BACKGROUND: For the past 30 years, the Sanger method has been the dominant approach and gold standard for DNA sequencing. The commercial launch of the first massively parallel pyrosequencing platform in 2005 ushered in the new era of high-throughput genomic analysis now referred to as next-generation sequencing (NGS).

CONTENT: This review describes fundamental principles of commercially available NGS platforms. Although the platforms differ in their engineering configurations and sequencing chemistries, they share a technical paradigm in that sequencing of spatially separated, clonally amplified DNA templates or single DNA molecules is performed in a flow cell in a massively parallel manner. Through iterative cycles of polymerase-mediated nucleotide extensions or, in one approach, through successive oligonucleotide ligations, sequence outputs in the range of hundreds of megabases to gigabases are now obtained routinely. Highlighted in this review are the impact of NGS on basic research, bioinformatics considerations, and translation of this technology into clinical diagnostics. Also presented is a view into future technologies, including real-time single-molecule DNA sequencing and nanopore-based sequencing.

SUMMARY: In the relatively short time frame since 2005, NGS has fundamentally altered genomics research and allowed investigators to conduct experiments that were previously not technically feasible or affordable. The various technologies that constitute this new paradigm continue to evolve, and further improvements in technology robustness and process streamlining will pave the path for translation into clinical diagnostics.

© 2009 American Association for Clinical Chemistry
Sanger sequencing, which is based on the electro-
phoretic separation of chain-termination products
produced in individual sequencing reactions. In NGS,
sequencing is performed by repeated cycles of
denatured deoxyribonucleotide triphosphate (dNTP)
inclusion in a well containing clonally amplified template
products. Pyrophosphate is produced in individual sequencing reactions. In NGS,
several hundred base pairs in length are end-repaired and ligated to adapter oligonucleo-
tides. The library is then diluted to single-molecule
concentration, denatured, and hybridized to individual
beads containing sequences complementary to
to generate 5′-phosphorylated blunt ends. The poly-
merase activity of Klenow fragment is used to add a
single A base to the 3′ end of the blunt phosphorylated
DNA fragments. This addition prepares the DNA frag-
ments for ligation to oligonucleotide adapters, which
have an overhang of a single T base at their 3′ end
to increase ligation efficiency. The adapter oligonucleo-
tides are complementary to the flow-cell anchors. Un-
with sequencing enzymes. Loaded into the GS FLX, the
picotiter plate functions as a flow cell wherein iterative pyrosequencing is performed by successive flow addi-
tion of the 4 dNTPs. A nucleotide-incorporation event in
a well containing clonally amplified template pro-
duces pyrophosphate release and picotiter-plate well-
localized luminescence, which is transmitted through
the fiber-optic plate and recorded on a charge-coupled
device camera. With the flow of each dNTP reagent,
wells are imaged, analyzed for their signal-to-noise ra-
tio, filtered according to quality criteria, and subse-
quent algorithmically translated into a linear se-
quence output. With the newest chemistry, termed
"Titanium," a single GS FLX run generates approximately $1 \times 10^8$ sequence reads, with read lengths of
$\geq$400 bases yielding up to 500 million base pairs (Mb)
of sequence. A recognized strength of the 454 technol-
ogy is the longer read length, which facilitates de novo
assembly of genomes (9). An outstanding concern has
been the accurate determination of homopolymers
$>3$–4 bases in length. A 6-base homopolymer should
theoretically yield twice the luminescence of a 3-base
homopolymer. Operationally, this luminescence yield
varies, and estimates of homopolymer length are less
accurate with increasing length (8, 10). 454 has re-
ported that the metal coating of the walls of picotiter
wells mentioned above improves the accuracy of ho-
mpolymer determination. Sequence coverage depth
and accuracy for the 454 technology is discussed below
in the NGS Data Analysis section.

**ILLUMINA/SOLEXA**

In 1997, British chemists Shankar Balasubramanian
and David Klenerman conceptualized an approach for
sequencing single DNA molecules attached to micro-
spheres. They founded Solexa in 1998, and their goal
during early development of sequencing single DNA
molecules was not achieved, requiring a shift toward
sequencing clonally amplified templates. By 2006, the
Solexa Genome Analyzer, the first "short read" se-
quecing platform, was commercially launched. Ac-
quired by Illumina (http://www.Illumina.com) in
2006, the Genome Analyzer uses a flow cell consisting
of an optically transparent slide with 8 individual lanes
on the surfaces of which are bound oligonucleotide an-
chors (Fig. 2). Template DNA is fragmented into
lengths of several hundred base pairs and end-repaired
to generate 5′-phosphorylated blunt ends. The poly-
merase activity of Klenow fragment is used to add a
single A base to the 3′ end of the blunt phosphorylated
DNA fragments. This addition prepares the DNA frag-
ments for ligation to oligonucleotide adapters, which
have an overhang of a single T base at their 3′ end
to increase ligation efficiency. The adapter oligonucleo-
tides are complementary to the flow-cell anchors. Un-

**ROCHE/454 LIFE SCIENCES**
The 454 technology (http://www.454.com) is derived
from the technological convergence of pyrosequencing
and emulsion PCR. In 1993, Nyren et al. described a
sequencing approach based on chemiluminescent de-
tection of pyrophosphate released during polymerase-
mediated deoxynucleotide triphosphate (dNTP) in-
corporation (4). Refinement by Ronaghi et al. served as
the foundation for the commercial development of py-
rosequencing (5, 6). On a separate front, Tawfik and
Griffiths described single-molecule PCR in microcom-
partments consisting of water-in-oil emulsions (7). In
2000, Jonathan Rothberg founded 454 Life Sciences,
which developed the first commercially available NGS
platform, the GS 20, launched in 2005. Combining
single-molecule emulsion PCR with pyrosequencing,
Margulies and colleagues at 454 Life Sciences per-
formed shotgun sequencing of the entire 580 069 bp of
the *Mycoplasma genitalium* genome at 96% coverage and
99.96% accuracy in a single GS 20 run (8). In 2007,
Roche Applied Science acquired 454 Life Sciences and
introduced the second version of the 454 instrument,
the GS FLX. Sharing the same core technology as the
GS 20, the GS FLX flow cell is referred to as a "picotiter
well" plate, which is made from a fused fiber-optic
bundle. In its newest configuration, approximately
$3.4 \times 10^6$ picoliter-scale sequencing-reaction wells are
etched into the plate surface, and the well walls have a
metal coating to improve signal-to-noise discrimina-
tion. For sequencing (Fig. 1), a library of template DNA is prepared by fragmentation via nebulization or soni-
cation. Fragments several hundred base pairs in length
are end-repaired and ligated to adapter oligonucleo-
tides. The library is then diluted to single-molecule
concentration, denatured, and hybridized to individual
beads containing sequences complementary to
adapter oligonucleotides. The beads are compartmental-
ized into water-in-oil microvesicles, where clonal
expansion of single DNA molecules bound to the beads
occurs during emulsion PCR. After amplification, the
emulsion is disrupted, and the beads containing clonally amplified template DNA are enriched. The
beads are again separated by limiting dilution, depos-
it into individual picotiter-plate wells, and combined

