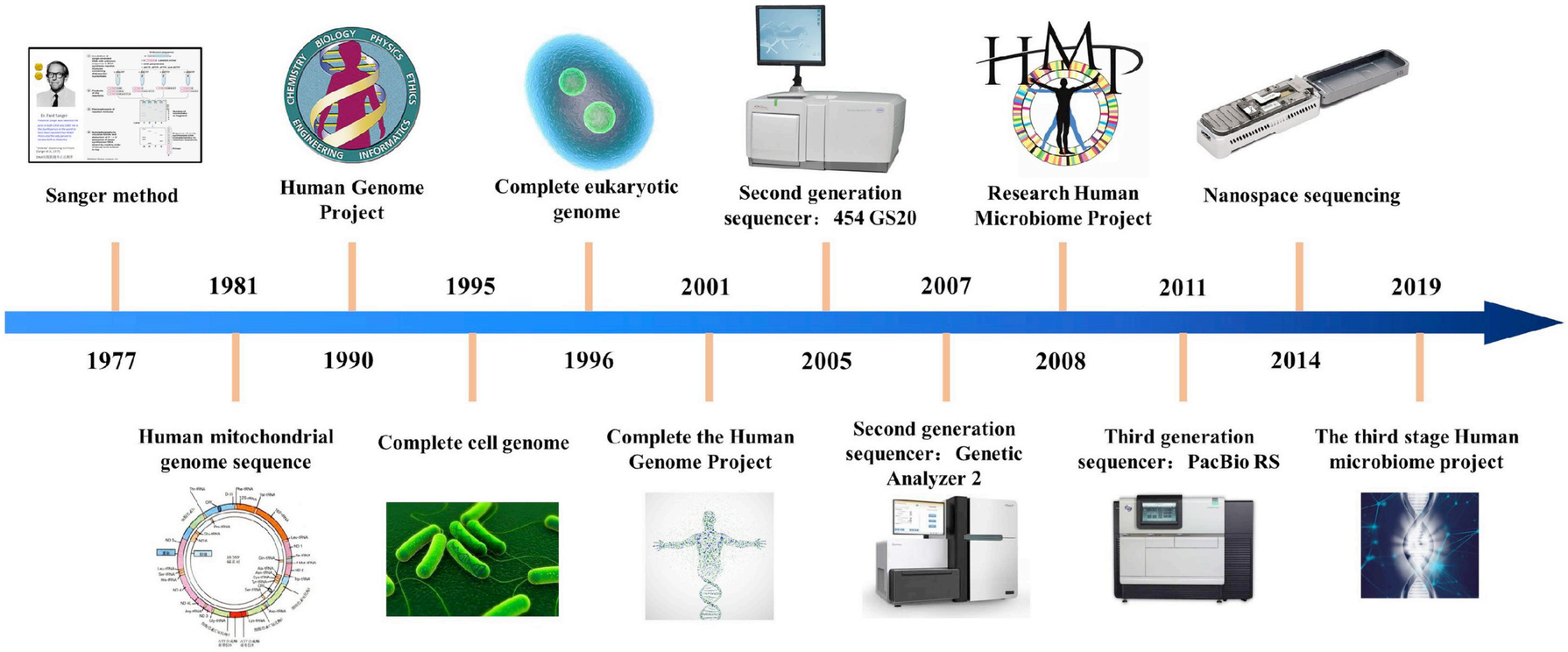
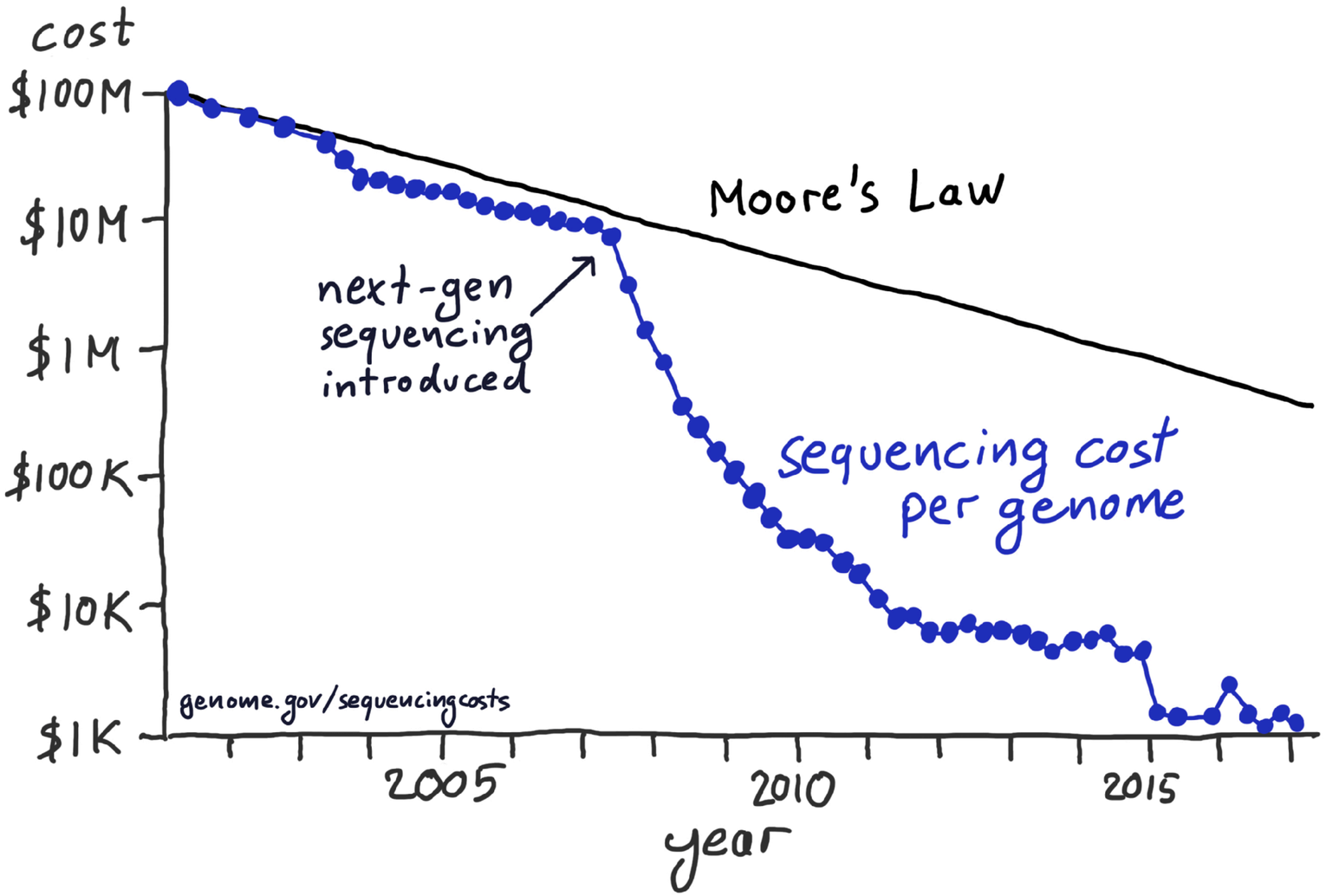
Introduction to Next Generation Sequencing

DNA Sequencing Technologies, Past & Present



ian.simpson@ed.ac.uk

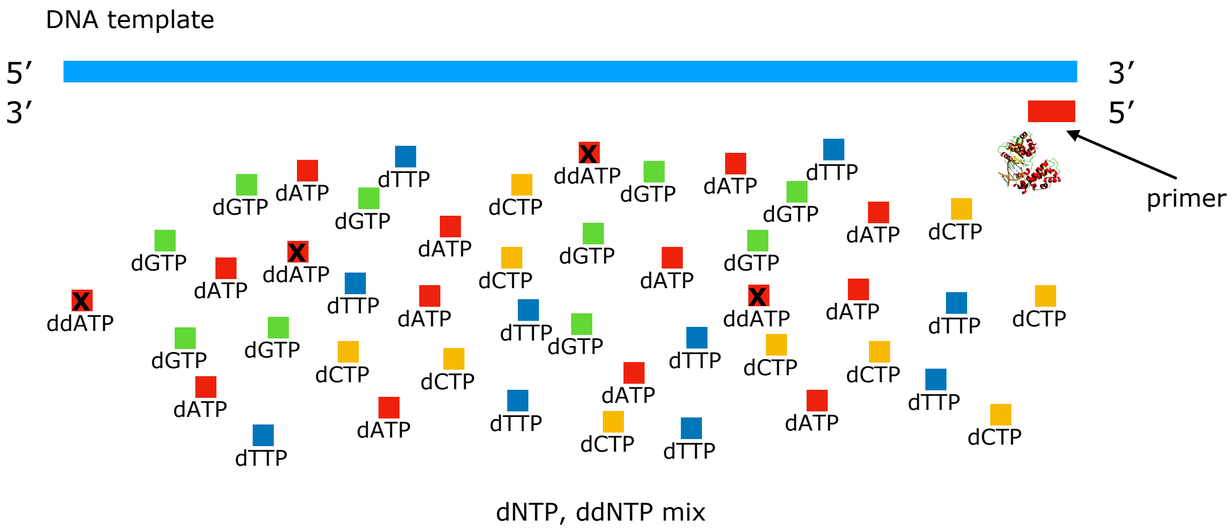
Sequencing Costs



ian.simpson@ed.ac.uk

Sanger Sequencing

DNA template
DNA primer
DNA polymerase (5′-3′ nuclease)
dNTPs (dATP, dCTP, dGTP & dTTP)
ddNTPs (lack 3′-OH, no extension)

