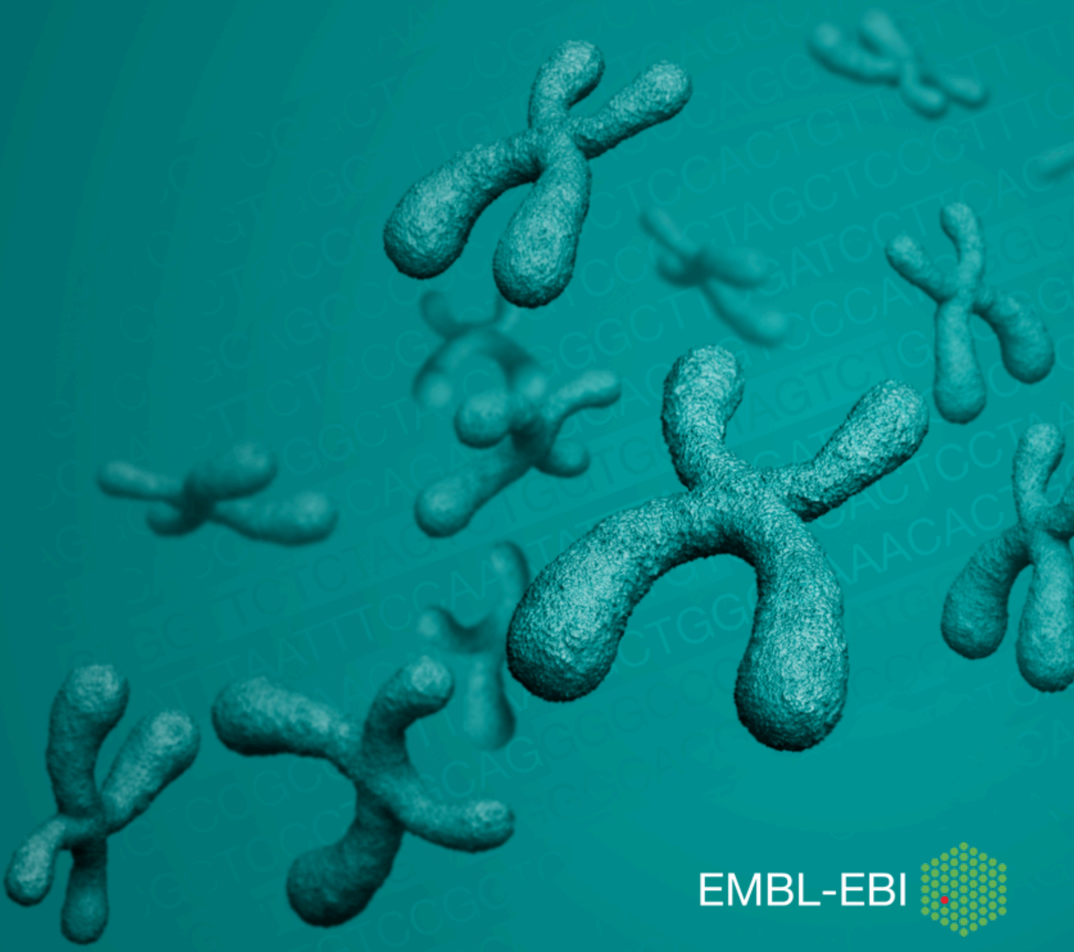
Challenges associated with analysis and storage of NGS data

Gabriella Rustici

Research and training coordinator

Functional Genomics Group

gabry@ebi.ac.uk

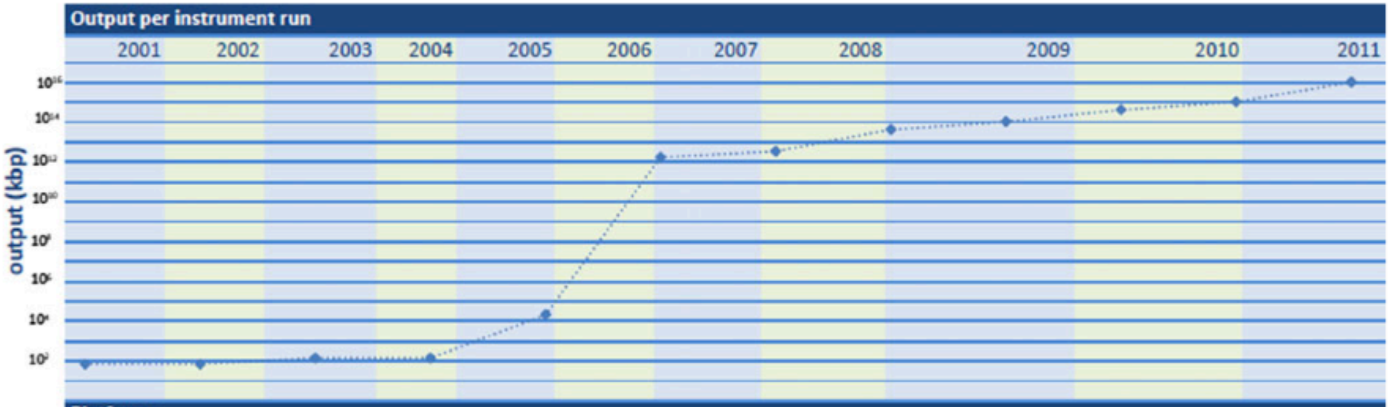


EMBL-EBI

Next-generation sequencing

- Next-generation sequencing (NGS) came to existence in the last decade
- NGS methods are highly parallelized processes that enable the sequencing of thousands to millions of molecules at once
- NGS has progressed beyond the analysis of DNA sequences
- Routinely used to analyze RNA, protein, as well as how they interact in complex networks
- The use of NGS in medical applications is a reality

NGS technology evolution



Platforms

	ABI 3730xl capillary sequencer		454 GS-20 pyrosequencer	Solexa/ Illumina sequence analyser	ABI SOLiD sequencer	Roche/454 Titanium, Illumina GAII	Illumina GAIIx, SOLiD 3.0	Illumina Hi-Seq 2000,	Danaher Motion Polonator model G.007, Heliscope™ single molecule sequencer, Single molecule real time (RNAP) sequencer	PacBio RS and SMRT TM sequencer, Oxford Nanopore GridION system, Sequencing by continuous synthesis in TOP chip's nanowell
2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Human genome draft	HapMap project begins	ENCODE project begins		Dog genome sequence publication	Chicken genome draft	1000 Genome, Human Microbiome project	Viral genome: Single-molecule DNA sequencing	1000 Genome pilot and HapMap3 publications	Human genetic syndromes publications	Single Molecule Sequencing of Cancer Transcriptomes
	Bos taurus genome draft	Swine genome draft				ENCODE project pilot publications	First tomour/normal publication	Real-Time DNA Sequencing from Single Polymerase Molecule	Decoding the genome of Haitian Cholera Pathogen	
		Finished Human genome sequence				Horse genome draft	Watson genome publication	Bos taurus genome assembly	Direct RNA Sequencing publication	
								Panda genome sequencing and assembling	Molecular motor for DNA strand nanopore sequencing	
								Complete genome sequencing of an individual by massively parallel DNA sequencing	DNA methylation during single-molecule, real-time sequencing	

Pareek at al, 2011. J Appl Genet. 2011 Nov;52(4):413-35

NGS advances

- DNA/RNA sequencing is cheaper and more efficient
- Innovative new experimental approaches for a deeper understanding of the molecular mechanisms of genome organization and cellular function
- For example, the ENCODE project:
  - Pilot phase: analyzed 1% of the human genome in un-precedent depth
  - With the introduction of NGS, expanded to the analysis of the entire genome (~ 1650 HT experiments)

Whole genome sequencing

- A recent estimate, counted 3920 bacterial and 854 eukaryotic genomes completely sequenced
- Challenges:
  - Different DNA sequencing platforms have different biases and abilities to call variants
  - Short indels (insertions and deletions) and larger structural variants are also difficult to call
  - *De novo* genome assembly can be attempted from short reads, but this remains difficult
- Increasing read length and accuracy will enhance the sequencing of genomes *de novo* and enable a more precise mapping of variants between individuals

