Introduction to next-generation sequencing

Simon van Heeringen
November 3, 2014

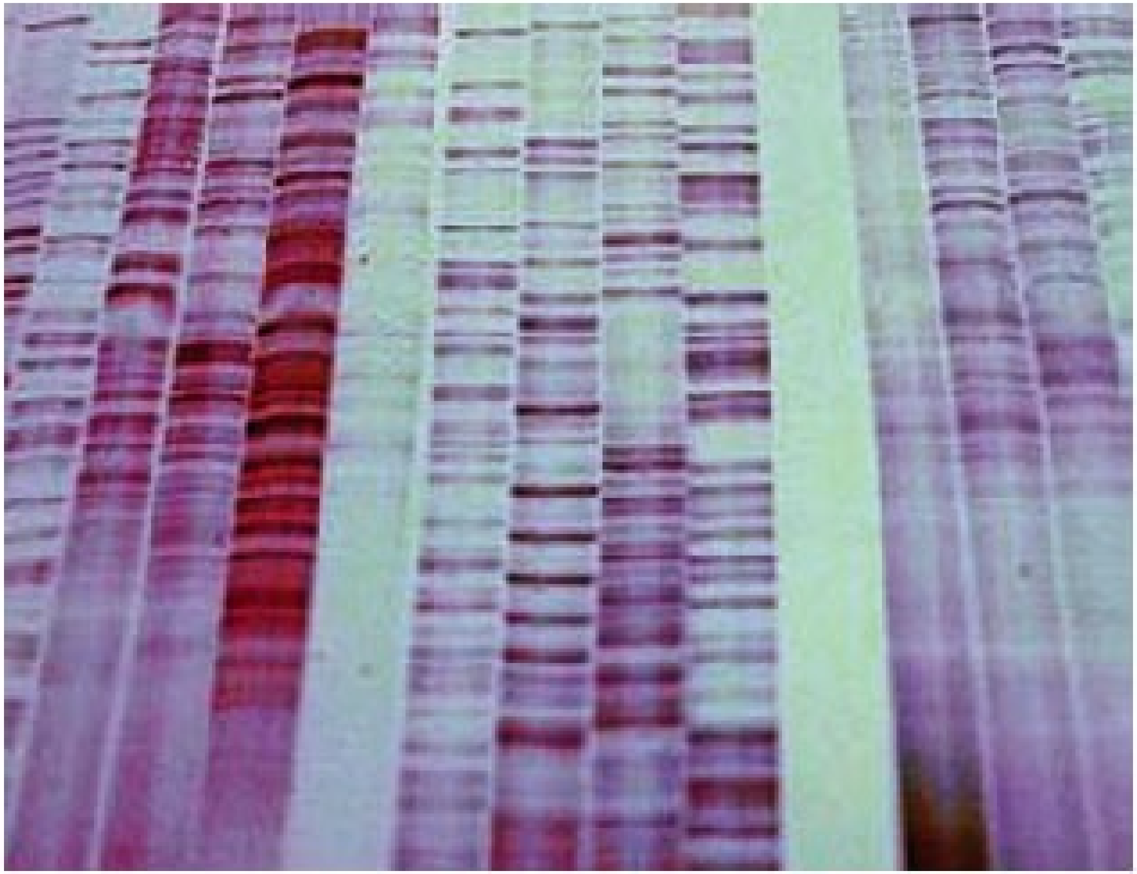
So, you want to do sequencing...?

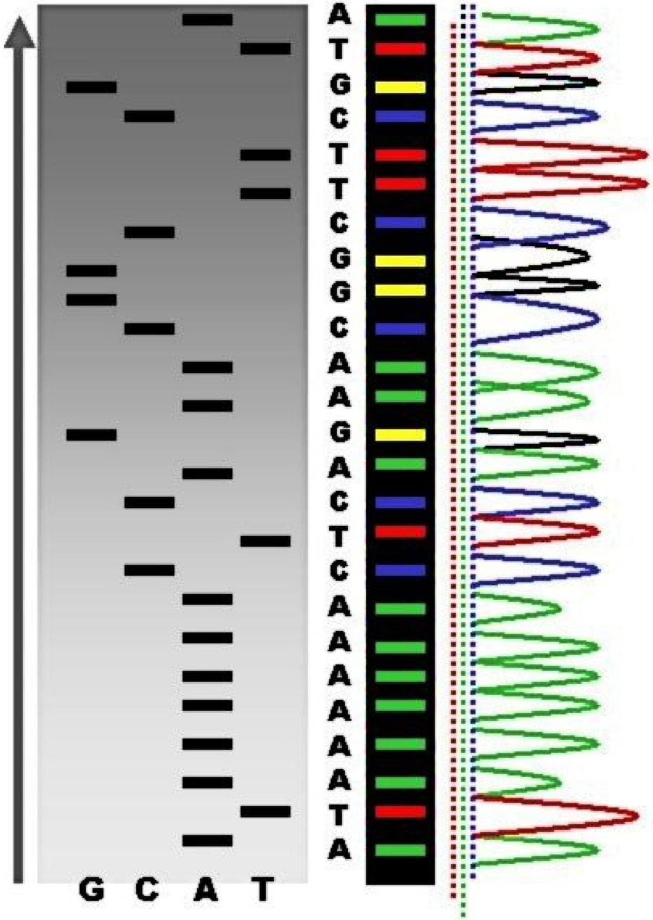
3'Seq 3-seq 3P-seq AHT-ChIP-seq ARS-seq ATAC-seq BOINC-seq BS-seq Bar-seq BisChIP-seq Bru-seq Bubble-seq CAB-seq CAGE-seq CHART-seq CLASH-seq CNV-seq CRE-seq Capture-C-seq Cel-seq ChIA-PET-seq ChIP-seq ChIRP-seq Chem-seq Chip-exo-seq Cir-seq DMS-seq DNAse-seq DNAseI-seq Dup-seq FAIRE-seq FRAG-seq FRT-seq Frac-seq Freq-seq GRO-seq GTI-seq HELP-seq HITS-KIN-seq Hi-C-seq HiTS-Flip-seq IMS-MDA-seq IN-seq Ig-seq Immuno-seq MeDIP-seq Methyl-seq Mu-seq NET-seq NOMe-seq Nascent-seq Novel-seq Nucleo-seq PAL-seq PAR-Clip-seq PARS-seq

http://liorpachter.wordpress.com/seq

0010110011101101000011101000010000000101000011000110010100100100000000110010001100000011000000111001001011010011001000110000001100
0110111101101110001000000111011001100001011011100010000001001000011001010110010101110010011010010110111001100111011001010110111
1000010110111001101000011001010110010101110010011010010110111001100111011001010110111001000000011011100110001101101101011011000

Sanger sequencing



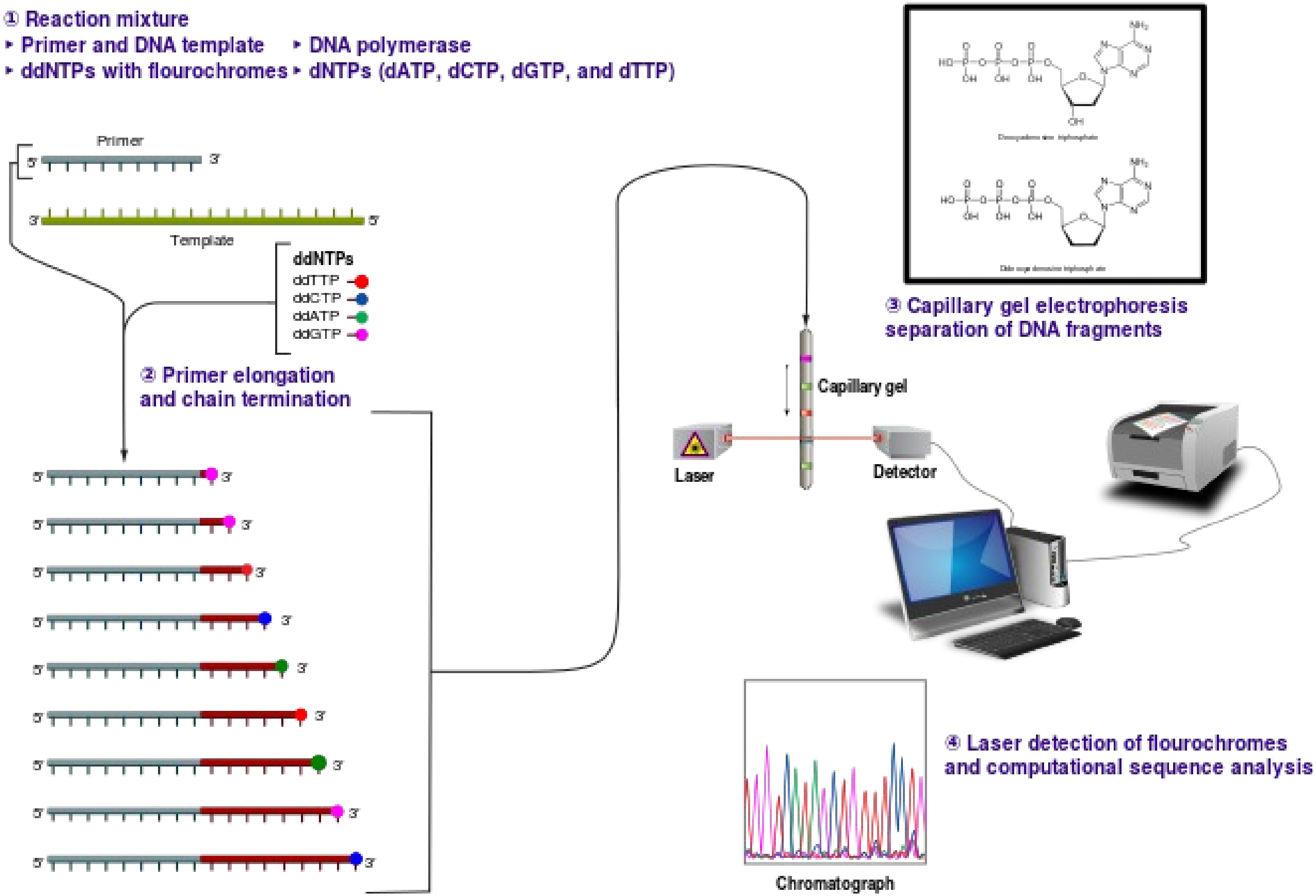


Nature Methods, 2008

Wikipedia

CAAAATACTGATATATACAACATGAACGAATGTCAGACAGTACATTGAAGGACAGAAGCCCGACAAAAATGAGCACATAATGTATGATTCCCC
GTGTAGTGGCACGATCTTGGCTCACTGCAACCTCTGCCTCCCGGGTTCAAGCGATTCTCCTGCCTCACCCTCCCGAATAGCTGGGATTACAG
GGGGATTCACCACGTTGGCCACGCTGGTCTGGAACTCCTATCCTCAAGTAATCCGCCCGCCTCGGCCTCCCAAAGTGCAGGCGTGAGCCAC
AAAATGGTTATGGAGATCAAAATAAAGGTGGGGTCGGGAATCGACTGGGAAGAGACGTGATGAAACGTTTCTGGGACGATGAAAAGGGTCTC

Sanger sequencing



① Reaction mixture

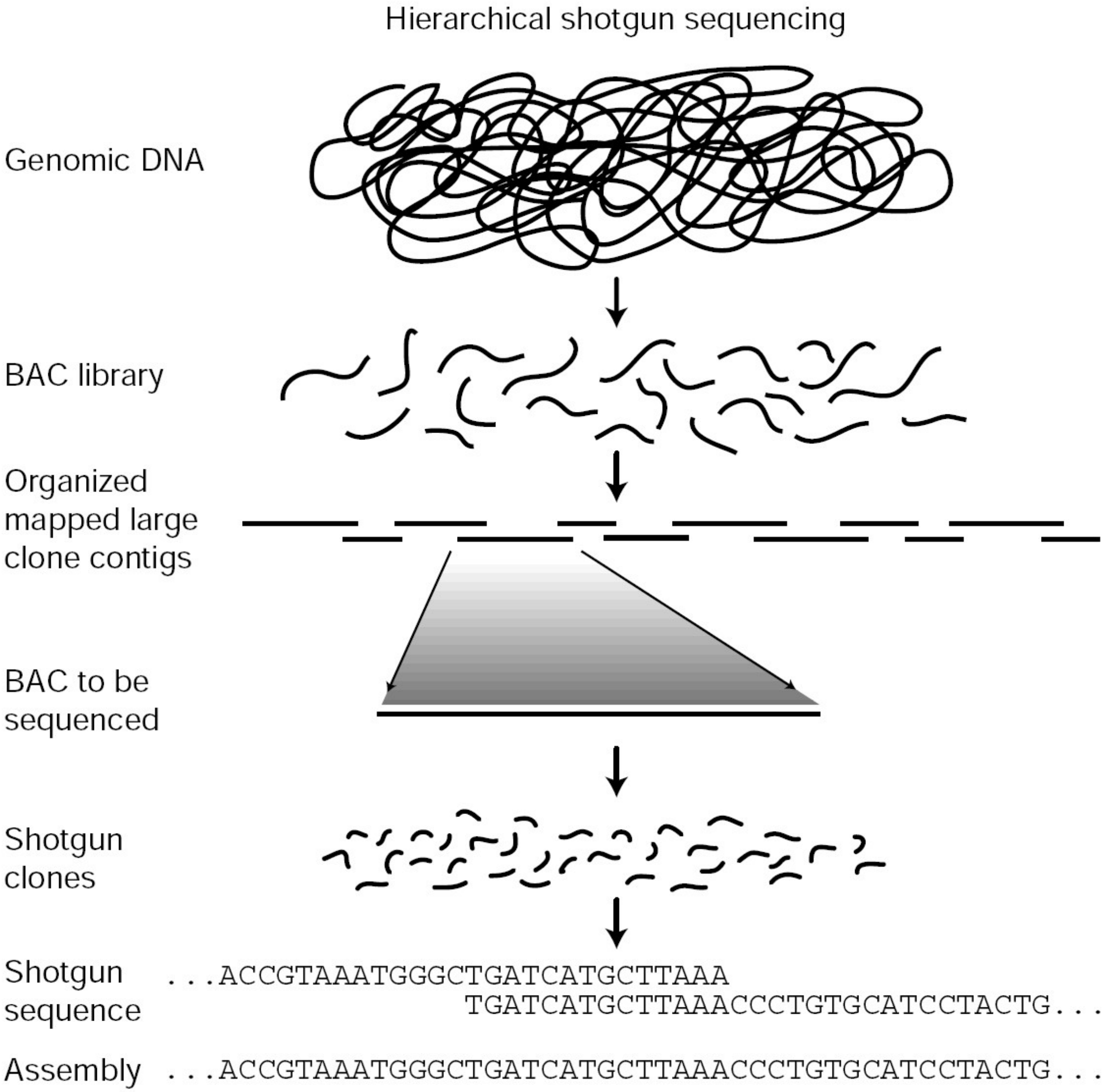
- Primer and DNA template
- ddNTPs with flourochromes
- DNA polymerase
- dNTPs (dATP, dCTP, dGTP, and dTTP)

② Primer elongation and chain termination

③ Capillary gel electrophoresis separation of DNA fragments

④ Laser detection of flourochromes and computational sequence analysis

Shotgun sequencing


Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Nature, 2001

2001: draft of the human genome

“Next-generation” sequencing

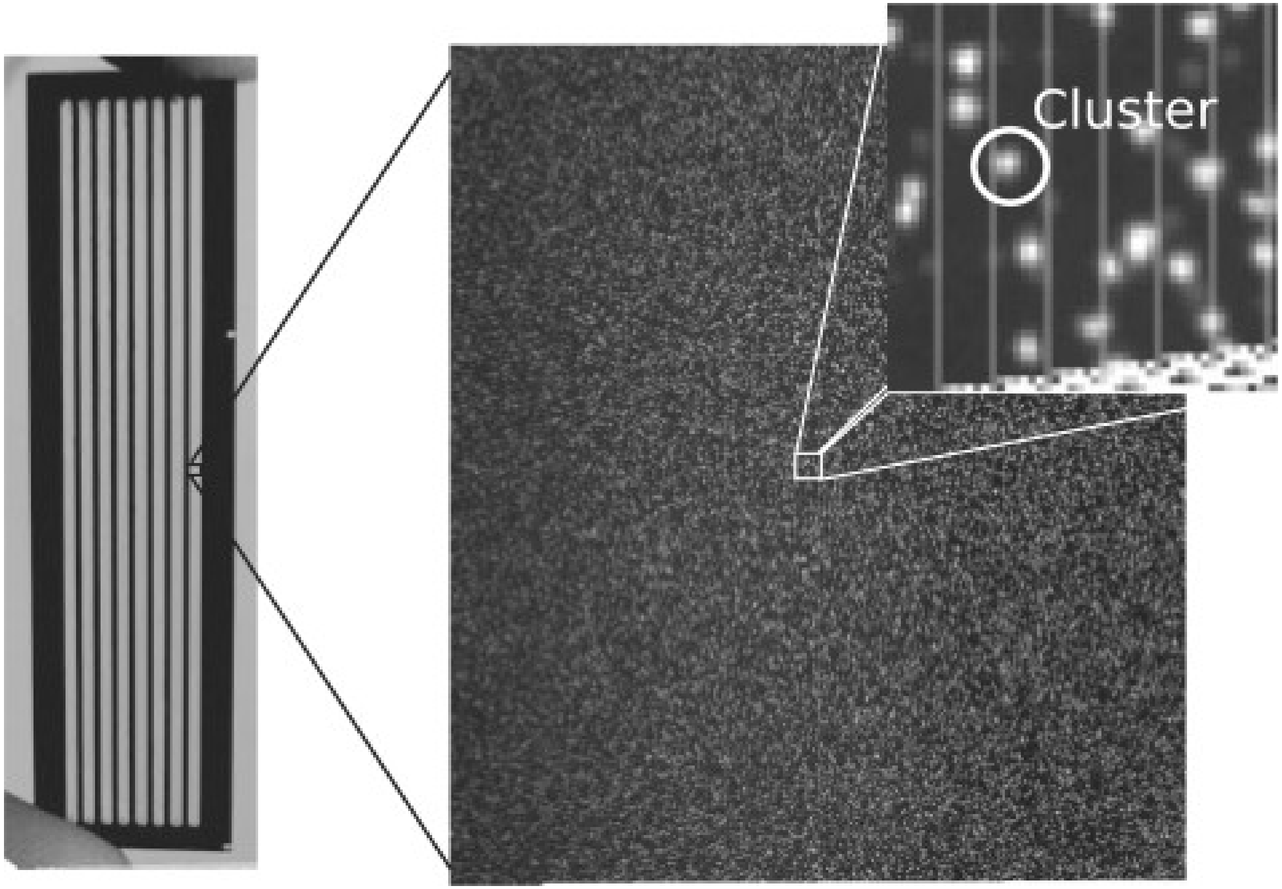
“Next-generation” sequencing

or

.. a whole lotta reads..