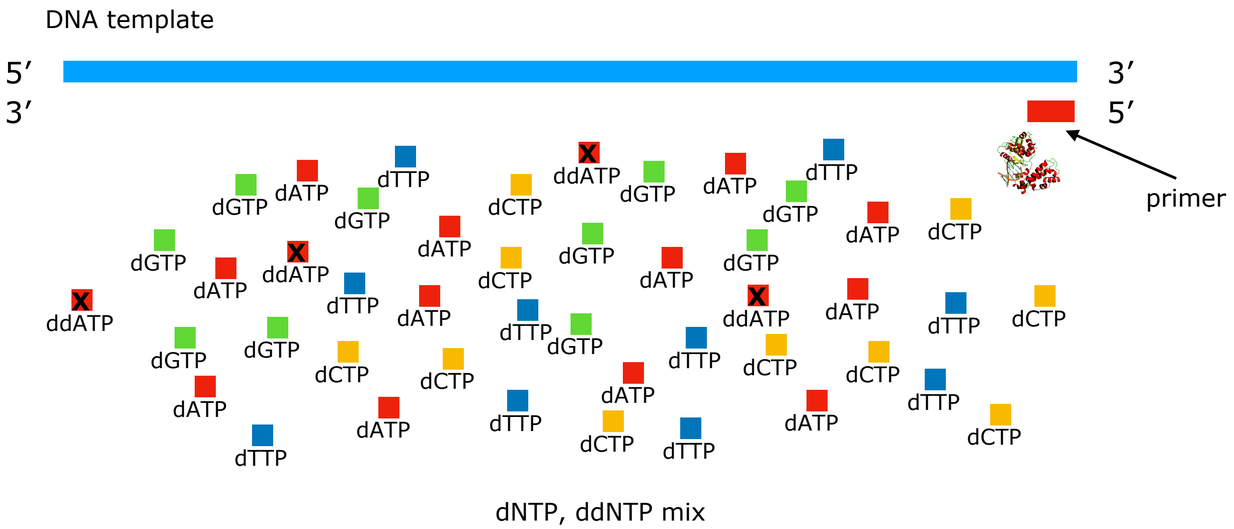


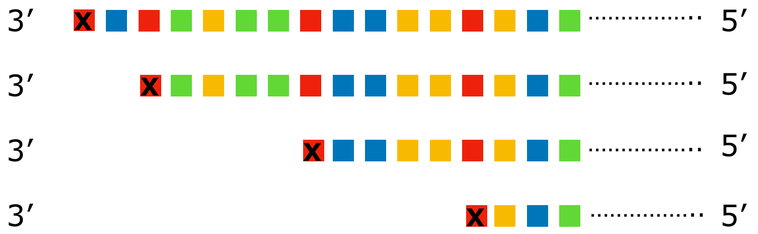
dNTP, ddNTP mix

DNA polymerase

ian.simpson@ed.ac.uk

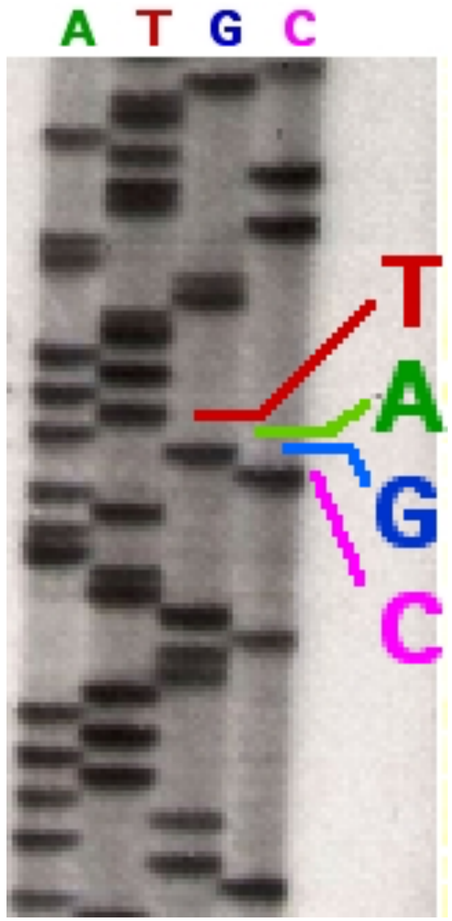
Sanger Sequencing



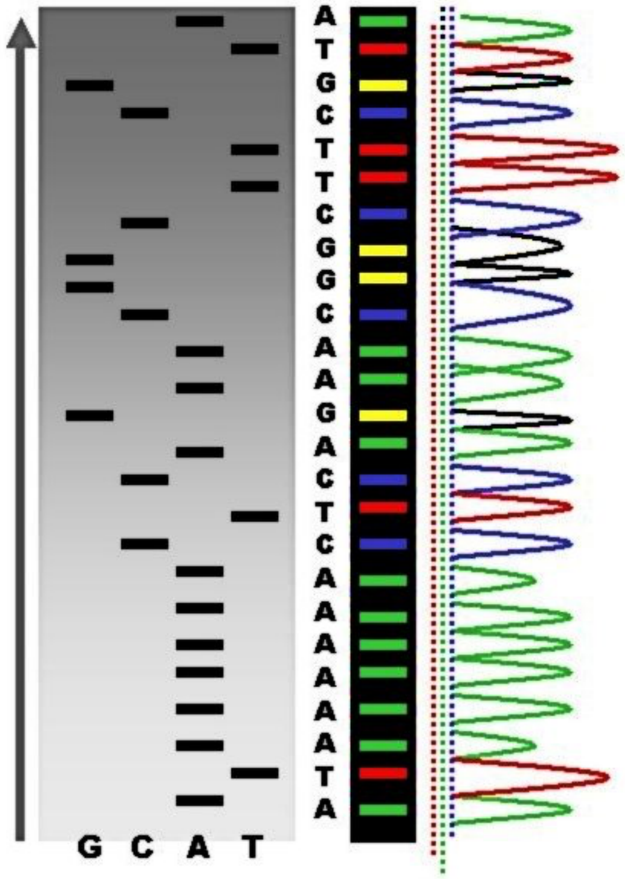


ian.simpson@ed.ac.uk

Sanger Sequencing



Radioactive sequencing
with 4 ddNTP lanes

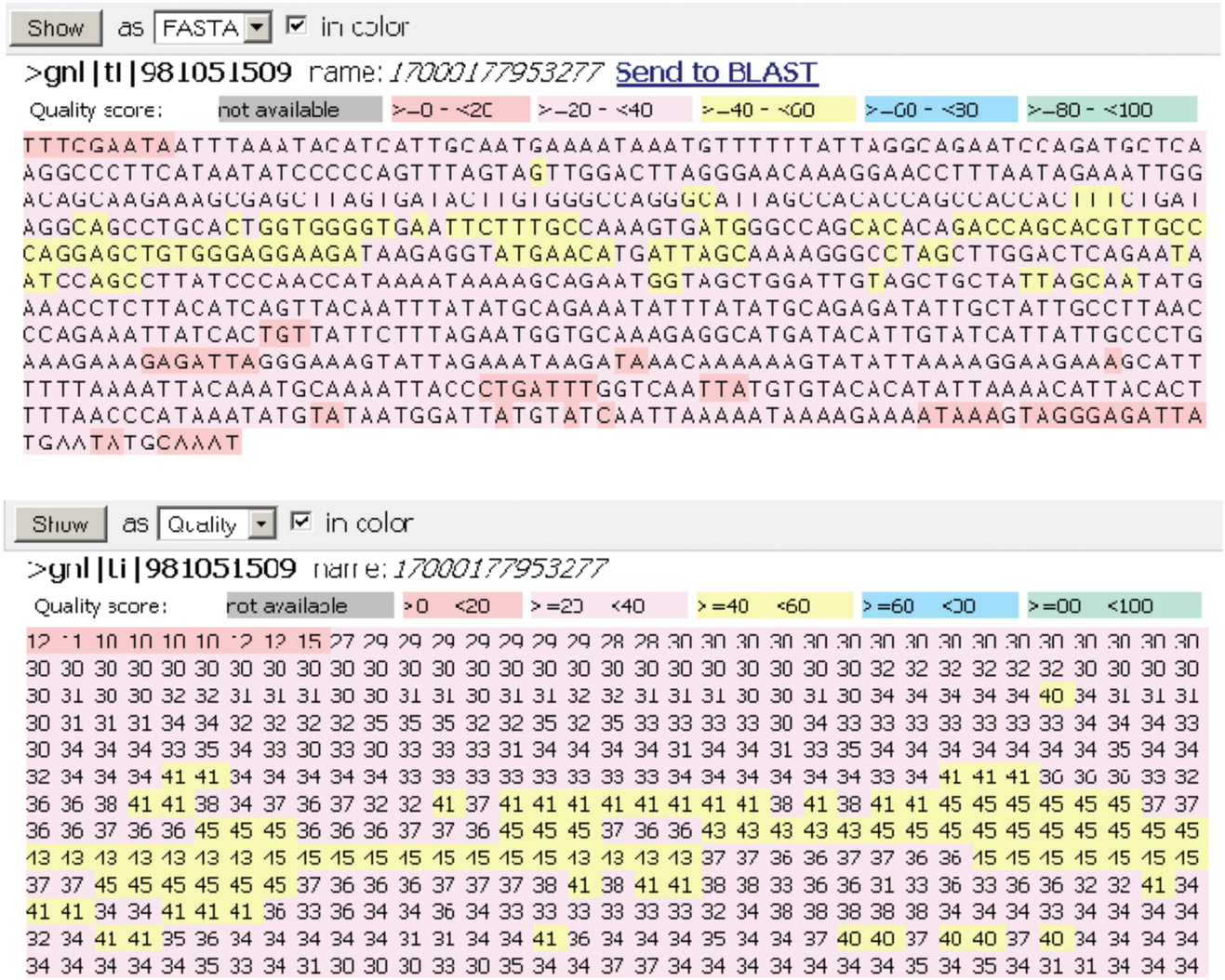


Automated fluorescent sequencing with
all 4 terminators in the same tube

ian.simpson@ed.ac.uk

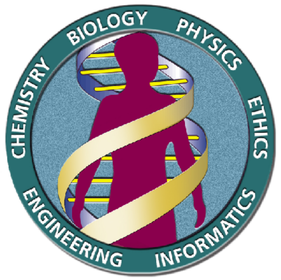
European Nucleotide Archive

https://www.ebi.ac.uk/ena/browser/home



ian.simpson@ed.ac.uk

Sanger Sequencing



CELERA
A PECorporation Business

Generally only possible to get 200bp-1kb of sequence per run
- generally due to length of gel/capillary used in the system

Low throughput, expensive, prone to failure

Produces lots of fragments that need to be assembled (ideally computationally)

Originally used to sequence the Human Genome

The Human Genome Project
- Strategy:-
  - Make a “large insert” library of fragments from the genome
  - From that make overlapping small insert fragment libraries
  - Sequence the small libraries and assemble to re-create the large insert
  - Map the large inserts with respect to each other to create a draft scaffold

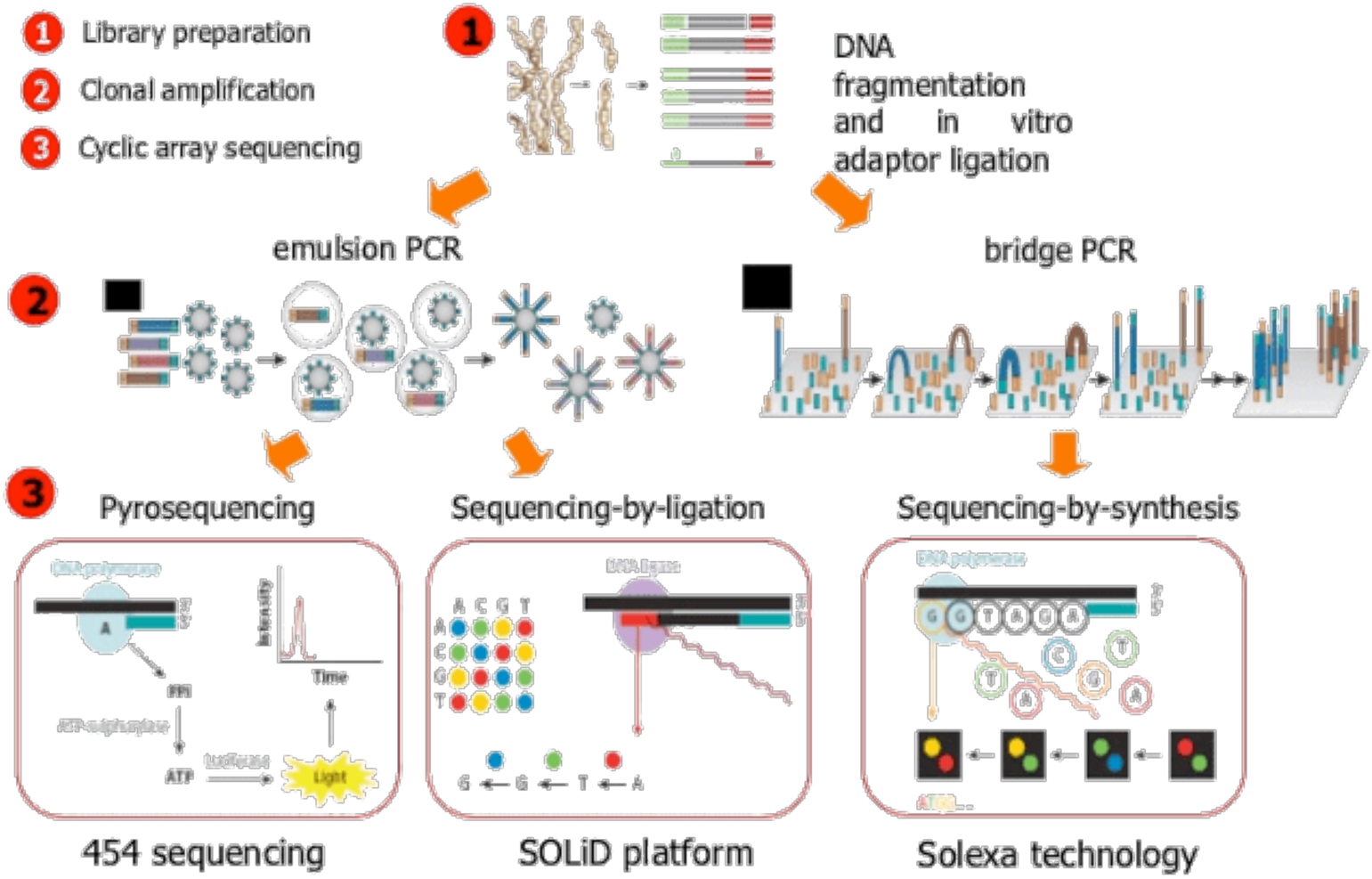
Expensive, time consuming, it took c.13 years and c.$3billion to complete (Feb15th 2001)

The Celera/Ventner (commercial) competition
- Strategy:-
  - Clone the genome straight into overlapping small insert fragment libraries
  - Sequence the small libraries and assemble to create a draft scaffold

Less expensive, more error prone, it took c.3 years and $0.3billion (Feb 16th 2001)