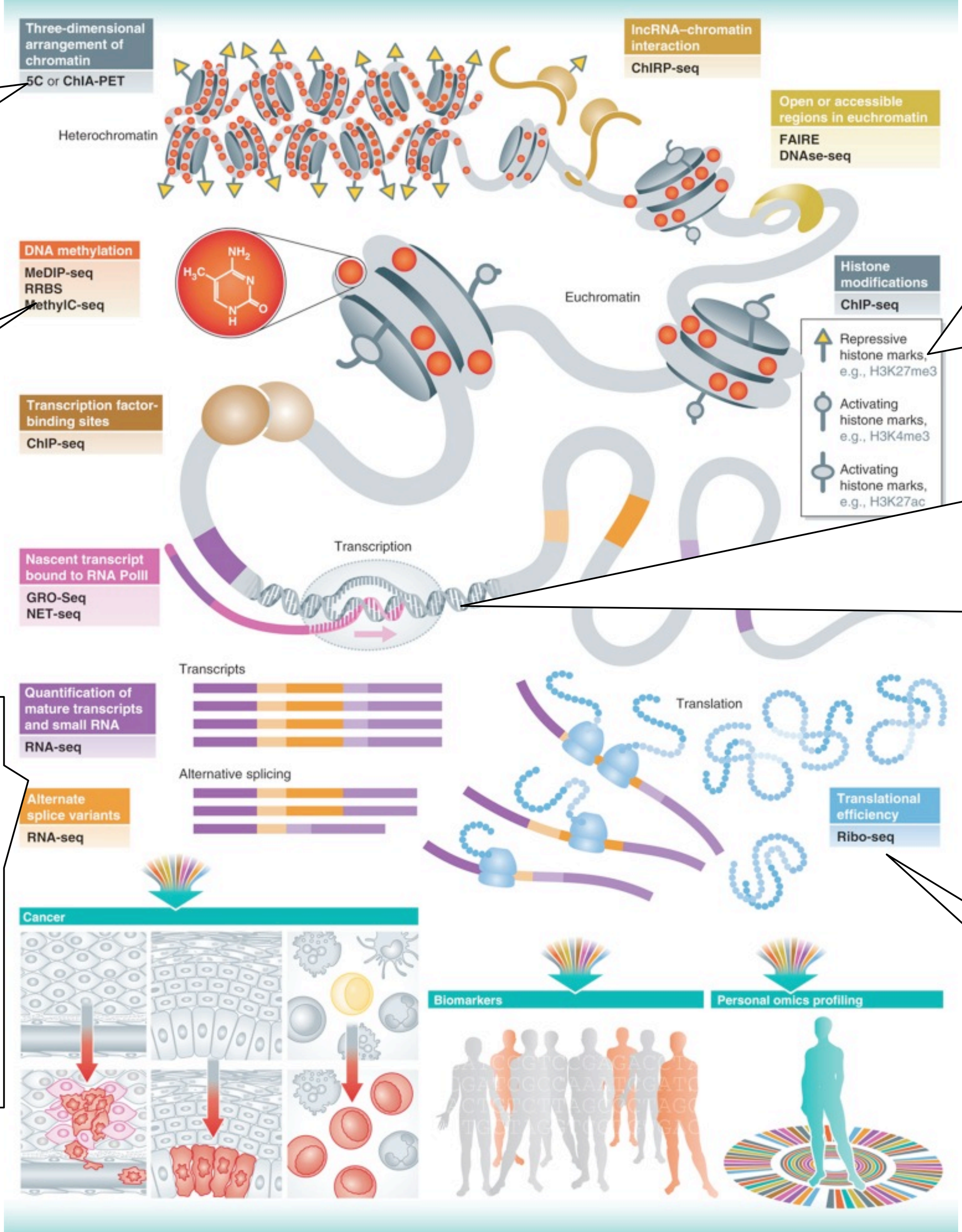
Medical genome sequencing

- Aims at identifying damaging polymorphisms in coding regions (exonic variants) and those present in functional regions
- Studies human genome variation by sequencing or genotyping large number of individuals
  - 1000 genome project (http://www.1000genomes.org/)
  - HapMap project (http://hapmap.ncbi.nlm.nih.gov/)
  - UK 10K project (http://www.uk10k.org/)
  - UK personal genome project and Genomics England (100K)
- So far 30M SNPs discovered from such projects
- Structural variations are much more prevalent than previously thought

Genome-wide mapping of chromosomal 3D structures

Mapping of epigenetic marks, known to silence part of the genome

RNA-seq: detection and quantification of transcripts, discovery of novel isoforms and linking expression to genomic variants



Three-dimensional arrangement of chromatin
5C or ChIA-PET
Heterochromatin
lncRNA–chromatin interaction
ChIRP-seq
Open or accessible regions in euchromatin
FAIRE
DNAse-seq
DNA methylation
MeDIP-seq
RRBS
MethylC-seq
NH2
H3C
N
O
H
Euchromatin
Histone modifications
ChIP-seq
Repressive histone marks, e.g., H3K27me3
Activating histone marks, e.g., H3K4me3
Activating histone marks, e.g., H3K27ac
Transcription factor-binding sites
ChIP-seq
Transcription
Nascent transcript bound to RNA PolII
GRO-Seq
NET-seq
Transcripts
Quantification of mature transcripts and small RNA
RNA-seq
Alternative splicing
Alternate splice variants
RNA-seq
Translation
Translational efficiency
Ribo-seq
Cancer
Biomarkers
Personal omics profiling

Mapping of histone marks, involved in gene regulation;
TF binding profiling

Measure production of nascent RNA, RNA-pol II bound transcripts, direction of transcription and rate of decay

Translation efficiency

Transcriptome analysis

- First were microarrays:
  - Limited to study known genes
  - Cross-hybridization issues
  - High noise level
  - Limited dynamic range (200 folds)
- Then came RNA-seq:
  - Little or no background noise
  - Large dynamic range (5000 folds)
  - Precise quantification of transcripts and exons
  - Analysis of transcript isoforms (still challenging due to transcriptome complexity in eukaryotes)
  - Allele specific expression
  - Identification of novel genes (fusion genes, etc…)

The real bottlenecks

- NGS, with its rapidly decreasing costs and increasing applications, is replacing many other technologies
- High resolution, low biases and detection power will make possible discoveries unachievable with previous technologies
- BUT…..significant challenges remain:
  - Data analysis: what biases do I have to take into consideration? What software tool is appropriate for my analysis needs? What analytical pipeline should I choose?
  - Storage: where and how are we going to store this data?

RNA-seq analysis core challenges

1. Experimental design
2. Mapping short RNA-seq reads
3. Identify expressed genes and isoforms
4. Estimate abundance of genes and isoforms
5. Analysis of differential expression

1. Experimental design

- Study design is very important – don’t try and do this post hoc!

- By randomizing samples appropriately across lanes / flow cells any biases that are introduced can be modeled

1. Experimental design – Read depth

- To obtain an in-depth view of every expressed transcript, it is necessary to sequence a sample to very high depth
- To obtain a more superficial summary of expression, far less depth may be necessary
- For normal RNA-seq analysis, I (John Marioni, EBI group leader) recommend around 10-20M reads per sample to collaborators

1. Experimental design – Number of samples

- Minimum of 3 per group to quantify variability accurately

- Statisticians always want more samples but this may not be possible in practice

- Again, it depends on the goal of the experiment – detecting smaller effects will require more samples

2. Mapping short RNA-seq reads

- Challenges:
    - Reads are short (~36-125 bases)
    - Large number of reads (hundreds of millions)
    - Many pieces don’t fit :
        - sequencing error/SNP/structural variant
    - Many pieces fit in many places:
        - low complexity region/microsatellite/repeat
    - Many reads span exon-exon junctions
- Mapping to either reference transcriptome or genome

2. Mapping short RNA-seq reads

- Many software tools are available
- “Unspliced read aligners” (i.e. MAQ, BWA, Bowtie)
  - Align reads to a reference without allowing any large gaps
  - Limited to identifying known exons and junctions and do not allow for the identification of spicing events involving new exons
- “Spliced aligners” (i.e. MapSplice, SpliceMap, TopHat, GSNAP)
  - Reads can be aligned to the entire genome, including intron-spanning reads that require large gaps for proper placement

Counting rules

- Count reads, not base-pairs
- Count each read at most once
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene

Do this using (e.g. HTSeq)

3. Identify expressed genes and isoforms

- Define a precise map of all transcripts and isoforms that are expressed in a particular sample
- Challenges:
  - Gene expression spans several orders of magnitude, with some genes represented by only a few reads
  - Reads originate from mature mRNA as well as the incompletely spliced precursor RNA
  - Reads are short, so which isoform produced each read?
- "genome-guided" (i.e. Cufflinks) vs. "genome independent" (i.e. transAbyss) methods
  - What is the biological question being asked?

3. Identify expressed genes and isoforms

- If a gene has a single transcript, this process is easy = sum the number of reads mapping to each of its constitutive exons
- If a gene has a multiple transcripts, the process is more difficult
  1. Reads spanning unique exon junctions or contained within unique exons are informative
  2. Various statistical techniques[1-4] to determine the expression of each isoform

