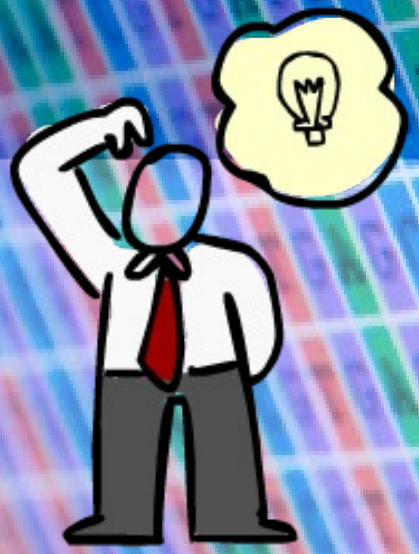


```
@M01264:15:000000000-D0BG0:1:1101:15111:1933 1:N:0:1
CACTTCAACCTCTGCCTCCCAGGTTCAAGTGATTCTCCTGCC
+
>11>AF@D@1B1B11BAFFE1CAFGFGGHGHGHHHHHHGHHHHH
@M01264:15:000000000-D0BG0:1:1101:18980:2056 1:N:0:1
GGAAACACCTTTCAACTGGAAGATATCATTAAACACAGG
+
AAAAAFFF?FFFGBFGFGGEGGHHHHHHHHHHHHHHHHH
```

NGS: from the design of the experiment to sequenced reads

Summer school Malta 13 September 2016

Train MALTA Horizon 2020

UNIVERSITY OF CAMBRIDGE

Experimental design

- The design of the experiment is the f irst 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 underling population you need statistics.
- Statistics needs biological replicates.

Experimental design

1)COVERAGE: How many reads do we need?

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

R_{length} = length in nucleotides of the reads (2XPE)
R_{num} = number of sequenced reads (2XPE)
A_{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.

Sims et al. Nature Reviews Genetics 15, 121–132 (2014)

Experimental design

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%

Experimental design

3)Read length: long or short reads?

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

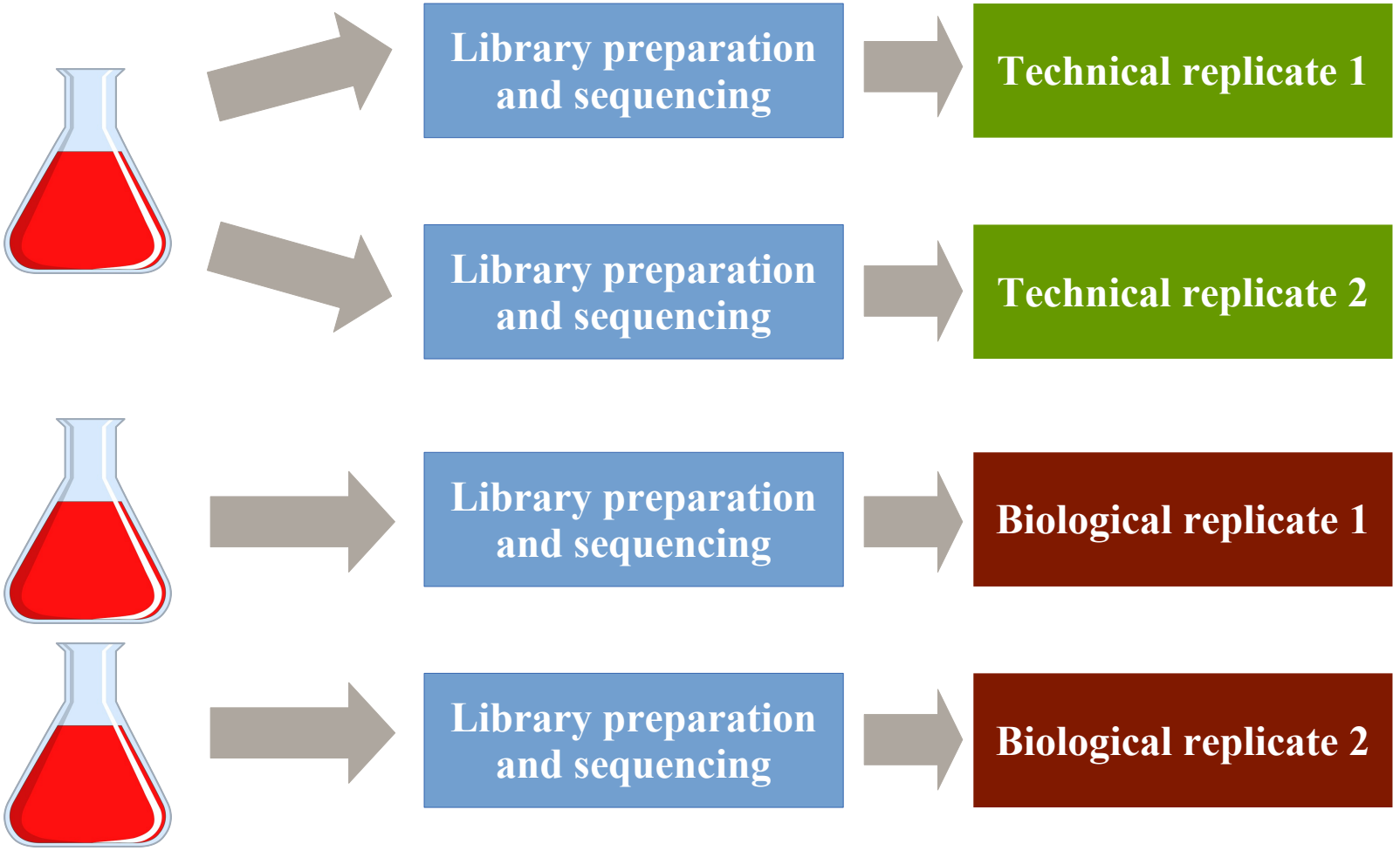
- GENOME RESEQUENCING
- De novo TRANSCRIPTOME
- TRANSCRIPTOME seq
- ChIP seq

Read length is invertionally 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.

