

```
@M01264:15:000000000-D0BG0:1:1101:15111:1933 1:N:0:1  
CACTTCAACCTCTGCCTCCCAGGTTCAAGTGATTCTCCTGCC  
+  
>11>AF@D@1B1B11BAFFE1CAFGFGGHGHHHHHHGHHHHH  
@M01264:15:000000000-D0BG0:1:1101:18980:2056 1:N:0:1  
GGAAACACCTTTCAACTGGAAGATATCATTAACACAGG  
+  
AAAAAFF?FFFGBFGFGGEGGHHHHHHHHHHHHHHHHHH
```

# NGS: from the design of the experiment to sequenced reads

Summer school Malta 13 September 2016

# Experimental design

- The design of the experiment is the first step and it is obviously determinant for all the downstream analyses.
- You have to evaluate all the potentialities and limitations of available technologies, designing the experiment according your goals.
- If you plan to use comparisons among samples to derive conclusions about the underlying population you need statistics.
- Statistics needs biological replicates.

## 1) COVERAGE: How many reads do we need?

The coverage is defined as  $C = (R_{\text{length}} \times R_{\text{num}}) / A_{\text{length}}$

$R_{\text{length}}$  = length in nucleotides of the reads (2XPE)

$R_{\text{num}}$  = number of sequenced reads (2XPE)

$A_{\text{length}}$  = number of nucleotides of sequenced subject (Genome, transcriptome, exome)

The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample.

## 2) ACCURACY: How much accurate has to be the experiment?

If we aim to identify SNP the accuracy has to be a priority: in order to identify a human SNP with frequency  $1/800$  we need accuracy rate of 99.9%.

A fewer accuracy is required for other goals (e.g. : identify known protein-coding genes; improve the annotation of the gene structure; quantitative transcript estimation; annotation of new genes...)

We can increase the accuracy rate by increasing the coverage (number of reads for sequenced transcript): 10 reads of the same transcribed RNA with an accuracy of 99.9% can effectively provide an accuracy level of 99.99%

## 3) Read length: long or short reads?

The answer depends again by the experiment we are going to project:

GENOME RESEQUENCING

De novo TRANSCRIPTOME

TRANSCRIPTOME seq

ChIP seq

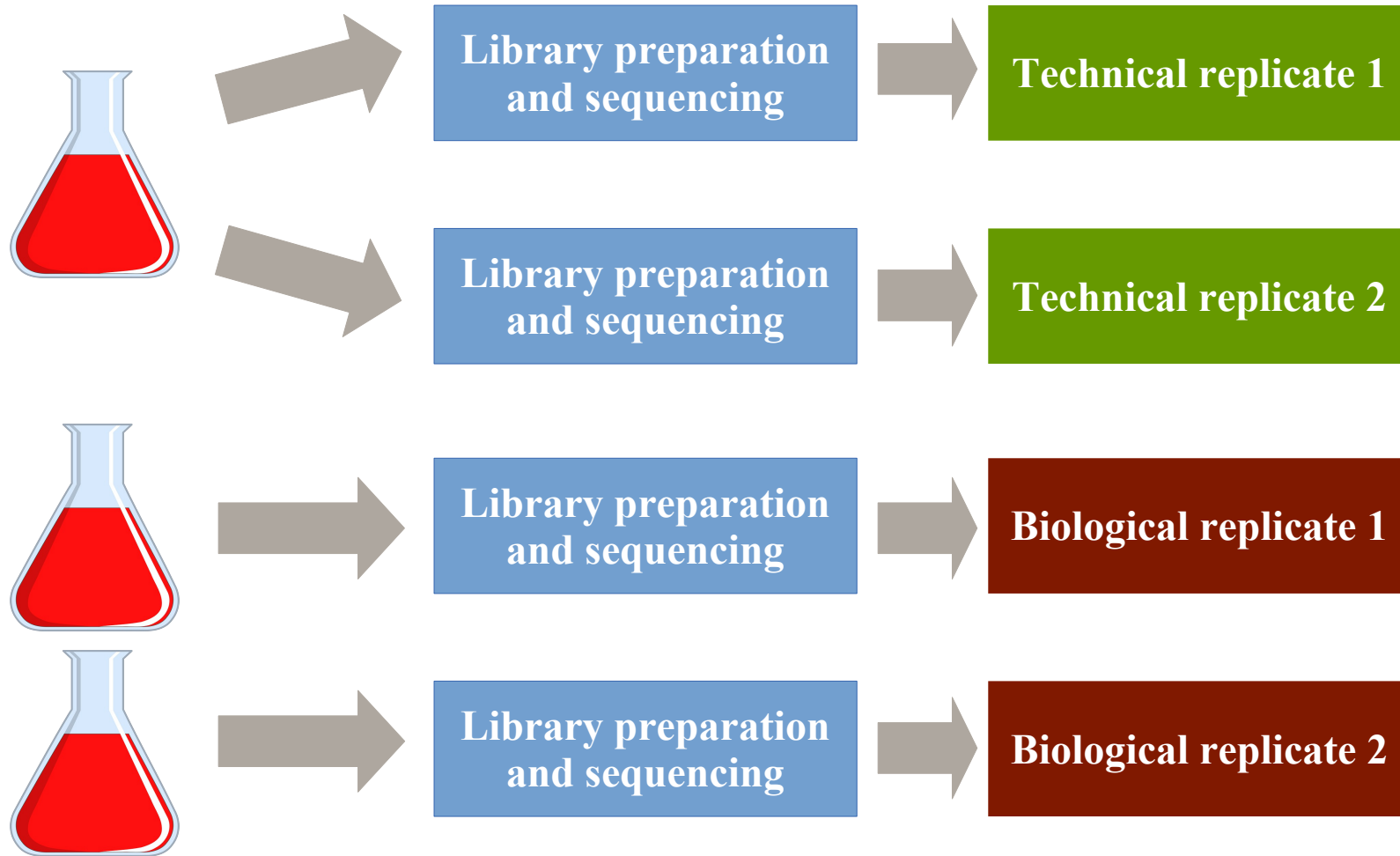
Read length is inversionally proportional to its multi-mappability, in a sample made by 50 nt reads there is a small fraction (<0.01 %) that can be mapped on multiple positions of the human genome.

# All we need are replicates!



Measurements are usually subject to variation and uncertainty. We need replicates to minimize fluctuations due measurement errors.

