

**BIO 337**

Tuesday, Feb 18 2014

# Fred Sanger

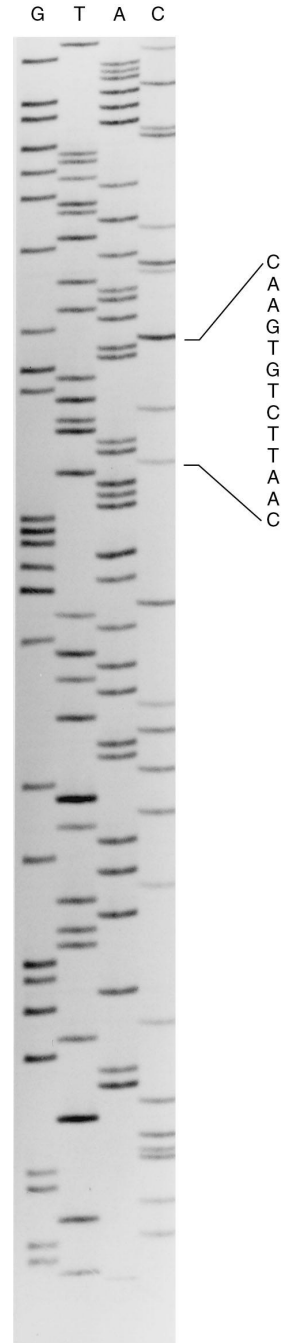
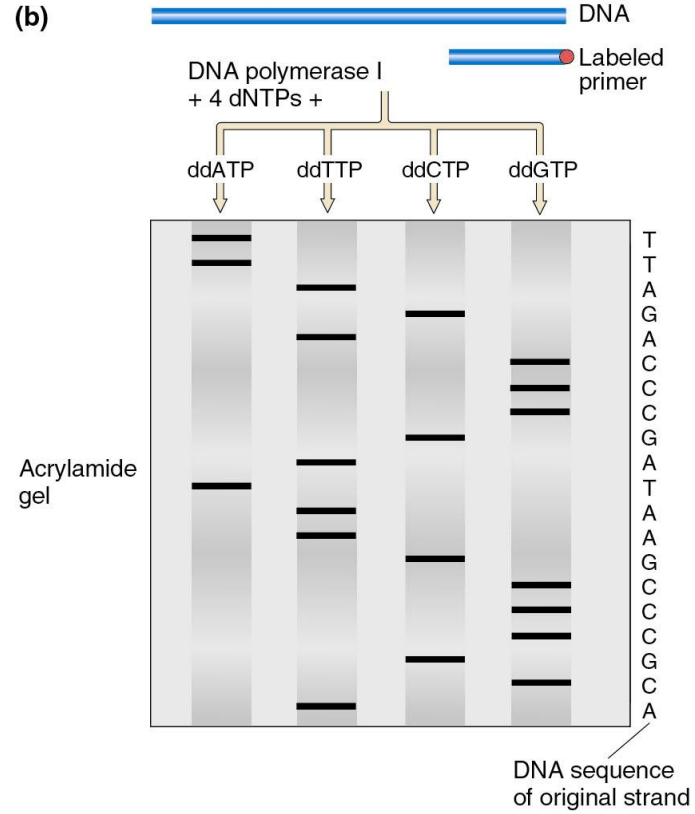
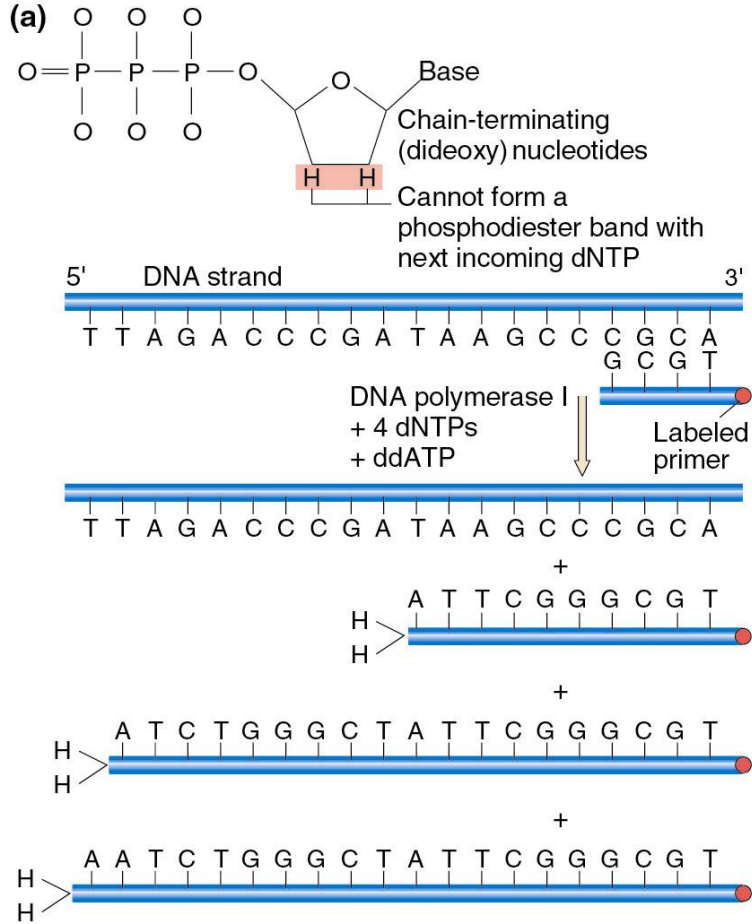
13 August 1918 – 19 November 2013



Nobel Prize in Chemistry, 1958 for protein sequencing (insulin)

Nobel Prize in Chemistry, 1980 for DNA sequencing

# Dideoxy sequencing



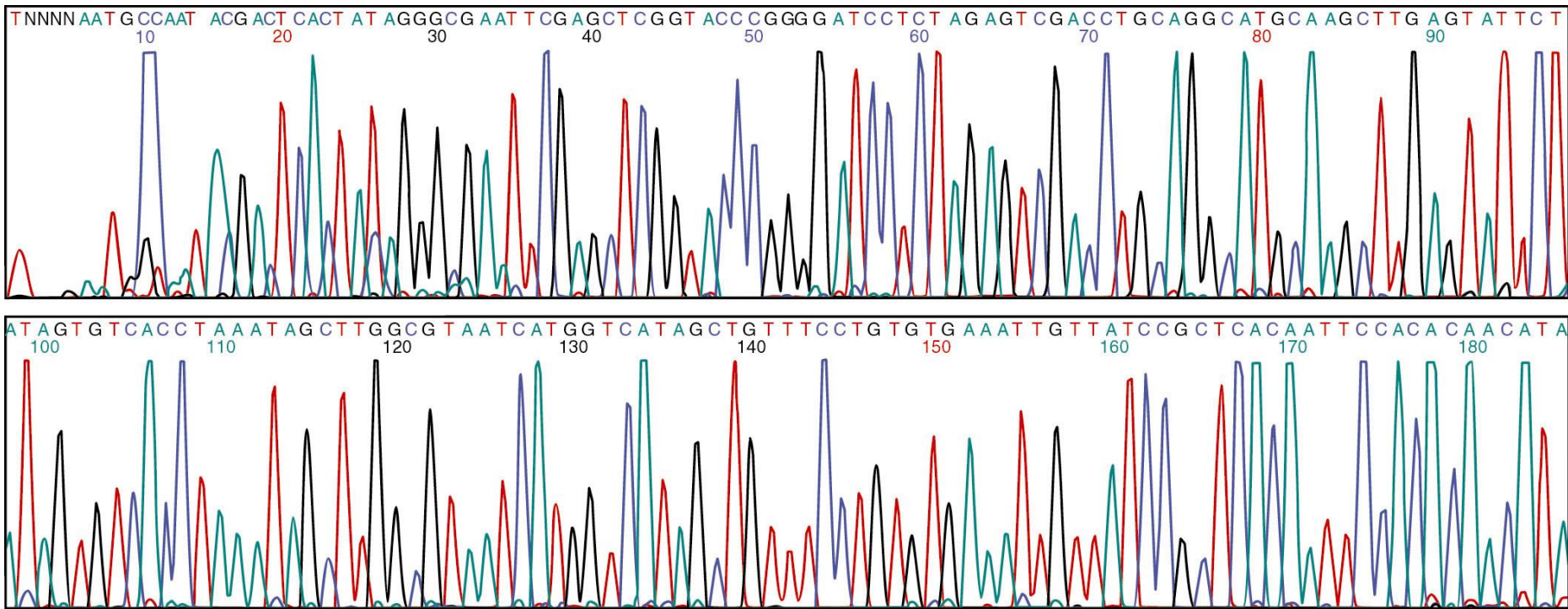
# Automated dye-terminator sequencing

4-fluorescently labelled dideoxy dye terminators

ddATP  
ddGTP  
ddCTP  
ddTTP

} pool and load in a single well or capillary

- scan with laser + detector specific for each dye
- automated base calling
- very long reads (~ 1000 bases)/run





February 2001



October 2010

# Functional genomics by sequencing



Used sequencing chemistry invented by Fred Sanger in 1977

In the last 3-5 years, radically new sequencing approaches have been invented and employed for functional genomics, termed

- Next-generation sequencing (NGS, 2<sup>nd</sup>, 3<sup>rd</sup> generation)
- Ultra high-throughput sequencing
- Single-molecule sequencing
- Deep sequencing

**Table 1 Next gen sequencing developers**

<b>Company</b>	<b>Technology overview</b>	<b>On market?</b>
Complete Genomics	Optical analysis of arrays of 'DNA nanoballs'	Yes
Genapsys Redwood City, California	Electronic detection of thermal/pH changes accompanying nucleotide addition	No
Genia Technologies	Pairing biological nanopores with semiconductor detection	No
GnuBio	Microfluidic system analyzes DNA nanodroplets with fluorescent primers	Alpha testing
Illumina	Sequencing by synthesis with fluorescently labeled reversible terminators	Yes
Lasergen Houston	Sequencing by synthesis with fluorescently labeled reversible terminators	No
Life Technologies (Ion Torrent)	Semiconductor sensor arrays detect protons released by nucleotide addition	Yes
NabSys Providence, Rhode Island	Single-molecule analysis revealing genomic location of sequencing probes	No
Noblegen Biosciences	Optical detection of 'expanded' DNA templates passing through synthetic pores	No
Oxford Nanopore Technologies	Detects changes in current as DNA strands pass through protein nanopores	No
Pacific Biosciences	Uses 'zero-mode waveguides' to optically detect real-time nucleotide addition	Yes
Qiagen (Intelligent Bio- Systems)	Sequencing by synthesis with fluorescently labeled reversible terminators	No
Roche (454)	Pyrosequencing of template-laden beads prepared by emulsion PCR	Yes
Stratos Genomics Seattle	Optical sequencing of fluorescently labeled, synthetically expanded templates	No

**Table 2 Next-generation DNA sequencing instruments**

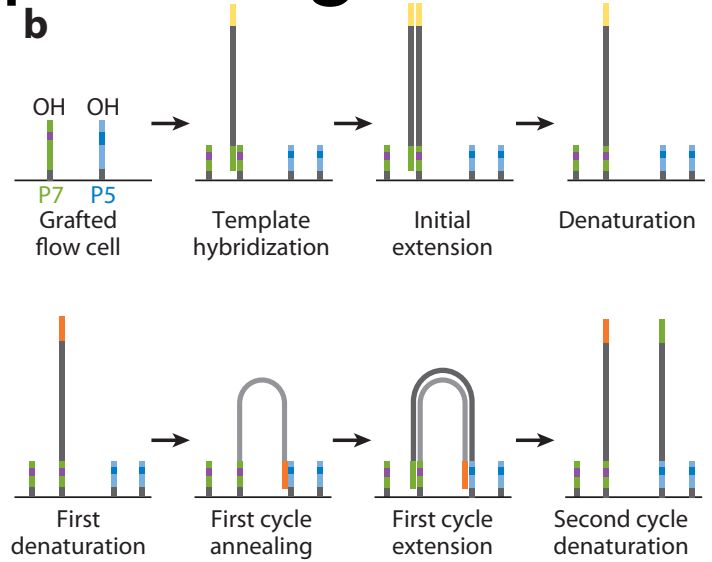
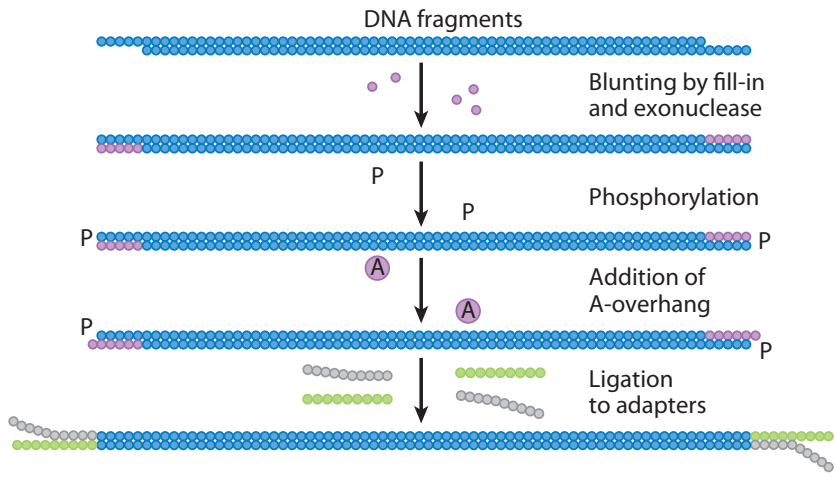
	Cost per base <sup>a</sup>	Read length (bp) <sup>b</sup>	Speed	Capital cost <sup>c</sup>
<b>Minimum cost per base</b>				
Complete Genomics	Low	Short	3 months	None (service)
HiSeq 2000 (Illumina)	Low	Mid	8 days	+++++++
SOLiD 5500xl (Life Technologies)	Low	Short	8 days	+++
<b>Maximum read length</b>				
454 GS FLX+ (Roche)	High	Long	1 day	+++++
RS (Pacific Biosciences)	High	Very long	<1 day	+++++++
<b>Maximum speed, minimum capital cost and minimum footprint</b>				
454 GS Junior (Roche)	High	Mid	<1 day	+
Ion Torrent PGM (Life Technologies)	Mid	Mid	<1 day	+
MiSeq (Illumina)	Mid	Long	1 day	+
<b>Combined prioritization of speed and throughput</b>				
Ion Torrent Proton (Life Technologies)	Low	Mid	<1 day	++
HiSeq 2500 (Illumina)	Low	Mid	2 days	+++++++

<sup>a</sup>'Low' is < \$0.10 per megabase, 'mid' is in-between and 'high' is > \$1 per megabase. <sup>b</sup>'Short' is < 200 bp, 'mid' is 200–400 bp, 'long' is > 400 bp and 'very long' is > 1,000 bp. <sup>c</sup>Each "+" corresponds to ~\$100,000. We list only commercialized instruments that can be purchased and for which performance data are publically available (as opposed to a comprehensive list of companies developing next-generation sequencing technologies). The categorizations refer to the aspect of sequencing performance to which the technology and/or its implementation in a specific instrument are primarily geared. These estimates were made at the time of publication, and the pace at which the field is moving makes it likely that they will be quickly outdated.