[1] Trapnell et al., *Nature Biotechnology*., 2010

[2] Li, Ruotti et al., *Bioinformatics*, 2010

[3] Turro et al., *Genome Biology*, 2011

[4] Glaus et al., *Bioinformatics,* 2013

4. Estimating transcript expression levels

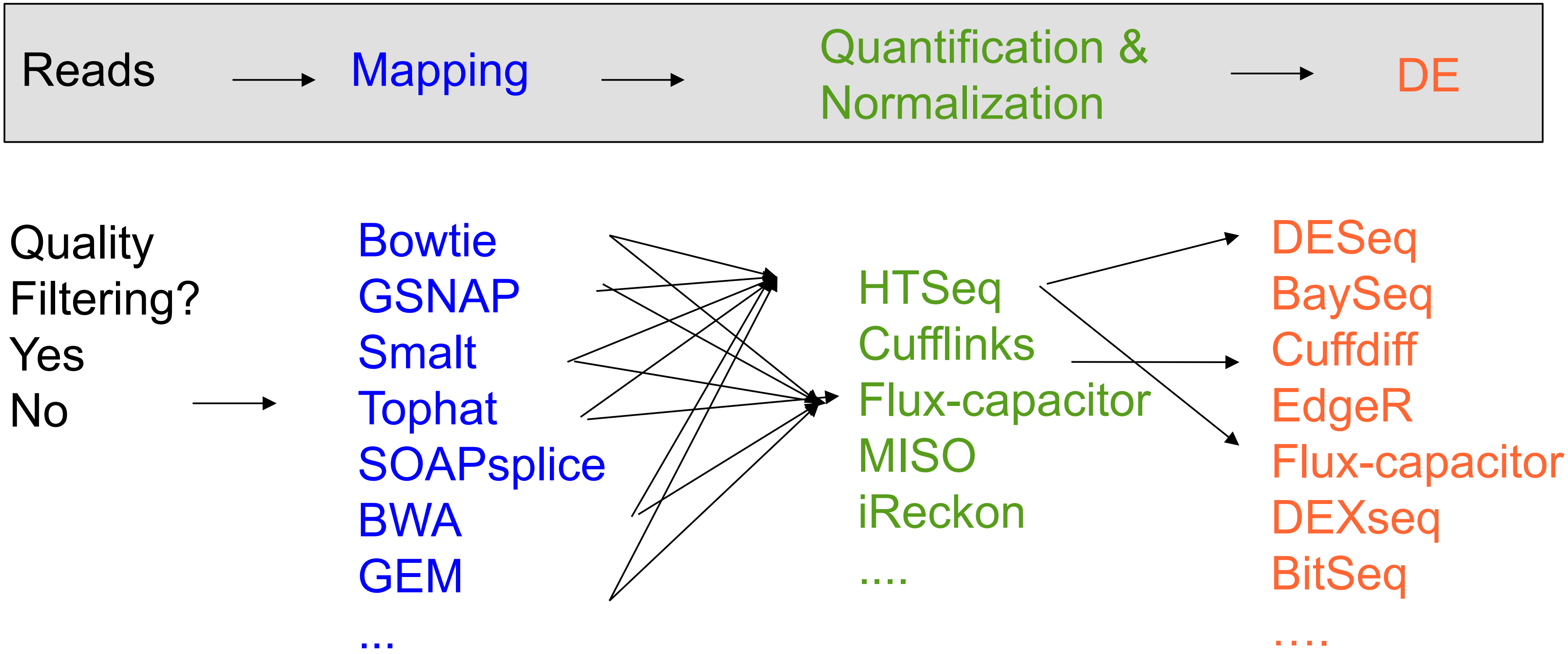
- Expression quantification requires proper normalization of read counts
- Challenges:
  - RNA fragmentation causes longer transcripts to generate more reads compared to shorter transcripts, present at the same abundance in the sample
  - The variability in the number of reads produced for each run causes fluctuations in the number of fragments mapped across samples
- The RPKM metric normalizes a transcript's read count by both its length and the total number of mapped reads in the sample

5. Analysis of differential expression

- How do expression levels differ across conditions?
- Challenges:
  - The power of detecting DE genes depends on sequencing depth of the sample, the expression of the gene and its length
  - Not enough replicates are available to model biological variability
  - Although variability is lower than in microarray data, measurements can vary due to different library preparation protocols and intrinsic variability in biological samples
- Bioconductor packages: edgeR, DEseq & DEXseq; Cuffdiff

RNA-Seq analysis

From reads to gene and differential expression (DE)



Reads → Mapping → Quantification & Normalization → DE

Quality Filtering?
Yes
No

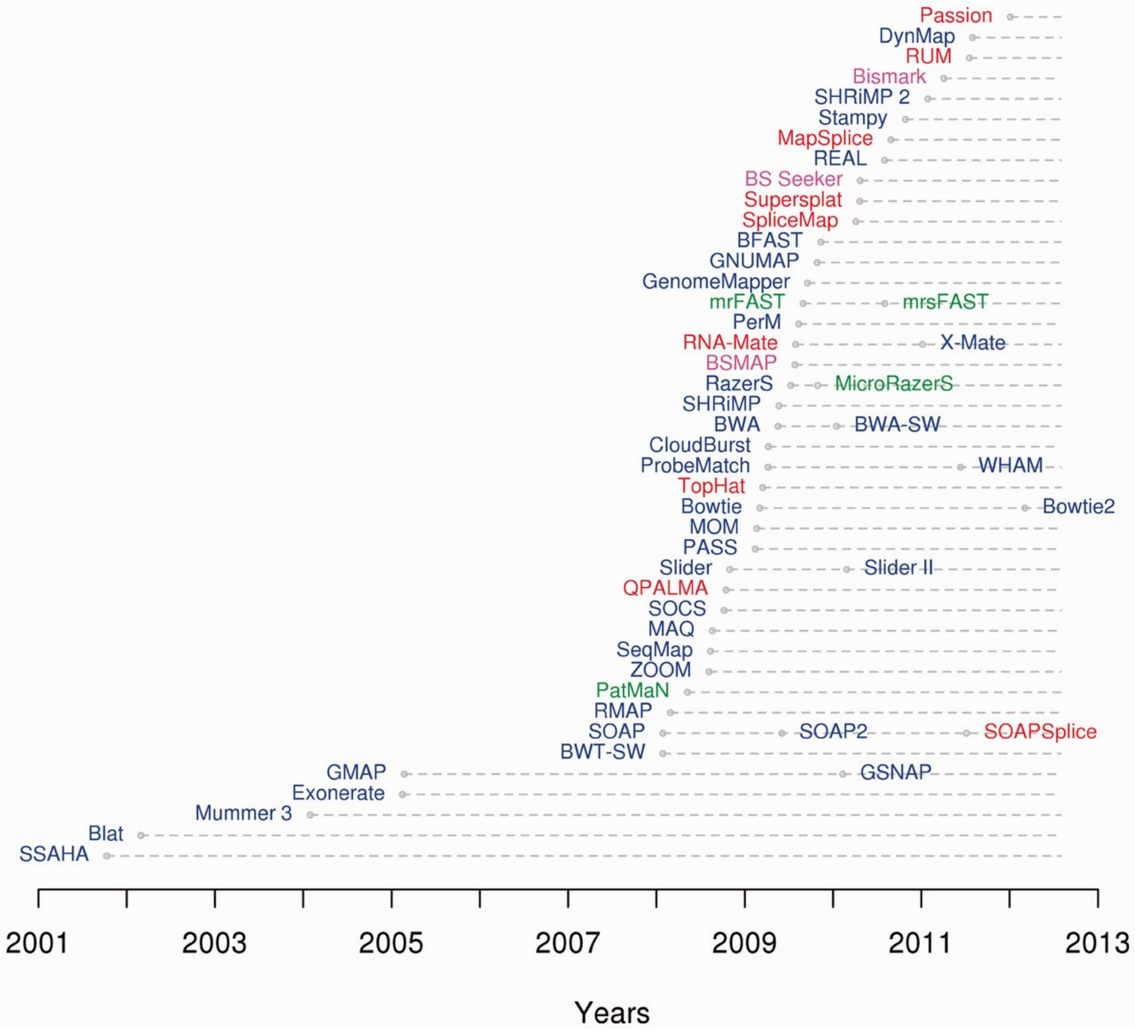
Bowtie
GSNAP
Smalt
Tophat
SOAPsplice
BWA
GEM
...

HTSeq
Cufflinks
Flux-capacitor
MISO
iReckon
....

DESeq
BaySeq
Cuffdiff
EdgeR
Flux-capacitor
DEXseq
BitSeq
....

What makes a difference?

