Chapter 1

An introduction to ultra-high-throughput DNA sequencing technologies and their application in genetics and functional genomics

1.1. Introduction

Eucalypt forest trees supply high quality raw material for the pulp, paper and wood industries, and have been identified as important role-players in the search for renewable energy resources. Eucalypts are hardy, fast growing and have a high dry matter production and resprouting potential, which makes them one of the most widely used tree species in industrial hardwood plantations (Forrest and Moore, 2008; Rengel *et al.*, 2009). In recent years, the global forestry industry has experienced a steady shift in location from the northern hemisphere to the tropics and subtropics, where it is actively competing with food crops for land space needed for expansion (Grattapaglia and Kirst, 2008). In South Africa, a recent report from the South African Department of Water Affairs and Forestry (DWAF) indicated that 1.25 million hectares (1.1%) of South Africa's total land area are covered by forestry plantations, of which 36% (450 000 hectares) are planted with *Eucalyptus* species (http://www2.dwaf.gov.za/webapp/Documents/FSA-Abstracts2009.pdf). The economic importance of plantation trees as renewable energy and biomass producing crops makes them excellent candidates for genetic improvement studies.

Eucalypts have a high fiber count of uniform nature, a sought after property that has created high demands in the pulp, paper and raw wood industries (Moore *et al.*, 2008). Large numbers of genes are affecting wood formation in forest trees, and have been actively investigated by various research groups focusing on key properties, such as wood density, pulp yield, cellulose content, fiber length and lignin content (for a review of the state of *Eucalyptus* breeding see Myburg *et al.*, 2005). Improvements to biomass yield and fiber quality with improved breeding programs and the direct application of biotechnology advances to crop development will play increasingly important roles in the future of the eucalypt forestry industry.

Woody biomass has been identified as important in the search for renewable energy resources. The United States Department of Energy (US-DOE) announced in 2007 their goal to reduce the usage of gasoline in the United States by 20% by the year 2017 (http://genomicscience.energy.gov/biofuels/). To achieve this, an expansion of the annual renewable fuel supply from a variety of plant materials, including grasses, woodchips and agricultural wastes needs to occur. The bioenergy initiative actively supported the research community in successfully determining the genomic sequence of the *Populus trichocarpa* genome (Tuskan *et al.*, 2006) and the *Eucalyptus grandis* genome (version 1.0 released in January 2011, http://www.phytozome.net) by the Joint Genome Institute (JGI). It is expected that fast growing, short-rotation woody crops such as *Eucalyptus* and *Populus* and their respective hybrids will contribute up to 30% of the biomass of the so-called "energy crops" (Hinchee *et al.*, 2009).

Advances in the fields of biotechnology, genetics and computer science have resulted in an unprecedented growth in the amount of biological data being generated on a daily basis by the scientific community. This aided the slow, but definitive paradigm shift from a hypothesis-driven scientific approach to a data-driven, explorative approach. Next-generation DNA sequencing technologies (NGS) have opened the floodgates in terms of biological sequence data generation. Since the first application of NGS by Margulies *et al.* (2005), various technological improvements have led to higher and higher base pair throughput from NGS platforms. As stated in the preamble of this document, the term ultra-high-throughput sequencing (uHTS) will be used in the rest of this manuscript to denote the different high throughput DNA

sequencing technologies (next generation sequencers, second generation sequencers and third generation sequencers, Werner, 2010).

High-throughput experiments now commonly investigate the range of gene expression products between different organisms, between tissues within organisms, or between tissues of the same organism in different disease states in order to investigate underlying molecular basis of a phenotype. Pyrosequencing technologies have effectively revolutionised the approach and turnover time needed to sequence and re-sequence genomes. Applications of uHTS technologies are evident in the advances made in the fields of mutation discovery, metagenomic characterisation, non-coding RNA and DNA-protein interaction discovery (Mardis, 2008). The data produced from these high-throughput experiments have resulted in a biological data glut, where gigabases of data are produced in a single experiment and biologists are now forced to design and follow efficient data management practices for experiments.

Sequencing large numbers of mRNAs from a sample forms the basis of the revolutionary expressed sequence tag method (EST) used for identifying genes during the human genome project (Adams *et al.*, 1991; Venter *et al.*, 2001). The costly nature, long experimental run time, low quality reads and general inability to detect transcripts expressed at a low level has hampered the technology from being widely used (Graveley, 2008). The parallel nature of next-generation sequencing makes it a ideal technology for transcriptome sequencing, generating hundreds of millions of short reads (35-350 base pairs (bp) long). Many research groups have employed a technology called mRNA-Seq (Section 1.3) to sequence at various levels of detail and complexity the transcriptomes of a diverse set of organisms (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008; Novaes *et al.*, 2008; Nagalakshmi *et al.*, 2008). Transcriptome studies have revealed, among others, differences in transcript abundance, efficiency of the machinery active during intron removal and detection of alternatively spliced transcripts between different tissues and/or organisms of interest. Improvements in the technology in terms of read length, the ability to perform paired-end sequencing, strand-specific sequencing and improved algorithms to assemble short reads will provide even greater insight into the transcriptome landscape (Graveley, 2008).

The following sections will focus on the different ultra-high-throughput DNA sequencing platforms

available in the market with specific focus on the applications of these technologies to the fields of genetics and functional genomics. A brief discussion regarding the data management issues involved in working with and analysing data from these platforms is then followed by a section dedicated to defining the main problem statement of this study. The final section of the chapter includes an outline of the specific aims and requirements in order to achieve the goals of this study.

1.2. Ultra-high-throughput DNA sequencing platforms

Ultra-high-throughput sequencing (uHTS) technologies have been categorically assigned to one of the following groups: microelectrophoretic methods, sequencing by hybridisation, real-time observation of single molecules and cyclic-array sequencing (Shendure *et al.*, 2004). The current technological advances made with cyclic-array sequencing has proven this to be the most successful approach by far, as is evident in the implementation of this technology in various commercial products. These products, defined in the literature as Next Generation Sequencing (NGS) platforms, or more recently Second Generation Sequencing (SGS) platforms (Kislyuk *et al.*, 2005), include the 454 Genome Sequencer (Roche Applied Science, Margulies *et al.*, 2005), Solexa technology (Illumina Genome Analyser, Fedurco *et al.*, 2006; Turcatti *et al.*, 2008) and the SOLiD platform (Applied Biosystems, Shendure *et al.*, 2005). Very recently, the term of Third Generation Sequencers (TGS) emerged with the advent of single molecule sequencers (Schuster, 2008). Of these systems, the most prolific commercial offerings include the Heliscope Single Molecule Sequencer (Helicos, Braslavsky *et al.*, 2003) and the Single Molecule Real Time (SMRT) sequencing platform from Pacific Biosciences (Eid *et al.*, 2009), but the nanoball sequencing platform from Complete Genomics (Drmanac *et al.*, 2010) and the innovative Ion Torrent (unpublished) platforms are also available.

1.2.1. Cyclic array sequencing applications

The first practical implementations of uHTS technologies included the *de novo* sequencing and assembly of the *Mycoplasma genitalium* genome (Margulies *et al.*, 2005), and the re-sequencing of an evolved

Escherichia coli strain (Shendure *et al.*, 2005). Since these seminal papers were published, different applications have been developed in which high-throughput technologies were employed in various biological scenarios which will be discussed in Section 1.3. Although the different uHTS platforms use diverse DNA sequencing biochemistry and follow different methodologies in terms of array generation, a general workflow common to most technologies can be envisioned. Most cyclic-array technologies rely on the random fragmentation of a target DNA library, followed by the *in vitro* ligation of a specific set of adaptor sequences. In the case of paired-end sequencing, a so-called "jumping" library of mate-pair tags with a controllable distance between them is generated (Ng *et al.*, 2005; Shendure *et al.*, 2005). Following amplification of the target sequences on a custom array, the sequencing process is achieved by alternative cycles of flushing enzymes across a target array in order to drive a biochemical process. At every step during the sequencing process an image capture device is used to record the chemical reaction taking place at every position on the array. Various downstream computational approaches are then available to produce a string of characters with associated quality or confidence values representing the DNA sequence hybridised to the specific position on the array.

454 GS FLX Pyrosequencing (Roche Applied Science)

The 454 pyrosequencer relies on the principle of 'pyrosequencing' which employs the biochemical cleavage of a pyrophosphate molecule released during nucleotide incorporation by DNA polymerase in order to set off a chain of reactions, which will ultimately produce a burst of light from the cleavage of oxyluciferin by luciferase (Margulies *et al.*, 2005). Initially developed by 454 Life Sciences, the technology was the first widely adopted high-throughput sequencing technology and has a well-established user community. As per the general protocol, sequencing libraries are constructed that give rise to a mixture of short, adaptor-flanked fragments. These fragments are then clonally amplified with emulsion PCR inside picoliter reactors on a custom array, with amplicons captured to the surface of 28-μm beads (Tawfik and Griffiths, 1998; Ghadessy *et al.*, 2001; Margulies *et al.*, 2005). A sequencing primer is hybridised to the universal adaptor at the appropriate position and orientation, and the pyrosequencing reaction initiated (Margulies *et al.*, 2005).

Several hundred cycles of pyrosequencing involves the inclusion of a single species of fluorescently-labeled nucleotides to the microtiter wells, and in wells where a base is incorporated, a pyrophosphate molecule is released. One reaction takes place for every base that is incorporated in the sequence, which leads to signal saturation when more than four or five bases are incorporated during homopolymer runs of the sequence (Margulies *et al.*, 2005). The nature of the technology results in asynchronous sequencing of the wells, in other words when the 'A'-base reaction takes place, multiple reactions might take place in some wells where more than one complimentary base is exposed. At the same time in wells where the template does not have a complimentary base no reaction will take place. The incorporation of bases is measured in sequence by a live capture of a charged coupled device (CCD, or camera) from the array.