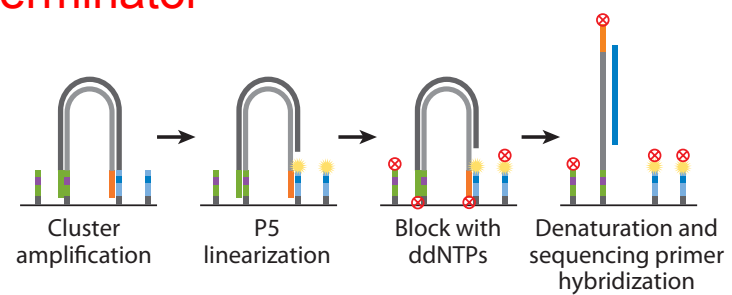
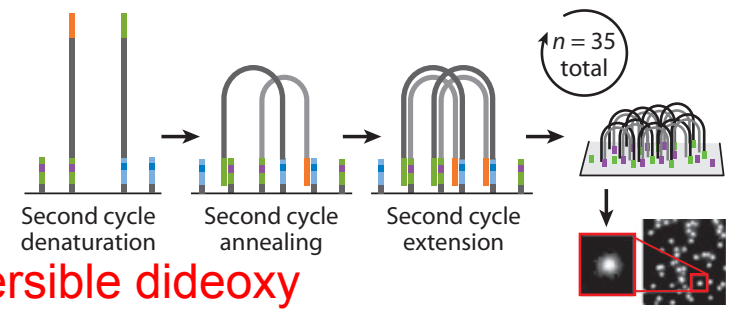
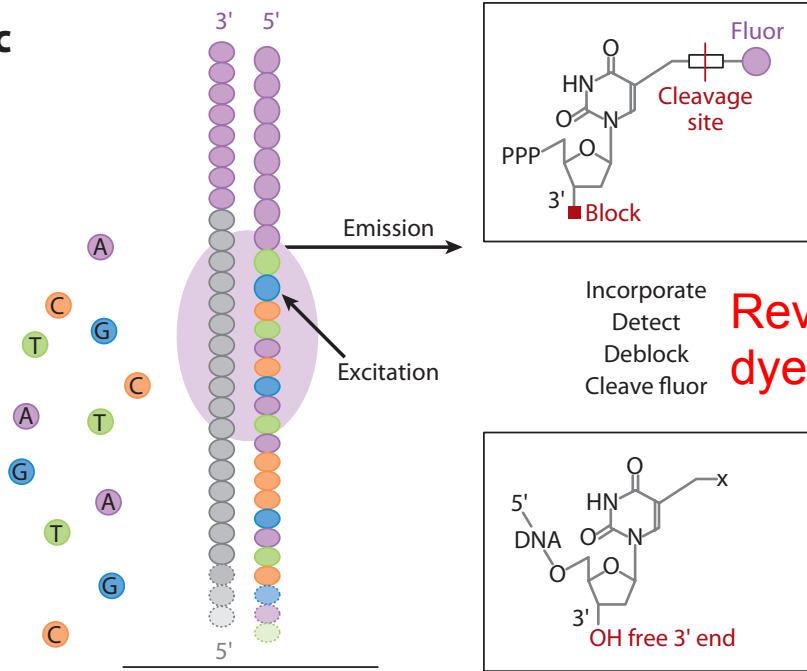


# Next Generation Sequencing: Illumina

**a** Illumina's library-preparation work flow

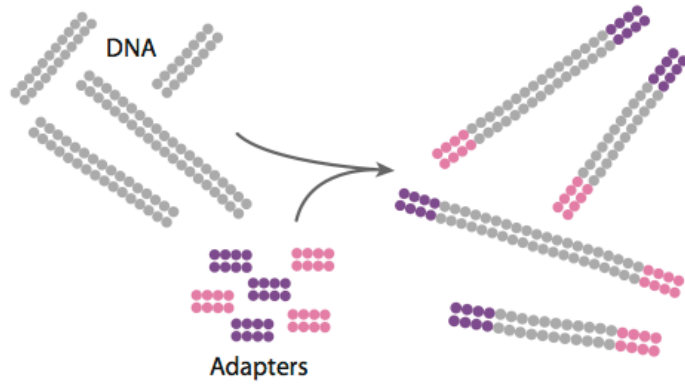


**c**



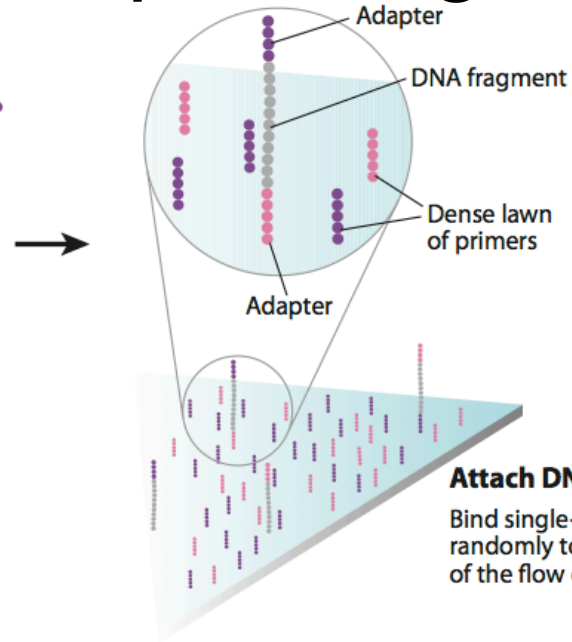
# Next Generation Sequencing: Illumina

a



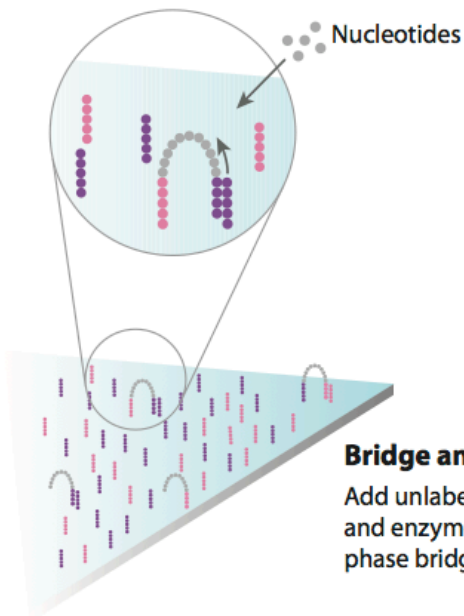
## Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.



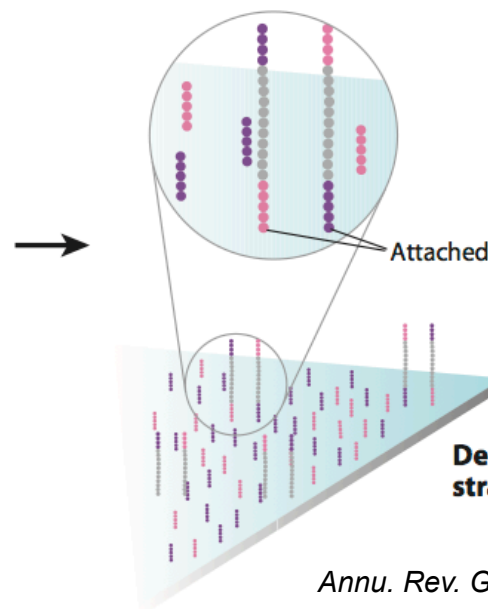
## Attach DNA to surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.



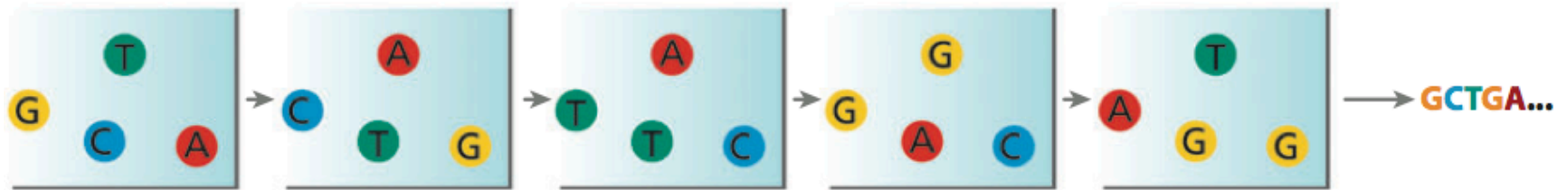
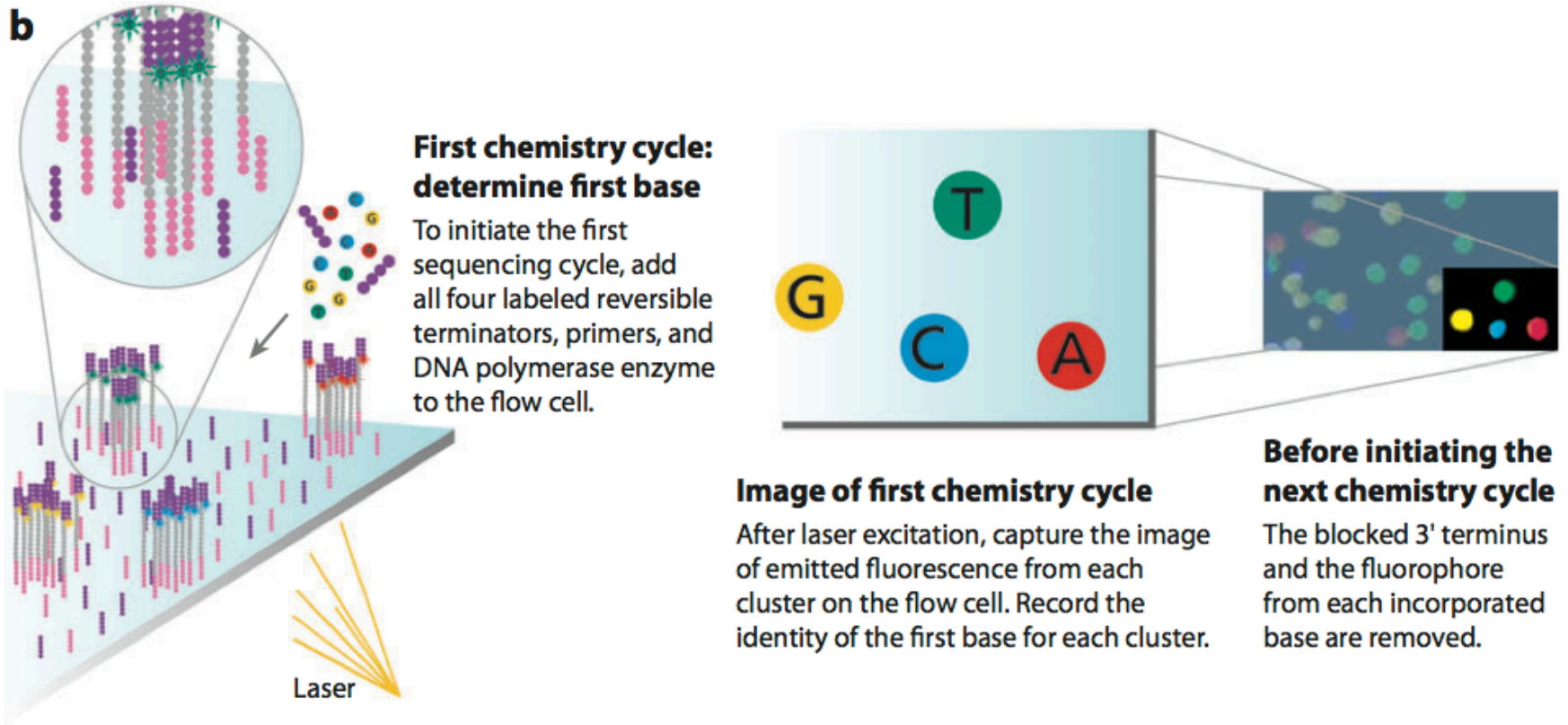
## Bridge amplification

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.



## Denature the double stranded molecules

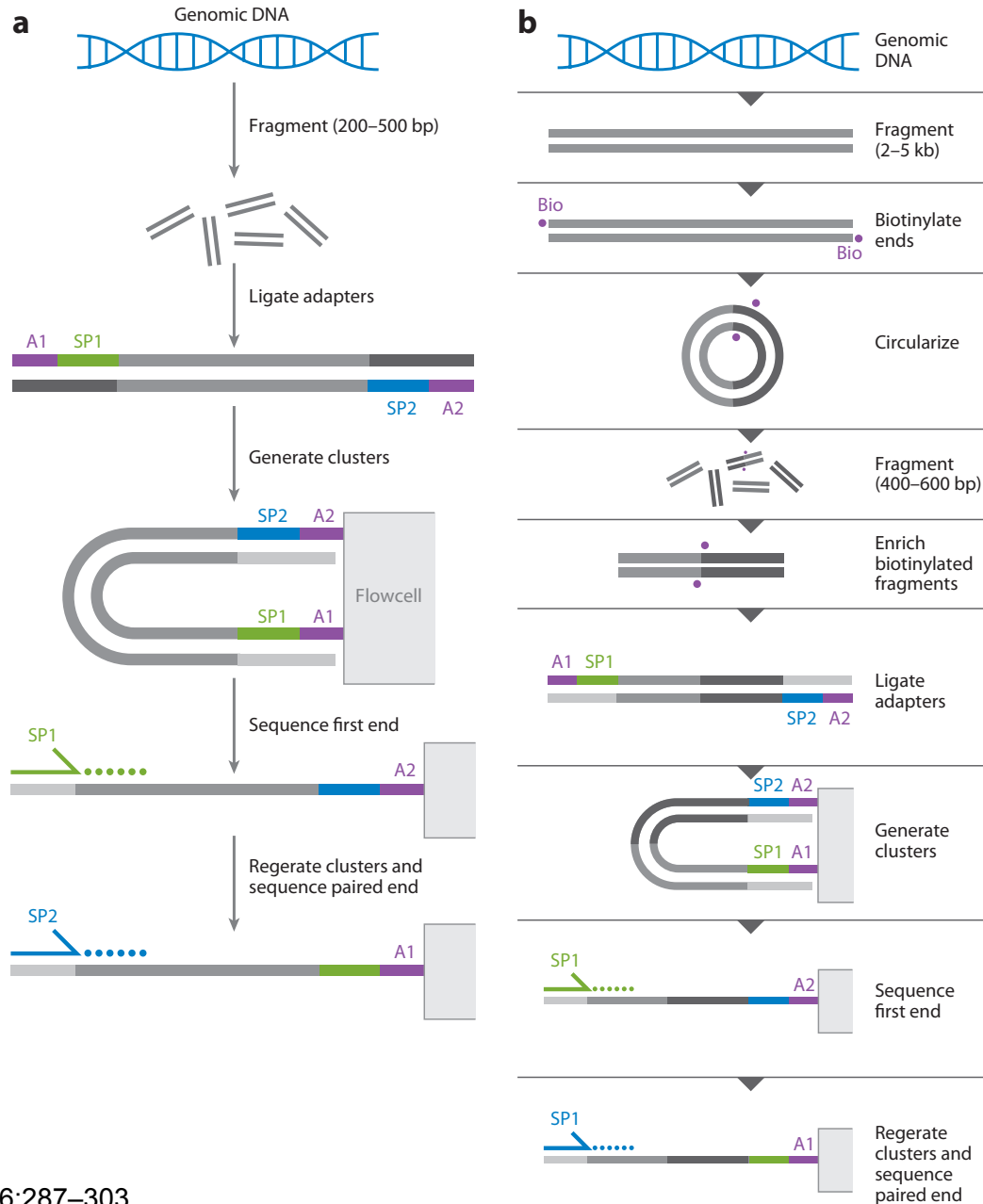
# Next Generation Sequencing: Illumina



## Sequence read over multiple chemistry cycles

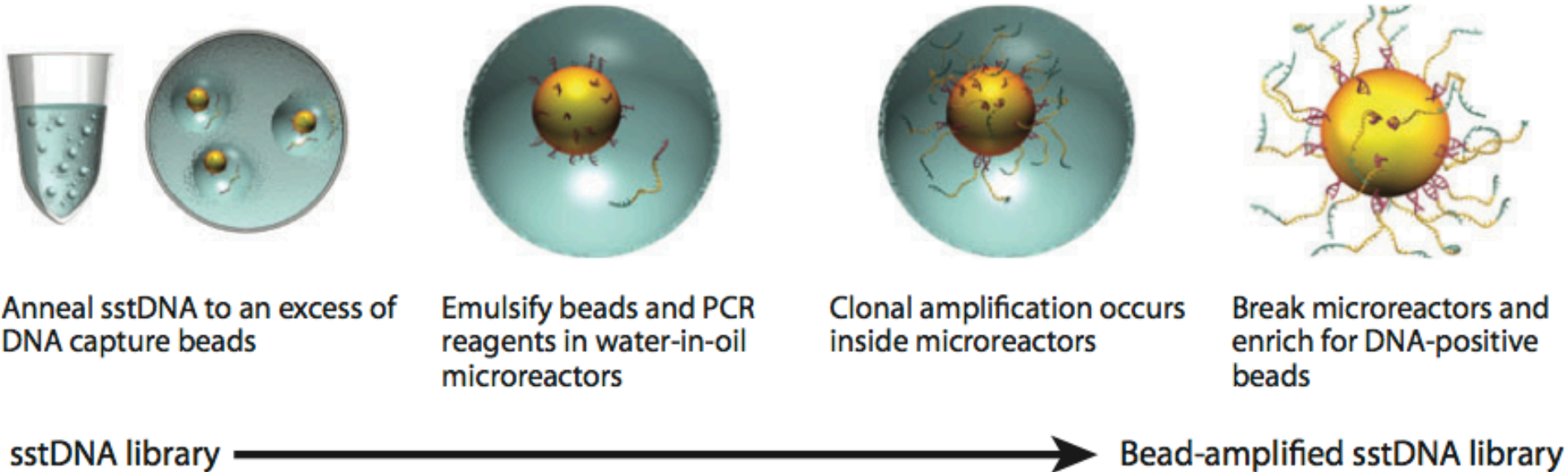
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

