Next Generation Sequencing

- DNA sequence represents a single format onto which a broad range of biological phenomena can be projected for high-throughput data collection
- Over the past three years, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by over two orders of magnitude, and democratizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators.
- Next-generation DNA sequencing has the potential to dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and interactomes to become inexpensive, routine and widespread

PMID: 18846087

New Sequencing Strategies

Over the past five years, the incentive for developing entirely new strategies for DNA sequencing has emerged on at least four levels:

1. In the wake of the Human Genome Project, there are few remaining avenues of optimization through which significant reductions in the cost of conventional DNA sequencing can be achieved.

2. The potential utility of short-read sequencing has been tremendously strengthened by the availability of whole genome assemblies for *Homo sapiens* and all major model organisms, as these effectively provide a reference against which short reads can be mapped.

3. A growing variety of molecular methods have been developed, whereby a broad range of biological phenomena can be assessed by high-throughput DNA sequencing (e.g., genetic variation, RNA expression, protein-DNA interactions and chromosome conformation).

4. General progress in technology across disparate fields, including microscopy, surface chemistry, nucleotide biochemistry, polymerase engineering, computation, data storage and others, have made alternative strategies for DNA sequencing increasingly practical to realize.

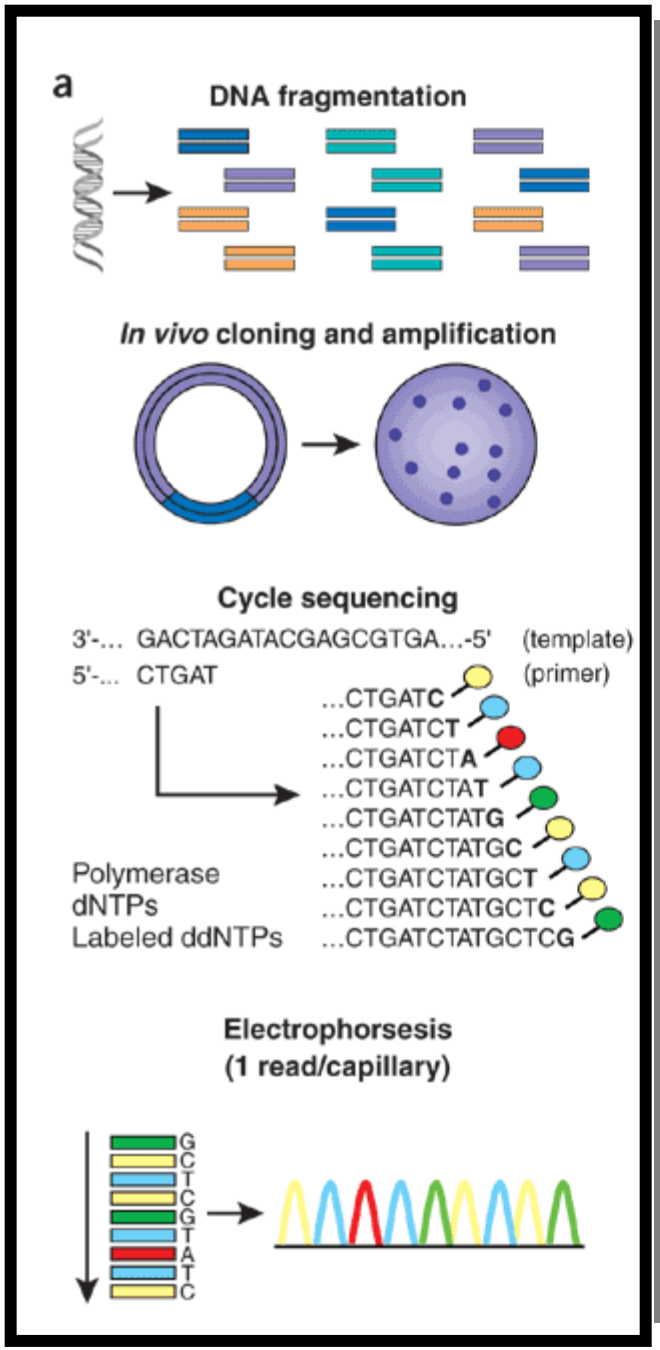
Sanger Sequencing

- Since the early 1990s, DNA sequence production has almost exclusively been carried out with capillary-based, semi-automated implementations of the Sanger biochemistry
- In high-throughput production pipelines, DNA to be sequenced is prepared by one of two approaches: first, for shotgun *de novo* sequencing, randomly fragmented DNA is cloned into a high-copy-number plasmid, which is then used to transform *Escherichia coli;* or second, for targeted resequencing, PCR amplification is carried out with primers that flank the target. The output of both approaches is an amplified template