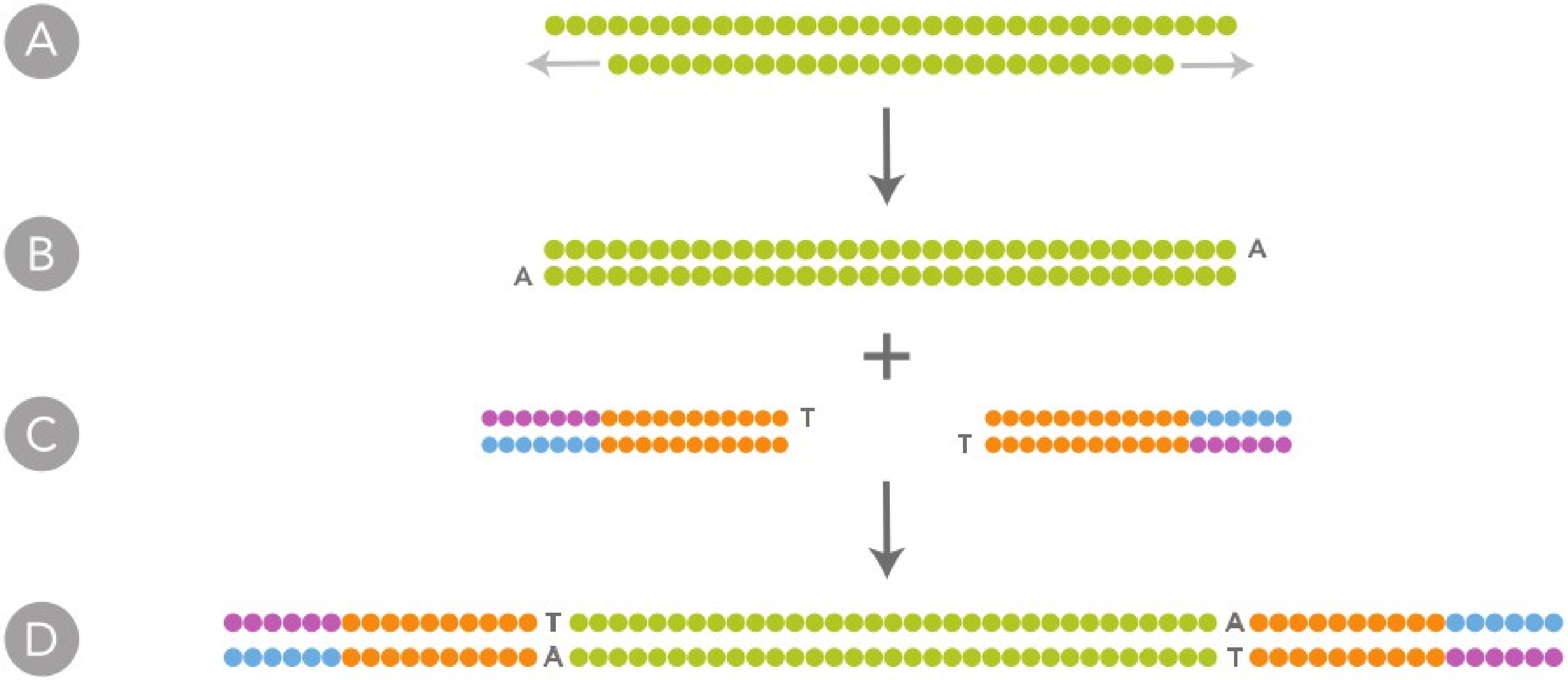
Illumina sequencing

An Illumina flowcell



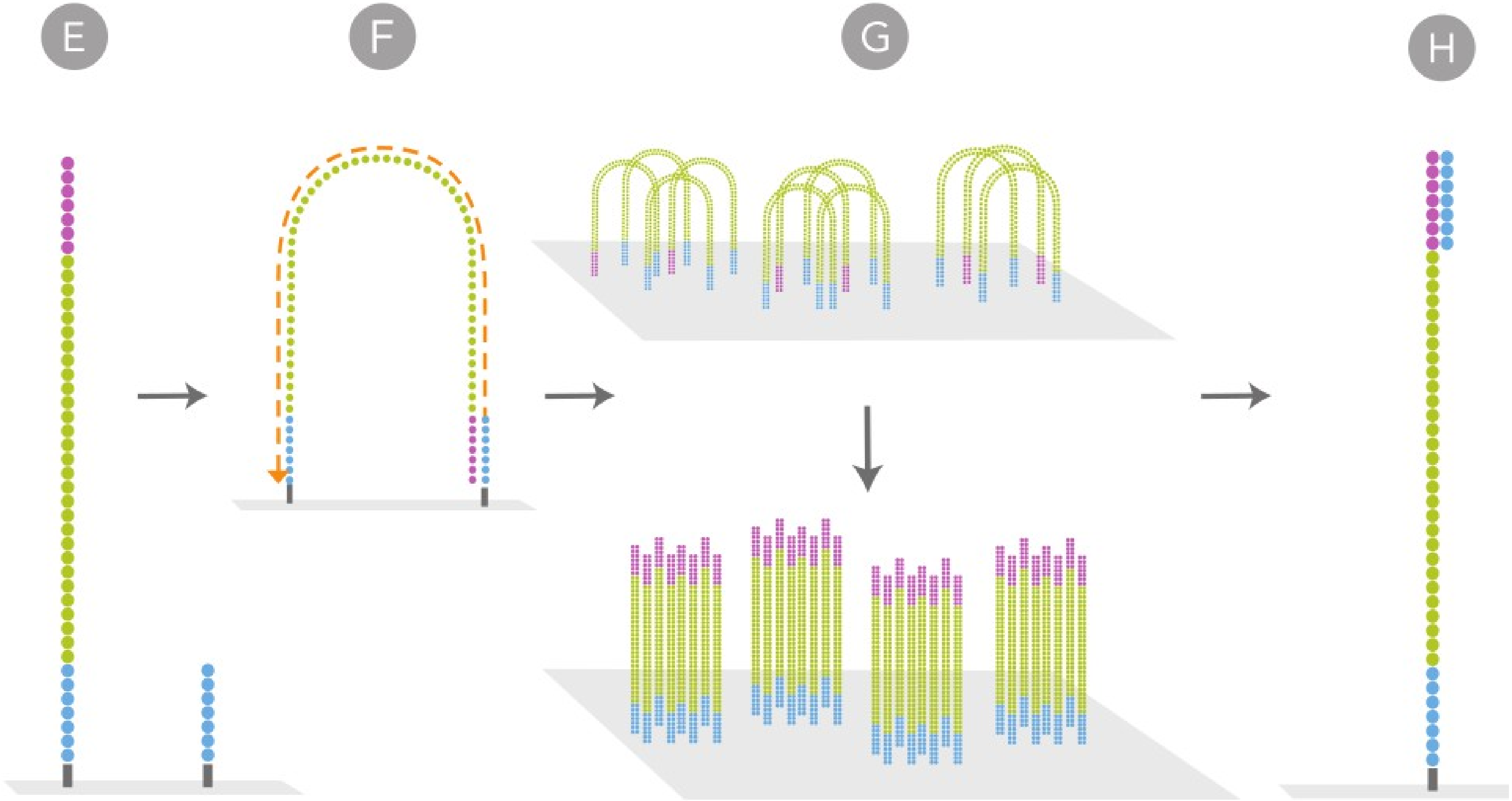
Whiteford et al., Bioinformatics, 2009

Sequencing by synthesis



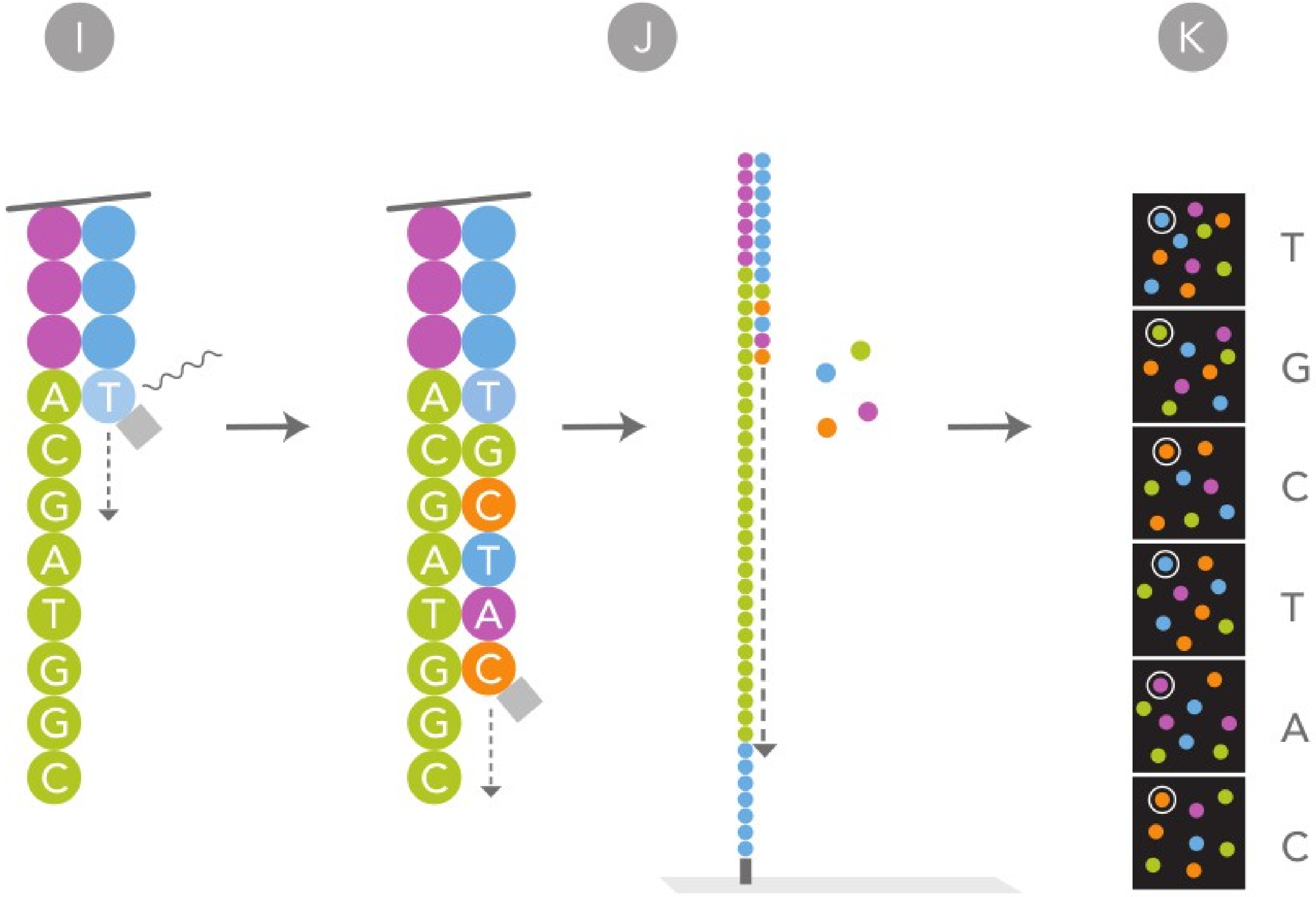
00101100111011010000111010000100000001010000110001100101001001000000011001000110000001100000011100100101101001100100011000000110
01101111011011100010000001110110011000010110111000100000010010000110010101100101011100100110100101101110011001110110010101101110
100001011011100110100001100101011001010111001001101001011011100110011101100101011011100100000001101110011000110110101101011011000

Sequencing by synthesis

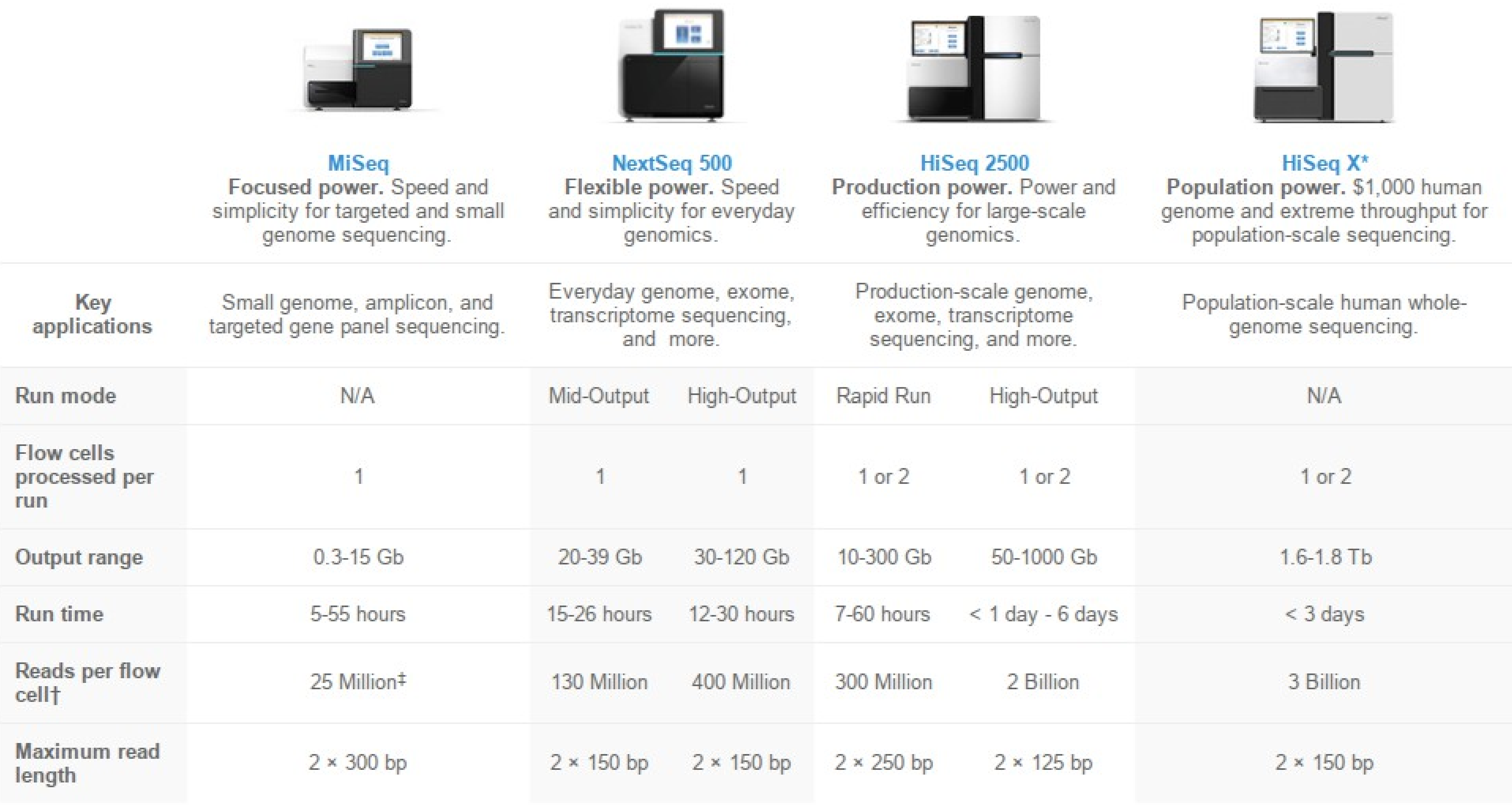


CAAAATACTGATATATACAACATGAACGAATGTCAGACAGTACATTGAAGGACAGAAGCCCGACAAAAATGAGCACATAATGTATGATTCCCC
GTGTAGTGGCACGATCTTGGCTCACTGCAACCTCTGCCTCCCGGGTTCAAGCGATTCTCCTGCCTCACCCTCCCGAATAGCTGGGATTACAG
GGGGATTCACCACGTTGGCCACGCTGGTCTGGAACTCCTATCCTCAAGTAATCCGCCCGCCTCGGCCTCCCAAAGTGCAGGCGTGAGCCAC
AAAATGGTTATGGAGATCAAAATAAAGGTGGGGTCGGGAATCGACTGGGAAGAGACGTGATGAAACGTTTCTGGGACGATGAAAAGGGTCTC

Sequencing by synthesis



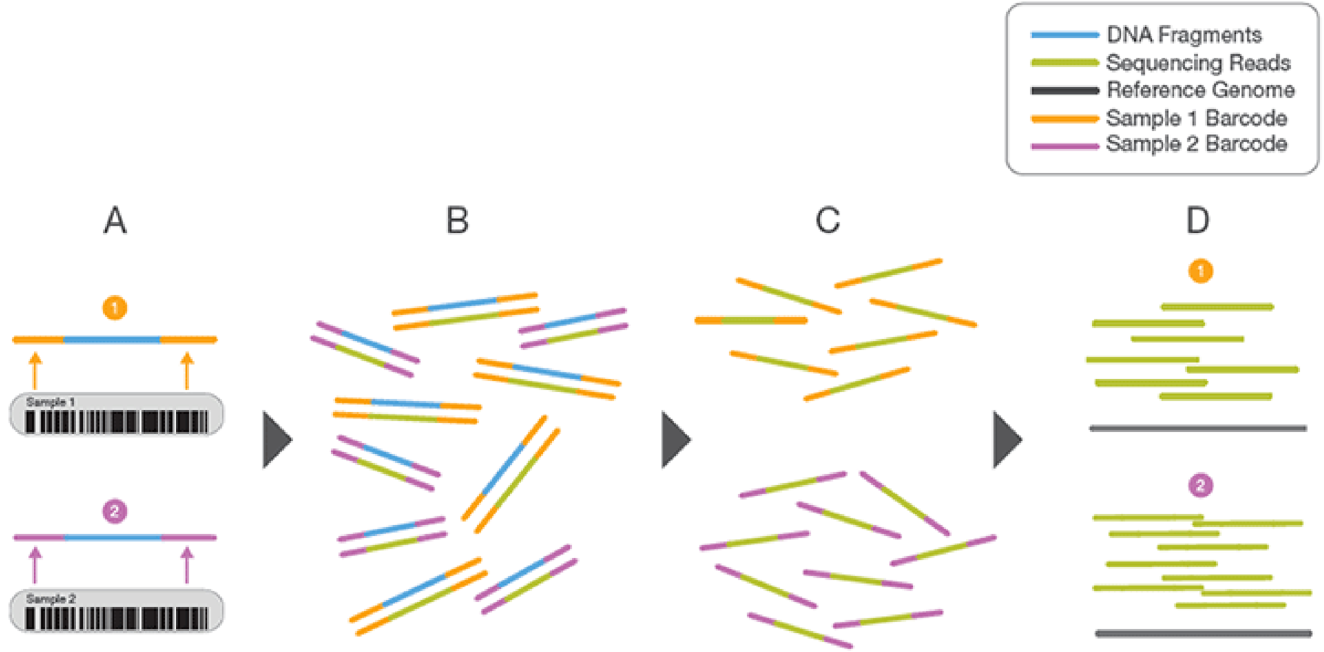
Illumina sequencing systems



	MiSeq **Focused power.** Speed and simplicity for targeted and small genome sequencing.	NextSeq 500 **Flexible power.** Speed and simplicity for everyday genomics.		HiSeq 2500 **Production power.** Power and efficiency for large-scale genomics.		HiSeq X* **Population power.** $1,000 human genome and extreme throughput for population-scale sequencing.
Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
Run mode	N/A	Mid-Output	High-Output	Rapid Run	High-Output	N/A
Flow cells processed per run	1	1	1	1 or 2	1 or 2	1 or 2
Output range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-300 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-55 hours	15-26 hours	12-30 hours	7-60 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	130 Million	400 Million	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp	2 × 125 bp	2 × 150 bp

Multiplexing

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.