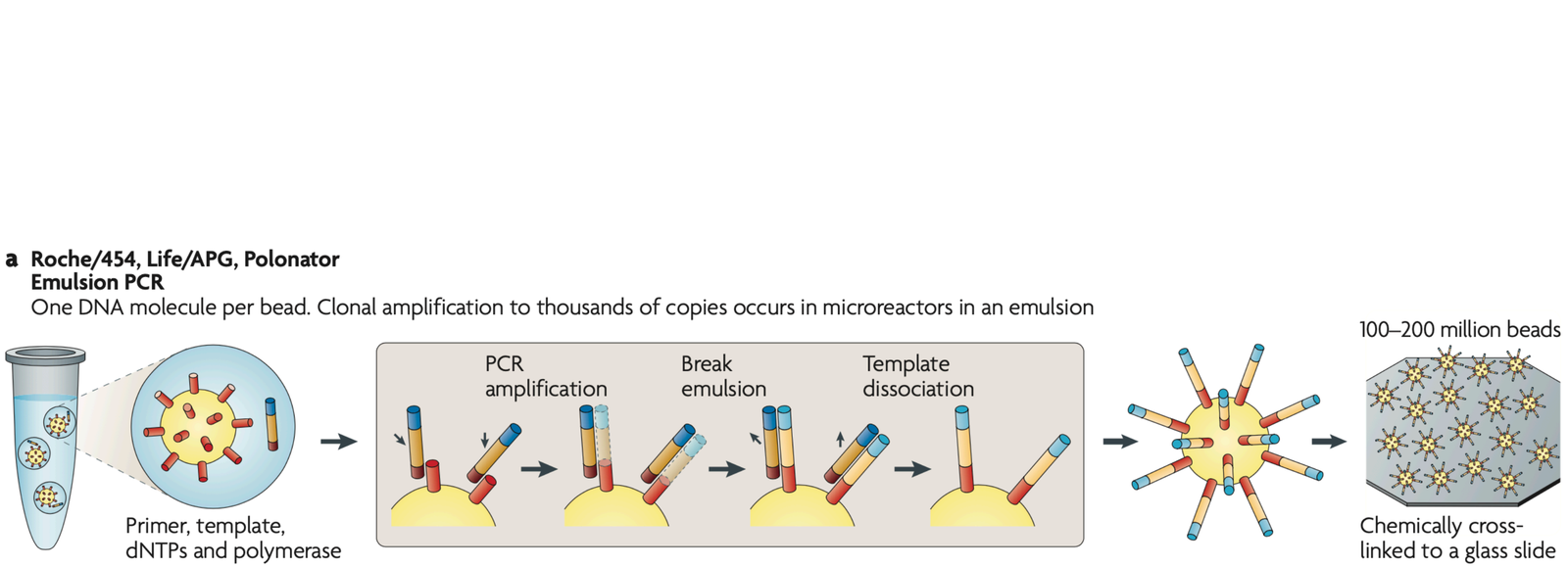
ian.simpson@ed.ac.uk

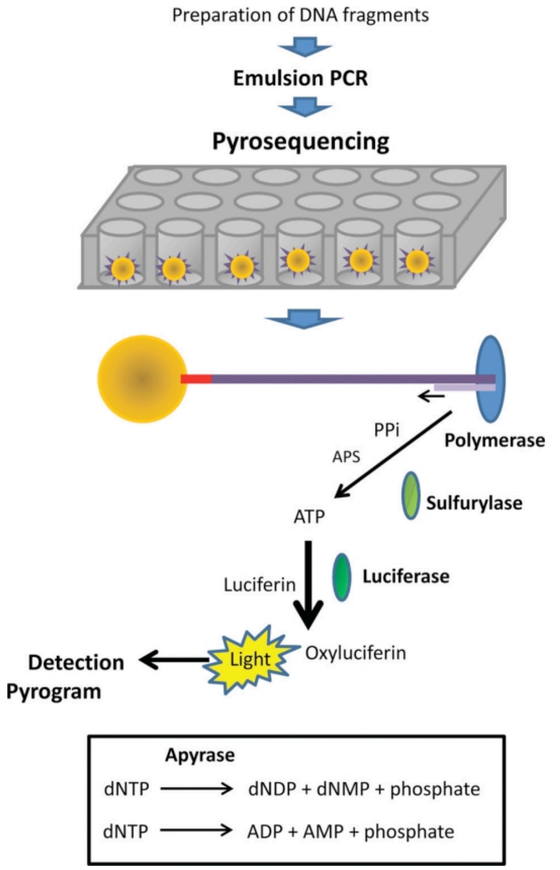
Next Generation Sequencing



ian.simpson@ed.ac.uk

454 Sequencing (Emulsion PCR - Pyrosequencing)



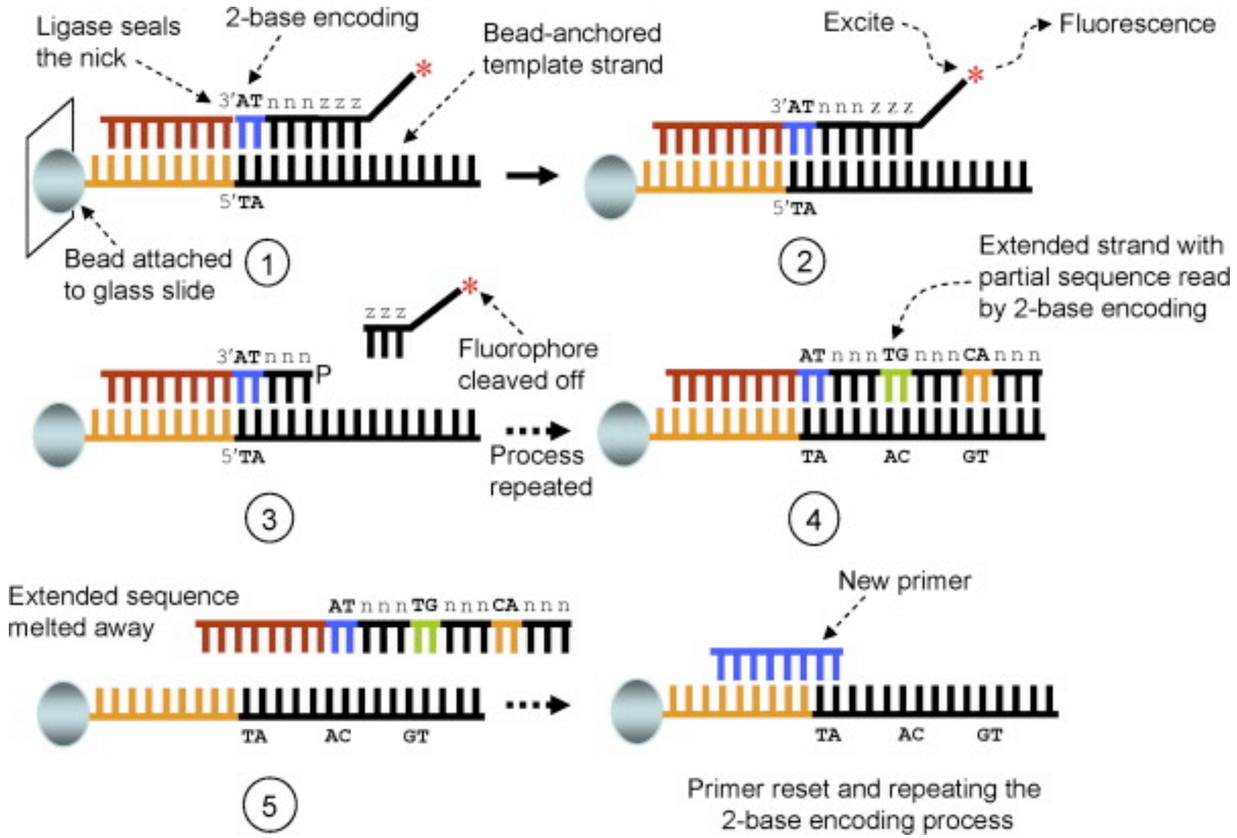


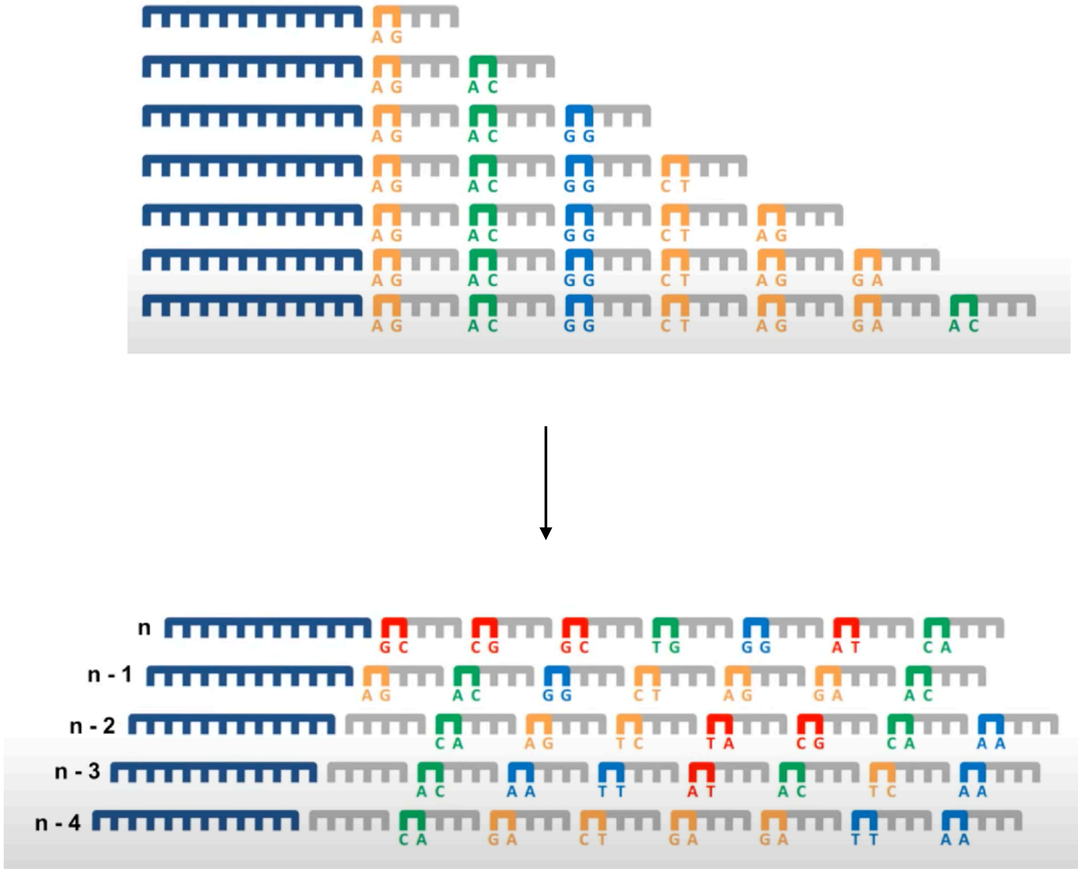
Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

https://www.youtube.com/watch?v=NJB-9HGSTfM

ian.simpson@ed.ac.uk

Solid Sequencing (Emulsion PCR – Sequencing by ligation)



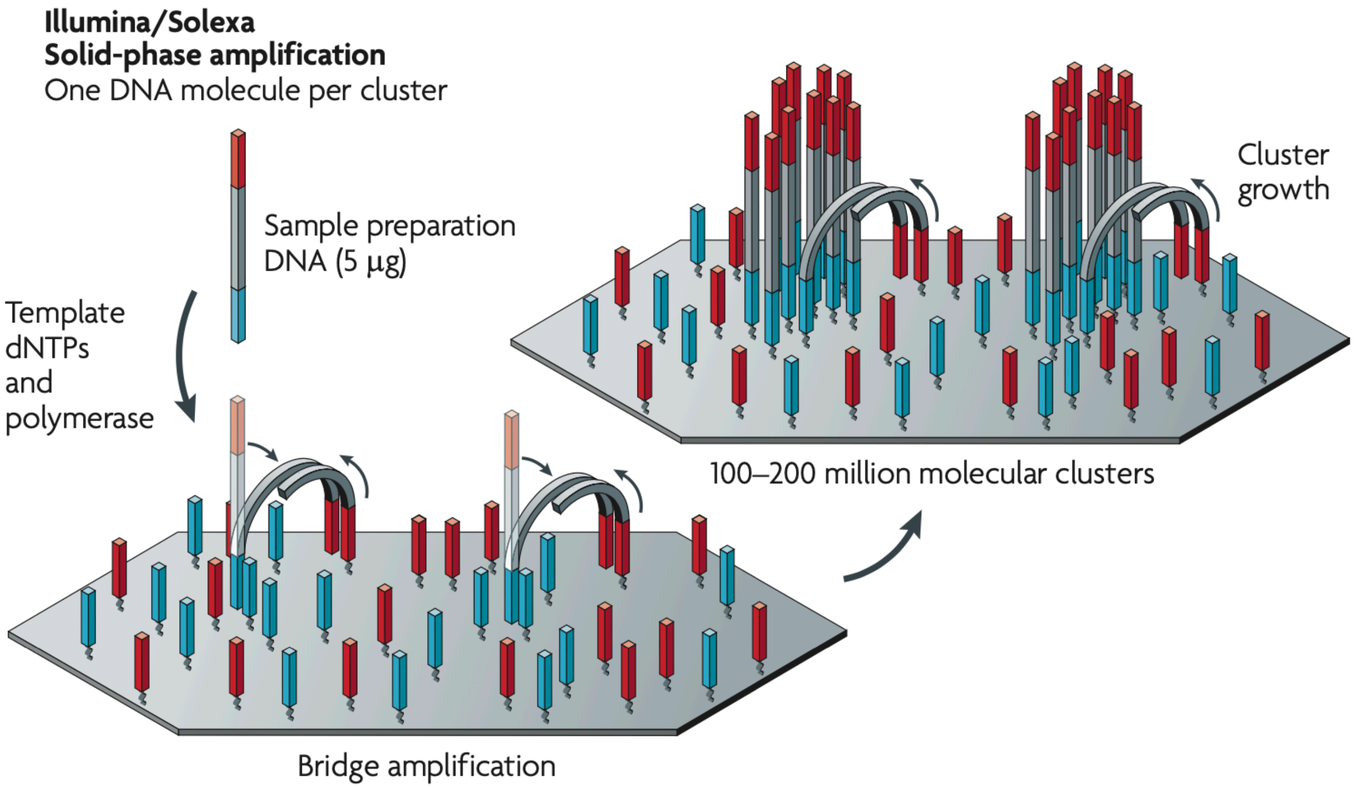


Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

overview - https://www.youtube.com/watch?v=nIvyF8bFDwM
detailed - https://www.youtube.com/watch?v=YLT-DUeaLms

ian.simpson@ed.ac.uk

Illumina Sequencing (Bridge PCR – Sequencing by synthesis)