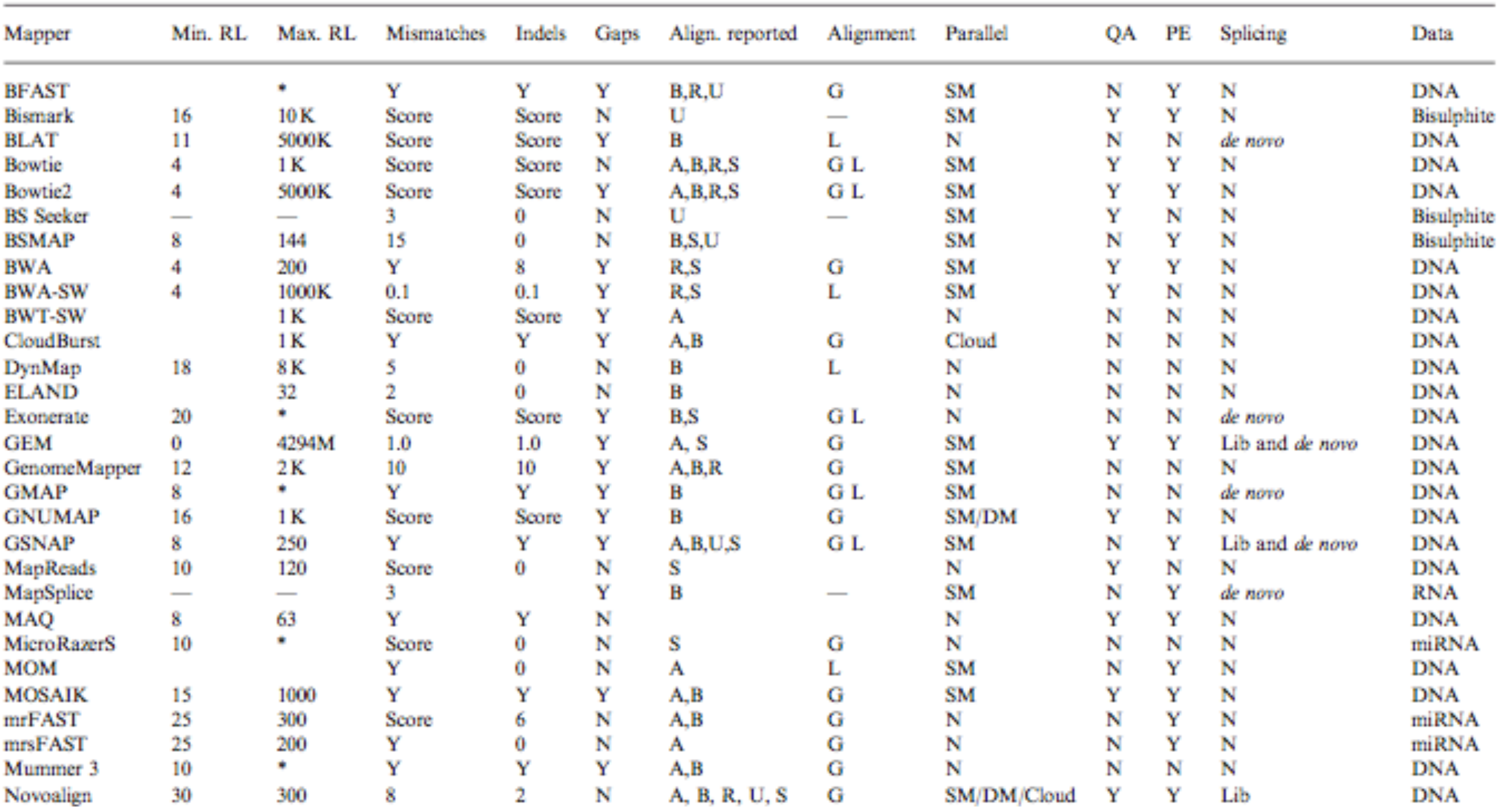
Mappers timeline (since 2001)



Nuno Fonseca

Fonseca at al, 2012. *Bioinformatics.* *28:* 3169-3177

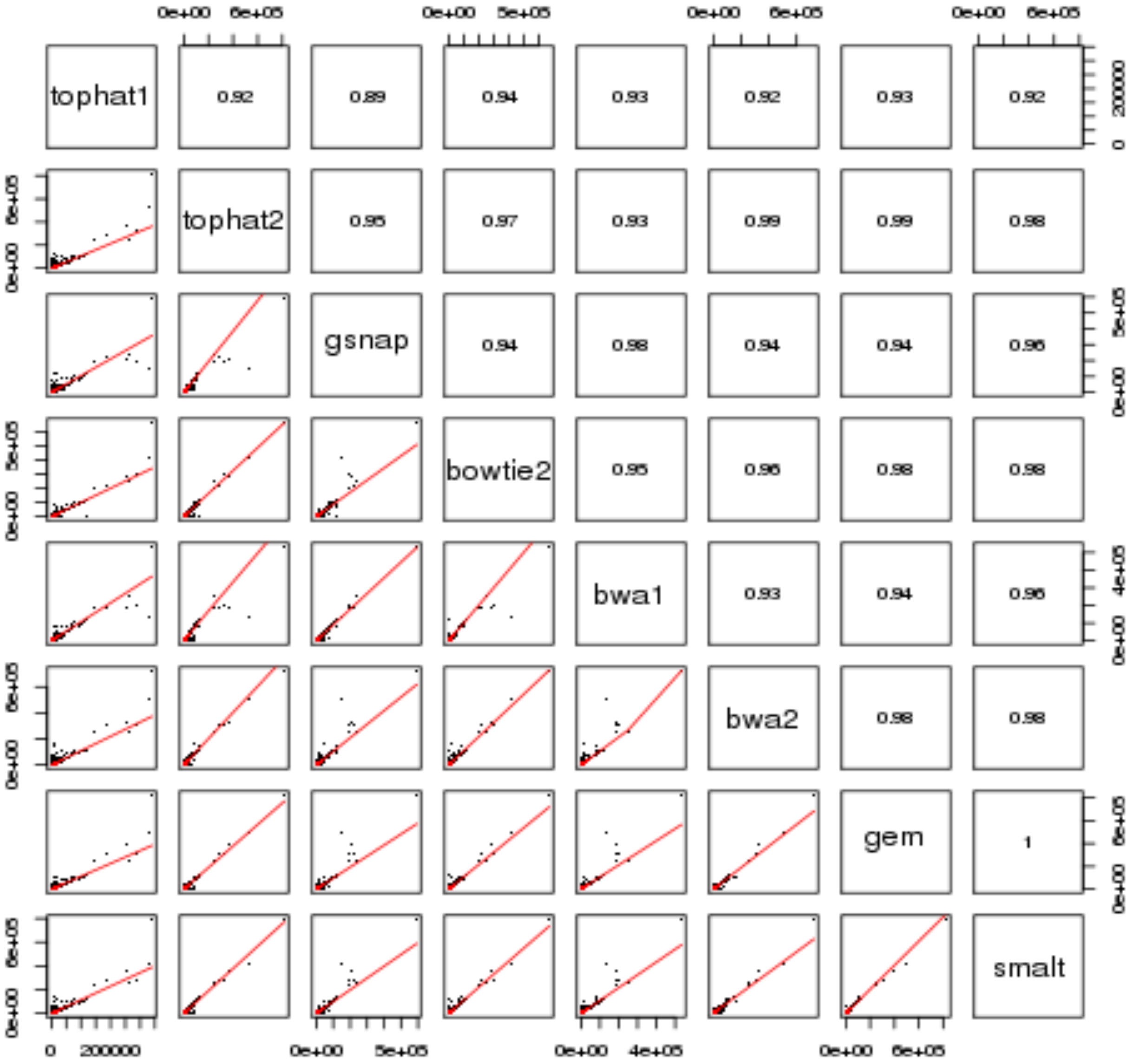
Mappers – features comparison



Mapper	Min. RL	Max. RL	Mismatches	Indels	Gaps	Align. reported	Alignment	Parallel	QA	PE	Splicing	Data
BFAST		*	Y	Y	Y	B,R,U	G	SM	N	Y	N	DNA
Bismark	16	10 K	Score	Score	N	U	—	SM	Y	Y	N	Bisulphite
BLAT	11	5000K	Score	Score	Y	B	L	N	N	N	*de novo*	DNA
Bowtie	4	1 K	Score	Score	N	A,B,R,S	G L	SM	Y	Y	N	DNA
Bowtie2	4	5000K	Score	Score	Y	A,B,R,S	G L	SM	Y	Y	N	DNA
BS Seeker	—	—	3	0	N	U	—	SM	Y	N	N	Bisulphite
BSMAP	8	144	15	0	N	B,S,U		SM	N	Y	N	Bisulphite
BWA	4	200	Y	8	Y	R,S	G	SM	Y	Y	N	DNA
BWA-SW	4	1000K	0.1	0.1	Y	R,S	L	SM	Y	N	N	DNA
BWT-SW		1 K	Score	Score	Y	A		N	N	N	N	DNA
CloudBurst		1 K	Y	Y	Y	A,B	G	Cloud	N	N	N	DNA
DynMap	18	8 K	5	0	N	B	L	N	N	N	N	DNA
ELAND		32	2	0	N	B		N	N	N	N	DNA
Exonerate	20	*	Score	Score	Y	B,S	G L	N	N	N	*de novo*	DNA
GEM	0	4294M	1.0	1.0	Y	A, S	G	SM	Y	Y	Lib and *de novo*	DNA
GenomeMapper	12	2 K	10	10	Y	A,B,R	G	SM	N	N	N	DNA
GMAP	8	*	Y	Y	Y	B	G L	SM	N	N	*de novo*	DNA
GNUMAP	16	1 K	Score	Score	Y	B	G	SM/DM	Y	N	N	DNA
GSNAP	8	250	Y	Y	Y	A,B,U,S	G L	SM	N	Y	Lib and *de novo*	DNA
MapReads	10	120	Score	0	N	S		N	Y	N	N	DNA
MapSplice	—	—	3		Y	B	—	SM	N	Y	*de novo*	RNA
MAQ	8	63	Y	Y	N			N	Y	Y	N	DNA
MicroRazerS	10	*	Score	0	N	S	G	N	N	N	N	miRNA
MOM			Y	0	N	A	L	SM	N	Y	N	DNA
MOSAIK	15	1000	Y	Y	Y	A,B	G	SM	Y	Y	N	DNA
mrFAST	25	300	Score	6	N	A,B	G	N	N	Y	N	miRNA
mrsFAST	25	200	Y	0	N	A	G	N	N	Y	N	miRNA
Mummer 3	10	*	Y	Y	Y	A,B	G	N	N	N	N	DNA
Novoalign	30	300	8	2	N	A, B, R, U, S	G	SM/DM/Cloud	Y	Y	Lib	DNA

