29

Deep Sequencing Data Analysis: Challenges and Solutions

Ofer Isakov and Noam Shomron
Sackler Faculty of Medicine,
Tel Aviv University,
Israel

1. Introduction

Ultra high throughput sequencing, also known as deep sequencing or Next Generation Sequencing (NGS), is revolutionizing the study of human genetics and has immense clinical implications. It has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude in just a few years, a trend which is guaranteed to rapidly accelerate in the near future (Metzker, 2010). Using deep sequencing, for example, it is now possible to discover novel disease causing mutations (Ley et al., 2008) and detect traces of pathogenic microorganisms (Isakov et al., 2011). For the first time, research fields such as personalized medicine for patient treatment are becoming tangible at genomic levels given advances in deep sequencing data integration.

The amount of data produced by a single ultra high throughput sequencing run is often tremendous and can reach hundreds of millions of reads in various lengths per experiment (Mardis, 2008). The storage, processing, querying, parsing, analyzing and interpreting of such an incredible amount of data is a significant task that holds many obstacles and challenges (Koboldt et al., 2010). In this chapter we will address some of the possibilities, potentials and questions raised during ultra high throughput sequencing data analysis. We will mainly focus on common pre-analysis concepts and crucial advanced considerations for alignment, assembly and variation detection. Currently, the deep sequencing user is faced with an abundance of deep sequencing data analysis tools, both publicly and commercially available. For each of the aforementioned analysis types, we will point out the various aspects to be considered when choosing a tool, and emphasize the relevant challenges and possible limitations in order to assist the user in picking the most suitable one. Since deep sequencing data analysis is a rapidly evolving field, our focus will be on fundamental concepts of the analysis process and the its challenges, allowing this read to be relevant amid additional published software.

Our first part will encompass a brief overview of current leading deep sequencing technologies with special attention to their features, strengths and possible drawbacks in regards to the different preliminary questions that one might ask when using ultra high throughput sequencing. The second part of the chapter introduces pre-analysis processes. These are common quality control and assurance methods that alleviate deep sequencing derived biases and improve the overall results of any down-stream analysis. In the third part of the chapter, we will go over the different aspects of the post-sequencing analysis,

specifically deep sequencing data alignment, assembly and variant detection. For each section we will cover leading methods and tools, quality evaluation and filtration and address the requirements, capabilities and limitations of these tools. The section on variation detection will cover both common variant detection considerations, variant specific challenges and currently available solutions.

2. Sequencing technologies

Sequencing technologies are evolving rapidly, with an overwhelming increase in efficiency and throughput (Mardis, 2008). This expeditious rate of change and improvement is accompanied by a variety of different sequencing platforms, with both great similarities and differences alike. Without going into the technology underlying each sequencing platform in detail, we will specify advantages and limitations both general and specific, that are relevant for deep sequencing experiment design. For this purpose, we will refer to the leading commercially available platforms produced by Roche/454 (Margulies et al., 2005a), Illumina/Solexa (Bentley et al., 2008), Life/APG (SOLiD) (McKernan et al., 2009) and Pacific Biosciences (Eid et al., 2009).

The initial step in the sequencing process is random fragmentation of the nucleotide sequence of interest, in order to increase throughput by simultaneously sequencing millions of fragments. These template fragments can then either undergo clonal amplification, in which they are ligated with adapters and amplified using common PCR (Polymerase Chain Reaction) primers (Roche; Illumina; Life), or they can be used as the sequencing templates themselves (single molecule templates; Pacific Biosciences). Clonal amplified template preparation requires a higher amount of initial DNA material. Since this technique relies on PCR amplification, errors might be introduced to the target before the sequencing process begins. The amount of introduced errors is related to the fidelity of the polymerase utilized in the reaction (Chan, 2009). These potential background errors could be considered actual sequence variants in the down stream analysis. PCR utilization might also result in amplification bias, misrepresenting high GC content areas. Such is the case in a recent study in which PCR introduced expression biases for GC rich chromosomes required additional assessment, hampering the uniformity of the results (Chiu et al., 2010). Simultaneously sequencing clonal amplified templates is further complicated by potential different extension rates that cause asynchronous sequencing (phasing), resulting in a higher background noise. Single molecule template sequencing (Schadt, Turner, & Kasarskis, 2010) does not require PCR amplification thus circumventing its derived amplification and clonal sequencing biases making it an appropriate tool to be used in quantification experiments (e.g RNA-seq, Chip-seq etc.) and in cases where the initial sample DNA content is scarce. Because sequencing is performed on a single molecule and sequences are inferred from extremely weak signals, the correcting effect of simultaneous same-sequence template sequencing is lost resulting in a higher error rate (Schadt et al., 2010). Therefore a higher sequencing fidelity is required (Metzker, 2010).

In addition to the aforementioned general template preparation and sequencing method associated biases, one needs to consider the inherent benefits and shortcomings of each sequencing technology. Pyrosequencing, for example, employed by the Roch 454's GS FLX platform, generates long reads (~400nts) and presents relatively unbiased coverage, enhancing de-novo genome assembly and improving alignment capabilities, thus making it an appropriate tool for SNP and structural variations discovery, demonstrating low false positive rates (Margulies et al., 2005a; Nothnagel et al., 2011). However, the technology's

susceptibility to insertion and deletion errors and higher rate of homo-polymer (e.g contiguous run of the same base pair) sequencing errors should be considered when performing a variation oriented research (Chan, 2009). Current reverse termination (Illumina's Genome Analyzer or HiSeq 2000) and sequencing by ligation (Life's SOLiD) technologies produce shorter reads (<200nts) but at a much higher throughput and are considered optimal for small scale variants detection (e.g SNPs and indels) due to very high detection resolution owed to massive read overlap and high coverage. However, short reads' inherent problems of ambiguous mapping and complicated assembly can result in higher false positive rates in variant discovery (Nothnagel et al., 2011), that could be alleviated by higher throughput and employment of paired-end sequencing (e.g sequencing both ends of a fragment template) (Medvedev et al., 2009; Metzker, 2010). Sequencing by ligation technology, employed by Life's SOLiD, reads the colors of fluorescently marked ligated primers and converts them into the template sequence. SOLiD is less susceptible to phasing errors and the unique conversion of color to sequence results in an inherent error correction and thus a more accurate SNP detection process. However, this reduced error rate requires utilization of a reference genome in the color conversion process (Kircher & Kelso, 2010). We also note that error rate increases across all platforms towards the end of a sequenced read (Dohm et al., 2008), due to reduced enzyme efficiency, loss of enzymes, increased phasing effect or incomplete dye removal. The different attributes of the variety of platforms can result in significantly different output data and performance, and it was demonstrated that the combination of more than one platform is potentially more cost effective and could yield higher fidelity and accuracy (Dalloul et al., 2010; Nothnagel et al., 2011). We will now discuss how to alleviate some of these inherent and general difficulties using pre-analysis processing.

3. Pre-analysis processing

In this section we will discuss the processing performed on deep sequencing output data prior to the specific experimental analysis. Mentioned above are examples of the vast cross platform differences that could affect the downstream analysis and thus the biological conclusions derived. These differences accompany the inherent bias in deep sequencing experiments (Dohm et al., 2008; Schwartz et al., 2011). In order to reduce these possibly confounding effects, platform manufacturers and developers provide the end-user with a sequencing quality scale for both automated and recommended manual quality based data filtration and refinement (Bentley et al., 2008; Harris et al., 2008; Margulies et al., 2005a; K. J. McKernan et al., 2009). We will suggest quality confirmation methods for the text based output end-users face after a sequencing run, and discuss common necessary pre-analysis processing steps that ensure data validity and proper utilization. For each platform's inherent quality assurance and control measures, one should address the specific platform's technical support and annotation.