- After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to ~1,000 bp, and per-base 'raw' accuracies as high as 99.999%. In the context of high-throughput shotgun genomic sequencing, Sanger sequencing costs on the order of $0.50 per kilobase.

Sanger Sequencing

- The sequencing biochemistry takes place in a 'cycle sequencing' reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to known sequence immediately flanking the region of interest.
- Each round of primer extension is stochastically terminated by the incorporation of fluorescently labeled dideoxynucleotides (ddNTPs). In the resulting mixture of end-labeled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position.
- Sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillary-based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-color detection of emission spectra, provides the readout that is represented in a Sanger sequencing 'trace'.
- Software translates these traces into DNA sequence, while also generating error probabilities for each base-call

Conventional versus second-generation sequencing



With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented, then cloned to a plasmid vector and used to transform *E. coli.* For each sequencing reaction, a single bacterial colony is picked and plasmid DNA isolated.

Each cycle sequencing reaction takes place within a microliter-scale volume, generating a ladder of ddNTP-terminated, dye-labeled products, which are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument.

As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace.

2nd Generation DNA Sequencing

- There are various alternative strategies for DNA sequencing, the most successful of which is *cyclic-array sequencing*
- The concept of cyclic-array sequencing can be summarized as the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection
- Cyclic-array sequencing has recently been realized in commercial products