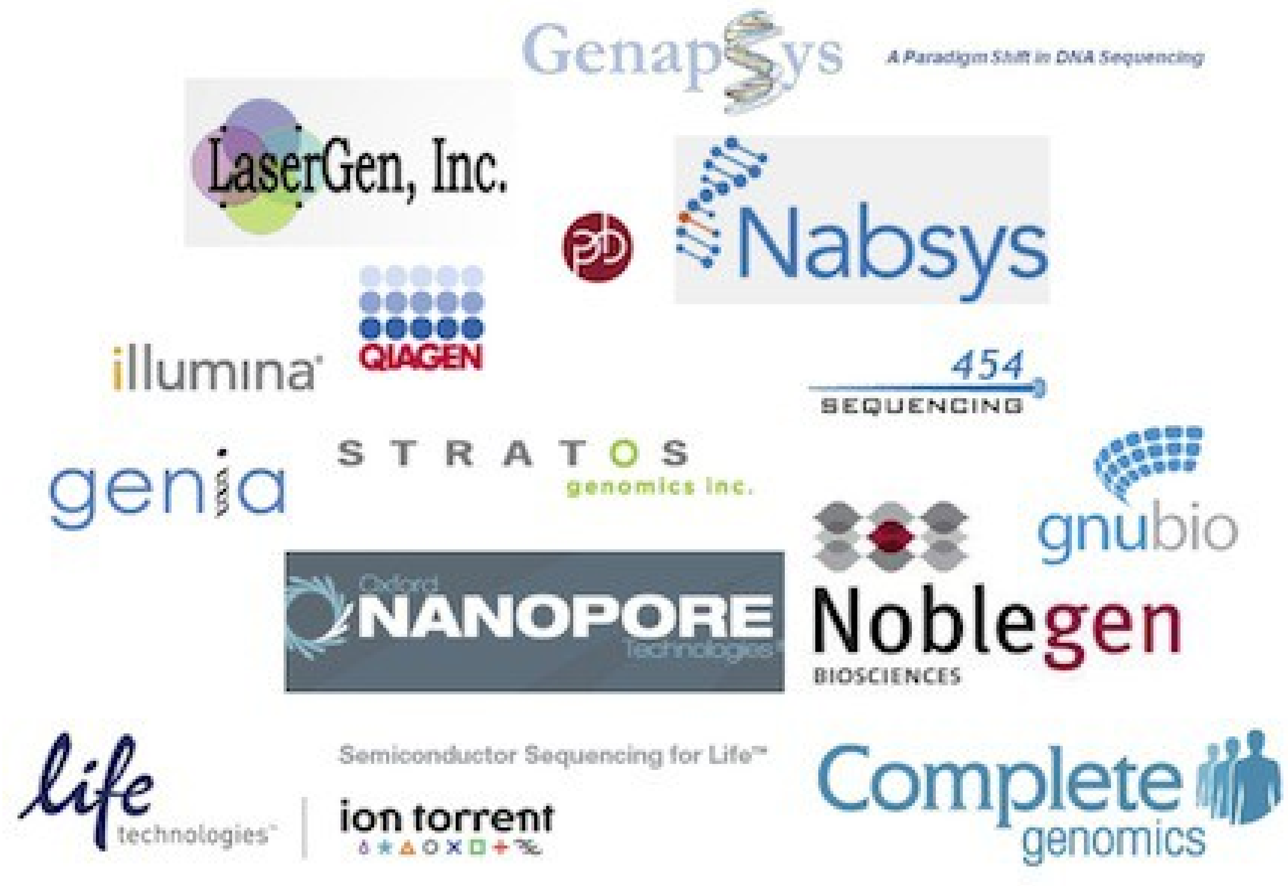
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.

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.

Illumina

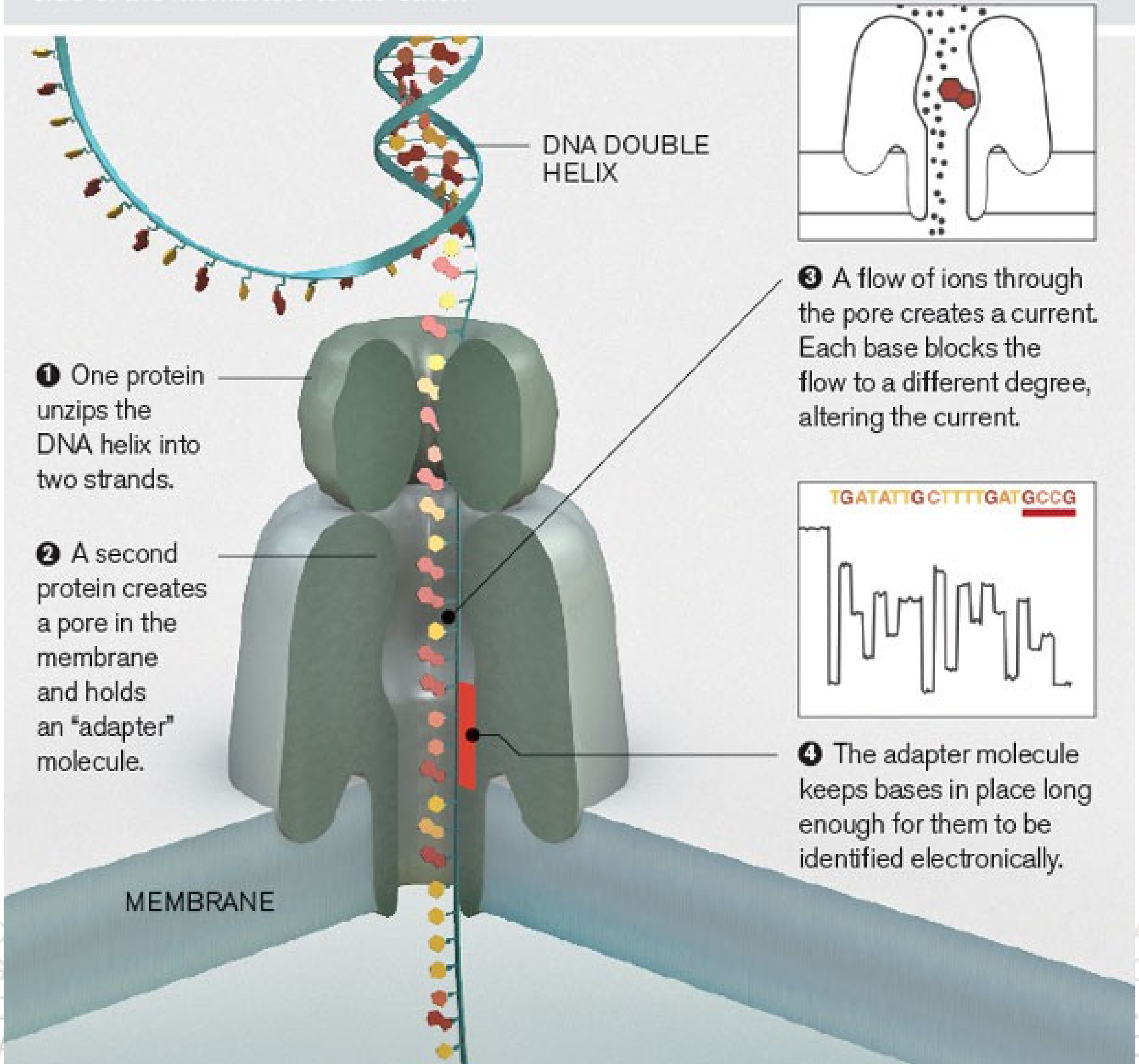
Other technologies...



http://nextgenseek.com

Nanopore sequencing

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



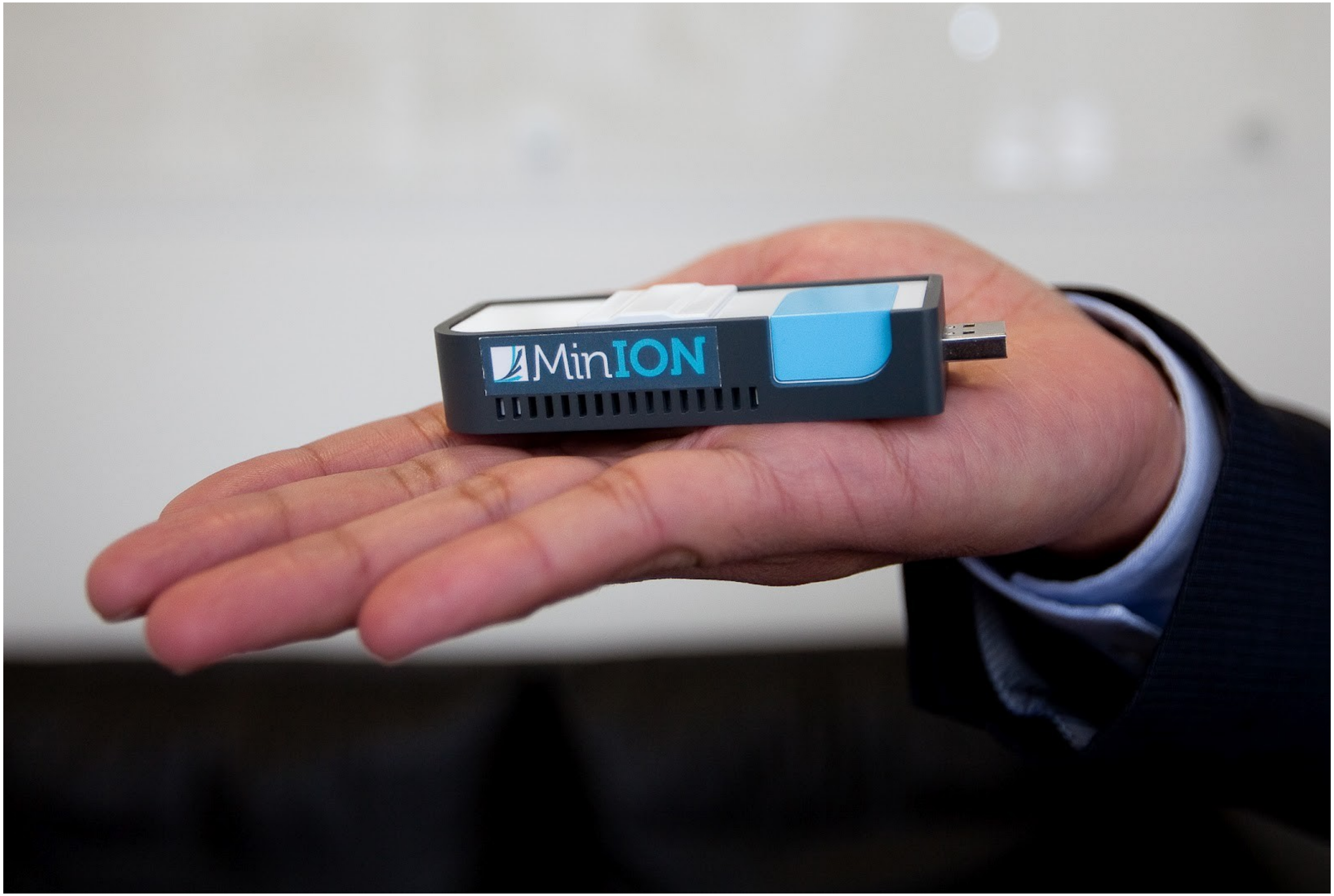
❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

The MinION



Pacific Biosciences

What's next?

- It's a rapidly evolving field
- Existing technologies continually improve

- Illumina started out as a small player 10 years ago (Solexa)
- Sequencing as a service?
- Benchtop (USB-sized?) sequencers?

The FASTQ format

```
@D256N5M1:31:C1B42ACXX:4:2305:3881:47605
ACCCCCCACAGGGACCCTTGTCACGTCCCCCTAACTCCCTGC
+
@?@FDFFFDFFFDBGIIIIGDGHIG@GHIIIGEF@@DFHGGI
```

FASTQ quality scores

Phred quality score

$$Q_{\text{sanger}} = -10 \log_{10} p$$

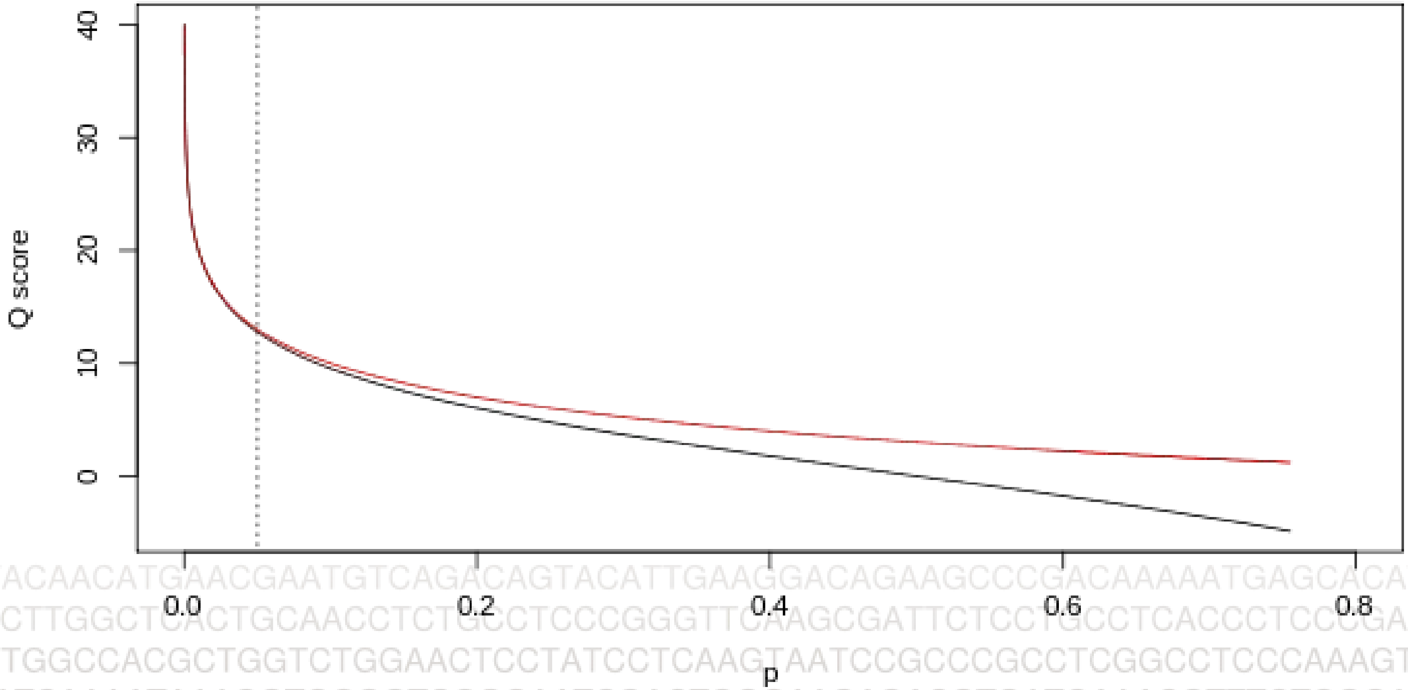
Q10	1 in 10	90%
Q20	1 in 100	99%
Q30	1 in 1000	99.9%
Q40	1 in 10000	99.99%

FASTQ quality scores

Phred quality score

$$Q_{\text{sanger}} = -10 \log_{10} p$$

Q10	1 in 10	90%
Q20	1 in 100	99%
Q30	1 in 1000	99.9%
Q40	1 in 10000	99.99%



FASTQ quality scores

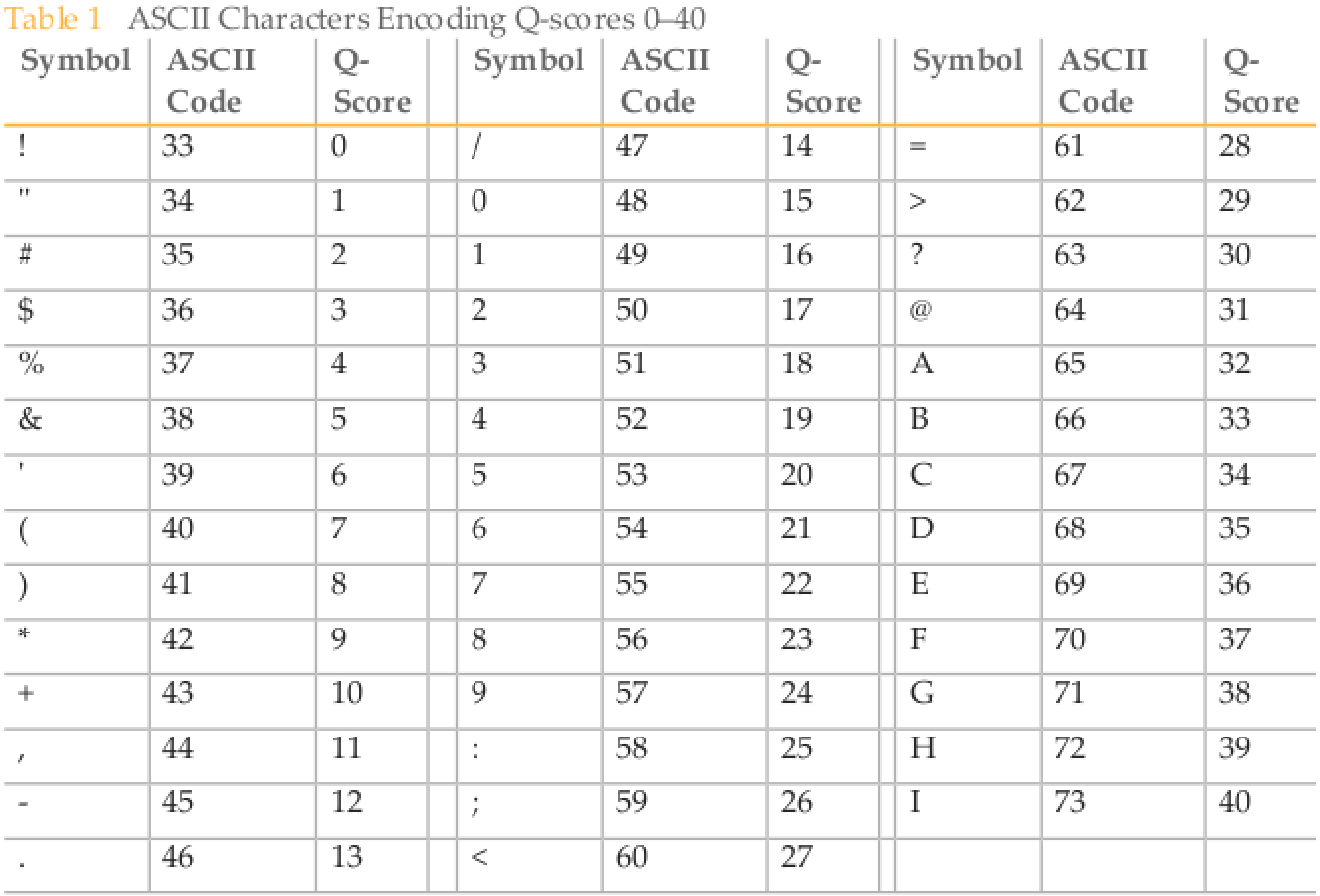


Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

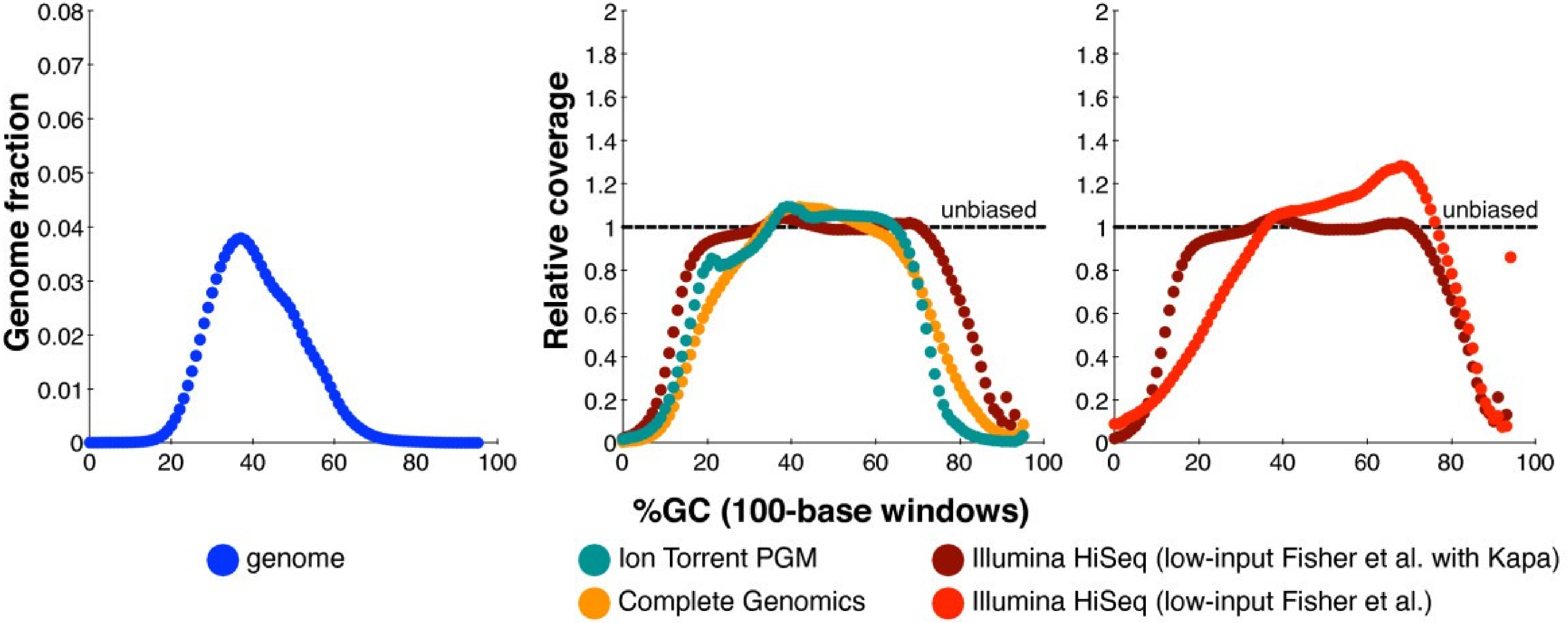
FASTQonfusion!

