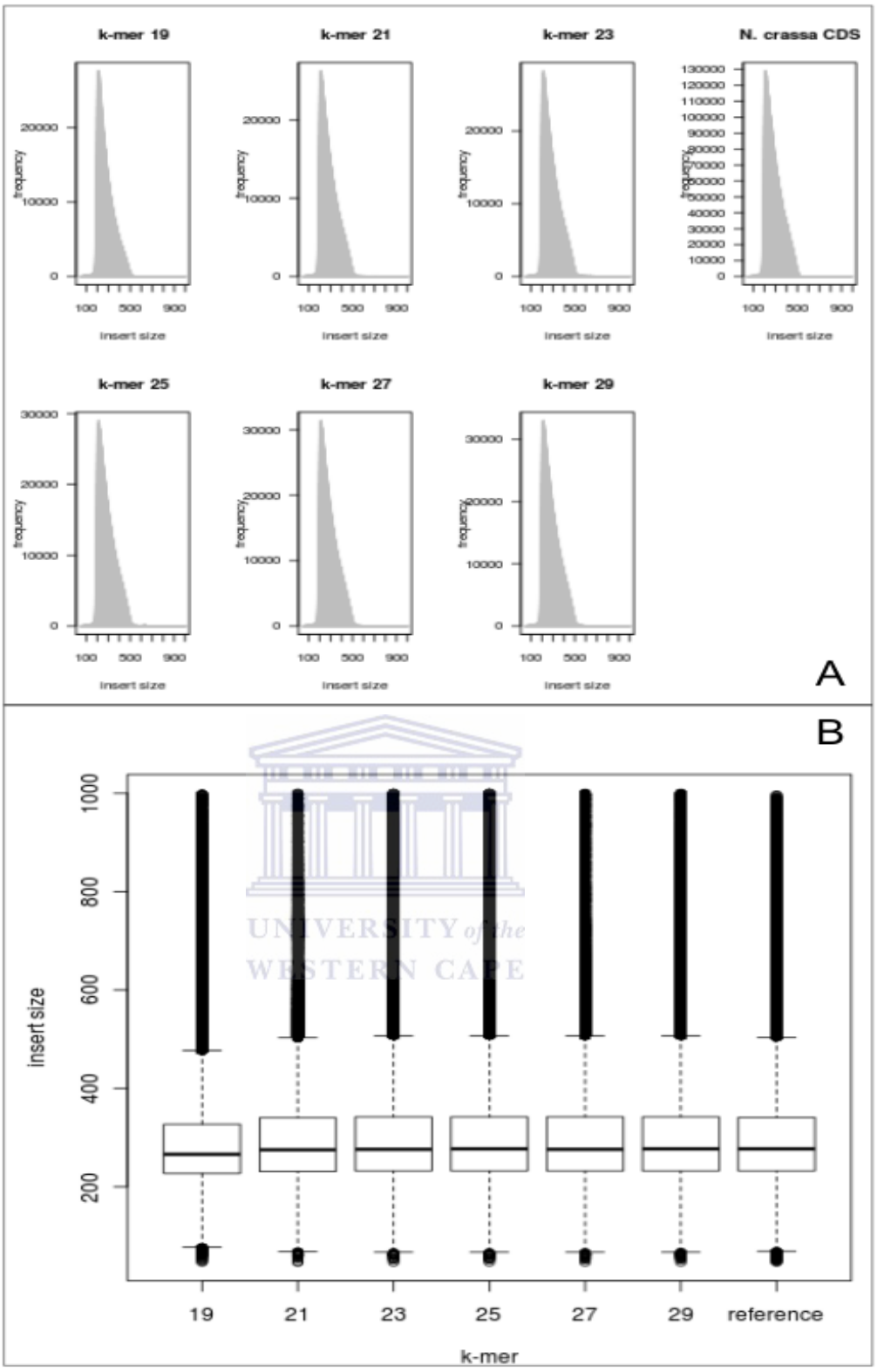


Figure 2.5 Distribution of insert sizes estimated from draft assemblies and *N. crassa* CDS.

The distribution distance (bp) between uniquely mapped read pairs is shown for each draft assembly (*k*-mers, 19-29) and *N. crassa* CDS. Panel A shows similar distribution of insert sizes. An unequal median insert size is easily noticed for *k*-mer 19 for panel B.

Kruskal–Wallis one-way analysis of variance suggests the insert size is significantly different across *N. crassa* assemblies (p-value < 2.2e-16). Multiple comparisons testing between insert sizes is shown in **Table 2.2**. Pairs of means with the same letter are not significantly different (P<0.01). *Post hoc* analysis indicates that insert size estimated from *k*-25 is statistically not different from that estimated from reference CDS. This indicates that *k*-25 is a suitable compromise between these two extremes (19-29). The idea of an intermediate *k*-mer has previously been mentioned by Surget-Groba and Montoya-Burgos, (2010), balancing diversity and contiguity. Furthermore, this supports the fixed *k*-mer value, 25 used in TRINITY (Grabherr et al., 2011) where the authors consider it to work very well for both highly and lowly expressed transcripts.

Table 2.2 *Post hoc* analysis of insert size.

Treatment	mean of the ranks	Groups at α = 0.01 (Post hoc)
9	19439998.353	E
21	20810019.398	D
23	20965022.201	C
25	20992962.558	B
27	20952503.889	C
29	21022899.552	A
reference	20987054.523	B

Table 2.3 Fractional coverage and *Post hoc* analysis normalised bit-score for draft *N. crassa* assemblies.

	Normalized Bit-scores		Fractional Coverage
Treatment	mean of the ranks	Groups (P<0.01)	Mean
9	14746.52	a	0.8950834
21	3139.55	d	0.9528985
23	3210.63	d	0.9566642
25	13518.89	cd	0.9572466
27	13769.44	cb	0.9532967
29	14057.94	b	0.8950834

The quality of the assemblies inferred by the alignment of their predicted proteins (measured by the bitscore and fractional coverage) is shown in **Table 2.3**. Overall, the mean fractional coverage of *k*-25 is slightly higher than that of the rest of the assemblies. The method for using this as a metric for evaluation *de novo* derived transcripts is described by Mbandi et al., (2014). The HSP ratio is a metric for scoring the highest segment pair from the alignments of the six frame translated proteins of each transfrag to a set of 'known' proteins. A poorly reconstructed transfrag would have truncated proteins and frame shifts that would produce low coverage alignments. On a transcriptome scale, the proportion of high scoring alignments is a good estimation of how good the assembly is, when compared to another generated with a different *k*-mer. Post hoc analysis of normalised bitscore suggests that *k*-25 assembly has intermediate quality for assemblies in the range 21-29.

2.5.2.2 Effect of insert size on assembly quality

The insert size estimated from the *k*-25 draft assembly was used to parameterize VELVET/OASES assembly. Given that the empirical distribution of the insert size data (**Figure 2.5**) is not bell-shaped, we compared assemblies generated with the median, mean and nominal insert size. The nominal value is provided as part of the data attributed on the SRA website for SRR100067. Some basic metrics describing the general attributes of the assemblies for unique transfrags is shown in **Table 2.4**. Each VELVET/OASES assembly is categorised based on the source of insert size estimate. The number of reconstructed transcripts depreciates substantially with increasing *k*-mer in both redundant and non-redundant (unique) assemblies irrespective of insert size category. De Bruijn graph based assemblers rely on absolute overlap between *k*-mers generated from RNA-Seq reads. The degree of overlap is inferior to and is specified by the magnitude of the *k*-mer. Intuitively, longer *k*-mers will unambiguously overlap at unique nodes that may be sufficiently long to represent entire exons or stretch beyond exon junctions (Martin and Wang, 2011). The specificity is inferior for low *k*-mer values which in combination with sequencing error might potentially elaborate transcript

variants. For each category, we observe that the combined assemblies (merge and pool) are more contiguous (larger N50 value). This corroborates with studies that suggest that combine assemblies generate longer transgfrags (Gibbons et al., 2009; Martin et al., 2010; Surget-Groba and Montoya-Burgos, 2010).

When assembly statistics are compared between the insert size categories, contiguity increases in the following order: default, nominal, mean and median. This observation suggests that sequence length increments are as a result of an optimal estimate of insert size that was used to parameterize paired-end assembly. Insert size in the context of OASES assembler is analogous to library size (Zerbino and Birney, 2008). Library size selection allows us to choose an appropriate insert length with precision, such that appropriate restrictions are implemented during de Bruijn graph construction and traversal. Imprecise insert estimation can occur during gel electrophoresis that may be subjective. Automated gel-less library construction methods are likely to minimise this effect (Rodrigue et al., 2010).

Table 2.4 Basic metrics describing the general size characteristics of the *N. crassa* assemblies.

Type	*K*-mer	№ of raw TF	№ of unique	Max TF	Medain TF	Mean TF	№ of TF > median	N50
Oases draft	19	45482	32986	47097	453	950	16509	2053
	21	24211	22931	16171	1016	1502	11467	2720
	23	22150	21686	14139	1033	1496	10845	2698
	25	21637	21394	14086	980	1450	10698	2640
	27	21360	21158	14086	891	1382	10583	2561
	29	21398	21302	14086	810	1325	10652	2499
Oases nominal	19	39199	30882	74230	597	1181	15446	2409
	21	23565	22817	14139	1293	1641	11409	2781
	23	22008	21510	14139	1220	1588	10762	2740
	25	21357	21122	14086	1150	1530	10564	2689
	27	21078	20819	14086	1038	1458	10412	2618
	29	21115	20952	14086	928	1391	10477	2562
	Oases-	69569	41764	74249	1675	1940	20890	2944
	Oases-P	148322	62769	74230	1453	1756	31386	2808
Oases mean*	19	39184	30383	45794	609	1201	15192	2459
	21	22271	21643	14138	1432	1747	10822	2900
	23	21012	20511	14139	1356	1683	10264	2849
	25	20455	20111	14086	1304	1633	10058	2796
	27	19607	19381	14086	1244	1575	9691	2738
	29	19581	19422	14086	1141	1513	9713	2701
	Oases-	71619	43914	45791	1780	2048	21960	3050
	Oases-P	142110	65642	45794	1563	1859	32835	2908
Oases median *	19	39517	30276	61473	610	1210	15143	2485
	21	21973	21349	14138	1506	1804	10675	2950
	23	20231	19799	14139	1467	1760	9904	2909
	25	19406	19203	16841	1425	1719	9603	2861
	27	18908	18680	14086	1357	1651	9342	2803
	29	18681	18522	14086	1281	1599	9262	2773
	Oases-	73215	45271	62895	1834	2107	22653	3091
	Oases-P	138716	67673	61473	1626	1918	33845	2971
Trinity	25	35720	35578	14086	240	880	17821	2441

*The mean and median OASES assemblies were performed with insert size estimates obtained by computing average and median mapped distances respectively using uniquely mapped paired-end reads to the *k*-mer 25 draft assemblies. The resulting transfrags were filtered for redundancy before computing basic assembly statistics.

2.5.2.3 Removing redundancy

For the purpose of the IFRAT pipeline, only the median *k*-25, merge pool and TRINITY were examined further. A summary on assembly statistics after removal of forward/reverse duplicates/substrings for the four assembly methods is shown in **Table 2.5**. When comparing the two single *k*-mer assembly approaches (TRINITY and OASES-25), we see that TRINITY produced twice as many transfrags as OASES-25, but at much shorter transfrags lengths. These two assemblies had very little redundant transfrags compared to multiple *k*-mer assemblies. Multiple *k*-mer assemblies produced a much higher number of transfrags than single *k*-mer assemblies, but between 38% and 56% were redundant. The median transfrag lengths for these assemblies were 7 fold greater than for the TRINITY assembly. To compare our filtering procedure (PERL script) with a typically applied post-assembly clustering method, we used CD-HIT-EST and generated a non-redundant assembly at 100 % global identity. At these settings, our filtering method produced comparable results.

Table 2.5 Attributes of *N. crassa* assemblies produced with different approaches.

Assembly	№ of TF	№ of unique TF (PERL)	Median unique TF length (PERL)	% Redundant TF PERL	№ of unique TF (CD-HIT	% Redundant TF CD-HIT
Trinity	35720	35578	240	0.4	35578	0.4
Oases-25	19406	19193	1426	1.09	19217	0.97
Oases-M	73215	45134	1839	38.35	45134	38.35
Oases-P	138716	61293	1749	55.81	61717	55.51

Typically, CD-HIT-EST is used at settings below 100% identity. The fraction of redundant transfrags at various identity thresholds for our *N. crassa* assemblies is shown in **Figure 2.6**. For the Oases-P assembly, at 80% identity nearly 90% of the transfrags are considered redundant by CD-HIT-EST. This represents nearly 46,000 transfrags that are lost for downstream analysis as compared to clustering at 100% identity. With this approach, we removed slightly more transfrags than

with CD-HIT-EST at 100% identity because this program does not properly process transfrags containing 'Ns' (author's personal communication). Our results suggest that single *k*-mer assemblies may not need this filtering step since the proportion of redundant transfrags in the TRINITY and Oases-25 datasets were only about 1%. In contrast, redundancy filtering is particularly important in multiple *k*-mer assemblies, considering that nearly half the transfrags in the Oases-M and Oases-P datasets were exact copies or substrings of other transfrags. It is unknown at what percent identity clustering results in significant loss of unique functional annotations. However, clustering without biological insight should be handled with caution, considering that our analysis indicates that already at 99% identity a significant subset of potentially unique transfrags is removed by CD-HIT-EST.

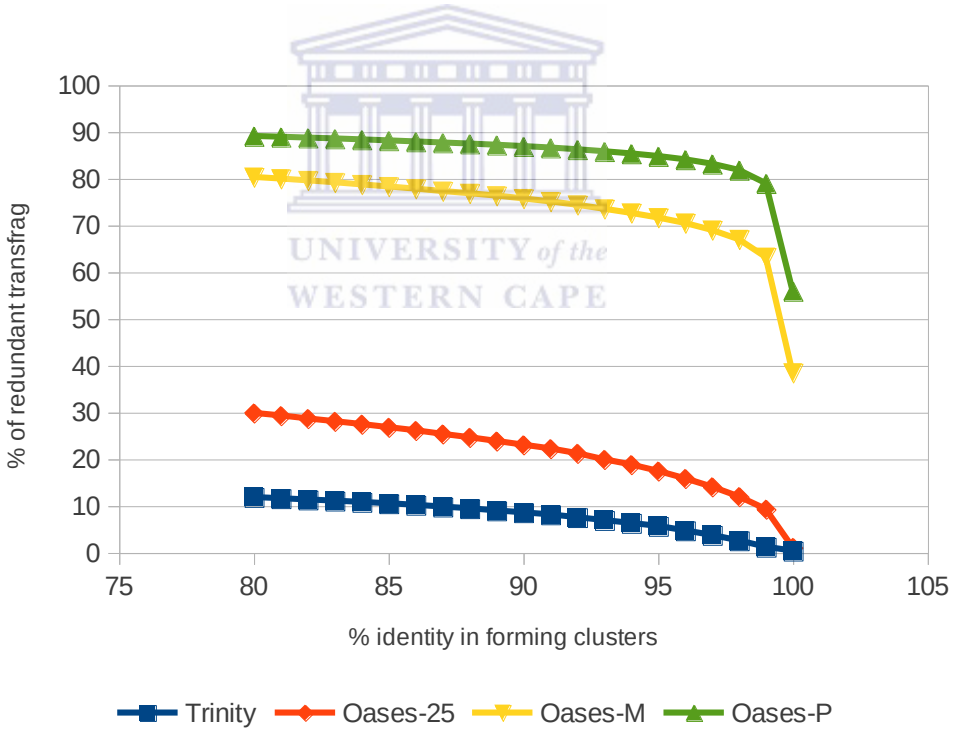


Figure 2.6 Comparing the ratio of redundant transfrags across all assemblies at each identity (%) threshold in creating clusters with CDHIT.

2.5.2.4 Selecting *bona fide* transfrags and their functional annotation

Each non-redundant assembly was separated into two categories: *bona fide* (coding with predicted ORF) and orphan (non-coding, coding without ORF); numbers are displayed in **Table 2.6**. In TRINITY, the proportion of orphan transfrags was higher (60%) than the proportion of *bona fide* transfrags. TRINITY also produced a considerably higher number of orphan tranfrags than any of the three OASES assemblies. As a result, the number of *bona fide* transfrags was very similar for the two single *k*-mer assemblies, and Oases-P generated the highest number of *bona fide* transfrags.

Table 2.6 Classification and annotation of *N. crassa* transfrags.

Assembly	№ of unique TF (UTF)	№ of orphan UTF	№ of *bona fide* UTF	№ of orphan UTF with blast hit	№ of *bona fide* UTF with blast hit	№ of *bona fide* UTF with multiPfam2go
Trinity	35578	20772	14806	266 (1.3%)	10320 (70%)	6523 (44%)
Oases-25	19193	5359	13834	160 (3%)	11438 (83%)	6944 (50.2%)
Oases-M	45134	7453	37681	412 (6%)	31311 (83.1%)	18173 (48.2%)
Oases-P	61293	10848	50445	646 (6%)	41383 (82%)	24393 (48.4%)

Figure 2.7 shows the distribution of tranfrag lengths between *bona fide* and orphans transfrags. Orphan transfrags were generally much shorter than *bona fide* transfrags. For the *bona fide* transfrags of the three OASES assemblies, the median transfrag length (~ 2 kb) and the distributions are very similar. We note that the OASES assemblies had a considerable number of *bona fide* transfrags that were substantially longer than 10 kb. The median transfrag length of *bona fide* transfrags assembled using TRINITY was 1.5 kb, and only a few of them were longer than 7.5 kb.

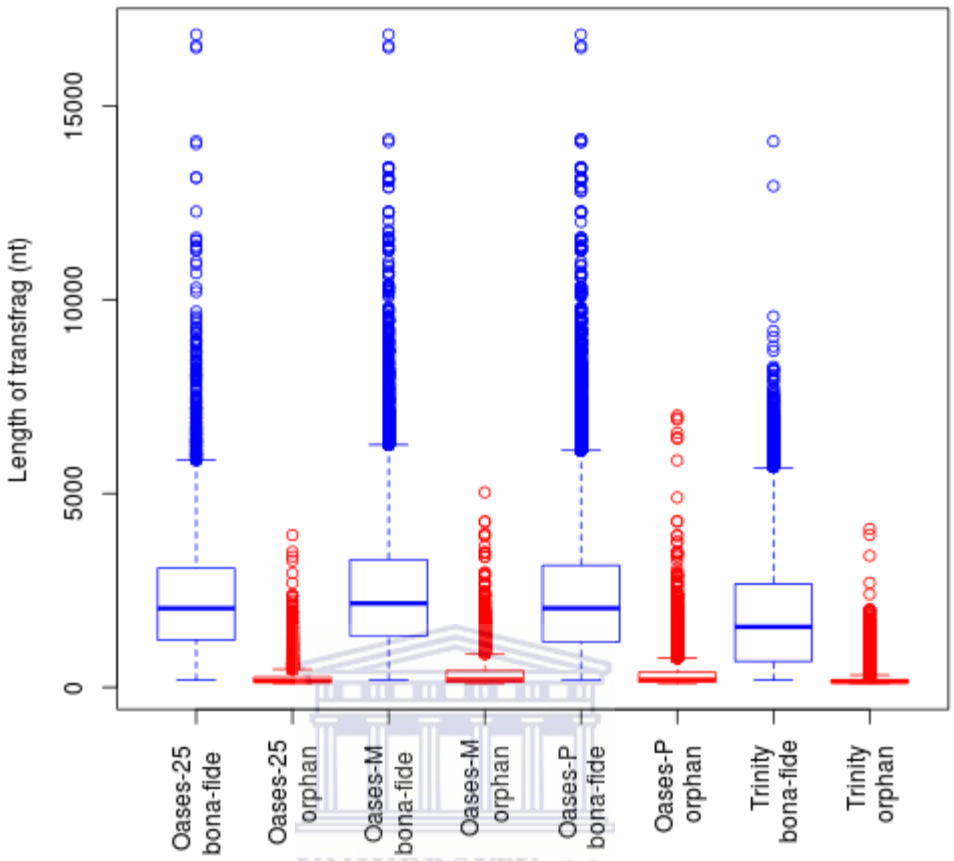


Figure 2.7 The distribution of transfrag length is shown for all assemblies. The blue and red boxes represent the bona-fide and orphan categories, respectively.

Non-redundant assemblies were annotated using BLAST and multiPfam2go (**Table 2.6**). We note that in all assemblies only a small proportion of orphan transfrags had a BLAST match. Despite the highest number of orphan transfrags, TRINITY had the least number of BLAST hits to transfrags in this category. In contrast, at least 70 % of *bona fide* transfrags from all assemblies had a BLAST hit. This number is higher than the ones typically reported in studies on *de novo* assembled transcriptomes (Franchini et al., 2011; Sun et al., 2010). In addition, *bona fide* transfrags were annotated with multiPfam2go. The fraction of transfrags that could be associated with gene ontology terms ranged from 33%-50%, which is also high for domain based annotation.

Our subsequent BLAST analysis corroborated this categorization, since 70-80% of *bona fide* transfrags had significant BLAST matches while this was true for only 1-6% of orphan tranfrags. We note that the median length of *bona fide* transfrags ranged from 1.5kb (TRINITY) to 2kb (OASES) which is consistent with the average coding sequence length in fungi (Galagan et al., 2005), while most of the orphan transfrags were short (med. 147-198 bp). However, our results confirmed previous findings that length is not the only indicator of coding potential (Frith et al., 2006) and 'non-blastable' transfrags (Logacheva et al., 2011; Sadamoto et al., 2012), since 6%-26% of the orphan transfrags with BLAST matches were less than 200 bp long.

The most profound observation for functional relatedness of *bona fide* transfrags is shown in **Figure 2.8**. Clearly, *bona fide* transfrags represents the majority of functional information that can be obtained from annotation by transference (e.g BLAST) which represents the most common approach for annotating assembly derived-transcripts.

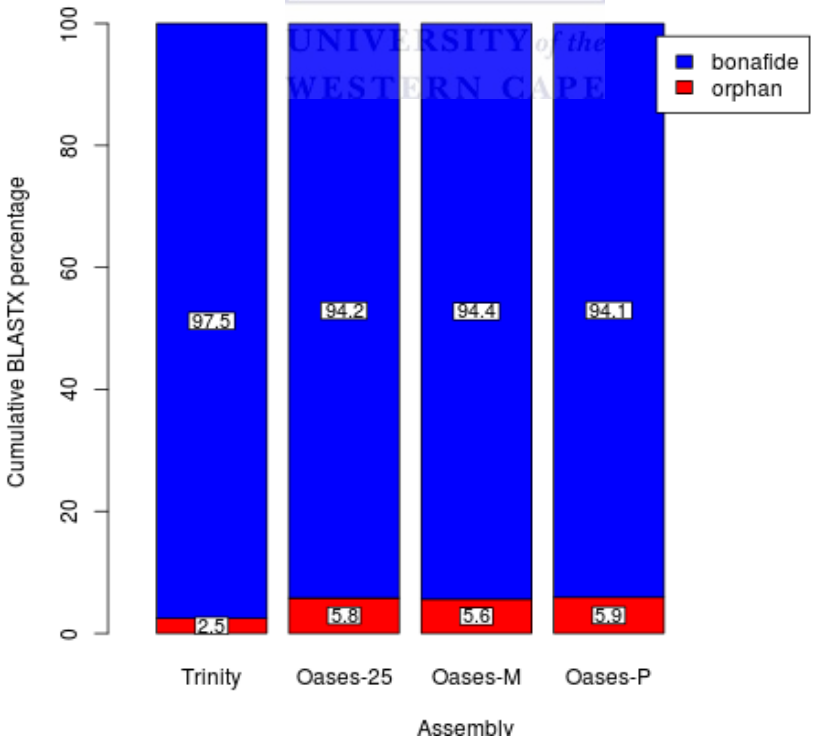


Figure 2.8 Distribution of BLASTx hits between *bona fide* and orphan transfrags. The *bona fide* transfrags are enriched with sequences that have a potential BLAST hit.

We integrated multi-domain co-occurrence architecture (Forslund and Sonnhammer, 2008) to complement BLAST annotation. This avoids non-specific annotation of promiscuous domains resulting from truncated transfrags. Between 44% and 50% of the *bona fide* transfrag-derived peptides from *N. crassa* were assigned with at least one GO term. Using IFRAT, we also improved annotation coverage of published transcriptome datasets from non-model organisms. The choice of database and the coverage filter threshold to a larger extent, accounts for small differences in the number of BLAST hits between *bona fide* transfrags and unfiltered assemblies. We attribute this high annotation coverage to the error tolerant CDS prediction (Shimizu et al., 2006) and selection of longer proteins with coding potential by IFRAT.

2.5.2.5 Assessing transfrag integrity and gene coverage

To evaluate the number of predicted genes represented by the *bona fide* transfrags, we aligned them to the predicted coding sequences (CDS) as well as to the genome of *N. crassa* (**Table 2.7**). Between 80% and 90% of the *bona fide* transfrags mapped to both datasets at high stringency. Although the numbers of *bona fide* transfrags between single and multiple *k*-mer assemblies is very different, the number of identified genes is very similar. Most strikingly, TRINITY identified the same number of predicted genes and putative novel *N. crassa* gene loci as Oases-P, independent of the dataset and the alignment thresholds. As a result, the number of *bona fide* tranfrags per gene is lower in single *k*-mer versus multiple *k*-mer assemblies. Orphan transfrags that mapped at the same stringency represented 15-40% of the known gene loci (**Table 2.7**), but ~ 90% were already identified by the longer *bona fide* category.

Table 2.7 Summary of *bona fide*† and orphan* transfrags integrity and validity

Assembly	№ of *bona fide* UTF	Alignment of TF to reference genes				Alignment of TF to reference genome		
		№ of TF Cov 50%, ID 50%	№ of Reference unigenes	№ of TF Cov 90%, ID 90%	№ of Reference unigenes	№ of TF uniquely mapped	№ of *N. crassa* genes identified by TF	№ of putative novel *N. crassa* gene loci
Trinity†	14806	12029	6594	5274	3378	13331	6968	1080
Oases-25†	13834	11485	6089	3983	2647	11675	6455	677
Oases-M†	37681	30152	6355	8590	2893	27874	6844	869
Oases-P†	50445	41074	6479	12882	3234	39626	6946	979
Trinity*	20772	6164	2381	4836	1846	18538	3960	7552
Oases-25*	5359	1887	1189	1320	834	4908	1455	2386
Oases-M*	7453	2555	1361	1348	840	6142	1659	2440
Oases-P*	10848	4105	1775	2142	1154	9186	2150	2927

All four assembly methods produced high quality datasets, as 80-90% of the transfrags mapped to the genome and the predicted CDS of *N. crassa* at high identity and coverage. *Bona fide* transfrags represented approximately 70% of the 9732 known gene loci in the *N. crassa* genome. In addition, they indicated the existence of 679-1080 unknown potentially coding gene locations. Orphan transfrags also mapped to known gene locations, but most of these locations were represented by longer *bona fide* transfrags. These orphan transfrags may represent biologically interesting data, such as truncated assemblies (e.g. rare exons, poorly expressed genes, transcript with under-sampled regions), or immature mRNA with intronic regions and long UTRs for which coding potential could not be predicted (Cui et al., 2010; Garber et al., 2011; Logacheva et al., 2011). Orphan transfrags that mapped to non-coding regions of the genomes could represent ribosomal or non-coding RNA (O'Neil et al., 2013), and also be of interest. In any case, it is advisable to verify the correct assembly of orphan transfrags and remove mis-assemblies using a suitable reference dataset, such as a reference genome or EST collection.

2.5.3 Selecting *bona fide* assembly-derived transcripts in other species

We also verified the suitability of the IFRAT pipeline for selecting reconstructed transcripts in non-model organisms. The analysis results for unique transfrags longer than 100 bp from each draft assembly are shown in **Table 2.8**. We predicted that up to 70% of the published transfrags do not code (are orphan). As before, the percentage of orphan transfrags with a BLAST hit was relatively low. In contrast, the proportion of *bona fide* tranfrags with significant BLAST matches was often

higher than in the unfiltered draft assemblies.