The most common initial form of output format is either a sequence FASTA file accompanied by a numerical quality QUAL file, describing the per-base probability of incorrect sequencing based on the PHRED quality score (Ewing and Green, 1998; Ewing et al., 1998), or the FASTQ format (Cock et al., 2010), containing sequences coupled with their quality stored as ASCII characters. Currently, Sanger FASTQ files use ASCII 33–126 to encode PHRED qualities from 0 to 93 (i.e. PHRED scores with an ASCII offset of 33), marking an error probability between 10^{0} and 10^{-93}. Up until the Genome Analyzer v1.3,

Illumina utilized a different scoring scale in their sequencing output, described in (Cock et al., 2010). Currently Illumina encodes PHRED scores with an ASCII offset of 64, and so can hold PHRED scores from 0 to 62 (ASCII 64–126). Life's SOLiD produce a color based FASTQ file (CSFASTQ) that utilizes the digits 0-3 to mark the sequenced color, the processing of which we will not cover in this section. Though these different scoring methods potentially contribute to misinterpretation and confusion, they can be easily converted and conformed (Cock et al., 2009; Goto et al., 2010; Holland et al., 2008; Stajich et al., 2002). Most current analysis tools are able to handle both scoring methods, though some require specific parameters to be set for dealing with each. When employing these analysis tools, one should mind the appropriate quality score is used.

Quality control of deep sequencing data refers to an overview on the base and quality distribution between lanes, tiles and cycles, and correlating the initial sequence data with expected length, GC content, ambiguous bases, sequence complexity and alignment ensuing location distributions which can hold information regarding possible sequencing bias, contamination or artifacts. Platform specific quality control tools.

(Cox et al., 2010; Dolan and Denver, 2008; Martinez-Alcantara et al., 2009) and more general quality assessment software (Dai et al., 2010; Schmieder and Edwards, 2011) can help circumvent such biases, by both raising awareness to implicating irregularities with textual and graphical data representation and by removing such low quality or aberrant sequences prior to the downstream analysis. The need for careful quality control is exemplified by deep sequencing data with a tile specific A base bias, leading to over-expression of the base in the sequences derived from that tile. When searching for rare sequence variants, such base over-expression should be considered when sequences supporting an A variant are derived from the aforementioned tile. A more common example is sequence duplication (Gomez-Alvarez et al., 2009),usually an artifact of PCR amplification and other library preparation processes, that cause over-representation of certain sequences. This creates a skewed coverage distribution that may subsequently bias the error model and thus substantially increase the number of false-positive SNP discoveries and tilt expression and metagenomic analysis results. Available quality control software allow the user to completely remove these duplicates (*FASTX - toolkit*; Li et al., 2009) or mark them for downstream analysis consideration (*PICARD*). Recently various algorithms utilizing suffix tree data structures were developed for sequencing error correction (Kelley et al., 2010; Zhao et al., 2010).

A common procedure in the pre-analysis process, following initial quality control, and prior to sequence duplication removal, is the compulsory tag / adapter removal (Lassmann et al., 2009; Schmieder et al., 2010) and optional quality trimming. Tags are used during the library preparation phase for amplification or differentiation processes (e.g multiplexing; Galan et al., 2010). If they are sequenced, they can profoundly affect the downstream analysis unless removed (e.g clipping). The clipping process, removes any tag remnants from the sequence reads, ridding the data from reads composed mainly or even solely of the tags. The user must set the minimal read length to be retained (according to the sequenced sample and experimental question) and consider possible sequence similarities between the sample and the adapters. Trimming, refers to the sequence removal from either the 5' or the 3' ends of a read where either the sequence complexity or quality does not pass user settings. It is often used for poly-A or poly-T removal, or removal of bases with significantly lower, bias introducing quality scores. Unlike clipping, which is mandatory for valid downstream analysis, trimming is only recommended to improve accuracy and performance in

subsequent analysis steps such as alignment and assembly. Following both clipping and trimming, the researcher may review the sequence data for size distribution, and verify concordance with the experimental context. For example, when performing microRNA sequencing experiments, one would expect the sequence size composition to be approximately 20-24 nts in length. If the majority of the data deviates from this range, a more careful examination of the information is in order and library preparation bias should be considered.

We urge the user to consider the sequencing data in the appropriate experimental context and utilize the aforementioned quality control and assurance methods prior to the downstream analysis to increase the experimental validity and accuracy and to ensure better, more reliable results.

4. Data analysis pathways

In the previous sections, we covered common deep sequencing data considerations and refinement, crucial and beneficial for all types of down stream analysis. In this section, we will go over the common data analysis pathways and possibilities, covering their appropriate utilization, the benefits and limitations of each pathway, and familiarizing the user with some of the common available analysis tools.

4.1 Alignment

Most of the analysis pathways specified below involve an initial step of mapping the deep sequencing reads against a reference genome of either the sequenced species, or a related organism with sufficient genetic resemblance. This step presents a computational challenge due to the sheer amount of short reads produced in deep sequencing experiments. It is further complicated by nucleotide and structural variance, sequencing errors, RNA editing and epigenetic modifications. When deep sequencing was initially introduced, established early-generation sequence alignment tools (Altschul et al., 1990; Kent, 2002) more suited for the query of a limited number of sequences were less appropriate for high throughput sequencing's millions of short sequence fragments mapping (Trapnell and Salzberg, 2009), requiring novel alignment algorithms and tools to be specifically designed. Current short read alignment tools.

(Langmead et al., 2009; Li and Durbin, 2009; Li et al., 2008; Li et al.,2009; Lin et al., 2008; *Novoalign*), utilize various heuristic techniques for alignment of millions of short sequences within an acceptable time requirement (Flicek and Birney, 2009). This section will not cover the underlying algorithms for each tool (Li and Homer, 2010). Instead, we will address a few imperative features to be considered when initiating data analysis and alignment.

When choosing an alignment tool, one needs to consider the memory and time requirements and limitations and the appropriateness of the tool to the exploratory question at hand. Some important features to be considered include:

Quality utilization and control - As we mentioned before, sequencing quality provides the user with initial assessment of the data. Some alignment tools, utilize these quality scores (Langmead et al., 2009; Li et al., 2008; *Novoalign*) and it was shown that such employment greatly improves the mapping performance (Frith et al., 2010; Li and Homer, 2010). Most common alignment software generate the alignment output in the Sequence Alignment Map (SAM) format (Li et al., 2009), with a multitude of supporting downstream analysis tools. This common format provides users with a simple and flexible common ground to evaluate

alignment results and easily extract and utilize data for further analysis. As for the sequencing output, so does the alignment output contain a PHRED based quality score for each of the aligned reads, describing the probability of per-base false alignment. Combination of this quality score together with other alignment parameters such as mismatches could and should be further assessed using specialized tools (Lassmann et al., 2011) in order to characterize mapped and unmapped reads for potential alignment improvement. These alignment quality scores can be re-assessed using currently available tools (McKenna et al., 2010; *Novoalign*), so that they better denote the probability of a mismatch between the aligned base and the reference sequence. This quality recalibration takes into account the given base and its quality score, the position within the read and the adjacent nucleotides to account for sequencing chemistry biases (Li, Li et al. 2009), and was shown to reduce the effect of sequencing technology derived biases and improve overall variant detection fidelity (DePristo et al., 2011).

Gapped alignment – An important feature one should be mindful of when choosing an alignment tool is whether the tool utilizes the gapped alignment algorithm. Since gapped alignment only mildly increases alignment sensitivity, it is not crucial to pick a supporting tool for many general purposes. However it is especially crucial for variant calling, specifically insertions and deletions (indels) detection (Krawitz et al., 2010) and it is highly recommended to choose a tool that implements gapped alignment (Li and Durbin, 2009; Li et al., 2008; *Novoalign*), when venturing on variant detection experiments, or when targeting known indel abundant areas.