- Illumina:
  - CASAVA <= 1.3 Solexa
  - CASAVA 1.3 – 1.7 Illumina
  - CASAVA >= 1.8 Sanger (the “standard”)

Data quality and bias

- Sequencing errors
  - Different for different techniques!
- Amplification in sample prep => duplicate reads
- GC bias
  - Sample prep
  - Bridge amplification

GC bias

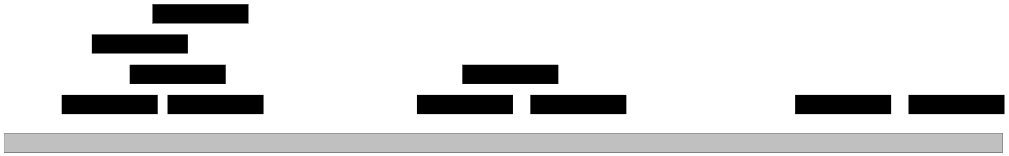


Ross et al, Genome Biology, 2013

Workflow

```
CTAGTGATTTATCATCTAGGCCAGTGAATACCAGTGGGTGGCAACCCTACC
GAATGCTCGAGCGTTCATGCGAACGATCCGAGCGCATTTTCGGCGCACGAC
CATGATGTGTAGGTAATGATTCTGAGACAAATTGCAATTGGTTTTCATTTT
ATATTGGGGTTTGGATAAACTGTTAAGCAGATTTGTCTTTCCTGAAACACT
TGTCAGTAACATTTTAAAAACAGTACAAGACATAATAGTGCCCATTGGGC
ATAACTGTCAAATGAAACAATCATCAAGTGAATTGAGTTTTAGTAGGAAT
CAGGGGTCTTGCCTGACAGAAGTGGGATTAACAGGATTCTAAAAAAAGCT
TTACAGATCGCCAAACCACAACAACAATAACAAAGGCATGGATAGGGATCC
TTAAAAAGCTTATCCCAATATGAATTGTTCCATATGGACCACTGTCAGAGG
GATTTTCCCGGGCTTAGGAAGGGGAGGAGCGAGCAAGACAGCCTACCTTTT
```

Mapping to reference



Identification of duplicates

Quantification

Peak calling

….

Workflow

```
CTAGTGATTTATCATCTAGGCCAGTGAATACCAGTGGGTGGCAACCCTACC
GAATGCTCGAGCGTTCATGCGAACGATCCGAGCGCATTTTCGGCGCACGAC
CATGATGTGTAGGTAATGATTCTGAGACAAATTGCAATTGGTTTTCATTTT
ATATTGGGGTTTGGATAAACTGTTAAGCAGATTTGTCTTTCCTGAAACACT
TGTCAGTAACATTTTAAAAACAGTACAAGACATAATAGTGCCCATTGGGC
ATAACTGTCAAATGAAACAATCATCAAGTGAATTGAGTTTTAGTAGGAAT
CAGGGGTCTTGCCTGACAGAAGTGGGATTAACAGGATTCTAAAAAAAGCT
TTACAGATCGCCAAACCACAACAACAATAACAAAGGCATGGATAGGGATCC
TTAAAAAGCTTATCCCAATATGAATTGTTCCATATGGACCACTGTCAGAGG
GATTTTCCCGGGCTTAGGAAGGGGAGGAGCGAGCAAGACAGCCTACCTTTT
```

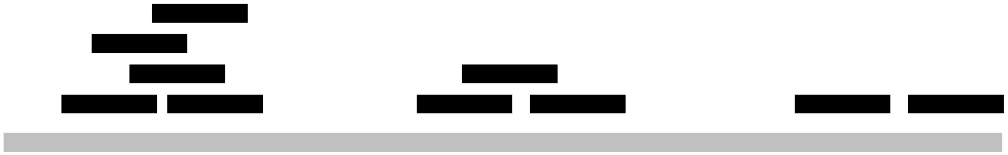
Identification of duplicates

Quantification

Peak calling

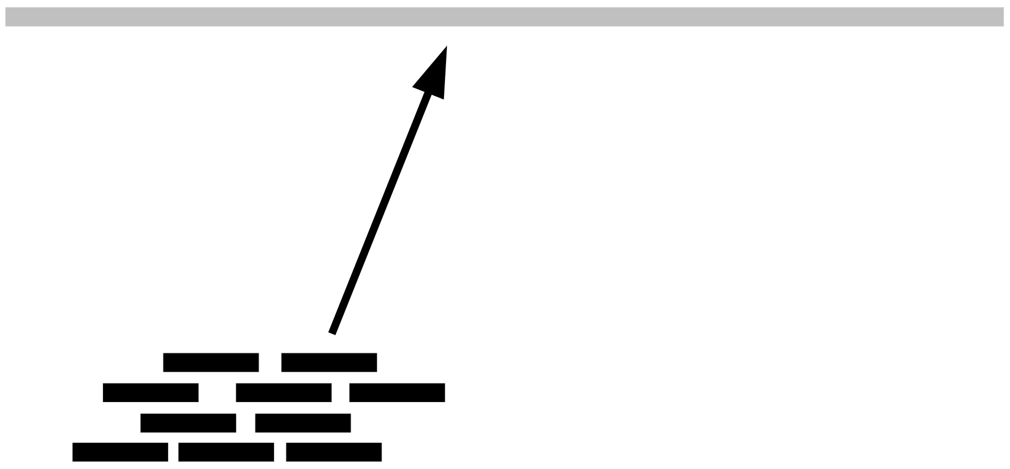
Assembly

….

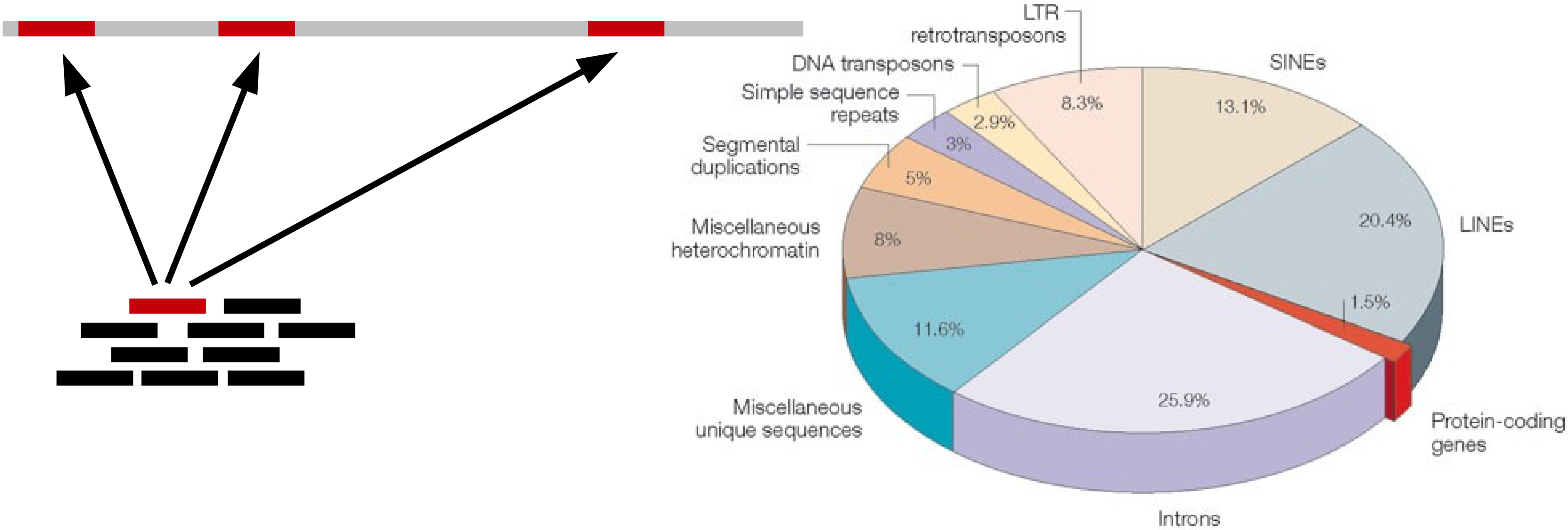


Mapping to reference

- Genome or transcriptome
  - Any other set of sequences



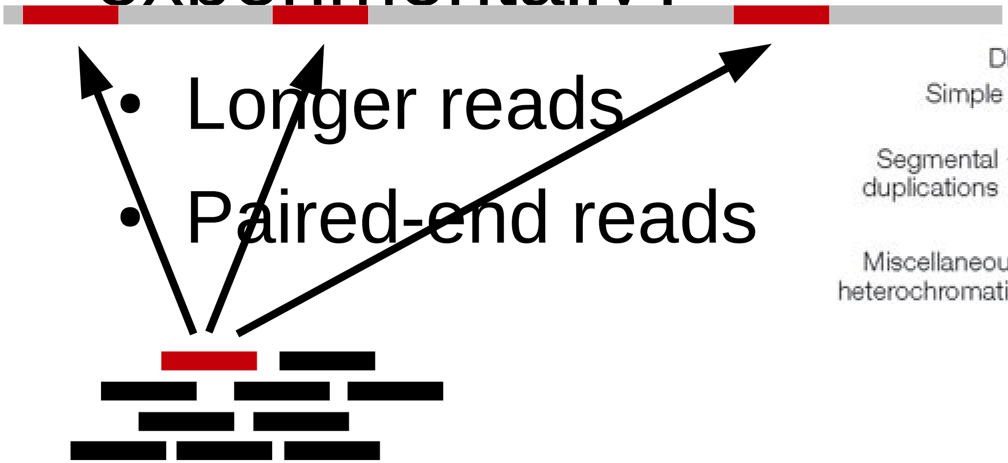
Repetitive sequence

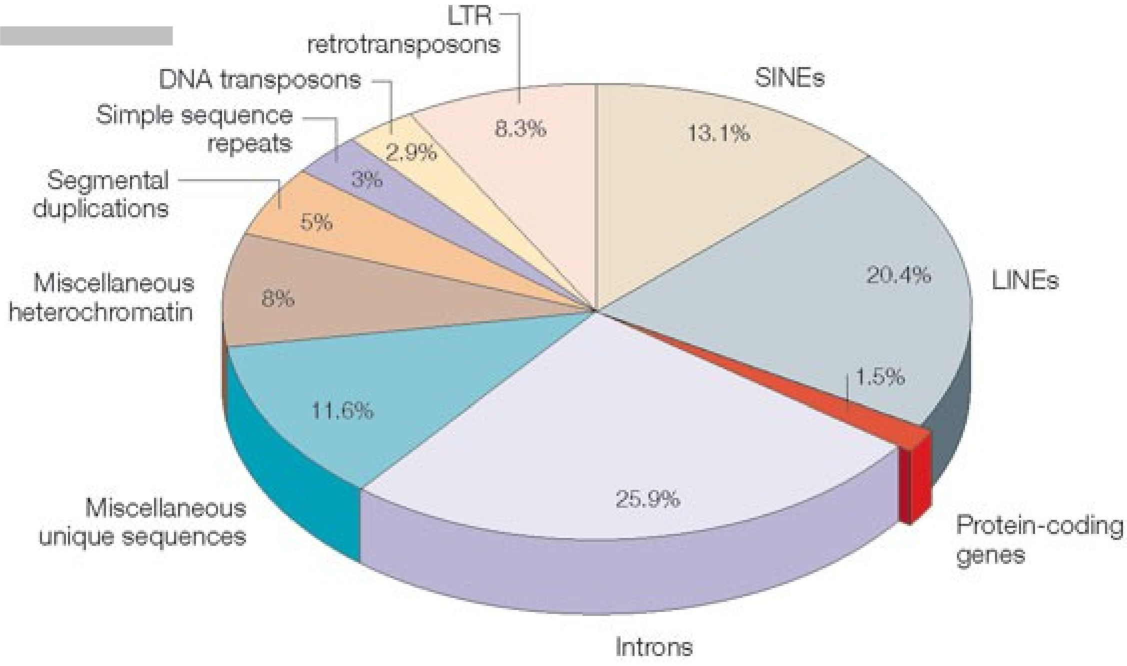


Copyright © 2005 Nature Publishing Group
Nature Reviews | Genetics

Repetitive sequence

- How to deal with repeats during mapping?
- What can be done experimentally?
  - Longer reads
  - Paired-end reads



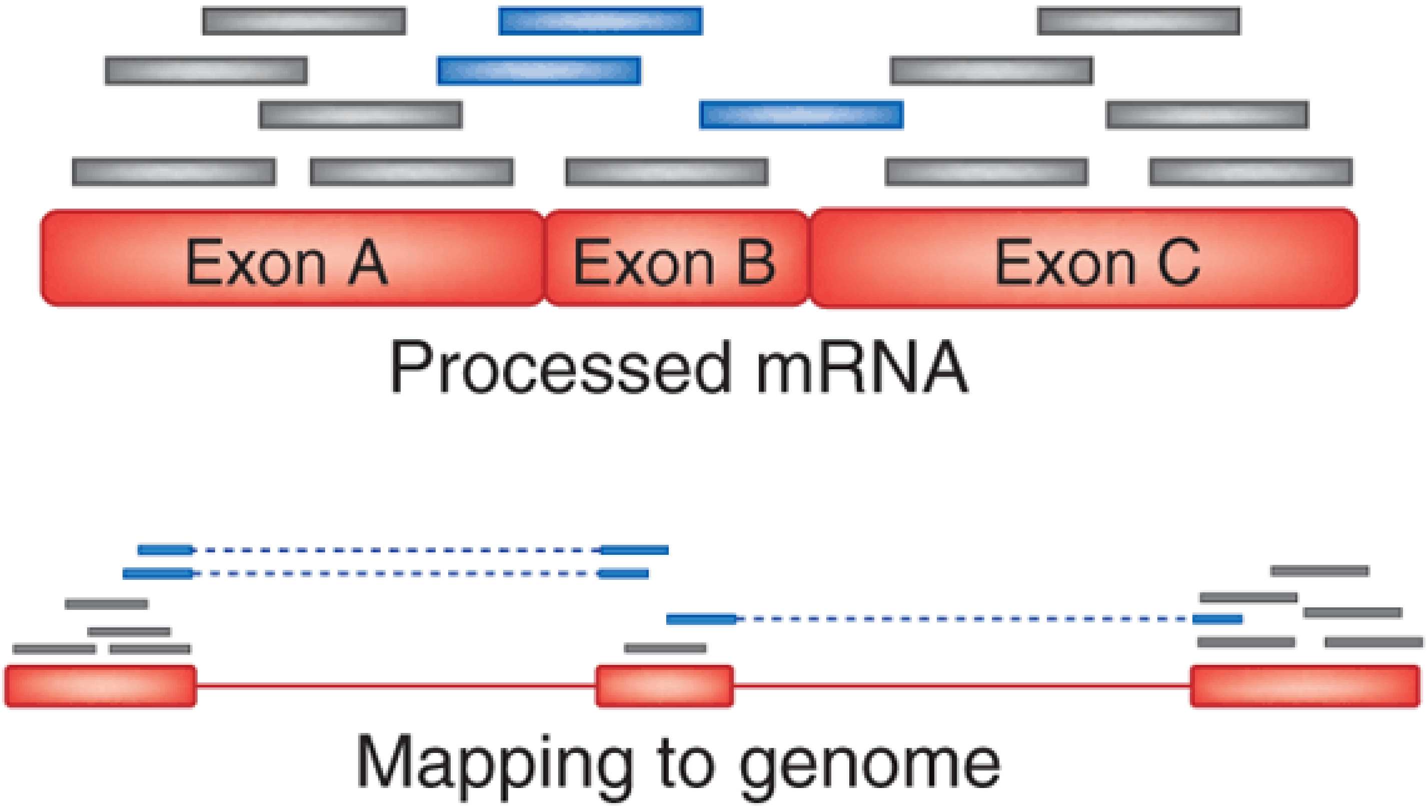


Copyright © 2005 Nature Publishing Group
Nature Reviews | Genetics

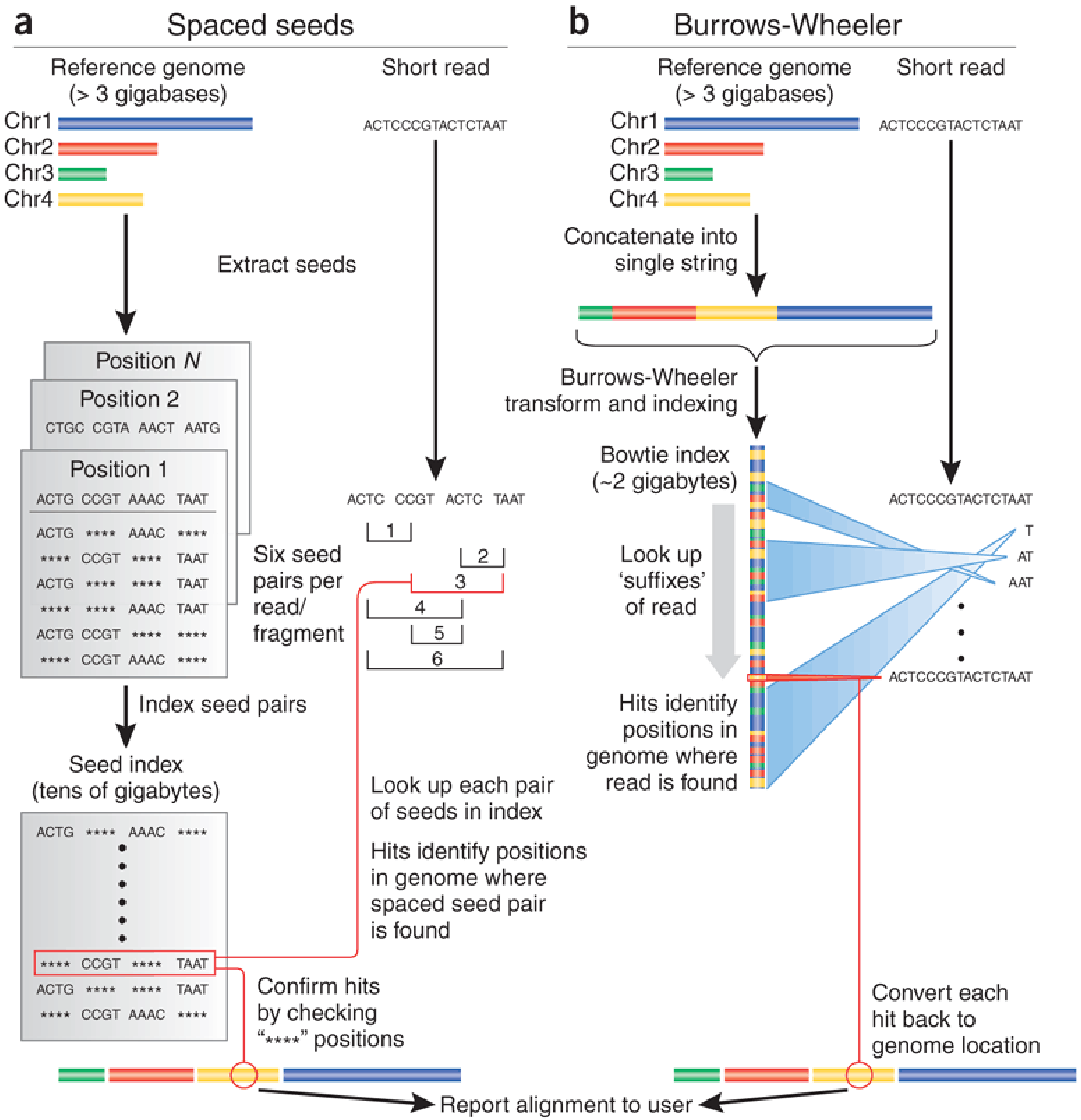
Difficulties

- Sequencing errors
- Errors in reference genome
- Polymorphisms
  - Insertions
  - Deletions
  - SNPs
- Spliced alignment

