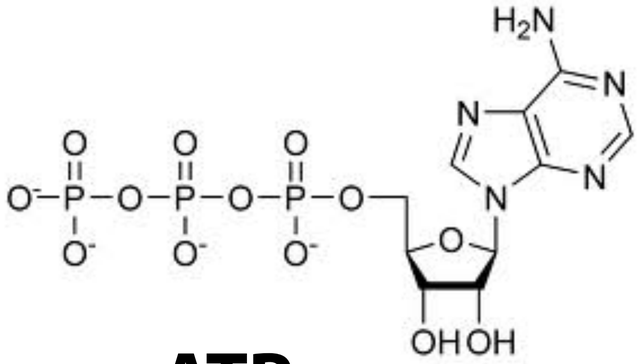
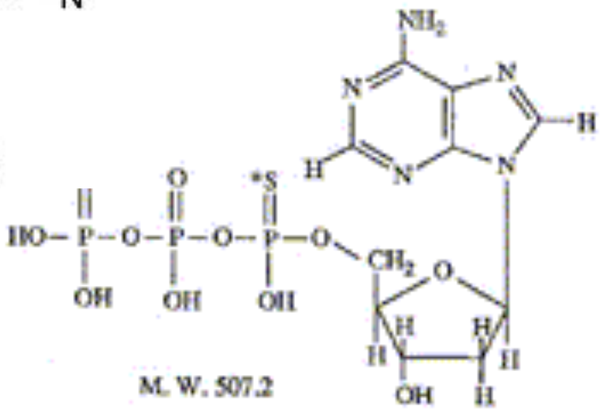
454

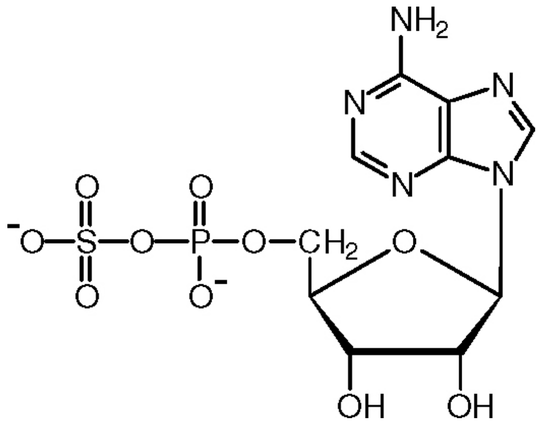
Three ATP-like derivatives



ATP



dATPαS



A 5' phosphosulfate (APS)

The 4 enzymes of pyrosequencing

$(DNA)_n + dXTP \xrightarrow{\text{DNA Polymerase}} (DNA)_{n+1} + PPi$

$PPi + APS \xrightarrow{\text{ATP Sulfurylase}} ATP + SO_4^{2-}$

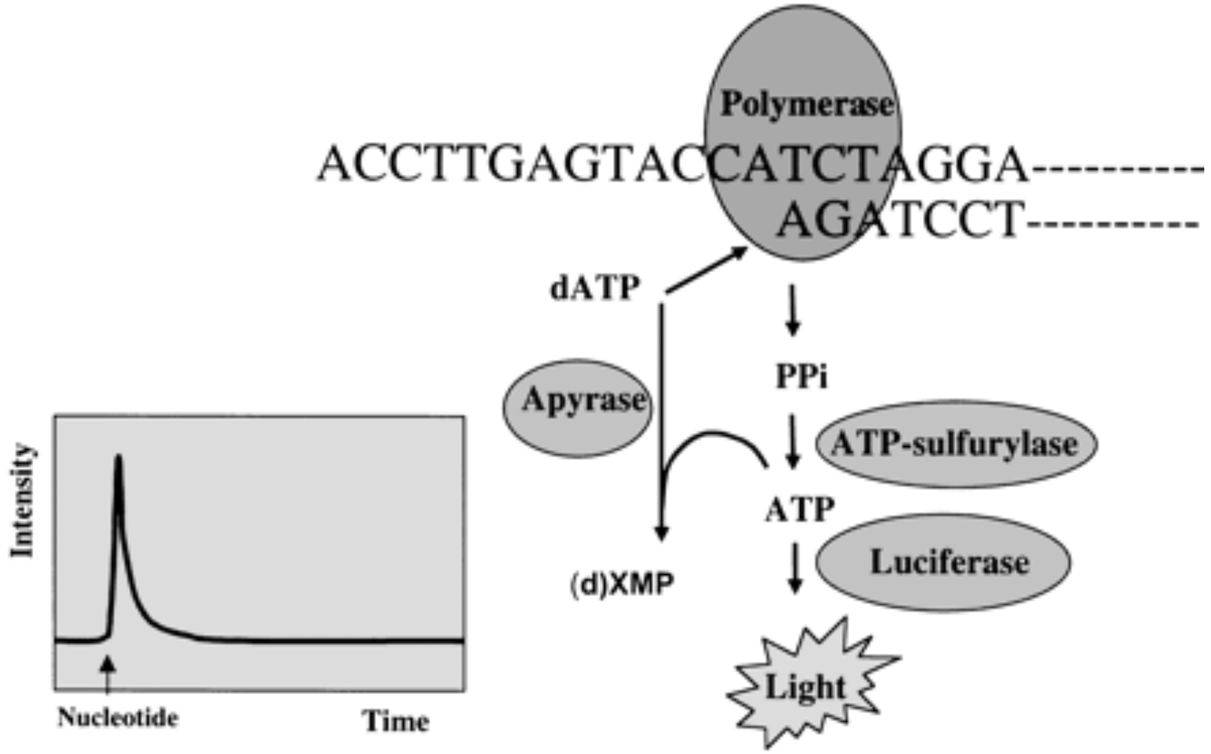
$ATP + Luciferin + O_2 \xrightarrow{\text{Luciferase}} AMP + PPi + Oxyluciferin + CO_2 + Light$