A. 454 sequencing: used in the 454 Genome Sequencers, Roche Applied Science; Basel

B. Solexa technology: used in the Illumina (San Diego) Genome Analyzer

C. SOLiD platform: Applied Biosystems; Foster City, CA, USA

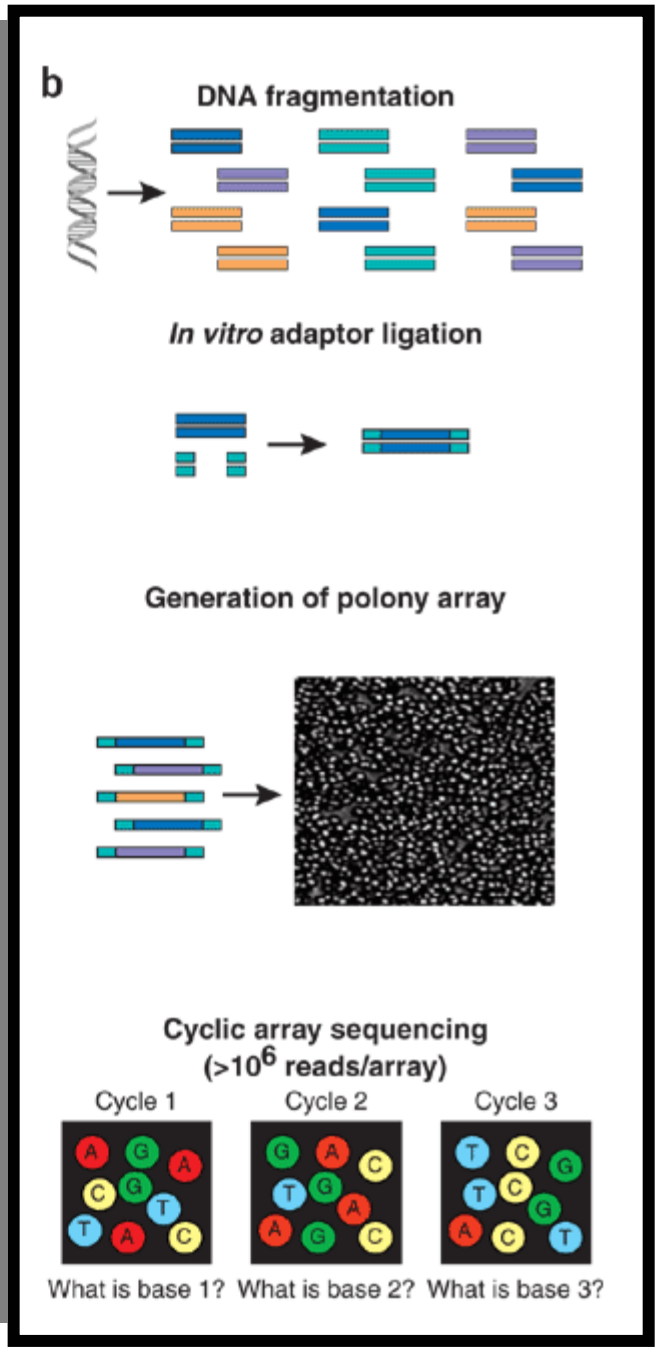
D. Polonator: Dover/Harvard

E. HeliScope Single Molecule Sequencer technology: Helicos; Cambridge, MA, USA.

- Although these platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their work flows are conceptually similar

Conventional versus second-generation sequencing

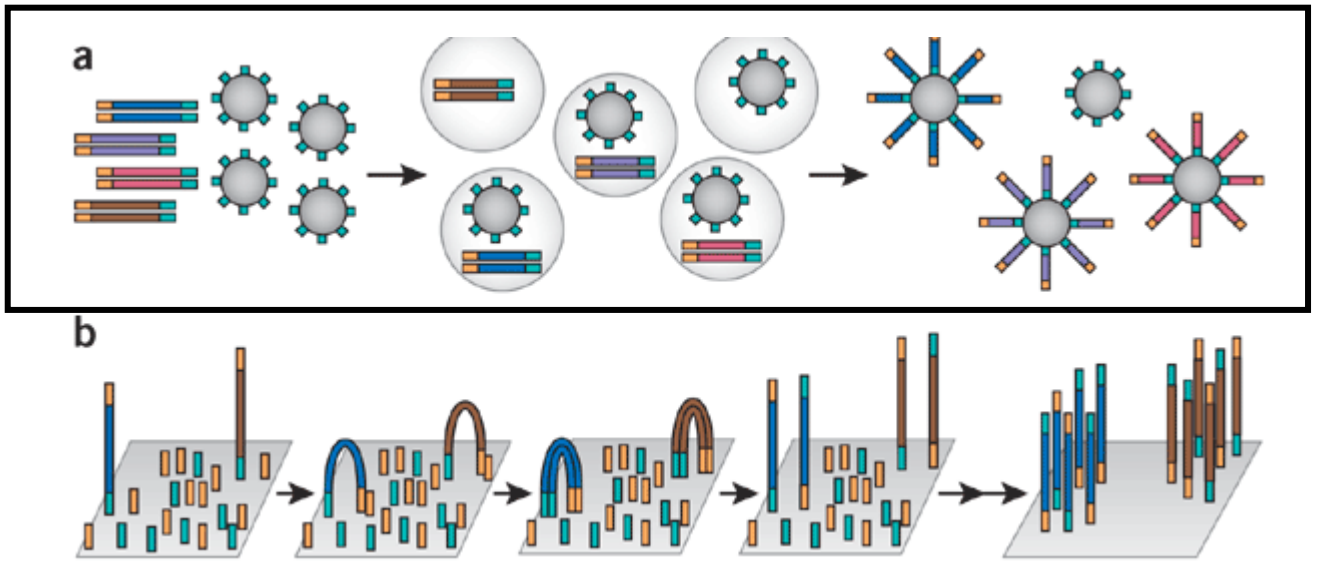
In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or 'polonies'. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature.