At the time of writing, approximately 800 papers had been published making use of 454 pyrosequencing, including very diverse applications in metagenomics, novel and re-sequenced genomes and plasmids, population diversity determination, RNA discovery and function inferences, epigenetic studies, transcriptome studies and genome structural variant investigations (for a review on the use of high-throughput sequencing technologies in functional genomics, see Section 1.3). The GS FLX Titanium series produce between 400 and 600 million high quality bases per run with an average read length of 400 bases, which amounts to just over 100 million high quality reads per run. The long read lengths make this technology ideal for *de novo* genome sequencing projects of various organisms (http://www.454.com). The issue with the homopolymer run base calls is an inherent feature of the technology, and can only be overcome by employing a more sensitive light intensity detection system (Rothberg and Leamon, 2008).

Illumina Genome Analysis (Illumina)

The development of the Illumina platform was derived from the initial work of Turcatti and colleagues on benzene-1,3,5-triacetic acid (BTA) and reversible deoxynucleotide terminators (Fedurco *et al.*, 2006; Turcatti *et al.*, 2008). The core methodology consists of adaptor-flanked DNA fragments of a couple of hundred base pairs that are amplified by a bridge PCR method. During this phase of the bridge PCR protocol, both forward and reverse primers are attached to a glass surface, in such a manner as to allow for the grouping of all amplified constructs from a single template in a cluster. During each

step of the bridge PCR, the reaction alternatively extends the template sequence with *Bst* polymerase and then denatures the double stranded sequence with formaldehyde (Turcatti *et al.*, 2008). After the amplification step, each cluster on the glass array should be represented by roughly 1 000 clonal amplicons, thus the initial concentration of the sequencing library needs to be known. The amplification process is highly parallelised, resulting in several million clusters amplified at distinct positions within each of the independent lanes on the array, or flow cell (Turcatti *et al.*, 2008). After cluster construction, the amplified constructs are denatured into single strands, and a sequencing primer is hybridised to the adaptor.

The sequencing process involves the single base-pair extension of the template sequence with a modified deoxynucleotide base. The deoxynucleotide base is modified in two ways; first, it is a reversible terminator, and secondly; it is fluorescently labeled to correspond to each of the four nucleotide bases. After incorporation of the modified deoxynucleotide base on the sequencing strand, chemical cleavage is needed to remove the 3' hydroxyl position, and the attached fluorescent molecule again starts a chain of reactions ending in the emission of a light signal. A CCD device captures the signal and the incorporated base is then computationally determined in downstream analysis of the images (with the Illumina analysis tools `Firecrest` and `Bustard`). The array is then prepared for the next cycle of base incorporation by enzymatically removing the blocking position of the incorporated base, and the next round of bases are flushed over the array. At every cycle of the sequencing process, only one base can be incorporated on the sequencing strand resulting in synchronous probe sequencing.

In contrast to the 454 sequencing, Illumina tends to focus on throughput rather than the lengths of the reads obtained from a sequencing run. At present, read lengths of up to 100 bp are possible, but there is a drop in quality of the reads as the read reaches the maximum read length. An example of the drop in quality of a 76 bp run of sequencing is presented in Figure 2.4, where a drop in base-quality can be observed from around base 68. The development of the paired-end protocol, where the both ends of the amplicons are sequenced, together with the extremely high-throughput (500 Gbp) on the HiSeq2000 platform, has made this technology ideal for genome re-sequencing and transcriptome studies where

the digital expression on a specific transcript can be measured (http://www.illumina.com). The factors limiting the technology to produce longer read lengths include the incomplete enzymatic cleavage of the fluorescent labels or terminal moieties, which leads to a decay in the detection signal and eventually leads to dephasing of the reaction (Shendure and Ji, 2008). Illumina technology suffers from a base substitution error, rather than an insertion or deletion as observed with the 454 platform. Average raw error rates have been reported to be in the order of 1-1.5%, but higher accuracy bases with error rates down to 0.1% can be achieved (Shendure and Ji, 2008).

SOLiD (Applied Biosystems)

The original work of Shendure *et al.* (2005) and patents by McKernan *et al.* (2006) directly led to the development of the unique two-base encoding methodology behind Applied Biosystem's SOLiD system. As with the other systems discussed thus far, a fragmented DNA library of adaptor-flanked regions serve as the starting point for this technology. Cloning of the fragments is achieved with emulsion PCR, with the amplicons captured to the surface of 1µm beads (Dressman *et al.*, 2003). After breaking the emulsion, the amplicon-containing beads are immobilised to a solid planar substrate in order to generate a dense, disordered array of beads (Shendure and Ji, 2008). After the addition of a universal primer that ligates to the amplicons, the rather complex sequencing process can begin.

A notable difference between the SOLiD and the methods mentioned previously is that the sequencing reaction is driven by a DNA ligase rather than a polymerase, and is achieved by ligating a degenerate fluorescent octamer to the template (Shendure *et al.*, 2005). The octamer mixture is structured so that the identity of a specific base in the octamer corresponds to the fluorescent label of the octamer. After ligation and image capture with a CCD, the octamer is chemically cleaved between positions three and six, removing the fluorescent label. In effect progressive rounds of octamer ligation results in the sequencing of every fifth base (Shendure and Ji, 2008). After several cycles, the extended primer is denatured and the system is reset to its original state. The process is repeated, each time sequencing a different position in the octamer by either using an initial primer of a different length or by using a different position in the octamer as the fluorescent label. An additional complication to the system is

that an error correction method is in place. Effectively two adjacent bases correspond to the selected fluorescent label, and each base position is then queried twice, once as the first base and once as the second base, during a given cycle. A graphical representation of the sequencing cycle with the two base encoding system can be viewed on the company's website (http://www.appliedbiosystems.com).

The result from the two-base encoding system is that very accurate base qualities (>99.94 % accuracy) are achieved with the SOLiD system (http://www.appliedbiosystems.com). Read lengths were initially limited to 36 bp, but steadily increased to 75 bp. The high quality of the reads, as well as the very high-throughput of 300 Gb per run from the SOLiD 5500xl System puts it in the same application space as the Illumina platform. The confidence in the quality of the reads also provides a good platform for polymorphism studies. Since the output from the SOLiD system is in "color space" and not "base space", decoding of the reads into base space needs to occur before any analysis can be performed on the results. Most widely used sequence mapping and assembly tools have been adapted to cater for working in "color space", and a variety of converters exists which will convert "color space" reads to "base space".

Complete Genomics (Complete Genomics)

Drmanac *et al.* (2010) described another DNA sequencing technology making use of self-assembling DNA nanoarrays and demonstrated it by re-sequencing three human genomes. The technology employs recursive restriction site cutting (type IIS restriction enzymes) and directional adaptor insertion methods to produce circled DNA replicated many times with a polymerase in order to create DNA nanoballs (Drmanac *et al.*, 2010). These nanoballs are attached to a photolithographic surface, and the sequence adjacent to the inserted directional adaptor sites sequenced using a high-accuracy combinatorial probe anchor ligation (cPAL) technology. cPAL uses degenerate anchors in order to read up to 10 bp adjacent to the inserted adaptor sites, with similar read accuracy across all the bases read. This method produced between 31-35 bp mate-paired reads.

Using nanoarray sequencing the average amount of sequence produced from three human genomes ranged from 124 Gb to 241 Gb, which corresponds to a coverage between 45X and 85X (Drmanac *et al.*, 2010). In terms of sequence quality and polymorphism calls, the authors achieved confident diploid calls

for up to 95% of the theoretical 98% of a Yoruban female genome (HapMap id: NA19240), with close to 94% of the SNP positions called (99.15% accuracy) in the HapMap phase I/II for the caucasian genome (NA07022).

Ssequencing-by-synthesis, and sequencing-by-ligation-based technologies use chained reads, where the substrate for cycle *N+1* depends on the product of cycle *N*. The ligation-based approach described by Drmanac *et al.* (2010) uses an unchained approach, where complete probes are ligated to the target sequences, and the sequencing process does not depend on driving the reaction to completion with high concentrations of labeled nucleotides as used in other methods. Because of the lack of high concentrations of purified fluorescently labeled substrates, the average cost per sequenced genome was reduced to under US$4 400. The short reads obtained from this technology and the late introduction of the commercial product to the market are some of the initial hurdles to overcome in order to ensure widespread adoption, but with the reduced cost this can be an attractive platform alternative to the Illumina and SOLiD platforms.

1.2.2. Single-molecule sequencing platforms

Single-molecule sequencers have been earmarked as the next big technological development aiming to achieve the target of sequencing a human genome for US$1 000. At the time of writing, only the Helicos Biosciences system was available as a commercial application, but the commercial launch of the Pacific Biosciences Single Molecule Real Time (SMRT™) system was imminent according to the company. The Ion Torrent system was first announced at the 2010 Advances in Genome Biology and Technology (AGBT, http://agbt.org) meeting, and received much attention that warrants its inclusion in the following section. Oxford Nanopore's sequencing system is still in development, and little information is available on the technical aspects of the system, and is therefore not covered in this review.

SMRT™ sequencing (Pacific Biosciences)

The technology that led to the development of Pacific Biosciences' single-molecule sequencer was first described by Eid *et al.* (2009). The technology also relies on the incorporation of a fluorescently-labeled

nucleotide complementary to the target strand being sequenced. A notable difference with the nature of the fluorescently-labeled nucleotide, is that the nucleotide is labeled on the phosphate group. This labeling strategy has the effect that the fluorescent label is naturally cleaved from the nucleotide together with the phosphate group during nucleotide incorporation into the synthesized strand. Another unique feature of the Pacific Biosystems system is that rather than fixing the DNA template to an array and flushing enzymes across it, the DNA polymerase enzyme is fixed to the array, with fragmented DNA and labeled nucleotides flowing over the array. The technology involves binding a DNA polymerase (Φ29) on a polyglycol-covered silica surface without direct interaction between the protein and the silica surface (Eid *et al.*, 2009). The seating of the polymerase protein occurs inside a zeptoliter (10^{-21} liter) well, which is small enough to allow a single fragmented DNA strand to enter, along with labeled nucleotides. Multiple wells are constructed in an aluminum cladding, known as the Zero-mode Waveguide (ZMW), in which the sequencing reaction occurs. Apart from functioning as a micro-reactor for the sequencing reaction, the ZMW reduces the background light noise which occurs in other wells on the ZMW, and allows for the detection of the light emitted from a single molecule of the fluorescently-labeled phosphate as nucleotides are incorporated by polymerase in real time (Single-molecule, real time (SMRT™) sequencing, Eid *et al.*, 2009). Since the whole process proceeds as fast as the DNA polymerase can incorporate bases into the template sequence, an average per base incorporation rate four orders of magnitude faster than second generation sequencers can be achieved. By simply manufacturing more wells on the ZMW, the reaction can occur in parallel, and comparable base pair throughput should be achievable in the future.