$ATP + dXTP \xrightarrow{\text{Apyrase}} ADP + dXDP + 2Pi$

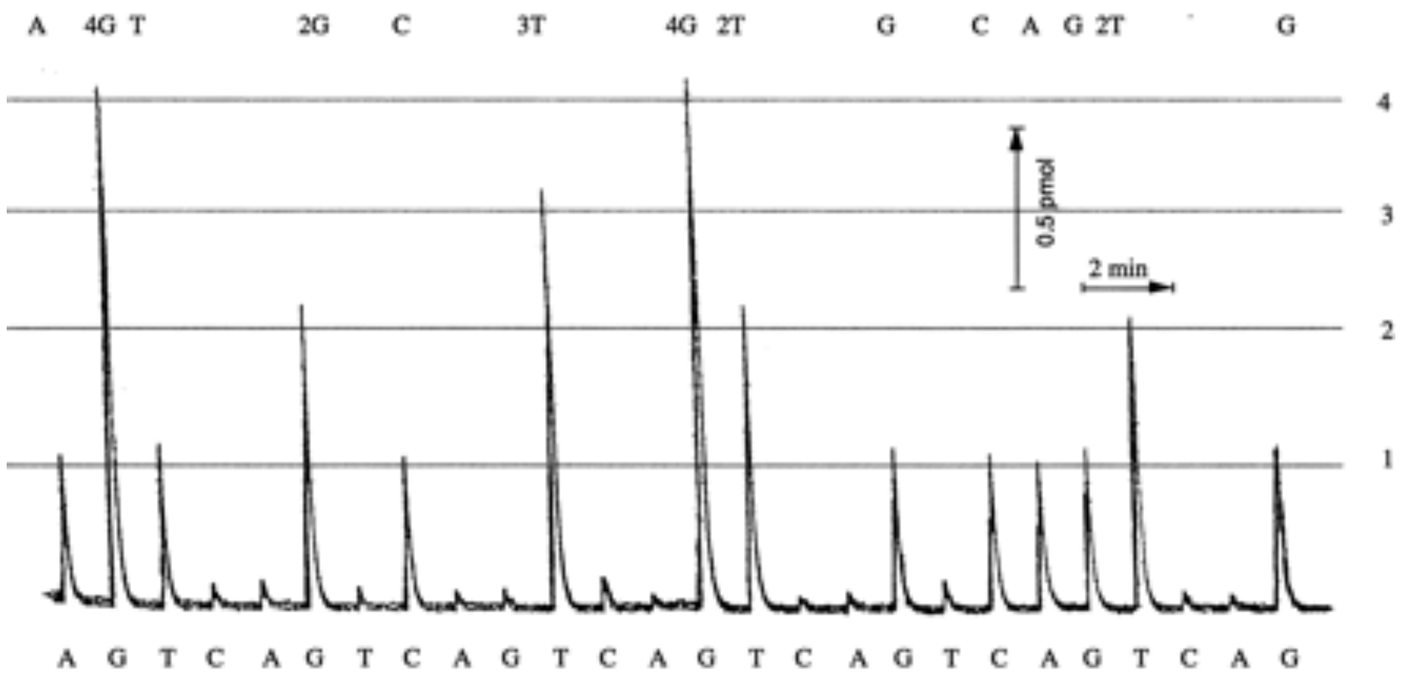
$ADP + dXDP \xrightarrow{\text{Apyrase}} AMP + dXMP + 2Pi$

Fig. 1. Schematic representation of progress of enzymatic reactions in Pyrosequencing. DNA template with hybridized primer and four enzymes involved in Pyrosequencing are added to a well of a microtiter plate. The four different nucleotides are added stepwise, and incorporation is followed using the enzyme ATP sulfurylase and luciferase. The unincorporated nucleotides of each addition are continuously degraded by apyrase allowing addition of subsequent nucleotide.

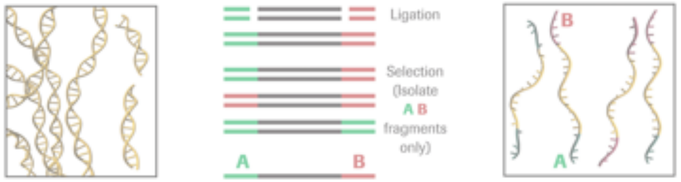
Pyrosequencing process



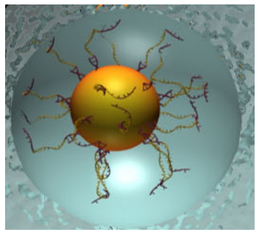
Output


A 4G T 2G C 3T 4G 2T G C A G 2T G
4
3
2
1
0.5 pmol
2 min
A G T C A G T C A G T C A G T C A G T C A G T C A G

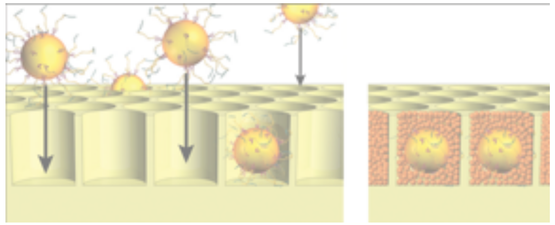
454



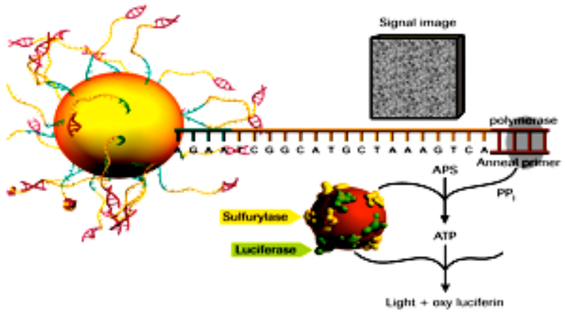
Library preparation



Emulsion PCR



Deposition



sequencing reaction

nature Vol 437|15 September 2005|doi:10.1038/nature03959

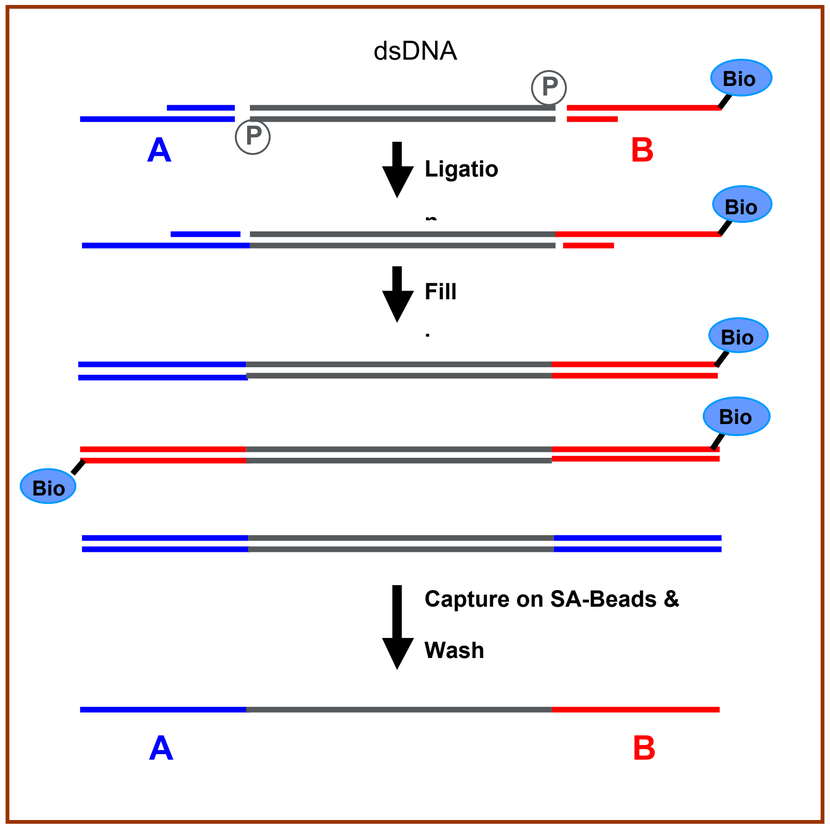
ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

