Next-generation sequencing technologies and applications for human genetic history and forensics

Eva C Berglund, Anna Kiialainen and Ann-Christine Syvänen*

Abstract
Rapid advances in the development of sequencing technologies in recent years have enabled an increasing number of applications in biology and medicine. Here, we review key technical aspects of the preparation of DNA templates for sequencing, the biochemical reaction principles and assay formats underlying next-generation sequencing systems, methods for imaging and base calling, quality control, and bioinformatic approaches for sequence alignment, variant calling and assembly. We also discuss some of the most important advances that the new sequencing technologies have brought to the fields of human population genetics, human genetic history and forensic genetics.

Background
Determining the DNA sequence is the most comprehensive way of obtaining information about the genome of any living organism. For decades, Sanger sequencing [1] using fluorescently labeled terminating nucleotides and electrophoresis has been the gold standard sequencing technology. Sanger sequencing made an early impact in the field of microbial genomics, with the first complete bacterial genome, *Haemophilus influenzae*, sequenced in 1995 [2]. Multicenter collaborations using numerous sequencing instruments and automated sample preparation also made it possible to use Sanger sequencing in the human genome project, which took more than 10 years and US$2.7 billion to complete [3,4].

In recent years, we have witnessed a rapid development of a new generation of DNA sequencing systems followed by a multitude of novel applications in biology and medicine. The major advantage of the new ‘second-generation’ or ‘massively parallel’ sequencing technologies, compared to Sanger sequencing, is their considerably higher throughput and thereby lower cost per sequenced base. On a second-generation sequencing (SGS) machine several human genomes can be sequenced in a single run in a matter of days. Here, we review recent technological advances of SGS technologies and discuss the bioinformatic and computational implications of the sequencing revolution. Finally we highlight some applications of SGS technology with a focus on human population genetics and genetic history, and genetic forensics.

Second-generation sequencing technologies
There are three major SGS systems that are routinely used in many laboratories today. The first system to become commercially available was the Genome Sequencer from 454 Life Sciences (Branford, CT, USA) (later acquired by Roche [5]) in 2005, which was also the first SGS technology to sequence a complete human genome, that of Dr. James D. Watson [6]. The Genome Analyzer, first conceived by Solexa and later further developed by Illumina (San Diego, CA, USA) [7] was launched in 2006, and the SOLiD system from Applied Biosystems [8] (now part of Life Technologies (Carlsbad, CA, USA)) in 2007. The key steps of a sequencing project are the same for all of these technologies: preparation and amplification of template DNA, distribution of templates on a solid support, sequencing and imaging, base calling, quality control and data analysis (Figure 1).

In terms of applications, there are two major types of projects, *de novo* sequencing and resequencing. In a *de novo* sequencing project, the genome of an organism is sequenced for the first time. In contrast, in resequencing...
applications, the genome or parts of it are sequenced of a species where a reference sequence is already available. This difference affects both the selection of sequencing strategy and the data analysis (further discussed below). In human forensics and population genetics the resequencing approach is used, but in microbial forensics both de novo sequencing and resequencing of microbial genomes may be required.
Two common measures of the amount of sequence data generated in a project are the sequencing depth and breadth. Sequencing depth, or coverage, is the average number of times each base in the genome is sequenced. For example, to sequence a 3 Gb human genome to 30 × coverage, 90 Gb of sequence data is needed. The coverage will be uneven over the genome however, and sequencing breadth, sometimes also referred to as genome coverage, is the percentage of the genome that is covered by sequence reads.

**DNA samples for sequencing**

High-quality DNA in sufficient quantity is the basis for any successful sequencing experiment. For most sequencing applications, 1 to 5 μg of purified DNA is needed, an amount that may not always be available. Whole genome amplification (WGA) has frequently been used to increase the amount of DNA for genotyping [9] and can be applied also in combination with SGS. Several microbial genomes have been sequenced using SGS after WGA (for example, the genome of uncultured bacterial symbionts of termites isolated from a single host cell [10]). WGA was also recently used to amplify DNA from single cells from primary breast tumors, and although sequence data was retrieved from only 6% of the genome of each cell, this genomic representation was enough to identify subpopulations of cancer cells by copy number variations [11].

SGS can also be used to detect rare and unknown variants in genomic regions of interest in a cost-efficient way, and in a larger number of samples than by whole genome sequencing. Another advantage of targeted sequencing is the reduced issue of sequence reads aligning to multiple locations in the genome. The most commonly used methods for enrichment of genomic regions for sequencing are either based on hybridization to biotinylated probes in solution or probes immobilized on microarrays, or on multiplexed amplification by PCR (reviewed in [12]). The recently developed selector probe technology, which is based on rolling circle amplification, provides efficient and highly multiplexed enrichment of small regions totaling up to 1 Mb in size is particularly useful for ultra-deep sequencing at low cost and high specificity [13]. Sequencing of human exomes enriched by hybridization-based capture in solution is becoming widely used, and has proven to be particularly successful for identification of mutated genes underlying monogenic disorders (see, for example, [14-17]). The methods for hybridization-based capture are also applicable to custom-selected genomic regions of interest [18].

**Preparation of sequencing libraries**

The DNA samples to be sequenced are first converted into one of two main types of sequencing libraries, fragment libraries or mate-pair libraries. The first step in the preparation of a sequencing library is to fragment the DNA sample, usually using sonication or nebulization. For preparation of fragment libraries, sequencing adapters are ligated to both ends of the DNA fragments, followed by PCR amplification using primers complementary to the adapters.

In the Illumina SGS technology, adapter-ligated DNA fragments are amplified directly in the flow cell subsequently used for sequencing. Each flow cell has eight channels (lanes) coated with oligonucleotides that are complementary to the adapters. The adapter-ligated DNA fragments are hybridized to the flow cell, in which they are distributed randomly and amplified by a process called bridge amplification. After amplification, DNA molecules are linearized to form clusters, each of which consists of about 1,000 copies of the original DNA molecule at that position.

