Chapter 9

Computational Analysis of High Throughput Sequencing Data

Steve Hoffmann

Abstract

The advent of High Throughput Sequencing (HTS) methods opens new opportunities for the analysis of genomes and transcriptomes. While the sequencing of a whole mammalian genome took several years at the turn of this century, today it is only a matter of weeks. The race towards the *thousand-dollar genome* is fueled by the – ethically challenging – idea of personalized genomic medicine. However, these methods allow new and interesting insights in many aspects such as the discovery of novel noncoding RNA classes, structural variants, or alternative splice sites to name a few. Meanwhile, several methods for HTS have been introduced to the markets. Here, an overview on the technologies and the bioinformatics analysis of HTS data is given.

Key words: High throughput sequencing, Short reads, Mapping, Assembly, SNP detection, 454, Illumina, Helicos, SOLiD

1. Introduction

When turning to High Throughput Sequencing (HTS), often also referred to as Next Generation Sequencing (NGS), one quickly realizes that the time of spreadsheet-bioinformatics is coming to an end. A single 3-day sequencer run confronts the scientists with terabytes of data: images, qualities, statistics, summaries, sequences, maps, and assemblies to name few. In the light of this quick and massive avalanche of data, deciding on data storage policies alone appears to be a difficult task. The deletion of any file will inevitably limit the possibilities of reanalysis or require cumbersome reevaluations. Especially for precious biological samples it might be worthwhile to store the data. For example, the reanalysis of images with alternative base calling tools may yield important improvements. Because the prices for whole-genome sequencing

Bernd Mayer (ed.), *Bioinformatics for Omics Data: Methods and Protocols*, Methods in Molecular Biology, vol. 719, DOI 10.1007/978-1-61779-027-0_9, © Springer Science+Business Media, LLC 2011

of mammalian genomes have dropped significantly below 20,000 US$ (1) and the time needed for sequencing does not exceed a couple of weeks, yet unknown amounts of data will accumulate in laboratories all over the world. Hence, whatever policies are adopted and irrespective of whether the sequencing itself is outsourced or not, a series of smaller HTS projects already requires a specialized infrastructure with arrays of hard disks, network architectures, or even tape storages. These issues may not be overstressed since they heavily affect everyday work with the HTS data. A few questions have to be answered carefully in each case. How to exchange data with the collaborators? How to set up pipelines for analyses? Is a version control of reanalyzed data necessary?

Picturing a tape storage library with loads of TB-cartridges in a molecular biology lab one immediately realizes that the analysis of HTS data is a problem of its own. Many standard algorithms and tools frequently used in genome informatics had or still have to be revised in order to contain the deluge of HTS data. Moreover, HTS offers the opportunity for new types of analysis such as transcription start site detection, ncRNA detection, or RNA structure probing requiring new algorithms and standards. This chapter intends to give an overview on the sequencing technologies as well as basic approaches to HTS data analysis. As the HTS methods are steadily improving and changing at a very fast pace, the focus will be on their basic principles rather than fugacious facts or specific pros and cons of single technologies. Despite their different error models, properties, and application areas, all HTS approaches gain their true beauty and fascination by their genuine combination of different techniques from various fields of science: molecular biology, chemistry, physics, material science, engineering, and computer science.

2. Materials

Current high throughput DNA sequencing methods may be subdivided in two major approaches: *sequencing by ligation* and *sequencing by synthesis*. In sequencing by synthesis a single-stranded, primer-probed DNA template is sequentially duplicated by a polymerase. During duplication, chemically modified nucleotides added to the newly synthesized strand allow the detection of each base of the template. On the other hand, sequencing by ligation does not use polymerases, but employs specifically binding primer probes. ABI’s SOLiD sequencing platform (2) puts this idea into practice. However, in the near future a direct read out of sequences using the physicochemical properties of nucleotides (such as charges) may become the state of the art.

In principle, all novel sequencing methods achieve high throughput by immobilizing large amounts DNA or cDNA fragments locally. Regardless whether the immobilization takes place on beads or plane surfaces – the idea is to spatially separate fragments sufficiently to perform the sequencing for all fragments simultaneously.

2.1. Sequencing Platforms

2.1.1. 454 Pyrosequencing

The 454 pyrosequencing system was the first HTS system introduced to the markets (3). The key idea of pyrosequencing is to trigger detectable chemiluminescent reactions during the sequencing step. To prepare a DNA library for the 454 platform, the double-stranded source material needs to be sheared into fragments of several hundred nucleotides. In a second step two different *linkers*, i.e. specific DNA sequences of known length and composition are ligated to the double-stranded shreds. A 5′-biotin tag attached to one of the linkers allows the immobilization of the fragments on DNA capture beads. An excess of beads ensures that the expected number of bound fragments per bead does not exceed one.

To generate detectable light signals during the synthesis step, an emulsion polymerase chain reaction (emPCR) clonally amplifies the fragment. During this process millions of copies are directly attached to the bead. Picotiter plates, i.e., plates with wells just large enough to contain a single bead, are used to trap and locally immobilize the beads.

