Mamm Genome (2009) 20:327–338
DOI 10.1007/s00335-009-9187-4

REVIEW

Next-generation sequencing of vertebrate experimental organisms

Daniel J. Turner · Thomas M. Keane · Ian Sudbery · David J. Adams

Received: 25 February 2009 / Accepted: 21 April 2009 / Published online: 19 May 2009

Abstract Next-generation sequencing technologies are revolutionizing biology by allowing for genome-wide transcription factor binding-site profiling, transcriptome sequencing, and more recently, whole-genome resequencing. While it is currently not possible to generate complete *de novo* assemblies of higher-vertebrate genomes using next-generation sequencing, improvements in sequence read lengths and throughput, coupled with new assembly algorithms for large data sets, will soon make this a reality. These developments will in turn spawn a revolution in how genomic data are used to understand genetics and how model organisms are used for disease gene discovery. This review provides an overview of the current next-generation sequencing platforms and the newest computational tools for the analysis of next-generation sequencing data. We also describe how next-generation sequencing may be applied in the context of vertebrate model organism genetics.

Introduction

When the Sanger (Sanger and Coulson 1975; Sanger et al. 1977) and Gilbert labs (Maxam and Gilbert 1977) first developed DNA sequencing, they would have been unlikely to have predicted the revolution that has ensued. Nowadays, rather than a sequencing experiment generating a single DNA sequence read of modest length taking days, millions of sequence reads, each several hundred base pairs in length, can be generated in a single experiment. From modest beginnings, with the first experimental organism sequenced being the phage ϕX174 (Sanger et al. 1978), progress has been inexorable with the sequence of viruses, including the human cytomegalovirus, following shortly afterward (Kouzarides et al. 1983, 1987), then numerous bacteria, with the human genome (Lander et al. 2001) and then the mouse (Waterston et al. 2002) being the first vertebrate genomes sequenced. The sequence of rat followed some years later (Gibbs et al. 2004). The Ensembl genome browser (Hubbard et al. 2009) now displays annotated genomes for 41 vertebrates. Apart from the human and mouse genomes, which were sequenced as part of an international consortium involving many sequencing centres, the majority of these genomes were sequenced by the Broad Institutes' Mammalian Genome Project (www.broad.mit.edu/node/296). The genomes displayed in Ensembl vary greatly in their quality and coverage, with many sequenced to just 2× coverage. Although these genomes represent a fabulous resource for comparative analysis, to make them a universal resource and to maximise their utility complete genome sequences are needed. In addition, to fully understand genome function and evolution, the complete sequence of multiple individuals or strains within a species will be required. In humans such an endeavour has already commenced, first with the publication of the complete genomes of four individuals (Bentley et al. 2008; Levy et al. 2007; Wang et al. 2008; Wheeler et al. 2008), and now with the 1000 genomes project (www.1000genomes.org), which is using next-generation sequencing to generate a high-resolution profile of genomic variation. Furthermore, The Mouse Genomes Project (www.sanger.ac.uk/modelorgs/mousegenomes) is in the process of sequencing the genomes of 17 key mouse strains. Indeed, while comparing genomes across the animal kingdom is a powerful way of

D. J. Turner · T. M. Keane · I. Sudbery · D. J. Adams (✉)
Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK
e-mail: da1@sanger.ac.uk

identifying conserved and divergent DNA sequences, aiding the identification of functionally relevant genomic regions, the evolutionary distance between organisms can make such comparisons difficult to interpret, particularly for subtle or quantitative phenotypes. This makes the generation and collation of sequence data from individuals or strains within a species desirable. Next-generation sequencing technologies are revolutionizing these sequence-gathering efforts and helping to obtain reference sequence for additional species such as Gorilla and Mammoth (Miller et al. 2008).

Significant efforts such as the collaborative cross (Churchill et al. 2004) and the heterogeneous stock cross (Solberg et al. 2006; Valdar et al. 2006) have been undertaken to generate genetically diverse mouse resources for complex trait analysis. Similar experiments are underway in rat (Aitman et al. 2008; Johannesson et al. 2009). The complete sequence of the progenitor strains of these experiments will be critical if we are to understand the molecular basis of the phenotypes that they reveal. Similarly, many labs have observed the partial penetrance of phenotypes when knockout alleles are bred onto different genetic backgrounds, and they are mapping these modifiers (Nadeau 2003). Similarly, numerous research groups have identified quantitative trait loci (QTLs) in mouse and rat that they wish to define. The sequence of a wide collection of mouse and rat strain genomes therefore will underpin the work of large genetic screens and also efforts ongoing in many mouse and rat labs throughout the world.