Table 2.8 Allocation of BLASTx hits between *bona fide* and orphan transfrags inferred with IFRAT

organism	№ of TF is publication	№ of TF with hit in publication	№ of UTF >= 100	№ of orphan UTF	№ of orphan UTF with blast hit	№ of *bona fide* UTF	№ of *bona fide* with blast hit
Hydra vulgaris	48909	17587 (36%)	44484	9806 (22%)	1086 (11.1%)	34717	15310 (44.1%)
Radix balthica	41590	7347 (17.7%)	38790	26846 (69%)	1360 (5.1%)	11944	6723 (56.3%)
Alipes grandidieri	66199	16688 (25.2%)	66297	31355 (47%)	1809 (5.8%)	34942	12253 (35.1%)
Cerebratulus marginatus	80865	11062 (13.7%)	81021	46345 (57%)	782 (1.7%)	34676	9995 (28.8%)
Chiton olivaceus	93879	24495 (26.1%)	93885	52461 (56%)	1692 (3.2%)	41424	11001 (26.6%)
Crella elegans	31703	13984 (44.1%)	31172	10930 (35%)	1364 (12.5%)	20242	7439 (36.8%)
Hormogaster samnitica	90928	25681 (28.2%)	90928	41271 (45%)	1003 (2.4%)	49657	15392 (31%)
Fagopyrum tataricum	25041	19072 (76.1%)	25040	5747 (23%)	1909 (33.2%)	19294	16326 (84.6%)

2.6 Conclusion

Transcriptome reconstruction in non-model species is increasingly reliant on approximate computationally intensive *de novo* approaches. At best, it is very difficult to assess correct assembly without a suitable reference and only a tiny fraction of the assembly is ascribed a functional label through annotation by transference. We propose a conceptual work-flow (IFRAT) that addresses key biological considerations in post-assembly analysis of eukaryotic transcriptomes without that need to perform extensive clustering. We critically evaluated its suitability for interrogating the transcriptome of non-model organisms. Our methods reduce the number of assembly derived-transcripts such that computational resources and annotation time is focused on the biologically relevant sequences. TRINITY performs better as a component of our pipeline for a single *k*-mer assembler with comparable performance to OASES multiple *k*-mer assemblies. IFRAT thus produces a *bona fide* collection of transfrags in non-model species that are suitable for functional inference.

CHAPTER 3

A glance at quality score: implication for *de novo* transcriptome reconstruction of Illumina reads

3.0 Abstract

Downstream analyses of short-reads from next-generation sequencing platforms are often preceded by a pre-processing step that removes uncalled and wrongly called bases. Standard approaches rely on their associated base quality scores to retain the read or a portion of it when the score is above a predefined threshold. It is difficult to differentiate sequencing error from biological variation without a reference using quality scores. The effects of quality score based trimming have not been systematically studied in *de novo* transcriptome assembly. Using RNA-Seq data produced from Illumina, we teased out the effects of quality score based filtering or trimming on *de novo* transcriptome reconstruction. We showed that assemblies produced from reads subjected to different quality score thresholds contain truncated and missing transfrags when compared to those from untrimmed reads. Our data supports the fact that *de novo* assembling of untrimmed data is challenging for de Bruijn graph assemblers. However, our results indicate that comparing the assemblies from untrimmed and trimmed read subsets can suggest appropriate filtering parameters and enable selection of the optimum *de novo* transcriptome assembly in non-model organisms.

3.1 Background

Ultra-high throughput or next generation sequencing (NGS) technologies generates a considerable amount of data. This is desirable for single-nucleotide resolution of the genome and underlying expressed transcriptional units. Their application in sequencing the transcriptome is facilitated by parallel development of reference free assembly algorithms that typically depend on the de Bruijn graph (Martin and Wang, 2011). This has resulted in an increase in the number of published transcriptome assemblies for non-model organisms. However, *de novo* assembly is based on approximate computation, which is impeded by random variations in sampling (bias in reads) and sequencing errors. Sequencing errors introduce false *k*-mers which increases the computational demands for graph resolution and the runtime of assembly algorithms. It is difficult to distinguish between sequencing errors from biological variation without a reference (Garber et al., 2011), since variation becomes dominant with volume of sequence data (Conway and Bromage, 2011). In addition, sampling methods aimed at enriching protein-coding (mRNA) transcripts are overwhelmed by bulk amounts of non-coding RNA (Cui et al., 2010) and immature mRNA with incompletely spliced introns (Garber et al., 2011). For researchers who outsource sequencing services, they do not have access to quality filtering tools embedded in NGS platforms (Cox et al., 2010). We can broadly identify two categories of pre-processing tools that address read usability: error correction and filtering/trimming algorithms which have emerged in response to low quality data. Error correction approaches have been largely applied on genomic reads, e.g, Coral (Salmela and Schröder, 2011) and Quake (Kelley et al., 2010) rely on multiple alignments in k-mer space and edit distance respectively to correct reads. Error correction maximizes the quantity of reads for downstream analyses but may reinforce errors and eliminate genuine reads with low frequency *k*-mer (Martin and Wang, 2011). Only recently has error correction been applied to RNA-Seq data where the SEECER algorithm relies on a k-mer profile Hidden Markov model (Le et al., 2013). However, MacManes and Eisen (2013) compared error correction tools on RNA-Seq data and showed that Reptile (Yang et al., 2010) performed best with *de novo*

transcriptome assembly of error corrected reads. Quality score based-trimming approaches are predominantly used, but they often lead to significant loss of data (Le et al., 2013) and are extremely subjective. Reads are often trimmed in varying modes: ConDeTri (Smeds and Künstner, 2011) trims the reads from the 5' , 3' or both ends over a defined number of bases (window) or per base and tools such as FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) will retain or discard an entire read after assessing quality over a fraction of bases in the read. On the other hand, read content artifacts such as sequencing adaptors and ribosomal RNA may need additional heuristics requirements for pre-processing. The effects of quality score based trimming and artifact removal have not been systematically addressed with respect to the quality of *de novo* assembly derived transcribed fragments. Using NGS data, we compared reference free transcriptome assemblies derived from various categories of quality trimmed reads: with and without artifact removal. Although, the RNA sample used is non-synthetic, we focused on the attributes of the assemblies rather than the biological relevance of the RNA source. We report our findings and propose that caution must be exercised when applying quality filters prior to *de novo* assemblies and that comparing the assemblies of untrimmed and subcategories of trimmed reads could provide an optimal quality score threshold for each read.

3.2 Materials and Method

3.2.1 Datasets

Publicly available RNA-Seq data (SRR100067) and the genome assembly (accession AABX00000000) for wild type *Neurospora crassa* 74-OR23-1VA were obtained from the NCBI, http://www.ncbi.nlm.nih.gov/Traces/sra and http://www.ncbi.nlm.nih.gov/Traces/wgsnih.gov/Traces/wgs respectively. Predicted coding sequences (CDS) for *N. crassa* were downloaded from http://fungidb.org release 2.0. In addition, the *Venturia inaequalis* draft genome assembly version 1.0 (Hesse et al., 2013), two lanes of 100 bp paired-end and one lane of 75 bp single-end Illumina RNA-Seq data were procured from a host free

culture of *V. inaequalis*. The datasets and scripts can be accessed via ftp://ftp.sanbi.ac.za/quality.trimming and https://bitbucket.org/Kimbung/hsp.ratio

3.2.2 Pre-Processing RNA-Seq Data

The raw RNA-Seq data from *N. crassa* was trimmed with a typically used minimum PHRED quality score threshold of 20 (Q20) and 10 (Q10) using ConDeTri, with modification (Smeds and Künstner, 2011) from the 3′-end to represent datasets one and two respectively. For *V. inaequalis*, we generated six categories of quality trimmed or filtered reads as follows: (i) Low quality bases were removed at the 3′-end of each read with a PHRED quality score below 20 or 10 representing datasets one and two respectively, (ii) Potential remnants of adapter sequences were removed using FLEXBAR (Dodt et al., 2012) followed by trimming low quality bases with a PHRED quality score below 20 or 10 that represents datasets three and four, (iii) adapter sequences only removed with FLEXBAR to create dataset five. A minimum read length of 36 bp was used for categories 1–5. A sixth category of pre-processed reads was obtained using the FASTX-toolkit by filtering reads where more than 80% of their bases have a PHRED quality less than 10.

3.2.3 *De Novo* Assembly

Reference free transcriptome reconstruction with the untrimmed and trimmed *N. crassa* datasets was performed with TRINITY (release 2012-06-08; *k*-mer 25; Grabherr et al., 2011). For comparison, OASES (version 0.2.06; Schulz et al., 2012) was used to generate assemblies with various *k*-mers (19–35). *V. inaequalis* datasets were assembled only with TRINITY. In all cases, only default assembly parameters were used. Transfrags (TF) ≥ 100 bp were kept for downstream analysis.

3.2.4 Comparing Assemblies

To avoid inflation in alignment or assembly statistics, each assembly was checked for redundant TF using a PERL script to remove exact matches. We aligned the TF from *N. crassa* generated with Q20 (one) and untrimmed reads to the genome with GMAP version 2013-10-04 (Wu and Watanabe, 2005). The following parameters described by Kupfer et al. (2004) were used: min-intron length = 20, max-intron length = 2000, total length = 5904. The total intron length per gene was estimated for *N. crassa* from http://fungi.ensembl.org release-17. The aligned TFs were filtered at high stringency of 95% identity and 95% coverage. TFs from untrimmed reads that did not overlap with those from trimmed reads were verified against predicted CDS loci and recorded as missing annotations using in house PERL scripts for post-processing GMAP alignments. TF derived for the *V. inaequalis* untrimmed and trimmed (category one) reads were aligned to the *V. inaequalis* draft genome using exonerate version 2.2.0 (Slater and Birney, 2005) with the following parameters: model est2genome, maxintron = 5000. Coordinates for best alignment locations were considered and visualized with Gbrowse (http://gmod.org/wiki/GBrowse). The proteins from UniProt Knowledgebase (FUNGI) release 2013_02 (The UniProt Consortium: http://www.uniprot.org) were searched against each customizable database of TF assembled from untrimmed and trimmed *V. inaequalis* reads with BLAST+ (Camacho et al., 2009). *N. crassa* TF produced with TRINITY from both trimmed and untrimmed reads were searched against UniProt *N. crassa* proteins (*E*-value: $10e^{-10}$). Counts of number of unique high scoring segment pairs (HSP) were computed. The ratio of the length of the HSP to known UniProt annotated proteins (hereafter referred to as HSP ratio) was generated with a series of in house PERL scripts and UNIX commands for each dataset. HSP ratio represents how well TF were reconstructed. Non-parametric analysis was applied to HSP ratios across read categories and the differences between the read pre-processing approaches was assessed *post hoc* using Agricolae package version 1.1-1 (de Mendiburu, 2012).

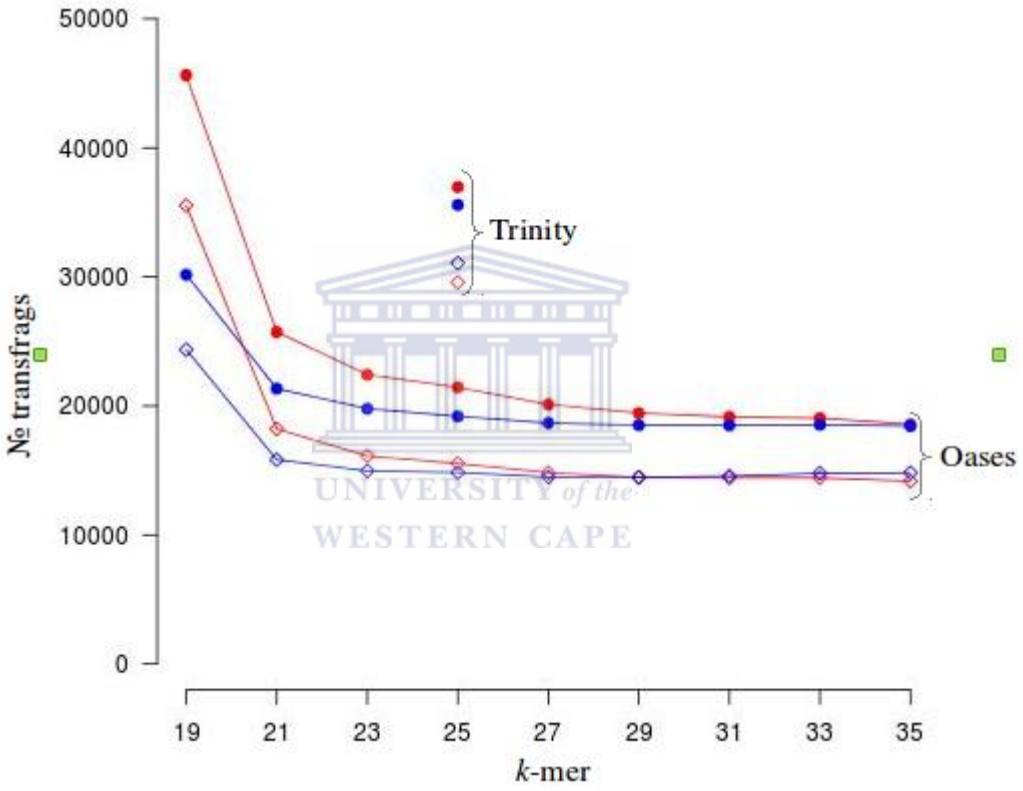
3.3 Results

To investigate the potential side effects of quality based trimming and artifact removal on *de novo* transcriptome assembly, we analysed datasets from a model (*N. crassa*) and non-model organism (*V. inaequalis*). A summary of read counts for each category of untrimmed and trimmed reads is shown in **Table 3.1**. More reads are removed when quality based trimming is preceded by adapter removal compared to doing the reverse.

Table 3.1 Attributes of transfrags produced with TRINITY

Organism	Read category	№ of reads retained	№ of unique TF	N50	№ unique HSP	Median HSP ratio	Mean HSP ratio	sd of HSP ratio	Groups at α = 0.01 (Post hoc)
N. crassa	Untrimmed	62,602,096	36964	2557	6773	0.999	0.862	0.212	a
	One	51,630,864	35578	2441	6668	0.992	0.845	0.226	b
	Two	55,155,297	35614	2532	6757	0.997	0.856	0.217	ab
V. inaequalis	Untrimmed	134,340,808	45449	1502	923328	0.964	0.859	0.205	a
	One	47,261,404	42325	540	686887	0.879	0.773	0.242	c
	Two	64,617,759	43832	696	760648	0.919	0.805	0.231	b
	Three	67,136,546	38645	979	810854	0.950	0.834	0.225	a
	Four	93,862,916	40311	1237	868775	0.960	0.848	0.214	a
	Five	92,491,510	46166	946	840307	0.949	0.835	0.220	a
	Six	101,402,320	43346	1402	907814	0.964	0.855	0.209	a

The percentage of trimmed reads ranged from 35 to 88%. Out of ~134 Gb *V. inaequalis* untrimmed reads, quality trimming preceded with adapter removal retained the smallest amount of reads. When comparing assemblies from various categories of reads, we note that the number of unique TF from untrimmed reads is always higher than those from trimmed reads irrespective of the assembler and dataset used (**Figure 3.1**).



Fingure 3.1 Distribution of unique (solid circles) and overlapping (diamond shaped) transfrags (TF) from *N. crassa*. TF from untrimmed and trimmed reads that map to common a genomic locus can be considered as overlapping. Below *k*-23, there is considerable difference in the number of unique and overlapping TF between the trimmed and untrimmed categories. TF from untrimmed and trimmed reads are represented in red and blue, respectively.

For *N. crassa* TFs, this is much more profound at lower *k*-mers. A similar trend is observed with the number of TFs, derived from untrimmed and trimmed reads that map to the same genomic loci. TFs produced with untrimmed reads recovered a higher number of known *N. crassa* proteins than those from the trimmed reads (**Table 3.1**). A total of 521 known gene loci were identified in *N. crassa* that overlapped with TFs derived from untrimmed but not trimmed reads. Transcriptome assembly statistics for each category of quality trimmed reads and the HSP ratios are shown in **Table 3.1**. The number of unique TFs is comparable among all assemblies for each organism. Untrimmed reads generated the largest number of TFs and identified the largest numbers of known UniProt proteins. Sequence similarity search identified 791 proteins that were present in all *V. inaequalis* assemblies. For *N. crassa*, 6218 proteins were common to all assemblies generated with TRINITY. Kruskal–Wallis one-way analysis of variance suggests that quality score base pre-processing had a significant effect on TF quality in both *N. crassa* ($p = 0.002999$) and *V. inaequalis* ($p < 2.2e\text{-}16$) data. The mean and median HSP ratios for TF from untrimmed reads were slightly higher than those from trimmed reads for both *N. crassa* and *V. inaequalis*. In addition, the untrimmed datasets has the least variation (**Table 3.1**). Multiple comparisons testing between HSP ratio is shown in **Table 3.1**. *Post hoc* analysis indicated that the more aggressive Q20 trimming, produced TFs of inferior quality compared to the Q10. TFs from Q10 and the untrimmed reads yielded no significant difference in HSP ratio. Groups with the same letters are not statistically different. Category one and two trimming strategies were significantly different to the other five categories ($p < 0.01$), for *V. inaequalis*. In both *N. crassa* and *V. inaequalis* datasets, TFs from untrimmed reads produced higher N50 values. Visual assessment of aligned *V. inaequalis* TFs from untrimmed and trimmed reads (category two), reveals missing TFs and incomplete TF reconstruction in the latter as shown in **Figure 3.2**.

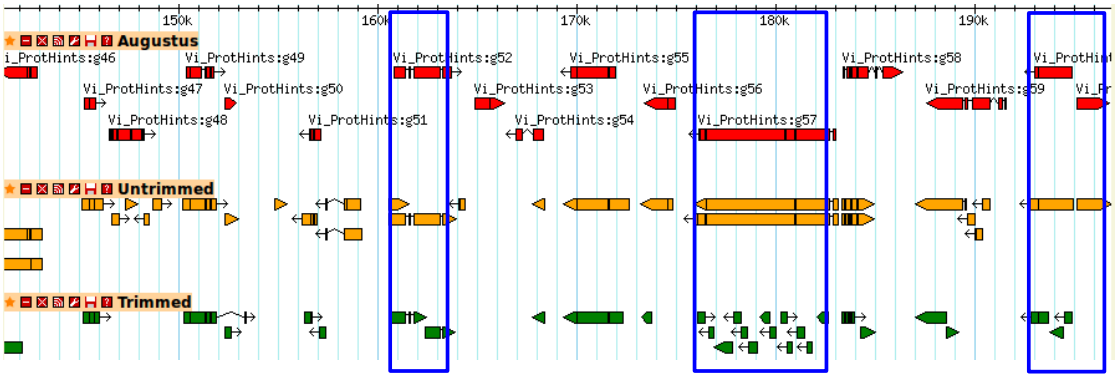


Figure 3.2 A GBrowse snapshot of predicted genes and transfrags (TFs) for *V. inaequalis*. Ab initio gene predictions are shown in red. TFs produced by Trinity with untrimmed and trimmed (category one) reads are shown in orange and green, respectively.

3.4 Discussion and conclusion

In this study, we teased apart the effects of quality based trimming and artifact removal on the quality of *de novo* transcriptome assembly. Quality based trimming approaches are routinely applied on reads generated from NGS platforms. Initial analysis by Garg et al. (2011) suggested that this procedure improved *de novo* transcriptome assembly. However the choice of per base quality score for trimming is subjective and there is no consensus on quality filtering/trimming thresholds since the quality score distribution is non-uniform across samples and the technologies for sequencing are constantly evolving. In addition, the study by Garg et al. (2011) employed a genome assembler which is not suitably optimized for transcriptome reconstruction and this could have had an impact on the interpretation of their results. We observed that, adapter removal was more efficient when performed prior to quality based-trimming. When reads are quality trimmed prior to adapter removal, the sequences may become too short for substring recognition. The higher median and mean HSP ratios and the number of UniProt identified *V. inaequalis* proteins, suggest that TF derived proteins from assembled untrimmed reads aligned with better quality than those from trimmed reads. Additional support for this observation is revealed by the

number of missing annotations in TFs from trimmed *N. crassa* reads. This corroborates anecdotal observation that quality trimming of reads can produce poor assemblies (Paszkiewicz and Studholme, 2010). Untrimmed reads result in more contiguous assemblies, which is probably due to a larger number of paired reads that provide support for connected edges in the de Bruijn graph. Quality trimming affects the quantity of usable reads and for each expression level there is a spectrum of parameters (typically *k*-mer) for optimal transcript assembly (Schulz et al., 2012). In non-model organisms, there is an optimal number of reads balancing coverage and errors (Francis et al., 2013) and aggressive trimming or filtering strategies are likely to affect the coverage dynamics. By applying various trimming or filtering approaches, the number of reads appropriate for assembly is achievable when gauged correctly with an suitable metric such as HSP ratio for evaluating the assembly. While quality based trimming is routinely applied prior to *de novo* transcriptome assembly, our analyses suggest that this could lead to missing annotations and incomplete transcript reconstruction. As such, caution must be exercised given that quality score thresholds for read trimming or filtering are subjective. Promiscuous application of quality score based trimming and or filtering should be gauged and additional effective heuristics assessment of transcript reconstruction be applied for each trimming criteria. Furthermore, our analyses demonstrate that HSP ratio in addition to N50 can assist in selecting the optimal transcriptome assembly.

CHAPTER 4

Identification of scab putative effector candidates and apple resistance genes: a comparison of *Venturia inaequalis* trancriptomes

4.0 Abstract

The availability of transcriptome data for non-model organisms has dramatically increased in the last decade due to the emergence of sequencing-by-synthesis technologies. Their usefulness greatly depends on the quality of the underlying assembly and annotation, which is expected to promote hypothesis driven research and expedite the genome annotation effort. Here we re-examine a recent reference transcriptome assembly of *V. inaequalis* (Indian isolate) and shed light on an important methodology concern, presenting complementary bioinformatics data analyses to ascertain the transcriptomic origin of assembled transcribed fragments. Identifying genes underlying pathogenicity and virulence in the pathogen and resistance in the host will illuminate our understanding of the *Venturia-Malus* pathosystem and facilitate rational design of control strategies.

We updated the published transcriptome of *V. inaequalis* with a method based on successive GMAP and BLAST alignment program iterations to distinguish scab from apple transfrags. Our *in silico* deconvolution approach binned ~50% of the transcriptome to plant origin, of which 233 (0.41%) segregated with *Malus spp* proteins through a non-redundant database search. Transfrags that specifically mapped to the apple genome and/or proteome encoded putative novel methalothioneins and defensins. Sequencing and assembly of a South African isolate produced 39,042 transfrags. About 40% transfrags from the Indian isolate and 46.5% transfrags from the S. Africa isolate had a BLAST match in UniProt. Among these hits, 62.5% (Indian isolate) and 78.6% (S. African isolate) could be mapped to gene ontology terms. Additionally, we predicted 420 (Indian isolate) and 514 (S. African isolate) secretory/signal peptides of which 40 and 30 respectively, had a *bona fide* N-terminal Y/F/WxC-effector motif. Both secretomes showed high numbers of predicted Cysteine-rich proteins and internal repeat sequences, but only 7 proteins in the combine dataset had a RxLR-like motif without a modular architecture.

We generated refined *V. inequalis* transcriptomes, free of host transfrags through a process of iterative sequence similarity screening. These datasets provide the basis of a re-constructed and characterised transcriptome resource specific to the

apple scab which will fuel the bench work and computational analyses that constitute the day-to-day operations for hypothesis driven research in *Venturia-Malus* pathosystem and prospective annotation of the *V. inaequalis* genome.

4.1 Introduction

Sequencing-by-synthesis (SBS) technologies, popularly known as next-generation sequencing technologies (NGS) are playing an increasingly important role in identifying fine scale biological variations and unravelling the genetic basis of host-pathogen interactions (Franchini et al., 2011; Kawahara et al., 2012). Specifically, RNA-Seq (applied in sequencing expressed transcribed fragments) is largely a valid and cost-effective approach for sequencing the transcriptome due to the high functional information content and less repetitive sequences (Li et al., 2012; Miller et al., 2012). By contrast, whole genome sequencing for most eukaryotes is still largely impractical (Parchman et al., 2010). The majority of NGS technologies produce short reads (Wilhelm and Landry, 2009) that are only amendable for comprehensive transcriptome profiling with parallel development of assembly algorithms that piece together the reads into transcript models (Gibbons et al., 2009). The availability of transcript models or EST-like sequences has been very instrumental in expediting gene discovery, annotation (Emrich et al., 2007) and phylogenomics inferences (Yang and Smith, 2013).

Access to transcriptomic sequence data of non-model organisms has dramatically increased during the last few years due to the plummeting cost per base of raw sequence data. Ultra-deep sequencing depths are attainable, enabling identification of lowly expressed transcripts, providing a near-complete snapshot of the transcriptome (Martin and Wang, 2011). Despite this level of unprecedented advantages, RNA-Seq has many computational challenges in managing large data and is not surprising that a large proportion of sequences stored in public repositories still require detailed re-annotation and re-analyses. Access to user-friendly analyses and integrated pipelines, allows scientists with limited bioinformatics training or computational competences to generate a large proportion of transcriptomic sequence (Cantacessi et al., 2010). There exists a problem for biologists carrying out those experiments that are fundamentally fuel by published transcriptome resources with flawed annotations compounded by contaminating sequences. This highlights the need for manual review of sequence

data as much as possible (Hadfield and Eldridge, 2014). However, it is hugely prohibitive to evaluate the fraction of published reference transcriptome datasets that are a miss-representation of their source organisms. The availability of public repositories mainly the National Centre for Biotechnology Information Short Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra), European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) and the DNA Database of Japan Sequence Read Archive (DRA, http://trace.ddbj.nig.ac.jp/dra/index_e.html) are creating avenues to re-purpose raw data and their assemblies. With the development of novel approaches and methodologies for interrogating short read data, it is expected that these datasets can be re-analysed.

A typical area for the application of NGS technologies is in exploring the transcriptomes of the host and the pathogens, and is probably the most cost effective approach to gain a full understanding of host–pathogen interaction at the molecular level (Kawahara et al., 2012; Zhu et al., 2013). Generating cDNA from a mixed library provides an opportunity to enrich genes that are specifically expressed in planta (Bowen et al., 2009; Kucheryava et al., 2008). Identifying such genes equips us with the knowledge of a subset of genes expressed during the life cycle transitions in hemibiotrophic pathogenes and facilitates the introgression of resistance into host varieties (Zhuang et al., 2012). Few studies have attempted to partition plant and fungus RNA-Seq data from mixed infection libraries of phytopathogen (Hsiang and Goodwin, 2003). Intuitively, it becomes imperative to trace the taxonomic origin of reads so that meaningful biological inferences are made without inflation. Discriminating transcribed fragments in mixed libraries is straight-forward when the reference genomes of the host and pathogens are available (Kawahara et al., 2012), and invariably depends on the genomes and sequence data quality. When one reference genome is present, an alignment followed by an "assembly of unmapped reads" strategy can be undertaken (Thakur et al., 2013). One drawback to this approach is that, unmapped reads of the target genome frustrate downstream analysis. This is because a number of variables may affect the quantity of unmapped reads: incomplete reference genome, sequencing errors, adapter remnants, sub-optimal

alignments thresholds etc. Thus successful partitioning of mixed RNA-Seq reads requires additional complementary heuristic approaches. Zhu and colleagues analysed the mixed library of poplar tissue infected with *Marssonina brunnea* through a variety of approaches that relied on the availability of sequence resources and reported different quantities of assembled transcribed fragments (Zhu et al., 2013). Furthermore, in some cases of well annotated genomes, a large proportion of assembled transcribed fragments fail to align for many reasons beyond the scope of the current study (see (Zhao et al., 2011), for a review). Thus the successful discrimination of assembled transcribed fragments from mixed libraries requires a method that integrates proteomic resources. Assembly derived transcripts have huge potential for phylogenomic inferences in non-model organisms (Yang and Smith, 2013) and the requirements for partitioning cannot be underestimated.

