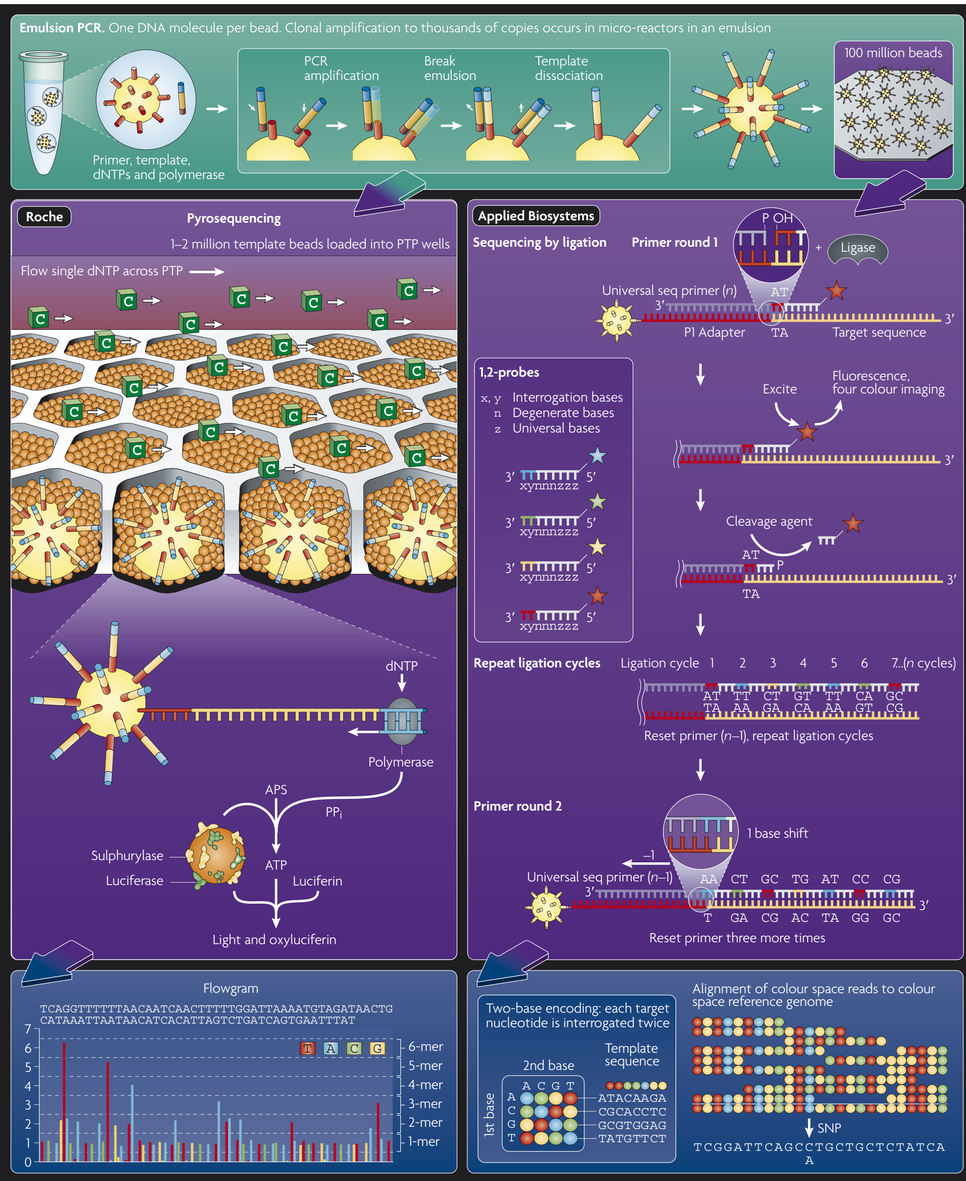
nature REVIEWS GENETICS

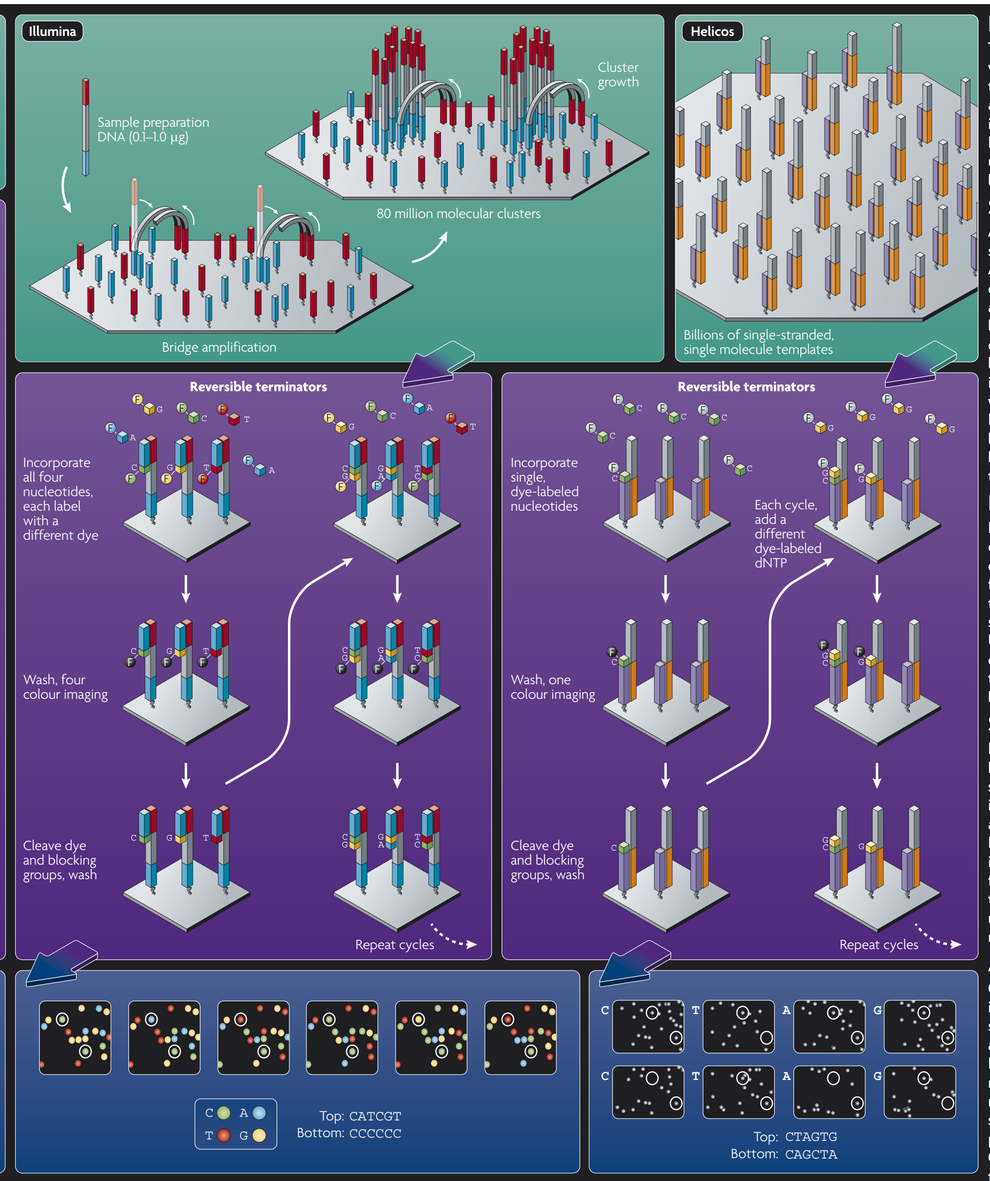
nature genetics

Sequencing technologies — the next generation

Michael L. Metzker

Never has the state of DNA sequencing technology been in greater flux than today. The steadfast approach of fluorescence-based Sanger sequencing appears to have reached its limit for technological improvements. It is being replaced by emerging technologies that promise faster and cheaper sequence information in far greater volumes than ever before. These next generation methodologies push back the limits of possibility, enabling research that would be impractical and too expensive using the Sanger paradigm. With this transition come new possibilities in the field of large-scale genomic science, coupled with new challenges in data storage and analysis. Here, the technical details of commercially available, next generation sequencing platforms are highlighted, along with their advantages and disadvantages. Gone are the days of a single platform capable of addressing the needs of most researchers. Investigators must now identify, from several DNA sequencing approaches, the one platform — or combination thereof — that best serves their application.





Pyrosequencing: Roche

Templates are prepared by emPCR[1], with 1–2 million beads deposited into PTP wells. Smaller beads with attached sulphurylase and luciferase surround the template beads. Individual dNTPs flow sequentially across the wells, dispensed in a predetermined order. On incorporation of the complement dNTP, released PP_i is converted to ATP, producing light from the oxidation of luciferin to oxyluciferin. Reads averaging 400 bases are recorded as flowgrams. For homopolymer repeats up to six nucleotides, the number of dNTPs added is directly proportional to the light signal. Insertions are the most common error type, followed by deletions.

