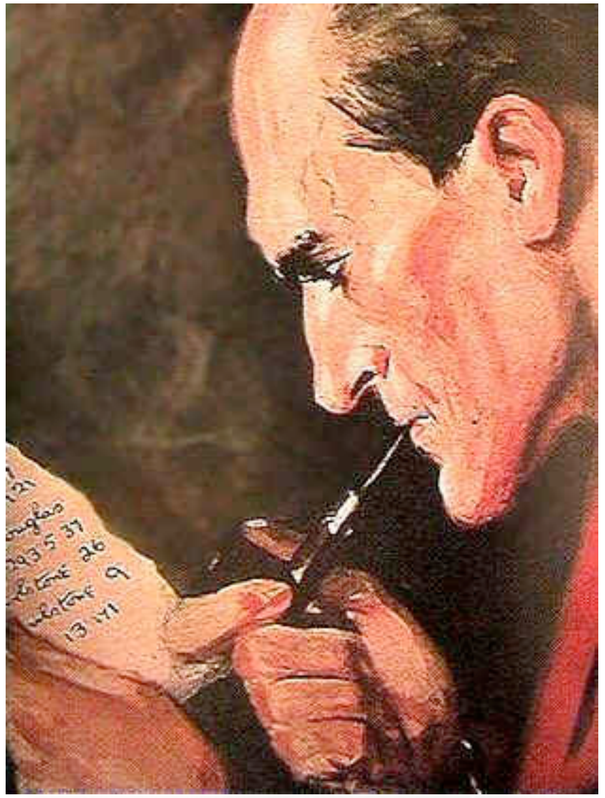
# Genomics

Malay K Basu (malay@uab.edu)

"All science is either physics or stamp collecting..."

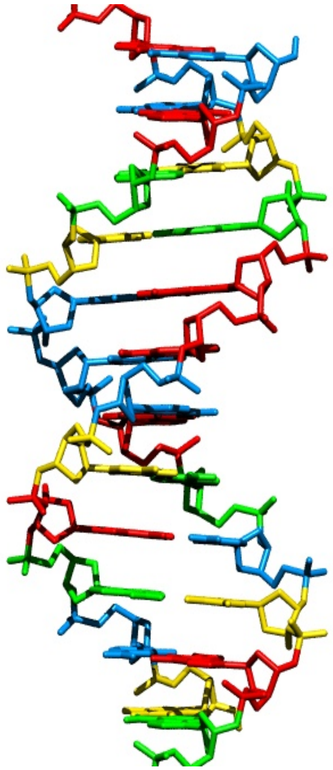
*- Ernest Rutherford*
*"As quoted in Rutherford at Manchester"*



http://www.perkydesigns.com

"Data! Data! Data! ...
I can't make bricks
without the clay"

*- Sherlock Holmes*
*"Adventures of Copper Beeches"*



...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT...

# Evolution of sequencing

# Archaic sequencing methods

## Early 70s: chromatography

# First nucleotide sequencing

Article

*Nature* **237**, 82-88 (12 May 1972) | doi:10.1038/237082a0

## Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein

W. MIN JOU, G. HAEGEMAN, M. YSEBAERT & W. FIERS

1. Laboratory of Molecular Biology and Laboratory of Physiological Chemistry, State University of Ghent, Belgium

**By characterization of fragments, isolated from a nuclease digest of MS2 RNA, the entire nucleotide sequence of the coat gene was established. A "flower"-like model is proposed for the secondary structure. The genetic code makes use of 49 different codons to specify the sequence of the 129 amino-acids long coat polypeptide.**

▲ Top

# First DNA sequencing

*Proc. Nat. Acad. Sci. USA*
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

## The Nucleotide Sequence of the *lac* Operator

**(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)**

WALTER GILBERT AND ALLAN MAXAM

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

*Communicated by J. D. Watson, August 9, 1973*

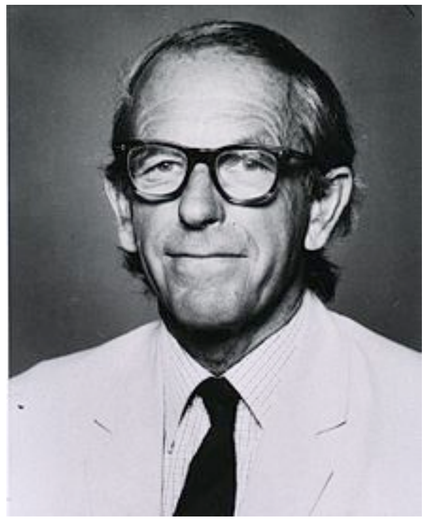
# First Genome Sequence

*(Reprinted from Nature, Vol. 260, No. 5551, pp. 500–507, April 8, 1976)*

# Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene

**W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert & M. Ysebaert**

Laboratory of Molecular Biology, University of Ghent, 9000 Ghent, Belgium



# Sanger dideoxy sequencging

First DNA genome sequenced in 1977: φX174.

*Nature Vol. 265 February 24 1977*

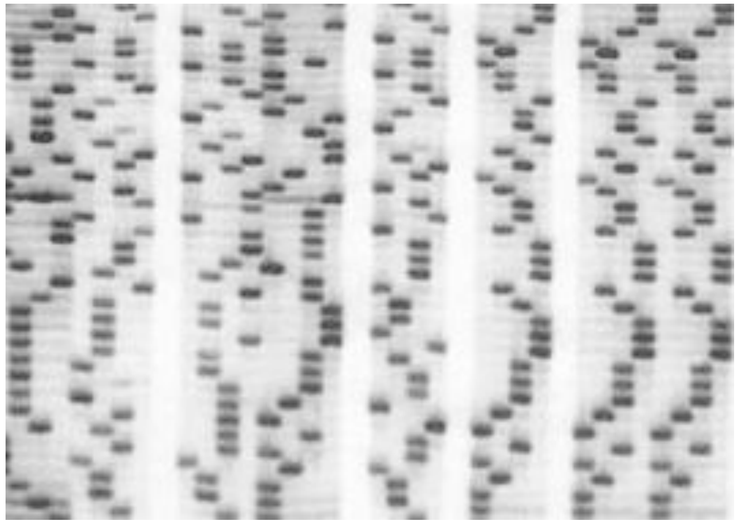


**articles**

## Nucleotide sequence of bacteriophage ΦX174 DNA

**F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith‘**

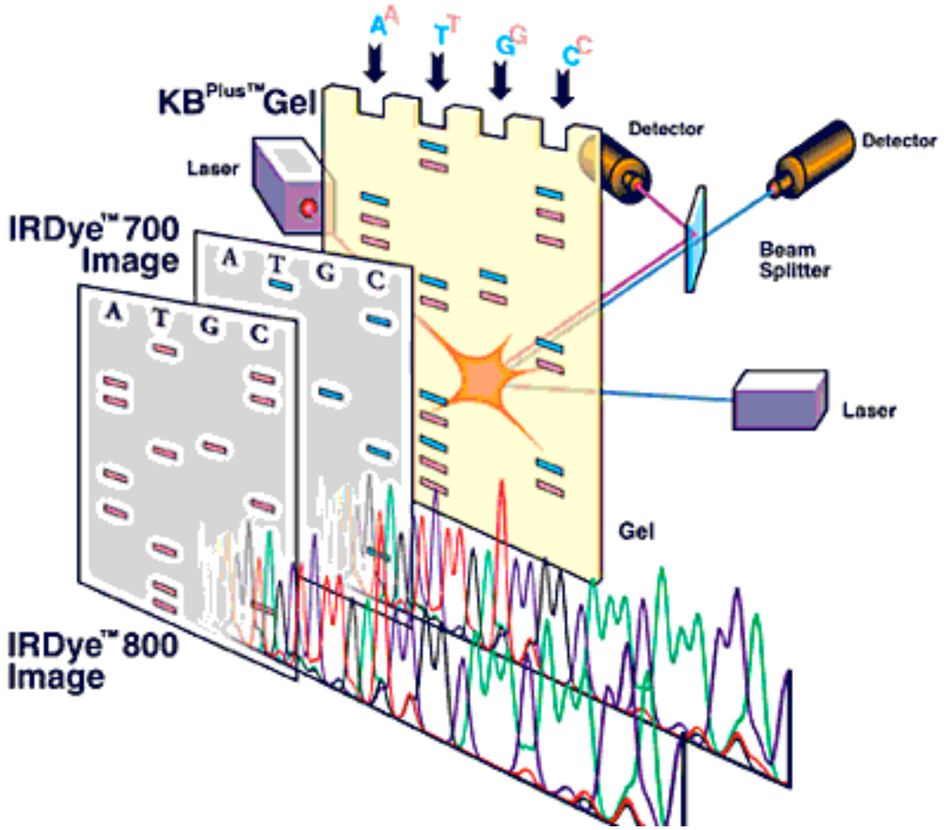
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK



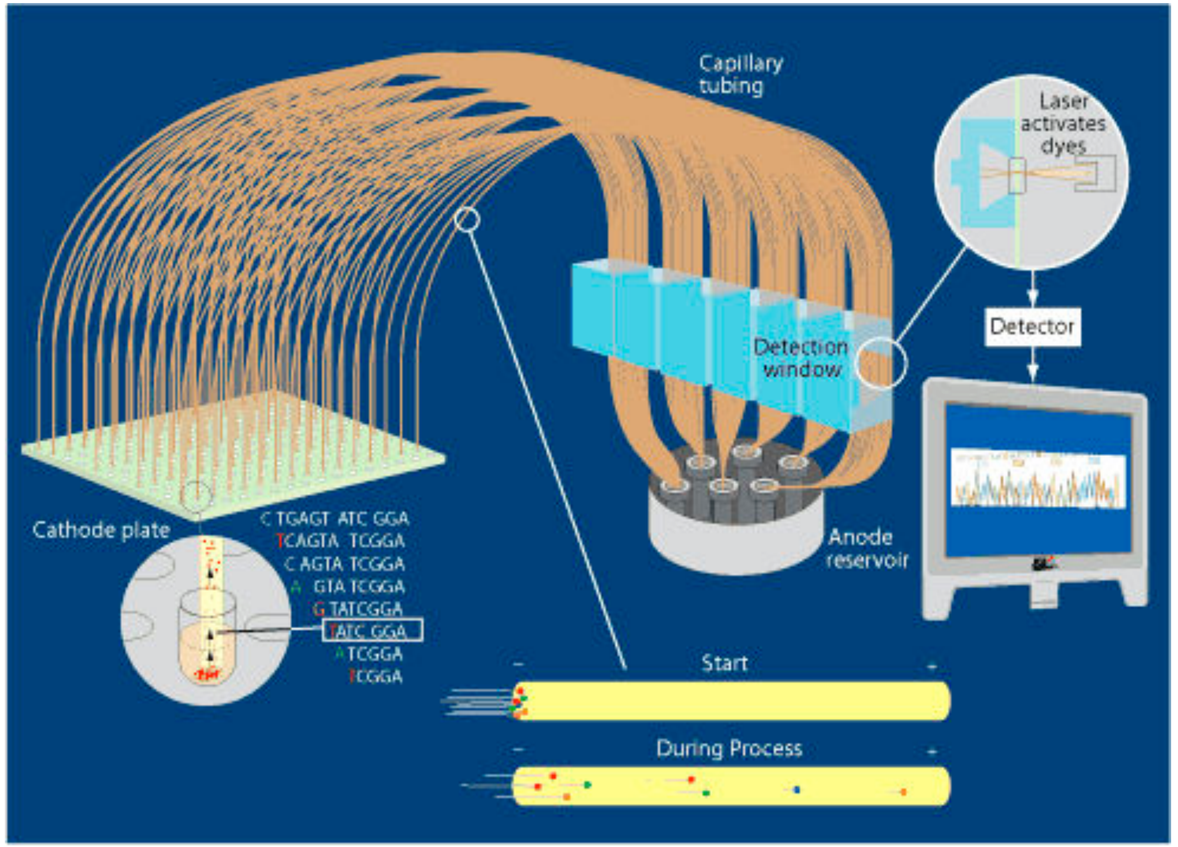
# 1990s: Large scale automated Sequencing

Generation 1: Gel based or capillary

# First automated sequencing



# Capillary Sequencing



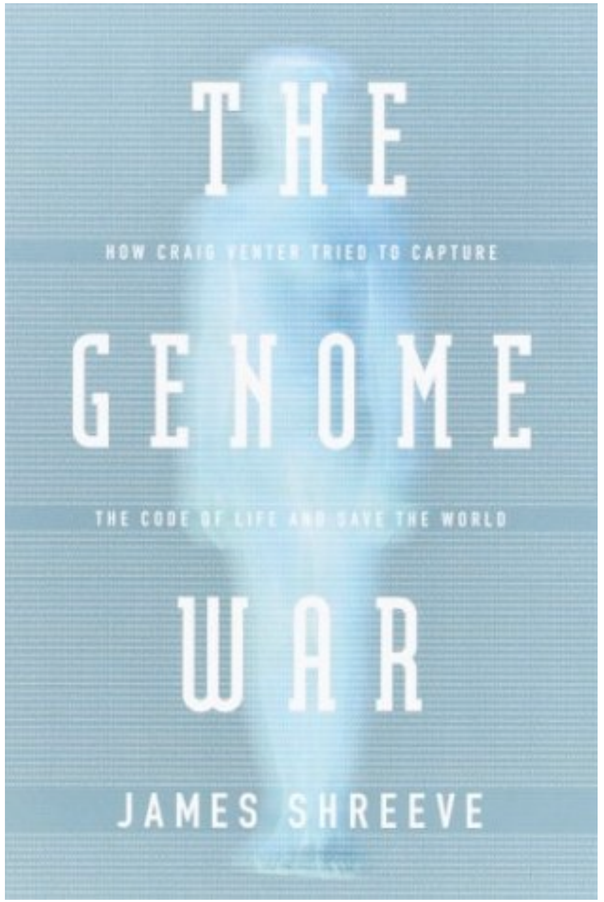
# 1995: *Haemophilus influenza*

RESEARCH ARTICLE

## Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.