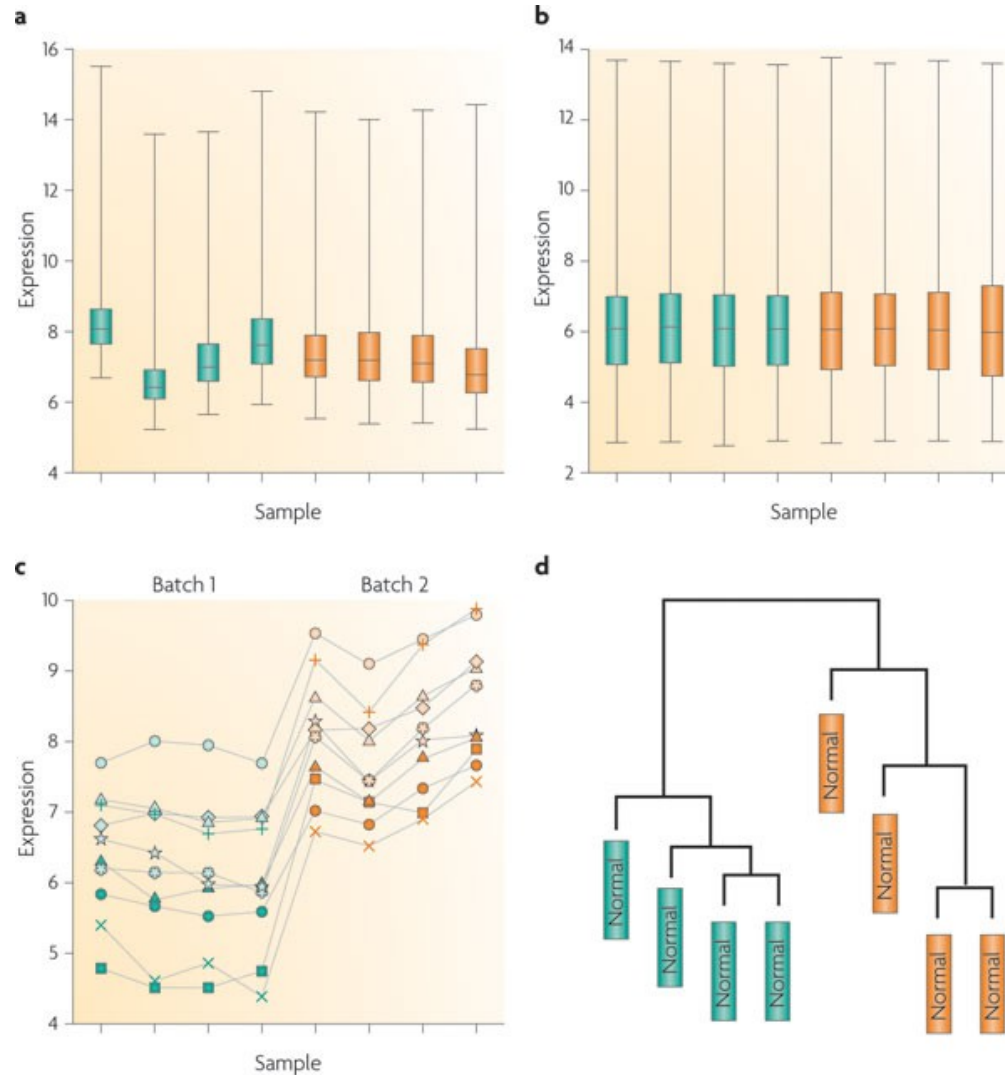
# Replicates



Biological replicates are samples that should be identical (as much as you can/want control) but are biologically separated (different cells, different organisms, different populations, colonies...)

# Avoiding batch effects

Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study.

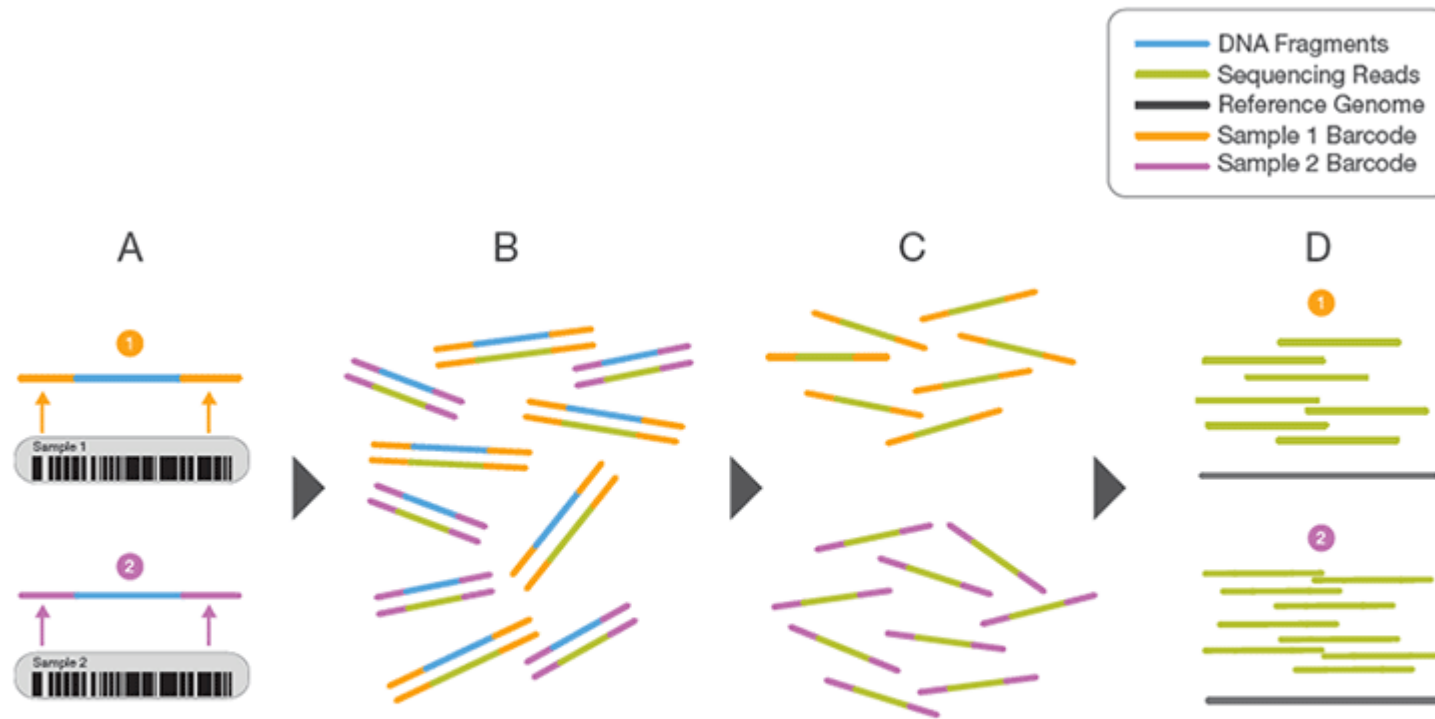


Nature Reviews | Genetics



# Multiplexing the samples can contrast batch effects

Figure 2: Conceptual Overview of Sample Multiplexing



- Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
- Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
- Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
- Each set of reads is aligned to the reference sequence.