In the 454 and SOLiD technologies, adapter-ligated fragments are hybridized to beads coated with an oligonucleotide that is complementary to one of the adapters for amplification in a water-in-oil emulsion PCR. Each water droplet constitutes a microreactor containing the PCR reagents and optimally a single bead with a single immobilized DNA fragment. Thus, multiple PCR can be performed in parallel in a single tube. After breaking the emulsion, the beads, which are now coated with thousands or millions of copies of the original DNA molecule, are loaded onto the solid support for sequencing. In the 454 system, the solid support is called PicotiterPlate and consists of wells that can fit a single DNA-coated bead each. SOLiD uses a glass slide to which the beads are distributed randomly.

The amplified fragments are then sequenced either from one end (single-end) or from both ends (paired-end). Paired reads allow more accurate alignment to a reference genome, and are also very useful to resolve repeats and improve assembly in *de novo* sequencing projects. The Illumina system generates sequence reads of the same length from both ends, whereas the second read from SOLiD is shorter (Table 1). The 454 system currently does not support paired-end sequencing of fragment libraries.

Mate-pair libraries are constructed by circularizing fragmented DNA, thereby bringing the two ends of the original DNA fragment adjacent to each other (Figure 2). After fragmentation of the circular DNA, the fragment containing the ends of the original linear DNA is selected using biotin capture. Sequencing both ends of the selected fragment will yield reads that are separated by the distance of the original fragment. In order to avoid chimeric sequence reads that span over both original fragment ends, the 454 and SOLiD systems include an internal adapter.
preparation, no internal adapter is used, and to mini-
mize the risk of sequencing over the original junction
the recommended read length is limited to 36 nucleo-
tides. Mate-pair libraries allow larger insert sizes (2 to
20 kb) than paired-end sequencing of fragment libraries.
Drawbacks of mate-pair sequencing are that the labora-
tory protocols are more complicated and that a substan-
tially larger amount of DNA (5 to 120 μg) is required.

In contrast to paired-end reads, which are oriented
towards each other, mate-pair reads are either both
oriented outwards from the original fragment or both
have the same orientation (Figure 2), which needs to be
accounted for in the data analysis. Large inserts are
especially valuable in de novo sequencing projects,
where they can substantially improve scaffolding(order-
ing of assembled contigs). Mate-pair sequencing is not
used as frequently in resequencing projects, where DNA
resources are often limited and the analysis is mainly
based on alignment to a reference genome.

Introducing an additional index tag (barcoding) to
each DNA fragment makes it possible to sequence
pooled samples that can be distinguished in silico after
sequencing. Multiplexing is useful in applications where
a relatively small amount of data is needed from each
sample, such as sequencing of small genomes or
enriched regions of large genomes (see, for example,
[19,20]). As the capacity of the sequencing instruments
has increased, multiplex sequencing of indexed samples
has become more and more important to minimize
sequencing costs. Indexing also decreases the risk of
sample mix-ups and contaminations during library pre-
paration. Currently, Illumina provides 24 different index
tags, Life Technologies 96 and Roche 12 (Table 1). Additional index tags for Illumina (48 tags in total) and
454 (120 additional tags) can be purchased from other
companies. It is also possible to use custom-designed
index tags [21,22].

Sequencing and imaging principles

Sequencing-by-synthesis

The Illumina and 454 technologies are based on sequen-
cing-by-synthesis. A DNA polymerase is used to extend
a sequencing primer by incorporating nucleotides that
form a growing sequence complementary to the tem-
plate DNA. In the Illumina system, fluorescent reversi-
ibly terminating nucleotides are used. All four
nucleotides are added at the same time, each with a
unique fluorescent label, which allows incorporation of
one base per cycle into each template molecule [23] (Figure 3a). After incorporation and fluorescence regis-
tration at four wavelengths, the terminating and fluores-
cent moieties are removed from the nucleotides to allow

Table 1 Characteristics of second-generation and third-generation sequencing instruments

| Instrument                          | Read length (nucleotides) | No. of readsa | Output (Gb)a | No. of samplesa, b | Runtime | Advantages                   | Disadvantages                                      |
|------------------------------------|---------------------------|---------------|--------------|--------------------|---------|------------------------------|---------------------------------------------------|
| Roche 454 GS FLX+                  | 700^                       | 1 × 10^9      | 0.7          | 192                 | 23 h    | Long reads, short run time   | Homopolymer errors, expensive                      |
| Illumina HiSeq2000                 | 100^                       | 3 × 10^9      | 600          | 384                | 11 days | High yield                   | No. of index tags limiting                         |
| Life Technologies SOLiD 5500xl     | 75^                         | 1.5 × 10^9    | 180          | 1,152              | 14 days | Inherent error correction    | Short reads                                      |
| Roche 454 GS Junior                | 400^                       | 1 × 10^9      | 0.035        | 132                | 9 h     | Long reads                   | Homopolymer errors, expensive                      |
| Illumina MiSeq                    | 150                        | 5 × 10^5      | 1.5          | 96                 | 27 h    | Short run time, ease of use  | Expensive per base                                |
| Ion Torrent PGM Ion 316 chip       | > 100^                     | 1 × 10^9      | 0.1          | 16                 | 2 h     | Short run time, low reagent cost | Not well evaluated                               |
| Helicos BioSciences HeliScope      | 35^                        | 1 × 10^9      | 35           | 4,800              | 8 days  | SMS, sequences RNA           | Short reads, high error rate                      |
| Pacific Biosciences PacBio RS      | > 1,000^                   | 1 × 10^9      | 0.1          | 1                  | 90 min  | SMS, long reads, short run time | High error rate, low yield                        |

Most of this information is subject to rapid change, and the aim of this table is not to present absolute numbers but to provide a general comparison between different sequencing systems.

^Numbers calculated for two flow cells on HiSeq2000 and SOLiD 5500xl.

^Calculated as no. of index tags (provided by the sequencing company) × no. of divisions on solid support.

^Average for single-end sequencing, paired-end reads are shorter.

^No. of reads decreases when the PicoTiterPlate is divided.

^36 nucleotides for mate-pair reads.