The use of SMRT™ sequencing has led to the development of a novel method of DNA circularisation, coined SMRTbell™, for consensus sequencing of the same molecule (Travers *et al.*, 2010). Using these circular templates which represents a linear DNA fragment, multiple passes of sequencing are performed, providing multiple copies of the same molecule. A demonstrative application of the technology was in re-sequencing a housekeeping gene (aroE132) with a single nucleotide difference between two strains of Multiple Resistance *Staphylococcus aureus* (MRSA, the FDA209 and Mu50 strains). By mixing the DNA fragments of the aroE132 gene from these two strains in different ratios, the robustness of the system to

detect the frequency of a single nucleotide difference within the samples was detemined (Travers *et al.*, 2010).

Flusberg *et al.* (2010) showed that detection of DNA methylation without bisulfite treatment was possible with SMRT™ sequencing, avoiding some of the drawbacks of bisulfite sequencing, which includes the costly sample preparations used in methylation studies, the constraints in primer design of a treated genome, and the ambiguities in alignments of the generated sequences to the reference genome. By measuring the pulse duration from the phosphate cleavage by DNA polymerase of the labled nucleotides, a difference in the polymer kinetics inside the ZMW well between methylated and non-methylated sites could be detected. The use of circular consensus sequencing aided in determining the parameters needed to measure methylated-adenosine sites, but methylated-cysteine and hydroxymethylcytosine detection needed additional kinetic sensitivity enhancements (Flusberg *et al.*, 2010).

Pacific Biosciences recently revealed read lengths up to 10 000 bp, and promises reads up to 50 000 bp in the near future. The high accuracy of the bases and confidence in detected variants of samples which are sequenced multiple times, are the major advantages of the technology, but the relatively low multiplexing capabillity of 3 000 ZMW wells in the commercial package is a drawback. However, the development system showcased at the 2010 AGBT meeting showed a massively parallel system, with over 80 000 ZMW wells capable of simultaneous sequencing in parallel. At the current sequencing speed of almost two nucleotides per second, this system has the potential to make real-time diagnostic sequencing a reality.

Heliscope Single Molecule Sequencer (Helicos Biosciences)

The Helicos sequencer is a single molecule cyclic array sequencer. It was developed based on the research by Braslavsky *et al.* (2003). The key advantage of this technology over cyclic array sequencers is that there is no amplification step required during the sequencing process, which implies that the each signal detected on the array originates from a single molecule, and not a cluster of amplicons. A highly sensitive fluorescent detection system is used to directly interrogate single DNA molecules *via* sequencing-by-synthesis. Poly-A tailed fragmented DNA template molecules are captured by a

surface-tethered poly-T array, yielding an array of primed, single sequencing templates. Fluorescently labled nucleotides and DNA polymerase are then systematically washed over the array, interspersed by chemical cleavage in order to detect the incorporated base *via* a CCD device.

Read lengths ranging from 35 bp to 70 bp have been reported with the system (Harris *et al.*, 2008; Pushkarev *et al.*, 2009), and read accuracy has been reported to be improved with a two-pass strategy in which the array of single molecules is sequenced, the original strand removed by denaturing, and the remaining strand re-sequenced (Harris *et al.*, 2008). This effectively yields a read in the opposite orientation from the template. This two-pass strategy can reduce the error rate from 2-7% to 0.2-1% (Shendure and Ji, 2008).

Due to the use of single molecules, a much higher density of unique fragments can fit on an array. Although the read length only ranges from 25 to 55 bases, the highly parallel nature of the technology allows it to achieve a throughput of between 21 and 35 Gb per run. The imaging system on the Helicos platform was designed for a theoretical throughput of 1Gb/hour, but this has not been achieved due to the practical constraints introduced by the chemical efficiency of the system. Functional genomic applications of the Helicos system have included the sequencing of a viral genome and BAC library (Harris *et al.*, 2008; Bowers *et al.*, 2009), digital gene expression of poly-A RNA transcripts generated by strand-specific reads (Lipson *et al.*, 2009; Ozsolak *et al.*, 2009) and ChIP-Seq applications (Goren *et al.*, 2010). The comparatively short average read length produced by the system, and the relatively late market introduction of the commercial application seem to be the major drawbacks in widespread adoption of the system.

Ion Torrent

At the 2010 Advances in Genome Biology and Technology Meeting the founder of 454 Life Sciences, Johnathan Rothberg, revealed an innovative approach of sequencing DNA using a semiconductor system to detect the change in pH (due to the release of an hydrogen) when a base gets incorporated during sequencing (http://www.agbt.org, http://www.iontorrent.com). This technology, described as "Post light sequencing with semiconductor chips" lowers the capital investment needed to acquire a

sequencer to below US$50 000, and the consumables for a run down to US$500 per sequencing run (http://www.iontorrent.com). As of the beginning of 2011, no research articles have been produced applying the Ion Torrent system in a research environment, and commercial instances of the sequencer have not been sold. However, this technology promises affordable high-throughput sequencing available without a large capital investment.

1.3. High-throughput DNA sequencing applications in genetics and functional genomics

The technological advances made with uHTS technologies have provided biologists with most of the required tools for a systematic approach to functional genomics. This has led to a gradual shift in focus from studying isolated parts of a system, to analysing DNA, RNA and proteins in context of the whole organism or cell. Genome re-sequencing efforts have led to better understanding and quantification of sequence and structural variation between individuals within species (Fullwood *et al.*, 2009; Pang *et al.*, 2010), and a more detailed blueprint of the genomic data organised in near complete chromosomes for most model organisms. Another consequential development was the understanding that the same physical blueprint, such as the genes embedded in a genome, exhibits massive variation in terms of functional post-transcriptional form and levels of transcript abundance (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008; Pan *et al.*, 2008; Sultan *et al.*, 2008). The study of genotypic variation present in transcription products gains merit when there is an observable effect of a mutation on a phenotype. This, together with the observation that there are distinct differences in the structure and abundance of transcripts in a cell, necessitates the study of transcriptomes not only in an individuals, but in a specific tissue and in many individuals in order to observe transcriptional differences that can be associated with a condition. Both these approaches are relying on the use of uHTS technologies to provide the primary data for genome and transcriptome wide studies.

De novo genome sequencing

Improvements in the chemistry used by sequencing platforms and the development of novel sequencing techniques such as paired-end sequencing have led to a gradual shift in sequencing applications from re-sequencing known genomes (Margulies *et al.*, 2005; Shendure *et al.*, 2005; Velicer *et al.*, 2006; Hofreuter *et al.*, 2006), to *de novo* sequencing and assembly of prokaryotic genomes (Tauch *et al.*, 2008; Reinhardt *et al.*, 2009), small eukaryotic genomes (DiGuistini *et al.*, 2010; Nowrousian *et al.*, 2010) and ultimately large eukaryotic genomes like that of the Giant Panda genome completely assembled from Illumina reads (Li *et al.*, 2010*b*). *De novo* genome sequencing with uHTS technologies has been thought an impossible task due the very short reads generated by these technologies, but mixing reads generated from different technologies which complement each other in terms of the read length, the quality of the bases in the reads, and the sequence throughput from these technologies have led to the development of cost-effective and *de novo* genome sequencing and assembly strategies (Aury *et al.*, 2008; DiGuistini *et al.*, 2010; Nowrousian *et al.*, 2010).

The most robust genome sequencing method is known as BAC-end sequencing. The fundamental apprach to BAC-end sequencing is to perform a shotgun fragmentation of chromosomal DNA, and making use of Bacterial Artifical Clones (BAC) as vectors to sequence around 500 bp of each end of the vector insertion point (Venter and Smith, 1996). BAC-end sequencing has been very successfully applied in large genome sequencing projects, including the human genome project (Venter *et al.*, 2001), and was a key improvement over the generation of overlapping Yeast artificial chromosomes (YACs, Venter and Smith, 1996). Making use of uHTS technologies has enabled the sequencing of the large, complex and highly repetitive genome of barley from BACs (Wicker *et al.*, 2006; Steurnagel *et al.*, 2009). Another sequencing approach in contrast to BAC-end sequencing is the whole genome shotgun sequencing (WGS) of the organism in a single approach using NGS. Uncertainty over the feasibility of using only uHTS technologies to sequence a large genome was laid to rest with the publication of the Giant Panda genome (Li *et al.*, 2010*b*). There are certain tradeoffs between WGS and BAC sequencing, for example the increase in bioinformatics costs to assemble a genome produced from uHTS technologies.

For large complex genomes full of repeat elements such as the cereal genomes, alternative methods to BAC and WGS approaches exist. These methods aim to sequence very specific, pre-selected regions of the genome. Some of these methods include restriction analysis, where genomic DNA is treated with a restriction endonuclease, and then fragmented to remove abundant repeat fractions (Van Tassell *et al.*, 2008). Another approach can be isolating specific chromosomes for sequencing by means of chromosome sorting (Dolezel *et al.*, 2007; Simková *et al.*, 2008*a*,*b*).

The application of uHTS technologies to sequence plant genomes is fast gaining momentum. Since the initial sequencing of the first plant genome, *Arabidopsis* (AGI, The Arabidopsis Genome Initiative, 2000), large genome sequencing projects including rice (Goff *et al.*, 2002; Yu *et al.*, 2002), poplar (Tuskan *et al.*, 2006), maize (Schnable *et al.*, 2009) and soybean (Schmutz *et al.*, 2010) genomes have been completed by using Sanger sequencing. One of the first agriculturally important crops to make use of uHTS technology (454 sequencing) to complete a genome sequence was the consortium to sequence a heterozygous grape variety (Velasco *et al.*, 2007). More examples of completed genome projects making use of a mixture of traditional and high-throughput technologies include the cucumber genome (Huang *et al.*, 2009*a*), BAC sequences of the barley genome (Steurnagel *et al.*, 2009), and a genomic survey of the perennial grass *Miscanthus* (*Miscanthus* x *giganteus*, Swaminathan *et al.* 2010). A recent report on the applications of uHTS technologies in plant genomics revealed that the sequencing of the cocao (*Threbroma cacao*), apple (*Malus domestica*) and strawberry (*Fragaria vesca*) genomes currently underway make use of a mixture of Sanger and uHTS approaches (Imelfort and Edwards, 2009).