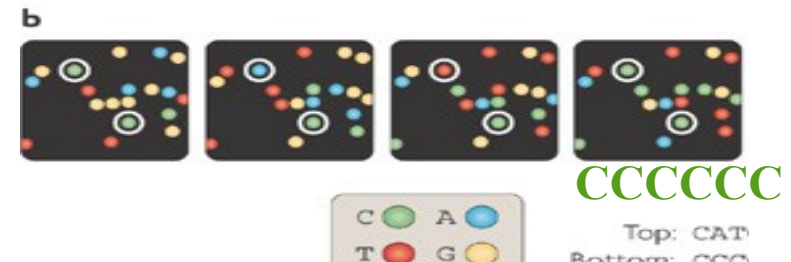
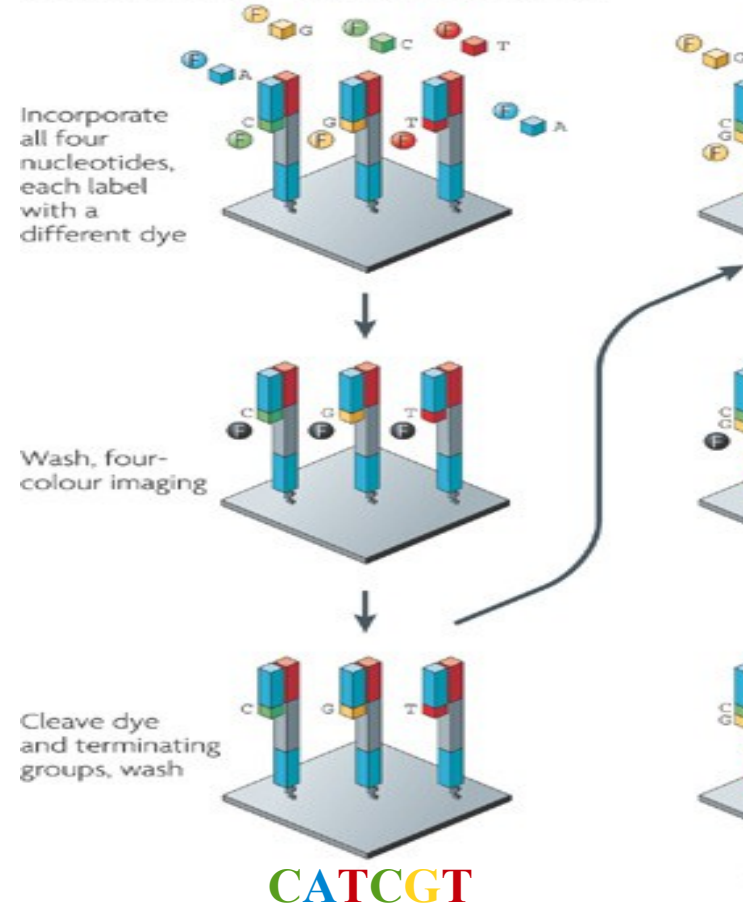
Which technology?

# Sequencing Technologies



The four-colour cyclic reversible termination (CRT) method uses 3'-O-azidomethyl reversible terminator chemistry using solid-phase-amplified template clusters

a Illumina/Solexa — Reversible terminators



# Sequencing Technologies

Focused power.



MiSeq Series

Small genome, amplicon, and targeted gene panel sequencing.

Flexible power.



NextSeq Series

Everyday genome, exome, transcriptome sequencing, and more.

Production power.



HiSeq Series

Production-scale genome, exome, transcriptome sequencing, and more.

Population power.



HiSeq X Series

Population- and production-scale human whole-genome sequencing.

	1 MiSeq	NextSeq	HiSeq	HiSeq X
Output Range	0.3-15 Gb	20-120 Gb	10-1000 Gb	1.6-1.8 Tb
Total reads	25 M	130-400 M	300 M – 2 B	3 B
Read length	Up to 300 b (x2)	Up to 150 b (x2)	Up to 150 b (x2)	Up to 150 b (x2)
Run time	5–65 h	15-30h	7 hr – 6 d	> 3 d

# Sequencing technologies



The trace is the hydrogen ion

Ion Proton



Ion PGM

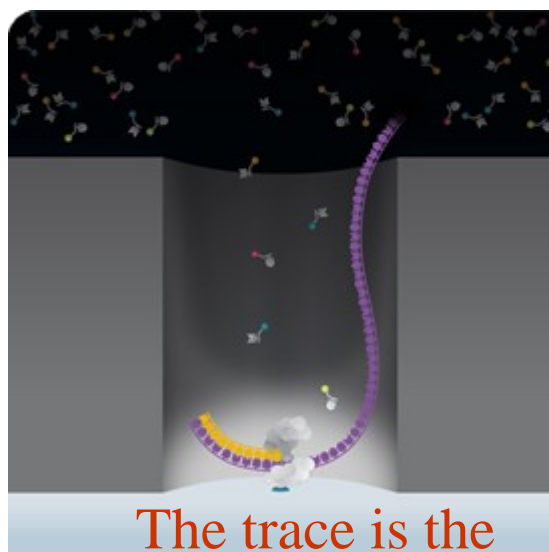


	Proton	Ion 314	Ion 316	Ion 318
Output Range	Up to 10 Gb	30-50 Mb;60-100 Mb	300-600 Mb;600 Mb-1 Gb	600 Mb-1 Gb; 1.2-2 Gb
Total reads	60–80 M	400-550 k	2-3 M	4-5.5 M
Read length	Up to 200 b	200; 400 b	200; 400 b	200; 400 b
Run time	2–4 hours	2.3 hr;3.7 hr	3.0 hr;4.9 hr	4.4 hr;7.3 hr