The mouse strain sequenced by the international mouse genome sequencing consortium was C57BL/6J (Waterston et al. 2002), which plays a central role in mouse genetics as founding stock for the generation of transgenic and knockout animals (Adams and van der Weyden 2008), as one of the eight strains used in the collaborative cross (Churchill et al. 2004), as a progenitor strain of the heterogeneous stock cross (Valdar et al. 2006), and as part of the mouse phenome project (Bogue and Grubb 2004; Bogue et al. 2007). In addition to the sequence of C57BL/6J, there are two large resources for the genomic sequence of inbred mouse strains. First, four laboratory strains were included by Celera in a whole-genome shotgun sequence of the mouse: A/J, DBA/2J, 129X1/SvJ, and 129S1/SvImJ (Marshall 2001). The data consist of 27.4 million capillary sequencing reads for a total of 5.3× coverage of the mouse genome. Sequences are from both ends of size-selected 2-, 10-, and 50-kb clones derived from randomly sheared mouse genomic DNA. Second, the National Institute of Environmental Health Sciences contracted Perlegen Sciences to resequence by hybridization 15 mouse inbred strains (Frazer et al. 2007). This set includes 11 classical strains (129S1/SvImJ, A/J, AKR/J, BALB/cBy, C3H/HeJ, DBA/2J, FVB/NJ, NOD/LtJ, BTBR T^{+}tf/J, KK/HlJ, and NZW/LacJ) and four strains derived from the wild (WSB/EiJ, PWD/PhJ, CAST/EiJ, and MOLF/EiJ), which represent the *M. m. domesticus, M. m. musculus*, *M. m. castaneus*, and *M. m. molossinus* subspecies. Unlike the Celera resource, the hybridization approach used by Perlegen does not generate sequence reads and can reliably detect only single nucleotide polymorphisms. Furthermore, the hybridization technology used by Perlegen queried only 1.49 billion bases of the reference genome (about 58% of non-repetitive sequence). The Perlegen approach was also found to have a false-negative rate as high as 50% (Yang et al. 2007). Celera did not generate enough sequence of any one strain sufficient for the generation of a *de novo* assembly and so only a hybrid assembly was generated, although their data have proved extremely useful for nucleotide variant discovery (Cunningham et al. 2006). Therefore, current resources lack the coverage and breadth of strains to make them a universal resource. The situation is similar for other model organisms, where many groups have initiated programmes to profile the nucleotide and structural variation between strains using technology such as array comparative hybridization, light shotgun sequencing, or sequencing by hybridization. These approaches, however, all come with compromises either because they require probes to be designed against a reference genome and are therefore unable to take into account novel nonreference sequence, or because they quickly become prohibitively expensive. In the case of hybridization-based approaches, there are limitations imposed because of genome complexity and repetitiveness. New sequencing technologies will play a vital role in deciphering the complete genomes of mouse and rat strains as well as the genomes of many other experimental organisms.

The goals for genome resequencing studies should include the following: (1) to identify nucleotide level variation between a reference and each strain, (2) to profile copy number variation between the reference and each strain, (3) to identify sequence that is unique to each strain, and (4) to ultimately generate complete *de novo* assemblies. Significant added value will be derived from comprehensive transcriptomics and the profiling of transcription factor binding sites. Collectively, these data will facilitate a systems biology approach to the study of phenotypes in model organisms providing us with unprecedented power to understand the genetic basis of traits. The technologies for achieving these goals are the focus of this review.

"Next-generation" sequencing platforms

Unlike capillary sequencing, in which a single sequencing reaction generates a single DNA sequence, next-generation sequencing generates hundreds of thousands of sequencing reactions in parallel. This allows vastly increased throughput and yield of data, enabling us to design genome-wide and

ultradeep sequencing projects that would not otherwise be possible because of their large size. In this section we discuss the various technologies that are available for next-generation sequencing. A summary of the various sequencing platforms and their throughput is shown in Table 1.

Roche 454

In 2005, the first of the next-generation DNA sequencers, the 454 GS20 (now Roche 454), became commercially available (Margulies et al. 2005). The Roche 454 uses bead-based emulsion polymerase chain reaction (em-PCR) to amplify copies of adapter-ligated template DNA molecules onto 20- or 28-μm beads, depending upon the model of the sequencer (Dressman et al. 2003). The ratio of template DNA to beads in the em-PCR is chosen to ensure that the majority of amplified beads become surrounded by amplicons derived from single template molecules. After em-PCR, hundreds of thousands of amplified beads are recovered and are deposited onto a PicoTiterPlate (PTP), which is a solid surface containing wells into which single beads can fit, along with packing beads and enzyme beads.

All amplified beads on the PTP are sequenced in parallel by pyrosequencing (Marsh 2007). In this reaction, nucleotides are flowed sequentially, in a fixed order, across the PTP. When a nucleotide that is complementary to the template strand flows across the PTP and enters a well, the polymerase incorporates that nucleotide, extending the existing DNA strand. The nucleotides do not possess blocking groups; thus, if the template strand contains two adjacent Ts, for example, two As are incorporated into the growing strand, so incorporation is asynchronous—strands extend at different rates. Nucleotide incorporation is accompanied by the release of pyrophosphate, which is used to generate a burst of light, the intensity of which is proportional to the number of nucleotides incorporated. This continues for a set number of cycles, and the signal intensity per nucleotide flow is recorded for each bead over time and is analysed to generate high-quality sequence.

In this way, throughput is increased and cost is reduced, compared to capillary sequencing, and cloning is avoided. The original model, the GS20, was capable of generating 20 megabases (Mb) of 100 base reads per run compared to fewer than 100 kb for a 96-well capillary machine, and the output has increased to several hundred megabases of 400–500 base reads per run with the Titanium version of the Roche 454 platform. A single Roche 454 sequencing run can thus generate sufficient data for many projects, particularly for the *de novo* assembly of bacterial genomes.

The major error mode with the use of pyrosequencing is that of sequencing homopolymers. The difference in signal intensity from the incorporation of, for example, eight of the same nucleotide compared with seven of the same

Table 1 Comparison of second-generation sequencing technologies

Sequencing platform	Sample requirements	Length of library prep/feature generation (days)	Method of feature generation	Sequencing chemistry	Read length (bases)	Run time	Throughput/run (Gb)	Throughput/day (Gb)
Roche 454 (FLX-Titanium)	1 μg for shotgun library, 5 μg for paired end	3–4	Bead-based/emulsion PCR	Pyrosequencing	400–500	10 h	0.4–0.5	~ 1
Illumina Genome Analyzer (GAII)	<1 μg for single or paired-end libraries	2	Isothermal 'bridge amplification' on flowcell surface	Reversible terminator SBS	35–75	2 days for 36-cycle single-end run, 4 days for 36-cycle paired-end run	3–6	1.5
ABI SOLiD	<2 μg for shotgun library, 5–20 μg for paired end	2–4.5	Bead-based/emulsion PCR	Ligation	25–75	6–7 days for fragment libraries, 8 days for 2 × 25 base paired-end libraries	10–20	1.7–2
Helicos tSMS	<2 μg, single end only	1	N/A (single molecule sequencing)	Virtual terminator SBS	25–50	8–9 days	21–28	2.5

nucleotide is much less than two versus one of the same nucleotide. This is compounded by the fact that there is some, albeit small, variation in the signal intensity between single incorporation events. In addition, the current cost of 454 sequencing per gigabase is considerably higher than short-read sequencing technologies, so short-read technologies tend to be used preferentially for sequencing applications that do not depend upon long reads.