Venturia inaequalis (Cooke) G. Wint, is a hemibiotrophic ascomycete fungus that affects apple cultivars primarily but can also infect *Malus* (crabapple). The disease is characterised by necrotic or chlorotic lesions on leaves, various structures of the flower and dark cocky lesions on young fruits. In mature fruit the infection causes "pin-point scab", which is characterised by small black spots. Severe infection can cause the fruit to become malformed, cracked and generally unsightly and therefore unmarketable. Severe early infection can cause defoliation of the blossom and fruit drop, which results in severe reductions in fruit yield (MacHardy et al., 2001). Access to and analysis of genomic resources is expected to illuminate our understanding of the host-pathogen interaction and rational design of control intervention. However, existing genomic resources are fragmentary (Bowen et al., 2009) or poorly represented in the public domain. For the latter case, a comparative analysis of the published transcriptome of *V. inaequalis* (Thakur et al., 2013) to that of a South African isolate, with the aim of collating a dedicated protein set to initiate a genome annotation jamboree produced a large category of unrelated assembled transcribed fragments. As a result of the ensuing refinement of the published transcriptome and coupled with previous findings of the role of secreted proteins in virulence by phythopathogens

(Dean et al., 2005; Morais do Amaral et al., 2012), the *V. inaequalis* secretome was revisited. The main objective of these ensuing analyses was to update the published transcriptome of *V. inaequalis* and perform comparative transcriptomic analysis with a local isolate. The first step towards achieving this goal was to ascertain the origin of assembled transcribed fragments from a mixed infection experiment. We examine the role of the secretome, identified functional domains induced in plant and discuss specific classes of genes that have been shown to be involved in plant pathogenesis.

4.2 Materials and Methods

4.2.1 Fungal isolate, library prep and DNA sequencing

Leaves with freshly sporulating lesions were collected from orchards at the Agricultural Research Council's Experimental Farm, Bien Donné, GPS coordinates (-33.843865, 18.978881) (http://www.onlinebrandambassadors.com/app/map/find-gps/), Simondium, South Africa. Disc sized lesions were excised using a 5 mm cork borer. The leaf discs were agitated thoroughly in 30 mL of sterile distilled water in a 90mm petri dish to release conidia. Approximately 200 µl of conidial suspension (8 x 10^3 conidia ml^{-1}) was spread evenly on 15 g/L water agar. After an overnight incubation at room temperature (~25 ºC), germinating spores were transferred to potato dextrose agar using a scalpel and allowed to grow for a month. Peripherally growing mycelia were inoculated into potato dextrose broth, supplemented with 25 mg/L oxy-tetracycline and incubated with agitation at 100 rpm for 3 months at 21 ºC.

The propagated mycelia were harvested as a pellet by centrifuging a 50 mL aliquot at 6000 x g for 30 minutes. The pellet was then re-suspended in 50 mL distilled water followed by centrifugation as described above and then stored at -80 ºC. Isolation of RNA was performed as described by Menhaj and colleagues (Menhaj et al., 1999) with minor modifications. In brief, 0.5g of frozen mycelia

was grounded to a powder in liquid nitrogen using a glass rod. This was re-suspended in 3 mL of lysis buffer (100 mM Tris-HCl pH 8.0, 600 mM NaCl, 20 mM EDTA, 4% SDS) and 3 mL of Phenol:Cholorform:Iso-amylalcohol (PCI mix in the ratio, 24:23:1). The mixture was placed on a shaker for 20 minutes and centrifuged at 12,000 x g for 10 minutes. The top phase was aspirated and mixed with 8 M LiCl then incubated overnight at 4 °C. The RNA pellet was recovered by centrifugation at 12 000 x g for 10 minutes in a 4 °C chilled environment and dissolved in 100 µl of DEPC treated distilled water. The total RNA was then re-precipitated by adding 1/10th volume 3 M Na-acetate (pH 5.2) and 2.5 volumes (v/v) ice cold absolute ethanol, followed by incubation for 1 hour at -20 °C. This was then pelleted by centrifugation at 12 000 x g for 10 minutes in a 4 °C pre-chilled centrifuge. The RNA pellet was then rinsed in ice-cold absolute ethanol, air-dried, re-suspended in 100 µl DEPC-treated distilled water and stored at -80 °C.

The barcoded RNA-seq paired-end sequencing libraries were prepared using the Illumina TruSeq™ RNA sample preparation kit (Illumina, Inc, San Diego, California). These were sequenced on the Illumina HiScanSQ (Illumina, Inc, San Diego, Carlifornia) to produce three libraries of 100 bp paired-end and one 75 bp single-end reads.

4.2.2 Availability of supporting data

The raw RNA-Seq data, custom PERL scripts (including details on protocol), filtered and untangled assemblies of the S. African isolate generated in this study are available on the South African National Bioinformatics Institute permanent data archive, ftp://ftp.sanbi.ac.za. The *V. inaequalis* draft genome assembly (3,088 contigs of a 37,685,262 bp) can be obtained on request (Celton et al., 2010). The draft genome sequence, CDS, and annotations of *Malus x domestica* were downloaded from http://genomics.research.iasma.it/download.html on 2010-11-01. The list of fungi proteomes that were used to create the 'mock' database is shown in **Table A4.1**, Appendix 1.

4.2.2 Bioinformatics analysis

4.2.2.1 *De novo* assembly

To remove reads potentially derived from ncRNA (Li et al., 2012), we performed BLAST+ (Camacho et al., 2009) against the Rfam database (Griffiths-Jones et al., 2003). Reads with no significant match (Evalue $\leq 10e^{-15}$) were removed using a BLAST lookup function (blastdbcmd) from a BLAST formatted database of raw reads. Due to side effects induced by quality trimming (Mbandi et al., 2014), untrimmed reads with no match to ncRNA were reordered into separate lists of shuffled pairs and singletons after a series of UNIX commands and in-house PERL scripts prior to assembly. *De novo* assembly was performed using TRINITY (release 2012-06-08; *k*-mer 25) (Grabherr et al., 2011) with default parameters and a minimum transfrag length of $\geq$ 100 bp.

4.2.2.2 Computational binning of plant (host) transfrags in the transcriptome assemblies

Homology based approaches have been successfully applied for taxonomic assignment of expressed sequence tags from mixed infection libraries (Hsiang and Goodwin, 2003; Kruger et al., 2002; Zhu et al., 2013). Both transfrags from the Indian and S. African isolates were each subjected through a process of iterative sequence based alignment screening. The origins of contaminating host transfrags was then assessed by visualizing the species distribution of best BLAST hit for each assembly against the non-redundant database (downloaded on 2011-11-17, ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz), E-value $\leq$ 1e-10 using the Blast2GO suite (Conesa et al., 2005; Götz et al., 2008). Then the assemblies were mapped to the apple genome (Velasco et al., 2010) using GMAP version 2013-10-04 (Wu and Watanabe, 2005) at 95% coverage and 95% identity. Given that transfrags may fail to align to the genome due to mis-assembly or suboptimal alignment parameters, the unmapped transfrags were screened against protein sequences. A

'mock' host-pathogen database was created that comprised the apple proteome (Velasco et al., 2010) and the proteomes of seven fungi that represented > 50% of best BLAST hits in the raw assemblies: *Pyrenophora tritici*, *Phaeosphaeria nodorum*, *Leptosphaeria maculans*, *Mycosphaerella graminicola*, *Botryotinia fuckeliana*, *Saccharomyces cerevisiae* and *Pyrenophora teres*. Unmapped transfrags were subsequently aligned to the 'mock' host-pathogen database using BLASTx (Camacho et al., 2009) with E-value ≤ 1e-10. To improve the fidelity of the remaining transfrags collection, we aligned them to the *S. cerevisiae* genome with alignment filtering at 95% coverage and 95% identity using alignment parameters obtained from summary statistics of intron distribution in Ascomycetes (Kupfer et al., 2004): min-intron length = 20, max-intron length = 2000. A pictorial representation of the workflow is shown in **Figure 4.1**. Sequences that did not segregate with the host and yeast proteome and/or genome were considered putative *V. inaequalis* transfrags. A BLASTn search (E-value 0.0001) and GMAP alignment (50% coverage and 50%) of the identified host transfrags against the draft genome assembly for *V. inaequalis* (Celton et al., 2010) was performed as an internal validation of the screening method.

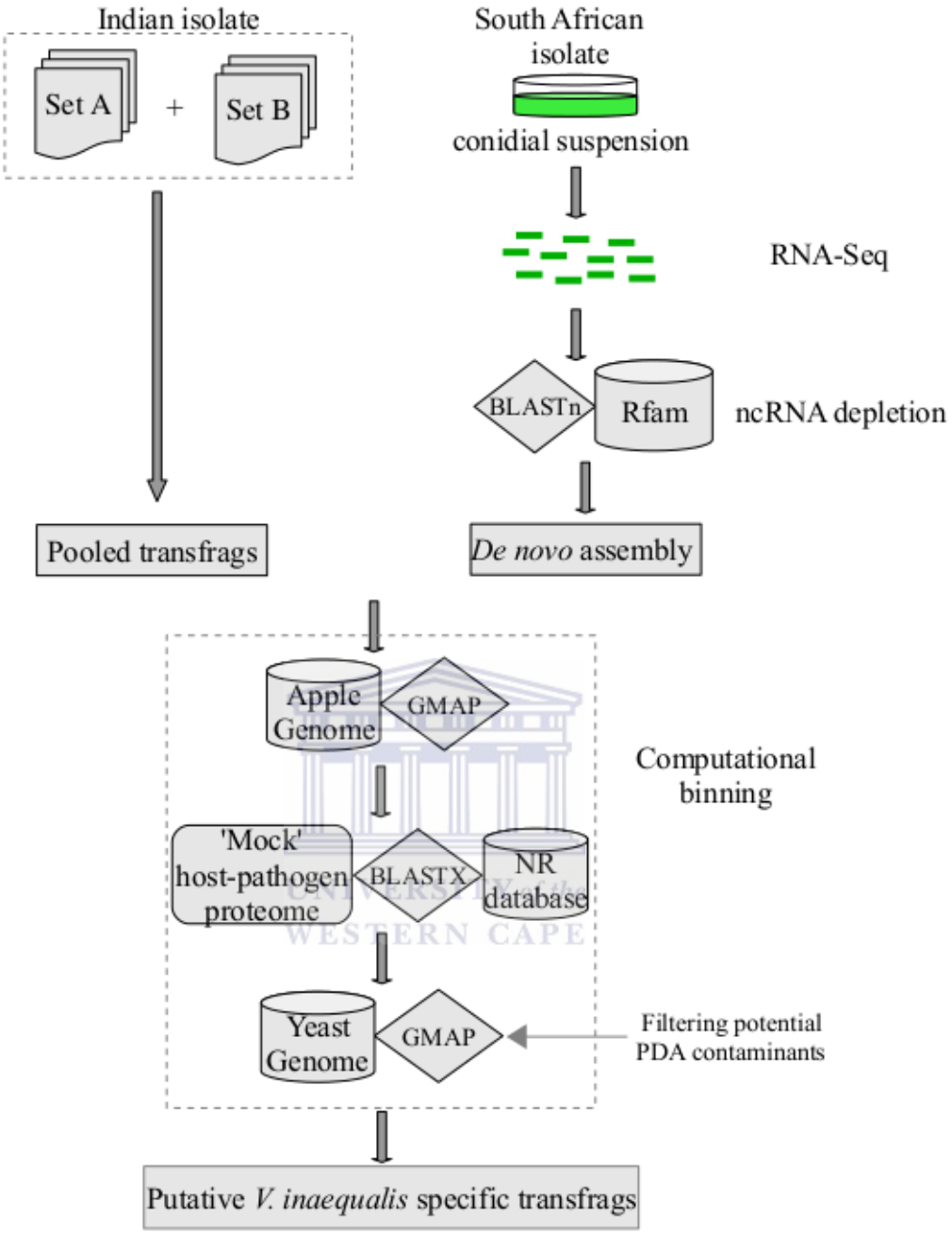


Figure 4.1 *In silico* separation of mixed host-pathogen transfrags. The origin of each transfrag is inferred by its segregation between the host and fungi sequences. Transfrags that do not align to the Apple genome nor produce significant hits with apple proteins are considered to be fungi related. The yeast genome served as additional filter for improving the fidelity and uniqueness of the transfrags.

4.2.2.3 *V. inaequalis* protein prediction

Transfrag derived proteins were obtained using a controlled conceptual translation for reliable prediction in three stages (**Figure 4.2**). Proteins with a BLAST match against the non-redundant database were translated in the frame of their best BLAST hit using EMBOSS (Rice et al., 2000) with in house PERL scripts. For transfrags without a BLAST match, we predicted the proteins using ESTSCAN version 3.0.3 (Iseli et al., 1999) with $hightaxo = "Dothideomycetes" for building the matrix with sequences from FUNGI Refseq release 60 and EMBL releases 116. Sequences for which we could not predict a protein using ESTSCAN were translated into six open reading frames (ORF) with EMBOSS and the longest ORF selected with a custom PERL script.

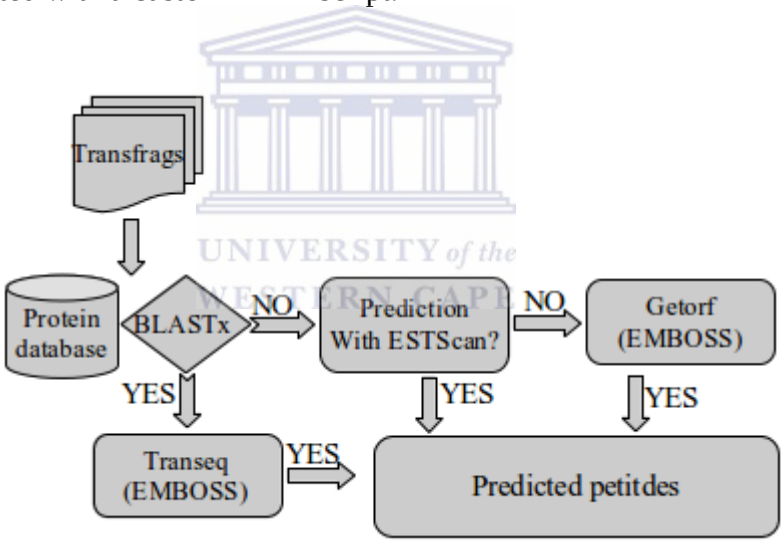


Figure 4.2 Schematic representation of protein prediction. Proteins are predicted in three stages, namely: Transeq (dependent of best blast hit frame), ESTScan (sequencing error and frameshift tolerance) and longest ORF (six frame translations). Subsequent stages are implemented if no protein is predicted upstream.

4.2.2.4 Assessing the transcriptome and inferring orthologous groups

An assessment of transcriptome completeness was performed using BLASTx (Camacho et al., 2009) with soft masking (-F "m S"), between putative *V. inaequalis* transfrags from the S. African and Indian isolates and an inventory of proteins belonging to 437 core genes of *Magnaporthe grisea* (http://korflab.ucdavis.edu/Datasets/genome_completeness/data, downloaded 2014-02-10). These low copy number genes are highly conserved and are a good proxy for estimating completeness of gene space in hemibiotrophic fungi. The number of high scoring segment pairs with minimum
E-value ≤ 1e-10 was indicative of completeness.

A complementary transcriptome content analysis was performed between predicted proteins from the transcriptomes of the S. African and Indian isolates of apple scab, and four proteomes of well-studied hemibiotrophic fungi (*Colletotrichum graminicola*, *Leptosphaeria maculans*, *Magnaporthe oryzae*, *Zymoseptoria tritici*, *Pyrenophora tritici-repentis*) using Inparanoid (version 4.1, Remm et al., 2001). Orthophylogenetic distances where computed using two complementary formulae; equation 2 (Ananthasubramanian et al., 2012) and equation 3 (Berglund et al., 2007) as follows:

Eq2

$$\text{Distance (A, B)} = 1 - \frac{(\text{№ proteins in A with orthologs in B} + \text{№ proteins in B with orthologs in A})}{\text{№ protein in A} + \text{№ proteins in B}}$$

Eq3

Average orthology distances (dAB + dBA)/2 where the distance from species A to B, dAB is,

$$\text{Distance (A, B)} = 1 - \frac{(\text{№ proteins in A} - \text{№ proteins in A with orthologs in B})}{\text{№ proteins in A}}$$

4.2.2.5 Annotating assembled sequences

The filtered or untangled assemblies were screened against the UniProt database (FUNGI) release-2013_11. Sequences with no BLASTx hits were searched against UniProt knowledgebase release-2013_11. The sequences were annotated by the best-BLAST annotation transfer method (Jones et al., 2005), where annotations are transferred to a query from its highest-scoring BLAST hit. To estimate the proportion of sequences that match to unique genes, we check for redundancy in BLAST hit accessions. Using a simple PERL script, we mapped GO terms from the gene association file version 2.0 obtained from the Gene Ontology Association database file (GOA, ftp://ftp.geneontology.org/go/gene-association). No annotation is transferred if there were no hits at the E-value cutoff 1e-3. The predicted peptides were subjected to domain analysis by InterProScan (Quevillon et al., 2005) with -appl = PfamA-27.0. The MEROPS database was interrogated for peptidase families using BLATCH BLAST with E-values 1e-4 (Rawlings and Morton, 2008).

4.2.2.6 Defining the secretome

Predicted proteomes were analysed for putative modulators of the host cells and immunity, using transfrag derived peptides ≥ 70 amino acids. A combination of tools is strongly recommended for defining the secretome (Klee and Ellis, 2005). Firstly, proteins were predicted as classical secreted proteins using SignalP 4.1 (Petersen et al., 2011). The existence of transmembrane helices (at least one) after the leader sequence cleavage site in proteins positive for SignalP was inferred with TMHMM2.0 (Krogh et al., 2001) and the prediction of subcellular location in Fungi was done with WolF PSORT v0.2 (Horton et al., 2007). Sequences without a transmembrane region showing a positive signal peptide cleavage site were considered as candidate secreted proteins provided it had 'extr' as one of the sites in the ranked list retrieved by WolF PSORT. The final set of candidate secreted proteins was searched against the list of known effectors in *V. inaequalis* (Bowen et al., 2009) with tBLASTn (E-value ≤ 1e-2). The secretome was scanned

for presence of the degenerative Y/F/WxC-motif (Godfrey et al., 2010) within the first 24 amino acids circumscribed after the signal peptide cleavage site. We also examine the presence of a RxLR pattern (Win et al., 2007) in the region between positions 31 and 57 from the N terminus for predicted secreted ORFs with a signal peptide cleavage site before amino acid position 30, using custom PERL scripts. In addition, we estimated the proportion of putative secreted proteins that are Cysteine-rich with custom PERL script and inferred the presence of internal tandem coding repeats using T-REKS (Jorda and Kajava, 2009) at a threshold similarity of 0.75 and zero indels.

4.2.2.7 Analysis of plant (apple) related sequences

We investigated the roles that transfrags, segregating with plant related material may play in the *Venturia-Malus* pathosystem. For this, the predicted apple CDS were mapped to the apple genome using GMAP version 2013-10-04 (Wu and Watanabe, 2005) at 95 % coverage and 95% identity thresholds. We extracted alignment coordinates of transfrags that aligned with the same level of stringency and overlaid them with those of the aligned CDS. Using an in house PERL script, we performed a pairwise comparison on mapped coordinates. The annotation of a transfrag is inferred from that of an overlapping CDS and then confirmed if the same transfrag produced a significant BLASTx hit with the corresponding ORF. Apple transfrags that neither overlap with predicted CDS nor produce significant BLASTx hit to a protein in the apple proteome were extract and clustered using BEDTools v2.14.2 (Quinlan and Hall, 2010). The regions enclosed by the outer boundaries of each cluster were labelled as novel putative gene loci. The transfrags constituting these loci were subjected to domain analysis by InterProScan (Quevillon et al., 2005).

4.3 Results

4.3.1 Sequencing yield and assembly

With the purpose of generating a reference transcript catalogue of *V. inaequalis* for comparative genomic analysis, 63,161,616 (100 bp, paired-end) and 8,017,576 (75 bp, single-end) Illumina RNA-seq reads were produced from apple scab that was isolated from South Africa. Of these, 53,940,202 (40.15%) reads with significant similarity to non-coding RNAs to Rfam database sequences where removed. The remainder, 80400606 (59.85%) longer than 74 bp were *de novo* assembled into 46,349 sequences. The transcriptome assembly generated 46,265 unique transfrags after removal of redundant sequences. **Figure 4.3** shows the composition of the transfrags as revealed by the species distribution of their best BLAST hits. We note that majority of hits against the non-redundant database; suggest extensive contamination of the Indian transcriptome assembly with plant related sequences.

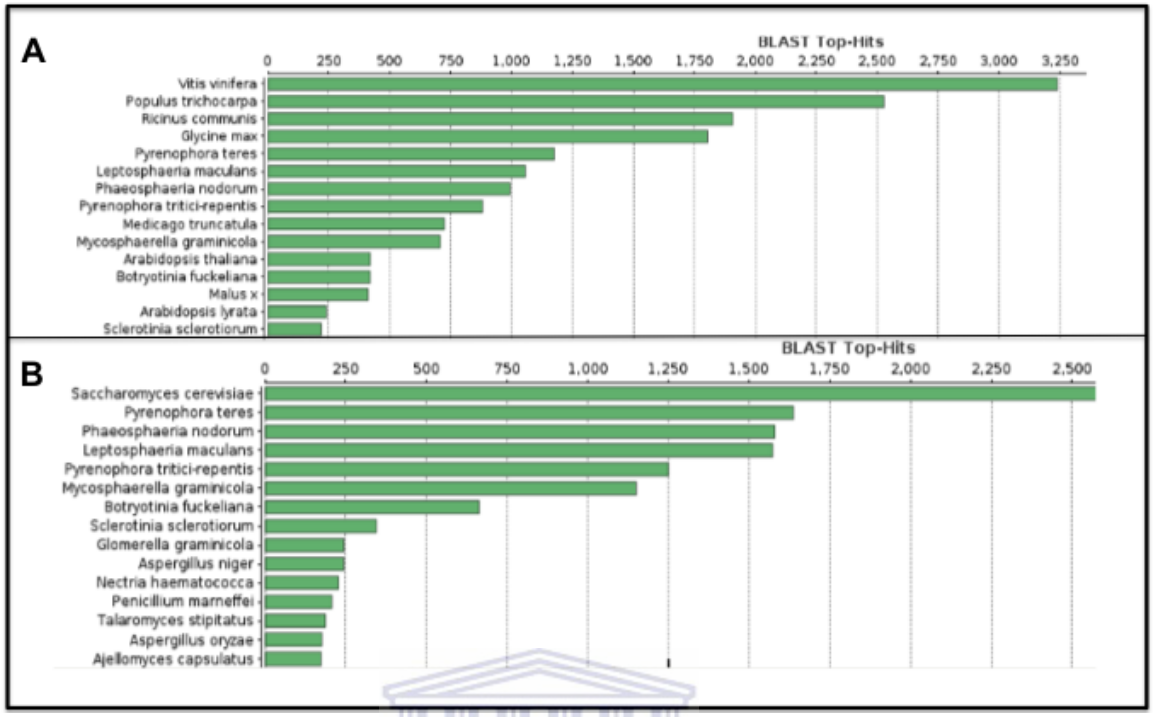


Figure 4.3 The distribution of species representing the best similarity hit in unfiltered assemblies. The best BLAST hits from unique transfrags in the Indian and S. African assemblies against NCBI non-redundant protein database are shown for the first top 15 species. The Indian transcriptome (panel A) shows very high proportion of hits to plant species, while the South African Isolate (panel A) is mainly of fungal origin.

4.3.2 Untangling the transcriptome assemblies

To determine the origin of sequences assembled from a mixed fungus-infected library such as the Indian isolates, we implemented a highly specific computational workflow for *in silico* depletion of transfrags. Transfrags are computationally binned based on genome alignment and their segregation according by affinities to the proteome of the host and potential growth media contaminant. The distribution of species from the best BLASTx hit after removal of apple and yeast related sequences is shown in **Figure 4.4**, with the largest number of hits from *P. teres*.

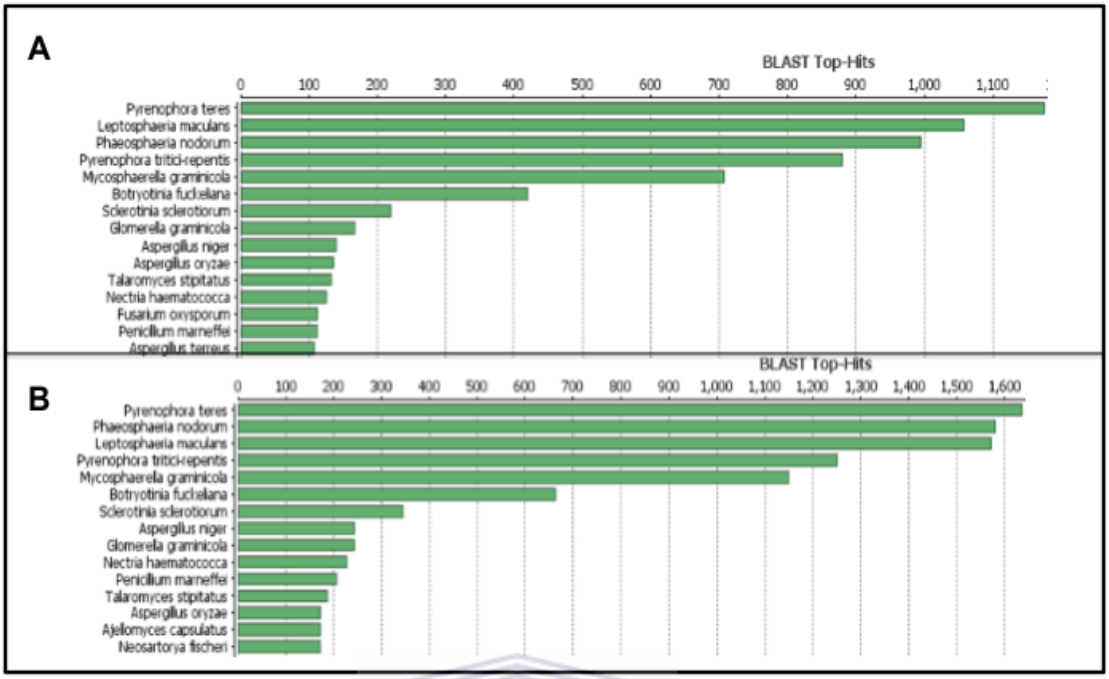


Figure 4.4 The distribution of species representing the best similarity hit in the filtered assemblies. The best blast hits from the assemblies against NCBI non-redundant protein database after depletion of apple and yeast transfrags is displayed for the first 15 species. The majority of hit in the Indian (top panel) and South Africa (lower panel) are from Dothideomycetes.

A summary of transfrag removed from the assemblies following *in silico* separation are shown in **Table 4.1.** A total of 28,298 (~ 50%) transfrags from the Indian assembly could be mapped to the apple genome. Of these, 1,199 (2.1%) produced significant BLASTx hits to apple proteins in the 'mock' host-pathogen database. By comparison, we applied the same screening procedure on the S. African isolate transcriptome assembly. A startling 2.14% of transfrags from the S. African isolate assembly could also be traced to apple. However, a large number of transfrags belonging to the latter produced significant alignments to the yeast genome (~13%) and proteins (0.51%). As an internal validation of the screening workflow, we screened the identified host transfrags against the *V. inaequalis* draft genome (Celton et al., 2010) at low stringency and found few hits: 442 (~1.6%) by BLASTn and 15 (0.05%) by GMAP from the Indian assembly.

Table 4.1 Summary of iterative depletion of contaminating transfrags in *V. inaequalis* transcriptome assemblies.