The nucleotides adenine (A), cytosine (C), thymine (T), and guanine (G) are sequentially washed over the plates in four recurring cycles. Polymerases along with other enzymes generate light signals when the nucleotides are incorporated into the DNA strand. More precisely, the enzyme luciferase is the major component of the chemiluminescent reaction (Fig. 1). After each cycle, a washing step is needed to remove the excess of nucleotides from the plate.

The description of the sequencing by synthesis step already reveals an important problem: sequencing of homopolymers. A stretch of two or more identical nucleotides in the DNA template will generate multiple subsequent *nonsynchronous* chemiluminescent reactions during a single cycle. Hence, the intensity of the light signal is the only way to determine the length of a homopolymer – making the use of rather complicated signal processing steps necessary. Additionally, despite the washing steps, nucleotides accumulate within the wells causing an increase of the background signal as the sequencing proceeds. The major advantage of the 454 method in comparison with other technologies is the longer read length making the sequences especially useful for assemblies (Table 1).

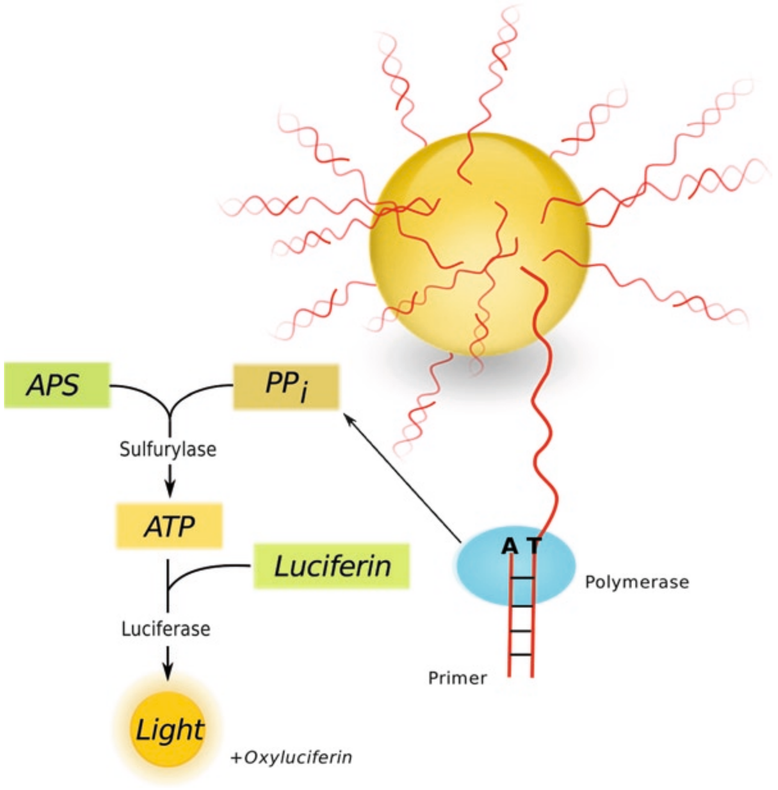


Fig. 1. 454 Pyrosequencing. After clonal amplification of DNA fragments at DNA capture beads, beads are trapped in small wells of a picotiter Plate. During the sequencing-by-synthesis the four nucleotides are washed in recurring cycles over the plate. Incorporation of nucleotides by the polymerase results in the production of ATP. In turn, the energy rich ATP triggers a luciferase reaction generating a light signal that is recorded by a CCD camera. 454 Sequencing © Roche Diagnostics. All rights reserved.

2.1.2. Illumina (Solexa)

With about 100 nucleotides, Illumina fragments are significantly shorter compared to those from 454 (4). However, due to its degree of parallelization this method allows a higher throughput with more than 20 gigabases per day. In contrast to 454, fragments ligated to two different adapters are immobilized on translucent plates, flow cells, densely coated with oligonucleotides complementary to the adapters. The key idea of this procedure is to clonally amplify the fragments in a circumscribed region of the flow cell. The complementary adapters on the plate act as PCR primers to generate clusters of millions of identical copies using a process called *bridge-amplification*. The prerequisite for a successful generation and detection of such clusters is that the initially binding fragments are well separated from each other (Fig. 2). In fact, separability and purity of clusters is one of the major challenges in the signal detection step.

The sequencing by synthesis step involves reversible dye-terminators. The polymerase incorporates differently labeled

Table 1
Comparison of high-throughput sequencing methods

Technology (platform)	Read length	bp per run	Accuracy[a]	Run time	Remarks[b]
Roche 454 (GS FLX w/ Titanium chemistry)	~400 bp	~400 MB	99.5%	7 h	Supports mate pair sequencing; indels in homopolymer stretches
Illumina/Solexa (HiSeq 2000)	36–100 bp	~200 GB (paired end/mate pair mode)	>98%	8 days	Supports paired end and mate pair sequencing
SOLiD (SOLiD 3.0)	~50 bp	~20–30 GB (paired-end)	99.94%	14 days	Supports mate pair sequencing
Helicos	22–55bp (35 bp)	~20–30 GB	>95%	8 days	Supports paired end sequencing

[a] According to the manufacturer

[b] Please note that mate pair sequencing refers to parallel sequencing of two DNA fragments with a known approximate physical distance. Paired end sequencing refers to the sequencing of a single DNA fragment from both sides