# Sequencing technologies



## SMRT TECHNOLOGY



The trace is the fluorophore incorporated into the 4 nucleotides

## PacBio RS II

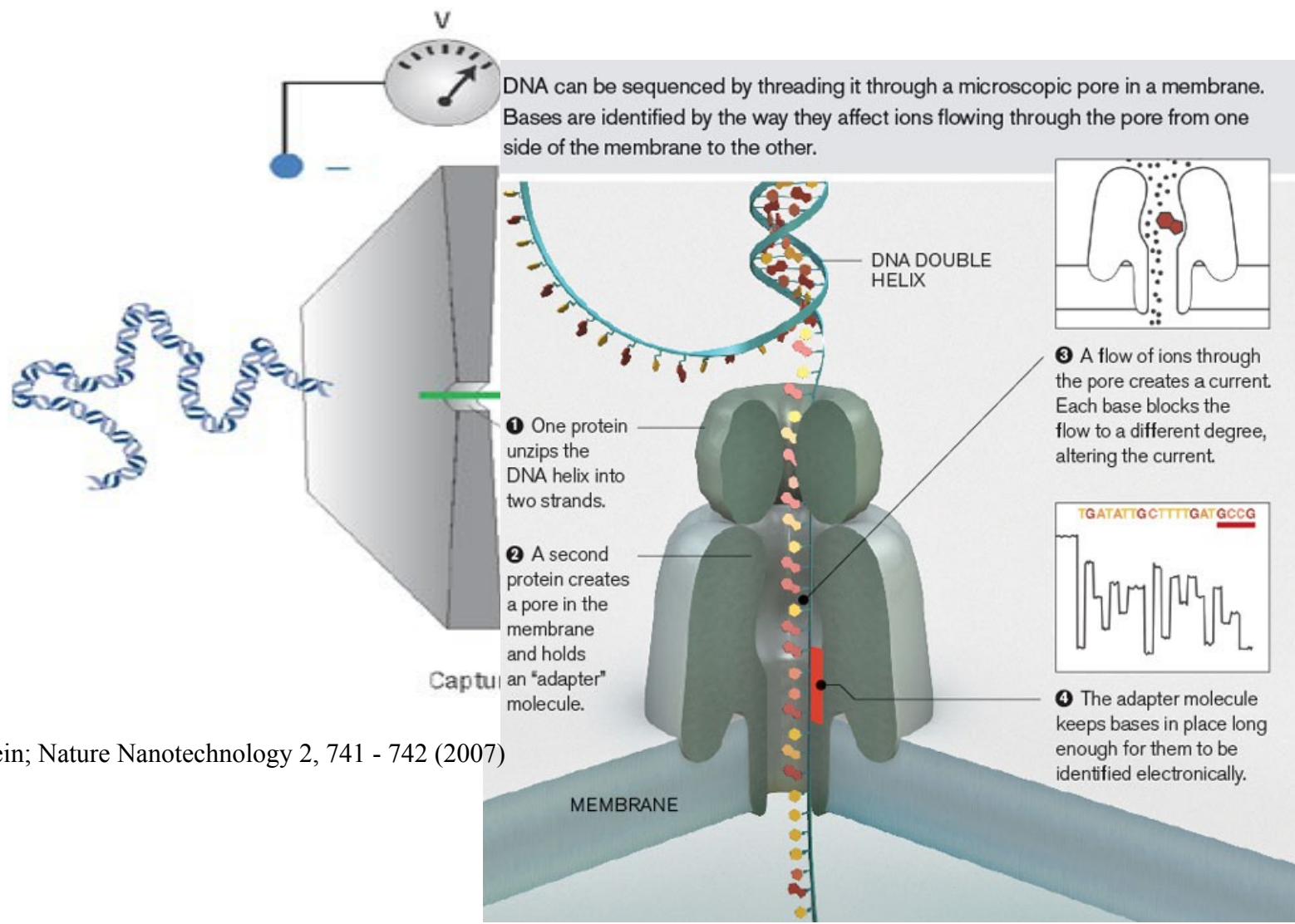


Single Molecule, Real-Time (SMRT) DNA Sequencing System

Read length 8.5 kb, up to 18 kb

It can detect DNA modifications (Methylations)