THE
HOW CRAIG VENTER TRIED TO CAPTURE
GENOME
THE CODE OF LIFE AND SAVE THE WORLD
WAR
JAMES SHREEVE

# 2001 Human Genome


# Human Genome

Not a single individual

Was  a hack job

Refined over the next 5 yrs

# Human Genome Assembly Information

*Metrics for the current genome assembly.*

Statistics for the current assembly are available below. Information on tiling path files (TPFs) for the human assembly is available at TPF Overview.

## Assembly Statistics for GRCh37.p12

Choose another assembly GRCh37.p1

Chromosome Lengths | Total Lengths | Ungapped Lengths | N50s | Gaps | Counts

Spanned gaps are found within scaffolds and there is some evidence suggesting linkage between the two sequences flanking the gap. Unspanned gaps are found between scaffolds and there is no evidence of linkage.

Primary Assembly | Information By Region

| | Spanned Gaps | | | Unspanned Gaps | | |
|---|---|---|---|---|---|---|
| chr | All Scaffolds | Placed Scaffolds | Unplaced Scaffolds | All Scaffolds | Placed Scaffolds | Unplaced Scaffolds |
| 1 | 19 | 19 | 0 | 22 | 22 | 0 |
| 2 | 3 | 3 | 0 | 15 | 15 | 0 |
| 3 | 0 | 0 | 0 | 7 | 7 | 0 |
| 4 | 1 | 1 | 0 | 12 | 12 | 0 |
| 5 | 1 | 1 | 0 | 6 | 6 | 0 |
| 6 | 6 | 6 | 0 | 8 | 8 | 0 |
| 7 | 9 | 9 | 0 | 8 | 8 | 0 |
| 8 | 1 | 1 | 0 | 9 | 9 | 0 |
| 9 | 15 | 15 | 0 | 29 | 29 | 0 |
| 10 | 8 | 8 | 0 | 12 | 12 | 0 |
| 11 | 4 | 4 | 0 | 11 | 11 | 0 |
| 12 | 1 | 1 | 0 | 8 | 8 | 0 |
| 13 | 0 | 0 | 0 | 1 | 1 | 0 |
| 14 | 0 | 0 | 0 | 5 | 5 | 0 |
| 15 | 2 | 2 | 0 | 10 | 10 | 0 |
| 16 | 1 | 1 | 0 | 10 | 10 | 0 |
| 17 | 2 | 2 | 0 | 5 | 5 | 0 |
| 18 | 2 | 2 | 0 | 7 | 7 | 0 |
| 19 | 1 | 1 | 0 | 8 | 8 | 0 |
| 20 | 2 | 2 | 0 | 9 | 9 | 0 |

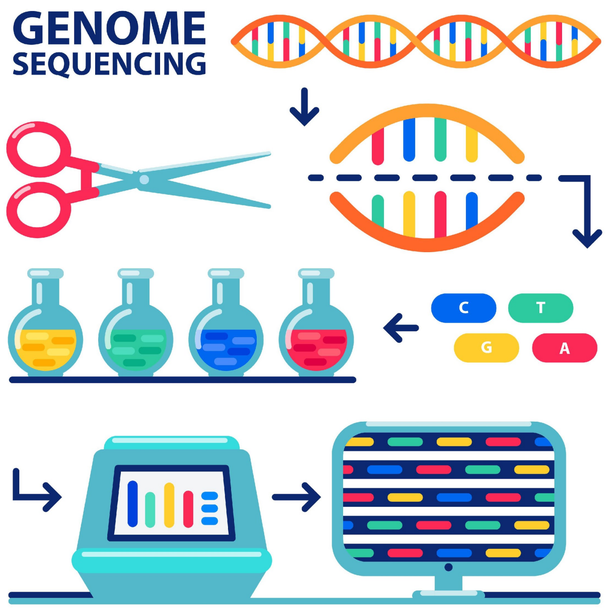
**Global stats for GRCh37.p12**

| | |
|---|---|
| General Info | |
| **Assembly Type** | haploid with alt loci |
| **Release Type** | patch |
| **Number of Assembly Units** | 12 |
| **Total Bases in Assembly** | 3,230,373,980 |
| **Total Non-N Bases in Assembly** | 2,987,105,853 |
| **Primary Assembly N50** | 46,395,641 |
| Region Information | |
| **Total number of defined regions** | 172 |
| **Number of Regions with Alternate Loci** | 3 |
| **Number of Regions with Fix Patches** | 110 |
| **Number of Regions with Novel Patches** | 60 |
| **Number of Regions as PAR** | 4 |
| Alternate Loci/PATCH Information | |
| **Total Number of Alternate Loci scaffolds** | 9 |
| **Number of Alternate Loci scaffolds aligned to the Primary Assembly** | 9 |
| **Number of FIX Patch scaffolds** | 121 |
| **Number of FIX Patch scaffolds aligned to the Primary Assembly** | 121 |
| **Number of NOVEL Patch scaffolds** | 73 |
| **Number of NOVEL Patch scaffolds aligned to the Primary Assembly** | 73 |

Reference assembly

# Next generation sequencing

# Short-read NGS




## Sequencing by Synthesis (SBS)

**Cyclic reversible termination (CRT)**

- HiSeq & MiSeq (Illumina)
- Genereader (Qiagen)

**Single nucleotide addition (SNA)**

- Ion Torrent (Thermo Fisher)
- 454 (Roche)

## Sequencing by Ligation (SBL)

**SOLiD (Applied Biosystems)**

# Massively Parallel Signature Sequencing (MPSS)

Early 1990s: created by Lynx technologies, purchased by Solexa/Illumina

APPLICATIONS OF NEXT-GENERATION SEQUENCING

# Coming of age: ten years of next-generation sequencing technologies

*Sara Goodwin[1], John D. McPherson[2] and W. Richard McCombie[1]*

Abstract | Since the completion of the human genome project in 2003, extraordinary progress has been made in genome sequencing technologies, which has led to a decreased cost per megabase and an increase in the number and diversity of sequenced genomes. An astonishing complexity of genome architecture has been revealed, bringing these sequencing technologies to even greater advancements. Some approaches maximize the number of bases sequenced in the least amount of time, generating a wealth of data that can be used to understand increasingly complex phenotypes. Alternatively, other approaches now aim to sequence longer contiguous pieces of DNA, which are essential for resolving structurally complex regions. These and other strategies are providing researchers and clinicians a variety of tools to probe genomes in greater depth, leading to an enhanced understanding of how genome sequence variants underlie phenotype and disease.

Read
The sequence of bases from a single molecule of DNA.

Sanger sequencing
An approach in which dye-labelled normal deoxynucleotides (dNTPs) and dideoxy-modified dNTPs are mixed. A standard PCR reaction is carried out and, as elongation occurs, some strands incorporate a dideoxy-dNTP, thus terminating elongation. The strands are then separated on a gel and the terminal base label of each strand is identified by laser excitation and spectral emission analysis.

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.*
[2]*Department of Biochemistry and Molecular Medicine; and the Comprehensive Cancer Center, University of California, Davis, California 95817, USA.*
*Correspondence to W.R.M.*
*mccombie@cshl.edu*

doi:10.1038/nrg.2016.49
Published online 17 May 2016

Starting with the discovery of the structure of DNA[1], great strides have been made in understanding the complexity and diversity of genomes in health and disease. A multitude of innovations in reagents and instrumentation supported the initiation of the Human Genome Project[2]. Its completion revealed the need for greater and more advanced technologies and data sets to answer the complex biological questions that arose; however, limited throughput and the high costs of sequencing remained major barriers. The release of the first truly high-throughput sequencing platform in the mid-2000s heralded a 50,000-fold drop in the cost of human genome sequencing since the Human Genome Project[3] and led to the moniker: next-generation sequencing (NGS). Over the past decade, NGS technologies have continued to evolve — increasing capacity by a factor of 100–1,000 (REF. 4) — and have incorporated revolutionary innovations to tackle the complexities of genomes. These advances are providing read lengths as long as some entire genomes, they have brought the cost of sequencing a human genome down to around US$1,000 (as reported by Veritas Genomics)[5], and they have enabled the use of sequencing as a clinical tool[3,6].

Although exciting, these advancements are not without limitations. As new technologies emerge, existing problems are exacerbated or new problems arise. NGS platforms provide vast quantities of data, but the associated error rates (~0.1–15%) are higher and the read lengths generally shorter (35–700 bp for short-read approaches)[7] than those of traditional Sanger sequencing platforms, requiring careful examination of the results, particularly for variant discovery and clinical applications. Although long-read sequencing overcomes the length limitation of other NGS approaches, it remains considerably more expensive and has lower throughput than other platforms, limiting the widespread adoption of this technology in favour of less-expensive approaches. Finally, NGS is also competing with alternative technologies that can carry out similar tasks, often at lower cost (BOX 1); it is not clear how these disparate approaches to genomics, medicine and research will interact in the years to come.

This Review evaluates various approaches used in NGS and how recent advancements in the field are changing the way genetic research is carried out. Details of each approach along with its benefits and drawbacks are discussed. Finally, various emerging applications within this field and its exciting future are explored.

## Short-read NGS

***Overview of clonal template generation approaches.*** Short-read sequencing approaches fall under two broad categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS). In SBL approaches, a probe sequence that is bound to a fluorophore hybridizes to a DNA fragment and is ligated to an adjacent oligonucleotide for imaging. The emission spectrum of the fluorophore indicates the identity of the base or bases complementary to specific positions within the probe. In SBS approaches, a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into





Chapter 8

# DNA Sequencing



Investigatory approaches that study whole or partial genomes, define the field of **genomics**. The last ten years has changed scientific discovery since the entire genomic sequence for thousands of different organisms are available for analysis. In fact, sequencing the human genome has gone from a monumental task that involved multiple sequencing centers and millions of dollars into a simple task that can be done for a few thousand US dollars and a few hours. In this chapter, we will survey the methods used to sequence DNA, and in the following chapter we will consider the assembly of whole genome sequences. This chapter will explain regular chain termination sequencing, a procedure used by many to double-check plasmid assembly, and sequence shorter segments of DNA. Then the chapter will focus on **next generation sequencing (NGS)** technologies, explaining the basic procedure for some of the top selling technologies in use as of writing. Technological advancements in NGS are very fast, and therefore, the discussion will focus more on a conceptual understanding and then introduce some details and nuances of the more popular technologies.

**genomics** Study of the structure, function, evolution, and mapping of genomes.
**next generation sequencing (NGS)** The term to describe experimental techniques to simultaneously decode the order of bases for millions of genomic DNA fragments.