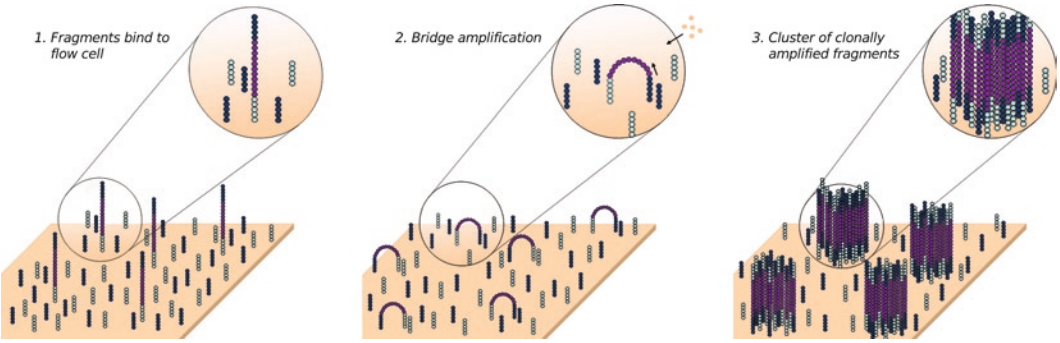

1. Fragments bind to flow cell
2. Bridge amplification
3. Cluster of clonally amplified fragments

Fig. 2. Illumina/Solexa cluster generation. Adapter-tagged fragments are immobilized on glass cover slips densely coated with reverse complementary adapters (1). Subsequently, fragments are amplified using a bridge amplification step (2), resulting in locally separated clusters of clonally amplified fragments (3). Separability and purity of clusters is a key prerequisite for the Illumina technology. During the sequencing step (using reversible terminator dye chemistry) the clusters generate sufficiently strong light signals to be detected by a camera. ©2010, Illumina, Inc. All rights reserved.

nucleotides that bring the sequencing to a sudden stop. Laser excitation of labels indicates the type of the incorporated nucleotide and cleaves the terminator activity. Subsequently, the next base can be called.

The Illumina approach intends to call bases synchronously, i.e. all light signals generated in the *n*th cycle belong to the *n*th nucleotide of all fragments in the flow cell. In practice, however, this does not always work. A failure to remove the terminator or the integration of nucleotides without a terminator activity, e.g., will result in a phase shift. This *phasing* increases noise and complicates the signal detection.

2.1.3. Helicos

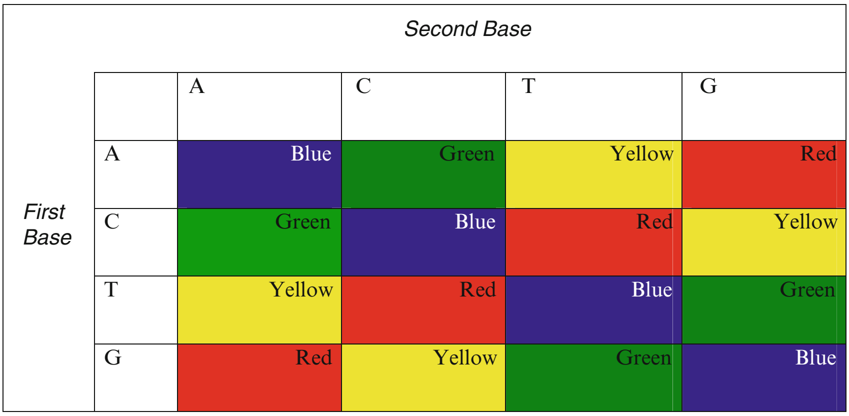
Helicos styles its single molecule sequencing figuratively as "DNA microscopy". Although the company's own interpretation is at least semantically questionable their system provides a solution that does not require amplification, which may introduce biases preceding the detection step (1, 5). Instead of specific adapter sequences, poly-A anchors are covalently attached to single fragments. Subsequently, glass cover slips covered with poly-T oligomers are used to capture the library fragments. During sequencing, the plate is incubated with only one Cy5-labled nucleotide at a time and a light signal is generated upon laser excitation. Because of the low signal strength, good background discrimination, and high-resolution image detection is required. And indeed, the error rates of Helicos are reported to be significantly higher in comparison with other HTS technologies (1), demanding more sensitive software in the downstream analysis.

2.1.4. SOLiD

Applied Biosystems' SOLiD technology employs a "sequencing by ligation" approach. Sheared fragments coupled to a universal primer (P1) are attached to beads and clonally amplified in an emPCR reaction. In contrast, beads are not captured in wells but covalently bound to glass plates.