Genome re-sequencing and variant discovery

Some of the first applications of uHTS technologies in a genomic context were the re-sequencing of the bacterial genomes of *Mycoplasma genitalium* (Margulies *et al.*, 2005), *Myxococcus xanthus* (Velicer *et al.*, 2006) and *Campylobacter jejuni* (Hofreuter *et al.*, 2006). In these projects, the microbes of interest were a lineage or strain that exhibits a biological phenotype different from the reference genome available for the species. These reference genomes served as template scaffolds onto which the generated sequences

were aligned, in order to detect single nucleotide polymorphism (SNP) and indel variations between the reference genome and the newly re-sequenced genome. The genomic differences were then related to the presence or absence of a biological phenotype, for instance antibiotic resistant genes or pathogenicity islands in the re-sequenced genomes.

Human cancer genomics has made great advances in terms of disease-specific re-sequencing efforts, revealing mutations in somatic tissues that are thought to contribute to tumor progression (Ley *et al.*, 2008; Mardis *et al.*, 2009; Pleasance *et al.*, 2010*a*). Exposure to detrimental environmental agents, such as tobacco smoke, has also lead to genome re-sequencing of tissues under mutational pressure from these exposures, providing insight into the genome-wide carcinogenic effect of these agents (Pleasance *et al.*, 2010*b*). Data from these studies led directly to the design of genome-wide association studies (GWAS), which have the basic aims to identify genetic markers which can be used to predict an individual's risk to disease, and secondly to highlight the molecular processes involved in a disease, with the ultimate aim of identifying potential therapeutic targets. A natural feedback of information is present in determining genetic variation, where polymorphism information produced from genome re-sequencing efforts leads to the design of population-based marker arrays, which in turn prompts investigation in very specific, personal-whole genomes (Mir, 2009). Re-sequencing of genomes of agricultural importance tends to focus on adaptive evolutionary traits and the detection of novel genetic markers, especially where large differences in phenotypes are present in a species. The detection of a selective genomic sweep shared by broiler populations involving metabolic regulation and reproductive genes in modern chickens is an excellent example of identifying the effects of adaptive evolution and selection pressure in populations (Rubin *et al.*, 2010). Variant discovery and domestication studies have also been investigated in the silkworm (Xia *et al.*, 2009; Li *et al.*, 2010*a*), soybean (anchoring markers on the genome by Hyten *et al.*, 2010), and rice (Huang *et al.*, 2009*b*).

In human genetics, the search for disease phenotypes and population genetic markers led to the establishment of the 1 000 Genomes Project (http://www.1000genomes.org). The latest release of the data generated by the 1 000 genomes projects (released on 21 June 2010) included the data from three of

the completed subprojects. This release included the data from nearly 700 human genomes, and aims to produce an extensive catalog of human genetic variation, including SNP and structural variants. The final project will contain data described as "genomes of about 2000 unidentified people.....will be sequenced using next generation sequencing technologies" (http://www.1000genomes.org). This achievement somewhat overshadows the phenomenal achievement of the completion of the first draft human genome in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001), and builds on the example set by the re-sequencing efforts of the human genome by various other research groups (Bentley *et al.*, 2008; Wang *et al.*, 2008; Wheeler *et al.*, 2008; Ahn *et al.*, 2009; Kim *et al.*, 2009; McKernan *et al.*, 2009; Pushkarev *et al.*, 2009; Drmanac *et al.*, 2010; Schuster *et al.*, 2010), it also serves as an excellent showcase of the advances made possible by next generation sequencing during the last decade.

The development of high-throughput genotyping methods make the use of SNPs highly attractive in especially agricultural applications (De la Vega *et al.*, 2005). High-density SNP markers in a genome are ideally suited for the construction of high-resolution genetic maps, the investigation of evolutionary history within a population or species, and the discovery of marker-trait associations to aid marker assisted selection (MAS) in breeding programs. During the discovery of marker-trait associations, a dense set of markers are needed to cover the genome of interest to discover a casual mutation, or a SNP which is in linkage disequilibrium with a casual mutation for the trait of interest (Aranzana *et al.*, 2005). The construction of high-density genetic maps requires the genotyping of a large number of individuals, and platforms with the ability to genotype a large number of samples at a large number of polymorphic sites are desired. Successful applications of high-throughput genotyping experiments include the design of a barley SNP assay using the Illumina GoldenGate™ technology, providing the barley community with a platform to investigate diversity with over 3 000 markers (Close *et al.*, 2009). High-throughput genotyping assays have also been developed for the unsequenced genomes of white and black spruce (*Picea glauca* and *Picea mariana*, Pavy *et al.*, 2008), the complex genome of soybean which contains a high proportion of paralogous genes (Hyten *et al.*, 2008) and the allohexaploid genome of wheat (Akhunov *et al.*, 2009). A future application of uHTS technologies in genotyping, would be designing SNP arrays

for an organism for which a genome is not yet available, but for which gene information derived from technologies such as mRNA-Seq can be useful. A large number of EST sequences from different lines or individuals have already been used for marker identification in maize (Barbazuk *et al.*, 2007) and *Eucalyptus* (Novaes *et al.*, 2008). The authors of the *Eucalyptus* article reported close to 24 000 SNPs, and validated a proportion of the data with a success rate of close to 85%. Two more popular approaches to SNP detection in portions of the genome is to make use of specific fragments produced from selective amplification with restriction enzymes as demonstrated by van Orsouw *et al.* (2007) and the sequencing of restriction-site associated DNA (RAD) tags (Baird *et al.*, 2008).

Genome re-sequencing efforts also provide insight into other genome structural variations, such as indels, copy number variation, inversions and translocations occurring between different genomes. Re-sequencing of two naturally inbred *Arabidopsis* strains has led to the discovery of more than 800 000 SNPs and almost 80 000 indels ranging from 1 to 3 base pairs (Ossowski *et al.*, 2008). Finding longer indels between the genomes was reported as a problematic issue with the short reads (36 bp in length), but the use of paired-end reads as implemented by most current high-throughput technologies has resolved the problem (Ng *et al.*, 2006; Fullwood *et al.*, 2009). Structural variation detection has also been successfuly employed in various human genome re-sequencing projects (McKernan *et al.*, 2009; Kim *et al.*, 2009; Pang *et al.*, 2010).

Transcriptome sequencing

The transcriptome of an organism can be defined as the complete set of mRNA transcripts produced at any time in a cell. The transcriptome is by nature not in a steady state and across cell types, during different conditions in the cell's lifecycle, and in response to external and internal stimuli. The use of expressed sequence tags (ESTs) has become a standard in obtaining information regarding the coding, or expressed regions of an organism for which a sequenced genome is not yet available. Recently, the use of uHTS technologies has been applied to sequencing the RNA landscape of a cell, by making use of a

technology now commonly known as mRNA-Seq (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008; Novaes *et al.*, 2008; Nagalakshmi *et al.*, 2008).

Various hybridisation-based methods have traditionally been used to study the transcriptome landscape, which have lately been complemented by sequence-based methods.. Traditionally, hybridisation-based methods involved labelling cDNA with a fluorescent dye, and then hybridysing the cDNA to a set of probes on a microarray. Specialised array chips, such as exon-arrays have been designed specifically to identify spliced isoforms (Clark *et al.*, 2002; Frey *et al.*, 2005; Singer *et al.*, 2006; Kapur *et al.*, 2007), while genomic tiling arrays have been used to identify novel transcripts of already sequenced organisms (Bertone *et al.*, 2004; Cheng *et al.*, 2005; David *et al.*, 2006). The development of parallelised sequencing technologies have increased the use of sequence-based approaches to gene expression profiling and the genome-wide evaluation of chromatin immunoprecipitation (ChIP-seq) experiments. Some of the limitations of hybridisation-based methods include the dependency on knowledge of the sequence of the studied genome in order to manufacture the probes, the occurance of inter-probe cross-hybridisation on the arrays, the presence of background noise and signal saturation, and some data-analysis issues in terms of normalisation of data between experiments (Eklund *et al.*, 2006; Okoniewski and Miller, 2006; Casneuf *et al.*, 2007; Royce *et al.*, 2007).

The development of tag-based sequencing methods which include cap analysis of gene expression (CAGE, Kodzius *et al.*, 2006), serial analysis of gene expression (SAGE, Velculescu *et al.*, 1995) and massively parallel signature sequencing (MPSS, Brenner *et al.*, 2000) allowed for the quantification of the amount of cDNA present in a biological sample. The advantages of these methods were that a unique hybridisation probe was not needed to detect each transcript and, in the case of SAGE analysis, multiple SAGE tags could be sequenced together providing several measurements simultaneously (Bertone *et al.*, 2005). The initial widespread adoption of these methods was hampered by the high cost of Sanger sequencing technology (Sanger *et al.*, 1977) used to determine the base pair composition of the sequence, and the technical problem that the very short tags (10-14 bp tags for SAGE analysis) generated by these technologies did not map uniquely to the reference genome (Bertone *et al.*, 2005; Wang *et al.*, 2009),

which made it very difficult to to distinguish transcript isoforms from each other. An improvement in read length (21 bp Long-SAGE, Saha *et al.*, 2002) overcame some of these limitations, but the use of SAGE was prohibitively expensive until the power of HTS technologies was employed (Deep-SAGE, Nielsen *et al.*, 2006).