^Run time depends on the read length, and on whether one or two flow cells are used.

^Second read in paired-end sequencing is limited to 35 nucleotides, and mate pair reads to 60 nucleotides.

^Average.

SMS = single molecule sequencing.
the next sequencing cycle. In 2010, Illumina released the HiSeq2000 instrument, which uses the same chemistry as the original Genome Analyzer instrument, but has improved imaging optics and can process two flow cells in parallel. The HiSeq2000 system has the highest throughput of all currently available SGS instruments, with around 600 Gb sequence produced per run (Table 1). Examples of what can be achieved with the current capacity of HiSeq2000 are shown in Table 2. Sequencing errors are primarily substitution errors and occur more frequently in the distal bases of a read.

In the 454 sequencing-by-synthesis reaction, natural non-terminating deoxynucleotides are added to the system sequentially (Figure 3b). In homopolymeric regions several bases will thus become incorporated in the same step. The 454 technology is based on the pyrosequencing principle [24], where pyrophosphate is released as a consequence of nucleotide incorporation and converted into ATP by sulfurylase. ATP is then used as a substrate for the production of light by luciferase, and the emission of light is registered by a charge-coupled device (CCD) camera. The major advantage of the 454 technology is the long read length. The misincorporation rate for the natural deoxynucleotides is low, resulting in low levels of nucleotide substitution errors. Insertion-deletion errors are frequent in homopolymeric regions, however, due to the non-linear light response when several nucleotides are incorporated simultaneously to the same molecule. Compared to the Illumina and SOLiD systems, the 454 technology is more expensive per base due to the lower capacity and the higher reagent cost associated with the multiple enzymes required. Thus, the 454 technology is mainly used in applications where long reads are desired, such as de novo sequencing.

Figure 2 Principles for construction of mate-pair sequencing libraries (a) Preparation of Illumina mate-pair libraries. Fragments are end-repaired using biotinylated nucleotides (1). After circularization, the two fragment ends (green and red) become located adjacent to each other (2). The circularized DNA is fragmented, and biotinylated fragments are purified by affinity capture. Sequencing adapters (A1 and A2) are ligated to the ends of the captured fragments (3), and the fragments are hybridized to a flow cell, in which they are bridge amplified. The first sequence read is obtained with adapter A2 bound to the flow cell (4). The complementary strand is synthesized and linearized with adapter A1 bound to the flow cell, and the second sequence read is obtained (5). The two sequence reads (arrows) will be directed outwards from the original fragment (6). (b) Preparation of Roche 454 paired-end libraries (these are called paired-end, but are based on the same principles as the mate-pair libraries in the other technologies). Original fragments (1) are end-repaired with unlabeled nucleotides, and biotin-labeled circularization adapters (CA) are ligated to the fragment ends (2). After circularization (3), fragmentation and affinity purification, library adaptors (LA1 and LA2) are ligated to the new fragment ends (4) and the fragments are amplified on beads by emulsion PCR. One single sequence read that covers the two original ends and the internal adapter is generated (5). Adapter sequence is removed in silico, and the sequence is split into two reads, which both have the same orientation (6). (c) Preparation of SOLiD mate-pair libraries. Steps 1 to 4 are analogous with preparation of Roche 454 paired-end libraries, with a biotin-labeled internal adapter (IA) and two sequencing adapters (P1 and P2). Sequencing is performed with two different primers, complementary to the P1 adapter and internal adapter, respectively (5). The resulting reads will have the same orientation (6).
projects, where the read length is the most important factor determining the quality of the assembly, and in metagenomics, where the sample contains a mix of different organisms.

**Sequencing-by-ligation**

The SOLiD technology is based on sequencing-by-ligation, where a DNA ligase is used to add probes to a growing oligonucleotide chain [25] (Figure 3c). The
probes consist of eight bases, five that are specific and complementary to the template and three that are universal and support hybridization to the DNA template. In the sequencing reaction, probes containing all possible combinations of the first five nucleotides are added. The probe that matches the template perfectly becomes hybridized and ligated to the sequencing primer or previous probe. The probes are fluorescently labeled according to the first two bases using a scheme for two-base encoding with four fluorophores. After imaging, the fluorescent label and the three universal bases are cleaved off, and a new set of probes is added. After the first round of sequencing, the newly built DNA strand is melted off, a new sequencing primer which starts one base off from the first primer binding site is hybridized, and the sequencing reaction is repeated now interrogating different positions. This process is repeated several times with different primers, so that all bases in the template become sequenced twice. Since each base is sequenced twice in the SOLiD system, most sequencing errors can be corrected during alignment, resulting in a low error rate of mapped data. It is possible to use different chemistries and read lengths in different lanes. The current read length is 75 nucleotides for fragment libraries, and the total yield per run (two glass slides) is around 180 Gb. Data analysis has traditionally been performed in color space, however, with the recent upgrade to the 5500xl system, it is now also possible to get error corrected reads in base space.

Sequencing-by-ligation is also used by Complete Genomics (Mountain View, CA, USA) [26], a company that sequences human genomes as a service. In their technology, the template DNA is first inserted into a single-stranded DNA circle, which is then copied several times to make up DNA nanoballs. The nanoballs are attached to arrays and sequenced by ligation reactions, which use multiple priming sites [27]. The current capacity of Complete Genomics is more than 600 genomes per month, and they are driving down the price of whole-genome sequencing.

**Base calling and quality control**

Intensities of light signals from the sequencing reactions are converted to bases (Illumina and 454 systems) or colors (SOLiD system). In addition to sequence data, base calling produces quality scores for each base, which are estimates of the probability of the call being erroneous. After base calling, reads with indications of mixed signals or other errors are filtered out. To facilitate troubleshooting and discrimination between problems caused by instrument/reagent factors and sample factors, the Illumina and 454 platforms include standardized control DNA in each run.