In the first sequencing round, a universal primer complementary to the 3′-end of P1 is used. Subsequently, fluorescently labeled 5 mer-probes are washed over the plate. Each label represents a set of four dinucleotide combinations specific only for the first and second position of the probe, while the other three nucleotides are random. Probes complementarily binding to the template sequence are ligated to the primer. After signal detection, the labels are simultaneously cleaved and a second ligation reaction elongates the extension product. At the end of the first ligation round with multiple ligation reactions the color information is obtained for pairs of nucleotides with a lag of 4nt, i.e. the 1st and 2nd, the 6th and 7th, the 11th and the 12th, and so on. Note that the color information from the first round is not sufficient to decode the bases. To decode a single base, the antecedent nucleotide (not only the color information) has to be known (Table 2). Hence, additional ligation rounds are necessary to translate the colors and to interrogate the remaining nucleotides. Therefore, the extension product is removed and a second primer complementary to the 3′-end of P1 with an offset of −1 binds to the template. In this ligation round, the first two interrogated nucleotides are the last base of P1 and the first base of the template. In total, it requires five ligation rounds to translate the whole template from the color space to sequence space. Due to the fact that each base is interrogated twice, a higher accuracy is expected. However, in case of a sequencing error, e.g., by an unspecific binding of a probe, correct translation to nucleotide space will fail for *all subsequent* bases. As a result, there are two basic philosophies for analyzing SOLiD data.

Table 2
The SOLiD color space coding table. Note that for each dinucleotide, the reverse, the complement, and the reverse complement always have the same color



		Second Base			
		A	C	T	G
First Base	A	Blue	Green	Yellow	Red
	C	Green	Blue	Red	Yellow
	T	Yellow	Red	Blue	Green
	G	Red	Yellow	Green	Blue

While some tools skip the translation and process the SOLiD data directly in color space (e.g., mapping the data to a color space version of the reference genome), others actually translate it to sequence space. Note, that the SOLiD color space has some interesting properties: for an unknown color *x*, the color signal *x-blue-blue-blue-blue* may translate to a poly-A, poly-C, poly-T or poly-G (Table 2). It was specifically designed so that for each dinucleotide the reverse, the complement and the reverse complement have the same color.

Another sequencing by ligation approach called combinatorial probe anchor ligation (cPAL) is used by Complete Genomics. This company currently offers sequencing services only. Complete Genomics was the first to announce the sequencing of a whole human genome with a coverage of 78-fold for less than 5,000 US$ (6).

2.1.5. Other Approaches to Sequencing

Nanopore sequencing is an alternative approach to HTS. The initial concept was based on threading a DNA molecule through a staphylococcal alpha-hemolysine. During passage, the DNA would alter a current applied to the molecule in a way that is specific for its nucleotide composition. Newer publications investigate the usage of covalently incorporated adapters to detect single nucleotides cleaved from DNA fragments upon exonuclease treatment by means of current modulations (7). Under optimal conditions this approach was reported to achieve acceptable error rates.

A different technology, currently developed by Pacific Biosciences, locally immobilizes the polymerases on a glass surface. Each polymerase is surrounded by a zero-mode waveguide (ZMW) that confines light in volumes (typically <10e-21 L) smaller than its wavelength (8). Using ZMWs light signals generated upon the integration of labeled nucleotides can be distinguished from the background. While the reads are reported to be significantly longer, error rates close to 10% make further improvements necessary before the technology is ready for the markets.

2.1.6. Paired-Ends

The sequencing of *paired-ends* helps to overcome the problem of ambiguity caused by short read lengths and improves the assembly results. By using DNA circularization during the library construction, shorter fragments can be obtained from both ends of a larger genomic fragment with a typical length of 2–5 kb. The pair of small DNA fragments is then sequenced in the same cluster or well. The prior knowledge of the approximate distance between the two fragments impedes the misplacement of both reads during alignment and helps to assemble repetitive regions. Furthermore, it is useful to detect structural variations such as copy number variations.

Note that there is some confusion about the terminology. In the Illumina protocol, the process described above is termed *mate-pair sequencing*, while the 454-protocol refers to it as paired-end sequencing. Illumina's paired-end sequencing does not require DNA circularization: a little modification of the single-end protocol allows the sequencing of a DNA fragment of about 200–500 bp from both ends simultaneously.

3. Methods

As mentioned above, the sheer amounts of data demand efficient and fast algorithms for the bioinformatics analysis. In principle, the analysis already begins with the base calling, i.e., the decoding of electromagnetic signals to nucleotide sequences and associated quality values. Meanwhile, several independent groups have proposed alternative base calling approaches with optimized results (9–11). These concepts vary from Bayesian base calling methods to approaches using support vector machines. Here, we only focus on some standard input and output formats as well as the two most common forms of sequence data analysis: mapping and assembly.

3.1. Basic Data Analysis and Evaluation

For further downstream analysis, base callers typically report the sequences together with their quality values. While the Illumina platform reports the sequences along with the quality values in a multiple FASTQ format, reads and quality values from the 454 GS FLX are often exported from the binary SFF format (Standard Flowgram Format) to two separate multiple FASTA files. However, the Genome Sequencer Data Analysis suite that comes with the sequencer offers several tools to process the SFF files directly.