# Paired-end and mate paired libraries



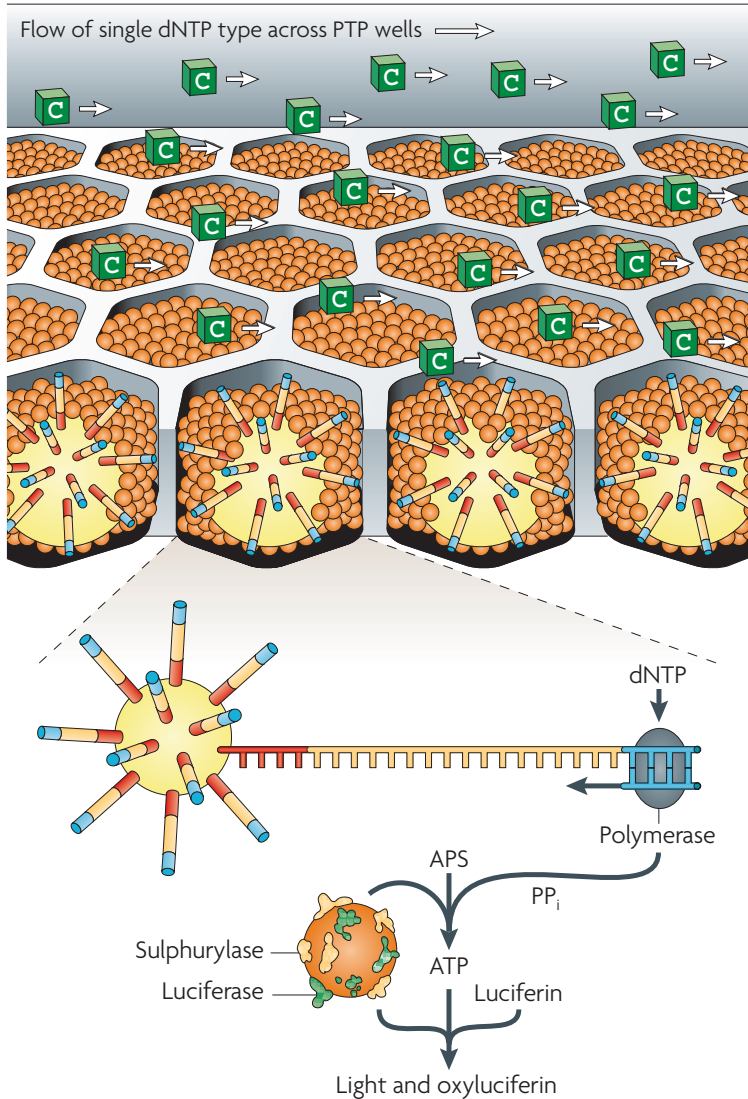
# Emulsion PCR for clonal amplification

Used in next-gen sequencing by Roche/454 and Life/Applied Biosystems platforms



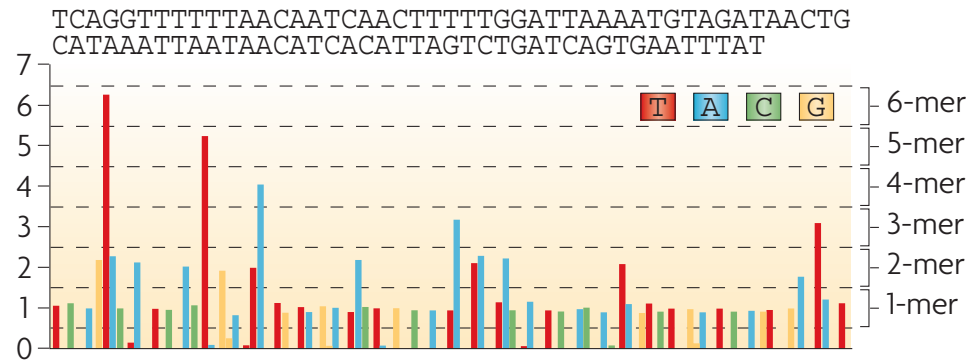
# Pyrosequencing

1-2 million template beads loaded into PTP wells

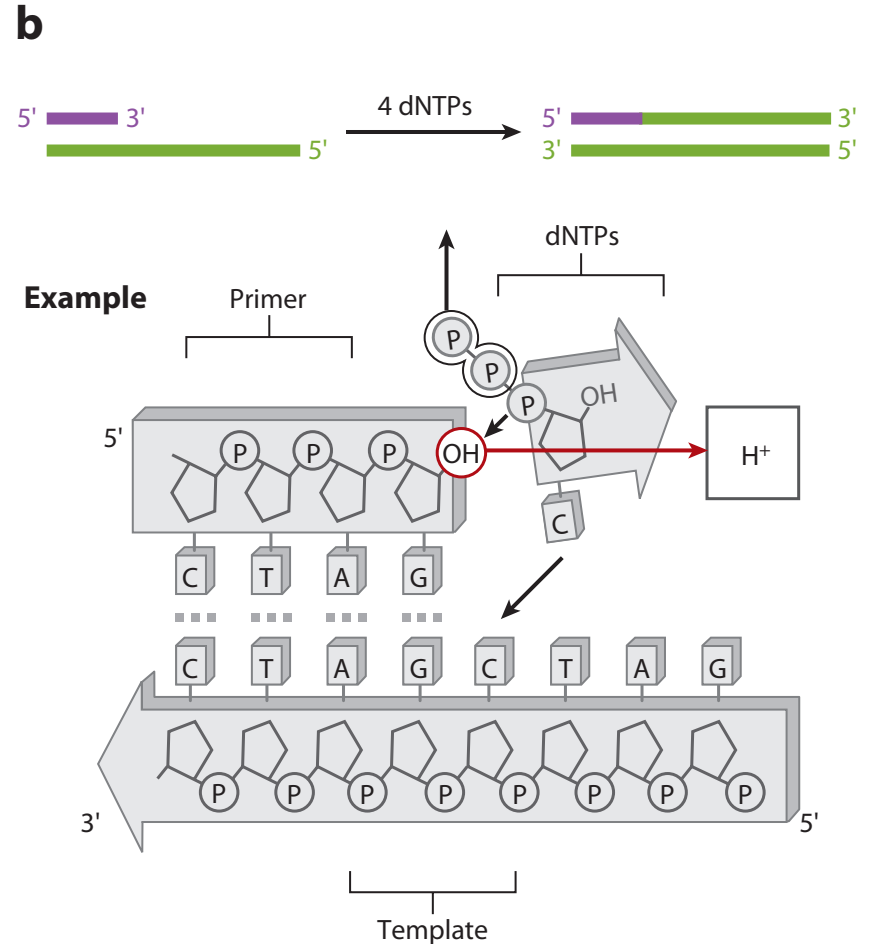
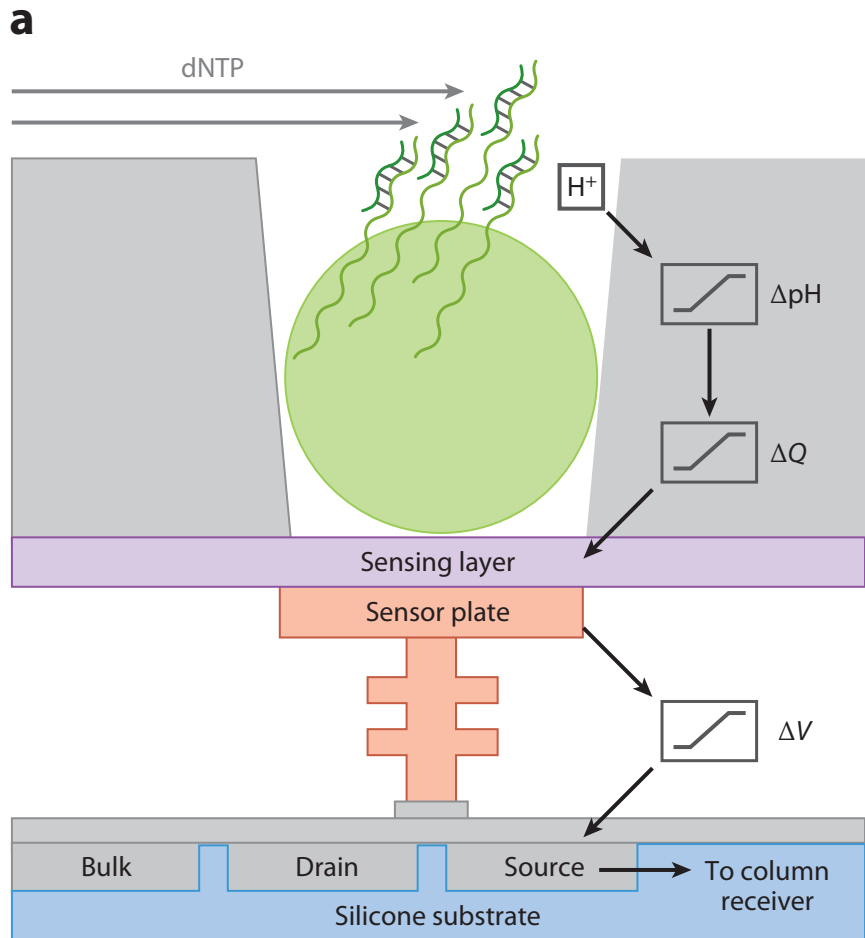