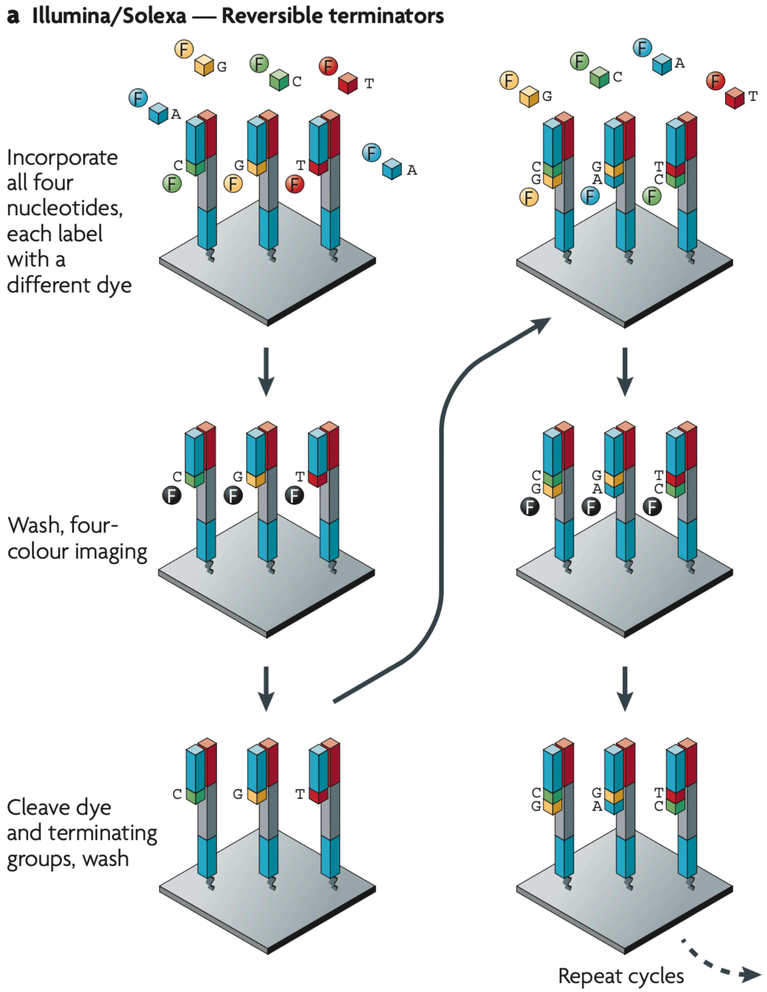


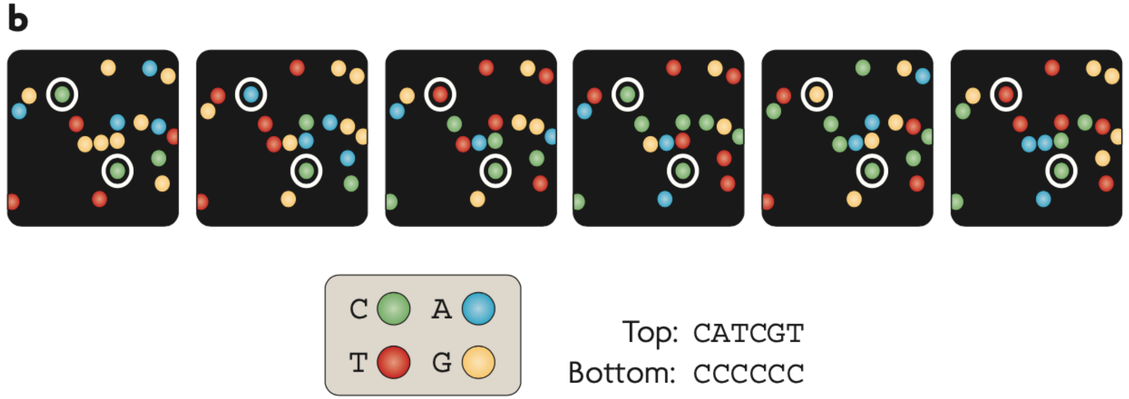
Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

https://www.youtube.com/watch?v=fCd6B5HRaZ8

ian.simpson@ed.ac.uk

Illumina Sequencing (Bridge PCR – Sequencing by synthesis)

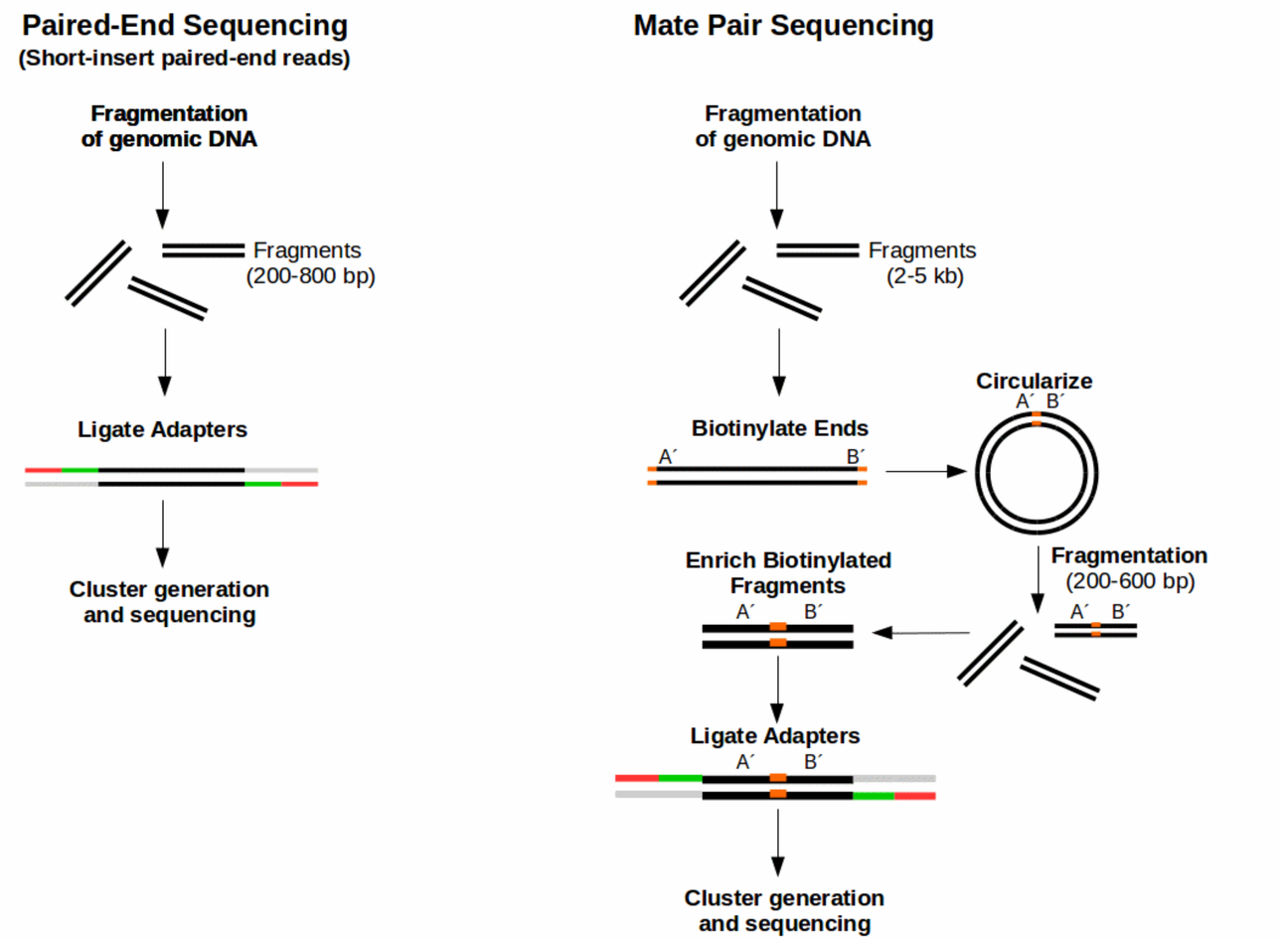




Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

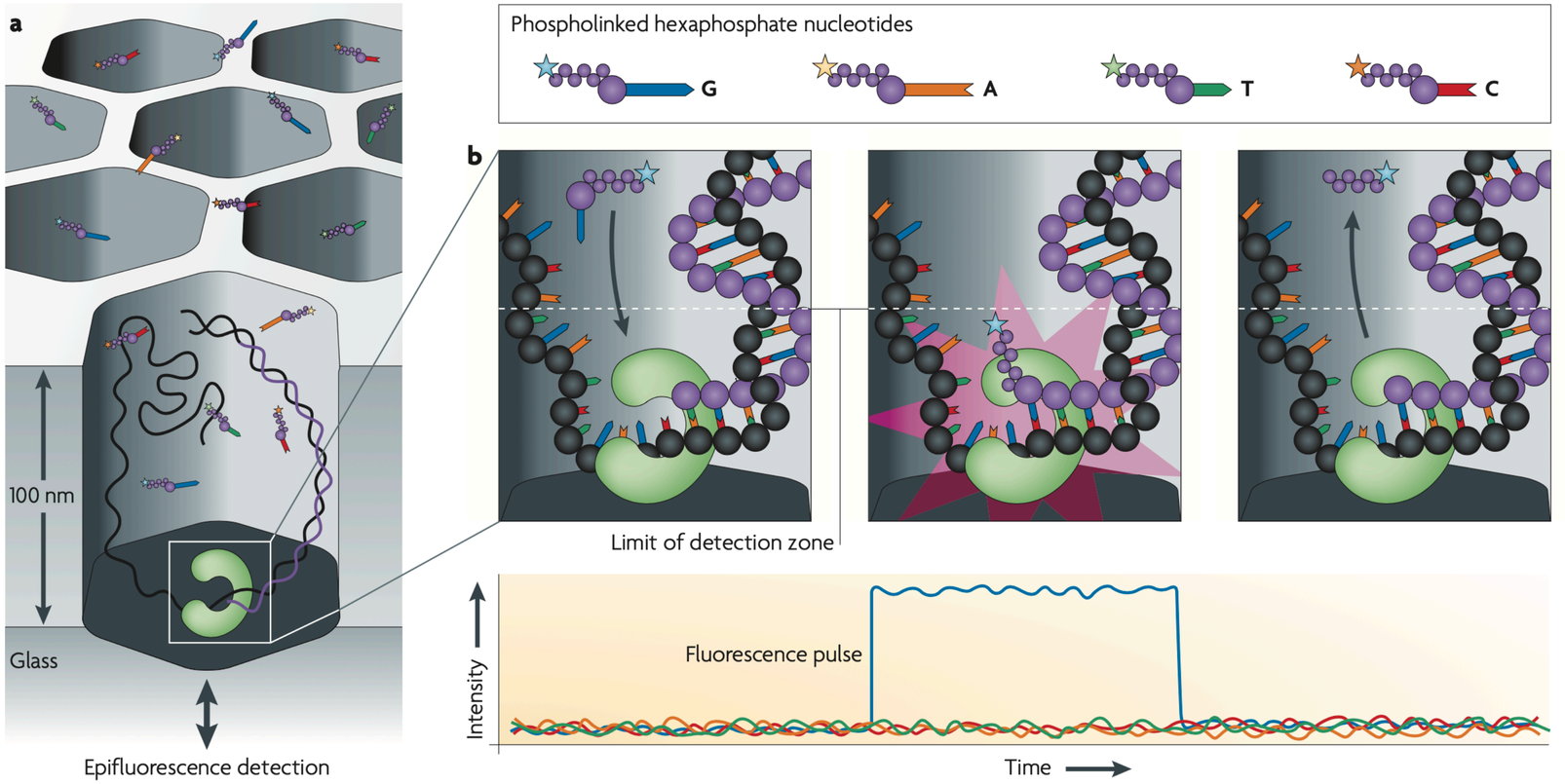
ian.simpson@ed.ac.uk

Paired-End & Mate-Paired Sequencing



ian.simpson@ed.ac.uk

Pacific Biosciences Sequencing - long read

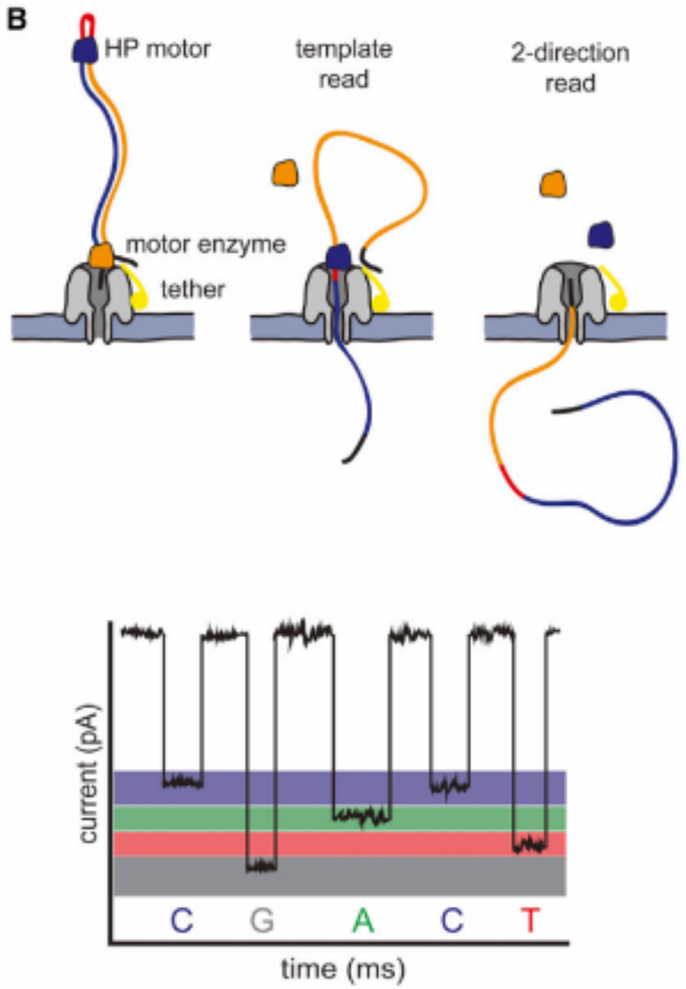


Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

https://www.youtube.com/watch?v=NHCJ8PtYCFc

ian.simpson@ed.ac.uk

Oxford Nanopore Sequencing (ONT) - long read



(MinION)