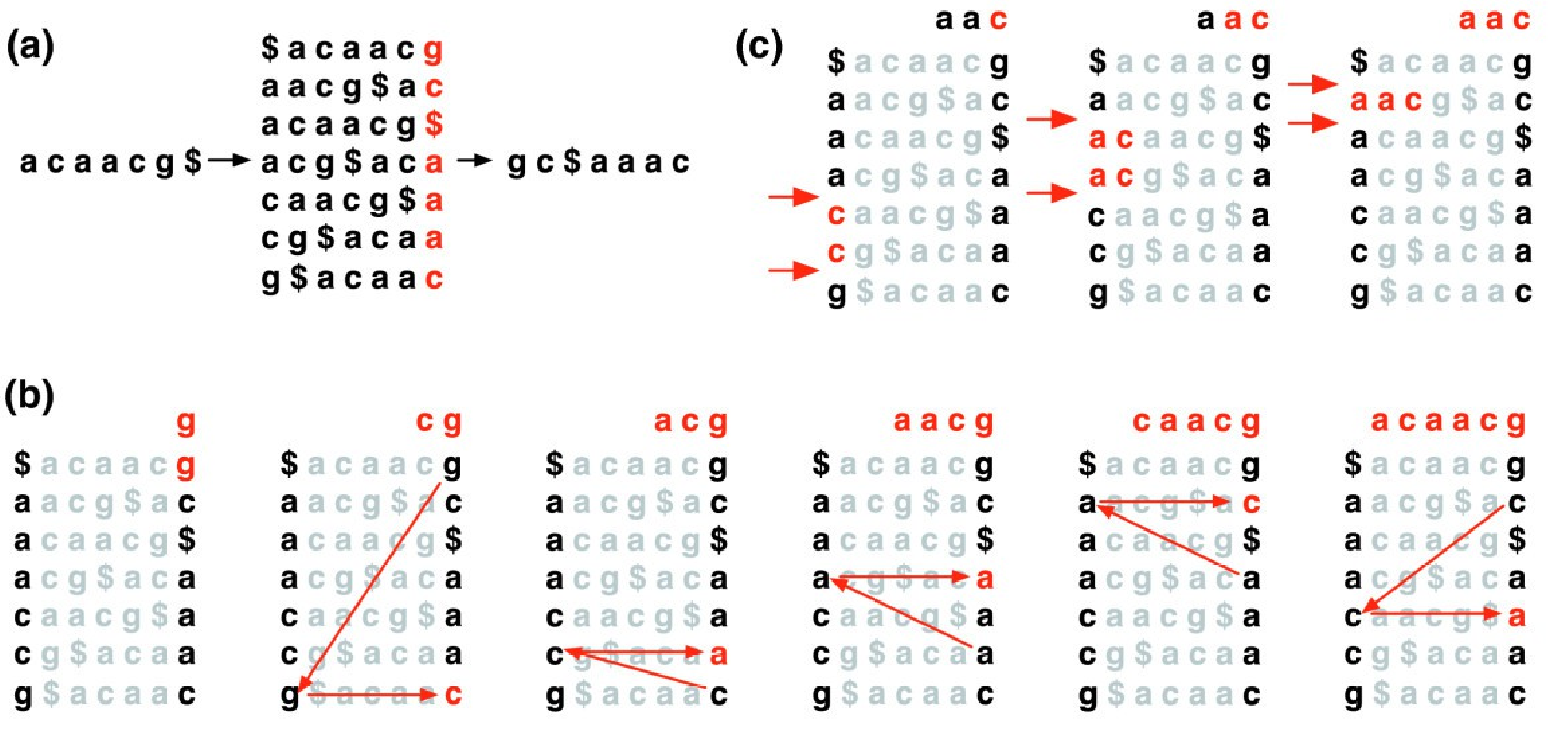
Mapping RNA-seq / spliced alignment



Approaches to mapping



Burrows-Wheeler transform



(a)

```
         $acaacg
         aacg$ac
         acaacg$
acaacg$→ acg$aca → gc$aaac
         caacg$a
         cg$acaa
         g$acaac
```

(c)

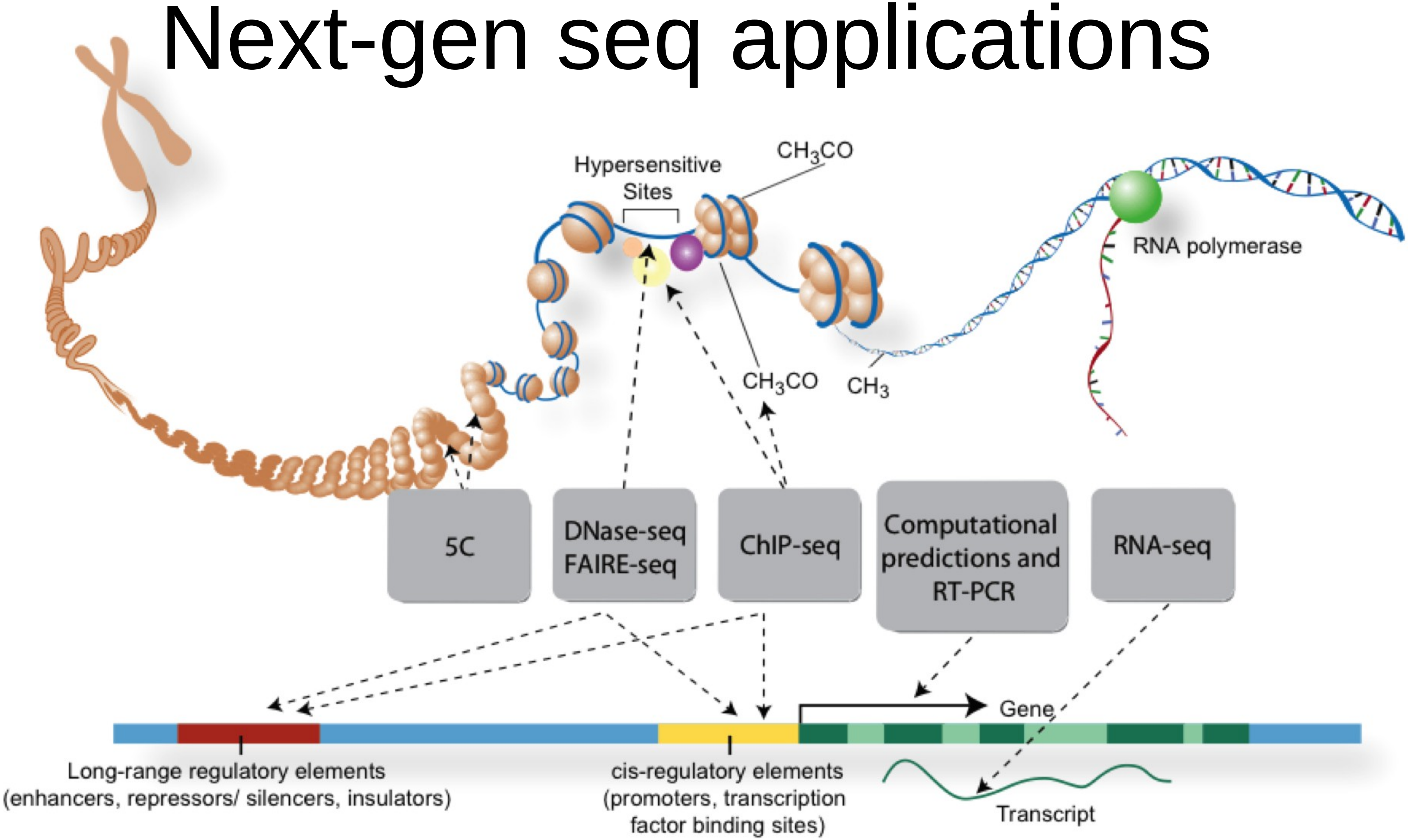
```
   aac        aac        aac
$acaacg    $acaacg    $acaacg
aacg$ac    aacg$ac    aacg$ac
acaacg$    acaacg$    acaacg$
acg$aca    acg$aca    acg$aca
caacg$a    caacg$a    caacg$a
cg$acaa    cg$acaa    cg$acaa
g$acaac    g$acaac    g$acaac
```

(b)

```
      g         cg        acg       aacg      caacg     acaacg
$acaacg    $acaacg    $acaacg    $acaacg    $acaacg    $acaacg
aacg$ac    aacg$ac    aacg$ac    aacg$ac    aacg$ac    aacg$ac
acaacg$    acaacg$    acaacg$    acaacg$    acaacg$    acaacg$
acg$aca    acg$aca    acg$aca    acg$aca    acg$aca    acg$aca
caacg$a    caacg$a    caacg$a    caacg$a    caacg$a    caacg$a
cg$acaa    cg$acaa    cg$acaa    cg$acaa    cg$acaa    cg$acaa
g$acaac    g$acaac    g$acaac    g$acaac    g$acaac    g$acaac
```

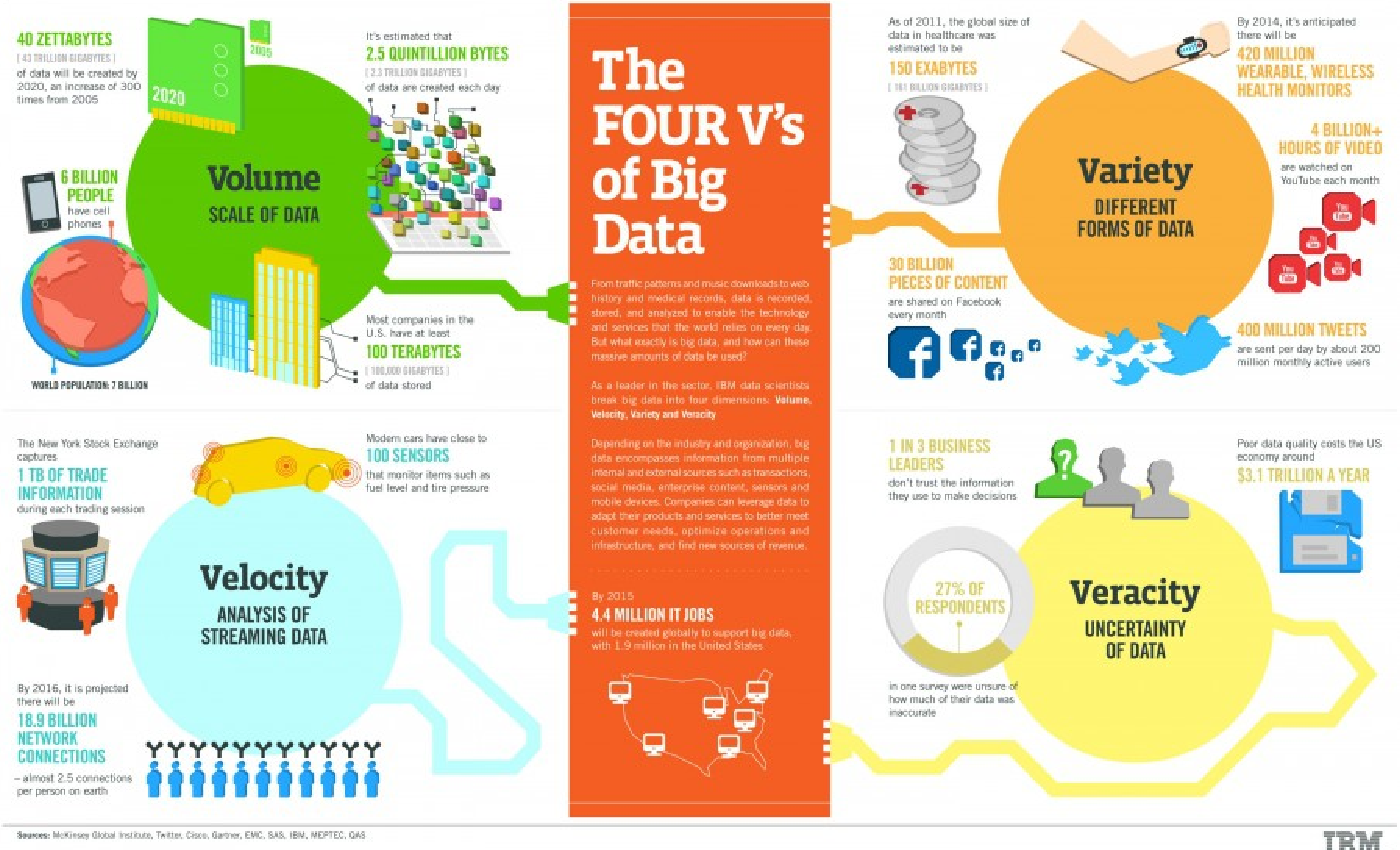
Langmead et al, 2009

Next-gen seq applications


Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq
FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

The ENCODE Consortium

Big Data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

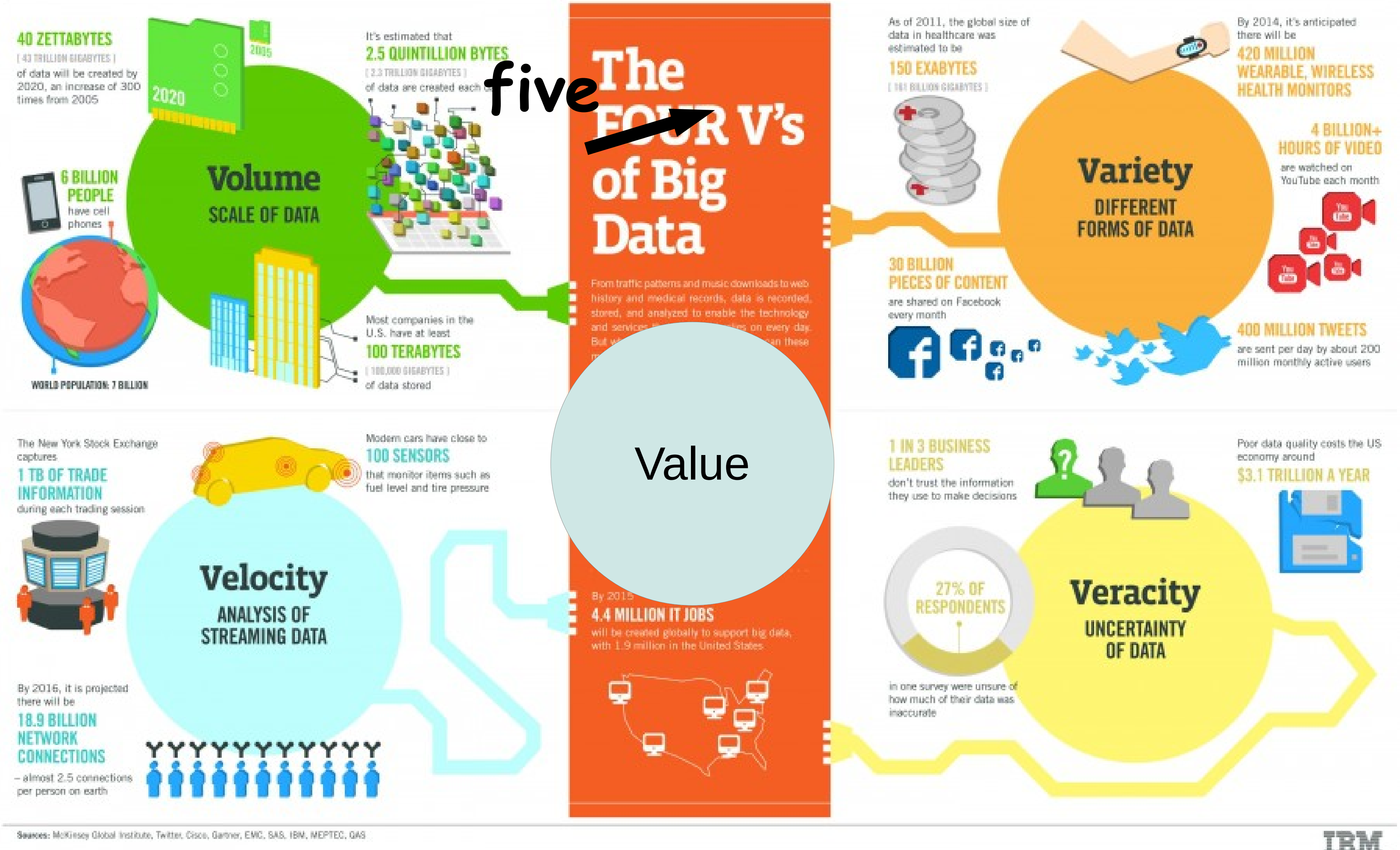
Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Big Data



five

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services th ies on every day. But wh can these m

Value

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT are shared on Facebook every month

