TRAINING MANUAL

DBT Sponsored Training Program

On

Next Generation Sequencing Data Analysis & Annotation

March 12 -16, 2013

INDIAN INSTITUTE OF SPICES RESEARCH

(Indian Council of Agricultural Research)
Kozhikode – 673012, KERALA

Published by

M. Anandaraj
Director

Organized by

Dr. Santhosh J Eapen, Co-ordinator
Dr. D. Prasath, Co-cordinator

Compiled & Edited by

Santhosh J Eapen
Rosana Babu
Dhanya K. P.
Faisal Moossa Athikkavil
Anoop T. A.

The manual is an in-house publication intended for training purposes only and is not for public circulation.

NEXT GENERATION SEQUENCING

Data Analysis &Annotation | 12-16 March 2013

GUEST FACULTY

	'Insights into NGS data analysis' by Ashwani Kumar Mishra *NxGenBio Life Sciences, Ne w Delhi*	13 March 2013 09.30 am ATIC Hall
	'NGS Data – Quality check, Assembly' by Rakesh Kumar Maurya *NxGenBio Life Sciences, New Delhi*	13 March 2013 02.00 pm ATIC Hall
	' Genotyping with RAD and ddRAD Sequencing' by Ramprasad V.L. *SciGenom Labs Pvt. Ltd., Chennai*	15 March 2013 01.30 pm ATIC Hall
	' Whole genome sequencing, assembly and annotation' by Aravind Ravindran *Texas A&M University, USA*	15 March 2013 03.00 pm ATIC Hall
	'Data Analysis Using CLC bio Genomic Workbench ' by Uday Deshpande, *CLC Bio India Pvt. Ltd., Visakhapatnam*	16 March 2013 09.30 am ATIC Hall
	'Insights from sequencing and annotation of insect genomes: Sanger vs. Next Generation Sequencing' by Tony Grace *Central University of Kerala, Kasaragod*	16 March 2013 03.00 pm Silver Jubilee Hall

CONTENTS

iisr सत्यमेव जयते DBT ICAR

DBT Sponsored Training Program

Next Generation Sequencing- Data Analysis and Annotation

March 12 - 16, 2013

Schedule

	Tuesday, 12 March 2013
09.30 am	Registration
10.00 am	Inauguration and address
11.00 am	Tea/Coffee & Group Photo
11.15 am	Pre-evaluation *by P. Rajeev, Senior Scientist, IISR, Calicut*
11.30 am	Next Generation Sequencing (L) *by A. I. Bhat, Principal Scientist, IISR, Calicut*
01.30 pm	Data Formats & QC analysis of read data (L&P) *by Rosana O. B., Research Scholar, IISR, Calicut*
	Wednesday, 13 March 2013
09.30 am	Insights in to NGS Data Analysis (L) *by Ashwani Kumar Mishra, NxGenBio, New Delhi*
01.30 pm	NGS Data Analysis – Quality check & Assembly (L&P) *by Rakesh Kumar Maurya, NxGenBio, New Delhi*
	Thursday, 14 March 2013
09.30 am	Genome Assembly & Variant Analysis (L&P) *by Rosana O. B., Research Scholar, IISR, Calicut*
01.30 pm	Structural & Functional Annotation (L&P) *by Rosana O. B., Research Scholar, IISR, Calicut*
	Friday, 15 March 2013
09.30 am	Plant Transcriptome Mapping (L&P) *by D. Prasath, Senior Scientist, IISR, Calicut*
01.30 pm	Genotyping with RAD & ddRAD Sequencing (L) *by Ramaprasad, Principal Scientist, SciGenome, Kochi*
03.00 pm	Whole genome sequencing, assembly and annotation of bacteria (L) *by Aravind Ravindran, Post doc fellow, Texas A&M University, USA*
	Saturday, 16 March 2013
09.30 am	Black Pepper Root Transcriptome Data Analysis Using CLC bio Genomic Workbench (L) *De novo* assembly (P) *by Uday Deshpande, Field Application Scientist, CLC Bio India Pvt. Ltd,. Visakhapatnam*
01.30 pm	Comparative Genomics (L&P) *by Santhosh J. Eapen, Principal Scientist, IISR, Calicut*
02.45 pm	Post evaluation & Feed back
03.00 pm *Plenary lecture*	Valedictory Insights from sequencing and annotation of insect genomes: Sanger Vs. Next Generation Sequencing *by Tony Grace, Central University of Kerala, Kasargod*

Chapter 1

Next Generation Sequencing Technologies

Dr. A. I. Bhat

Introduction

NGS technologies have demonstrated the capacity to sequence DNA at unprecedented speed, thereby enabling previously unimaginable scientific achievements and novel biological applications. But, the massive data produced by NGS also presents a significant challenge for data storage, analyses, and management solutions. Advanced bioinformatics tools are essential for the successful application of NGS technology. As evidenced throughout this review, NGS technologies will have a striking impact on genomic research and the entire biological field. With its ability to tackle the unsolved challenges unconquered by previous genomic technologies, NGS is likely to unravel the complexity of the human genome in terms of genetic variations, some of which may be confined to susceptible loci for some common human conditions. This chapter reviews basic concepts, general applications, and the potential impact of next- generation sequencing (NGS) technologies on genomics, with particular reference to currently available and possible future platforms and bioinformatics.

Since the time DNA was discovered as the code to all biological life on earth, man has sought to unravel its mysteries. If the genetic code could be sequenced or "read", the origins of life itself may be revealed. Although this thought might not be entirely true, the efforts to date made have certainly revolutionized the biological field. The "original" sequencing methodology, known as Sanger chemistry, uses specifically labeled nucleotides to read through a DNA template during DNA synthesis. This sequencing technology requires a specific primer to start the read at a specific location along the DNA template, and record the different labels for each nucleotide within the sequence. After a series of technical innovations, the Sanger method has reached the capacity to read through 1000–1200 basepair (bp); however, it still cannot surpass 2 kilo basepair (Kbp) beyond the specific sequencing primer.

In order to sequence longer sections of DNA, a new approach called shotgun sequencing was developed during Human Genome Project (HGP). In this approach, genomic DNA is enzymatically or mechanically broken down into smaller fragments and cloned into sequencing vectors in which cloned DNA fragments can be sequenced individually. The complete sequence of a long DNA fragment can be eventually generated by these methods by alignment and reassembly of sequence fragments based on partial sequence overlaps. Shotgun sequencing was a significant advantage from HGP, and made sequencing the entire human genome possible. The core philosophy of massive parallel sequencing used in next- generation sequencing (NGS) is adapted from shotgun sequencing.

New NGS technologies read the DNA templates randomly along the entire genome. This is accomplished by breaking the entire genome into small pieces, then ligating those small pieces of DNA to designated adapters for random read during DNA synthesis (sequencing- by-synthesis). Therefore, NGS technology is often called massively parallel sequencing.

The read length (the actual number of continuous sequenced bases) for NGS is much shorter than that attained by Sanger sequencing. At present, NGS only provides 50–500 continuous basepair reads, which is why sequencing results are defined as short reads. These short reads are a major limitation in current technology; however, developing NGS technologies, such as single-molecule sequencing, may surpass Sanger methodologies and have the potential to read several continuous kilo basepairs (Kbps) (Table 1). Since next-generation technologies currently produce short reads, coverage is a very important issue. Coverage is defined as the number of short reads that overlap each other within a specific genomic region. For example, a 30-fold coverage for CYP2D6 gene means that every nucleotide within this gene region is represented in at least 30 distinct and overlapping short reads. Sufficient coverage is critical for accurate assembly of the genomic sequence. In addition to the need for adequate coverage, short reads create many sequences that cannot be interpreted or “mapped” to any reference DNA or be accurately assembled. This is simply because some of the short reads are too short and may match with many different regions of the genome and are not unique to any specific region of the sequence. Short-read sequences that can be assembled and matched with a reference sequence are generally called “mappable reads”. NGS is a rapidly evolving technology that is changing on an almost daily basis. The purpose of this review is to highlight these advances and bring the reader up to date on the latest technological achievements in DNA sequencing technologies, particularly as related to genomics. Following completion of the HGP, a new approach, genome-wide association study (GWAS), was widely applied to genomics. Although several early GWAS studies reported potentially promising results, the majority of GWAS studies were disappointing because of inadequate sample size, limitation of arrays for certain genetic variations, and/or heterogeneity in phenotype. These obstacles may be overcome by new genomic technology, i.e., next-generation sequencing (NGS), also known as massively parallel sequencing or multiplex cyclic sequencing. Since many genetic variants which contribute to many human conditions are still unknown, unbiased whole-genome sequencing will help to identify these genetic variants, including single nucleotide variants (SNVs) or single nucleotide polymorphisms (SNPs), small insertions and deletions (indels, 1–1000 bp), and structural and genomic variants (>1000 bp).

Previously, DNA sequencing was performed almost exclusively by the Sanger method, which has excellent accuracy and reasonable read length but very low throughput. Sanger sequencing was used to obtain the first consensus sequence of the human genome in 2001 and the first individual human diploid sequence. Shortly thereafter, the second complete individual genome was sequenced using next-generation technology, which marked the first human genome sequenced with new NGS technology. Since then, several additional diploid human genomes have been sequenced with NGS utilizing a variety of related techniques to rapidly sequence genomes with varying degrees of coverage. A common strategy for NGS is to use DNA synthesis or ligation process to read through many different DNA templates in parallel. Therefore, NGS reads DNA templates in a highly parallel manner to generate massive amounts of sequencing data but, as mentioned above, the read length for each DNA template is relatively short (35–500 bp) compared to traditional Sanger sequencing (1000–1200 bp).

Several NGS methods recently developed allow larger-scale DNA sequencing. The number of large short-read sequences from NGS is increasing at exponential rates. Currently, five NGS platforms are commercially available, including the Roche GS-FLX 454 Genome Sequencer (originally 454 sequencing), the Illumina Genome Analyzer (originally Solexa technology), the ABI SOLiD analyzer, Polonator G.007 and the Helicos HeliScope platforms. These NGS instruments generate different base read lengths, different error rates, and different error profiles relative to Sanger sequencing data and to each other. NGS technologies have increased the speed and throughput capacities of DNA sequencing and, as a result, dramatically reduced overall sequencing costs.

History of DNA Sequencing

First- and second-generation sequencing technologies have led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances, including enabling a more complete understanding of whole genome sequences and the information encoded therein, a more complete characterization of the methylome and transcriptome and a better understanding of interactions between proteins and DNA. Nevertheless, there are sequencing applications and aspects of genome biology that are presently beyond the reach of current sequencing technologies, leaving fertile ground for additional innovation in this space. We also describes a new generation of single-molecule sequencing technologies (third generation sequencing) that is emerging to fill this space, with the potential for dramatically longer read lengths, shorter time to result and lower overall cost.

First-generation sequencing

First-generation sequencing was originally developed by Sanger in 1975 (the chain-termination method) and in parallel by Maxam and Gilbert in 1977 (a chemical sequencing method). From these first-generation methods, Sanger sequencing ultimately prevailed given it was less technically complex and more amenable to being scaled up. For Sanger sequencing practiced today, during sample preparation, different-sized fragments of DNA are generated each starting from the same location. Each fragment ends with a particular base that is labeled with one of four fluorescent dyes corresponding to that particular base. Then all of the fragments are distributed in the order of their length via capillary electrophoresis. Information regarding the last base is used to determine the original sequence. This method results in a read length that is _800 bases on average, but may be extended to above 1000 bases. While fully automated implementations of this approach were the mainstay for the original sequencing of the human genome, their chief limitation was the small amounts of DNA that could be processed per unit time, referred to as throughput, as well as high cost, resulting in it taking roughly 10 years and three billion dollars to sequence the first human genome .

Second-generation sequencing

Commercial SGS tools emerged in 2005 in response to the low throughput and high cost of first-generation methods. To address this problem, SGS tools achieve much higher throughput by sequencing a large number of DNA molecules in parallel. With most SGS technologies, tens of thousands of identical strands are anchored to a given location to be read in a process consisting of successive washing and scanning operations. The 'wash-and-scan' sequencing process involves sequentially flooding in reagents, such as labeled nucleotides, incorporating nucleotides into the DNA strands, stopping the incorporation reaction, washing out the excess reagent, scanning to identify the incorporated bases and finally treating the newly incorporated bases to prepare the DNA templates for the next 'wash-and-scan' cycle. This cycle is repeated until the reaction is no longer viable. The array of DNA anchor locations can have a very high density of DNA fragments, leading to extremely high overall throughput and a resultant low cost per identified base when such instruments are run at high capacity. For example, Illumina's HiSeq 2000 instrument can generate upwards of 300 or more gigabases of sequence data in a single run. The time-to-result for these SGS methods is generally long (typically taking many days), due to the large number of scanning and washing cycles required. Furthermore, because step yields for the addition of each base are, 100%, a population of molecules becomes more asynchronous as each base is added. This loss of synchronicity (called dephasing) causes an increase in noise and sequencing errors as the read extends, effectively limiting the read length produced by the most widely used SGS systems to significantly less than the average read lengths achieved by Sanger sequencing. Further, in order to generate this large number of DNA molecules, PCR amplification is required. The amplification process can introduce errors in the template sequence as well as amplification bias. The effects of these pathologies are that neither the sequences nor the frequencies with which they appear are always faithfully preserved. In addition, the process of amplification increases the complexity and time associated with sample preparation. Finally, the massively high throughput achieved by SGS technologies per run generates mountains of highly informative data that challenge data storage and informatics operations, especially in light of the shorter reads (compared with Sanger sequencing) that make alignment and assembly processes challenging. First-generation sequencing and SGS technologies have led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances. Nevertheless there are sequencing applications and aspects of genome biology that are presently beyond the reach of current sequencing technologies, leaving fertile ground for additional innovation in this space.

Next-generation sequencing platforms

Among the five commercially available platforms, the Roche/454 FLX, the Illumina/Solexa Genome Analyzer, and the Applied Biosystems (ABI) SOLiD Analyzer are currently dominating the market. The other two platforms, the Polonator G.007 and the Helicos HeliScope, have just recently been introduced and are not widely used. Additional platforms from other manufacturers are likely to become available within the next few years and bring NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript new and exciting technologies, faster sequencing speed, and a more affordable price. Methodologies used by each of the current available NGS systems are discussed below.

(i) Roche GS-FLX 454 Genome Sequencer

The Roche GS-FLX 454 Genome Sequencer was the first commercial platform introduced in 2004 as the 454 Sequencer. The second complete genome of an individual was sequenced with this platform). The 454 Genome Sequencer uses sequencing-by-synthesis technology known as pyrosequencing. The key procedure in this approach is emulsion PCR in which single-stranded DNA binding beads are encapsulated by vigorous vortexing into aqueous micelles containing PCR reactants surrounded by oil for emulsion PCR amplification. During the pyrosequencing process, light emitted from phosphate molecules during nucleotide incorporation is recorded as the polymerase synthesizes the DNA strand. Initially, the 454 Sequencer had a read length of 100 bp but now can produce an average read length of 400 bp. The maximum ~600 bp capacity of 454 systems approaches the halfway of current Sanger sequencing capacities (~1200 bp). At 600 bp, the 454 Sequencer has the longest short reads among all the NGS platforms; and generates ~400–600 Mb of sequence reads per run; critical for some applications such as RNA isoform identification in RNA-seq and de novo assembly of microbes in metagenomics. Raw base accuracy reported by Roche is very good (over 99%); however, the reported relatively error-prone raw data sequence, especially associated with insertion-deletions, is a major concern. Low yield of sequence reads could translate into a much higher cost if additional coverage is needed to define a genetic mutation.

(ii) Illumina/Solexa Genome Analyzer

The Illumina/Solexa Genome Analyzer was the second platform to reach market, and currently is the most widely used system. The Illumina platform uses sequencing-by- synthesis approach in which all four nucleotides are added simultaneously into oligo-primed cluster fragments in flow-cell channels along with DNA polymerase. Bridge amplification extends cluster strands with all four fluorescently labeled nucleotides for sequencing. The Genome Analyzer is widely recognized as the most adaptable and easiest to use sequencing platform. Superior data quality and proper read lengths have made it the system of choice for many genome sequencing projects. To date, the majority of published NGS papers have described methods using the short sequence data produced with the Genome Analyzer. At present, the new Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2 × 100 basepairs (pair-end reads), and generates about 200 giga basepair (Gbp) of short sequences per run. The raw base accuracy is greater than 99.5%.

(iii) ABI SOLiD platform

The ABI SOLiD platform uses a unique sequencing-byligation approach in which it uses an emulsion PCR approach with small magnetic beads to amplify the DNA fragments for parallel sequencing. During SOLiD sequencing, DNA ligation is carried out to link specific fluorescent labeled 8-mer oligonucleotides for “dinucleotide-encoding”, whose 4^{th} and 5th bases are encoded by specific fluorescence. Each fluorescent marker on a 8-mer identifies a two-base combination, which can be further distinguished with a universal primer offsetting scheme. The primer offsetting scheme allows a universal primer that is offset by one base from the adapter-fragment position to hybridize to DNA templates in five cycle sets permitting the entire fragment to be sequenced and each base position sequenced twice during each cycle. Each ligation step is followed by fluorescence detection and another round of ligation. SOLiD4 analyzer has a read length of up to 50 bp and can produce 80–100 Gbp of mappable sequences per run. The latest model, 5500×l solid system (previously known as SOLiD4hq) can generate over 2.4 billion reads per run with a raw base accuracy of 99.94% due to its 2-base encoding mechanism. This instrument is unique in that it can process two slides at a time; one slide is receiving reagents while the other is being imaged. The SOLiD4 platform probably provides the best data quality as a result of its sequencing-by-ligation approach but the DNA library preparation procedures prior to sequencing can be tedious and time consuming. The newly marketed EZ-Bead system may provide some resolution to this problem.

(iv) Danaher/Dover/Azco Polonator G.007

The Danaher/Dover/Azco Polonator G.007 is a new platform on the market with emphasis on competitive pricing. The Polonator platform employs a sequencing-by-ligation approach using a randomly arrayed, bead-based, emulsion PCR to amplify DNA fragments for parallel sequencing. The short-read length is 26 bp, and 8–10 Gbp of sequence reads are generated per run, with 92% of the reads mappable. The random bead-based array will likely be replaced with their patented rolonies technology (rolling circle colonies) on an ordered array to increase accuracy and improve read length.

(v) Helicos HeliScope

The Helicos HeliScope platform is the first single molecular sequencing technology available that uses a highly sensitive fluorescence detection system to directly detect each nucleotide as it is synthesized. The distinct characteristic of this technology is its ability to sequence single DNA molecules without amplification, defined as Single-Molecule Real Time (SMRT) DNA sequencing. The short-read length ranges from 30 bp to 35 bp at present time, with a raw base accuracy greater than 99%, and 20–28 Gbp of potential sequence reads per run in the near future.

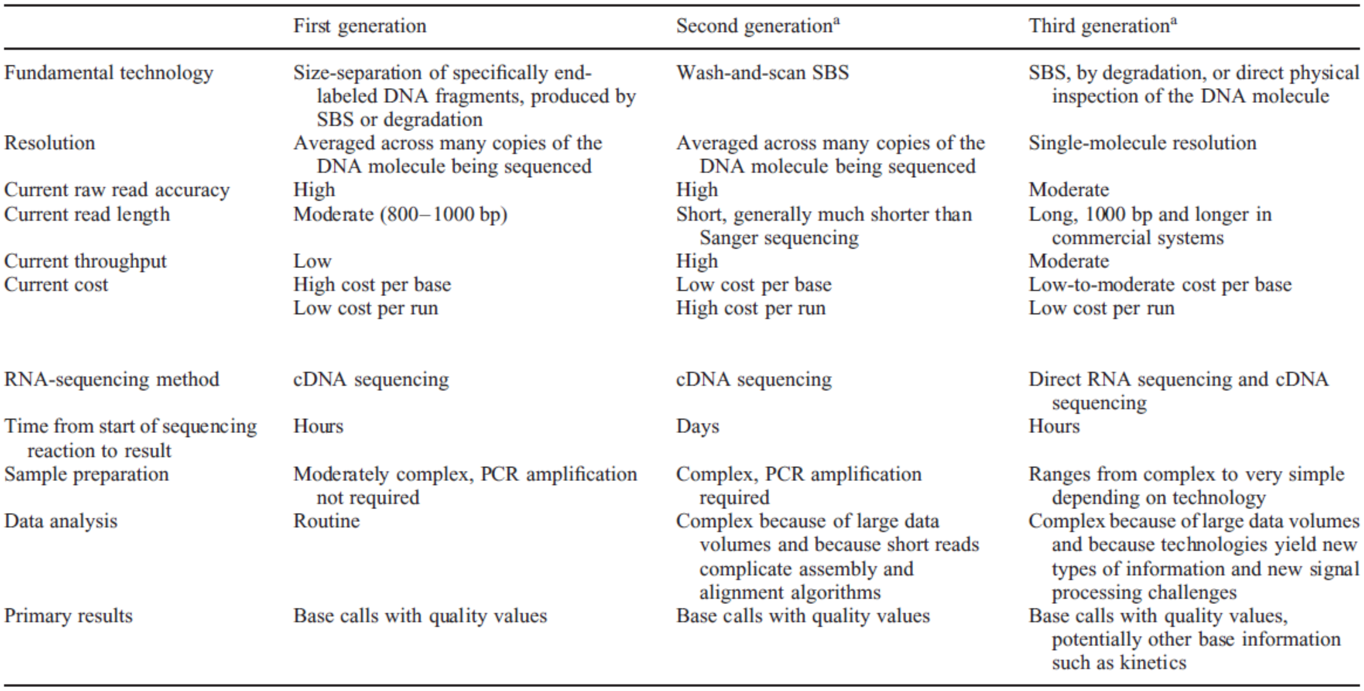
The advantage of single-molecule DNA sequencing technology is its potential to read extremely long sequences and fast sequencing speed, which could translate into a dramatic reduction in overall sequencing cost. As such, advanced single DNA molecule sequencing technology has been defined as the next-NGS technology. However, the basic philosophy of massive parallel sequencing is still the same and the term next- generation sequencing (NGS) will only be used in this review. More detailed technical description of these platforms is available elsewhere.

Transitioning from SGS to TGS

The genomics community has been enormously enabled by first- and second-generation sequencing (SGS) technologies in comprehensively characterizing DNA sequence variation, de novo sequencing of a number of species, sequencing of microbiomes, detecting methylated regions of the genome, quantitating transcript abundances, characterizing different isoforms of genes present in a given sample and identifying the degree to which mRNA transcripts are being actively translated. One of the hallmark features of the SGS technologies is their massive throughput at a modest cost, with hundreds of gigabases of sequencing now possible in a single run for several thousand dollars. Despite the recent and rapid acceptance of SGS technologies, a new generation of single-molecule sequencing (SMS) technologies is emerging. Unlike major SGS sequencing by synthesis (SBS) technologies that rely on PCR to grow clusters of a given DNA template, attaching the clusters of DNA templates to a solid surface that is then imaged as the clusters are sequenced by synthesis in a phased approach, the new generation of SBS technologies interrogate single molecules of DNA, such that no synchronization is required (a limitation of SGS), thereby overcoming issues related to the biases introduced by PCR amplification and dephasing. More importantly, this new generation of sequencing technologies has the potential to exploit more fully the high catalytic rates and high processivity of DNA polymerase or avoid any biology or chemistry altogether to radically increase read length (from tens of bases to tens of thousands of bases per read) and time to result (from days to hours or minutes). The promises then of this new, third generation of sequencing technologies in offering advantages over current sequencing technologies are (i) higher throughput; (ii) faster turnaround time (e.g. sequencing metazoan genomes at high fold coverage in minutes); (iii) longer read lengths to enhance de novo assembly and enable direct detection of haplotypes and even whole chromosome phasing; (iv) higher consensus accuracy to enable rare variant detection; (v) small amounts of starting material (theoretically only a single molecule may be required for sequencing); and (vi) low cost, where sequencing the human genome at high fold coverage for less than $100 is now a reasonable goal for the community.

There may not yet be consensus on what constitutes a third generation, or next–next-generation sequencing instrument, given advances are being made on rapid time scales that do not easily fit into generational time scales. However, SMS without the need to halt between read steps (whether enzymatic or otherwise), where reads from SMS instruments represent sequencing of a single molecule of DNA. SMS technologies that do not purposefully pause sequencing reaction after each base incorporation represent the most thoroughly explored TGS approaches in hopes of increasing sequencing rates, throughput and read lengths, lowering the complexity of sample preparation and ultimately decreasing cost. However, as a result of using these criteria to define TGS, a number of exciting technologies do not fit neatly into this definition, but are nevertheless exciting in terms of how they complement current SGS technologies.

Table 1. Comparison of First generation, second generation and third generation sequencing techniques.



	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Next- Next-generation sequencing platforms (Third Generation Sequencing)

SMS technologies can roughly be binned into three different categories: (i) SBS technologies in which single molecules of DNA polymerase are observed as they synthesize a single molecule of DNA; (ii) nanopore-sequencing technologies in which single molecules of DNA are threaded through a nanopore or positioned in the vicinity of a nanopore, and individual bases are detected as they pass through the nanopore; (iii) direct imaging of individual DNA molecules using advanced microscopy techniques etc. Each of these technologies provides novel approaches to sequencing DNA and has advantages and disadvantages with respect to specific applications. These technologies are at varying stages of development, making the writing of a review on TGS difficult given there is still much to prove regarding the utility of many of the TGS technologies.

However, if the full potential of these technologies is realized, in several years time, whole genome sequencing will likely be fast enough and inexpensive enough to resequence genomes as needed for any application. Here we discuss many of the emerging TGS technologies that have the potential to make such stunning advances possible.

Since single DNA molecule sequencing technology can read through DNA templates in real time without amplification, it provides accurate sequencing data with potentially long-reads and efforts have focused recently in this new direction. Several unique single- molecule DNA sequencing technologies are currently under development; however, little information has been made publically available.

(i) SMS sequencing by synthesis

Single-molecule real-time sequencing: The single-molecule real-time (SMRT) sequencing approach developed by Pacific Biosciences is the first TGS approach to directly observe a single molecule of DNA polymerase as it synthesizes a strand of DNA, directly leveraging the speed and processivity of this enzyme to address many of the shortcomings of SGS. Given that a single DNA polymerase molecule is of the order of 10 nm in diameter, two important obstacles needed to be overcome to enable direct observation of DNA synthesis as it occurs in real time are: (i) confining the enzyme to an observation volume that was small enough to achieve the signal-to-noise ratio needed to accurately call bases as they were incorporated into the template of interest; and (ii) labeling the nucleotides to be incorporated in the synthesis process such that the dye–nucleotide linker is cleaved after completion of the incorporation process so that a natural strand of DNA remains for continued synthesis and so that multiple dyes are not held in the confinement volume at a time (something that would destroy the signal-to-noise ratio). The problem of observing a DNA polymerase working in real time, detecting the incorporation of a single nucleotide taken from a large pool of potential nucleotides during DNA synthesis, was solved using zero-mode waveguide (ZMW) technology.

(ii) Nano-technologies for single-molecule sequencing

Thousands of nano-tunnels on a chip can be used to monitor the movement of a polymerase molecule on a single DNA strand during replication to perform single-molecule DNAsequencing-by-synthesis. Nano-technologies have long been considered a cutting-edge technology for single-molecule DNA sequencing and several nanopore sequencing concepts and technologies are currently under development. One concept is based on the observation that when a DNA strand is pulled through a nanopore by an electrical current, each nucleotide base (A, T, C, G) creates a unique pattern in the electrical current. This unique nanopore electrical current fingerprint can be used for nanopore sequencing.

(iii) Fluorescence-based single-molecule sequencing

Pacific BioSciences is developing a single-molecule real time (SMRT) DNA sequencing technology. This approach performs single-molecule sequencing by identifying nucleotides which are phospholinked with distinctive colors. During the synthesis process, fluorescence emitted as the phosphate chain is cleaved and the nucleotide is incorporated by a polymerase into a single DNA strand.

(iv) Electronic detection for single-molecule sequencing

Reveo is developing a technology to stretch out DNA molecules on conductive surfaces for electronic base detection. A stretched and immobilized strand of DNA will be read through by multiple nano-knife edge probes. Each nano-knife edge probe specifically recognizes only one nucleotide for single-molecule sequencing. Intelligent Biosystems is also developing a platform using the electronic detection approach which will allow for high speed and high sensitivity single-molecule analysis with decreased background noise.

(v) Electron microscopy for single-molecule sequencing

Electron microscopy (EM) was the first proposed and attempted approach to sequence DNA molecules before the Sanger sequencing was established and this concept has recently been reevaluated with the emergence of new technologies. Since scanning tunneling microscopy (STM) can reach atomic resolution, STM for single-molecule sequencing is being explored. LightSpeed Genomics is developing a microparticle approach by capturing sequence data with optical detection technology and new sequencing chemistry from a large field of view to reduce the time consuming sample and detector rearrangement. Halcyon Molecular is developing a DNA sequencing technology by atom-by-atom identification and EM analysis. The key advantage of this technology is very long read lengths. ZS Genetics is also developing EM-based technologies for single-molecule DNA sequencing.

Other approaches for single-molecule sequencing

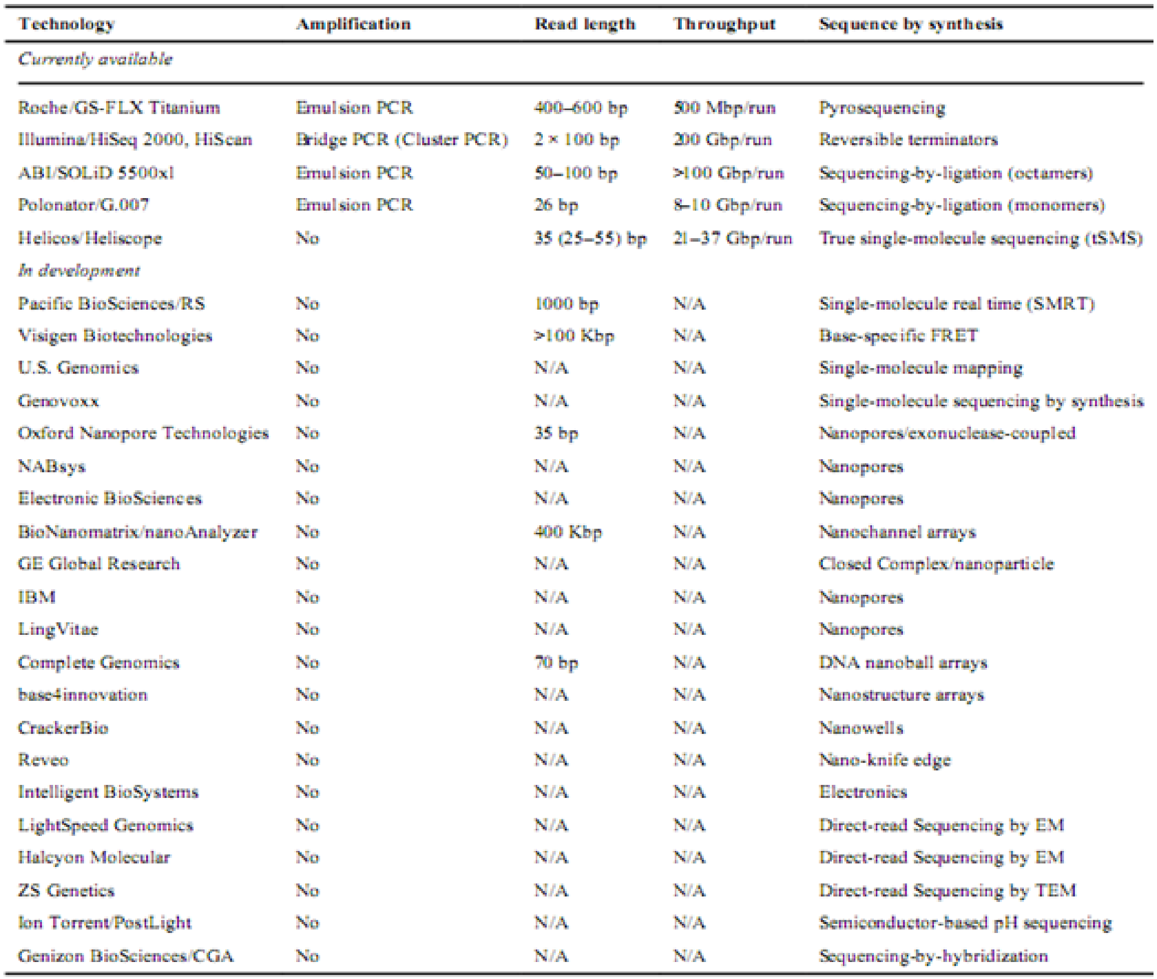
Ion Torrent developed an entirely new approach to sequencing based on the well- characterized biochemistry that when a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct. They have developed an ion sensor that can detect hydrogen ions and directly convert the chemical information to digital sequence information. In essence, their NGS platform can be defined as the world's smallest solid-state pH meter.

Focusing on resequencing specific sections of the human genome combined with genome- region enrichment, Genizon BioSciences is developing a sequencing-by-hybridization technology based on known reference sequences. Avantome (acquired by Illumina) is also exploring the single-molecule sequencing technologies.

Road to the personal genome project

Since the initiation of 1000 genome project the cost of sequencing an individual genome has been rapidly decreasing and will likely reach $1000 per person within a short period of time, making personalized medicine become a possible reality. In genomics, the personal genome era made available by NGS technologies will mark a significant milestone in entire genomic research field in the foreseeable future. It is not clear which NGS technology will eventually dominate the genomic research field, but it is almost certain that further reductions in cost, rapid increases in sequencing speed with improved accuracy, and the advantages conferred by these new technologies will assure that NGS will become an essential molecular tool affecting all aspects of the biological sciences. Detailed information of the NGS technologies and platform discussed above is summarized in Table 2.

Table 2. Platforms and detailed information on NGS technologies

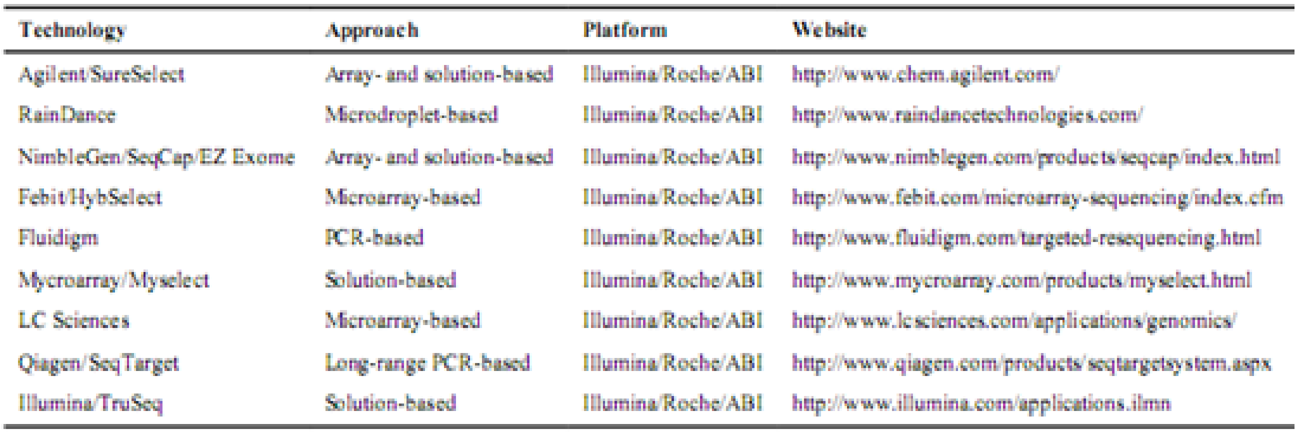


Technology	Amplification	Read length	Throughput	Sequence by synthesis
Currently available				
Roche/GS-FLX Titanium	Emulsion PCR	400–600 bp	500 Mbp/run	Pyrosequencing
Illumina/HiSeq 2000, HiScan	Bridge PCR (Cluster PCR)	2 × 100 bp	200 Gbp/run	Reversible terminators
ABI/SOLiD 5500xl	Emulsion PCR	50–100 bp	>100 Gbp/run	Sequencing-by-ligation (octamers)
Polonator/G.007	Emulsion PCR	26 bp	8–10 Gbp/run	Sequencing-by-ligation (monomers)
Helicos/Heliscope	No	35 (25–55) bp	21–37 Gbp/run	True single-molecule sequencing (tSMS)
In development				
Pacific BioSciences/RS	No	1000 bp	N/A	Single-molecule real time (SMRT)
Visigen Biotechnologies	No	>100 Kbp	N/A	Base-specific FRET
U.S. Genomics	No	N/A	N/A	Single-molecule mapping
Genovoxx	No	N/A	N/A	Single-molecule sequencing by synthesis
Oxford Nanopore Technologies	No	35 bp	N/A	Nanopores/exonuclease-coupled
NABsys	No	N/A	N/A	Nanopores
Electronic BioSciences	No	N/A	N/A	Nanopores
BioNanomatrix/nanoAnalyzer	No	400 Kbp	N/A	Nanochannel arrays
GE Global Research	No	N/A	N/A	Closed Complex/nanoparticle
IBM	No	N/A	N/A	Nanopores
LingVitae	No	N/A	N/A	Nanopores
Complete Genomics	No	70 bp	N/A	DNA nanoball arrays
base4innovation	No	N/A	N/A	Nanostructure arrays
CrackerBio	No	N/A	N/A	Nanowells
Reveo	No	N/A	N/A	Nano-knife edge
Intelligent BioSystems	No	N/A	N/A	Electronics
LightSpeed Genomics	No	N/A	N/A	Direct-read Sequencing by EM
Halcyon Molecular	No	N/A	N/A	Direct-read Sequencing by EM
ZS Genetics	No	N/A	N/A	Direct-read Sequencing by TEM
Ion Torrent/PostLight	No	N/A	N/A	Semiconductor-based pH sequencing
Genizon BioSciences/CGA	No	N/A	N/A	Sequencing-by-hybridization

Current strategies for the NGS project

To ensure the correct identification of genetic variants, short-read coverage must be sufficient to ensure the complete and accurate sequence assembly. Currently, at least 30× coverage is recommended in whole-genome scans for rare genetic variants in human genomes, which is a burden on computer resources and cost management. Although the cost of whole-genome sequencing has dropped substantially, the cost remains a major obstacle; whole-genome sequencing of a single individual currently costs approximately $100,000.00. By targeting specific regions of interest, selective DNA enrichment techniques improve the overall cost and efficiency of NGS; however, targeted enrichment must maintain uniform coverage, high reproducibility, and no allele bias for any genomic region. Targeted sequencing generally focuses on all protein-coding subsequences (the functional exome), which only requires ~5% as much sequencing compared to that required for the entire human genome. This strategy currently reduces the overall cost to around $10,000 or less for the sequencing of a single individual. An important consideration to the cost of such experiments is the depth of sequence coverage required to achieve a desired sensitivity and specificity of at least 25-fold nominal sequence coverage. The most common techniques for targeted sequence enrichment are either microarray-based or solution hybrid- based. Several targeted selection technologies have been marketed and successfully applied in different NGS projects with variable success and may become the tools of choice to lower the burden of time and cost. For example, using targeted selection strategy, the mutations in DHODH from four individuals from three unrelated families with Miller syndrome have been successfully identified, illustrating that selective DNA enrichment techniques will dramatically reduce overall cost and accelerate discovery of genetic variants that cause rare and yet to be discovered genetic disorders. Other genetic loci for rare diseases have also been successfully identified through exome sequencing, further validating this strategy. Commercially available products for targeted sequence-enrichment include Agilent's SureSelect and NimbleGen's SeqCap/EZ Exome (both array- and solution-based technologies), RainDance and Illumina's TruSeq (solution-based technology), Febit's HybSelect and LC Sciences (microarray-based strategy), Qiagen and Fluidigm (PCR-based method) (Table 3).