The development of a technology to sequence the transcriptome content of a biological sample has been achieved by the major high-throughput sequence technology companies (see Section 1.2 for a overview of the technologies). The premise of these technologies is the fragmentation of a population of RNA (total RNA, polyA-selected RNA), which is converted to a library of cDNA fragments with adapters attached to one or both ends. Each RNA molecule can then be sequenced in a high-throughput manner from one (single end sequencing) or both ends, resulting in reads that can vary from 35-450 bp in length depending on the technology used. Prior to sequencing, the RNA or cDNA molecules can be amplified, but sequencing of RNA without amplification has the added advantage of providing expression information in addition to the transcript sequences (Wilhelm and Landry, 2009). RNA-Seq technology is slowly reaching maturity, and it offers some key advantages over hybridisation-based technologies, with longer sequences than tag-based technologies, and a lower cost per base pair than traditional EST sequencing technologies. It has also been shown that RNA-Seq detects differential gene expression with greater sensitivity than expression (Li *et al.*, 2008*a*; Marioni *et al.*, 2008) and tiling microarrays (Hiller *et al.*, 2009).

Findings obtained with genome-wide analysis of transcribed sequences and potential transcriptional start sites indicated that the traditional genome-centric view of the protein coding regions of the genome needed to be replaced by a more complex transcript-centric view (Bertone *et al.*, 2004; Johnson *et al.*, 2005; Carninci *et al.*, 2006). These findings brought the idea that there is a defined set of isolated loci transcribed independently into doubt, and indicated that numerous overlapping coding and non-coding transcripts span the entire genome, and that those transcripts are of biological importance in the cell system, which in turn led to a renewed research interest in transcription and transcription-related products in a cell. Recently, with the use of RNA-Seq to determine the proportion of the genome which is

transcribed, evidence suggests that the initial estimation of transcription might have been excessively overestimated (van Bakel *et al.*, 2010). The earlier studies were based on tiling microarray data, and the recent studies indicated that the microarray platform is susceptible to a high rate of false positives (van Bakel *et al.*, 2010). In the recent study, most of the transcripts not mapping to exonic regions, mapped to introns, raising the posibillity that these RNA-Seq fragments belong to pre-mRNAs (van Bakel *et al.*, 2010). This study indicated that most of the genome is not appreciably transcribed in levels associated with gene expression, but still leaves the question of what the function of low-level transcribed genomic regions are.

One of the initial applications of mRNA-seq derived data was the discovery of novel transcripts, with the simultaneous estimation of transcript abundance (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008). Cloonan *et al.* (2008) sequenced poly-A captured RNA transcripts from two different mouse tissues, and demonstrated that alternative splice forms from transcriptionally active tissues were readily detectable with mRNA-seq. The sequencing approach they followed (not normalising the sequence libraries) led to the elucidation of transcript expression values, an approach initially proposed by Mortazavi *et al.* (2008) for mouse transcripts. Mortazavi *et al.* (2008) developed a measure of gene expression, measured in reads per kilobase of exon per million mapped sequence reads (RPKM), which is a normalised measure of exonic read density. The use of RPKM values was widely adopted, and various software packages utillise this measure to report gene expression. Furthermore, Cloonan *et al.* (2008) demonstrated that the *de novo* detection of gene models is possible with high levels of expression and alluded that allele specific expression detection is a near-certain posibillity in transcript expression studies. In order to perform *de novo* prediction of gene models from a genome using mRNA-Seq, Denoeud *et al.* (2008) developed a software package `G-Mo.R-Se`, and applied it to the recently sequenced *Vitis vinifera* genome. The authors used mRNA-Seq (175 million Illumina reads) from four different tissues and identified new exons in known loci and alternative splice forms, as well as entirely new loci in the *Vitis* genome.

Data obtained from mRNA-Seq experiments led to investigations into the alternative splice complex-

ity of genes active in different tissues. Previous methods using microarray profiling and cDNA sequencing lacked the sensitivity or confidence due insufficient coverage needed to validate multiple splice events. In the human genome, alternatively spliced transcripts were estimated to occur in two thirds of the genes, but studies using mRNA-Seq estimated that 95% of multi-exon human genes in major human tissues showed evidence of alternative splicing (Pan *et al.*, 2008). Similar results were obtained in human embryonic kidney and B cell line tissues, where an average of 7.2 splice junctions per gene was identified, but employing a very lenient measure of one matched sequence to validate a synthetic splice junction (Sultan *et al.*, 2008). In *Arabidopsis,* the percentage of alternatively spliced genes was estimated at 42% for multi-exon genes (Filichkin *et al.*, 2010), which also surpasses the previous estimates of between 22% and 33% (Campbell *et al.*, 2006; Wang and Brendel, 2006; Chen *et al.*, 2007; Barbazuk *et al.*, 2008). Intron retention was the most prevalent form of alternative splicing in *Arabidopsis,* and was frequently associated with specific abiotic stresses of the plants, which led the authors to postulate the existance of a functional transcript regulation mechanism similar to the regulated unproductive splicing and translation (RUST) mechanism in animals (Lewis *et al.*, 2003; Filichkin *et al.*, 2010). These discussions regarding different splice forms being actively transcribed in a cell under certain conditions raised the question regarding in what quantities these splice forms are distributed across tissue types. In previous studies to quantify transcript expression from mRNA-Seq data, reads were not allocated to specific isoforms, but this feature was implemented in the `Cufflinks` software package (Trapnell *et al.*, 2010). The authors of `Cufflinks` detected 330 genes present in mouse myoblast tissue, which switched their dominant transcription start site or splice isoform during a time-series experiment. `Cufflinks` also no longer relies on any *a priori* information regarding the gene models of an organism, and is able to infer the gene models directly from the combination of mRNA-Seq data and a genome.

Antisense transcription has been shown to play an important regulatory role in the eukaryotic genome. A simple modification to the RNA-Seq method enabled the method to yield strand-specific transcripts (ssRNA-seq, Parkhomchuk *et al.*, 2009; Perkins *et al.*, 2009). The method incorporated a deoxyUTP during the second strand cDNA synthesis, followed by the destruction of the uridine-containing strand in

the sequencing library, thus allowing the polarity of the transcripts to be known. The method was applied to the yeast and mouse model organism datasets, yielding new information regarding promotor-associated and antisense transcription (Parkhomchuk *et al.*, 2009). Another genome-wide investigation of the transcriptional landscape using ssRNA-seq revealed the presence of subtle regulatory RNA and small RNA sequences in the genome of the bacterial pathogen *Salmonella enterica* serovar Typhi (Perkins *et al.*, 2009). The mapping of strand-specific reads to the *S. enterica* Typhi genome provided a single base pair resolution map of active transcriptional elements, resolving overlapping annotated transcripts previously made. The utilisation of ssRNA-seq data derived from large eukaryotic genomes will shed light on the content of the pervasively transcribed transcriptome in future studies.

The combination of high-density genome-wide genetic markers with expression profiling data to identify trait-associated gene expression patterns, or expression Quantitative Trait Loci (eQTL) in mapping populations is fast becomming a reality with the use of HTS technologies. Using data from 60 human Caucasian participants in the HapMap project, Montgomery *et al.* (2010) investigated the occurence of detectable eQTLs from genome-wide collections of SNPs. The authors were also able to detect allele-specific expression from the same expression dataset, which would certainly form the basis for expression studies in hybrid mapping populations. According to the authors, a dataset of 10 million mappable fragments are required in order to quantify alternative and highly abundant transcripts (Montgomery *et al.*, 2010). A similar study of 69 lymphoblastoid cell lines derived from Nigerian HapMap participants identified over a thousand genes where genetic variation contributes to variation in expression and splicing (Pickrell *et al.*, 2010). Results from these studies confirm the observation that most eQTLs are located close to the gene's transcriptional start site, and that most eQTLs influence expression in a *cis* fashion (as oppose to *trans*-regulated expression). In addition to the ability to quantify the expression of different transcript isoforms, these studies also improved the annotation of the genome by detecting previously unannotated exons (Pickrell *et al.*, 2010).

mRNA-Seq has been shown to produce accurate measurements of the expression landscape of the genome with unprecedented accuracy. Data derived from mRNA-Seq experiments has been used to

detect the expression of known and previously unknown transcripts, to assemble transcriptomes from organisms with no genomic information, to detect allele specific expression patterns, and identify novel splice forms. Bioinformatics algorithms and data management approaches to handle these datasets are evolving at a rapid pace in order to to handle mRNA-Seq data, and it is not uncommon for a software package to undergo several version updates in a short period of time as the nuances of these datasets are better understood. The computational needs of processing mRNA-Seq, or any uHTS dataset for that matter, varies according to the intended applications, from a large number of CPUs needed in loosely-coupled homology searches of tens of thousands of genes against public datasets in parallel, to the massive memory requirements of *de novo* assemblers, and must be considered when a high-throughput experiment is planned.

1.4. Core analyses associated with ultra-high-throughput Illumina sequence mRNA-Seq data

One of the strengths of ultra-high-throughput sequencing platforms is in the various practical applications it has in genetic and genomic studies. For each of these applications, there exists a core set of data analysis methods performed with the data in order to address the underlying biological questions. The core data analysis tools range from estimating the quality of the bases received from sequencing facilities, assembling of reads into larger contigs (transcriptomes or genomes), and mapping of reads to a target sequence in order to detect structural variation, evaluate transcript expression, and perform SNP mining or structural variation detection.

Determining the quality of Illumina mRNA-Seq data

`Illumina` results are generally presented to researchers in the `FASTQ` format, The preprocessing of the images is performed by the sequencing facility, since it uses the proprietry `Illumina Pipeline` to perform the base-calling from the image sources. The output from the `Illumina Pipeline`, or to be more specific, the `BUSTARD` tool, is a `FASTQ` formatted quality `FASTA` file (Figure 1.1). The `FASTQ` quality

values differ from the standardised `Phred` quality values prepared by Sanger-based sequencing machines and software pipelines, and also differs depending on the version of `Illumina Pipeline` that was used to perform the base calling. `Phred`-based quality scores are calculated by $Q_{Phred} = -10log_{10}(\frac{1}{\$error\_prob})$, where *$error_prob* is the probability of the base call being wrong (Ewing *et al.*, 1998; Ewing and Green, 1998). In order to present the score of a base in a single character, the Q_{Phred} score is converted to a corresponding American Standard Code for Information Interchange (ASCII) character. ASCII is an 8-bit character set defining alphanumeric characters widely used in the computer industry. Since ASCII 32 is the whitespace (spacebar) character, `Phred` scores use ASCII characters 32-126 to represent qualities from 0-93. The dynamic range of a `Phred` score ranges from 1.0 (a completely wrong base), through to $10^{-9.3}$, an extremely accurate base (Cock *et al.*, 2010). This is also known as the `fastq-sanger` format.