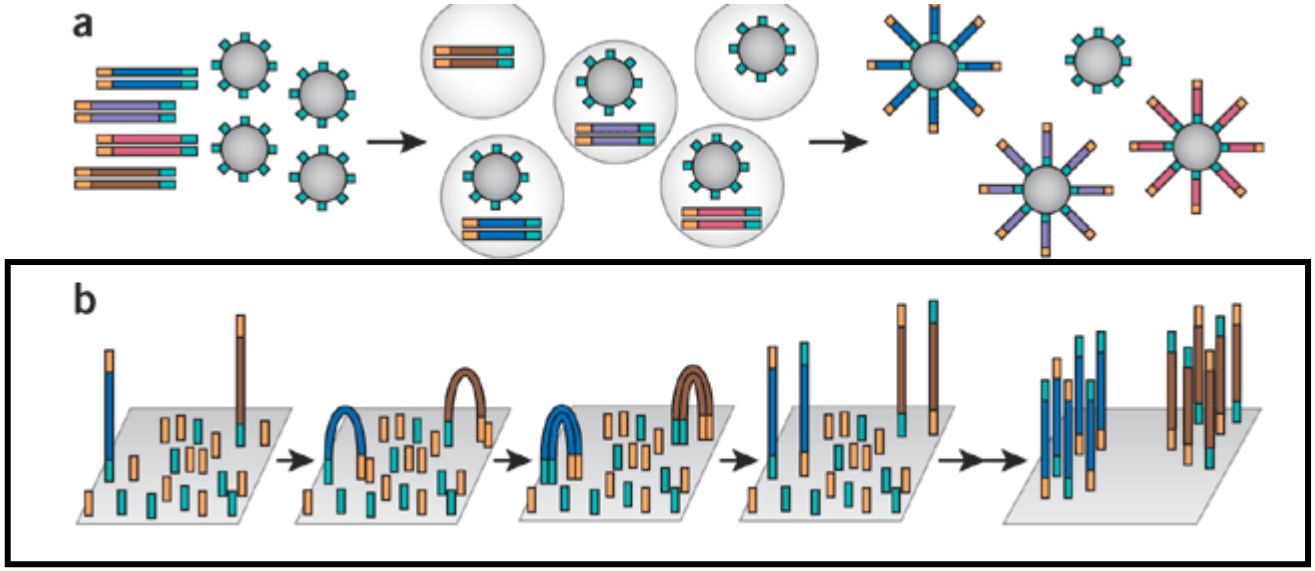
2nd Generation DNA Sequencing

- Library preparation is accomplished by random fragmentation of DNA, followed by *in vitro* ligation of common adaptor sequences.
- The generation of clonally clustered amplicons to serve as sequencing features can be achieved by several approaches, including *in situ* polonies, emulsion PCR or bridge PCR.
- What is common to these methods is that PCR amplicons derived from any given single library molecule end up spatially clustered, either to a single location on a planar substrate (*in situ* polonies, bridge PCR), or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR).
- The sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition.
- The platforms that are discussed here all rely on sequencing by synthesis, that is, serial extension of primed templates, but the enzyme driving the synthesis can be either a polymerase or a ligase.