Fonseca at al, 2012. *Bioinformatics. 28:* 3169-3177

RNA-Seq Mappers



RNA-Seq – iRAP pipeline

Reads → Mapping → Quantification & Normalization → DE

Filtering/QC

No

Yes

FASTQC

FASTX

Check for contamination

Tophat1
Tophat2
Bowtie1
Bowtie2
SMALT
GSNAP
GEM
BWA1
BWA2
SoapSplice
Star
BFAST

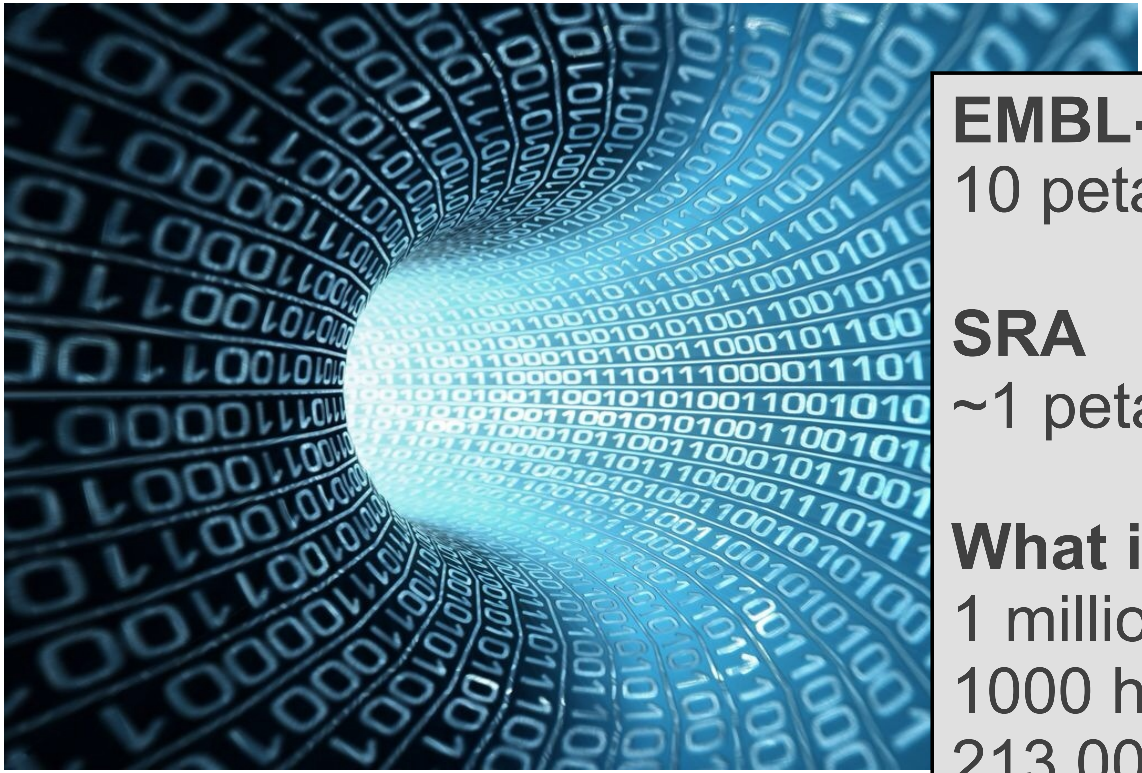
Cufflinks1
Cufflinks2
HTSeq
Flux-capacitor
Basic counting per exon

Scripture

Cuffdiff1
Cuffdiff2
DESeq
EdgeR
Flux-capacitor
DEXseq

Fonseca, N.A. et al (2013) iRAP – an integrated RNA-seq Analysis Pipeline, *Bioinformatics*, submitted

NGS data storage



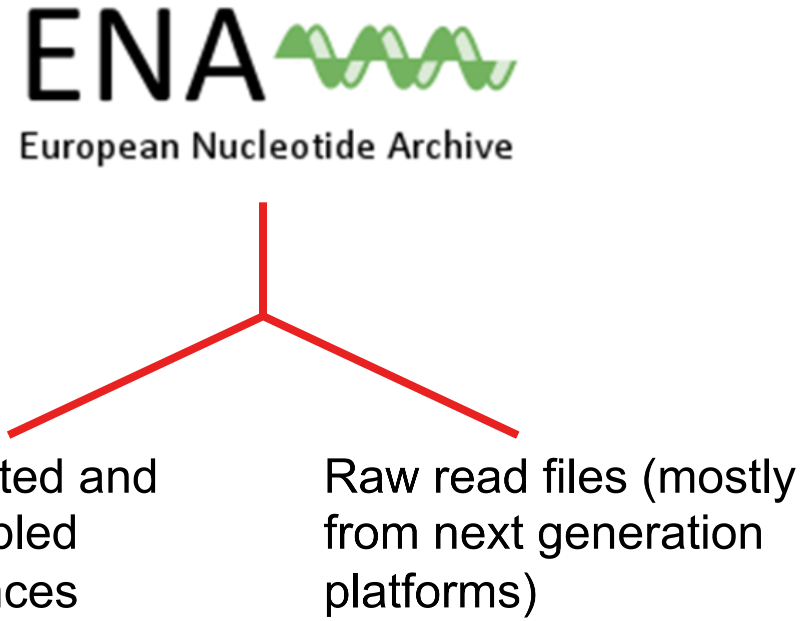
EMBL-EBI
10 petabytes

SRA
~1 petabytes

What is a petabyte?
1 million gigabytes
1000 hard drives (1TB)
213.000 DVDs

Complete Genomics
0.5 TB for a single file

ENA archives raw sequence data

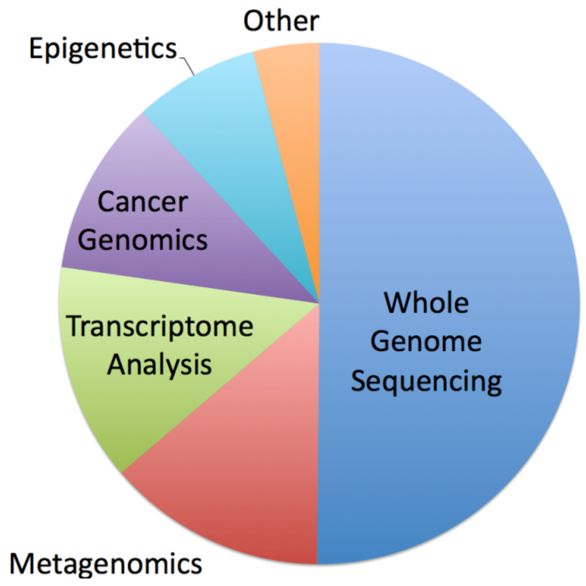


ENA

European Nucleotide Archive

Annotated and assembled sequences

Raw read files (mostly from next generation platforms)



- This is a global initiative, coordinated by the International Nucleotide Sequence Database Collaboration (INSDC)
- Other archives at DDBJ and NCBI
- All archives are mirrored for consistency across the INSDC

ENA supports other EBI services

EBI Metagenomics

https://www.ebi.ac.uk/metagenomics/
Environmental sample /Community sequencing
MiXS and MIMARKS standards by GSC