https://www.youtube.com/watch?v=E9-Rm5AoZGw

ian.simpson@ed.ac.uk

3rd Generation sequencing - Oxford Nanopore

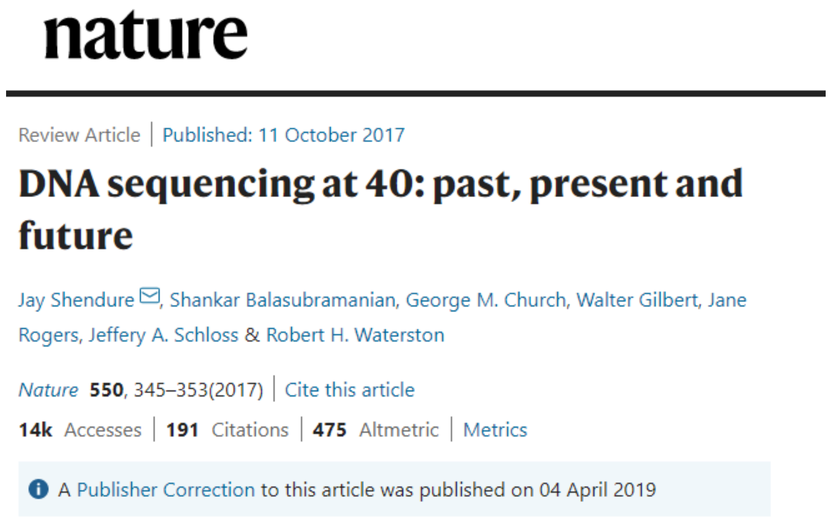
ian.simpson@ed.ac.uk

Summary of Next Generation Sequencing Platforms

Instrument	Amplification	Run time	Millions of Reads/run	Bases / read	Reagent Cost/run	Reagent Cost/Gb	Reagent Cost/Mread	bp/run	Gbp/run	cost/Gb
Applied Biosystems 3730 (capillary)	PCR, cloning	2 hrs.	0.000096	650	$144	$2,307,692.31	$1,500,000.00	62,400	0	$2,307,692.31
454 FLX+	emPCR	20 hrs.	1	650	$6,200	$9,538.46	$6,200.00	650,000,000	0.65	$9,538.46
Illumina GA IIx - v5 PE	bridgePCR	14 days	640	288	$17,978	$97.54	$28.09	184,320,000,000	184.32	$97.54
Illumina MiSeq v3	bridgePCR	55 hrs.	22	600	$1,442	$109.24	$65.55	13,200,000,000	13.2	$109.24
Illumina NextSeq 500	BridgePCR	30 hrs.	400	300	$4,000	$33.33	$10.00	120,000,000,000	120	$33.33
Illumina HiSeq 2500 - high output v4	BridgePCR	6 days	2000	250	$14,950	$29.90	$7.48	500,000,000,000	500	$29.90
Illumina HiSeq X (2 flow cells)	BridgePCR	3 days	6000	300	$12,750	$7.08	$2.13	1,800,000,000,000	1,800.00	$7.08
Ion Torrent – PGM 318 chip	emPCR	7.3 hrs.	4.75	400	$874	$460.00	$184.00	1,900,000,000	1.9	$460.00
Ion Torrent - Proton I	emPCR	4 hrs.	70	175	$1,000	$81.63	$14.29	12,250,000,000	12.25	$81.63
Ion Torrent - Proton III (forecast)	emPCR	6 hrs.	500	175	$1,000	$11.43	$2.00	87,500,000,000	87.5	$11.43
Life Technologies SOLiD – 5500xl	emPCR	8 days	1410	110	$10,503	$67.72	$7.45	155,100,000,000	155.1	$67.72
Pacific Biosciences RS II	None - SMS	2 hrs.	0.03	3000	$100	$1,111.11	$3,333.33	90,000,000	0.09	$1,111.11
Oxford Nanopore MinION (forecast)	None - SMS	≤6 hrs.	0.1	9000	$900	$1,000.00	$9,000.00	900,000,000	0.9	$1,000.00
Oxford Nanopore GridION 2000 (forecast)	None - SMS	varies	4	10000	$1,500	$37.50	$375.00	40,000,000,000	40	$37.50
Oxford Nanopore GridION 8000 (forecast)	None - SMS	varies	10	10000	$1,000	$10.00	$100.00	100,000,000,000	100	$10.00

ian.simpson@ed.ac.uk

References



nature

Review Article | Published: 11 October 2017

DNA sequencing at 40: past, present and future

Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss & Robert H. Waterston

Nature **550**, 345–353(2017) | Cite this article

14k Accesses | **191** Citations | **475** Altmetric | Metrics

A Publisher Correction to this article was published on 04 April 2019

https://www.nature.com/articles/nature24286

Trends in Genetics

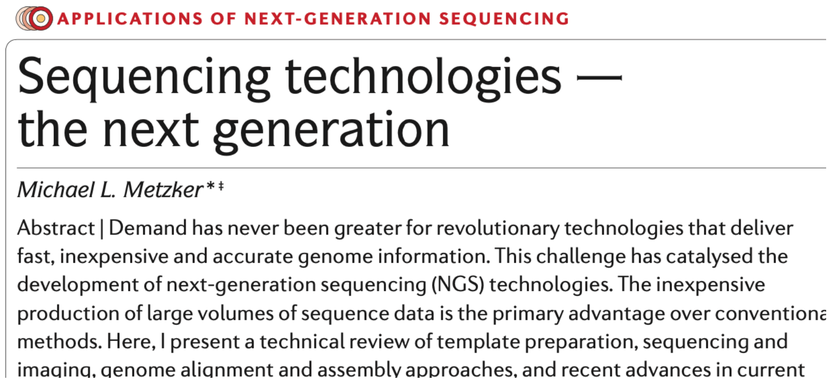
Review

The Third Revolution in Sequencing Technology

Erwin L. van Dijk,[1,*] Yan Jaszczyszyn,[1] Delphine Naquin,[1] and Claude Thermes[1]

Trends in Genetics, September 2018, Vol. 34, No. 9

https://doi.org/10.1016/j.tig.2018.05.008



APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

*Michael L. Metzker**‡

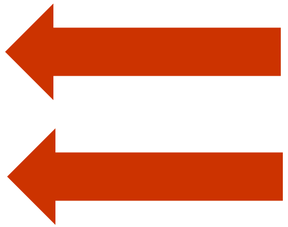
Abstract | Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventiona methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current

Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

The FASTQ Sequence Format

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;;9;7;;.7;393333
```



DNA read

Base quality score

ian.simpson@ed.ac.uk

FASTQ quality scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

$P=10^{-Q/10}$ $Q=-10\log_{10}(P)$

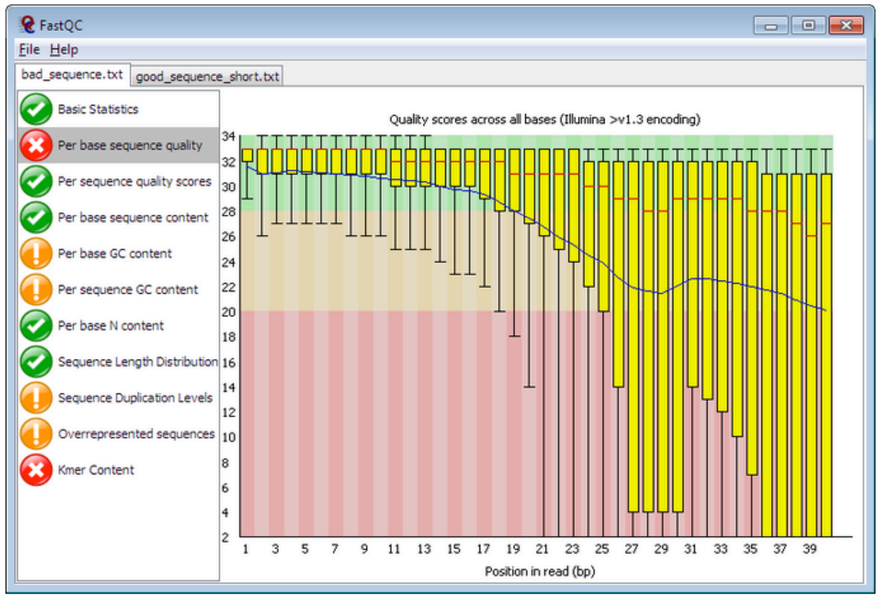
ian.simpson@ed.ac.uk

Quality Control - FastQC

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

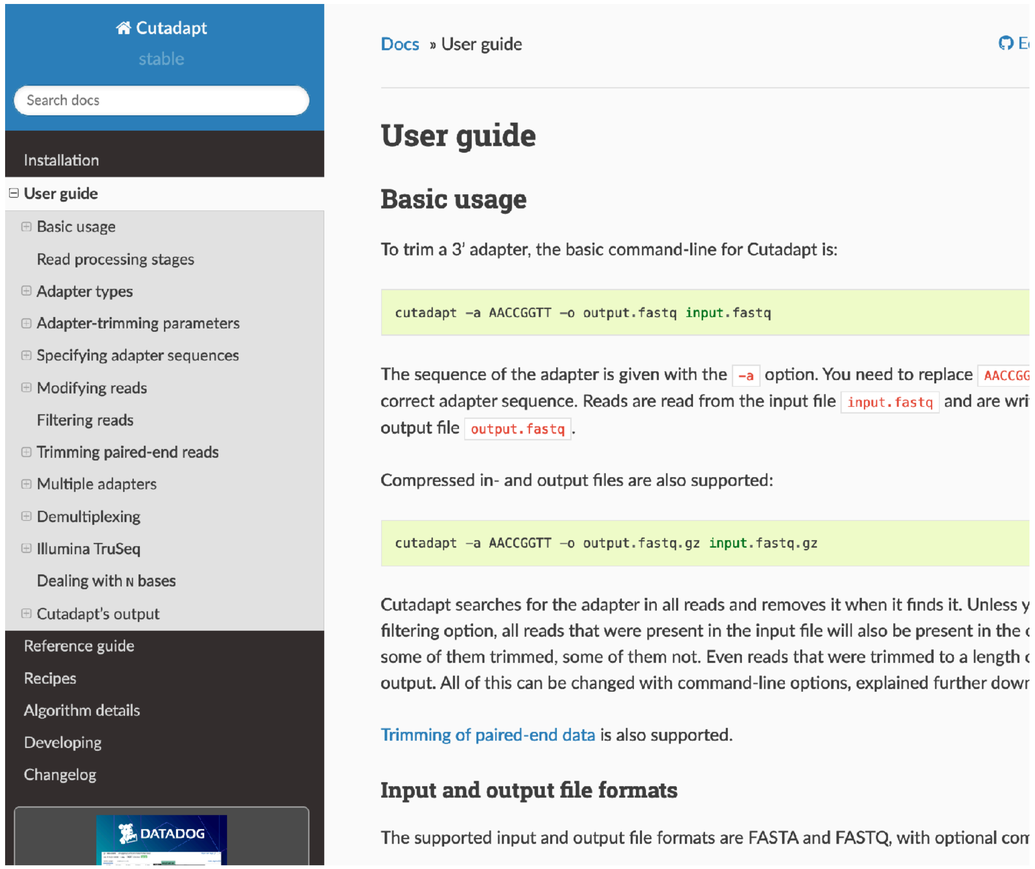
Download Now



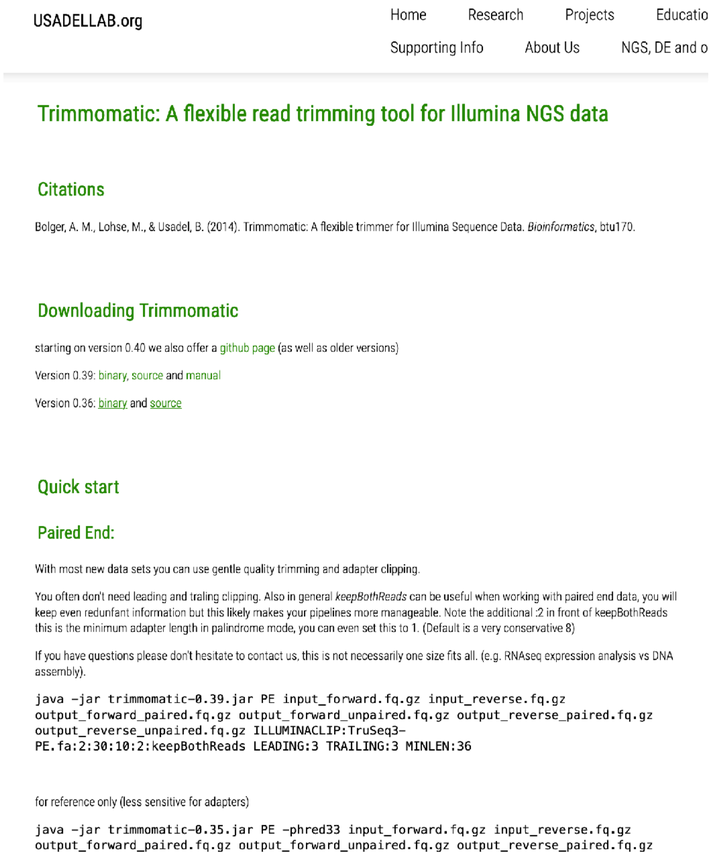
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

ian.simpson@ed.ac.uk

Adapter Trimming



https://cutadapt.readthedocs.io/en/stable/guide.html



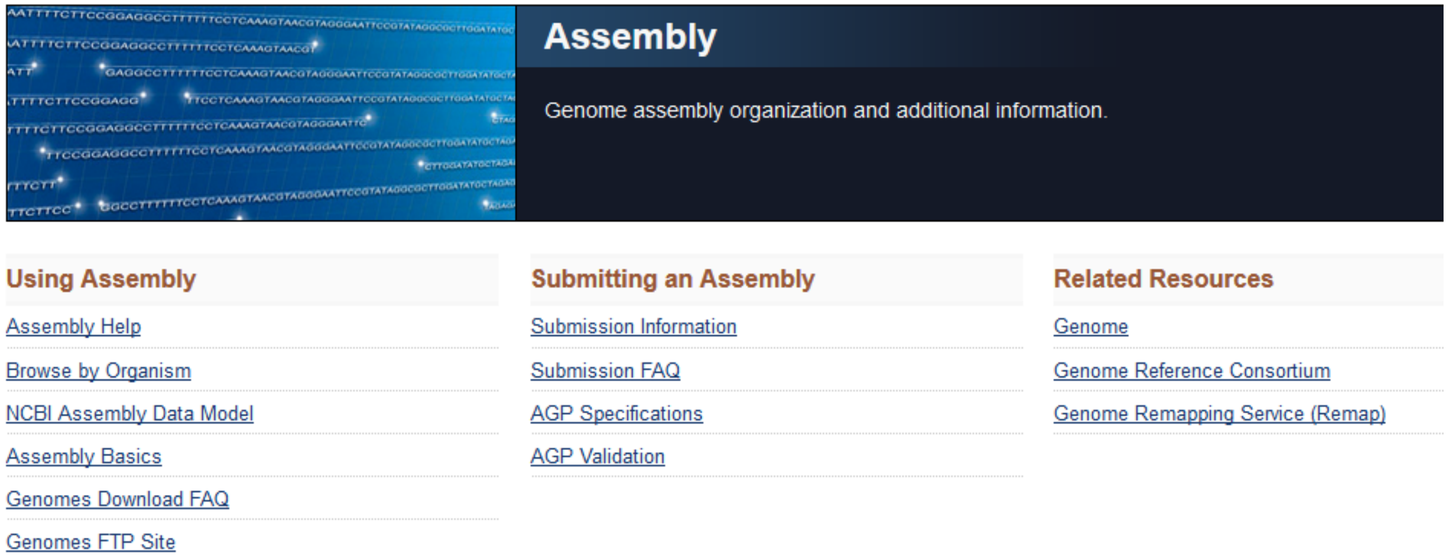
http://www.usadellab.org/cms/?page=trimmomatic

ian.simpson@ed.ac.uk

Genome Assembly

Genome assembly is the process of converting short reads into a detailed set of sequences corresponding to the chromosome(s) of an organism.

- http://www.ncbi.nlm.nih.gov/assembly/
- http://www.ncbi.nlm.nih.gov/assembly/basics/



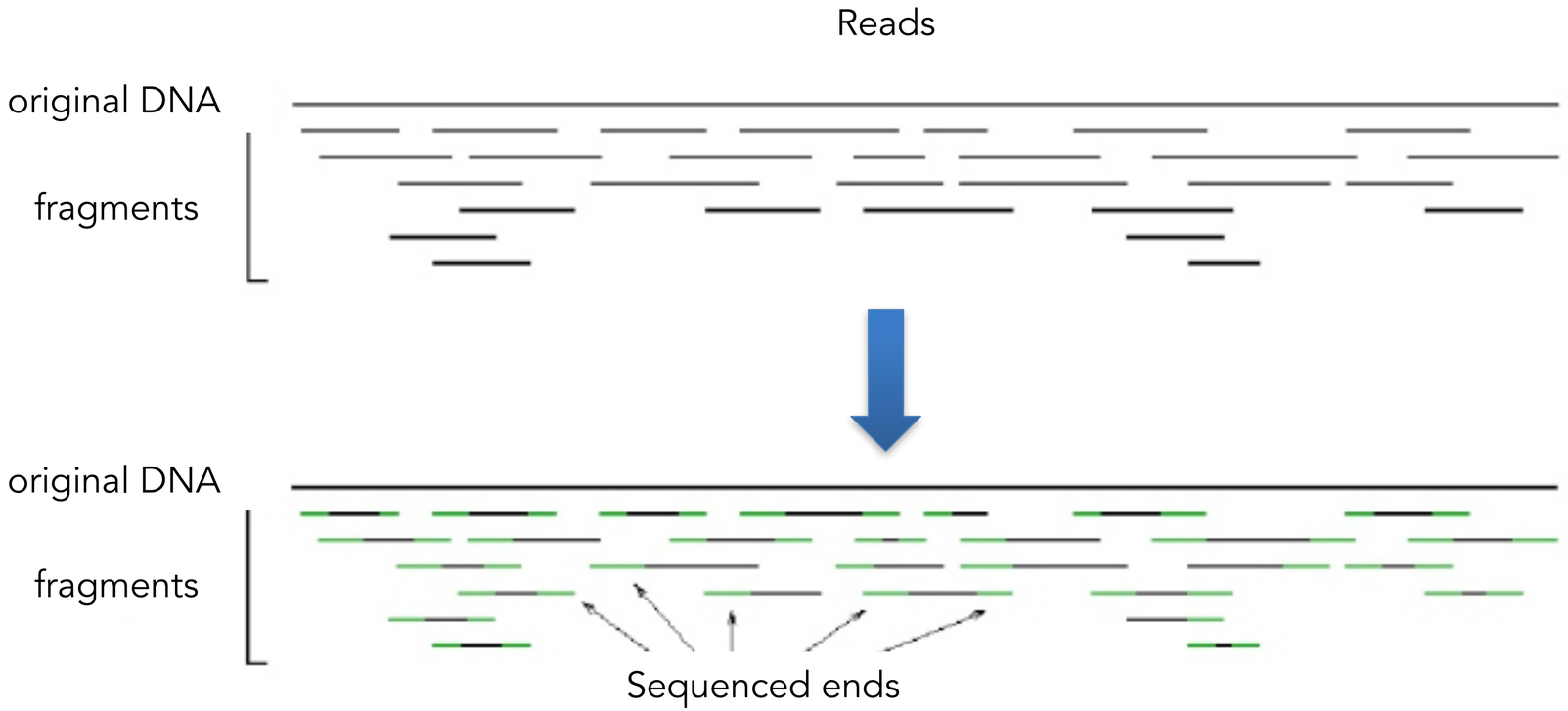
ian.simpson@ed.ac.uk

Genome Assembly

- Genome assembly is needed when a genome is first sequenced. We can relate reads to chromosomes.
- For the human genome, the assembly is "frozen" as a snapshot every few years. The current assembly is GRCh38. (GRC refers to Genome Reference Consortium at http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)
- For most human genome work we do not need to do "de novo" (from anew) assembly. Instead we map reads to a reference genome—one that is already assembled.
- Genome assembly is a crucial behind-the-scenes part of calling human genome (or other) variants.

ian.simpson@ed.ac.uk

Sequencing Strategies



http://www.cbcb.umd.edu/research/assembly_primer

ian.simpson@ed.ac.uk

Building Sequence Contigs

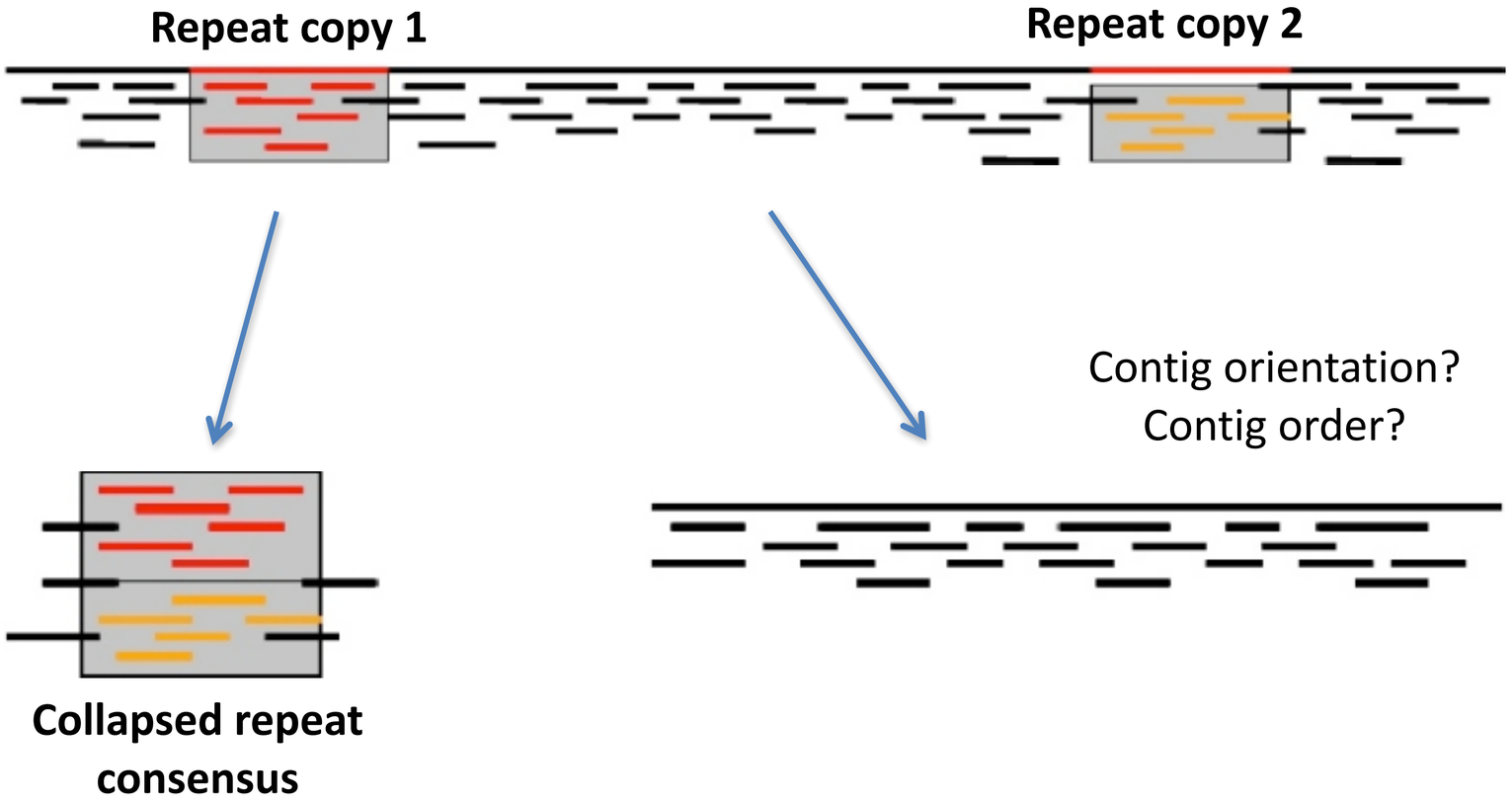
reads

contigs

scaffolds

ian.simpson@ed.ac.uk

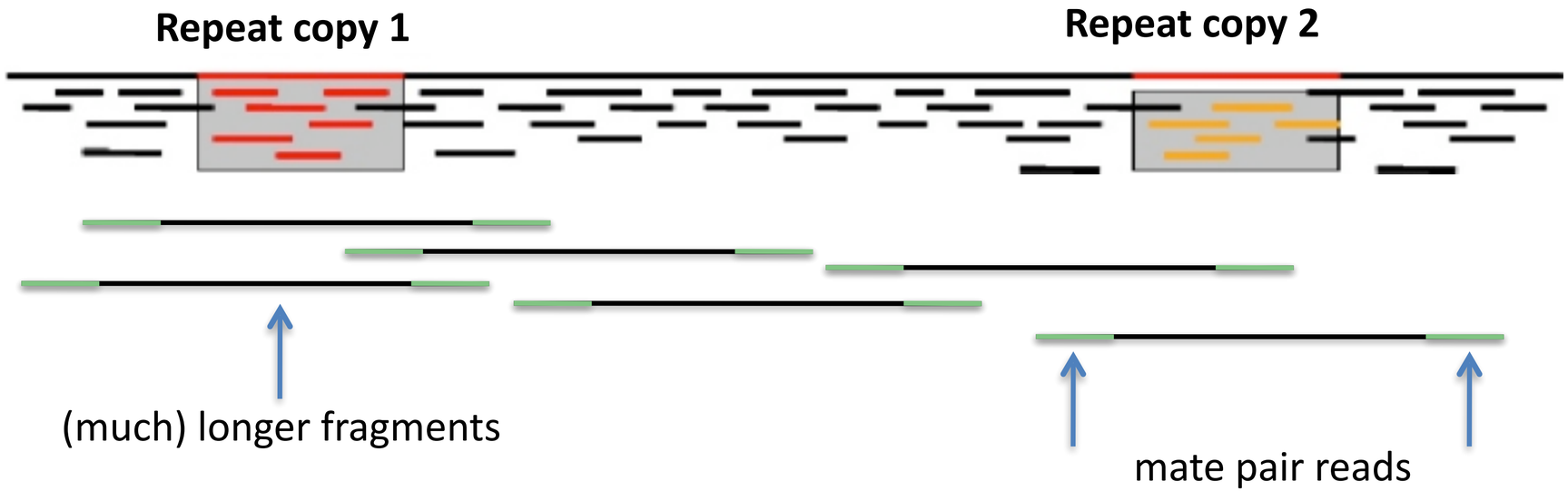
Building Sequence Contigs



reads
contigs
scaffolds

ian.simpson@ed.ac.uk

Long Reads Resolve Repeat Placement



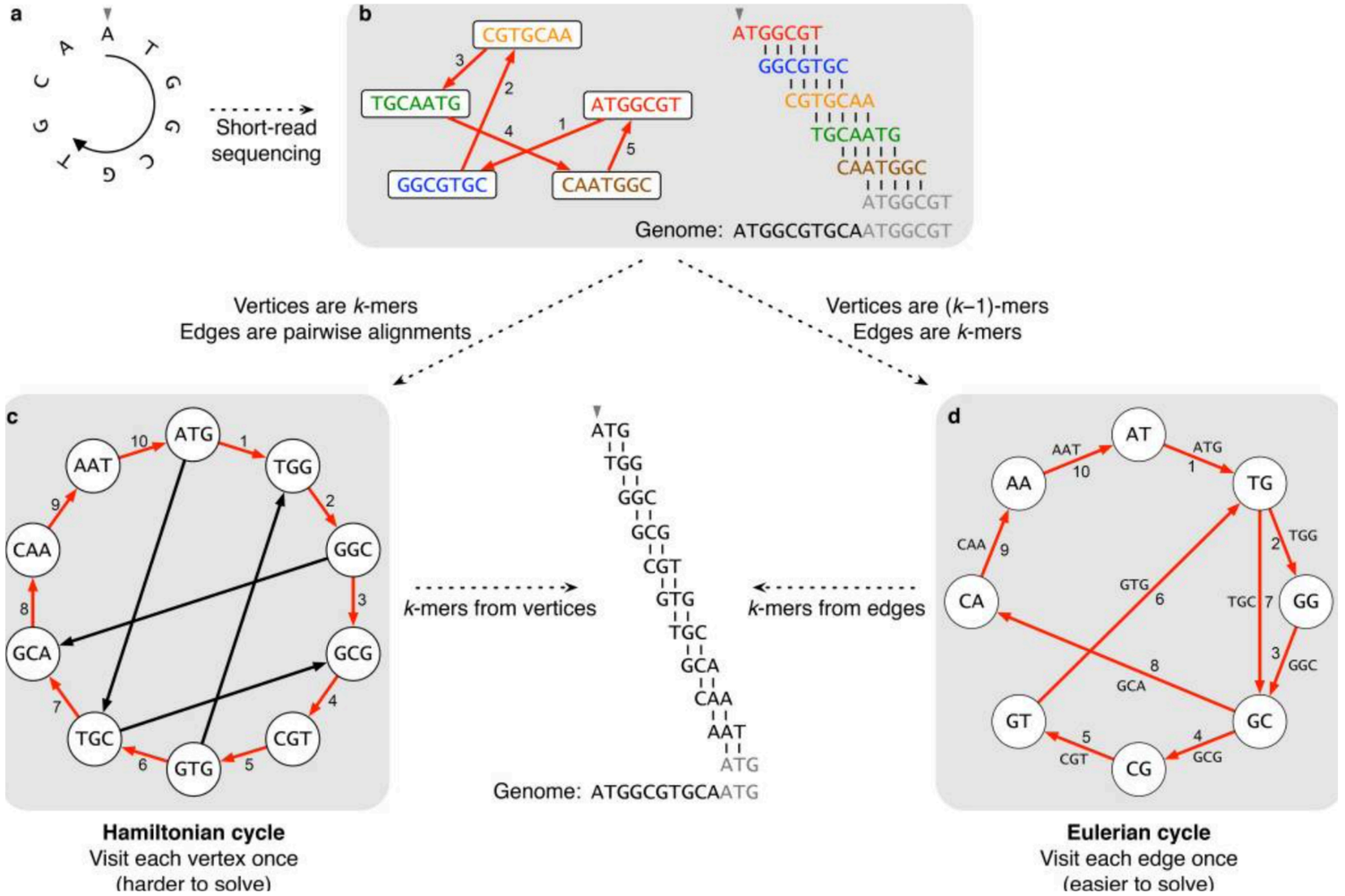
ian.simpson@ed.ac.uk

Software for Genome Assembly

Assembler	Reference	URL
ABySS	Simpson *et al.* (2009)	http://www.bcgsc.ca/platform/bioinfo/software
ALLPATHS-LG	Gnerre *et al.* (2011)	http://www.broadinstitute.org/software/allpaths-lg/blog/
Bambus2	Koren *et al.* (2011)	http://www.cbcb.umd.edu/software
CABOG	Miller *et al.* (2008)	http://www.jcvi.org/cms/research/projects/cabog/overview/
SGA	Simpson and Durbin (2012)	https://github.com/jts/sga
SOAPdenovo	Luo *et al.* (2012)	http://soap.genomics.org.cn/soapdenovo.html
Velvet	Zerbino and Birney (2008)	http://www.ebi.ac.uk/~zerbino/velvet/

ian.simpson@ed.ac.uk