Clonal amplification of sequencing features

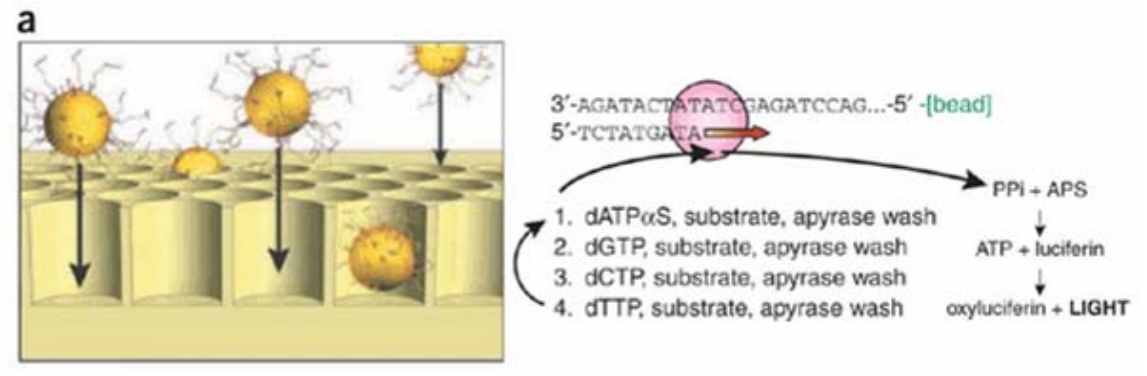


(**a**) The 454, the Polonator and SOLiD platforms rely on emulsion PCR to amplify clonal sequencing features. In brief, an *in vitro*–constructed adaptor-flanked shotgun library (shown as gold and turquoise adaptors flanking unique inserts) is PCR amplified (that is, multi-template PCR, not multiplex PCR, as only a single primer pair is used, corresponding to the gold and turquoise adaptors) in the context of a water-in-oil emulsion. One of the PCR primers is tethered to the surface (5'-attached) of micron-scale beads that are also included in the reaction. A low template concentration results in most bead-containing compartments having either zero or one template molecule present. In productive emulsion compartments (where both a bead and template molecule is present), PCR amplicons are captured to the surface of the bead. After breaking the emulsion, beads bearing amplification products can be selectively enriched. Each clonally amplified bead will bear on its surface PCR products corresponding to amplification of a single molecule from the template library.

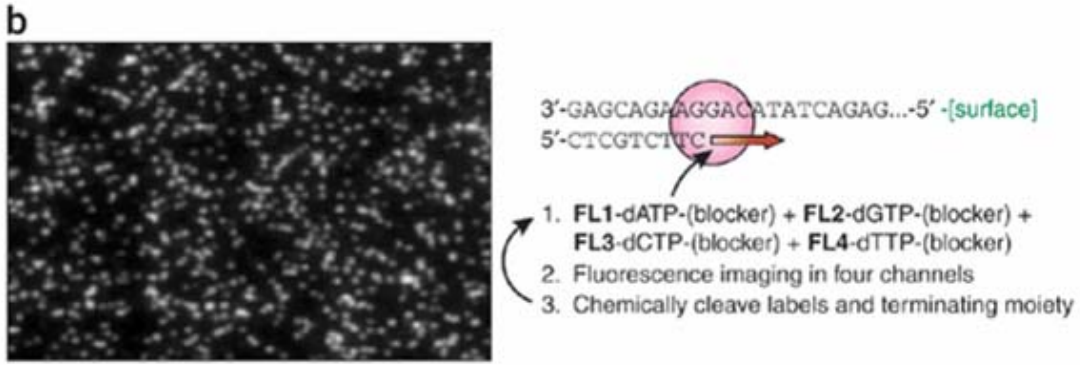


(**b**) The Solexa technology relies on bridge PCR (aka 'cluster PCR') to amplify clonal sequencing features. In brief, an *in vitro*–constructed adaptor-flanked shotgun library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5' ends by a flexible linker. As a consequence, amplification products originating from any given member of the template library remain locally tethered near the point of origin. At the conclusion of the PCR, each clonal cluster contains 1,000 copies of a single member of the template library. Accurate measurement of the concentration of the template library is critical to maximize the cluster density while simultaneously avoiding overcrowding.