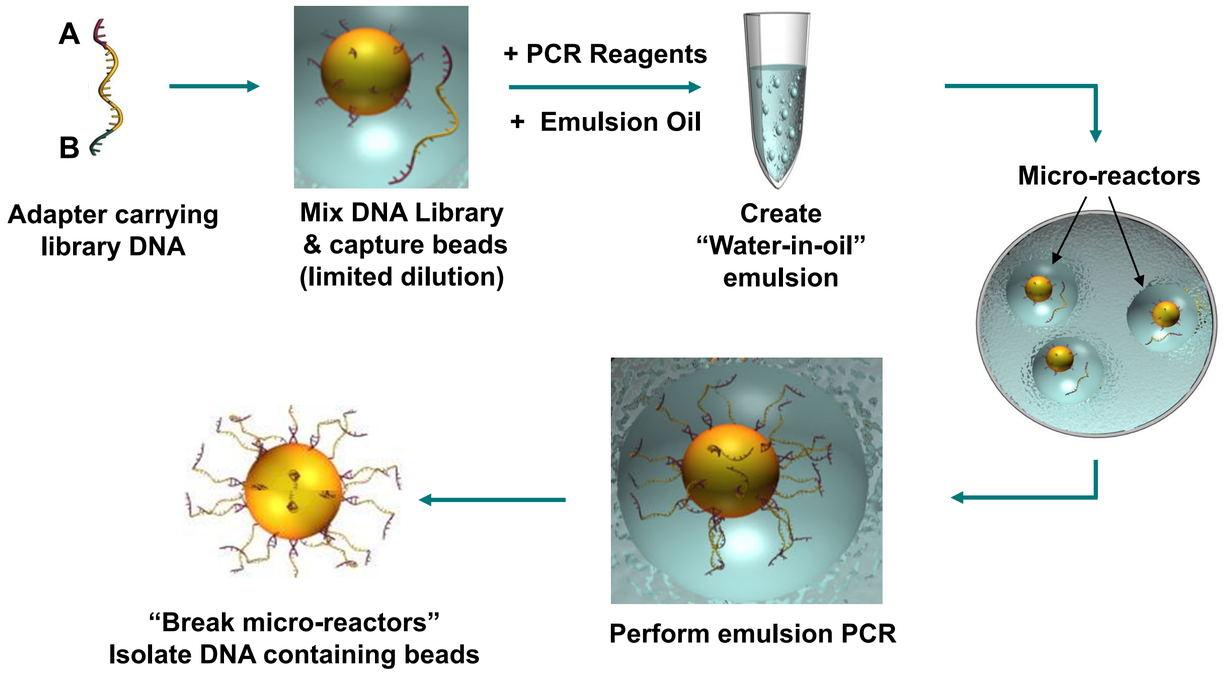
The proliferation of large-scale DNA-sequencing projects in recent years has driven a search for alternative methods to reduce time and cost. Here we describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel fibre-optic slide of individual wells and is able to sequence 25 million bases, at 99% or better accuracy, in one four-hour run. To achieve an approximately 100-fold increase in throughput over current Sanger sequencing technology, we have developed an emulsion method for DNA amplification and an instrument for sequencing by synthesis using a pyrosequencing protocol optimized for solid support and picolitre-scale volumes. Here we show the utility, throughput, accuracy and robustness of this system by shotgun sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome with 96% coverage at 99.96% accuracy in one run of the machine.

Library preparation



Standard Library

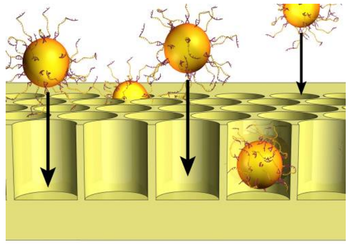
Emulsion PCR



- Generation of millions of clonally amplified sequencing templates on each bead
- No cloning and colony picking

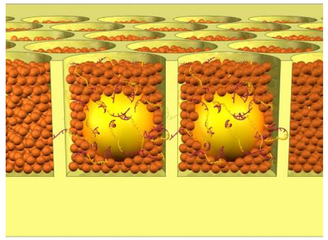
Deposition

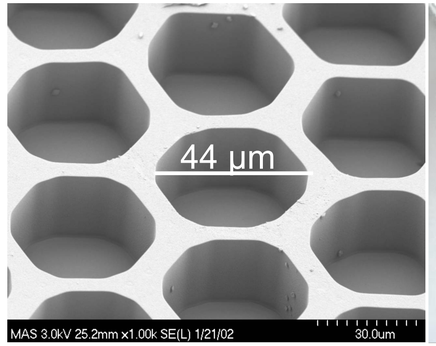
Load beads into PicoTiter™ Plate

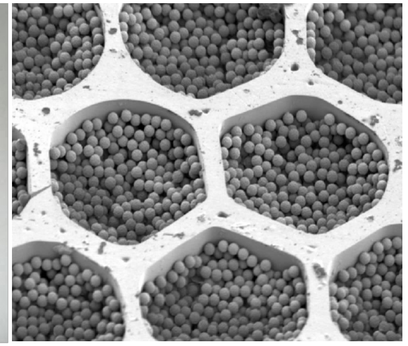


Load Enzyme Beads

Centrifuge Step

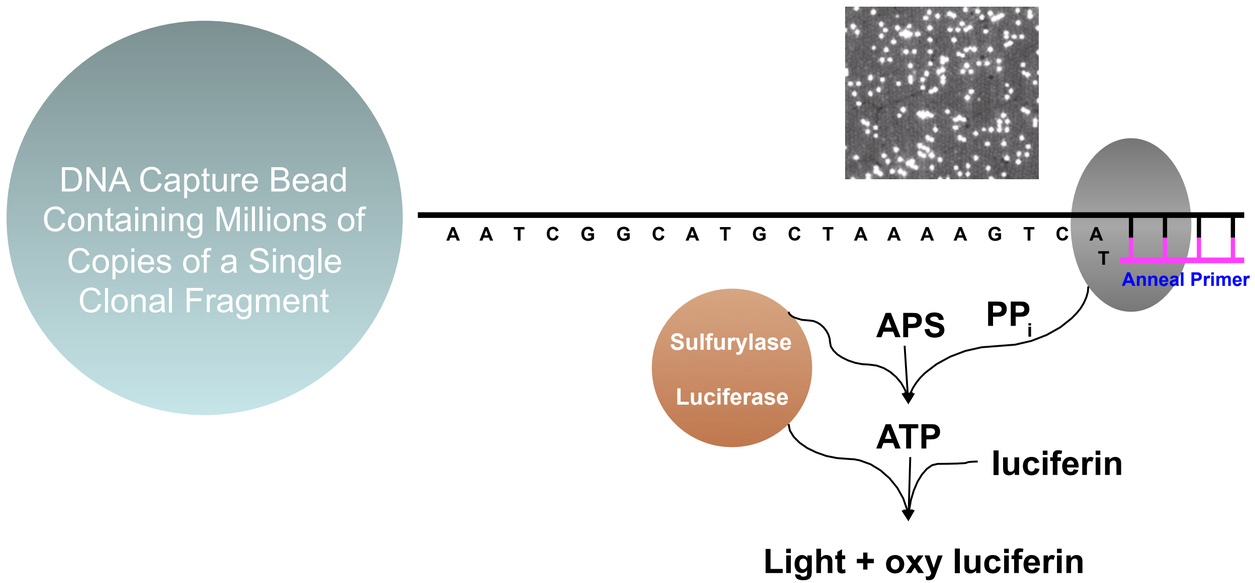






Sequencing

- Simultaneous sequencing of the entire genome in hundreds of thousands of picoliter-size wells
- Pyrophosphate signal generation