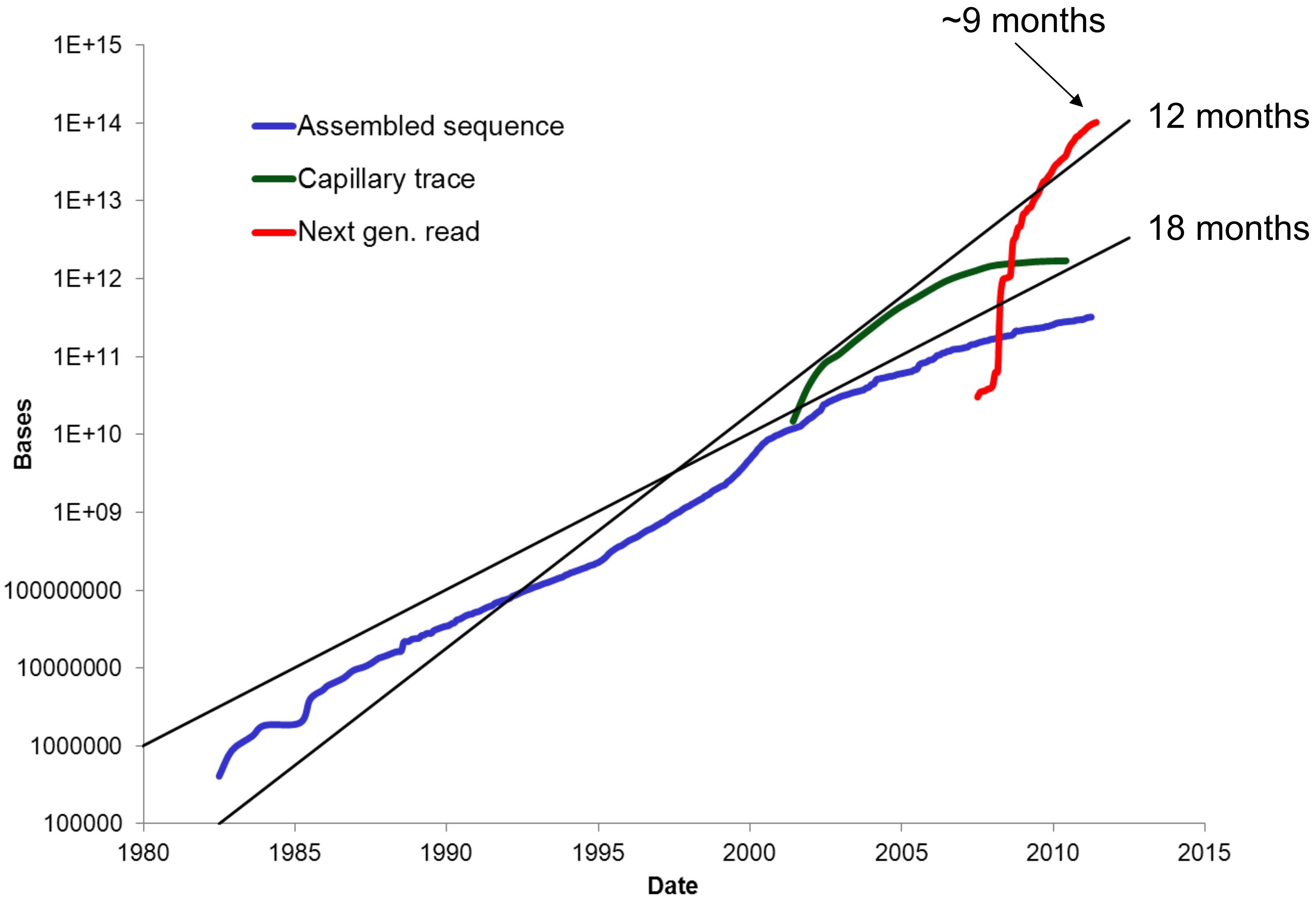
ARRAYEXPRESS

http://www.ebi.ac.uk/arrayexpress/
Expression studies benefit from MIAME (Minimum Information About a Microarray Experiment) related standards (MINSEQE)

european genome-phenome archive

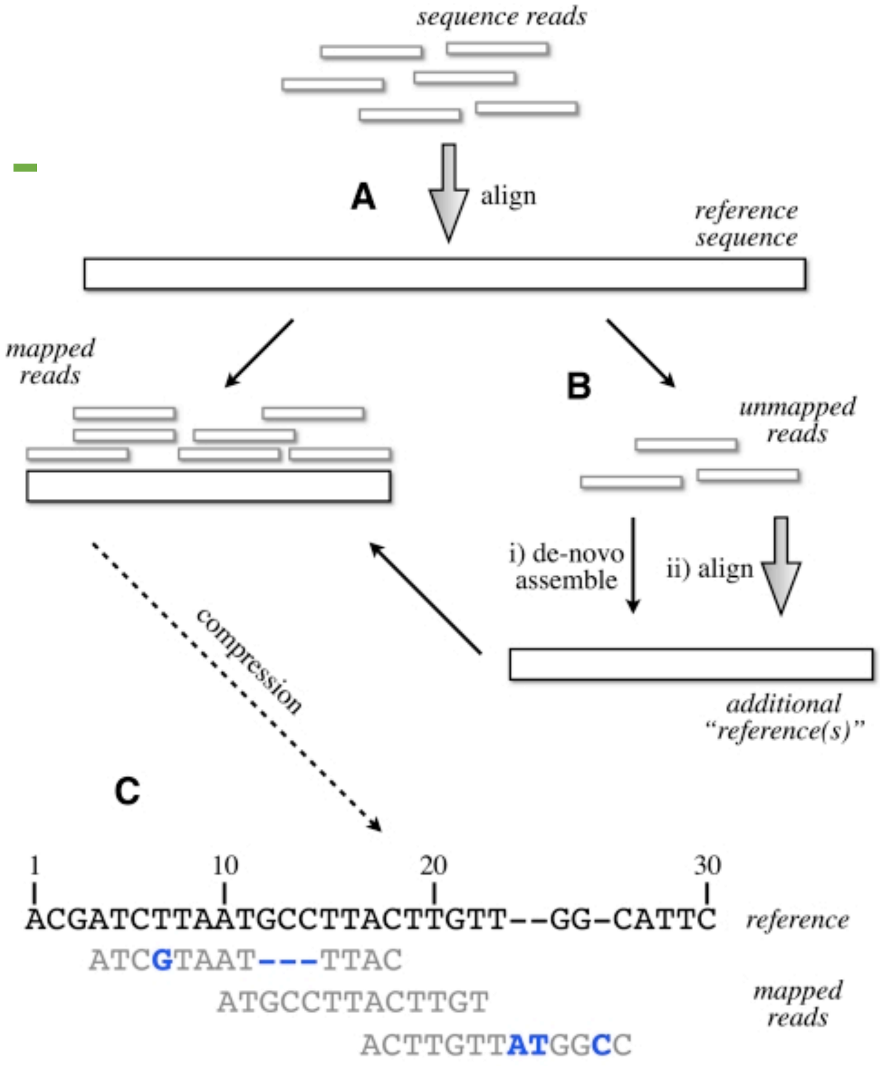
https://www.ebi.ac.uk/ega/
Access to data controlled by submitter nominated data access committee (DAC)

The need for compression



Reference-based compression technique - CRAM

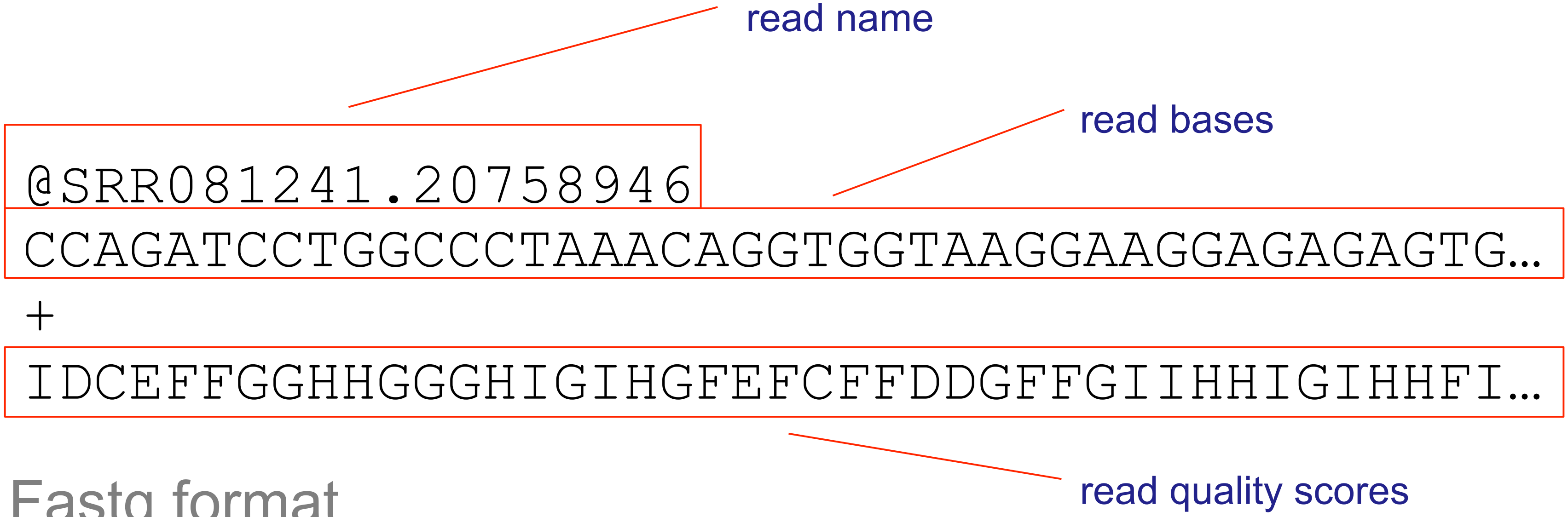
1. Reads are first aligned to the reference
2. Unaligned reads are pooled to create a specific “compression framework” for this data set
3. The base pair information is stored using specific offsets of reads on the reference, with additional information



Position	Strand	Substitutions	Insertions	Deletions
4	+	4-G	*none*	5-3
6	+	*none*	*none*	*none*
7	+	*none*	8-AT 4-C	*none*

Fritz et al, 2011. *Genome Res*. 21:734-740

What is a Read?

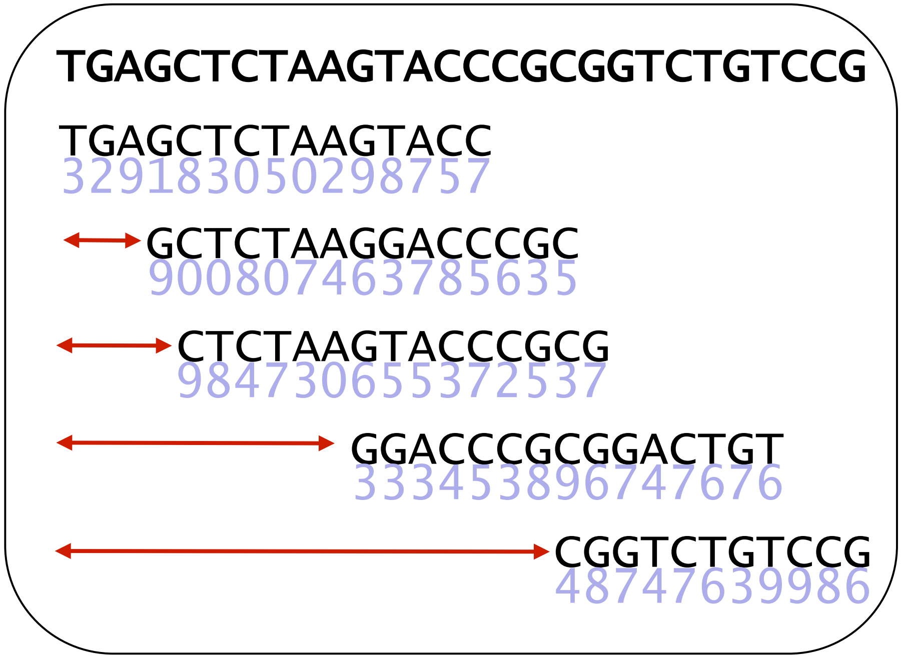


```
@SRR081241.20758946
CCAGATCCTGGCCCTAAACAGGTGGTAAGGAAGGAGAGAGTG…
+
IDCEFFGGHHGGGHIGIHGFEFCFFDDGFFGIIHHIGIHHFI…
```

Fastq format

- ✓ Usually 50-100 bp long
- ✓ Quality score is a measure of how certain the machine was about the observed base.