**Molecular Biology. DOI:** https://doi.org/10.1016/B978-0-12-813288-3.00008-2

# Illumina Video

https://www.youtube.com/watch?v=womKfikWlxM

# Long-read NGS technologies

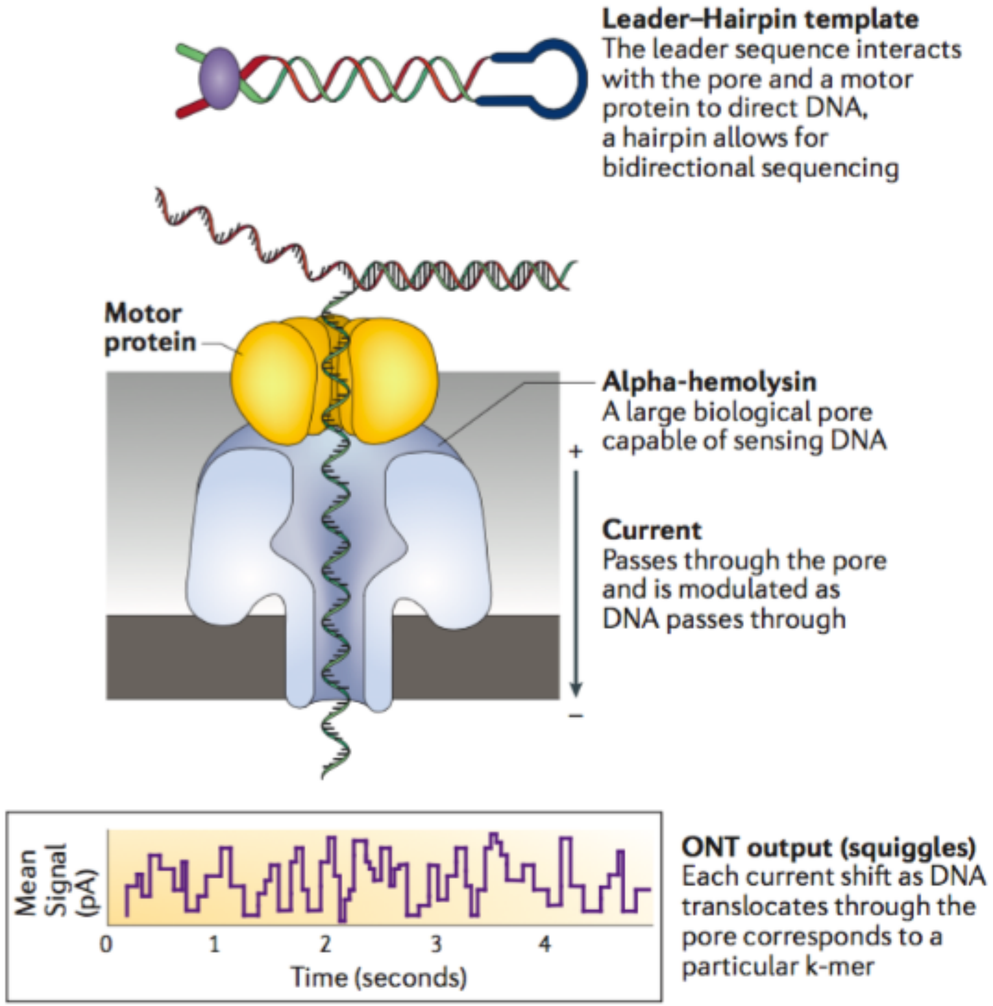
Table 1 | **Long-range sequencing and mapping platforms**

| Platform | General characteristics and costs | Major applications | Bioinformatics challenges |
|---|---|---|---|
| PacBio SMRT sequencing | Single-molecule long reads averaging ~10 kb with some approaching 100 kb; several fold more expensive than short reads | De novo genome assembly, structural variant detection, gene isoform resolution and epigenetic modifications | Raw reads have high error rates dominated by false insertions; requires new alignment and error correction algorithms |
| Oxford Nanopore sequencing | Single-molecule long reads averaging ~10 kb with some >1 Mb; several fold more expensive than short reads | De novo genome assembly, structural variant detection, gene isoform resolution and epigenetic modifications | Raw reads have high error rates dominated by false deletions and homopolymer errors; requires new alignment and error correction algorithms |
| 10X Genomics Chromium | Linked reads spanning ~100 kb derived from a collection of short-read sequences; moderately more expensive than short reads | De novo genome assembly and scaffolding, phasing, detection of large structural variants (>10 kb) and single-cell gene expression | Sparse sequencing rather than true long reads; more complicated to align, with poorer resolution of locally repetitive sequences |
| Hi-C-based analysis | Pairs of short reads (<100 bp) formed from crosslinking chromatin interactions; moderately more expensive than short reads | Genome scaffolding and phasing | Sparse sequencing with highly variable genomic distance between pairs (1 kb to 1 Mb or longer) |
| BioNano Genomics optical mapping | Optical mapping of long DNA molecules (~250 kb or longer) labelled with fluorescent probes; less expensive than short reads | Genome scaffolding and detection of large structural variants (>10 kb) | Limited algorithms to discover high-confidence alignment between an optical map and a sequence assembly |

PacBio SMRT, Pacific Biosciences single-molecule real time.

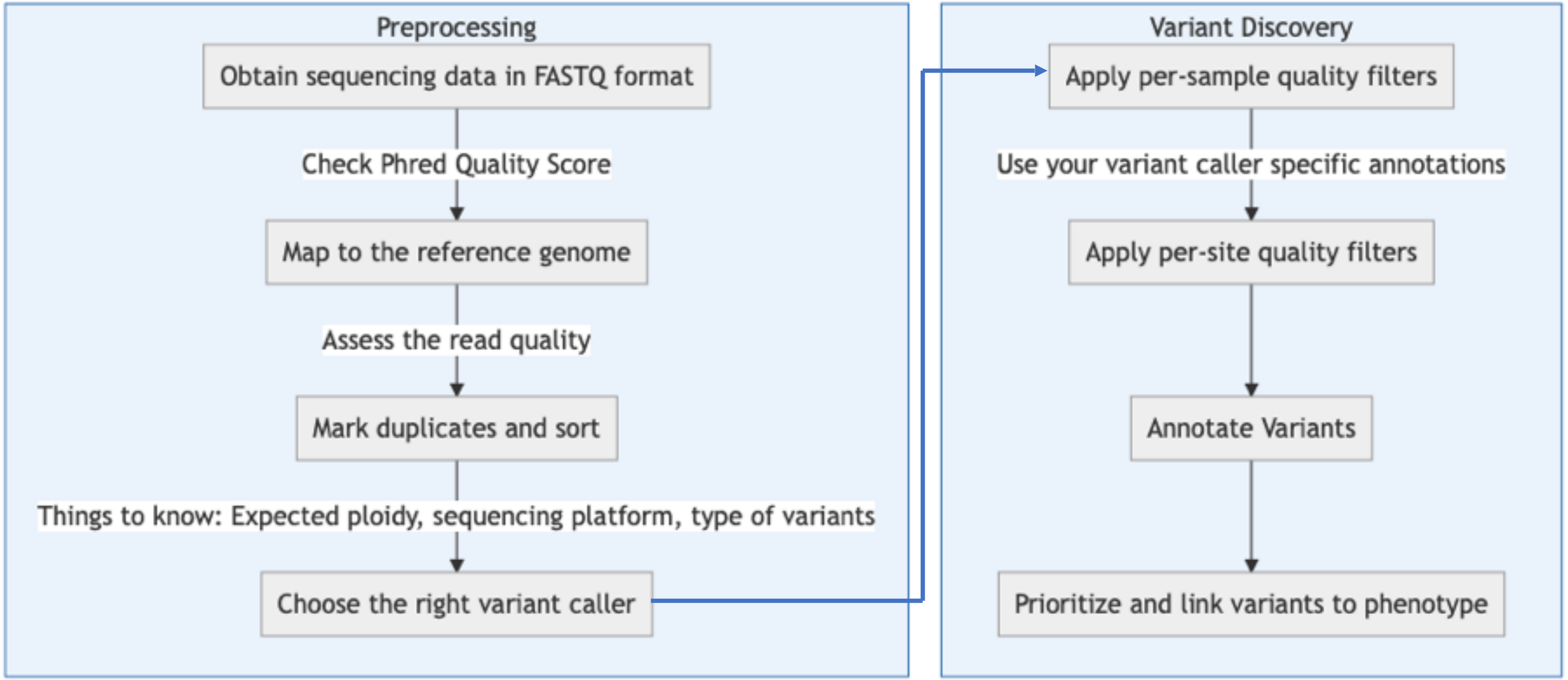
*(Sedlazeck et al., Nat Rev Genet, 2018)*

# Oxford Nanopore sequencing




*(Goodwin et al., Nat Rev Genet, 2016)*

# NGS analysis workflow methodology overview



# Steps to Assemble a Genome

**Some Terminology**

***read*** a 500-900 long word that comes out of sequencer

***mate pair*** a pair of reads from two ends of the same insert fragment

***contig*** a contiguous sequence formed by several overlapping reads with no gaps

***supercontig*** (scaffold) an ordered and oriented set of contigs, usually by mate pairs

***consensus sequene*** sequence derived from the multiple alignment of reads in a contig

..ACGATTACAATAGGTT..

# Definition of Coverage


Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

**Definition:** Coverage **C = N L / G**

How much coverage is enough?

**Lander-Waterman model:** **Prob[ not covered bp ] = $e^{-C}$**
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

# Draft sequencing of full genome

6 to 8X coverage

# SNP finding

>= 20x coverage

# Assembly

Join reads to larger sequence: "contigs".

Reference based assembly

De Novo assembly

# Publicly available de novo assemblers

Phrap (www.phrap.org)

Celera (wgs-assembler.sf.net)

Paracel (www.paracel.com)

Arachne (ftp://ftp.broadinstitute.org/pub/crd/ARACHNE/)

CAP3 (http://seq.cs.iastate.edu/)

# Gene prediction

Evidence based gene calling: BLAST

Ab initio gene calling; no homolog required:
GeneMark, Glimmer, MetaGene.

# ORFans

Open Reading Frame (ORFs) with no similarity to any sequence in the database.

# Annotation

Finding function of a gene

# Next-gen sequencing

Whole Genome Sequencing

RNA-Seq

Exome

Chip-Seq

Methylation (Bisulfite sequencing)