The multiple FASTA format contains multiple sequences that are preceded by a header line. The header starts either with the symbol ">" or ";". It holds the basic identifiers and sequence information. All subsequent lines hold the sequence itself. For 454 reads, the header line starts with a unique identifier followed by the read length, the coordinates of the bead and the date of the sequencing run.

```
>FW8YT1Q01B9VMY length=380 xy=0815_2008 region=1 run=R_XXXX
TGATCTTACTATCATTGCAAAGCCACTTAAAGAC CACACACTACGTCACTGGAAAAGAGT
TCAATAGAGGCCTCCTACGAGTAACACCCTTACAC TTCTGCTACAGAAACTACACCTTTT
```

Quality values for 454 reads are given in a separate file. The assignment of reads and quality values is possible via header line.

```
>FW8YT1Q01B9VMY length=380 xy=0815_2008 region=1 run=R
```

```
37 39 39 39 39 28 28 31 33 37 37 35 35 35 35 35 35 37 27 31
31 37 37 37 37 37 37 38 37 37 37 37 39 39 39 39 39 39 39 39
39 37 37 33 32 32 30 32 32 32 19 19 15 15 15 15 23 16 30 29 …
```

The FASTAQ format is a slight modification of the FASTA format.

````
@solexaY:7:1:6:1669/1
GCCAGGNTCCCCACGAACGTGCGGTGCGTGACGGGC
+solexaY:7:1:6:1669/1
``a`aYDZaa``aa_Z_`a[`````a`_P][[\_\V
````

The FASTQ header begins with an "@" symbol. Again, all following lines hold the sequence itself. For Illumina reads, the header informs about the name of the instrument (solexaY), the flowcell lane (7), the tile number within the flow cell (1), its *x*- and *y*-coordinates (6:1,669) and has a flag indicating whether the read is single-end (/1) or belongs to a mate-pair or paired-end run (/2). The "+" sign followed by the same sequence identifier indicates the beginning of the quality value string. Note that the qualities are given in ASCII.

The quality values give an estimate on the accuracy of the base calling. Nowadays, most sequencing platforms report a Phred quality score. The score, originally developed in the context of the Human Genome project, is given by

$$Q = -10 \cdot \log_{10} p$$

where, p is the probability that the reported base is *incorrect*. Illumina initially decided to deviate from this scoring and instead used the formula

$$Q_{Solexa} = -10 \cdot \log_{10} \frac{p}{1-p}.$$

While the Illumina quality score Q_{Solexa} is asymptotically identical to Q for low error probabilities, it is typically smaller for higher error probabilities. Since the Illumina quality scores can become negative, a conversion to real phred scores using

$$Q = 10 \cdot \log_{10} \left(1 + 10^{Q_{Solexa}/10}\right)$$

may be necessary. While high Illumina quality scores have been reported to overestimate the base calling accuracy, low scores underestimated the base calling accuracy (10, 12). Since version 1.3, the proprietary Solexa pipeline uses Phred scores. It is important to note that also the encoding of the quality string has been subject to changes. The new pipeline encodes the phred qualities from 0 to 62 in a non-standard way using the ASCII characters 64 to 126. Due to the fact that Phred scores from 0 to 93 are normally encoded using ASCII characters 33–126, a conversion might be necessary.

3.2. Mapping

In genome informatics the *mapping* describes the process of generating a (mostly heuristic) alignment of query sequences to reference genomes. It is the basis for qualitative as well as quantitative analysis. To map HTS sequences, the algorithms have to address three different problems at once. In addition to the tremendous amounts sequences, the methods have to deal with a lower data quality and shorter read lengths. Particularly for short (erroneous) reads it is often not possible to decide for its original position in the reference genome since the reads may align equally well to several genomic locations. Sequencing of repetitive regions complicates this problem even more. The methods presented here apply different mapping policies to tackle those problems.

To address the problem of the huge amount of data, most of the short read alignment programs use index structures either for the reads or the reference.

3.2.1. Mapping with Hash Tables

Heng Li et al. developed one of the first read mappers, MAQ, for Illumina sequences based on hash tables. Although the tool is no longer supported, a look into the core of this approach reveals some basic principles and policies of short read mapping. The focus of MAQ is to incorporate quality values to facilitate and improve read mapping (13). By default, MAQ indexes only the first 28 bp of the reads (the seed) in six different hash tables ensuring that all reads with at most two mismatches may be found in the genome. Equivalently for a seed of 8 bp the hash tables are built from three pairs of complementary templates, 11110000, 00001111, 11000011, 00111100, 11001100, and 00110011, where a 1 indicates a base that is included in the hash key generation. After this indexing step, MAQ proceeds by scanning the reference sequence once for each pair of complementary template. Each time a seed hit is encountered MAQ attempts to extend the hit beyond the seed and scores it according to the quality values. It has been reported earlier that the use of quality values during the read alignment can improve the mapping results substantially (14). By default, MAQ reports all hits with up to two mismatches– but its algorithm is able to find only 57% of the reads with three mismatches. Hits with insertions and deletions (indels) are not reported. Furthermore, for reads with multiple equally scoring best hits only one hit is reported.