Mismatches and Gap penalties – Most alignment tools allow the user to set the number of allowed mismatches between the read and a reference location and the scoring scale for gap opening and extension. Allowing more mismatches results in a higher portion of mapped reads but at the cost of increased ambiguity and reduced confidence of these alignments. Mismatch allowance should be set while considering the specific experiment at hand. For example, when undergoing microRNA expression profiling, one will want an accurate estimate of the abundance of each microRNA , and should not allow a high mismatch rate if any. On variant calling experiments however, the user should consider the possible expected size range of the variants before setting the allowed mismatch and gap penalty parameters (e.g, if one aims to find a >5nt long deletion, the mismatch limitation should allow it).

Multiple mapping - In theory, unique alignment, mapping a read to a single unique loci on the reference genome is expected by most reads longer than 30 nts when aligning against a large human scale reference. Usually, a portion of the reads will remain unmapped due to contaminant origin or sequencing errors, or more commonly, they will ambiguously map to several different locations (multiple mapping) due to sequence homology and repetitiveness. Different alignment tools flag these multiply mapped reads, and provide the user with the option to either randomly assign them to one loci (Li and Durbin, 2009) or just output all of them (*Novoalign*). Researchers may choose to incorporate only uniquely mapped reads into their downstream analysis, or set a maximal number of different mapping locations for incorporated reads. Discarding multiply mapped reads results in loss of a substantial portion of the data, with potential crucial effects on the following analysis. Currently, there are several approaches for allocation of these multiply mapped reads. One method is to count each read as if originating from each of the mapped loci, potentially over-estimating the expression or coverage of some, since the same read could not have originated from more than one loci. Another method is to divide each read count between

all its mapped loci, adding a small equal portion to each. This could have the opposite effect of under-estimating expression and coverage, especially for low complexity loci. Several methods utilize heuristics for dividing these reads amongst their mapped loci according to the uniquely mapped reads in those regions (Hashimoto et al., 2009; Mortazavi et al., 2008). A fairly novel approach utilizes probabilistic models such as maximum likelihood to compute the most likely origin of each read greatly improving the results of quantitative deep sequencing experiments and differential expression (Paşaniuc et al., 2011).
Since each parameter can greatly affect various performance attributes, considering the aforementioned features is crucial when initiating deep sequencing data alignment. The user should always mind the alignment tool's inherent limitations and implement parameters settings according to the experiment at hand and the expected possible downstream analysis, picking an appropriate tool and tuning necessary features for optimal alignment results.

4.2 Assembly

Assembly refers to the process of piecing together short DNA/RNA sequences into longer ones (e.g contigs) which are then grouped to form scaffolds for computationally reconstructing a sample's genetic component. When the assembly process is performed with the assistance of a reference genome, it is referred to as mapping assembly, if no reference is available it is called *de novo* assembly. Original computational assembly tools were designed to use capillary-based sequencing's 800 base pairs long sequences in order to deduce the original full sequence through examination of overlapping segments. Deep sequencing data presents a more compound assembly problem due to higher amounts of sequences that are significantly shorter. Though it adds complexity to the process, this significant increase in throughput enables the successful realization of whole mammalian genome *de novo* assembly as shown in (Li, Fan, et al., 2010; Li, Zhu, et al., 2010). Sequencing errors, uneven genome coverage and reads too short to be informative in repeated regions required a new breed of assembly tools designed specifically for short reads (Butler et al., 2008; Chaisson and Pevzner, 2008; Dohm et al., 2007; Jeck et al., 2007; Li, Zhu, et al., 2010; Margulies et al., 2005b; Simpson et al., 2009; Treangen et al., 2011; Zerbino and Birney, 2008). These tools mainly rely on two algorithms, and differ mostly in the way they deal with sequencing errors and inconsistencies and sequence repeats. Since tools utilized today could be either deprecated or significantly changed in the near future, we will not address the underlying advantages and disadvantages for each specific tool. We will, however, cover some of the more general inherent challenges of deep sequencing data assembly and recommended optimization methods.
Assembly Algorithms - Currently there are two main models for deep sequencing data assembly, Overlap-Layout-Consensus (OLC) (Myers, 1995) which calculates overlaps by (computationally expensive) pairwise alignments, and de Bruijn graph-based (DBG) which creates a shared k-mer dictionary for the assembly process. K is often set by the user and it is recommended that it be set large enough so that most overlaps are true and do not occur by chance, and short enough so as to allow overlap between related sequences. Since comprehensive reviews are available on these algorithms (Miller et al., 2010), we will focus more on specific algorithm related considerations for tool selection. A recent overview comparing the performance of a variety of tools for assembly under different conditions (Zhang et al., 2011), recommended the use of OLC based assemblers (Hernandez et al., 2008; Margulies et al., 2005b) for small scale (e.g microorganisms) genome assemblies While

reserving the use of the less computationally demanding DBG based tools for the assembly of large (eukaryote) genomes (Butler et al., 2008; Li, Zhu, et al., 2010; Simpson et al., 2009; Zerbino and Birney, 2008). Another consideration is the read size, with OLC being most appropriate for a limited number of fairly long reads (~100-800 bp) and DBG more suited for the assembly of millions of short reads (25-100 bp) (Miller et al., 2010). One should note that DBG based tool's implementation of specific heuristics reduces CPU demand but at the cost of higher sensitivity to sequencing errors that could result in a much higher memory requirement. We therefore urge the user to run a more strict quality assessment and filtration when embarking on DBG based assembly. We also note that some of the assembly challenges such as identical repeat regions longer than the sequenced reads length, remain insurmountable by computational and algorithmic improvements and must be alleviated by technical means such as longer reads or paired-end sequencing (Cahill et al., 2010).

Quality Assessment - An assembly's quality is measured by it's contiguity and cumulative size and the accuracy of the assembly. The contiguity is assessed using length statistics such as contig and scaffold maximal and average length, combined total length and N50 (The length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly (Miller et al., 2010)). An assembly's accuracy is more difficult to assess and external data is usually needed to reveal both miss-assembly (e.g sequences that are inaccurately joined) and per base accuracy (e.g contigs with nucleotide mismatches). One way to estimate fidelity is by utilizing paired end reads, re-aligning them against the assembled contigs to reveal discrepancies in insert size which probably indicate wrong assembly. When there are available reference sequences they should be utilized for further validation of the assembled contigs, matching sequences and marking possible mismatches and chimeras (non-related sequences assembled into one contig). If no reference sequence is available, it has been shown that the sequence of available closely related organism (e.g comparative assembly (Pop et al., 2004)) could be utilized for the same purpose and for contig adjacency assessment (Gnerre, et al., 2009; Husemann and Stoye, 2010; Meader et al., 2010). A crucial aspect of assembly quality assurance is the sequence quality. Erroneous sequence reads result in higher computer memory requirements (especially in DBG based tools (Miller et al., 2010)) and either no assembly output or wrong inaccurate contigs. As part of the assembly related quality assurance, it is recommended to discard all reads with ambiguous bases (e.g N) and reads composed entirely of homo-polymer sequences to alleviate this increase in computational demand (Paszkiewicz and Studholme, 2010). It is also good practice to trim low quality bases from read edges and of course remove adapters prior to assembly.

Assembly represents one of the more challenging computational tasks at present and it is further complicated when implemented on deep sequencing data. General considerations mentioned in this chapter will help the user to both better understand the challenges inherent in the sequence data and to match a selected tool's underlying implemented algorithm with the data at hand and the assembly goals. Moreover, assembly quality could now be better assessed using the aforementioned parameters, such as N50 and fidelity, in order to compare assembly tools performance for both existing and future software.