---

**Reviews**

Sanger sequencing, which is based on the electro-
phoretic separation of chain-termination products
produced in individual sequencing reactions. In NGS,
sequencing is performed by repeated cycles of polymerase-mediated nucleotide extensions or, in one
format, by iterative cycles of oligonucleotide ligation.
As a massively parallel process, NGS generates hun-
dreds of megabases to gigabases of nucleotide-
sequence output in a single instrument run, depending
on the platform. These platforms are reviewed next.
Fig. 1. Roche 454 GS FLX sequencing.
Template DNA is fragmented, end-repaired, ligated to adapters, and clonally amplified by emulsion PCR. After amplification, the beads are deposited into picotiter-plate wells with sequencing enzymes. The picotiter plate functions as a flow cell where iterative pyrosequencing is performed. A nucleotide-incorporation event results in pyrophosphate (PPi) release and well-localized luminescence. APS, adenosine 5’-phosphosulfate.
Fig. 2. Illumina Genome Analyzer sequencing.
Adapter-modified, single-stranded DNA is added to the flow cell and immobilized by hybridization. Bridge amplification generates clonally amplified clusters. Clusters are denatured and cleaved; sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Postincorporation fluorescence is recorded. The fluor and block are removed before the next synthesis cycle.
under limiting-dilution conditions, adapter-modified, single-stranded template DNA is added to the flow cell and immobilized by hybridization to the anchors. In contrast to emulsion PCR, DNA templates are amplified in the flow cell by “bridge” amplification, which relies on captured DNA strands “arching” over and hybridizing to an adjacent anchor oligonucleotide. Multiple amplification cycles convert the single-molecule DNA template to a clonally amplified arching “cluster,” with each cluster containing approximately 1000 clonal molecules. Approximately \(50 \times 10^6\) separate clusters can be generated per flow cell. For sequencing, the clusters are denatured, and a subsequent chemical cleavage reaction and wash leave only forward strands for single-end sequencing. Sequencing of the forward strands is initiated by hybridizing a primer complementary to the adapter sequences, which is followed by addition of polymerase and a mixture of 4 differently colored fluorescent reversible dye terminators. The terminators are incorporated according to sequence complementarity in each strand in a clonal cluster. After incorporation, excess reagents are washed away, the clusters are optically interrogated, and the fluorescence is recorded. With successive chemical steps, the reversible dye terminators are unblocked, the fluorescent labels are cleaved and washed away, and the next sequencing cycle is performed. This iterative, sequencing-by-synthesis process requires approximately 2.5 days to generate read lengths of 36 bases. With \(50 \times 10^6\) clusters per flow cell, the overall sequence output is \(>1\) billion base pairs (Gb) per analytical run (11). The newest platform, the Genome Analyzer II, has optical modifications enabling analysis of higher cluster densities. Coupled with ongoing improvements in sequencing chemistry and projected read lengths of 50-plus bases, further increases in output should be realized. Illumina and other NGS technologies have devised strategies to sequence both ends of template molecules. Such “paired-end” sequencing provides positional information that facilitates alignment and assembly, especially for short reads (12, 13). A technical concern of Illumina sequencing is that base-call accuracy decreases with increasing read length (14). This phenomenon is primarily due to “dephasing noise.” During a given sequencing cycle, nucleotides can be under- or overincorporated, or block removal can fail. With successive cycles, these aberrations accumulate to produce a heterogeneous population in a cluster of strands of varying lengths. This heterogeneity decreases signal purity and reduces precision in base calling, especially at the 3’ ends of reads. Modifications in sequencing chemistry and algorithms for data-image analysis and interpretation are being pursued to mitigate dephasing (15). Investigators at the Wellcome Trust Sanger Institute, who have extensive experience with the Illumina platform, have published a series of technical improvements for library preparation, including methods for increasing the reproducibility of fragmentation by adaptive focused acoustic wave sonication, enhanced efficiency of adapter ligation by use of an alternate ligase, and reducing the G+C bias that has been observed in Illumina reads via a modified gel-extraction protocol (16).

APPLIED BIOSYSTEMS/SOLID

The SOLiD (Supported Oligonucleotide Ligation and Detection) System 2.0 platform, which is distributed by Applied Biosystems (http://www.solid.appliedbiosystems.com), is a short-read sequencing technology based on ligation. This approach was developed in the laboratory of George Church and reported in 2005 along with the resequencing of the Escherichia coli genome (17). Applied Biosystems refined the technology and released the SOLiD instrumentation in 2007. Sample preparation shares similarities with the 454 technology in that DNA fragments are ligated to oligonucleotide adapters, attached to beads, and clonally amplified by emulsion PCR. Beads with clonally amplified template are immobilized onto a derivitized-glass flow-cell surface, and sequencing is begun by annealing a primer oligonucleotide complementary to the adapter at the adapter-template junction (Fig. 3). Instead of providing a 3’ hydroxyl group for polymerase-mediated extension, the primer is oriented to provide a 5’ phosphate group for ligation to interrogation probes during the first “ligation sequencing” step. Each interrogation probe is an octamer, which consists of (in the 3’-to-5’ direction) 2 probe-specific bases followed by 6 degenerate bases with one of 4 fluorescent labels linked to the 5’ end. The 2 probe-specific bases consist of one of 16 possible 2-base combinations (for example TT, GT, and so forth). In the first ligation-sequencing step, thermostable ligase and interrogation probes representing the 16 possible 2-base combinations are present. The probes compete for annealing to the template sequences immediately adjacent to the primer. After annealing, a ligation step is performed, followed by wash removal of unbound probe. Fluorescence signals are optically collected before cleavage of the ligated probes, and a wash is performed to remove the fluor and regenerate the 5’ phosphate group. In the subsequent sequencing steps, interrogation probes are ligated to the 5’ phosphate group of the preceding pentamer. Seven cycles of ligation, referred to as a “round,” are performed to extend the first primer. The synthesized strand is then denatured, and a new sequencing primer offset by 1 base in the adapter sequence \((n - 1)\) is annealed. Five rounds total are performed, each time with a new primer with a successive offset \((n - 2, n - 3, \text{and so on})\). By this approach, each template nucleo-
Fig. 3. Applied Biosystems SOLiD sequencing by ligation.