In the Illumina system, the clusters are identified during the first four cycles of sequencing. Intensities are registered for each cluster in every cycle and converted to nucleotide sequence. If the initial recognition of clusters was not perfect, some clusters may contain more than one original template molecule. It is also possible that some clusters contain many molecules that have incorporated fewer (phasing) or more (prephasing) nucleotides than the number of cycles. Such clusters are filtered out by a so-called chastity filter, which is based on the ratio of signal intensities of the bases with the strongest and the second strongest intensity. Control DNA from the phage phiX is sequenced in each flow cell. In the analysis pipeline, phiX reads are identified by comparison to the phiX genome and the error rate is determined and used as a measure of the quality of the run.

During QC filtering of data from the 454 system, possible polyclonal beads and beads with no template are identified based on the number of positive and negative flows. In addition, reads that do not start with a specific key sequence, which is part of the adapter, and reads that have a high number of off-peak signal intensities (indicative of homopolymer errors) are filtered out. Sequence reads are also trimmed from the 3’ end to remove adapter sequence and bases of low quality, arising from phasing/prephasing issues and loss of signal intensity. Beads with control DNA, labeled with a different key sequence, are included in each run. With the aid of the key sequence, these sequence reads are identified and aligned to a reference sequence, and the percentage of reads that match with 95, 98 and 100% similarity is reported.

After color calling in the SOLiD system, possible polyclonal reads and reads with color combinations that do not make sense according to the two-base encoding scheme are filtered out. No control DNA is used, but the quality of a run is assessed from the color intensity distribution.

Despite the standard QC steps, not all obtained data will be of high quality. To recognize potential problems and biases it is useful to apply additional quality control measures. A good resource for assessing the quality of sequence data is the FastQC software [28], which reports, for example, distributions of base qualities, GC content, redundancy and over-representation of adapter or primer sequence.

**Trends and upcoming technologies**

**Low-capacity sequencing systems**

While the sequencing companies compete to increase throughput, they have also launched systems with lower capacity. The reason for this trend is that the high capacity of the original systems is not always required, and the current multiplexing possibilities do not match their throughput, which results in a much higher coverage (and cost) than needed for many applications.
Roche 454 Technologies was first to launch the Genome Sequencer Junior system in 2010 and Illumina launched the MiSeq system in 2011. The capacity of these systems is 35 Mb and 1.5 Gb per run, respectively. Life Technologies recently acquired the company Ion Torrent, whose technology has a concept similar to that of the 454 system. However, detection is based on pH changes caused by release of electrons upon nucleotide incorporation rather than pyrophosphate release. Since no enzymes are used for detection, the reagent cost for this instrument is low.

Compared to the high-capacity instruments, the cost per base is high for the smaller machines, but they are suitable for sequencing small genomes, amplicons or DNA enriched by targeted capture. Since the run time for these systems is short, they are also useful for technology development runs and to optimize reaction conditions for a larger run. These machines have been referred to as ‘personal sequencers’, meaning that they are easily obtainable also by laboratories with smaller resources that want to have rapid access to a sequencing instrument, but need relatively small amounts of data.

**Single molecule sequencing**

In single molecule sequencing, sometimes also referred to as third-generation sequencing, no amplification of the template molecules is performed prior to sequencing. These technologies provide improved quantitative accuracy by eliminating the risk of biases introduced during preparation of sequencing libraries. Single molecule sequencing also allows direct sequencing of RNA molecules, detection of chemically modified bases such as DNA methylation, and increased read lengths. Longer reads will be useful in *de novo* sequencing projects and open up perspectives for experimental phasing (determination of which variant alleles are on the same chromosome), in contrast to statistical phasing that is used today.

In the Heliscope Single Molecule Sequencer system from Helicos Biosciences (Cambridge, MA, USA) [29] single stranded poly(dA)-tailed templates are attached to poly(dT) oligonucleotide primers that are anchored on a flow cell. In each sequencing cycle one type of reversibly terminating fluorescently labeled nucleotides is added and incorporated by a polymerase, the slide is washed and imaged, and the dye labels are cleaved off [30,31]. This technology generates around 35 Gb per run, and the read length is 35 nucleotides on average (Table 1).

Pacific Biosciences (Menlo Park, CA, USA) [32] has developed a system called single molecule real time (SMRT) sequencing, which uses a DNA polymerase anchored on a glass surface and nucleotides with phospholinked fluorescent labels that are cleaved off when the nucleotides are incorporated. The sequencing reaction takes place on zero-mode waveguide nanostructure arrays [33]. The incorporation of the fluorescently labeled nucleotides is monitored in real time, which results in very short run times. This system has read lengths over a thousand bases, but error rates are high and throughput is currently limited to 0.1 Gb per run (Table 1). The Pacific Biosciences system was successfully used for rapid analysis of the *Cholera* strains in the outbreak in Haiti in 2010 [34].

Several new technologies for single molecule sequencing are under development. Nanopore sequencing technologies (Oxford Nanopore (Oxford, UK) [35], NABsys (Providence, RI, USA) [36]) are based on detecting natural electric or chemical differences between nucleotides, and do not require labeling of DNA. The Starlight system (Life Technologies) is a real-time technology that uses a quantum-dot-labeled polymerase and distinctly labeled fluorescent nucleotides. The reaction and detection principles underlying the third-generation sequencing systems are demanding, and today it is not obvious which of these systems that will reach the capacity and accuracy required for practical ‘real life’ applications.

**Bioinformatic analysis of sequence data**

**Alignment**

For resequencing applications, where a reference sequence is already available, the first step of the analysis is usually to align the sequence reads to the reference genome (Figure 4a). In recent years, several novel alignment programs have been developed that are adapted to the shorter read lengths, different error distributions and larger data amounts obtained by SGS technologies (reviewed in [37]). These programs have different properties in terms of, for example, their ability to perform gapped alignment, how base qualities are used during alignment and how reads aligning to repeated regions are treated. Some aligners can handle data from any sequencing platform, whereas others are specific to one platform.