The Illumina `FASTQ` format encode base qualities in two different scoring systems. `Illumina Pipeline` ($<$ version 1.3) defined a new scoring formula to determine the quality score: $Q_{Solexa} = -log_{10}(\frac{\$error\_prob}{1-\$error\_prob})$. The after-effect of this non-standard scoring formula resulted in a change of the ASCII-offset used to represent a base score. Since the Q_{Solexa} score's lower limit is -5, assuming a random read error probability of 0.75, a very low quality base will result in a whitespace character representing the quality score (this occurs because ASCII characters 0-32 are all whitespace characters). Due to the fact that whitespace characters can be interpreted differently by some computer operating systems, which should be avoided in setting a standard where the quality values are aimed to be represented in a single line of a text file (for example, the newline character is also a whitespace character), the ASCII offset of 64 was chosen. This resulted that ASCII 59-126 was used, providing the Q_{Solexa} score a dynamic range from -5 through to 62 inclusive (this format is generally known as the `fastq-solexa` format). After version 1.3 of the `Illumina Pipeline`, the scoring function changed to be compatable with the `Phred` standard, but the ASCII offset of +64 remained, and the format is now known as the `fastq-illumina` format (Illumina, 2008). For a review of the complete history of the `FASTQ` format, and also the introduction of the `ABI Solid CSFASTQ` format (in color-space, not base or sequence space), please see Cock *et al.* (2010). The discussion above was required to introduce the concept of format conversions of raw Illumina

A)

```
@HWI-EAS121  0005  FC61APKAAXX:1:1:2358:16565#0/1
GTAGTAACTTGNCATTTGCTAGTGTGCTTGTGACATGTAGTTTTAGGTCATTTATTNATCTTTTACTCTCAGGAATTCAG
+HWI-EAS121  0005  FC61APKAAXX:1:1:2358:16565#0/1
cdddddeeeeebKbbbcccccdeeeeeeeeedeeeeeeeeeeeeeeeddda^bbeeeeee`D]``bb^````aacdaTdaedec`

@HWI-EAS121  0005  FC61APKAAXX:1:1:2358:4461#0/1
TTTTGATGTTGNCAGGATTACAAGGAACAGCCATTTTCTCTAGTGTGTTACTAGGGNGAGCAATACAGGAATTAAATGGC
+HWI-EAS121  0005  FC61APKAAXX:1:1:2358:4461#0/1
YRa\\T\\a]]FVXZURVRVRZQZX]__bU_VUU]a]Va\X[[Y[QZOZZ]RT[SVDZZR`Z`ZTGa]T\[KK`KaBBBB

@HWI-EAS121  0005  FC61APKAAXX:1:1:2358:19891#0/1
CAAGCGCAGGANGCCATGTGGACAATCAAGTCAACAACACGGGAACTGTAGCCCCANTCATTGTCGTACCATGAGACCAG
+HWI-EAS121  0005  FC61APKAAXX:1:1:2358:19891#0/1
dddTddddddbKbbb]b`^`ddddaffffeeeeeeffffacdc^bbdad]b_`GWIYYX\[VXab[__aaa_a_^_Z^B

@HWI-EAS121  0005  FC61APKAAXX:1:1:2358:1852#0/1
GCAATACATGCNGTTACAAATACTTGATTGGAATGCATTCATTGTGCACGTGGGTANACTGCGGTGTGGGGAATCAGCCT
+HWI-EAS121  0005  FC61APKAAXX:1:1:2358:1852#0/1
dddadeeeeebKbbb__ba  fffff^ffcceffecfdaffffffLdfffffbddY`\XHW[VZYUYRa^Hab^a^^^acca\

@HWI-EAS121  0005  FC61APKAAXX:1:1:2358:7138#0/1
CTGGTGTGCTTNCAAATGCTCCTTTCATGCTGAACTTGGATTGTGACCACTACATANATAACAGCAAGGCCGTCCGCGAG
+HWI-EAS121  0005  FC61APKAAXX:1:1:2358:7138#0/1
dddaadbdddblbbbaccccfffffffeffffefefefeedffdbddfdfffffdbaGaaa``]]ad_`bcbbbcbcdba^
```

B)

@HWI-EAS121 0005 FC61APKAAXX:	Instrument name
1:	Flowcell lane
1:	Tile number within flowcell
2358:	X-coordinate of cluster
7138:	Y-coordinate of cluster
#0:	Index number (if multiplexed)
/1:	Member of paired-read (1 or 2)

Figure 1.1: An example of an Illumina FASTQ formatted mRNA-Seq file. The example presented above represents five 80 bp reads and the quality values associated with the reads (a). The sequence and quality header lines are denoted by the @ and + symbols, while the line following the header line represent the bases and the qualities associated with the specific base pair. Note that the whitespace lines in between the reads were inserted to improve readability of the format. The header file contains the following information separated by colons; the unique instrument name, the flowcell lane, the tile number within the flowcell, the 'x' coordinate and 'y' coordinate of the cluster within the tile, the index number for a multiplexed sample and if paired, the first or second member of a pair (b).

data. Some assembly tools perform the conversion between the Illumina formats (both `fastq-illumina` and `fastq-solexa`) to the traditional `fastq-sanger` formats if the input type is specified at run time. There are also standalone conversion tools available to translate between the different formats for use in analysis tools that do not provide the conversion ability.

De Bruijn graph-based genome and transcriptome assembly

The short reads produced by uHTS technologies are not suited to be assembled by the same sequence assemblers as traditional Sanger sequencing reads. With longer Sanger reads, the assembly process relied on the overlapping of reads which fit together to generate a consensus sequence, or contig. Very short reads are not suited for the traditional overlap-layout-consensus based method of assembly (Zerbino and Birney, 2008). Because of the large numbers of reads that are produced, short reads have a much higher coverage over a specific region. An overlap-based method, where the actual reads are stored to generate a consensus sequence, has computational limitations when handling billions of reads where large numbers of reads have an overlap of all but one base pair. With overlap-based methods, each read forms a node of a graph, and the nodes are connected by an overlay metric between the nodes (Batzoglou, 2005).

A fundamental shift in the methodology behind aligning short reads was introduced in 2001, with the adaptation of *de Bruijn* graphs to represent and organise the relation between reads using an Eulerian path approach to assemble sequence reads (Pevzner *et al.*, 2001). In essence, *de Bruijn* graphs do not represent whole reads as nodes in a graph, but rather break the reads into words of a pre-defined length (length k, henceforth known as kmer(s)), and the reads are then organised in paths through the graph in a determined order. By using kmers rather than reads, the redundancy of the graph is inherently handled by the structure of the graph, without increasing the number of nodes in the graph. Every node in the graph thus represents a single k-mer (non-redundant), and have explicit links to the neighbors, or start and end positions of the kmer in a read (Pevzner *et al.*, 2001). Various research groups have since investigated the use of *de Bruijn* graphs in short read assembly software programs (Shah *et al.*, 2004; Bokhari and Sauer, 2005; Myers, 2005; Jiang *et al.*, 2007; Zerbino and Birney, 2008).

The `Velvet` program was one of the first *de novo* short read assemblers implementing the *de Bruijn*

graph assembly strategy. While transcriptome-specific assemblers were developed towards the end of this study, during the initial phases of this project `Velvet` was the only assembler found to produce cDNA contigs of reasonable length and quantity. Analysis with `Velvet` consists of two phases, first the indexing of the input reads with the desired kmer, and secondly the traversing and tracking of the kmers to construct the contigs. `Velvet` relies on coverage per kmer to eliminate erronous nodes, resolve repeated kmers and find the path between the nodes which is most represented by coverage and constructs the output sequence (Zerbino and Birney, 2008). `Velvet` is an example of a memory hungry application, with massive memory requirements needed to store and traverse the kmer graphs. A recent experiment of a single lane of 76 bp paired sequence (≈40 million reads), consumed close to 45 GB of RAM during assembly with a kmer of 41 bp. The developers of the `Velvet` package are continuously improving the memory footprint of the algorithms used.

Alternative assemblers which utilise the *de Bruijn* graph assembly approach include but is not limited to the `ABySS` (Simpson *et al.*, 2009) and `OASES` (Zerbino *et al.*, unpublished) assemblers. `ABySS` was used to succesfully assemble the human transcriptome of a patient with follicular lymphona (Birol *et al.*, 2009). Using `ABySS`, the authors assembled ≈65 000 contigs representing close to 30 Mb of the human transcriptome. The `OASES` assembler was developed as an extension to the `Velvet` assembler with the purpose of focusing on splice variant assembly of transcripts. The source code of the project was made public early in 2010, and at the time of writing no peer reviewed publications had been published using the appllication. These applications are viable alternatives for transcriptome assembly projects.

Mapping mRNA-Seq reads to a reference dataset

The requirements of a short read mapper can be seperated into a strategic requirement in terms of alignment accuracy, and a more practical requirement in terms of a time constraint (Trapnell and Salzberg, 2009). Firstly, the use of high-throughput sequence technologies for variant discovery in whole genomes requires the accurate, high confidence alignment of the short read to the target genome. In this application, the presence of repeat regions in the genome, as well as natural variation that occurs between the reference genome and the re-sequenced genome needs to be accounted for, and the short read

mapper needs to be robust enough to handle these issues confidently. Traditional alignment programs, such as `BLAST` (Altschul *et al.*, 1990) and `BLAT` (Kent, 2002) are also able to align short sequences to a target genome, but the algorithms used in these aligners are not optimised for very short reads (35-76 bp), and the time required by these aligners to perform billions of alignments hampers these programs from being serious contenders for high-throughput alignments.