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

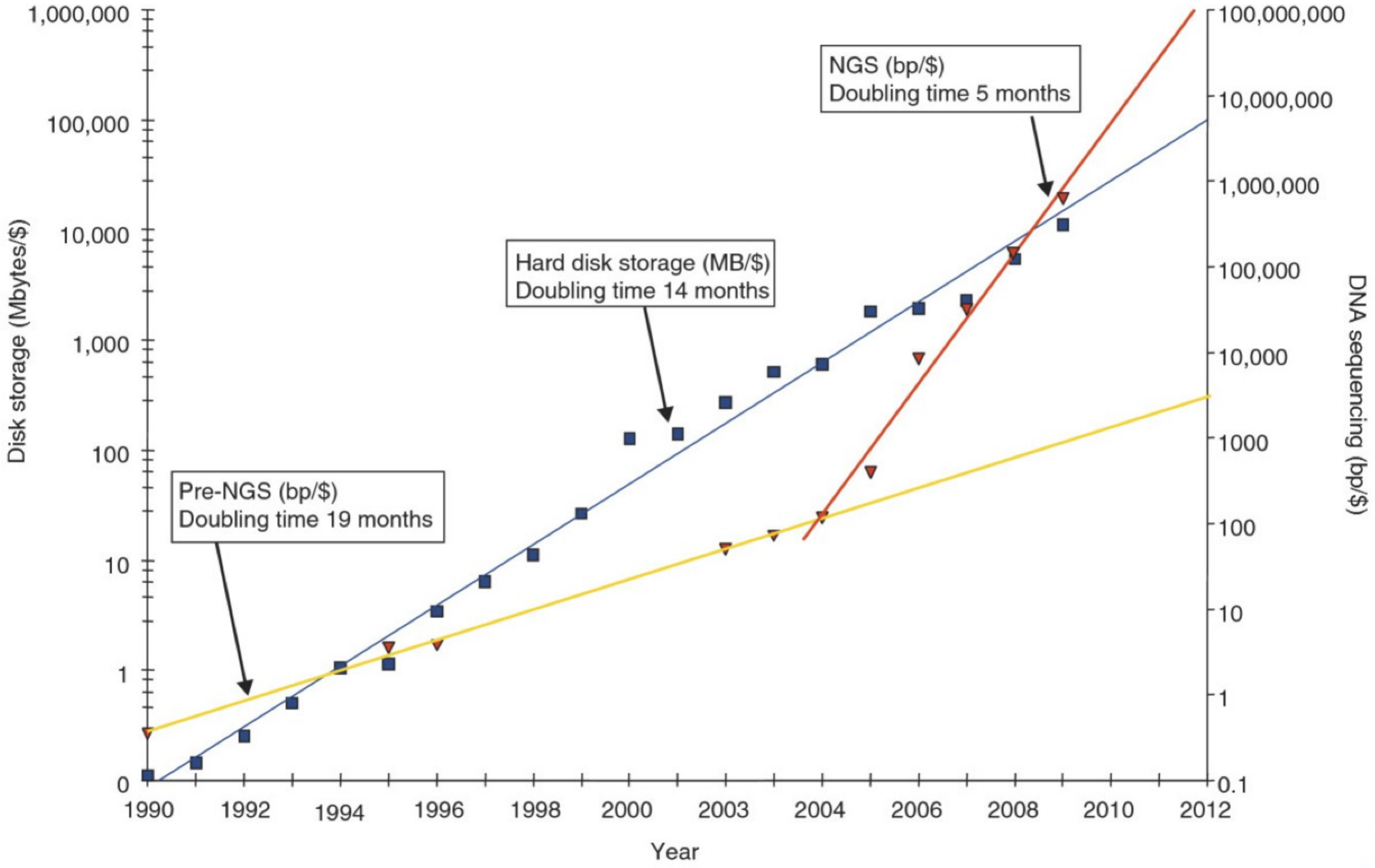
27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Challenges

- Storage
  - From terabytes to petabytes
- Analysis
  - Computational resources
    - Memory
    - CPU
- Sharing
  - Bandwidth

NextGen Sequencing a Game-Changer



NGS (bp/$)
Doubling time 5 months

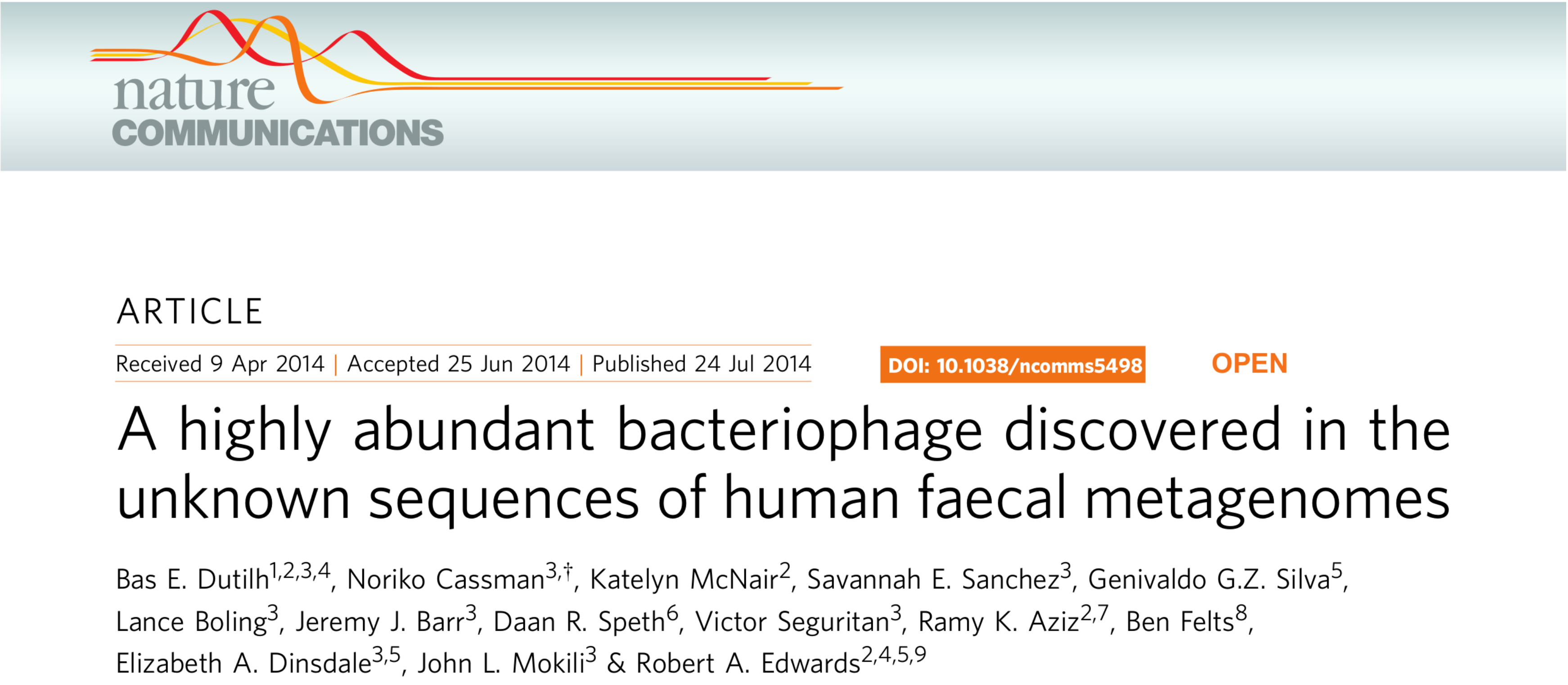
Hard disk storage (MB/$)
Doubling time 14 months

Pre-NGS (bp/$)
Doubling time 19 months

Lincoln Stein (via C. Titus Brown)

Opportunities

- A wealth of data in public databases
- Computational analysis:
  - Reproducible
  - Not dependent on lab / materials
- Cloud-based analysis
  - Amazon AWS
  - National HPC and e-Science resources
    - The Netherlands: SURFsara



nature COMMUNICATIONS

ARTICLE

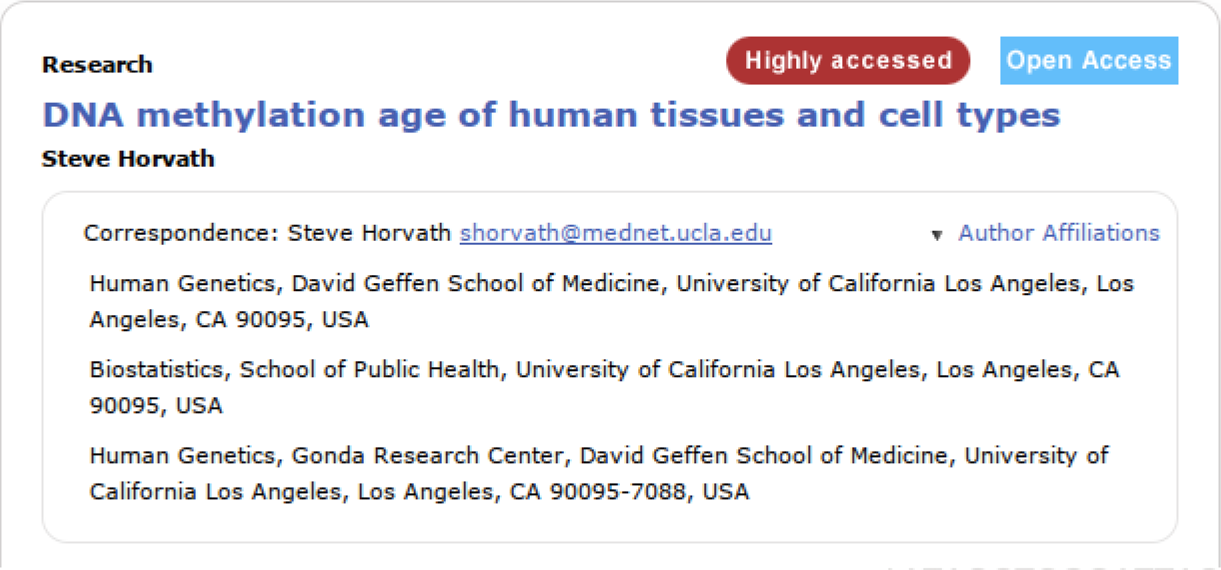
Received 9 Apr 2014 | Accepted 25 Jun 2014 | Published 24 Jul 2014

DOI: 10.1038/ncomms5498

OPEN

A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes

Bas E. Dutilh[1,2,3,4], Noriko Cassman[3,†], Katelyn McNair[2], Savannah E. Sanchez[3], Genivaldo G.Z. Silva[5], Lance Boling[3], Jeremy J. Barr[3], Daan R. Speth[6], Victor Seguritan[3], Ramy K. Aziz[2,7], Ben Felts[8], Elizabeth A. Dinsdale[3,5], John L. Mokili[3] & Robert A. Edwards[2,4,5,9]



DNA methylation age of human tissues and cell types

Steve Horvath

Correspondence: Steve Horvath shorvath@mednet.ucla.edu

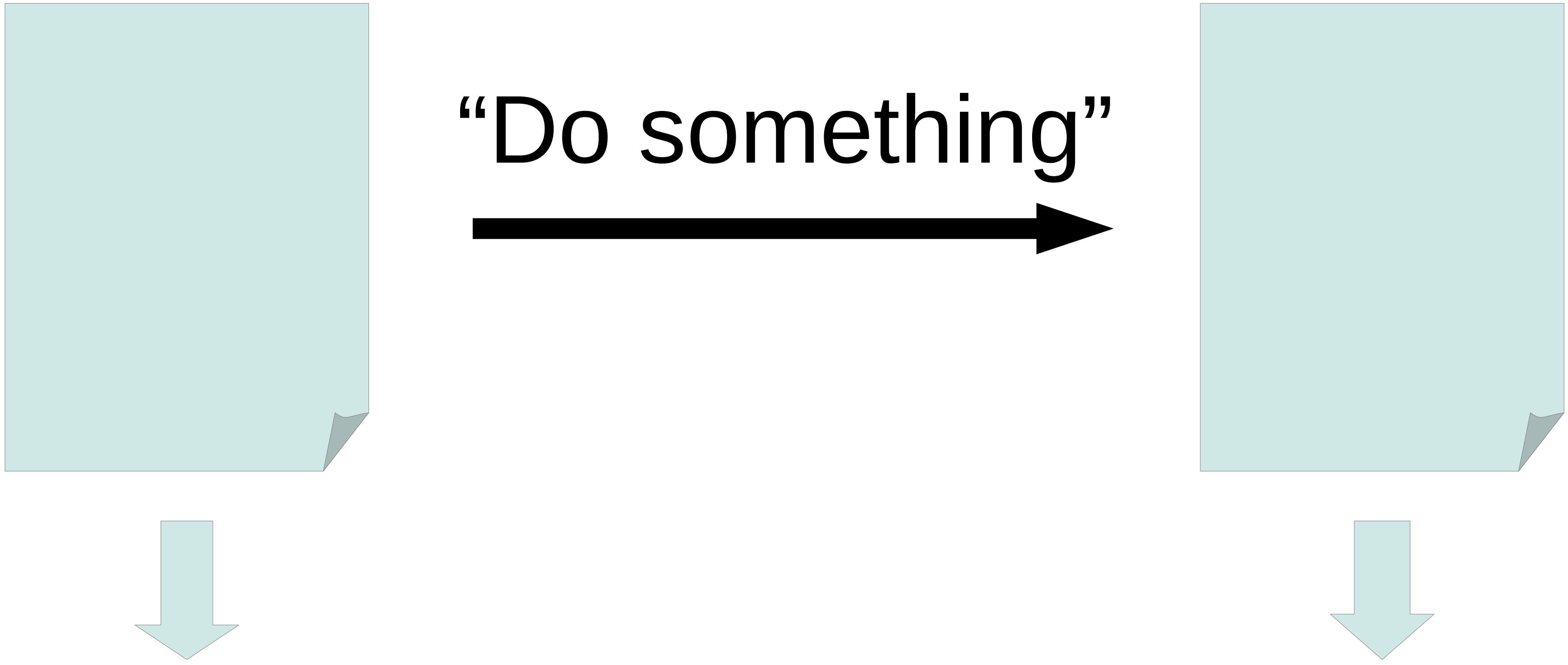
▾ Author Affiliations

Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA

Biostatistics, School of Public Health, University of California Los Angeles, Los Angeles, CA 90095, USA

Human Genetics, Gonda Research Center, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095-7088, USA

Bioinformatic analysis in a nutshell



“Do something”

Mostly text:
- Sequences
- Genomic coordinates
- Etc.

Mostly text:
- Sequences
- Genomic coordinates
- Etc.

A side note...

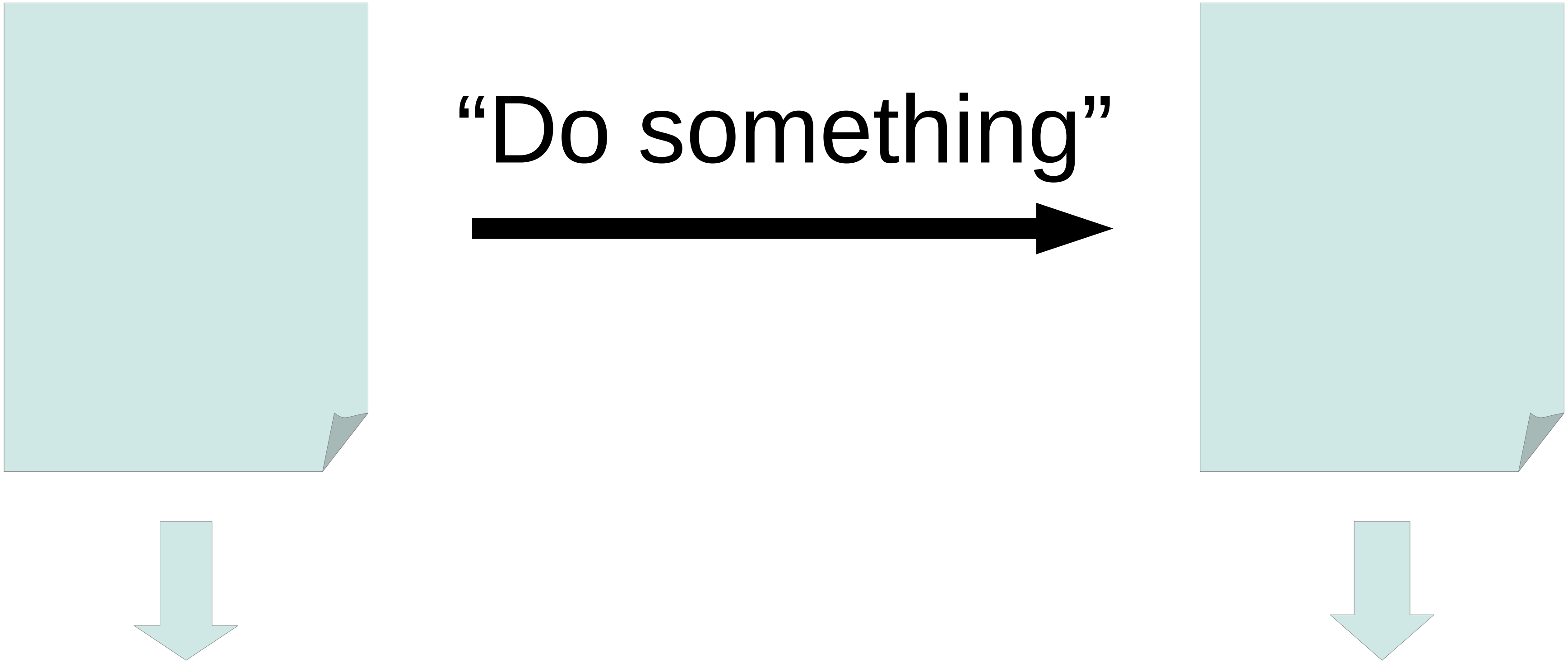
Bioinformatician? Computational Biologist? Data analyst? Data curator? Database developer? Statistician? Mathematical Modeler? Software Developer? Ontologist? Programmer?