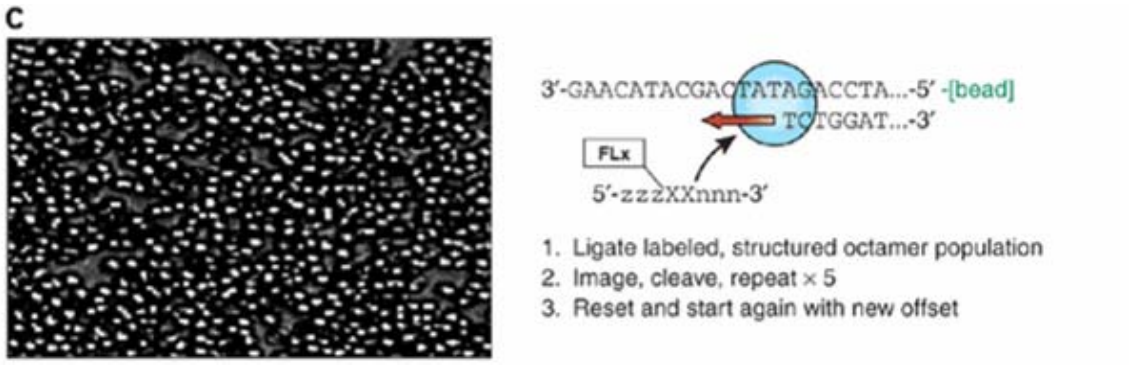
Strategies for cyclic array sequencing



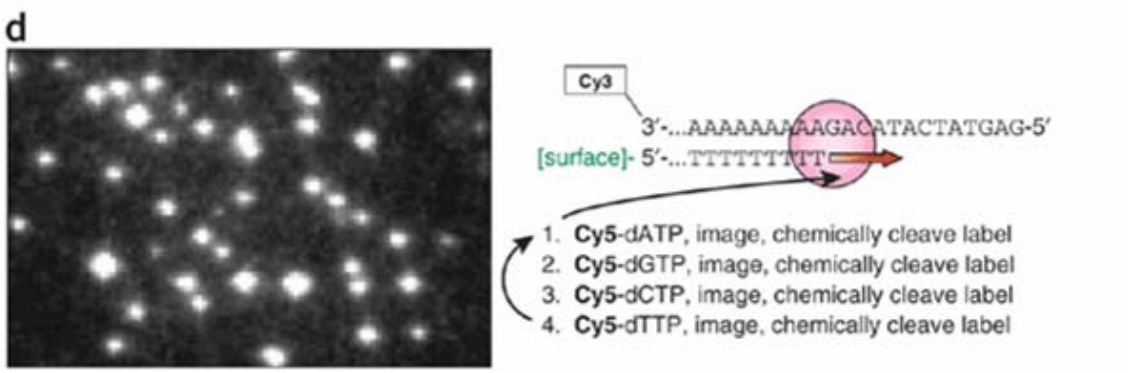
With the **454 platform**, clonally amplified 28- m beads generated by emulsion PCR serve as sequencing features and are randomly deposited to a microfabricated array of picoliter-scale wells. With pyrosequencing, each cycle consists of the introduction of a single nucleotide species, followed by addition of substrate (luciferin, adenosine 5'-phosphosulphate) to drive light production at wells where polymerase-driven incorporation of that nucleotide took place. This is followed by an apyrase wash to remove unincorporated nucleotide. Image from Margulies *et al*. (2005)



With the **Solexa technology**, a dense array of clonally amplified sequencing features is generated directly on a surface by bridge PCR (aka cluster PCR). Each sequencing cycle includes the simultaneous addition of a mixture of four modified deoxynucleotide species, each bearing one of four fluorescent labels and a reversibly terminating moiety at the 3' hydroxyl position. A modified DNA polymerase drives synchronous extension of primed sequencing features. This is followed by imaging in four channels and then cleavage of both the fluorescent labels and the terminating moiety.



With the **SOLiD and the Polonator platforms**, clonally amplified 1- m beads are used to generate a disordered, dense array of sequencing features13. Sequencing is performed with a ligase, rather than a polymerase13, 24, 26, 27, 28. With SOLiD, each sequencing cycle introduces a partially degenerate population of fluorescently labeled octamers. The population is structured such that the label correlates with the identity of the central 2 bp in the octamer (the correlation with 2 bp, rather than 1 bp, is the basis of two-base encoding)26. After ligation and imaging in four channels, the labeled portion of the octamer (that is, 'zzz') is cleaved via a modified linkage between bases 5 and 6, leaving a free end for another cycle of ligation. Several such cycles will iteratively interrogate an evenly spaced, discontiguous set of bases. The system is then reset (by denaturation of the extended primer), and the process is repeated with a different offset (e.g., a primer set back from the original position by one or several bases) such that a different set of discontiguous bases is interrogated on the next round of serial ligations.



With the **HeliScope platform**, single nucleic acid molecules are sequenced directly, that is, there is no clonal amplification step required. Poly-A–tailed template molecules are captured by hybridization to surface-tethered poly-T oligomers to yield a disordered array of primed single-molecule sequencing templates. Templates are labeled with Cy3, such that imaging can identify the subset of array coordinates where a sequencing read is expected. Each cycle consists of the polymerase-driven incorporation of a single species of fluorescently labeled nucleotide at a subset of templates, followed by fluorescence imaging of the full array and chemical cleavage of the label.