Target	Indian	South African
Malus x *domestica* genome	27,099	901
Malus x *domestica* proteome	1,199	88
Saccharomyces cerevisiae genome	5	5997
Saccharomyces cerevisiae proteome	134	170
Saccharomyces sp proteins in non-redundant (NCBI) database	1	67
№ of putative *V. inaequalis* TF	28,444	39,042

To assess the gene space coverage we align the 28,444 and 39,042 filtered transfrags from the Indian and S. African isolates respectively to a set of core eukaryotic genes (CEGs). Although initially used to assess gene space in newly sequenced genomes, CEGs are reliable indicator of completeness of gene space in *de novo* transcriptome assemblies of eukaryotic species (Chow et al., 2014). Both transcriptome assemblies detected all 437 CEG proteins from *M. grisea* (e-value ≤ 1e-10). However, the number of transfrags aligning with ORF coverage ≥ 50% is marginally higher for the Indian (97.9%) than the S. African (94.5%) assembly. We note that the filtered Indian assembly is more contiguous with a larger N50 value (2,170) as shown in **Table 4.2**. Thus, these results support the completeness and suitability of the fungi enriched transcriptome assemblies for downstream analysis.

Table 4.2 Transcriptome completeness and assembly attributes of putative *V. inaequalis* transfrags

Assembly	Largest transfrag	Median	Mean	N50	№ of CEG hits	№ of CEG hits where ORF coverage ≥ 50%
Indian	27,669	228	831	2,170	437	428 (97.9%)
S. African	29,034	256	651	1,574	437	413 (94.5%)

4.3.3 Analysis of plant (apple) genes involved in resistance

Given that genes are expressed, up or down regulated in response to fungal material, we annotate the plant related transfrags obtained from the Indian transcriptome in search of genes that may play a role in the host-pathogen interaction. We identified 3,722 transfrags that aligned to the apple genome without overlapping a CDS and did not produce a significant BLAST hit to the apple proteome. Genome based clustering of these sequences generated 3,527 novel putative gene loci. About 65 transfrags constituting these gene loci could be annotated with InterProScan. Of particular interest are transfrags with meaningful descriptions namely: 5 metallothioneins, 3 defensins and 5 C-terminal Cysteine knots which may be involved in resistance. The results of this analysis are presented in the **Table A4.2**, Appendix 1.

4.3.4 Secretome analysis and annotation of fungi (*Venturia*) transfrags

In the absence of an annotated reference, transfrags were compared by BLASTx against UniProt database of peptide sequences, the most comprehensive and well annotated collection of proteins, thus identifying 11,373 (40%) and 18,134 (46,5%) transfrags from the Indian (IN) and South African (SA) isolates respectively with significant similarity to known proteins as shown in **Table 4.3**.

Table 4.3 Annotation summary of transcriptome assemblies and secretome for two *Venturia* isolates.

	Indian	South African
BLAST annotation		
№ of TF with at least one BLAST match	11,373	18,134
№ of unique genes	10,824	14,262
№ of TF with GO	7,126	11,146
№ of predicted proteins with interpro domain	6,583	8,386
Attributes of the secretomes		
№ of TF with Signal Cleavage site (+)	558	730
№ of TF with Transmembrane (-)	460	575
№ of candidate TF constituting the secretomes	420	514
№ of TF with candidate Y/F/WxC-motif	30	40
№ of TF with Cysteine > 5% in mature protein	82	97
№ of TF with internal tandem repeats	57	56
№ of TF with candidate RXLR effector pattern	5	2

Although this is less than 50%, BLASTx searches identified a total of approximately 10,824 (IN) and 14,262 (SA) unique protein accessions, indicating that the untangled transcriptomes represented a substantial fraction of *V. inaequalis* genes. The orthophylogenic distances suggest that *V. inaequalis* is most related to *P. tritici-repentis* in comparison to all the hemibiotrophic fungi use in this study, **Table A4.3**, Appendix 1.

By inferring the presence of a leader sequence cleavage site, absence of transmembrane helix and evidence of extracellular deposition, we were able to define the secretome from conceptually translated peptides, as shown in **Table 4.4**.

Table 4.4 Comparison of protease distribution between *V. inaequalis* and selected hemibiotrophic fungi

	Peptidase family						
species	Aspartic	Cysteine	Glutamic	Metallo	Serine	Threonine	Unknown
V. inaequalis (SA)	18	64	0	95	117	15	4
V. inaequalis (IN)	19	47	0	73	95	11	3
P. tritici repentis	20	65	1	125	141	19	9
L. maculans	15	62	1	118	123	20	9
C. graminicola	17	64	1	142	176	21	9
M. oryzae	22	58	2	128	159	21	9
Z. tritici	28	52	4	114	165	19	8

A total of 420 (IN) and 514 (SA) transfrags-derived proteins were identified as *bona fide* secreted peptides. When the secretomes of *V. inaequalis* isolates are compared with those of selected hemibiotrophic fungi, we realised that it is about half the size of the latter (**Figure 4.5**).

The predicted secretomes had 40 (IN) and 30 (SA) transfrag-derived proteins containing the Y/F/WxC-effector motif. Both isolates showed the same preponderance of internal tandem repeats in their secretomes. At > 5% Cys threshold, 82 (IN) and 97 (SA) proteins were identified. Amongst the list of PFAM domains with more than one occurrence, PF12296 (Hydrophobic surface binding protein A) and PF06985 (Heterokaryon incompatibility protein, HET) were exclusively found in the Indian isolate secretome. Worthy of mention is PF00172 (Fungal Zn(2)-Cys(6) binuclear cluster domain) which occurs multiple times in the Indian and only once in the South Africa isolate secretome.

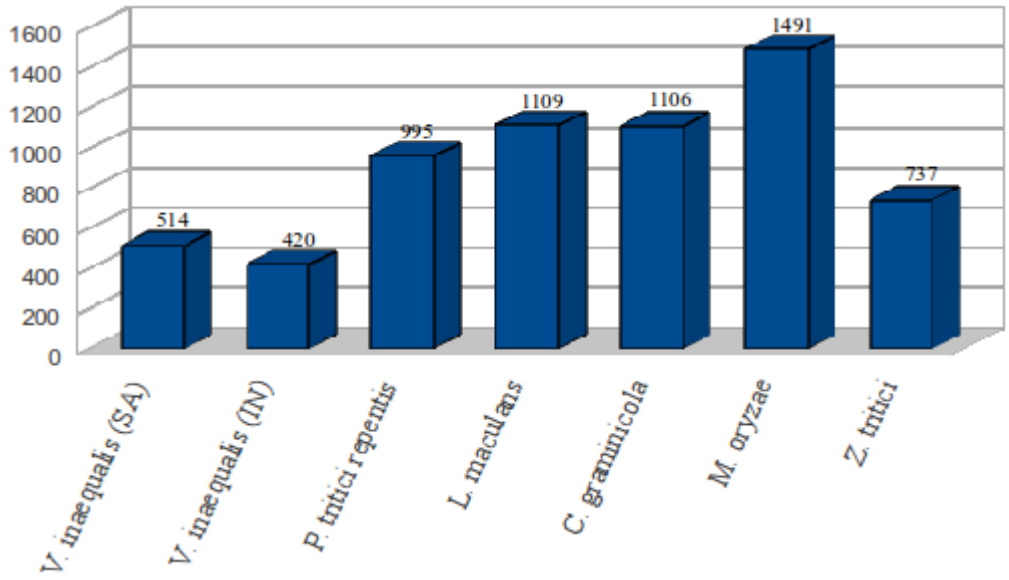


Figure 4.5 Distribution of secreted proteins in *V. inaequalis* and selected fungi. The secretome of *V. inaequalis* is compared with those of other fungi, suggesting that only half of the secretome may be represent in both isolates.

Searches against the MEROPS database identified the seven categories of peptidases in *V. inaequalis*. **Table 4.4** shows a comparative cumulative distribution of peptidase families in *V. inaequalis* and selected hemibiotrophic fungi. The major contributors of the peptidase families are serine, metallo and cysteine peptidases. However, glutamic peptidase was not found in *V. inaequalis* and yet ranged from 1 (*P. tritici repentis*) to 4 (*Z. tritici*) suggesting their rare occurrence in hemibiotrophic fungi.

The number of peptides inferred from the transcriptome assemblies which could be assigned InterPro domain, reveal markers potentially crucial in fungal biology (**Table A4.4**, appendix 1). Prominent amongst the ranked list of interPro domains are the ATP-binding-cassette (ABC)-type transporters and cytochrome P450 gene families, as these are implicated in aspects of host defence and chemical detoxification.

4.4 Discussion

The *Venturia-Malus* pathosystem is arguably the most studied host-pathogen interaction in woody plant species (Cova et al., 2010). In addition, the disease has been known since the renaissance (MacHardy et al., 2001) and there is still no comprehensive genomic/transcriptomic resource to drive the development of new strategies in apple breedomics. South Africa is a relatively small apple grower in terms of global hectares, but is a major volume exporter in global terms (http://www.daff.gov.za, 2011-12), experiencing challenges with apple scab infection. To this end, we explore the decreased costs of procuring sequence data from the rapidly evolving sequencing-by-synthesis technologies to generate RNA-Seq data from a South African isolate of *V. inaequalis*. During the course of this work, Thakur and colleagues, (Thakur et al., 2013) released a 'reference' transcriptome assembly of apple scab. As a result, we aimed at establishing a comprehensive collection of transcript models that will expedite the genome annotation effort (Celton et al., 2010).

Our examination of the Indian isolate assembly data (Thakur et al., 2013) did not reflect what would normally be observed in fungi. The authors claimed that only unmapped reads to the apple genome were assembled, but we identified ~50% transcripts that had plant origin. Discriminating the host and pathogen in a mixed pool of infected tissues has well established procedures (Zhu et al., 2013). We designed an *in silico* de-convolution approach incorporating existing genomic and proteomic resources to delineate the taxonomic origin of assembled transcribed fragments, from the inadequately filtered RNA-Seq data of pooled reads derived from the host-free and in planta libraries. The high specificity of our *in silico* subtraction approach is reflected in the species distribution of best-blast hits before (**Figure 4.3**) and after (**Figure 4.4**) removing plant and putative contaminating transfrags. In addition, an insignificant number of identified plant transfrag aligned at very low stringency to the outsourced *V. inaequalis* draft genome (Celton et al., 2010). At this threshold, some uniquely mapped transfrags had > 12 putative exons suggesting that they are spurious alignments. This reveals

the strength of our *in silico* subtraction approach which can be readily applied in situations of mixed RNA-seq library analyses. The presence of host transfrags in the Indian transcriptome (Thakur et al., 2013) is probably due to suboptimal alignment thresholds during mapping of RNA-seq reads to the apple genome (Velasco et al., 2010). However, we found traces of host transfrags in the SA assembly possibly from inadvertent contamination at the sequencing facility that is involved in sequencing pomaceous fruits. Hadfield and Eldridge (Hadfield and Eldridge, 2014), suggested that in a sequencing facility handling more than one target organism, there is a high chance of cross-contamination.

The evidence of finding host transfrags comprising approximately 50% of the Indian assembly strongly indicates almost every analysis carried out on the Indian assembly (Thakur et al., 2013) especially annotations relating to protein families, secreted peptides and petidase families are inflated and do not correlate with what is likely the biology of apple scab. For example, a cursory glance into the content of merged putative plant transfrags (IN and SA) lead to the identification of transfrags involved in resistance, signalling and pathogenesis with near absolute sequence similarity and alignment coverage to apple genes. For example, MDP0000466190 was identified in plants transfrags and has recently been associated with the development of ontogenic resistance in apple (Gusberti et al., 2013). In addition, we identified 3,722 novel genomic loci on the apple genome that did not overlap with CDS prediction (Velasco et al., 2010). These transfrags may represent cognate genomic contaminants. However, 65 of them had predicted interPro domains some encoding metallothioneins and defensins. In a case where they represent *bona fide* transcribed fragments, predicting genes from their mapped loci will provide a detailed glimpse into the compendium of existing apple genes which may be specifically expressed in the *Venturia-Malus* interaction.

The study by Thakur and colleagues (Thakur et al., 2013) suggested that the peptidase inhibitor family was exclusively present in *V. inaequalis*. In contrast, our analyses show that this family occurs in *P. tritici-repentis* and *M. oryzae*.

Querying the MEROPS database of proteome-wide annotation of the peptidase family reveal that the inhibitor families are well represented in *P. tritici-repentis* and *M. oryzae*. Amongst the selected fungi used in this study, *V. inaequalis* is much more closely related to *P. tritici-repentis* than the other fungi in this study, which is in line with previous findings (Thakur et al., 2013). Contaminating transfrags-derived peptides would rarely have a one-to-one relationship in a reciprocal BLAST which is the main criterion for selecting candidate orthologs for single linkage clustering (Remm et al., 2001) and would not inflate the orthologous groups.

The interpro domain architecture between the IN and SA transcriptomes are very similar and amongst the list of highly represented signatures, two domains: ABC transporters and cytochrome p450 are very well represented. ABC transporters have been strongly implicated in resistance to toxic compounds of natural and artificial origin (Zwiers et al., 2003). Long-term and extensive fungicide use has led to decreased sensitivity of *V. inaequalis* to multiple fungicide modes of action. Similarly, cytochrome p450 genes play a role in secondary metabolism and have been implicated in resistance by some field resistant strains of *V. inaequalis* (Schnabel and Jones, 2001). The occurrence of P450s and ABC transporters represent interesting candidates for functional characterization in *V. inaequalis* interactions with apple defence chemicals. The number of unique BLAST hits and retrievable GO terms is higher for the SA isolate than the IN isolate, albeit the later was generated from a pooled mixed infection library. We note that the Indian assembly (Thakur et al., 2013) was generated after a series of extensive post assembly *de novo* clustering. *De novo* clustering approaches may assign transfrags to clusters with biologically unrelated representatives, thereby reducing the number of retrievable unique BLAST hits (Haznedaroglu et al., 2012).

The number of transfrags in the predicted secretome was comparably lower in the Indian (420) and S. African (514) isolates compared to other fungi analysed in this study. Given that the transcriptome is dynamic, not all genes responsible for the secretome would be sampled. Because of premature termination of ORFs during

protein prediction, some secreted proteins may have been eliminated for being < 70 amino acids in length. A suitable comparison would require a proteome-wide comparison of predicted genes from the *V. inaequalis* genome (Celton et al., 2010). We predicted a lower number of secreted proteins from the Indian tramscriptome compared to previous observation (Thakur et al., 2013) because the eventual location of proteins was used to define proteins that have a high probability of being secreted into the extracellular spaces. A tBLASTx search of the predicted transcriptome identified only 8 of the known *Venturia* effectors (Bowen et al., 2009). This number is however lower than that identified by Thakur and colleagues (Thakur et al., 2013) due to our post-assembly filtering of the plant related transfrags and the stringency in our secretome analysis but also important is that they had used the nucleotide sequence, some of which produced shorter peptides for reliable prediction of signal peptide and transmembrane regions. In addition, one such effector (contig_57585, **Figure A4.1**, Appendix 1) is not of fungal origin and as such, cannot be a candidate effector. To provide some annotation for the predicted secretome we surveyed for functional domains and the presence of the consensus and/or degenerative RxLR-like and Y/F/WxC effector motifs. Although we observed a preponderance of Y/F/WxC candidates, there was a depletion of the RxLR motif. We attribute the low frequency of RxLR-like motifs to have occurred by chance which is in accordance with the observation that the RxLR pattern may not be a suitable marker for effector identification in *V. inaequalis* (Thakur et al., 2013). We note however that, the small number of candidate Y/F/WxC-like effectors in *V. inaequalis* is not uncommon, given that non-haustoria forming fungi have an obvious under-representation of genes encoding this effector (Godfrey et al., 2010). Both secretomes, had a high number of cysteine-rich proteins and contain near perfect internal repeats suggesting a major contribution to virulence akin to the secretome repertoire observed in *Piriformospora indica* with no obvious inclination to a particular effector delivery motif (Rafiqi et al., 2013). The high frequency of zinc cluster domain in the Indian isolate compared with the South African isolate, may suggest an up-regulation in response to plant material which will be necessary in modulating *hyphal growth*. The unique presence of zinc cluster or binuclear

proteins in fungi (MacPherson et al., 2006) may represent an interesting anti-fungal target and characterization of the function of this transcription factor in virulence and identification of its target genes may constitute a promising basis for the better understanding of *Venturia-Malus* interactions. Further observation suggesting in planta expressed assembled-transcribed fragments is the exclusive presence of PF06985 and PF12296 domains in the Indian isolate, probably required for morphogenetic differentiation in response to plant material. Homologs of hydrophobic surface binding proteins such as *HsbA* Pfam motif (PF06985) are differentially up-regulated throughout appressorium development in *M. Oryzae* (Soanes et al., 2012).

4.5 Conclusion

Transcriptome annotations fuel the bench work for genome annotation and starting point for operations in molecular biology and bioinformatics. This study has greatly resolved our comprehension of *V. inaequalis* secretome and identified novel putative Y/F/WxC-like effector genes which may potentially modulate the host environment. In addition, we found putative novel gene loci on the the apple genome encoding methalothionins and defensins that could participate in plant defence and pathogen attack. They remain an interesting find and future work will predict the gene structure and function in these novel loci. Once the genomic sequences of *V. inaequalis* is available and other fungal species are published, the secretome predictions can be further refined. The transcriptomes of these isolates have advanced our still very limited understanding of how plant pathogens recognize and respond to the host environment.

CHAPTER 5

Identification of genes required for invasion and Melanin Biosynthesis: An annotation and analysis of the draft genome sequence of *V. inaequalis*

5.0 Abstract

Venturia inaequalis is a heterothalic, hemibiotrophic Ascomycetes fungus which forms subcuticular hyphae and derives nutrients from the underlying host tissue. Mechanical and enzymatic processes play a crucial role in the initial establishment of infection. Its subcuticular pathogenic lifestyle is unique and the genes controlling this are largely unknown. The availability of a draft genome is an ideal recourse in identifying virulence factors involve in pathogenesis. Such genes will facilitate rational design and implementation of control measures.

The IFRAT pipeline (see chapter 2) rendered 17,773 and 17,171 *bona fide* transfrags from transcriptome assemblies a South African and Indian isolate respectively. As a result, 27,578 transfrags (South African (16,387) and Indian,(11,191) that aligned at high stringency to the draft *V. inaequalis* geneome were used by AUGUSTUS to predict 11,692 protein-coding genes. Of these, 9,319 ORFs retrieved BLAST hits against the UniProt database resulting in 8,992 unique accessions. Clustering of 24 fungi proteomes with *V. inaequalis* generated 13,575 ortholog groups that included 8,867 from the latter. Comparative genomic analysis identified two hydrophobin genes with characteristic idiosyncratic cysteine residues that are required in disulfide bridge formation. The arsenal of carbohydrate degrading enzymes is reminiscent of a hemibiotrophic lifestyle and included 601 carbohydrate-binding modules and enzymes, grouped in 86 CAZyme families. Comparison of the number of cell wall degrading enzymes amongst biotrophs, hemibiotophs and necrotrophs, shows an expansion in *V. inaequalis* polysaccharide lyase, PL4 (6) and carbohydrate esterase, CE5 (28) families that are required for saccharification. No glutamic peptidases were found in the predicted proteome. The PKS and accessory proteins necessary for melanin biosynthesis were also identified. InterPro annotation of the latter suggest that the biosynthesis of melanin is via 1,8-dihydroxynaphthalene.

The identification of these genes represents novel targets for the rational design and dissemination of control strategies. Simultaneous, multiple genes silencing that utilize these marker genes, will enable the development of high through-put screening for functional genomics.

5.1 Introduction

The heterothalic, hemibiotrophic Ascomycetes fungus *Venturial inaequalis* is a fungal pathogen that causes scab infection in Apple (*Malus* X *domesticus*) and other members of Maloideae (Bowen et al., 2009). Scab infection is characterised by precocious foliage fall and deformation in fruit size and shape, leading to serious losses in crop yields and marketability worldwide. The cost of breading and selecting resistant cultivars together with routine chemical input and their associated environmental hazards aggravates the negative economic impact in growing saleable apples. The association between the host and the pathogen has evolved over a long-time: infection is non-lethal and the host can produce fruits despite numerous bouts of infection (MacHardy et al., 2001).

Apple scab resistance is one of the most well characterised phytopathosystems in woody species (Cova et al., 2010; Bowen et al., 2011). Insight into the genetic basis for the association between *Malus* and *V. inaequalis* exemplifies a gene-for-gene relationship. In this model, the outcome of the interaction is controlled by a combination of inheritable resistant (*R*) gene of the host and corresponding virulence determinants of the pathogen, avirulence (*Avr*) gene (Bénaouf and Parisi, 2000). Efforts made at introducing resistant genes in the host are usually time consuming and challenged by the subtle biology and fitness rate of the pathogen. The catalogue of R genes known to date is small (Broggini et al., 2007) and much of the disease versus resistant pathway is still unknown.

The apple genome sequence is expected to reinforce the development of molecular tools that would enable high-throughput segregation analysis in genetic studies and precise quantitative and qualitative traits selection for appealing varieties (Velasco et al., 2010; Zhang et al., 2012). However, a more comprehensive breeding program should incorporate combining resistance to disease as well as maintaining desirable fruit characteristics. It therefore seems essential that, the availability of the pathogen genomic resources is invaluable in providing a more holistic approach in apple breedomics. Construction of

expressed sequence tag libraries followed by bioinformatics analysis provided the first large scale analysis of putative candidate gene effectors (Bowen et al., 2009) in *V. inaequalis*. However, they are not publicly available to initiate community directed annotation and revision. This has limited the molecular biological research for this parasite.

V. inaequalis is a classic member of a group of phytopathogenic fungi that form subcuticular hyphae and derive nutrients from the underlying host tissue (Yepes, 1993). The pathogenic phase of disease starts with germination of ascospores (sexual spore) mainly from leaf litter that serve as the primary source of inoculum (MacHardy et al., 2001). Microscopic studies of germination and eventual penetration of the host have illuminated our limited knowledge of the events that prelude colonization. The similarity in the penetration pathway amongst hemibiotrophs allows us to anticipate the involvement of hydrophobins as surface interactors facilitating fungal cutinase activity (Skamnioti and Gurr, 2008). Cutinases act as surface sensors, mediating appressorium differentiation and penetration peg formation. Two cutinase-like genes were identified that are induced upon morphogenetic differentiation (Kucheryava et al., 2008). The identification of additional cutinase-like domains in RNA-seq studies (Thakur et al., 2013) raise the question whether other such genes exit within the genome. In addition, the observation that a melanized appressorial ring structure fades upon application of a melanin inhibitor adds another key player in the infection pathway. The fact that trihydroxynaphthalene reductase-silenced transformants exhibited a distinctive light brown phenotype (Fitzgerald et al., 2004) suggest that melanin is produced via 1,8-dihydroxynaphthalene (DHN). Assess to the draft genome assembly for *V. inaequalis* (Hesse et al., 2013) is a valuable recourse for unravelling the molecular mechanisms of pathogenesis and adaptation which are crucial for rational design of control strategies. High quality assembled transcripts will greatly enhance the annotation of genes when the genome sequence is available. In the following analysis, we applied the IFRAT protocol (**Chapter 2**) to create a set of *bona fide* transfrags from the South African and Indian apple scab isolates for gene prediction. We describe a unique resource to facilitate

fundamental and applied molecular investigation in the mechanism of apple scab pathogenesis. In addition to identifying key markers of the life cycle transition (dormancy to biotroph), this chapter provides supporting and complementary information to the transcriptomic resource, established in **Chapter 4**. The ensuing analysis provides a basis for virulence protein discovery via follow-up genomics based approaches.

5.2 Materials and Methods

5.2.1 Datasets

We obtained the unfiltered non-redundant transcriptome assemblies of the Indian and South African isolate. Details about generation of these assemblies are described by Thakur et al., (2013), and in **Chapter 4** of this thesis. The protein-coding genes from the *Venturia* draft genome assembly (Hesse et al., 2013) were procured from an in-house annotations analysis, in which the transcriptome assemblies were used as hints for gene prediction (see subsection 5.2.3 for an abridged summary).

5.2.2 Preprocessing the *V. inaequalis* transcriptome assemblies

The non-redundant transcriptome assemblies: Indian and South African isolates; were post-processed using IFRAT (developed in **Chapter 2**). *Bona fide* transfrags were aligned to the draft genome sequence of *V. inaequalis* (South African, SA isolate) using BLAST (E-value 1e-4) to identity high scoring segment pairs (HSP). The overall identity and coverage for each aligned sequence was computed using HSP tilling with in-house PERL scripts. Transfrags producing low quality alignments at minimum threshold of 80% coverage and 80% identity and those that failed to aligned were searched against Genbank non-redundant database (NR) using BLASTx (E-value 1e-4). Transfrags that did not align to the *V. inaequalis* genome were aligned to the apple genome (Velasco et al., 2010) using BLASTn (E-value 1e-4). Transfrags that aligned to the apple scab genome at low

quality and did not align to the apple genome were considered to be of fungal origin.

5.2.3 Gene prediction

High quality assembly-derived *bona fide* transcripts from both isolates were integrated in a gene-calling pipeline (Hesse et al., 2013). In brief, the draft genome was masked using RepeatMasker Open-3.0.1996–2004 (http://www.repeatmasker.org). We probed the approximate location of *bona fide* transfrags on the genome using BLASTn with an e-value threshold of 1e-5. Using an in house PERL script, we excise a substring of the genome bordering all high scoring segment pairs per aligned transfrag. A precise alignment between the genome substring contigs and *bona fide* transfrags was performed with Exonerate version 2.2.0 (Slater and Birney, 2005), with the following parameters: model est2genome and maxintron = 5000. Only transcripts that aligned at a minimum identity and coverage of 80% and 80% respectively, where retained as qualifying sequences for gene prediction with Augustus version 2.5.5 (Stanke et al., 2006). A search of the predicted genes against a database of the core conserved eukaryotic (CEG) genes from *Magnaporthe grisea* using BLASTp (E-value 1e-6), verify the gene space completeness (Parra et al., 2007).

5.2.4 Gene fragmentation analysis using low coverage transfrags with fungal hits

Only transfrags that aligned to the *V. inaequalis* genome below the threshold for genome annotation are used in this section. Peptides of an arbitrary length (100 amino acids) were generated from the 5' and 3' ends of the transfrag-derived proteins with hits to fungi protein in NR. Using BLASTp, we used these short peptides to search against the predicted genes (proteins) of the *V. inaequalis* genome. Alignment of each fragment was achieved by combining soft filtering (-F "m S"), optimal local Smith-Waterman and BLOSUM80 as the scoring matrix. The best hit for each pair of 5' and 3' cognate peptides were aligned together

(bl2seq) to check for any sequence similarity. If both hits have no sequence similarity, the original protein from which the 5' and 3' peptides were obtained is used to search against known fungal proteins. The final list of truncated genes is those transfrag-derived proteins which produce a BLASTp hit to a UniProt (FUNGI) with a minimum coverage of 80%.

5.2.5 Bioinformatic analyses of the secretome

The availability of near-complete set of proteins from *Venturia* allows us to redefine the secretome. We developed a semi-automated secretome prediction pipeline which is an extension of the procedure in Chapter 4, using BASH, AWK and PERL scripts. A protein was considered to have a N-terminal signal peptide if "SignalP D-score = Y", obtained from SignalP version 4.1 (Petersen et al., 2011). The portion of the signal peptide was excised and the remainder of the protein scanned for transmembrane spanning regions (TMs) using TMHMM2.0 (Krogh et al., 2001) and all proteins with '0 Tms' in the output were kept. The original protein sequences satisfying the aforementioned criteria were screened for extracellular localization using "runWolfPsortSummary fungi" in the WoLF PSORT v0.2 package (Horton et al., 2007). The final set of candidate secreted proteins (secretome) contains 'extr' as one of the sites in the ranked list retrieved by WolF PSORT. The secretome was scanned for the presence of the degenerative Y/F/WxC-motif (Godfrey et al., 2010) within the first 24 amino acids after the signal peptide cleavage site. In addition, custom Perl scripts were used to screen candidate RxLR effectors in the secreted translations (Win et al., 2007).