3.2.2. Mapping with Suffix Arrays and the Burrows–Wheeler Transform

A second approach to short read alignment uses the Burrows–Wheeler Transform (BWT). In brief, the BWT is a sorted cyclic permutation of some text *T*, e.g., a reference genome. Its main advantage is that the BWT of *T* contains stretches of repetitive symbols – making the compression of the *T* more effective. The backward search algorithm (15) on a compressed BWT simulates a fast traversal of a prefix tree for *T* – without explicitly representing the tree in the memory. It only requires two arrays to efficiently access the compressed BWT, which is the key to the speed and the low memory footprint of read aligners such as BWA (16), Bowtie (17) and SOAP2 (18). Because the backward search only finds exact matches, additional algorithms for inexact searches had to be devised. BWA, for example, solves this problem by enumerating alternative nucleotides to find mismatches, insertions and deletions, while SOAP2 employs a split alignment strategy. Here, the read is split into two parts, to allow a single mismatch, insertion or deletion. The mismatch can exist in at most one of the two fragments at the same time. Likewise, the read is split into three fragments to allow two mismatches and so forth. Other tools such as Bowtie do not allow short read alignments with gaps.

BWT-based read mappers are the speed champions of short aligners – with an exceptionally low memory footprint. However, for all of the tools described above, the user has to carefully choose a threshold for a maximum number of acceptable errors. For error thresholds >2 mismatches, insertions, or deletions, the speed decreases significantly. While these thresholds seem to be sufficient for mapping of genomic DNA, mapping of transcriptome data or data that contains contaminations (e.g., linkers) may be more difficult. In contrast the tool *segemehl* (19), based on enhanced suffix arrays, aims to find a best local alignment with increased sensitivity. In a first step, exact matches of all substrings of a read and the reference genome are computed. The exact substring matches are then modified by a limited number of mismatches, insertions, and deletions. The set of exact and inexact substring matches is subsequently evaluated using a fast accurate alignment method. While the program shows good recall rates of 80% for high error rates of around 10%, it has a significantly larger memory footprint in comparison with the BWT and hashing methods. A practical example is given at the end of this chapter (see Note 1).

The selection of an appropriate mapping method depends on various criteria (see Note 2). Due to the different indexing techniques some short read aligners are limited to certain read lengths. These tools may not be used if long reads or reads of different sizes need to be aligned. Furthermore, for speed reasons some aligners report only one hit per read – regardless of whether multiple equally good hits could be obtained. This may be a problem if repetitive regions are sequenced. The user has to assess carefully which degree of sensitivity is needed. A method that discards reads with multiple hits (sometimes a random hit is reported) or high error rates may be suitable for SNP detection, while mapping of transcriptome (RNAseq) data may require a higher sensitivity. A selection of mapping tools is given at the end of the chapter (see Note 3).

3.2.3. SAM/BAM Mapping Output Format

Because most of the read mapping tools have their own output formats, a standard output format for short read aligners was developed in the context of the 1000 Genomes Project (http://www.1000genomes.org) (20). The Sequence Alignment/Map (SAM) is a human readable tab-delimited format. A binary equivalent (BAM) is intended to facilitate the parsing with computer programs. The SAM format contains a header and an alignment section. A typical header section starts with a mandatory header line (@HD) that holds the file format version (VN:1.0). Sequence dictionaries (@SQ) hold the names (SN:chr20) and the lengths (LN:62435964) of the reference sequences to which the reads in the alignment section are mapped to.

```
@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
```

To identify different biological samples, the SAM file may also hold one or more read groups (@RG). Each group has to have a unique identifier. (ID:L1, ID:L2) and the name of the sample (SM:NA12991) from which the reads were obtained. Additionally, the platform unit (PU:SC_1_10), e.g. the lane of the Illumina flowcell, or the library name (LB:SC_1) can be given.

The alignment section holds all read alignments. A typical alignment line like

```
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195
AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:
<9/,&,22;;<<<
NM:i:1 RG:Z:L1
```

has the format

```
<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR>
<MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>
[<TAG>:<VTYPE>:<VALUE> [...]]
```

where, the field <QNAME> holds the name of the query sequence (or sequence pair), <RNAME> the name of the reference sequence and <POS> the position in the reference sequence. The mapping quality value is store in the <MAPQ> field. The extended <CIGAR> string is a representation of the read alignment. It is comprised of a series of operation lengths plus the operation types. While the conventional <CIGAR> format only allows for three types of operations (M for match or mismatch, I for insertion, and D for deletion), the extended <CIGAR> also identifies clipping, padding, and splicing operations. Finally, the fields <SEQ> and <QUAL> hold the read sequence and the corresponding quality values. A complete description of the SAM and BAM formats can be obtained from http://samtools.sourceforge.net.

This output format has quickly advanced to a standard and many read mappers offer a SAM/BAM compatible output.

3.3. Assembly of Short Read Data

The advent of HTS has raised hopes to quickly and inexpensively perform de novo assemblies of large genomes. However, shorter read lengths and higher error rates have spoiled all too optimistic expectations.

One of the first tools for short read assembly, SSAKE (21), employs a greedy method to build larger contigs from short Illumina reads. After building a hash table holding unique read sequences, a prefix tree indexes all such sequences. The assembly starts with the most abundant unique sequence. All 3′ *most k-mers*, i.e. substrings of length k at most m characters from the 3′ end of the sequence apart, are looked up in the prefix tree. All hits are used to build the first consensus contig. This consensus is then used to find the next set of *k-mers*. This process is iterated until all possibilities of the contig extension are exhausted. While such a simple method works well for small genomes, the assembly of larger genomes from short reads is rather cumbersome. Zerbino et al. (22) used a more complicated de Bruijn-Graph approach in their program called "Velvet". Each node in the graph represents a series of overlapping *k-mers*, such that two adjacent *k*-mers overlap by $k-1$ characters. Two nodes A and B are connected by a directed edge if the last k-mer of the node A overlaps with the first of B. Hence, not only the reads, but a whole series of overlapping reads can be modeled as a path through the graph. The authors report that Velvet is capable of assembling bacterial genomes with N50 contig lengths of up to 50 kb. In simulations with 5-Mb regions of large mammalian genomes, contigs were ~3 kb long. If available, both applications make use of mate-pair information.