**d**

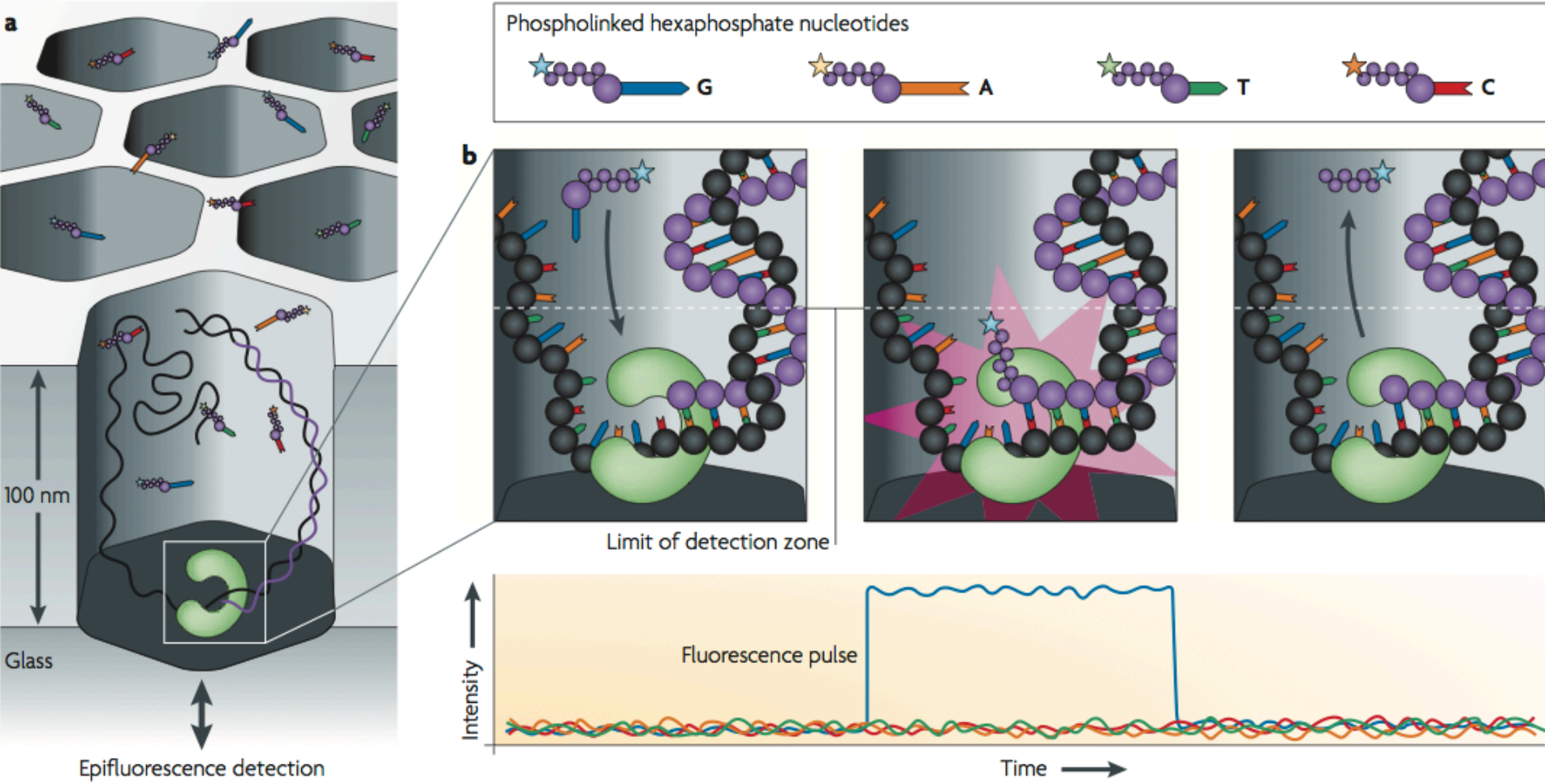
Flowgram



# Ion Torrent pH based sequencing

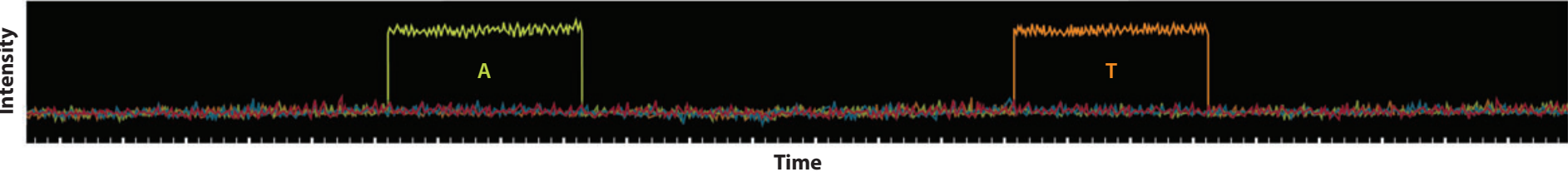


# Single-molecule sequencing: Pacific Biosystems

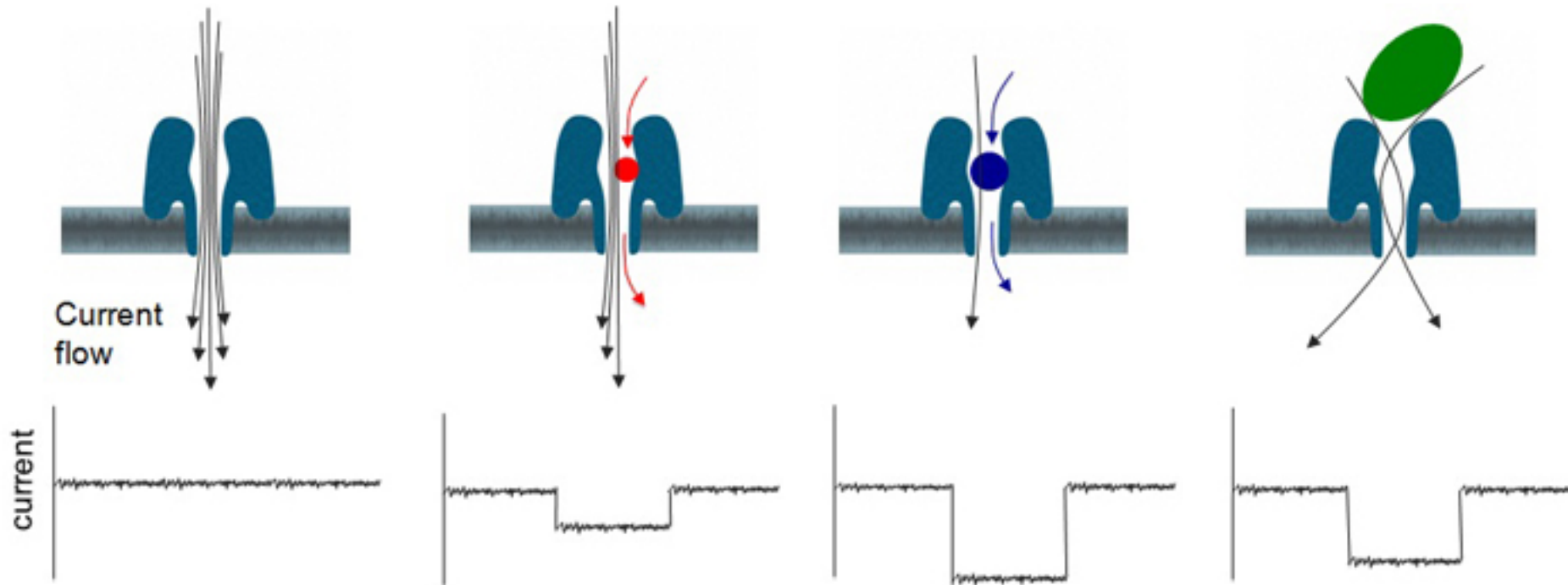




# Single-molecule sequencing: Pacific Biosystems



# Nanopore sequencing



Nanopores can be

- biological: formed by a pore-forming protein in a membrane such as a lipid bilayer
- solid-state: formed in synthetic materials such as silicon nitride or graphene
- hybrid: formed by a pore-forming protein set in synthetic material

# Features of next-gen sequencing

- Short reads (35 bp – 400 bp)
- Millions of reads per run ( $10^7$  –  $5 \times 10^8$ )
- Higher error rate per basepair raw
- No cloning in *E. coli*
- Huge amounts of data per experiment (20 GB primary/2 TB raw)
- Large data storage and computational analysis requirements

**NGS data**

```
graph TD; A[NGS data] --> B[Counting]; A --> C[Variants]; A --> D[Assembly]; B --- B1[RNA-seq]; B --- B2[ChIP-seq]; B --- B3[etc.]; C --- C1[Cancer genomes]; C --- C2[Genetic variation]; C --- C3[etc.]; D --- D1[New genomes, transcriptomes]; D --- D2[etc.];
```

**Counting**

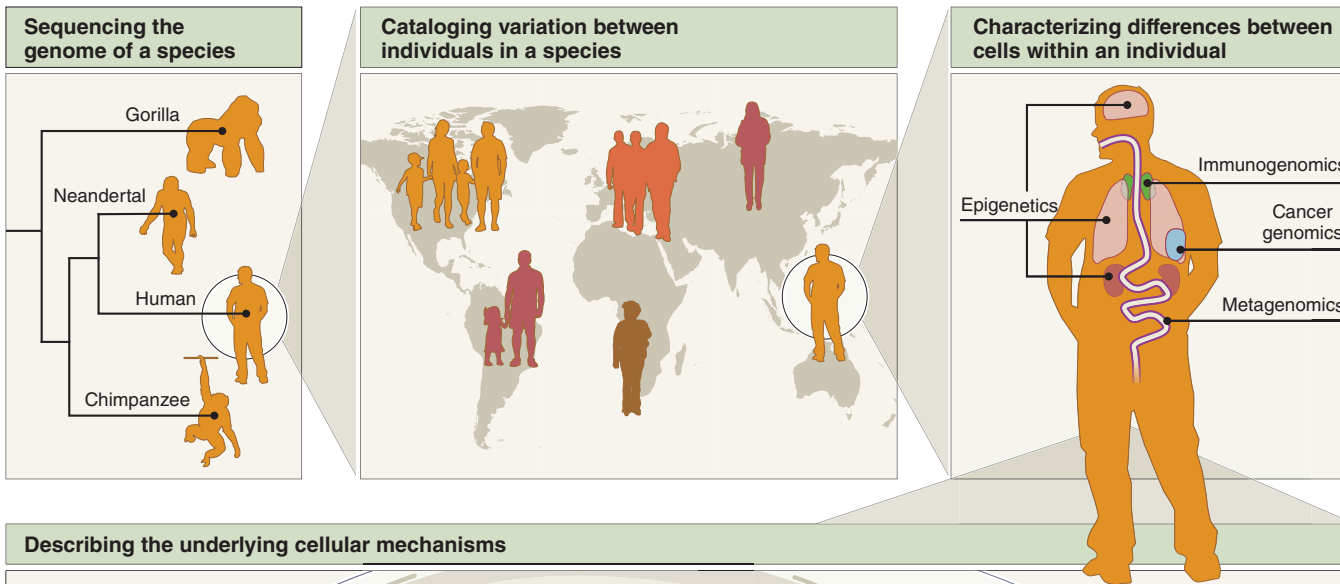
RNA-seq  
ChIP-seq  
etc.

**Variants**

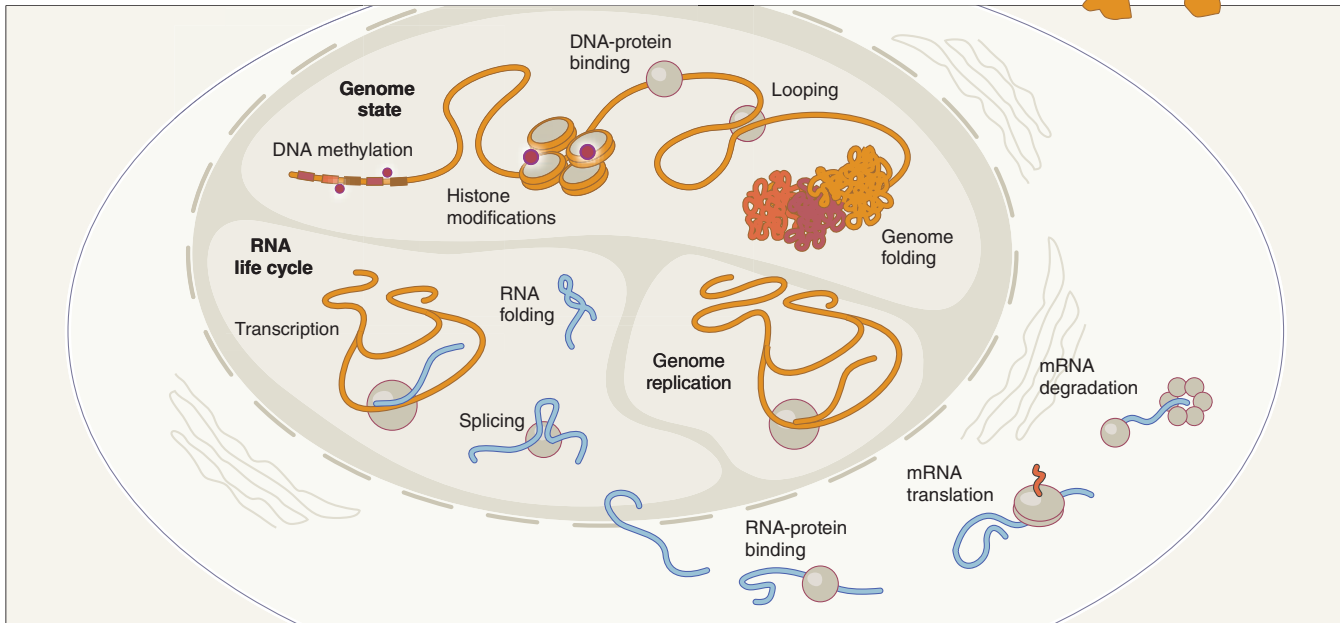
Cancer genomes  
Genetic variation  
etc.

**Assembly**

New genomes,  
transcriptomes  
etc.



**Describing the underlying cellular mechanisms**



# Some applications of next-gen sequencing

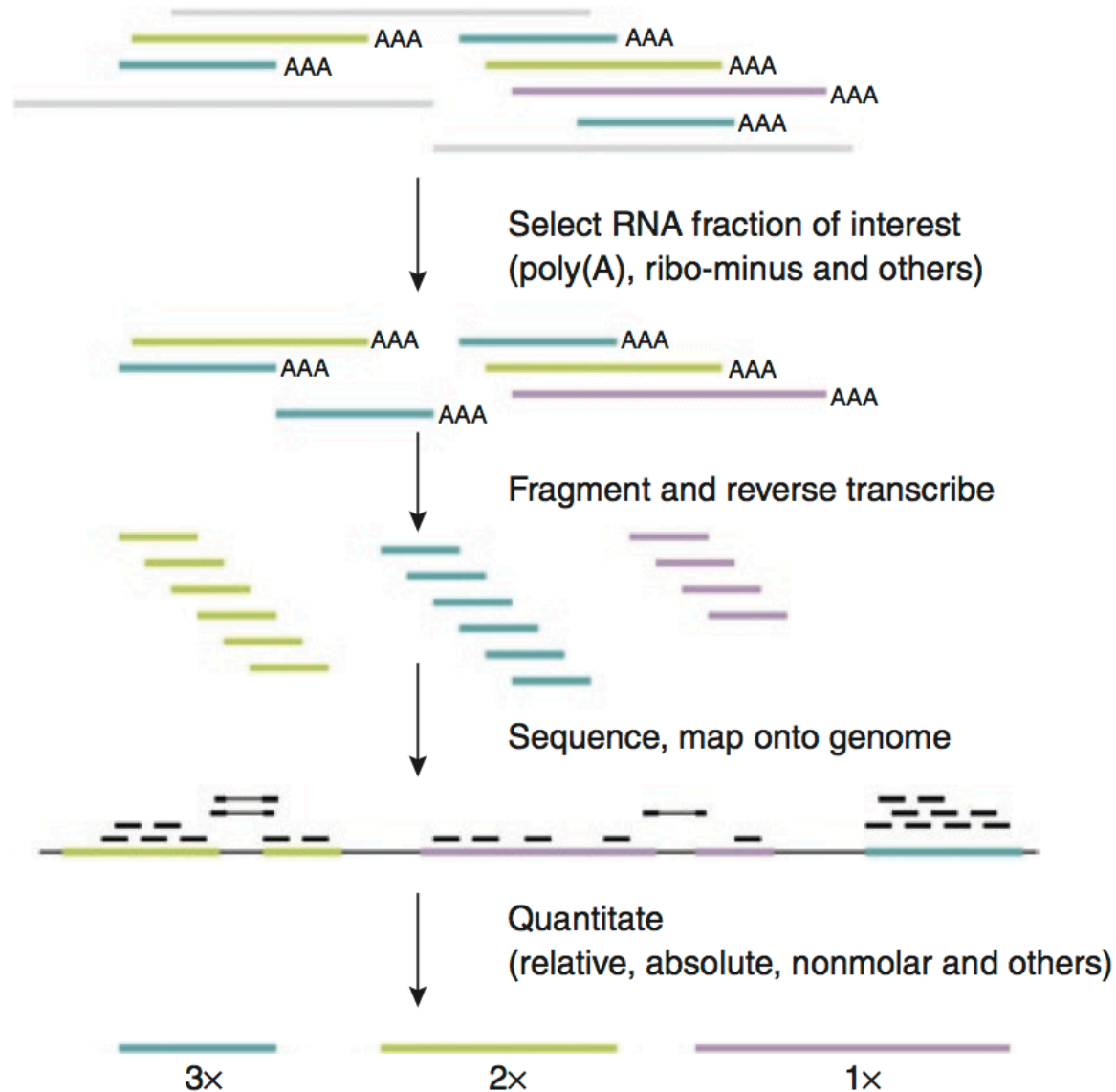
- Genome sequencing and variant discovery
- de novo assembly of bacterial and other small genomes
- DNA-protein interactions ChIP-seq
- Chromatin and epigenetics Methyl-seq
- RNA expression levels (profiling) RNA-seq
- ncRNA/small RNA discovery and profiling
- Metagenomics
- Sequencing extinct species (museomics)

**Table 1 Applications of next-generation DNA sequencing**

Method	Sequencing to determine:	Example reference	'Subway' route as defined in Figure 3
DNA-Seq	A genome sequence	57	Comparison, 'anatomic' (isolation by anatomic site), flow cytometry, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing
Targeted DNA-Seq	A subset of a genome (for example, an exome)	20	Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing
Methyl-Seq	Sites of DNA methylation, genome-wide	34	Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing
Targeted methyl-Seq	DNA methylation in a subset of the genome	129	Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing
DNase-Seq, Sono-Seq and FAIRE-Seq	Active regulatory chromatin (that is, nucleosome-depleted)	113	Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
MAINE-Seq	Histone-bound DNA (nucleosome positioning)	130	Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)	131	Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions	46	Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing
RNA-Seq	RNA (that is, the transcriptome)	39	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
FRT-Seq	Amplification-free, strand-specific transcriptome sequencing	119	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing
NET-Seq	Nascent transcription	41	Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing
Hi-C	Three-dimensional genome structure	71	Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing
Chia-PET	Long-range interactions mediated by a protein	73	Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing
Ribo-Seq	Ribosome-protected mRNA fragments (that is, active translation)	48	Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing
TRAP	Genetically targeted purification of polysomal mRNAs	132	Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
PARS	Parallel analysis of RNA structure	42	Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing
Synthetic saturation mutagenesis	Functional consequences of genetic variation	93	Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing
Immuno-Seq	The B-cell and T-cell repertoires	86	Perturbation, 'anatomic', DNA extraction, PCR and sequencing
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants	95	Variation, genetic manipulation, phage display, <i>in vitro</i> competitive binding, DNA extraction, PCR and sequencing
PhIT-Seq	Relative fitness of cells containing disruptive insertions in diverse genes	92	Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing

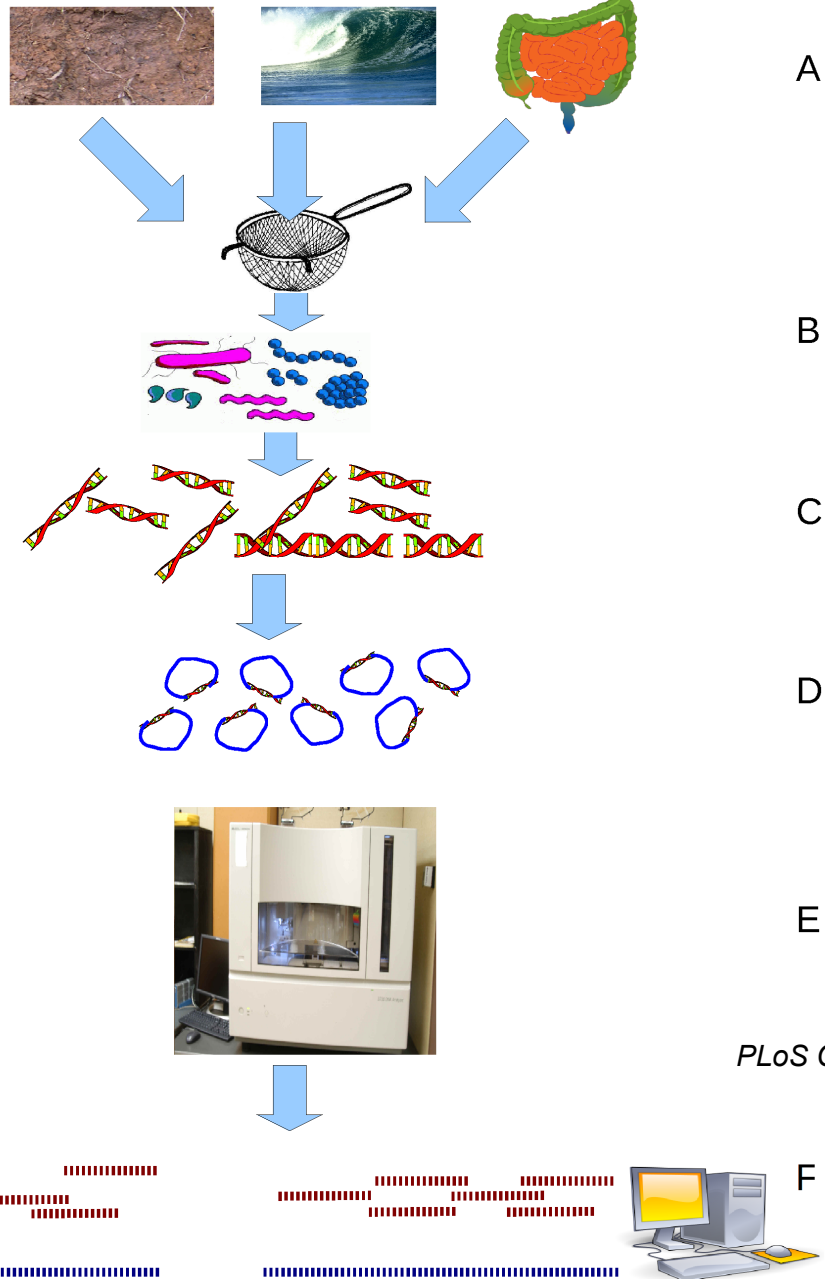
FAIRE-seq, formaldehyde-assisted isolation of regulatory elements—sequencing. MAINE-Seq, MNase-assisted isolation of nucleosomes—sequencing; RIP-Seq, RNA-binding protein immunoprecipitation—sequencing; CLIP-Seq, cross-linking immunoprecipitation—sequencing; HITS-CLIP, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation; FRT-Seq, on-flowcell reverse transcription—sequencing. NET-Seq, native elongating transcript sequencing. TRAP, translating ribosome affinity purification. PhIT-Seq, phenotypic interrogation via tag sequencing.

# Gene expression profiling with RNA-seq



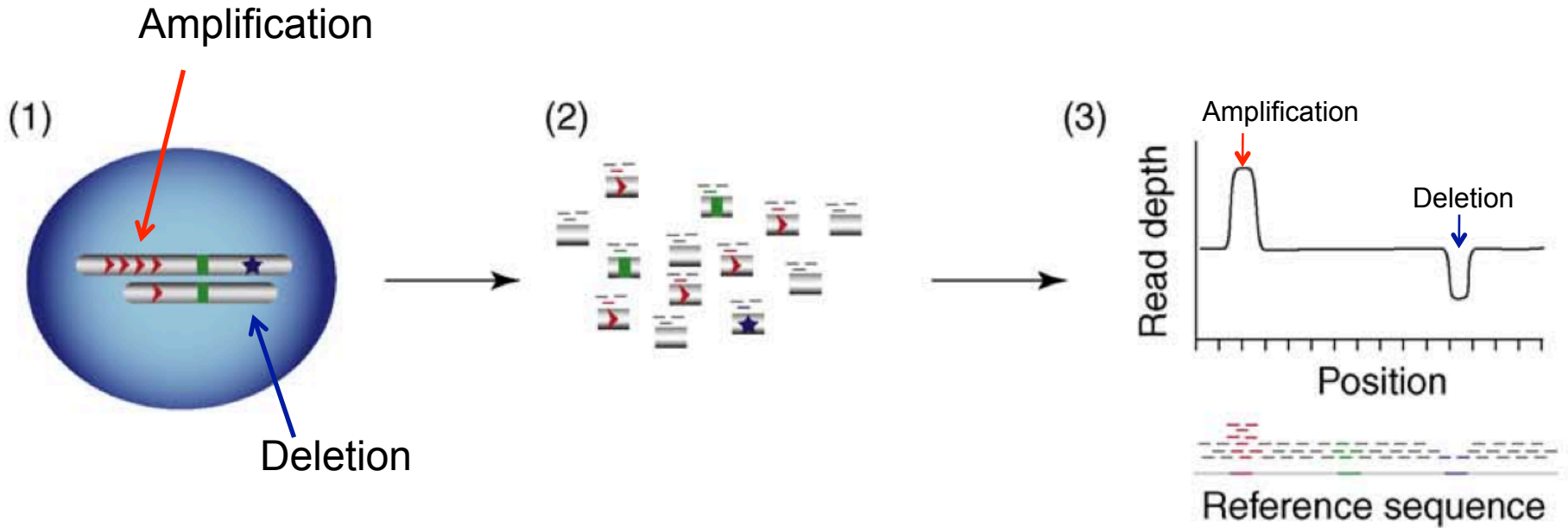


# Metagenomics



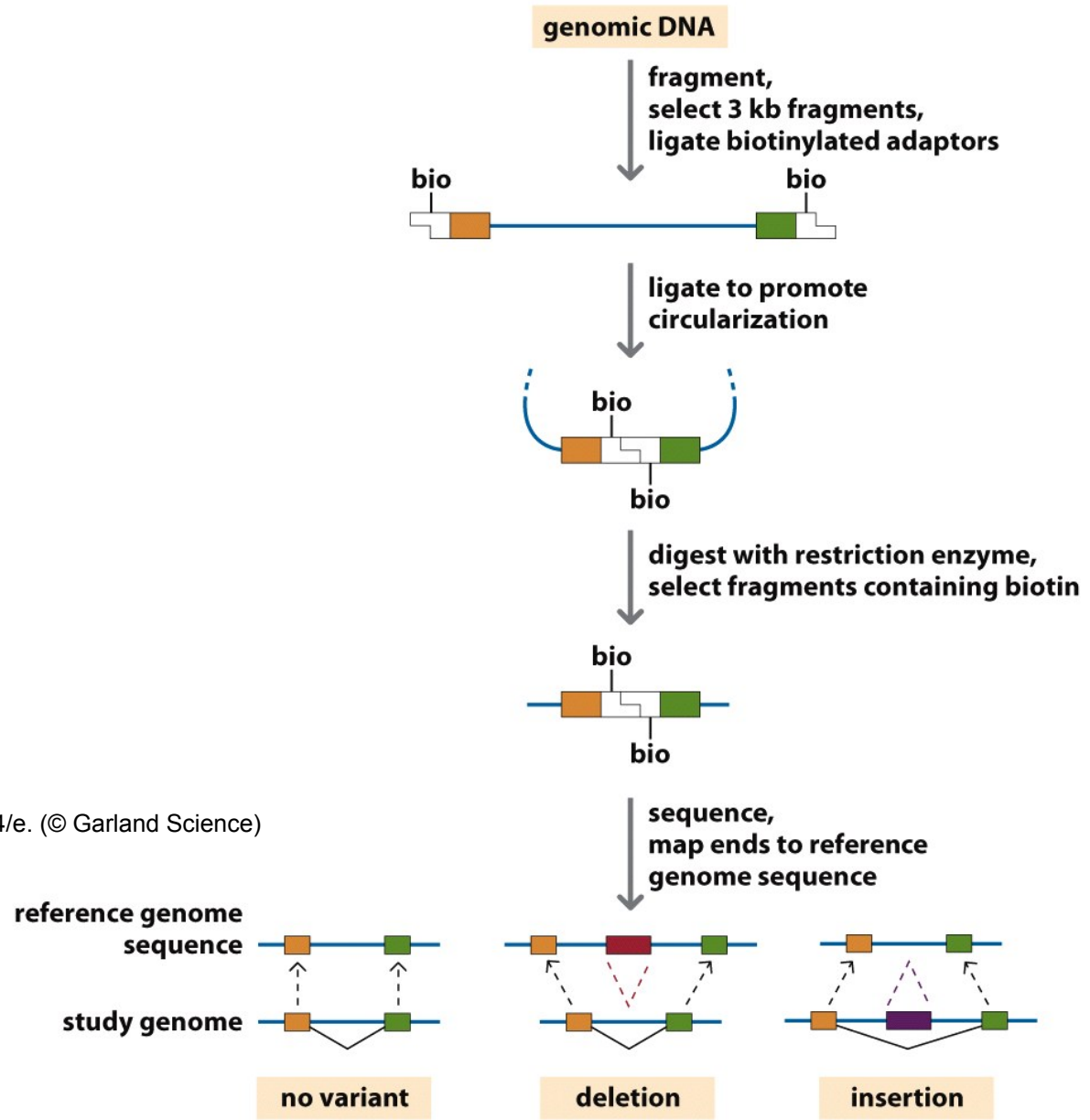
*PLoS Comput. Biol.* (2010) 6(2): e1000667

# Finding copy number variants with NGS



*Trends Biotechnol* (2009) **27**: 448-54

# Deletions and amplifications with paired end sequencing



Human Molecular Genetics, 4/e. (© Garland Science)

# Whole-genome sequencing to identify disease gene

The NEW ENGLAND JOURNAL of MEDICINE

## ORIGINAL ARTICLE

### Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy

James R. Lupski, M.D., Ph.D., Jeffrey G. Reid, Ph.D., Claudia Gonzaga-Jauregui, B.S., David Rio Deiros, B.S., David C.Y. Chen, M.Sc., Lynne Nazareth, Ph.D., Matthew Bainbridge, M.Sc., Huyen Dinh, B.S., Chyn Jing, M.Sc., David A. Wheeler, Ph.D., Amy L. McGuire, J.D., Ph.D., Feng Zhang, Ph.D., Pawel Stankiewicz, M.D., Ph.D., John J. Halperin, M.D., Chengyong Yang, Ph.D., Curtis Gehman, Ph.D., Danwei Guo, M.Sc., Rola K. Irikat, B.S., Warren Tom, B.S., Nick J. Fantin, B.S., Donna M. Muzny, M.Sc., and Richard A. Gibbs, Ph.D.

## ABSTRACT

### BACKGROUND

Whole-genome sequencing may revolutionize medical diagnostics through rapid identification of alleles that cause disease. However, even in cases with simple patterns of inheritance and unambiguous diagnoses, the relationship between disease

From the Department of Molecular and Human Genetics (J.R.L., J.G.R., C.G.-J., M.B., F.Z., P.S., D.M.M., R.A.G.), the Hu-

### METHODS

We identified a family with a recessive form of Charcot–Marie–Tooth disease for which the genetic basis had not been identified. We sequenced the whole genome of the proband, identified all potential functional variants in genes likely to be related to the disease, and genotyped these variants in the affected family members.

.edu.

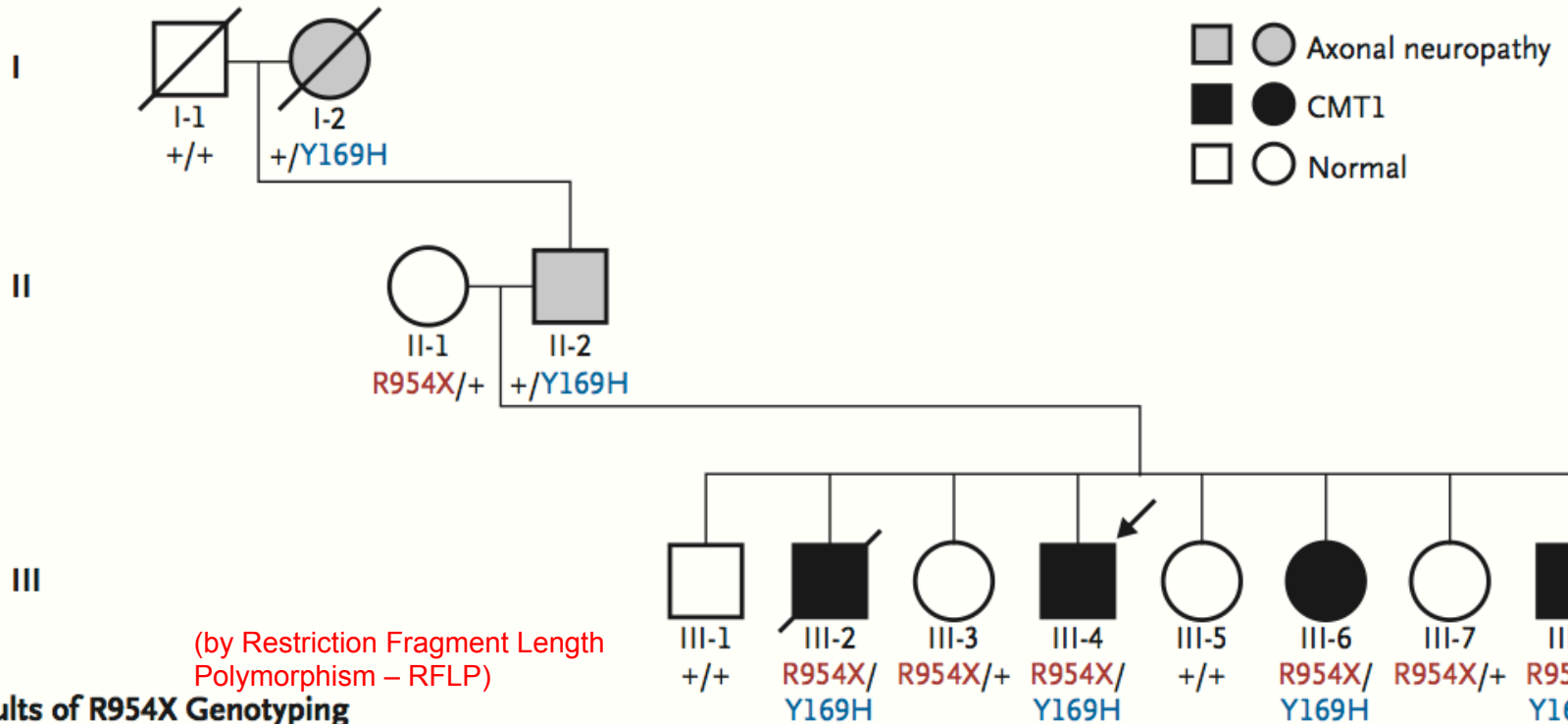
This article (10.1056/NEJMoa0908094) was published on March 10, 2010, at NEJM.org.

N Engl J Med 2010;362:1181-91.

Copyright © 2010 Massachusetts Medical Society.