Top: SOLiD color-space coding. Each interrogation probe is an octamer, which consists of (3’-to-5’ direction) 2 probe-specific bases followed by 6 degenerate bases (nnnzzz) with one of 4 fluorescent labels linked to the 5’ end. The 2 probe-specific bases consist of one of 16 possible 2-base combinations. Bottom: (A), The P1 adapter and template with annealed primer (n) is interrogated by probes representing the 16 possible 2-base combinations. In this example, the 2 specific bases complementary to the template are AT. (B), After annealing and ligation of the probe, fluorescence is recorded before cleavage of the last 3 degenerate probe bases. The 5’ end of the cleaved probe is phosphorylated (not shown) before the second sequencing step. (C), Annealing and ligation of the next probe. (D), Complete extension of primer (n) through the first round consisting of 7 cycles of ligation. (E), The product extended from primer (n) is denatured from the adapter/template, and the second round of sequencing is performed with primer (n – 1). With the use of progressively offset primers, in this example (n – 1), adapter bases are sequenced, and this known sequence is used in conjunction with the color-space coding for determining the template sequence by deconvolution (see Fig. 1 in the online Data Supplement). In this technology, template bases are interrogated twice.
Table 1. Comparison of NGS platforms.

<table>
<thead>
<tr>
<th></th>
<th>Roche 454 GS FLX</th>
<th>Illumina Genome Analyzer</th>
<th>Applied Biosystems SOLID</th>
<th>Sanger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing method</td>
<td>Pyrosequencing</td>
<td>Reversible dye terminators</td>
<td>Sequencing by ligation</td>
<td>Dye terminators</td>
</tr>
<tr>
<td>Read lengths</td>
<td>400 bases</td>
<td>36 bases</td>
<td>35 bases</td>
<td>800 bp</td>
</tr>
<tr>
<td>Sequencing run time</td>
<td>10 h</td>
<td>2.5 days</td>
<td>6 days</td>
<td>3 h</td>
</tr>
<tr>
<td>Total bases per run</td>
<td>500 Mb</td>
<td>1.5 Gb</td>
<td>4 Gb</td>
<td>800 bp</td>
</tr>
</tbody>
</table>

The first single-molecule sequencing platform, the HeliScope, is now available from Helicos BioSciences (http://www.helicosbio.com) with a company-reported sequence output of 1 Gb/day. This technology stems from the work of Braslavsky et al., published in 2003 (18). Having obviated clonal amplification of template, the method involves fragmenting sample DNA and polyadenylation at the 3’ end, with the final adenosine fluorescently labeled. Denatured polyadenylated strands are hybridized to poly(dT) oligonucleotides immobilized on a flow-cell surface at a capture density of up to $100 \times 10^6$ template strands per square centimeter. After the positional coordinates of the captured strands are recorded by a charge-coupled device camera, the label is cleaved and washed away before sequencing. For sequencing, polymerase and one of 4 Cy5-labeled dNTPs are added to the flow cell, which is imaged to determine incorporation into individual strands. After label cleavage and washing, the process is repeated with the next Cy5-labeled dNTP. Each sequencing cycle, which consists of the successive addition of polymerase and each of the 4 labeled dNTPs, is termed a “quad.” The number of sequencing quads performed is approximately 25–30, with read lengths of up to 45–50 bases having been achieved. The Helicos platform was used to sequence the 6407-base genome of bacteriophage M13 (19). This study demonstrated both the potential and important technical issues that may be relevant to all single-molecule sequencing methods that are based on sequencing by synthesis. First, sequencing accuracy was appreciably improved when template molecules were sequenced twice (“2-pass” sequencing). Second, the accuracy of sequencing homopolymers was compromised by the polymerase adding additional bases of the same identity in a homopolymeric stretch in a given dNTP addition. Helicos has since developed proprietary labeled dNTPs, termed “virtual terminators,” which the company reports reduce polymerase processivity so that only single bases are added, improving the accuracy of homopolymer sequencing. Interestingly, the percentage of strands in which longer read lengths can be achieved (e.g., 50 nucleotides) is substantially lower than that obtained with shorter (e.g., 25 nucleotides) read-length sequencing, possibly reflecting secondary structures (e.g., hairpins) assumed by the template molecules.

The Impact of NGS on Basic Research

In the short 4 years since the first commercial platform became available, NGS has markedly accelerated multiple areas of genomics research, enabling experiments that previously were not technically feasible or affordable. We describe major applications of NGS and then review the analysis of NGS data.

Clinical Chemistry 55:4 (2009) 647
GENOMIC ANALYSIS

The high-throughput capacity of NGS has been leveraged to sequence entire genomes, from microbes to humans (3, 8, 9, 11, 20–24), including the recent sequencing of the genome of cytogenetically normal acute myeloid leukemia cells, which identified novel, tumor-specific gene mutations (25). The longer read lengths of the 454 technology, compared with the Illumina and SOLiD short-read technologies, facilitate the assembly of genomes in the absence of a reference genome (i.e., de novo assembly). For resequencing, both long- and short-read technologies have been used successfully. In one comparative study, the 454, Illumina, and SOLiD technologies all accurately detected single-nucleotide variations when coverage depth was ≥15-fold per allele (20) (the critical issue of coverage depth is discussed further in the NGS Data Analysis section). The 454 read lengths provide nucleotide haplotype information over a range of several hundred base pairs and are predicted to be better suited for detecting larger insertions and deletions and for producing alignments in areas containing repetitive sequences. Further studies are needed to compare technology performance for detecting insertions and deletions. Each platform has an optional strategy for sequencing both ends of DNA libraries (paired-end sequencing). In addition to effectively doubling sequence output, knowing that reads are associated with each other on a given fragment augments alignment and assembly, especially for short reads. Paired-end sequencing has been used to map genomic structural variation, including deletions, insertions, and rearrangements (12, 13, 26, 27). The ability to sequence complete human genomes at a substantially reduced cost with NGS has energized an international effort to sequence thousands of human genomes over the next decade (http://www.1000genomes.org), which will lead to the characterization and cataloging of human genetic variation at an unprecedented level.