Table 3. Targeted sequence- enrichment technologies for NGS.



Technology	Approach	Platform	Website
Agilent/SureSelect	Array- and solution-based	Illumina/Roche/ABI	http://www.chem.agilent.com/
RainDance	Microdroplet-based	Illumina/Roche/ABI	http://www.raindancetechnologies.com/
NimbleGen/SeqCap/EZ Exome	Array- and solution-based	Illumina/Roche/ABI	http://www.nimblegen.com/products/seqcap/index.html
Febit/HybSelect	Microarray-based	Illumina/Roche/ABI	http://www.febit.com/microarray-sequencing/index.cfm
Fluidigm	PCR-based	Illumina/Roche/ABI	http://www.fluidigm.com/targeted-resequencing.html
Mycroarray/Myselect	Solution-based	Illumina/Roche/ABI	http://www.mycroarray.com/products/myselect.html
LC Sciences	Microarray-based	Illumina/Roche/ABI	http://www.lcsciences.com/applications/genomics/
Qiagen/SeqTarget	Long-range PCR-based	Illumina/Roche/ABI	http://www.qiagen.com/products/seqtargetsystem.aspx
Illumina/TruSeq	Solution-based	Illumina/Roche/ABI	http://www.illumina.com/applications.ilmn

Bioinformatics for NGS data

The parallel short-read strategy of NGS opens many challenges for bioinformatics to interpret the short reads and the genetic variations in human genomes. The full benefit of NGS will not be achieved until bioinformatics are able to maximally interpret and utilize these short-read sequences, including alignment, assembly, etc. Typically, tens or hundreds of Gbp short reads can be generated during each run in any given NGS platform. As a result, the average NGS experiment generates terabytes of raw data, making data analysis and management of data problematic. Given the vast amount of data produced by NGS, developing a massive data storage and management solution and creating informatics tools to effectively analyze data will be essential to the successful application of NGS technology. Further adding to the bioinformatics problems, there are differences among the various NGS platforms in term of data format, length of reads, etc., which results in the need for diversity in bioinformatics including sequence quality scoring, alignment, assembly, and data processing.

The benefits of NGS sequencing will not be fully appreciated until extremely high- performance computing and intensive bioinformatics support is available. The information accrued by NGS may lead to a paradigm shift in the way that genetics and bioinformatics converge. Since NGS technology is in an early stage of development, a variety of software tools are under development and many are available online for NGS data analysis. Their functions fit into several general categories: (1) alignment of reads to a reference sequence; (2) de novo assembly, (3) reference-based assembly; (4) base-calling and/or genetic variation detection (such as SNV, Indel); (5) genome annotation, and (6) utilities for data analysis.

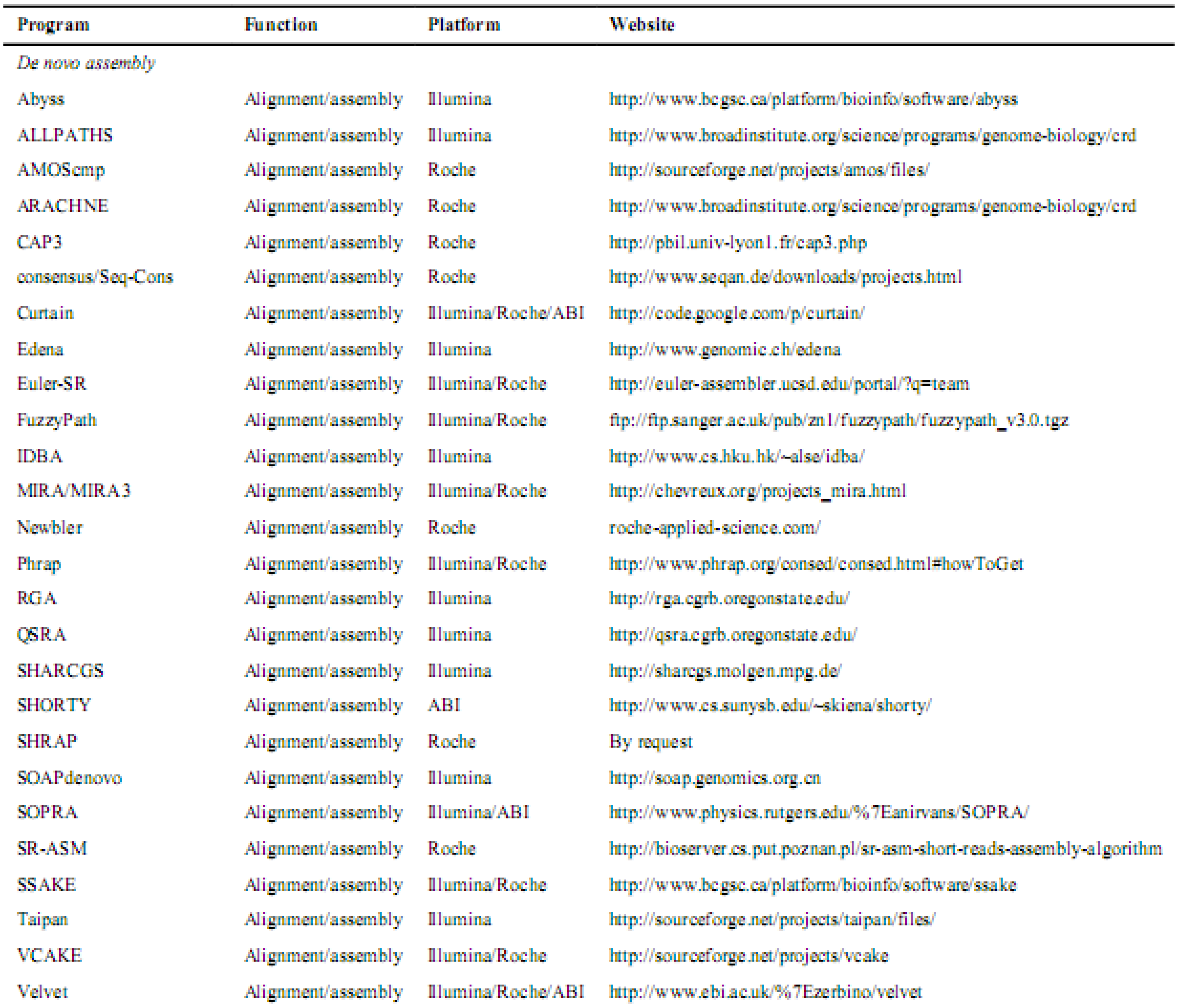
Alignment and assembly

Despite the sequencing power of NGS, the short-read length strategy creates serious limitations in many biological applications. Efforts to date have focused on overcoming the limitation of short reads for genome-wide analysis, but unfortunately, current available bioinformatics ability and computing power is lagging far behind the needs for NGS sequencing data analysis. In genomics, reference-based assembly is often performed to map the number of short reads to a human reference genome which creates challenges for the algorithms and computing of alignment. Since repetitive sequences are widely distributed across the entire human genome, some short reads will align equally to multiple chromosomal locations. This is one of the reasons multiple-fold coverage of a given region is required for NGS and why further resequencing with Sanger methodology is often needed to ascertain the genetic variant detected in short reads.

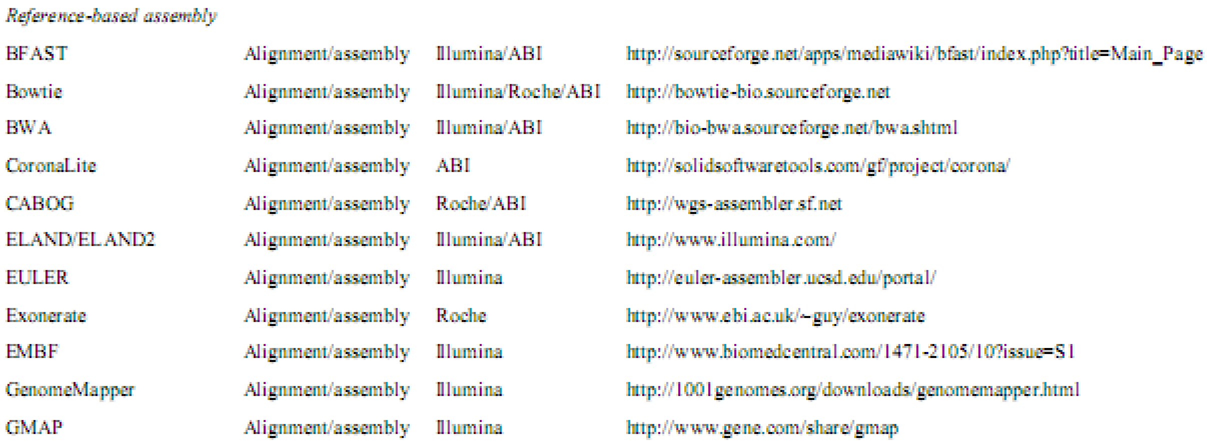
The most important step in NGS data analysis is successful alignment or assembly of short reads to a reference genome. It is a challenge to efficiently align short reads to a reference genome, especially when developing new algorithms to handle ambiguities or lack of accuracy during the alignment. Based on the mapping quality concept, MAQ (Mapping and Assembly with Quality), a very popular NGS software program, was developed that can efficiently map short reads to a reference genome and derive genotype calls to the consensus sequence with quality scores. MAQ is one of the first reference guided assembly programs. It is accurate, efficient, versatile, and user-friendly, and has been successfully applied to several NGS projects. ELAND (Efficient Large-Scale Alignment of Nucleotide Databases), another NGS program designed to search DNA files for short DNA reads allowing up to 2 errors per match, has also been successfully used in several NGS projects. Benchmarks comparing ELAND with other popular NGS software, such as MAQ, BLAST (Basic Local Alignment Search Tool), SOAP (Short Oligonucleotide Alignment Program), and SeqMap etc. (Table 4), generally place ELAND as one of the fastest available programs. Compared to reference-based assembly with very short-read length sequences, de novo assembly is even more challenging. Currently de novo assembly with NGS data is generally limited to microbial genome projects (Metagenomics) due to the small bacterial genome size. The primary goal of current algorithms and computing for short-read assembly with NGS technologies is to increase read length. This goal will likely be achieved by the development of single-molecule sequencing technologies. Certain improvements in existing NGS technologies, such as mate-paired short reads, may also make this goal attainable. Individual human genomes (one Asian and one African) have been successfully sequenced and assembled using the Illumina Genome Analyzer (read lengths ranged from 35 to 75 basepairs) with a modified SOAP program, SOAP denovo. Available bioinformatic tools for short-read alignment, de novo and reference-based assembly for NGS are listed in Table 4. Since many of the programs are open source, additional programming may be needed to modify the program to the needs of a specific NGS project. Some online utility programs, such as EagleView or LookSeq also provide some additional assistance on NGS data analysis and interpretation (Table 4). In bioinformatics, sequence assembly refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original order of the sequence. Genome assembly is simply putting sequenced fragments of DNA into their correct chromosomal positions. This is needed as DNA sequencing technology cannot read whole genomes in one go, but rather reads small pieces of between 20 and 1000 bases, depending on the technology used. Typically the short fragments, called reads, result from sequencing genomic DNA, or gene transcript (ESTs).

The first sequence assemblers began to appear in the late 1980s and early 1990s as variants of simpler sequence alignment programs to piece together vast quantities of fragments generated by automated sequencing instruments called DNA sequencers. As the sequenced organisms grew in size and complexity (from small viruses over plasmids to bacteria and finally eukaryotes), the assembly programs used in these genome projects needed to increasingly employ more and more sophisticated strategies to handle: Terabytes of sequencing data which need processing on computing clusters; Identical and nearly identical sequences (known as *repeats*) which can, in the worst case, increase the time and space complexity of algorithms exponentially; Errors in the fragments from the sequencing instruments, which can confound assembly.

Table 4. Alignment, Assembly and Utility Bioinformatics tools

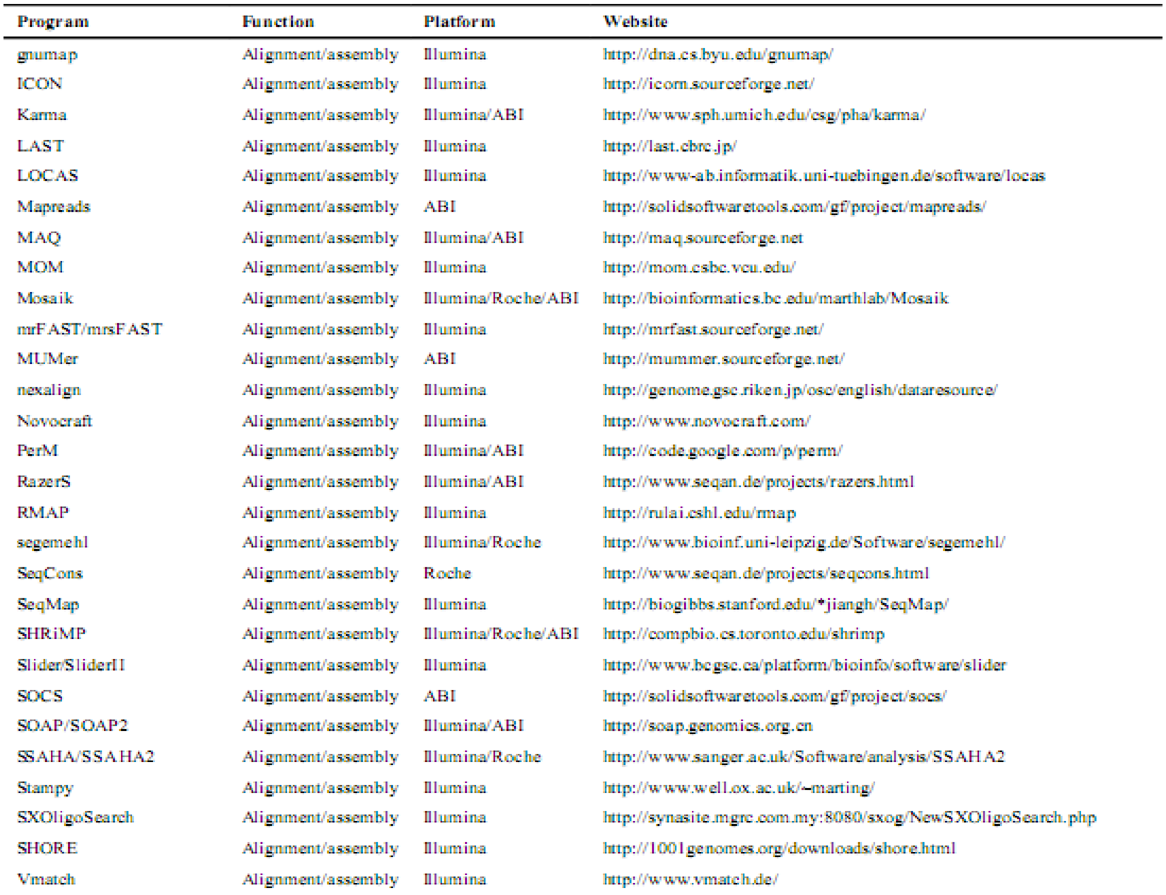


Program	Function	Platform	Website
De novo assembly			
Abyss	Alignment/assembly	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATHS	Alignment/assembly	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
AMOScmp	Alignment/assembly	Roche	http://sourceforge.net/projects/amos/files/
ARACHNE	Alignment/assembly	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
CAP3	Alignment/assembly	Roche	http://pbil.univ-lyon1.fr/cap3.php
consensus/Seq-Cons	Alignment/assembly	Roche	http://www.seqan.de/downloads/projects.html
Curtain	Alignment/assembly	Illumina/Roche/ABI	http://code.google.com/p/curtain/
Edena	Alignment/assembly	Illumina	http://www.genomic.ch/edena
Euler-SR	Alignment/assembly	Illumina/Roche	http://euler-assembler.ucsd.edu/portal/?q=team
FuzzyPath	Alignment/assembly	Illumina/Roche	ftp://ftp.sanger.ac.uk/pub/zn1/fuzzypath/fuzzypath_v3.0.tgz
IDBA	Alignment/assembly	Illumina	http://www.cs.hku.hk/~alse/idba/
MIRA/MIRA3	Alignment/assembly	Illumina/Roche	http://chevreux.org/projects_mira.html
Newbler	Alignment/assembly	Roche	roche-applied-science.com/
Phrap	Alignment/assembly	Illumina/Roche	http://www.phrap.org/consed/consed.html#howToGet
RGA	Alignment/assembly	Illumina	http://rga.cgrb.oregonstate.edu/
QSRA	Alignment/assembly	Illumina	http://qsra.cgrb.oregonstate.edu/
SHARCGS	Alignment/assembly	Illumina	http://sharcgs.molgen.mpg.de/
SHORTY	Alignment/assembly	ABI	http://www.cs.sunysb.edu/~skiena/shorty/
SHRAP	Alignment/assembly	Roche	By request
SOAPdenovo	Alignment/assembly	Illumina	http://soap.genomics.org.cn
SOPRA	Alignment/assembly	Illumina/ABI	http://www.physics.rutgers.edu/%7Eanirvans/SOPRA/
SR-ASM	Alignment/assembly	Roche	http://bioserver.cs.put.poznan.pl/sr-asm-short-reads-assembly-algorithm
SSAKE	Alignment/assembly	Illumina/Roche	http://www.bcgsc.ca/platform/bioinfo/software/ssake
Taipan	Alignment/assembly	Illumina	http://sourceforge.net/projects/taipan/files/
VCAKE	Alignment/assembly	Illumina/Roche	http://sourceforge.net/projects/vcake
Velvet	Alignment/assembly	Illumina/Roche/ABI	http://www.ebi.ac.uk/%7Ezerbino/velvet

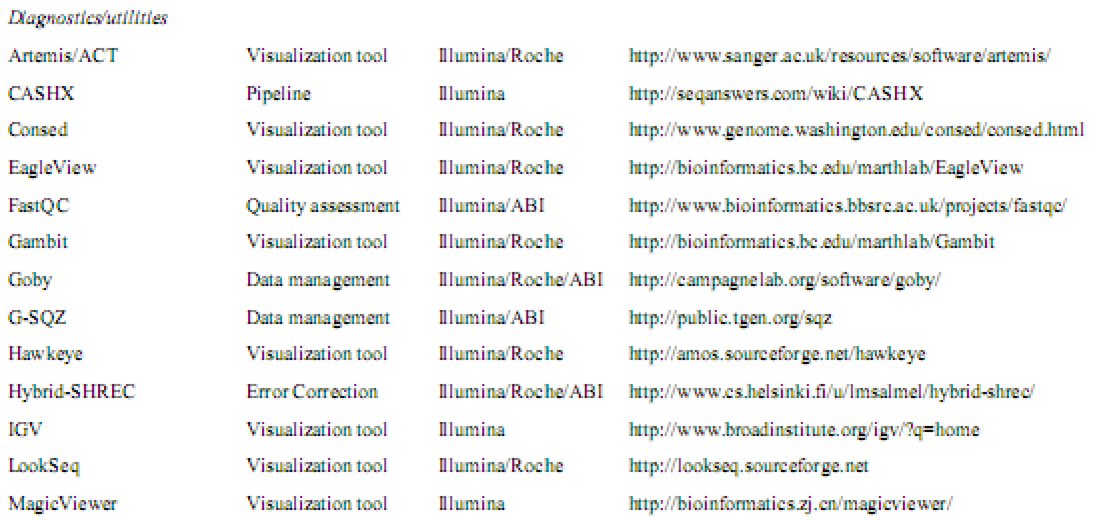


Reference-based assembly

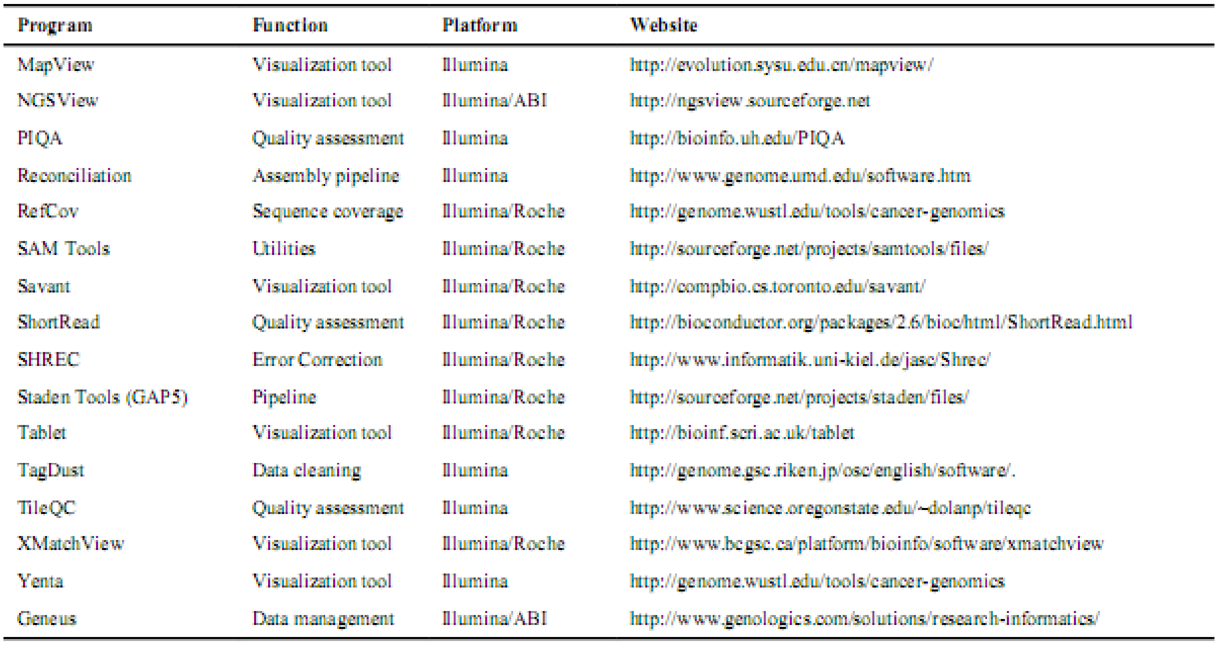
BFAST	Alignment/assembly	Illumina/ABI	http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main_Page
Bowtie	Alignment/assembly	Illumina/Roche/ABI	http://bowtie-bio.sourceforge.net
BWA	Alignment/assembly	Illumina/ABI	http://bio-bwa.sourceforge.net/bwa.shtml
CoronaLite	Alignment/assembly	ABI	http://solidsoftwaretools.com/gf/project/corona/
CABOG	Alignment/assembly	Roche/ABI	http://wgs-assembler.sf.net
ELAND/ELAND2	Alignment/assembly	Illumina/ABI	http://www.illumina.com/
EULER	Alignment/assembly	Illumina	http://euler-assembler.ucsd.edu/portal/
Exonerate	Alignment/assembly	Roche	http://www.ebi.ac.uk/~guy/exonerate
EMBF	Alignment/assembly	Illumina	http://www.biomedcentral.com/1471-2105/10?issue=S1
GenomeMapper	Alignment/assembly	Illumina	http://1001genomes.org/downloads/genomemapper.html
GMAP	Alignment/assembly	Illumina	http://www.gene.com/share/gmap



Program	Function	Platform	Website
gnumap	Alignment/assembly	Illumina	http://dna.cs.byu.edu/gnumap/
ICON	Alignment/assembly	Illumina	http://icorn.sourceforge.net/
Karma	Alignment/assembly	Illumina/ABI	http://www.sph.umich.edu/csg/pha/karma/
LAST	Alignment/assembly	Illumina	http://last.cbrc.jp/
LOCAS	Alignment/assembly	Illumina	http://www-ab.informatik.uni-tuebingen.de/software/locas
Mapreads	Alignment/assembly	ABI	http://solidsoftwaretools.com/gf/project/mapreads/
MAQ	Alignment/assembly	Illumina/ABI	http://maq.sourceforge.net
MOM	Alignment/assembly	Illumina	http://mom.csbc.vcu.edu/
Mosaik	Alignment/assembly	Illumina/Roche/ABI	http://bioinformatics.bc.edu/marthlab/Mosaik
mrFAST/mrsFAST	Alignment/assembly	Illumina	http://mrfast.sourceforge.net/
MUMer	Alignment/assembly	ABI	http://mummer.sourceforge.net/
nexalign	Alignment/assembly	Illumina	http://genome.gsc.riken.jp/osc/english/dataresource/
Novocraft	Alignment/assembly	Illumina	http://www.novocraft.com/
PerM	Alignment/assembly	Illumina/ABI	http://code.google.com/p/perm/
RazerS	Alignment/assembly	Illumina/ABI	http://www.seqan.de/projects/razers.html
RMAP	Alignment/assembly	Illumina	http://rulai.cshl.edu/rmap
segemehl	Alignment/assembly	Illumina/Roche	http://www.bioinf.uni-leipzig.de/Software/segemehl/
SeqCons	Alignment/assembly	Roche	http://www.seqan.de/projects/seqcons.html
SeqMap	Alignment/assembly	Illumina	http://biogibbs.stanford.edu/*jiangh/SeqMap/
SHRiMP	Alignment/assembly	Illumina/Roche/ABI	http://compbio.cs.toronto.edu/shrimp
Slider/SliderII	Alignment/assembly	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slider
SOCS	Alignment/assembly	ABI	http://solidsoftwaretools.com/gf/project/socs/
SOAP/SOAP2	Alignment/assembly	Illumina/ABI	http://soap.genomics.org.cn
SSAHA/SSAHA2	Alignment/assembly	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/SSAHA2
Stampy	Alignment/assembly	Illumina	http://www.well.ox.ac.uk/~marting/
SXOligoSearch	Alignment/assembly	Illumina	http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
SHORE	Alignment/assembly	Illumina	http://1001genomes.org/downloads/shore.html
Vmatch	Alignment/assembly	Illumina	http://www.vmatch.de/



Diagnostics/utilities			
Artemis/ACT	Visualization tool	Illumina/Roche	http://www.sanger.ac.uk/resources/software/artemis/
CASHX	Pipeline	Illumina	http://seqanswers.com/wiki/CASHX
Consed	Visualization tool	Illumina/Roche	http://www.genome.washington.edu/consed/consed.html
EagleView	Visualization tool	Illumina/Roche	http://bioinformatics.bc.edu/marthlab/EagleView
FastQC	Quality assessment	Illumina/ABI	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
Gambit	Visualization tool	Illumina/Roche	http://bioinformatics.bc.edu/marthlab/Gambit
Goby	Data management	Illumina/Roche/ABI	http://campagnelab.org/software/goby/
G-SQZ	Data management	Illumina/ABI	http://public.tgen.org/sqz
Hawkeye	Visualization tool	Illumina/Roche	http://amos.sourceforge.net/hawkeye
Hybrid-SHREC	Error Correction	Illumina/Roche/ABI	http://www.cs.helsinki.fi/u/lmsalmel/hybrid-shrec/
IGV	Visualization tool	Illumina	http://www.broadinstitute.org/igv/?q=home
LookSeq	Visualization tool	Illumina/Roche	http://lookseq.sourceforge.net
MagicViewer	Visualization tool	Illumina	http://bioinformatics.zj.cn/magicviewer/

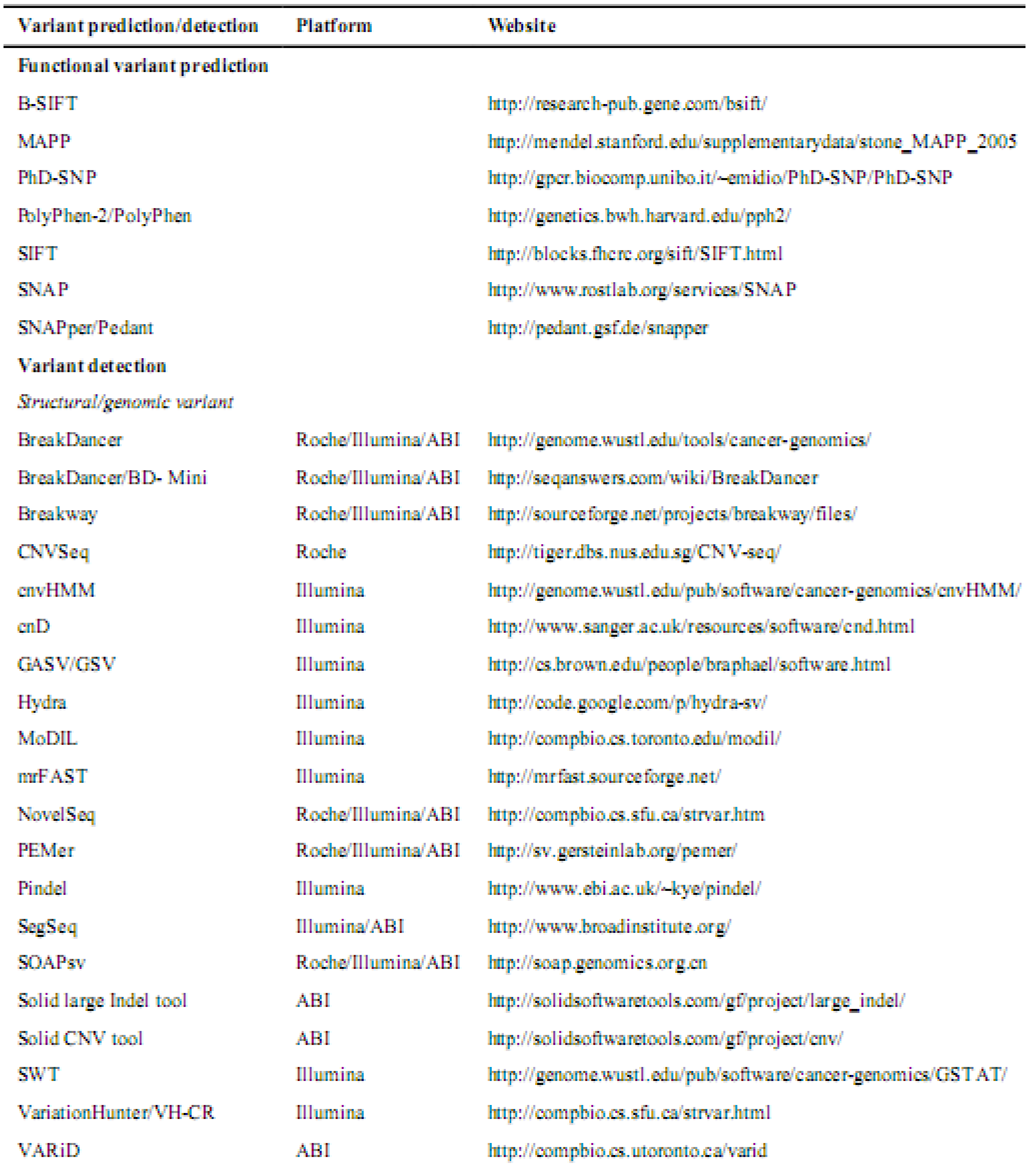


Program	Function	Platform	Website
MapView	Visualization tool	Illumina	http://evolution.sysu.edu.cn/mapview/
NGSView	Visualization tool	Illumina/ABI	http://ngsview.sourceforge.net
PIQA	Quality assessment	Illumina	http://bioinfo.uh.edu/PIQA
Reconciliation	Assembly pipeline	Illumina	http://www.genome.umd.edu/software.htm
RefCov	Sequence coverage	Illumina/Roche	http://genome.wustl.edu/tools/cancer-genomics
SAM Tools	Utilities	Illumina/Roche	http://sourceforge.net/projects/samtools/files/
Savant	Visualization tool	Illumina/Roche	http://compbio.cs.toronto.edu/savant/
ShortRead	Quality assessment	Illumina/Roche	http://bioconductor.org/packages/2.6/bioc/html/ShortRead.html
SHREC	Error Correction	Illumina/Roche	http://www.informatik.uni-kiel.de/jasc/Shrec/
Staden Tools (GAP5)	Pipeline	Illumina/Roche	http://sourceforge.net/projects/staden/files/
Tablet	Visualization tool	Illumina/Roche	http://bioinf.scri.ac.uk/tablet
TagDust	Data cleaning	Illumina	http://genome.gsc.riken.jp/osc/english/software/.
TileQC	Quality assessment	Illumina	http://www.science.oregonstate.edu/~dolanp/tileqc
XMatchView	Visualization tool	Illumina/Roche	http://www.bcgsc.ca/platform/bioinfo/software/xmatchview
Yenta	Visualization tool	Illumina	http://genome.wustl.edu/tools/cancer-genomics
Geneus	Data management	Illumina/ABI	http://www.genologics.com/solutions/research-informatics/

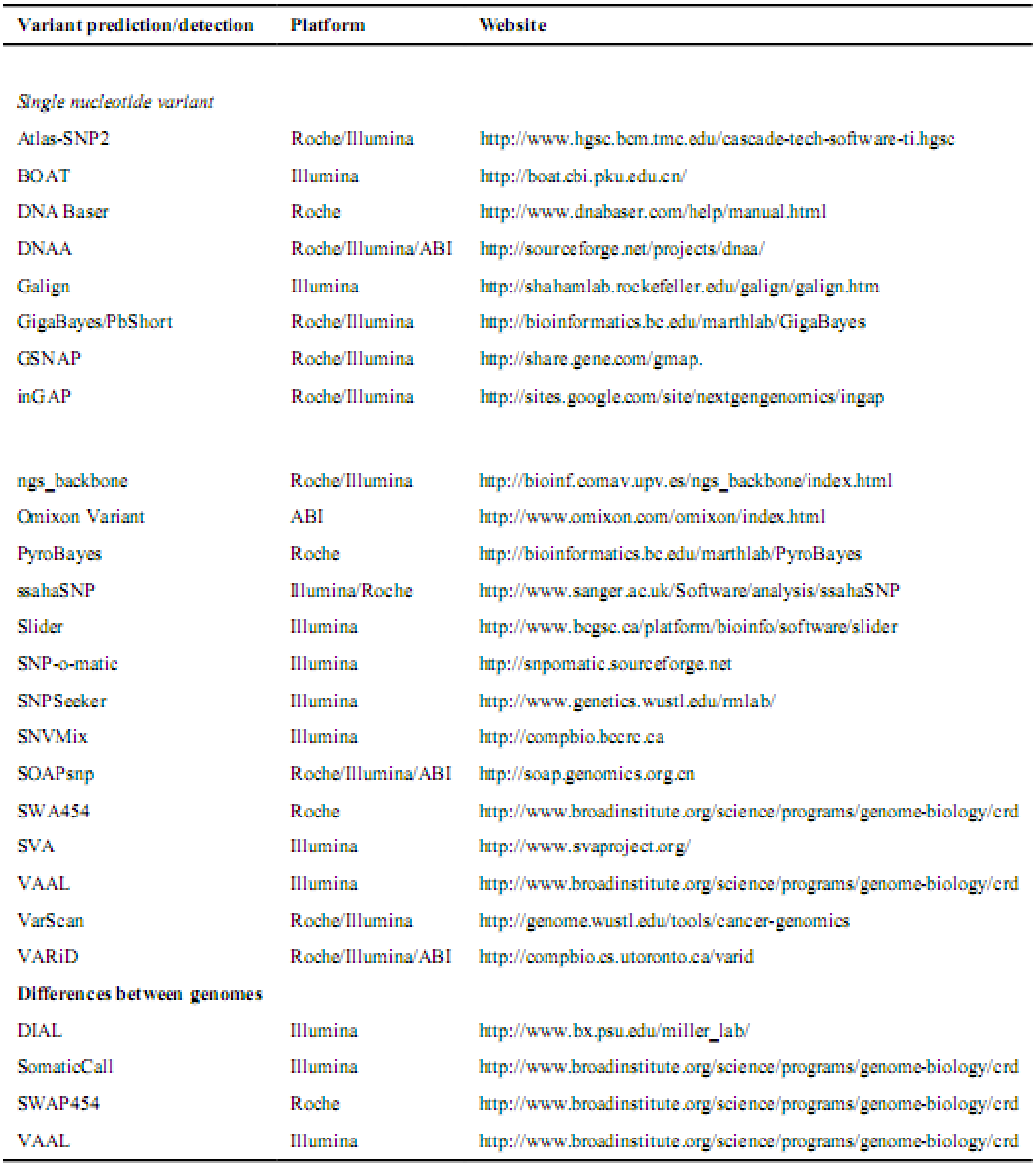
Annotation and functional prediction

After successful alignment and assembly of NGS data, the next challenge is to interpret the large number of apparently novel genetic variants (or mutations) present by chance in any single human genome, making it difficult to identify which variants are causal, even when considering only non-synonymous variants. Many novel genetic variants/variations have been discovered for each sequenced genome, resulting in approximately 400 function- altering variants for protein-coding sequences per individual genome. Recognition of functional variants is at the center of the NGS data analysis and bioinformatics. It is challenging to develop software with the ability to distinguish low-frequency alleles descendent from ancient ancestors from de novo or extremely rare mutations recently introduced into the population. Available bioinformatics tools for annotation and functional prediction of NGS data are listed in Table 5. SIFT (Sorting Intolerant From Tolerant) is used to predict whether an amino acid substitution affects protein function based on sequence homology and the physical properties of the amino acid can be applied to find non-synonymous polymorphisms within NGS data. By considering the physiochemical variations presented in protein sequence alignment and the property of variations, Multivariate Analysis of Protein Polymorphism (MAPP) can predict the impact of all possible amino acid substitutions on the function of the protein. Predictor of human Deleterious Single Nucleotide Polymorphisms (PhD-SNP) is an optimized program to predict if a given single point protein mutation can be classified as disease-related or as neutral polymorphism based on protein sequence and profile information. Polymorphism Phenotyping (PolyPhen) and updated PolyPhen-2 are tools which predict the possible impact of an amino acid substitution on the structure and function of a specific protein using straightforward comparative physical methods. Variation detection software, which includes screening genomes for structural and single nucleotide variants and the differences between genomes are generally integrated with alignment and assembly processes and are listed in Table 5.