UCSC Sequencing Center

The machine

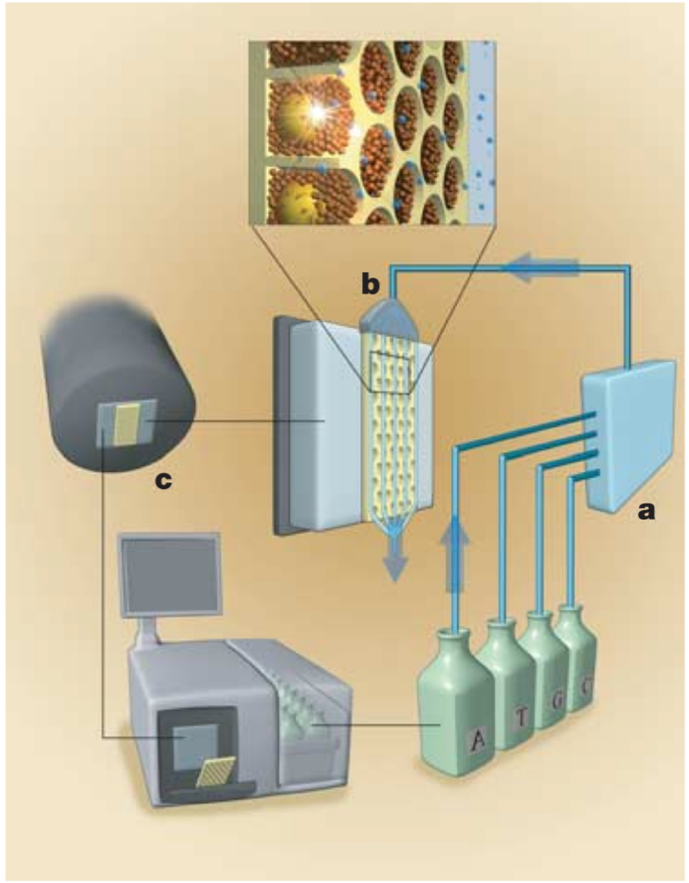


Figure 2 | Sequencing instrument. The sequencing instrument consists of the following major subsystems: a fluidic assembly (**a**), a flow chamber that includes the well-containing fibre-optic slide (**b**), a CCD camera-based imaging assembly (**c**), and a computer that provides the necessary user interface and instrument control.

- dNTP flow (dATPαS instead of dATP)
- Substrate flow (D-luciferin, APS etc)
- Apyrase flow (destroys triphosphates)

Margulies et al. 2005

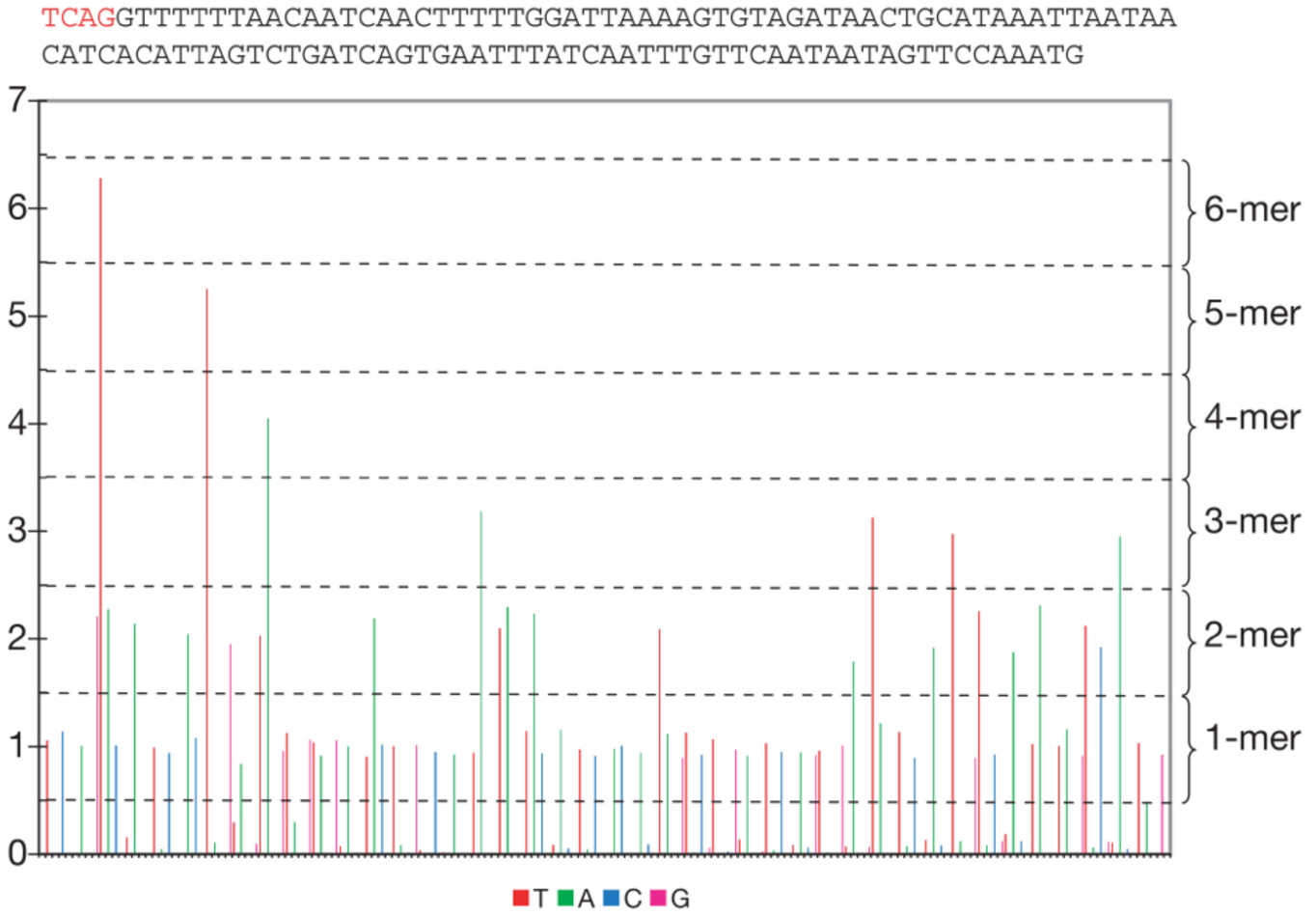
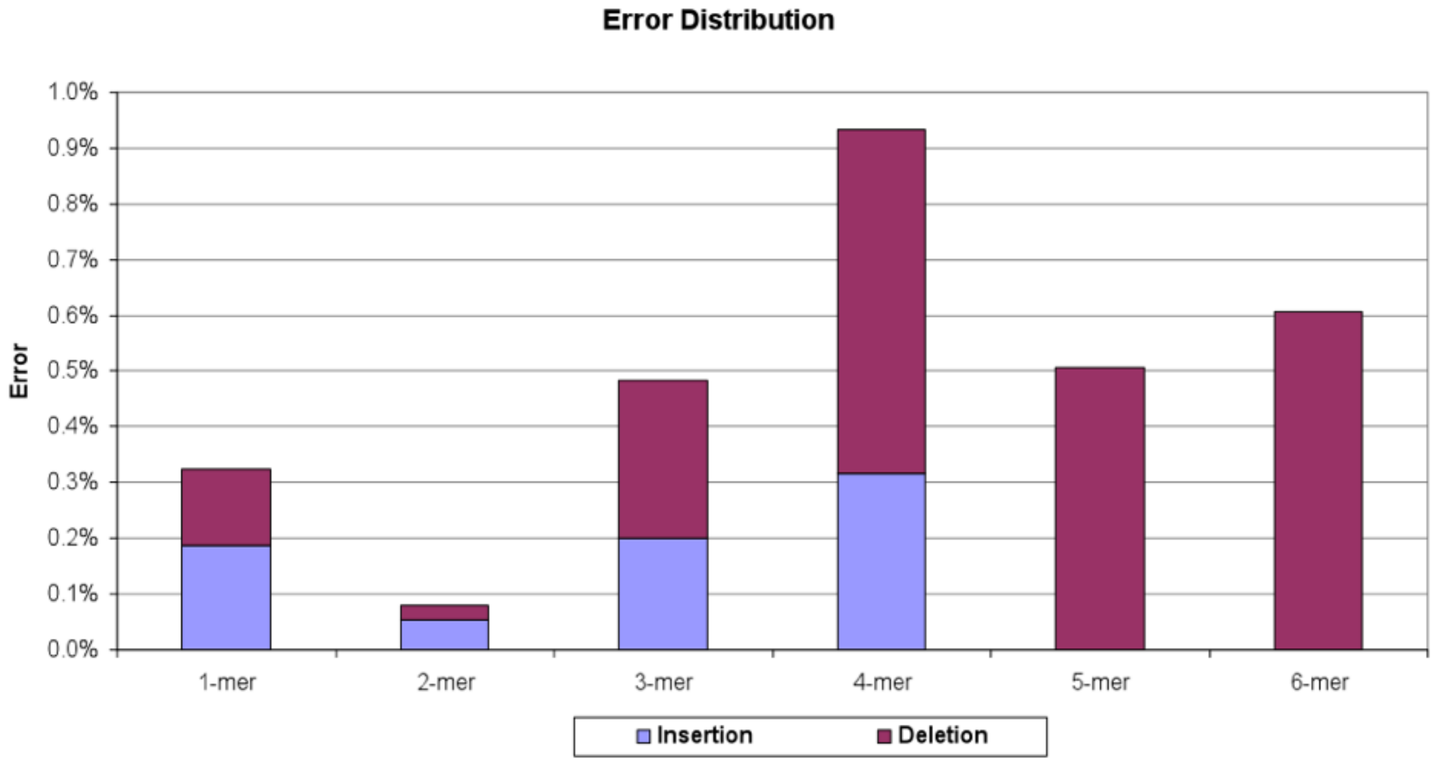