# Lior Pachter's list

https://liorpachter.wordpress.com/seq/

# Read alignment

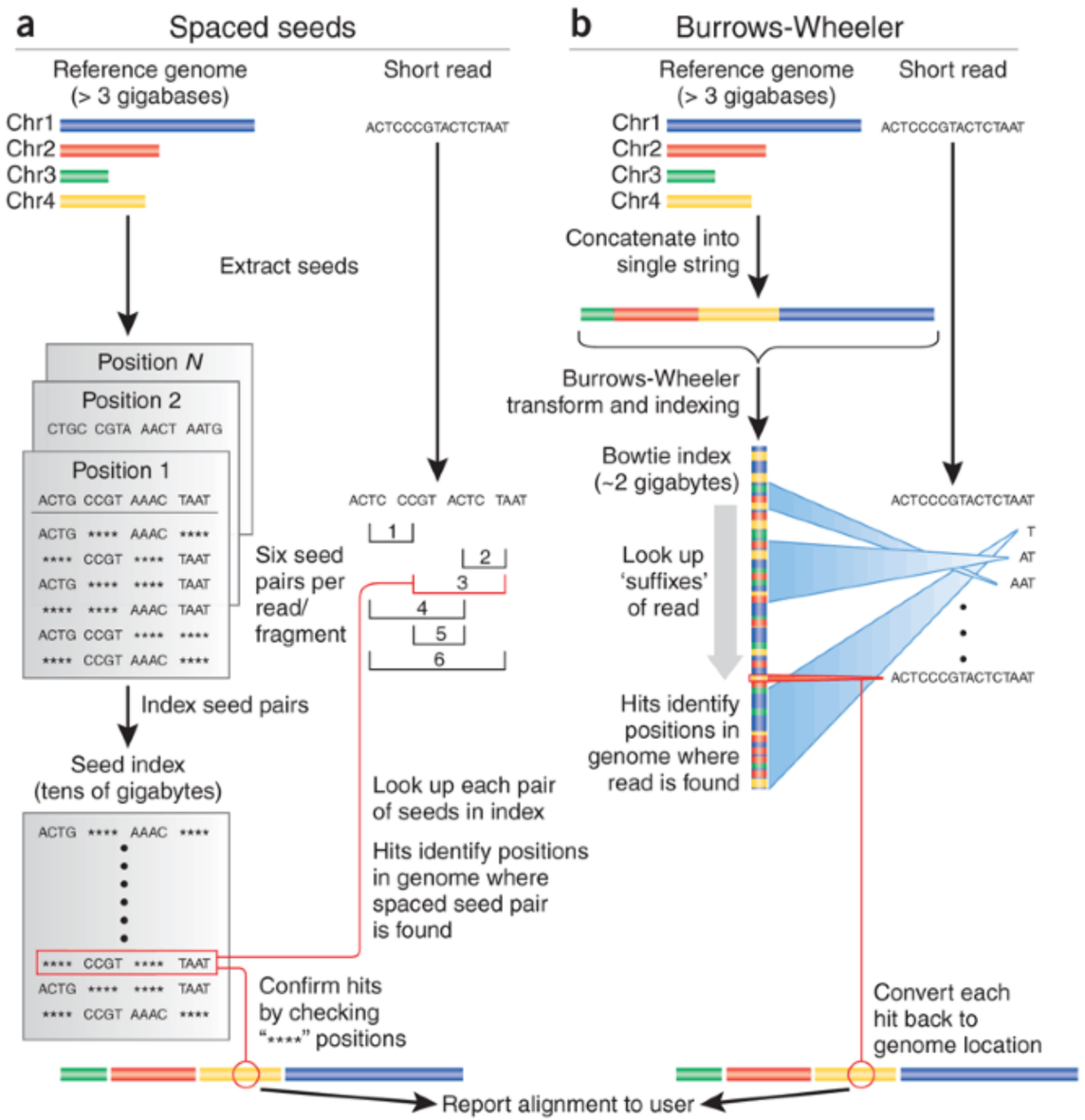
Input: FASTQ

Output: SAM/BAM Format

Methods: Seeds, Suffix Array, Burrows-Wheeler Transform and LF-mapping

Purpose: Map our reads

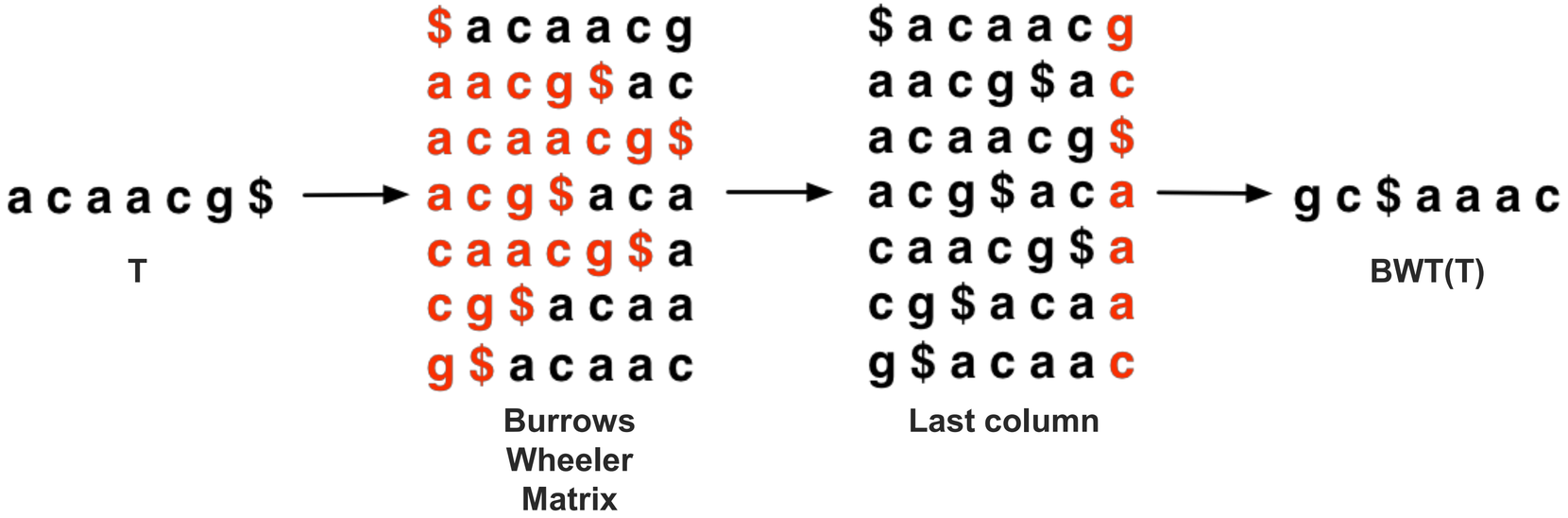
# Read alignment




Trapnell & Salzberg, 2009

# The Burrows-Wheeler Transform is a reversible representation with handy properties

- Sort all the possible rotations of original string

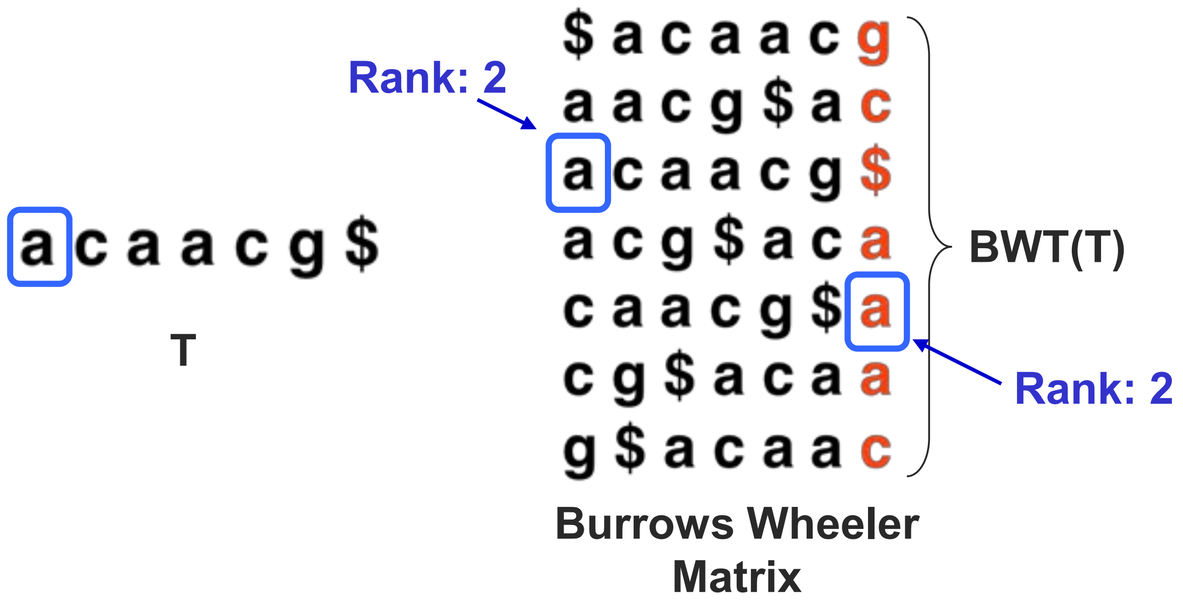



- Once BWT(T) is built, *all else shown here is discarded*
  - Matrix will be shown for illustration only

Burrows M, Wheeler DJ: **A block sorting lossless data compression algorithm**. *Digital Equipment Corporation, Palo Alto, CA* 1994, Technical Report 124; 1994

# A text occurrence has the same rank in the first and last columns

- When we rotate left and sort, the first character retains its rank. Thus the same text occurrence of a character has the same rank in the **L**ast and **F**irst columns.



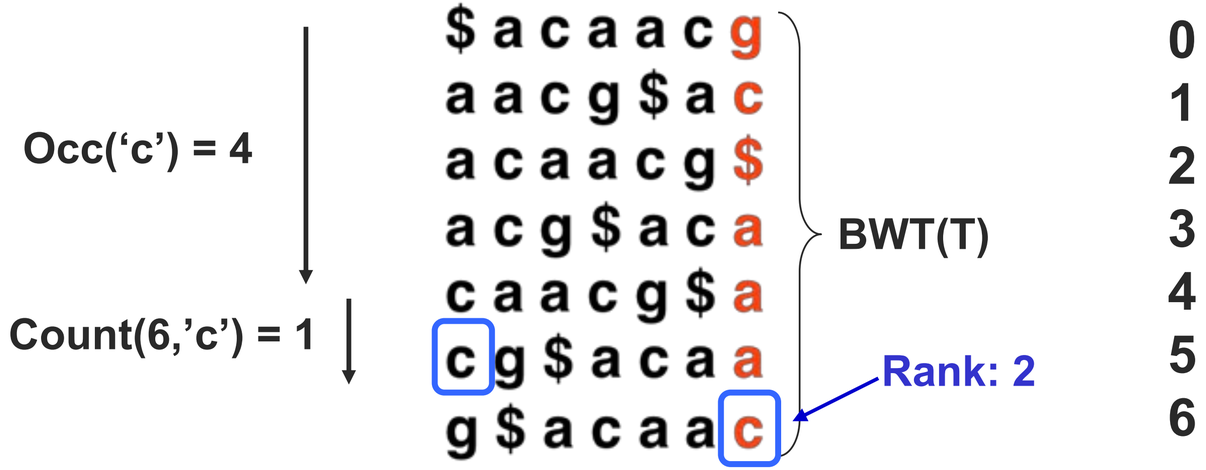

Courtesy of Ben Langmead. Used with permission.



http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

# The Last to First (LF) function matches character and rank

**LF(6, ‘c’) = Occ(‘c’) + Count(6,’c’) = 5**




**Occ(qc) – Number of characters lexically smaller than qc in BWT(T)**

**Count(idx, qc) – Number of qc characters before position idx in BWT(T)**

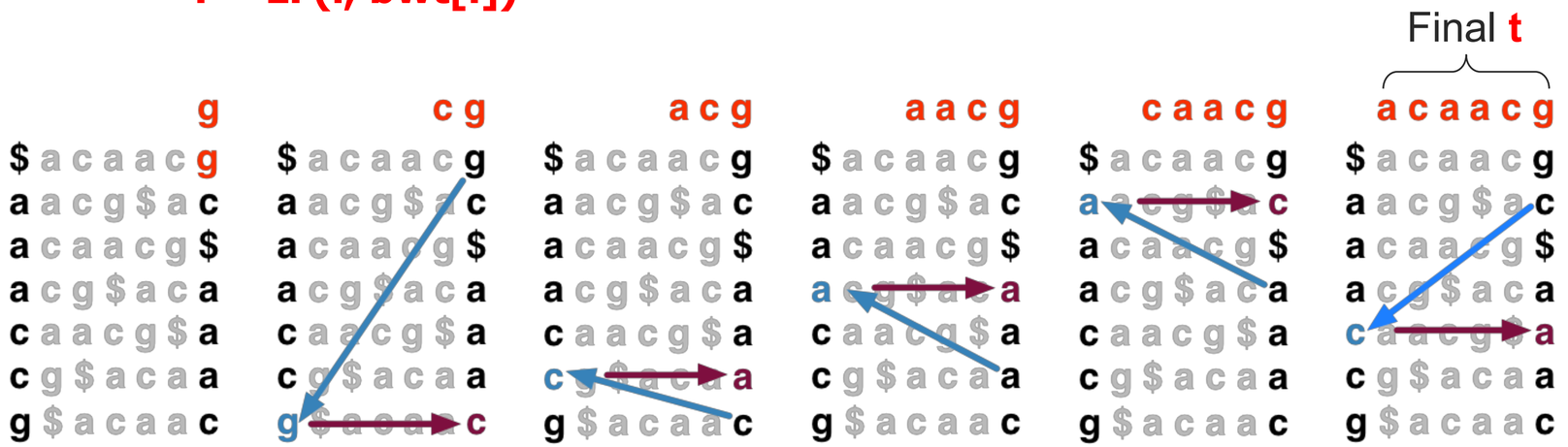
Courtesy of Ben Langmead. Used with permission.



http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

# The Walk Left Algorithm inverts the BWT

```
i = 0
t = ""
while bwt[i] != '$':
    t = bwt[i] + t
    i = LF(i, bwt[i])
```




Courtesy of Ben Langmead. Used with permission.

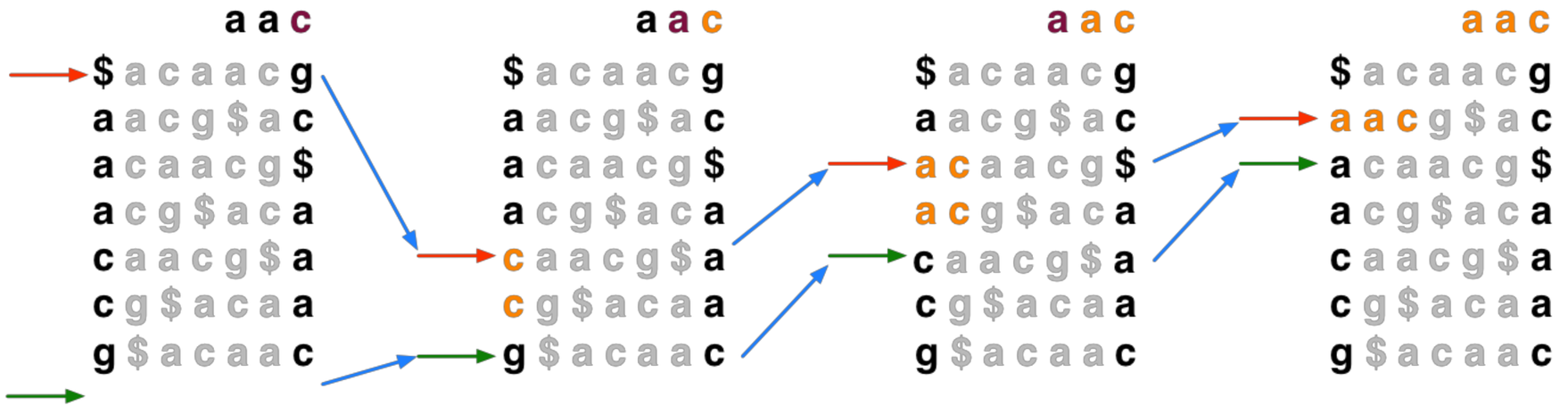


http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

# Exact Matching with FM Index

q = "aac"
top = 0
bot = len(bwt)
for qc in reverse(q):
  top = LF(top, qc)
  bot = LF(bot, qc)

**In each iteration top & bot delimit the range of rows beginning with progressively longer suffixes of q**

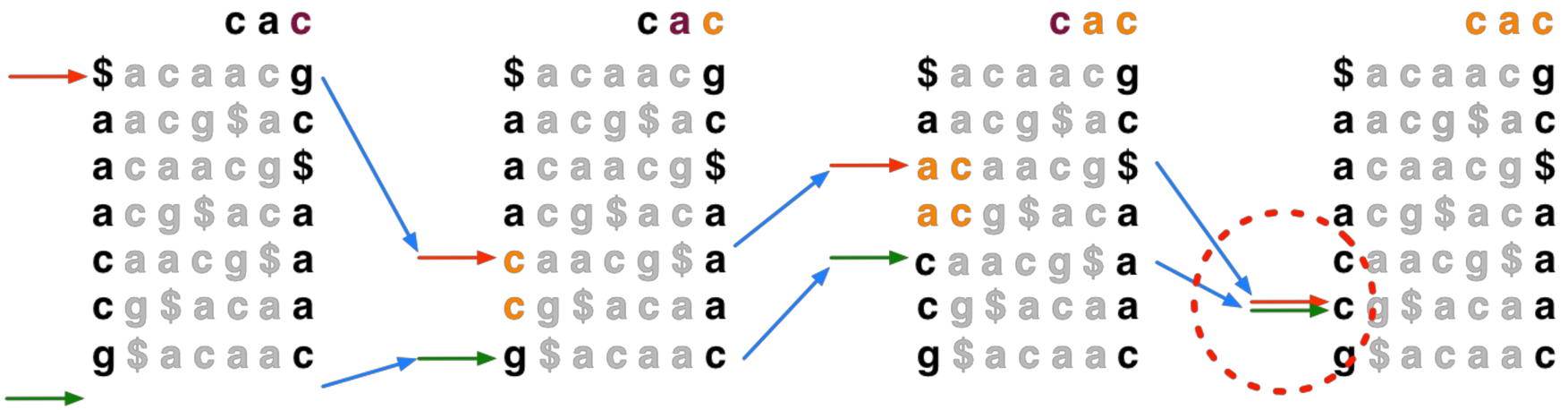



Courtesy of Ben Langmead. Used with permission.