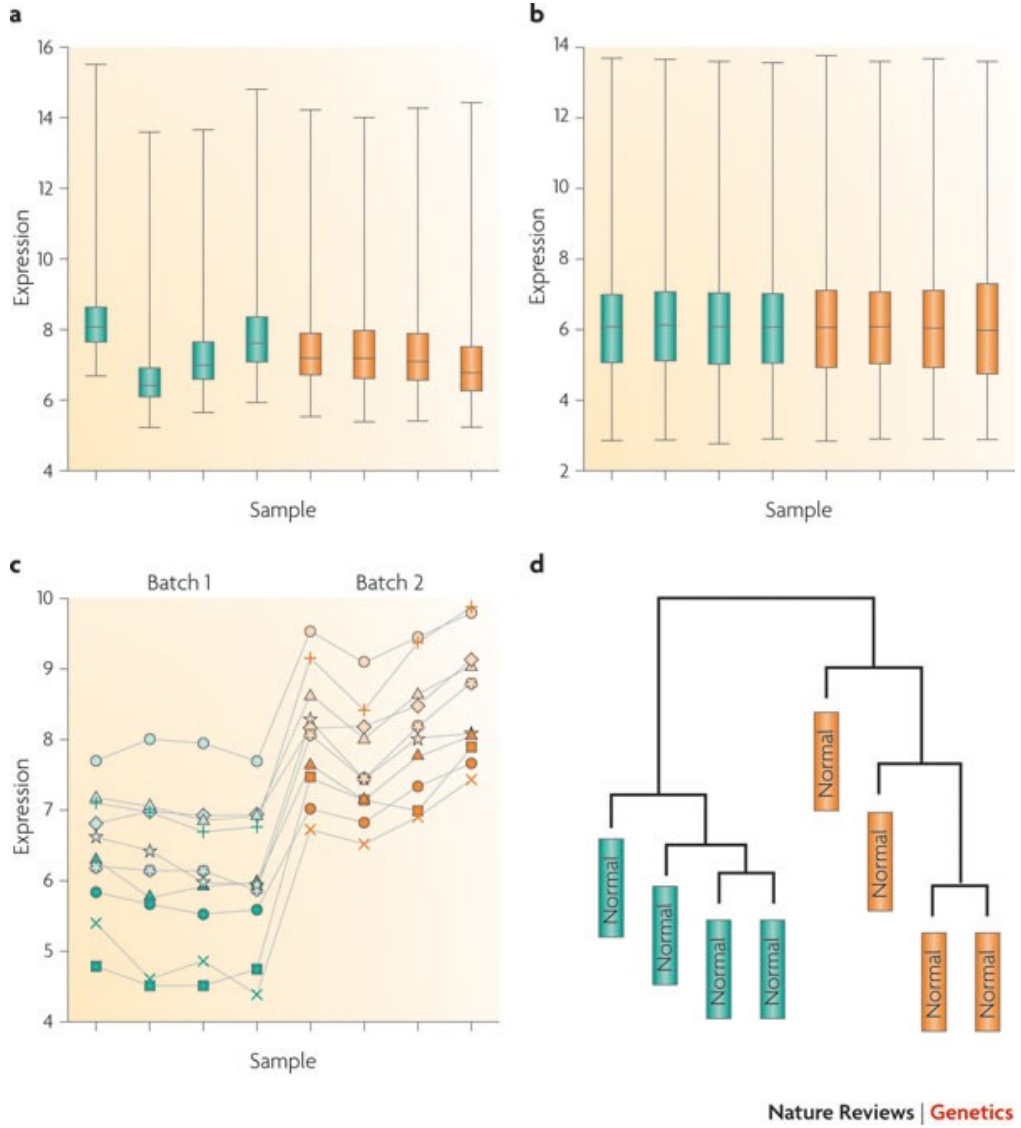
Replicates



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

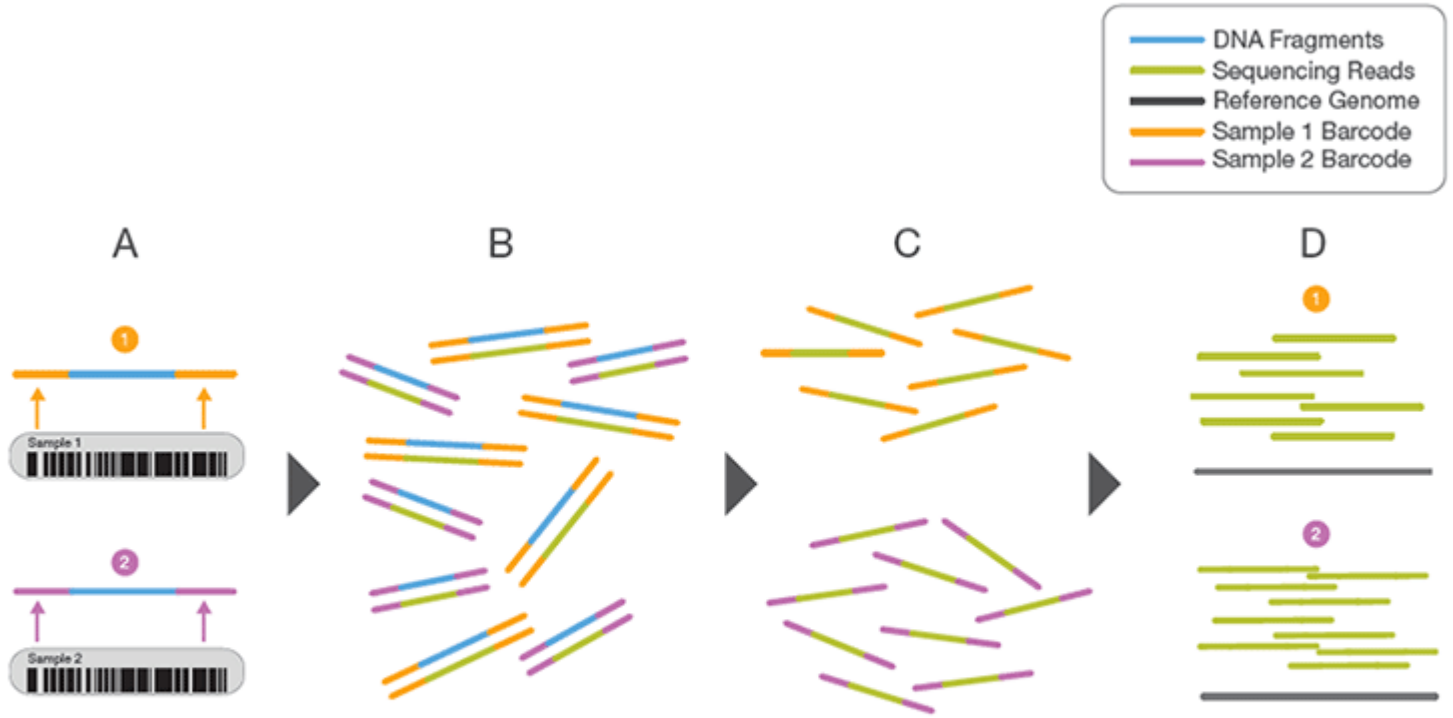
Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study.