To save computational time, the alignment is usually performed in two steps: (1) a limited number of candidate positions are identified by fast heuristic approaches, and (2) candidate positions are evaluated by more accurate methods, such as the Smith-Waterman algorithm [37]. The first step of the alignment is typically based either on hash tables or on the FM index (named for its creators, Paolo Ferragina and Giovanni Manzini). Usually, the algorithms do not map the entire read at once, but start with a seed sequence consisting of the first 20 to 40 bases. The seed is made up of consecutive bases, or allows for mismatches at certain positions (spaced seeds). To improve sensitivity, several different spaced seeds can be used. To improve speed, some programs discard reads with too many candidate positions.

Regardless of which aligner is selected, there is always a trade off between speed and sensitivity. As the amount
of data has increased, algorithms based on the FM index (for example, Bowtie \[38\] and Burrows-Wheeler Aligner (BWA) \[39\]) have become increasingly popular. This index is based on the Burrows Wheeler transform, which was originally used in algorithms for data compression and is efficient for large amounts of data. The expected divergence between the reference sequence and the reads also influences the efficiency of alignment programs. While the BWA loses rapidly in speed when the difference between the reads and the reference is more than 4%, the use of several spaced seeds in BFAST \[40\] makes it highly tolerant to divergent reads.

**Variant calling**

There are multiple programs for identifying single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and structural variants between sequence reads and a reference genome (reviewed in \[41\]). These programs aim to distinguish authentic variants from
sequencing errors and incorrect alignments by evaluating different parameters, such as base qualities, coverage and the number of reads supporting the variant. Different combinations of aligners and variant callers typically yield different results, and there is no consensus in the scientific community on which algorithms are most appropriate to use. An approach that was used in the 1000 Genomes Project is to run several variant callers, and report the variants detected by at least two different methods.

Many variant callers tend to overcall, and it is often necessary to filter the detected variants. It is also possible to improve the results by local realignment around candidate variants and recalibration of quality scores before variant calling, using the software suite GATK [42]. During realignment, all reads around potential indels (insertions and deletions) are extracted and subsequently aligned again in a multiple alignment with exact methods. This can remove many false positive variants, since alignments are especially error prone around indels. During recalibration, different types of information, such as base quality scores, position in the read and dinucleotide content at positions present in dbSNP is used to improve the base quality scores.

De novo assembly

In a de novo sequencing project, sequence reads are aligned to each other without the guidance of a reference sequence, with the aim to assemble as long contiguous sequences (contigs) as possible (Figure 4b). This process is computationally much more challenging than alignment to a reference sequence. For example, running an assembly of a human genome requires around 150 GB of RAM using SOAPdenovo [43] and around 512 GB of RAM using Allpaths-LG [44]. In contrast, aligning paired-end sequences to a reference using BWA only requires 3.5 GB of RAM, and can be done on a standard desktop computer for a limited number of samples. Long reads and paired-end or mate-pair data help to resolve repeated regions, and will greatly improve the assembly. Since throughput, read length and error distributions differ between the platforms, a combination of data can often be useful for de novo sequencing, such as 454 data to obtain long reads and Illumina data to increase the coverage and correct homopolymer errors. As of October 2011, the number of published genomes sequenced with SGS is approximately the same as those sequenced with Sanger sequencing according to the Genomes OnLine Database (GOLD) [45]. However, among approximately 1,000 listed ongoing projects, Sanger sequencing is only used in 10%, with most of the genomes being sequenced with Illumina (49%) or 454 (40%). 454 sequencing is more frequently used for bacterial and archaeal genomes than for eukaryotes.

Algorithms developed for de novo assembly of short reads are based on the de Bruijn graph rather than the overlap graph, which is used in most assemblers of Sanger sequencing data. In the de Bruijn graph, reads are decomposed into shorter pieces of length k (k-mers), and overlaps between the k-mers are identified [46]. Memory and time requirements increase rapidly with the number of unique k-mers, which in turn is heavily influenced by sequencing errors. Therefore, some programs attempt to correct sequencing errors by substituting k-mers with very low incidence with the most similar high-incidence k-mer. Despite novel algorithms, de novo assembly of large genomes remains very difficult, and it is also hard to evaluate the quality of an assembly. As an attempt to advance this field, the Genome Center at University of California Davis and researchers at University of California Santa Cruz recently launched a genome assembly competition called The Assemblathon, based on both real and simulated data [47].

Computer resources

The increasing output of the new sequencing technologies continually raises the requirements for computer resources and bioinformatic competence to handle the data [48]. Furthermore, the rate of generation of sequence data increases much faster than the capacity of storing data, suggesting that storage might become a major bottleneck. To circumvent this, there are ongoing efforts to develop a more efficient file format, the CRAM format, which is based on storing the differences compared to a reference rather than the reads themselves [49]. The CRAM format requires five times less storage space than the BAM format, which is the standard today. In lack of sufficient computer power for data analysis, it is possible to analyze data using internet-based so-called cloud computing and pay for the service by the hour. Several public data sets are already available on clouds; it is for example possible to analyze the data from the 1000 Genomes Project on the Amazon Web Services Cloud [50]. The use of hardware acceleration solutions such as graphics processing units to decrease computational time is not widely employed in bioinformatics yet, but may increase in the future. An example of a program that uses this approach is BarraCUDA [51], a sequence aligner based on the BWA algorithm, which uses the Nvidia CUDA architecture.

A useful resource for information regarding software and analysis is SEQuancers [52], a community for discussion on new sequencing technologies. SEQuancers also lists many of the programs available for data analysis [53]. In addition to open-source software, there are also commercial software suites, like CLCbio [54], which may be more accessible for users without programming experience.
Population genetics

Genetic information provides valuable insights on human origin and migration, and genotyping studies have revealed a distinct geographic substructure of human populations [55,56]. SGS technologies, including sequencing of complete genomes of many individuals, will allow population genetic studies at substantially increased resolution and improve detection of variation at the rare end of the scale. However, they will at the same time increase the complexity of interpretation.