http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

# Exact Matching with FM Index




- If range becomes empty (**top** = **bot**) the query suffix (and therefore the query) does not occur in the text

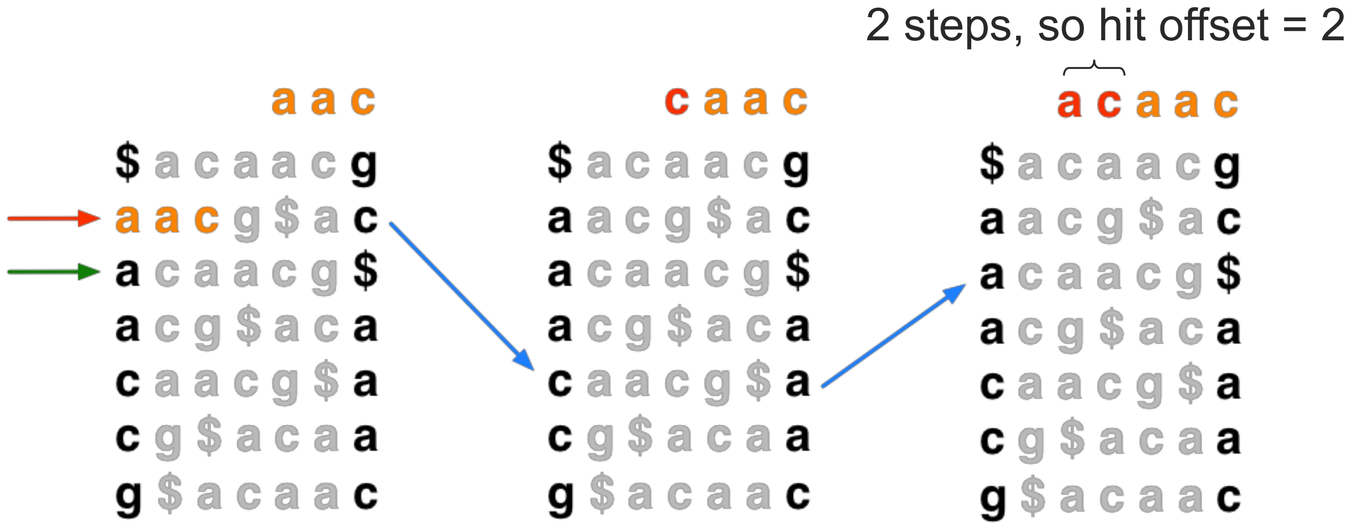
Courtesy of Ben Langmead. Used with permission.



http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

# Rows to Reference Positions

- Naïve solution 1: Use “walk-left” to walk back to the beginning of the text; number of steps = offset of hit




- Linear in length of text in general – too slow

Courtesy of Ben Langmead. Used with permission.



http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

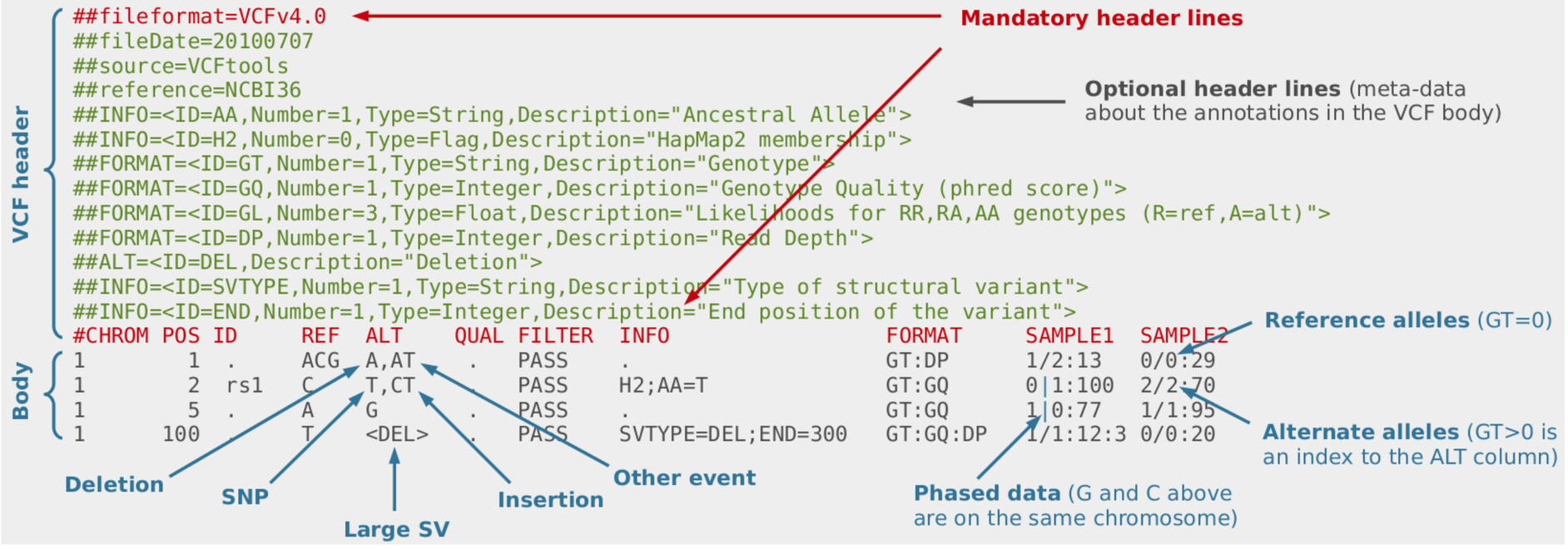
# Variant calling

Input: SAM/BAM

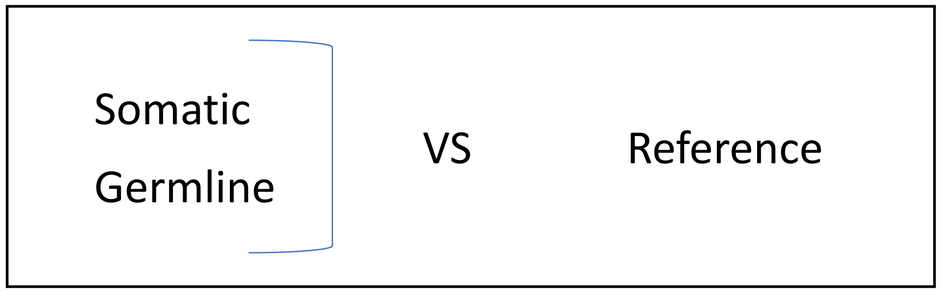
Output: VCF, Variant Call Format

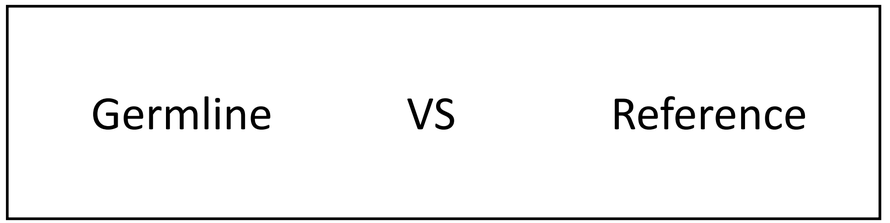
Methods: Naive Variant Calling, Bayesian Variant Calling, Heuristic Variant Calling

Purpose: Clean the noise



# Somatic vs Germline Variant Calling



# Data Formats

FASTQ

- Unaligned read sequences with base qualities

SAM/BAM

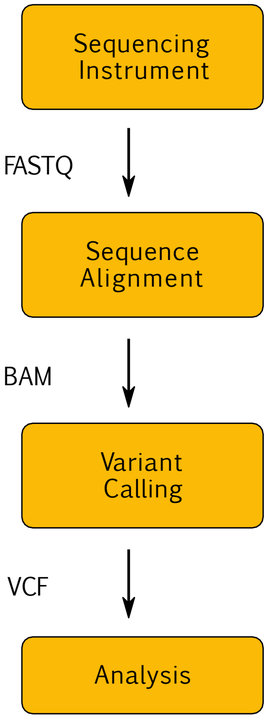
- Unaligned or aligned reads
- Text and binary formats

CRAM

- Better compression than BAM

VCF/BCF

- Flexible variant call format
- Arbitrary types of sequence variation
- SNPs, indels, structural variations




Specifications maintained by the Global Alliance for Genomics and Health

# FASTQ

- Simple format for raw unaligned sequencing reads
- Extension to the FASTA file format
- Sequence and an associated per base quality score

```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBBABAAAABBBBB=@>B
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC
+
BBABBBABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A
```

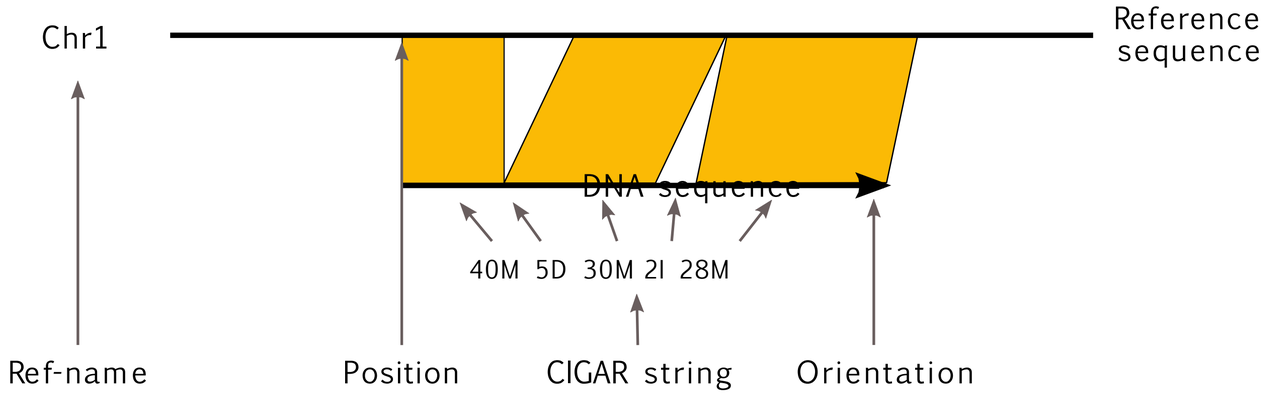
- Quality encoded in ASCII characters with decimal codes 33-126
  - ASCII code of “A” is 65, the corresponding quality is Q=65-33=32
  - Phred quality score: $P = 10^{-Q/10}$

    ```
    perl -e 'printf "%d\n",ord("A")-33;'
    ```
- Beware: multiple quality scores were in use!
  - Sanger, Solexa, Illumina 1.3+
- Paired-end sequencing produces two FASTQ files

# SAM / BAM

SAM (Sequence Alignment/Map) format

- Unified format for storing read alignments to a reference genome
- Developed by the 1000 Genomes Project group (2009)
- One record (a single DNA fragment alignment) per line describing alignment between fragment and reference
- 11 fixed columns + optional key:type:value tuples




Note that BAM can contain

- unmapped reads
- multiple alignments of the same read
- supplementary (chimeric) reads

# SAM

## SAM fields

| | | |
|---|---|---|
| 1 | QNAME | Query NAME of the read or the read pair |
| 2 | FLAG | Bitwise FLAG (pairing, strand, mate strand, etc.) |
| 3 | RNAME | Reference sequence NAME |
| 4 | POS | 1-Based leftmost POSition of clipped alignment |
| 5 | MAPQ | MAPping Quality (Phred-scaled) |
| 6 | CIGAR | Extended CIGAR string (operations: MIDNSHPX=) |
| 7 | MRNM | Mate Reference NaMe ('=' if same as RNAME) |
| 8 | MPOS | 1-Based leftmost Mate POSition |
| 9 | ISIZE | Inferred Insert SIZE |
| 10 | SEQ | Query SEQuence on the same strand as the reference |
| 11 | QUAL | Query QUALity (ASCII-33=Phred base quality) |
| 12- | OTHER | Optional fields |

```
$ samtools view -h file.bam | less

@HD VN:1.0  GO:none SO:coordinate

@SQ SN:1        LN:249250621        UR:hs37d5.fa.gz AS:NCBI37           M5:1b22b98cdeb4a9304cb5d48026a85128 SP:Human
@SQ SN:2        LN:243199373        UR:hs37d5.fa.gz AS:NCBI37           M5:a0d9851da00400dec1098a9255ac712e SP:Human
@RG ID:1        PL:ILLUMINA PU:13350_1  LB:13350_1  SM:13350_1  CN:SC
@PG ID:bwa  PN:bwa  VN:0.7.10-r806          CL:bwa mem hs37d5.fa.gz 13350_1_1.fq 13350_1_1.fq
1:2203:10256:56986  97  1   9998            0   106M45S =  10335       0 \
     CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAACCCT[...]CAAACCCACCCCCAAACCCAAAACCTCACCAC                    \
     FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJJJFF[...]<---F-----A7-J-<J-A--77AF---J7--    \       MD:Z:1G24C2A76
     PG:Z:MarkDuplicates RG:Z:1  NM:i:3  MQ:i:0  AS:i:94 XS:i:94
```