Leek et al. Nature Reviews Genetics 11, 733-739 (October 2010) | doi:10.1038/nrg2825

Multiplexing the samples can contrast batch effects

Figure 2: Conceptual Overview of Sample Multiplexing



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

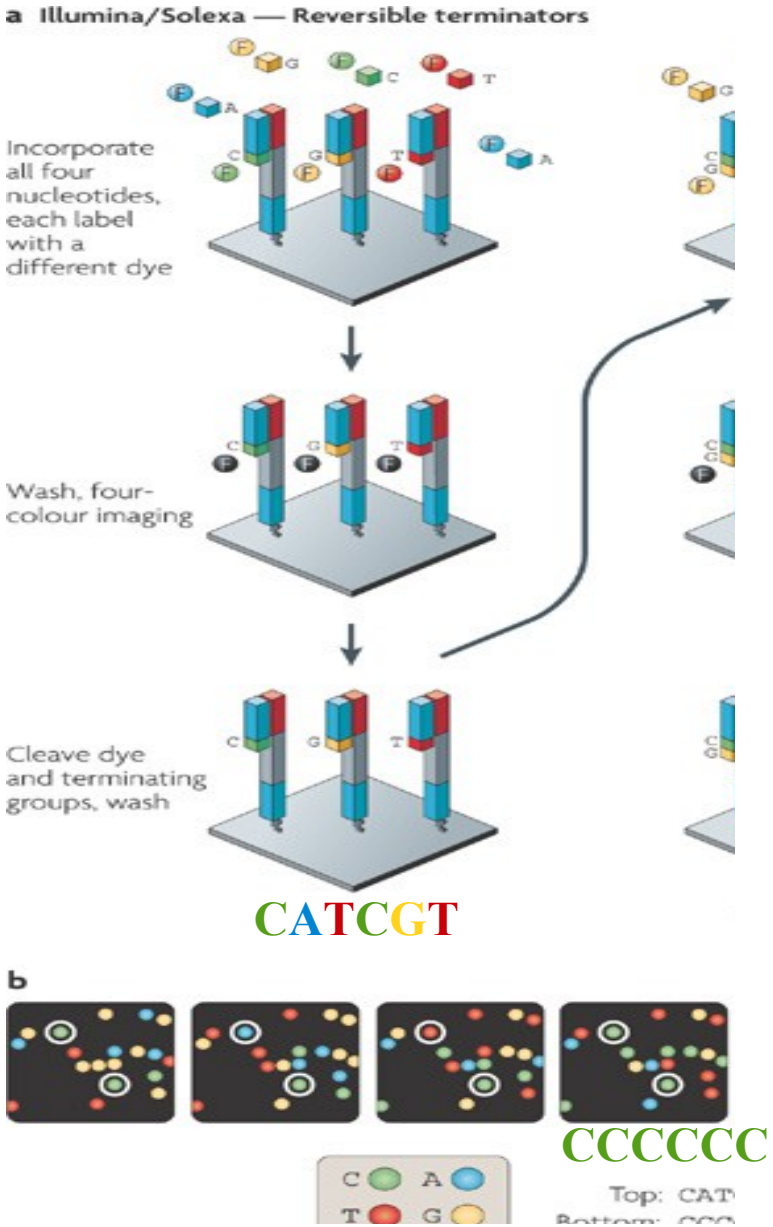
Question

Which technology?

Sequencing Technologies

illumina®

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



Michael L. Metzker; Nature Reviews Genetics 11, 31-46 (January 2010)

Sequencing Technologies

illumina®

	Focused power.	Flexible power.	Production power.	Population power.
				
	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series
	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.	Production-scale genome, exome, transcriptome sequencing, and more.	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