Table 5. Genetic Variant Prediction and detection bioinformatics programs for NGS data analysis



Variant prediction/detection	Platform	Website
Functional variant prediction		
B-SIFT		http://research-pub.gene.com/bsift/
MAPP		http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005
PhD-SNP		http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP
PolyPhen-2/PolyPhen		http://genetics.bwh.harvard.edu/pph2/
SIFT		http://blocks.fhcrc.org/sift/SIFT.html
SNAP		http://www.rostlab.org/services/SNAP
SNAPper/Pedant		http://pedant.gsf.de/snapper
Variant detection		
Structural/genomic variant		
BreakDancer	Roche/Illumina/ABI	http://genome.wustl.edu/tools/cancer-genomics/
BreakDancer/BD- Mini	Roche/Illumina/ABI	http://seqanswers.com/wiki/BreakDancer
Breakway	Roche/Illumina/ABI	http://sourceforge.net/projects/breakway/files/
CNVSeq	Roche	http://tiger.dbs.nus.edu.sg/CNV-seq/
cnvHMM	Illumina	http://genome.wustl.edu/pub/software/cancer-genomics/cnvHMM/
cnD	Illumina	http://www.sanger.ac.uk/resources/software/cnd.html
GASV/GSV	Illumina	http://cs.brown.edu/people/braphael/software.html
Hydra	Illumina	http://code.google.com/p/hydra-sv/
MoDIL	Illumina	http://compbio.cs.toronto.edu/modil/
mrFAST	Illumina	http://mrfast.sourceforge.net/
NovelSeq	Roche/Illumina/ABI	http://compbio.cs.sfu.ca/strvar.htm
PEMer	Roche/Illumina/ABI	http://sv.gersteinlab.org/pemer/
Pindel	Illumina	http://www.ebi.ac.uk/~kye/pindel/
SegSeq	Illumina/ABI	http://www.broadinstitute.org/
SOAPsv	Roche/Illumina/ABI	http://soap.genomics.org.cn
Solid large Indel tool	ABI	http://solidsoftwaretools.com/gf/project/large_indel/
Solid CNV tool	ABI	http://solidsoftwaretools.com/gf/project/cnv/
SWT	Illumina	http://genome.wustl.edu/pub/software/cancer-genomics/GSTAT/
VariationHunter/VH-CR	Illumina	http://compbio.cs.sfu.ca/strvar.html
VARiD	ABI	http://compbio.cs.utoronto.ca/varid

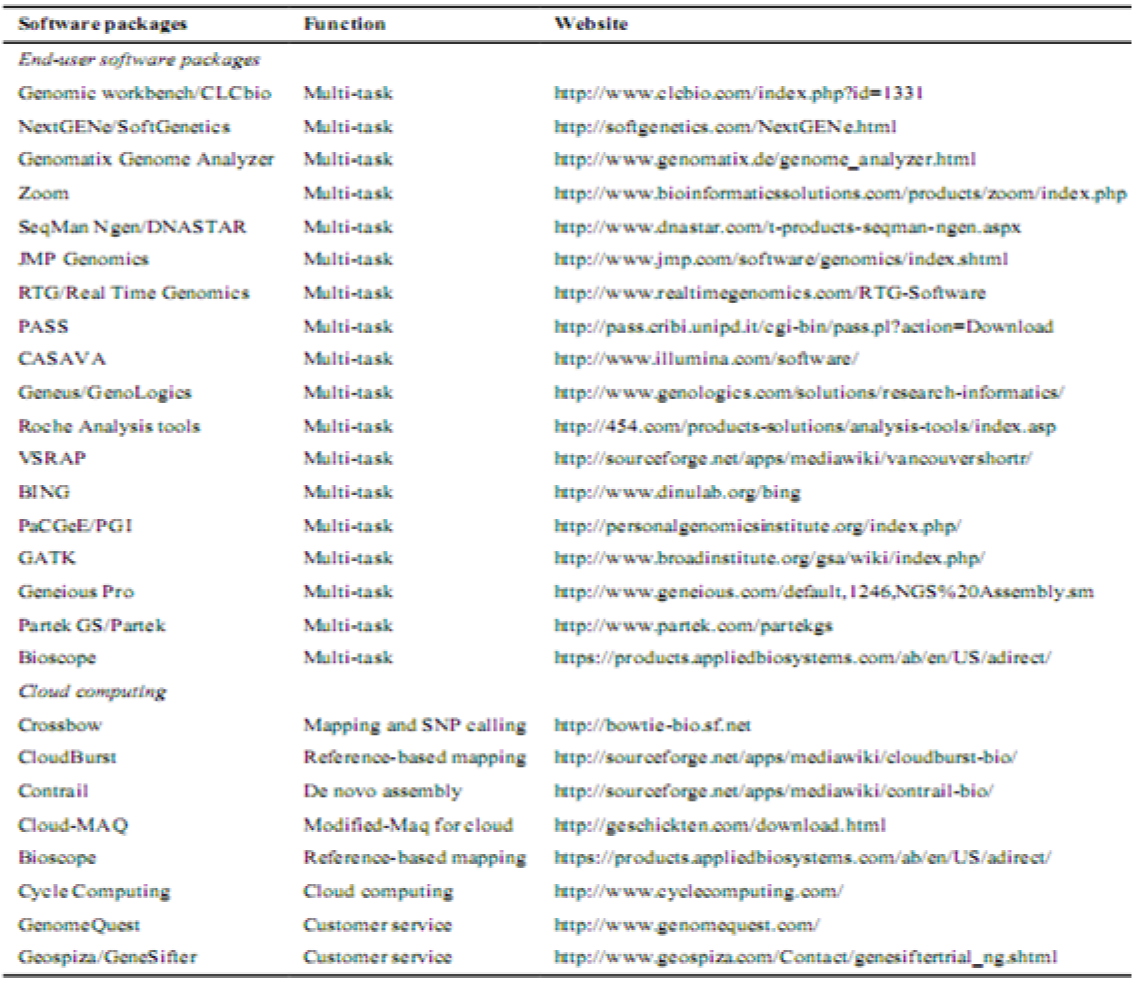


Variant prediction/detection	Platform	Website
Single nucleotide variant		
Atlas-SNP2	Roche/Illumina	http://www.hgsc.bcm.tmc.edu/cascade-tech-software-ti.hgsc
BOAT	Illumina	http://boat.cbi.pku.edu.cn/
DNA Baser	Roche	http://www.dnabaser.com/help/manual.html
DNAA	Roche/Illumina/ABI	http://sourceforge.net/projects/dnaa/
Galign	Illumina	http://shahamlab.rockefeller.edu/galign/galign.htm
GigaBayes/PbShort	Roche/Illumina	http://bioinformatics.bc.edu/marthlab/GigaBayes
GSNAP	Roche/Illumina	http://share.gene.com/gmap.
inGAP	Roche/Illumina	http://sites.google.com/site/nextgengenomics/ingap
ngs_backbone	Roche/Illumina	http://bioinf.comav.upv.es/ngs_backbone/index.html
Omixon Variant	ABI	http://www.omixon.com/omixon/index.html
PyroBayes	Roche	http://bioinformatics.bc.edu/marthlab/PyroBayes
ssahaSNP	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/ssahaSNP
Slider	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slider
SNP-o-matic	Illumina	http://snpomatic.sourceforge.net
SNPSeeker	Illumina	http://www.genetics.wustl.edu/rmlab/
SNVMix	Illumina	http://compbio.bccrc.ca
SOAPsnp	Roche/Illumina/ABI	http://soap.genomics.org.cn
SWA454	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
SVA	Illumina	http://www.svaproject.org/
VAAL	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
VarScan	Roche/Illumina	http://genome.wustl.edu/tools/cancer-genomics
VARiD	Roche/Illumina/ABI	http://compbio.cs.utoronto.ca/varid
Differences between genomes		
DIAL	Illumina	http://www.bx.psu.edu/miller_lab/
SomaticCall	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
SWAP454	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
VAAL	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd

End-user packages

End-user software packages which provide a user-friendly interface, easy to use data input and output formats, and integrates multiple computing programs into one software package, may be the best solution for most biomedical researchers. Based on our experience, among available end-user packages, Genomic Workbench from CLC Bio appears to be the most widely used. NextGENe from SoftGenetics is excellent for candidate-gene resequencing projects, but it cannot handle very large datasets and may not be suitable for large genome sequencing projects. SeqMan Ngen from DNASTAR is under development but currently unavailable. Although commercial end-user packages tend to carry a hefty price, some are available free online as detailed in Table 6.

Table 6. End-User software packages and cloud computing software for NGS data analysis



Software packages	Function	Website
End-user software packages		
Genomic workbench/CLCbio	Multi-task	http://www.clcbio.com/index.php?id=1331
NextGENe/SoftGenetics	Multi-task	http://softgenetics.com/NextGENe.html
Genomatix Genome Analyzer	Multi-task	http://www.genomatix.de/genome_analyzer.html
Zoom	Multi-task	http://www.bioinformaticssolutions.com/products/zoom/index.php
SeqMan Ngen/DNASTAR	Multi-task	http://www.dnastar.com/t-products-seqman-ngen.aspx
JMP Genomics	Multi-task	http://www.jmp.com/software/genomics/index.shtml
RTG/Real Time Genomics	Multi-task	http://www.realtimegenomics.com/RTG-Software
PASS	Multi-task	http://pass.cribi.unipd.it/cgi-bin/pass.pl?action=Download
CASAVA	Multi-task	http://www.illumina.com/software/
Geneus/GenoLogics	Multi-task	http://www.genologics.com/solutions/research-informatics/
Roche Analysis tools	Multi-task	http://454.com/products-solutions/analysis-tools/index.asp
VSRAP	Multi-task	http://sourceforge.net/apps/mediawiki/vancouvershortr/
BING	Multi-task	http://www.dinulab.org/bing
PaCGeE/PGI	Multi-task	http://personalgenomicsinstitute.org/index.php/
GATK	Multi-task	http://www.broadinstitute.org/gsa/wiki/index.php/
Geneious Pro	Multi-task	http://www.geneious.com/default,1246,NGS%20Assembly.sm
Partek GS/Partek	Multi-task	http://www.partek.com/partekgs
Bioscope	Multi-task	https://products.appliedbiosystems.com/ab/en/US/adirect/
Cloud computing		
Crossbow	Mapping and SNP calling	http://bowtie-bio.sf.net
CloudBurst	Reference-based mapping	http://sourceforge.net/apps/mediawiki/cloudburst-bio/
Contrail	De novo assembly	http://sourceforge.net/apps/mediawiki/contrail-bio/
Cloud-MAQ	Modified-Maq for cloud	http://geschickten.com/download.html
Bioscope	Reference-based mapping	https://products.appliedbiosystems.com/ab/en/US/adirect/
Cycle Computing	Cloud computing	http://www.cyclecomputing.com/
GenomeQuest	Customer service	http://www.genomequest.com/
Geospiza/GeneSifter	Customer service	http://www.geospiza.com/Contact/genesiftertrial_ng.shtml

As previously mentioned, the fact that high-performance computing and intensive bioinformatic support is needed for NGS, it is difficult for many research laboratories to successfully conduct NGS projects due to the high level of information technology support required. A possible solution is cloud computing. In cloud computing, a user can use a virtual operating system (or “cloud”) to process data on a computer cluster for high parallel tasks. CrossBow is the first cloud computing software capable of performing alignment and single nucleotide polymorphism analysis on multiple whole- human datasets. CloudBurst is another new parallel read-mapping cloud algorithm optimized for mapping NGS data to a human reference genome, SNP discovery, genotyping and personal genomics. Data generated on Applied Bio-systems' SoLiD platform uses a two colored system which makes it unsuitable for analysis by many available software packages. The Bioscope package, developed by ABI, is devoted to their SoLiD data and can be used as a single software package or for cloud computing, likewise, CASAVA package developed by Illumina is utilized for Genome Analyzer data. Available NGS cloud computing technologies are listed in Table 6. GenomeQuest, Complete Genomics and Geospiza/GeneSifter provide online customer oriented NGS data analysis services, which is a little different from cloud computing by definition.

Conclusion

NGS technologies will certainly enable us to identify all the causative variants including “rare variants” within individual human subjects. It is anticipated that whole- genome sequencing (or exome sequencing) will make significant contributions to our understanding of the genetic etiologies that contribute to complex human disease, as well as the genetic basis of genomics.

Further reading

1. *Ansorge WJ. Next-generation DNA sequencing techniques. Nat. Biotechnol. 2009; 25:195–203.*
2. *Bau S, Schracke N, Kranzle M, Wu H, Stahler PF, Hoheisel JD, Beier M, Summerer D. Targeted next- generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Anal. Bioanal. Chem. 2009; 393:171–175. [PubMed: 18958448]*
3. *Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res. 2009; 19:336–346. [PubMed: 19056694]*
4. *Chistoserdova L. Recent progress and new challenges in metagenomics for biotechnology. Biotechnol. Lett. 2010; 32:1351–1359. [PubMed: 20495950]*
5. *Daly AK. Genome-wide association studies in pharmacogenomics. Nat. Rev. Genet. 2010a; 11:241– 246. [PubMed: 20300088]*
6. *Daly AK. Pharmacogenetics and human genetic polymorphisms. Biochem. J. 2010b; 429:435–449. [PubMed: 20626352]*
7. *Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat. Methods. 2009; 6:S6–S12. [PubMed: 19844229]*
8. *Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. 2009; 10:241–251. [PubMed: 19293820]*

Chapter 2

NGS Data Formats & Quality control (*QC*)

Rosana O. Babu

Introduction- NGS Data Formats

Sequence Sources & Read Data Formats

NGS data sets are very large, but as the cost per MB is now so low we expect to see them generated for many large-scale experiments where the mapping and analysis of the short read sets will be the key to interpretation of results.

For further analysis of NGS assembled data, whether a de novo assembly or using a set of NGS reads mapped to a reference sequence. Our initial aim is to rapidly produce a set of utility functions that can be used to read and interpret NGS assemblies, for example MAQ .map files, and to calculate statistics based on position, gene an

notations, etc. These will then be used to build new applications and to demonstrate to end users, and to other developers, the ease of programming and the added value of the many interfaces already available.

NGS data volumes require new storage structures. As well as reading MAQ .map files and other formats for 454 sequencing and for new instruments) either by reading their formats or by using existing utilities to convert them. The instrument manufacturers show encouraging signs of using open data format and we will keep a close watch on emerging formats for a brief understanding

(i) NGS Data Sources

- NCBI
- EMBL
- DDBJ
- MINSEQE Minimum Information about a high-throughput Sequencing Experiment

(ii) Sequences

Sequences can be read and written in a variety of formats. These can be very confusing for users, but EMBOSS aims to make life easier by automatically recognizing the sequence format on input. That means that if you are converting from using another sequencing package to EMBOSS and you have your existing sequences in a format that is specific for that package, for example GCG format, you will have no problem reading them in. If you don't hold your sequence in a recognized standard format, you will not be able to analyze your sequence easily. Sequencers; read aligners; genome assemblers; scientific results from sequencing data; databases of reads; databases of genomes, variation, or other results. When all of these output data and require inputs in their own formats, we can be sure that there will be enormous amount of annoying, boring, and disappointing work while dealing with sequencing data.

(iii) What is sequence format?

Sequence formats are ASCII TEXT. They are the required arrangement of characters, symbols and keywords that specify what things such as the sequence, ID name, comments, etc. look like in the sequence entry and where in the entry the program should look to find them. There are generally no hidden, unprintable 'control' characters in any sequence format (there are none in those that EMBOSS supports). All standard sequence formats can be printed out or viewed simply by displaying their file.

(iv) Why so many formats?

There are at least a couple of dozen sequence formats in existence at the moment. Some are much more common than others. Formats were designed so as to be able to hold the sequence data and other information about the sequence. Nearly every sequence analysis package written since programs were first used to read and write sequences has invented its own format. Nearly every collection of sequences that dares call itself a database has stored its data in its own format. Most formats allow you to hold other description, annotation and comments like Fasta format. Other formats have specific fields for holding information such as references, keywords, associated entries in other databases and feature tables.

(v) Identification

Most sequence formats include at least one form of ID name, usually placed somewhere at the top of the sequence format. The simple format **FASTA** has the ID name as the first word on its title line. For example the ID name 'xyz':

```
>xyz
ttcctctttctcgactccatcttcgcggtagctgggaccgccgttcagtcgccaatatgcagctctttgtccgcgcccaggagctaca
caccttcgaggtgaccggccaggaaacggtcgcccagatcaaggctcatgtagcctcactggagggcatt
```

(vi) Sequence ID’s and Accessions

An entry in a database must have some way of being uniquely identified in that database. Most sequence databases have two such identifiers for each sequence - an ID name and an Accession number.

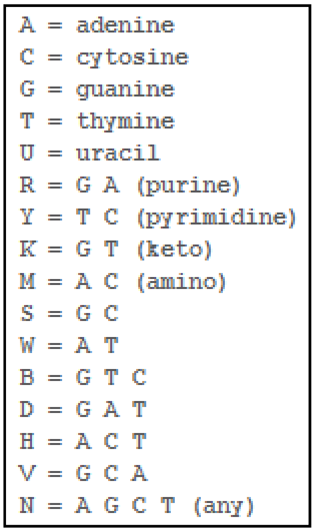
Why are there two such identifiers? The ID name was originally intended to be a human-readable name that had some indication of the function of its sequence. In EMBL and GenBank the first two (or three) letters indicated the species and the rest indicated the function, for example 'hsfau' is the 'Homo Sapiens FAU pseudogene'. This naming scheme started to be a problem when the number of entries added each day was so vast that people could not make up the ID names fast enough. Instead, the Accession numbers were used as the ID name. Therefore you will now find ID names like 'AF061303', the same as the Accession number for that sequence in EMBL. ID names are not guaranteed to remain the same between different versions of a database (although in practice they usually do).

Accession numbers are unique alphanumeric identifiers that are guaranteed to remain with that sequence through the rest of the life of the database. If two sequences are merged into one, then the new sequence will get a new Accession number and the Accession numbers of the merged sequences will be retained as 'secondary' Accession numbers. EMBL, GenBank and SwissProt share an Accession numbering scheme - an Accession number uniquely identifies a sequence within these three databases.

(vii) The Sequence

Nucleotide (DNA or RNA) sequences are usually stored in the IUBMB standard codes (International Union of Biochemistry and Molecular Biology*)*.

Similarly, protein sequences are usually stored in the IUPAC standard one-letter codes (*International Union of Pure and Applied Chemistry*).



```
A = adenine
C = cytosine
G = guanine
T = thymine
U = uracil
R = G A (purine)
Y = T C (pyrimidine)
K = G T (keto)
M = A C (amino)
S = G C
W = A T
B = G T C
D = G A T
H = A C T
V = G C A
N = A G C T (any)
```

Sequence Database Formats

Some of the most widespread sequence formats apart from **fasta** are those used by the major sequence databases.

- EMBL
- GenBank
- SwissProt
- PIR

(i) Sequence Files

Files can hold sequences in standard recognized formats. Files can also hold sequences in non-standard unrecognizable ways. Do not save your sequences in a word-processor format file.

NGS Data Formats

To date, the following sequence formats are accepted as input for NGS tools.

(i) Read Formats (sample are given below)

- **FASTQ** format is a common format for short reads with quality scores. It is supported in EMBOSS 6.1.0 as a sequence format. Quality scores are also used if the format is more explicitly named in EMBOSS: fastqsanger or fastqillumina
- **SFF** Standard Flowgram Format, to hold the "trace" data for 454 reads
- **SRF** Sequence Read Format (also called Short Read Format), format specification has not been updated since February 2008. Applied Biosystems SRF Conversion Tool (solid2srf) converts SOLiD™ system reads into SRF format.
- **SCARF**, Solexa Compact ASCII Read Format. This format contains all information for one read in a single line. From left to right each line contains the read name, nucleotide sequence, quality scores for each position, and more information. Illumina’s pipeline can produce SCARF files with quality scores in ASCII or numeric format.
- **SCF** first version was described in 1992, since then it has undergone several important changes such as a major reorganization of the ordering of the data items in the file and also in the way they are represented

(ii) Assembly Formats

- MAQ .map format (a compressed binary file format designed for short read alignment)
- MAF, MIRA Assembly Format
- AMOS A Modular Open-Source Assembler assembly format, used by velvet

- SAM/BAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments

Sample Sequence and Read formats

Sequence formats

(i) Plain sequence format
A sequence in plain format may contain only IUPAC characters and spaces (no numbers!).

Note: A file in plain sequence format may only contain one sequence, while most other formats accept several sequences in one file.

An example sequence in plain format is

```
acaagatgccattgtcccccggcctcctgctgctgctgctctccggggccacggccaccgctgccctgcccctggagggtggccc
cggaataaggaaaagca
```

(ii) EMBL format
A sequence file in EMBL format can contain several sequences. One sequence entry starts with an identifier line ("ID"), followed by further annotation lines. The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

An example sequence in EMBL format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for preprocortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
acaagatgccattgtcccccggcctcctgctgctgctgctctccggggccacggccaccg          60
ctgccctgcccctggagggtggccccaccggccgagacagcgagcatatgcaggaagcgg         120
caggaataaggaaaagcagcctcctgactttcctcgcttggtggtttgagtggacctccc         180
aggccagtgccgggcccctcataggagaggaagctcgggaggtggccaggcggcaggaag         240
gcgcacccccccagcaatccgcgcgccgggacagaatgccctgcaggaacttcttctgga         300
agaccttctcctcctgcaaataaaacctcacccatgaatgctcacgcaagtttaattaca         360
gacctgaa                                                             368
//
```

(iii) FASTA format

A sequence file in FASTA format can contain several sequences. Each sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column.

An example sequence in FASTA format is:

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for preprocortistatin like peptide,
complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGG
CCACCGCTGCCCTGCCCCTGGAGGGTGGC
```

(iv) GCG format

A sequence file in GCG format contains exactly one sequence, begins with annotation lines and the start of the sequence is marked by a line ending with two dot ("..") characters. This line also contains the sequence identifier, the sequence length and a checksum. This format should only be used if the file was created with the GCG package.

An example sequence in GCG format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for preprocortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
AB000263  Length: 368  Check: 4514  ..
1 acaagatgccattgtcccccggcctcctgctgctgctgctctccggggccacggccaccg
61ctgccctgcccctggagggtggccccaccggccgagacagcgagcatatgcaggaagcgg
```

(v) GCG-RSF (rich sequence format)

The new GCG-RSF can contain several sequences in one file. This format should only be used if the file was created with the GCG package.

(vi) GenBank format

A sequence file in GenBank format can contain several sequences. One sequence in GenBank format starts with a line containing the word LOCUS and a number of annotation lines. The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("//").

An example sequence in GenBank format is:

```
LOCUS       AB000263                368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for preprocortistatin like peptide, complete
cds.
ACCESSION   AB000263
ORIGIN
    1  acaagatgccattgtcccccggcctcctgctgctgctgctctccggggccacggccaccg
    61 tgccctgcccctggagggtggccccaccggccgagacagcgagcatatgcaggaagcgg
    121ggaataaggaaaagcagcctcctgactttcctcgcttggtggtttgagtggacctccc
    181gccagtgccgggcccctcataggagaggaagctcgggaggtggccaggcggcaggaag
    241gcacccccccagcaatccgcgcgccgggacagaatgccctgcaggaacttcttctgga
    301accttctcctcctgcaaataaaacctcacccatgaatgctcacgcaagtttaattaca
    361cctgaa
//
```

(vii) Genomatix annotation syntax

Some Genomatix tools, e.g. Gene2Promoter or GPD allow the extraction of sequences. Genomatix uses the following syntax to annotate sequence information: each information item is denoted by a keyword, followed by a "=" and the value. These information items are separated by a pipe symbol "|".

The keywords are the following:

loc	The Genomatix Locus Id, consisting of the string "GXL_" followed by a number.
sym	The gene symbol. This can be a (comma-separated) list.
geneid	The NCBI Gene Id. This can be a (comma-separated) list.
acc	A unique identifier for the sequence. E.g. for Genomatix promoter regions, the Genomatix Promoter Id is listed in this field.
taxid	The organism's Taxon Id
spec	The organism name
chr	The chromosome within the organism.
ctg	The NCBI contig within the chromosome.
str	Strand, (+) for sense, (-) for antisense strand.
start	Start position of the sequence (relative to the contig).
end	End position of the sequence (relative to the contig).
len	Length of the sequence in basepairs.
tss	A (comma-separated list of) UTR-start/TSS position(s). If there are several TSS/UTR-starts, this means that several transcripts share the same promoter (e.g. when they are splice variants). The positions are relative to the promoter region.
probe	A (comma-separated list of) Affymetrix Probe Id(s).
unigene	A (comma-separated list of) UniGene Cluster Id(s).
homgroup	An identifier (a number) for the homology group (available for promoter sequences

	only). Orthologously related sequences have the same value in this field.
promset	If the sequence is a promoter region, the promoter set is denoted here.
descr	The gene description. If several genes (i.e. NCBI gene ids) are associated with the sequence, the descriptions for all of the genes are note, separated by ";"
comm	A comment field, used for additional annotation. For promoter sequences, this field contains information about the transcripts associated with the promoter. For each transcript the Genomatix Transcript Id, accession number, TSS position and quality is listed, separated by "/". For GenomatixCompGen promoters no transcripts are assigned, in this case the string "CompGen promoter" is denoted.

(viii) IG format

A sequence file in IG format can contain several sequences, each consisting of a number of comment lines that must begin with a semicolon (";"), a line with the sequence name (it may not contain spaces!) and the sequence itself terminated with the termination character '1' for linear or '2' for circular sequences.

An example sequence in IG format is:

```
; comment
; comment
AB000263
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGG
CCACCGCTGCCCTGCC
```

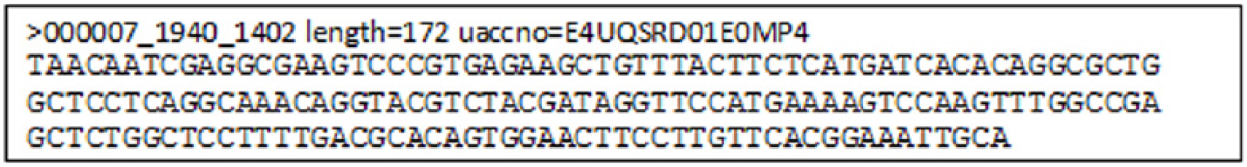
Read Formats

Sequencers use their own output formats, and even multiple different formats per technology.

(i) Output formats from Roche 454 sequencer

.sff is 454’s binary (=compressed) format, including the raw data. Not the very raw images, but flow values (flowgram), base calls, base-call qualities, read trimming and quality. There are many simple tools that “unzip” .sff into some readable textual format (see the black example).

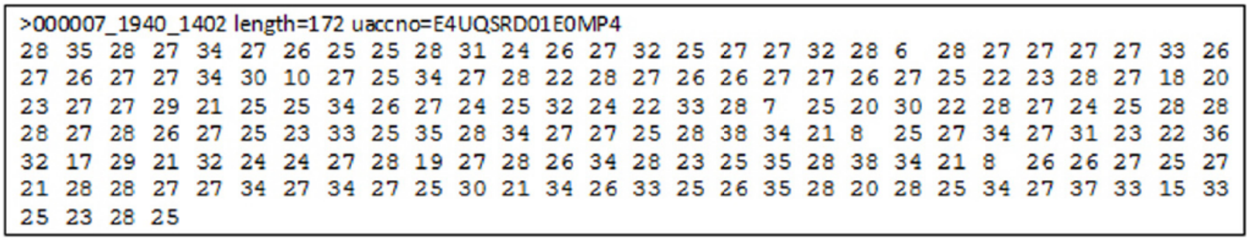
.fna is a FASTA (or FASTA-like) format including the base-called reads.



```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
TAACAATCGAGGCGAAGTCCCGTGAGAAGCTGTTTACTTCTCATGATCACACAGGCGCTG
GCTCCTCAGGCAAACAGGTACGTCTACGATAGGTTCCATGAAAAGTCCAAGTTTGGCCGA
GCTCTGGCTCCTTTTGACGCACAGTGGAACTTCCTTGTTCACGGAAATTGCA
```

Read with .fna format.

.qual includes qualities corresponding to bases in .fna.



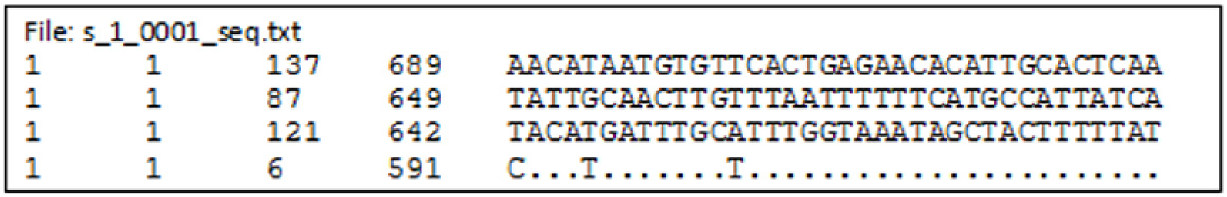
```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
28 35 28 27 34 27 26 25 25 28 31 24 26 27 32 25 27 27 32 28 6  28 27 27 27 27 33 26
27 26 27 27 34 30 10 27 25 34 27 28 22 28 27 26 26 27 27 26 27 25 22 23 28 27 18 20
23 27 27 29 21 25 25 34 26 27 24 25 32 24 22 33 28 7  25 20 30 22 28 27 24 25 28 28
28 27 28 26 27 25 23 33 25 35 28 34 27 27 25 28 38 34 21 8  25 27 34 27 31 23 22 36
32 17 29 21 32 24 24 27 28 19 27 28 26 34 28 23 25 35 28 38 34 21 8  26 26 27 25 27
21 28 28 27 27 34 27 34 27 25 30 21 34 26 33 25 26 35 28 20 28 25 34 27 37 33 15 33
25 23 28 25
```

Read with .qual format.

(ii) Output formats from Illumina

Illumina is over-creative with their own formats.

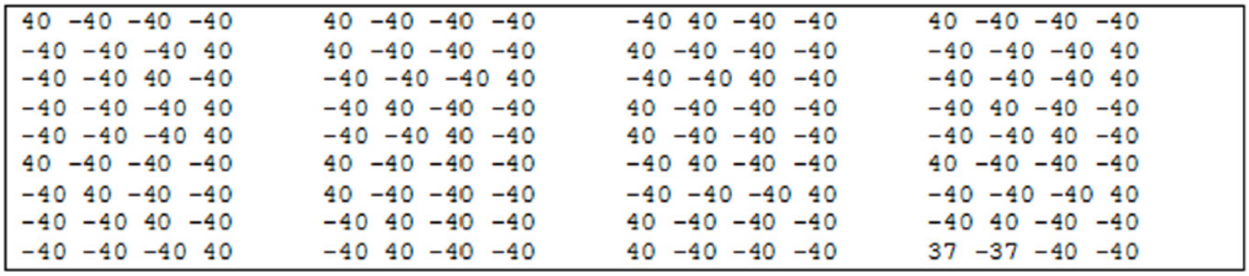
.seq.txt: base-called sequences



```
File: s_1_0001_seq.txt
1       1       137     689     AACATAATGTGTTCACTGAGAACACATTGCACTCAA
1       1       87      649     TATTGCAACTTGTTTAATTTTTTCATGCCATTATCA
1       1       121     642     TACATGATTTGCATTTGGTAAATAGCTACTTTTTAT
1       1       6       591     C...T.......T.......................
```

Read with .seq.txt format.

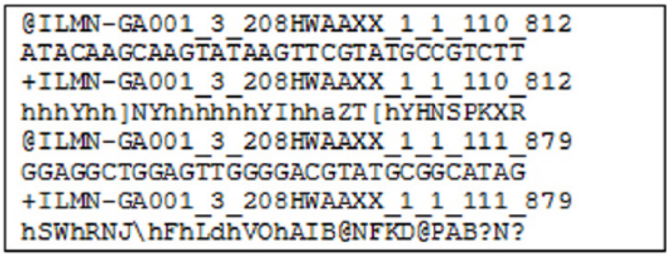
.prb.txt: qualities for each nucleotide/each cycle (is some kind of semi-raw data?)



```
40 -40 -40 -40      40 -40 -40 -40      -40 40 -40 -40      40 -40 -40 -40
-40 -40 -40 40      40 -40 -40 -40      40 -40 -40 -40      -40 -40 -40 40
-40 -40 40 -40      -40 -40 -40 40      -40 -40 40 -40      -40 -40 -40 40
-40 -40 -40 40      -40 40 -40 -40      40 -40 -40 -40      -40 40 -40 -40
-40 -40 -40 40      -40 -40 40 -40      40 -40 -40 -40      -40 -40 40 -40
40 -40 -40 -40      40 -40 -40 -40      -40 40 -40 -40      40 -40 -40 -40
-40 40 -40 -40      40 -40 -40 -40      -40 -40 -40 40      -40 -40 -40 40
-40 -40 40 -40      -40 40 -40 -40      40 -40 -40 -40      -40 40 -40 -40
-40 -40 -40 40      -40 40 -40 -40      40 -40 -40 -40      37 -37 -40 -40
```

Read with .prb.txt format.

Illumina FASTQ is not real FASTQ, but looks similarly (to cause confusion). The quality score is an own Illumina score, computed by subtracting 64 from the ASCII code of the character. For example the ‘h’ in the example is ord (‘h’) – 64 = 104 – 64 = 40.



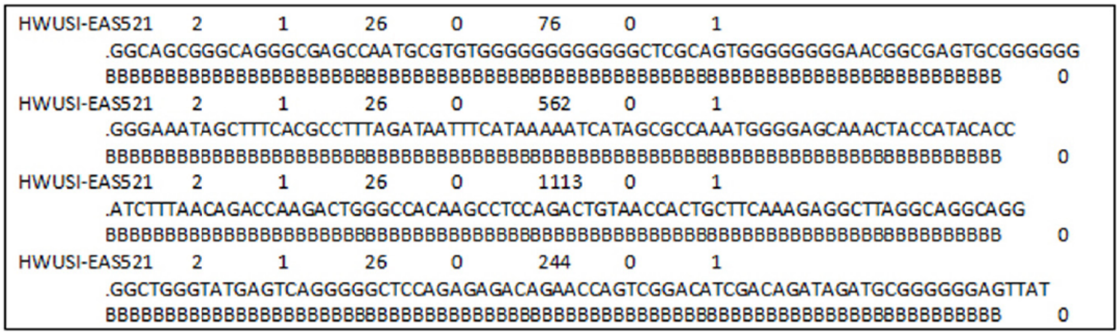
```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

Read with .fastq format.

See http://www.asciitable.com/ for ASCII codes (Dec is the ‘ord’ in decimal numbers).