# CIGAR string

## Compact representation of sequence alignment

| | |
|---|---|
| M | alignment match or mismatch |
| = | sequence match |
| X | sequence mismatch |
| I | insertion to the reference |
| D | deletion from the reference |
| S | soft clipping (clipped sequences present in SEQ) |
| H | hard clipping (clipped sequences NOT present in SEQ) |
| N | skipped region from the reference |
| P | padding (silent deletion from padded reference) |

## Examples:

```
Ref:                    ACGTACGTACGTACGT
Read:       ACGT----ACGTACGA
Cigar:      4M 4D 8M

Ref:                    ACGT----ACGTACGT
Read:       ACGTACGTACGTACGT
Cigar:      4M 4I 8M

Ref:        ACTCAGTG--GT
Read:       ACGCA-TGCAGTtagacgt
Cigar:      5M 1D 2M 2I 2M 7S
```

# Flags

| Hex | Dec | Flag | Description |
|---|---|---|---|
| 0x1 | 1 | PAIRED | paired-end (or multiple-segment) sequencing technology |
| 0x2 | 2 | PROPER_PAIR | each segment properly aligned according to the aligner |
| 0x4 | 4 | UNMAP | segment unmapped |
| 0x8 | 8 | MUNMAP | next segment in the template unmapped |
| 0x10 | 16 | REVERSE | SEQ is reverse complemented |
| 0x20 | 32 | MREVERSE | SEQ of the next segment in the template is reversed |
| 0x40 | 64 | READ1 | the first segment in the template |
| 0x80 | 128 | READ2 | the last segment in the template |
| 0x100 | 256 | SECONDARY | secondary alignment |
| 0x200 | 512 | QCFAIL | not passing quality controls |
| 0x400 | 1024 | DUP | PCR or optical duplicate |
| 0x800 | 2048 | SUPPLEMENTARY | supplementary alignment |

## Bit operations made easy

- python

  0x1 | 0x2 | 0x20 | 0x80 .. 163

  bin(163) .. 10100011

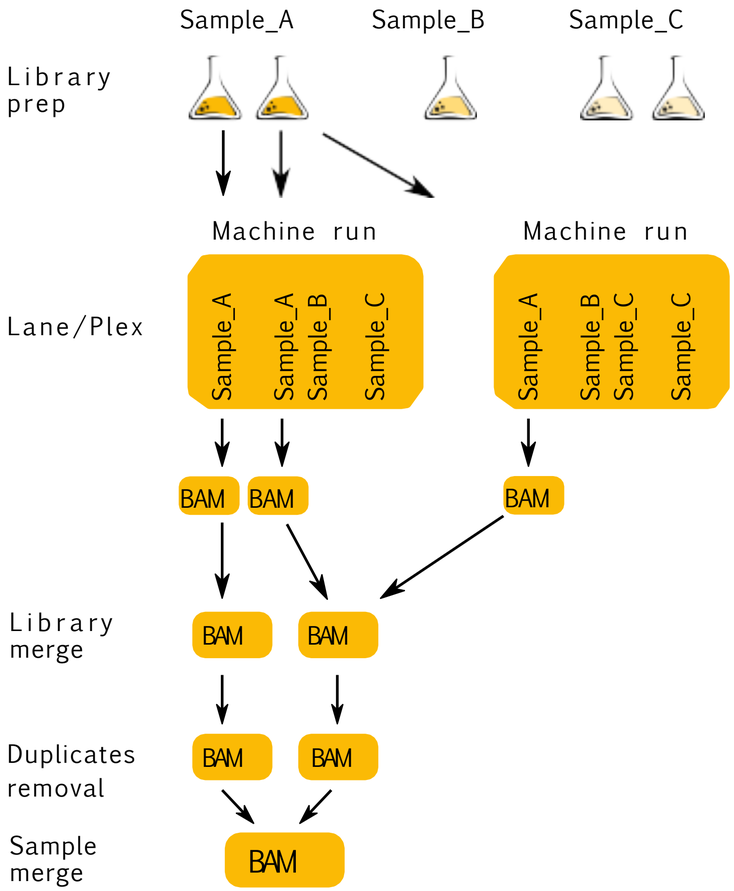
- samtools flags

  0xa3 163 PAIRED,PROPER_PAIR,MREVERSE,READ2

# Optional tags

Each lane has a unique RG tag that contains meta-data for the lane RG tags

- ID: SRR/ERR number
- PL: Sequencing platform
- PU: Run name
- LB: Library name
- PI: Insert fragment size
- SM: Individual
- CN: Sequencing center



# BAM

**BAM (Binary Alignment/Map) format**

- Binary version of SAM
- Developed for fast processing and random access
  - BGZF (Block GZIP) compression for indexing

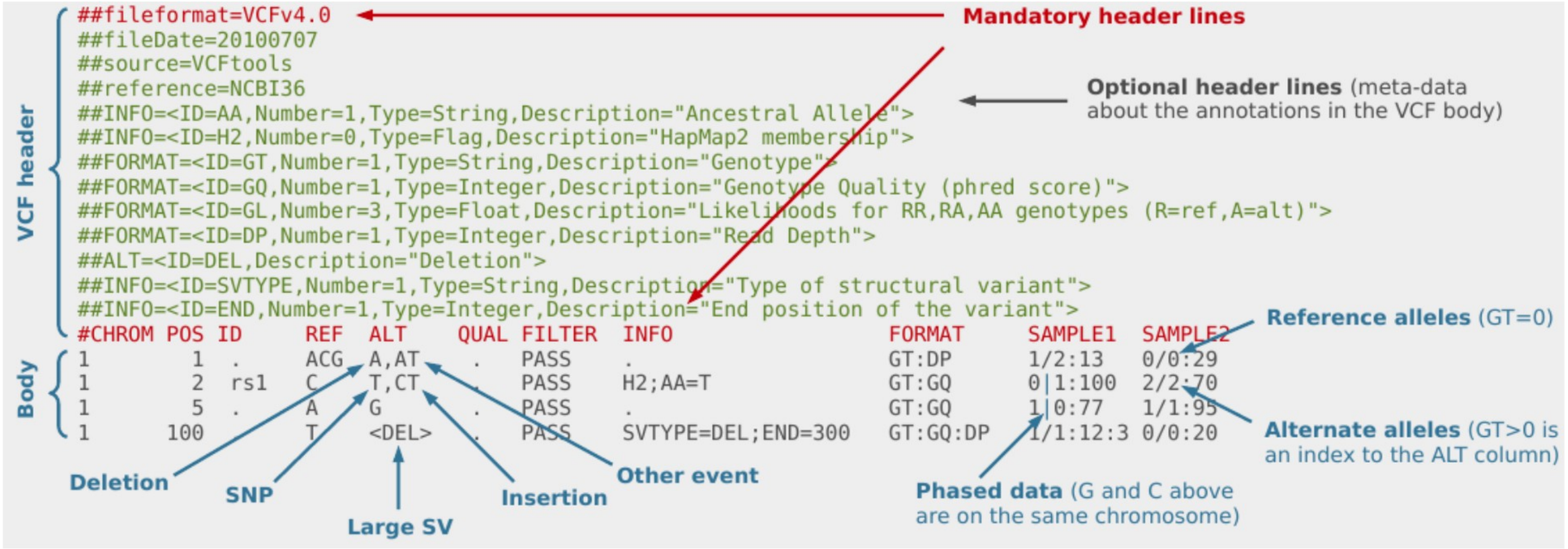
**Key features**

- Can store alignments from most mappers
- Supports multiple sequencing technologies
- Supports indexing for quick retrieval/viewing
- Compact size (e.g. 112Gbp Illumina = 116GB disk space)
- Reads can be grouped into logical groups e.g. lanes, libraries, samples
- Widely supported by variant calling packages and viewers

# VCF: Variant Call Format

File format for storing variation data

- Tab-delimited text, parsable by standard UNIX commands
- Flexible and user-extensible
- Compressed with BGZF (bgzip), indexed with TBI or CSI (tabix)



# VCF / BCF

VCFs can be very big

- compressed VCF with 3781 samples, human data:
  - 54 GB for chromosome 1
  - 680 GB whole genome

VCFs can be slow to parse

- text conversion is slow
- main bottleneck: FORMAT fields

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1  3  .  A  G  .  PASS  AC=67;AN=5400;DP=2809  GT:PL:DP:GQ  1/1:0,9,73:26:22    0/0:0,9,73:13:31    0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15
1  4  .  A  T  .  PASS  AC=15;AN=6800;DP=6056  GT:PL:DP:GQ  0/0:0,9,73:13:31    1/0:255,0,75:32:15  0/0:0,2,80:14:90  1/1:0,9,73:26:22    0/0:0,9,73:13:31
1  5  .  C  T  .  PASS  AC=20;AN=6701;DP=5234  GT:PL:DP:GQ  1/0:255,0,75:32:15  0/0:0,2,170:14:90   1/1:0,9,73:13:31  0/0:0,6,50:13:80    0/0:0,2,80:14:90
1  6  .  A  G  .  PASS  AC=67;AN=5400;DP=2809  GT:PL:DP:GQ  1/1:0,9,73:26:22    0/0:0,9,73:13:31    0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15
1  7  .  A  T  .  PASS  AC=15;AN=6800;DP=6056  GT:PL:DP:GQ  0/0:0,9,73:13:31    1/0:255,0,75:32:15  0/0:0,2,80:14:90  1/1:0,9,73:26:22    0/0:0,9,73:13:31
```

BCF

- binary representation of VCF
- fields rearranged for fast access

```
#CHROM POS ID REF ALT QUAL FILTER INFO          FORMAT      SAMPLE1           SAMPLE2           SAMPLE3           SAMPLE4             SAMPLE5
1      6   .  A   G   .    PASS   AC=67;AN=540  GT:PL:DP:GQ  1/1:0,9,73:26:22  0/0:0,9,73:13:31  0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15