Applied Biosystems' SOLiD

The SOLiD was released in late 2007 and like the 454 relies on em-PCR to amplify fragmented DNA onto beads clonally. After em-PCR, amplified beads are recovered and the amplicon strands are modified at their 3′ ends to allow covalent attachment to a glass slide (Dressman et al. 2003). However, with the SOLiD beads are considerably smaller than with the 454, i.e., 1 μm rather than 28 μm, which allows a higher density of beads to be packed into the same area. The current density range is in the region of 100 million beads per sequencing run, although a large number of these beads are not analysed because they have more than one template amplified onto them giving a "mixed read" which is filtered out in the analysis.

Rather than using pyrosequencing, the SOLiD platform uses sequencing-by-ligation (Shendure et al. 2005). After hybridization of the sequencing primer to bead-bound amplicons, 16 random 8-mer probes are added. Probes are labelled using one of four fluorescent dyes, which are assigned based on the nucleotides at the first and second positions at the 3′ end of the probe. The first and second 3′ bases of one of the 16 probes will be complementary to the template strands around a bead. This oligo ligates to the sequencing primer and the slide is imaged. The probe is cleaved, removing the fluorescent label but leaving five bases of probe ligated to the sequencing primer, and the random probe set is added once more. Several rounds of ligation and imaging generate a colour profile of every fifth dinucleotide.

The extended sequencing primer is removed, and a second sequencing primer, one nucleotide shorter than the first, is hybridized; ligation proceeds for the same number of rounds. The process is repeated using a total of five sequencing primers, at the end of which a series of colours is obtained for each bead. Because only four colours are used and because each colour represents four dinucleotides, it is not possible to decipher the identity of the nucleotides without knowing the first base in the sequence. This is achieved by sequencing one base of the adapter.

The SOLiD is currently capable of producing approximately 20 Gb of short-read sequence data per run (25–50 bases) and so is more suited to resequencing than *de novo* assembly, although optimized protocols for long-insert read pairs up to 10 kb are available.

Illumina Genome Analyzer

Solexa sequencers first became available for beta testing in late 2006. They were rebranded as Illumina Genome Analyzers (GA) before their wider release following Solexa's takeover. Unlike the two next-generation sequencing platforms described above, the GA does not rely on em-PCR to clonally amplify template strands. Instead, adapter ligated template molecules flow into a hollow glass slide, or flow cell, at a low concentration using a fluidic pumping device termed a cluster station. The interior surfaces of the flow cells are coated with polyacrylamide to which a random "lawn" of forward and reverse primers is attached. Template DNA hybridizes to the primers and is copied onto the flow-cell surface by extension of the flow-cell primer to which it hybridized. This generates a reverse complementary copy of the template strand that is covalently attached to the flow-cell surface. These newly synthesized strands serve as templates for an isothermal amplification reaction, resulting in clusters of amplified strands, each of which was derived from a single template molecule and is immobile.

Amplified clusters consist of double-stranded DNA. One strand is selectively removed before sequencing primer hybridization and the sequencing reaction itself, so that all strands within each cluster are sequenced in the same direction, from the same end. The flow cell is then transferred to a Genome Analyzer, where the single-stranded clusters undergo a sequencing-by-synthesis reaction using reversible fluorescent terminator deoxyribonucleotides (Bentley et al. 2008). Being terminator nucleotides, each DNA strand within a cluster incorporates a single nucleotide during each chemistry cycle, and being clonal, each strand within a cluster incorporates the same nucleotide. Clusters are imaged, blocking groups and fluorophores on the newly incorporated nucleotides are removed simultaneously by chemical cleavage, and the next round of nucleotide incorporation begins. Sequence length is identical for all clusters because it is governed by the number of cycles of nucleotide incorporation, imaging, and cleavage. Images are analysed, generating a separate sequence for each cluster.

An Illumina Genome Analyzer is currently capable of producing an up to 10-Gb "purity-filtered" sequence per 76-cycle paired-end run. Beyond this length, the frequency of substitution errors currently increases significantly as a result of signal decay and cluster phasing and prephasing. Long insert protocols are available, as are array- and solution-based targeted resequencing and multiplexing protocols.

Helicos' true single-molecule-sequencing (tSMS) technology

Launched in 2008, Helicos' tSMS sequencing platform is the first of what could be considered the next-next

generation (or 3rd generation) of sequencing platform and is based on a technology that was published in 2003 (Braslavsky et al. 2003). Sequencing takes place on millions of templates in parallel, but tSMS differs from other currently available sequencing technologies. Whereas Illumina amplifies single template molecules to make clusters, and 454 and SOLiD use em-PCR to amplify copies of a single template molecule onto a bead, tSMS does not amplify templates in any way before sequencing. As a consequence, library preparation is simple and rapid, requiring only the addition of a poly-A tail and a fluorescent label. Tailed template strands hybridize to poly-T oligonucleotides on the flow-cell surface, and these single molecules are detected by their fluorescent label.