# Genotyping to confirm disease allele

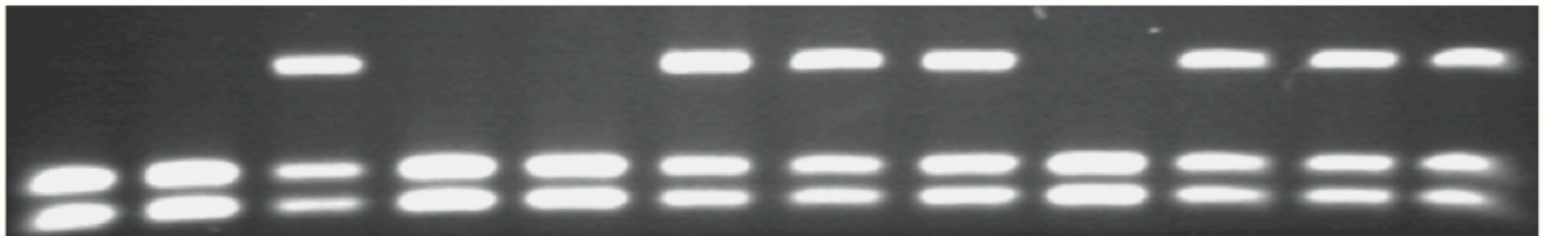
## A SH3TC2 Genotype and Phenotype



## B Results of R954X Genotyping

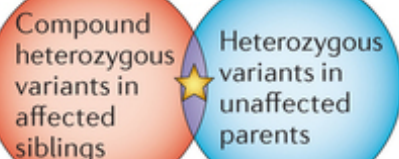
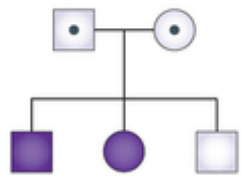
G→A mutant  
(R954X)

Wild type

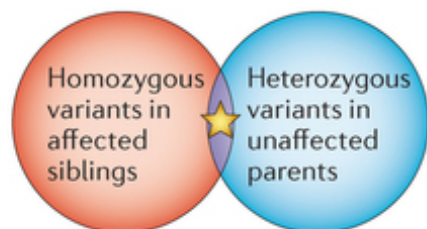
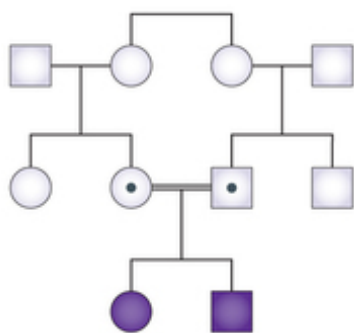


**a Inherited mutations**

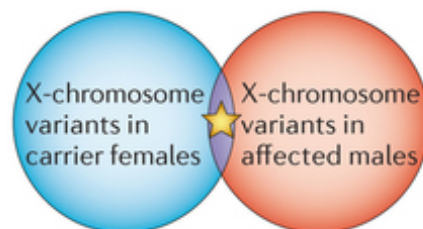
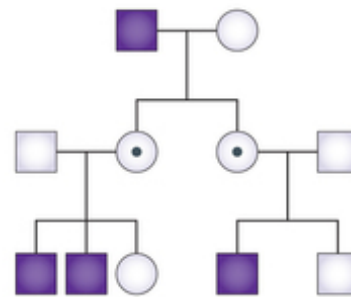
Autosomal recessive



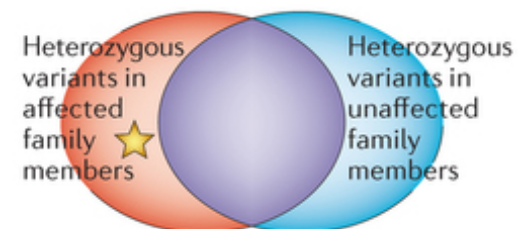
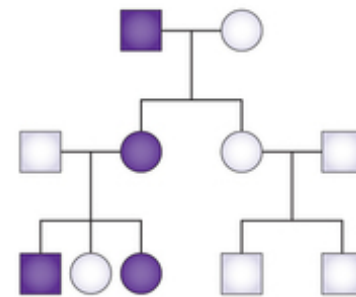
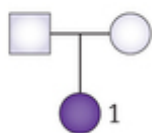
Consanguineous autosomal recessive



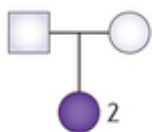
X-linked recessive



Autosomal dominant

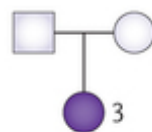
**b De novo dominant mutations**

Patient 1 heterozygous variants

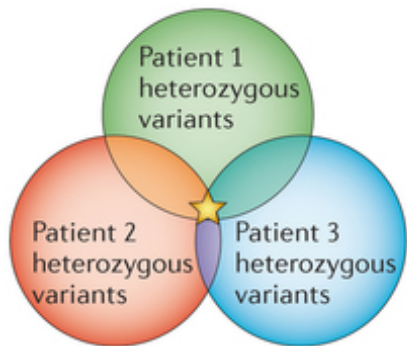
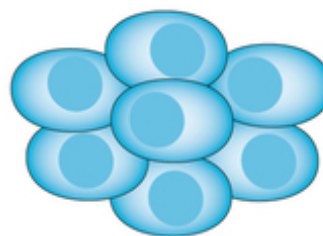


Patient 2 heterozygous variants

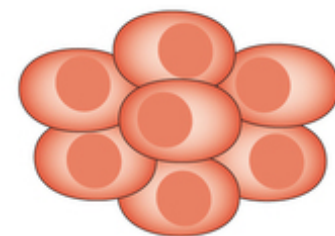
Patient 3 heterozygous variants



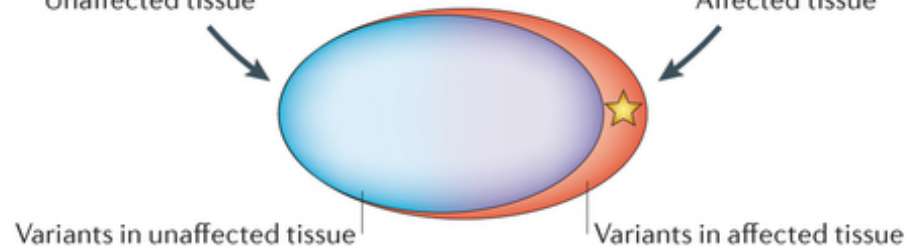
3

**c Mosaic mutations**

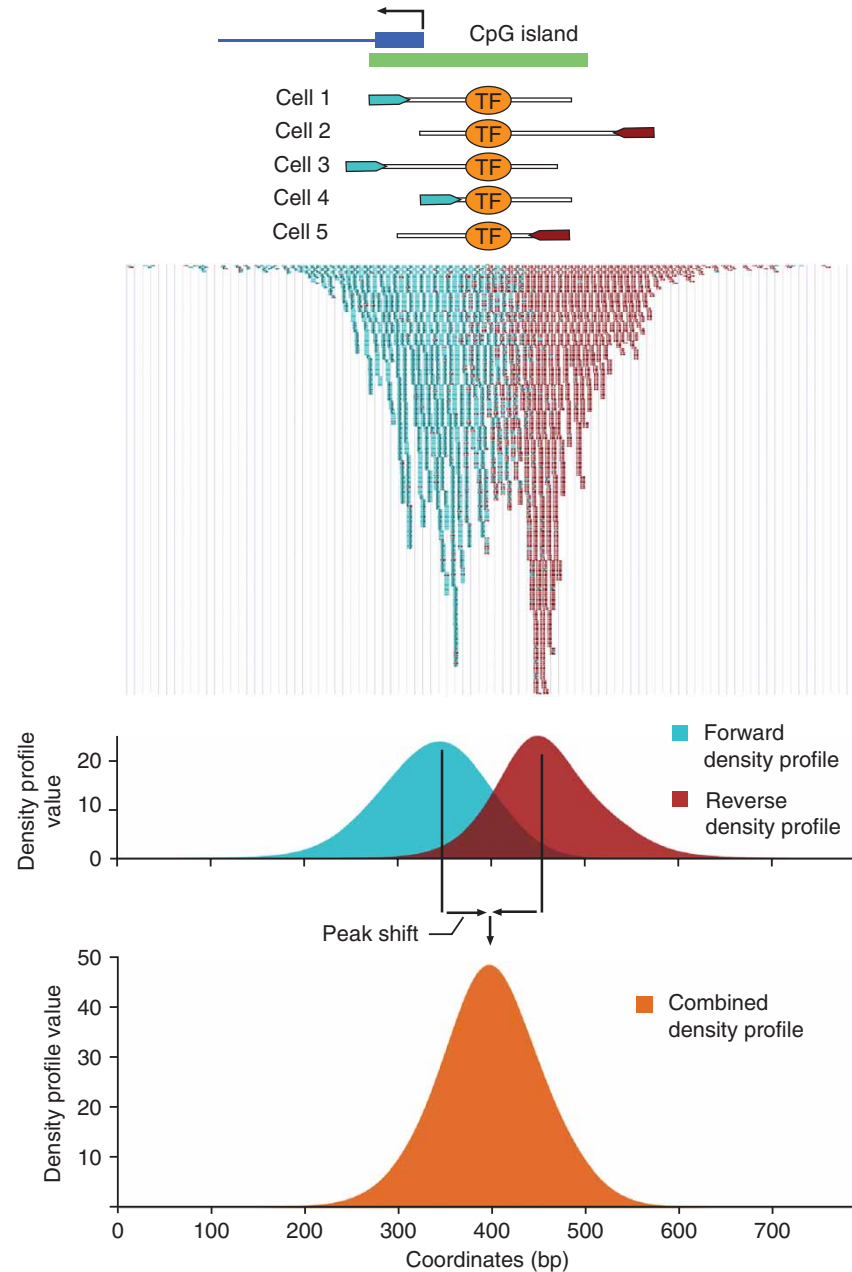
Unaffected tissue

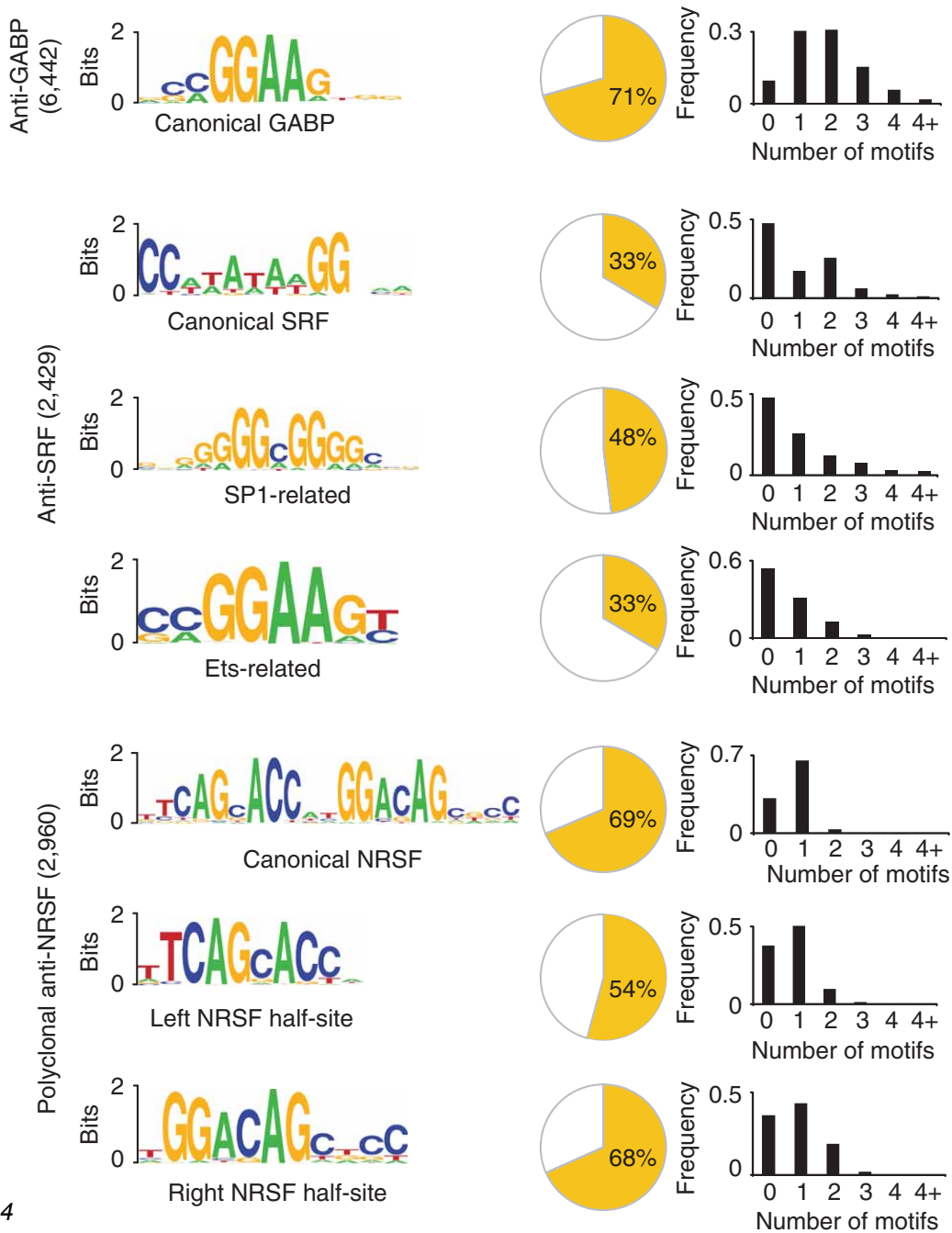


Affected tissue



# Chromatin immunoprecipitation (ChIP)-seq





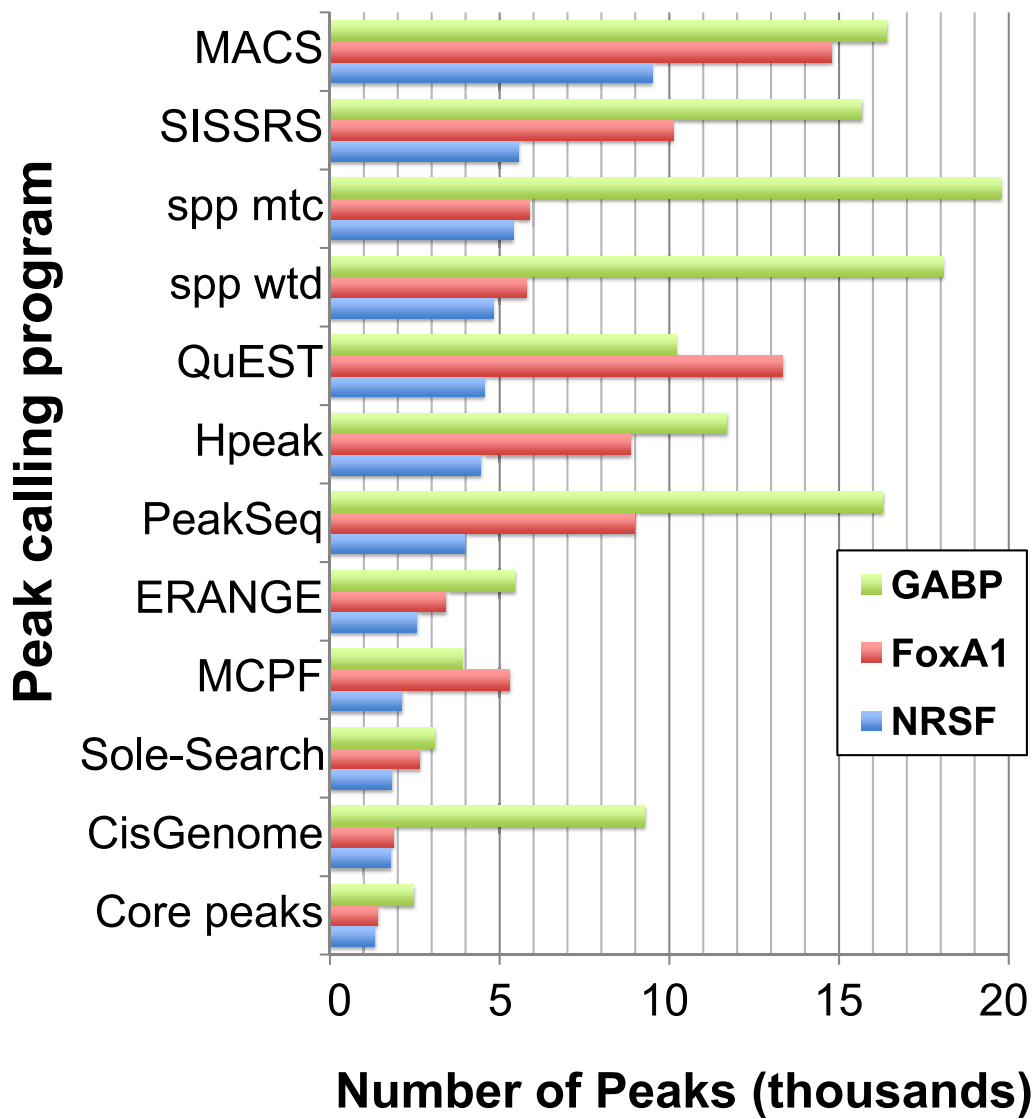


Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific density	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X			X	X		X		X		conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1		X			X				X			
E-RANGE	27	3.1		X			X				X	X		chromosome scale Poisson dist.
MACS	13	1.3.5		X			X			X		X		local Poisson dist.
QuEST	14	2.3			X		X			X**		X		chromosome scale Poisson dist.
HPeak	29	1.1		X			X					X		Hidden Markov Model
Sole-Search	23	1	X	X			X		X			X		One sample t-test
PeakSeq	21	1.01		X			X					X		conditional binomial model
SISSRS	32	1.4		X		X					X			
spp package (wtd & mtc)	31	1.7		X		X		X	X'	X				
				<b>Generating density profiles</b>			<b>Peak assignment</b>		<b>Adjustments w. control data</b>		<b>Significance relative to control data</b>			