Sequencing by ligation: Applied Biosystems

Around 100 million emPCR-prepared template beads are deposited onto a glass slide. On annealing of a universal primer, a library of 1,2-probes is added. Appropriate conditions enable selective hybridization and ligation of probes to complementary positions. The first (Y) and second (Z) positions of the 1,2-probes are designed as interrogation bases, such that the 16 dinucleotides are encoded by four dyes. Following four-colour imaging, the ligated 1,2-probes are chemically cleaved to generate a 5′-PO_4 group (P). The cycle of hybridization, ligation, imaging, and cleavage is repeated six more times. The extended primer is then stripped from the templates, and a second ligation round is performed with an n–1 primer, which resets the interrogation bases one position to the left. Interrogating each base twice improves the accuracy of the colour call. Seven ligation cycles ensue, followed by three more ligation rounds. A string of 35 data bits, encoded in colour space, are then aligned to a reference genome to decode the DNA sequence. Substitutions are the most common error type.

Reversible terminators: Illumina

Bridge amplification of DNA fragments is randomly distributed across eight channels of a glass slide, to which high-density forward and reverse primers are covalently attached. The solid-phase amplification produces ~80 million MCs from individual ssDNA templates. A primer is annealed to the free ends of templates in each MC. The polymerase extends and then terminates DNA synthesis from a set of four RTs, each labeled with a different dye. Unincorporated RTs are washed away, base identification is performed by four-colour imaging, and blocking and dye groups are removed by chemical cleavage to permit the next cycle. Colour images for a given MC provide reads of ~45 bases. Substitutions are the most common error type.

Single molecule sequencing with RTs: Helicos

Billions of unamplified ssDNA templates are prepared with poly(dA) tails that hybridize to poly(dT) primers covalently attached to a glass slide. For one-pass sequencing, this primer–template complex is sufficient. Two-pass sequencing involves copying the template strand, removing the original template, and annealing a primer directed towards the surface (not shown). Unlike Illumina's RTs, the four Helicos RTs are labeled with the same dye and dispensed individually in a predetermined order. An incorporation event results in a fluorescent signal. The problem of dephasing, in which thousands of copied templates within a given MC do not extend their primers efficiently, is eliminated using single molecules. Deletions, the most common error type, can be greatly reduced by two-pass sequencing providing ~25 base consensus reads.

Applications and challenges

Over 100 papers have described the fruits of these innovations. While improvements continue, read length limitations, and error types and frequencies significantly impact assembly strategies. For short (<100-base) read platforms, assembly is guided by mapping to a reference genome. Combining Sanger and Roche data (100-base reads) improves *de novo* assemblies[2], and as pyrosequencing read lengths have improved, *de novo* assemblies using mixed Roche (250-base reads) and Illumina data have been described[3]. The first personalized genome sequencing projects were recently reported using the Roche[4] and Illumina platforms. Roche, Illumina, and AB platforms are being utilized in the 1,000 Genomes Project to produce a detailed map of human genetic variation and in the Human Microbiome Project to correlate microbiome dynamics with human health. Applications are not limited to sequencing genomes. Consensus counting assays[5] have recently emerged, enabling global profiling of transcription factor binding, mRNA splicing, DNA methylation, small RNAs, chromatin structure, and DNase hypersensitivity sites. Paired-end sequencing protocols have also been developed by Roche, Illumina, and AB. These are not only important for *de novo* assemblies, but for identifying structural variation and mapping mRNA splice isoforms. Looking to the future, the development of platforms from companies such as Pacific Biosciences, Dover Systems (Polonator G.007), Visigen Biotechnologies, LaserGen, Inc., Intelligent Bio-Systems, Complete Genomics, and Oxford Nanopore Technology — as well as those described here — is expected to further improve throughput and accuracy while lowering cost.

References

[1]See Abbreviations. | [2]Goldberg, S.M.D . *et al.* A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. U S A* **103**, 11240-11245 (2006). | [3]Holt, K.E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nature Genet.* **40**, 987-993 (2008). | [4]Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008). | [5]Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nature Methods* **5**, 19-21 (2008).

Abbreviations

APS, adenosine 5′-phosphosulphate; ATP, adenosine triphosphate; dNTP, 2′-deoxyribonucleoside triphosphate; emPCR, emulsion PCR; MC, molecular cluster; PP_i, inorganic pyrophosphate; PTP, PicoTiterPlate; RT, reversible terminator; SNP, single nucleotide polymorphism; ssDNA, single-stranded DNA.

Acknowledgements

Michael L. Metzker is at the Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, Texas, USA. Edited by Louisa Flintoft; designed by Patrick Morgan.