Advantages of 2nd Generation DNA Sequencing

1. *in vitro* construction of a sequencing library, followed by *in vitro* clonal amplification to generate sequencing features, circumvents several bottlenecks that restrict the parallelism of conventional sequencing (that is, transformation of *E. coli* and colony picking).
2. Array-based sequencing enables a much higher degree of parallelism than conventional capillary-based sequencing. As the effective size of sequencing features can be on the order of 1 µm, hundreds of millions of sequencing reads can potentially be obtained in parallel by rastered imaging of a reasonably sized surface area.
3. Because array features are immobilized to a planar surface, they can be enzymatically manipulated by a single reagent volume

Disadvantages of 2nd Generation DNA Sequencing

The most prominent of these include read-length (for all of the new platforms, read-lengths are currently much shorter than conventional sequencing) and raw accuracy (on average, base-calls generated by the new platforms are at least tenfold less accurate than base-calls generated by Sanger sequencing).

Table 1. Second-generation DNA sequencing technologies

▲ Figures and tables index Next table ▸

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

Table 2. Applications of next-generation sequencing

◂ Previous table ▲ Figures and tables index Next table ▸

Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46,47,48,49,50,51,52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56,57,58,59,60,61,62,63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67,68,68,69,70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multipleindividuals	61,71

Software and bioinformatics tools for data analysis

A variety of software tools are available for analyzing next-generation sequencing data. Their functions fit into several general categories, including:

1. alignment of sequence reads to a reference
2. base-calling and/or polymorphism detection
3. *de novo* assembly, from paired or unpaired reads
4. genome browsing and annotation

Table 3. Bioinformatics tools for short-read sequencing

◂ Previous table ▴ Figures and tables index

Program	Categories	Author(s)	Reference	URL
Cross_match	Alignment	Phil Green, Brent Ewing and David Gordon		http://www.phrap.org/phredphrapconsed.html
ELAND	Alignment	Anthony J. Cox		http://www.illumina.com/
Exonerate	Alignment	Guy S. Slater and Ewan Birney	72	http://www.ebi.ac.uk/~guy/exonerate
MAQ	Alignment and variant detection	Heng Li	37	http://maq.sourceforge.net
Mosaik	Alignment	Michael Strömberg and Gabor Marth		http://bioinformatics.bc.edu/marthlab/Mosaik
RMAP	Alignment	Andrew Smith, Zhenyu Xuan and Michael Zhang	73	http://rulai.cshl.edu/rmap
SHRiMP	Alignment	Michael Brudno and Stephen Rumble		http://compbio.cs.toronto.edu/shrimp
SOAP	Alignment	Ruiqiang Li *et al.*	35	http://soap.genomics.org.cn
SSAHA2	Alignment	Zemin Ning *et al.*	36	http://www.sanger.ac.uk/Software/analysis/SSAHA2
SXOligoSearch	Alignment	Synamatix		http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
ALLPATHS	Assembly	Jonathan Butler *et al.*	38	
Edena	Assembly	David Hernandez *et al.*	74	http://www.genomic.ch/edena
Euler-SR	Assembly	Mark Chaisson and Pavel Pevzner	75	
SHARCGS	Assembly	Juliane Dohm *et al.*	76	http://sharcgs.molgen.mpg.de
SHRAP	Assembly	Andreas Sundquist *et al.*	39	
SSAKE	Assembly	René Warren *et al.*	40	http://www.bcgsc.ca/platform/bioinfo/software/ssake
VCAKE	Assembly	William Jeck	77	http://sourceforge.net/projects/vcake
Velvet	Assembly	Daniel Zerbino and Ewan Birney	41	http://www.ebi.ac.uk/%7Ezerbino/velvet
PyroBayes	Base caller	Aaron Quinlan *et al.*	34	http://bioinformatics.bc.edu/marthlab/PyroBayes
PbShort	Variant detection	Gabor Marth		http://bioinformatics.bc.edu/marthlab/PbShort
ssahaSNP	Variant detection	Zemin Ning *et al.*		http://www.sanger.ac.uk/Software/analysis/ssahaSNP