X\* = Windows-only GUI or cross-platform command line interface

X\*\* = optional if sufficient data is available to split control data


X' = method excludes putative duplicated regions, no treatment of deletions



**A**  Sequence of the human genome  
One dimension

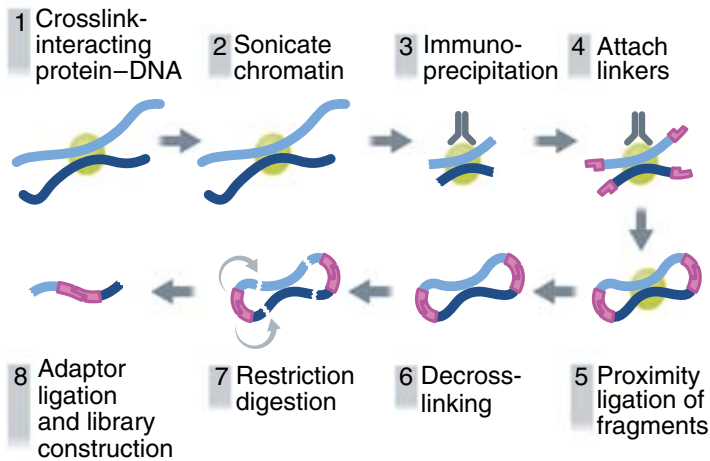
```

ATCGATCCGTCCGAGACCTAGTC
GATCGATCGCCAAATCGATCGGA
TCGACTGTCTTAGCGCTAGCCGA
GATCTGCTAGGTCGTGTGACAAA
    
```

**B**  Genomic rearrangements by paired-end sequencing  
Two dimensions

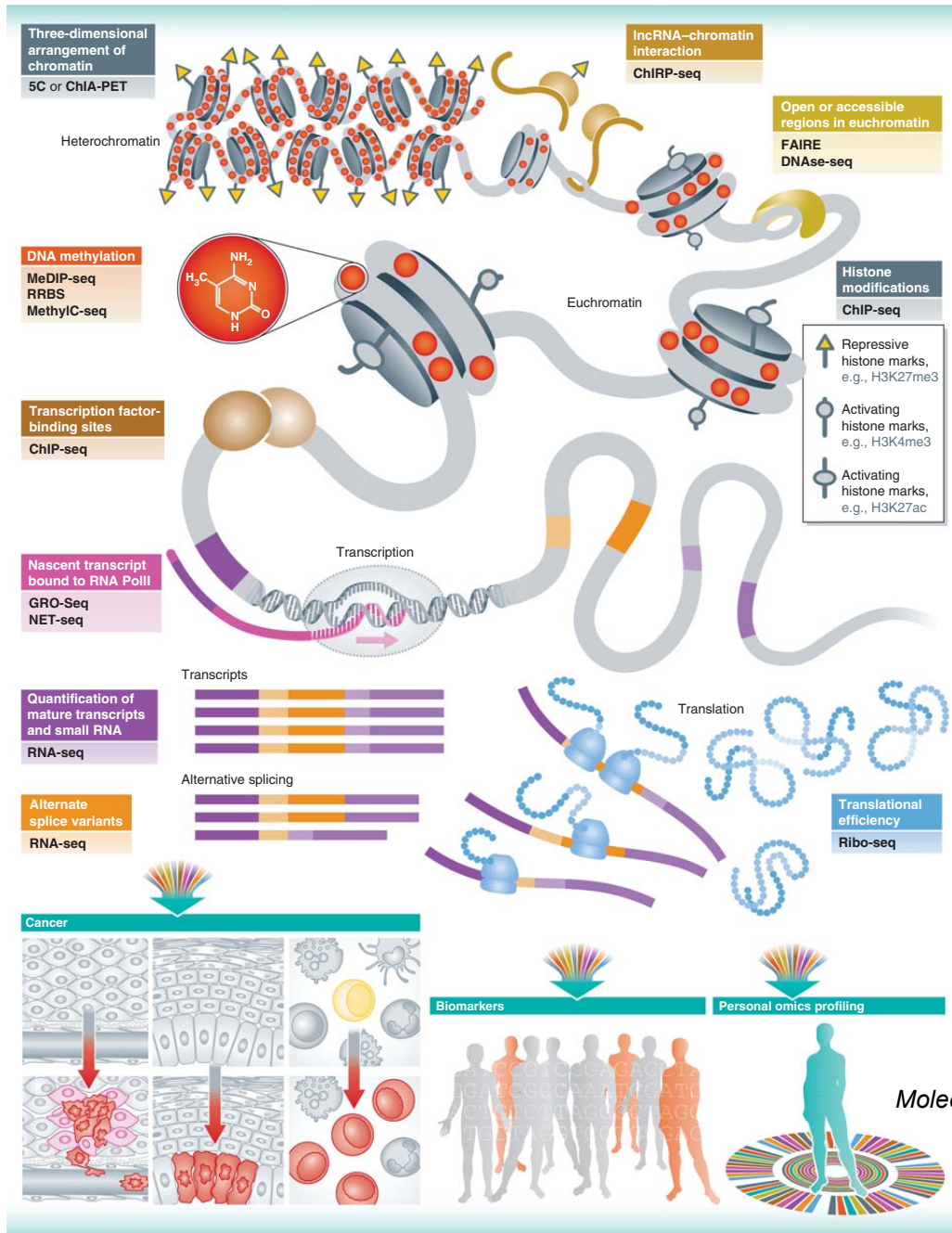


**C**  Chromosome conformations by HiC and ChIA-PET  
Three dimensions



**D**  Longitudinal sequencing  
Four dimensions





**Table I** The various NGS assays employed in the ENCODE project to annotate the human genome

Feature	Method	Description	Reference
Transcripts, small RNA and transcribed regions	RNA-seq CAGE	Isolate RNA followed by HT sequencing HT sequencing of 5'-methylated RNA	(Waern <i>et al</i> , 2011) (Kodzius <i>et al</i> , 2006)
	RNA-PET ChIRP-Seq	CAGE combined with HT sequencing of poly-A tail Antibody-based pull down of DNA bound to lncRNAs followed by HT sequencing	(Fullwood <i>et al</i> , 2009c) (Chu <i>et al</i> , 2011)
	GRO-Seq	HT sequencing of bromouridinated RNA to identify transcriptionally engaged PolII and determine direction of transcription	(Core <i>et al</i> , 2008)
	NET-seq	Deep sequencing of 3' ends of nascent transcripts associated with RNA polymerase, to monitor transcription at nucleotide resolution	(Churchman and Weissman, 2011)
	Ribo-Seq	Quantification of ribosome-bound regions revealed uORFs and non-ATG codons	(Ingolia <i>et al</i> , 2009)
Transcriptional machinery and protein-DNA interactions	ChIP-seq	Antibody-based pull down of DNA bound to protein followed by HT sequencing	(Robertson <i>et al</i> , 2007)
	DNase footprinting	HT sequencing of regions protected from DNase1 by presence of proteins on the DNA	(Hesselberth <i>et al</i> , 2009)
	DNase-seq	HT sequencing of hypersensitive non-methylated regions cut by DNase1	(Crawford <i>et al</i> , 2006)
	FAIRE	Open regions of chromatin that is sensitive to formaldehyde is isolated and sequenced	(Giresi <i>et al</i> , 2007)
	Histone modification	ChIP-seq to identify various methylation marks	(Wang <i>et al</i> , 2009a)
DNA methylation	RRBS	Bisulfite treatment creates C to U modification that is a marker for methylation	(Smith <i>et al</i> , 2009)
Chromosome-interacting sites	5C	HT sequencing of ligated chromosomal regions	(Dostie <i>et al</i> , 2006)
	ChIA-PET	Chromatin-IP of formaldehyde cross-linked chromosomal regions, followed by HT sequencing	(Fullwood <i>et al</i> , 2009a)



[www.illumina.com](http://www.illumina.com)

The HiSeq X™ Ten, **composed of 10 HiSeq X Systems**, is the first sequencing platform that breaks the \$1000 barrier for a 30x human genome.

# Raw NGS data (FASTQ file)

## 4 lines per sequence

```
@HWI-ST1097:104:D13TNACXX:4:1101:18100:2240 1:Y:0:CAACTA  
TGAGGCAAACCCAACCTTATATGGGTCAATATAATGGTAAAGAAGGTTTAAA  
+  
=7=<+2<AACAA<A+<A97AB7<7+2?ABBA@@B4A1?7A<*:.;00=AAA  
@HWI-ST1097:104:D13TNACXX:4:1101:18326:2181 1:N:0:CAACTA  
CATACATCAAATTTTACAAAACCTCGAATCTCGGTGGTATTATTCCGACAG  
+  
CCCFHHHHGJJJIIIGJGHJIHGIGHGDEIGIHFHIIIIGG>?DH6  
@HWI-ST1097:104:D13TNACXX:4:1101:18259:2224 1:N:0:CAACTA  
CAGGTGGAGGGACCGGGTAGTGCCGGATCAAGTAGTGTAGTATTTATTGTA  
+  
@C@DBDFDHHGHHI<DHGH1CFGHIIIBHBIIIEHIIIGFDGAGE>GGHGC@E  
@HWI-ST1097:104:D13TNACXX:4:1101:18256:2243 1:N:0:CAACTA  
GATAGGTTTGTATGATCTAATTGGTGGCAACTGGGTCCCTCCCATCCTAGC  
+  
@@@FFFDDDFDFHGIJGIJJIICGHIJJAHGGGHGBFGGIHIFGGEAADHG  
@HWI-ST1097:104:D13TNACXX:4:1101:18728:2073 1:N:0:CAACTA  
TTTCTTTCGAAGGAACCCCTCTTCTCATGCTTTGTGCTACTCTGAGGCAA  
+  
@@@DDDDHHD1<AEHGGGG<FHGIEHEH9CDDA*??D<DDHHAG<?1?1?
```

Read ID

Sequence

Optional Read ID

Base quality scores

Filesize: few hundred Mb to 1 or 2 Gb, ~100 million lines for one experiment!

# Encoding quality scores

Quality character	!"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
ASCII Value	33 43 53 63 73
Base Quality (Q) (ASCII-33)	0 10 20 30 40