1      6   .  A   G   .    PASS   AC=67;AN=540  GT:1/1:0/0:0/0:1/0:1/0  PL:0,9,73:0,9,73:0,9,73:255,0,75:255,0,75 DP:26:13:48:32:32 GQ:22:31:99:15:15
```

https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format

# gVCF

Often it is not enough not know *variant* sites only

- was a site dropped because of a reference call or because of missing data?
- We need evidence for both variant and non-variant positions in the genome

gVCF

- blocks of reference-only sites can be represented in a single record using the INFO/END tag
- symbolic alleles <*> for incremental calling
  - raw, “callable” gVCF
  - calculate genotype likelihoods only once (an expensive step)
  - then call incrementally as more samples come in

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | SAMPLE |
|---|---|---|---|---|---|---|---|---|---|
| 19 | 9902 | . | G | <*> | . | . | END=9915;MinDP=0 | PL:DP | 0,0,0:0 |
| 19 | 9916 | . | C | <*> | . | . | END=9922;MinDP=5 | PL:DP | 0,15,137:5 |
| 19 | 9923 | . | G | <*> | . | . | END=9948;MinDP=10 | PL:DP | 0,30,214:10 |
| 19 | 9949 | . | G | A, <*> | . | . | DP=28 | PL:DP | 0,60,255,78,255,255:27 |
| 19 | 9950 | . | C | <*> | . | . | END=9958;MinDP=28 | PL:DP | 0,84,255:28 |
| 19 | 9959 | . | G | T, <*> | . | . | DP=34 | PL:DP | 0,82,255,99,255,255:34 |
| 19 | 9960 | . | C | <*> | . | . | END=9969;MinDP=34 | PL:DP | 0,102,255:34 |

**Symbolic "unobserved" allele**
Represents any other possible alternate allele

**A block of 10 sites** with at least 34 reference reads

**Genotype likelihoods**
for CC, C*, **

# Quality Control

Biases in sequencing

- Base calling accuracy
- Read cycle vs. base content
- GC vs. depth
- Indel ratio

Biases in mapping

Genotype checking

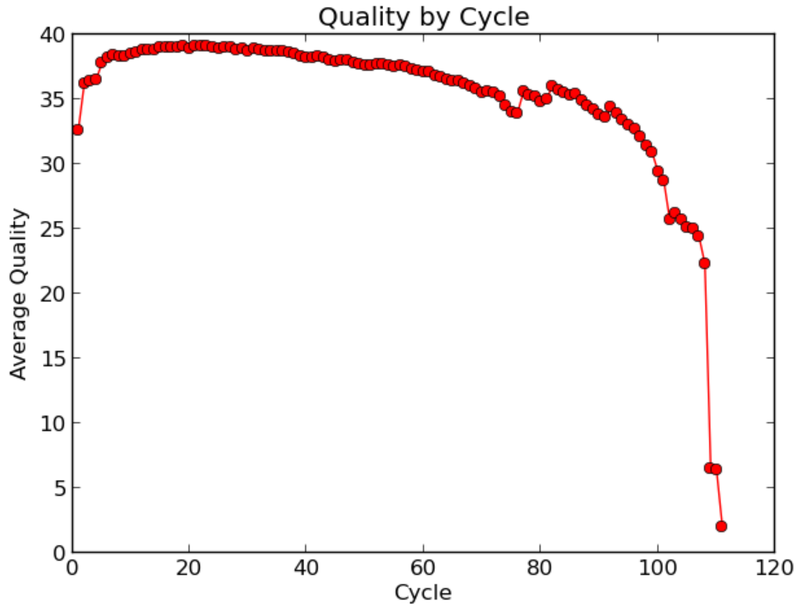
- Sample swaps
- Contaminations

# Base quality

Sequencing by synthesis: dephasing

- growing sequences in a cluster gradually desynchronize
- error rate increases with read length

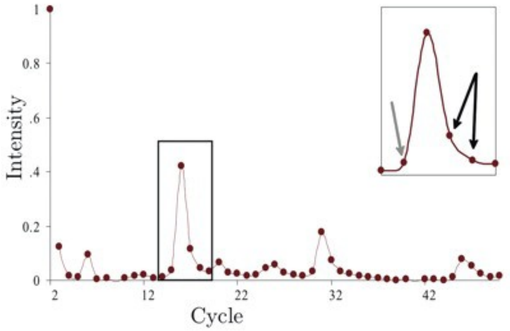
Calculate the average quality at each position across all reads



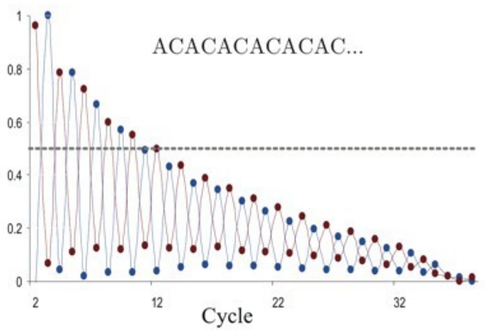