5.2.6 Annotating the predicted coding sequences (genes)

Automated homology based search using BLASTp against a variety of annotated proteins in specialised protein databases was predominantly used for functional annotation transference. In certain cases, functional transference was reinforced by manual annotation, including visual inspection of protein sequences

alignments. Machine learning tools specifically trained for a protein family of interest were used to infer functional relationships in genes with no significant sequence similarity to known proteins.

A total of 11, 692 *V. inaequalis* predicted open reading frames were subjected to InterProScan analysis (Quevillon et al., 2005) and screened against NR using BLASTp with criterion imposed at a cut-off expectation value of 1e-5. The BLATCH BLAST service was used to interrogate the MEROPS database of peptidase families at a E-value of 1e-4 (Rawlings and Morton, 2008).

5.2.6.1 Hydrophobins

UniProt Fungi was initially searched using the keywords "hydrophobin". A customizable database of the retrieved sequences was created. The proteome of *V. inaequalis* was searched against the former with BLASTp (E-value: 1e-10). Alignments with a minimum of 50% coverage (query and hit) and 50% identity were manually verified in producing a meaningful "hydrophobin" hit in the non-reducing database. The Interpro results were also scanned using the text search term "hydrophobin" to confirm the presence of a domain in queries with similarity to the hydrophobin domain. This procedure was repeated with the same set of genes after removal of signal peptides as described by Littlejohn et al., 2012. Candidate hydrophobin genes were compared to the following Class II hydrophonins procured from UniProtKB/TrEMBL (http://www.uniprot.org): *Trichoderma reesei* (P52754 and P79073), *Ophiostoma ulmi* (Q06153); *Cryphonectria parasitica* (P52753), *Cladosporium fulvum* (Q9C2X0), *Magnaporthe grisea* (O94196), *Gibberella moniliformis* (Q6YF29), *Trichoderma harzianum* (P79072). Multiple alignments were performed with T-Coffee (http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi), (Notredame et al., 2000).

5.2.6.2 Carbohydrate Active enzyme

The CAZymes-encoding genes, Carbohydrate-degrading enzymes or Carbohydrate Active enzyme (CAZymes) were identified through a modular annotation procedure that scans the entire *V. inaequalis* proteome with profile Hidden Markov carbohydrate models derived from CAZy database (Cantarel et al., 2009). The automated carbohydrate-active enzyme annotation procedure makes use of the Hmmscan program in HMMER version 3.0 package (Eddy, 2011) to screen family-specific HMM profiles of CAZymes procured from the dbCAN database, downloaded on 18-02-2014 (Yin et al., 2012). The HMM primary result was filtered using the hmmscan-parser script provide by the dbCAN such that if the alignment length > 80 amino acid, use E-value < 1e-5, otherwise use E-value < 1e-3. For confirmatory purposes, we augmented the annotation procedure above with a rigorous simmilarity based approach available at CAZymes Analysis Toolkit, CAT (Park et al., 2010), using the standard parameter procedures previously described (Adhikari et al., 2013). Briefly, the predicted proteome of *V. inaequalis* is subjected to a bi-directional BLASTp search against the entire non-redundant sequences of the Carbohydrate-Active Enzymes database (Cantarel et al., 2009).

5.2.6.3 Cluster analysis of fungi proteomes

Screening for homologs or orthologs between multiple proteomes with sequence similarity methods is fast but scales with large number of proteomes due to the time complexity of aligning sequences. To overcome this, we chose a single-linkage clustering method that has been suitably optimized to make this process more flexible and allow user specific automation. In addition, this method incorporates confidence values for the predicting orthologs.

We explore the increasing availability of complete proteomes for proteome-scale prediction of orthologous groups. Orthologous groups were predicted using 24 proteomes representing filamentous fungi from various lifestyles: *Alternaria*

brassicicola, *Albugo laibachii*, *Aspergillus nidulans*, *Epichloe festucae*, *Candida albicans* , *Chaetomium globosum*, *Cladosporium fulvum*, *Cochliobolus sativus*, *Colletotrichum graminicola*, *Fusarium graminearum*, *Histoplasma capsulatum*, *Hysterium pulicare*, *Leptosphaeria maculans*, *Magnaporthe oryzae*, *Neurospora crassa* *Phaeosphaeria nodorum*, *Phytophthora infestans*, *Pyrenophora tritici-repentis* , *Puccinia graminis*, *Rhytidhysteron rufulum*, *Saccharomyces cerevisiae*, *Venturia inaequalis*, *Verticillium dahliae*, *Ustilago maydis*, *Zymoseptoria tritici* (*Mycosphaerella graminicola*). Ortholog clusters were generated with MultiParanoid (Alexeyenko et al., 2006). MultiParanoid applies single-linkage clustering to merge multiple pairwise ortholog groups from InParanoid (Remm et al., 2001) into multi-species orthologous groups. InParanoid was invoked to search orthologs between (N*(N+1)/2)-N pairs of proteomes using and *ad hoc* PERL script. 'N' represents the number of proteomes to be clustered. Homologs from each proteome are considered by using BLAST to perform a self-search of homologs within each respective proteome. The resulting pairwise ortholog clusters from Inparanoid were directly transferred to MultiParanoid. The collection of proteomes used for this analysis is shown in **Table A5.1**, Appendix 2.

To identify homologs of genes for melanin biosynthesis in *V. inaequalis*, we procured the corresponding protein sequences in GenBank database under accession no. AF025541 (*alb1*), U95042 (*arp1*), AF099736 (*arp2*), AF116901 (*abr1*), AF116902 (*ayg1*), and AF104823 (*abr2*) from *Aspergillus fumigatus* (Tsai et al., 1999). Using InParanoid (Remm et al., 2001) we identified the orthologs in the proteome of *V. inaequalis*. Only homologs at 100% InParanoid confidence score were considered as putative orthologs.

5.3 RESULTS AND DISCUSSION

We present the preliminary analyses of the draft genome assembly en route to a finished sequence of apple scab that has emerged as a model system in woody plant species for understanding phytopathosystem interactions because of its economic value. We highlight essential features for pathogen invasion.

5.3.1 Selecting suitable transfrags for gene-calling

The IFRAT pipeline (developed in chapter 2) was used to screen for putative protein-coding transfrags. In **Table 5.1**, the number of unique transfrags between the orphan and *bona fide* categories are shown for the South African (SA) and Indian (IN) isolates. We note that the number of *bona fide* transfrags is almost the same from both isolates. However, previous analyses (**Chapter 4**) suggest that a significant proportion of the Indian assembly contain contaminating host sequences. With the availability of a draft *Venturia* genome sequence, we verified this by aligning the *bona fide* transfrags (**Table 5.2**). About 598 (SA, ~3.4%) and 5,214 (IN, 30.4%) *bona fide* transfrags failed to align to the *Venturia* genome. Of these, 44 (SA, ~7.4%) and 4027 (IN, ~77.2%) produced significant BLAST hits to plant sequences in a non-redundant database (NR) search (**Table 5.3**). Transfrags that do not align to the genome but have a BLAST match to fungi sequences in NR may represent genes in regions of the genome not represented in the current assembly. In controlled experiments involving de Bruijn graph transcriptome assemblers, Zhao and colleagues, (2011) showed that more than 10% transfrags fail to align to the finished reference Yeast genome and in certain cases, the unique sequences accounted for more than 60% of all unmapped transfrags. Alignment parameters often represent a compromise between aligning transfrags to their true loci, while avoiding spurious alignments. About 788 (SA) and 766 (IN) transfrags align at low quality (80% identity and 80% coverage threshold) (**Table 5.3**). Transfrags may also fail to align because *de novo* transcriptome reconstruction often produces miss-assemblies (Surget-Groba and Montoya-Burgos, 2010) which do not satisfy alignment criteria. Some transfrags align to

the genome but do not have a BLAST match. The existence of such "non-blastable" transfrags could be the result of contamination by genomic DNA and wrong choice of search parameters. These transfrags may also represent lineage specific genes with sufficient sequence divergence (Logacheva et al., 2011).

Table 5.1 Identification of bonafide transfrags using IFRAT.

Transcriptome assembly	South African	Indian
№ of Orphan* TF	28,492	39,711
№ of *bona fide* TF	17,773	17,171
№ unique TF (≥ 100 bp)	46,265	56,882

*Represent transfrags with no protein coding potential or those that are predicted as coding but with no reliable protein prediction by IFRAT

Table 5.2 Summary of transfrag alignments to the genome.

	South African	Indian
№ of *bona fide* TF used for gene prediction	16,387	11,191
№ of *bona fide* TF that did not align	598	5,214
№ of *bona fide* TF with low align	788	766
Total (bona fide TF)	17773	17171

Table 5.3 Categorizing unqualifying transfrags by top BLAST match to NR.

Transcriptome assembly	South African		Indian	
	BLAST against scab genome (S. Africa)			
	yes	no	yes	no
№ of TF with match to Fungi	170	195	280	57
№ of TF with match to Plant	12	44	64	4027
№ of TF to others (virus and microbes)	8	9	9	46
№ of TF with no match	598	350	413	1084
Total № of TF	788	598	766	5214

5.3.2 The general genome features

The genome of *Venturia inaequalis* was sequenced through a whole-genome shotgun approach (Hesse et al., 2013). The genome assembly contains 48,259,067 bp distributed over 44,196 contigs with an N50 value of 26,001 and a GC content of 45.48 % (**Table 5.4**). About 39,718,033 bp (~82.3 %) are effectively represented in 3,106 contigs ≥ 1000 bp with a maximum contig size of ~251.3 Mbp. The genome size had previously been estimated to be approximately 100 Mb (Broggini et al., 2007) and therefore larger than many ascomycetes genomes such as *Aspergillus nidulans* (Galagan et al., 2003) estimated to be 30 Mb, and that of *M. grisea* to be 42 Mb (Dean et al., 2005). However, the draft assembly is smaller than that of the rust fungi *Melampsora larici-populina,* 101-Mb and *Puccinia graminis* f. sp. *Tritici*, 89Mb (Duplessis et al., 2011). Using evidence from assembled transcribed fragments (where necessary), 11,692 protein-coding genes were predicted of which only 20 located in 41,090 contigs < 1000 bp. Given that the protein-coding genes represent ~35% of the genome, this indicates that our assembly largely represents gene-containing contigs and the unaccounted remainder is likely low complexity regions or distributed repetitive elements which are generally challenging for *de novo* assembly from short reads (Miller et al., 2010). Repeat sequences are ubiquitous components of fungal genomes with "typical" repeat content ranging from between 3% (e.g., *A. nidulans, A. fumgatus,* and *A. oryzae*) to 10% (e.g., *Neurospora, Magnaporthe*) (Galagan et al., 2005). Efforts are being made to incorporate mate pairs which will facilitate contig ordering, scaffolding and gap closure.

Table 5.4 Summary statistics of assembly and annotation of the genome of the apple scab pathogen *Venturia inaequalis* South Africa isolate.

Cumulative size of contigs	48,259,067 (~ 48,2% of expected size)
№ of contigs	44,196
Longest contig	251,291
Fraction of N's in assembly	0.52%
N50 contig length	26,001
L50 contig length	498
Gene space completeness (CEG, *M. grisea*)	433
GC content	45.48%
Protein-coding genes	11,692
Mean contig size	1,091
Medain contig size	145
№ of contigs ≥ 100 kbp	22
№ of contigs ≥ 10 kbp	1,053
№ of contigs ≥ 1 kbp	3,106

Gene space completeness indicates that 433 out of 437 conserved eukaryotic genes (CEGMA) from *Magnaporthe grisea* are contained in the genome annotation. The L50 contig count indicates the number of contigs larger than the N50 length.

Having been compiled from short-read data only, the *V. inaequalis* assembly is fragmented, resulting in some genes being split into two or more gene models, with others being truncated versions of the complete genes. An *ad hoc* procedure for reciprocal identification of common gene (hit) based on BLASTx search with the flanking regions of each assembly-derived transcript was implemented. A typical example of the house-keeping gene, (Glyceraldehyde-3-phosphate) retrieved by the method is shown in **Figure 5.1**.

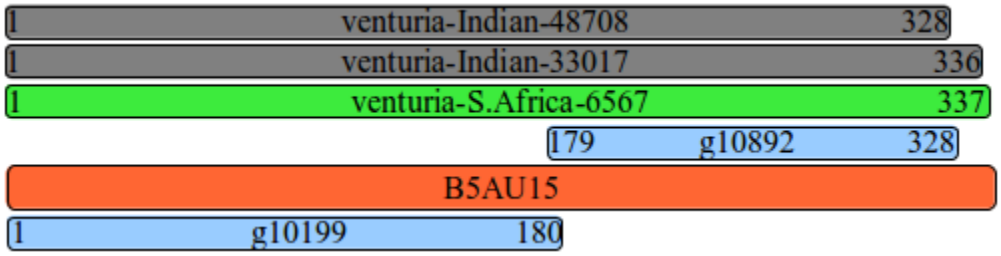


Figure 5.1 A representation of predicted GAPDH-like genes (blue) and proteins (grey and green), aligned to a UniProt GAPDH (B5AU15). The genes and proteins were predicted from the genome and transcriptome respectively. Flanking numbers indicate amino acid positions in B5AU15 that are involved in the alignment. The quality of the alignment is indicated by the percentage identity of the high scoring segment pair.

The consequence of gene fragmentation and how it affects the composition of ortholog clusters cannot be easily assessed as this can vary with the structure of the predicted gene. For example, we found that gene g10892 segregated in a cluster of histone proteins because Inparanoid (Remm et al., 2001) imposes criterion for the global matched region no less than 50% of the longer gene protein sequence. A closer examination using BLASTp revealed that the gene prediction might have merged two neighbouring coding regions into one gene model as show in **Figure 5.2**. This simple gene fragmentation analysis work-flow which we discussed above can be explored in many examples of WGS assemblies from short reads, to flag special cases of genes for manual examination. Despite these shortcomings, a search for the CEG set of 437 conserved eukaryotic genes of *M. grisea* (Parra et al., 2007) found 98.6% at a minimum coverage of 50% in the predicted gene set. About 321 were present at > 90% length indicating a high level of completeness for the predicted ORFs.

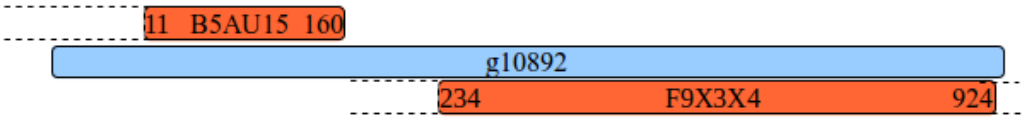


Figure 5.2 Representation of a putative merge gene model.
Gene g10892 (blue) produce two significant hits (B5AU15 and F9X3X4) that segregate in separate clusters and are biologically unrelated.

5.3.3 Annotation of predicted ORFs reveals pathogenesis and virulence arsenal

Predicted ORFs were searched against the non-redundant and UniProt databases from which 9,185 (~78.6%) and 9,319 (79.7%) had at least one significant blast hit which contain 8,858 and 8,992 unique accessions respectively. By mapping the best BLAST hit to gene ontology association (Jones et al., 2005), 5,746 functional terms could be retrieved from the later. ORFs with no hit could represent divergent genes that have been tailored to suit the unique biological adaptation of the apple scab vis-à-vis pathogenesis and virulence.

5.3.3.1 Secreted proteins

Access to the *V. inaequalis* draft genome sequence enabled us to revisit and re-defined the secretome. **Figure 5.3** shows an updated distribution of secreted proteins in selected fungi. About 1,061 putative ORF were predicted as secretory proteins (1,054 are ≥ 70 amino acids in length). We note that this is twice the number predicted for the transcriptome assemblies (**Chapter 4**). The size of the secretome is comparable to that of other fungi. An examination of the secreted ORFs, after removal of signal peptide, for motifs associated with virulence reveal only 2 candidate RxLR-like motifs, probably as random hits. The scarcity of RxLR pattern in *Venturia* is consistent with their independent evolution and expansion in Oomycetes such as *Phytophthora infestans* albeit absent in *Pythium* species (Adhikari et al., 2013). The mechanism of RxLR-like protein delivery has been speculated to occur at haustoria interphase (Talbot, 2007) and incidentally, *Venturia* does not form haustoria (Jha et al., 2010). However, we identified 19

secreted ORFs with an N-terminal Y/F/WxC-motif. This number is fewer than that found in the transcriptome assemblies (**chapter 4**) possibly due to splice variants. Delivery of Y/F/WxC-effectors may also depend on haustoria formation although non-haustoria forming fungi have a small number of predicted proteins with a Y/F/WxC-motif (Godfrey et al., 2010). Additional features of the secretome are highlighted in subsequent sections described hereafter.

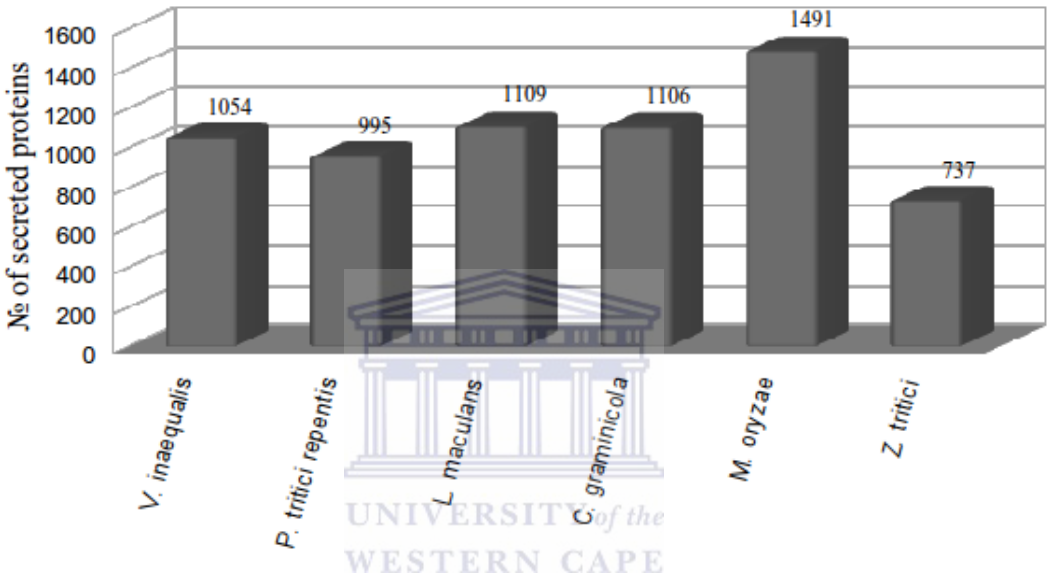


Figure 5.3 Distribution of secreted proteins in *V. inaequalis* and selected fungi. The secretome of *V. inaequalis* is compared with those of other fungi. This is twice as much as that predicted for the transcriptome (chapter 4).

5.3.3.2 Genes involved in cuticle penetration

Successful establishment of primary infection in *Venturia* depends on optimal environmental conditions (leaf wetness duration and temperature) for ascospore production and infection of apple leaves (MacHardy and Jeger 1983; MacHardy, 1996). This has been exploited for disease prognosis and establishing appropriate fungicide schedules (Rossi et al., 2003). Upon contact with plant surfaces, hydrophobic-like proteins are implicated in the surface interaction during infection-related development. These proteins have the unique ability to self assemble and they function at the fungal wall-air interface, limiting desiccation

and providing protection against both chemical and enzymatic attack (Wösten, 2001). Two characteristic hydrophobic-like proteins (g6671 and g9216) were identified in the entire predicted proteome of apple scab with significant BLAST hits to fungi hydrophobins in UniProt. The alignment in **Figure 5.4** shows the characteristic idiosyncratic eight Cys residues implicated in the formation of disulfide bridges. The eight Cys residue is missing in g9216 possibly as a result of a gene prediction error. There is indeed alignment in and around the cysteine residues with little conservation between other parts of the sequence. This is reflected in the generalised structure of the hydrophobins where the intervening amino acids between cysteines are variable in number and type (Kershaw and Talbot, 1998).

Although recent bioinformatics studies of *Aspergillus* hydrophobins suggest that intermediate/different forms can also exist with distinct physicochemical characteristics (Littlejohn et al., 2012), the presence of only two hydrophobin genes may be sufficient in fulfilling the needs of the apple scab. Based on the InterPro analysis we categorises the *Venturia* hydrophobins as Class II hydrophobin cerato-ulmin. In addition, both hydrophobins were predicted as secreted proteins. This feature is consistent with their role as extracellular surfactants. The analyses of fungal genomse indicated that hydrophobins generally exist as small gene families of two to ten members, with a few species containing more members (e.g., *Coprinus cinereus* displays 33 members (Sunde et al., 2008). Hydrophobins are suggested to function as a developmental sensor for appresorium formation, since it is involved in the interaction with hydrophobic leaf surfaces. The Expression of MHP1 was highly induced during plant colonization and conidiation, but could hardly be detected during mycelial growth *M. grisea*. In addition, targeted disruption of MHP1 resulted in a reduced frequency of appressorium formation (Kim et al., 2005).

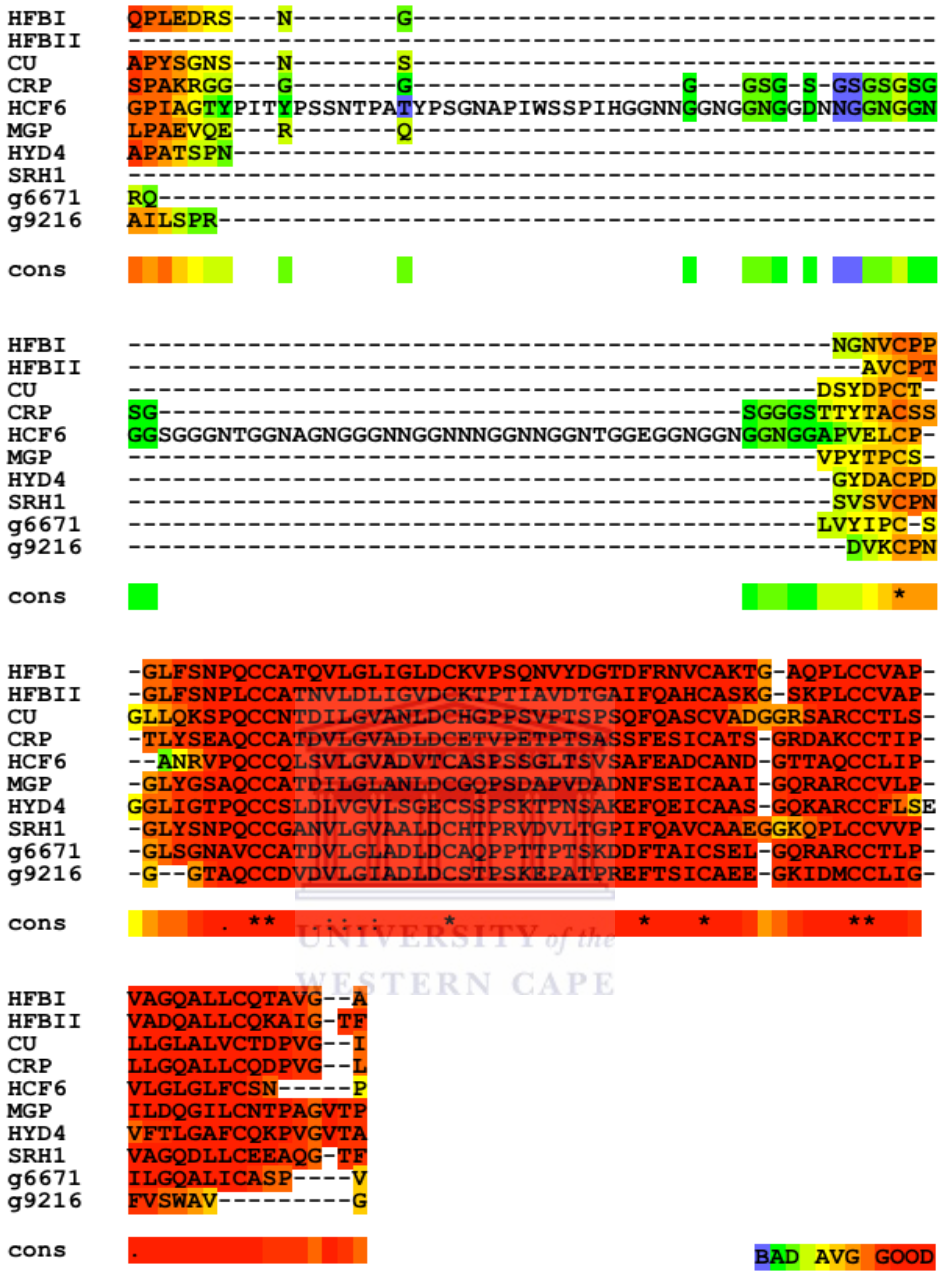


Figure 5.4 Amino acid sequence comparison of class II hydrophobins.
Multiple sequence alignment of hydrophobins represented by 6 fungi. Only amino acids between the signal cleavage site and last amino residues are shown. The relatively low degree of sequence conservation is apparent. The abbreviations used are: HFBI, T. reesei (accession P52754); HFBII, T. reesei (accession P79073); CU, O. ulmi (accession Q06153); CRP, C. parasitica (accession P52753); HCF6, C. fulvum (accession Q9C2X0); MGP, M. grisea (accession O94196); HYD4, G. moniliformis (accession Q6YF29) and SRH1, T. harzianum (accession P79072).

Pathogens synthesize and secrete various peptides/proteins that modulate host physiology and block host responses in order to persist in the host (Yike, 2011). Biotrophic and necrotrophic fungi share enzymatic elements but these may have different purposes when causing disease (Meinhardt et al., 2014). While necrotrophs apply brute force causing rapid cell death and overwhelming the plant defenses, biotrophic pathogens appear to evade plant defenses with stealthy methods (Latunde-Dada, 2001). Hemibiotrophs such as *Venturia* possesses properties of both groups. This can be exemplified in the distribution of peptidase families of emzymes as shown in **Table 5.5**. A total of 451 peptidases were identified in the entire proteome with 87 predicted to be secreted. The number of cysteine (Cys), metallo, serine (Ser) and threonine (Thr) peptidases are higher in *Venturia* when compare to typical fungi phytopathogens. However, the sequence simmilarity method used did not retrieve any Glutamic (Glu) peptidases. We note the Glutamic proteases (G1) are a characteristically rare group of peptidases otherwise thought to be limited to the filamentous fungal species of the Ascomycota phylum and generally range between 1-4 per genome (Sims et al., 2004). In the MEROPS peptidase database version 9.1 (Rawlings et al., 2012), sixty out of the sixty-six putative G1 peptidases were from Ascomycetes of which six are supposedly non-peptidase homologs lacking one or both catalytic residues, thereby reducing the total number of peptidases of Ascomycete origin to fifty-four (Jensen et al., 2010). Contrary to the ubiquitous occurrence of core groups of proteases, the paucity of G1 proteases both within and between the sequenced fungal genomes and in nature exemplifies their non-essentiality in fungal growth /survival (Sims et al., 2004).

Table 5.5 Comparison of protease distribution in *V. inaequalis* and selected fungi with the same nutritutional status.