The Illumina score can be transformed to Phred score. Given a character ch, the following expression should give the Phred quality q (please double-check:) q = 10 * log(1 + 10 ^ (ord(ch) - 64) / 10.0)) / log(10)

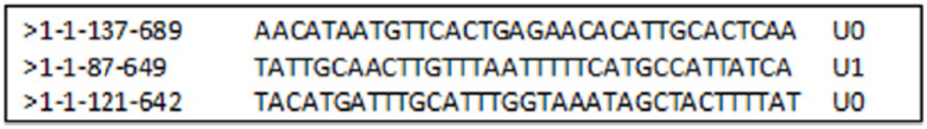
Qseq: Not looking like FASTQ, but the scores are proper Phred scores. Their encoding is however still ASCII – 64.



```
HWUSI-EAS521    2       1       26      0       76      0       1
        .GGCAGCGGGCAGGGCGAGCCAATGCGTGTGGGGGGGGGGGGCTCGCAGTGGGGGGGGAACGGCGAGTGCGGGGGG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
HWUSI-EAS521    2       1       26      0       562     0       1
        .GGGAAATAGCTTTCACGCCTTTAGATAATTTCATAAAAATCATAGCGCCAAATGGGGAGCAAACTACCATACACC
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
HWUSI-EAS521    2       1       26      0       1113    0       1
        .ATCTTTAACAGACCAAGACTGGGCCACAAGCCTCCAGACTGTAACCACTGCTTCAAAGAGGCTTAGGCAGGCAGG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
HWUSI-EAS521    2       1       26      0       244     0       1
        .GGCTGGGTATGAGTCAGGGGGCTCCAGAGAGACAGAACCAGTCGGACATCGACAGATAGATGCGGGGGGAGTTAT
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB      0
```

Read with .Qseq format.

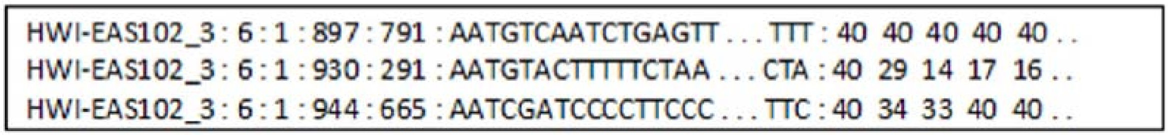
Illumina “**single line format**” and **SCARF**: base-called sequences and quality scores in a bit more handy format.



```
>1-1-137-689    AACATAATGTTCACTGAGAACACATTGCACTCAA    U0
>1-1-87-649     TATTGCAACTTGTTTAATTTTTCATGCCATTATCA   U1
>1-1-121-642    TACATGATTTGCATTTGGTAAATAGCTACTTTTAT   U0
```

Read with single line format

.

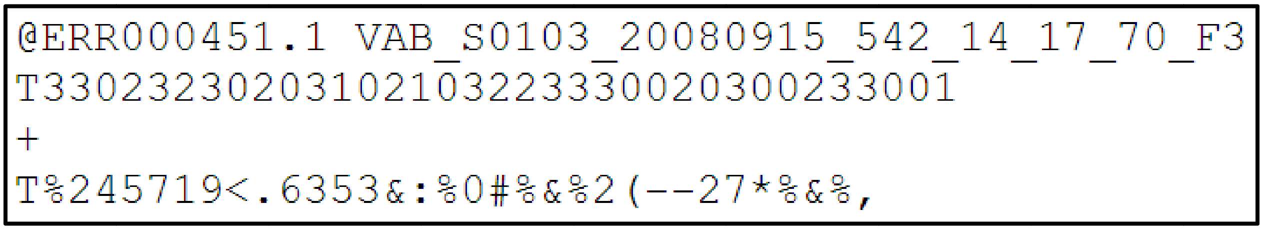


```
HWI-EAS102_3:6:1:897:791:AATGTCAATCTGAGTT...TTT:40 40 40 40 40..
HWI-EAS102_3:6:1:930:291:AATGTACTTTTTCTAA...CTA:40 29 14 17 16..
HWI-EAS102_3:6:1:944:665:AATCGATCCCCTTCCC...TTC:40 34 33 40 40..
```

Read with SCARF format

(iii) Output formats from SOLiD

SOLiD outputs **CSFASTA**. It's a semi-raw format similar to FASTQ. Read starts with the first base, but colour calls follow.



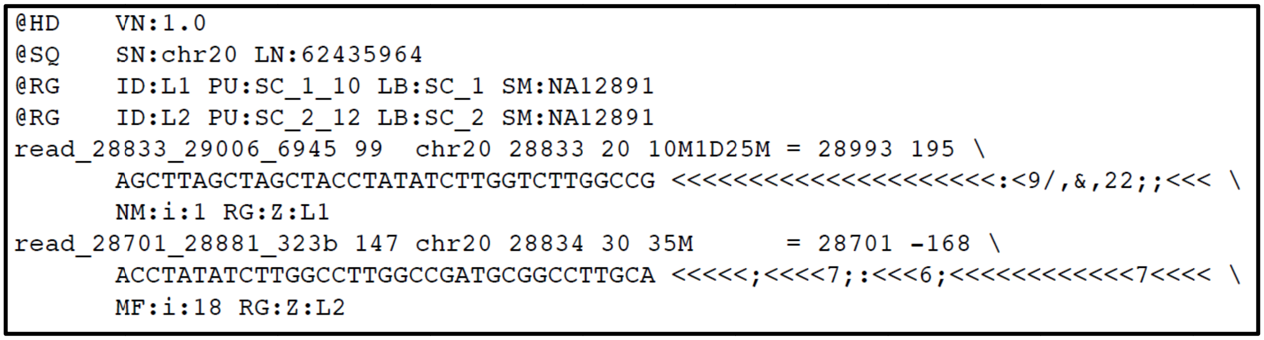
```
@ERR000451.1 VAB_S0103_20080915_542_14_17_70_F3
T330232302031021032233300203002330001
+
T%245719<.6353&:%0#%&%2(--27*%&%,
```

Read with CsFASTA format

Assembly Format Sample

Read aligners output a number of more or less similar own formats. However many of them output data in the common SAM/BAM format, what is good news.

For the typically large data, BAM is currently the most recommended and most "standard" format.



```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

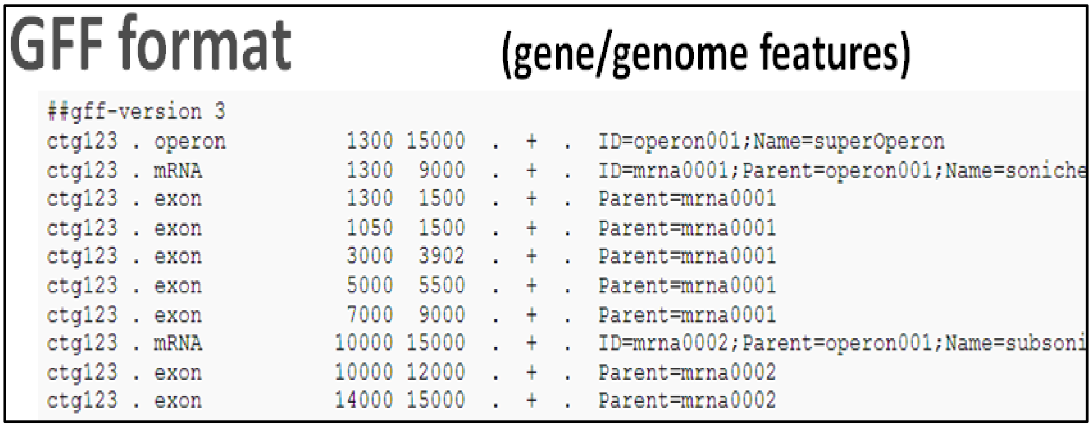
Figure 10: Read with SAM format

Genome & sequence annotation formats

Genome & sequence annotation formats can store some of the diverse kinds of data resulting from a sequencing study.BED format is a plain-text (tabular) format that can directly be displayed for example in the UCSC Genome Browser or similar.

GFF format, latest version GFF3. It's a tabular plain-text format for genome or sequence annotation, can contain also the sequences, alignments, dependencies between features. Is extensible. Currently most recommended & quite "standard" format.

BioXSD is a new set of structured, "object-oriented" formats for exchange of sequence data, any kind of sequence/genome annotation, and related. It is under development and optimised for use with (but not limited to) Web services, and aims to becoming a standard format in future. Very generic and extensible thanks to the use of XML Schema (XSD).



GFF format (gene/genome features)

```
##gff-version 3
ctg123 . operon      1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA        1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=soniche
ctg123 . exon        1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon        1050  1500  .  +  .  Parent=mrna0001
ctg123 . exon        3000  3902  .  +  .  Parent=mrna0001
ctg123 . exon        5000  5500  .  +  .  Parent=mrna0001
ctg123 . exon        7000  9000  .  +  .  Parent=mrna0001
ctg123 . mRNA       10000 15000  .  +  .  ID=mrna0002;Parent=operon001;Name=subsoni
ctg123 . exon       10000 12000  .  +  .  Parent=mrna0002
ctg123 . exon       14000 15000  .  +  .  Parent=mrna0002
```

Read with SAM format

Metadata Storage

Nucleotide sequence databases (GenBank, EMBL, DDBJ, gathered under the INSDC consortium) store a number of types of metadata about the genomic sequences.

However, additional metadata is crucial to enable efficient future research using the submitted genome & metagenome-sequencing data. GSC consortium has therefore defined the MIGS and MIMS standards for additional metadata that should be submitted to a public repository.

"The minimum information about a genome sequence (MIGS) specification" .Extended to the Minimimum Information about a Metagenomic Sequence/Sample: MIMS. MIENS (environmental sequencing) follow. GCDML is an XML format for MIGS/MIMS/MIENS metadata, (Using XML Schema and SAWSDL annotations.)

The two main public read archives and their two main formats

It's however crucial for each data set to decide what data should be deleted, and after what time! When and whether to delete the raw data, flows, colours, when and whether the base-called reads & qualities, alignments. The data is often enormously huge and it will cost you and/or the national/global institutes a lot of money and energy to store and back them up. Hence we use SRA format (XML), SRF format.

NCBI SRA submission

SRA submission using the interactive submission interface

Researchers often need to submit their original sequence data to NCBI SRA (short read archive) for review. Sequence data to be submitted have to be accompanied by metadata which hold information of the study, experimental design, samples, run parameters, etc. In general there are two ways to submit metadata to the archive:

- Understand SRA terminology: While you are waiting for your account, go through SRA starter quick guide: http://www.ncbi.nlm.nih.gov/books/NBK47529/ Basically, in each submission, you can include one or multiple studies. You will be asked for an abstract describing each study. Associated with each study are one or multiple samples. A sample describes the nature and origin of the material which has been sequenced. You can have multiple experiments associated with one particular sample within a study. The experiment holds details on how the sequencing library was prepared. For each experiment, you can have multiple sequencing runs (NCBI suggest to include only one lane in each "run").

- Use the interactive submission interface on their web site: http://trace.ncbi.nlm.nih.gov/Traces/sra_sub/sub.cgi?&m=submissions&s=default. The interactive submission interface allows tracking of your submissions and informs you about errors in the submission process.

- Generate a set of XML files which hold all relevant data and submit them using ftp or ascp (something similar to and faster than ftp). In anycases the actual sequencing data are submitted via ftp or ascp.

- If you have just a few lanes of sequencing data we suggest using the interactive tool for generation of the metadata and submission of your files. If you have large data sets we provide help with the generation of XML files and the transfer via ascp.

Quality Check (QC)

Introduction- QC

Next generation sequencing (NGS) technologies provide a high-throughput means to generate large amount of sequence data. However, quality check (QC) of sequence data generated from these technologies is extremely important for meaningful downstream analysis. Further, highly efficient and fast processing tools are required to handle the large volume of datasets. Modern high throughput sequencers can generate tens of millions of sequences in a single run. Before analyzing this sequence to draw biological conclusions you should always perform some simple quality control checks to ensure that the raw data looks good and there are no problems or biases in your data which may affect how you can usefully use it.

Most sequencers will generate a QC report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself. QC aims to get a QC report which can spot problems which originate either in the sequencer or in the starting library material.

After obtaining reads (Fast Q, Fast A) from sequencer machine, first step is to check for QC data

- Next generation high throughput sequencers generate huge amount of sequencing data (100Mb to few GB's).
- Quality check and primary analysis of raw sequence data is vital prior to the in-depth analysis.
- QC and primary analysis of the huge sequencing data using ordinary methods / machines is close to impossible.

Q C analysis using Tool: FastQC

FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material. FasQC can be run in one of two modes. It can either run as a standalone interactive application for the immediate analysis of small numbers of FastQ files, or it can be run in a non-interactive mode where it would be suitable for integrating into a larger analysis pipeline for the systematic processing of large numbers of files.

Basic Operations

(i) Opening a Sequence file

To open one or more Sequence files interactively simply run the program and select File > Open. You can then select the files you want to analyze.

Newly opened files will immediately appear in the set of tabs at the top of the screen. Because of the size of these files it can take a couple of minutes to open them. FastQC operates a queuing system where only one file is opened at a time, and new files will wait until existing files have been processed.

FastQC supports files in the following formats

- FastQ (all quality encoding variants)
- CasavaFastQ files*
- ColorspaceFastQ
- GZip compressed FastQ
- SAM
- BAM
- SAM/BAM Mapped only (normally used for color space data)

* Casavafastq format is the same as regular fastq except that the data is usually split across multiple files for a single sample. In this mode the program will merge the files in a sample group and present a single report for each sample. Also Casavafastq files contain poor quality sequences which have been flagged to be removed. In Casava mode the program will exclude these flagged sequences from the report.

By default FastQC will try to guess the file format from the name of the input file. Anything ending in .sam or .bam will be opened as a SAM/BAM file (using all sequences, mapped and unmapped), and everything else will be treated as FastQ format. If you want to override this detection and specify the file format manually then you can use the drop down file filter in the file chooser to select the type of file you're going to load. You need to use the drop down selector to make the program use the Mapped BAM or Casava file modes as these won't be selected automatically.

(ii) Evaluating Results

The analysis in FastQC is performed by a series of analysis modules. The left hand side of the main interactive display or the top of the HTML report show a summary of the modules which were run, and a quick evaluation of whether the results of the module seem entirely normal (green tick), slightly abnormal (orange triangle) or very unusual (red cross).

It is important to stress that although the analysis results appear to give a pass/fail result, these evaluations must be taken in the context of what you expect from your library. A 'normal' sample as far as FastQC is concerned is random and diverse. Some experiments may be expected to produce libraries which are biased in particular ways. You should treat the summary evaluations therefore as pointers to where you should concentrate your attention and understand why your library may not look random and diverse.

(iii) Saving a Report

In addition to providing an interactive report FastQC also has the option to create an HTML version of this report for a more permanent record. This HTML report can also be generated directly by running FastQC in non-interactive mode. To create a report simply select File > Save Report from the main menu. By default a report will be created using the name of the fastq file with _fastqc.zip appended to the end. The report will be created for whichever file tab was active when the menu option was selected. The report file which is created is actually a zip file. When uncompressed this will create a folder with the same name as the report file. Inside this will be a series of files, but the one to look at is the fastqc_report.html file which links to the other files in the report.

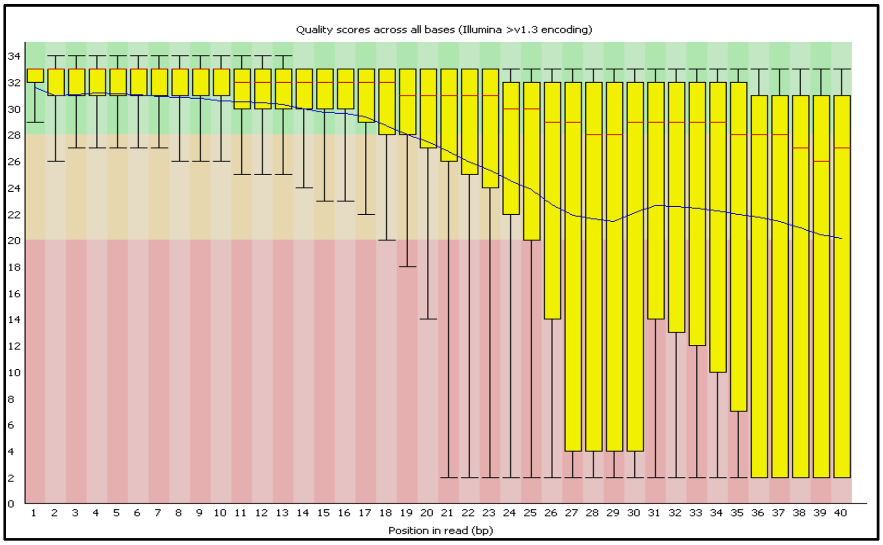
Analysis modules

(i) Basic Statistics

The Basic Statistics module generates some simple composition statistics for the file analyzed.

- Filename: The original filename of the file which was analyzed
- File type: Says whether the file appeared to contain actual base calls or color space data which had to be converted to base calls
- Encoding: Says which ASCII encoding of quality values was found in this file.
- Total Sequences: A count of the total number of sequences processed. There are two values reported, actual and estimated. At the moment these will always be the same. In the future it may be possible to analyze just a subset of sequences and estimate the total number, to speed up the analysis, but since we have found that problematic sequences are not evenly distributed through a file we have disabled this for now.
- Filtered Sequences: If running in Casava mode sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will the number of sequences actually used for the rest of the analysis.
- Sequence Length: Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
- %GC: The overall %GC of all bases in all sequence

(ii) Per Base Sequence Quality



This view shows an overview of the range of quality values across all bases at each position in the FastQ file.

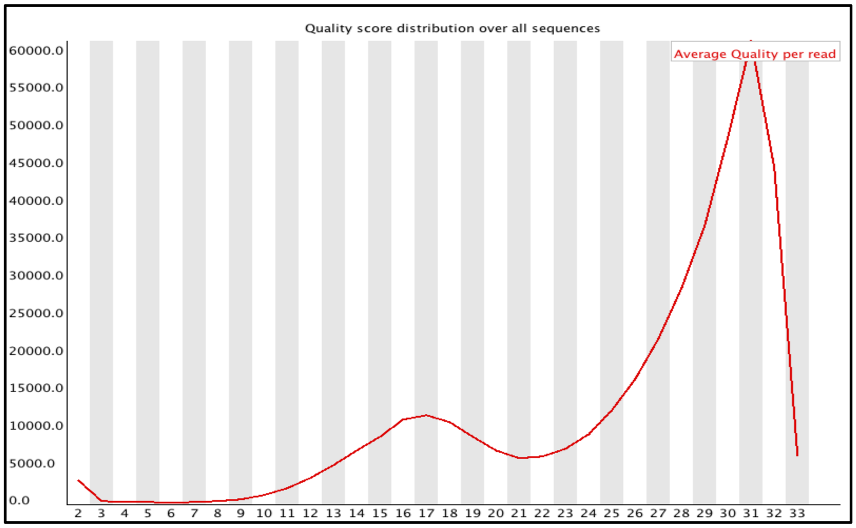
For each position a Box Whisker type plot is drawn. The elements of the plot are as follows:

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read. It should be mentioned that there are number of different ways to encode a quality score in a Fast Q file. FastQC attempts to automatically determine which encoding method was used, but in some very limited datasets it is possible that it will guess this incorrectly (ironically only when your data is universally very good!). The title of the graph will describe the encoding FastQC thinks your file used. A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25.This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

(iii) Per Sequence Quality Scores

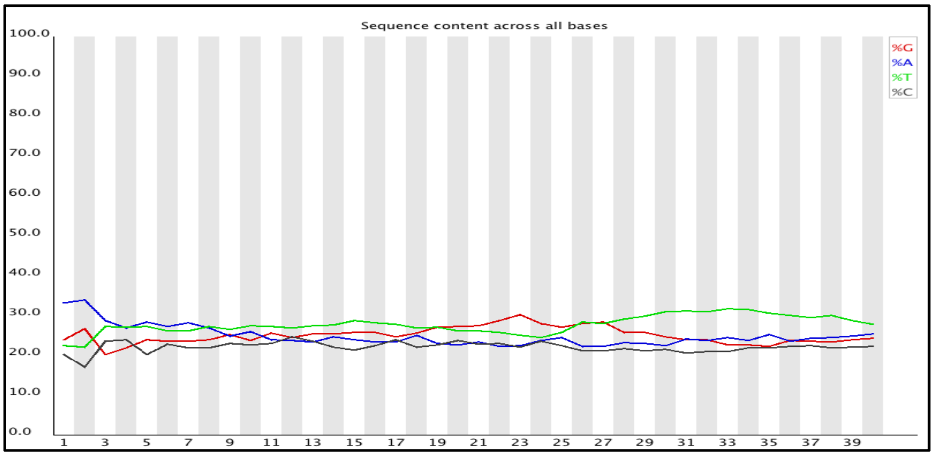
The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences.



If a significant proportion of the sequences in a run have overall low quality then this could indicate some kind of systematic problem - possibly with just part of the run (for example one end of a flowcell). A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate. An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

(iv) Per Base Sequence Content

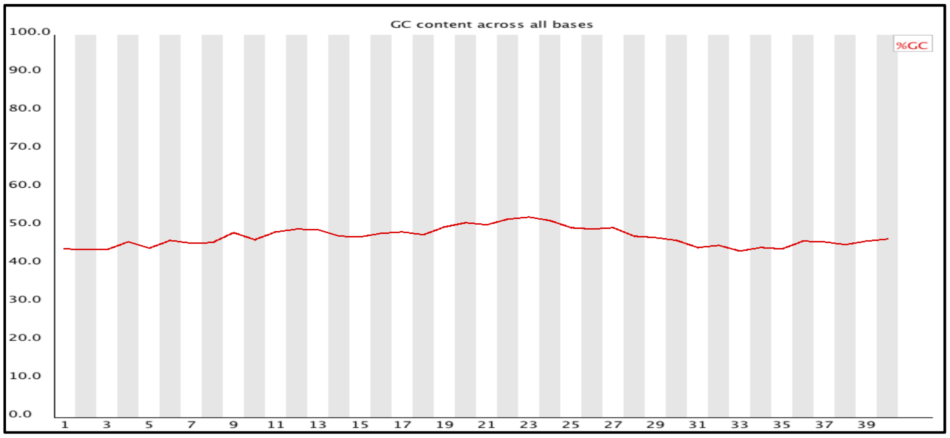
Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.



In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall amount of these bases in your genome, but in any case they should not be hugely imbalanced from each other. If you see strong biases which change in different bases then this usually indicates an overrepresented sequence which is contaminating your library. A bias which is consistent across all bases either indicates that the original library was sequence biased, or that there was a systematic problem during the sequencing of the library. This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position. This module will fail if the difference between A and T, or G and C is greater than 20% in any position.

(v) Per Base GC Content

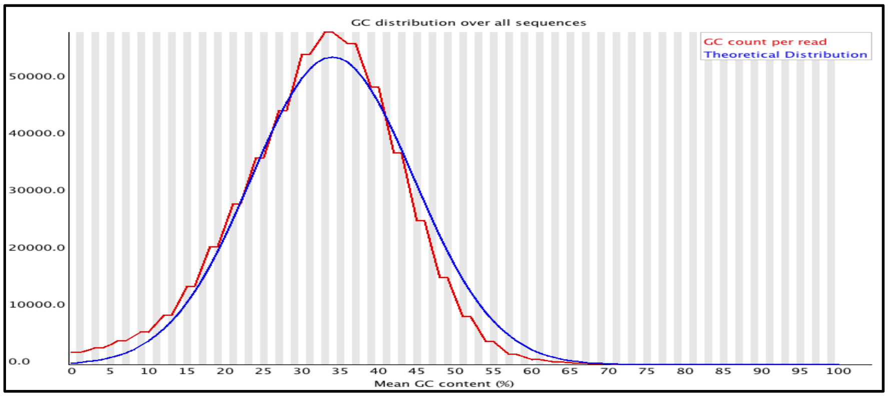
Per Base GC Content plots out the GC content of each base position in a file. In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the line in this plot should run horizontally across the graph. The overall GC content should reflect the GC content of the underlying genome.



If you see a GC bias which changes in different bases then this could indicate an over presented sequence which is contaminating your library. A bias which is consistent across all bases either indicates that the original library was sequence biased, or that there was a systematic problem during the sequencing of the library. This module issues a warning it the GC content of any base strays more than 5% from the mean GC content. This module will fail if the GC content of any base strays more than 10% from the mean GC content.

(vi) Per Sequence GC Content

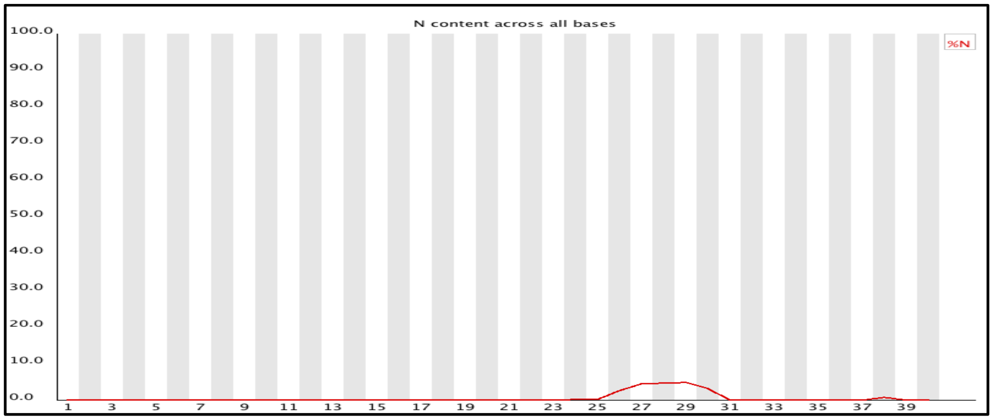
This module measures the GC content across the whole length of each sequence in a file and compares it to a modeled normal distribution of GC content.



In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what your genome's GC content should be. A warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads. This module will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads.

(vii) Per Base N Content

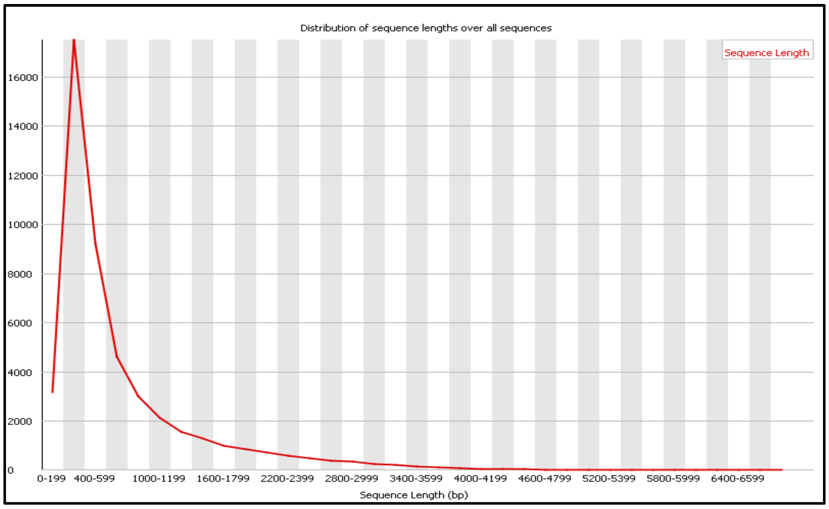
If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called.



It's not unusual to see a very low proportion of Ns appearing in a sequence, especially nearer the end of a sequence. However, if this proportion rises above a few percent it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls. This module raises a warning if any position shows an N content of >5%. This module will raise an error if any position shows an N content of >20%.

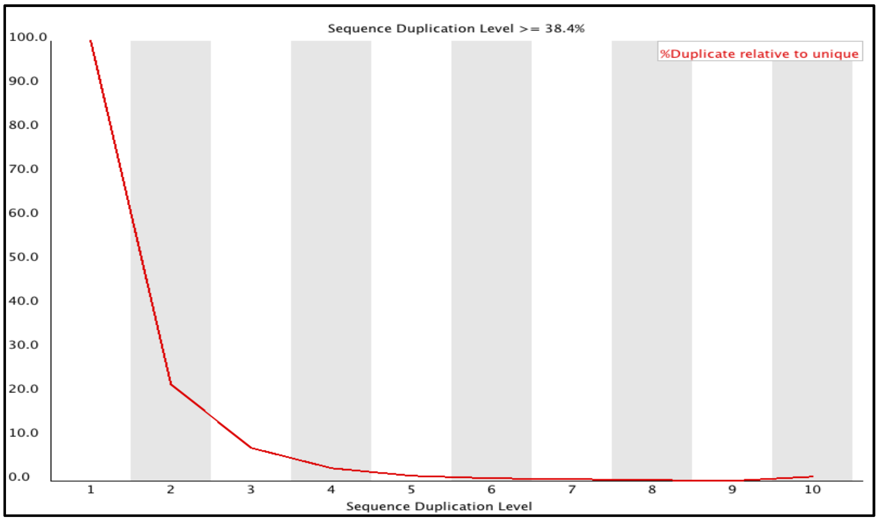
(viii) Sequence Length Distribution

Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end. This module generates a graph showing the distribution of fragment sizes in the file which was analyzed. In many cases this will produce a simple graph showing a peak only at one size, but for variable length FastQ files this will show the relative amounts of each different size of sequence fragment. This module will raise a warning if all sequences are not the same length. This module will raise an error if any of the sequences have zero length.



(ix) Duplicate Sequences

In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification). This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication.



To cut down on the memory requirements for this module only sequences which occur in the first 200,000 sequences in each file are analyzed, but this should be enough to get a good impression for the duplication levels in the whole file. Each sequence is tracked to the end of the file to give a representative count of the overall duplication level. To cut down on the amount of information in the final plot any sequences with more than 10 duplicates are placed into the 10 duplicates category - so it's not unusual to see a small rise in this final category. If you see a big rise in this final category then it means you have a large number of sequences with very high levels of duplication. Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to under represent highly duplicated sequences.This module will issue a warning if non-unique sequences make up more than 20% of the total. This module will issue an error if non-unique sequences make up more than 50% of the total.

(x) Overrepresented Sequences

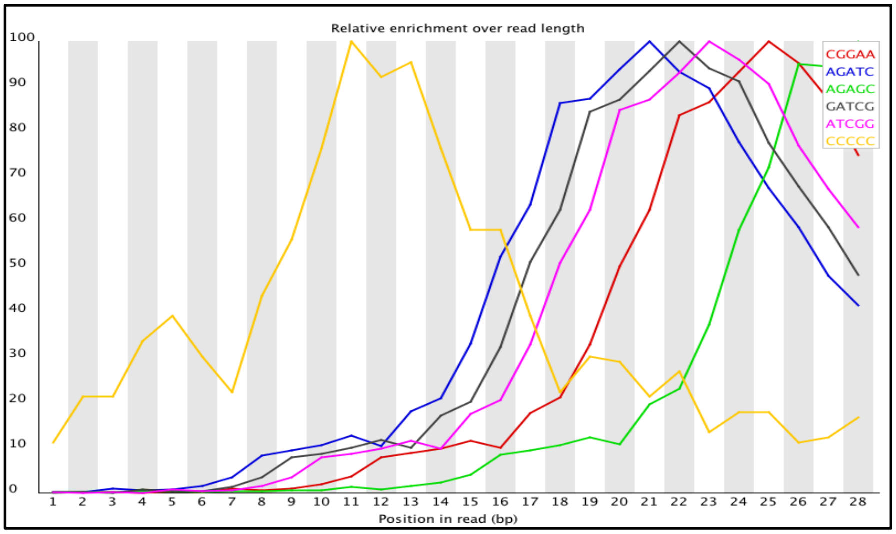
A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected. This module lists all of the sequence which makes up more than 0.1% of the total. To conserve memory only sequences which appear in the first 200,000 sequences are tracked to the end of the file. It is therefore possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module. For each overrepresented sequence the program will look for matches in a database of common contaminants and will report the best hit it finds. Hits must be at least 20bp in length and have no more than 1 mismatch. Finding a hit doesn't necessarily mean that this is the source of the contamination, but may point you in the right direction. Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to under represent highly duplicated sequences. This module will issue a warning if any sequence is found to represent more than 0.1% of the total. This module will issue an error if any sequence is found to represent more than 1% of the total.

(xi) Over represented K-mers

The analysis of overrepresented sequences will spot an increase in any exactly duplicated sequences, but there is a different subset of problems where it will not work.

- If you have very long sequences with poor sequence quality then random sequencing errors will dramatically reduce the counts for exactly duplicated sequences.
- If you have a partial sequence which is appearing at a variety of places within your sequence then this won't be seen either by the per base content plot or the duplicate sequence analysis.

This module counts the enrichment of every 5-mer within the sequence library. It calculates an expected level at which this k-mer should have been seen based on the base content of the library as a whole and then uses the actual count to calculate an observed/expected ratio for that k-mer. In addition to reporting a list of hits it will draw a graph for the top 6 hits to show the pattern of enrichment of that Kmer across the length of your reads. This will show if you have a general enrichment, or if there is a pattern of bias at different points over your read length.



Any k-mer showing more than a 3 fold overall enrichment or a 5 fold enrichment at any given base position will be reported by this module. To allow this module to run in a reasonable time only 20% of the whole library is analyzed and the results are extrapolated to the rest of the library. This module will issue a warning if any k-mer is enriched more than 3 fold overall, or more than 5 fold at any individual position. This module will issue an error if any k-mer is enriched more than 10 fold at any individual base position.

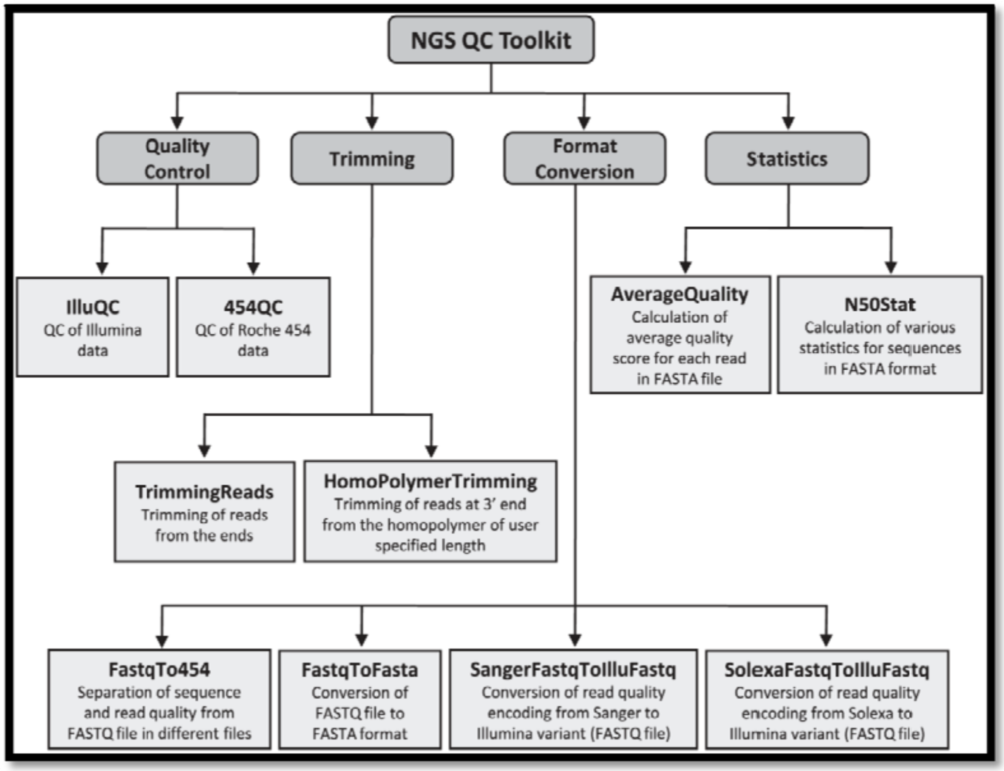
Practical

QC analysis using Tool: FastQC

Load the whole genome sequence read data to the software to analyze and save results.

QC analysis using NGSQCToolkit_v2.3

Run perl script using terminal, within the folder containing whole genome sequencing read data, and analyze the result**.** This tool performs quality check and filtering of the sequencing data generated using Illumina technology and 454 reads. Input to this tool is FASTQ files containing read and corresponding quality scores. The tool helps to understand detailed QC statistics of the complete reads as below.


NGS QC Toolkit
Quality Control
Trimming
Format Conversion
Statistics
IlluQC
QC of Illumina data
454QC
QC of Roche 454 data
AverageQuality
Calculation of average quality score for each read in FASTA file
N50Stat
Calculation of various statistics for sequences in FASTA format
TrimmingReads
Trimming of reads from the ends
HomoPolymerTrimming
Trimming of reads at 3' end from the homopolymer of user specified length
FastqTo454
Separation of sequence and read quality from FASTQ file in different files
FastqToFasta
Conversion of FASTQ file to FASTA format
SangerFastqToIlluFastq
Conversion of read quality encoding from Sanger to Illumina variant (FASTQ file)
SolexaFastqToIlluFastq
Conversion of read quality encoding from Solexa to Illumina variant (FASTQ file)

Tools in NGS QC toolkit

(i) QC

- IlluQC.pl: Tool for quality control of sequencing data generated using Illumina platform (FASTQ format)
- IlluQC_PRLL.pl: This tool has the same functionality as IlluQC.pl. However, it provides an additional option to use multiple CPUs to speed up the analysis
- 454QC.pl: Tool for quality control of sequencing data generated using 454 platform (read and quality in FASTA format)
- 454QC_PRLL.pl: Tool performs same quality control analysis as 454QC.pl and helps to analyze data using multiple CPUs
- 454QC_PE.pl: Tool for quality control of paired-end sequencing data generated using 454 platform (read and quality in FASTA format)

(ii) Format-converter

- SangerFastqToIlluFastq.pl: To convert fastq-sanger variant to fastq-illumina variant of FASTQ format

- SolexaFastqToIlluFastq.pl: To convert fastq-solexa variant to fastq-illumina variant of FASTQ format
- FastqTo454.pl: To convert FASTQ format (any variant) to 454 format (two files in FASTA format: one for reads/sequences (.fna) and another for quality (.qual))
- FastqToFasta.pl: To convert FASTQ format file to FASTA format file for reads/sequences

(iii) Trimming

- TrimmingReads.pl: Tool for trimming reads from 5’ and/or 3’ end of the read (FASTQ or FASTA format)
- HomoPolymerTrimming.pl: Tool for trimming 3’ end of the reads from the first base of homopolymer of given length
- AmbiguityFiltering.pl: Tool for filtering reads containing ambiguous bases or trimming flanking ambiguous bases

(iv) Statistics

- AvgQuality.pl: Tool to calculate average quality score for each read and overall quality score for the given FASTA quality file
- N50Stat.pl: Tool to generate statistics for read/sequence data given in FASTA format

Running Illu QC.pl

IlluQC.pl: This tool performs quality check and filtering of the sequencing data generated using Illumina technology. Input to this tool is FASTQ files (any variant) containing read and corresponding quality scores. Following are the options available with IlluQC.pl

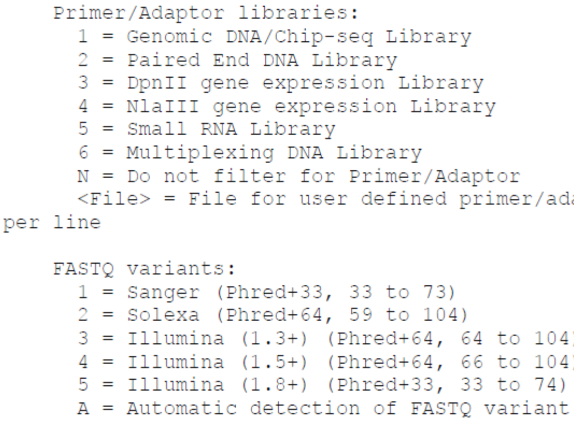
Usage: perl IlluQC.pl <options>
Perl IlluQC.pl read1.fastq read2.fastq 2 A

Paired End reads

-pe <Forward reads file> <Reverse reads file> <Primer/Adaptor library> <FASTQ variant>

Paired-end read files (FASTQ) with primer/adaptor library and FASTQ variant

User may choose from the provided primer/adaptor library or can give a file containing primer/adaptor sequences, one per line. Multiple libraries can be given using multiple '-pe' options