TARGETED GENOMIC RESEQUENCING

Sequencing of genomic subregions and gene sets is being used to identify polymorphisms and mutations in genes implicated in cancer and in regions of the human genome that linkage and whole-genome association studies have implicated in disease (28, 29). Especially in the latter setting, regions of interest can be hundreds of kb’s to several Mb in size. To best use NGS for sequencing such candidate regions, several genomic-enrichment steps, both traditional and novel, are being incorporated into overall experimental designs. Overlapping long-range PCR amplicons (approximately 5–10 kb) can be used for up to several hundred kb’s, but this approach is not practical for larger genomic regions. More recently, enrichment has been achieved by hybridizing fragmented, denatured human genomic DNA to oligonucleotide capture probes complementary to the region of interest and subsequently eluting the enriched DNA (30–33). Capture probes can be immobilized on a solid surface (Roche NimbleGen, http://www.nimblegen.com; Agilent Technologies, http://www.agilent.com; and Febit, http://www.febit.com) or used in solution (Agilent). Current NimbleGen arrays contain 350 000 oligonucleotides of 60–90 bp in length that are typically spaced 5–20 nucleotides apart, with oligonucleotides complementary to repetitive regions being excluded. For enrichment, 5–20 μg of genomic DNA is fragmented and ligated to oligonucleotide linkers containing universal PCR priming sites. This material is denatured, hybridized to an array for 3 days, and eluted, with the enriched DNA amplified by the PCR before NGS library preparation. In reported studies, up to 5 Mb of sequence has been captured on the 350K array, with 60%–75% of sequencing reads mapping to targeted regions; other reads mapping to nontargeted regions reflect nonspecific capture. In development by NimbleGen is the use of an array of 2.1 × 10^6 features for capturing larger genomic regions. Agilent’s solution-based technology uses oligonucleotides up to 170 bases in length, with each end containing sequences for universal PCR priming and with primer sites containing a restriction endonuclease—recognition sequence. The oligonucleotide library is amplified by the PCR, digested with restriction enzymes, and ligated to adapters containing the T7 polymerase promoter site. In vitro transcription is performed with biotinylated UTP to generate single-stranded biotinylated cRNA capture sequences. For capture, 3 μg of fragmented, denatured genomic DNA is hybridized with cRNA sequences for 24 h in solution. After hybridization, duplexes consisting of single-stranded DNA and cRNA are bound to streptavidin-coated magnetic beads; the cRNA is then enzymatically digested, leaving enriched single-stranded DNA that is subsequently processed for NGS. An alternative enrichment approach developed by RainDance Technologies (http://www.raindancetechnologies.com) uses a novel microfluidics technology in which individual pairs of PCR primers for the genomic regions of interest are segregated in water in emulsion droplets and then pooled to create a “primer library.” Separately, emulsion droplets containing genomic DNA and PCR reagents are prepared. Two separate droplet streams are created, one with primer-library droplets and the other with droplets containing genomic DNA/PCR reagents. The 2 streams are merged and primer-library droplets and genomic DNA/PCR reagent droplets are paired in a 1:1 ratio. As paired droplets proceed through the microfluidic channel, they pass an electrical impulse that causes them to physically coalesce. The
Next-Generation Sequencing

coalesced droplets containing individual primer pairs and genomic DNA/PCR reagents are deposited in a 96-well plate and amplified by the PCR. After amplification, the emulsions are disrupted, and the amplicons are pooled and processed for NGS.

METAGENOMICS

NGS has had a tremendous impact on the study of microbial diversity in environmental and clinical samples. Operationally, genomic DNA is extracted from the sample of interest, converted to an NGS library and sequenced. The sequence output is aligned to known reference sequences for microorganisms that are predicted to be present in the sample. Closely related species can be discerned, and more distantly related species can be inferred. In addition, de novo assembly of the data set can yield information to support the presence of known and potentially new species. Qualitative genomic information is obtained, and analysis of the relative abundance of the sequence reads can be used to derive quantitative information on individual microbial species. To date, most NGS-based metagenomic analyses have used the 454 technology and its associated longer read lengths to facilitate alignment to microbial reference genomes and for de novo assembly of previously uncharacterized microbial genomes. Examples of metagenomic studies include the analysis of microbial populations in the ocean (34, 35) and soil (36), the identification of a novel arenavirus in transplantation patients (37), and the characterization of microflora present in the human oral cavity (38) and the guts of obese and lean twins (39).

TRANSCRIPTOME SEQUENCING

NGS has provided a powerful new approach, termed “RNA-Seq,” for mapping and quantifying transcripts in biological samples. Total, ribosomal RNA–depleted, or poly(A)+ RNA is isolated and converted to cDNA. A typical protocol would involve the generation of first-strand cDNA via random hexamer–primed reverse transcription and subsequent generation of second-strand cDNA with RNase H and DNA polymerase. The cDNA is then fragmented and ligated to NGS adapters. For small RNAs such as microRNAs (miRNAs) and short interfering RNAs, preferential isolation via a small RNA–enrichment method, size selection on an electrophoresis gel, or a combination of these approaches is commonly used. RNA ligase is used to join adapter sequences to the RNA; this step is often followed by a PCR amplification step before NGS processing. After sequencing, reads are aligned to a reference genome, compared with known transcript sequences, or assembled de novo to construct a genome-scale transcription map. Although RNA-Seq is in its early stages as a technology, it has already shown some advantages over gene expression arrays (40). First, arrays depend on tiling existing genomic sequences, whereas RNA-Seq is not constrained by this limitation, allowing characterization of transcription without prior knowledge of the genomic sites of transcription origin. RNA-Seq is capable of single-base resolution and, compared with arrays, demonstrates a greater ability to distinguish RNA isoforms, determine allelic expression, and reveal sequence variants. Expression levels are deduced from the total number of reads that map to the exons of a gene, normalized by the length of exons that can be uniquely mapped. Results obtained with this approach have shown close correlation with those of quantitative PCR and RNA-spiking experiments. The dynamic range of RNA-Seq for determining expression levels is 3–4 orders of magnitude, compared with 2 orders of magnitude for expression arrays. In this context, RNA-Seq has shown improved performance for the quantitative detection of both highly produced transcripts and transcripts produced at low levels. RNA-Seq is being used to confirm and revise gene annotation, including 5’ and 3’, and exon/intron boundaries; the latter is achieved by mapping reads to exon junctions defined by GT-AG splicing consensus sites. Both qualitative and quantitative information regarding splicing diversity can be deduced. RNA-Seq has been applied to a variety of organisms, including Saccharomyces cerevisiae, Arabidopsis thaliana, mice, and human cells (40–51).

MAPPING OF DNA-BINDING PROTEINS AND CHROMATIN ANALYSIS

The delineation of regulatory proteins associated with genomes was substantially accelerated by the introduction of chromatin immunoprecipitation and microarray hybridization (ChIP-on-chip) technology (52). In this approach, proteins in contact with genomic DNA are chemically cross-linked (typically with mild formaldehyde treatment) to their binding sites, and the DNA is fragmented by sonication or digestion with micrococcal nuclease. The proteins cross-linked with DNA are immunoprecipitated with antibodies specific for the proteins of interest. The DNA in the immunoprecipitate is purified and hybridized to an oligonucleotide array consisting of sequences from the genome, allowing identification of the protein-binding sites. This approach has been successfully used to identify binding sites for transcription factors and histone proteins. ChIP-on-chip technology is now being supplanted in a variety of experimental settings with ChIP-Seq, in which the DNA harvested from the immunoprecipitate is converted into a library for NGS. The obtained reads are mapped to the reference genome of interest to generate a genome-wide protein-binding map (53–55). Studies to date that have exam-
ined the genomic-binding sites of the human NRSF (neuron restrictive silencer factor) and STAT1 (signal transducer and activator of transcription 1) proteins indicate the resolution of ChIP-Seq to be greater than for ChIP-on-chip, as evidenced by confirmation of previously identified binding sites and identification of novel binding sites (56, 57). Analogous to RNA-Seq, ChIP-Seq has the important advantage of not requiring prior knowledge of genomic locations of protein binding. In addition to the study of transcription factors, NGS is being used to map genomic methylation. One approach involves traditional bisulfite conversion of DNA followed by NGS, which has been applied to the study of entire genomes or genomic subregions (58, 59). Ongoing studies are attempting to develop a variant of ChIP-Seq in which genomic methylation is assayed by coupling immunoprecipitation with a monoclonal antibody directed against methylated cytosine and subsequent NGS (60).