So you want to be a computational biologist? Loman & Watson, 2013

Computational Biology

- It's about the biology
- It's research
- The computer is just the tool used to answer interesting questions
- Iterative, collaborative process between wet-lab and dry-lab

Bioinformatic analysis in a nutshell



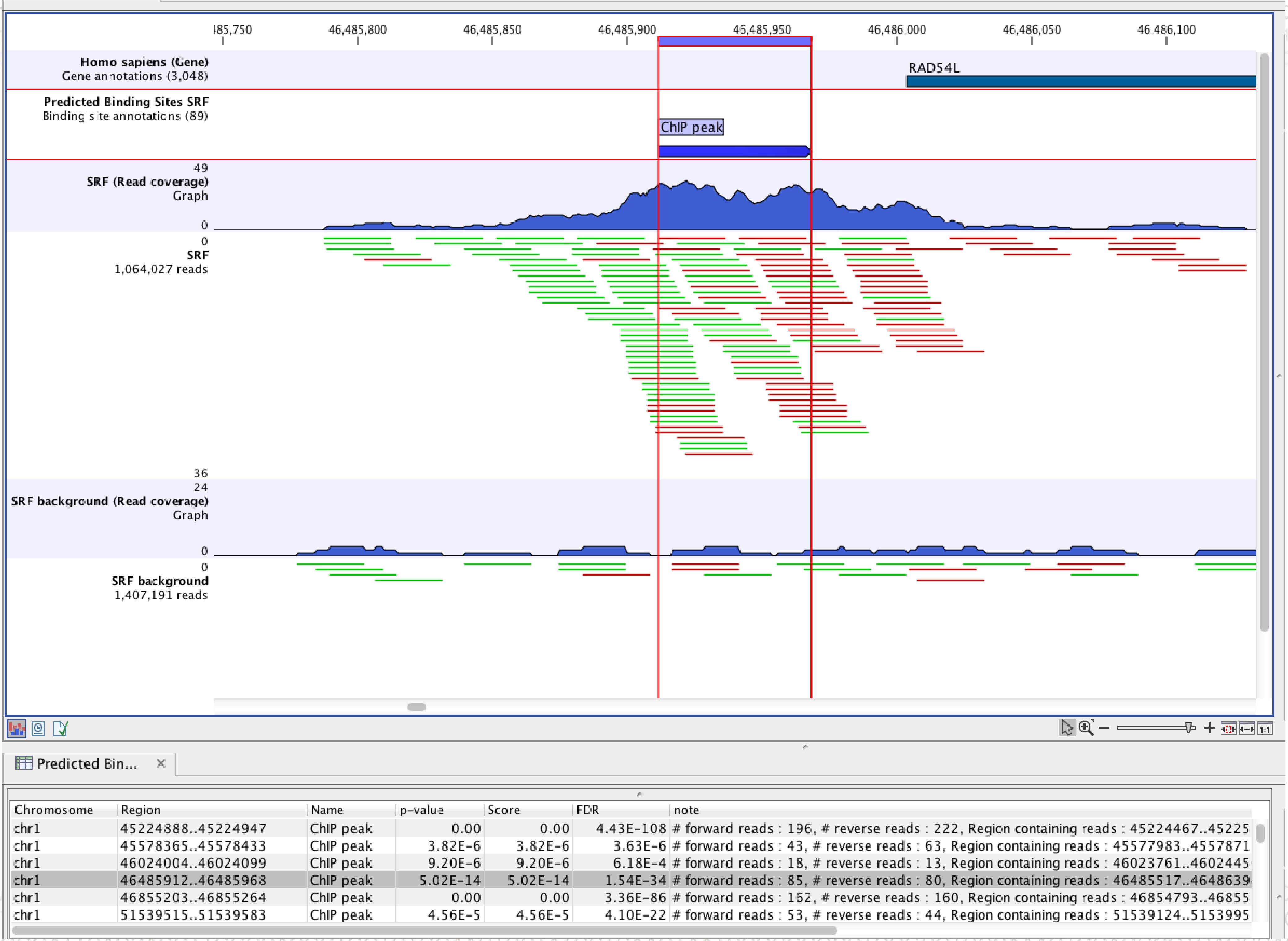
“Do something”

Mostly text:

- Sequences
- Genomic coordinates
- Etc.

Mostly text:

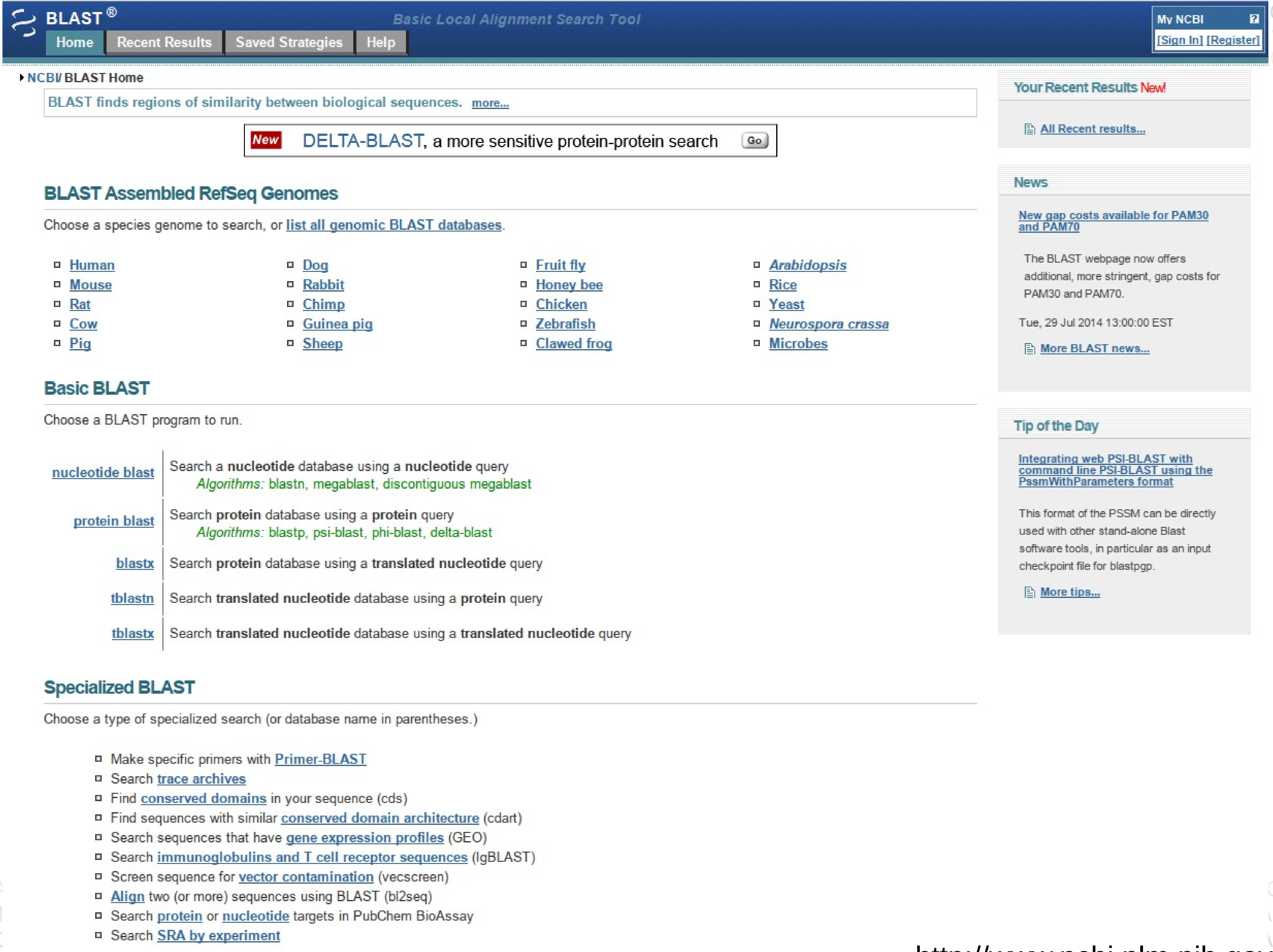
- Sequences
- Genomic coordinates
- Etc.



Chromosome	Region	Name	p-value	Score	FDR	note
chr1	45224888..45224947	ChIP peak	0.00	0.00	4.43E-108	# forward reads : 196, # reverse reads : 222, Region containing reads : 45224467..45225
chr1	45578365..45578433	ChIP peak	3.82E-6	3.82E-6	3.63E-6	# forward reads : 43, # reverse reads : 63, Region containing reads : 45577983..4557871
chr1	46024004..46024099	ChIP peak	9.20E-6	9.20E-6	6.18E-4	# forward reads : 18, # reverse reads : 13, Region containing reads : 46023761..4602445
chr1	46485912..46485968	ChIP peak	5.02E-14	5.02E-14	1.54E-34	# forward reads : 85, # reverse reads : 80, Region containing reads : 46485517..4648639
chr1	46855203..46855264	ChIP peak	0.00	0.00	3.36E-86	# forward reads : 162, # reverse reads : 160, Region containing reads : 46854793..46855
chr1	51539515..51539583	ChIP peak	4.56E-5	4.56E-5	4.10E-22	# forward reads : 53, # reverse reads : 44, Region containing reads : 51539124..5153995

GGGGATTCACCACGTTGGCCACGCTGGTCTGGAACTCCTATCCTCAAGTAATCCGCCCGCCTCGGCCTCCCAAAGTGCAGGCGTGAGCCAC
AAAATGGTTATGGAGATCAAAATAAAGGTGGGGTCGGGAATCGACTGGGAAGAGACGTGATGAAACGTTT

http://www.clcbio.com



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Sheep
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- *Neurospora crassa*
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins and T cell receptor sequences (IgBLAST)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA by experiment

Your Recent Results New!

All Recent results...

News

New gap costs available for PAM30 and PAM70

The BLAST webpage now offers additional, more stringent, gap costs for PAM30 and PAM70.

Tue, 29 Jul 2014 13:00:00 EST

More BLAST news...

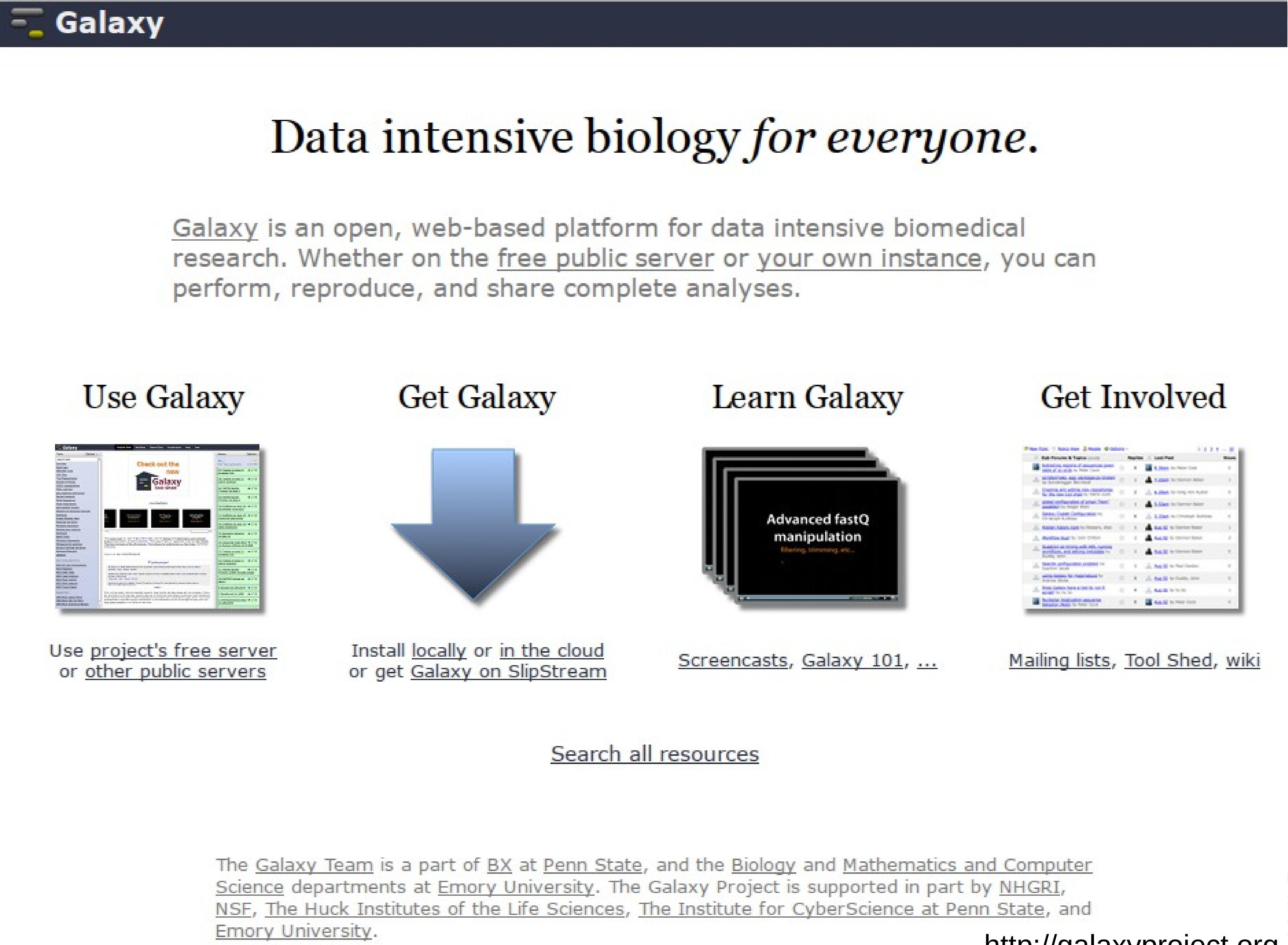
Tip of the Day

Integrating web PSI-BLAST with command line PSI-BLAST using the PssmWithParameters format

This format of the PSSM can be directly used with other stand-alone Blast software tools, in particular as an input checkpoint file for blastpgp.

More tips...

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



Galaxy

Data intensive biology *for everyone*.

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

Mailing lists, Tool Shed, wiki

Search all resources

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

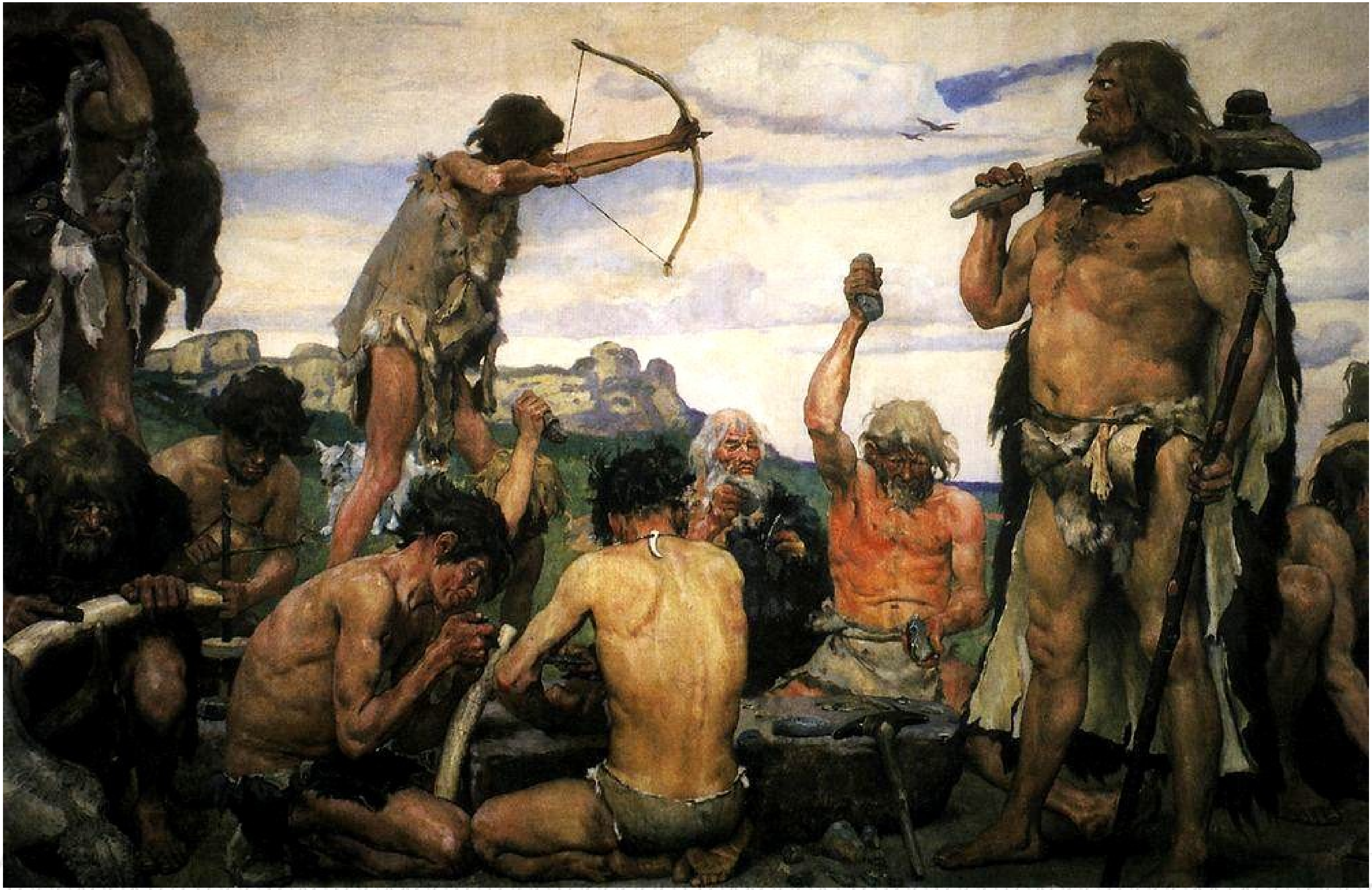
http://galaxyproject.org

An alternative

The command line

```
simon@gimme ~ $
```

Why?



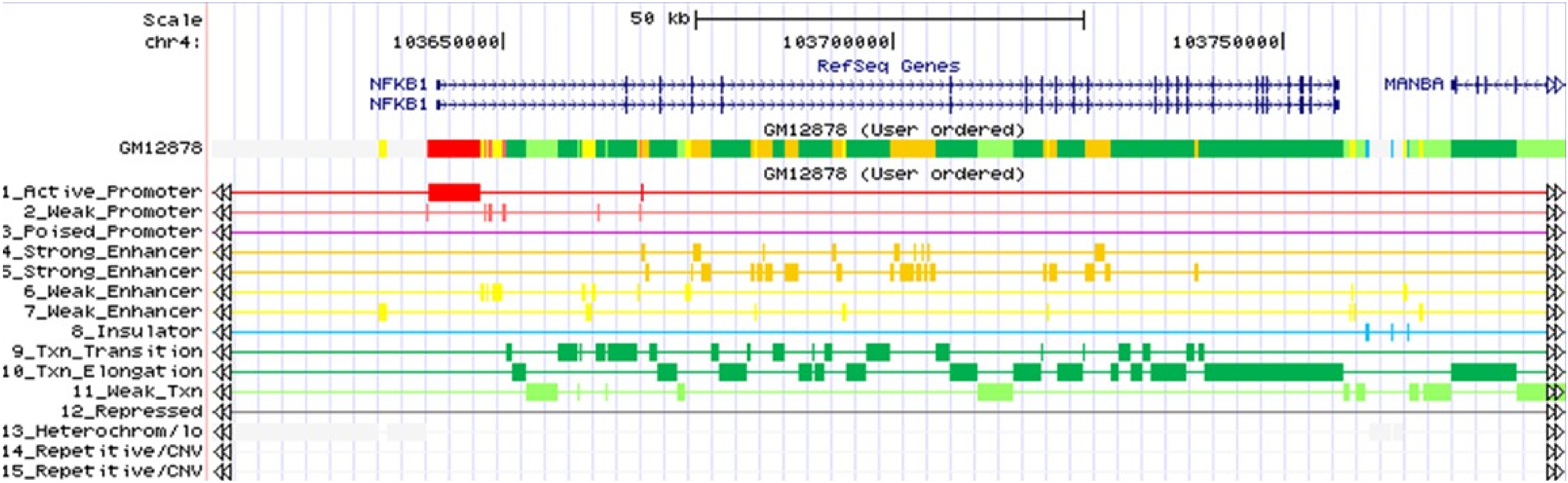
Viktor M. Vasnetsov - Wikipedia

The command line

- Powerful
- Great control over what you're doing
- Run multiple jobs
  - at once
  - hundreds of 'em!
- Many computational tools don't have a GUI
- Reproducible, reusable research
  - (In theory...)