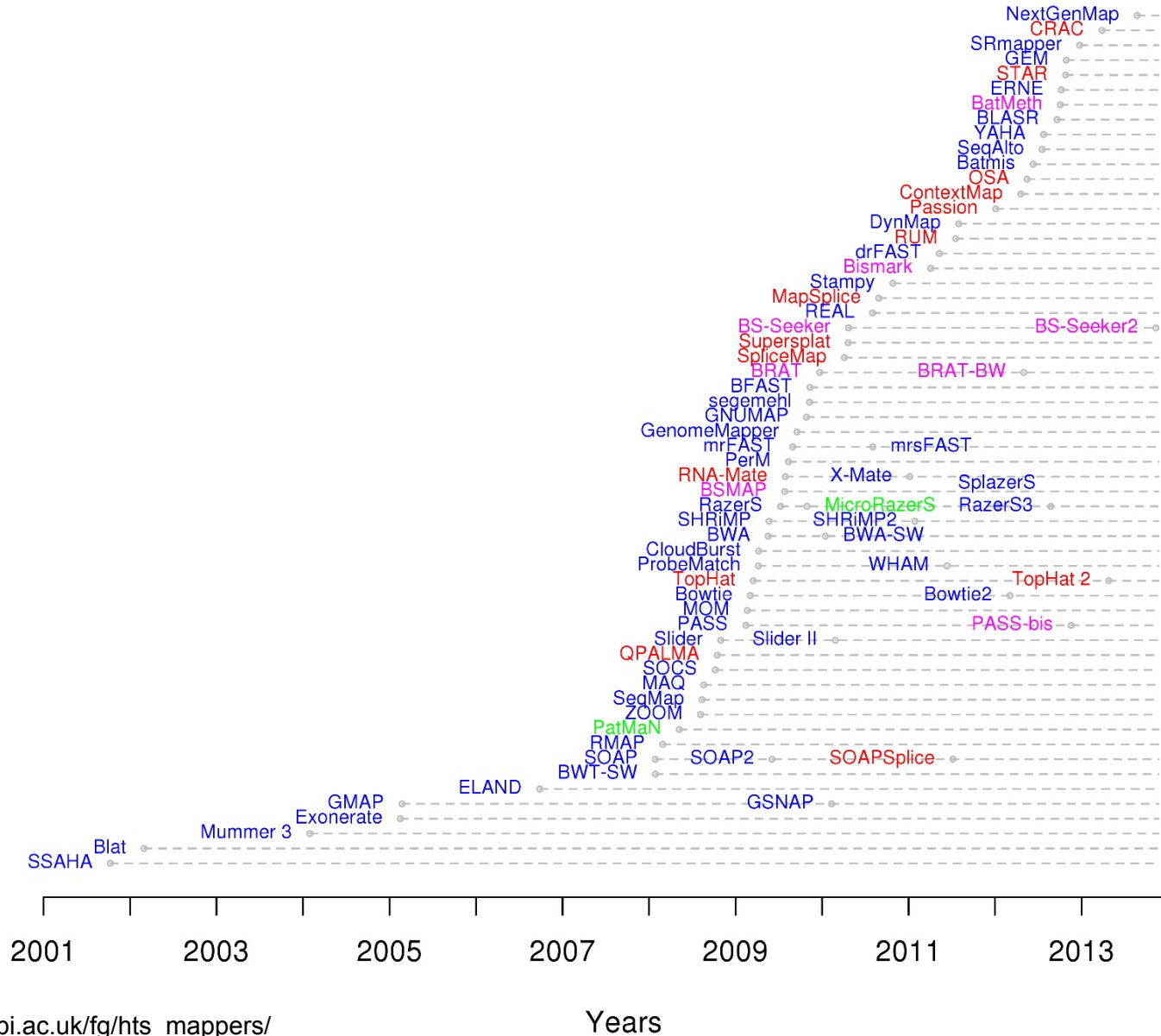
Probability of error =  $10^{-Q/10}$

This is a **Phred** score, a standard measure of sequencing quality

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%



# NGS aligners

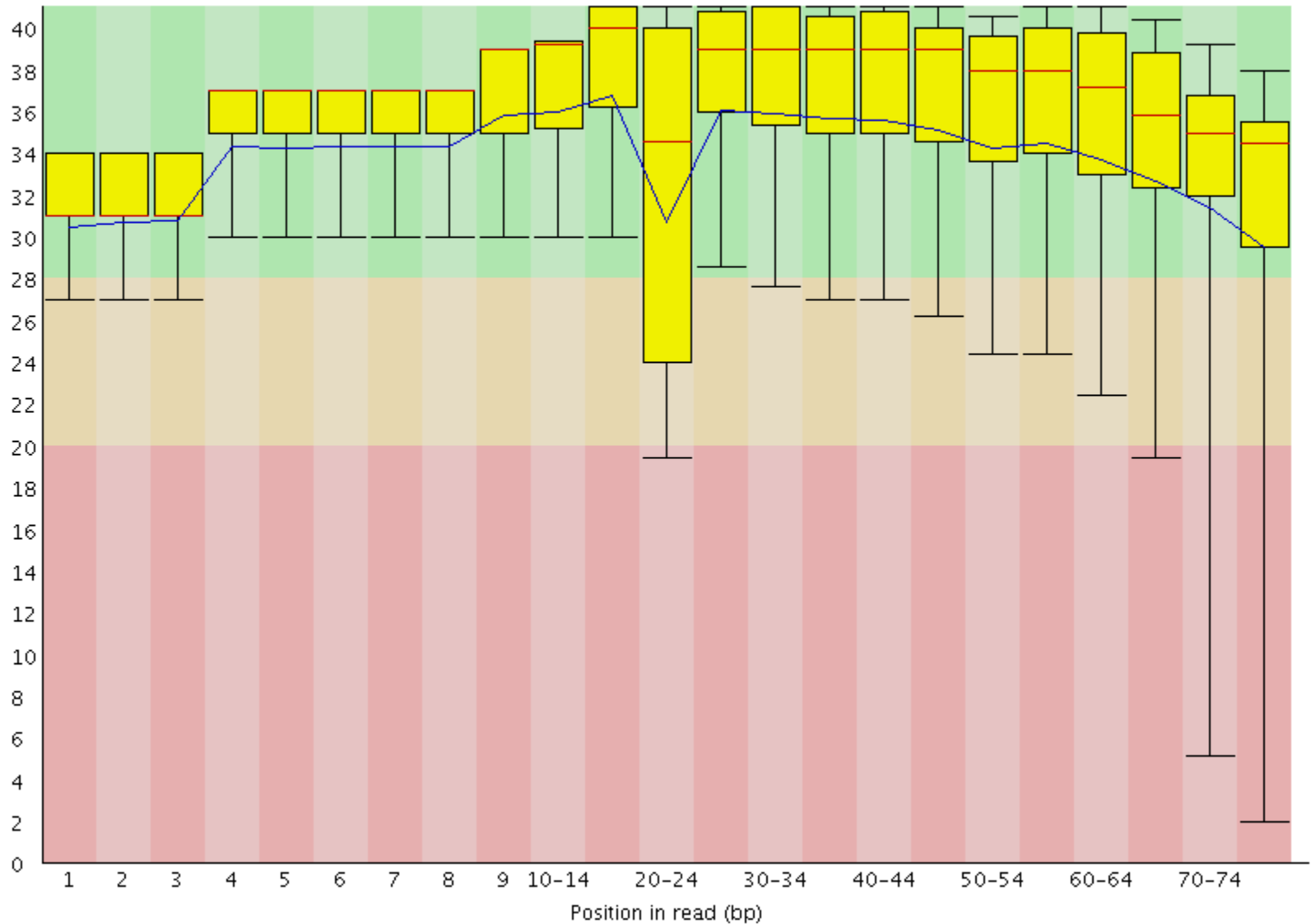


# NGS aligners

Mapper	O.S.	Seq. Plat.	Input	Output	Min. RL	Max. RL	Mis-matches	Indels	Gaps	Splicing
<b>BFAST</b>	Linux,Mac	I,So,4, Hel	(C)FAST(A/Q)	SAM TSV		*	Y	Y	Y	N
<b>Blat</b>	Linux,Mac	N	FASTA	TSV BLAST		115000K	Score	Score	Y	De novo
<b>Bowtie</b>	Linux,Mac,Windows	I,So,4,Sa,P	(C)FAST(A/Q)	SAM TSV		41K	Score	Score	N	N
<b>Bowtie2</b>	Linux,Mac,Windows	I,4,Ion	FASTA/Q	SAM TSV		45000K	Score	Score	Y	N
<b>BS-Seeker2</b>	Linux, Unix, Mac	I	FASTA/Q, qseq	SAM BAM	10	200	Score	Score	Y	N
<b>BWA</b>	Linux,Mac,Windows	I,So,4,Sa,P	FASTA/Q	SAM	4	200	Y		8Y	N
<b>CloudBurst</b>	Linux,Mac,Windows	N	FASTA	TSV		1K	Y	Y	Y	N
<b>ELAND</b>	Linux, Unix, Mac	I	FASTA		15	150		2Score	N	N
<b>GMAP</b>	Linux,Unix,Mac,Windows	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM GFF Native		8*	Y	Y	Y	De novo
<b>MapReads</b>	Linux,Mac,Windows	So	FASTA/Q	TSV	10	120	Score		0N	N
<b>MAQ</b>	Linux,Mac	I,So	(C)FAST(A/Q)	TSV	8	63	Y	Y	N	N
<b>MOSAİK</b>	Linux,Unix,Mac,Windows	I,So,4,Sa,Hel,Ion,P	(C)FAST(A/Q)	BAM	15	1000	Y	Y	Y	N
<b>mrFAST</b>	Linux,Unix	I	FASTA/Q	SAM DIVET	25	1000	Score		4N	N
<b>Novoalign(CS)</b>	Linux	I,So,4,Hel,Ion	(C)FAST(A/Q) Illumina	SAM Native	1	250	Y	Y	Y	Lib
<b>RMAP</b>	Linux,Mac	I,So,4	(C)FAST(A/Q)	BED	11	10K	Y		0N	N
<b>SHRiMP2</b>	Linux, Unix, Mac	I,So,4	FASTA/Q	SAM	30	1K	Y	Score	N	N
<b>SOAP2</b>	Linux	I	FASTA/Q	SAM TSV	27	1K		2	0Y	N
<b>SOAPSplice</b>	Linux,Mac	I,4	FASTA/Q	TSV	13	3K		5	2Y	De novo
<b>SSAHA2</b>	Linux,Mac	I,4,Sa	FASTA/Q	SAM	15	48K	Score	Score	N	N
<b>TopHat 2</b>	Linux,Mac	I	FASTA/Q	BAM					N	De novo
<b>VMATCH</b>	Linux,Mac	N	FASTA	TSV			Score	Score	Y	N

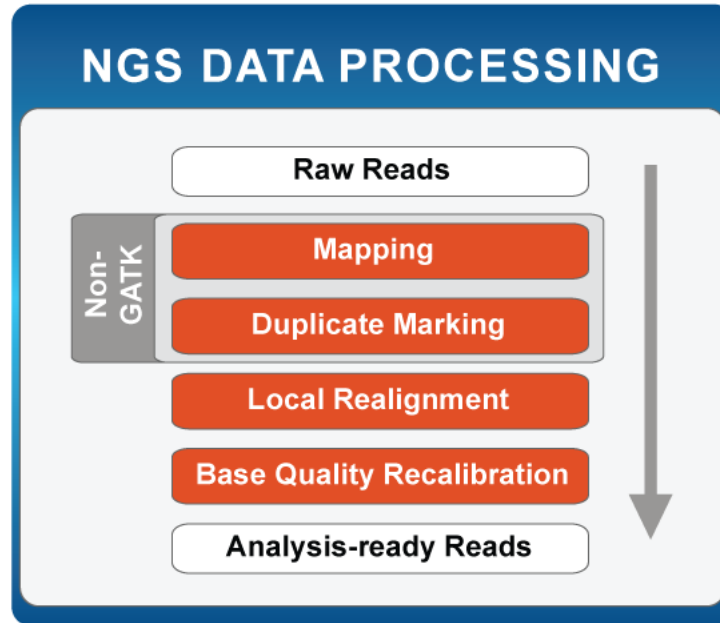
# FASTQC

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

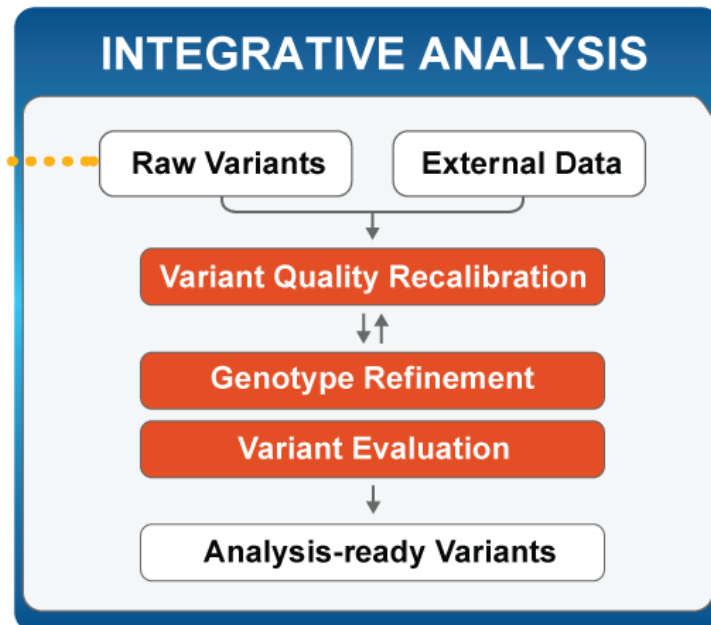
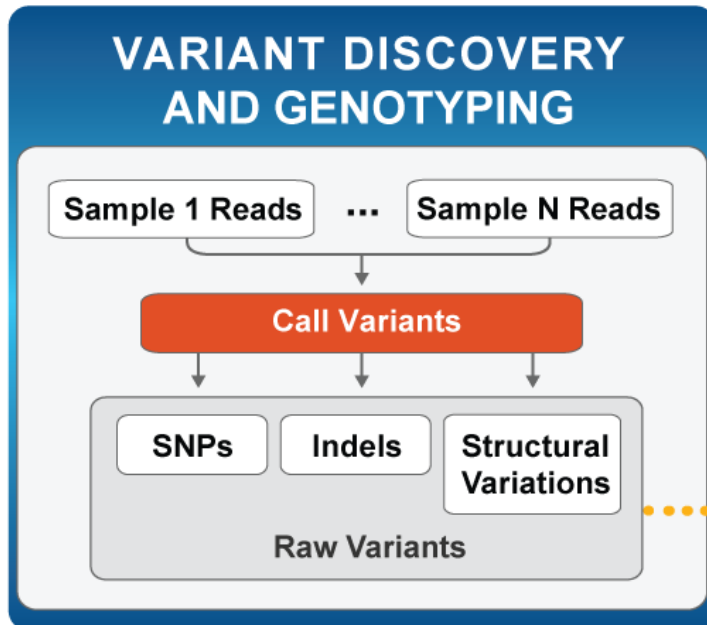




# Variant calling with GATK

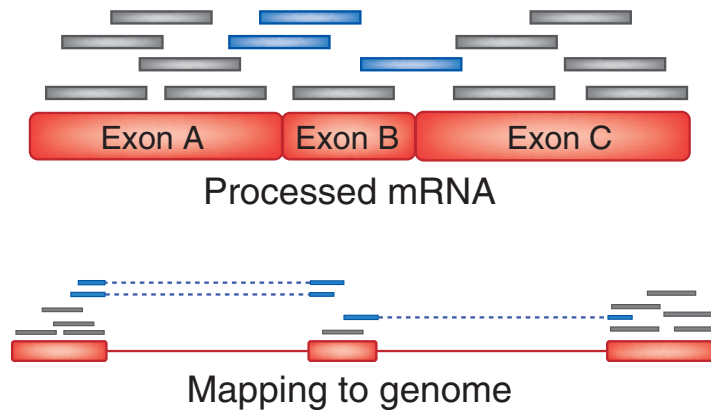


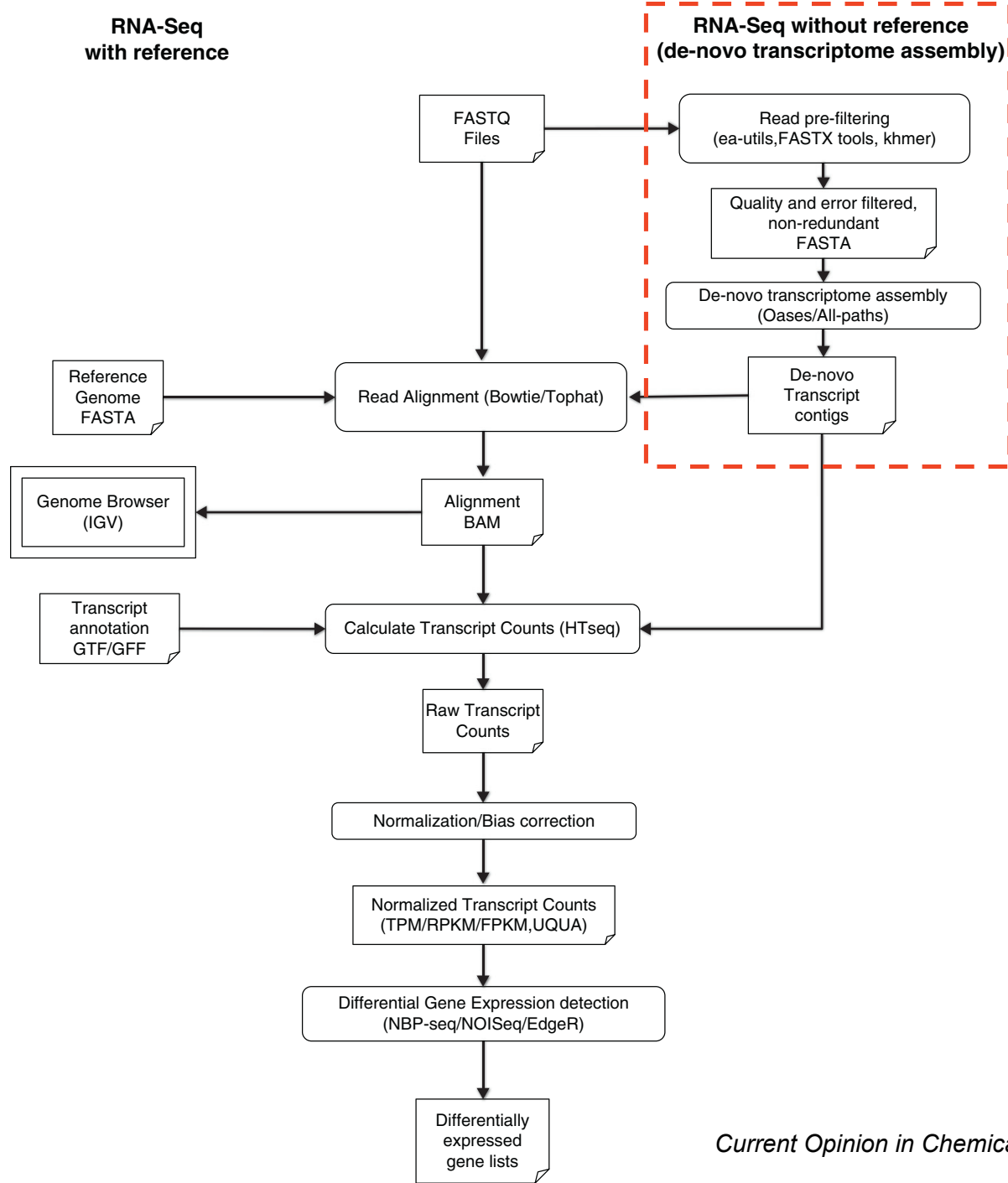
[www.broadinstitute.org/gatk/](http://www.broadinstitute.org/gatk/)





# RNA-seq alignments







# What lies ahead...?

- End-to-end genome sequencing
- Sequencing entire pedigrees
- Sequencing within intact cells
- Single-cell genomes, transcriptomes, epigenomes
- Protein-protein interactions by sequencing
- Cell fate mapping
- Single molecule protein sequencing