# Sequencing technologies



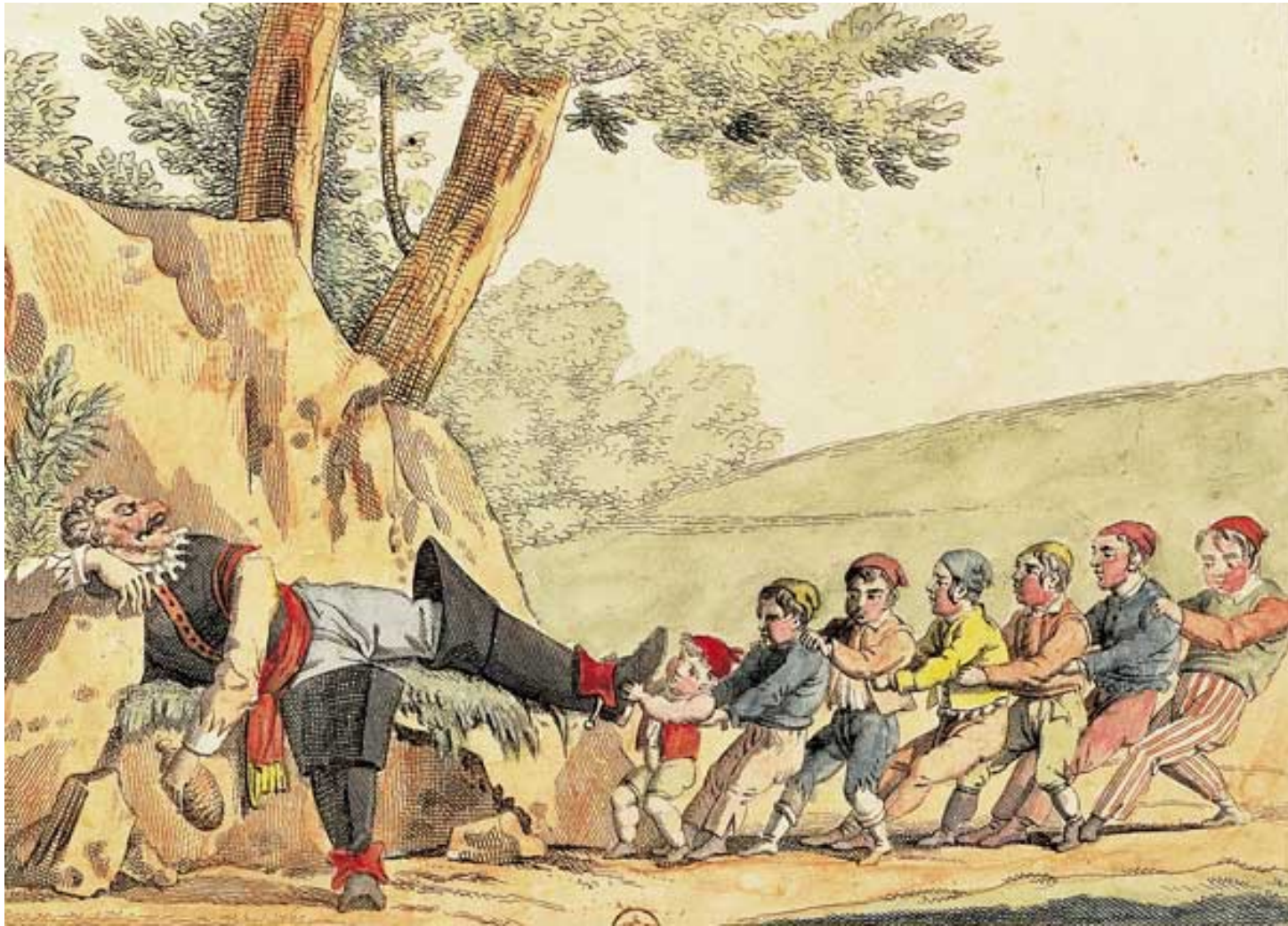
Stein; Nature Nanotechnology 2, 741 - 742 (2007)

# Sequencing Technologies

	Read length	Accuracy	Reads per run	Time per run	Advantages	Disadvantages
<b>Ion semiconductor (Ion Torrent sequencing)</b>	up to 400 b	98%	up to 80 million	2 hours	Less expensive equipment. Fast.	Homopolymer errors.
<b>Single-molecule real-time sequencing (Pacific Bio)</b>	avg 14,000 b maximum read length >40,000	99.999; 87%	50,000 per SMRT cell, or 500–1000 meg	30 minutes to 4 ho	Longest read length. Fast. Detects 4mC, 5mC, 6	Moderate throughput. Equipment can be very expensive.
<b>Sequencing by synthesis (Illumina)</b>	50 to 300 b	98%	up to 3 billion	1 to 10 days	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.



**Together is better!!!!!!!**



**Thousand of datasets freely available and just waiting for you!!!!**

# NCBI Sequencing Read Archive

<http://www.ncbi.nlm.nih.gov/sra>

NCBI Resources How To Sign in to NCBI

SRA SRA Search

Advanced Help



## SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

### Getting Started

[Understanding and Using SRA](#)

[How to Submit](#)

[Login to Submit](#)

[Download Guide](#)

### Tools and Software

[Download SRA Toolkit](#)

[SRA Toolkit Documentation](#)

[SRA-BLAST](#)

[SRA Run Browser](#)

[SRA Run Selector](#)

### Related Resources

[dbGaP Home](#)

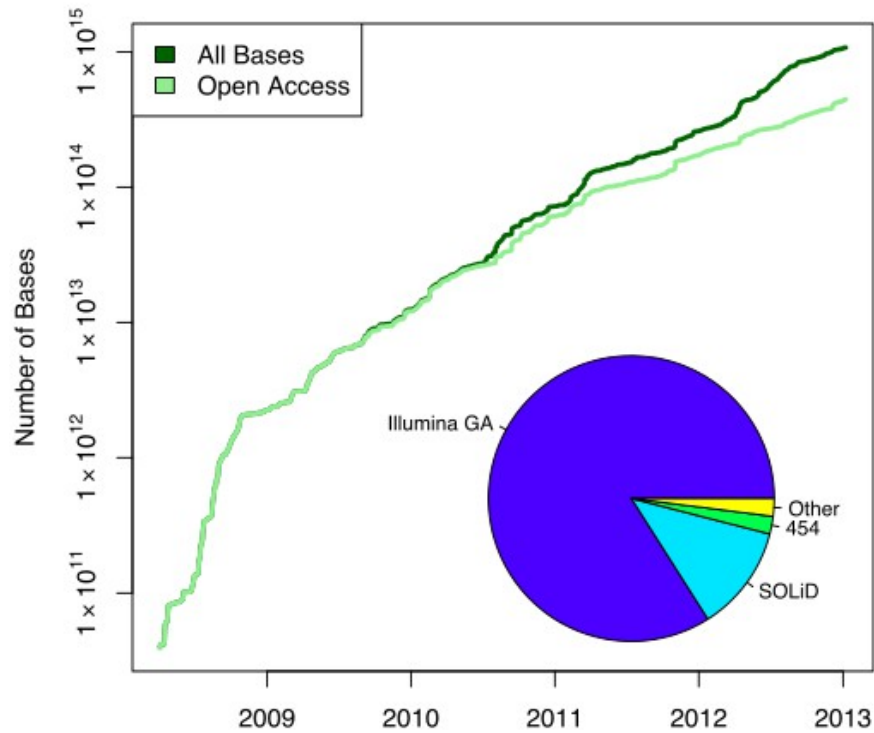
[Trace Archive Home](#)

[BioSample](#)

[GenBank Home](#)

Stores Raw sequencing data from high-throughput sequencing platforms  
Data format is SRA, a compressed one easily convertible in Fastq

# NCBI Sequencing Read Archive



Moore B., History (and predicted future) size of the Sequence Read Archive

**Numbers to 01/02/2015**

Studies: 48005 records

Samples: 677374 records

**SRA Toolkit**

Handbook available at:

<http://www.ncbi.nlm.nih.gov/books/NBK242621>

# NCBI Gene Expression Omnibus

- <http://www.ncbi.nlm.nih.gov/geo/>

The screenshot shows the NCBI Gene Expression Omnibus (GEO) website homepage. At the top, there is a navigation bar with the NCBI logo, links for 'Resources' and 'How To', and a 'Sign in to NCBI' button. Below this is a secondary navigation bar with 'GEO Home', 'Documentation', 'Query & Browse', and 'Email GEO'. The main heading is 'Gene Expression Omnibus', followed by a brief description: 'GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.' To the right of the description is the GEO logo and a search box labeled 'Keyword or GEO Accession' with a 'Search' button. The page is organized into three main columns: 'Getting Started', 'Tools', and 'Browse Content'. 'Getting Started' includes links for Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, and How to Download Data. 'Tools' includes links for Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, GEO BLAST, Programmatic Access, and FTP Site. 'Browse Content' includes a Repository Browser and a table of statistics: DataSets: 3848, Series: 54610, Platforms: 13886, and Samples: 1331166. At the bottom, there is an 'Information for Submitters' section with links for Login to Submit, Submission Guidelines, Update Guidelines, MIAME Standards, Citing and Linking to GEO, Guidelines for Reviewers, and GEO Publications.

Getting Started	Tools	Browse Content
<a href="#">Overview</a>	<a href="#">Search for Studies at GEO DataSets</a>	<a href="#">Repository Browser</a>
<a href="#">FAQ</a>	<a href="#">Search for Gene Expression at GEO Profiles</a>	<a href="#">DataSets: 3848</a>
<a href="#">About GEO DataSets</a>	<a href="#">Search GEO Documentation</a>	<a href="#">Series: 54610</a>
<a href="#">About GEO Profiles</a>	<a href="#">Analyze a Study with GEO2R</a>	<a href="#">Platforms: 13886</a>
<a href="#">About GEO2R Analysis</a>	<a href="#">GEO BLAST</a>	<a href="#">Samples: 1331166</a>
<a href="#">How to Construct a Query</a>	<a href="#">Programmatic Access</a>	
<a href="#">How to Download Data</a>	<a href="#">FTP Site</a>	

Information for Submitters		
<a href="#">Login to Submit</a>	<a href="#">Submission Guidelines</a>	<a href="#">MIAME Standards</a>
	<a href="#">Update Guidelines</a>	<a href="#">Citing and Linking to GEO</a>
		<a href="#">Guidelines for Reviewers</a>
		<a href="#">GEO Publications</a>

A public functional genomics data repository where array and sequence-based data are accepted.

# NCBI Gene Expression Omnibus

It includes different types of experiments (RT-PCR, arrays, SAGE, tiling arrays, mass spec, etc...)

## Numbers to 01/02/2015

H.T. sequencing studies: 113328 records

*Homo sapiens* (37337)

*Mus musculus* (33583)

*Saccharomyces cerevisiae* (4347)

*Arabidopsis thaliana* (4120)

*Drosophila melanogaster* (4061)

*Caenorhabditis elegans* (3760)

*Rattus norvegicus* (1416)

*Danio rerio* (1284)

*Escherichia coli* (1268)

.....

# European Nucleotide Archive

- <http://www.ebi.ac.uk/ena>

The screenshot shows the ENA website interface. At the top, there is a navigation bar with links for 'Services', 'Research', 'Training', and 'About us'. Below this is the ENA logo and a search bar with a 'Search' button. The search bar includes examples: 'BN000065, histone' and links for 'Advanced search' and 'Sequence'. A secondary navigation bar contains links for 'Home', 'Search & Browse', 'Submit & Update', 'Software', 'About ENA', and 'Support'. The main content area features a large heading 'European Nucleotide Archive' followed by a descriptive paragraph: 'The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)'. Below this, it states: 'Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.' There are two search sections: 'Text Search' and 'Sequence Search'. The 'Text Search' section has a search input field, examples 'BN000065, histone', a 'Search' button, and a link to 'Advanced search'. The 'Sequence Search' section has a text area with the prompt 'Enter or paste a nucleotide sequence or accession number', a 'Search' button, and a link to 'Advanced search'. On the right side, there is a 'Popular' section with a list of links: 'Submit and update', 'Sequence submissions', 'Genome assembly submissions', 'Submitting environmental sequences', 'Citing ENA data', 'Rest URLs for data retrieval', and 'Rest URLs to search ENA'. Below that is a 'Latest ENA news' section with three news items: '09 Dec 2014: ENA release 122', '12 Nov 2014: Simplification of data release procedures', and '11 Nov 2014: ENA/EMG Sample Record Annotation Workshop'.

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.

# And many other data !!!

Encyclopedia Of DNA Elements (ENCODE) Project (Data collection, integrative analysis, and a comprehensive catalog of all sequence-based functional elements)

<http://www.genome.gov/10005107><http://encodeproject.org/>

Epigenomics (NIH Common Fund) Data collection, integrative analysis and a resource of human epigenomic data

<http://www.roadmapepigenomics.org/><https://commonfund.nih.gov/epigenomics/>

International Human Epigenome Consortium (IHEC) Data collection and reference maps of human epigenomes for key cellular states relevant to health and diseases

<http://www.ihec-epigenomes.org/>

BLUEPRINT :Epigenome Data collection on the epigenome of blood cells

<http://www.blueprint-epigenome.eu/><http://www.nature.com/nbt/journal/v30/n3/full/nbt.2153.html>

FANTOM5 Project Large collection of CAGE based expression data across multiple species (time-series and perturbations)

<http://fantom.gsc.riken.jp/>[http://fantom.gsc.riken.jp/5/sstar/Data\\_source](http://fantom.gsc.riken.jp/5/sstar/Data_source)