Because single molecules are the substrate for the sequencing reaction, the flow cell can be packed to a very high density, so billions of strands can potentially fit onto a single flow cell. Fluorescent nucleotides are added singly. These are not terminator nucleotides as such, but virtual terminators that rely upon steric hindrance to deter the incorporation of more than one nucleotide per cycle. After incorporation the flow cell is visualized to identify strands that incorporated that particular nucleotide, the fluorescence is removed, and the next nucleotide is added. Strands incorporate nucleotides in an asynchronous way, the rate of extension being governed by how the sequence of a particular template strand corresponds to the order in which nucleotides are added. This results in sequences that differ in length but which are typically 25–50 bases.

Sequencing single molecules avoids the problems of phasing encountered by 454, Illumina, and SOLiD, where some members of the group of templates being sequenced do not incorporate a nucleotide at a given cycle and so lag behind the others. There is relatively little information available about the error rate of the tSMS platform, but it is conceivable that sequencing single molecules causes problems with sensitivity. Reported causes of error are problems with long homopolymers (particularly runs of poly-C) and a high incidence of deletions. These can be reduced from 2 to 7% to below 1% by reading the same strand twice, but this is achieved at the expense of doubling the running time and increasing the length of the library prep (Harris et al. 2008). There is currently no protocol available for performing paired-end sequencing on the tSMS platform.

Emerging sequencing technologies

Pacific Biosciences has recently showed promising early results using single-molecule real-time DNA sequencing (SMRT technology) (Eid et al. 2009). There are two key underlying proprietary technologies: (1) phospholinked nucleotides, where each nucleotide is labeled with a different fluorophore that is attached to the γ-phosphate. The fluorophore is thus removed upon incorporation of the base by a polymerase. (2) "Zero mode waveguides" enable individual molecules to be visualised without noise from the background of unincorporated nucleotides.

The polymerase pauses as incorporation occurs, during which time the fluorophore attached to the incorporated nucleotide becomes excited and emits fluorescence. The technology has the potential to produce very long reads at a rate of around 10 bases per second, with thousands of reactions proceeding in parallel. Instruments are expected to be available in 2010.

Dover Systems' Polonator was announced early 2008 and arose from collaboration between George Church's laboratory at MIT and the Danaher Corporation. Perhaps the most appealing aspect of this platform is that it is "open source," thus users are free to purchase reagents from any supplier allowing flexibility since users are not committed to using a single sequencing chemistry, although the Polonator was developed using bead-based em-PCR (Dressman et al. 2003) and sequencing by ligation (Shendure et al. 2005).

The instrument has an excellent potential throughput rate of approximately 3 Gb/day, though read lengths are currently short (2× 14-base paired end reads), which will make mapping sequence reads to a vertebrate-size genome difficult. It should be inexpensive to run and currently has a list price that is considerably lower than the other systems.

Oxford Nanopore Technologies is developing a label-free, single-molecule sequencing technology called BASE in which a processive exonuclease enzyme and a α-haemolysin nanopore are set into a lipid bilayer that lies above a microwell (Howorka et al. 2001; Maglia et al. 2008). Many of these wells are arrayed on a silicon chip. DNA is digested by the exonuclease, after which the released nucleotides pass through the nanopore. Transient binding of nucleotides to a cyclodextrin ring within the nanopore generates a change in the conductivity across the pore, which is characteristic of that nucleotide.

Once commercialized, the technology will be marketed, sold, and distributed exclusively by Illumina, which has also made a major equity investment in the company. Other nanopore-base sequencing technologies are being developed by NABsys and Sequenom.

Intelligent Bio-systems was founded by Jingyue Ju of Columbia University and is based on a proprietary sequencing-by-synthesis technology, which utilizes four fluorescent reversible terminator nucleotides (Ju et al. 2006). No instrument is available yet, but the company claims that its technology will allow millions of sequencing reactions to take place in parallel, with high accuracy and speed and low cost, albeit on amplified DNA.

VisiGen Biotechnologies is developing a real-time, single-molecule sequencing technology based on fluorescence resonance energy transfer (FRET) between an

immobilized, fluorescently labelled polymerase and fluorescent nucleotides. Massively parallel arrays of these polymerases would allow very high sequencing rates.

More distant sequencing technologies are under development by Affymetrix, Reveo, Base4innovation, Genome Corp, and Complete Genomics.

Computational analysis of new technology sequencing data

As illustrated above, next-generation sequencing technologies are capable of generating vast quantities of data, and with impending release of next-next or third-generation sequencing technology, the yield of data is set to skyrocket. Next we outline the currently available computational tools that may be used for analysis of sequencing data generated on these platforms.

Mapping sequencing reads to the genome

By aligning short-read sequences back to a reference genome we can detect a range of different types of sequence variation, including single nucleotide polymorphisms (SNPs), short insertion/deletions (indels), and structural and copy number variants. The single most important task in variant discovery using short reads is to align the individual reads (or read pairs) with the correct location on the reference genome. The ability to map short sequence reads to the correct location is dependent on a number of factors such as the complexity of the reference genome, length of the sequence reads, error rates of the reads, and the diversity of the individual or strain compared to the reference (Li et al. 2008).