4.3 Variant calling

Variant calling refers to the identification of single nucleotide polymorphisms (SNPs), insertions and deletions (indels), copy number variations (CNVs) and other types of structural variations (e.g inversions, translocations etc.) in a sequenced sample (Durbin et

al., 2010). Detection of these variants from deep sequencing data requires in most cases both a reference genetic sequence to compare the sequence data against (Li, Li, et al., 2009), and a specialized variant calling software that utilizes probabilistic methods for correctly inferring variants. The process is complicated by areas of low coverage, sequencing errors, misalignment caused by either low complexity and repeat regions or adjacent variants and library preparation biases (e.g PCR duplicates) (Chan, 2009). Variant calling depends on an efficient combination between an accurate alignment and sophisticated inference of variance from it. Since alignment optimization was already discussed in a previous section, in this section our focus will be more on aspects of variant deduction. We will cover the basic common challenges and difficulties both general and specific for each variant type, Present leading bioinformatic tools and databases and their contributions to the field and provide the user with critical considerations and solutions for some of the aforementioned challenges.

After initial alignment, certain factors can critically alter the results of variant detection. One should consider them prior to downstream analysis and implement the appropriate modifications if necessary.

Depth of coverage – Previous studies demonstrated positive correlation between variant calling sensitivity and increased read depth (Krawitz et al., 2010). Depth can be increased by either reducing the size of the selected or enriched target region, performing a higher number of sequencing cycles to produce longer reads to cover the target region or simply assigning more sequencing lanes. Each method has its benefits and drawbacks. For example, assigning an additional lane to sequence the same sample requires a higher financial investment but allows better noise filtration and sequencing errors recognition. Targeting a specific region increases the coverage and sensitivity at the selected segment, but at the cost of information loss at the areas outside. After the sequencing process is complete, upper and lower depth thresholds should be applied on the sequencing data before variant calling is performed. Setting a lower coverage limit removes erroneous mismatches caused by sequencing errors and thus supported by very few reads (Durbin et al., 2010; Li, Li, et al., 2009). Although it is recommended on most tools, setting a lower limit has been shown to reduce sensitivity without increasing specificity in some tools (Goya et al., 2010) and therefore should be considered in the context of the utilized tool. Setting an upper limit removes mismatches caused by copy number variations, PCR duplicates introduced by library preparation (Gomez-Alvarez et al., 2009) and reads mapping to paralogous sequences. The limit should be set according to the initial coverage and we recommend setting the limit to ~10 times the average coverage. PCR duplicates should be further assessed, removed and marked using specialized tools (Li et al., 2009; *PICARD*) as mentioned in the pre-analysis processing section.

Mapping quality and Quality recalibration – Some reads mapping to under represented regions in the genome, especially low complexity and repetitive regions will be inaccurately mapped with a low mapping quality. SNPs derived from these reads have higher chance of being false-positives (Durbin et al., 2010) and should be more carefully examined, setting a more strict quality and coverage threshold if possible. As mentioned in the prior section of alignment considerations, quality recalibration increases the validity of the alignment qualities so that they better denote the probability of a mismatch between the base and the reference. Naturally, these re-calibrated qualities improve the efficiency of variant detection tools that incorporate alignment qualities into their calling algorithms (Koboldt et al., 2009; Li, Li, et al., 2009; McKenna et al., 2010; Qi, et al., 2010).

Cross-lane comparison – It is good practice, when different-lane same-sample sequences are available, to compare the amount of SNPs, insertions and deletions detected for each lane. If one of the lanes has a significantly higher amount of detected variants, it is probable that it will introduce false-positives to the analysis and exclusion of that lane from downstream variant calling is recommended. Another possible data validation option is comparison against a SNP chip if available (Koboldt et al., 2010). Going over each annotated SNP provides the user with more than a million checkpoints to ascertain both the validity and fidelity of the sequencing process, and the chromosomal representation (e.g haploid or diploid).
We will now address a few more variant specific considerations and applications.

4.3.1 Single nucleotide polymorphisms

After aligning deep sequencing reads against a reference genome, SNPs can be naively inferred from the results by simply denoting each base that is inconsistent between reference and read as a SNP. This straightforward inference of mismatches results in a massive amount of alleged SNPs, many of which suffer from some sort of inaccuracy such as: calling a mismatch in the wrong location, homozygousity and heterozygousity discrepancies and even calling a mismatch in the correct location but with the wrong base. Currently most SNP calling tools (Koboldt et al., 2009; Li et al., 2009; 2008; Li, Li, et al., 2009; McKenna et al., 2010; Qi et al., 2010) apply different probabilistic based considerations and heuristics such as quality assessment and recalibration, SNP filtration, local realignment, coverage assessment, prior probability based on known SNPs, genotype based likelihood and even cancer genomics (Goya et al., 2010) to elucidate SNPs from alignment results. The user should be familiar with these considerations and be aware of the tools that apply each when performing SNP calling. We will go over some of them and discuss their effects and benefits.
Local realignment – Current mapping tools align reads independently of the alignment region context. If a read's beginning or end maps to a region containing an indel, a mismatch will be called instead of an indel due to alignment scoring considerations. Adding a secondary, local alignment that considers reads that support the presence of an indel in the vicinity of either detected SNPs or known SNP sites retrieved from dbSNP (Day, 2010), results in a significant reduction in false positive SNPs (Durbin et al., 2010; McKenna et al., 2010). This local realignment is highly recommended prior to SNP analysis and is either performed inherently in some tools (Qi et al., 2010) or can be specifically performed using other available tools (McKenna et al., 2010).
Base Alignment Quality – Since local realignment is a computationally intensive process that depends on correctly denoting insertions and deletions, another method for increased SNP detection accuracy is purposed (Li, 2011). Implementing a per-base alignment quality recalibration for re-evaluation of misalignment probability using profile hidden markov models. This quality recalibration can be performed using SAMtools (Li et al., 2009).
Transition / Transversion Ratio (Ti/Tv) – The expected ratio between transitions (e.g purine purine substitutions) and transversions (e.g purine pyrimidine substitutions) can be elucidated from empirical data retrieved from the 1000 Genomes project Durbin et al., 2010). This ratio could be utilized as an initial quality assessment standard. Currently the expected Ti/Tv ratio is ~2.3 for whole-genome sequencing and around 3.3 for whole-exome sequencing (coding regions only) (DePristo et al., 2011). When detected SNPs demonstrate a ratio closer to the expected ratio for random substitutions, with transversions twice as

common as transitions (e.g ~0.5), low quality variant calling or data is implied and quality thresholds should be reassessed.

dbSNP validation – After producing a list of detected SNPs, it is highly recommended to compare it against dbSNP, the largest repository of SNP data found within the National Center for Biotechnology Information database. Detected SNPs present in the database are considered as known, and the ones not found are considered novel (Li and Stockwell, 2010). The portion of novel SNPs detected in a deep sequencing experiment should range between 1 and 10 percent (DePristo et al., 2011). If this proportion is higher, a high rate of false positive variants is suggested and we recommended reevaluating the detection process and possibly implementing a more strict variation inclusion criteria.