Species	Peptidase family						
	Asp	Cys	Glu	Metallo	Ser	Thre	Unknown
V. inaequalis	27	77	0	136	179	24	8
P. tritici repentis	20	65	1	125	141	19	9
C. graminicola	17	64	1	142	176	21	9
L. maculans	15	62	1	118	123	20	9
M. oryzae	22	58	2	128	159	21	9
Z. tritici	28	52	4	114	165	19	8

The CAZYome of *Venturia* portrays an impressive toolbox for the degradation of carbohydrates. Based on the dbCAN classification scheme (Yin et al., 2012), we identified 601 CAZymes-encoding ORFs belonging to 86 CAZyme families (**Table 5.6**). Amongst these, 28 CAZymes belong to 9 families of Carbohydrate Binding Modules (CMB): CBM19 (2), CBM32 (2), CBM48 (2), CBM67 (2), CBM13 (3), CBM1 (3), CBM63 (3), CBM50 (5), CBM18 (6). Of all the identified CAZymes in *Venturia*, 258 were predicted as secreted. The number of CAZyme families is comparable to those identified in hemibiotrophs and necrotrophs, and is larger than those in biotrophs. The diversity of CAZyme families correlates with adaptation to lifestyles (van den Brink and de Vries, 2011) and also reflects host preference among plant pathogenic fungi (King et al., 2011). Possessing a variety of CAZymes could explain in part the ability of *Venturia* to infect a range of hosts despite the high diversity of *Malus* species (Leroy et al., 2013). Fungi are able to produce all kinds of CAZymes and particular interest is placed on plant cell wall degrading enzymes (CWDE) that have received special attention because of the role in penetration, colonization and successful establishment of infection. For simplicity, only the CWDE group of CAZymes which represent pectinases, cellulases and hemicellulases are displayed in **Table 5.8** (see Zhao et al., 2014 for a detailed analysis of CAZymes families across phyla). A fair comparison can be

made with these results to that by Zhao and colleagues, (2014) since the methods used in identifying the CAZyme families were the same. Albeit not topping the list of CAZyme families, *Venturia* had the largest number of PL4 (6 polysaccharide lyase) and CE5 (28 cutinases) amongst all the fungus use in this study. Given that the presence of CE5 was further confirmed by sequence similarity (Park et al., 2010), *V. inaequalis* has the largest number of cutinases when compared to those (103 fungi) in the study by Zhao et al., (2013). Another phythopathogen with an expansion in CE5 family is the hemibiotrophic rice blast fungus *Magnaporthe orzea* with 19 cutinases. Despite the stringency of assessing secreted proteins, 23 CE5 were identified in the refined secretome of *Venturia*.

Table 5.6 Variability in plant cell degrading enzymes across fungi of differing nutritional lifestyles

Nutritional type	Fungi	№ CAZyme families	Pathogen-related cell wall degrading enzymes														
			PL1	PL3	PL4	PL9	PL10	GH28	GH78	GH88	CE5	GH1	GH3	GH5	GH45	GH74	GH10	GH11
Biotroph	*A. laibachii*	41	0	0	0	0	0	3	0	0	4	0	4	22	0	0	0	0
	C. fulvum	87	3	4	2	0	0	14	5	3	11	3	20	15	0	2	2	2
	E. festucae	61	0	0	0	0	0	2	0	0	4	0	7	6	0	5	0	0
	P. graminis	62	2	0	0	0	0	0	0	0	9	0	2	29	0	2	5	0
	U. maydis	49	0	0	0	0	0	0	0	0	4	0	3	13	3	2	2	0
Hemibiotroph	*C. sativus*	94	6	5	4	0	0	4	5	0	14	3	15	15	3	4	5	5
	C. graminicola	96	7	4	3	0	0	7	4	0	13	0	18	15	2	7	9	6
	F. graminearum	92	9	7	3	0	0	6	7	0	13	3	22	14	0	5	5	2
	L. maculans	87	8	7	4	0	0	7	0	0	8	3	13	14	2	2	3	2
	M. oryzae	88	2	0	0	0	0	4	4	0	19	2	18	13	0	5	6	5
	P. infestans	72	18	44	3	0	0	22	5	0	5	18	20	21	0	0	3	0
	V. inaequalis	86	4	0	6	0	0	12	2	0	28	2	12	12	2	2	0	3
	Z. tritici	68	2	0	0	0	0	2	3	0	6	2	16	9	0	4	2	0
Necrotroph	*A. brassicicola*	89	8	12	3	0	0	7	0	0	9	2	12	15	2	5	5	4
	P. nodorum	89	4	2	4	0	0	5	4	0	11	2	15	18	2	5	8	7
	P. tritici-repentis	92	3	3	4	0	0	6	4	0	10	3	12	13	3	4	4	3
	V. dahlia	92	17	12	5	2	0	12	9	4	14	2	16	13	2	8	4	5

PL, polysaccharide lyase; CE, carbohydrate esterase and GH, glycoside hydrolase

5.3.3.3 Identification of melanin biosynthesis genes

Clustering of 286,779 proteins sequences from twenty five fungal proteomes was determined by InParanoid (Remm et al., 2001) and Multiparanoid (Alexeyenko et al., 2006). As a result, we clustered 209,857 proteins into 13,575 orthologous gene groups (**Figure 5.5**). About 2,825 from *V. inaequalis* did not cluster with any of the proteomes under default criteria and were considered *Venturia* specific. These ORFs may be encoded by potentially novel genes that have evolved in the *Venturia* lineage. However, we speculate that this large number proteome specific ORFs may contain functional orthologs that are replaced by genes with complementary roles in the other proteomes. The method of finding orthologs used herein is dependent on sequence simmilarity. As such, these proteome specific ORFs could have a high level of sequence divergence, precluding efficient recognition of orthologs (Logacheva et al., 2011). The *Venturia* specific proteome contains 400 predicted secreted proteins.

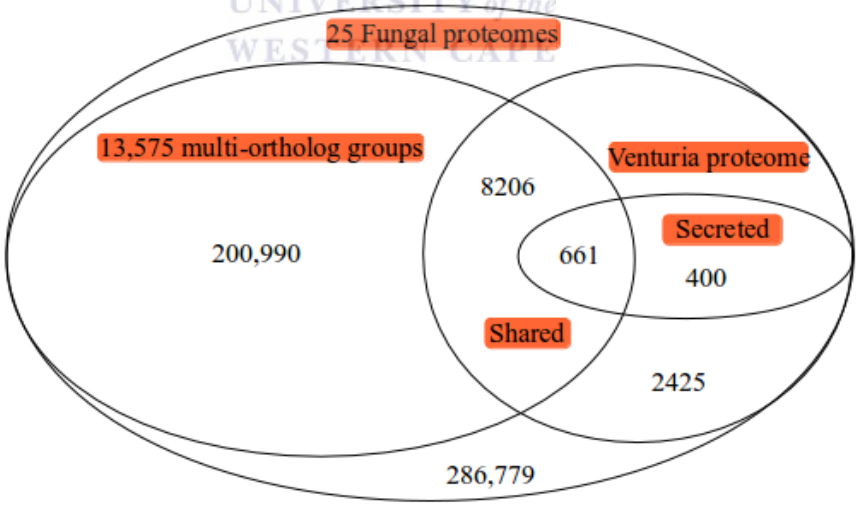


Figure 5.5 Distribution of *V. inaequalis* proteins in orthologous groups. The proteins specific proteome of *V. inaequalis* contain a large number of secreted proteins.

Treatment of *V. inaequalis* and *Diplocarpon rosae* with melanin biosynthesis inhibitors (MBIs), prevented melanization of appressoria and positively correlated with reduced infection rate (Gachomo et al., 2010; Steiner and Oerke, 2007). Unlike *M. grisea* and *Colletotrichum lagenarium* where melanin is localized in the inner layer of the cell wall, it is found to be restricted at the penetration pore in a melanized appressorial ring structure (MARS). Appressoria without MARS are not able to infect plant (Steiner and Oerke, 2007). The majority of fungi, synthesize melanin from endogenous substrate via a 1,8-dihydroxynaphthalene (DHN) intermediate (Eisenman and Casadevall, 2012). Amongst the steps leading to DHN, the first involves formation of 1,3,6, 8-tetrahydroxynaphthalene (1,3,6,8-THN) catalyzed by a polyketide synthase (PKS). Following this is a series of reductions (catalysed by trihydroxynaphthalene reductase, THN) and dehydration reactions. In *A. fumigatus*, the melanin biosynthesis gene clusters consists of six genes and spans 19 kb pairs of DNA (Eisenman and Casadevall, 2012). Homologs of these genes were found via one-to-one ortholog analysis with InParanoid at 100% confidence (**Table 5.7**). Each gene had a meaningful Interpro description that confirmed their enzymatic roles, similar to those of *A. fumigatus*. However, these genes were found on different contigs in the current assembly of *V. inaequalis*. Unlike the cluster model of melanin biosynthetic genes in *A. fumigatus*, three homologs identified in brown and black fungi are not closely linked in the genome of *Colletotrichum lagenarium* (Keller and Hohn, 1997).

Table 5.7 *A. fumigatus* genes of the Melanin biosynthetic pathway and corresponding orthologs in *V. inaequalis*.

V. inaequalis ortholog	Interpro annotation	*A. fumigatus* (accessions)
g7420	Polyketide synthase dehydratase	AAC39471.1
g8299	Alpha/beta hydrolase of unknown function	AAF03354.1
g7475	Multicopper oxidase	AAF03353.1
g8063	Short-chain dehydrogenase/reductase	AAF03314.1
g10071	Scytalone reductase	AAC49843.1
g3511	Multicopper oxidase	AAF03349.1

The role of melanin as a virulence factor is by enabling the fungal cell to retain glycerol thereby reducing the porosity of the appressorial wall. Albino *Magnaporthe* mutants produce an abnormally permeable and morphologically abnormal appressoria unable to generate high pressure (Howard and Ferrari, 1989). As a proof of principle, a partial cDNA sequence for the *V. inaequalis* THN gene homolog was identified from an expressed sequence tag (EST) database. A hairpin construct was used to achieve RNA-mediated gene silencing (quelling) which produces transformants exhibiting a distinctive light brown phenotype (Fitzgerald et al., 2004).

5.4 Conclusion

The availability of the draft genome sequence of apple scab has facilitated the identification of virulence factors and expanded our understanding of the mechanisms underlying infection. This does not fully inform the range of possibilities that exists for a given species or the ways in which these components will manifest in response to the environment. We chose to focus our analyses on gene families that can be implicated in establishing an infection in the host and identified 2 hydrophobins, 85 cell wall degrading enzymes and 6 homologs of melanin biosynthesis. These genes are an interesting find as they may represent novel targets for the rational design and dissemination of control strategies. Silencing, utilising the marker genes involved melanin biosynthesis, will enable the development of high through-put screening for functional genomics.

CHAPTER 6

Summary, Limitations and Future Work

6.1 Summary

Venturia inaequalis continues to be the most important pathogen of domesticated apples (*Malus* x *domestica*) and also infects other members comprising species in the genera *Crataegus, Sorbus, Pyracantha, Eriobotrya, Kageneckia*, and *Heteromeles* (Bus et al., 2011). Several studies into the interaction of *V. inaequalis* and *Malus* spp have established the *Venturia-Malus* pathosystem as a model system for the gene-for-gene interaction in woody plant species (Vincent G M Bus et al., 2005; Kucheryava et al., 2008; Paris et al., 2009; Bowen et al., 2011; Leroy et al., 2013). In such an interaction, a defence reaction that completely prevents infection is initiated by a resistance (*R*) gene product in the host after recognizing an "effector" (avirulence, *Avr*) gene product in the pathogen (Dodds and Rathjen, 2010; Van Der Biezen and Jones, 1998). A few studies have identified *R* gene containing loci from apple cultivars (Bénaouf and Parisi, 2000; Broggini et al., 2007; V. G. M. Bus et al., 2005) and attempts have been made to characterize the *Avr* genes of *V. inaequalis* (Köller, 1991; Kollar, 1998; Valsangiacomo and Gessler, 1992; Bowen et al., 2009; Kucheryava et al., 2008). Knowledge from these interactions has facilitated the introgression of scab resistance gene in apple breeding (King et al., 1999; Belfanti et al., 2004). Despite the pernicious agriculture importance of apple scab, there is still much to understand about its mechanisms of pathogenesis and virulence. Eradicating the well established invasive apple scab fungus is infringed by the development of resistance. Paucity of genomic resources has limited the shift from conventional breeding which are usually time-consuming. The availability of genomic resources can expedite: 1) development of whole genome mutagenesis screens, 2) the design of standardized transformation in *Venturia*, 3) understanding the *R* gene mediated resistance break down and 4) illuminating defence responses elicited in apple (Jha et al., 2010).

Recent advances in DNA sequencing technology have reduced transcriptome sequencing to a fraction of the cost and time previously required by Sanger

sequencing in procuring large amounts of sequence data (Parchman et al., 2010). In the absence of a suitable reference against which a comparison can be made, *de novo* sequence assembly is paramount (Paszkiewicz and Studholme, 2010). The characteristic diminutive sequence length and high base-calling error rates of NGS technologies render exclusive assembly of RNA-seq data computationally challenging: paving the way for developing optimal data preprocessing and post-processing strategies. We harness the power of RNA-Seq data to generate sequence data for *V. inaequalis*. The work in this thesis describes a number of intuitive approaches that greatly enhance the application of RNA-Seq in non-model organisms. These heuristic approaches improve the quality and quantity of *V. inaequalis* reconstructed transcripts.

De novo reconstruction of transcribed fragments (Gibbons et al., 2009) especially those from the Illumina platform typically results in large assemblies that are redundant, with differing levels of fragmented or miss-assembled transfrags. Post-assembly processing approaches that are intended to reduce redundancy typically involves reassembly (Feldmeyer et al., 2011) or clustering of assembled sequences (Surget-Groba and Montoya-Burgos, 2010). However, these methods are mostly based on common word heuristics (Hazelhurst et al., 2008) that create clusters of biologically unrelated sequences (Bragg and Stone, 2009), resulting in loss of unique transfrag annotations (Haznedaroglu et al., 2012) and propagation of miss-assemblies. In addition, transfrags may represent incompletely spliced transcripts given that reads originate from mature mRNA and precursor mRNA (Garber et al., 2011). Such transfrags are likely to be responsible for the significant number of non- BLAST searchable sequences that are encountered in *de novo* transcriptome assemblies. Intuitively, truncated and chimeric transfrags can affect domain integrity and frustrate domain based annotations typically performed with InterProScan (Quevillon et al., 2005; Zdobnov and Apweiler, 2001) associated databases.

6.1.1 Selecting suitable transfrags for downstream analyses

We propose a post-assembly approach that mitigates these problems in a pipeline architecture known as IFRAT (Inferring Functionally Relevant Assembly-derived Transcripts). The major highlights are that, IFRAT incorporates an assessment of protein coding potential and eliminates redundancy without clustering to select *bona fide* transfrags. The usefulness of IFRAT was evaluated on *Neurospora crassa* Illumina reads. The non-redundant output of IFRAT produced the same the number of unique gene ontology terms when compare to the unfiltered assembly. The *bona fide* fraction of these assemblies also represented over 94% of retrievable BLAST hits. When IFRAT is used to process transcriptome assemblies of published datasets, we show that it can provide a reference transcriptome enriched with functionally relevant assembly-derived transcripts.

A fundamental feature used to predict coding potential that has high concordance with sophisticated discrimination, is ORF length (Frith et al., 2006; Liu et al., 2006). Truncated transfrags will invariably result in short ORFs. ORF length is the primary criterion and has good discriminant power in almost all coding-potential prediction methods (Wang et al., 2013). It becomes evident that, optimal assembly is required for effective coding-potential assessment.

6.1.2 Selecting appropriate threshold for quality score based filtering

In a non-model organism, there is an optimal number of reads balancing coverage and errors (Francis et al., 2013) and aggressive trimming or filtering strategies are likely to affect the coverage dynamics. Quality score based filtering or trimming results in bulk data loss and affect the quantity of reads available for *de novo* assembly (Le et al., 2013). There is no consensus on quality filtering threshold and it is difficult to filter biological variation from sequencing errors without a reference using quality score. We showed that quality score based filtering is subjective and very stringent filtering can be injurious to the overall assembly quality. By generating sub-samples

of reads using varying quality score filtering thresholds, it is possible to arrive at a threshold that is optimally suitable for assembly. For example, assemblies produced from reads subjected to different quality score thresholds contain truncated and missing transfrags when compared to those from untrimmed reads (Mbandi et al., 2014). However, such a strategy must be gauged with an appropriate transcriptome assembly quality assessment metric such as the HSP ratio. Quality assessment and filtering should remain a prerequisite for downstream analysis; however each dataset must be handled differently. We recommend variable quality score based trimming or filtering after artefact removal to guide the choices for optimal prepossessing parameters. Indeed, these findings are in line with the view that for *de novo* transcriptome assembly gentle trimming or no trimming at all is preferable (MacManes, 2014).

6.1.3 Untangling transfrags in mixed host and pathogen transcriptome assemblies

In chapter 4, quality score dependent filtering was applied to generate 39, 042 transfrags of a host free culture of *V. inaequalis*, South African isolate. We took advantage of a recently published transcriptome assembly of an Indian isolate generated from a mixed library (Thakur et al., 2013) to perform comparative transcriptomics. A key interest in pathogen-host interaction is the identification of pathogen secreted proteins that are induced in planta. They perform important tasks such as nutrient acquisition and block plant defenses. RNA-Seq data from mixed library produce reads that originate from the pathogen and the host (Zhu et al., 2013). It thus becomes very important to ascertain the origin of transcribed fragments (Soderlund, 2009). To achieve this, we propose an iterative alignment procedure that takes advantage of the published apple genome and proteome (Velasco et al., 2010) to partition fungi (Indian isolate) and plant transfrags. The *in silico* deconvolution approach binned ~50 % of transfrags of the Indian isolate to plant origin, of which 233 (0.41%) segregated with *Malus spp* proteins through a non-redundant database

search. Transfrags that specifically mapped to the apple genome and or proteome encoded putative novel methalothioneins and defensins that are associated with resistance. About 40% transfrags from the Indian isolate and 46.5% transfrags from the S. Africa isolate had a BLAST match in UniProt knowlegebase. Additionally, we predicted 420 (Indian isolate) and 514 (S. African isolate) secretory/signal peptides of which 40 and 30 respectively, had a *bona fide* N-terminal Y/F/WxC-effector motif. The scarcity of RxLR-like motif in both transcriptomes suggest that other mechanisms of effector delivery may exist for this non-haustoria forming fungus.

6.1.4 Identification of genes involved in pathogenesis

In a separate endeavour, IFRAT was used to generate a set of *bona fide* transfrags from the Indian (17,171) and South Africa (17,773) isolates. Gene prediction with *bona fide* transfrags (27,578) as hints that effectively align to the *V. inaequalis* draft genome, produced 11,692 protein coding genes. Unmapped transfrags were either of plant origin or miss-assemblies. From the predicted CDS we identify two hydrophobin-like genes whose general structures conform to that of class II hydrophobins. Both hydrophobins contain the majority of the characteristics cysteine residues involved in the formation of disulphide bridges which is crucial for their surfactant-like properties. Comparative CAZyome analyses reveal an expansion of the polysaccharide lyase (6) and carbohydrate esterase (28) cazyme families. When compared with the study of 103 CAZyomes of representative fungi from Ascomycota, Basidiomycota, Chytridiomycota, and Zygomycota (Zhao et al., 2014, 2013), *V. inaequalis* has the largest number of cutinases. We identified the full compliment of enzyme homologs (that includes a polyketide synthase and accessory proteins) in the melanin biosynthetic pathway via a 1,8-dihydroxynaphthalene (DHN), a product that is subsequently reduced and dehydrated to 1,3,6,8-tetrahydroxynaphthalene (1,3,6,8-THN). Our finding corroborates those of previous studies that confirmed the role of melanin in MARS formation (Steiner and Oerke,

2007) and RNA*i* of THN reductase induce albino-like phenotype (Fitzgerald et al., 2004).

6.2 Possible limitation and recommendations

In this thesis, we have exclusively applied a modified version of PORTRAIT (Arrial et al., 2009) in pipeline architecture (IFRAT) in selecting transfrags for downstream analysis without a suitable reference genome. PORTRAIT depends on ANGLE (Shimizu et al., 2006) for the prediction of coding regions. A perceived limitation in this approach is the accurate prediction of coding potential from unknown miss-assembled transfrags and truncated transfrags. The current version of ANGLE does not incorporate a model of boundary sites (start and stop sites) which can potentially hinder the predictive power on transfrags which a significant portion of 5'-UTR and 3'-UTR. For example, a small percentage of orphan transfrags had a BLAST match although the majority was from truncated transfrags akin to the observations by Tao et al., (2012). In addition, the method for RNA-Seq library preparation does not fully distinguish coding from non-coding RNA fragments. As such, reconstructed transcripts may contain intronic sequences that dilute the predictive power of coding-potential assessment. Indeed the analysis in Chapter 2 confirmed that orphan transfrags with BLAST hit often overlap with intronic or UTR regions of the genome. It may be necessary to preformed ncRNA depletion prior to assembly (Cui et al., 2010; Li et al., 2012; O'Neil et al., 2013).

Preprocessing may often involve the removal of artefacts that are remnants of adapters. We observed that adapter removal is much more efficient when performed prior to quality filtering (Mbandi et al., 2014). Trimming invariably mutates the NGS reads that mitigate string recognition during adapter removal.

The HSP ratio that I introduced in Chapter 3 is a metric for scoring the highest segment pair from the alignments of the six frame translated proteins from each

transfrag to a set of 'known' proteins. A poorly reconstructed transfrag would have truncated proteins and or frame shifts that would produce low coverage alignments. On a transcriptome scale, the proportion of high scoring alignments is a good estimation of how well the assembly is when compared to another assembly performed in a different way. The known proteins do not necessarily have to come from a closely related species. Intuitively, housekeeping genes are fairly similar across taxa and will give a good approximation. A chimeric transfrag may produce two independent high scoring local alignments similar to that of the disjointed transfrags. When computing the HSP ratio, a separate evaluation on the level of miss-assemblies may provide further details about the quality of the assemblies.

De novo transcriptome assembly will continue to remain an arduous task when assembling reads that originate from precursor and mature mRNA. Sequencing of ribosome-nascent-chain (RNC) has hardly been performed on non-model organisms. RNC-Seq exhibit unique advantages against traditional RNA-Seq in that only reads from mature mRNA will be sampled (Zhang et al., 2014). This so-called RNC-Seq (translatome sequencing) should be the method of choice for interrogating protein-coding genes in non-model organisms.

6.3 Future work envisaged

Despite the many impressive feats that RNA-seq has helped accomplish, the majority of studies of *de novo* transcriptome sequencing have been largely descriptive (Strickler et al., 2012). This may partly be due to the fact that high-throughput gene screening models do not exist for the many organisms whose transcriptomes are currently being sequenced. The transcript and gene models generated within the framework of this thesis ushers a revival in genetic manipulation and molecular biology research into the biology of this pathogen. The following investigations are under-way:

1. With the existing high resolution genetics map of *V. inaequalis*, it is possible to generate mate-pair libraries and additional sequencing for gap closure and contig ordering.
2. Real time and RNA*i* of homologous enzymes in the melanin biosynthesis pathways followed by whole parasite melanin inhibition studies.
3. The current study has only provided a snapshot of the *V. inaequalis* transcriptome in host free culture. We envisage a time course expression profiling throughout the life cycle.
4. Elucidate the proteome changes in response to biotic stress from fungal invasion aided by the predicted transcript and CDS models and the molecular basis for specificity of different *Venturia* species.

Completion of these studies will identify proteins that should be investigated for their biological relevance, thus providing a promising avenue that advance control and mangement of *V. inaequalis*.

References

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S., Elliston, K.O., 1996. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. Genome Res. 6, 829–845. doi:10.1101/gr.6.9.829

Adhikari, B.N., Hamilton, J.P., Zerillo, M.M., Tisserat, N., Lévesque, C.A., Buell, C.R., 2013. Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes. PLoS ONE 8, e75072. doi:10.1371/journal.pone.0075072

Alexeyenko, A., Tamas, I., Liu, G., Sonnhammer, E.L.L., 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22, e9–e15. doi:10.1093/bioinformatics/btl213

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. doi:10.1093/nar/25.17.3389

Ananthasubramanian, S., Metri, R., Khetan, A., Gupta, A., Handen, A., Chandra, N., Ganapathiraju, M., 2012. Mycobacterium tuberculosis and Clostridium difficille interactomes: demonstration of rapid development of computational system for bacterial interactome prediction. Microb. Inform. Exp. 2, 4. doi:10.1186/2042-5783-2-4

Arrial, R.T., Togawa, R.C., Brigido, M.M., 2009. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. BMC Bioinformatics 10, 239. doi:10.1186/1471-2105-10-239

Bai, Y., Dougherty, L., Xu, K., n.d. Towards an improved apple reference transcriptome using RNA-seq. Mol. Genet. Genomics 1–12. doi:10.1007/s00438-014-0819-3

Belfanti, E., Silfverberg-Dilworth, E., Tartarini, S., Patocchi, A., Barbieri, M., Zhu, J., Vinatzer, B.A., Gianfranceschi, L., Gessler, C., Sansavini, S., 2004. The

HcrVf2 gene from a wild apple confers scab resistance to a transgenic cultivated variety. Proc. Natl. Acad. Sci. U. S. A. 101, 886–890. doi:10.1073/pnas.0304808101

Bénaouf, G., Parisi, L., 2000. Genetics of Host-Pathogen Relationships Between Venturia inaequalis Races 6 and 7 and Malus Species. Phytopathology 90, 236–242. doi:10.1094/PHYTO.2000.90.3.236

Bentley, D.R., 2006. Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 16, 545–552. doi:10.1016/j.gde.2006.10.009

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59. doi:10.1038/nature07517

Berglund, A.-C., Sjolund, E., Ostlund, G., Sonnhammer, E.L.L., 2007. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 36, D263–D266. doi:10.1093/nar/gkm1020

Bhardwaj, J., Chauhan, R., Swarnkar, M.K., Chahota, R.K., Singh, A.K., Shankar, R., Yadav, S.K., 2013. Comprehensive transcriptomic study on horse gram (Macrotyloma uniflorum): De novo assembly, functional characterization and comparative analysis in relation to drought stress. BMC Genomics 14, 647. doi:10.1186/1471-2164-14-647

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., the Galaxy Team, 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics 26, 1783–1785. doi:10.1093/bioinformatics/btq281

Bouck, A., Vision, T., 2007. The molecular ecologist's guide to expressed sequence tags. Mol. Ecol. 16, 907–924. doi:10.1111/j.1365-294X.2006.03195.x

Bowe, A., Onodera, T., Sadakane, K., Shibuya, T., 2012. Succinct de Bruijn Graphs, in: Raphael, B., Tang, J. (Eds.), Algorithms in Bioinformatics, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 225–235.

Bowen, J.K., Mesarich, C.H., Bus, V.G.M., Beresford, R.M., Plummer, K.M.,

Templeton, M.D., 2011. Venturia inaequalis: the causal agent of apple scab. Mol. Plant Pathol. 12, 105–122. doi:10.1111/j.1364-3703.2010.00656.x

Bowen, J.K., Mesarich, C.H., Rees-George, J., Cui, W., Fitzgerald, A., Win, J., Plummer, K.M., Templeton, M.D., 2009. Candidate effector gene identification in the ascomycete fungal phytopathogen Venturia inaequalis by expressed sequence tag analysis. Mol. Plant Pathol. 10, 431–448. doi:10.1111/j.1364-3703.2009.00543.x

Bragg, L.M., Stone, G., 2009. k-link EST clustering: evaluating error introduced by chimeric sequences under different degrees of linkage. Bioinformatics 25, 2302–2308. doi:10.1093/bioinformatics/btp410

Broggini, G.A.L., Le Cam, B., Parisi, L., Wu, C., Zhang, H.-B., Gessler, C., Patocchi, A., 2007. Construction of a contig of BAC clones spanning the region of the apple scab avirulence gene AvrVg. Fungal Genet. Biol. 44, 44–51. doi:10.1016/j.fgb.2006.07.001

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., Brom, T.H., 2012. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. ArXiv12034802 Q-Bio.

Buetow, K.H., Edmonson, M.N., Cassidy, A.B., 1999. Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet. 21, 323–325. doi:10.1038/6851

Burke, J., Davison, D., Hide, W., 1999. d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences. Genome Res. 9, 1135–1142.