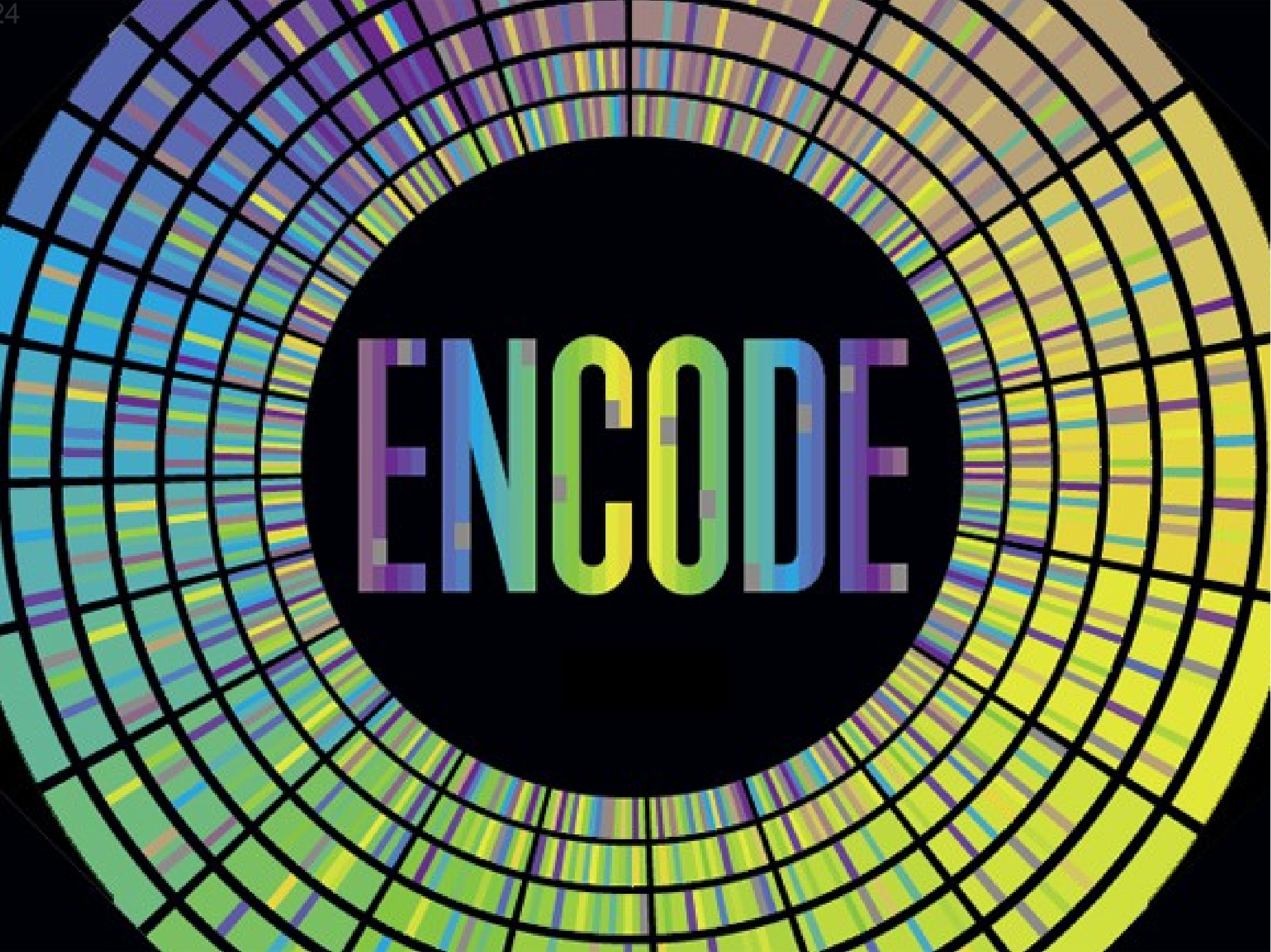
Genomic of drug sensitivity in cancer

<http://www.cancerrxgene.org/>

Geuvadis RNA sequencing project of 1000 Genomes samples: mRNA and small RNA sequencing on 465 lymphoblastoid cell line (LCL) .

<http://www.geuvadis.org/web/geuvadis>

**ENCODE**





# What is ENCODE?

OPEN  ACCESS Freely available online

PLOS **BIOLOGY**

## A User's Guide to the Encyclopedia of DNA Elements (ENCODE)

The ENCODE Project Consortium<sup>1</sup> \*

# ENCODE

The great novelty resides into the 'standardized catalog' approach to create public free resources