NGS Data Analysis

NGS experiments generate unprecedented volumes of data, which present challenges and opportunities for data management, storage, and, most importantly, analysis (61). NGS data begin as large sets of tiled fluorescence or luminescence images of the flow-cell surface recorded after each iterative sequencing step (Fig. 4). This volume of data requires a resource-intensive data-pipeline system for data storage, management, and processing. Data volumes generated during single runs of the 454 GS FLX, Illumina, and SOLiD instruments are approximately 15 GB, 1 TB, and 15 TB, respectively. The main processing feature of the data pipeline is the computationally intensive conversion of image data into sequence reads, known as base calling. First, individual beads or clusters are identified and localized in an image series. Image parameters such as intensity, background, and noise are then used in a platform-dependant algorithm to generate read sequences and error probability–related quality scores for each base. Although many researchers use the base calls generated by the platform-specific data-pipeline software, alternative base-calling programs that use more advanced software and statistical techniques have been developed. Features of these alternative programs include the incorporation of ambiguous bases into reads, improved removal of poor-quality bases from read ends (62), and the use of data sets for software training (15). Incorporation of these features has been shown to reduce read error and improve alignment, especially as platforms are pushed to generate longer reads. These advantages, however, must be weighed against the substantial computer resources required by the large volumes of image data.

The quality values calculated during NGS base calling provide important information for alignment, assembly, and variant analysis. Although the calculation of quality varies between platforms, the calculations are all related to the historically relevant phred score, introduced in 1998 for Sanger sequence data (63, 64). The phred score quality value, \( q \), uses a mathematical scale to convert the estimated probability of an incorrect call, \( e \), to a log scale:

\[
q = -10 \cdot \log_{10}(e).
\]

Miscall probabilities of 0.1 (10%), 0.01 (1%), and 0.001 (0.1%) yield phred scores of 10, 20, and 30, respectively. The NGS error rates estimated by quality values depend on several factors, including signal-to-noise levels, cross talk from nearby beads or clusters, and dephasing. Substantial effort has been made to understand and improve the accuracy of quality scores and the underlying error sources (10, 14), including inaccuracies in homopolymer run lengths on the 454 platform and base-substitution error biases with the Illumina format. Study of these error traits has led to examples of software that require no additional base calling but that improve quality-score accuracies and thus improve sequencing accuracy (65, 66). Quality values are an important tool for rejecting low-quality reads, trimming low-quality bases, improving alignment accuracy, and determining consensus-sequence and variant calls (67).

Alignment and assembly are substantially more difficult for NGS data than for Sanger data because of the shorter reads lengths in the former. One limitation of short-read alignment and assembly is the inability to uniquely align large portions of a read set when the read length becomes too short. Similarly, the number of uniquely aligned reads is reduced when aligning to larger, more complex genomes or reference sequences because of their having a higher probability of repetitive sequences. A case in point is a modeling study that indicated that 97% of the E. coli genome can be uniquely aligned with 18-bp reads but that only 90% of the human genome can be uniquely aligned with 30-bp reads (68, 69). Unique alignment or assembly is reduced not only by the presence of repeat sequences but also by shared homologies within closely related gene families and pseudogenes. Nonunique read alignment is handled in software by read distribution between multiple alignment positions or leaving alignment gaps. De novo assembly will reject these reads, leading to shorter and more numerous assembled contigs. These factors are relevant when choosing an appropriate sequencing platform with its associated read length, particularly for de novo assembly (9).

Error rates for individual NGS reads are higher than for Sanger sequencing. The higher accuracy of
Sanger sequencing reflects not only the maturity of the chemistry but also the fact that a Sanger trace peak represents highly redundant, multiple terminated extension reactions. Accuracy in NGS is achieved by sequencing a given region multiple times, enabled by the massively parallel process, with each sequence contributing to “coverage” depth. Through this process, a “consensus” sequence is derived. To assemble, align, and analyze NGS data requires an adequate number of overlapping reads, or coverage. In practice, coverage across a sequenced region is variable, and factors other than the Poisson-like randomness of library preparation that may contribute to this variability include differential ligation of adapters to template sequences and differential amplification during clonal template generation (11, 70). Beyond sequence errors, inadequate coverage can cause failure to detect actual nucleotide variation, leading to false-negative results for heterozygotes (3, 11). Studies have shown that coverages of less than 20- to 30-fold begin to reduce the accuracy of single-nucleotide polymorphism calls in data on the 454 platform (65). For the Illumina system, higher

**Fig. 4. Pseudocolor image from the Illumina flow cell.**

Each fluorescence signal originates from a clonally amplified template cluster. Top panel illustrates 4 emission wavelengths of fluorescent labels depicted in red, green, blue, and yellow. Images are processed to identify individual clusters and to remove noise or interference. The lower panel is a composite image of the 4 fluorescence channels.
coverage depths (50- to 60-fold) have been used in an effort to improve short-read alignment, assembly, and accuracy, although coverage in the 20- to 30-fold range may be sufficient for certain resequencing applications (14). As noted above, one comparative study of a yeast genome showed that the 454, Illumina, and SOLiD technologies all accurately detected single-nucleotide variations when the coverage depth was ≥15-fold per allele (20). Coverage gaps can occur when sequences are not aligned because of substantial variance from a reference. Alignment of repetitive sequences in repeat regions of a target sequence can also affect the apparent coverage. Reads that align equally well at multiple sites can be randomly distributed to the sites or in some cases discarded, depending on the alignment software. In de novo–assembly software, reads with ambiguous alignments are typically discarded, yielding multiple aligned read groups, or contigs, with no information regarding relative order.

A large variety of software programs for alignment and assembly have been developed and made available to the research community (see Table 1 in the online Data Supplement). Most use the Linux operating system, and a few are available for Windows. Many require a 64-bit operating system and can use ≥16 MB of RAM and multiple central-processing unit cores. The range of data volumes, hardware, software packages, and settings leads to processing times from a few minutes to multiple hours, emphasizing the need for sufficient computational power. Although a growing set of variations in alignment and assembly algorithms are available, there remains the trade-off between speed and accuracy in which many but not all possible alignments are evaluated, with a balance having to be struck between ideal alignment and computational efficiency.

NGS software features vary with the application and in general may include alignment, de novo assembly, alignment viewing, and variant discovery programs. In addition some NGS statistical data-analysis tools are being developed (such as JMP Genomics; SAS Institute). Software packages available for alignment and assembly to a reference sequence include Zoom (71), MAQ (67), Mosaic (72), SOAP (73), and SHRIMP (http://compbio.cs.toronto.edu/shrimp/), which supports SOLiD color-space analysis. Software for de novo assembly includes Edina (70), EULER-SR (74), SHARCGS (75), SSAKE (69), Velvet (76), and SOAPdenovo (http://soap.genomics.org.cn/). Recently released commercial software for alignment and de novo assembly includes packages from DNASTar (www.dnastar.com), SoftGenetics (www.softgenetics.com), and CLC bio (www.clcbio.com) that feature data viewers that allow the user to see read alignments, coverage depth, genome annotations, and variant analysis. Fig. 5 presents some examples of NGS data viewed in 2 different software systems.