Bus, V.G.M., Laurens, F.N.D., van de Weg, W.E., Rusholme, R.L., Rikkerink, E.H.A., Gardiner, S.E., Bassett, H.C.M., Kodde, L.P., Plummer, K.M., 2005. The Vh8 locus of a new gene-for-gene interaction between Venturia inaequalis and the wild apple Malus sieversii is closely linked to the Vh2 locus in Malus pumila R12740-7A. New Phytol. 166, 1035–1049. doi:10.1111/j.1469-8137.2005.01395.x

Bus, V.G.M., Rikkerink, E.H.A., Caffier, V., Durel, C.-E., Plummer, K.M., 2011. Revision of the Nomenclature of the Differential Host-Pathogen Interactions

of Venturia inaequalis and Malus. Annu. Rev. Phytopathol. 49, 391–413. doi:10.1146/annurev-phyto-072910-095339

Bus, V.G.M., Rikkerink, E.H.A., Weg, W.E. van de, Rusholme, R.L., Gardiner, S.E., Bassett, H.C.M., Kodde, L.P., Parisi, L., Laurens, F.N.D., Meulenbroek, E.J., Plummer, K.M., 2005. The Vh2 and Vh4 scab resistance genes in two differential hosts derived from Russian apple R12740-7A map to the same linkage group of apple. Mol. Breed. 15, 103–116. doi:10.1007/s11032-004-3609-5

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., Jaffe, D.B., 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res. 18, 810–820. doi:10.1101/gr.7337908

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. doi:10.1186/1471-2105-10-421

Cantacessi, C., Jex, A.R., Hall, R.S., Young, N.D., Campbell, B.E., Joachim, A., Nolan, M.J., Abubucker, S., Sternberg, P.W., Ranganathan, S., Mitreva, M., Gasser, R.B., 2010. A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. Nucleic Acids Res. 38, e171–e171. doi:10.1093/nar/gkq667

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 37, D233–238. doi:10.1093/nar/gkn663

Celton, J.-M., Christoffels, A., Sargent, D.J., Xu, X., Rees, D.J.G., 2010. Genome-wide SNP identification by high-throughput sequencing and selective mapping allows sequence assembly positioning using a framework genetic linkage map. BMC Biol. 8, 155. doi:10.1186/1741-7007-8-155

Chagné, D., Crowhurst, R.N., Troggio, M., Davey, M.W., Gilmore, B., Lawley, C., Vanderzande, S., Hellens, R.P., Kumar, S., Cestaro, A., Velasco, R., Main, D.,

Rees, J.D., Iezzoni, A., Mockler, T., Wilhelm, L., Van de Weg, E., Gardiner, S.E., Bassil, N., Peace, C., 2012. Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple. PLoS ONE 7, e31745. doi:10.1371/journal.pone.0031745

Chang, Z., Wang, Z., Li, G., 2014. The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study. PLoS ONE 9, e94825. doi:10.1371/journal.pone.0094825

Cheung, F., Haas, B.J., Goldberg, S.M.D., May, G.D., Xiao, Y., Town, C.D., 2006. Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. BMC Genomics 7, 272. doi:10.1186/1471-2164-7-272

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E.G., Wetter, T., Suhai, S., 2004. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. Genome Res. 14, 1147–1159. doi:10.1101/gr.1917404

Chiara, M., Horner, D.S., Spada, A., 2013. De Novo Assembly of the Transcriptome of the Non-Model Plant Streptocarpus rexii Employing a Novel Heuristic to Recover Locus-Specific Transcript Clusters. PLoS ONE 8, e80961. doi:10.1371/journal.pone.0080961

Chikhi, R., Rizk, G., 2013. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol. Biol. AMB 8, 22. doi:10.1186/1748-7188-8-22

Clarke, K., Yang, Y., Marsh, R., Xie, L., Zhang, K.K., 2013. Comparative analysis of de novo transcriptome assembly. Sci. China Life Sci. 56, 156–162. doi:10.1007/s11427-013-4444-x

Clifton, S.W., Mitreva, M., 2009. Strategies for Undertaking Expressed Sequence Tag (EST) Projects, in: Parkinson, J. (Ed.), Expressed Sequence Tags (ESTs), Methods in Molecular Biology. Humana Press, pp. 13–32.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38, 1767–1771.

doi:10.1093/nar/gkp1137

Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. Nat Biotech 29, 987–991. doi:10.1038/nbt.2023

Conesa, A., Götz, 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, 3674–3676. doi:10.1093/bioinformatics/bti610

Conway, T.C., Bromage, A.J., 2011. Succinct data structures for assembling large genomes. Bioinformatics 27, 479–486. doi:10.1093/bioinformatics/btq697

Costa, V., Angelini, C., De Feis, I., Ciccodicola, A., 2010. Uncovering the Complexity of Transcriptomes with RNA-Seq. BioMed Res. Int. 2010. doi:10.1155/2010/853916

Coupland, P., Chandra, T., Quail, M., Reik, W., Swerdlow, H., 2012. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. BioTechniques 53. doi:10.2144/000113962

Cova, V., Paris, R., Passerotti, S., Zini, E., Gessler, C., Pertot, I., Loi, N., Musetti, R., Komjanc, M., 2010. Mapping and functional analysis of four apple receptor-like protein kinases related to LRPKm1 in HcrVf2-transgenic and wild-type apple plants. Tree Genet. Genomes 6, 389–403. doi:10.1007/s11295-009-0257-2

Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485. doi:10.1186/1471-2105-11-485

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., Hu, S., Yu, J., 2010. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics 96, 259–265. doi:10.1016/j.ygeno.2010.07.010

de Mendiburu, F. (2012). Statistical Procedures for Agricultural Research. Package "Agricolae", Comprehensive R Archive Network. Vienna, Austria: Institute for Statistics and Mathematics. Available at: http://cran.r-

project.org/web/packages/agricolae/agricolae.pdf

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.-R., Pan, H., et a., 2005. The genome sequence of the rice blast fungus Magnaporthe grisea. Nature 434, 980–986. doi:10.1038/nature03449

Dodds, P.N., Rathjen, J.P., 2010. Plant immunity: towards an integrated view of plant–pathogen interactions. Nat. Rev. Genet. 11, 539–548. doi:10.1038/nrg2812

Dodt, M., Roehr, J., Ahmed, R., Dieterich, C., 2012. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology 1, 895–905. doi:10.3390/biology1030895

Droege, M., Hill, B., 2008. The Genome Sequencer FLX™ System—Longer reads, more applications, straight forward bioinformatics and more complete data sets. J. Biotechnol., Genome Research in the Light of Ultrafast Sequencing Technologies 136, 3–10. doi:10.1016/j.jbiotec.2008.03.021

Duan, J., Xia, C., Zhao, G., Jia, J., Kong, X., 2012. Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics 13, 392. doi:10.1186/1471-2164-13-392

Duplessis, S., Cuomo, C.A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D.L., Hacquard, S., Amselem, J., Cantarel, B.L., et al., 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc. Natl. Acad. Sci. 201019315. doi:10.1073/pnas.1019315108

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7, e1002195. doi:10.1371/journal.pcbi.1002195

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133–138. doi:10.1126/science.1162986

Eisenman, H.C., Casadevall, A., 2012. Synthesis and assembly of fungal melanin. Appl. Microbiol. Biotechnol. 93, 931–940. doi:10.1007/s00253-011-3777-2

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107, 1–15. doi:10.1038/hdy.2010.152

Ekblom, R., Slate, J., Horsburgh, G.J., Birkhead, T., Burke, T., 2012. Comparison between Normalised and Unnormalised 454-Sequencing Libraries for Small-Scale RNA-Seq Studies. Comp. Funct. Genomics 2012, 1–8. doi:10.1155/2012/281693

El-Metwally, S., Hamza, T., Zakaria, M., Helmy, M., 2013. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PLoS Comput Biol 9, e1003345. doi:10.1371/journal.pcbi.1003345

Emrich, S.J., Barbazuk, W.B., Li, L., Schnable, P.S., 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res. 17, 69–73. doi:10.1101/gr.5145806

Eren, A.M., Vineis, J.H., Morrison, H.G., Sogin, M.L., 2013. A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. PLoS ONE 8, e66643. doi:10.1371/journal.pone.0066643

Erlich, Y., Mitra, P.P., delaBastide, M., McCombie, W.R., Hannon, G.J., 2008. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat. Methods 5, 679–682. doi:10.1038/nmeth.1230

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.

Feldmeyer, B., Wheat, C.W., Krezdorn, N., Rotter, B., Pfenninger, M., 2011. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12, 317. doi:10.1186/1471-2164-12-317

Fitzgerald, A., van Kan, J.A.L., Plummer, K.M., 2004. Simultaneous silencing of multiple genes in the apple scab fungus, Venturia inaequalis, by expression of RNA with chimeric inverted repeats. Fungal Genet. Biol. 41, 963–971.

doi:10.1016/j.fgb.2004.06.006

Forslund, K., Sonnhammer, E.L.L., 2008. Predicting protein function from domain content. Bioinformatics 24, 1681–1687. doi:10.1093/bioinformatics/btn312

Franchini, P., Merwe, M. van der, Roodt-Wilding, R., 2011. Transcriptome characterization of the South African abalone Haliotis midae using sequencing-by-synthesis. BMC Res. Notes 4, 59. doi:10.1186/1756-0500-4-59

Francis, W.R., Christianson, L.M., Kiko, R., Powers, M.L., Shaner, N.C., Haddock, S.H.D., 2013. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics 14, 167. doi:10.1186/1471-2164-14-167

Fraser, B.A., Weadick, C.J., Janowitz, I., Rodd, F.H., Hughes, K.A., 2011. Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome. BMC Genomics 12, 202. doi:10.1186/1471-2164-12-202

Frith, M.C., Bailey, T.L., Kasukawa, T., Mignone, F., Kummerfeld, S.K., Madera, M., Sunkara, S., Furuno, M., Bult, C.J., Quackenbush, J., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Pesole, G., Mattick, J.S., 2006. Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol. 3, 40–48.

Gachomo, E.W., Seufferheld, M.J., Kotchoni, S.O., 2010. Melanization of appressoria is critical for the pathogenicity of Diplocarpon rosae. Mol. Biol. Rep. 37, 3583–3591. doi:10.1007/s11033-010-0007-4

Gahlan, P., Singh, H.R., Shankar, R., Sharma, N., Kumari, A., Chawla, V., Ahuja, P.S., Kumar, S., 2012. De novo sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major transcriptome adjustments. BMC Genomics 13, 126. doi:10.1186/1471-2164-13-126

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., et al., 2003. The genome sequence of the filamentous fungus Neurospora crassa. Nature 422, 859–868. doi:10.1038/nature01554

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 Res. 15, 1620–1631. doi:10.1101/gr.3767105

Galperin, M.Y., Walker, D.R., Koonin, E.V., 1998. Analogous Enzymes: Independent Inventions in Enzyme Evolution. Genome Res. 8, 779–790. doi:10.1101/gr.8.8.779

Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Meth 8, 469–477. doi:10.1038/nmeth.1613

Garg, R., Patel, R.K., Tyagi, A.K., Jain, M., 2011. De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. DNA Res. 18, 53–63. doi:10.1093/dnares/dsq028

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., Nekrutenko, A., 2005. Galaxy: A platform for interactive large-scale genome analysis. Genome Res. 15, 1451–1455. doi:10.1101/gr.4086505

Gibbons, J.G., Janson, E.M., Hittinger, C.T., Johnston, M., Abbot, P., Rokas, A., 2009. Benchmarking Next-Generation Transcriptome Sequencing for Functional and Evolutionary Genomics. Mol. Biol. Evol. 26, 2731–2744. doi:10.1093/molbev/msp188

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., Martin, J.-F., 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12, 245. doi:10.1186/1471-2164-12-245

Gladieux, P., Zhang, X.-G., Afoufa-Bastien, D., Valdebenito Sanhueza, R.-M., Sbaghi, M., Le Cam, B., 2008. On the Origin and Spread of the Scab Disease of Apple: Out of Central Asia. PLoS ONE 3, e1455. doi:10.1371/journal.pone.0001455

Godfrey, D., Böhlenius, H., Pedersen, C., Zhang, Z., Emmersen, J., Thordal-Christensen, H., 2010. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. BMC Genomics 11, 317. doi:10.1186/1471-2164-11-317

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3435. doi:10.1093/nar/gkn176

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. doi:10.1038/nbt.1883

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R., 2003. Rfam: an RNA family database. Nucleic Acids Res. 31, 439–441. doi:10.1093/nar/gkg006

Gusberti, M., Gessler, C., Broggini, G.A.L., 2013. RNA-Seq Analysis Reveals Candidate Genes for Ontogenic Resistance in Malus-Venturia Pathosystem. PLoS ONE 8, e78457. doi:10.1371/journal.pone.0078457

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512. doi:10.1038/nprot.2013.084

Hadfield, J., Eldridge, M.D., 2014. Multi-genome alignment for quality control and contamination screening of next-generation sequencing data. Bioinforma. Comput. Biol. 5, 31. doi:10.3389/fgene.2014.00031

Harbers, M., Carninci, P., 2005. Tag-based approaches for transcriptome research and genome annotation. Nat. Methods 2, 495–502. doi:10.1038/nmeth768

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., Xie, Z., 2008. Single-Molecule DNA Sequencing of a Viral Genome. Science 320, 106–109. doi:10.1126/science.1150427

Hazelhurst, S., Hide, W., Lipták, Z., Nogueira, R., Starfield, R., 2008. An overview of the wcd EST clustering tool. Bioinformatics 24, 1542–1546. doi:10.1093/bioinformatics/btn203

Haznedaroglu, B.Z., Reeves, D., Rismani-Yazdi, H., Peccia, J., 2012. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics 13, 170. doi:10.1186/1471-2105-13-170

Haznedaroglu, B.Z., Reeves, D., Rismani-Yazdi, H., Peccia, J., 2012. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics 13, 170. doi:10.1186/1471-2105-13-170

Hesse, U., Mafofo, J., Mbandi, S., Oreetseng, M., Heusden, P., Husselmann, L., Rees DJG., and Christoffels A., 2013. “Genome of Venturia inaequalis–the causal agent of apple scab,” a Poster presented at the join Conference of Intelligent Systems for Molecular Biology and European Conference on Computational Biology, 21–23 July, Berlin, Germany.

Hoon, S., Ratnapu, K.K., Chia, J., Kumarasamy, B., Juguang, X., Clamp, M., Stabenau, A., Potter, S., Clarke, L., Stupka, E., 2003. Biopipe: A Flexible Framework for Protocol-Based Bioinformatics Analysis. Genome Res. 13, 1904–1915. doi:10.1101/gr.1363103

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., Nakai, K., 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35, W585–W587. doi:10.1093/nar/gkm259

Hou, R., Yang, Z., Li, M., Xiao, H., 2013. Impact of the next-generation sequencing

data depth on various biological result inferences. Sci. China Life Sci. 56, 104–109. doi:10.1007/s11427-013-4441-0

Howard, R.J., Ferrari, M.A., 1989. Role of melanin in appressorium function. Exp. Mycol. 13, 403–418. doi:10.1016/0147-5975(89)90036-4

Hsiang, T., Goodwin, P.H., 2003. Distinguishing plant and fungal sequences in ESTs from infected plant tissues. J. Microbiol. Methods 54, 339–351. doi:10.1016/S0167-7012(03)00067-8

Huang, X., Madan, A., 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9, 868–877. doi:10.1101/gr.9.9.868

Hutchison, C.A., 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. 35, 6227–6237. doi:10.1093/nar/gkm688

Ilie, L., Fazayeli, F., Ilie, S., 2011. HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics 27, 295–302. doi:10.1093/bioinformatics/btq653

Iseli, C., Jongeneel, C.V., Bucher, P., 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol. 138–148.

Jensen, K., Østergaard, P.R., Wilting, R., Lassen, S.F., 2010. Identification and characterization of a bacterial glutamic peptidase. BMC Biochem. 11, 47. doi:10.1186/1471-2091-11-47

Jha, G., Thakur, K., Thakur, P., 2010. The *Venturia* Apple Pathosystem: Pathogenicity Mechanisms and Plant Defense Responses. BioMed Res. Int. 2009. doi:10.1155/2009/680160

Jones, C.E., Baumann, U., Brown, A.L., 2005. Automated methods of predicting the function of biological sequences using GO and BLAST. BMC Bioinformatics 6, 272. doi:10.1186/1471-2105-6-272

Jones, D.C., Ruzzo, W.L., Peng, X., Katze, M.G., 2012. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. Nucleic Acids Res. doi:10.1093/nar/gks754

Jorda, J., Kajava, A.V., 2009. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics 25, 2632–2638. doi:10.1093/bioinformatics/btp482

Kawahara, Y., Oono, Y., Kanamori, H., Matsumoto, T., Itoh, T., Minami, E., 2012. Simultaneous RNA-Seq Analysis of a Mixed Transcriptome of Rice and Blast Fungus Interaction. PLoS ONE 7, e49423. doi:10.1371/journal.pone.0049423

Keller, N.P., Hohn, T.M., 1997. Metabolic Pathway Gene Clusters in Filamentous Fungi. Fungal Genet. Biol. 21, 17–29. doi:10.1006/fgbi.1997.0970

Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11, R116.

Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Res. 12, 656–664. doi:10.1101/gr.229202

Kershaw, M.J., Talbot, N.J., 1998. Hydrophobins and Repellents: Proteins with Fundamental Roles in Fungal Morphogenesis. Fungal Genet. Biol. 23, 18–33. doi:10.1006/fgbi.1997.1022

Kim, S., Ahn, I.-P., Rho, H.-S., Lee, Y.-H., 2005. MHP1, a Magnaporthe grisea hydrophobin gene, is required for fungal development and plant colonization. Mol. Microbiol. 57, 1224–1237. doi:10.1111/j.1365-2958.2005.04750.x

King, B.C., Waxman, K.D., Nenni, N.V., Walker, L.P., Bergstrom, G.C., Gibson, D.M., 2011. Arsenal of plant cell wall degrading enzymes reflects host preference among plant pathogenic fungi. Biotechnol. Biofuels 4, 4. doi:10.1186/1754-6834-4-4

King, G.J., Tartarini, S., Brown, L., Gennari, F., Sansavini, S., 1999. Introgression of the Vf source of scab resistance and distribution of linked marker alleles within the Malus gene pool. Theor. Appl. Genet. 99, 1039–1046. doi:10.1007/s001220051412

Klee, E.W., Ellis, L.B., 2005. Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics 6, 256. doi:10.1186/1471-2105-6-256

Koestler, T., Haeseler, A. von, Ebersberger, I., 2010. FACT: Functional annotation transfer between proteins with similar feature architectures. BMC

Bioinformatics 11, 417. doi:10.1186/1471-2105-11-417

Kollar, A., 1998. Characterization of an endopolygalacturonase produced by the apple scab fungus, Venturia inaequalis. Mycol. Res. 102, 313–319. doi:10.1017/S0953756297005194

Köller, W., 1991. Role of Cutinase in the Penetration of Apple Leaves by *Venturia inaequalis*. Phytopathology 81, 1375. doi:10.1094/Phyto-81-1375

Korban SS, Skirvin RM., 1984. Nomenclature of the cultivated apple. HortScience 19:177–80.

Korban, S.S., 1986. Interspecific hybridization in *Malus*. HortScience 21, 41–48.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol. 305, 567–580. doi:10.1006/jmbi.2000.4315

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol. 305, 567–580. doi:10.1006/jmbi.2000.4315

Kruger, W.M., Pritsch, C., Chao, S., Muehlbauer, G.J., 2002. Functional and Comparative Bioinformatic Analysis of Expressed Genes from Wheat Spikes Infected with *Fusarium graminearum*. Mol. Plant. Microbe Interact. 15, 445–455. doi:10.1094/MPMI.2002.15.5.445

Kucheryava, N., Bowen, J.K., Sutherland, P.W., Conolly, J.J., Mesarich, C.H., Rikkerink, E.H.A., Kemen, E., Plummer, K.M., Hahn, M., Templeton, M.D., 2008. Two novel Venturia inaequalis genes induced upon morphogenetic differentiation during infection and in vitro growth on cellophane. Fungal Genet. Biol. 45, 1329–1339. doi:10.1016/j.fgb.2008.07.010

Kumar, S., Blaxter, M.L., 2010. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11, 571. doi:10.1186/1471-2164-11-571

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., Murphy, J.W., 2004. Introns and Splicing Elements of Five Diverse Fungi. Eukaryot. Cell 3, 1088–1100. doi:10.1128/EC.3.5.1088-1100.2004

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-

efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. doi:10.1186/gb-2009-10-3-r25

Latunde-Dada, A.O., 2001. Colletotrichum: tales of forcible entry, stealth, transient confinement and breakout. Mol. Plant Pathol. 2, 187–198. doi:10.1046/j.1464-6722.2001.00069.x

Le, H.-S., Schulz, M.H., McCauley, B.M., Hinman, V.F., Bar-Joseph, Z., 2013. Probabilistic error correction for RNA sequencing. Nucleic Acids Res. 41, e109–e109. doi:10.1093/nar/gkt215

Leroy, T., Lemaire, C., Dunemann, F., Cam, B.L., 2013. The genetic structure of a Venturia inaequalis population in a heterogeneous host population composed of different Malus species. BMC Evol. Biol. 13, 64. doi:10.1186/1471-2148-13-64

Li, D., Deng, Z., Qin, B., Liu, X., Men, Z., 2012. De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (Hevea brasiliensis Muell. Arg.). BMC Genomics 13, 192. doi:10.1186/1471-2164-13-192

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. doi:10.1093/bioinformatics/btl158

Lisacek, F.C., Traini, M.D., Sexton, D., Harry, J.L., Wilkins, M.R., 2001. Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. Proteomics 1, 186–193. doi:10.1002/1615-9861(200102)1:2<186::AID-PROT186>3.0.CO;2-G

Littlejohn, K.A., Hooley, P., Cox, P.W., 2012. Bioinformatics predicts diverse Aspergillus hydrophobins with novel properties. Food Hydrocoll. 27, 503–516. doi:10.1016/j.foodhyd.2011.08.018

Liu, J., Gough, J., Rost, B., 2006. Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. PLoS Genet 2, e29. doi:10.1371/journal.pgen.0020029

Lizotte-Waniewski, M., Tawe, W., Guiliano, D.B., Lu, W., Liu, J., Williams, S.A.,

Lustigman, S., 2000. Identification of Potential Vaccine and Drug Target Candidates by Expressed Sequence Tag Analysis and Immunoscreening of Onchocerca volvulus Larval cDNA Libraries. Infect. Immun. 68, 3491–3501.

Logacheva, M.D., Kasianov, A.S., Vinogradov, D.V., Samigullin, T.H., Gelfand, M.S., Makeev, V.J., Penin, A.A., 2011. De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum). BMC Genomics 12, 30. doi:10.1186/1471-2164-12-30

Lu, B., Zeng, Z., Shi, T., 2013. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci. China Life Sci. 56, 143–155. doi:10.1007/s11427-013-4442-z

MacHardy, W. E. 1996. Apple Scab, Biology, Epidemiology and Management. American Phytopathological Society, St. Paul, MN.

MacHardy, W. E., and Jeger, M. 1983. Integrating control measures for the management of primary apple scab, Venturia inaequalis (Cke.) Wint. Prot. Ecol. 5:103-125.

MacHardy, W.E., Gadoury, D.M., Gessler, C., 2001. Parasitic and Biological Fitness of Venturia inaequalis: Relationship to Disease Management Strategies. Plant Dis. 85, 1036–1051. doi:10.1094/PDIS.2001.85.10.1036

MacManes, M.D., 2014. On the optimal trimming of high-throughput mRNA sequence data. Bioinforma. Comput. Biol. 5, 13. doi:10.3389/fgene.2014.00013

MacManes, M.D., Eisen, M.B., 2013. Improving transcriptome assembly through error correction of high-throughput sequence reads. PeerJ 1. doi:10.7717/peerj.113

MacPherson, S., Larochelle, M., Turcotte, B., 2006. A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins. Microbiol. Mol. Biol. Rev. 70, 583–604. doi:10.1128/MMBR.00015-06

Maher, C.A., Palanisamy, N., Brenner, J.C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T.R., Grasso, C., Yu, J., Lonigro, R.J., Schroth, G., Kumar-Sinha, C., Chinnaiyan, A.M., 2009. Chimeric transcript discovery by

paired-end transcriptome sequencing. Proc. Natl. Acad. Sci. doi:10.1073/pnas.0904720106

Mardis, E.R., 2009. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med. 1, 40. doi:10.1186/gm40

Mardis, E.R., 2013. Next-Generation Sequencing Platforms. Annu. Rev. Anal. Chem. 6, 287–303. doi:10.1146/annurev-anchem-062012-092628

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380. doi:10.1038/nature03959

Martin, J., Bruno, V.M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., Wang, Z., 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11, 663. doi:10.1186/1471-2164-11-663

Mbandi, S.K., Hesse, U., Rees, D.J.G., Christoffels, A.G., 2014. A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. Bioinforma. Comput. Biol. 5, 17. doi:10.3389/fgene.2014.00017

Meinhardt, L.W., Costa, G.G.L., Thomazella, D.P., Teixeira, P.J.P., Carazzolle, M.F., Schuster, S.C., Carlson, J.E., Guiltinan, M.J., Mieczkowski, P., Farmer, A., Ramaraj, T., Crozier, J., Davis, R.E., Shao, J., Melnick, R.L., Pereira, G.A., Bailey, B.A., 2014. Genome and secretome analysis of the hemibiotrophic fungal pathogen, Moniliophthora roreri, which causes frosty pod rot disease of cacao: mechanisms of the biotrophic and necrotrophic phases. BMC Genomics 15, 164. doi:10.1186/1471-2164-15-164

Melsted, P., Pritchard, J.K., 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics 12, 333. doi:10.1186/1471-2105-12-333

Menhaj, A.R., Mishra, S.K., Bezhani, S., Kloppstech, K., 1999. Posttranscriptional control in the expression of the genes coding for high-light-regulated HL#2 proteins. Planta 209, 406–413. doi:10.1007/s004250050743

Metzker, M.L., 2005. Emerging technologies in DNA sequencing. Genome Res. 15, 1767–1776. doi:10.1101/gr.3770505

Miller, H.C., Biggs, P.J., Voelckel, C., Nelson, N.J., 2012. De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (Sphenodon punctatus). BMC Genomics 13, 439. doi:10.1186/1471-2164-13-439

Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327. doi:10.1016/j.ygeno.2010.03.001

Morais do Amaral, A., Antoniw, J., Rudd, J.J., Hammond-Kosack, K.E., 2012. Defining the Predicted Protein Secretome of the Fungal Wheat Leaf Pathogen Mycosphaerella graminicola. PLoS ONE 7, e49904. doi:10.1371/journal.pone.0049904

Morozova, O., Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255–264. doi:10.1016/j.ygeno.2008.07.001

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628. doi:10.1038/nmeth.1226

Mount, D.W., 2007. Using the Basic Local Alignment Search Tool (BLAST). Cold Spring Harb. Protoc. 2007, pdb.top17. doi:10.1101/pdb.top17

Murray, I.A., Clark, T.A., Morgan, R.D., Boitano, M., Anton, B.P., Luong, K., Fomenkov, A., Turner, S.W., Korlach, J., Roberts, R.J., 2012. The methylomes of six bacteria. Nucleic Acids Res. gks891. doi:10.1093/nar/gks891

Nagaraj, S.H., Deshpande, N., Gasser, R.B., Ranganathan, S., 2007. ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Res. 35, W143–W147. doi:10.1093/nar/gkm378

Nagaraj, S.H., Gasser, R.B., Ranganathan, S., 2006. A hitchhiker's guide to expressed

sequence tag (EST) analysis. Brief. Bioinform. 8, 6–21. doi:10.1093/bib/bbl015

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, e91776. doi:10.1371/journal.pone.0091776

Needleman,S.B. andWunsch,C.D. (1970)A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453