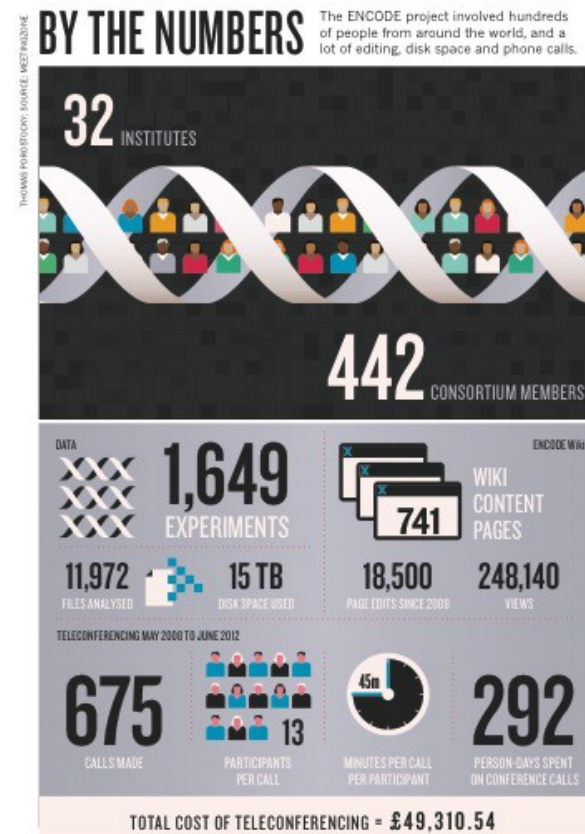
+1,600 experiments

147 cell types

225 antibodies

~450 authors from

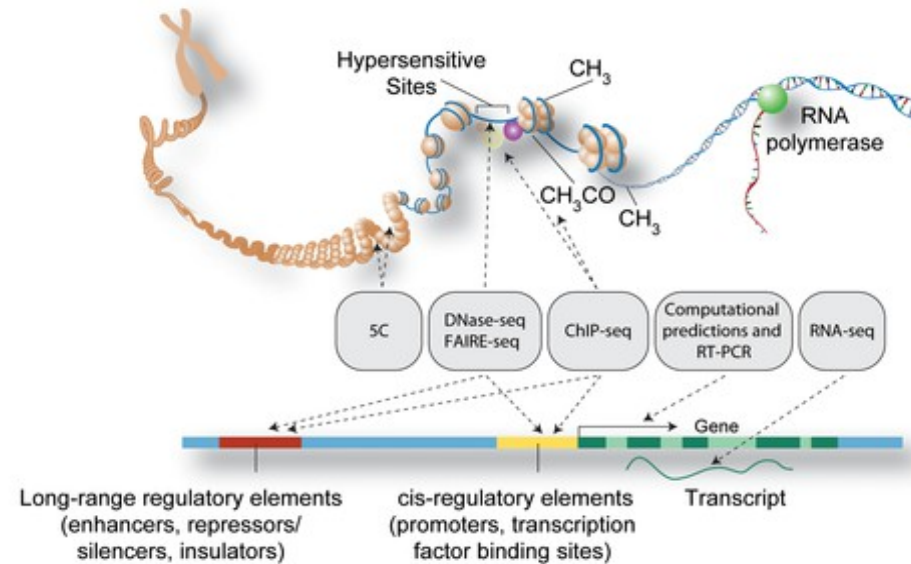
32 institutions



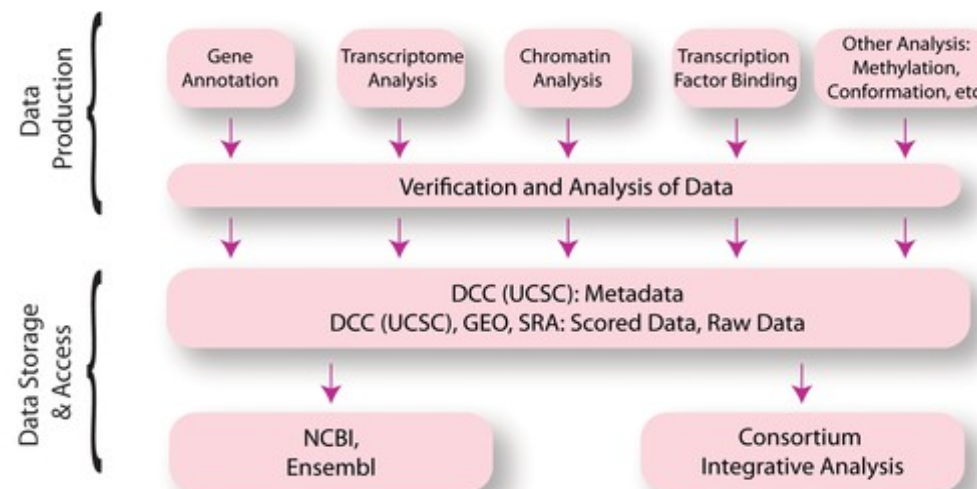
# ENCODE

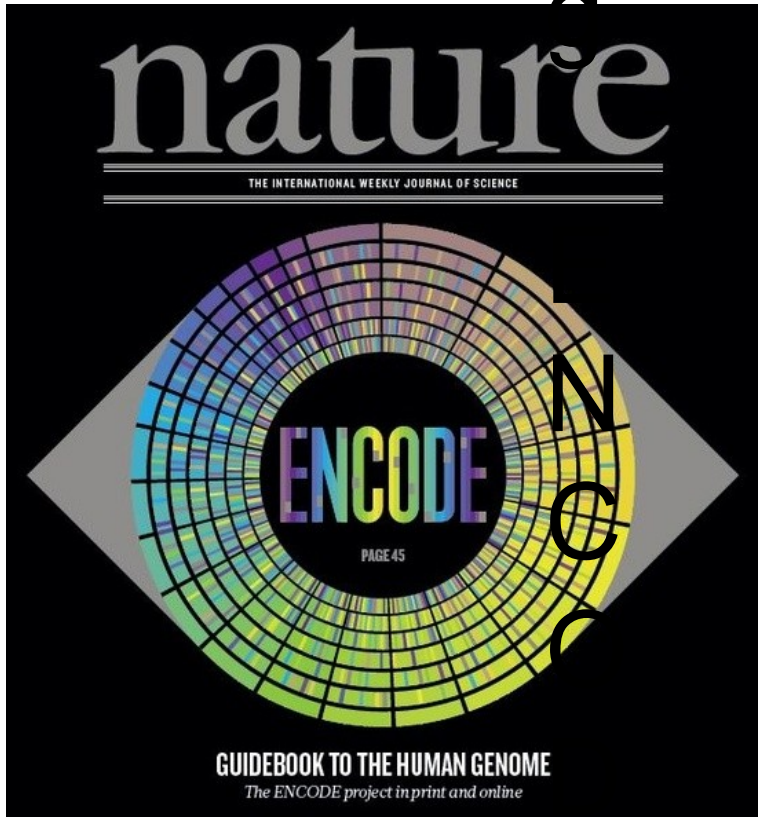
## The consortium workflow

A.



B.





## CONTENTS

### FEATURE

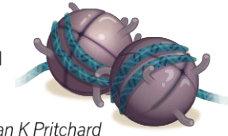
- 46 **The human encyclopaedia**  
*Brendan Maher*

### COMMENT

- 49 **Lessons for big-data projects**  
*Ewan Birney*

### NEWS & VIEWS

- 52 **ENCODE explained**  
*Joseph R Ecker;*  
*Wendy A Bickmore;*  
*Inês Barroso; Jonathan K Pritchard*  
& *Yoav Gilad; Eran Segal*



### ARTICLES

- 57 **An integrated encyclopedia of DNA elements in the human genome**  
*The ENCODE Project Consortium*
- 75 **The accessible chromatin landscape of the human genome**  
*R E Thurman et al.*
- 83 **An expansive human regulatory lexicon encoded in transcription factor footprints**  
*S Neph et al.*
- 91 **Architecture of the human regulatory network derived from ENCODE data**  
*M B Gerstein et al.*
- 101 **Landscape of transcription in human cells**  
*S Djebali et al.*

### LETTER

- 109 **The long-range interaction landscape of gene promoters**  
*A Sanyal et al.*

**MORE ONLINE**

#### NATURE ENCODE EXPLORER

Nature ENCODE Explorer offers you a way to explore the wealth of data across all 30 ENCODE papers. By linking relevant paragraphs, figures and tables from the papers, the 'threads' allow you to examine different themes

[nature.com/encode](http://nature.com/encode)



GENOME RESEARCH

Genome Biology

BMC Genetics

Available on the App Store

The free Nature ENCODE app for the iPad features all 30 papers plus videos and comment



**Accessing through ENCODE:**

**Go to the encode webpage ( <https://www.encodeproject.org/> ) and explore the experiment matrix**

**How many DNaseSeq are available for the GM12878 cell lines?**

**Which files can you download?**