In the future, alternative approaches that combine the high coverage provided by short read sequencers such as Illumina with longer 454 and Sanger reads may prove to be more effective when it comes to the assembly of larger vertebrate genomes. However, another tool that employs de Bruijn-Graphs, SOAPdenovo, was used to successfully assemble mammalian genomes from single-end and mate-pair Illumina sequences only (23, 24). A list of selected tools is given at the end of the chapter (see Note 4).

3.4. Other Applications

The most important goal of personalized genomics is the detection of variations such as single nucleotide polymorphisms (SNP). HTS for the first time offers the opportunity to detect previously unknown SNPs with minor allelic variants on a large scale. Therefore, some vendors of HTS platforms such as Illumina provide tools to call SNPs directly from the sequencing data.

Alternative methods for SNP calling are, for example, offered by MAQ (13) or SOAPsnp (25).

Primary prerequisite for the SNP calling is a high quality of the sequencing data. To assure this, typically all reads with multiple hits, reads with low overall qualities, and reads with more than one mismatch are discarded. The success of SNP calling in HTS data depends not only on the quality of the reads but also on the coverage.

After mapping the reads to a reference, the cross-section at sufficiently covered (>10) genomic positions is checked for polymorphisms. To do this many SNP callers employ Bayesian statistics. For example, SNPsoap assumes a set of ten different genotypes

$$T_i = H_m H_n = \{\mathrm{AA}, \mathrm{CC}, \mathrm{GG}, \mathrm{TT}, \mathrm{AC}, \mathrm{AG}, \mathrm{AT}, \mathrm{CG}, \mathrm{CT}, \mathrm{GT}\}$$

where, H_m and H_n denote the two haplotypes of the genotype T_i at some position i of the genome. To obtain an estimate on the conditional probability of the data one may calculate

$$P(D \mid T) = \prod_{k=1}^{l} \frac{P(d_k \mid H_m) + P(d_k \mid H_n)}{2}$$

where, l is the number of observed alleles in the cross-section and $p(d_k \mid \mathrm{H})$ is the probability of observing the allele d_k under the hypothesis H . The posterior probability for a genotype T_i given the HTS data D is then evaluated with

$$P(T_i \mid D) = \frac{P(T_i) P(D_i \mid T_j)}{\sum_{x=1}^{10} P(T_x) P(D \mid T_x)}$$

where the probability of a genotype $P(T)$ is usually calculated using prior knowledge. To reduce false-positive SNP calling SOAPsnp additionally considers quality values. A similar approach was chosen for the Helicos pipeline (5).

However, an important drawback of the approach sketched above is that the successful base calling depends on an equally successful coverage of both haplotypes and it may only be used for single individuals but not for pooled samples. It furthermore assumes that only two nucleotides segregate per site so that autosomal mutations may be missed.

An interesting alternative maximum-likelihood approach for analyzing pooled samples was published by Michael Lynch (26).

4. Notes

1. Meanwhile there are several tools available to map reads to a reference genome. The selection of the appropriate read mapper depends on several criteria such as accuracy or speed. Here, a mapping run with segemehl is given as an example. The program can be downloaded at http://www.bioinf.uni-leipzig.de/Software/segemehl. The program should compile on all LINUX systems with a C99-compatible C compiler and 2 GB of free memory. To generate an executable binary, type

   `tar -xvzf segemehl_0_0_*.tar.gz` and then `call make`

   to compile the program.

   To start a set of short reads and a reference genome is needed. Reads for *Arabidopsis. thaliana* may be obtained at http://www.bioinf.uni-leipzig.de/~steve/omics/arabidopsis.fna. The *A. thaliana* reference sequence is available at the website of the plant genome database (http://www.plantgdb.org/download/Download/xGDB/AtGDB/ATgenomeTAIR9.171). To map the reads just call

   ```
   ./segemehl.x -x ATgenomeTAIR9.171.idx -d
    ATgenomeTAIR9.171 -q arabidopsis.fa > ara-
    bidopsis.map
   ```

   With the same call, segemehl generates an index (ATgenomeTAIR9.171.idx) of the Arabidopsis genome. In this example most of the time is spent on the index construction. If the index was already built segemehl may be called with

   ```
   ./segemehl.x -i ATgenomeTAIR9.171.idx -d
    ATgenomeTAIR9.171 -q arabidopsis.fa > ara-
    bidopsis.map
   ```

   To increase the sensitivity the option –D 2 may be given. In case the program is run on a multi-core architecture –threads 4 will parallelize the task in four threads. The minimum required accuracy of the alignment may be changed using the –A parameter.