```
    Primer/Adaptor libraries:
      1 = Genomic DNA/Chip-seq Library
      2 = Paired End DNA Library
      3 = DpnII gene expression Library
      4 = NlaIII gene expression Library
      5 = Small RNA Library
      6 = Multiplexing DNA Library
      N = Do not filter for Primer/Adaptor
      <File> = File for user defined primer/ada
per line

    FASTQ variants:
      1 = Sanger (Phred+33, 33 to 73)
      2 = Solexa (Phred+64, 59 to 104)
      3 = Illumina (1.3+) (Phred+64, 64 to 104)
      4 = Illumina (1.5+) (Phred+64, 66 to 104)
      5 = Illumina (1.8+) (Phred+33, 33 to 74)
      A = Automatic detection of FASTQ variant
```

For eg.: -pe r1.fq r2.fq 3 1 -pe t1.fq t2.fq 2 A

Singe End reads

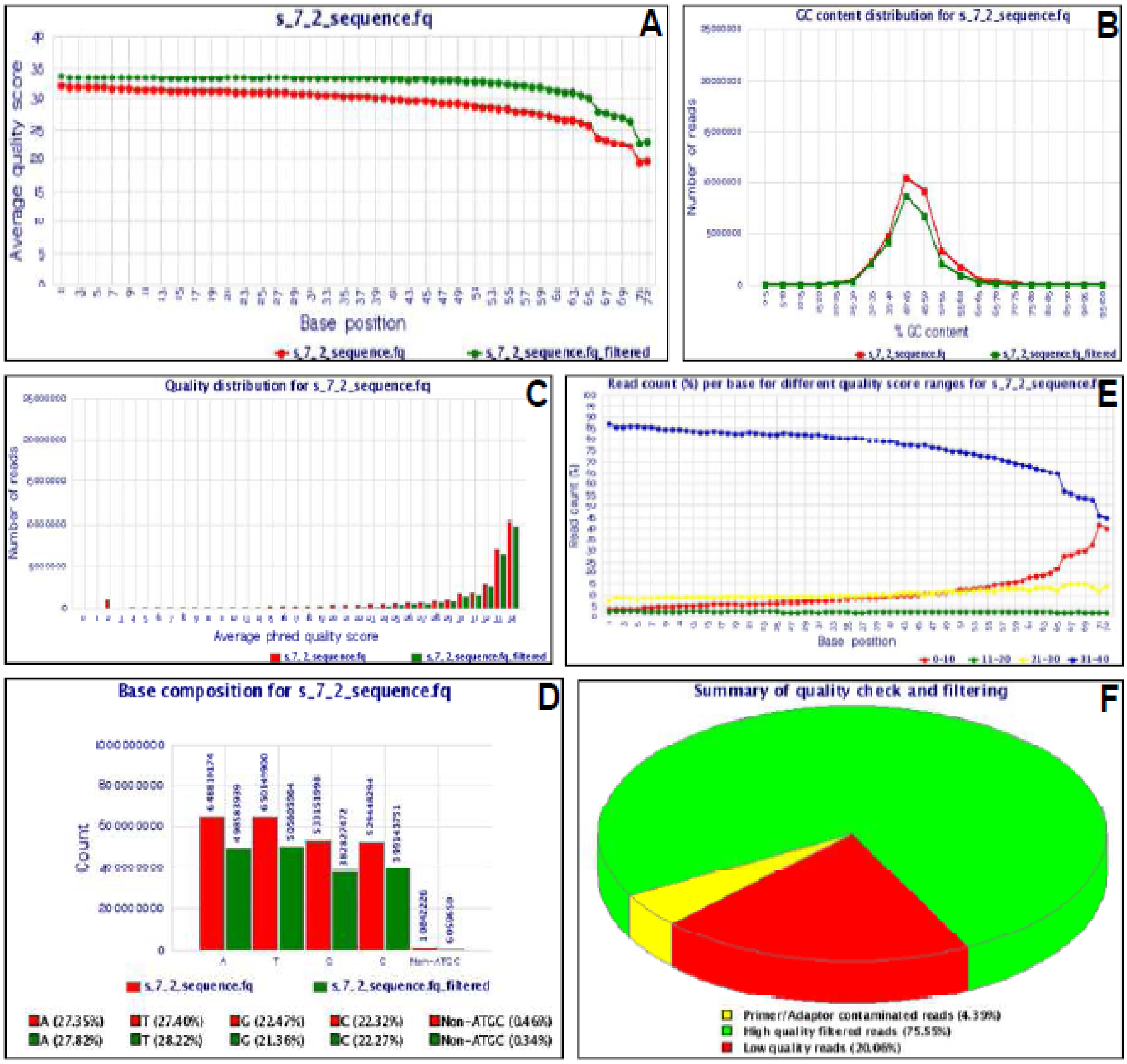
-se <Reads file> <Primer/Adaptor library> <FASTQ variant>

Single-end read file (FASTQ) with primer/adaptor library and FASTQ variant

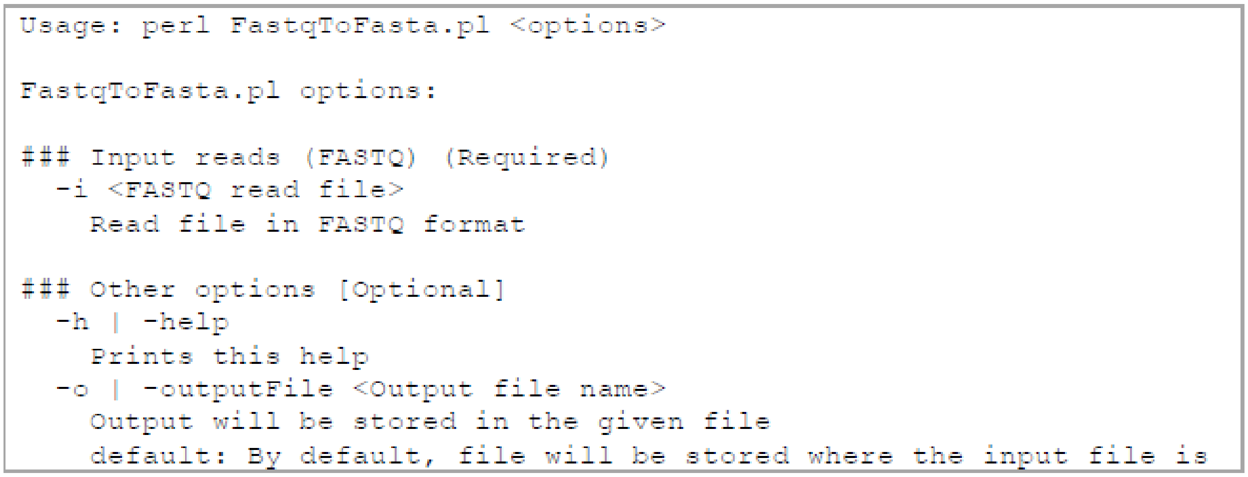
Multiple libraries can be given using multiple '-se' options

For eg.: -se r1.fq 3 2 -se t2.fq 2 2

Output: IlluQC.pl generates statistics for quality check and filtering steps along with quality of input and high-quality filtered data in the form of text files and graphs. The sample output graphs showing the average quality score at each base position (A), GC content distribution (B), average quality distribution (C) and base composition (D) for input and filtered reads. (E) percentage of reads for different quality score ranges at each base position. The pie chart with summary of quality control analysis (F).


A
s_7_2_sequence.fq
Average quality score
Base position
s_7_2_sequence.fq
s_7_2_sequence.fq_filtered
B
GC content distribution for s_7_2_sequence.fq
Number of reads
% GC content
s_7_2_sequence.fq
s_7_2_sequence.fq_filtered
C
Quality distribution for s_7_2_sequence.fq
Number of reads
Average phred quality score
s_7_2_sequence.fq
s_7_2_sequence.fq_filtered
E
Read count (%) per base for different quality score ranges for s_7_2_sequence.fq
Read count (%)
Base position
0-10
11-20
21-30
31-40
D
Base composition for s_7_2_sequence.fq
Count
648819174
498583939
650149900
505805564
533151998
382827472
529448294
399143751
10842226
6050659
A
T
G
C
Non-ATGC
s_7_2_sequence.fq
s_7_2_sequence.fq_filtered
A (27.35%)
T (27.40%)
G (22.47%)
C (22.32%)
Non-ATGC (0.46%)
A (27.82%)
T (28.22%)
G (21.36%)
C (22.27%)
Non-ATGC (0.34%)
F
Summary of quality check and filtering
Primer/Adaptor contaminated reads (4.39%)
High quality fibered reads (75.55%)
Low quality reads (20.06%)

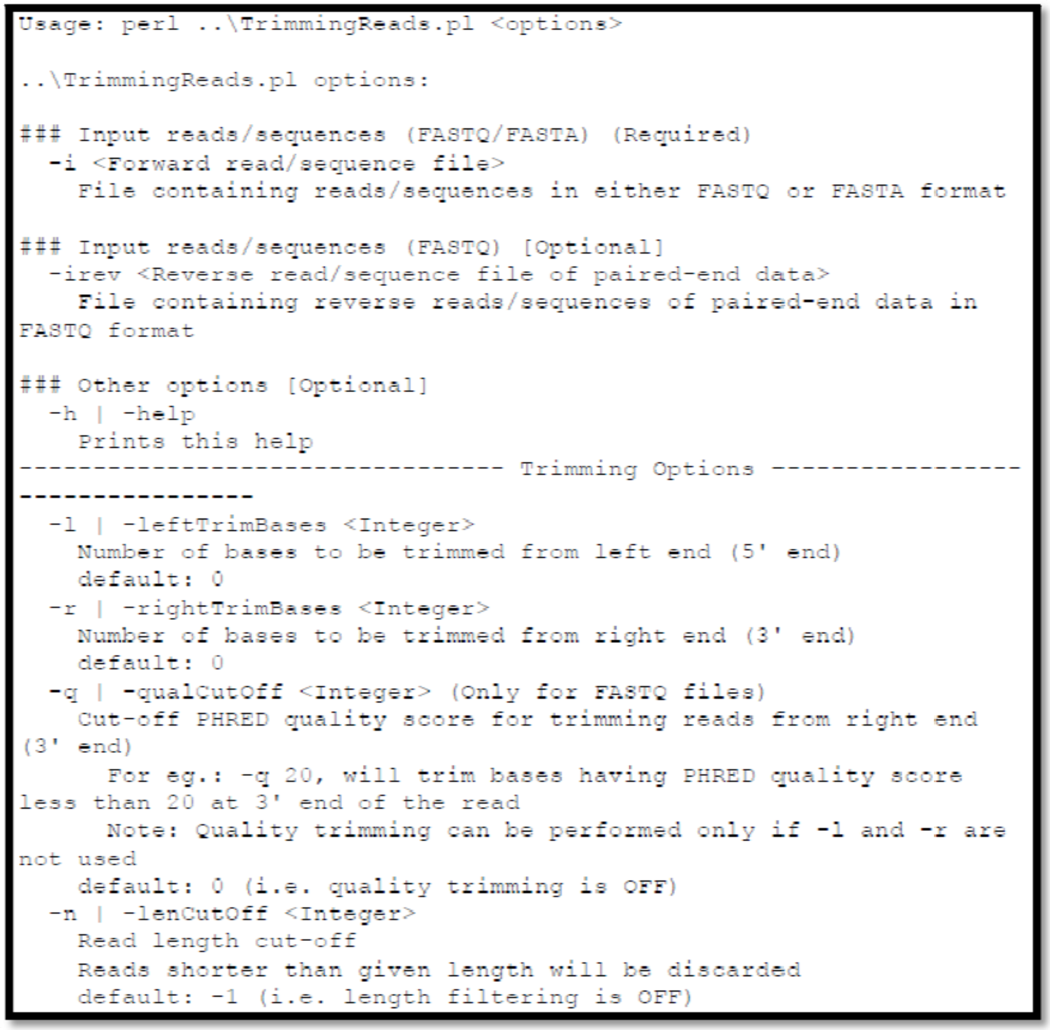
- **FastqToFasta.pl:** It exports reads/sequences from the FASTQ file to the FASTA format file. Following are the options available with the tool.



```
Usage: perl FastqToFasta.pl <options>

FastqToFasta.pl options:

### Input reads (FASTQ) (Required)
  -i <FASTQ read file>
    Read file in FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, file will be stored where the input file is
```

Trimming Tools

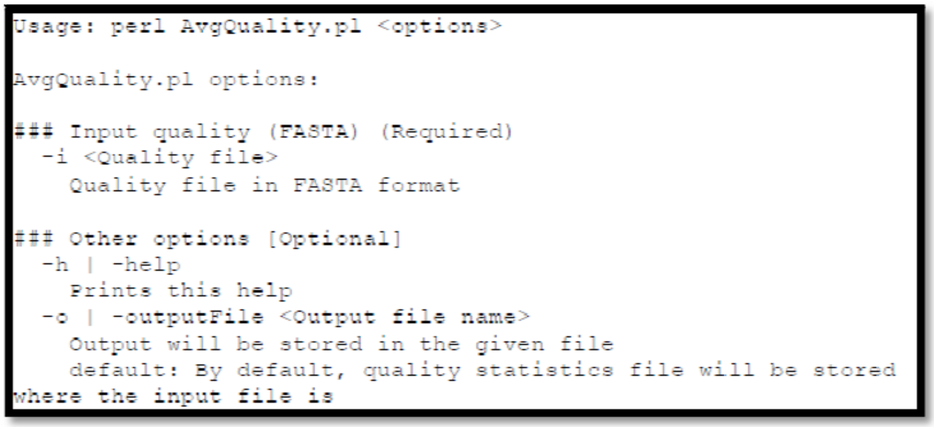
TrimmingReads.pl: This tool trims the reads/sequences and their quality scores (in case of FASTQ file) in two ways. First, it trims fixed (user-specified) number of bases from 5’ and/or 3’ end of the reads and corresponding qualities from the input FASTQ file. Second, it trims low quality bases from 3’ end of the read using user-defined threshold value of quality score. Input to this tool is either FASTQ or FASTA format file. Options are provided to specify the number of bases to be trimmed and the quality threshold for quality based trimming.



```
Usage: perl ..\TrimmingReads.pl <options>

..\TrimmingReads.pl options:

### Input reads/sequences (FASTQ/FASTA) (Required)
  -i <Forward read/sequence file>
    File containing reads/sequences in either FASTQ or FASTA format

### Input reads/sequences (FASTQ) [Optional]
  -irev <Reverse read/sequence file of paired-end data>
    File containing reverse reads/sequences of paired-end data in
FASTQ format

### Other options [Optional]
  -h | -help
    Prints this help
--------------------------------- Trimming Options -----------------
----------------
  -l | -leftTrimBases <Integer>
    Number of bases to be trimmed from left end (5' end)
    default: 0
  -r | -rightTrimBases <Integer>
    Number of bases to be trimmed from right end (3' end)
    default: 0
  -q | -qualCutOff <Integer> (Only for FASTQ files)
    Cut-off PHRED quality score for trimming reads from right end
(3' end)
      For eg.: -q 20, will trim bases having PHRED quality score
less than 20 at 3' end of the read
      Note: Quality trimming can be performed only if -l and -r are
not used
    default: 0 (i.e. quality trimming is OFF)
  -n | -lenCutOff <Integer>
    Read length cut-off
    Reads shorter than given length will be discarded
    default: -1 (i.e. length filtering is OFF)
```

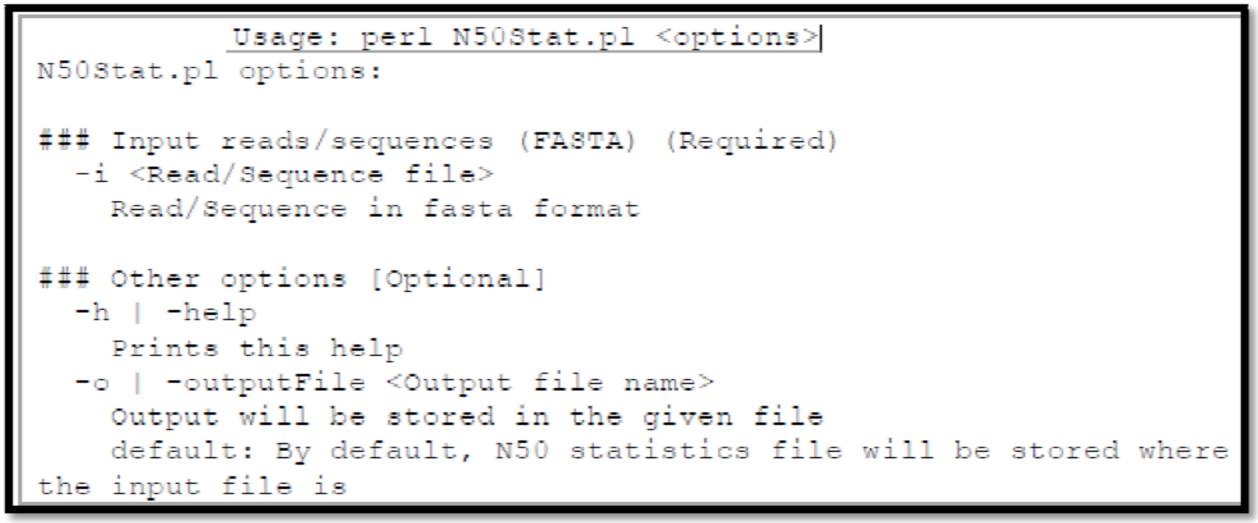
Statistics Tools

AvgQuality.pl: Tool calculates average quality score for each read and overall average quality score for the given file. This tool takes a quality file in FASTA format as an input.



```
Usage: perl AvgQuality.pl <options>

AvgQuality.pl options:

### Input quality (FASTA) (Required)
  -i <Quality file>
    Quality file in FASTA format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, quality statistics file will be stored
where the input file is
```

N50Stat.pl: This tool calculates different statistics for read file given in FASTA format. It calculates total number of reads/sequences, total and individual (A,T,C,G and N) number of bases, G+C and A+T counts, and minimum, maximum, average, median, N25, N50, N75, N90 and N95 read/sequence length. Following are the options provided for this tool.



```
          Usage: perl N50Stat.pl <options>
N50Stat.pl options:

### Input reads/sequences (FASTA) (Required)
  -i <Read/Sequence file>
    Read/Sequence in fasta format

### Other options [Optional]
  -h | -help
    Prints this help
  -o | -outputFile <Output file name>
    Output will be stored in the given file
    default: By default, N50 statistics file will be stored where
the input file is
```

Output: QC statistics

- File name 98-93454AllContigs.fna
- Total number of reads 9831
- Total number of trimmed reads containing homopolymer 0
- Total number of trashed reads (<100 bp in length after trimming) 0
- Total number of low quality reads (excluding <100 reads) 0
- Total number of HQ reads 9831
- Percentage of HQ reads 100.00%
- Total number of bases 39869019
- Total number of bases in HQ reads 39869019
- Total number of HQ bases in HQ reads 39753807
- Percentage of HQ bases in HQ reads 99.71%
- Number of Primer/Adaptor trimmed reads NA
- Total number of HQ filtered reads 9831
- Percentage of HQ filtered reads 100.00%

Detailed QC statistics

- File name 98-93454AllContigs.fna_filtered
- Total number of reads 9831 9831
- Minimum read length 100 100
- Maximum read length 180159 180159
- Average read length 4055.44 4055.44
- Median read length 903 903
- N25 length 36897 36897
- N50 length 17867 17867
- N75 length 5960 5960
- N90 length 2000 2000
- N95 length 1006 1006
- Total number of bases 39869019 39869019
- Total number of HQ bases 39753807 39753807
- Percentage of HQ bases 99.71% 99.71%

- Average quality score (Overall) 63.39 63.39

Further Reading

1. *Patel RK, Jain M (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. PLoS ONE, 7(2): e30619.*

Chapter 3

Insights in to NGS Data Analysis

Ashwani Kumar Mishra & Rakesh Kumar Maurya

Genome Sequencing is only the beginning. Further analysis are necessary to convert raw sequence data into understanding of functional biology. Various NGS platforms in the market such as Illumina, SOLiD, and Roche, offer unprecedented ability to apply massively parallel sequencing of transcriptomic (RNA) and genomic (DNA) samples to understand disease and health . The enormous data pose a fundamental problem of management and analysis. First, errors in the original raw data from multiple platforms and diverse applications must be checked. Then, as computational methods for assembly, alignment, and variation detection continue to advance, a broad range of genetic analysis applications including: comparative genomics, high-throughput polymorphism detection, analysis of coding and non-coding RNAs, and identifying mutant genes in disease pathways can be addressed.

Experience the power of High Throughput SNP Genotyping services at affordable prices on India's first Fluidigm EP1 platform only at NxGenBio Life Sciences.

Practical

NGS Data Analysis – Quality check & Assembly

Chapter 4

Genome Alignment, Assembly & Variant Analysis

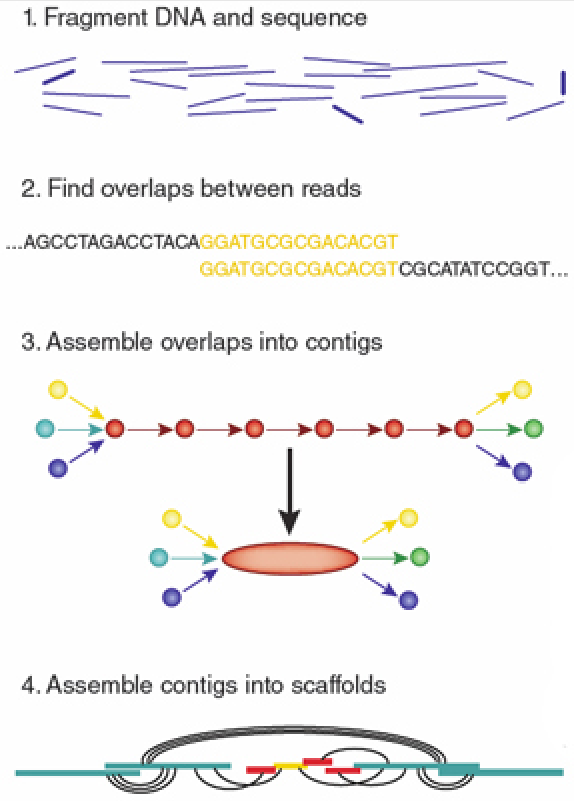
Rosana O. Babu

Introduction

In bioinformatics, sequence assembly refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original order of the sequence. Genome assembly is simply putting sequenced fragments of DNA into their correct chromosomal positions. This is needed as DNA sequencing technology cannot read whole genomes in one go, but rather reads small pieces of between 20 and 1000 bases, depending on the technology used. Typically the short fragments, called reads, result from sequencing genomic DNA, or gene transcript (ESTs). It is the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. In a shotgun sequencing project, the entire DNA from a source (usually a single organism, anything from a bacterium to a mammal) is first fractured into millions of small pieces. These pieces are then "read" by automated sequencing machines, which can read up to 1000 nucleotides or bases at a time. (The four bases are adenine, guanine, cytosine, and thymine, represented as AGCT.) A genome assembly algorithm works by taking all the pieces and aligning them to one another, and detecting all places where two of the short sequences, or *reads*, overlap. These overlapping reads can be merged, and the process continues.

Genome assembly is a very difficult computational problem, made more difficult because many genomes contain large numbers of identical sequences, known as repeats. These repeats can be thousands of nucleotides long, and some occur in thousands of different locations, especially in the large genomes of plants and animals. The resulting (draft) genome sequence is produced by combining the information sequenced contigs and then employing linking information to create scaffolds. Scaffolds are positioned along the physical map of the chromosomes creating a "golden path".

To assemble a genome, computer programs typically use data consisting of single and paired reads. Single reads are simply the short sequenced fragments themselves; they can be joined up through overlapping regions into a continuous sequence known as a 'contig'. Repetitive sequences, polymorphisms, missing data and mistakes eventually limit the length of the contigs that assemblers can build.



1. Fragment DNA and sequence

2. Find overlaps between reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds

Michael Schatz, Cold Spring Harbor

Genome assembly stitches together a genome from short sequenced pieces of DNA.

Given a set of sequence fragments the object is to find the shortest common super sequence.

1. calculate pairwise alignments of all fragments
2. choose two fragments with the largest overlap
3. merge chosen fragments
4. repeat step 2. and 3. until only one fragment is left
5. Paired reads typically are about the same length as single reads, but they come from either end of DNA fragments that are too long to be sequenced straight through.

Depending on the preparation technique, that distance can be as short as 200 base pairs or as large as several tens of kilobases. Knowing that paired reads were generated from the same piece of DNA can help link contigs into 'scaffolds', ordered assemblies of contigs with gaps in between. Paired-read data can also indicate the size of repetitive regions and how far apart contigs are.

Genome Assemblers

The first sequence assemblers began to appear in the late 1980s and early 1990s as variants of simpler sequence alignment programs to piece together vast quantities of fragments generated by automated sequencing instruments called DNA sequencers. Originally, most large-scale DNA sequencing centers developed their own software for assembling the sequences that they produced. However, this has changed as the software has grown more complex and as the number of sequencing centers has increased. An example of such assembler *Short Oligonucleotide Analysis Package* developed by BGI for de novo assembly of human-sized genomes, alignment, SNP detection, resequencing, indel finding, and structural variation analysis. As the sequenced organisms grew in size and complexity from small viruses over plasmids to bacteria and finally eukaryotes, the assembly programs used in these genome projects needed to increasingly employ more and more sophisticated strategies to handle:

- Terabytes of sequencing data which need processing on computing clusters;
- Identical and nearly identical sequences (known as *repeats*) which can, in the worst case, increase the time and space complexity of algorithms exponentially;
- Errors in the fragments from the sequencing instruments, which can confound assembly.

De-novo vs. mapping (reference) assembly

In sequence assembly, two different types can be distinguished:

- **De-novo**: assembling short reads to create full-length (sometimes novel) sequences. Overlapping reads are presumed to be from the same area of the genome, for *de-novo* assembly, read depth should be more.

- **Mapping or Reference:** assembling reads against an existing backbone sequence taken as reference, building a sequence that is similar but not necessarily identical to the backbone sequence. Combines all sequencing reads into contigs based on sequence similarity between reads and reference.

In terms of complexity and time requirements, de-novo assemblies are orders of magnitude slower and more memory intensive than mapping assemblies. This is mostly due to the fact that the assembly algorithm needs to compare every read with every other read (an operation that has a complexity of O (n^2) but can be reduced to O(n log(n)). Referring to the comparison drawn to shredded books in the introduction: while for mapping assemblies one would have a very similar book as template (perhaps with the names of the main characters and a few locations changed), the de-novo assemblies are more hardcore in a sense as one would not know beforehand whether this would become a science book, or a novel, or a catalogue etc.

Reference Assembly

One of the most complex and computationally intensive tasks of genome sequence analysis is genome assembly. Even today, few centers have the resources, in both software and hardware, to assemble a genome from the thousands or millions of individual sequences generated in a whole-genome shotgun sequencing project. With the rapid growth in the number of sequenced genomes has come an increase in the number of organisms for which two or more closely related species have been sequenced. This has created the possibility of building a comparative genome assembly algorithm, which can assemble a newly sequenced genome by mapping it onto a reference genome. Reference assisted assembly requires the use of a reference sequence, as a model, to assist in the assembly of the novel genome. The standard method for identifying the best reference sequence for the assembly of a novel genome aims at counting the number of reads that align to the reference sequence, and then choosing the reference sequence which has the highest number of reads aligning to it.

De novo assembly

A new generation of sequencing technologies is revolutionizing molecular biology. Illumina's Solexa and Applied Biosystems' SOLiD generate gigabases of nucleotide sequence per week. However, a perceived limitation of these ultra-high-throughput technologies is their short read-lengths. De novo assembly of sequence reads generated by classical Sanger capillary sequencing is a mature field of research. Unfortunately, the existing sequence assembly programs were not effective for short sequence reads generated by Illumina and SOLiD platforms. Early studies suggested that, in principle, sequence reads as short as 20-30 nucleotides could be used to generate useful assemblies of both prokaryotic and eukaryotic genome sequences, albeit containing many gaps. The early feasibility studies and proofs of principle inspired several bioinformatics research groups to implement new algorithms as freely available software tools specifically aimed at assembling reads of 30-50 nucleotides in length. This has led to the generation of several draft genome sequences based exclusively on short sequence Illumina sequence reads, recently culminating in the assembly of the 2.25-Gb genome of the giant panda from Illumina sequence reads with an average length of just 52 nucleotides.

Velvet is a novel set of de Bruijn graph-based sequence assembly methods for very short reads that can both remove errors and, in the presence of read pair information, resolve a large number of repeats. With unpaired reads, the assembly is broken when there is a repeat longer than the *k*-mer length. With the addition of short reads in read pair format, many of these repeats can be resolved, leading to assemblies similar to draft status in bacteria and reasonably long (∼5 kb) SCSCs in eukaryotic genomes. Velvet can convert high-coverage very short reads into reasonably sized contigs with no additional information. With additional paired read information to resolve small repeats, almost complete genomes can be assembled. We believe the Velvet framework will provide a rich set of different algorithmic options tailored to different tasks and thus provide a platform for cheap de novo sequence assemblies, eventually for all genomes.

Assessing Assembly Quality

Common measures of quality are:

- number and sizes of contigs
- Assumption: few large contigs is better than many small contigs.
- True because there are fewer gaps in the former, but, does not account for the possibility of misassembling.

Tools/ Software's for assembly

- TIGR Assembler
- Velvet (Denovo)
- Maq (Reference)
- Reference assembly& Alignment using BWA tool and Visualization of alignment using SAM

Applications of Genome assembly

- Generating and interpreting alignment status and reports
- Genome variation calling (finding SNP's, *indels*)
- Variation annotation and Viewing

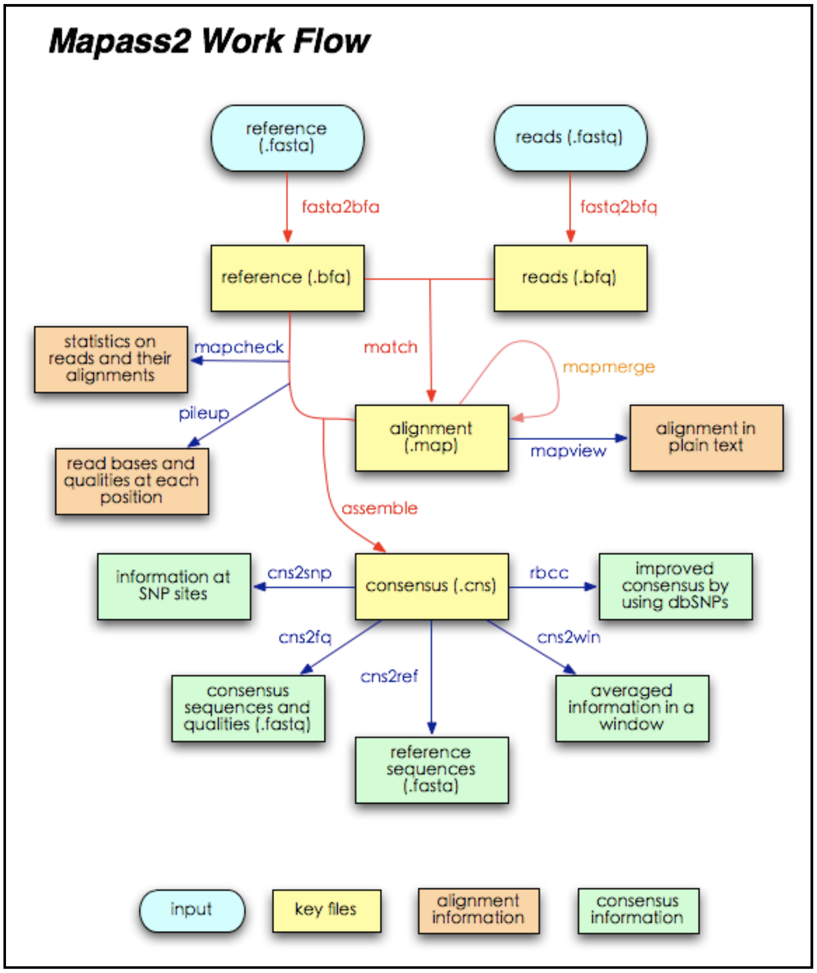
Practical

(i) Genome Alignment & Assembly with reference using MAQ tool

Maq is a software that builds mapping assemblies from short reads generated by the next-generation sequencing machines. It is particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has preliminary functions to handle ABI SOLiD data.

Maq first aligns reads to reference sequences and then calls the consensus. At the mapping stage, maq performs **ungapped** alignment. For single-end reads, maq is able to find all hits with up to 2 or 3 mismatches, depending on a command-line option; for paired-end reads, it always finds all paired hits with one of the two reads containing up to 1 mismatch. At the assembling stage, maq calls the consensus based on a statistical model. It calls the base which maximizes the posterior probability and calculates a phred quality at each position along the consensus. Heterozygotes are also called in this process.

Maq Workflow



MAQ Procedure

- Convert the reference sequences to the binary fasta format
  maq fasta2bfa reference.fasta reference.bfa

- Convert the reads1 to the binary fastq format
  maq fastq2bfq pairedfiltered1.fastq pairedfiltered1.bfq

- Convert the reads2 to the binary fastq format
  maq fastq2bfq pairedfiltered2.fastq pairedfiltered2.bfq

- Align the reads to the reference
  maq match pairedfiltered1.map reference.bfa pairedfiltered1.bfq
  maq match pairedfiltered2.map reference.bfa pairedfiltered2.bfq

- Merge a batch of read alignments together.

  ```
  maq mapmerge out.aln.map pairedfiltered1.map pairedfiltered2.map
  ```

- Statistics from the alignment

  ```
  maq mapcheck reference.bfa pairedfiltered1.map >>mapcheck1.txt
  maq mapcheck reference.bfa pairedfiltered2.map >>mapcheck2.txt
  maq mapcheck reference.bfa out.aln.map >>mapcheckmerged.txt
  ```

- Build the mapping assembly (writing consensus of sequences without variations)

  ```
  maq assemble consensus.cns reference.bfa pairedfiltered1.map >>assemble.log
  maq assemble consensus.cns reference.bfa pairedfiltered2.map >>assemble.log
  maq assemble consensus.cns reference.bfa out.aln.map >>assemble_merged.log
  ```

- Extract consensus sequences and qualities

  ```
  maq cns2fq consensus.cns >>cns.fq
  ```

- Extract list of SNPs

  ```
  maq cns2snp consensus.cns >>cns.snp
  ```

- Extract list of InDels

  ```
  maq indelpe reference.bfa out.aln.map >> cns.indelpe
  ```

- Show detailed information at all sites. The output format is identical to cns2snp report.