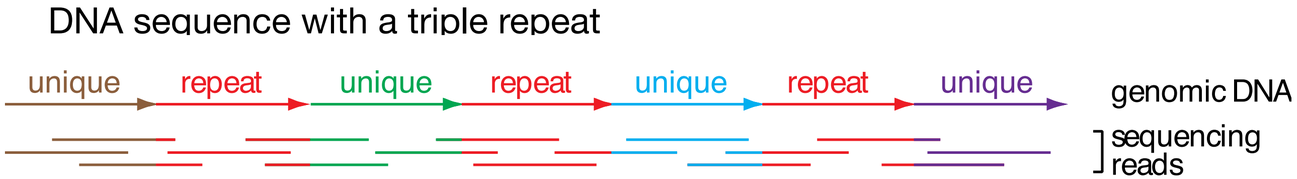
de Bruijn Graphs & Genome Assembly

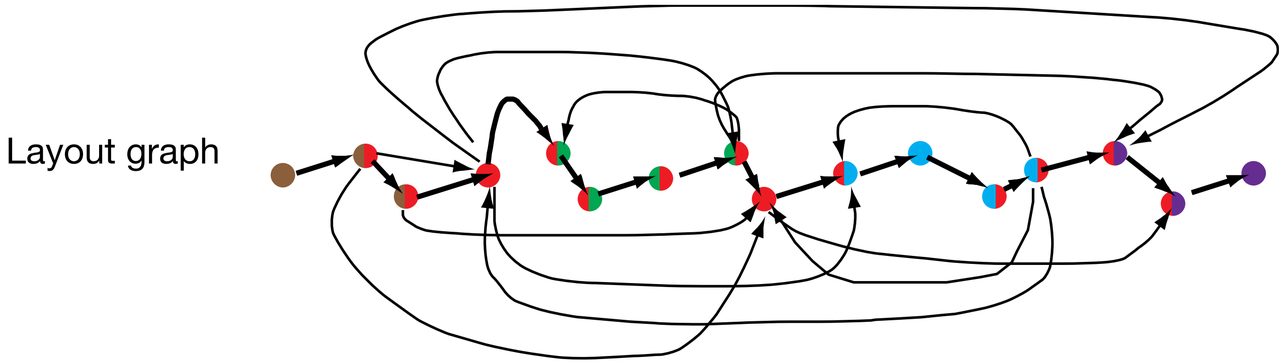


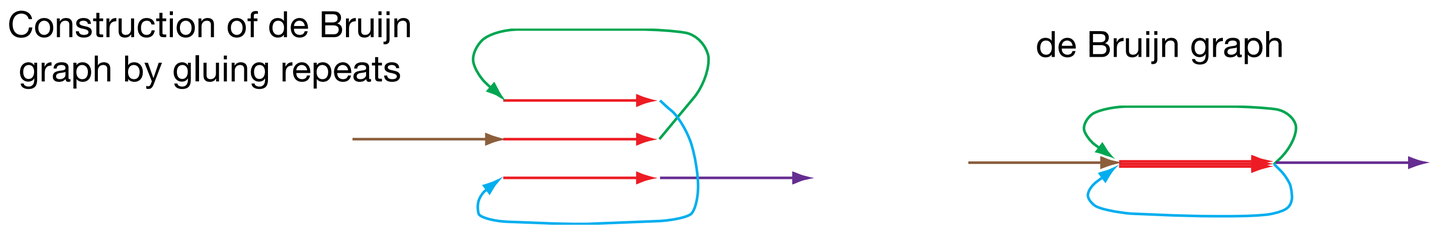
Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov 8;29(11):987-91. doi: 10.1038/nbt.2023. PMID: 22068540; PMCID: PMC5531759.

ian.simpson@ed.ac.uk

de Bruijn Graphs & Genome Assembly



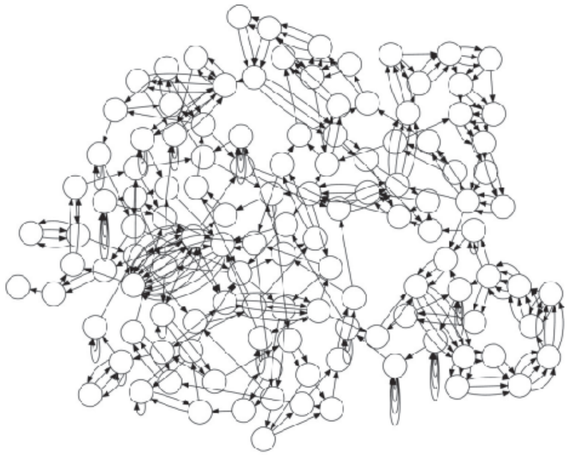




ian.simpson@ed.ac.uk

de Bruijn Graphs Resolve Assembly with Higher k-values

E. coli K12 (k=50)



E. coli K12 (k=1,000)

E. coli K12 (k=5,000)

ian.simpson@ed.ac.uk

Mapping Reads to Reference Genomes

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

From: Nat Biotechnol. Author manuscript; available in PMC 2010 May 1.
Published in final edited form as:
Nat Biotechnol. 2009 May; 27(5): 455–457.
doi: 10.1038/nbt0509-455.

- In order to USE the sequences we have generated we need to be able to align them to reference sequences such as genomes to allow us to interpret and analyse our data.
- Recent software tools allow the mapping (alignment) of millions or billions of short reads to a reference genome.
- For the human genome, this would take thousands of hours using BLAST.
- Reads may come from regions of repetitive DNA (exacerbated by sequencing errors)

ian.simpson@ed.ac.uk

Alignment Considerations

- Speed and Sensitivity

- For all software we measure error rates: using some gold standard we define true positive (TP) and true negative (TN) results, and we then define sensitivity and specificity.

- A standard format has been introduced for aligned sequences called Sequence Alignment/Map (**SAM**). Its binary version (which is compressed) is called **BAM**

- Aligned **BAM** files are available at repositories (Sequence Read Archive at NCBI, ENA at Ensembl)

- SAMTools is a software package commonly used to analyse SAM/BAM files (http://samtools.sourceforge.net/)

ian.simpson@ed.ac.uk

Anatomy of a SAM file

(1) The query name of the read is given (M01121...)

(2) The flag value is 163 (1+2+32+128)

(3) The reference sequence name, chrM, refers to the mitochondrial genome

(4) Position 480 is the left-most coordinate position of this read

(5) The Phred-scaled mapping quality is 60 (an error rate of 1 in 10^6)

(6) The CIGAR string (148M2S) shows 148 matches and 2 soft-clipped (unaligned) bases

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571      163     chrM
480     60      148M2S  =       524     195     AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG  BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFH00E@EGFGGEEE1FFEEEHBGEFFFGGGG@</0
1BG212222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-
RG:Z:Sample7    XC:i:148        XT:A:U  NM:i:3  SM:i:37
AM:i:37 X0:i:1  X1:i:0  XM:i:3  XO:i:0  XG:i:0  MD:Z:19C109C0A17
```