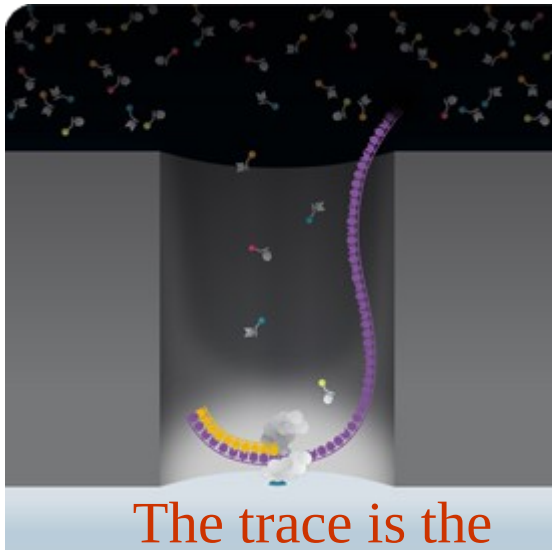
Ion PGM

	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

PACIFIC BIOSCIENCES™

SMRT TECHNOLOGY



The trace is the f luorophore incorporated into the 4 nucleotides

PacBio RS II

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

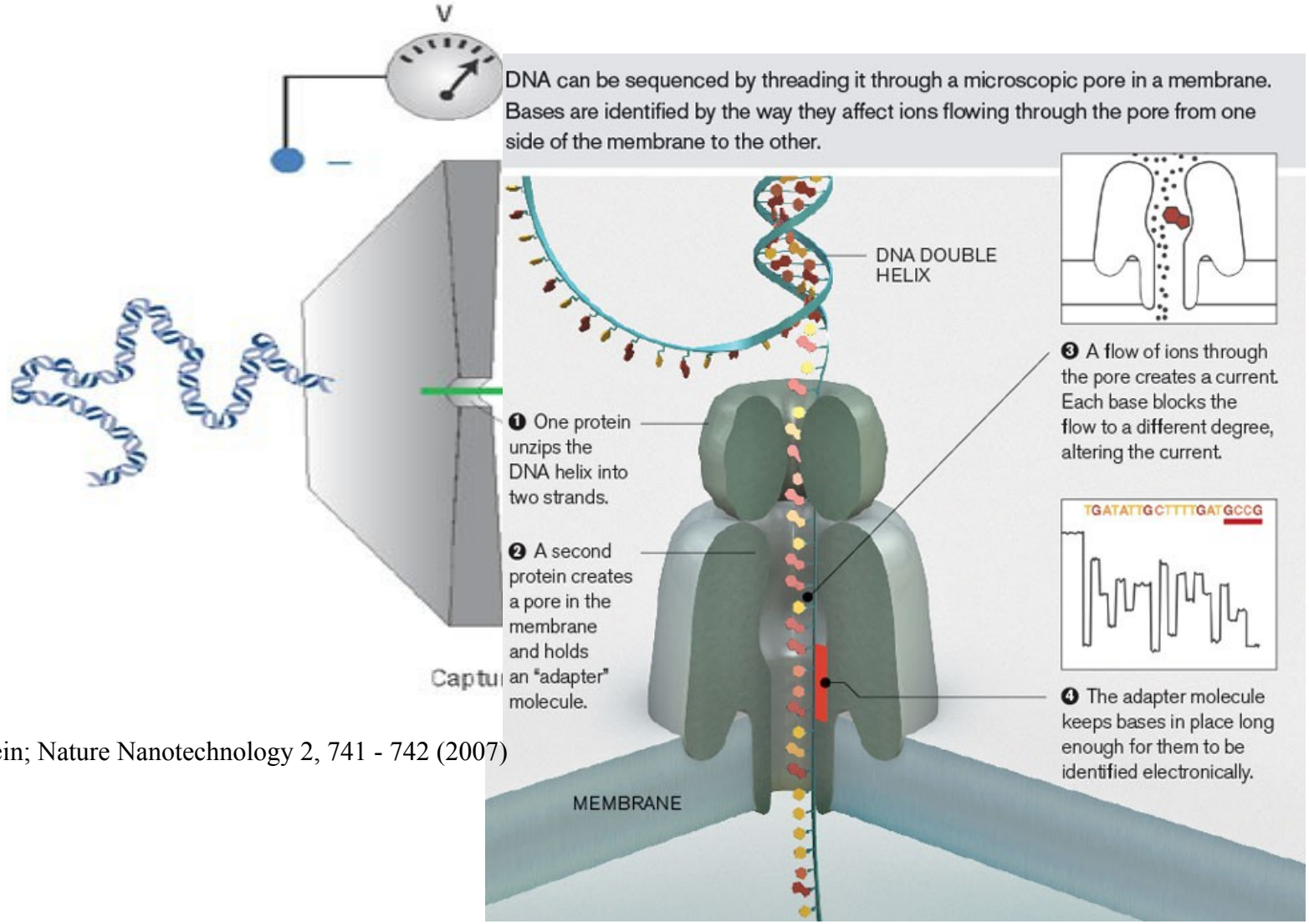
Read length 8.5 kb, up to 18 kb

It can detect DNA modifications (Methylations)

Sequencing technologies

Oxford NANOPORE Technologies

miniION

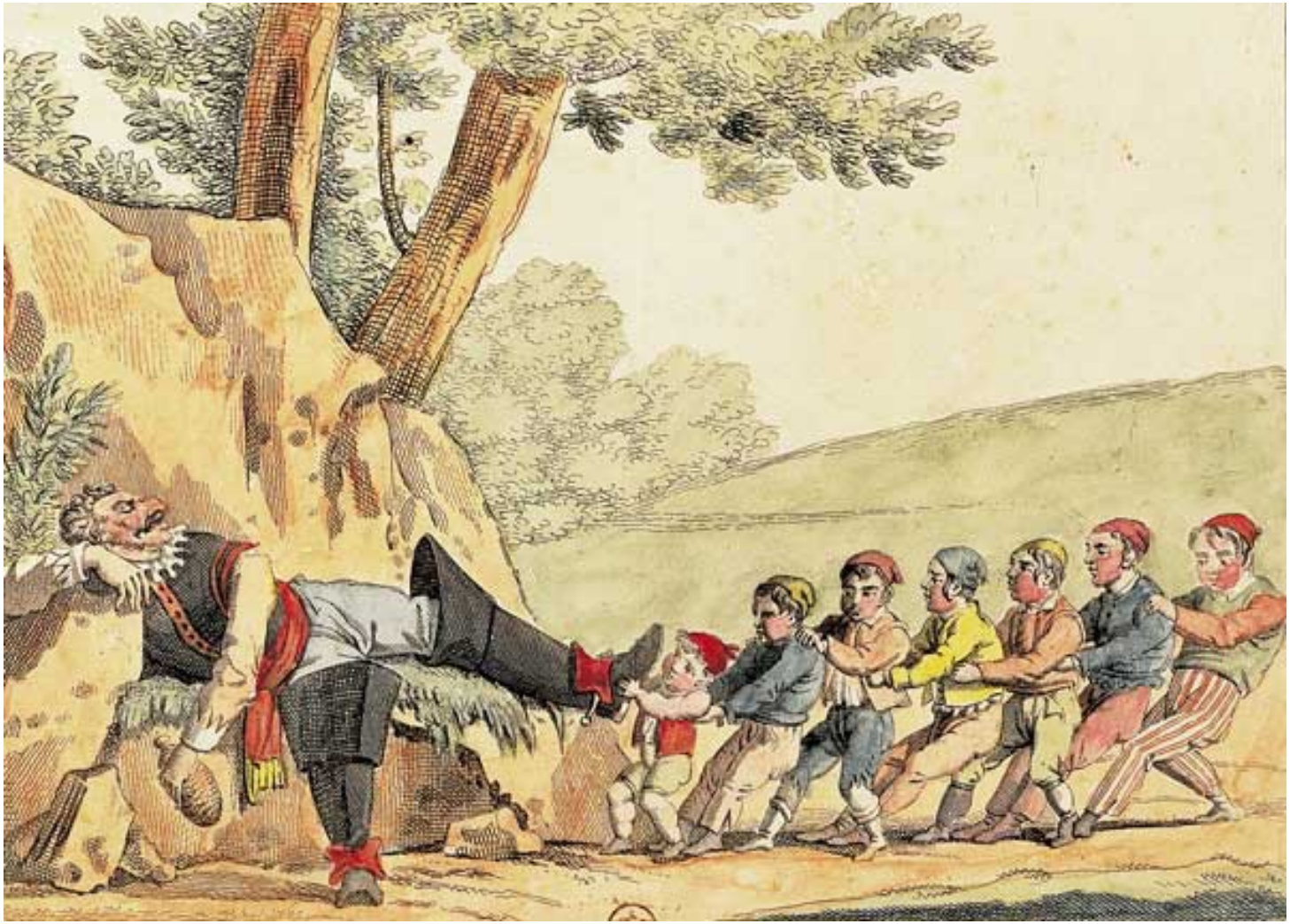


Stein; Nature Nanotechnology 2, 741 - 742 (2007)

Sequencing Technologies

	Read length	Accuracy	Reads per run	Time per run	Advantages	Disadvantages
Ion semiconductor (Ion Torrent sequencing)	up to 400 b	98%	up to 80 million	2 hours	Less expensive equipment. Fast.	Homopolymer errors.
Single-molecule real-time sequencing (Pacific Bio)	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.
Sequencing by synthesis (Illumina)	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.