4.3.2 Insertions and deletions (Indels)

Indels are the second most common type of polymorphism and the most common structural variant, in this sub-section we will address only short indels as the next section will deal with the larger (>1000kb) structural variants. Most indels range between 2-16 bases in length (Mullaney, et al.,2010) (also referred to as micro-indels) and their frequency has been shown to vary across the genome with lower rate in conserved and functional regions and an increased rate in hot spots for genetic variation. The average indel rate is approximately one indel in 5.1 to 13.2 kb of DNA (Mills et al., 2006). Their presence implicates on the pathogenesis of disease, gene expression and functionality, viral disease forms identification and they can be used as genetic markers in natural populations. Indels occur in an estimated rate that is eight-fold lower than SNPs (Durbin et al., 2010). This rate varies extensively between sequenced individuals, usually due to variability between mapping and detection tools. Reads covering an indel are generally more difficult to map since their correct alignment either involves complex gapped alignment or paired-end sequencing inference. Optimal indel detection is performed by combining application of an appropriate alignment software and variant detection tool (Albers et al., 2010; Koboldt et al., 2009; Li et al., 2009; McKenna et al., 2010; Qi et al., 2010; Kai et al., 2009), and careful adjustment of their parameters according to the suspected variants. As mentioned before in the alignment section, it is highly recommended to perform indel calling with alignment tools that implement gapped alignment (Krawitz et al., 2010; Li and Durbin, 2009; Li et al., 2008; *Novoalign*). A few considerations when addressing insertion-deletion detection:

Read length – Increasing the read length has been shown to improve the ability to map and detect insertion related reads. Sequence reads 36 bases long, such as the ones produced by the Illumina GAIIx, have been shown to be inefficient for detection of insertions longer than 3 bases with a complete inability to detect insertions longer than 7 bases. Hence the length of the sequenced reads should be considered according to the insertion size range suspicion and adjusted appropriately. Naturally, when insertion size is expected to surpass the read length it is impossible to detect them using single-end sequencing. Increasing the read length has also been shown to improve micro-indel (<10 bases) detection sensitivity without significantly affecting specificity, demonstrating a more efficient method for increasing coverage than simply producing more reads.

Paired-end reads – Indel detection greatly improves when based on paired end reads deep sequencing data (Mullaney et al., 2010). Both alignment (Li and Durbin, 2009; Li et al.,2008) and variant detection tools (Kai Ye et al., 2009) utilize paired-end reads so that one of the reads is used to pinpoint the pair's loci in the reference while the other read can be subjected to gapped alignment and indel inference. Furthermore, the insert (e.g the unsequenced gap

between a read pair) can also be used to deduce the presence of an indel (discussed in the next section).

4.3.3 Structural variants

Structural variants (Feuk et al., 2006) are defined as genomic alterations that involve segments of DNA that are larger than 1 kb. They include: (1) Copy number variations (CNV), which are sections in the DNA with a variable copy number when comparing to a reference genome. Insertions, deletions and duplications are types of CNVs. (2) Segmental duplications, several copies of DNA segments that are almost identical (>90%) that can appear in a variable number of copies, also considered a type of CNV. (3) Inversions, segments in the DNA that are reversed in orientation. (4) Translocations, an intra or inter chromosomal location shift in a DNA segment without changing the total DNA content. (5) Segmental uniparental disomy, where a diploid individual's pair of homologous chromosomes originated from a single parent. Since current deep sequencing platforms do not produce reads that span the length of structural variants, utilization of paired end mapping is necessary for their exact elucidation. The quality of structural variation detection using deep sequencing can be assessed by the accuracy of break point localization, copy number count and variation size estimation (Medvedev et al., 2009).

Paired-end mapping (PEM) - Paired-end sequencing refers to the process of sequencing a cloned DNA fragment on both ends, resulting in two associated sequence reads with an unsequenced insert between them. The insert length varies between several bases to several thousands of bases and thus appropriate for the detection process of the aforementioned large scale structural variants. Structural variants are often detected indirectly through associated paired-end deep sequencing data patterns (Bashir et al., 2008; Korbel et al., 2009; Medvedev et al., 2009). Some of these patterns approximate the location of the structural breakpoints, and some provide an exact localization. For example, the signature of an insertion or deletion can be easily inferred by comparing the expected read pair distance according to the reference with the expected insert size, if the reference distance is longer or shorter than the insert size, the presence of a deletion or insertion can be inferred respectively but the deduction of the exact location of the indel from these signatures is more difficult. However, in an "anchored split mapping" signature, when one read from a pair is perfectly aligned against the reference and its pair cannot be aligned against its designated reference location, the unaligned read can be utilized in order to pinpoint the exact location of existing large deletions or small insertions (Medvedev et al., 2009). SV PEM signatures improve all aspects of SV detection quality and so PEM is highly recommended for this purpose.

Insert size – insert size, set by the size of the DNA fragments introduced by library preparation can affect the outcome of SV detection. If the experimental goal is to detect as many structural variants as possible, a larger insert length is suggested. If however, a more precise localization of the variants is necessary, a shorter insert length is recommended, though resulting in an overall lower variant discovery sensitivity (e.g if you find the variant, there is a greater probability of precise localization) (Bashir et al., 2008).

Depth of coverage – Coverage can also be utilized for SV elucidation, specifically large scale deletions and duplications (Yoon et al., 2009). As we expect reads mapping to each region to follow a Poisson distribution, deviations from the expected coverage suggest the presence of a duplication or deletion. SV detection benefits from combining increased coverage with abundance of paired-end reads with a significant increase in specificity (Bashir et al., 2008). Coverage cannot be utilized however to elucidate the exact location of these SVs, only to

suggest the expected region in some cases. Coverage biases, as mentioned in previous sections, can also misconstrue the SV detection process.

Clustering methods – After SV signatures have been detected, a calculated inference process is crucial. Most methods utilize some sort of clustering for all the pairs supporting a variant, and deduce variant information from that cluster. Since most methods rely on PEM and coverage for SV detection, they differ mainly on their clustering methods. It is important to familiarize with these methods since some are more suited for certain variation types. Since describing each clustering method is beyond the scope of this chapter, we will only point out certain aspects that are both easy to implement and have significant effects on detection quality. The standard clustering method (Korbel et al., 2009) utilizes only uniquely mapped read pairs, and discards the ones mapped to multiple loci. It also utilizes a set standard deviation limit for the difference between known insert size range and the observed mapped reference distance. These "hard" filters reduce the effectiveness of both homozygousity/heterozygousity inference and small scale variation detection (Medvedev et al., 2009). Soft (Hormozdiari et al., 2010) and Distribution (K. Chen et al., 2009; McKernan et al., 2009) based clustering methods consider multiply mapped reads and assigns them according to their supporting context, thus increasing sensitivity for the presence of small indels and heterozygousity and should be considered when experimentally relevant.

Validation by assembly – It is recommended to combine *de novo* assembly with structural variant detection in order to validate detected variants. Once the assembly process is complete, a search for supporting and conflicting sequence contigs should be performed (Koboldt et al., 2010).

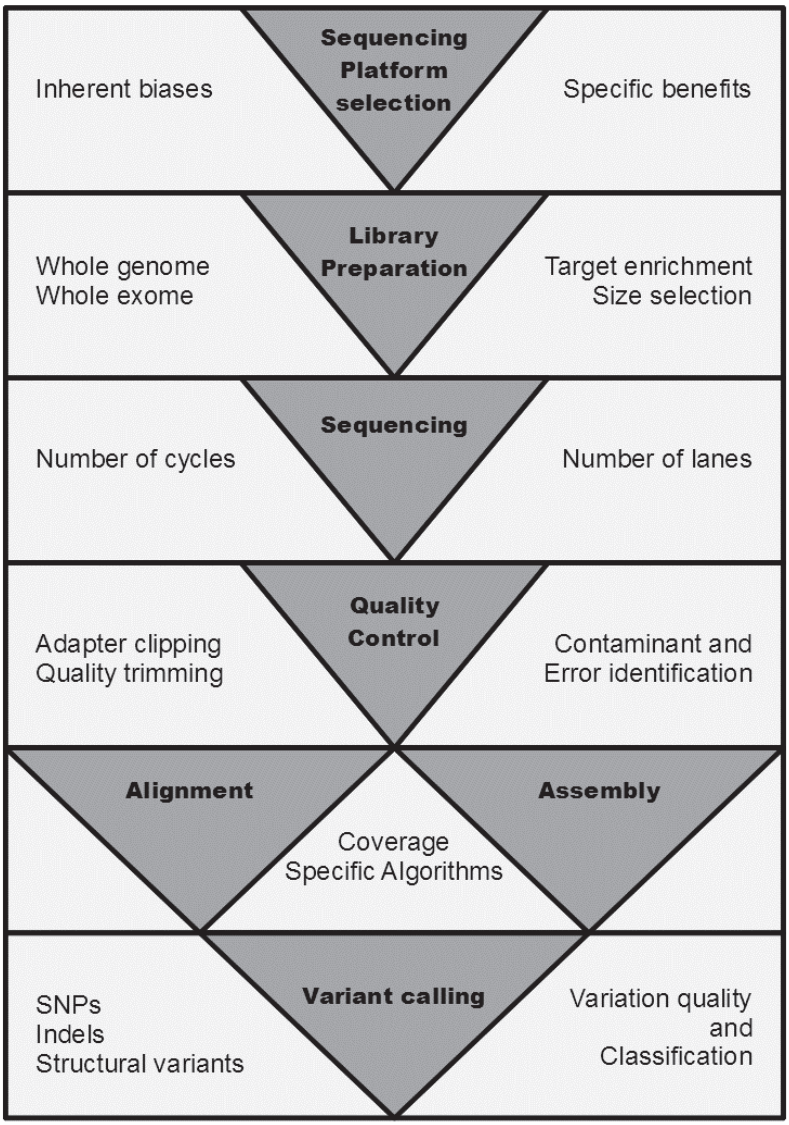
We note that the field of structural variation deduction from deep sequencing data is still in its infancy and both false-positive and false-negative rates are far from satisfactory (Hormozdiari et al., 2009). As both sequencing technology improves, raising coverage and read length, and the algorithmic utilization of such improvements continues, we expect greater utilization of deep sequencing for SV detection.