| Quality | Probability of error | Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1000 | 99.9% |
| 40 (Q40) | 1 in 10000 | 99.99% |

# Base calling errors

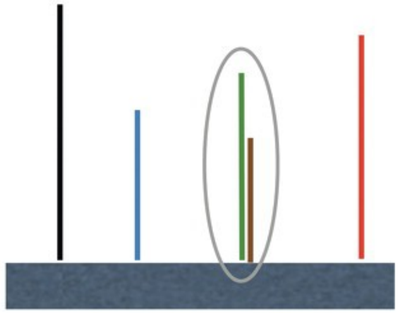
Phasing noise $\phi$



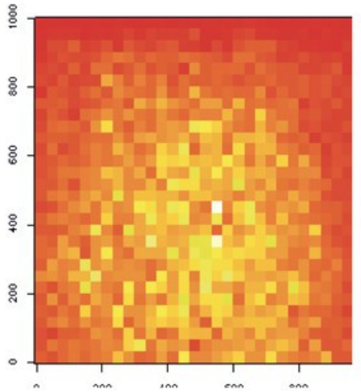

Signal Decay $\delta$



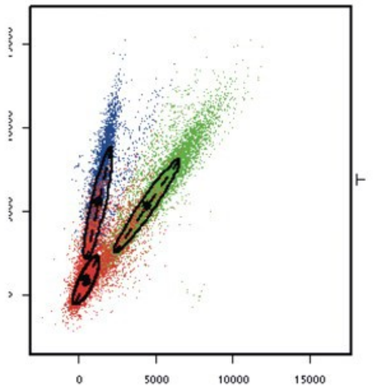

Mixed Cluster $\mu$



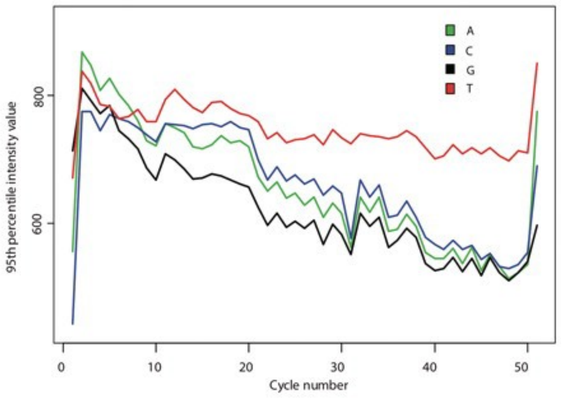
Boundary effects $\omega$



Cross-talk $\Sigma$

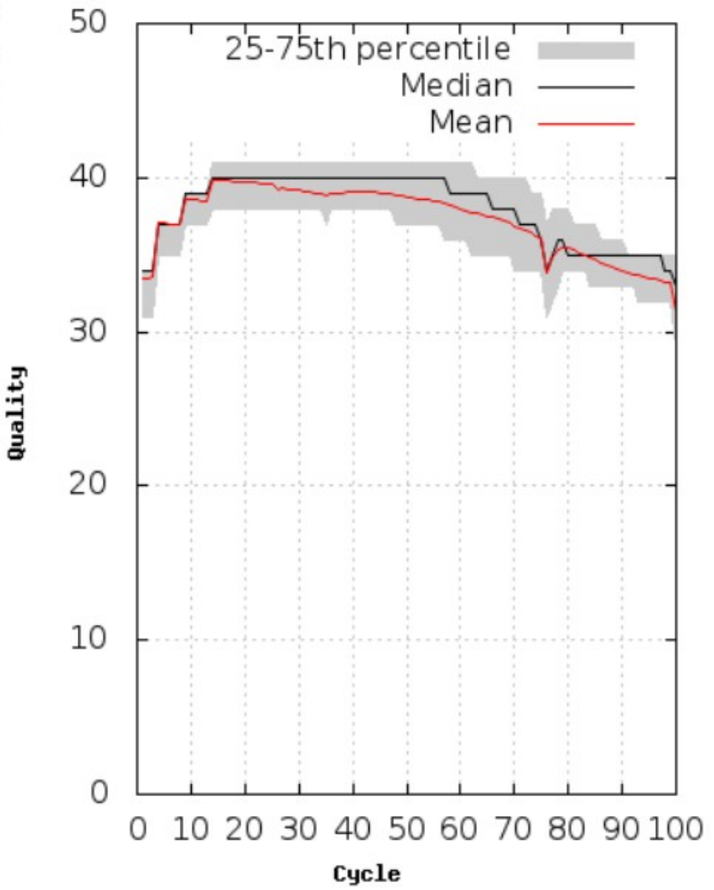


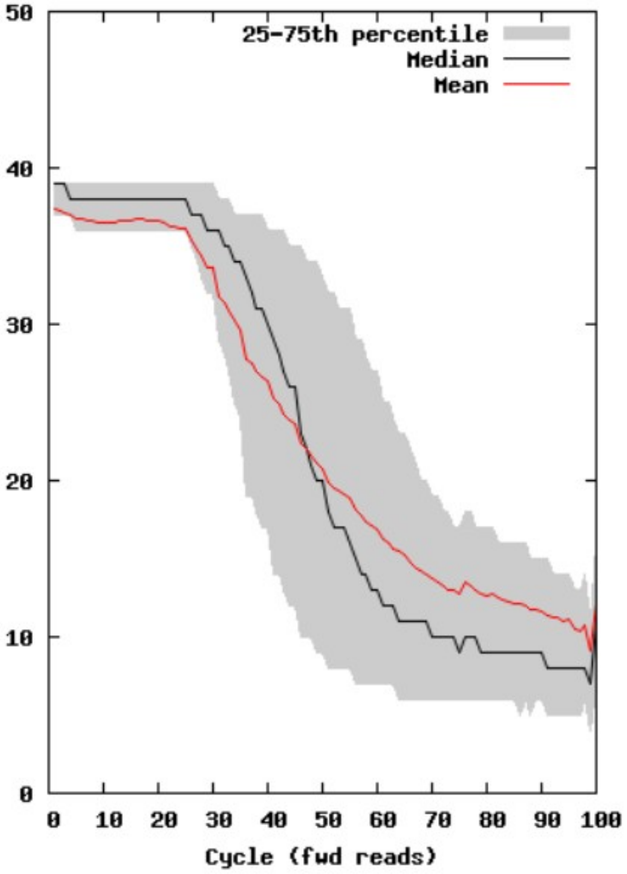

T fluophore accumulation $\tau$




Base-calling for next-generation sequencing platforms, doi: 10.1093/bib/bbq077

# Base quality

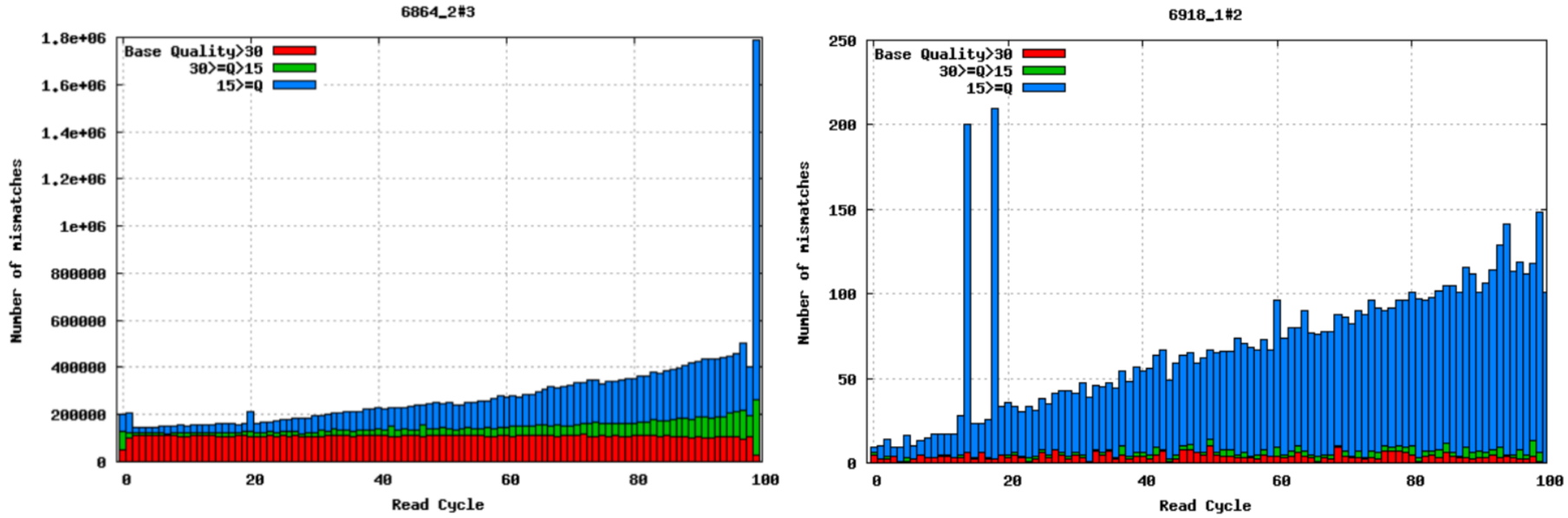


# Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

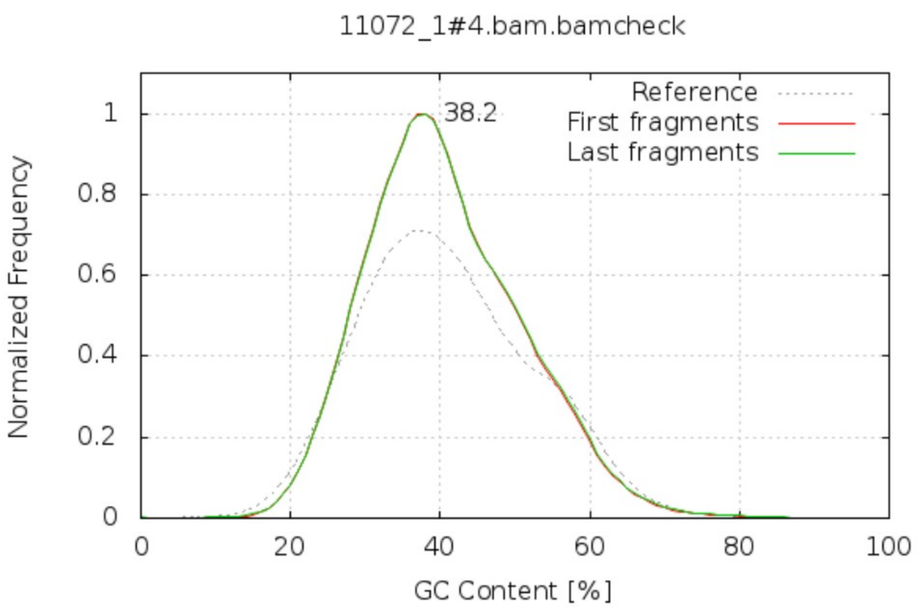
- detect cycle-specific errors
- Base qualities are informative!

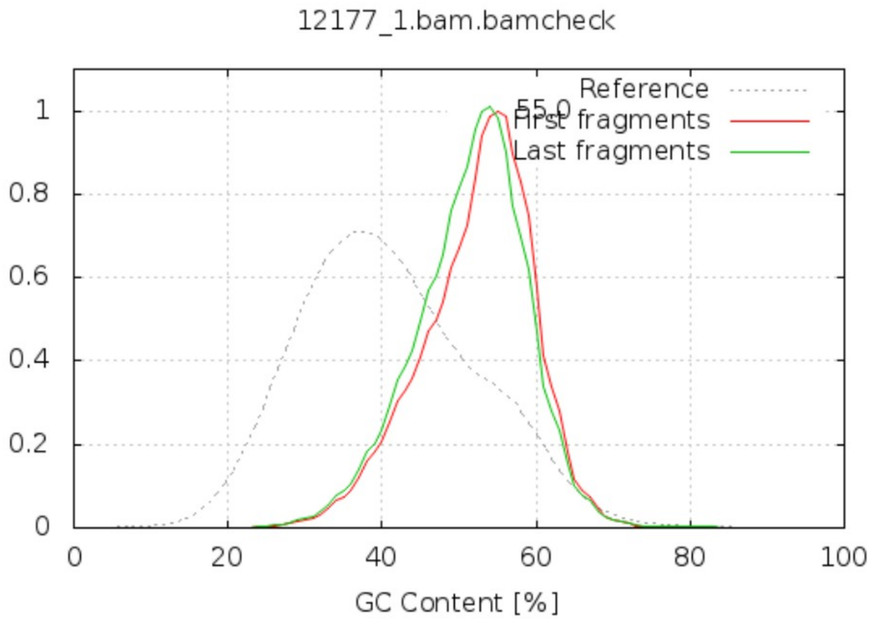


# GC bias

GC- and AT-rich regions are more difficult to amplify

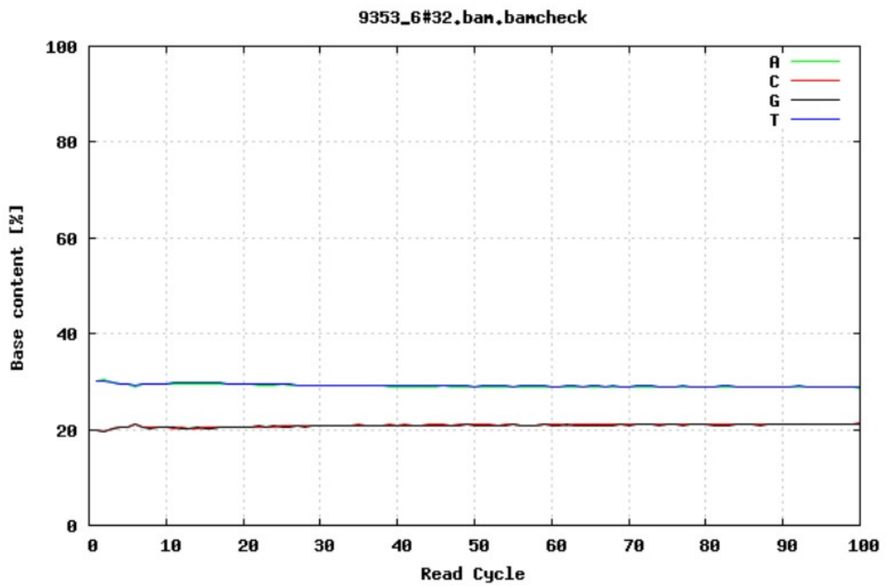
- compare the GC content against the expected distribution (reference sequence)

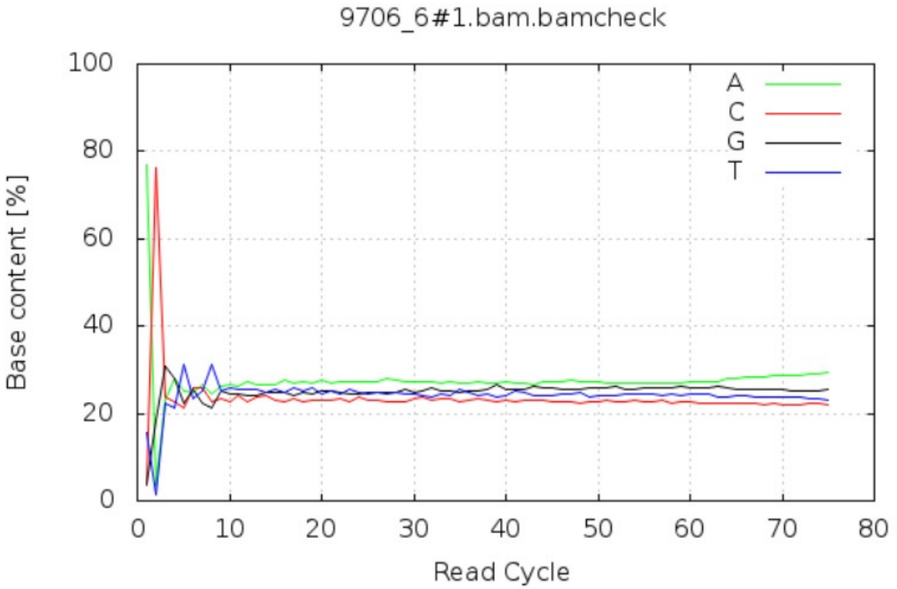


# CC content by cycle

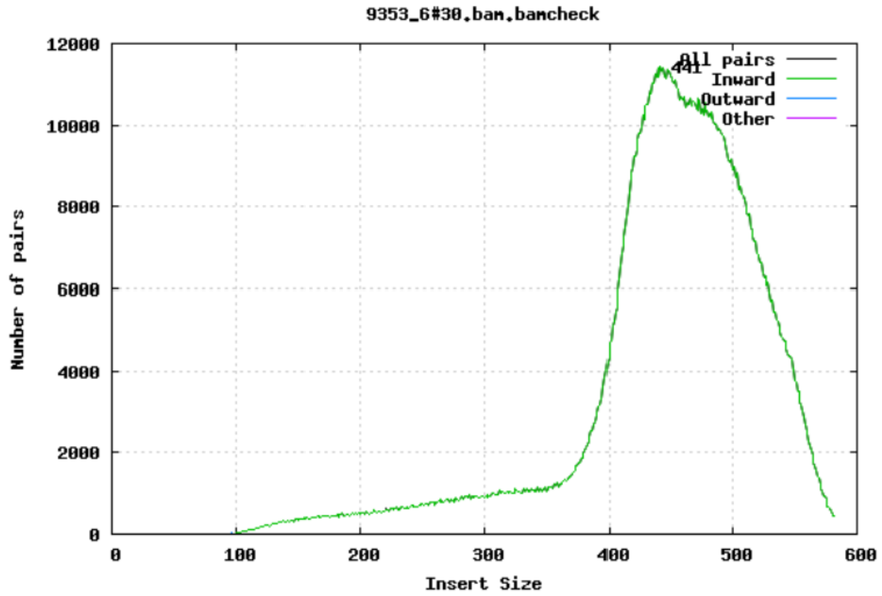
Was the adapter sequence trimmed?

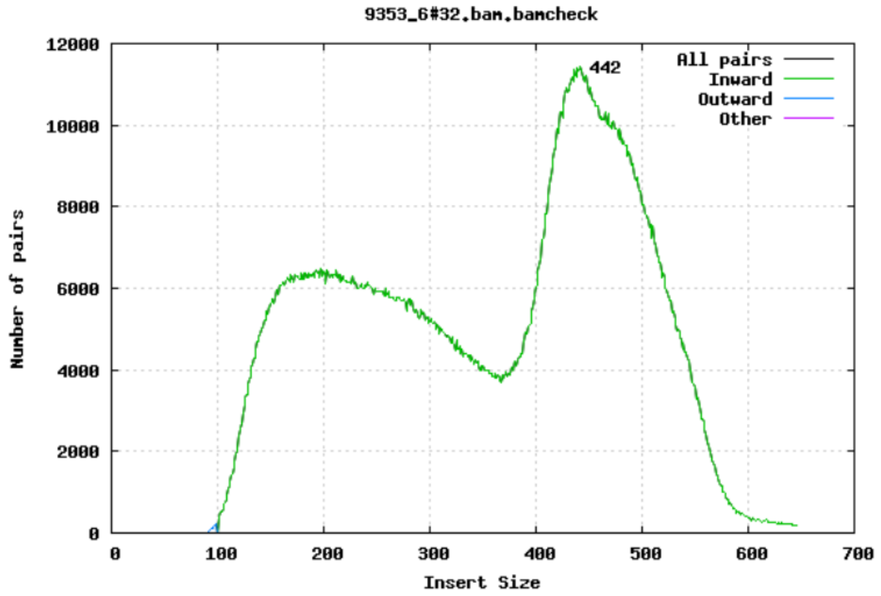


# Fragment size

Paired-end sequencing: the size of DNA fragments matters

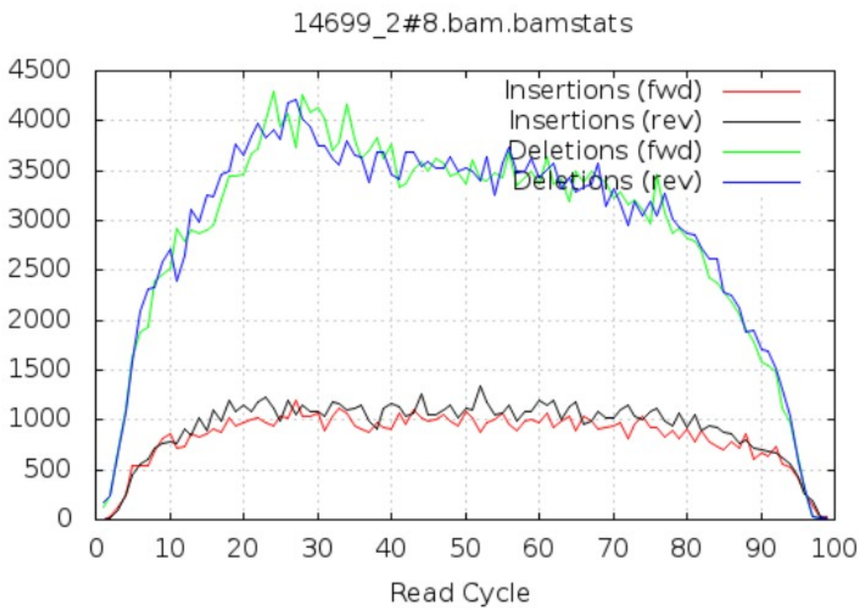


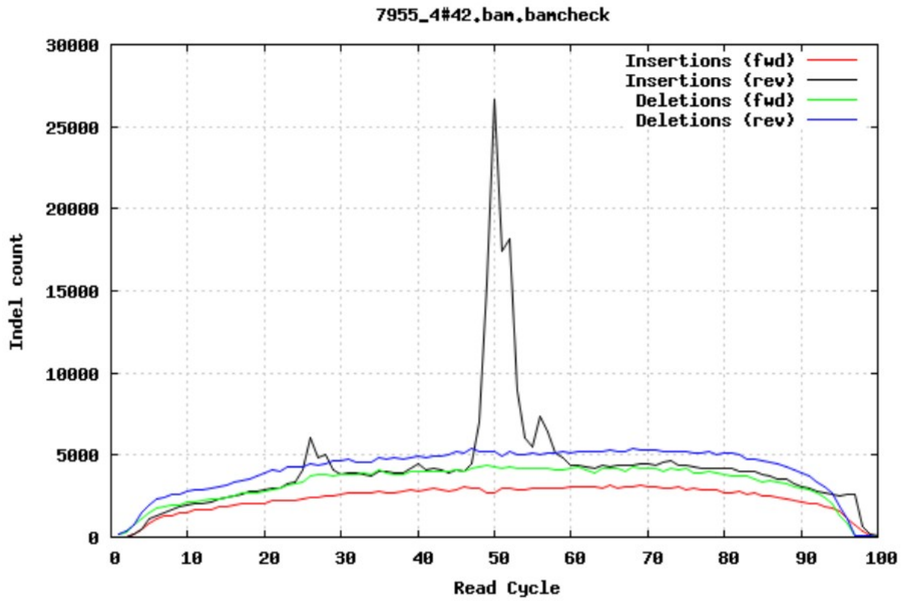



# Insertions / Deletions per cycle

False indels

- air bubbles in the flow cell can manifest as false indels



# Auto QC tests

A suggestion for human data:

| | |
|---|---|
| Minimum number of mapped bases | 90% |
| Maximum error rate | 0.02% |
| Maximum number of duplicate reads | 5% |
| Minimum number of mapped reads which are properly paired | 80% |
| Maximum number of duplicated bases due to overlapping read pairs | 4% |
| Maximum in/del ratio | 0.82 |
| Minimum in/del ratio | 0.68 |
| Maximimum indels per cycle, factor above median | 8 |
| Minimum number of reads within 25% of the main peak | 80% |