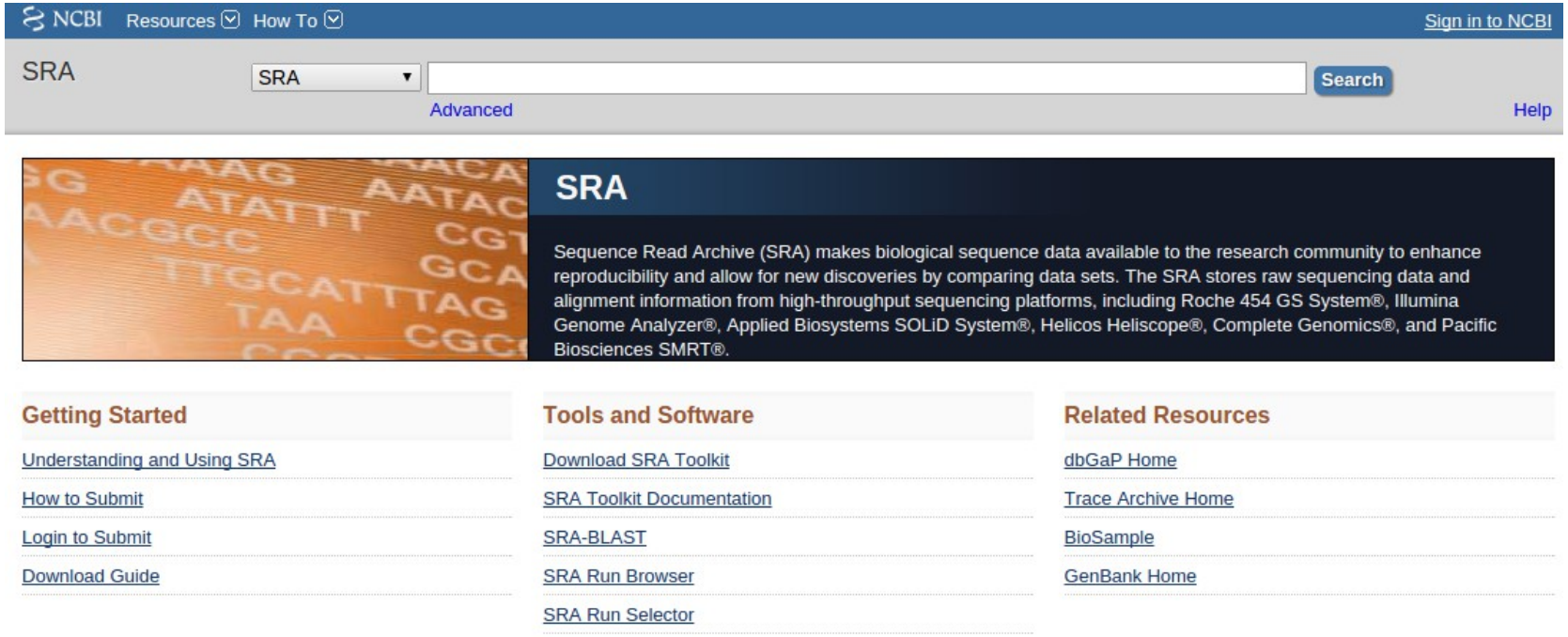
Together is better!!!!!!!