Figure 3 | Flowgram of a 113-bases read from an *M. genitalium* run. Nucleotides are flowed in the order T, A, C, G. The sequence is shown above the flowgram. The signal value intervals corresponding to the various homopolymers are indicated on the right. The first four bases (in red, above the flowgram) constitute the 'key' sequence, used to identify wells containing a DNA-carrying bead.

Margulies et al. 2005



Supplementary Figure 5. Detailed error rates in sequencing a mixture of 6 test fragments, as a function of homopolymer length. Single base error rates are referenced to the total number of *single* bases sequenced. For each homopolymer, the error rate is referenced to the total number of bases sequenced that belong to homopolymers of that length.

Margulies et al. 2005

came from a homopolymer of length two or greater. Since the probability of measuring a signal, *given a homopolymer length*, was empirically established, Bayes' Theorem can be used to determine the probability that a particular homopolymer length produced the observed signal, as follows:

$$P(n \mid s) = \frac{P(s \mid n)P(n)}{\sum_j P(s \mid j)P(j)}$$

where s is the observed signal and n is the length of the homopolymer that produced the signal. As described above, the probability P(s|n) of measuring signal s given a homopolymer of length n follows a Gaussian distribution. For a random nucleotide sequence, the probability P(n) of encountering a homopolymer of length n is simply $1/4^n$ (ignoring a multiplicative normalization constant). The quality score assigned to each base called for each fragment can then be reported as a *phred*-equivalent using the following transformation:

$$Q = -10 \log_{10}[P(\geq n \mid s)]$$

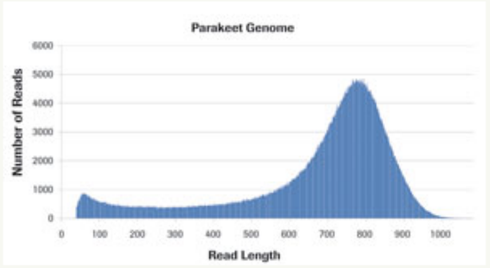
We verified the validity of this approach by correlating calculated *phred* scores and observed *phred* scores, sequencing known genomes other than those used to establish the distribution of signals (Supplementary Figure 10). Our correlation shows excellent correspondence up to *phred* 50 and compares favorably to that established for Sanger sequencing and capillary eletrophoresis [3].

Margulies et al. 2005

GS FLX+ System		
Sequencing Kit	**New! GS FLX Titanium XL+**	**GS FLX Titanium XLR70**
Read Length	Up to 1,000 bp	Up to 600 bp
Mode Read Length	700 bp	450 bp
Throughput Profile	- 85% of total bases from reads >500 bp - 45% of total bases from reads >700 bp	- 85% of total bases from reads > 300 bp - 20% of total bases from reads > 500 bp
Typical Throughput	700 Mb	450 Mb
Reads per Run	~1,000,000 shotgun	~1,000,000 shotgun, ~700,000 amplicon
Consensus Accuracy*	99.997%	99.995%
Run Time	23 hours	10 hours
Sample Input	gDNA or cDNA	gDNA, cDNA, or amplicons (PCR products)
Multiplexing	Multiplex Identifiers (MIDs): 132 Gaskets: 2, 4, 8, 16 regions	
Computing	GS FLX+ Computing Station available	
Dimensions	Upper assembly 74.3 cm × 69.8 cm × 36.1 cm (29.25" × 27.5" × 14.2") (W × D × H), incl. monitor 82.5 cm (32.5") H	
	Lower assembly 75.2 cm × 90.8 cm × 92.7 cm (29.62" × 35.75" × 36.5") (W × D × H)	
Weight	532 lbs (242 kg)	
Power Supply	120 VAC 50/60 Hz 1250 VA 230 VAC 50/60 Hz 1250 VA	

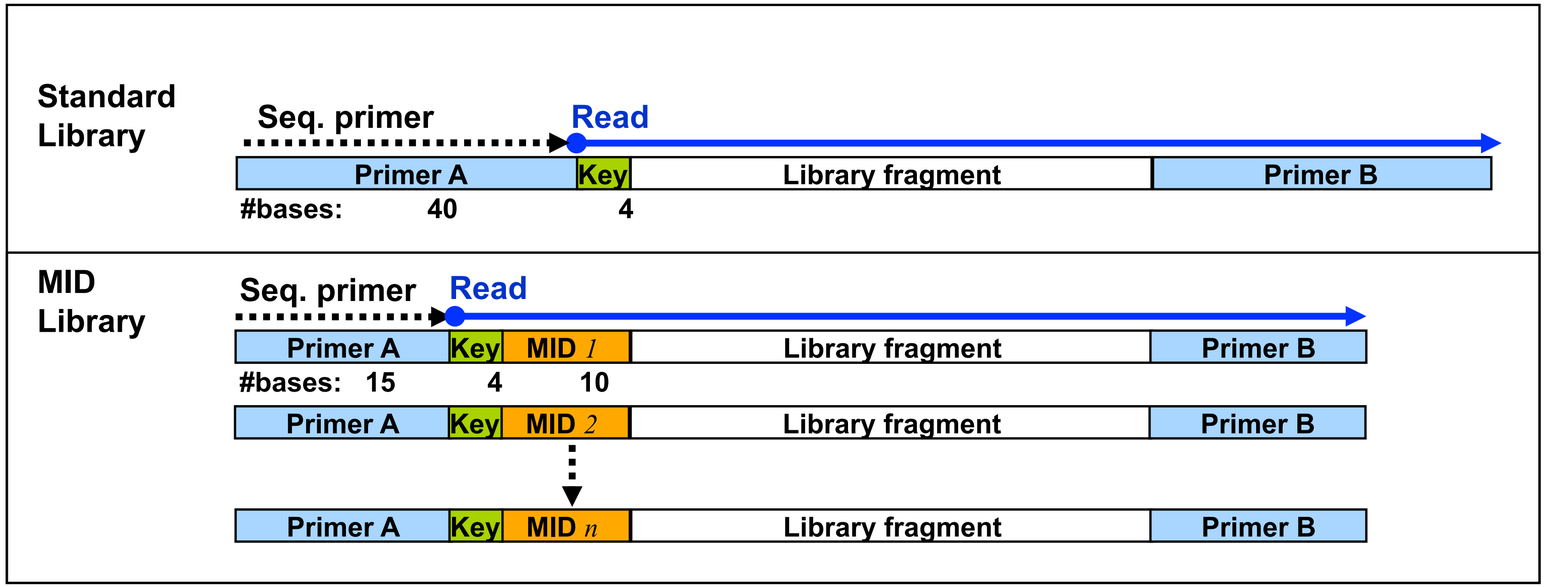
Actual results depend on specific sample and genomic characteristics.