4.3.4 Variant classification

Calling variants using deep sequencing data often results in a multitude of detected variations, even after strict and effective quality filtration as denoted earlier, deep sequencing data reveals thousands to millions of different variations (Imelfort et al., 2009). These variations can result in biological effects through introduction of different amino acids into protein sequences, early termination of coding sequences and alteration of regulatory elements and splice sites. A natural step following the variant calling process is annotating the detected variants and elucidating their effect and biological significance, separating clinically, scientifically and medically relevant variations from neutral, non functional ones. In a large list spanning this many variants, manual annotation of each variant effect is neither feasible or accurate. We will therefore cover currently leading principles for computational classification and prioritization of detected variants.

Initial prioritization - The first step in variation characterization is basic variation properties deduction. Variant properties such as it's location, whether in a known coding sequence, non coding transcript, promoter, splice site etc. Once a variation is localized in a coding sequence, a subsequent analysis of it's frame effect and whether its synonymous (e.g changing an amino acid) or non-synonymous should be performed. These basic properties allow initial prioritization of the variation list, considering that coding sequence non-sense

mutations are more likely to be functionally relevant than mutations in an unexpressed genomic sequence. When dealing with an annotated genome, computational tools should be utilized for this purpose (Conde et al., 2006; Li and Stockwell, 2010; McLaren et al., 2010; Yuan et al., 2006). We recommend checking the dbSNP version utilized by chosen annotation tools and strive to employ the most up-to-date version available so as to increase the availability of variant annotations.



Sequencing strategy decision flowchart

Coding sequence variants - In order to ascertain the most likely phenotype affiliated coding sequence variation from a given list, current variation profiling methods utilize biochemical and physical properties of both amino acids and proteins considering both structure (Ramensky et al., 2002) and function (Bromberg and Rost, 2007; Calabrese et al., 2009) and utilizing various probability algorithm (Mi, et al., 2007). Possible incorporated characteristics include: molecular mass, polarity, acidity, basicity, aromaticity, conformational flexibility and hydrophobicity of amino acids (Ng and Henikoff, 2006) and hydrogen bond breaks, introduction of a buried polar residue, loss of salt bridge, insertion of proline into α-helix, and the breaking of disulfide bonds in proteins (Wang and Moult, 2001). Some available tools (Ashkenazy et al., 2010; Kumar et al., 2009; Li et al., 2009) utilize the fact that functionally crucial amino acids are evolutionary conserved, by employing multiple sequence alignment based conservation scores in order to prioritize given variations. Utilizing orthologous sequences for this purpose demonstrates higher efficiency than incorporation of paralogous, since the latter represents proteins with slight differences in both sequence and function and is less informative for conservation analysis. It was shown that conservation degree is in fact the most reliable method for predicting possible pathogenicity of a missense variant (Flanagan et al., 2010).
For the purpose of both prioritization and functional analysis optimization, we recommend combining available annotation tools that employ a variety of prioritization features (George et al., 2008; Lee and Shatkay, 2008). A recent study implemented some of these variation classification methods on recorded SNPs in a target gene, in order to elucidate possible cancer causing mutations, reducing the initial number of suspected SNPs from thousands to less than 30 (Choura and Rebai, 2009). Another study utilized bioinformatic tools to classify known non synonymous mutations in colon cancer and was able to pinpointed four SNPs already known as related to increased cancer risk (Doss and Sethumadhavan, 2009). However, a recent comprehensive review (Karchin, 2009), implemented and compared leading variant classification tools on three different studies (Doecke et al., 2008; Fatemi et al., 2008; Van Deerlin et al., 2008) associating both exonic and intronic, novel and known SNPs with a variety of disease, and demonstrated that a combination of several tools can possibly result in conflicting annotations and functional effects deduction. Another comparison (Thusberg et al., 2011), that tested the performance of several of the aforementioned tools in predicting pathogenicity using test data retrieved from dbSNP, demonstrated the sensitivity characterizing these tools to range between 0.59 to 0.9, with the preferred tools for their analysis to be SNPs&GO and MutPred (Calabrese et al., 2009; Li et al., 2009). Both studies agree that inference of functionality and pathogenicity is not a fully automatic pathway and educated interpretation of the results must be conducted.

5. Conclusion

Deep-sequencing data analysis is a growing field with many computational challenges. A normal deep sequencing run outputs a massive amount of data which require complex computational processing and interpretation. The overflow of available bioinformatic tools and software for each of the optional analysis steps presents a challenge for the researcher aiming to evaluate and interpret deep sequencing data. In this chapter we familiarized the reader with crucial concepts and considerations for preparation, refinement, analysis and elucidation of valid and accurate conclusions. The field is rapidly evolving both in hardware and sequencing platform technology and in computational techniques, algorithms, software

and tools. It is crucial to understand the various challenges involved in deep sequencing experiments, and the current available solutions, both in concept and in practice. The concepts presented in this chapter are aimed towards optimizing deep sequencing experiments, concentrating on initial steps of data preparation and quality refinement and covering several possible analysis pathways while denoting some of the currently available and leading tools, and some of their underlying methods.

The first section of this chapter, introduced deep sequencing technology's available platforms in regards to their advantages and limitations, emphasizing that although they are all considered high throughput sequencing platforms, they present different capabilities and proficiencies. When a choice between platforms is available, one can improve data retrieval and validity simply by matching the most appropriate platform with the specific experimental needs.

The second section covered the concept of deep sequencing data quality control. Using bioinformatic tools, based on both empirical and probabilistic deduction, sequencing derived errors can be reduced which otherwise would be incorporated into downstream analysis. We described current quality scales, with methods for their assessment and their relevance for improved data retrieval. Employment of these quality control and assurance methods can assist in uncovering biased sequencing lanes and recurring errors and contaminants that could significantly alter deep sequencing results. We therefore strongly urge users to utilize them prior to any following experimental evaluation, making their incorporation a standard in deep sequencing experiments.

The third and subsequent sections covered specific and very common analysis pathways: alignment, assembly and variant calling. The chapter introduced basic challenges faced in each type of analysis, their current limitations and considerations pivotal for preferential experimental planning. A description of each challenge was accompanied by delineation of current methods, tools and solutions when available. Familiarized with these challenges, the user can now conduct better analytic decisions and employ the most appropriate tools and techniques. Understanding the exact edge of each analytic pathway can help the user to perform their deep sequencing experiments in the most effective manner employing both current and future software for optimal variant calling.