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

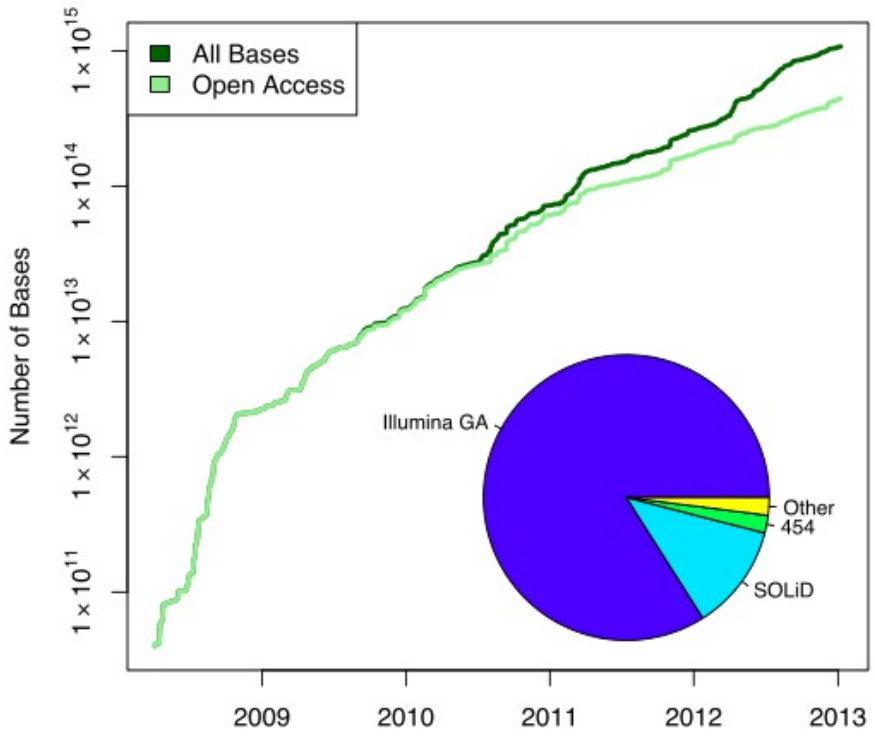
NCBI Sequencing Read Archive

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



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

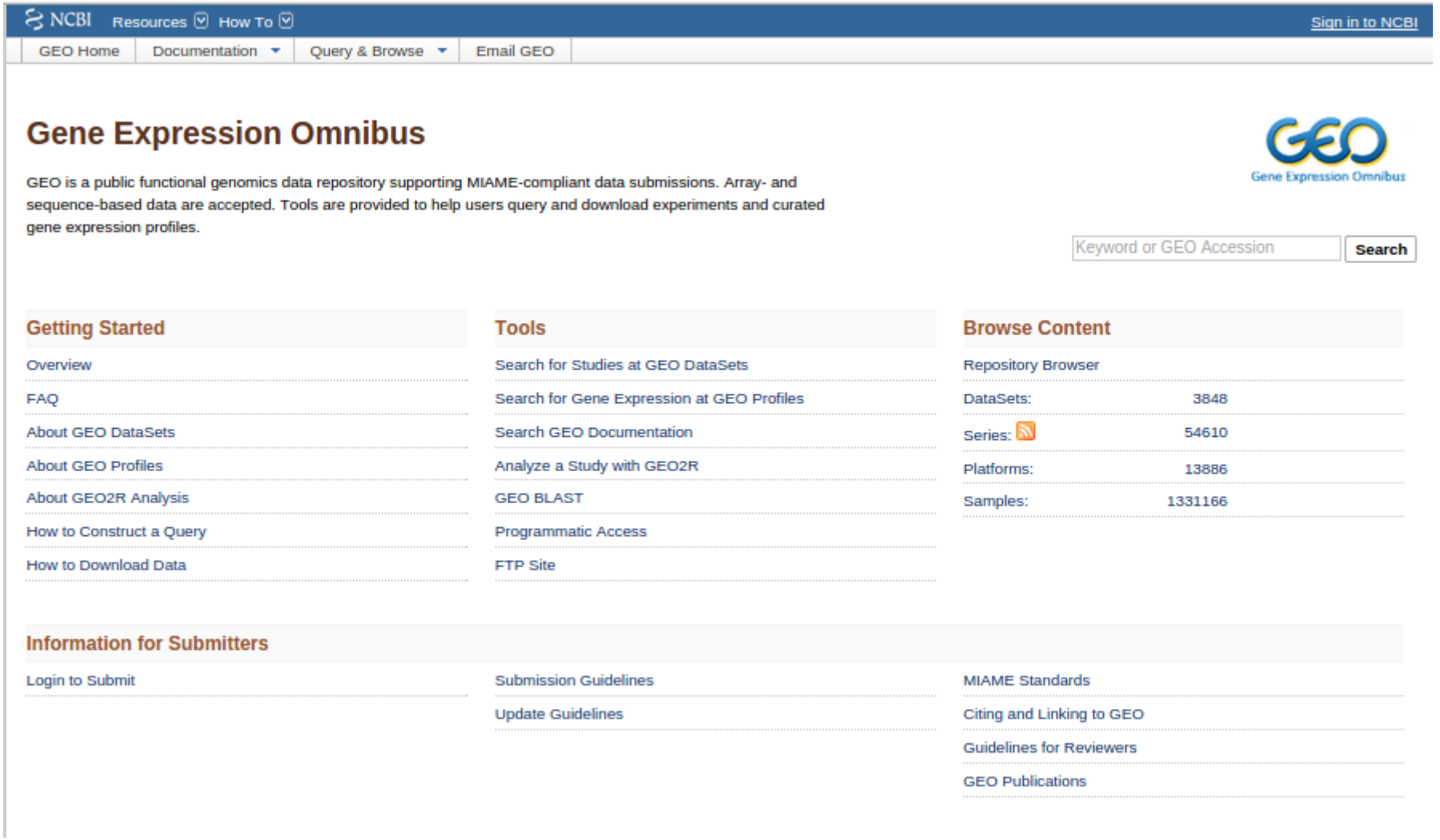
Numbers to 01/02/2015

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/



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...)

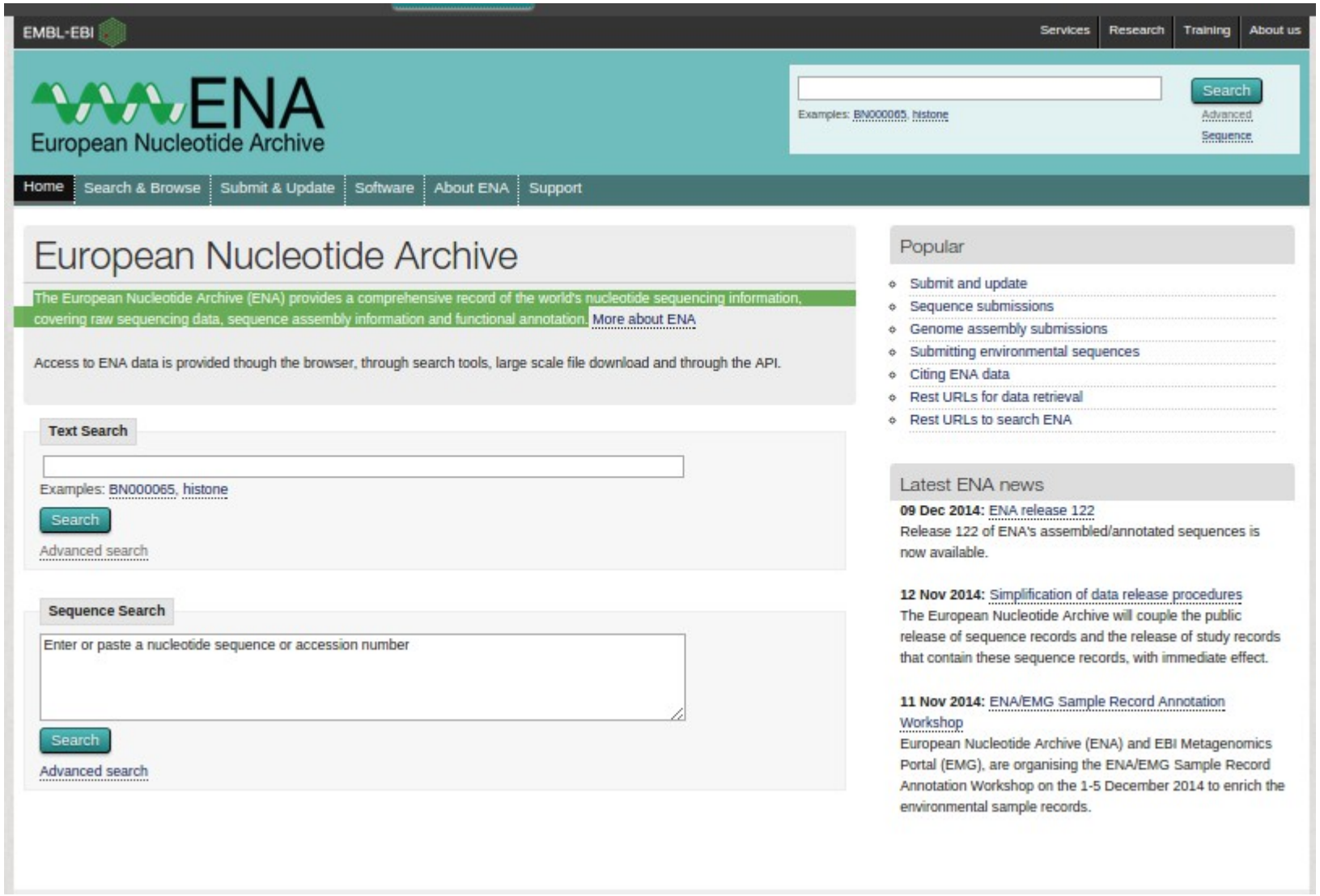
Numbers to 01/02/2015

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 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 ofall sequence-based functional elements)

http://www.genome.gov/10005107http://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 keycellular 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)