For small genomes such as closely related bacterial strains, the task of aligning reads is relatively easy because even with very short sequence read lengths, a very high proportion of the reads will align to only one location due to the uniqueness of the reference (Fig. 1). At read lengths of 30 bp, only 85% of the mouse reference genome is unique enough to call high confidence variants with error-free reads. The presence of read pairs can increase the accuracy of read alignment because "mate pair" information can be used to place reads where only one of the mates aligns with high confidence to the genome (often referred to as the mapping quality) but where the read or "mate" at the other end of the sequenced molecule maps to several possible genomic locations. One additional factor that can decrease the accuracy of read alignment is incorrect base calls due to sequencing errors that can result in reads being more similar to the wrong location on the genome. To overcome these problems, most short-read mapping programs use some combination of base quality scores,

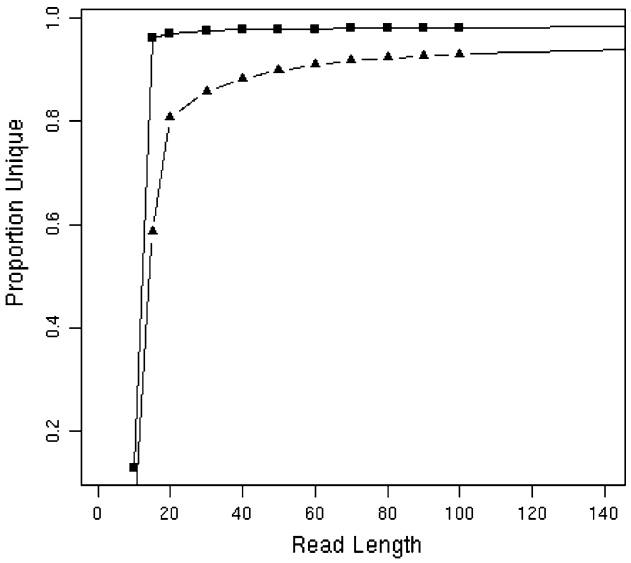

Fig. 1 The proportion of unique sequence in the *Streptococcus suis* (squares) and *Mus musculus* (triangles) genomes for varying read lengths. This graph indicates that read length has a critical affect on the ability to place reads uniquely to the genome

mapping quality of the reads, and the number of reads, calling the variant at each position in the genome to assess the quality of a SNP call. To meet these unique challenges, a number of new tools have been developed specifically to align short-read sequences to genomes (Table 2).

MAQ was one of the early short-read alignment tools (Li et al. 2008). MAQ uses a k-mer hash table approach for indexing sequence reads and chooses the place in the genome where the read aligns with the minimum sum of the base qualities of the mismatched bases. This helps to overcome the problem of erroneous base calls, making a read more similar to the wrong location on the genome. MAQ attempts to align all of the reads regardless of whether they fall into repeat regions of the genome. Clearly, this could lead to false-positive variant calls where a read maps to multiple places equally well on the genome. To overcome this, MAQ assigns a phred-like (Stein 2003) mapping quality to each read alignment, which is related to the confidence in the alignment of the read. Importantly, MAQ is very quick to map large numbers of sequence reads. MAQ gets its speed primarily by using a hybrid approach to aligning reads. It first tries to match the reads by a simple ungapped alignment and carries only a Smith-Waterman alignment (an accurate but slow algorithm for producing local alignments) on unmapped reads where the mate is already mapped. Another mapping application specifically developed for short reads is SHRiMP. SHRiMP, developed at the University of Toronto, is a more general-purpose alignment tool as it carries out seeded Smith-Waterman alignments and can be used to align reads of any length or type. However, it does not calculate

Table 2 A summary of short-read alignment tools

	Illumina	454	SOLiD	S	I	URL
Bowtie	Y	Y	N	Y	N	http://bowtie-bio.sourceforge.net
ELAND	Y	N	N	N	N	http://www.illumina.com
Exonerate	Y	Y	N	N	Y	http://www.ebi.ac.uk/~guy/exonerate/
GMAP	Y	N	N	N	N	http://www.gene.com/share/gmap
MOSAIK	Y	Y	Y	Y	Y	http://bioinformatics.bc.edu/marthlab/Mosaik
MAQ	Y	N	Y	Y	Y	http://maq.sourceforge.net
MUMer	Y	Y	N	Y	Y	http://mummer.sourceforge.net/
Novocraft	Y	N	N	Y	Y	http://www.novocraft.com/
RMAP	Y	N	N	N	N	http://rulai.cshl.edu/rmap/
SeqMap	Y	N	N	N	I	http://biogibbs.stanford.edu/~jiangh/SeqMap/
SHRiMP	Y	Y	Y	Y	Y	http://compbio.cs.toronto.edu/shrimp/
SOAP	Y	N	N	Y	Y	http://soap.genomics.org.cn/
SSAHA2	Y	Y	N	Y	Y	http://www.sanger.ac.uk/Software

S outputs SNPs, I outputs short insertion deletions (indels)

mapping qualities or use paired-end information during the alignment process, a limitation of this tool. Mosaik, developed by Gabor Marth's lab at Boston College, is one of the most general-purpose short-read tools. It can align and assemble reads generated from all sequencing platforms, along with legacy Sanger reads, and importantly it is proficient at detecting short indels, an important function lacking in other software tools.

Each of the sequencing machine vendors has also produced its own mapping tools. Newbler, from 454, is capable of read mapping and sequence assembly. One of the benefits of 454 technology is the increased read length compared with that of other platforms, which makes reliable read mapping easier. Also, Newbler is specifically designed to handle these longer reads. ELAND is the short-read alignment tool developed by Illumina for use with its GA platform and is included in the processing pipeline with the instrument. ELAND is extremely fast to run and outputs all of the alternative places a read can be mapped onto the genome. One drawback is that it is difficult to run ELAND independent of the Illumina GAPipeline.