6. References

Albers, C. A., Lunter, G., Macarthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2010). Dindel: Accurate indel calls from short-read data. *Genome Research*. doi:10.1101/gr.112326.110

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-410. doi:10.1006/jmbi.1990.9999

Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38, W529-533 (2010).

Bashir, A., Volik, S., Collins, C., Bafna, V., & Raphael, B. J. (2008). Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Computational Biology*, *4*(4), e1000051. doi:10.1371/journal.pcbi.1000051

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53-59. doi:10.1038/nature07517

Bromberg, Y., & Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823-3835. doi:10.1093/nar/gkm238

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., et al. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research*, *18*(5), 810-820. doi:10.1101/gr.7337908

Cahill, M. J., Köser, C. U., Ross, N. E., & Archer, J. A. C. (2010). Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. *PloS One*, *5*(7), e11518. doi:10.1371/journal.pone.0011518

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, *30*(8), 1237-1244. doi:10.1002/humu.21047

Chaisson, M. J., & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes. *Genome Research*, *18*(2), 324-330. doi:10.1101/gr.7088808

Chan, E. Y. (2009). Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery. *Methods in Molecular Biology (Clifton, N.J.)*, *578*, 95-111. doi:10.1007/978-1-60327-411-1_5

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Meth*, *6*(9), 677-681. doi:10.1038/nmeth.1363

Chiu, R. W. K., Sun, H., Akolekar, R., Clouser, C., Lee, C., McKernan, K., Zhou, D., et al. (2010). Maternal Plasma DNA Analysis with Massively Parallel Sequencing by Ligation for Noninvasive Prenatal Diagnosis of Trisomy 21. *Clin Chem*, *56*(3), 459-463. doi: 10.1373/clinchem.2009.136507

Choura, M., & Rebaï, A. (2009). Applications of computational tools to predict functional SNPs effects in human ErbB genes. *Journal of Receptor and Signal Transduction Research*, *29*(5), 286-291. doi:10.1080/10799890902911948

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422-1423. doi:10.1093/bioinformatics/btp163

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767-1771. doi:10.1093/nar/gkp1137

Conde, L., Vaquerizas, J. M., Dopazo, H., Arbiza, L., Reumers, J., Rousseau, F., Schymkowitz, J., et al. (2006). PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Research*, *34*(Web Server issue), W621-625. doi:10.1093/nar/gkl071

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

Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., Omenn, G., et al. (2010). NGSQC: cross-platform quality analysis pipeline for

deep sequencing data. *BMC Genomics*, *11 Suppl 4*, S7. doi:10.1186/1471-2164-11-S4-S7

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., Bouffard, P., et al. (2010). Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biology*, *8*(9). doi:10.1371/journal.pbio.1000475

Day, I. N. M. (2010). dbSNP in the detail and copy number complexities. *Human Mutation*, *31*(1), 2-4. doi:10.1002/humu.21149

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, *advance online publication*. doi:10.1038/ng.806

Doecke, J., Zhao, Z. Z., Pandeya, N., Sadeghi, S., Stark, M., Green, A. C., Hayward, N. K., et al. (2008). Polymorphisms in MGMT and DNA repair genes and the risk of esophageal adenocarcinoma. *International Journal of Cancer. Journal International Du Cancer*, *123*(1), 174-180. doi:10.1002/ijc.23410

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2007). SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, *17*(11), 1697-1706. doi:10.1101/gr.6435207

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105. doi:10.1093/nar/gkn425

Dolan, P. C., & Denver, D. R. (2008). TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics*, *9*, 250. doi:10.1186/1471-2105-9-250

Doss, C. G. P., & Sethumadhavan, R. (2009). Investigation on the role of nsSNPs in HNPCC genes--a bioinformatics approach. *Journal of Biomedical Science*, *16*, 42. doi:10.1186/1423-0127-16-42

Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061-1073. doi:10.1038/nature09534

Eid, J. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133 -138 (2009).

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, *8*(3), 186-194.

Ewing, B., Hillier, L., Wendl, M C, & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, *8*(3), 175-185.

FASTX - toolkit. (n.d.). Retrieved from http://hannonlab.cshl.edu/fastx_toolkit/

Fatemi, S. H., King, D. P., Reutiman, T. J., Folsom, T. D., Laurence, J. A., Lee, S., Fan, Y.-T., et al. (2008). PDE4B polymorphisms and decreased PDE4B expression are associated with schizophrenia. *Schizophrenia Research*, *101*(1-3), 36-49. doi:10.1016/j.schres.2008.01.029

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews. Genetics*, *7*(2), 85-97. doi:10.1038/nrg1767

Flanagan, S. E., Patch, A.-M., & Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers*, *14*(4), 533-537. doi:10.1089/gtmb.2010.0036

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, *6*(11 Suppl), S6-S12. doi:10.1038/nmeth.1376

Frith, M. C., Wan, R., & Horton, P. (2010). Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Research*, *38*(7), e100. doi:10.1093/nar/gkq010

Galan, M., Guivier, E., Caraux, G., Charbonnel, N., & Cosson, J.-F. (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, *11*, 296. doi:10.1186/1471-2164-11-296

George Priya Doss, C., Sudandiradoss, C., Rajasekaran, R., Choudhury, P., Sinha, P., Hota, P., Batra, U. P., et al. (2008). Applications of computational algorithm tools to identify functional SNPs. *Functional & Integrative Genomics*, *8*(4), 309-316. doi:10.1007/s10142-008-0086-7

Gnerre, S., Lander, E. S., Lindblad-Toh, K., & Jaffe, D. B. (2009). Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biology*, *10*(8), R88. doi:10.1186/gb-2009-10-8-r88

Gomez-Alvarez, V., Teal, T. K., & Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal*, *3*(11), 1314-1317. doi:10.1038/ismej.2009.72

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., & Katayama, T. (2010). BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics (Oxford, England)*, *26*(20), 2617-2619. doi:10.1093/bioinformatics/btq475

Goya, R., Sun, M. G. F., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., Senz, J., et al. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics (Oxford, England)*, *26*(6), 730-736. doi:10.1093/bioinformatics/btq040

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, *320*(5872), 106-109. doi:10.1126/science.1150427

Hashimoto, T., de Hoon, M. J. L., Grimmond, S. M., Daub, C. O., Hayashizaki, Y., & Faulkner, G. J. (2009). Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics (Oxford, England)*, *25*(19), 2613-2614. doi:10.1093/bioinformatics/btp438

Hernandez, D., François, P., Farinelli, L., Osterås, M., & Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*, *18*(5), 802-809. doi:10.1101/gr.072033.107

Holland, R. C. G., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., Foisy, S., et al. (2008). BioJava: an open-source framework for bioinformatics. *Bioinformatics (Oxford, England)*, *24*(18), 2096-2097. doi:10.1093/bioinformatics/btn397

1000 Genomes Project, http://www.1000genomes.org/

Hormozdiari, F., Alkan, C., Eichler, E. E., & Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, *19*(7), 1270-1278. doi:10.1101/gr.088633.108

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., et al. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, *26*(12), i350-357. doi:10.1093/bioinformatics/btq216

Husemann, P., & Stoye, J. (2010). Phylogenetic comparative assembly. *Algorithms for Molecular Biology: AMB*, *5*, 3. doi:10.1186/1748-7188-5-3

Imelfort, M., Duran, C., Batley, J., & Edwards, D. (2009). Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal*, *7*(4), 312-317. doi:10.1111/j.1467-7652.2009.00406.x

Isakov O, Modai S, Shomron N. Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics*. 2011 Aug 1;27(15):2027-30.

Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., et al. (2007). Extending assembly of short DNA sequences to handle error. *Bioinformatics (Oxford, England)*, *23*(21), 2942-2944. doi:10.1093/bioinformatics/btm451

Karchin, R. (2009). Next generation tools for the annotation of human SNPs. *Briefings in Bioinformatics*, *10*(1), 35-52. doi:10.1093/bib/bbn047

Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, *11*(11), R116. doi:10.1186/gb-2010-11-11-r116

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656-664. doi:10.1101/gr.229202. Article published online before March 2002

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524-536. doi:10.1002/bies.200900181

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England)*, *25*(17), 2283-2285. doi:10.1093/bioinformatics/btp373

Koboldt, D. C., Ding, L., Mardis, E. R., & Wilson, R. K. (2010). Challenges of sequencing human genomes. *Briefings in Bioinformatics*, *11*(5), 484 -498. doi:10.1093/bib/bbq016

Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Zhengdong, Snyder, M., et al. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, *10*(2), R23. doi:10.1186/gb-2009-10-2-r23

Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S., & Robinson, P. N. (2010). Microindel detection in short-read sequence data. *Bioinformatics (Oxford, England)*, *26*(6), 722-729. doi:10.1093/bioinformatics/btq027

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2009). TagDust--a program to eliminate artifacts from next generation sequencing data. *Bioinformatics (Oxford, England)*, *25*(21), 2839-2840. doi:10.1093/bioinformatics/btp527

Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2011). SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics (Oxford, England)*, *27*(1), 130-131. doi:10.1093/bioinformatics/btq614

Lee, P. H., & Shatkay, H. (2008). F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Research*, *36*(Database issue), D820-824. doi:10.1093/nar/gkm904

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al., DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008 Nov 6;456(7218):66-72.

Li, Biao, Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., et al. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics (Oxford, England)*, *25*(21), 2744-2750. doi:10.1093/bioinformatics/btp528

Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics*, *27*(8), 1157 -1158. doi:10.1093/bioinformatics/btr076

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473-483. doi:10.1093/bib/bbq015

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851-1858. doi:10.1101/gr.078212.108

Li, K., & Stockwell, T. (2010). VariantClassifier: A hierarchical variant classifier for annotated genomes. *BMC Research Notes*, *3*(1), 191. doi:10.1186/1756-0500-3-191

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature*, *463*(7279), 311-317. doi:10.1038/nature08696

Li, R., Li, Y., Fang, X., Yang, H., Wang, Jian, Kristiansen, K., & Wang, Jun. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124-1132. doi:10.1101/gr.088013.108

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., & Wang, Jun. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)*, *25*(15), 1966-1967. doi:10.1093/bioinformatics/btp336

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265-272. doi:10.1101/gr.097261.109

Lin, H., Zhang, Zefeng, Zhang, M. Q., Ma, B., & Li, M. (2008). ZOOM! Zillions of oligos mapped. *Bioinformatics (Oxford, England)*, *24*(21), 2431-2437. doi:10.1093/bioinformatics/btn416

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics: TIG*, *24*(3), 133-141. doi:10.1016/j.tig.2007.12.007

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., & Chen, Z. (2005a). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380. doi:10.1038/nature03959

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., & Chen, Z. (2005b). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380. doi:10.1038/nature03959

Martínez-Alcántara, A., Ballesteros, E., Feng, C., Rojas, M., Koshinsky, H., Fofanov, V. Y., Havlak, P., et al. (2009). PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics (Oxford, England)*, *25*(18), 2438-2439. doi:10.1093/bioinformatics/btp429

McKenna, A., Hanna, Matthew, Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303. doi:10.1101/gr.107524.110

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527-1541. doi:10.1101/gr.091868.109

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527-1541. doi:10.1101/gr.091868.109

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069 -2070. doi:10.1093/bioinformatics/btq330

Meader, S., Hillier, L. W., Locke, D., Ponting, C. P., & Lunter, G. (2010). Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Research*, *20*(5), 675-684. doi:10.1101/gr.096966.109

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, *6*(11 Suppl), S13-20. doi:10.1038/nmeth.1374

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31-46. doi:10.1038/nrg2626

Mi, H., Guo, N., Kejariwal, A., & Thomas, P. D. (2007). PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research*, *35*(Database issue), D247-252. doi:10.1093/nar/gkl869

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327. doi:10.1016/j.ygeno.2010.03.001

Mills, Ryan E, Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W Stephen, & Devine, Scott E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, *16*(9), 1182-1190. doi:10.1101/gr.4565806

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621-628. doi:10.1038/nmeth.1226

Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131-R136. doi:10.1093/hmg/ddq400

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *2*(2), 275-290.

Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*, *7*, 61-80. doi:10.1146/annurev.genom.7.080505.115630

Nothnagel, M., Herrmann, A., Wolf, A., Schreiber, S., Platzer, M., Siebert, R., Krawczak, M., et al. (2011). Technology-specific error signatures in the 1000 Genomes Project data. *Human Genetics*. doi:10.1007/s00439-011-0971-3

Novoalign. (n.d.).. Retrieved from http://www.novocraft.com/main/page.php?s=novoalign

Paşaniuc, B., Zaitlen, N., & Halperin, E. (2011). Accurate Estimation of Expression Levels of Homologous Genes in RNA-seq Experiments. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *18*(3), 459-468. doi:10.1089/cmb.2010.0259

Paszkiewicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, *11*(5), 457-472. doi:10.1093/bib/bbq020

PICARD. (n.d.).. Retrieved from http://picard.sourceforge.net/index.shtml

Pop, M., Phillippy, A., Delcher, A. L., & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in Bioinformatics*, *5*(3), 237-248.

Qi, J., Zhao, F., Buboltz, A., & Schuster, S. C. (2010). inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics (Oxford, England)*, *26*(1), 127-129. doi:10.1093/bioinformatics/btp615

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, *30*(17), 3894-3900.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227-240. doi:10.1093/hmg/ddq416

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, *27*(6), 863-864. doi:10.1093/bioinformatics/btr026

Schmieder, R., Lim, Y. W., Rohwer, F., & Edwards, R. (2010). TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, *11*, 341. doi:10.1186/1471-2105-11-341

Schwartz, S., Oren, R., & Ast, G. (2011). Detection and removal of biases in the analysis of next-generation sequencing reads. *PloS One*, *6*(1), e16685. doi:10.1371/journal.pone.0016685

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117-1123. doi:10.1101/gr.089532.108

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, *12*(10), 1611-1618. doi:10.1101/gr.361602

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation*, *32*(4), 358-368. doi:10.1002/humu.21445

Trapnell, C., & Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nature Biotechnology*, *27*(5), 455-457. doi:10.1038/nbt0509-455

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., & Pop, M. (2011). Next Generation Sequence Assembly with AMOS. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis... [et Al*, *Chapter 11*, Unit11.8. doi:10.1002/0471250953.bi1108s33

Van Deerlin, V. M., Leverenz, J. B., Bekris, L. M., Bird, T. D., Yuan, W., Elman, L. B., Clay, D., et al. (2008). TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *Lancet Neurology*, *7*(5), 409-416. doi:10.1016/S1474-4422(08)70071-1

Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res 38*, e164 (2010).

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263-270. doi:10.1002/humu.22

Ye, Kai, Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865-2871. doi:10.1093/bioinformatics/btp394

Yoon, S., Xuan, Z., Makarov, V., Ye, Kenny, & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, *19*(9), 1586 -1592. doi:10.1101/gr.092981.109

Yuan, H.-Y., Chiou, J.-J., Tseng, W.-H., Liu, C.-H., Liu, C.-K., Lin, Y.-J., Wang, H.-H., et al. (2006). FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Research*, *34*(Web Server issue), W635-641. doi:10.1093/nar/gkl236

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821-829. doi:10.1101/gr.074492.107

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., & Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS One*, *6*(3), e17915. doi:10.1371/journal.pone.0017915

Zhao, X., Palmer, L. E., Bolanos, R., Mircean, C., Fasulo, D., & Wittenberg, G. M. (2010). EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *17*(11), 1549-1560. doi:10.1089/cmb.2010.0127