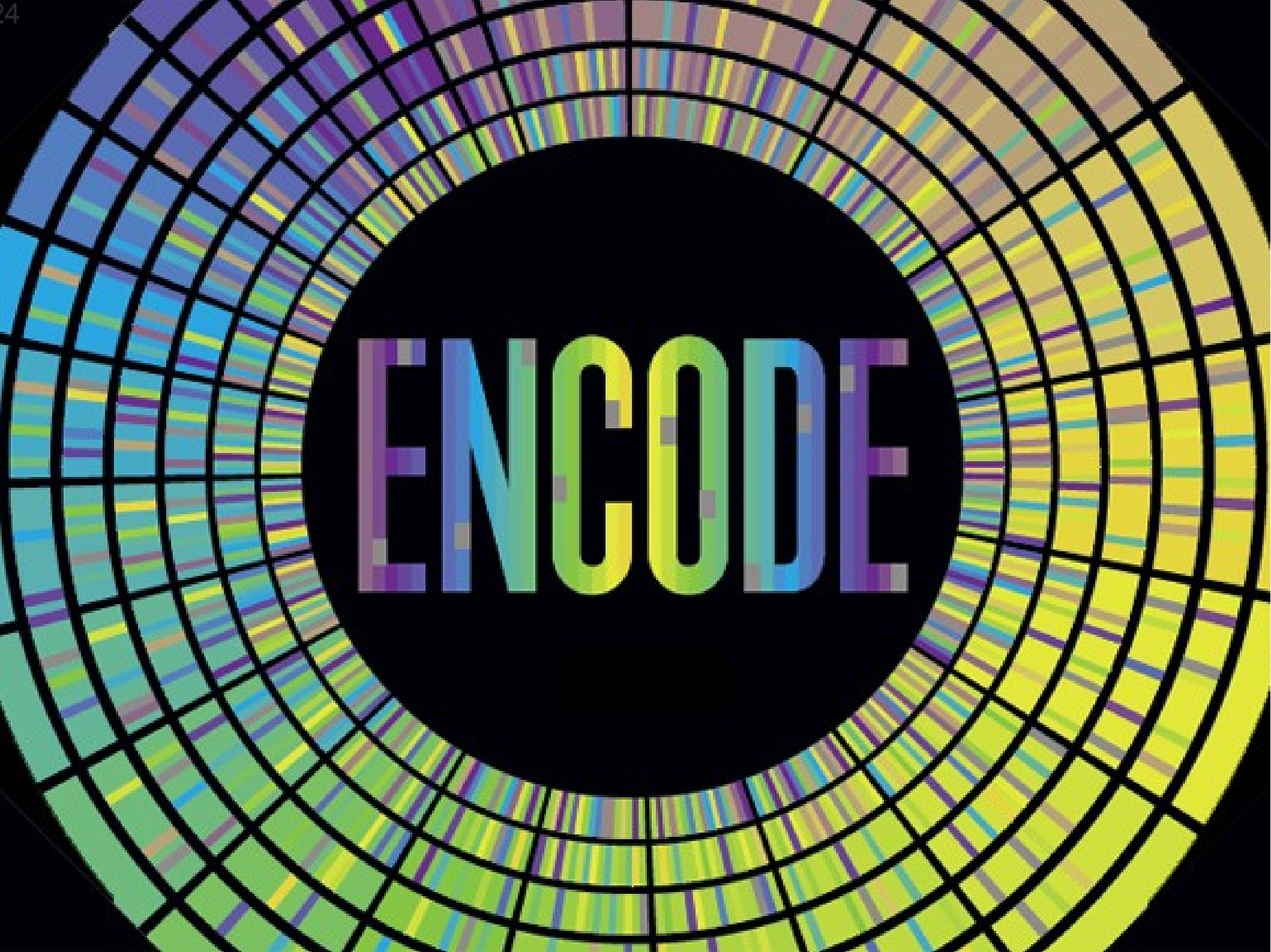
http://fantom.gsc.riken.jp/http://fantom.gsc.riken.jp/5/sstar/Data_source