RNA-Seq data analysis poses unique challenges and requires sequence alignment across spliced regions of transcripts as well as poly(A) tails. Current software has made strong inroads, however, with incorporation of motif recognition at splice junctions and identification of intron–exon borders through regions of low alignment coverage (41). Deciphering multiple transcript isoforms involves mapping reads to known and putative splicing junctions and, in one approach, requires that each isoform be supported by multiple independent splice-junction reads with independent start sites (51). ERANGE software has been used in the analysis of the mouse transcriptome (43). ERANGE maps unique reads to their genomic site of origin and maps reads that match to several sites, or multireads, to a most likely site of origin. Reads that do not map to a known exon are grouped together by homology into candidate exons or parts of exons. The near proportional nature of NGS transcriptome data allows quantification of RNA production from the coverage of the assembled or aligned data. ERANGE uses normalized counts of unique reads, spliced reads, and multireads to quantify transcripts. Additional analytical considerations are needed for miRNA studies, including RNA secondary-structure analysis for hairpins, alignment to known miRNA databases, and searches within the NGS data set for complementary miRNA strands, as described in studies of developing rice grains (77) and chicken embryos (78).

Research with ChIP-Seq has led to analysis methods and software that exploit the advantages over ChIP-in-chip, namely a larger, more information-rich data set. The single-base resolution of the data allows improved estimation of binding-site positions in the programs QuEST (79) and MACS (80). Aligned data at the protein–binding regions typically have 2 characteristic offset peaks, each of which is populated by only forward or only reverse reads. These peaks are hallmarks of the immunoprecipitated short ChIP-Seq DNA fragments with a binding site near the center and are used by the software to estimate binding-site location near the mean peak position. Additional program features include advancements in statistical analysis to minimize miscalled binding sites, error probability estimation, and motif analysis (see Table 1 in the online Data Supplement).

A Clinical Future for NGS

From the impact that NGS has made at the basic-research level, we can anticipate its translation into molecular diagnostics. Key issues that will need to be addressed in this transition will include complexity of
Fig. 5. Examples of NGS data viewed in 2 different software systems.

(A), Roche Amplicon Variant Analyzer software displaying GS FLX data from the CFTR gene [cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)]. Lower pane shows reference sequence (green) above 18 of 68 aligned reads. Column highlighted in yellow and blue shows a heterozygous single-nucleotide polymorphism (SNP). Single T/A insertions (red) may represent errors. Upper pane shows percent variation from reference (vertical bars) and coverage (pale blue line). (B), DNASTar SeqMan Pro software displaying Illumina data from Mycobacterium massiliense. Lower pane shows reference sequence above aligned reads. Green and red arrows show direction of sequencing; base calls at variance with reference are indicated (red). Three columns in agreement (red) indicate presumptive SNPs. Other bases in red may be errors. Upper pane shows read coverage, with relative alignment positions above the graph. See the Acknowledgments for disclosure information on the CFTR-gene analysis performed on residual, deidentified DNA.
Reviews

New single-molecule sequencing technologies in development may decrease sequencing time, reduce costs, and streamline sample preparation. Real-time sequencing by synthesis is being developed by VisiGen (http://www.visigenbio.com) and Pacific Biosciences (http://www.pacificbiosciences.com). VisiGen’s approach uses DNA polymerase modified with a fluorescent donor molecule. Attached to a glass slide surface, the polymerase directs strand extension from primed DNA templates. Nucleotides are modified with fluorescent acceptor molecules, and light energy is used during incorporation to invoke fluorescence resonance energy transfer between polymerase and nucleotide fluorescent moieties, the latter being in the
Next-Generation Sequencing

Thomas R. Gingeras

Pacific Biosciences performs single-molecule real-time sequencing and uses phospholinked fluorescently labeled dNTPs. DNA sequencing is performed in thousands of reaction wells 50–100 nm in diameter that are fabricated with a thin metal cladding film deposited on an optical waveguide consisting of a solid, transparent silicon dioxide substrate. Each reaction well is a nanophotonic chamber in which only the bottom third is visualized, producing a detection volume of approximately 20 zL ($20 \times 10^{-21}$ L). DNA polymerase/template complexes are immobilized to the well bottoms, and 4 differently labeled dNTPs are added. As the DNA polymerase incorporates complementary nucleotides, each base is held within the detection volume for tens of milliseconds, orders of magnitude longer than the amount of time it takes for a nucleotide to diffuse in and out of the detection volume. Laser excitation enables the incorporation events in individual wells to be captured through the optical waveguide, with the fluorescent color detected reflecting the identity of the dNTP incorporated. For sequencing, Pacific Biosciences uses a modified phi29 DNA polymerase that has enhanced kinetic properties for incorporating the system’s phospholinked fluorescently labeled dNTPs. In addition, phi29 DNA polymerase is highly processive, with strand-displacement activity. By taking advantage of these properties, Pacific Biosciences has demonstrated sequencing reads exceeding 4000 bases when a circularized single-stranded DNA molecule is used as template. In this configuration, the phi29 DNA polymerase carries out multiple laps of DNA strand-displacement synthesis around the circular template. The mean DNA-synthesis rate was determined to be approximately 4 bases/s. The observed errors, including deletions, insertions, and mismatches can be addressed by developing a consensus sequence read derived from the multiple rounds of template sequencing. Further refinement of the chemistries and platform instrumentation are ongoing, with a 2010 target date for commercial launch (92–94).

Farther out toward the horizon is sequencing based on monitoring the passage of DNA molecules through nanopores 2–5 nm or greater in diameter. Nanopores are being fabricated in inorganic membranes (solid-state nanopores), assembled from protein channels in lipid membranes, or configured in polymer-based nanofluidic channels. In some configurations, current is applied across nanopore membranes to drive the translocation of negatively charged DNA molecules through pores while monitoring changes in membrane electrical conductance measured in the picoampere range. NABsys (http://www.nabsys.com) is pursuing a combination of nanopores with sequencing by hybridization in which single-stranded DNA molecules are hybridized with a library of hexamers of known sequence. The hybridized DNA is interrogated through a nanopore, with the current changes being different in regions of hexamer hybridization. The patterns of hybridization are used to map annealing regions and determine sequence. Oxford Nanopore Technologies (http://www.nanoporetech.com) is developing nanopore-based sequencing that uses an α-hemolysin protein channel in reconstituted lipid bilayers. The nanopores are situated in individual array wells, and single DNA molecules are introduced into the wells and progressively digested by exonuclease. The released single-nucleotide bases enter the nanopore and alter the electrical current, creating a characteristic current change for each individual base (95, 96). Although the technology is seemingly futuristic, considerable NHGRI funding is being directed toward a variety of nanopore technologies under development as part of the goal of achieving the $1000 genome. For further descriptions of nanopore technologies, the reader is referred to recent reviews (97, 98).