The 1000 Genomes Project, which was launched in 2007 as a collaborative effort between three sequencing centers in the USA, the Sanger Institute in the UK and the Beijing Genomics Institute in China, was the first attempt to apply SGS in population genetics [57]. Analysis of the first set of data from three pilot projects revealed that each person carries an estimated 250 to 300 loss-of-function variants in annotated genes, and the de novo germline substitution rate was estimated to 10^{-8} per base and generation [58]. Additionally, several hundred thousand SNPs with allele frequencies that differ between populations were identified. The enrichment of non-synonymous changes at these sites suggested an action of local adaptation [58]. Most variable sites have a low minor allele frequency, and are not shared among diverged populations [59]. The complete 1000 Genomes Project will include low-coverage genome sequencing, exome sequencing and genotyping of 2,500 individuals from 27 populations in Europe, Africa, Asia and the Americas, and will be an invaluable resource for studies of human population genetics.

Two recent studies developed new statistical methods for population genetic analysis of whole genome sequence data, to infer human demographic parameters [60-62]. The results showed that European and Chinese populations have similar population size histories, and both populations experienced a more severe bottleneck than African populations 10,000 to 60,000 years ago [60]. It was also found that the Eurasians diverged from the Africans 38,000 to 64,000 years ago and that the southern African San population diverged from other human populations even earlier, 108,000 to 157,000 years ago [61]. High-throughput sequencing of enriched mitochondrial DNA (mtDNA) of 109 randomly selected samples from three Filipino groups suggested that previous estimates of population size history of these populations, obtained from a biased sampling and Sanger sequencing of the hypervariable region 1, were erroneous, thus demonstrating the advantage of SGS methods [63]. As more and more genome sequences from diverse populations become available, it will be possible to decipher human history in much more detail than ever before.

Human genetic history

Although population genetic studies have been powerful for studying human history, the studies have mostly concentrated on extant populations. Yet, humans have always been fascinated by extinct cultures and species. Where did they come from? What did they look like? How are they related to contemporary species? Why did they become extinct? With recent progress in retrieving and analyzing DNA from ancient archeological findings [64], the answers to these questions are now emerging.

Ancient DNA has been isolated and sequenced from several extinct animals, including mammoth, cave bear, ground sloth and moas, and has provided insights into the lifestyle of these species [65-70]. The first ancient, although relatively recent, human genome to be sequenced was from a 4,000 years old permafrost-preserved hair tuft from a Palaeo-Eskimo from Qeqertasussuk in western Greenland [71]. The sample belonged to a male of the Saqqaq culture, the first known inhabitants of Greenland. The genome was sequenced to a coverage of 20 × with 79% of the genome represented. The analysis revealed that the Saqqaq man was more closely related to populations from far eastern Siberia than to the current inhabitants of Greenland [71], suggesting that the Saqqaq culture arose from an independent migration from Siberia about 5,500 years ago. Several phenotypic characteristics of the Saqqaq man were deduced from the genome sequence, including dark and thick hair and a metabolism and body mass index adapted to a cold climate [71].

The complete mitochondrial DNA of the Neanderthals, who are the closest known relatives of present-day humans and lived in Europe and western Asia circa 400,000 to 30,000 years ago [72,73], has been sequenced from bones found in Spain, Germany, Croatia and Russia [74,75]. A draft sequence at 1.3 × coverage of the complete Neanderthal genome was generated from DNA isolated from three approximately 40,000-year-old bones found in Vindija Cave in Croatia [76]. The analysis suggested that the Saqqaq man was more closely related to populations from far eastern Siberia than to the current inhabitants of Greenland [71], suggesting that the Saqqaq culture arose from an independent migration from Siberia about 5,500 years ago. Several phenotypic characteristics of the Saqqaq man were deduced from the genome sequence, including dark and thick hair and a metabolism and body mass index adapted to a cold climate [71].

The complete mitochondrial DNA of the Neanderthals, who are the closest known relatives of present-day humans and lived in Europe and western Asia circa 400,000 to 30,000 years ago [72,73], has been sequenced from bones found in Spain, Germany, Croatia and Russia [74,75]. A draft sequence at 1.3 × coverage of the complete Neanderthal genome was generated from DNA isolated from three approximately 40,000-year-old bones found in Vindija Cave in Croatia [76]. The analysis suggested that the Saqqaq man was more closely related to populations from far eastern Siberia than to the current inhabitants of Greenland [71], suggesting that the Saqqaq culture arose from an independent migration from Siberia about 5,500 years ago. Several phenotypic characteristics of the Saqqaq man were deduced from the genome sequence, including dark and thick hair and a metabolism and body mass index adapted to a cold climate [71].

The mtDNA [77] and complete genome at 1.9 × coverage [78] of a 30,000 to 50,000 years old finger bone from an unknown hominin group, found in Denisova Cave in southern Siberia has also been sequenced. The analysis suggested that this population, called Denisovans, was a sister group to the Neanderthals. It was estimated that Denisovans contributed to approximately 5% of the genomes of the present-day population of Melanesia, the island region north and northeast of Australia.

Working with ancient DNA requires special consideration, since DNA degrades into shorter fragments...
Forensic genetics

DNA profiling by typing short tandem repeats (STRs) is today the standard method for forensic analysis, and the Y chromosome is the most informative marker for forensic genotypes. The standard STR typing method provides sufficient discrimination power for most applications, the use of forensic markers derived from large-scale sequencing. The major problem with such a replacement would be that forensic markers are already widely in use in the existing forensic databases are based on STR profiles, and in many cases forensic samples have not been retained for long-term analyses. Unfortunately, the current SCS technologies are relatively poor at resolving STRs, and this replacement would require the databases as well as collection of redundant SNP information during a transition phase. The benefits of analyzing degraded samples present in small quantities, simplified laboratory procedures, larger amount of data and accurate variant calling using SCS in forensic analysis.

In recent years, genome-wide association studies have used single nucleotide polymorphisms (SNPs) to help identify forensic markers. This has helped to improve resolution in forensic analyses. In a recent study, two Y chromosomes separated by approximately 13 generations were sequenced using SCS technology [83,84]. Approximately one mutation per generation was identified, suggesting that every individual Y chromosome can be relatively accurately (80% to 90%) predicted. However, the current SCS technologies are relatively poor at resolving STRs, and this replacement would be relatively accurately (80% to 90%) predicted. The major problem with such a replacement would be the increase in amount of data and accurate variant calling using SCS during a transition phase. The benefits of analyzing degraded samples present in small quantities, simplified laboratory procedures, larger amount of data and accurate variant calling using SCS in forensic analysis.
mtDNA material by target capture, similar to how Neanderthal DNA was enriched from contaminated samples using primer extension capture [74] and array-based capture [87].