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* *

ENCODE

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

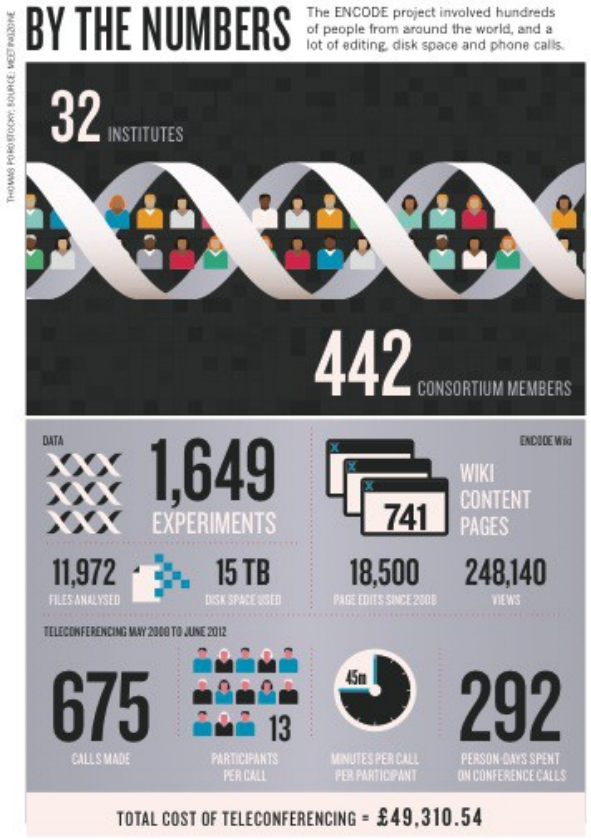
+1,600 experiments

147 cell types

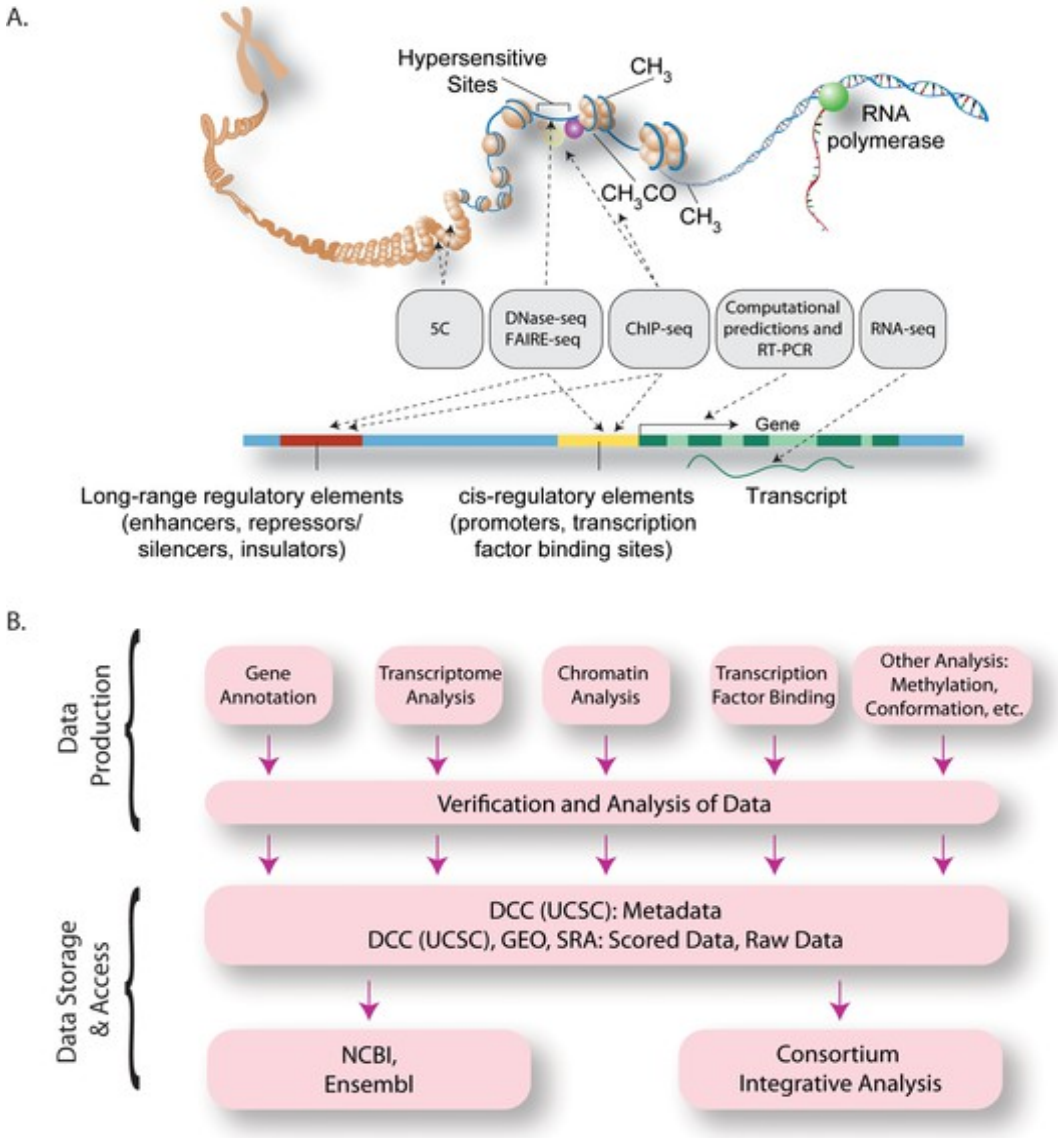
225 antibodies

~450 authors from

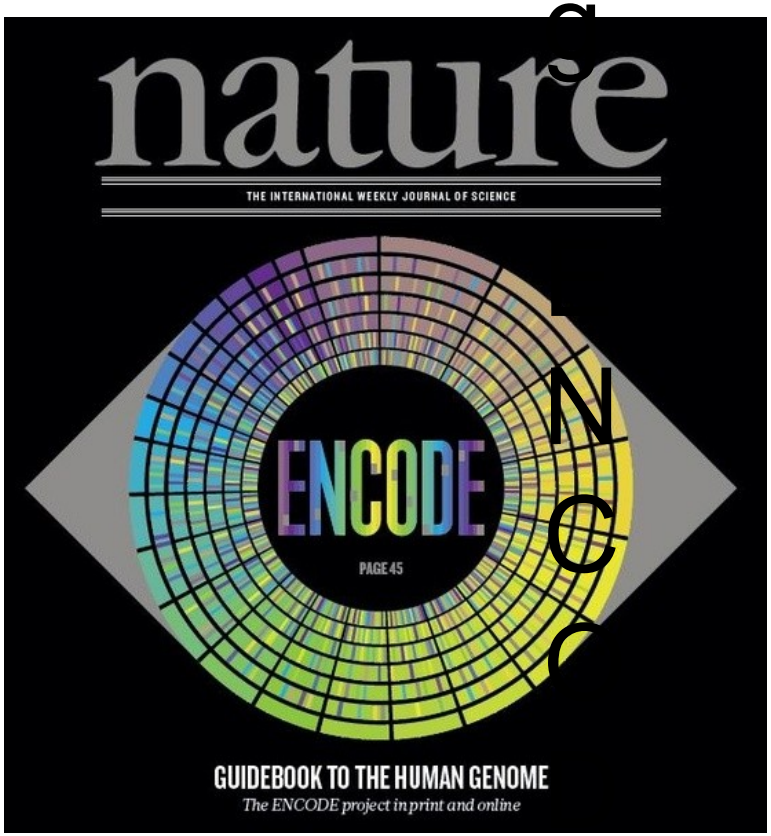
32 institutions



The consortium workflow

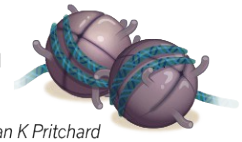


ENCODE



ing ENCODE

CONTENTS



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

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

ENCODE

Accessing through ENCODE:

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

How many DNASeq are available for the GM12878 cell lines?

Which files can you download?