Bowtie is the first of a new generation of the short-read aligners that use the Burrows-Wheeler transform, which is an indexing system with a very low memory footprint (Burrows and Wheeler 1994). Bowtie is considerably faster than other short-read alignment tools. However, it is only effective when the reads are extremely similar to the reference. One useful feature of Bowtie is that its output can be imported into MAQ and hence one can utilise the variant calling tools included with MAQ. This method has also been implemented in BWA, a new developmental alignment tool (http://maq.sourceforge.net/bwa-man.shtml) from the same authors as MAQ.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the simplest type of sequence variation to profile since in an alignment they appear as high-quality single base differences between sequence reads and the reference genome. A number of papers have demonstrated very high SNP calling accuracy from new sequencing technologies (Hillier et al. 2008; Li et al. 2008; Wang et al. 2008). Because of the variations in sequence quality across individual runs and across technologies, most studies use another method (e.g., genotyping) to calibrate the SNP-calling cutoffs (e.g., mapping quality, minimum depth, cumulative base quality). For example, Hillier et al. (2008) used the Illumina platform to sequence a *C. elegans* strain and obtained a SNP validation rate of 96.3% using PCR-targeted capillary sequencing on a subset of the total SNPs found. When critically reviewing claims about the accuracy of SNP detection algorithms, it is extremely important to consider where in the genome the test or candidate SNPs are located. SNPs located in non-repetitive regions of the genome are generally considerably easier to validate than SNP located in more complex regions.

Short insertion/deletions (indels)

To detect short indels, a short-read alignment tool must be able to carry out gapped alignments. As seen in Table 2, the majority of short-read alignment tools have the ability to detect short indels. Since indels are harder to validate than SNPs, few of the indel calling algorithms have been rigorously validated.

Structural variation

Almost all of the new sequencing technologies are capable of generating paired-end reads with varying insert sizes. Currently, the Illumina Genome Analyzer and the Roche FLX can produce libraries with insert sizes up to 3 kb, while the SOLiD system can generate insert sizes up to 10 kb. However, it should be noted that the exact insert size is very library dependent and it is often necessary to "discover" the exact insert size of the library by aligning the reads to the reference genome. A number of recent studies have used insert size discrepancies to discover high-quality structural variants (e.g., large insertions, deletions, and inversions). Campbell et al. (2008) recently showed how it is possible to use Illumina sequencing to identify useful information from incorrectly aligned read pairs to identify somatically acquired structural rearrangements in lung cancer. Another study used 454 paired-end sequencing to find structural variants (SVs) in the human genome and was able to experimentally validate a high number of the SVs found (Korbel et al. 2007). A number of tools have been developed to visualize SVs. For example, MAQ and Mosaik come with viewing applications that show the alignment of the reads on the reference genome, allowing them to be easily visualized (Huang and Marth 2008).

Sequence assembly

There are two main types of sequence assembly from new sequencing technology reads. In the first case, the reads are mapped back to the reference genome and a consensus sequence is generated by calling a base at each position where reads have mapped along the reference (referred to as a mapped assembly). This is generally not regarded as a true *de novo* assembly because the resulting assembly could be structurally biased by the reference genome. However, this bias can be reduced by using the mate-pair information to confirm the structure of the contigs. MAQ, Mosaik, and Newbler are capable of carrying out mapped based assemblies.

One of the major challenges of this short-read sequencing era is the quest for software that can generate true *de novo* assemblies of a vertebrate genome. The algorithmic and computational challenges posed by this problem have led to the development of several new assemblers (Table 3). The huge volumes of data and the small size of the sequence reads generated on new-technology sequencing platforms has meant that traditional approaches of sequence assembly, such as computing all possible read overlaps (e.g., phrap) (Stein 2003), are not computationally feasible. Almost all of the new assembly algorithms utilise some form of a graph traversal approach such as the de Bruijn graph (Pevzner et al. 2001). One measure of the performance of any assembly algorithm is the N50. The N50 of an assembly is the length where 50% of bases in the assembly are found in contigs with at least this length.

Velvet is a *de novo* assembler based on the de Bruijn graph approach (Zerbino and Birney 2008) and has been used to assemble BACs with an N50 of 2 kb. Butler et al. (2008) tested their assembler ALLPATHS by generating simulated reads from ten finished genomes ranging from bacteria, fungi, and a 10-Mb section of the human genome. They achieved impressive N50s with some of the genomes assembling into a single contig. However, the lack of real sequence data in this study makes it difficult to assess this assembler's true performance (Butler et al. 2008).

Most *de novo* assemblers perform well on bacteria and small eukaryotes but the challenge is on to develop an assembler that can handle a higher-vertebrate genome. At the moment, Abyss is the only assembler capable of assembling vertebrate-sized genomes (http://www.bcgsc.ca/platform/bioinfo/software/abyss). It achieves this by distributing and processing the de Bruijn graph over a computer cluster and therefore requires considerable computational resources. It is expected that as read lengths

Table 3 A summary of assemblers developed for use with next-generation sequencing data

	Illumina	454	SOLiD	D	M	E	URL
Abyss	Y	N	N	Y	N	Y	http://www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATHS	Y	N	N	Y	N	N	
MAQ	Y	N	Y	N	Y	N	http://maq.sourceforge.net
MIRA2	N	Y	N	Y	N	N	http://chevreux.org/projects_mira.html
Newbler	N	Y	N	Y	Y	N	https://www.roche-applied-science.com
SSAKE	Y	N	N	Y	N	N	http://www.bcgsc.ca/platform/bioinfo/software/ssake
SHARCGS	Y	N	N	Y	N	N	http://sharcgs.molgen.mpg.de/
VCAKE	Y	N	Y	Y	N	N	http://sourceforge.net/projects/vcake
Velvet	Y	N	N	Y	N	N	http://www.ebi.ac.uk/~zerbino/velvet/

D capable of *de novo* assembly, *M* mapped assembly, *E* can assemble experimental organisms

increase, it will become more feasible to perform whole-genome assemblies of experimental organisms.