Conclusions

The past few years have witnessed the emergence of NGS technologies that share a common basis, massively parallel sequencing of clonally amplified DNA molecules. In 2008, the first NGS platform based on single-molecule DNA sequencing was launched. On the horizon are real-time single-molecule DNA-sequencing technologies and approaches based on nanopores. NGS has had a substantial impact on basic genomics research in terms of scale and feasibility. Over the next several years, NGS is anticipated to transition into clinical-diagnostic use. Essential elements to make this transition successful will be the requirement of streamlining the processes, especially sample preparation, coupled with improvements in technology robustness and characterization of accuracy through validation studies. The large amounts of sequence-data output will pose a bioinformatics challenge for the clinical laboratory. In addition to data processing, the interpretation of sequencing results will require further characterization of the genomic variation present in the regions analyzed. Although considerable work lies ahead to implement NGS into clinical diagnostics, the potential applications are exciting and numerous.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 re-
Role of Sponsor: The funding organizations played a direct role in the preparation of the manuscript and in the final approval of the manuscript.

Acknowledgments: The analysis of the CFTR gene illustrated in Fig. 5 was performed on residual, deidentified DNA under the approval of University of Utah Institutional Review Board, human subjects protocol number 7275.

References

Next-Generation Sequencing


Supplemental Data Figure 1.

<table>
<thead>
<tr>
<th>Position</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>N</td>
<td>A</td>
<td>T</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>Round 2</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Round 3</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>Round 4</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>C</td>
<td>T</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Round 5</td>
<td>A</td>
<td>N</td>
<td>T</td>
<td>C</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Template</td>
<td>T</td>
<td>T/T</td>
<td>A/A</td>
<td>G/G</td>
<td>A/A</td>
<td>C/C</td>
<td>A</td>
</tr>
</tbody>
</table>

Table showing the nucleotide positions and cycles for primer binding and sequencing:

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Primer n</th>
<th>Primer n-1</th>
<th>Primer n-2</th>
<th>Primer n-3</th>
<th>Primer n-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>Primer n</td>
<td>Primer n-1</td>
<td>Primer n-2</td>
<td>Primer n-3</td>
<td>Primer n-4</td>
</tr>
<tr>
<td></td>
<td>Primer n</td>
<td>Primer n-1</td>
<td>Primer n-2</td>
<td>Primer n-3</td>
<td>Primer n-4</td>
</tr>
<tr>
<td></td>
<td>Primer n</td>
<td>Primer n-1</td>
<td>Primer n-2</td>
<td>Primer n-3</td>
<td>Primer n-4</td>
</tr>
<tr>
<td></td>
<td>Primer n</td>
<td>Primer n-1</td>
<td>Primer n-2</td>
<td>Primer n-3</td>
<td>Primer n-4</td>
</tr>
<tr>
<td></td>
<td>Primer n</td>
<td>Primer n-1</td>
<td>Primer n-2</td>
<td>Primer n-3</td>
<td>Primer n-4</td>
</tr>
</tbody>
</table>
Supplemental Figure 1. SOLiD Color Space Coding and Sequence Deconvolution. Top. Left) Interrogation of a template with sequence TAGACA (positions 1-6 with adapter sequence T at position 0). Right) Color space coding is reshown as reference. Bottom. Sequencing by ligation and deconvolution of template sequence. Extension of primer (n) interrogates template positions 1 and 2, 6 and 7, 11 and 12, and so forth. No sequence information can be determined during the first round since each fluor represents four two base possible combinations. The extension of primer (n-1) allows for the deconvolution of the base at position 1 with the knowledge of the P1 adapter base “T” at position 0. The extension of primers (n-2) and (n-3) provide additional positional color space information prior to deconvolution of the template performed after round 5 (primer n-4). The ”N” bases show the positions along the template that have been sequenced twice prior to deconvolution.
## Supplementary Data Table 1 Next Generation Sequencing Software Summary