*Consensus accuracy at 15x coverage.

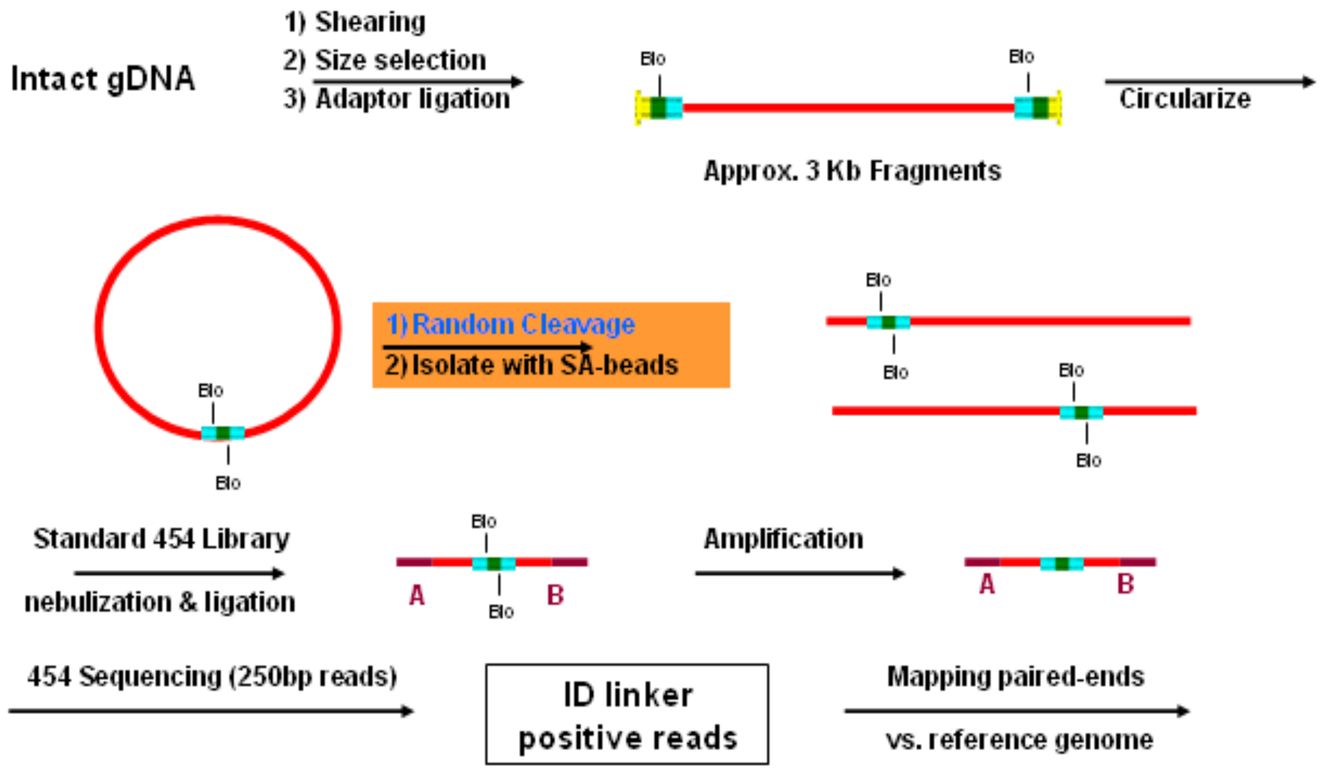


Read length distribution of 1,293,269 reads from a GS FLX+ run of the parakeet genome. Mode read length is 793 bp.