(7) An = sign shows that the mate reference matches the reference name

(8) The 1-based left position is 524

(9) The insert size is 195 bases

(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

(11) Each base is assigned a quality score (from BBBBB ending FHC.-)

(12) This read has additional, optional fields that accompany the MiSeq analysis

Flag Value	Meaning	Flag Sum
1	read is paired	1
32	read2 was reverse complemented	33
64	read1	97
2048	Supplementary alignment	2145

NB a BAM file is a binary version of a SAM file

ian.simpson@ed.ac.uk

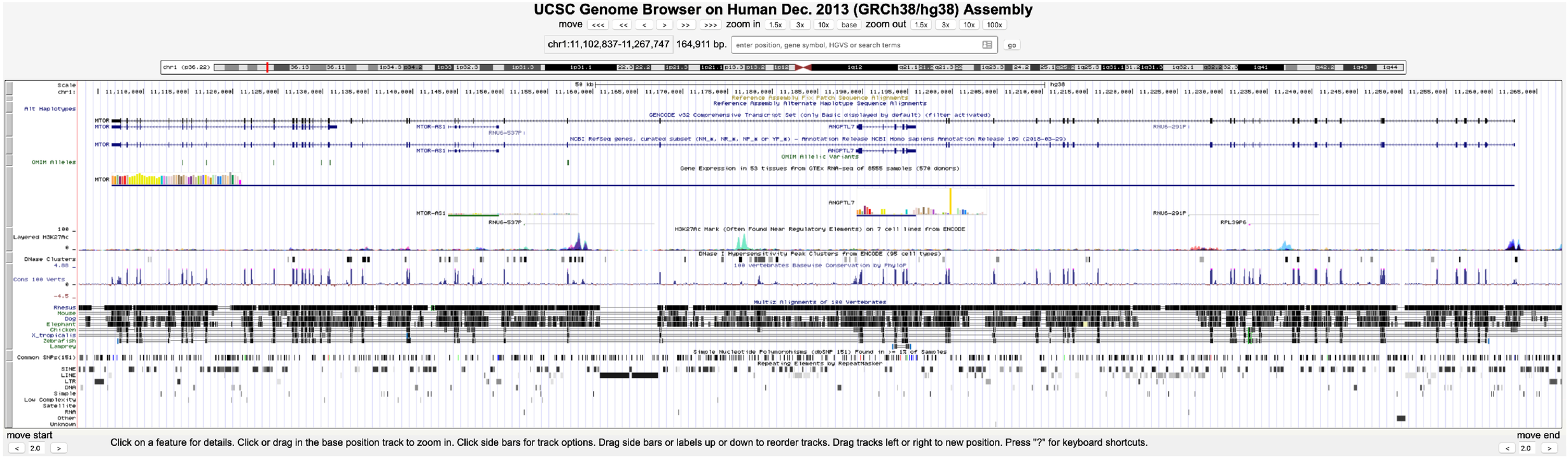
Multi-Mapping

ian.simpson@ed.ac.uk

Visualising NGS Data

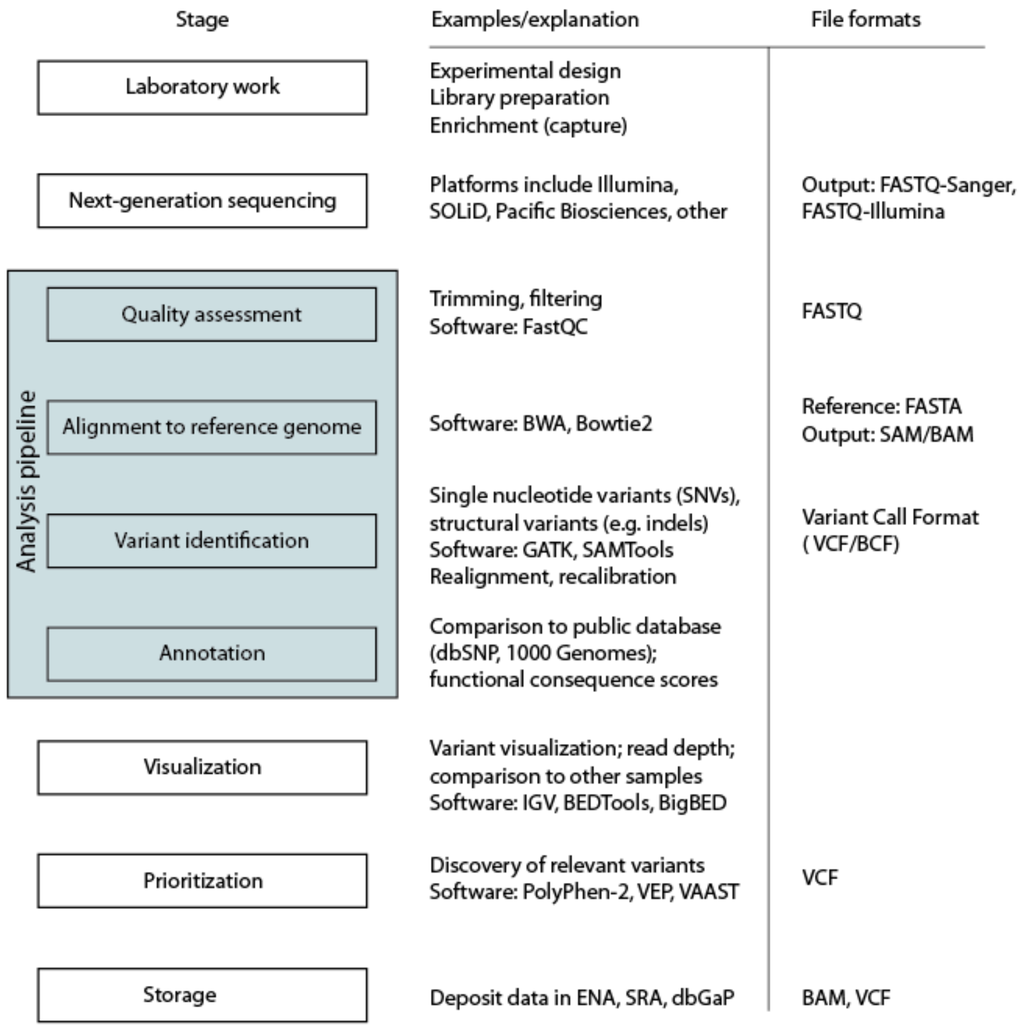
BAM files

- Genome Workbench from NCBI - https://www.ncbi.nlm.nih.gov/tools/gbench/
- Upload BAM file to a server and point to it using the UCSC Genome Browser - https://genome.ucsc.edu
- Use Integrative Genomics Viewer (IGV) - http://software.broadinstitute.org/software/igv/
- Use samtools tview



ian.simpson@ed.ac.uk

A Next-Generation Sequencing Workflow



ian.simpson@ed.ac.uk

Human Genome Sequencing

We currently obtain whole genome sequences at 30x to 50x depth of coverage.

For a typical individual:

- 2.8 billion base pairs are sequenced
- ~3-4 million single nucleotide variants
- ~600,000 insertions/deletions (indels)
- Cost (research basis) is <$2000 per genome

We also can enrich the collection of exons ("whole exome sequencing"). For a typical individual:

- 60 million base pairs are sequenced
- There are ~80,000 variants
- There are ~11,000 non-synonymous SNPs

ian.simpson@ed.ac.uk

Neutral Versus Deleterious Variation

- For each genome, we can expect to identify ~4 million variants that are exonic, intronic, or intergenic. We first focus on exonic variants. Of these, there are ~11,000 synonymous SNPs (not changing the amino acid specified by the codon; likely to be benign) and ~11,000 non-synonymous SNPs.

- We also consider indels (some of which introduce stop codons), homozygous deletions, splice site mutations, or other changes that may disrupt gene function.

ian.simpson@ed.ac.uk

Distinguishing Neutral From Deleterious Variants

Most DNA is under neutral selection (not under positive or negative selection)

Some variants are deleterious. How can we classify 11,000 non-synonymous SNPs in a genome?

- Conservation: determine conservation of an amino acid across species
- Structure: determine (or predict) effect of a variant on protein structure
- True positives: train algorithms on a database of known disease-associated mutations (OMIM)
- True negatives: train algorithms of a set of variants in ‘apparently normal’ individuals (1000 Genomes)

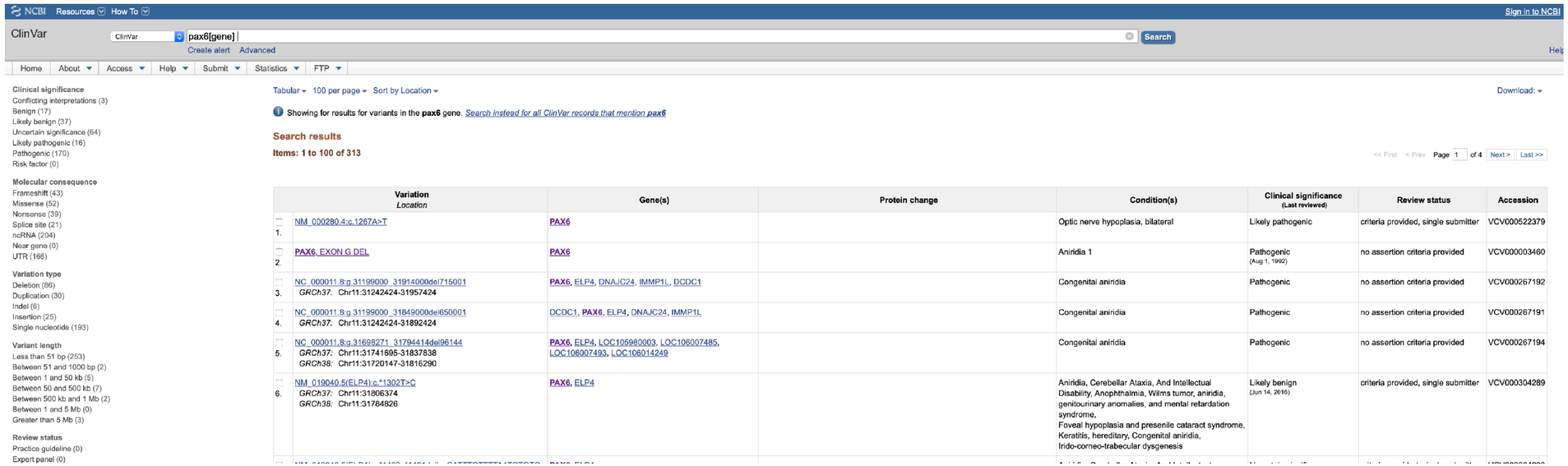
ian.simpson@ed.ac.uk

Software To Distinguish Neutral From Deleterious Variants

- PolyPhen2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. http://genetics.bwh.harvard.edu/pph2/
- SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. http://sift.jcvi.org/
- VAAST (Variant Annotation, Analysis & Search Tool) is a probabilistic search tool used to identify disease-causing variants
- VAAST calculates amino acid substitution frequencies for healthy genomes and disease genomes (both of these differ from standard BLOSUM62).

NCBI clinVar

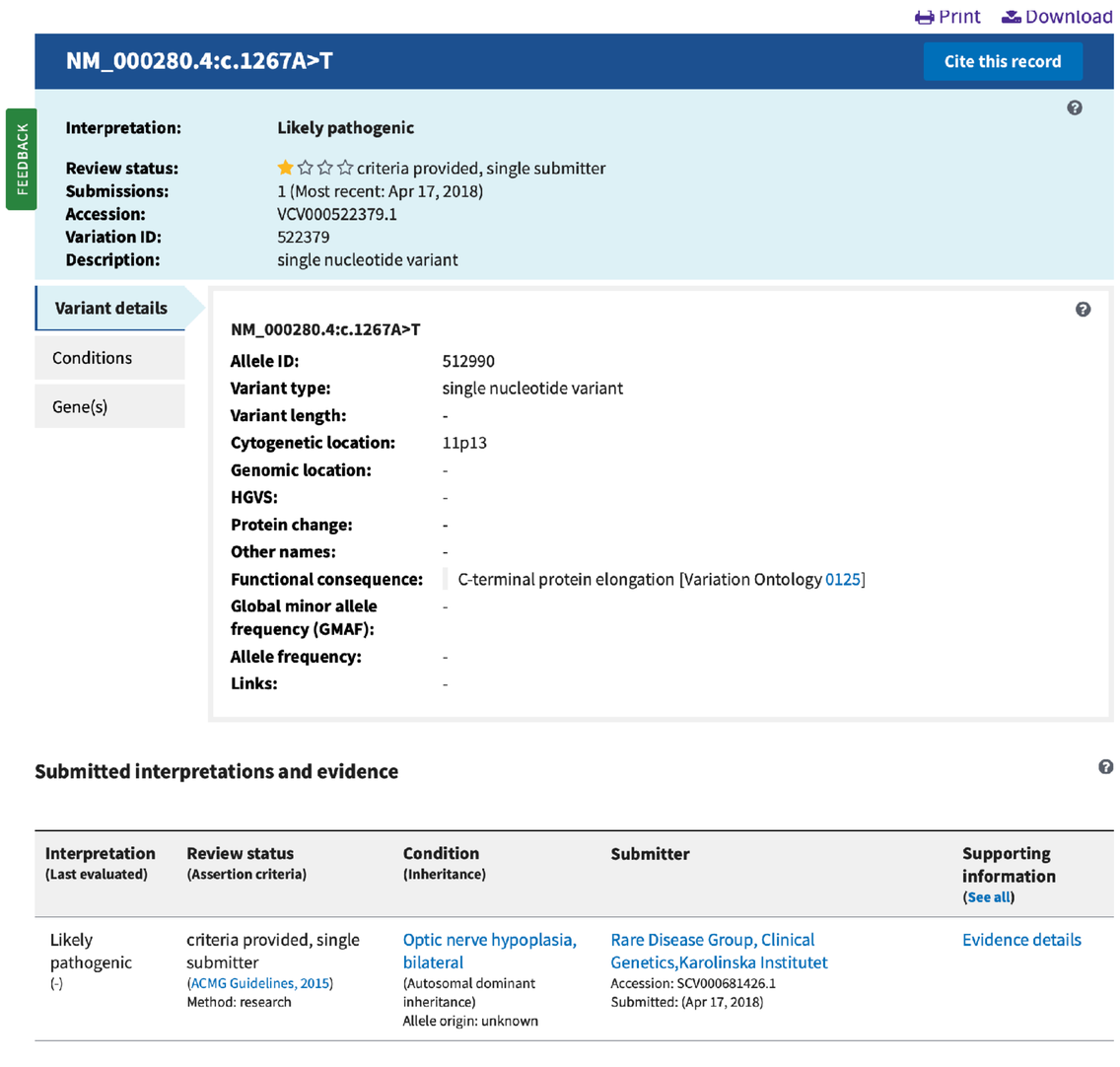
https://www.ncbi.nlm.nih.gov/clinvar/



- ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes with supporting evidence
- Facilitates access to and communication about the relationships asserted between human variation and observed health status
- Alleles described in submissions are mapped to reference sequences, and reported according to the HGVS standard (http://varnomen.hgvs.org)

den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016 Jun;37(6):564-9. doi: 10.1002/humu.22981.

clinVar Report



Print Download

NM_000280.4:c.1267A>T Cite this record

FEEDBACK

Interpretation:	**Likely pathogenic**
Review status:	★☆☆☆ criteria provided, single submitter
Submissions:	1 (Most recent: Apr 17, 2018)
Accession:	VCV000522379.1
Variation ID:	522379
Description:	single nucleotide variant

Variant details | Conditions | Gene(s)

NM_000280.4:c.1267A>T

Allele ID:	512990
Variant type:	single nucleotide variant
Variant length:	-
Cytogenetic location:	11p13
Genomic location:	-
HGVS:	-
Protein change:	-
Other names:	-
Functional consequence:	C-terminal protein elongation [Variation Ontology 0125]
Global minor allele frequency (GMAF):	-
Allele frequency:	-
Links:	-

Submitted interpretations and evidence

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Likely pathogenic (-)	criteria provided, single submitter (ACMG Guidelines, 2015) Method: research	Optic nerve hypoplasia, bilateral (Autosomal dominant inheritance) Allele origin: unknown	Rare Disease Group, Clinical Genetics,Karolinska Institutet Accession: SCV000681426.1 Submitted: (Apr 17, 2018)	Evidence details

ian.simpson@ed.ac.uk

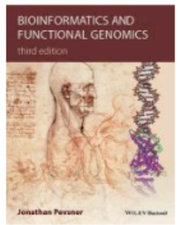
Summary

- Next-generation sequencing (NGS) technology is revolutionising biology. We are now able to catalog genetic variation at unprecedented depth
- There is rapid growth in the technologies used for NGS. There are also vast numbers of software solutions for quality control, sequence alignment, genome assembly, variant calling (including single nucleotide variants, indels, and structural variants), and variant prioritisation
- Key file formats include FASTQ ("raw" reads), BAM/SAM (aligned reads), and VCF (variant calls). Many tools are available for the generation, analysis, and visualisation of these types of files.

ian.simpson@ed.ac.uk

Further Reading

If you would like to read more about next generation sequencing then please look at the following chapter and references from the Bio1 course “Resource List”



BOOK **Bioinformatics and functional genomics**

Pevsner, Jonathan, 1961-, Third edition.., Hoboken, John Wiley & Sons, Incorporated, 2015

Note: *Please read Chapter 9 - Analysis of Next-Generation Sequencing Data.*

Add tags to item

Complete Check availability >

ian.simpson@ed.ac.uk