RNA-derived reads can be mapped to a target sequence with different objectives; firstly, a fully sequenced, annotated genome where gene models are already predicted, and the mapped reads are used to calculate gene expression values; secondly an un-annotated or newly sequenced genome to detect gene models or infer new genes; or thirdly, a set of genes or coding regions from a unknown genome (typically the results from a *de novo* transcriptome assembly project). Several short read mapping software packages are available, some of the first mappers include `ZOOM!` (Lin *et al.*, 2008), `MAQ` (Li *et al.*, 2008*b*), `Mosaik` (Stromberg and Marth, 2008), `SOAP` (Li *et al.*, 2008*d*), `SHRiMP` (Rumble *et al.*, 2009) and `Bowtie` (Langmead *et al.*, 2009), with more recent updates to the algorithms implemented in `SOAP2` (Li *et al.*, 2009*b*) and the successor to `Bowtie`, `BWA` (Li and Durbin, 2009, Table 1.1). These short read mappers typically works by selecting a defined wordsize usually from the beginning of the short read, and then requiring some number of these words to fit perfectly to the target to find a match, while mismatches are allowed to occur within the rest of the words (Li *et al.*, 2008*d*; Langmead *et al.*, 2009; Li *et al.*, 2009*b*; Li and Durbin, 2009). Another common approach is to create a subsequence, or a spaced seed, along the high quality 5' end of the short read sequence, and again with some mismatch threshold allowed, the seeds are aligned to the target (Lin *et al.*, 2008; Li *et al.*, 2009*b*; Rumble *et al.*, 2009). The next section describes in detail the difference in these two approaches, as implemented by the `Bowtie` and `MAQ` aligners.

Mapping reads with the spaced seed approach

`MAQ` employs a spaced seed indexing strategy in order to align segments of a short read to a genome. A short read is effectively divided into four sets of words of equal length, called a spaced seed. By default, `MAQ` uses the first 28 bp of a short read for seed generation, and uses a word size of six to

Table 1.1: A selected list of short read sequence alignment tools currently available for academic use. These software tools perform essentialy the same function in aligning reads generated from uHTS technologies to a target genome, but implementing different mathematical, statistical and programmatic approaches to achieve this goal.

Program name	Description	Reference
BFAST	BLAT-like Fast Accurate Search Tool for aligning re-sequence data to a genome. The program returns an accurate alignment for a candidate alignment location where the short read corresponds to the genome. It also includes support for two-base encoding sequences from the SOLiD platform.	Homer *et al.* (2009*a*,*b*)
Bowtie	A very efficient short read aligner implementing the Burrows-Wheeler transform in order to be memory efficient. Bowtie can align up to 25 million 35 bp reads per CPU hour.	Langmead *et al.* (2009)
BWA	An update of the MAQ package, based on a backward search with Burrows-Wheeler transform, effectively elimitating the alignment of repeated short reads.	Li and Durbin (2009)
ERANGE	Mapping mRNA-Seq data to genomes for quantification of transcript expression. Makes use of the Bowtie aligner.	Mortazavi *et al.* (2008)
Genome Mapper	Simultaneously aligning reads to multiple genomes by collapsing the corresponding regions of the genomes into a single graph structure. Used by the 1001 genomes project (http://1001genomes.org) consortium.	Schneeberger *et al.* (2009)
RMAP	Used base quality scores in deciding the appropriate map position of a read on a reference sequence.	Smith *et al.* (2008)
Slider and SliderII	Specifically developed for the Illumina platform, and uses the probability files instead of the sequence files in order to perform the alignment to the reference sequence.	Malhis *et al.* (2009)
SOAP and SOAP2	Introduced gapped and ungapped alignments, and the use of a paired- end module. SOAP2 update of SOAP, implementing a Burrows-Wheeler transform algorithm.	Li *et al.* (2008*d*, 2009*b*)
TopHat	Uses BWA to perform multiple alignments to a genome with mRNA-Seq data in order to detect splice junctions.	Trapnell *et al.* (2009)
MAQ	One of the first short read aligners to implement mapping quality to the target genome. Not as computationally efficient as some of the other programs.	Li *et al.* (2008*c*)
Mosaik	Produces gapped alignments using the Smith-Waterman alignment algorithm, and forms part of a software suite which includes SNP calling.	Stromberg and Marth (2008)

generate the spaced seeds. If a perfect match between the read and the target sequence exists, then all of the spaced seeds will match the target. If, however, a mismatch is present in the target sequence, then one or possibly more of the spaced seeds will not match perfectly. When two mismatches are present between the short read and the target sequence, at most two of the spaced seeds will not have a perfect match (only one space seed will show a mismatch if the mismatches are close to each other, and do not span a space seed boundary). By aligning pairs of spaced seeds (there are six possible pairs for the 4 seeds) to the target, it is possible to identify the possible locations on the entire target sequence where the complete short read will match, allowing for at most two seed mismatches. The resulting list of candidate positions are then compared to the complete read extending from position 28 onwards without gaps to identify the correct mapping position. The sum of the qualities of the mismatched bases are then calculated and stored together with a random number and the hit positions in an index. When two short read sequences are mapped with the same mismatch quality scores, the one with the smallest random number is selected as the best possible alignment. `MAQ` can be configured to use up to 20 spaced seeds, and is then able to find all 28 bp seeds with up to 3 bp mismatches, although this means a mismatch ratio of more than 10% between the seed and the target sequence.

Mapping reads with the Burrows-Wheeler transform approach

The Burrows-Wheeler transform (BWT) is a much more complicated method, but has the advantage of running substantially faster (up to 35x when compared to `MAQ`) than an index-based method, and with a smaller memory footprint (Langmead *et al.*, 2009). Originally developed for lossless file compression (Burrows and Wheeler, 1994), the transform involves building an extremely efficient transformation of the target sequence, and then mapping a short read one base at a time to the BWT target. This is achieved by combining the BWT with some opportunistic data structures and the building of a reverse index to minimize backtracking, to allow for an efficient search space (Ferragina and Manzini, 2000, 2001). Each new successively aligned character allows the algorithm to narrow down the possible location where a short read might match perfectly. It has been shown that the original implementation of `MAQ` and `SOAP` would take 35x and 300x longer than the corresponding `Bowtie` alignment (Langmead *et al.*, 2009).

Since the original publications of `MAQ` (development discontinued and replaced by `BWA`, Li and Durbin, 2009) and `SOAP` (updated as `SOAP2`, Li *et al.* 2009*b*), both of these these programs have been updated to utilise the BWT algorithm for building a transformed target sequence. The much smaller memory footprint (1.3 GB for the entire human genome), and the general 30x speedup of the BWT algorithm has made this approach currently the most widely used tool for mapping short reads to a target sequence.

Mapping high-throughput genomic reads to a genome

High-throughput DNA sequencing is ideally suited for genome re-sequencing projects where variant discovery is the main focus (see section 1.3 for a review of re-sequencing projects). The fraction of short reads which map to the reference genome depends on several factors. If there is a minimal amount of variation between the reference and the re-sequenced genome, the alignment algorithms improved are capable to align from around 70-75% of single end reads to the reference genome, up to 85% with the `BWA` aligner, and up to 98% with paired-end reads (Langmead *et al.*, 2009; Li and Durbin, 2009). The quality of the sequencing library, the amount of repeat regions in the reference genome, the length of the reads and the insert size in the case of paired-end reads all influence the mappability of a short read. Paired-end reads improves the mappability of a sequenced fragment by having two reads with a known destance associated with the fragment. Paired-reads are specifically useful for improving fragment mappability in cases where one of the reads aligns to a repeat region in the genome sequence. It has been calculated that with 35 bp reads, the fraction of the human genome that is re-sequenceable is 85%, and with paired-end reads with an insert of 170 bp, this fraction increases to 93% (Li *et al.*, 2008*b*). Any additional increase in short read mappability could only be obtained with an increase in read length and having datasets of varying insert sizes available.

Mapping mRNA-Seq reads to a genome

RNA-derived reads, such as those produced by mRNA-Seq, strand specific RNA-Seq and total-RNA-Seq protocols provided by Illumina require gapped alignments across gene splice juctions in order to map sequenced reads to eukaryotic genomes. The computational approach to map reads to exon-exon bound-

aries is different to genome derived short read mapping due to the possibility of a single read spanning across two exons that were joined during transcript processing. The first approach to solve this problem was to utilise the structure of known genes in determining the intron-exon boundaries of a gene, such as implemented in the `ERANGE` package (Mortazavi *et al.*, 2008). Another approach is to extract possible junction sequences from the aligned genomic sequence with some form of machine learning algorithm, for example a logistic regression classifier (Pan *et al.*, 2008) and a support vector machine-like approach (Schulze *et al.*, 2007; De Bona *et al.*, 2008). Unfortunately these methods only work for organisms for which gene models are available, as the gene models serve as a required input to delineate the intron-exon boundaries together with training data sets.

Because of the reliance on known gene models to map the RNA-Seq reads to fully sequenced genomes as mentioned before, these methods are limited in detecting novel splice junctions. Another approach to splice juction mapping was proposed and implemented by the two software packages `TopHat` (Trapnell *et al.*, 2009) and `G-Mo.R-Se` (Denoeud *et al.*, 2008). These packages utilise the power of a BWT mapping tool (initially only `Bowtie`, but `Bowtie` and `BWA` are now supported) to detect possible exons, and then by joining the exons which share transcripts, remap the data in order to detect possible splice junctions. Of the two packages, `TopHat` package is currently being actively maintained.

1.5. High-throughput DNA sequencing data management

Recent calculations from the Ontario Institute for Cancer Research indicated that since the advent of uHTS, the cost of sequencing a base has been dropping faster than the cost associated with storing a byte of data on a computational storage meduim (Stein, 2010). The author investigated the historical trends in data storage prices *vs.* DNA sequencing costs, and found that the doubling time in sequenced base pair per dollar was less than six months, exceeding the drop in disk storage cost on a logarithmic scale. One of the fundamental problems in terms of sequence storage, is that a single base has multiple bytes associated with it. During a uHTS run where the bases incorporated during the sequencing process is captured by a CCD, the image needs to be converted from an image to a string representation, usually in

basespace, but colorspace is also gaining prevalence in order to prepare the data for input into a variety of analysis programs. A quality score is usually associated with the each base call, effectively doubling the storage space needed for a base. Format incompatablilities, such as the case of the `FASTQ` format (Section 1.4 on page 25) can require various duplicate versions of the same data to be stored as input files. Different analysis tools produce various output files, which can be thought of as different representations of a base, highlighting different features of the base, or the surrounding bases in terms of biological relevance. The problem in terms of storage cost and expansion capabilities is thus compounded by the already exponential growth of uHTS base throughput, and the non-linear relationship between a base of sequence and the space required to store the biologial relevance of that base.