Applications to mammalian genome sequencing

Although next-generation sequencing technology is still novel, particularly with respect to the tool necessary to analyse the data, strides have already been made using it. A total of four complete diploid genomes of human individuals have been sequenced to data. The first was achieved using traditional Sanger/capillary sequencing methods (Levy et al. 2007). The second of these genomes was that of the scientist James Watson, which was sequenced using the Roche 454 technology to 7.5× genome coverage (Wheeler et al. 2008). The reads were aligned to the NCBI reference sequence using a combination of the BLAT and Smith-Waterman algorithms (Wheeler et al. 2008). The sequence differed from the reference at 3.32 Mb, of which 2.7 Mb were known differences. The sequences of the other two human genomes, that of a Chinese individual and an African, were done using the Illumina Genome Analyzer platform and sequenced to greater than 30× genome coverage (Bentley et al. 2008; Wang et al. 2008). For both genomes reads were aligned to the NCBI human reference sequence revealing approximately 3 million SNPs. In both cases around 74% of these SNPs were previously known. The accuracy of these studies, as assessed by array-based genotyping, was similar to that of the Watson Roche 454 genome. In each of the four genomes sequenced, novel "nonreference" sequence was discovered.

While next-generation sequencing technology has shown itself to be useful for identifying variations between individuals, it has also been useful for decoding novel genomes. The latest assembly of the gorilla genome sequence is a hybrid between sequence data derived from traditional capillary read sequencing and next-generation sequencing. First, the genome was sequenced to 2× genome coverage using a traditional capillary shotgun sequencing approach, then Illumina reads were added to facilitate the identification of novel sequences and the stitching together of contigs. This approach has produced a cleaner more accurate sequence compared to other 2× coverage genomes (http://www.ensembl.org/Gorilla_gorilla/Info/Index). Higher Illumina sequence read coverage is currently being generated to produce a high-quality *de novo* assembly of gorilla. Another interesting application of next-generation sequencing has been the sequencing of nearly 3 Gb of the mammoth genome (Miller et al. 2008). Ancient DNA samples were extracted from hair shafts, sequenced using Roche 454 technology, and aligned to the genome of a modern elephant. It should be noted that in all these applications, sequence generated by next-generation technology has been aligned to some reference or scaffold generated using capillary sequence. These mapped to reference assemblies vary greatly in there coverage and quality and by definition are blind to most novel sequences.

Next-generation sequencing technology as a replacement for microarrays

Since the advent of high-throughput DNA microarrays, it has been possible to interrogate levels of thousands of individual nucleotide species (such as transcripts or DNA fragments recovered from CHiP experiments) simultaneously. In principle, most applications of microarray technology can also be achieved using next-generation sequencing, as the levels of any given nucleotide species can be inferred from the number of times it is identified in a sequencing experiment. One recent application of new sequencing technologies, called RNA-Seq (Marioni et al. 2008; Mortazavi et al. 2008; Wang et al. 2009), directly sequences expressed transcripts in order to directly interrogate levels of transcription. The RNA is isolated from a particular cell type and reverse transcribed, and the resulting cDNA is subjected to next-generation sequencing. The subsequent sequencing reads are aligned back to the reference genome and the sequencing depth is used as the measure of expression levels. Unlike array-based approaches, RNA-Seq offers a complete and unbiased view of the full repertoire of transcripts (Pan et al. 2008). One of the greatest advantages is that it allows for the detection of transcripts that are expressed at very low levels because of the high sequencing depth that can be achieved with new sequencing technologies. Marioni et al. (2008) found that the method was highly replicable and had very little technical variation across different runs of the sequencer. t Hoen et al. (2008) found that the changes in expression observed by deep sequencing were larger than observed by microarrays or quantitative PCR. They were able to detect processes such as calmodulin-dependent protein kinase activity, vesicle transport along microtubules, and antisense transcription that were not observable with an array-based approach.

Challenges and future directions

Next-generation sequencing technologies afford many opportunities but will also pose considerable challenges. As indicated in this review, the major challenges of these technologies revolve around the management and analysis of the sequence data. Because every DNA sequencing run on a next-generation sequencing platform generates many gigabytes of data that must be analysed and archived, considerable computational resources are required. At present there are few institutions, besides the genome

centres, that have the computational resources, or indeed the knowledge, to handle the amount of data that these machines are capable of generating. This is certainly the case for more complex tasks such as the generation of *de novo* assemblies. There is likely to be an evolution in the software tools available for analysing sequencing data, and just as occurred with microarray analysis tools, this will place the power of new sequencing technologies more within reach of the average research lab and researcher. As indicated in this review, many of the sequencing machine vendors already provide software modules and pipelines that facilitate some of the common sequence analysis tasks. It is likely that the vendors will develop software to keep pace with the development of their machines to support read mapping, nucleotide variant calling, and the generation of *de novo* assemblies. The research community should expect and demand that these software tools be completely open source so that anyone can have access to them.

One of the great success stories of the genome project era was that all of the sequence data generated from these projects was freely available, and essentially released as it was generated. This is an expectation for all future genome projects that the community and funding bodies should enforce. Similarly, because it is likely that many groups will start to generate assemblies of vertebrate genomes using next-generation sequencing, it is important that the standard set for the release and publication of these assemblies be high. This can be managed in part by the free release of sequence data, which can in turn be reanalysed and scrutinized by the community.