CRAM lossless model: sequence information



TGAGCTCTAAGTACCCGCGGTCTGTCCG

TGAGCTCTAAGTACC
329183050298757

GCTCTAAGGACCCGC
900807463785635

CTCTAAGTACCCGCG
984730655372537

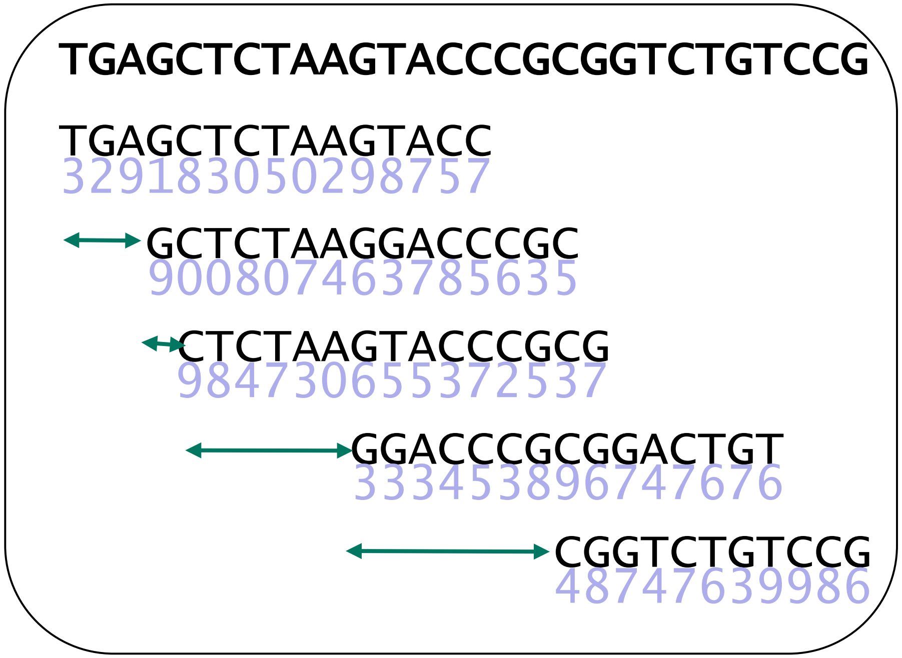
GGACCCGCGGACTGT
333453896747676

CGGTCTGTCCG
48747639986

Start	Sequence
0	TGAGCTCTAAGTACC
3	GCTCTAAGGACCCGC
4	CTCTAAGTACCCGCG
10	GGACCCGCGGACTGT
17	CGGTCTGTCCG

- Store start positions
- This is one possibility, but we can do better!

CRAM lossless model: sequence information



TGAGCTCTAAGTACCCGCGGTCTGTCCG

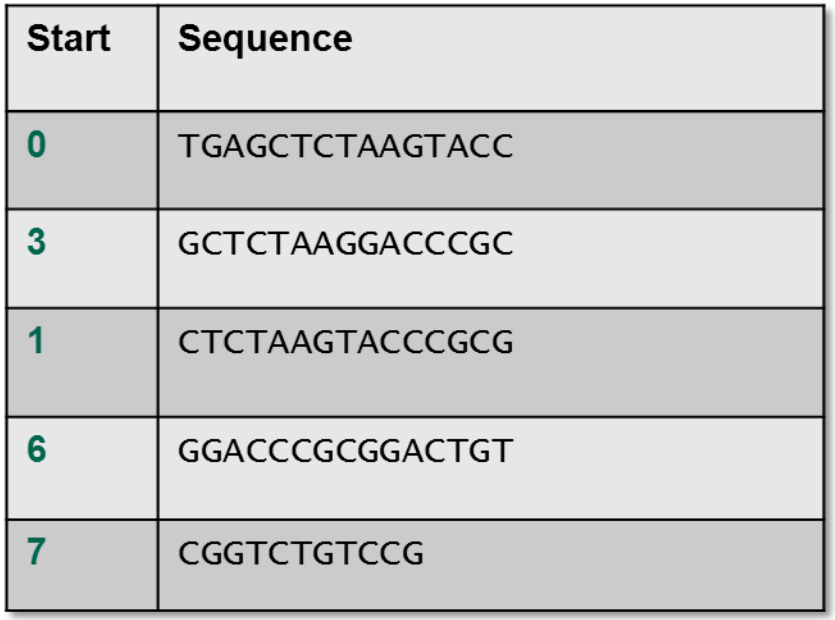
TGAGCTCTAAGTACC
329183050298757

GCTCTAAGGACCCGC
900807463785635

CTCTAAGTACCCGCG
984730655372537

GGACCCGCGGACTGT
333453896747676

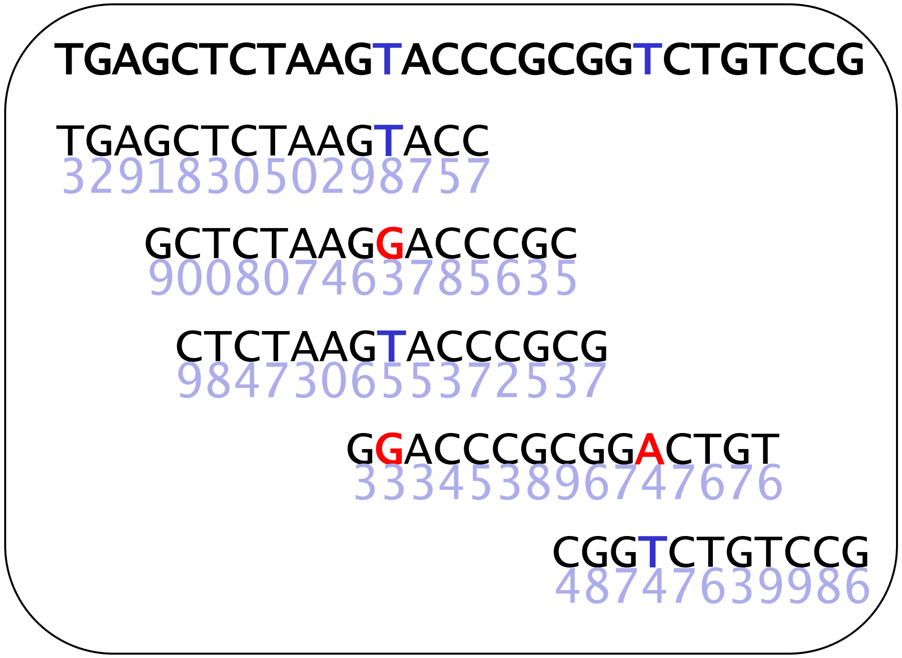
CGGTCTGTCCG
48747639986



Start	Sequence
0	TGAGCTCTAAGTACC
3	GCTCTAAGGACCCGC
1	CTCTAAGTACCCGCG
6	GGACCCGCGGACTGT
7	CGGTCTGTCCG

- Store start offsets

CRAM lossless model: sequence information

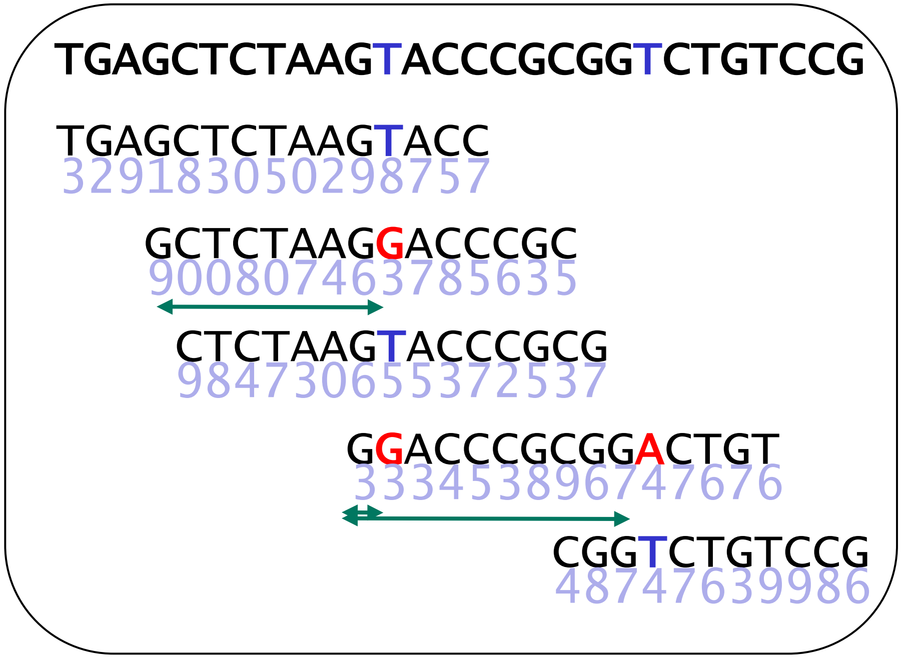


```
TGAGCTCTAAGTACCCGCGGTCTGTCCG

TGAGCTCTAAGTACC
329183050298757

   GCTCTAAGGACCCGC
   900807463785635

    CTCTAAGTACCCGCG
    984730655372537

          GGACCCGCGGACTGT
          333453896747676

                 CGGTCTGTCCG
                 48747639986
```