In comparison with human samples, where a limited set of predesigned DNA markers is usually enough for forensic analysis, microbial samples provide an additional challenge since the markers are unknown and specific for each case. Microbial forensics has previously benefited from Sanger sequencing, for example for determination of the origin of the bacteria used in the well known 2001 anthrax attack in the USA [88,89]. To resolve this case, thousands of colonies were inspected manually, and a number of colonies with slightly different morphology were selected for whole genome sequencing [88,89]. Based on the identified polymorphisms, four tests were designed and used to screen more than 1,000 isolates from different laboratories. The cost of each anthrax genome in 2001 was estimated to US $140,000 [90]. In comparison, four Bacillus anthracis strains were recently sequenced on the SOLiD system with a reagent cost per genome of less than US$1,000 [91]. According to the current specifications of Illumina’s HiSeq2000 instrument, approximately 300 × coverage of 384 B. anthracis isolates can be obtained from a single run at a reagent cost of less than US$50 per genome.

Conclusions

The impressive advances of sequencing technologies in recent years have enabled a range of new applications, including population genomics based on the complete genome sequences of a large number of individuals and sequencing of the complete genomes of highly contaminated samples of ancient DNA from humans and other species. The technical development of current SGS systems shows no sign of halting, with major upgrades from both the Illumina and SOLiD platforms in 2011. In forensic genetics, SGS is not yet widely employed, although a number of studies have highlighted emerging possibilities. With further decreased sequencing cost and miniaturized equipment, it will become possible for in principle any experimental research group to perform whole genome sequencing of both large and small genomes on a bench-top DNA sequencer. While the resources for data storage and analysis will initially be limiting, new simplified solutions for data storage and analysis are also underway. Before long, DNA sequencing is likely to become an easily accessible routine method, like the PCR technique is today.

References

1. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977, 74:5463-5467.
2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott JT, Sherry R, Liu L, Glodek A, Kelley JM, Weidman JJ, Phillips CA, Spigots TG, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, et al: Whole-genome random sequencing and analysis of Haemophilus influenzae Rd. Science 1995, 269:409-512.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine J, Livi V, Mergen M, Myers W, Norris J, O’roka C, cancelling 2001 anthrax attack in the USA [88,89]. To benefit from Sanger sequencing, for example for determination of the origin of the bacteria used in the well known 2001 anthrax attack in the USA [88,89]. To resolve this case, thousands of colonies were inspected manually, and a number of colonies with slightly different morphology were selected for whole genome sequencing [88,89]. Based on the identified polymorphisms, four tests were designed and used to screen more than 1,000 isolates from different laboratories. The cost of each anthrax genome in 2001 was estimated to US $140,000 [90]. In comparison, four Bacillus anthracis strains were recently sequenced on the SOLiD system with a reagent cost per genome of less than US$1,000 [91]. According to the current specifications of Illumina’s HiSeq2000 instrument, approximately 300 × coverage of 384 B. anthracis isolates can be obtained from a single run at a reagent cost of less than US$50 per genome.

Conclusions

The impressive advances of sequencing technologies in recent years have enabled a range of new applications, including population genomics based on the complete genome sequences of a large number of individuals and sequencing of the complete genomes of highly contaminated samples of ancient DNA from humans and other species. The technical development of current SGS systems shows no sign of halting, with major upgrades from both the Illumina and SOLiD platforms in 2011. In forensic genetics, SGS is not yet widely employed, although a number of studies have highlighted emerging possibilities. With further decreased sequencing cost and miniaturized equipment, it will become possible for in principle any experimental research group to perform whole genome sequencing of both large and small genomes on a bench-top DNA sequencer. While the resources for data storage and analysis will initially be limiting, new simplified solutions for data storage and analysis are also underway. Before long, DNA sequencing is likely to become an easily accessible routine method, like the PCR technique is today.

References

1. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977, 74:5463-5467.
2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott JT, Sherry R, Liu L, Glodek A, Kelley JM, Weidman JJ, Phillips CA, Spigots TG, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, et al: Whole-genome random sequencing and analysis of Haemophilus influenzae Rd. Science 1995, 269:409-512.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine J, Livi V, Mergen M, Myers W, Norris J, O’roka C, etc.
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huison GD, Wortman JR, Zhang Q, Kodira CD, Zheng YH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GI, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al: The sequence of the human genome. Science 2001, 291:1304-1315.
5. Roche: Roche 454.http:[www.my454.com].
6. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niaz F, turcotte CL, Itoh GP, Luik PS, Chaintaut C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothenberg JM: The complete genome of an individual by massively parallel DNA sequencing. Nature 2008, 452:872-876.
7. Illumina: Illumina homepage.http:[www.illumina.com].
8. Applied Biosystems: Applied Biosystems homepage.http:[www. appliedbiosystems.com].
9. Lowmar L, Syvami AC: Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. Hum Mutat 2006, 27:630-614.
10. Hongh Y, Sharma VK, Prakash T, Noda S, Taylor TD, Kudo T, Sakaki Y, Shibata T, Hanada M, Ohmoto M, Ohkuma M: Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. Proc Natl Acad Sci USA 2008, 105:5555-5560.
11. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McInodoo J, Cook K, Stepaniak A, Levy D, Espoito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M: Tumour evolution inferred by single-cell sequencing. Nature 2011, 472:90-94.
12. Turner GH, Nig SB, Nickerson DA, Shendure J: Methods for genomic partitioning. Annu Rev Genomics Hum Genet 2009, 10:263-284.
13. Johansson IS, Isaksson M, Sorqvist ET, Hoos NF, Stenberg J, Sjoblom T, Pettersson K, Weidman JF, Phillips CA, Spigots TG, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, et al: Whole-genome random sequencing and analysis of Haemophilus influenzae Rd. Science 1995, 269:409-512.
14. Piersk SB, Walsh T, Chisolm KM, Lee MK, thompson AM, Fumaura A, Opitz JM, Levy-Lahad E, Klein KB, King MC: Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perraun Syndrome. Am J Hum Genet 2010, 87:282-288.