Multiplexing



Mate-pair with 454



UCSC Sequencing Center

Flow value separation

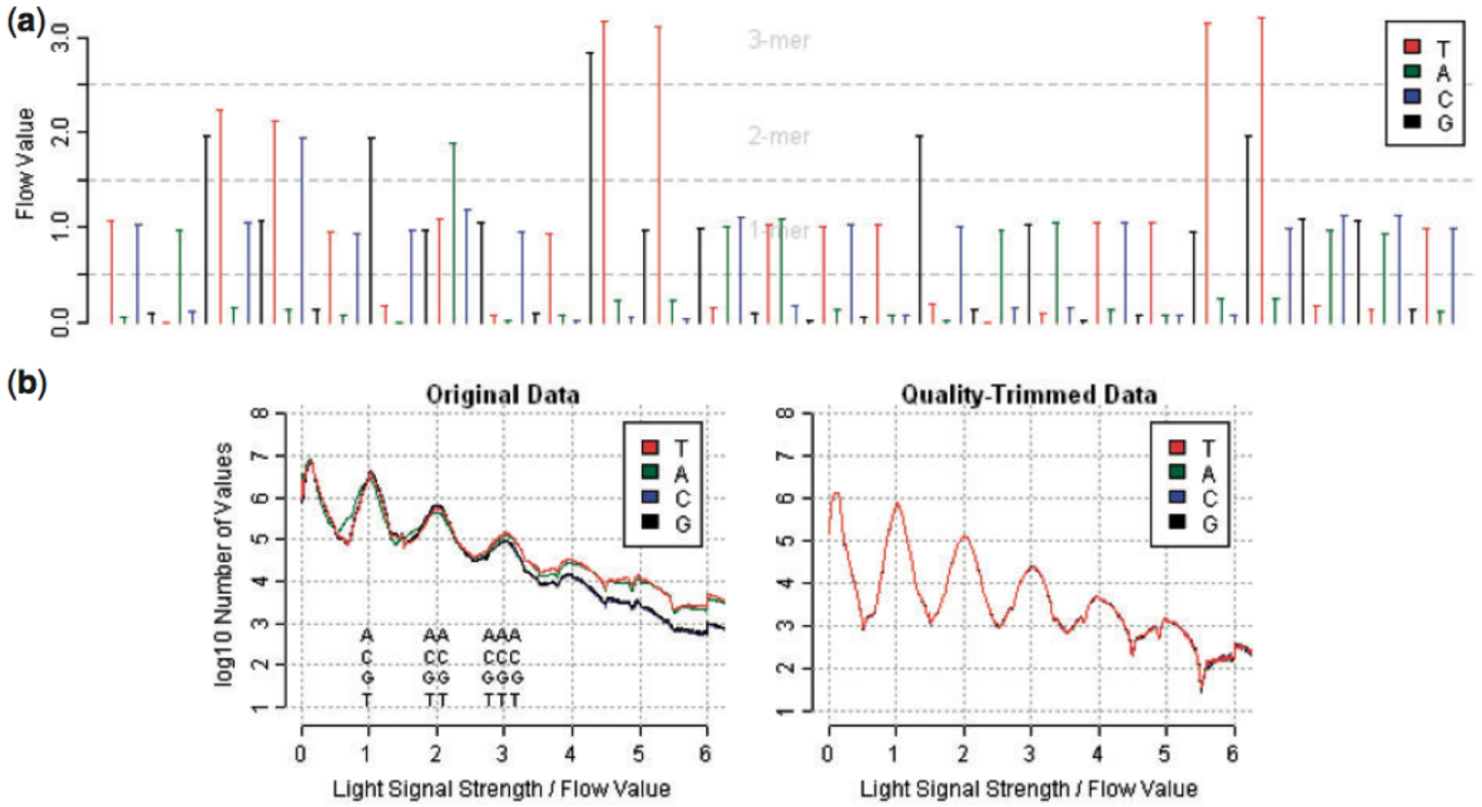
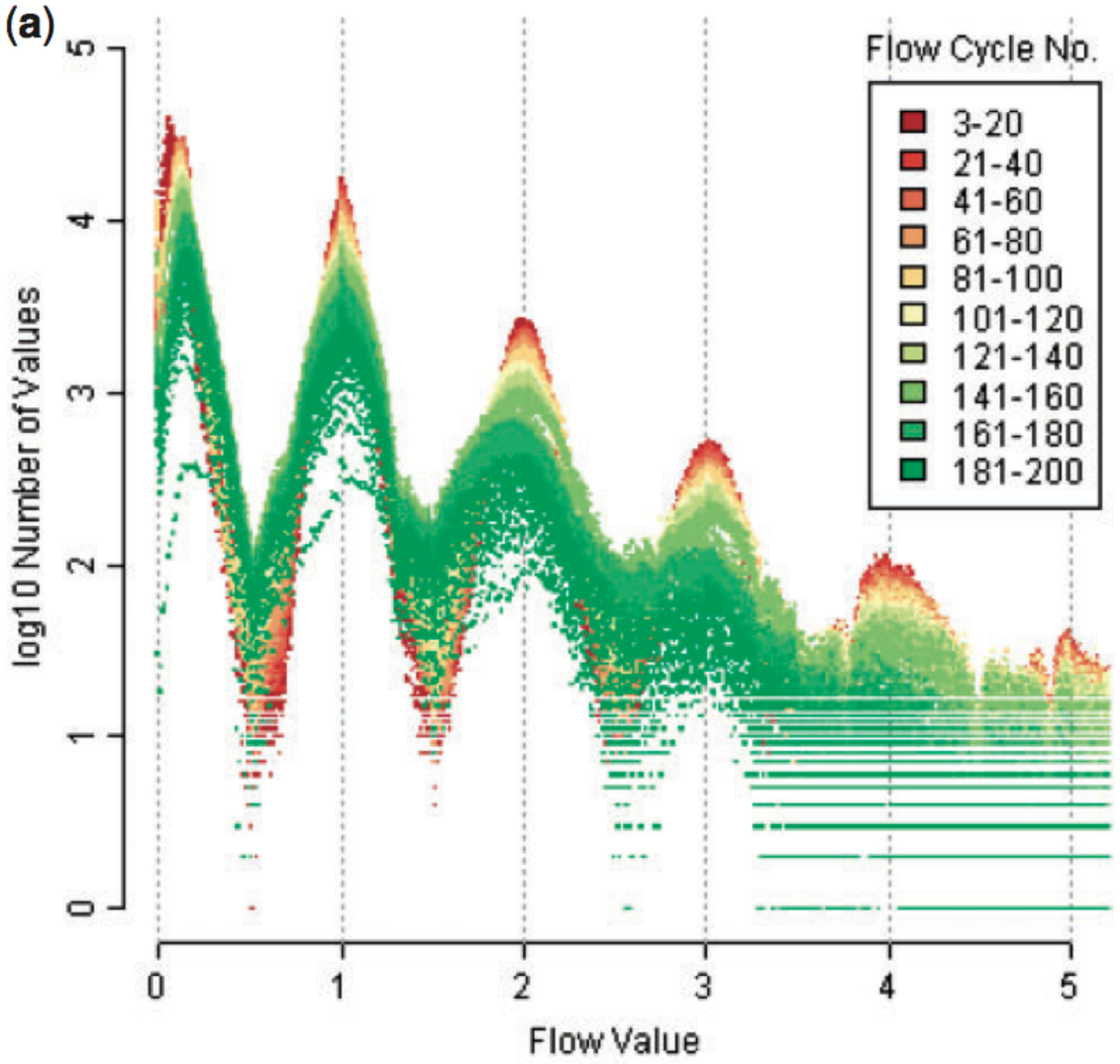


Fig. 1. (**a**) A 454 flowgram: cyclic flowing during one read. The light signal strengths (flow values) are directly translated into homopolymer runs. (**b**) Absolute frequencies of flow values (*E.coli*). Left: original data, no quality-trimming; right: quality-trimmed. The trimming algorithm enhances the separation of the homopolymer length distributions and levels out discrepancies between the nucleotides such that the curves for the four nucleotides are nearly identical.

Balzer et al. 2010

Flow value separation



Balzer et al. 2010

Errors in 454 data

Table 1

Types of error

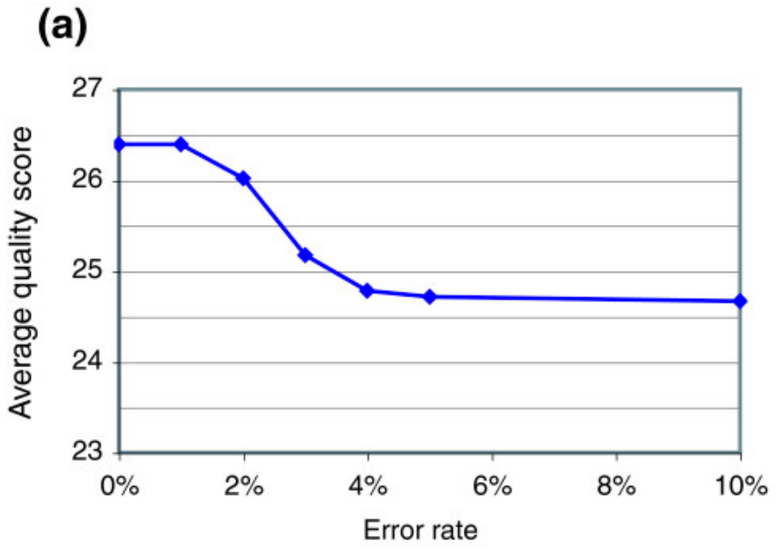
Error type	Number of occurrences	Percent of errors	Error rate
Insertions	58,337	36%	0.18%
Homopolymer extension and CAFIE	32,858	20%	0.10%
Not associated with homopolymers	25,479	16%	0.08%
Deletions	43,107	27%	0.13%
Incomplete homopolymer extension	13,868	9%	0.04%
Not associated with homopolymers	29,239	18%	0.09%
Mismatches	25,281	16%	0.08%
Homopolymer extension and CAFIE	16,725	10%	0.05%
Not associated with homopolymers	8,556	5%	0.03%
Ambiguous base calls (N)	34,184	21%	0.10%
Read errors	**Number of occurrences**	**Cumulative percent of reads**	**Percent of reads**
Reads with no errors (perfect match)	279,468	82%	82%
Reads with no more than one error	35,813	93%	11%
Reads with no more than two errors	11,651	96%	3%
Reads with more than two errors	13,218	100%	4%

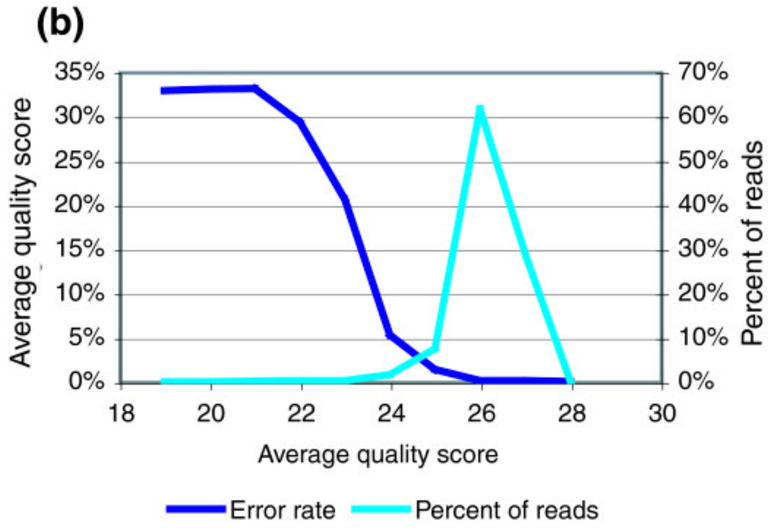
Errors were classified as insertions, deletions, mismatches (substitutions) and ambiguous base calls (Ns). We further classified insertions, deletions and mismatches by their association with homopolymer effects. Deletions corresponding to an adjacent base are considered incomplete extension. Insertions and mismatches of the same base as an adjacent base are extensions. Insertions and mismatches that are the same as an upcoming homopolymer with no more than two intervening bases are considered carry forward errors.