Start	Mismatch location	Mismatch call
0	–	
3	11	G
1	–	
6	11 20	G A
7	–	

- Store start offsets
- Store mismatch positions and calls

CRAM lossless model: sequence information



TGAGCTCTAAGTACCCGCGGTCTGTCCG

TGAGCTCTAAGTACC
329183050298757

GCTCTAAGGACCCGC
900807463785635

CTCTAAGTACCCGCG
984730655372537

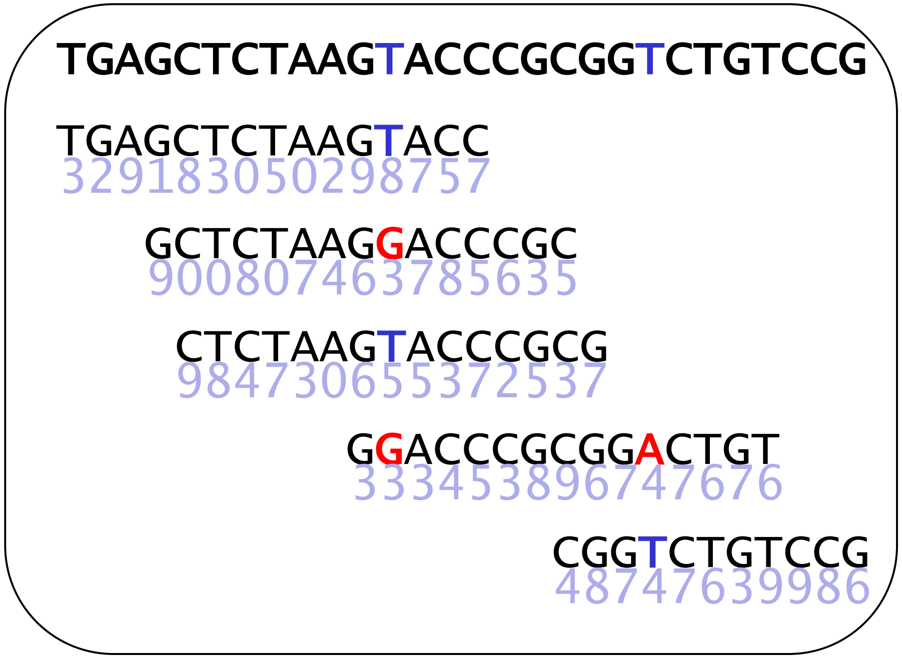
GGACCCGCGGACTGT
333453896747676

CGGTCTGTCCG
48747639986

Start	Mismatch location	Mismatch call
0	–	
3	8	G
1	–	
6	1 10	G A
7	–	

- Store start offsets
- Store mismatch offsets and calls

CRAM lossless model: sequence information



TGAGCTCTAAGTACCCGCGGTCTGTCCG

TGAGCTCTAAGTACC
329183050298757

GCTCTAAGGACCCGC
900807463785635

CTCTAAGTACCCGCG
984730655372537

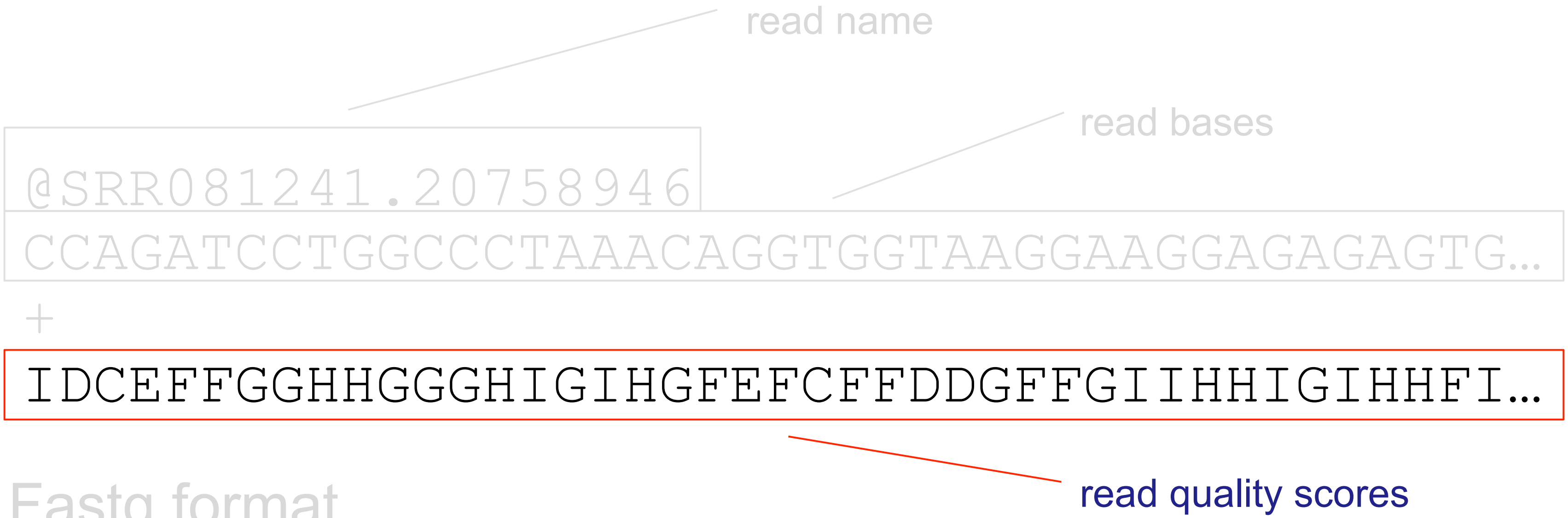
GGACCCGCGGACTGT
333453896747676

CGGTCTGTCCG
48747639986

Start	Sequence
0	TGAGCTCTAAGTACC
3	GCTCTAAGGACCCGC
4	CTCTAAGTACCCGCG
10	GGACCCGCGGACTGT
17	CGGTCTGTCCG

Start	Mismatch location	Mismatch call
0	–	
3	8	G
1	–	
6	1 10	G A
7	–	

What is a Read?



read name

read bases

```
@SRR081241.20758946
CCAGATCCTGGCCCTAAACAGGTGGTAAGGAAGGAGAGAGTG...
+
IDCEFFGGHHGGGHIGIHGFEFCFFDDGFFGIIHHIGIHHFI...
```

read quality scores

Fastq format

CRAM lossy model - Quality scores

- All the quality scores of positions showing variation are stored
- In addition, a user defined percentage of quality positions (that are identical to the reference) can be stored
- Percentage specific to classes of data and, potentially, specific data sets
- By allowing this, the compression can place more value on some data sets than others

CRAM – a technology for raw sequence data compression

- This technology offers:
  - lossless compression, in which read sequence and per-base quality information is faithfully preserved, and
  - lossy models, in which data are selectively reduced to reach an optimal balance between data preservation and compression
- Focused on compressing whole genome sequences as this will be the largest component of sequence archives growth for the next decade
- Can be applied to RNA-seq and ChIP-seq but attention should be paid to aspects as unaligned data

Data reproducibility is crucial

- How do you store your data? How do you document it? If you leave, how easy is it for coworkers to continue your progress? If you stop for a while, how easy is it to restart?
- Bioconductor focuses on:
  - ✓ open-source, open-development
  - ✓ versioned packaging of data, metadata, and analytic software. Past experiments can be replicated using the exact version of software that was used for the actual analysis
  - ✓ high-quality coding and documentation standards (i.e. package vignette)

in order to foster reproducible analysis in genome scale biology.

Future NGS developments and challenges

- Data processing and storage needs to keep up to date with emerging new technologies (i.e. single cell sequencing)
- Genome interpretation: understanding the significance of variants in individual genomes on human phenotypes and diseases
- Cost-benefit analyses of sequencing applications in the clinic have to be conducted before actual medical application
- Ethical issues will emerge with the commonalization of personal genomes

Acknowledgments

Nuno Fonseca, Brazma’s group, EMBL-EBI

John Marioni, EMBL-EBI

Rajesh Radhakrishnan, ENA, EMBL-EBI

Vadim Zalunin, ENA, EMBL-EBI

More information:

- http://www.ebi.ac.uk/ena/about/cram_toolkit
- http://wwwdev.ebi.ac.uk/fg/hts_mappers/
- http://www.ebi.ac.uk/training/
- http://www.ebi.ac.uk/training/online/