  ```
  maq cns2view consensus.cns > >out.view
  ```

MAQ Output

Several files will be generated by the commands. Endusers may be interested in these ones:

cns.fq	**Consensus sequences and their qualities**
cns.snp	**List of SNPs (format described in the reference manual).**
assemble.log	**Statistics about the coverage and SNPs**

Genome variation calling (finding SNP's, indels)

Variation annotation & Analysis

- Download reference structural annotation file in gff format
- Edit the Mapping_of_Variants_to_Genes.pl with corresponding name of SNP file and gff file
- Keep the **Mapping_of_Variants_to_Genes.pl, consensus.fasta,** structural annotation file in gff format in single folder and Run script

USAGE: **perl Mapping_of_Variants_to_Genes.pl > variation_annotation.txt**

- Open the variation_annotation.txt to see annotations

Further Reading

1. *Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20: 1165-1173*

2. *Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res 18: 810–820.*

3. *Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13: 91–96.*

Chapter 5

Structural & Functional Annotation

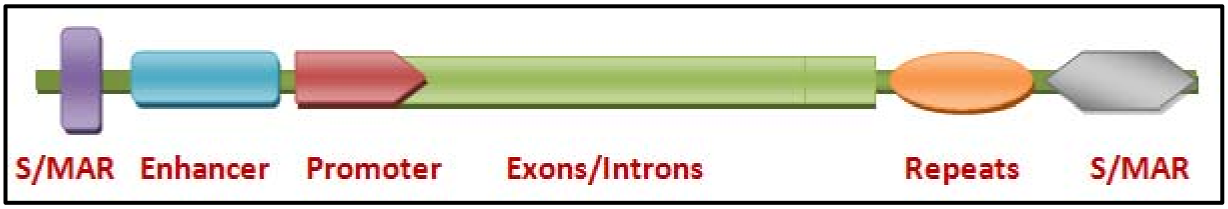
Rosana O. B.

Introduction

Obtaining the sequence from whole genome sequencing and assembling is not the end of a genome project, however finding and attaching the structural elements and its related function are the next major steps, which are called “Genome Annotation”. The process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it. Annotation is the process of adding pertinent information about the raw DNA sequences to the genome or process of attaching biological information to DNA or Protein sequences by describing different regions of the code and identifying which regions can be called genes and thereby its products and functions. This include spotting locations of genes, total number of genes, coding regions, intron-exon structure, start and stop codons, intron lengths, alternative splicing, SNP’s, InDels and untranslated regions (UTRs) as well as and determining what those genes do along with the gene product and functional information. Once a genome is sequenced, it needs to be annotated to make sense of it. With many genomes now sequenced, computational annotation methods to characterize genes and proteins from their sequence are increasingly important. It consists of two main steps:

- Identifying elements on the genome- gene structure prediction
- Attaching biological information to these elements- gene function prediction

There are various parts within the gene with different functions, some may code for protein, others may contain regulatory information, some may form introns and will not be translated and their function is still unclear. The diagram shown below represents fragment of DNA, with single hypothetical gene. Each region has to be annotated from DNA sequences based on similarity searches or literature reviews.



Obviously computer programs are essential to this process; however, human brains are often required to evaluate computer-generated gene models. Several Automatic annotation tools are available that are highly accurate. Annotation tools can perform all this by computer analysis, as opposed to manual annotation which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

The basic level of annotation uses BLAST for finding similarities, and then annotating genomes based on that. However, much additional information is available to annotation platform nowadays. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases such as Ensemble rely on both curated data sources as well as a range of different software tools in their automated genome annotation pipeline.

Annotation Methods

- *Ab-Initio* **-** Statistics-based methods
- Homology based methods

The first hurdle for any functional annotation process is to define 'function'. In general however, the problem is multi-dimensional: a protein can have a molecular function, a cellular role, and be part of a functional complex or pathway (these are the distinctions used in the Gene Ontology. Furthermore, certain aspects of molecular function can be illustrated by multiple descriptive levels (for example, the coarse 'enzyme' category versus a more specific 'protease' assignment). Even the more detailed definition would not reveal the cellular role of the protein (apoptosis, metabolism, blood coagulation, and so on). Most function-prediction methods, both sequence and structure based, rely on inferring relationships between proteins that permit the transfer of functional annotations and binding specificities from one to the other. A notable challenge here is deciphering the connection between the detected similarities (structural or in sequence) and the actual level of functional relatedness. Function is often associated with domains, and another problem is the identification of functional domains from sequence alone. The accuracy of current methods for predicting domain boundaries is not yet completely satisfactory. Several methods provide reliable predictions if a structural template for the protein is available, but when this is not the case, one is left with the problem of whether the experimental annotation used for the inference refers to the same domain for which the sequence similarity/motif is established.

The function of a protein can also be inferred from its evolutionary relationship with proteins of known function, provided that the relationship is properly inspected. Orthologous proteins in different species most often share function, but paralogy (that is, divergence following duplication of the original gene) does not guarantee common function. Distinguishing between orthology and paralogy can be attempted on the basis of observed sequence-similarity patterns, by analyzing the specific conservation pattern of residues responsible for function in the family, or on the basis of the protein structure (either experimentally determined or modeled). In all cases, this requires the clustering of proteins into evolutionary families, which can be achieved using similarity-detection tools such as BLAST or profiling tools based on multiple sequence alignments, for example, PSI-BLAST. Several available resources provide pre-compiled family assignments for proteins on a genomic scale, based only on their sequence. Resources can be subdivided into those that consider full-length sequences and those based on domains or motifs that map to certain sub-sequences. In both cases, the degree of granularity of the classification is important, as this is related to the level of functional features that a group of proteins is expected to share.

> The accuracy of sequence-based methods is affected by the type and amount of information on the specific protein family but, overall, they seem to be reasonably accurate. Their success rate has been shown to be greater than 70% when tested on a limited dataset (all structures solved by the Midwest Center for Structural Genomics during the first five years of the Protein Structure Initiative)

Structural & Functional Genome Annotation

Structural annotation consists of the identification of genomic elements.

- Open reading frame and their localization
- Gene structure
- Coding regions
- Location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- Biochemical function
- Biological function
- Involved regulation and interactions
- Expression

These steps may involve both biological experiments and *in-silico* analysis. A variety of software tools have been developed to permit scientists to view and share genome annotations.

- Structural annotation for prokaryotes & eukaryotes – PRODIGAL & AUGUSTUS
- Functional annotation (based gene level homology and protein level homology)

Practical

Structural Annotation

1. **Eukaryotes- AUGUSTUS (Gene structural prediction with a genome model)**

AUGUSTUS is used in many genome annotation projects. Accuracy measure use *sensitivity* (Sn) and *specificity* (Sp) for a feature (coding base, exon, transcript, gene); the sensitivity is defined as the number of correctly predicted features divided by the number of annotated features. The specificity is the number of correctly predicted features divided by the number of predicted features. A predicted exon is considered correct if both splice sites are at the annotated position of an exon. A predicted transcript is considered correct if all exons are correctly predicted and no additional exons not in the annotation. A predicted gene is considered correct if any of its transcripts are correct, i.e. if at least one isoform of the gene is exactly as annotated in the reference annotation.

Usage:

```
~/Programs/augustus.2.5.5/bin/augustus --strand=both --genemodel=partial --singlestrand=true --alternatives-from-evidence=true --alternatives-from-sampling=true --progress=true --gff3=on --uniqueGeneId=true --species=magnaporthe_grisea Input_genome.fasta > structural_annotation.gff
```

2. **Prokaryotes – PRODIGAL**

Prodigal (Prokaryotic Dynamic Programming **G**enefinding **Al**gorithm) is a microbial (bacterial and archaeal) gene finding program. Prodigal is an extremely fast gene recognition tool (written in very vanilla C). It can analyze an entire microbial genome in 30 seconds or less. Prodigal is a highly accurate gene finder. It correctly locates the 3' end of every gene in the experimentally verified Ecogene data set (except those containing introns). It possesses a very sophisticated ribosomal binding site scoring system that enables it to locate the translation initiation site with great accuracy (96% of the 5' ends in the Ecogene data set are located correctly). Prodigal's false positive rate compares favorably with other gene identification programs, and usually falls under 5%. Prodigal performs well even in high GC genomes, with over a 90% perfect match (5'+3') to the *Pseudomonas aeruginosa* curated annotations. Prodigal can run in metagenomic mode and analyze sequences even when the organism is unknown. Prodigal can be run in one step on a single genomic sequence or on a draft genome containing many sequences. It does not need to be supplied with any knowledge of the organism, as it learns all the properties it needs to on its own. Prodigal source code is freely available under the General Public License.

Usage:

```
~/Programs/prodigal.v2_60.linux        -a protein_file.fa -g 11 –d  nucleotide_exon_seq.fa

-f gff -i contigs.fa -o genes_quality.txt -s genes_score.txt -t genome_training_file.txt
```

Functional Annotation

Steps for functional annotation

Functional annotation requires creating database for homology searches. First download complete swissprot database and Complete NCBI and setup database for homology search using following commands.

Run separately each command for nucleotide and protein

1. setup blast database for nucleotide

```
formatdb -i ncbi.fasta -p F -o F -n database1
```

2. setup blast database for protein

```
formatdb -i uniprot_sprot.fasta -p T -o F -n database2
```

3. Blasting the annotations

(i) For nucleotide

```
blastall -p blastn -i sequence.fasta -d database1 -a 2 -m 8 -o blast_result_nucleotide.txt
```

(ii) For Protein

```
blastall -p blastx -i contigs.fa -d database2 -a 2 -m 8 -o blast_result_protein.txt
```

4. Sorting for blast min E-value

(i) For nucleotide

```
sort -k 1,1 -k 11,11g blast_result.txt > sorted_blast_result_nucleotide.txt
```

(ii) For Protein

```
sort -k 1,1 -k 11,11g blast_resul_prot.txt > sorted_blast_result_protein.txt
```

5. Further filtering for best blast hit

(i) For nucleotide

```
awk '{if($1!=id){print};id=$1}' sorted_blast_result_nucleotide.txt>best_blast_result1.txt
```

(ii) For Protein