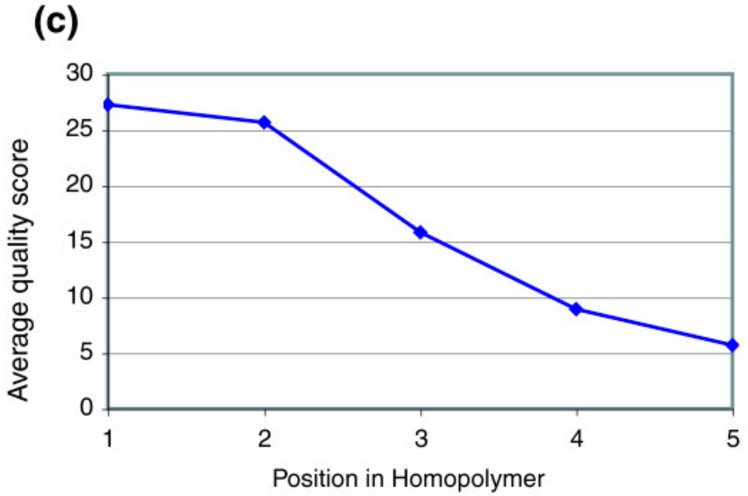
Errors in 454 data

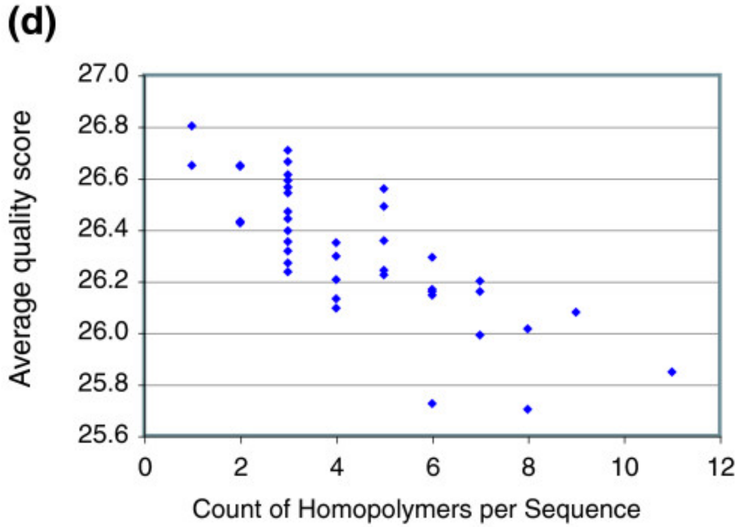
- incomplete homopolymer extension
- carry forward incomplete extension (CAFIE)

Base/Read qualities







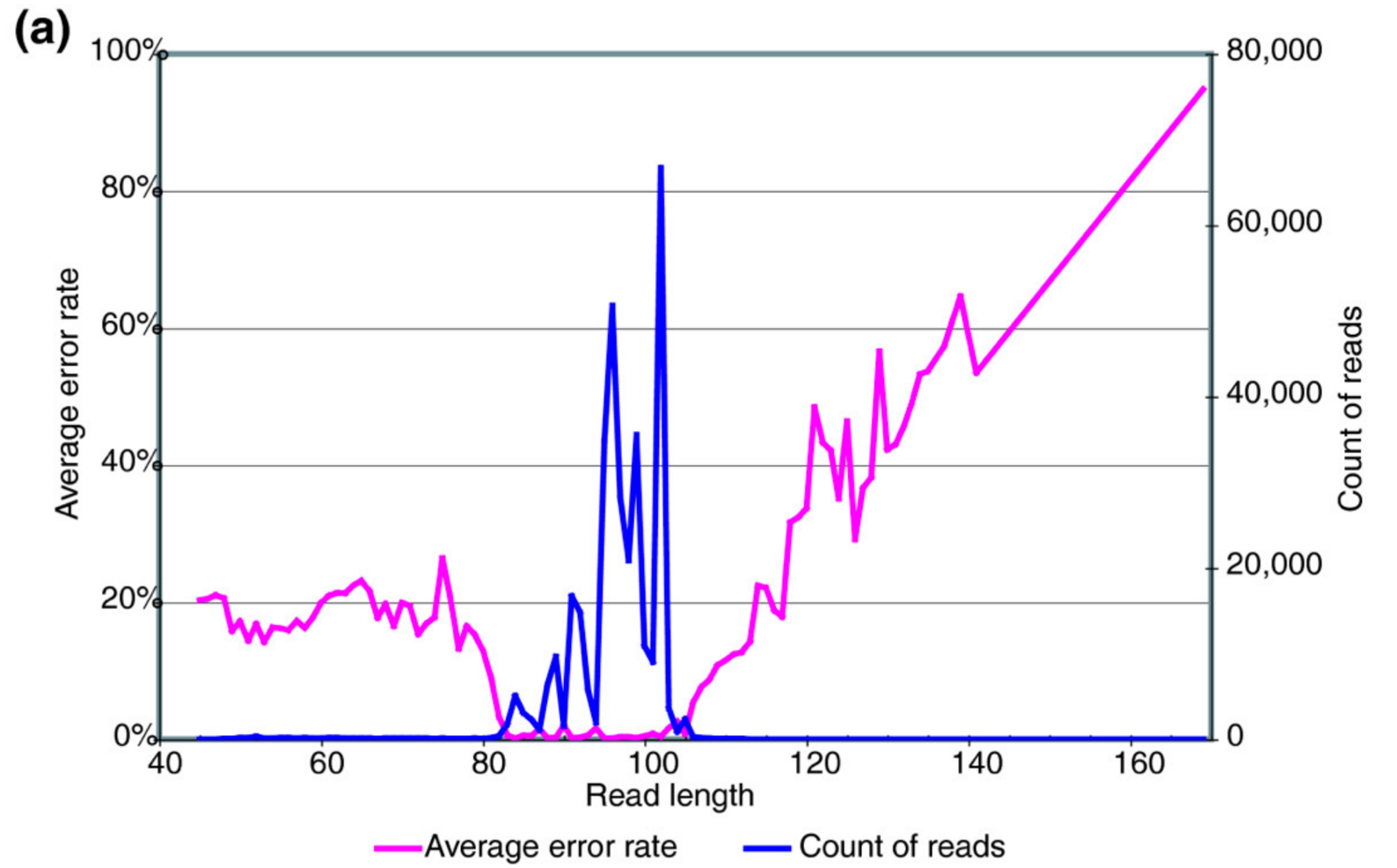


Huse et al. 2007

Length versus Error rate

Figure 3.

Resolution: standard /



Huse et al. 2007