Acknowledgements

Our research group in Molecular Medicine is funded by the Swedish Research Council (VT-NT and VR-MH contract numbers 90559401 and 90560401), the Foundation for Strategic Research in Sweden (BSc08-0008), the Swedish Foundations for Cancer Research and Pediatric Cancer Research and the European Commission (FP7 projects ENGAGE no 201413, eEUvadds no 261123, ESGI 262055). We thank Olof Karlberg, Kristina Larsson and Ulfik Lijiedahl for input to the manuscript and Magnus Lundgren for assistance with graphical work.
58.  Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Halti NE, McVean GA. A map of human genome variation from population-scale sequencing. Nature 2010, 467:1061-1067.

59.  Gravel S, Heinitz BM, Guttenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA 2011, 108:11983-11988.

60.  Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature 2011, 475:493-496.

61.  Gronau J, Hubisz MJ, Guld MD, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 2011, 43:1031-1034.

62.  Pritchard JK. Whole-genome sequencing data offer insights into human demography. Nat Genet 2011, 43:923-925.

63.  Gunnnardottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M. Out of Africa: modern human origins special feature: the full Neandertal genome. Science 2010, 329:1212-1213.

64.  Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Alberts WR, Good JM, Hedges E, Green RE, Krause J, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkann C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Tallermo S, Shunkov MV, Derevianko AP, Hübner JJ, Kelso J, Stankiewicz SB, Paabo S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 2010, 468:1053-1060.

65.  Kayser M, de Knijff P. The hare and the tortoise: one small step for four SNPs, one giant leap for SNP-kind. Forensic Sci Int Genet 2010, 4:59-61.

66.  Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Dettner JC, Paabo S, Rubin EM. Genomic sequencing of Pleistocene cave bears. Science 2005, 309:597-599.

67.  Cooper A, Lalueza-Fox C, Anderson SW, Rambaut A, Austin J, Ward R. Complete mitochondrial genome sequences of two extinct moas clarify rattle evolution. Nature 2001, 409:74-77.

68.  Hoss M, Dilling A, Currant A, Paabo S. Molecular phylogeny of the extinct ground sloth Mylodon darwini. Proc Natl Acad Sci USA 1996, 93:181-185.

69.  Krajewki C, Buckley L, Westerman M. DNA phylogeny of the marsupial wolf resolved. Proc Biol Sci 1997, 264:911-917.

70.  Huynen L, Durand EY, Malaspinas AS, Jensen JD, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Wilson AC, Hedges E, Green RE, Krause J, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkann C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Tallermo S, Shunkov MV, Derevianko AP, Hübner JJ, Kelso J, Stankiewicz SB, Paabo S. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 2009, 325:318-319.

71.  Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Paabo S. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 2010, 464:894-897.

72.  Reich D, Green RE, Kitcher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkann C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Tallermo S, Shunkov MV, Derevianko AP, Hübner JJ, Kelso J, Stankiewicz SB, Paabo S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 2010, 468:1053-1060.

73.  Hammer MP, Woerner AE, Mendez FL, Watkins JC, Wall JD. Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci USA 2011, 108:15123-15128.

74.  Perl P, Ivanov PL, Kimpson C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K. Identification of the remains of the Romanov family by DNA analysis. Nat Genet 1994, 6:130-135.

75.  He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Veluculescu VE, Diaz LA Jr, Kinzler KW, Vogelstein B, Papadopoulos N. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 2010, 464:610-614.

76.  Bandelt HJ, Salas A. Current next generation sequencing technology may not meet forensic standards. Forensic Sci Int Genet 2011.

77.  Xue Y, Tyler-Smith C. The hare and the tortoise: one small step for four SNPs, one giant leap for SNP-kind. Forensic Sci Int Genet 2010, 4:59-61.

78.  Xue Y, Wang Q, Long Q, Ng BL, Seward H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y, MacArthur DG, Quail MA, Carter NP, Yang H, Tyler-Smith C. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol 2009, 19:1453-1457.

79.  Ballantyne KN, Goebelmoed M, Fang R, Schapa O, Lao D, Wollstein A, Tao Y, van Dijk J, Vermuelen M, Brauer S, Decorte R, Poetsch M, von Wumbrand-Schwarz N, de Knijff P, Labuda D, Vezina H, Knoblauch H, Lessig R, Roever L, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet 2010, 87:341-353.

80.  Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet 2011, 12:179-192.

81.  Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Marroco JP, Xu Z, Rodos M, Bhatacharjee A, Brzulka Z, Alberts WR, de la Rasilla M, Forteza J, Rosas A, Lachmann M, Hannon GJ, Paabo S. Targeted investigation of the Neandertal genome by array-based sequence capture. Science 2010, 328:723-725.

82.  Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch JD, Smith KL, Schupp JM, Solomon D, Kirmi P, Fraser CM. Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. Science 2003, 296:2028-2033.

83.  Ensrin L. Anthrax investigation. Full-genome sequencing paved the way from spores to a suspect. Science 2008, 321:898-899.

84.  Cummings CA, Relman DA. Genomics and microbiology. Microbial forensics—“cross-examining pathogens”. Science 2002, 296:1976-1979.

85.  Cummings CA, Bormann Chung CA, Fang R, Barker M, Brzoska P, Williamson PC, Beaudry J, Matthews M, Schupp J, Wagner DM, Birdsell D, Vogler AJ, Furtado MR, Kirmi P, Budowiele A. Accurate, rapid and high-throughput detection of strain-specific polymorphisms in Bacillus anthracis and Yersinia pestis by next-generation sequencing. Investig Genet 2010, 1:5.

Cite this article as: Berglund et al. Next-generation sequencing technologies and applications for human genetic history and forensics. Investigative Genetics 2011 2:23.