Several databases have been established to collect next-generation sequencing data. These include the European Short Read Archive (ERA) (www.ebi.ac.uk/embl/Documentation/ENA-Reads.html), based at the European Bioinformatics Institute in Cambridge, and the NIH Short Read Archive (www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi), based as the National Center for Biotechnology Information, Bethesda, Maryland. These databases are likely to face significant challenges in the management and storage of data from new-technology sequencing projects as projections suggest that storage and memory technology will be pushed to the limit to accommodate all of the next-generation sequencing data that are being generated and which will require archiving. As sequencing read lengths are increasing dramatically and read numbers appear to also be increasing on most platforms, these data management and storage issues will require significant investments in computational infrastructure; this will have significant financial implications.

For mouse and rat genetics, it is a golden age. Within the next few years it is likely that we will have high-quality assemblies for many of the strains that are commonly used in our research laboratories. The sequence of these strains will finally allow us to gain a complete picture of the genetic variation between strains, and will greatly facilitate the identification of the causal variants responsible for the QTLs and modifier alleles that many of us have spent years mapping. It is also likely that there will be a renaissance in techniques such as the use of *N*-ethyl-*N*-nitrosourea (ENU) (Justice 2000; Kile et al. 2003) and other mutagens for forward genetic screens. When we reach the point of the $1,000 genome for mouse and rat, touted as a looming landmark in human genetics, finding causal nucleotide variants from mutagenesis screens in the mouse will no longer be the challenge that it currently is. Indeed, in yeast and *C. elegans* next-generation sequencing technologies are resulting in the demise of positional cloning-based approaches because it is now considerably easier to just resequence the entire genome of the mutant yeast or worm rather than map the mutation (Hillier et al. 2008; Schacherer et al. 2007). The same will soon be the case for vertebrate experimental organisms such as mouse and rat.

One thing that is clear is that in the future, when we look back on this era of new-sequencing technology development, we will wonder how we ever lived without it.

Acknowledgments DJA was supported by Cancer Research-UK and the Wellcome Trust. This project was supported by the Medical Research Council – UK and the Wellcome Trust Sanger Institute. We thank Louise van der Weyden for critically reviewing the manuscript.

References

Adams DJ, van der Weyden L (2008) Contemporary approaches for modifying the mouse genome. Physiol Genomics 34:225–238

Aitman TJ, Critser JK, Cuppen E, Dominiczak A, Fernandez-Suarez XM et al (2008) Progress and prospects in rat genetics: a community view. Nat Genet 40:516–522

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bogue MA, Grubb SC (2004) The Mouse Phenome Project. Genetica 122:71–74

Bogue MA, Grubb SC, Maddatu TP, Bult CJ (2007) Mouse Phenome Database (MPD). Nucleic Acids Res 35:D643–D649

Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 100:3960–3964

Burrows M, Wheeler D (1994) A block sorting lossless data compression algorithm. Digital Equipment Corporation Technical Report 124

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK et al (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res 18:810–820

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40:722–729

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD et al (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137

Cunningham F, Rios D, Griffiths M, Smith J, Ning Z et al (2006) TranscriptSNPView: a genome-wide catalog of mouse coding variation. Nat Genet 38:853

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA 100:8817–8822

Eid J, Fehr A, Gray J, Luong K, Lyle J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA et al (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448:1050–1053

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J et al (2008) Single-molecule DNA sequencing of a viral genome. Science 320:106–109

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G et al (2008) Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods 5:183–188

Howorka S, Cheley S, Bayley H (2001) Sequence-specific detection of individual DNA strands using engineered nanopores. Nat Biotechnol 19:636–639

Huang W, Marth G (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res 18:1538–1543

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K et al (2009) Ensembl 2009. Nucleic Acids Res 37:D690–D697

Johannesson M, Lopez-Aumatell R, Stridh P, Diez M, Tuncel J et al (2009) A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock. Genome Res 19:150–158

Ju J, Kim DH, Bi L, Meng Q, Bai X et al (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci USA 103:19635–19640

Justice MJ (2000) Capitalizing on large-scale mouse mutagenesis screens. Nat Rev Genet 1:109–115

Kile BT, Hentges KE, Clark AT, Nakamura H, Salinger AP et al (2003) Functional genetic analysis of mouse chromosome 11. Nature 425:81–86

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Kouzarides T, Bankier AT, Barrell BG (1983) Nucleotide sequence of the transforming region of human cytomegalovirus. Mol Biol Med 1:47–58

Kouzarides T, Bankier AT, Satchwell SC, Weston K, Tomlinson P et al (1987) Sequence and transcription analysis of the human cytomegalovirus DNA polymerase gene. J Virol 61:125–133

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL et al (2007) The diploid genome sequence of an individual human. PLoS Biol 5:e254

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Maglia G, Restrepo MR, Mikhailova E, Bayley H (2008) Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. Proc Natl Acad Sci USA 105:19720–19725

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517

Marsh S (2007) Pyrosequencing applications. Methods Mol Biol 373:15–24

Marshall E (2001) Genome sequencing. Celera assembles mouse genome; public labs plan new strategy. Science 292:822

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Miller W, Drautz DI, Ratan A, Pusey B, Qi J et al (2008) Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456:387–390

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628

Nadeau JH (2003) Modifier genes and protective alleles in humans and mice. Curr Opin Genet Dev 13:290–295

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA 98:9748–9753

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441–448

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG et al (1978) The nucleotide sequence of bacteriophage phiX174. J Mol Biol 125:225–246

Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D et al (2007) Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. PLoS ONE 2:e322

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Solberg LC, Valdar W, Gauguier D, Nunez G, Taylor A et al (2006) A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. Mamm Genome 17:129–146

Stein L (2003) Large scale sequencing. Curr Protoc Bioinformatics Chapter 11, Unit 11.1

t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH et al (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res 36:e141

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P et al (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 38:879–887

Wang J, Wang W, Li R, Li Y, Tian G et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F (2007) On the subspecific origin of the laboratory mouse. Nat Genet 39:1100–1107

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829