   The output file arabidopsis.map now contains the mapping information. Note that unlike other tools segemehl by default also reports reads that map to multiple sites. The mapping file contains a description of the fields in the header line. This file may be used to generate BED or other file formats to visualize the mapping data in genome browsers.

2. Although HTS methods are already well established some issues may still be a nuisance and a common source of error in data analysis. The error models of the technologies are very different. While mismatches are the major error type in

Illumina sequences, 454 sequences suffer from insertions and deletions – especially in homopolymers. As mentioned earlier, the rate of indels significantly increases along homopolymer stretches making it cumbersome to call SNPs in those regions. Some sequencers such as Helicos are reported to have much higher error rates. Moreover, also the reference sequences are not free of errors.

Due to the additional RT-PCR step RNAseq data is less accurate than genomic data. Therefore, especially small RNAs such as miRNA require additional sensitivity. Before mapping RNA sequences should be scanned for poly-A tails. Clipping of these tails facilitates the mapping process as e.g. A-rich genomic stretches may misguide the mapping procedure.

As pointed out earlier, short read lengths make the assembly of larger contigs very difficult. Hence, assemblies with HTS data have to be planned carefully. An approach that uses technologies that provide longer reads or a combination of different sequencing methods is more likely to succeed.

3. Selected mapping programs

   - `MAQ` (13) is one of the first tools for mapping of HTS reads with hash tables. The tool only considers mismatches and was developed for Illumina reads only (http://maq.sourceforge.net/maq-man.shtml).
   - `BWA` (16) is a very fast short read aligner based on the Burrows-Wheeler transform that also allows the detection of insertions and deletions. It is not limited to a specific platform or read length. The tool typically allows only a few errors per read (http://bio-bwa.sourceforge.net/bwa.shtml).
   - `Bowtie` (17) is also based on a Burrows-Wheeler transform. This tool currently does not support the detection of indels and works for Illumina reads only. Only a few errors per read are allowed (http://bowtie-bio.sourceforge.net/tutorial.shtml).
   - `SOAP2` (18) is an alternative to BWA (http://soap.genomics.org.cn/).
   - `segemehl` (19) is a sensitive read aligner with indel detection support. The program does not depend on fixed read lengths and is platform-independent. It allows mapping of sequences with higher error rates but has a large memory footprint (http://www.bioinf.uni-leipzig.de/Software/segemehl; http://ngslib.genome.tugraz.at/node/36).

4. Selected assembly programs

   - `SSAKE` (21) is a sequence assembly program for Illumina reads based on a greedy hashing method. The program is capable of assembling short genomes (http://www.bcgsc.ca/platform/bioinfo/software/ssake).

- `Velvet` (22) is based on a de Bruijn-Graph method (http://www.ebi.ac.uk/~zerbino/velvet).
- `SHARCGS` (27) is one of the first sequence assemblers for Illumina reads. It uses a similar strategy like SSAKE (http://sharcgs.molgen.mpg.de/).

Acknowledgments

Many thanks go to Maribel Hernandez Rosales, Dulce Palafox, Ishaan Gupta, Sven Findeis, Dominic Rose, and Jörg Hackermüller for fruitful discussions and proof-reading the manuscript. This publication is supported by LIFE-Leipzig Research Center for Civilization Diseases, Universitaet Leipzig. This project was funded by means of the European Social Fund and the Free State of Saxony.

References

1. Pushkarev, D., Neff, N. F., and Quake, S. R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**, 847–52.
2. Pandey, V., and Nutter, P. E. (2008) Next-generation genome sequencing: towards personalized medicine. Wiley, New York.
3. Margulies, M., Egholm, M., Altman, W. E. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–80.
4. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P. et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–9.
5. Harris, T. D. et al (2009) Single-molecule DNA sequencing of a viral genome. *Science* **302**, 106–9.
6. Drmanac, R., Sparks, A. B., Callow, M. J. et al. (2009) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81.
7. Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**, 265–70.
8. Eid, J., Fehr, A., Gray, J. et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–8.
9. Quinlan, A. R., Steward, D. A., Stromberg, M. P., and Marth, G. T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5**, 454–57.
10. Kircher, M., Stenzel, U., and Kelso, J. (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**, R83.
11. Erlich, Y., Mitra, P. P., de la Bastide, M., McCombie, W. R., and Hannon, G. J. (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* **5**, 679–82.
12. Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, e105.
13. Li, H., Ruan, J., and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851–8.
14. Smith, A. D., Xuan, Z., and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinform* **9**, 128.
15. Ferragina, P., and Manzini, G. (2000) Opportunistic data structures with applications. *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 390–8.
16. Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–60.
17. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.
18. Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009) SOAP2:

an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–7.

19. Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermuller, J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol* **5**, e1000502.
20. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9.
21. Warren, R. L., Sutton, G. G., Jones, S. J. M., and Holt, R. A. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500–1.
22. Zerbino, D. R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–9.
23. Li, R., Zhu, H., and Wang, J. (2009) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* doi:10.1101/gr.097261.109.
24. Li, R. et al. (2009) The sequence and de novo assembly of the giant panda genome. *Nature* doi:10.1038/nature08696.
25. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**, 1124–32.
26. Lynch, M. (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182**, 295–301.
27. Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res* **17**, 1697–706.