The nature of uHTS data requires a disciplined and structural approach to data management. The different file formats required by software packages require that the data be duplicated between analysis steps, increasing the data storage and computational cost associated with uHTS analysis. Tools developed for uHTS analysis are being made available to the community at a rapid pace, and an analysis environment where these tools can be distributed to various users for immediate use and implementation in data analysis workflows is essential.

1.5.1. Widely-used bioinformatics workflow systems

During the last decade, many bioinformatics research groups have dedicated resources to develop mature automated and semi-automated analysis environments. The implementations of these systems are as varied as the number of programming languages used to develop the system, and include executing complex analysis on local resources (`Ergratis` Orvis *et al.*, 2010; `Kepler` Ludäscher *et al.*, 2005; `Galaxy` Goecks *et al.*, 2010), on remote systems through web-services access (`Taverna`, Oinn *et al.*, 2004), or making use of distributed grid systems (`Taverna, Galaxy`). To evaluate different workflow systems, one needs to critically evaluate the the relative strengths and weaknesses of these cyberinfrastructure implementations.

Using dedicated, local resources for high-throughput data analysis has the the advantage of having

complete control over the number of CPU cyles dedicated to a project. The downside of local resources is firstly the cost of the resource, the cost of installing and maintaining a diverse set of analysis tools and systems on the servers, and the investment in human capacity to fully utilise and maintain the hardware components.

Web-services, grid and cloud computing offer attractive alternatives to overcome the initial capital investment in hardware (Stein, 2010). One of the fundamental requirements of utilising a remote resource for computing, is the access to fast and cheap network bandwith to the remote server for data transfer, but this requirement often precludes the use of remote services from some institutions or research groups. Access to these remote computing sites is also limited to the availability of CPUs at the remote sites at any given time.

Taverna

`Taverna` (Oinn *et al.*, 2004) was developed as part of the $^{my}Grid$ initiative for the composition and execution of workflows in the life sciences domain. `Taverna` relies on the Simplified conceptual workflow language (`Scufl`) to represent each step of a workflow as a single task. A graphical user interface (GUI) was developed and packaged as part of `Taverna` which acts as a container in which `Scufl`-based workflows can be constructed, without the need to learn the `Scufl` language. The workflows in `Taverna` rely on the availabillity of programmatic access to bioinformatics repositories, such as `GenBank`, and analysis tools, such as the `EMBOSS` suite of tools at the European Bioinformatics Institute (EBI), `SOAPlab` (Senger *et al.*, 2003) and `BioMOBY` (Wilkinson and Links, 2002). Access to the tool or repository is granted through a web-service interface (Stein, 2002), which allows the consumer (the `Taverna` client) to query a database or start an analysis tool on the host server remotely. The advantages of this type of architecture is that data stored in large datacenters, such as the EBI, NCBI and DDBJ, are accessible to users accross the world through a simple, standardised interface. Centers with access to large computational resources can also expose analysis web-services to the community, and therefore allow smaller research groups with limited resources to execute jobs with large computational requirements remotely. This service-oriented

design of `Taverna` also allows it to connect to services that can submit jobs on a grid-like environment for distributed computing.

Taverna has been successfully employed by many research groups, the biggest and most prominent is the integration of `Taverna` into the cancer Biomedical Informatics Grid (caBIG) project, where `Taverna` and the Web-service-Business Process Execution Language (WS-BPEL) are used in a service-oriented data analysis environment (Tan *et al.*, 2008, 2009; Missier *et al.*, 2009). As explained above, the service-oriented nature of `Taverna` relies on the abillity to connect to a host server to interact with the data, but when the data is not mirrored on the host server, the data needs to be transfered to the compute elements. This requires that either a reliable, fast and inexpensive network connection is needed to connect to the remote services, or a duplication of the services needs to be present on a local network where the data is already present.

The nature of uHTS data in general does not lend it to be readily distributed to various computing locations. In most cases, the prohibitive factor is the cost and time needed to duplicate multi-GB datasets accross many locations in order to perform analysis in parallel. Although the South African Research Network (SANREN, http://meraka.org.za/sanren.htm) has made great progress in terms of providing a fast and reliable cyberinfrastructure between South African research institutes and the rest of the world, the availability of reliable bandwith at a high enough data throughput is still a major hurdle to overcome.

Kepler

The `Kepler`-project (Ludäscher *et al.*, 2005, https://kepler-project.org) is an example of a data-driven, scientific data analysis and knowledge discovery pipeline. This `JAVA`-based application is very similar to the web-service-based implementation of `TAVERNA`, but relies on the `Ptolemy II` open-source software framework which support an actor-oriented pipeline design (Eker *et al.*, 2003). An actor can be seen as a step in the analysis pipeline, where multiple actors can be connected to each other *via* data channels. The `Ptolemy II` system was designed with heterogenous data in mind, and has been very successfuly implemented in automated pipelines by scientific groups (Lee and Zheng, 2005; Lee, 2009; Leung *et al.*, 2009).

Ergatis

`Ergatis` is a workflow management system optimised for paralellised analysis of constructed pipelines making use of the `Sun Grid Engine` (SGE, Orvis *et al.*, 2010). It is a workflow management system targeted for working with genome sequence data, where analysis pipelines can be executed on a single server, or distributed across large computing clusters. `Ergatis` was developed making use of standard ontologies in bioinformatics, and supports input files in the Bioinformatics Sequence Markup Language format (http://www.bsml.org), the Sequence Ontology for sequence feature annotation (Eilbeck *et al.*, 2005), and the Gene Ontology format for functional annotations (Gene Ontology Consortium, 2001). The workflow system has the added capabillity of exporting results into a `CHADO`-based database (Mungall *et al.*, 2007), making it compatible with the `GMOD` set of tools (Stein *et al.*, 2002). The `Ergatis` system executes scripts or tools locally and does not require a web-service as interface, in contrast to `TAVERNA` and `Kepler`, and offers a flexible user interface to manage and control executing workflows.

Galaxy

The `Galaxy` workflow system (Goecks *et al.*, 2010) has been used by several research groups for biological data analysis (Kosakovsky Pond *et al.*, 2009; Gaulton *et al.*, 2010; Peleg *et al.*, 2010). The goal of `Galaxy` is to serve as a layer of abstration on top of a myriad of underlying tools, and serve them to regular users through an intuitive web interface. The inputs and results from various programs, as well as the parameters used for each of these programs are stored in a history of a project or analysis step, which can be shared with collaborators, used as a workflow for similair analysis steps, or archived for publications. Almost any scriptable piece of software, including custom `Python`, `PERL` and `R` scripts can be wrapped in the `Galaxy` interface allowing for the easy extension of the framework to include custom tools. `Galaxy` hides the underlying complexity of the programs imbedded in it allowing users to focus on scientific hypotheses, rather than technical issues associated with the software needed to perform the analysis used to adress the biological questions.

1.6. Problem Statement

The hypothesis is formulated that by making use of data from Illumina mRNA-Seq deep sequencing data, the transcriptome of a complex eukaryotic organism like *Eucalyptus* can be successfully assembled and characterised to such an extend that biologically relevant and accurate information can be obtained regarding transcriptional control of growth and development.

In order to test the hypothesis, a structured approach is needed to first identify a suitable data management and data analysis framework to aid in the analysis of uHTS data. The data analysis framework will then be used to test the different parameters and settings of the software packages used to assemble and annotate the *Eucalyptus* transcriptome. The framework should be readily extendible with additional software tools that are not already implemented in the framework to aid in the analysis and construct automated workflows to perform the data analysis steps.

The workflows developed should then be used to perform a *de novo* assembly and homology-based annotation of the transcriptome of a *Eucalyptus grandis* x *Eucalyptus urophylla* plantation tree from deep sequenced mRNA-Seq data. The assembly should be validated as far as possible without the aid of the draft *Eucalyptus grandis* genome sequence, to validate that a *de novo* transcriptome assembly is indeed possible. The assembled gene catalog should be characterised and annotated with homologs from other angiosperm transcriptomes, and used to identify genes differentially expressed between xylogenic and phytosynthetic tissues.

To allow access to the assembled gene catalog, a web-based system should be developed that stores the contigs and corresponding annotations, and allows users to browse and search for contigs based on the annotations assigned to the contigs. The gene expression (FPKM) of the contig in each of the sampled tissues used perform the assembly should additionally be made available in the user interface.

1.7. Specific research questions and aims

- With the current selection of open-source uHTS data management and analysis packages available, is it possible to develop automated software workflows that perform DNA sequence analysis? In each of the developed workflows, identify the key parameters that have an effect on the results from a workflow. Where software tools are not present in the selected data management system, these tools should either be developed or added to the system to successfully perform a *de novo* assembly and annotation of a transcriptome dataset.
- To what extent can a transcriptome of a complex organism like *Eucalyptus* be assembled and evaluated using only mRNA-Seq data? The workflows developed in the previous aim should be use to completely assemble and annotate a large eukyryotic transcriptome. The assembled transcriptome should be evaluated for contig contiguity and the presence of full-length contigs in the dataset, without the aid of the *Eucalyptus* genome sequence. Functional annotation of the transcripts should be made in an automated fashion, and the transcript dataset should be compared to other angiosperm datasets in terms of the number and diversity of the assembled contigs. Finally, the gene expression profiles (FPKM) values of the transcripts should be used to identify a set of differentially expressed genes in xylogenetic and phytosynthetic tissues.
- Development of an intuitive, web-based *Eucalyptus* specific transcriptome resource that enables users to query and browse the assembled transcriptome dataset based on annotations? The web-resource should serve as a central repository for the data generated in the previous aims, and should be considered as a development platform and extension point for future whole genome mRNA-Seq based transcriptome sequencing and expression studies in *Eucalyptus*.