<table>
<thead>
<tr>
<th>Program Name</th>
<th>Alignment to Reference</th>
<th>De Novo Assembly</th>
<th>SNP/Indel Discovery</th>
<th>CHiP/BS-Seq</th>
<th>RNA-Seq</th>
<th>Alternate Base Caller</th>
<th>Alignment Viewer</th>
<th>Statistical Analysis</th>
<th>References</th>
<th>Authors</th>
<th>Web Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLPATHS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td>Jonathan Butler et al.</td>
<td></td>
</tr>
<tr>
<td>Alta-cyclic</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2)</td>
<td>Yaniv Erlich et al.</td>
<td>hannonlab.cshl.edu/Alta-Cyclic/main.html</td>
</tr>
<tr>
<td>Bowtie</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(3)</td>
<td>Ben Langmead and Cole Trapnell</td>
<td>bowtie-bio.sourceforge.net</td>
</tr>
<tr>
<td>BS-Seq</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(4)</td>
<td>S.J. Cokus et al</td>
<td>epigenomics.mdb.ucla.edu/BS-Seq/download.html</td>
</tr>
<tr>
<td>CHiPSeq</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(5)</td>
<td>D.S. Johnson et al.</td>
<td>woldlab.caltech.edu/html/chipseq_peak_finder</td>
</tr>
<tr>
<td>CLCbio Genomics Workbench</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Windows</td>
<td></td>
<td></td>
<td></td>
<td>(6)</td>
<td></td>
<td><a href="http://www.clcbio.com">www.clcbio.com</a></td>
</tr>
<tr>
<td>EagleView</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(7)</td>
<td>W. Huang, G. Marth</td>
<td>bioinformatics.bc.edu/marthlab/EagleView</td>
</tr>
<tr>
<td>Erange</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(9)</td>
<td>Ali Mortazavi et al.</td>
<td>bioinfo.cgrb.oregonstate.edu/docs/solexa</td>
</tr>
<tr>
<td>Euler-SR</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(10)</td>
<td>Mark J. Chaisson and Pavel A. Pevzner</td>
<td>woldlab.caltech.edu/maseq</td>
</tr>
<tr>
<td>Exonerate</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(11)</td>
<td>Guy St, C Slater and Ewan Birney</td>
<td><a href="http://www.ebi.ac.uk/~guy/exonrate">www.ebi.ac.uk/~guy/exonrate</a></td>
</tr>
<tr>
<td>FindPeaks</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(12)</td>
<td>Anthony Fejes</td>
<td>vancouvershortr.sourceforge.net</td>
</tr>
<tr>
<td>GMAP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(13)</td>
<td>Thomas Wu and Colin Watanabe</td>
<td></td>
</tr>
<tr>
<td>JMP Genomics</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(14)</td>
<td>SAS Institute Inc.</td>
<td><a href="http://www.jmp.com/software/genomics">www.jmp.com/software/genomics</a></td>
</tr>
<tr>
<td>MAQ</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(15)</td>
<td>Heng Li</td>
<td>maq.sourceforge.net/index.shtml</td>
</tr>
<tr>
<td>MIRA2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(16)</td>
<td>Bastien Chevreux</td>
<td>chevreux.org/projects_mira.html</td>
</tr>
<tr>
<td>MOSAIK</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(17)</td>
<td>Michael Strömberg and Gabor Marth</td>
<td></td>
</tr>
<tr>
<td>NextGENe</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Windows</td>
<td></td>
<td>(19)</td>
<td>SoftGenetics</td>
<td><a href="http://www.softgenetics.com">www.softgenetics.com</a></td>
</tr>
<tr>
<td>NOVOALIGN</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Linux, Mac OS X</td>
<td></td>
<td>(20)</td>
<td>Novocraft Technologies</td>
<td><a href="http://www.novocraft.com/products.html">www.novocraft.com/products.html</a></td>
</tr>
<tr>
<td>PBShort</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Linux</td>
<td></td>
<td>(21)</td>
<td>Gabor Marth</td>
<td>bioinformatics.bc.edu/marthlab/PbShort</td>
</tr>
<tr>
<td>PyroBayes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Linux</td>
<td></td>
<td>(22)</td>
<td>A.R. Quinlan, D.A. Stewart, M.P. Strömberg, G.T. Marth</td>
<td><a href="http://bioinformatics.bc.edu/marthlab/PyroBayes">http://bioinformatics.bc.edu/marthlab/PyroBayes</a></td>
</tr>
<tr>
<td>Program Name</td>
<td>Alignment to Reference</td>
<td>De Novo Assembly</td>
<td>SNP Indel Discovery</td>
<td>CHIP Bisulphite-Seq</td>
<td>RNA-Seq</td>
<td>Alternate Base Caller</td>
<td>Alignment Viewer</td>
<td>Statistical Analysis</td>
<td>References</td>
<td>Authors</td>
<td>Web Site</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------</td>
<td>------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------</td>
<td>----------------------</td>
<td>----------------</td>
<td>--------------------</td>
<td>------------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>QuEST</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(17)</td>
<td>Anton Valouev</td>
<td></td>
<td></td>
<td>mendel.stanford.edu/sidowlab/downloads/quest</td>
</tr>
<tr>
<td>RMAP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(18)</td>
<td>Andrew D. Smith and Zhenyu Xuan</td>
<td></td>
<td></td>
<td><a href="http://rulai.cshl.edu/rmap">http://rulai.cshl.edu/rmap</a></td>
</tr>
<tr>
<td>SAM</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rene Warren, Yaron Butterfield, Asim Siddiqui and Steven Jones</td>
<td></td>
<td></td>
<td><a href="http://www.bcgsc.ca/platform/bioinfo/software/sam">www.bcgsc.ca/platform/bioinfo/software/sam</a></td>
</tr>
<tr>
<td>SeqMan Ngen with SeqMan Pro</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(20)</td>
<td>DNAStar</td>
<td></td>
<td></td>
<td><a href="http://www.dnastar.com">www.dnastar.com</a></td>
</tr>
<tr>
<td>SeqMap</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(21)</td>
<td>H. Jiang and W.H. Wong</td>
<td></td>
<td></td>
<td>biogibbs.stanford.edu/~jiangh/SeqMap</td>
</tr>
<tr>
<td>SHARCGS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(22)</td>
<td>Juliane C. Dohm, Claudio Lottaz, and Heinz Himmelbauer</td>
<td></td>
<td></td>
<td>sharcgs.molgen.mpg.de</td>
</tr>
<tr>
<td>SHRAP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(23)</td>
<td>Andreas Sundquist et al.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHRIMP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Michael Brudno and Stephen Rumble</td>
<td></td>
<td></td>
<td><a href="http://compbio.cs.toronto.edu/shrimp">http://compbio.cs.toronto.edu/shrimp</a></td>
</tr>
<tr>
<td>SISSRs</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Raja Johi</td>
<td></td>
<td></td>
<td>dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs</td>
</tr>
<tr>
<td>Slider</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(24)</td>
<td>N. Malhis, Y. Butterfield, M. Ester, S.J. Jones</td>
<td></td>
<td></td>
<td><a href="http://www.bcgsc.ca/platform/bioinfo/software/slider">http://www.bcgsc.ca/platform/bioinfo/software/slider</a></td>
</tr>
<tr>
<td>SOAPaligner/soap2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(25)</td>
<td>T.W. Lam, Alan Tam, Simon Wong, Edward Wu and S.M. Yiu, Ruiqiang Li</td>
<td>soap.genomics.org.cn/soapaligner.html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(26)</td>
<td>T.W. Lam, Alan Tam, Simon Wong, Edward Wu and S.M. Yiu, Ruiqiang Li</td>
<td>soap.genomics.org.cn/#</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSaha2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(27)</td>
<td>Zemin Ning, Anthony Cox and James Mullikin</td>
<td><a href="http://www.sanger.ac.uk/Software/analysis/SSAHA">www.sanger.ac.uk/Software/analysis/SSAHA</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssahaSNP</td>
<td>X</td>
<td></td>
<td></td>
<td>Linux, Solaris, Mac</td>
<td></td>
<td></td>
<td></td>
<td>Adam Spargo and Zemin Ning</td>
<td><a href="http://www.sanger.ac.uk/Software/Analysis/ssahaSNP">www.sanger.ac.uk/Software/Analysis/ssahaSNP</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSAKE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(28)</td>
<td>Rene Warren, Granger Sutton, Steven Jones and Robert Holt</td>
<td><a href="http://www.bcgsc.ca/platform/bioinfo/software/ssake">http://www.bcgsc.ca/platform/bioinfo/software/ssake</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCAKE</td>
<td>X</td>
<td></td>
<td></td>
<td>Perl</td>
<td></td>
<td></td>
<td>(29)</td>
<td>W. Jeck</td>
<td></td>
<td></td>
<td><a href="http://sourceforge.net/projects/vcake">http://sourceforge.net/projects/vcake</a></td>
</tr>
<tr>
<td>Velvet</td>
<td>X</td>
<td></td>
<td></td>
<td>Linux, Mac</td>
<td></td>
<td></td>
<td></td>
<td>Daniel Zerbino and Ewan Birney</td>
<td><a href="http://www.ebi.ac.uk/~zerbino/velvet/">http://www.ebi.ac.uk/~zerbino/velvet/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XMatchView</td>
<td>X</td>
<td></td>
<td></td>
<td>Linux, Windows</td>
<td></td>
<td></td>
<td></td>
<td>Rene Warren and Steven Jones</td>
<td><a href="http://www.bcgsc.ca/platform/bioinfo/software/xmatchview">http://www.bcgsc.ca/platform/bioinfo/software/xmatchview</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