```
awk '{if($1!=id){print};id=$1}' sorted_blast_result_protein.txt>best_blast_result2.txt
```

Further Reading

1. *Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25:25-29.*

2. *Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo JH, Lee J, Baker D, Chivian D, Kim D, Ezkurdia I: Assessment of predictions submitted for the CASP7 domain prediction category.Proteins 2007, 69(Suppl 8):137-151.*

3. *Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J MolBiol 1990, 215:403-410.*

4. *Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: Towards fully automated structure-based function prediction in structural genomics: a case study. J MolBiol 2007, 367:1511-1522.*

Chapter 6

Plant Transcriptome Mapping

Dr. D. Prasath & Rosana O. Babu

Introduction

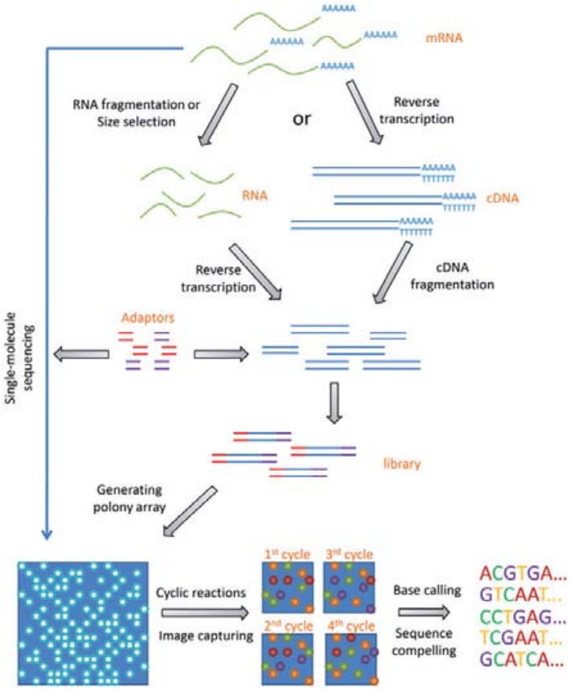
The transcriptome encompasses the set of transcripts from a cell or a population of cells, which include protein-coding mRNAs and non-coding small RNAs (e.g. ribosomal, tRNA, miRNA). Traditionally, transcriptome profiling, or transcriptomics, has focused on quantifying gene expression. With the advent of Ultra high-throughput sequencing (UHTS) technologies, it is now possible to obtain highly resolved structural information of RNA populations on a high-throughput platform. This includes mapping transcript initiation and termination sites, splice junctions and post-transcriptional modifications. Such information will lead to a better understanding of the functional elements within the genome and the discovery of novel developmental or environmental regulatory networks.

EST sequencing has traditionally been the core technology used for the discovery of reference transcripts. However, it has some inherent limitations, such as low throughput, high cost and a long experimental cycle. In recent years, researchers have developed a high-throughput sequencing technology called Next Generation Sequencing (NGS). Various platforms utilize NGS, such as the Illumina Genome Analyzer, the Roche/454 Genome Sequencer FLX Instrument, and the ABI SOLiD System; these have proven to be powerful and cost-effective tools for advanced research in many areas, including genome sequencing, genome resequencing, miRNA expression profiling, DNA methylation analysis, and especially the de novo transcriptome sequencing of non-model organisms. This method of transcriptome analysis is fast and simple because it does not require bacterial cloning of the cDNAs. Instead, direct cDNA sequencing generates an extraordinary depth of short reads. It is a more comprehensive and efficient way to measure transcriptome composition, obtain RNA expression patterns, and discover new genes. In addition, this approach is very sensitive, and thus allows the detection of low-abundance transcripts. Illumina genome analyzer based sequencing technology (Illumina, USA) yields huge amount of short reads with high coverage. Assembling such short reads is a challenging task, more so in the absence of reference sequences. A few bioinformatics tools have been developed for de novo assembly using short-read sequence data which vary in their success and application, and depends upon data specific strategies.

RNA sequencing

RNA-seq refers to the deep-sequencing of RNA pools. While UHTS often refers to deep sequencing of mRNAs, any RNA population can be analyzed. Two methods are typically used to capture and sequence RNA pools. In both methods mRNA pools are enriched by capturing the molecules through the polyadenylated tails, and a ribosomal RNA removal step is often added before or after the mRNA purification. In one method, mRNA enriched pools are then fragmented into roughly equal lengths and then reverse-transcribed using random hexamers to generate a cDNA library. Alternatively, RNA is reverse transcribed using an oligo-dT adapter and the resulting cDNA is fractionated. The former method has the advantage of more uniform representation across the coding region of the transcript, but may result in the under-representation 50 and 30 sequences. The later method provides good coverage of 30 sequences but biases against the body of the transcript. The cDNAs are then fitted with adaptors at one or both ends through a ligation step(s). It is desirable to add these adaptors during the single-strand stage (RNA or cDNA) synthesis step in order to retain strand specificity in the final sequence reads. The tagged cDNA library is subsequently amplified through PCR before being sequenced.

For a typical RNA-seq experiment, mRNA is isolated and reverse-transcribed (RT) into cDNA libraries with homogeneous lengths. This is achieved by either RNA or cDNA fragmentation. Recently, single-molecule capture methods have been developed (left) that obviate the need for a RT step. In the case of smRNA studies, total RNA is fractionated on acrylamide gels and smRNAs excised. Adaptors at one or both ends of the RNA are added prior to cDNA amplification and library construction. For the Solexa platform, cDNA molecules are anchored onto a polony array (flow cell) surface, which are then subjected to PCR amplification. Images are taken after each cycle for base calling and sequence generation. Currently for the Illumina platform, ~120 million single or paired-end reads of 32 - 80 nt are generated on a single flow cell (8 lanes/flow cell) which is then processed further depending on the research goals.



Overview of RNA-seq experimental procedures.

RNA sequencing analysis

Overview of RNA-seq data analysis

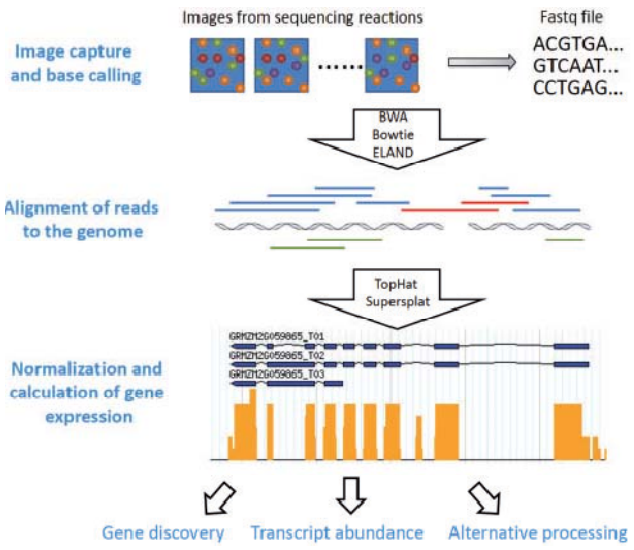


Image capture and base calling
Images from sequencing reactions
Fastq file
ACGTGA...
GTCAAT...
CCTGAG...
BWA
Bowtie
ELAND
Alignment of reads to the genome
TopHat
Supersplat
Normalization and calculation of gene expression
Gene discovery
Transcript abundance
Alternative processing

Although the volume of data from RNA-seq experiments is often burdensome, it can provide enormous insight. Just as cDNA sequencing with Sanger sequencers drastically expanded our catalog of known human genes, RNA-seq reveals the full repertoire of alternative splice isoforms in our transcriptome and sheds light on the rarest and most cell- and context-specific transcripts6. Furthermore, because the number of reads produced from an RNA transcript is a function of that transcript's abundance, read density can be used to measure transcript and gene expression with comparable or superior accuracy to expression microarrays. RNA-seq experiments must be analyzed with robust, efficient and statistically principled algorithms. Fortunately, the bioinformatics community has been hard at work developing mathematics, statistics and computer science for RNA-seq and building these ideas into software tools.

RNA-seq analysis tools generally fall into three categories:

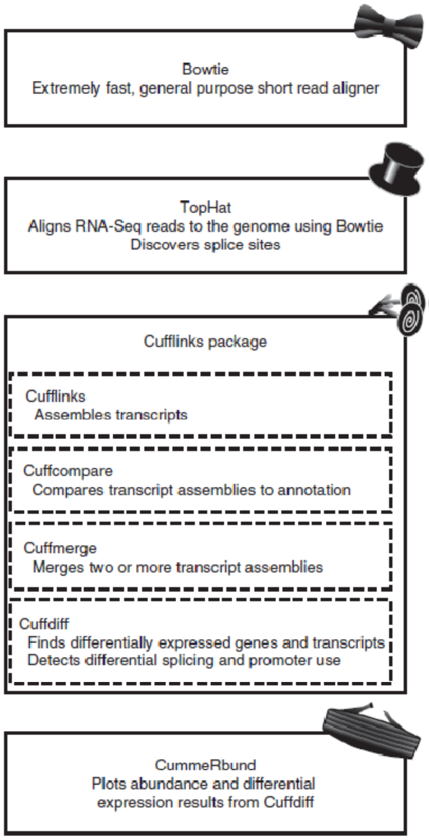
(i) those for read alignment;

(ii) those for transcript assembly or genome annotation; and

(iii) those for transcript and gene quantification.

Differential gene and transcript expression analysis

The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNAsequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. TopHat and Cufflinks performs such analyses, also covers several


Bowtie
Extremely fast, general purpose short read aligner
TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites
Cufflinks package
Cufflinks
Assembles transcripts
Cuffcompare
Compares transcript assemblies to annotation
Cuffmerge
Merges two or more transcript assemblies
Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use
CummeRbund
Plots abundance and differential expression results from Cuffdiff

accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results.

Ginger transcriptome analysis: A case study

Genomic tools are now being developed to accelerate the identification of resistance genes and the development of bacterial wilt resistant ginger. The present study describes the first global analysis of ginger and mango ginger *R. solanacearum* challenge inoculated regimes, which would serve as a blueprint of gene expression profile.

Ginger (*Zingiber officinale*) is affected by many diseases. Among them, bacterial wilt (*Ralstonia solanacearum)* is one of the most important production constraints in tropical, sub tropical and warm temperature regions of the world. In spite of extensive search, no resistance source could be located in ginger. However, mango ginger *(Curcuma amada)*, which is resistant to *R. solanacearum*, is a potential donor, if the exact mechanism of resistance is understood. Although ESTs and other cDNA sequences are among the most reliable evidences for the identification of gene-rich regions in a genome, gene identification and genome annotation, very less effort has been made for ginger-pathogen interaction. Our objective is deep sequencing of transcriptomes of both genotypes followed by *de novo* assembly and annotation.

The transcriptome has been sequenced from rhizome tissue samples after challenge inoculation with *R. solanacearum* using next generation technologies for short read (Illumina platform). It involves analysis of GC content, repeat content, putative functions, gene families, transcription factor encoding genes and its comparative analysis with and *C. amada*. The identification of many defense related genes differentially expressed provides many insights to resistance mechanism to *R. solanacearum* and for studying pathways involved in responses to pathogen. We also identified several candidate genes that may underline the difference in resistance to *R. solanacearum* between ginger and mango ginger.

Conclusion

Despite these challenges, UHTS-based transcriptomics approaches promise ‘never-before’ opportunities to explore plant transcriptomes. As improvements to the sequencing chemistry, sequencing hardware and software and statistical methods of analysis continue to progress, the expectations for transcriptomics studies will continue to increase. It has been speculated that the cost of sequencing a complete genome or transcriptome will not be a limiting factor in the foreseeable future. This may allow experiments that were deemed as economically unfeasible in the past to be routinely performed. In summary, UHTS-based approaches have clearly demonstrated their advantages over previously developed methods and are becoming the new standard for transcriptomics studies.

Practical

1. Blast2GO analysis of Transcriptome
2. Conserved Domain Search

Further Reading

1. *Annadurai RS, Neethiraj R, Jayakumar V, Damodaran AC, Rao SN, et al. (2013) De Novo Transcriptome Assembly (NGS) of Curcuma longa L. Rhizome Reveals Novel Transcripts Related to Anticancer and Antimalarial Terpenoids. PLoS ONE 8(2): e56217. doi:10.1371/journal.pone.0056217*

2. *Barakat A, DiLoreto D, Zhang Y, Smith C, Baier K, Powell W, Wheeler N, Sederoff R, Carlson J. 2009. Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. BMC Plant Biology, 9(1): 51.*

3. *Garg R, Patel RK, Jhanwar S, Priya P, Bhattacharjee A, Yadav G, Bhatia S, Chattopadhyay D, Tyagi AK, Jain M (2011). Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol., 156: 1661-1678.*

4. *Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 15: 29(7): 644-52.*

5. *Jhanwar S, Priya P, Garg R, Parida SK, Tyagi AK, Jain M (2012). Transcriptome sequencing of wild chickpea as a rich resource for marker development. Plant Biotechnology J., 10: 690-702*

6. *Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics *12: 323.*

7. *Trapnell, C. et al. 2012. Differential gene and transcript expression analysis of RNA-seqexperiments with TopHatand Cufflinks. Nature Protocols 7: 562–578.*

8. *Wang L, Li P, Brutnell TP (2010) Exploring plant transcriptomes using ultra high-throughput sequencing. Briefings in Functional Genomics 9: 118–128.*

9. *Tophat Manual: http://tophat.cbcb.umd.edu/manual.html*

10. *Cufflinks Manual: http://cufflinks.cbcb.umd.edu/manual.html*

Chapter 7

Genotyping with RAD and ddRAD Sequencing

Dr. VL Ramprasad

Genotyping requires thousands of genomes to be compared in a reliable, consistent way. Restriction site associated DNA sequencing (RAD-Seq) interrogates a fraction of the genome across many individuals, an ideal method for genotyping. By using restriction enzyme digestion and sequencing the regions adjacent to restriction sites, researchers can examine the same subset of genomic regions for thousands of individuals and identify many genetic markers along the genome. Other NGS methods examine a larger portion of the genome and offer more data, but they are costly and cannot be used to study the thousands of individuals required for genotyping. RAD-Seq applications include: Genetic marker discovery, Local genome assembly, QTL mapping, Linkage mapping.

SciGenom uses double digest RAD-Seq (ddRAD-Seq), a variation of RAD-Seq, for genotyping. Traditional RAD-Seq uses one restriction enzyme and random shearing to generate fragments from genomic DNA. However, these are high DNA loss steps and offer little control over the fragments that are sequenced. For organisms without a reference genome, a significant portion of the RAD-Seq data has been discarded due to sequence read errors and the presence of variable sites.ddRAD-Seq was designed to address RAD-Seq short-comings. In ddRAD-Seq, genomic DNA is digested with two restriction enzymes, and the resulting fragments undergo adaptor ligations and precise size selection before sequencing. Only a very small fraction of the fragments will be sequenced. These fragments are naturally selected to be from the same genomic regions across individuals.

Further, ddRAD requires half as many reads to achieve high confidence SNP calling, because the chance of obtaining duplicate reads from the same restriction site are very low. Due to these modifications, ddRAD has become a more economical method to genotype thousands of individuals, and has been used for SNP discovery between two *Peromyscus*species that have no reference sequence.

Chapter 8

Whole genome sequencing, assembly and annotation of bacteria

Aravind Ravindran & Dennis C. Gross

Abstract

The whole-genome sequencing of bacteria used two high-throughput sequencing techniques, namely 454 pyrosequencing and Illumina Solexa sequencing. Single and paired-end reads were generated on a 454 GS-FLX Titanium sequencer, and Illumina sequence reads were obtained using Illumina genome analyzer. Newbler 2.0 software was used for *de novo* assembles of 454 sequences into contigs and scaffolds. CLC bio Genomics Workbench version 4.9 software were used to align Illumina contigs against the 454 scaffolds to confirm the orientations and integrity of the assembled sequences and to close gaps and link contigs together within the scaffold. OpGen technologies (Madison, WI) generated a *de novo* assembled optical map using *KpnI* restriction sites of the genome. *In silico KpnI* restriction maps of the scaffolds were constructed and aligned to the optical map according to their restriction fragment pattern by using MapSolver v.3.1 software (OpGen Technologies, Inc.). PCR primers were designed to close the gaps by amplified the PCR products and sequenced; these sequences were helpful in joining the overlapping contigs between the scaffolds. Assembly was correlated with the optical map for further validation, and the assembled whole genome sequence was submitted to the Integrated Microbial Genomes Expert Review (IMG/ER) system for annotation by gene calling. The annotated genomes were submitted to the JGI GenePRIMP pipeline for identification of the short, long, broken and interrupted genes to be corrected using Artemis. The Final whole-genome was submitted in IMG/ER for comparative genome studies.

Chapter 9

Data Analysis Using CLC bio Genomic Workbench

Dr. Uday Deshpande

Black Pepper Root Transcriptome

The dataset of raw reads from the root transcriptome of black pepper deposited in NCBI database under SRA047721 accession number (Gordo et al., 2012) was reanalyzed using CLC bio Genomic Workbench ver 6. The black pepper root transcriptome was sequenced on SoLiD NGS platform and contains about 71878419 reads 50 bp in length totaling 3.5 Gbp.

Reads were assembled using De Novo assembly tool in CLC bio Genomic Workbench. Blast2Go tools were used to annotate 1unigenes. The predicted proteins showed significant homology with the Arabidopsis proteome. This presentation on black pepper root transcriptome data analysis explains the transcriptome data analysis pipeline and demonstrates the capability of tools in CLC bio Genomic workbench for the genomic data analysis. Finally comparative benchmarking results CLC bio Genomic Workbench with respect to original published results is presented.

Further Reading

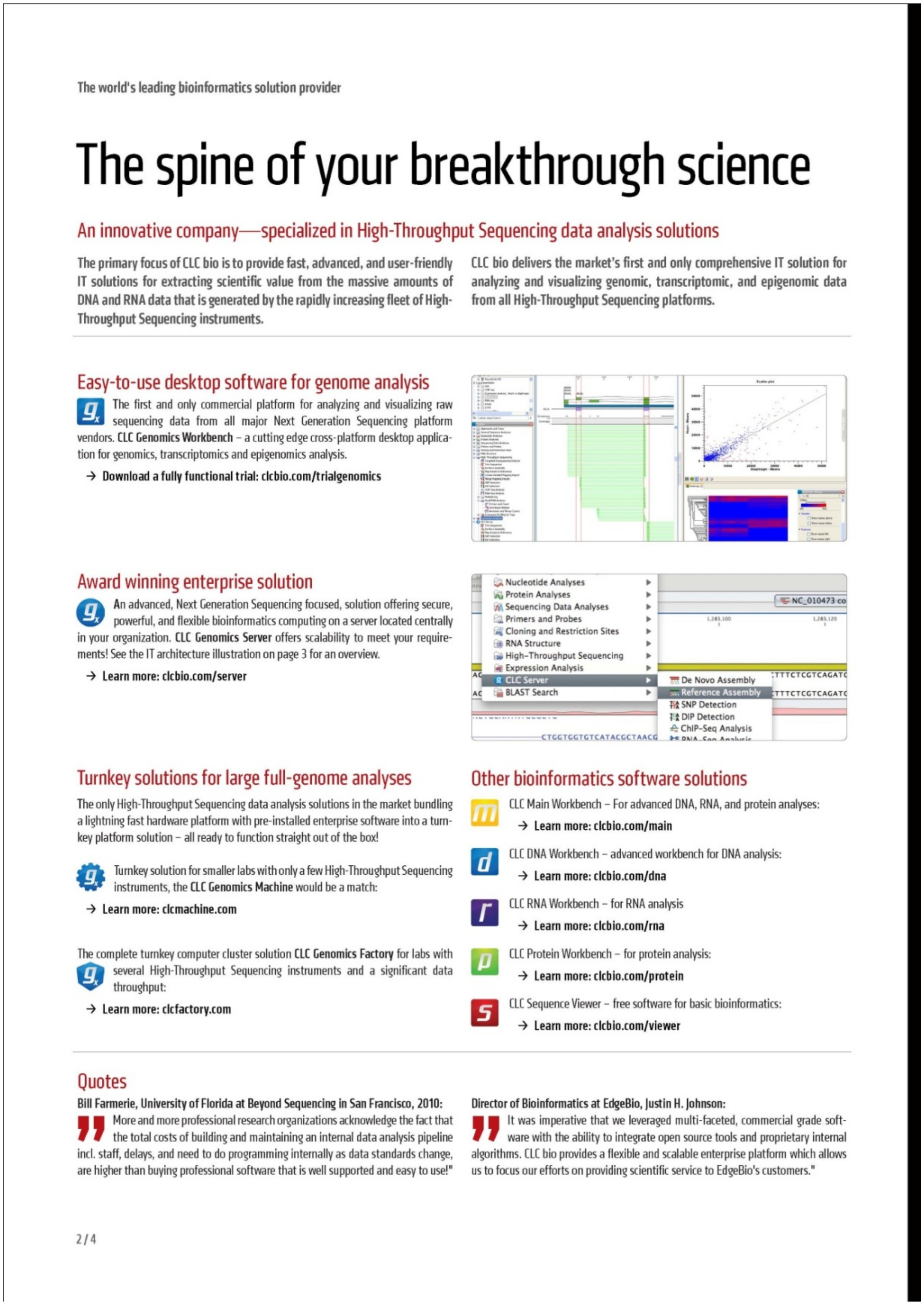
Gordo et al. BMC Plant Biology 2012, 12:168, 2 -9

http://www.biomedcentral.com/1471-2229/12/168

Practical on *CLC Work Bench*

Beyond imagination

CLC bio – helping you discover new science



The world's leading bioinformatics solution provider

The spine of your breakthrough science

An innovative company—specialized in High-Throughput Sequencing data analysis solutions

The primary focus of CLC bio is to provide fast, advanced, and user-friendly IT solutions for extracting scientific value from the massive amounts of DNA and RNA data that is generated by the rapidly increasing fleet of High-Throughput Sequencing instruments.

CLC bio delivers the market's first and only comprehensive IT solution for analyzing and visualizing genomic, transcriptomic, and epigenomic data from all High-Throughput Sequencing platforms.

Easy-to-use desktop software for genome analysis

The first and only commercial platform for analyzing and visualizing raw sequencing data from all major Next Generation Sequencing platform vendors. **CLC Genomics Workbench** – a cutting edge cross-platform desktop application for genomics, transcriptomics and epigenomics analysis.

→ **Download a fully functional trial: clcbio.com/trialgenomics**

Award winning enterprise solution

An advanced, Next Generation Sequencing focused, solution offering secure, powerful, and flexible bioinformatics computing on a server located centrally in your organization. **CLC Genomics Server** offers scalability to meet your requirements! See the IT architecture illustration on page 3 for an overview.

→ **Learn more: clcbio.com/server**

Turnkey solutions for large full-genome analyses

The only High-Throughput Sequencing data analysis solutions in the market bundling a lightning fast hardware platform with pre-installed enterprise software into a turnkey platform solution – all ready to function straight out of the box!

Turnkey solution for smaller labs with only a few High-Throughput Sequencing instruments, the **CLC Genomics Machine** would be a match:

→ **Learn more: clcmachine.com**

The complete turnkey computer cluster solution **CLC Genomics Factory** for labs with several High-Throughput Sequencing instruments and a significant data throughput:

→ **Learn more: clcfactory.com**

Other bioinformatics software solutions

CLC Main Workbench – For advanced DNA, RNA, and protein analyses:

→ **Learn more: clcbio.com/main**

CLC DNA Workbench – advanced workbench for DNA analysis:

→ **Learn more: clcbio.com/dna**

CLC RNA Workbench – for RNA analysis

→ **Learn more: clcbio.com/rna**

CLC Protein Workbench – for protein analysis:

→ **Learn more: clcbio.com/protein**

CLC Sequence Viewer – free software for basic bioinformatics:

→ **Learn more: clcbio.com/viewer**

Quotes

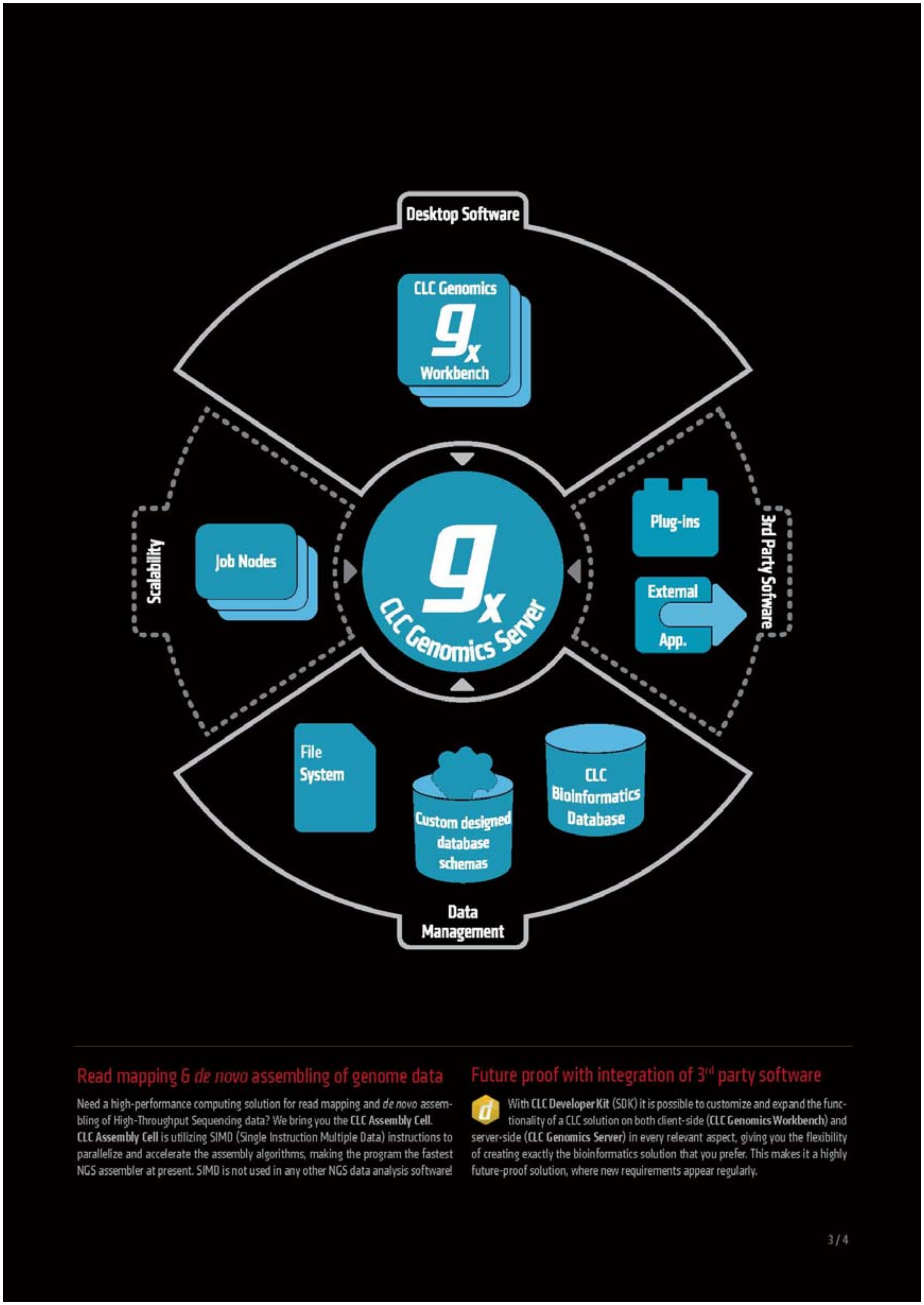
Bill Farmerie, University of Florida at Beyond Sequencing in San Francisco, 2010:

" More and more professional research organizations acknowledge the fact that the total costs of building and maintaining an internal data analysis pipeline incl. staff, delays, and need to do programming internally as data standards change, are higher than buying professional software that is well supported and easy to use!"

Director of Bioinformatics at EdgeBio, Justin H. Johnson:

" It was imperative that we leveraged multi-faceted, commercial grade software with the ability to integrate open source tools and proprietary internal algorithms. CLC bio provides a flexible and scalable enterprise platform which allows us to focus our efforts on providing scientific service to EdgeBio's customers."

2 / 4



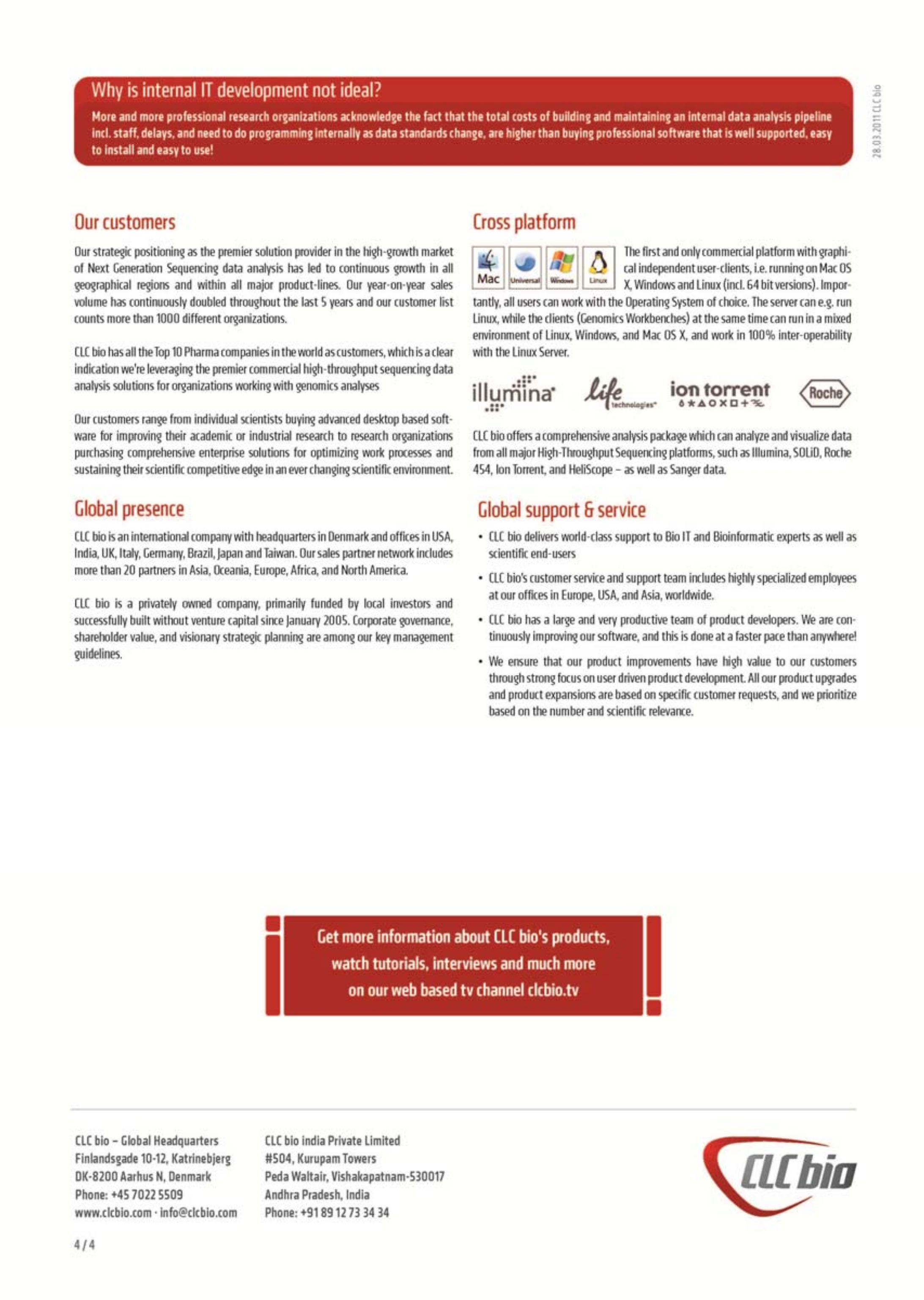
Read mapping & *de novo* assembling of genome data

Need a high-performance computing solution for read mapping and *de novo* assembling of High-Throughput Sequencing data? We bring you the **CLC Assembly Cell**. **CLC Assembly Cell** is utilizing SIMD (Single Instruction Multiple Data) instructions to parallelize and accelerate the assembly algorithms, making the program the fastest NGS assembler at present. SIMD is not used in any other NGS data analysis software!

Future proof with integration of 3rd party software

With **CLC Developer Kit** (SDK) it is possible to customize and expand the functionality of a CLC solution on both client-side (**CLC Genomics Workbench**) and server-side (**CLC Genomics Server**) in every relevant aspect, giving you the flexibility of creating exactly the bioinformatics solution that you prefer. This makes it a highly future-proof solution, where new requirements appear regularly.

3 / 4



Why is internal IT development not ideal?

More and more professional research organizations acknowledge the fact that the total costs of building and maintaining an internal data analysis pipeline incl. staff, delays, and need to do programming internally as data standards change, are higher than buying professional software that is well supported, easy to install and easy to use!

28.03.2011 CLC bio

Our customers

Our strategic positioning as the premier solution provider in the high-growth market of Next Generation Sequencing data analysis has led to continuous growth in all geographical regions and within all major product-lines. Our year-on-year sales volume has continuously doubled throughout the last 5 years and our customer list counts more than 1000 different organizations.

CLC bio has all the Top 10 Pharma companies in the world as customers, which is a clear indication we're leveraging the premier commercial high-throughput sequencing data analysis solutions for organizations working with genomics analyses

Our customers range from individual scientists buying advanced desktop based software for improving their academic or industrial research to research organizations purchasing comprehensive enterprise solutions for optimizing work processes and sustaining their scientific competitive edge in an ever changing scientific environment.

Global presence

CLC bio is an international company with headquarters in Denmark and offices in USA, India, UK, Italy, Germany, Brazil, Japan and Taiwan. Our sales partner network includes more than 20 partners in Asia, Oceania, Europe, Africa, and North America.

CLC bio is a privately owned company, primarily funded by local investors and successfully built without venture capital since January 2005. Corporate governance, shareholder value, and visionary strategic planning are among our key management guidelines.

Cross platform

The first and only commercial platform with graphical independent user-clients, i.e. running on Mac OS X, Windows and Linux (incl. 64 bit versions). Importantly, all users can work with the Operating System of choice. The server can e.g. run Linux, while the clients (Genomics Workbenches) at the same time can run in a mixed environment of Linux, Windows, and Mac OS X, and work in 100% inter-operability with the Linux Server.

CLC bio offers a comprehensive analysis package which can analyze and visualize data from all major High-Throughput Sequencing platforms, such as Illumina, SOLiD, Roche 454, Ion Torrent, and HeliScope – as well as Sanger data.

Global support & service

- CLC bio delivers world-class support to Bio IT and Bioinformatic experts as well as scientific end-users
- CLC bio's customer service and support team includes highly specialized employees at our offices in Europe, USA, and Asia, worldwide.
- CLC bio has a large and very productive team of product developers. We are continuously improving our software, and this is done at a faster pace than anywhere!
- We ensure that our product improvements have high value to our customers through strong focus on user driven product development. All our product upgrades and product expansions are based on specific customer requests, and we prioritize based on the number and scientific relevance.

Get more information about CLC bio's products, watch tutorials, interviews and much more on our web based tv channel clcbio.tv

CLC bio – Global Headquarters
Finlandsgade 10-12, Katrinebjerg
DK-8200 Aarhus N, Denmark
Phone: +45 7022 5509
www.clcbio.com · info@clcbio.com

CLC bio india Private Limited
#504, Kurupam Towers
Peda Waltair, Vishakapatnam-530017
Andhra Pradesh, India
Phone: +91 89 12 73 34 34

4 / 4

Chapter 10

Comparative Genomics

Dr. Santhosh J. Eapen & Rosana O. B.

Introduction

What is comparative genomics?

Comparative genomics is an exciting new field of biological research in which the genome sequences of different species - human, mouse and a wide variety of other organisms from yeast to chimpanzees - are compared. Comparative genomics is the study of the relationship of genome structure and function across different biological species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes. While it is still a young field, it holds great promise to yield insights into many aspects of the evolution of modern species. The sheer amount of information contained in modern genomes (3.2 gigabases in the case of humans) necessitates that the methods of comparative genomics are automated. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome.

By comparing the finished reference sequence of the human genome with genomes of other organisms, researchers can identify regions of similarity and difference. This information can help scientists better understand the structure and function of human genes and thereby develop new strategies to combat human disease. Comparative genomics also provides a powerful tool for studying evolutionary changes among organisms, helping to identify genes that are conserved among species, as well as genes that give each organism its unique characteristics.

Human FOXP2 gene and evolutionary conservation is shown in and multiple alignment (at bottom of figure) in this image from the UCSC Genome Browser. Note that conservation tends to cluster around coding regions (exons).

Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved.

One of the important goals of the field is the identification of the mechanisms of eukaryotic genome evolution. It is however often complicated by the multiplicity of events that have taken place throughout the history of individual lineages, leaving only distorted and superimposed traces in the genome of each living organism. For this reason comparative genomics studies of small model organisms (for example the model *Caenorhabditis elegans* and closely related *Caenorhabditis briggsae*) are of great importance to advance our understanding of general mechanisms of evolution.

Having come a long way from its initial use of finding functional proteins, comparative genomics is now concentrating on finding regulatory regions and siRNA molecules. Recently, it has been discovered that distantly related species often share long conserved stretches of DNA that do not appear to code for any protein (see conserved non-coding sequence). One such ultra-conserved region, that was stable from chicken to chimp has undergone a sudden burst of change in the human lineage, and is found to be active in the developing brain of the human embryo.

Computational approaches to genome comparison have recently become a common research topic in computer science. A public collection of case studies and demonstrations is growing, ranging from whole genome comparisons to gene expression analysis. This has increased the introduction of different ideas, including concepts from systems and control, information theory, strings analysis and data mining. It is anticipated that computational approaches will become and remain a standard topic for research and teaching, while multiple courses will begin training students to be fluent in both topics.

Benefits of comparative genomics

Dramatic results have emerged from the rapidly developing field of comparative genomics. Comparison of the fruit fly genome with the human genome reveals that about sixty percent of genes are conserved (Adams *et al.* 2000). That is, the two organisms appear to share a core set of genes. Researchers have also found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly.

Using computer-based analysis to zero in on the genomic features that have been preserved in multiple organisms over millions of years, researchers will be able to pinpoint the signals that control gene function, which in turn should translate into innovative approaches for treating human disease and improving human health. In addition to its implications for human health, comparative genomics may benefit the broader animal world and ecological studies as well. As sequencing technology grows easier and less expensive, it will find wide applications in agriculture, biotechnology, and zoology as a tool to tease apart the often-subtle differences among animal and plant species. Such efforts might also lead to the rearrangement of our understanding of some branches of the evolutionary "tree of life," as well as point to new strategies for conserving rare and endangered species.

Why is there an increased interest in genomics?

Although living creatures look and behave in many different ways, all of their genomes consist of DNA, the chemical chain that makes up the genes that code for thousands of different kinds of proteins. Precisely which protein is produced by a given gene is determined by the sequence in which four chemical building blocks - adenine (A), thymine (T), cytosine (C) and guanine (G) - are laid out along DNA's double-helix structure.

In order for researchers to use an organism's genome most efficiently in comparative studies, data about its DNA must be in large, contiguous segments, anchored to chromosomes and, ideally, fully sequenced. Furthermore, the data needs to be organized to allow easy access for researchers using sophisticated computer software to conduct high-speed analyses.

The successful completion of the Human Genome Project in April 2003 has demonstrated that large-scale sequencing projects can generate high-quality data at a reasonable cost. As a result, the interest in sequencing the genomes of many other organisms has risen dramatically.

What other genomes have been sequenced?

In addition to sequencing the 3 billion letters in the human genetic instruction book, researchers involved in the Human Genome Project have already sequenced the genomes of a number of important model organisms that are commonly used as surrogates in studying human biology. These are the chimpanzee, the mouse, the rat, two puffer fish, two fruit flies, two sea squirts, two roundworms, baker's yeast and the bacterium *Escherichia coli*. Currently, sequencing centers supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) are close to completing working drafts of the chicken, the dog, the honey bee, the sea urchin and a set of four fungi. In the summer of 2003, the centers also began sequencing the genome of the rhesus macaque monkey, and many other organisms are in the sequencing pipeline.

Major outcomes of comparative genomics

The rapidly emerging field of comparative genomics has already yielded dramatic results. For example, a March 2000 study comparing the fruit fly genome with the human genome discovered that about 60 percent of genes are conserved between fly and human. Or, to put it simply, the two organisms appear to share a core set of genes.

Researchers have found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. Even more surprisingly, when scientists inserted a human gene associated with early-onset Parkinson's disease into fruit flies, they displayed symptoms similar to those seen in humans with the disorder, raising the possibility the tiny insects could serve as a new model for testing therapies aimed at Parkinson's.

More recently, a comparative genomic analysis of six species of yeast prompted scientists to significantly revise their initial catalog of yeast genes and to predict a new set of functional elements thought to play a role in regulating genome activity.

How Are Genomes Compared?

A simple comparison of the general features of genomes such as genome size, number of genes, and chromosome number presents an entry point into comparative genomic analysis. Data for several fully-sequenced model organisms is shown in Table 1. The comparisons highlight some striking findings. For example, while the tiny flowering plant Arabidopsis thaliana has a smaller genome than that of the fruit fly Drosophila melanogaster (157 million base pairs v. 165 million base pairs, respectively) it possesses nearly twice as many genes (25,000 v. 13,000). In fact A. thaliana has approximately the same number of genes as humans (~25,000). Thus, a very early lesson learned in the "genomic era" is that genome size does not correlate with evolutionary status, nor is the number of genes proportionate to genome size.

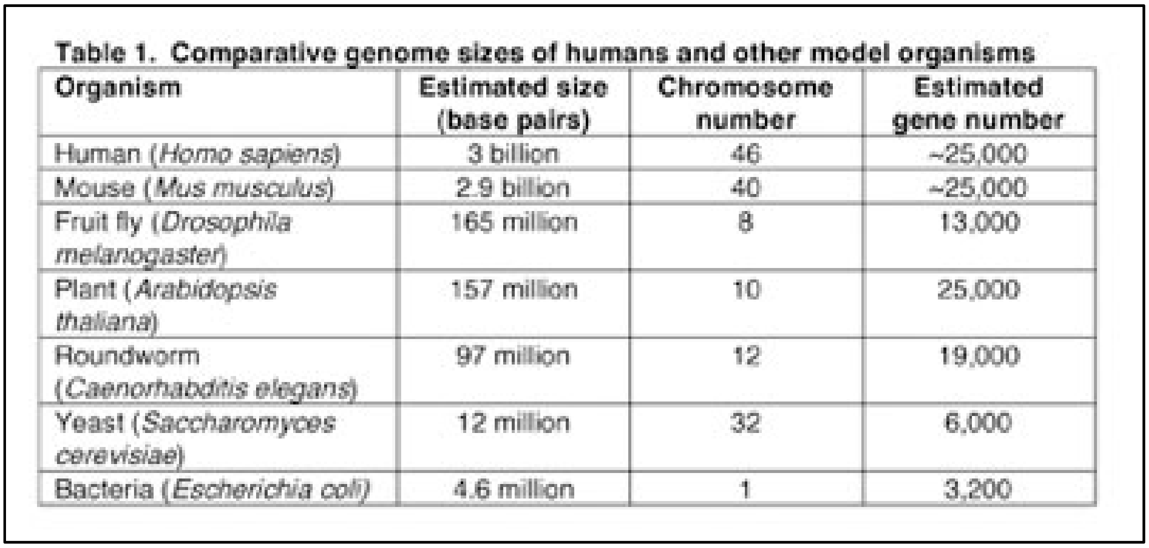


Table 1. Comparative genome sizes of humans and other model organisms

Organism	Estimated size (base pairs)	Chromosome number	Estimated gene number
Human (*Homo sapiens*)	3 billion	46	~25,000
Mouse (*Mus musculus*)	2.9 billion	40	~25,000
Fruit fly (*Drosophila melanogaster*)	165 million	8	13,000
Plant (*Arabidopsis thaliana*)	157 million	10	25,000
Roundworm (*Caenorhabditis elegans*)	97 million	12	19,000
Yeast (*Saccharomyces cerevisiae*)	12 million	32	6,000
Bacteria (*Escherichia coli*)	4.6 million	1	3,200

Table 1: Comparative genome sizes of humans and other model organisms

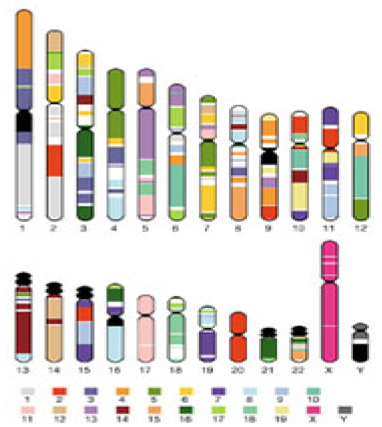


Figure 1: Conserved segments in the human and mouse genome

Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as color blocks. Each color corresponds to a particular mouse chromosome. Centromeres, are arranged in similar blocks in different species. The nature and extent of conservation of synteny differs substantially among chromosomes. For example, the X chromosomes are represented as single, reciprocal syntenic blocks. Human chromosome 20 corresponds entirely to a portion of mouse chromosome 2, with nearly perfect conservation of order along almost the entire length, disrupted only by a small central segment. Human chromosome 17 corresponds entirely to a portion of mouse chromosome 11. Other chromosomes, however, show evidence of more extensive inter chromosomal rearrangement. Results such as these provide an extraordinary glimpse into the chromosomal changes that have shaped the mouse and human genomes since their divergence from a common ancestor 75–80 million years ago.

Comparison of discrete segments of genomes is also possible by aligning homologous DNA from different species. An example of such an alignment is shown in Figure 2, where a human gene (pyruvate kinase: PKLR) and the corresponding PKLR homologs from macaque, dog, mouse, chicken, and zebrafish are aligned. Regions of high DNA sequence similarity with human across a 12-kilobase region of the PKLR gene are plotted for each organism. Notice the high degree of sequence similarity between human and macaque (two primates) in both PKLR exons (blue) as well as introns (red) and untranslated regions (light blue) of the gene. In contrast, the chicken and zebra fish alignments with human only show similarity to sequences in the coding exons; the rest of the sequence has diverged to a point where it can no longer be reliably aligned with the human DNA sequence. Using such computer-based analysis to zero in on the genomic features that have been preserved in multiple organisms over millions of years, researchers are able to locate the signals that represent the location of genes, as well as sequences that may regulate gene expression. Indeed, much of the functional parts of the human genome have been discovered or verified by this type of sequence comparison (Lander *et al.* 2001) and it is now a standard component of the analysis of every new genome sequence.

subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black. (International Human Genome Sequencing Consortium; Lander, E. S. et al. 2001)

Finer-resolution comparisons are possible by direct DNA sequence comparisons between species. Figure 1 depicts a chromosome-level comparison of the human and mouse genomes that shows the level of synteny between these two mammals. Synteny is a situation in which genes

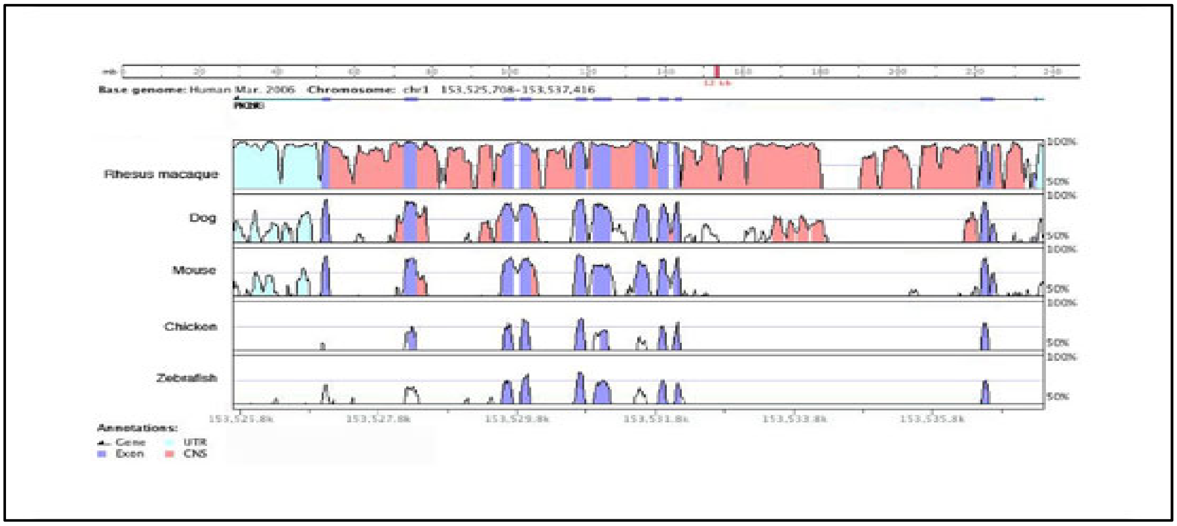


Figure 2: Human PKLR gene region compared to the macaque, dog, mouse, chicken, and zebra fish genomes

Numbers on the vertical axis represent the proportion of identical nucleotides in a 100-bp window for a point on the plot. Numbers on the horizontal axis indicate the nucleotide position from the beginning of the 12-kilobase human genomic sequence. Peaks shaded in blue correspond to the PKLR coding regions. Peaks shaded in light blue correspond to PKLR mRNA untranslated regions. Peaks shaded in red correspond to conserved non-coding regions (CNSs), defined as areas where the average identity is > 75%. Alignment was generated using the sequence comparison tool VISTA (http://pipeline.lbl.gov).

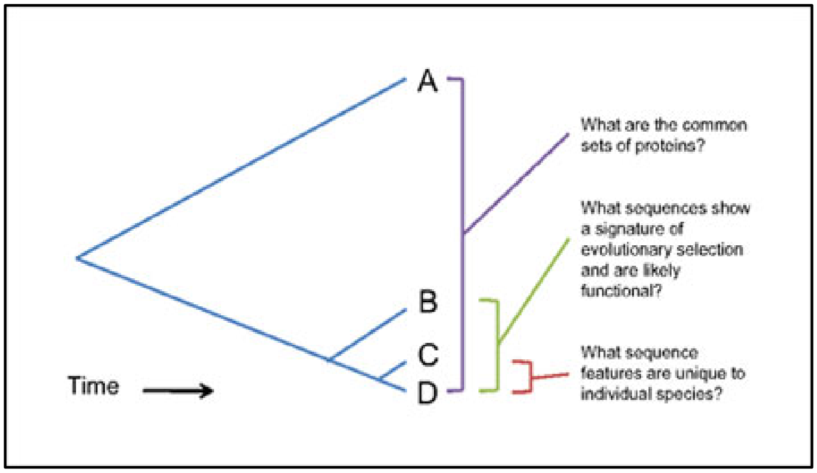

A
B
C
D
Time
What are the common sets of proteins?
What sequences show a signature of evolutionary selection and are likely functional?
What sequence features are unique to individual species?

Figure 3: Comparisons of genomes at different phylogenetic distances serve to address specific questions.

We have learned from homologous sequence alignment that the information that can be gained by comparing two genomes together is largely dependent upon the phylogenetic distance between them. Phylogenetic distance is a measure of the degree of separation between two organisms or their genomes on an evolutionary scale, usually expressed as the number of accumulated sequence changes, number of years, or number of generations. The distances are often placed on phylogenetic trees, which show the deduced relationships among the organisms (Figure 3). The more distantly related two organisms are, the less sequence similarity or shared genomic features will be detected between them. Thus, only general insights about classes of shared genes can be gathered by genomic comparisons at very long phylogenetic distances (e.g., over one billion years since their separation). Over such very large distances, the order of genes and the signatures of sequences that regulate their transcription are rarely conserved. .

At closer phylogenetic distances (50–200 million years of divergence), both functional and non-functional DNA is found within the conserved segments. In these cases, the functional sequences will show signatures of selection by virtue of their sequences having changed less, or more slowly than, non-functional DNA. Moreover, beyond the ability to discriminate functional from non-functional DNA, comparative genomics is also contributing to the identification of general classes of important DNA elements, such as coding exons of genes, non-coding RNAs, and some gene regulatory sites. In contrast, very similar genomes separated by about 5 million years of evolution (such as human and chimpanzee) are particularly useful for finding the sequence differences that may account for subtle differences in biological form. These are sequence changes under directional selection, a process whereby natural selection favors a single phenotype and continuously shifts the allele frequency in one direction. Comparative genomics is thus a powerful and promising approach to biological discovery that becomes more and more informative as genomic sequence data accumulate.

Practical

VISTA Tool for comparative genomics

Comparison of DNA sequences from different species is a fundamental method for identifying functional elements in genomes. VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species.

VISTA tools are created to assist biologists in carrying out this task. First VISTA server at http://www-gsd.lbl.gov/vista/ was launched in the summer of 2000 and was designed to align long genomic sequences and visualize these alignments with associated functional annotations. Currently the VISTA site includes multiple comparative genomics tools and provides users with rich capabilities to browse pre-computed whole-genome alignments of large vertebrate genomes and other groups of organisms with VISTA Browser, to submit their own sequences of interest to several VISTA servers for various types of comparative analysis and to obtain detailed comparative analysis results for a set of cardiovascular genes. We illustrate capabilities of the

VISTA site by the analysis of a 180 kb interval on human chromosome 5 that encodes for the kinesin family member 3A (*KIF3A*) protein.

VISTA comparative genomics packages

(i) mVISTA

Align and compare your sequences from multiple species mVISTA. mVISTA is designed to perform pairwise alignments of DNA sequences up to mega bases long from two or more species and to visualize these alignments together with annotations. AVID is the alignment engine behind mVISTA, and it allows the global alignment of DNA sequences of arbitrary length. To use mVISTA for comparative sequence analysis, two or more sequences in FASTA format (plain text only) or GenBank accession numbers together with a gene annotation file are submitted to the Web server.

(ii) rVISTA

rVISTA (regulatory VISTA) combines searching the major transcription factor binding site database TRANSFAC™ Professional from Biobase with a comparative sequence analysis. It can be used directly or through links in mVISTA, GenomeVISTA and VISTA Browser.The visualization program for rVISTA allows the user to look at binding sites for a single transcription factor and/or various combinations of transcription factor binding sites, which allows the user easily to examine the clustering of binding sites for factors that are believed to interact with one another. Both global (AVID) and local (BLASTZ) alignment algorithms are incorporated into rVISTA.

(iii) Genome VISTA

Compare your sequences with several whole genome assemblies. It will automatically find the ortholog, obtain the alignment and VISTA plot.

- **wgVISTA**

  Align sequences up to 10Mb long (finished or draft) including microbial whole-genome assemblies.

- **Phylo-Vista**

  The Phylo-VISTA program with its associated web server presents a novel method for the visualization and analysis of conservation in multiple sequence alignments by providing several significant extensions to VISTA tools. Analyze multiple DNA sequence alignments of sequences from different species while considering their phylogenic relationships.

Installation and Usage

VISTA is an application that allows the visualization of long sequence alignments with annotation information. The VISTA program uses the file or files (to visualize several related alignments) produced by any procedure of global or local alignment (such as BLAST, Gap (GCG), etc.) of two DNA sequences and parsed by the user according to 'Alignment_file' format. In a short time we are planning to release the stand-alone global alignment software to use as the first step before visualization.

The VISTA plot is based on moving a user-specified window over the entire alignment and calculating the percent identity over the window at each base pair. The X-axis represents the base sequence; the Y-axis represents the percent identity. If the user supplies an annotation file, genes and exons are marked above the plot. The direction of genes is indicated by an arrow, while the coding exons and UTRs are marked with rectangles of different color. Conserved regions are highlighted under the curve, with red indicating a conserved non-coding region and blue indicating a conserved exon. Conserved UTRs are colored turquoise. The colors can be modified by the user. A conserved region is defined with percentage and length cutoffs. Conserved segments with percent identity X and length Y are defined to be regions in which every contiguous sub-segment of length Y was at least X% identical to its paired sequence. These segments are merged to define the conserved regions.

VISTA can be configured for visualizing alignments of various lengths by changing several parameters: the number of pages on which the output appears, the number of frames per page, the window size, and the resolution at which the alignment is plotted. VISTA allows one to easily create figures for various documents.

(i) Installation:

Create a directory and copy Vista.jar and retepPDF2.jar to it. Then change your CLASSPATH environment variable to include references to these two files.

(ii) Example I (Windows):

1. mkdir c:\vista

2. copy source_path\Vista.jar c:\vista

3. copy source_path\retepPDF2.jar c:\vista

4. set CLASSPATH=c:\vista\Vista.jar;c:\vista\retepPDF2.jar

(iii) Example II (UNIX, csh/tcsh):

```
1. mkdir target_path/vista
2. cp source_path/Vista.jar target_path/vista
3. cp source_path/retepPDF2.jar target_path/vista
4. setenv CLASSPATH "target_path/vista/Vista.jar:target_path/vista/retepPDF2.jar"
```

(iv) Usage: java Vista [-options] plot_file

where "plot_file" is the name of a file containing plot parameters,

and options include:

```
  -q   turn on quiet mode
  -d   turn on debug mode
```

Further Reading

1. *Stein LD, et al. (2003). "The genome sequence of Caenorhabditisbriggsae: a platform for comparative genomics". PLoS Biology**1** (2): E45. doi:10.1371/journal.pbio.0000045. PMC 261899.PMID 14624247.*
2. *"Newly Sequenced Worm a Boon for Worm Biologists". PLoS Biology**1** (2): e4–e4. 2003. doi:10.1371/journal.pbio.0000044.edit*
3. *Bejerano, Gill; Michael Pheasant, Igor Makunin, Stuart Stephen, W James Kent, John S Mattick, David Haussler (2004-05-28). "Ultraconserved elements in the human genome".Science (New York, N.Y.)304 (5675): 1321-1325. doi:10.1126/science.1098119. ISSN 1095-9203.Retrieved 2012-01-13.*
4. *Pollard, Katherine S.; Sofie R. Salama, Nelle Lambert, Marie-Alexandra Lambot, Sandra Coppens, Jakob S. Pedersen, Sol Katzman, Bryan King, Courtney Onodera, Adam Siepel, Andrew D. Kern, Colette Dehay, Haller Igel, Manuel Ares, Pierre Vanderhaeghen, David Haussler (2006). "An RNA gene expressed during cortical development evolved rapidly in humans". Nature 443 (7108): 167-172. doi:10.1038/nature05113. ISSN 0028-0836. Retrieved 2012-01-13.*
5. *Cristianini N and Hahn M (2006). Introduction to Computational Genomics.Cambridge University Press.ISBN 0-5216-7191-4.*

APPENDIX

Software packages for next generation sequence analysis

Integrated solutions

CLCbio Genomics Workbench	*de novo* and reference assembly of Sanger, Roche FLX, Illumina, Helicos, and SOLiD data. Commercial next-gen-seq software that extends the CLCbio Main Workbench software. Includes SNP detection, CHiP-seq, browser and other features. Commercial. Windows, Mac OS X and Linux.	http://www.clcbio.com/index.php?id=1240
GALAXY	Galaxy = interactive and reproducible genomics. A job webportal.	http://wiki.g2.bx.psu.edu/
GenoMatix	Integrated Solutions for Next Generation Sequencing data analysis.	http://www.genomatix.de/en/produkte.html
JMP Genomics	Next gen visualization and statistics tool from SAS. They are working with NCGR to refine this tool and produce others.	http://www.jmp.com/software/genomics/
NextGENe	de novo and reference assembly of Illumina, SOLiD and Roche FLX data. Uses a novel Condensation Assembly Tool approach where reads are joined via "anchors" into mini- contigs before assembly. Includes SNP detection, CHiP-seq, browser and other features. Commercial. Win or MacOS.	http://softgenetics.com/NextGENe.html
SeqMan Genome Analyser	Software for Next Generation sequence assembly of Illumina, Roche FLX and Sanger data integrating with Lasergene Sequence Analysis software for additional analysis and visualization capabilities. Can use a hybrid templated/de novo approach. Commercial. Win or Mac OS X.	http://www.dnastar.com/products/SMGA.php
SHORE	SHORE, for Short Read, is a mapping and analysis pipeline for short DNA sequences produced on a Illumina Genome Analyzer. A suite created by the 1001 Genomes project. Source for POSIX.	http://1001genomes.org/software/shore.html
SlimSearch	Fledgling commercial product.	http://www.realtimegenomics.com/

Sequence Format Conversion and QC

Galaxy tools	sequence format convertion tools	https://main.g2.bx.psu.edu/root/tool_menu
FAstQC	FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FASTX-Toolkit	The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.	hannonlab.cshl.edu/fastx_toolkit/

Alignment and Assembly (Reference & De-novo)

Alignment and Assembly to reference

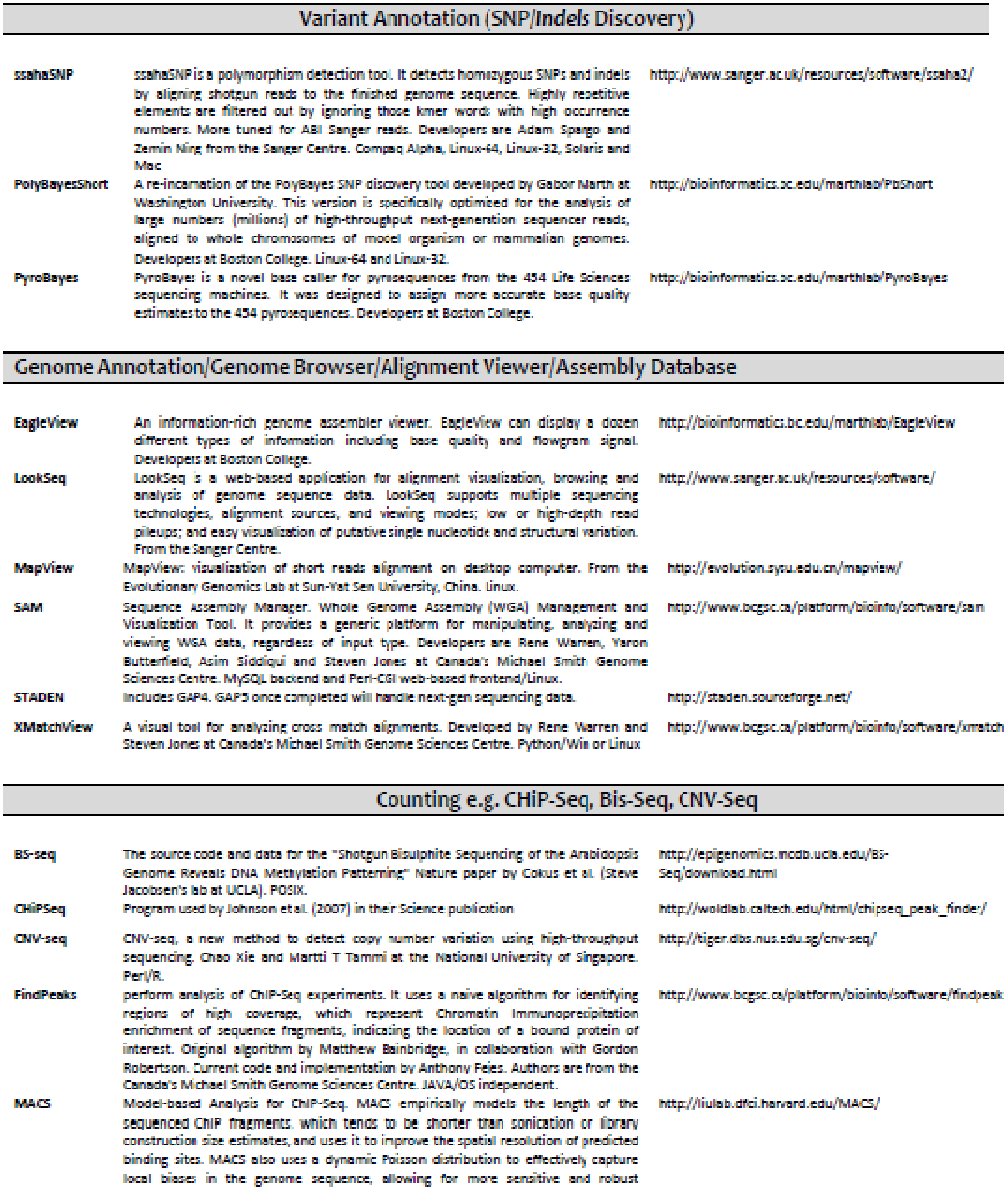
BFAST	Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.	https://secure.genome.ucla.edu/index.php/BFAST
Bowtie	Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.	http://bowtie-bio.sourceforge.net/index.shtml
BWA	Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome.	http://maq.sourceforge.net/
ELAND	Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony J. Cox for the Solexa 1G machine.	http://bioinfo.cgrb.oregonstate.edu/docs/solexa/

Exonerate	Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.	http://www.ebi.ac.uk/~guy/exonerate/
GenomeMapper	GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments. A tool created by the 1001 Genomes project. Source for POSIX.	http://1001genomes.org/downloads/genomemapper.html
GMAP	GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C/Perl for Unix.	http://www.gene.com/share/gmap/
gnumap	The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.	dna.cs.byu.edu/gnumap/
MAQ	Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle ABI SOLiD data. Written by Heng Li from the Sanger Centre. Features extensive supporting tools for DIP/SNP detection, etc. C++ source	http://sourceforge.net/projects/maq/
MOSAIK	MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FLX, Illumina, SOLiD, and Helicos. Written by Michael Strömberg at Boston College. Win/Linux/MacOSX	http://bioinformatics.bc.edu/marthlab/Mosaik
mrFAST & mrsFAST	mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies; in a fast and memory-efficient manner. Robust to INDELs and MrsFAST has a bisulphite mode. Authors are from the University of Washington. C as source.	http://mrfast.sourceforge.net/
MUMmer	MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient suffix tree library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. POSIX OS required.	http://mummer.sourceforge.net/
Novocraft	Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Seq. Commercial. Available free for evaluation, educational use and for use on open not-for-profit projects. Requires Linux or Mac OS X.	http://www.novocraft.com/main/index.php
PASS	It supports Illumina, SOLiD and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spaced seed intial filter, then NW dynamic algorithm to a SW(like) local alignment. Authors are from CRIBI in Italy. Win/Linux.	http://pass.cribi.unipd.it/cgi-bin/pass.pl
RMAP	Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics). POSIX OS required.	http://rulai.cshl.edu/rmap/
SeqMap	Supports up to 5 or more bp mismatches/INDELs. Highly tunable. Written by Hui Jiang from the Wong lab at Stanford. Builds available for most OS's.	biogibbs.stanford.edu/%7Ejiangh/SeqMap/
SHRiMP	Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Brudno and Stephen Rumble at the University of Toronto. POSIX.	http://compbio.cs.toronto.edu/shrimp/
Slider	An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences. Authors are from BCGSC.	http://www.bcgsc.ca/platform/bioinfo/software/slider
SOAP	SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences.	http://soap.genomics.org.cn/
SSAHA	SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using a hash table. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.	http://www.sanger.ac.uk/resources/software/ssaha2/
SOCS	Aligns SOLiD data. SOCS is built on an iterative variation of the Rabin-Karp string search algorithm, which uses hashing to reduce the set of possible matches, drastically increasing search speed. Authors are Ondov B, Varadarajan A, Passalacqua KD and Bergman NH.	http://www.solidsoftwaretools.com/gf/project/socs/

SWIFT	The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT — fast local alignment search, guaranteeing to find epsilon-matches between two sequences. SWIFT BALSAM — a very fast program to find semiglobal non-gapped alignments based on k-mer seeds. Authors are Kim Rasmussen (SWIFT) and Wolfgang Gerlach (SWIFT BALSAM)	http://bibiserv.techfak.uni-bielefeld.de/swift/welcome.html
SXOligoSearch	SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq RNA or NCBI genome builds for a number of organisms. Web Portal. OS independent.	synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
Vmatch	A versatile software tool for efficiently solving large scale sequence matching tasks. 'Vmatch subsumes the software tool REPuter, but is much more general, with a very flexible user interface, and improved space and time requirements. Essentially a large string matching toolbox. POSIX.	http://www.vmatch.de/'
ZOOM	ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerged by next-generation sequencing technology, back to the reference genomes, and carry out post-analysis. ZOOM is developed to be highly accurate, flexible, and user-friendly with speed being a critical priority. Commercial. Supports Illumina and SOLiD data.	http://www.bioinfor.com/products/zoom/index.php

De novo Alignment/Assembly

ABySS	Assembly By Short Sequences. ABySS is a de novo sequence assembler that is designed for very short reads. The single-processor version is useful for assembling genomes up to 40-50 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes. By Simpson JT and others at the Canada's Michael Smith Genome Sciences Centre. C++ as source.	www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATHS	ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is a whole genome shotgun assembler that can generate high quality assemblies from short reads. Assemblies are presented in a graph form that retains ambiguities, such as those arising from polymorphism, thereby providing information that has been absent from previous genome assemblies. Broad Institute.	ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/
Edena	Edena (Exact DE Novo Assembler) is an assembler dedicated to process the millions of very short reads produced by the Illumina Genome Analyzer. Edena is based on the traditional overlap layout paradigm. By D. Hernandez, P. François, L. Farinelli, M. Osteras, and J. Schrenzel. Linux/Win.	http://www.genomic.ch/edena.php
EULER-SR	Short read *de novo* assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research). Uses a de Bruijn graph approach.	http://euler-assembler.ucsd.edu/portal/
MIRA 3	MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing technology (GS20 or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.	http://chevreux.org/projects_mira.html
SEQAN	A Consistency-based Consensus Algorithm for De Novo and Reference-guided Sequence Assembly of Short Reads. By Tobias Rausch and others. C++, Linux/Win.	http://www.seqan.de/projects/consensus/
SHARCGS	De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecular Genetics.	http://sharcgs.molgen.mpg.de/
SSAKE	The Short Sequence Assembly by K-mer search and 3' read Extension (SSAKE) is a genomics application for aggressively assembling millions of short nucleotide sequences by progressively searching for perfect 3'-most k-mers using a DNA prefix tree. Authors are René Warren, Granger Sutton, Steven Jones and Robert Holt from the Canada's Michael Smith Genome Sciences Centre. Perl/Linux.	http://www.bcgsc.ca/platform/bioinfo/software/ssake
SOAP3	Part of the SOAP suite	http://soap.genomics.org.cn/
VCAKE	De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.	http://sourceforge.net/projects/vcake/
VELVET	Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X coverage and paired reads. Developed by Daniel Zerbino and Ewan Birney at the	http://www.ebi.ac.uk/~zerbino/velvet/



Variant Annotation (SNP/Indels Discovery)

ssahaSNP	ssahaSNP is a polymorphism detection tool. It detects homozygous SNPs and indels by aligning shotgun reads to the finished genome sequence. Highly repetitive elements are filtered out by ignoring those kmer words with high occurrence numbers. More tuned for ABI Sanger reads. Developers are Adam Spargo and Zemin Ning from the Sanger Centre. Compaq Alpha, Linux-64, Linux-32, Solaris and Mac	http://www.sanger.ac.uk/resources/software/ssaha2/
PolyBayesShort	A re-incarnation of the PolyBayes SNP discovery tool developed by Gabor Marth at Washington University. This version is specifically optimized for the analysis of large numbers (millions) of high-througput next-generation sequencer reads, aligned to whole chromosomes of model organism or mammalian genomes. Developers at Boston College. Linux-64 and Linux-32.	http://bioinformatics.bc.edu/marthlab/PbShort
PyroBayes	PyroBayes is a novel base caller for pyrosequences from the 454 Life Sciences sequencing machines. It was designed to assign more accurate base quality estimates to the 454 pyrosequences. Developers at Boston College.	http://bioinformatics.bc.edu/marthlab/PyroBayes

Genome Annotation/Genome Browser/Alignment Viewer/Assembly Database

EagleView	An information-rich genome assembler viewer. EagleView can display a dozen different types of information including base quality and flowgram signal. Developers at Boston College.	http://bioinformatics.bc.edu/marthlab/EagleView
LookSeq	LookSeq is a web-based application for alignment visualization, browsing and analysis of genome sequence data. LookSeq supports multiple sequencing technologies, alignment sources, and viewing modes; low or high-depth read pileups; and easy visualization of putative single nucleotide and structural variation. From the Sanger Centre.	http://www.sanger.ac.uk/resources/software/
MapView	MapView: visualization of short reads alignment on desktop computer. From the Evolutionary Genomics Lab at Sun-Yat Sen University, China. linux.	http://evolution.sysu.edu.cn/mapview/
SAM	Sequence Assembly Manager. Whole Genome Assembly (WGA) Management and Visualization Tool. It provides a generic platform for manipulating, analyzing and viewing WGA data, regardless of input type. Developers are Rene Warren, Yaron Butterfield, Asim Siddiqui and Steven Jones at Canada's Michael Smith Genome Sciences Centre. MySQL backend and Perl-CGI web-based frontend/Linux.	http://www.bcgsc.ca/platform/bioinfo/software/sam
STADEN	Includes GAP4. GAP5 once completed will handle next-gen sequencing data.	http://staden.sourceforge.net/
XMatchView	A visual tool for analyzing cross match alignments. Developed by Rene Warren and Steven Jones at Canada's Michael Smith Genome Sciences Centre. Python/Win or Linux	http://www.bcgsc.ca/platform/bioinfo/software/xmatch

Counting e.g. CHiP-Seq, Bis-Seq, CNV-Seq

BS-seq	The source code and data for the "Shotgun Bisulphite Sequencing of the Arabidopsis Genome Reveals DNA Methylation Patterning" Nature paper by Cokus et al. (Steve Jacobsen's lab at UCLA). POSIX.	http://epigenomics.mcdb.ucla.edu/BS-Seq/download.html
CHiPSeq	Program used by Johnson et al. (2007) in their Science publication	http://woldlab.caltech.edu/html/chipseq_peak_finder/
CNV-seq	CNV-seq, a new method to detect copy number variation using high-throughput sequencing. Chao Xie and Martti T Tammi at the National University of Singapore. Perl/R.	http://tiger.dbs.nus.edu.sg/cnv-seq/
FindPeaks	perform analysis of ChIP-Seq experiments. It uses a naive algorithm for identifying regions of high coverage, which represent Chromatin Immunoprecipitation enrichment of sequence fragments, indicating the location of a bound protein of interest. Original algorithm by Matthew Bainbridge, in collaboration with Gordon Robertson. Current code and implementation by Anthony Fejes. Authors are from the Canada's Michael Smith Genome Sciences Centre. JAVA/OS independent.	http://www.bcgsc.ca/platform/bioinfo/software/findpeak
MACS	Model-based Analysis for ChIP-Seq. MACS empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust	http://liulab.dfci.harvard.edu/MACS/

	prediction.	
PeakSeq	PeakSeq: Systematic Scoring of ChIP-Seq Experiments Relative to Controls. a two-pass approach for scoring ChIP-Seq data relative to controls. The pass identifies putative binding sites and compensates for variation in the mappability of sequences across the genome.	
QUEST	Quantitative Enrichment of Sequence Tags. Sidow and Myers Labs at Stanford.	http://mendel.stanford.edu/sidowlab/downloads/quest/
SISSRs	Site Identification from Short Sequence Reads. BED file input. Raja Jothi @ NIH. Perl.	http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/

Alternate Base Calling

ROLEXA	R-based framework for base calling of Solexa data.	svitsrv25.epfl.ch/R-doc/library/Rolexa/html/00Index.html
Alta-cyclic	"a novel Illumina Genome-Analyzer (Solexa) base caller"	http://hannonlab.cshl.edu/Alta-Cyclic/main.html

Transcriptomics

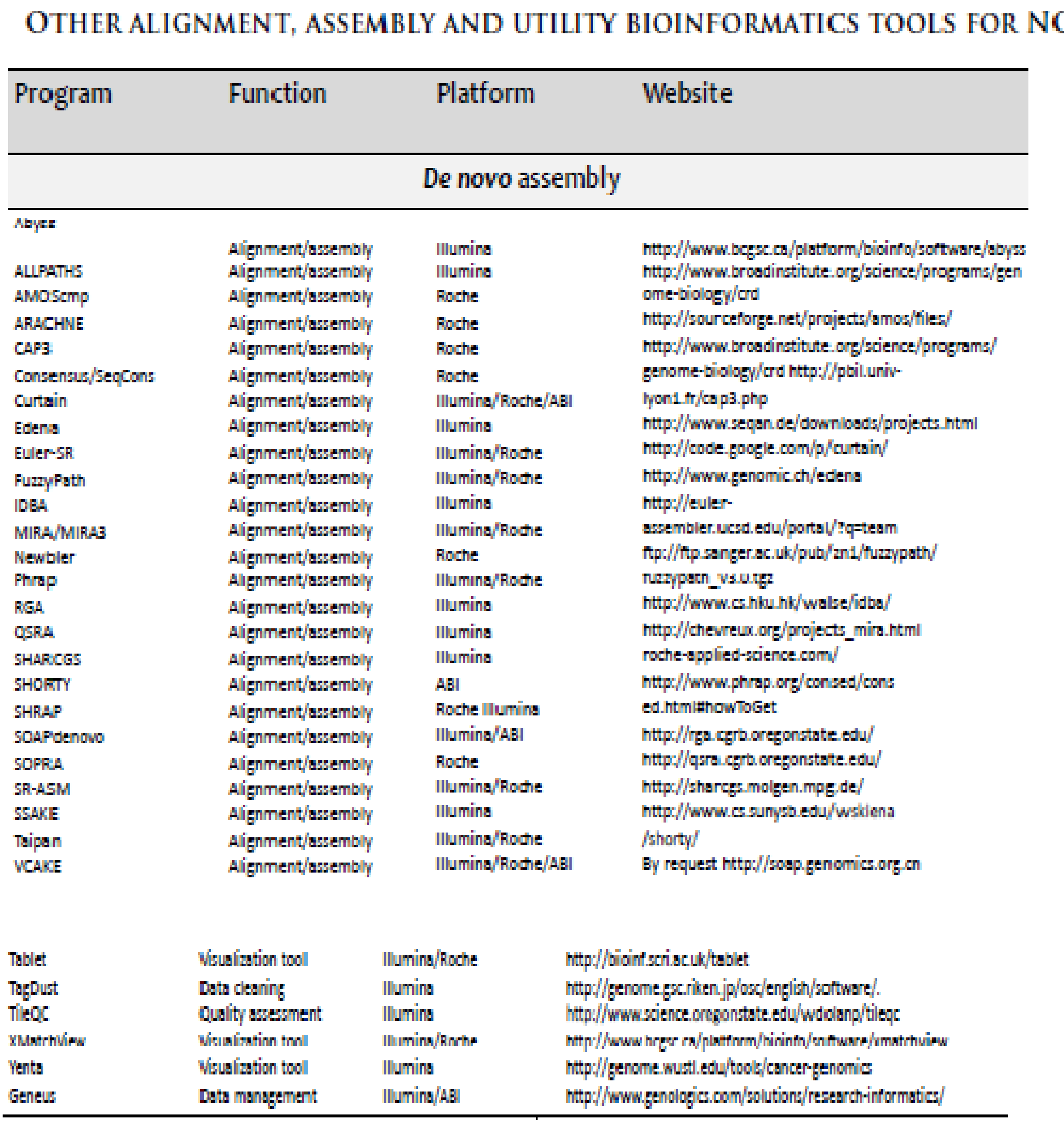
ERANGE	Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. Supports Bowtie, BLAT and ELAND. From the Wold lab	http://woldlab.caltech.edu/rnaseq/
G-Mo.R-Se	G-Mo.R-Se is a method aimed at using RNA-Seq short reads to build de novo gene models. First, candidate exons are built directly from the positions of the reads mapped on the genome (without any ab initio assembly of the reads), and all the possible splice junctions between those exons are tested against unmapped reads. From CNS in France.	http://www.genoscope.cns.fr/externe/gmorse/
MapNext	MapNext: A software tool for spliced and unspliced alignments and SNP detection of short sequence reads. From the Evolutionary Genomics Lab at Sun-Yat Sen University, China.	http://evolution.sysu.edu.cn/english/software/mapnext.h
QPalma	Optimal Spliced Alignments of Short Sequence Reads. Authors are Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch.	www.fml.tuebingen.mpg.de/raetsch/suppl/qpalma
RSAT	RSAT: RNA-Seq Analysis Tools. RNASAT is developed and maintained by Hui Jiang at Stanford University.	biogibbs.stanford.edu/%7Ejiangh/rsat/
TopHat	TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. TopHat is a collaborative effort between the University of Maryland and the University of California, Berkeley	http://tophat.cbcb.umd.edu/

Genome Visualization

Artemis	Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation.	http://www.sanger.ac.uk/resources/software/artemis/
GenomeView	GenomeView is a next-generation stand-alone genome browser and editor initiated in the BSB group at VIB and currently developed at Broad Institute.	http://genomeview.org/
IGV	The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.	http://www.broadinstitute.org/igv/
Gbrowse	GBrowse is a combination of database and interactive web pages for manipulating and displaying annotations on genomes.	http://gmod.org/wiki/GBrowse

THE TARGETED SEQUENCE-ENRICHMENT TECHNOLOGIES FOR NGS

Technology	Approach	Platform	Website
Agilent/SureSelect	Array- and solution-based	Illumina/Roche/ABI	http://www.chem.agilent.com/
RainDance	Microdroplet-based	Illumina/Roche/ABI	http://www.raindancetechnologies.com/
NimbleGen/SeqCap/EZ Exome	Array- and solution-based	Illumina/Roche/ABI	http://www.nimblegen.com/products/seqcap/index.ht
Febit/HybSelect	Microarray-based	Illumina/Roche/ABI	http://www.febit.com/microarray-
Fluidigm	PCR-based	Illumina/Roche/ABI	http://www.fluidigm.com/targeted-resequencing.html
Mycroarray/Myselect	Solution-based	Illumina/Roche/ABI	http://www.mycroarray.com/products/myselect.html
LC Sciences	Microarray-based	Illumina/Roche/ABI	http://www.lcsciences.com/applications/genomics/
Qiagen/SeqTarget	Long-range PCR-based	Illumina/Roche/ABI	http://www.qiagen.com/products/seqtargetsystem.as
Illumina/TruSeq	Solution based	Illumina/Roche/ABI	http://www.illumina.com/applications.ilmn

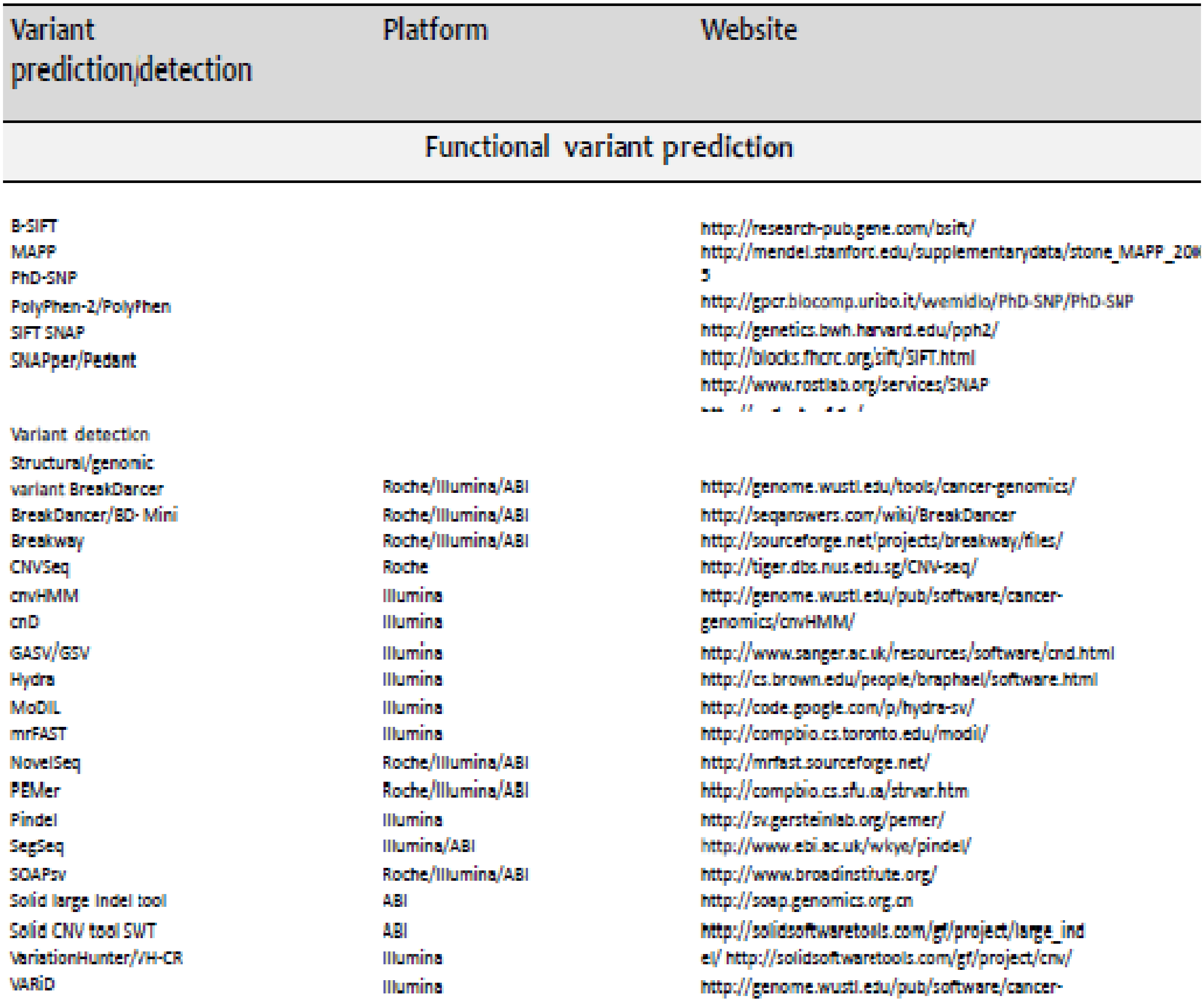


OTHER ALIGNMENT, ASSEMBLY AND UTILITY BIOINFORMATICS TOOLS FOR NC

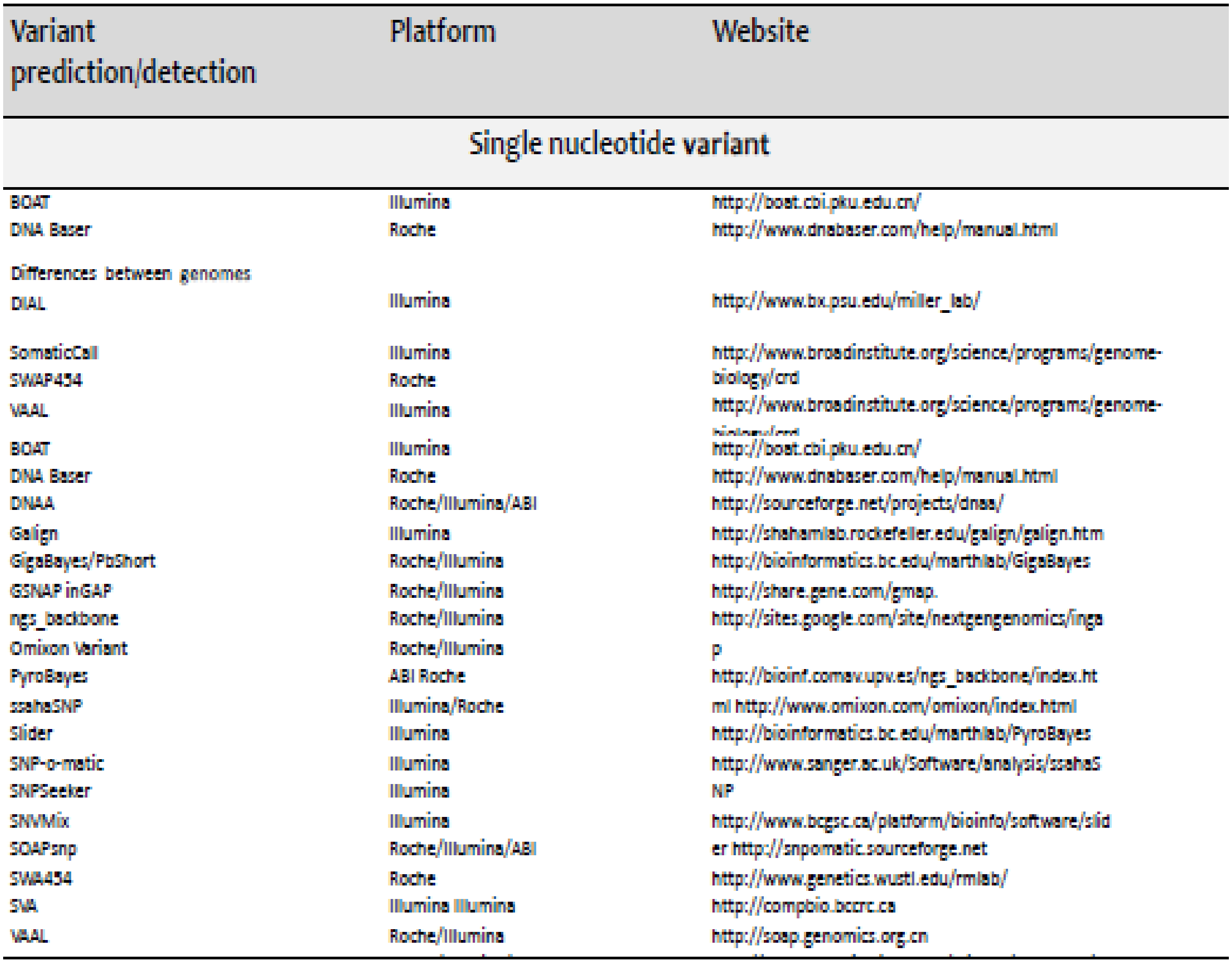
Program	Function	Platform	Website
De novo assembly			
Abyss			
	Alignment/assembly	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATHS	Alignment/assembly	Illumina	http://www.broadinstitute.org/science/programs/gen
AMOScmp	Alignment/assembly	Roche	ome-biology/crd
ARACHNE	Alignment/assembly	Roche	http://sourceforge.net/projects/amos/files/
CAP3	Alignment/assembly	Roche	http://www.broadinstitute.org/science/programs/
Consensus/SeqCons	Alignment/assembly	Roche	genome-biology/crd http://pbil.univ-
Curtain	Alignment/assembly	Illumina/Roche/ABI	lyon1.fr/cap3.php
Edena	Alignment/assembly	Illumina	http://www.seqan.de/downloads/projects.html
Euler-SR	Alignment/assembly	Illumina/Roche	http://code.google.com/p/curtain/
FuzzyPath	Alignment/assembly	Illumina/Roche	http://www.genomic.ch/edena
IDBA	Alignment/assembly	Illumina	http://euler-
MIRA/MIRA3	Alignment/assembly	Illumina/Roche	assembler.ucsd.edu/portal/?q=team
Newbler	Alignment/assembly	Roche	ftp://ftp.sanger.ac.uk/pub/zn1/fuzzypath/
Phrap	Alignment/assembly	Illumina/Roche	fuzzypath_v3.0.tgz
RGA	Alignment/assembly	Illumina	http://www.cs.hku.hk/walse/idba/
QSRA	Alignment/assembly	Illumina	http://chevreux.org/projects_mira.html
SHARCGS	Alignment/assembly	Illumina	roche-applied-science.com/
SHORTY	Alignment/assembly	ABI	http://www.phrap.org/consed/cons
SHRAP	Alignment/assembly	Roche Illumina	ed.html#howToGet
SOAPdenovo	Alignment/assembly	Illumina/ABI	http://rga.cgrb.oregonstate.edu/
SOPRA	Alignment/assembly	Roche	http://qsra.cgrb.oregonstate.edu/
SR-ASM	Alignment/assembly	Illumina/Roche	http://sharcgs.molgen.mpg.de/
SSAKE	Alignment/assembly	Illumina	http://www.cs.sunysb.edu/wskiena
Taipan	Alignment/assembly	Illumina/Roche	/shorty/
VCAKE	Alignment/assembly	Illumina/Roche/ABI	By request http://soap.genomics.org.cn
Tablet	Visualization tool	Illumina/Roche	http://bioinf.scri.ac.uk/tablet
TagDust	Data cleaning	Illumina	http://genome.gsc.riken.jp/osc/english/software/.
TileQC	Quality assessment	Illumina	http://www.science.oregonstate.edu/wdiolanp/tileqc
XMatchView	Visualization tool	Illumina/Roche	http://www.bcgsc.ca/platform/bioinfo/software/xmatchview
Yenta	Visualization tool	Illumina	http://genome.wustl.edu/tools/cancer-genomics
Geneus	Data management	Illumina/ABI	http://www.genologics.com/solutions/research-informatics/

Program	Function	Platform	Website
		Reference assembly	
Bowtie	Alignment/assembly	Illumina/Roche/ABI	http://bowtie-bio.sourceforge.net
BWA CoronaLite	Alignment/assembly	Illumina/ABI ABI	http://bio-
Bowtie	Alignment/assembly	Illumina/Roche/ABI	http://bowtie-bio.sourceforge.net
BWA CoronaLite	Alignment/assembly	Illumina/ABI ABI	http://bio-
CABOG	Alignment/assembly	Roche/ABI	bwa.sourceforge.net/bwa.shtml
ELAND/ELAND2	Alignment/assembly	Illumina/ABI	http://solidsoftwaretools.com/gf/pro
EULER Exonerate	Alignment/assembly	Illumina Roche	ject/corona/ http://wgs-
EMBF	Alignment/assembly	Illumina Illumina	assembler.sf.net
GenomeMapper	Alignment/assembly	Illumina Illumina	http://www.illumina.com/
GMAP	Alignment/assembly	Illumina	http://euler-assembler.ucsd.edu/portal/
gnumap	Alignment/assembly	Illumina/ABI	http://www.ebi.ac.uk/wguy/exonerate
ICON Karma	Alignment/assembly	Illumina Illumina	http://www.biomedcentral.com/1471-
LAST LOCAS	Alignment/assembly	ABI	2105/10?issue=S1
Mapreads	Alignment/assembly	Illumina/ABI Illumina	http://1001genomes.org/downloads/genome
MAQ MOM	Alignment/assembly	Illumina/Roche/ABI	mapper.html
Mosaik	Alignment/assembly	Illumina	http://www.gene.com/share/gmap
mrFAST/mrsFAST	Alignment/assembly	ABI Illumina	http://dna.cs.byu.edu/gnumap/
MUMer	Alignment/assembly	Illumina	http://icom.sourceforge.net/
Nexalign	Alignment/assembly	Illumina/ABI	http://www.sph.umich.edu/csg/pha/karma/
Novocraft	Alignment/assembly	Illumina/ABI	http://last.cbrc.jp/
PerM RazerS	Alignment/assembly	Illumina	http://www-ab.informatik.uni-
RMAP	Alignment/assembly	Illumina/Roche	tuebingen.de/software/locas
segemehl	Alignment/assembly	Roche	http://solidsoftwaretools.com/gf/project/mapr
SeqCons	Alignment/assembly	Illumina	eads/ http://maq.sourceforge.net
SeqMap	Alignment/assembly	Illumina/Roche/ABI	http://mom.csbc.vcu.edu/
SHRiMP	Alignment/assembly	Illumina	http://bioinformatics.bc.edu/marthlab/M
SSAHA/SSAHA2	Alignment/assembly	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/SSAHA2
Stampy	Alignment/assembly	Illumina	http://www.well.ox.ac.uk/wmarting/
SXOligoSearch	Alignment/assembly	Illumina	http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
SHORE	Alignment/assembly	Illumina	http://1001genomes.org/downloads/shore.html
Vmatch	Alignment/assembly	Illumina	http://www.vmatch.de/
Diagnostics/utilities			
Artemis/ACT	Visualization tool	Illumina/Roche	http://www.sanger.ac.uk/resources/software/artemis/
CASHX	Pipeline	Illumina	http://seqanswers.com/wiki/CASHX
Consed	Visualization tool	Illumina/Roche	http://www.genome.washington.edu/consed/consed.html
EagleView	Visualization tool	Illumina/Roche	http://bioinformatics.bc.edu/marthlab/EagleView
FastQC	Quality assessment	Illumina/ABI	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
Gambit	Visualization tool	Illumina/Roche	http://bioinformatics.bc.edu/marthlab/Gambit
Goby	Data management	Illumina/Roche/ABI	http://campagnelab.org/software/goby/
G-SQZ	Data management	Illumina/ABI	http://public.tgen.org/sqz
Hawkeye	Visualization tool	Illumina/Roche	http://amos.sourceforge.net/hawkeye
Hybrid-SHREC	Error Correction	Illumina/Roche/ABI	http://www.cs.helsinki.fi/u/lmsalmel/hybrid-shrec/
IGV	Visualization tool	Illumina	http://www.broadinstitute.org/igv/?q¼home
LookSeq	Visualization tool	Illumina/Roche	http://lookseq.sourceforge.net
MagicViewer	Visualization tool	Illumina	http://bioinformatics.zj.cn/magicviewer/
MapView	Visualization tool	Illumina	http://evolution.sysu.edu.cn/mapview/
NGSView	Visualization tool	Illumina/ABI	http://ngsview.sourceforge.net
PIQA	Quality assessment	Illumina	http://bioinfo.uh.edu/PIQA
Reconciliation	Assembly pipeline	Illumina	http://www.genome.umd.edu/software.htm
RefCov	Sequence coverage	Illumina/Roche	http://genome.wustl.edu/tools/cancer-genomics
SAM Tools	Utilities	Illumina/Roche	http://sourceforge.net/projects/samtools/files/
Savant	Visualization tool	Illumina/Roche	http://compbio.cs.toronto.edu/savant/
ShortRead	Quality assessment	Illumina/Roche	http://bioconductor.org/packages/2.6/bioc/html/ShortRead.html
SHREC	Error Correction	Illumina/Roche	http://www.informatik.uni-kiel.de/jasc/Shrec/
Staden Tools (GAP5)	Pipeline	Illumina/Roche	http://sourceforge.net/projects/staden/files/

THE GENETIC VARIANT PREDICTION AND DETECTION BIOINFORMATIC PROGRAMS FOR NGS DATA ANALYSIS



Variant prediction/detection	Platform	Website
Functional variant prediction		
B-SIFT		http://research-pub.gene.com/bsift/
MAPP		http://mendel.stanford.edu/supplementarydata/stone_MAPP_20
PhD-SNP		5
PolyPhen-2/PolyPhen		http://gpcr.biocomp.unibo.it/wemidio/PhD-SNP/PhD-SNP
SIFT SNAP		http://genetics.bwh.harvard.edu/pph2/
SNAPper/Pedant		http://blocks.fhcrc.org/sift/SIFT.html
		http://www.rostlab.org/services/SNAP
		[illegible]
Variant detection		
Structural/genomic		
variant BreakDancer	Roche/Illumina/ABI	http://genome.wustl.edu/tools/cancer-genomics/
BreakDancer/BD- Mini	Roche/Illumina/ABI	http://seqanswers.com/wiki/BreakDancer
Breakway	Roche/Illumina/ABI	http://sourceforge.net/projects/breakway/files/
CNVSeq	Roche	http://tiger.dbs.nus.edu.sg/CNV-seq/
cnvHMM	Illumina	http://genome.wustl.edu/pub/software/cancer-
cnD	Illumina	genomics/cnvHMM/
GASV/GSV	Illumina	http://www.sanger.ac.uk/resources/software/cnd.html
Hydra	Illumina	http://cs.brown.edu/people/braphael/software.html
MoDIL	Illumina	http://code.google.com/p/hydra-sv/
mrFAST	Illumina	http://compbio.cs.toronto.edu/modil/
NovelSeq	Roche/Illumina/ABI	http://mrfast.sourceforge.net/
PEMer	Roche/Illumina/ABI	http://compbio.cs.sfu.ca/strvar.htm
Pindel	Illumina	http://sv.gersteinlab.org/pemer/
SegSeq	Illumina/ABI	http://www.ebi.ac.uk/wkye/pindel/
SOAPsv	Roche/Illumina/ABI	http://www.broadinstitute.org/
Solid large Indel tool	ABI	http://soap.genomics.org.cn
Solid CNV tool SWT	ABI	http://solidsoftwaretools.com/gf/project/large_ind
VariationHunter/VH-CR	Illumina	el/ http://solidsoftwaretools.com/gf/project/cnv/
VARiD	Illumina	http://genome.wustl.edu/pub/software/cancer-



Variant prediction/detection	Platform	Website
	Single nucleotide variant	
BOAT	Illumina	http://boat.cbi.pku.edu.cn/
DNA Baser	Roche	http://www.dnabaser.com/help/manual.html
Differences between genomes		
DIAL	Illumina	http://www.bx.psu.edu/miller_lab/
SomaticCall	Illumina	http://www.broadinstitute.org/science/programs/genome-
SWAP454	Roche	biology/crd
VAAL	Illumina	http://www.broadinstitute.org/science/programs/genome- biology/crd
BOAT	Illumina	http://boat.cbi.pku.edu.cn/
DNA Baser	Roche	http://www.dnabaser.com/help/manual.html
DNAA	Roche/Illumina/ABI	http://sourceforge.net/projects/dnaa/
Galign	Illumina	http://shahamlab.rockefeller.edu/galign/galign.htm
GigaBayes/PbShort	Roche/Illumina	http://bioinformatics.bc.edu/marthlab/GigaBayes
GSNAP inGAP	Roche/Illumina	http://share.gene.com/gmap.
ngs_backbone	Roche/Illumina	http://sites.google.com/site/nextgengenomics/inga
Omixon Variant	Roche/Illumina	p
PyroBayes	ABI Roche	http://bioinf.comav.upv.es/ngs_backbone/index.ht
ssahaSNP	Illumina/Roche	ml http://www.omixon.com/omixon/index.html
Slider	Illumina	http://bioinformatics.bc.edu/marthlab/PyroBayes
SNP-o-matic	Illumina	http://www.sanger.ac.uk/Software/analysis/ssahaS
SNPSeeker	Illumina	NP
SNVMix	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slid
SOAPsnp	Roche/Illumina/ABI	er http://snpomatic.sourceforge.net
SWA454	Roche	http://www.genetics.wustl.edu/rmlab/
SVA	Illumina Illumina	http://compbio.bccrc.ca
VAAL	Roche/Illumina	http://soap.genomics.org.cn