Nguyen-Dumont, T., Pope, B.J., Hammet, F., Mahmoodi, M., Tsimiklis, H., Southey, M.C., Park, D.J., 2013. Cross-platform compatibility of Hi-Plex, a streamlined approach for targeted massively parallel sequencing. Anal. Biochem. 442, 127–129. doi:10.1016/j.ab.2013.07.046

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217. doi:10.1006/jmbi.2000.4042

O'Neil, D., Glowatz, H., Schlumpberger, M., 2013. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity, in: Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K. (Eds.), Current Protocols in Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Oono, Y., Kobayashi, F., Kawahara, Y., Yazawa, T., Handa, H., Itoh, T., Matsumoto, T., 2013. Characterisation of the wheat (triticum aestivum L.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in Pi-stressed wheat. BMC Genomics 14, 77. doi:10.1186/1471-2164-14-77

Ozsolak, F., Ting, D.T., Wittner, B.S., Brannigan, B.W., Paul, S., Bardeesy, N., Ramaswamy, S., Milos, P.M., Haber, D.A., 2010. Amplification-free digital gene expression profiling from minute cell quantities. Nat. Methods 7, 619–621. doi:10.1038/nmeth.1480

Palmieri, N., Schlötterer, C., 2009. Mapping accuracy of short reads from massively

parallel sequencing and the implications for quantitative expression profiling. PloS One 4, e6323. doi:10.1371/journal.pone.0006323

Pang, T., Ye, C.-Y., Xia, X., Yin, W., 2013. De novo sequencing and transcriptome analysis of the desert shrub, Ammopiptanthus mongolicus, during cold acclimation using Illumina/Solexa. BMC Genomics 14, 488. doi:10.1186/1471-2164-14-488

Papanicolaou, A., Stierli, R., Ffrench-Constant, R.H., Heckel, D.G., 2009. Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics 10, 447. doi:10.1186/1471-2105-10-447

Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W., Buerkle, C.A., 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11, 180. doi:10.1186/1471-2164-11-180

Paris, R., Cova, V., Pagliarani, G., Tartarini, S., Komjanc, M., Sansavini, S., 2009. Expression profiling in HcrVf2-transformed apple plants in response to Venturia inaequalis. Tree Genet. Genomes 5, 81–91. doi:10.1007/s11295-008-0177-6

Park, B.H., Karpinets, T.V., Syed, M.H., Leuze, M.R., Uberbacher, E.C., 2010. CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology 20, 1574–1584. doi:10.1093/glycob/cwq106

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., Soldatov, A., 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 37, e123–e123. doi:10.1093/nar/gkp596

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinforma. Oxf. Engl. 23, 1061–1067. doi:10.1093/bioinformatics/btm071

Paszkiewicz, K., Studholme, D.J., 2010. De novo assembly of short sequence reads. Brief. Bioinform. 11, 457–472. doi:10.1093/bib/bbq020

Patel, R.K., Jain, M., 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. PLoS ONE 7, e30619. doi:10.1371/journal.pone.0030619

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786. doi:10.1038/nmeth.1701

Pettersson, E., Lundeberg, J., Ahmadian, A., 2009. Generations of sequencing technologies. Genomics 93, 105–111. doi:10.1016/j.ygeno.2008.10.003

Pevzner, P.A., Tang, H., Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. 98, 9748–9753. doi:10.1073/pnas.171285098

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341. doi:10.1186/1471-2164-13-341

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120. doi:10.1093/nar/gki442

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al., 2013. A large-scale evaluation of computational protein function prediction. Nat. Methods 10, 221–227. doi:10.1038/nmeth.2340

Rafiqi, M., Jelonek, L., Akum, N.F., Zhang, F., Kogel, K.-H., 2013. Effector candidates in the secretome of Piriformospora indica, a ubiquitous plant-associated fungus. Plant-Microbe Interact. 4, 228. doi:10.3389/fpls.2013.00228

Rawlings, N.D., Barrett, A.J., Bateman, A., 2012. MEROPS: the database of

proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 40, D343–D350. doi:10.1093/nar/gkr987

Rawlings, N.D., Morton, F.R., 2008. The MEROPS batch BLAST: a tool to detect peptidases and their non-peptidase homologues in a genome. Biochimie 90, 243–259. doi:10.1016/j.biochi.2007.09.014

Remm, M., Storm, C.E.V., Sonnhammer, E.L.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052. doi:10.1006/jmbi.2000.5197

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277. doi:10.1016/S0168-9525(00)02024-2

Riesgo, A., Andrade, S.C.S., Sharma, P.P., Novo, M., Pérez-Porro, A.R., Vahtera, V., González, V.L., Kawauchi, G.Y., Giribet, G., 2012. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Front. Zool. 9, 33. doi:10.1186/1742-9994-9-33

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., et al., 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7, 909–912. doi:10.1038/nmeth.1517

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., Chisholm, S.W., 2010. Unlocking Short Read Sequencing for Metagenomics. PLoS ONE 5, e11840. doi:10.1371/journal.pone.0011840

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., Nyrén, P., 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242, 84–89. doi:10.1006/abio.1996.0432

Rossi, V., Giosuè, S., Bugiani, R., 2003. A model simulating deposition of Venturia inaequalis ascospores on apple trees*. EPPO Bull. 33, 407–414. doi:10.1111/j.1365-2338.2003.00665.x

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al., 2011. An

integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352. doi:10.1038/nature10242

Sadamoto, H., Takahashi, H., Okada, T., Kenmoku, H., Toyota, M., Asakawa, Y., 2012. De Novo Sequencing and Transcriptome Analysis of the Central Nervous System of Mollusc Lymnaea stagnalis by Deep RNA Sequencing. PLoS ONE 7, e42546. doi:10.1371/journal.pone.0042546

Salmela, L., Schröder, J., 2011. Correcting errors in short reads by multiple alignments. Bioinformatics 27, 1455–1461. doi:10.1093/bioinformatics/btr170

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467.

Schnabel, G., Jones, A.L., 2001. The 14α-Demethylasse(*CYP51A1*) Gene is Overexpressed in *Venturia inaequalis* Strains Resistant to Myclobutanil. Phytopathology 91, 102–110.

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, 1086–1092. doi:10.1093/bioinformatics/bts094

Sharov, A.A., Dudekula, D.B., Ko, M.S.H., 2005. Genome-wide assembly and analysis of alternative transcripts in mouse. Genome Res. 15, 748–754. doi:10.1101/gr.3269805

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145. doi:10.1038/nbt1486

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., Church, G.M., 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science 309, 1728–1732. doi:10.1126/science.1117389

Shimizu, K., Adachi, J., Muraoka, Y., 2006. ANGLE: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. J. Bioinform. Comput. Biol. 04, 649–664. doi:10.1142/S0219720006002260

Shrestha, R., Lubinsky, B., Bansode, V.B., Moinz, M.B., McCormack, G.P., Travers, S.A., 2014. QTrim: a novel tool for the quality trimming of sequence reads

generated using the Roche/454 sequencing platform. BMC Bioinformatics 15, 33. doi:10.1186/1471-2105-15-33

Shulaev, V., Korban, S.S., Sosinski, B., Abbott, A.G., Aldwinckle, H.S., Folta, K.M., Iezzoni, A., Main, D., Arús, P., Dandekar, A.M., Lewers, K., Brown, S.K., Davis, T.M., Gardiner, S.E., Potter, D., Veilleux, R.E., 2008. Multiple Models for Rosaceae Genomics. Plant Physiol. 147, 985–1003. doi:10.1104/pp.107.115618

Sims, A.H., Dunn-Coleman, N.S., Robson, G.D., Oliver, S.G., 2004. Glutamic protease distribution is limited to filamentous fungi. FEMS Microbiol. Lett. 239, 95–101. doi:10.1016/j.femsle.2004.08.023

Skamnioti, P., Gurr, S.J., 2008. Cutinase and hydrophobin interplay. Plant Signal. Behav. 3, 248–250.

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinform. 6:31.

Smeds, L., Künstner, A., 2011. ConDeTri - A Content Dependent Read Trimmer for Illumina Data. PLoS ONE 6, e26314. doi:10.1371/journal.pone.0026314

Smit AFA, Hubley R, Green P: RepeatMasker Open-3.0. 1996-2010. [http://www.repeatmasker.org]

Soanes, D.M., Chakrabarti, A., Paszkiewicz, K.H., Dawe, A.L., Talbot, N.J., 2012. Genome-wide Transcriptional Profiling of Appressorium Development by the Rice Blast Fungus Magnaporthe oryzae. PLoS Pathog 8, e1002514. doi:10.1371/journal.ppat.1002514

Soderlund, C., 2009. Computational techniques for elucidating plant–pathogen interactions from large-scale experiments on fungi and oomycetes. Brief. Bioinform. 10, 654–663. doi:10.1093/bib/bbp053

Song, C.-X., Clark, T.A., Lu, X.-Y., Kislyuk, A., Dai, Q., Turner, S.W., He, C., Korlach, J., 2012. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat. Methods 9, 75–77. doi:10.1038/nmeth.1779

Staden, R., 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 6, 2601–2610.

Stajich, J.E., Harris, T., Brunk, B.P., Brestelli, J., Fischer, S., Harb, O.S., Kissinger, J.C., Li, W., Nayak, V., Pinney, D.F., Stoeckert, C.J., Roos, D.S., 2011. FungiDB: an integrated functional genomics database for fungi. Nucleic Acids Res. 40, D675–D681. doi:10.1093/nar/gkr918

Stanke, M., Schöffmann, O., Morgenstern, B., Waack, S., 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62. doi:10.1186/1471-2105-7-62

Steiner, U., Oerke, E.-C., 2007. Localized Melanization of Appressoria Is Required for Pathogenicity of *Venturia inaequalis*. Phytopathology 97, 1222–1230. doi:10.1094/PHYTO-97-10-1222

Strickler, S.R., Bombarely, A., Mueller, L.A., 2012. Designing a transcriptome next-generation sequencing project for a nonmodel plant species1. Am. J. Bot. 99, 257–266. doi:10.3732/ajb.1100292

Sun, C., Li, Y., Wu, Q., Luo, H., Sun, Y., Song, J., Lui, E.M., Chen, S., 2010. De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11, 262. doi:10.1186/1471-2164-11-262

Sunde, M., Kwan, A.H.Y., Templeton, M.D., Beever, R.E., Mackay, J.P., 2008. Structural analysis of hydrophobins. Micron 39, 773–784. doi:10.1016/j.micron.2007.08.003

Surget-Groba, Y., Montoya-Burgos, J.I., 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 20, 1432–1440. doi:10.1101/gr.103846.109

Talbot, N.J., 2007. Plant pathology: Deadly special deliveries. Nature 450, 41–43. doi:10.1038/450041a

Tao, X., Gu, Y.-H., Wang, H.-Y., Zheng, W., Li, X., Zhao, C.-W., Zhang, Y.-Z., 2012. Digital Gene Expression Analysis Based on Integrated De Novo Transcriptome Assembly of Sweet Potato [Ipomoea batatas (L.) Lam.]. PLoS ONE 7, e36234. doi:10.1371/journal.pone.0036234

Thakur, K., Chawla, V., Bhatti, S., Swarnkar, M.K., Kaur, J., Shankar, R., Jha, G.,

2013. De Novo Transcriptome Sequencing and Analysis for Venturia inaequalis, the Devastating Apple Scab Pathogen. PLoS ONE 8, e53937. doi:10.1371/journal.pone.0053937

Tsai, H.-F., Wheeler, M.H., Chang, Y.C., Kwon-Chung, K.J., 1999. A Developmentally Regulated Gene Cluster Involved in Conidial Pigment Biosynthesis in Aspergillus fumigatus. J. Bacteriol. 181, 6469–6477.

Tseng, G.C., Ghosh, D., Feingold, E., 2012. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 40, 3785–3799. doi:10.1093/nar/gkr1265

Valsangiacomo, C., Gessler, C., 1992. Purification and characterization of an exo-polygalacturonase produced by Venturia inaequalis, the causal agent of apple scab. Physiol. Mol. Plant Pathol. 40, 63–77. doi:10.1016/0885-5765(92)90072-4

Van den Brink, J., de Vries, R.P., 2011. Fungal enzyme sets for plant polysaccharide degradation. Appl. Microbiol. Biotechnol. 91, 1477–1492. doi:10.1007/s00253-011-3473-2

Van Der Biezen, E.A., Jones, J.D.G., 1998. Plant disease-resistance proteins and the gene-for-gene concept. Trends Biochem. Sci. 23, 454–456. doi:10.1016/S0968-0004(98)01311-5

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., et al., 2010. The genome of the domesticated apple (Malus × domestica Borkh.). Nat. Genet. 42, 833–839. doi:10.1038/ng.654

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.

Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I., Marden, J.H., 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. 17, 1636–1647. doi:10.1111/j.1365-294X.2008.03666.x

Vogel, C., Berzuini, C., Bashton, M., Gough, J., Teichmann, S.A., 2004. Supra-

domains: Evolutionary Units Larger than Single Protein Domains. J. Mol. Biol. 336, 809–823. doi:10.1016/j.jmb.2003.12.026

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., Li, W., 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 41, e74–e74. doi:10.1093/nar/gkt006

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484

Wasmuth, J.D., Blaxter, M.L., 2004. prot4EST: Translating Expressed Sequence Tags from neglected genomes. BMC Bioinformatics 5, 187. doi:10.1186/1471-2105-5-187

Wenger, Y., Galliot, B., 2013. RNAseq versus genome-predicted transcriptomes: a large population of novel transcripts identified in an Illumina-454 Hydra transcriptome. BMC Genomics 14, 204. doi:10.1186/1471-2164-14-204

Wilhelm, B.T., Landry, J.-R., 2009. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48, 249–257. doi:10.1016/j.ymeth.2009.03.016

Win, J., Morgan, W., Bos, J., Krasileva, K.V., Cano, L.M., Chaparro-Garcia, A., Ammar, R., Staskawicz, B.J., Kamoun, S., 2007. Adaptive Evolution Has Targeted the C-Terminal Domain of the RXLR Effectors of Plant Pathogenic Oomycetes. Plant Cell 19, 2349–2369. doi:10.1105/tpc.107.051037

Wösten, H.A.B., 2001. HYDROPHOBINS: Multipurpose Proteins. Annu. Rev. Microbiol. 55, 625–646. doi:10.1146/annurev.micro.55.1.625

Wu, T.D., Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859–1875. doi:10.1093/bioinformatics/bti310

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G.K.-S., Wang, J., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. doi:10.1093/bioinformatics/btu077

Xu, D.-L., Long, H., Liang, J.-J., Zhang, J., Chen, X., Li, J.-L., Pan, Z.-F., Deng, G.-

B., Yu, M.-Q., 2012. De novo assembly and characterization of the root transcriptome of Aegilops variabilis during an interaction with the cereal cyst nematode. BMC Genomics 13, 133. doi:10.1186/1471-2164-13-133

Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342. doi:10.1038/nrg3174

Yang, Y., Smith, S.A., 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14, 328. doi:10.1186/1471-2164-14-328

Ye, C., Ma, Z., Cannon, C.H., Pop, M., Yu, D.W., 2012. Exploiting sparseness in de novo genome assembly. BMC Bioinformatics 13, S1. doi:10.1186/1471-2105-13-S6-S1

Yepes, L.M., 1993. Pathogenesis of *Venturia inaequalis* on Shoot-Tip Cultures and on Greenhouse-Grown Apple Cultivars. Phytopathology 83, 1155. doi:10.1094/Phyto-83-1155

Yike, I., 2011. Fungal Proteases and Their Pathophysiological Effects. Mycopathologia 171, 299–323. doi:10.1007/s11046-010-9386-2

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y., 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445–451. doi:10.1093/nar/gks479

Zdobnov, E.M., Apweiler, R., 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848. doi:10.1093/bioinformatics/17.9.847

Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829. doi:10.1101/gr.074492.107

Zhang, G., Wang, T., He, Q., 2014. How to discover new proteins—translatome profiling. Sci. China Life Sci. 57, 358–360. doi:10.1007/s11427-014-4618-1

Zhang, L., Yan, H.-F., Wu, W., Yu, H., Ge, X.-J., 2013. Comparative transcriptome analysis and marker development of two closely related Primrose species (Primula poissonii and Primula wilsonii). BMC Genomics 14, 329.

doi:10.1186/1471-2164-14-329

Zhang, Q., Ma, B., Li, H., Chang, Y., Han, Y., Li, J., Wei, G., Zhao, S., Khan, M.A., Zhou, Y., Gu, C., Zhang, X., Han, Z., Korban, S.S., Li, S., Han, Y., 2012. Identification, characterization, and utilization of genome-wide simple sequence repeats to identify a QTL for acidity in apple. BMC Genomics 13, 537. doi:10.1186/1471-2164-13-537

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., Hao, P., 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12, S2. doi:10.1186/1471-2105-12-S14-S2

Zhao, Z., Liu, H., Wang, C., Xu, J.-R., 2013. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 14, 274. doi:10.1186/1471-2164-14-274

Zhao, Z., Liu, H., Wang, C., Xu, J.-R., 2014. Correction: Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 15, 6. doi:10.1186/1471-2164-15-6

Zhao, Z., Tan, L., Dang, C., Zhang, H., Wu, Q., An, L., 2012. Deep-sequencing transcriptome analysis of chilling tolerance mechanisms of a subnival alpine plant, Chorispora bungeana. BMC Plant Biol. 12, 222. doi:10.1186/1471-2229-12-222

Zhou, X., Rokas, A., 2014. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. Mol. Ecol. 23, 1679–1700. doi:10.1111/mec.12680

Zhu, S., Dai, Y.-M., Zhang, X.-Y., Ye, J.-R., Wang, M.-X., Huang, M.-R., 2013. Untangling the transcriptome from fungus-infected plant tissues. Gene 519, 238–244. doi:10.1016/j.gene.2013.02.023

Zhuang, X., McPhee, K.E., Coram, T.E., Peever, T.L., Chilvers, M.I., 2012. Rapid transcriptome characterization and parsing of sequences in a non-model host-pathogen interaction; pea-Sclerotinia sclerotiorum. BMC Genomics 13, 668. doi:10.1186/1471-2164-13-668

Zwiers, L.-H., Stergiopoulos, I., Gielkens, M.M.C., Goodall, S.D., Waard, M.A.D., 2003. ABC transporters of the wheat pathogen Mycosphaerella graminicola function as protectants against biotic and xenobiotic toxic compounds. Mol. Genet. Genomics 269, 499–507. doi:10.1007/s00438-003-0855-x

Appendix 1

Table A4.1 Sources of Fungi proteomes used in this study

Fungi	№ of proteins	Source - date of download
P. tritici	12169	http://www.broadinstitute.org – October 2013
P. nodorum	12379	http://www.broadinstitute.org – October 2013
L. maculans	12469	ftp://ftp.ensemblgenomes.org release 21 - December 2013
M. graminicola	10933	http://genome.jgi-psf.org/Mycgr3/Mycgr3.download.ftp.html
B. fuckeliana	10351	ftp://ftp.ensemblgenomes.org 21 - December 2013
P. teres	11799	ftp://ftp.ensemblgenomes.org release 21 - January 2014
S. cerevisiae	5381	http://www.broadinstitute.org – October 2013

Table A4.2 Putative Apple resistance gene identified through InterProScan analysis

Transfrag ID	Interpro ID	Database	Domain ID
Metallothionein			
contig_12263	IPR001008	PRINTS	PR00875
contig_14277	IPR001008	PRINTS	PR00875
contig_18548	IPR000006	PRINTS	PR00860
contig_51843	IPR001396	PRINTS	PR00873
contig_5953	IPR001008	PRINTS	PR00875
Defensins			
contig_23890	IPR006081	ProSitePatterns	PS00269
contig_31864	IPR006081	ProSitePatterns	PS00269
contig_50614	IPR006081	ProSitePatterns	PS00269
Cystine knots			
contig_19800	IPR006207	ProSitePatterns	PS01185
contig_5488	IPR006207	ProSitePatterns	PS01185
contig_58884	IPR006207	ProSitePatterns	PS01185
contig_61488	IPR006207	ProSitePatterns	PS01185
contig_6995	IPR006207	ProSitePatterns	PS01185

Table A4.3 Orthophylogram distances between Venturia and selected hemibiotrophic fungi

Fungi species	№ of o*rthologs*		Ananthasubramanian et al., 2012		Berglund et al., 2007	
	Indian	S. African	Indian	S. African	Indian	S. African
C. graminicola	3066	3146	0.8612	0.8431	0.8119	0.8101
L. maculans	3066	3136	0.8656	0.8469	0.8237	0.8187
M. oryzae	2903	2943	0.8725	0.8543	0.8338	0.8290
P. tritici repentis	3182	3259	0.8589	0.8399	0.8118	0.8079
Z. tritici	2838	2948	0.8698	0.8527	0.8158	0.8153

Table A4.4 Selected most represented (InterPro) protein domains inferred from predicted peptides from the Indian (ordered list) and South African (unordered list) isolates

Interpro ID	Interpro description	Number of predicted peptides per isolate	
		Indian	South African
IPR001680	WD40 repeat	201	248
IPR011701	Major facilitator superfamily	122	105
IPR000719	Protein kinase domain	91	127
IPR001138	Zn(2)-C6 fungal-type DNA-binding domain	76	67
IPR018108	Mitochondrial substrate/solute carrier	61	89
IPR001128	Cytochrome P450	55	54
IPR001650	Helicase, C-terminal	55	53
IPR000504	RNA recognition motif domain	52	67
IPR002198	Short-chain dehydrogenase/reductase SDR	51	75
IPR003439	ABC transporter-like	51	42
IPR007219	Transcription factor domain, fungi	47	57
IPR000873	AMP-dependent synthetase/ligase	35	54
IPR005828	General substrate transporter	35	56
IPR010730	Heterokaryon incompatibility	34	34
IPR011545	DNA/RNA helicase, DEAD/DEAH box type, N-terminal	32	35

```
>contig_57585
Paths (1):
  Path 1: query 1..294 (294 bp) => genome MDC003623.651:13,892..13,601 (-292 bp)
    cDNA direction: indeterminate
    Genomic pos: malus.x.domestica:144,162,992..144,162,701 (- strand)
    Accessions: MDC003623.651:13,601..13,892 (out of 22003 bp)
    Number of exons: 1
    Coverage: 100.0 (query length: 294 bp)
    Trimmed coverage: 100.0 (trimmed length: 294 bp, trimmed region: 1..294)
    Percent identity: 98.0 (288 matches, 4 mismatches, 2 indels, 0 unknowns)
    Non-intron gaps: 1 openings, 2 bases in cdna; 0 openings, 0 bases in genome
    Translation: 2..294 (97 aa)
    Amino acid changes:

Alignments:
  Alignment for path 1:

    -MDC003623.651:13892-13601  (1-294)   97%

                    0     .    :    .    :    .    :    .    :    .    :
aa.g                1   K  N  Q  S  T  C  S  P  S  L  S  L  S  L  L  L  P
 -MDC003623.651:13892 AAAAAATCAATCTACTTGTTCTCCATCTCTCTCTCTCTCTGCTTCTCC
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
                    1 AAAAAATCAATCTACTTGTTCTCCATCTCTCTCTCTCTCTGCTTCTCC
aa.c                1   K  N  Q  S  T  C  S  P  S  L  S  L  S  L  L  L  P

                   50     .    :    .    :    .    :    .    :    .    :
aa.g               18     L  S  L    L  C  S  A  K  L  E  G  P  K  E  K  K
 -MDC003623.651:13842 CTCTCTCTCTC  CTCTGCTCTGCGAAACTCGAGGGTCCGAAAGAGAAGA
                      |||||||||||--||||||||||||||||||||||||||||||||||||||
                   51 CTCTCTCTCTCTCCTCTGCTCTGCGAAACTCGAGGGTCCGAAAGAGAAGA
aa.c               18     L  S  L  S  S  A  L  R  N  S  R  V  R  K  R  R

                  100     .    :    .    :    .    :    .    :    .    :
aa.g               34     N  K  Q  T  N  S  Q  N  E  T  A  N  R  R  K  T
 -MDC003623.651:13794 AGAACAAACAAACGAATTCACAAAACGAAACAGCAAATCGACGCAAAACC
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
                  101 AGAACAAACAAACGAATTCACAAAACGAAACAGCAAATCGACGCAAAACC
aa.c               33 R  T  N  K  R  I  H  K  T  K  Q  Q  I  D  A  K  P

                  150     .    :    .    :    .    :    .    :    .    :
aa.g               50 Q  L  P  I  L  A  Q  T  P  T  H  L  S  V  F  G  S
 -MDC003623.651:13744 CAACTCCCAATCCTCGCGCAGACCCCAACCCACCTCTCTGTTTTCGGATC
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
                  151 CAACTCCCAATCCTCGCGCAGACCCCAACCCACCTCTCTGTTTTCGGATC
aa.c               50  N  S  Q  S  S  R  R  P  Q  P  T  S  L  F  S  D  Q

                  200     .    :    .    :    .    :    .    :    .    :
aa.g               67  S  D  P  D  P  P  I  R  P  T  Q  L  S  P  S  D  S
 -MDC003623.651:13694 AAGTGATCCGGATCCTCCGATCCGACCAACCCAACTCTCCCCATCCGATT
                      ||||||||||||||||||||||||||||||||||| ||| |||||||||
                  201 AAGTGATCCGGATCCTCCGATCCGACCAACCCAACTCGCCCGATCCGATT
aa.c               67    V  I  R  I  L  R  S  D  Q  P  N  S  P  D  P  I

                  250     .    :    .    :    .    :    .    :
aa.g               84    I  F  C  I  T  D  D  S  E  V  L  A  L  Q
 -MDC003623.651:13644 CCATTTTCTGCATTACCGATGACTCGGAGGTGCTCGCATTGCAG
                      ||||||||  |||||  ||||||||||||||||||||||||||||||
                  251 CCATTTTCCGCATTGCCGATGACTCGGAGGTGCTCGCATTGCAG
aa.c               83 P  F  S  A  L  P  M  T  R  R  C  S  H  C
```

Figure A4.1 Alignment of transfrag, contig_57585 to the apple genome

Appendix 2

Table A5.1 Sources of Fungi proteomes used in this study

Fungi	№ of proteins	Source - date of download
A. laibachii	13,804	ftp://ftp.ensemblgenomes.org release 21 - Decembe 2013
A. brassicicola	10,688	http://genome.jgi-psf.org/Altbr1/Altbr1.home.html 11.01.2012
A. nidulans	10,560	http://www.broadinstitute.org - October 2013
C. albicans	5,931	http://www.broadinstitute.org - October 2013
C. globosum	11,124	http://www.broadinstitute.org - October 2013
C. fulvum	14,127	http://genome.jgi.doe.gov/Clafu1/Clafu1.home.html 29.09.2013
C. sativus	12,250	http://genome.jgi-psf.org - 11.01.2012
C. graminicola	12,006	http://www.broadinstitute.org - 11.01.2012
E. festucae	9,273	http://csbio-l.csr.uky.edu/m3 - 29.09.2013
F. graminearum	13,321	http://www.broadinstitute.org - September 2013
H. capsulatum	9,251	http://www.broadinstitute.org - October 2013
H. pulicare	12,352	http://genome.jgi-psf.org - 29.09.2013
L. maculans	12,469	ftp://ftp.ensemblgenomes.org release 21 - December 2013
M. oryzae	12,991	http://www.broadinstitute.org - 29.09.2013
N. crassa	10,785	http://www.broadinstitute.org - 29.09.2013
P. nodorum	12,379	http://www.broadinstitute.org - October 2013
P. infestans	18,140	http://www.broadinstitute.org -13.11.2013
P. graminis	15,979	http://www.broadinstitute.org - 07.10.2013
P. tritici-repentis	12,169	http://www.broadinstitute.org - October 2013
R. rufulum	12,117	http://genome.jgi-psf.org - 29.09.2013
S. cerevisiae	5,381	http://www.broadinstitute.org - October 2013
M. graminicola	10,933	http://www.jgi.doe.gov/Mgraminicola - October 2013
U. maydis	6,522	http://www.broadinstitute.org - 07.10.2013
V. dahliae	10,535	http://www.broadinstitute.org - 07.10.2013