The command line

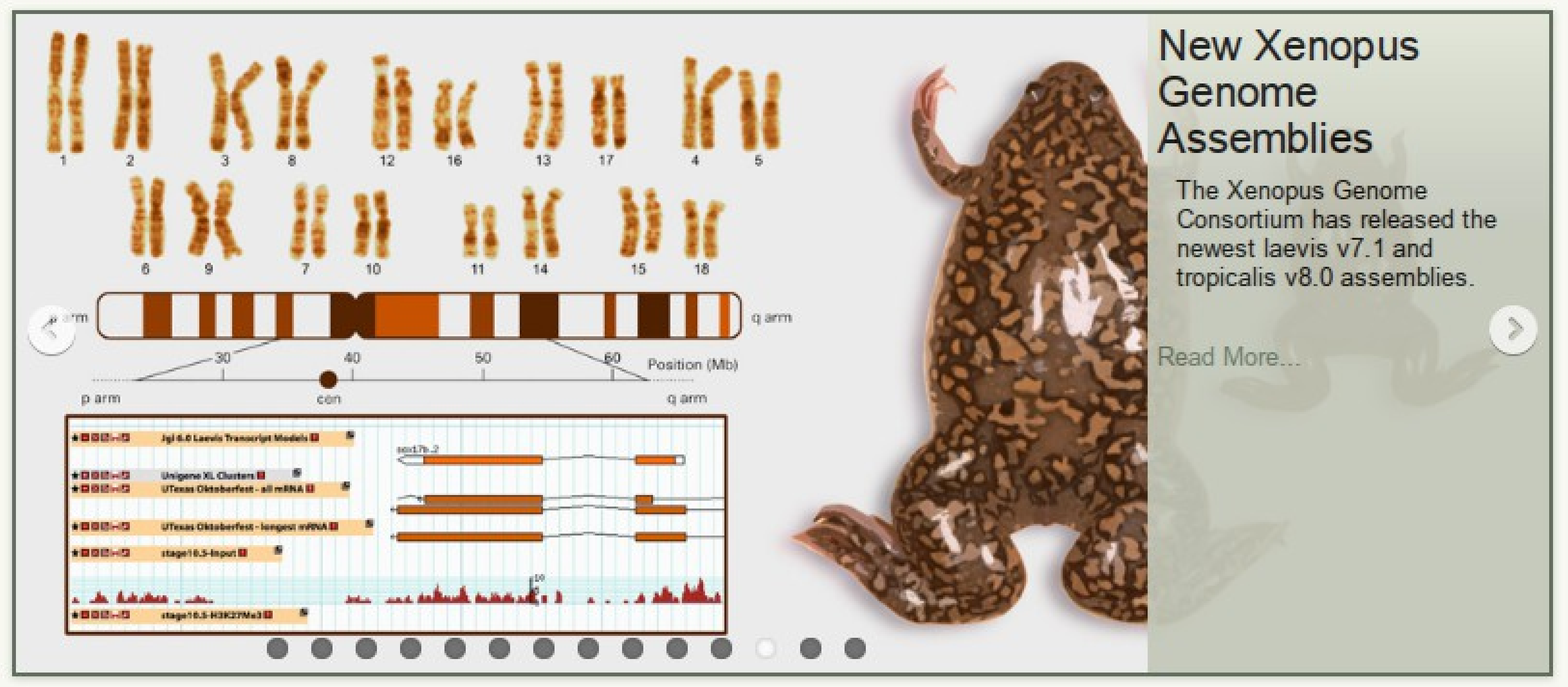
An example



ChromHMM, Ernst & Kellis, 2012

An example

- Studying development in *Xenopus tropicalis*
- 10 different assays in 5 different stages of development
- Next-gen sequencing data
- Analysis steps:
  - 1) Mapping + “peak-calling”
  - 2) Combine data in ChromHMM and learn model
  - 3) Run analyses and make figures

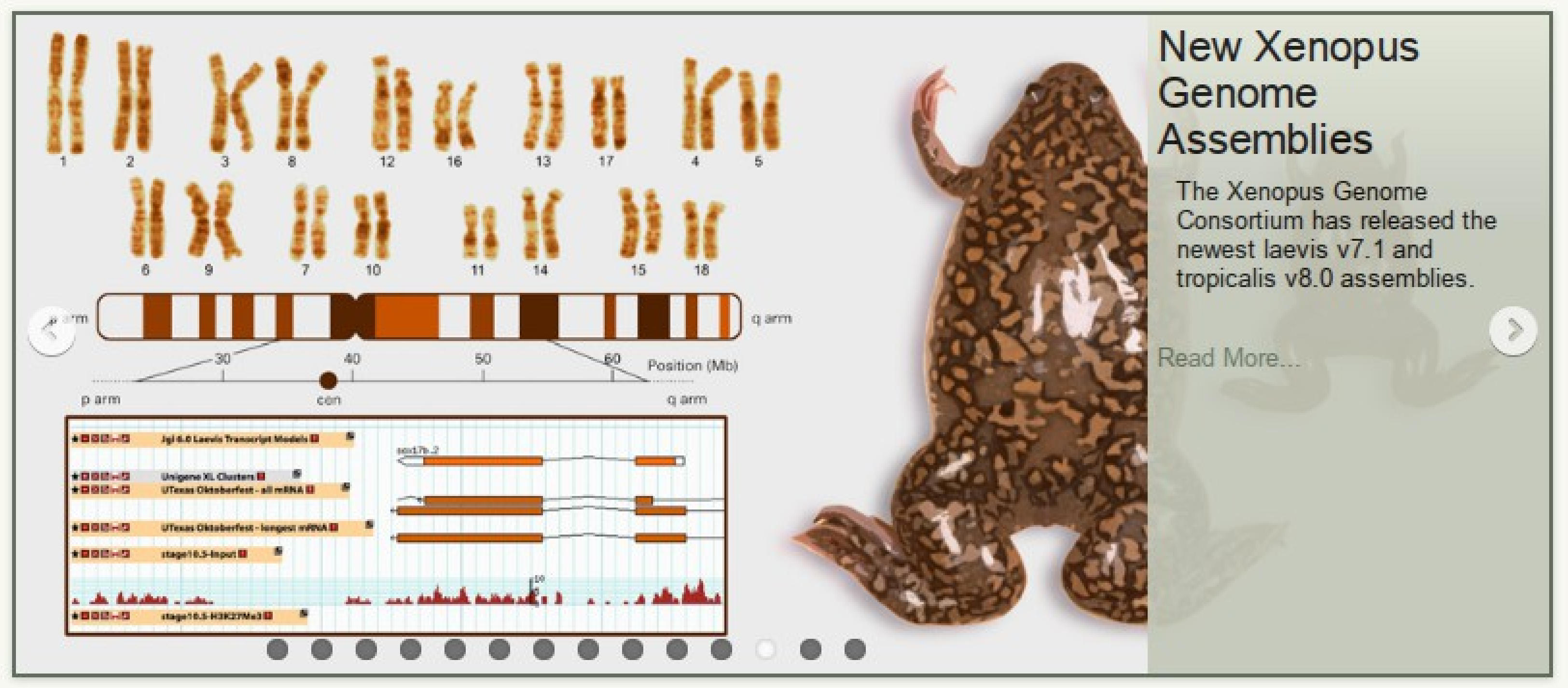


New Xenopus Genome Assemblies

The Xenopus Genome Consortium has released the newest laevis v7.1 and tropicalis v8.0 assemblies.

Read More...

- Analysis steps:
  - 1) Mapping + “peak-calling”
  - 2) Combine data in ChromHMM and learn model
  - 3) Run analyses and make figures



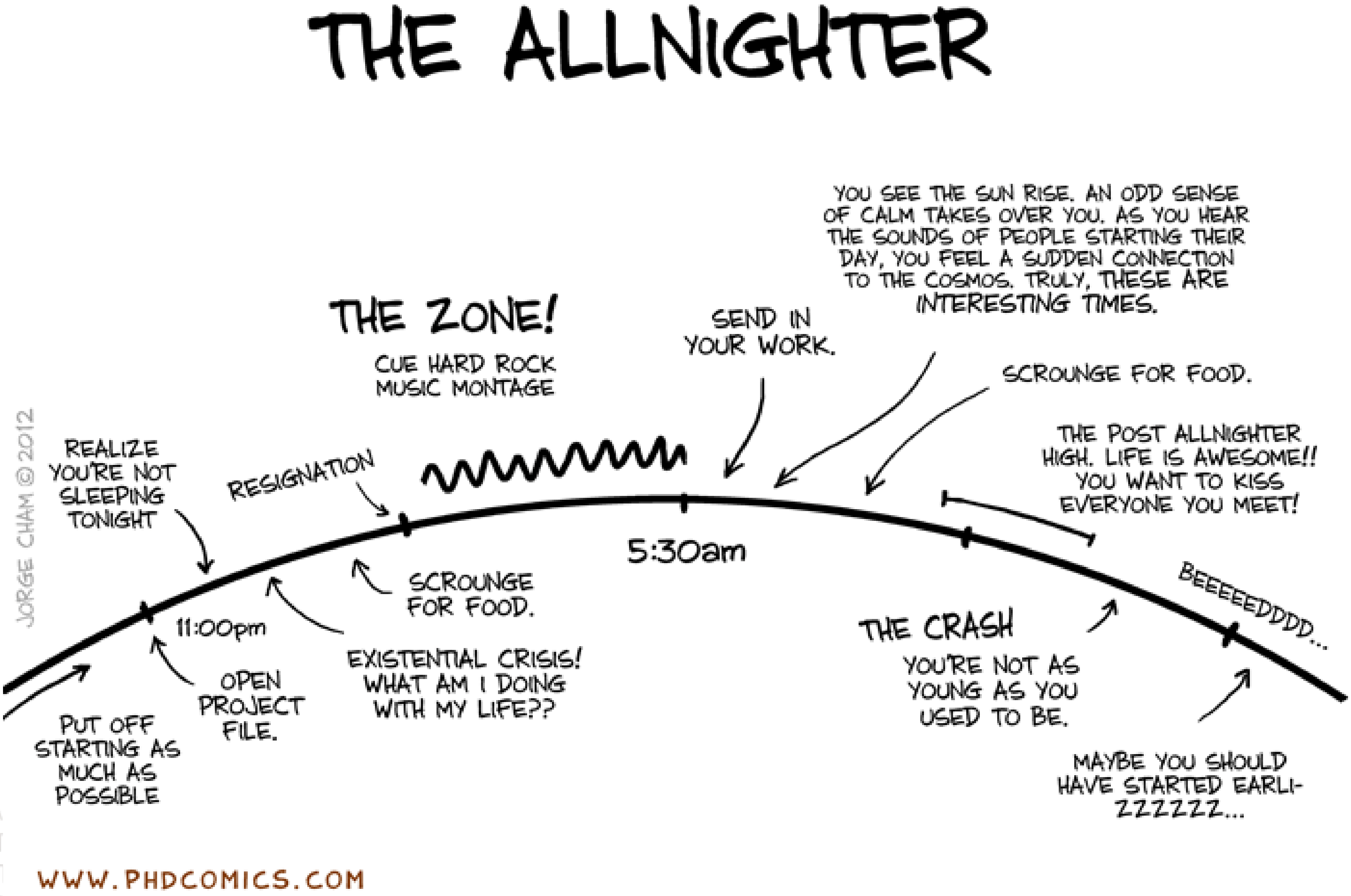
New Xenopus Genome Assemblies

The Xenopus Genome Consortium has released the newest laevis v7.1 and tropicalis v8.0 assemblies.

Read More...

- Analysis steps:
  - 1) **Mapping** + “peak-calling”
  - 2) Combine data in ChromHMM and learn model
  - 3) Run analyses and make figures

That's one unhappy PhD student..


THE ALLNIGHTER
JORGE CHAM © 2012
PUT OFF STARTING AS MUCH AS POSSIBLE
REALIZE YOU'RE NOT SLEEPING TONIGHT
11:00pm
OPEN PROJECT FILE.
EXISTENTIAL CRISIS! WHAT AM I DOING WITH MY LIFE??
SCROUNGE FOR FOOD.
RESIGNATION
THE ZONE!
CUE HARD ROCK MUSIC MONTAGE
5:30am
SEND IN YOUR WORK.
YOU SEE THE SUN RISE. AN ODD SENSE OF CALM TAKES OVER YOU. AS YOU HEAR THE SOUNDS OF PEOPLE STARTING THEIR DAY, YOU FEEL A SUDDEN CONNECTION TO THE COSMOS. TRULY, THESE ARE INTERESTING TIMES.
SCROUNGE FOR FOOD.
THE POST ALLNIGHTER HIGH. LIFE IS AWESOME!! YOU WANT TO KISS EVERYONE YOU MEET!
THE CRASH
YOU'RE NOT AS YOUNG AS YOU USED TO BE.
BEEEEEEDDDD...
MAYBE YOU SHOULD HAVE STARTED EARLI-ZZZZZZ...
WWW.PHDCOMICS.COM

Instead...

```
$ sed -i 's/JGI_7.1/JGI_8.0/' config.txt

$ ./run_analysis.sh config.txt
```

- Change configuration file
- Start script
- ~~Go home and watch Netflix~~ Continue with new, exciting analysis

Other advantages

- Load-whole-file versus streaming

- Data doesn't always all fit into memory
- A lot of biological data is just text
- Can be processed line by line

0010110011101101000011101000010000000101000011000110010100100100000000110010001100000011000000111001001011010011001000110000001100
0110111101101110001000000111011001100001011011100010000000100100001100101011001010111001001101001011011100110011101100101011011110
1000010110111001101000011001010110010101110010011010010110111001100111011001010110111001000000011011100110001101101101011011000

General considerations

- Understand your goals
- Step-by-step, don't try to do it all at once
- Try to break your own scripts
- Choose appropriate methods and tools

CAAAATACTGATATATACAACATGAACGAATGTCAGACAGTACATTGAAGGACAGAAGCCCGACAAAAATGAGCACATAATGTATGATTCCCCG
GTGTAGTGGCACGATCTTGGCTCACTGCAACCTCTGCCTCCCGGGTTCAAGCGATTCTCCTGCCTCACCCTCCCGAATAGCTGGGATTACAG
GGGGATTCACCACGTTGGCCACGCTGGTCTGGAACTCCTATCCTCAAGTAATCCGCCCGCCTCGGCCTCCCAAAGTGCAGGCGTGAGCCAC
AAAATGGTTATGGAGATCAAAATAAAGGTGGGGTCGGGAATCGACTGGGAAGAGACGTGATGAAACGTTTCTGGGACGATGAAAAGGGTCTG

TRUST NO ONE

0101100111011010000111010000100000001010000110001100101001001000000011001000110000001100000011100100101101001100100011000000110
0110111101101110001000000111011001100001011011100010000001001000011001010110010101110010011010010110111001100111011001010110111
1000010110111001101000011001010110010101110010011010010110111001100111011001010110111001000000011011100110001101101101011011000

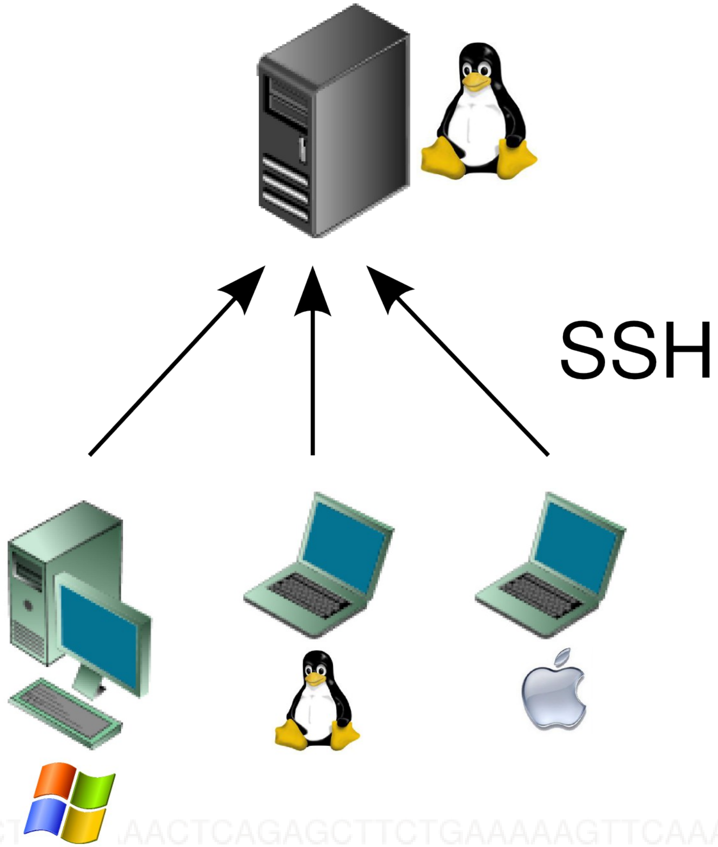
Today

- Familiarize yourself with the Linux command line
- Next-gen file formats
  - FASTQ
  - BAM
- Mapping (?)

CAAAATACTGATATATACAACATGAACGAATGTCAGACAGTACATTGAAGGACAGAAGCCCGACAAAAATGAGCACATAATGTATGATTCCCC
GTGTAGTGGCACGATCTTGGCTCACTGCAACCTCTGCCTCCCGGGTTCAAGCGATTCTCCTGCCTCACCCTCCCGAATAGCTGGGATTACAG
GGGGATTCACCACGTTGGCCACGCTGGTCTGGAACTCCTATCCTCAAGTAATCCGCCCGCCTCGGCCTCCCAAAGTGCAGGCGTGAGCCAC
AAAATGGTTATGGAGATCAAAATAAAGGTGGGGTCGGGAATCGACTGGGAAGAGACGTGATGAAACGTTTCTGGGACGATGAAAAGGGTCTC

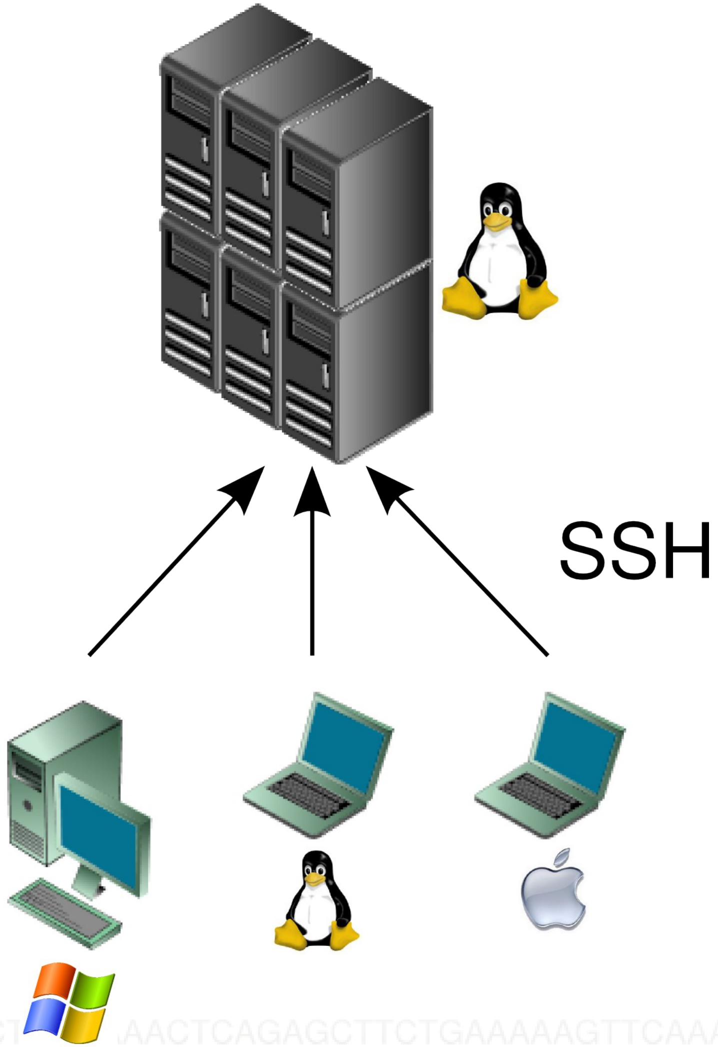
SSH (Secure Shell)

- Connect to server
- Clients available for every OS
- Perform analysis remotely

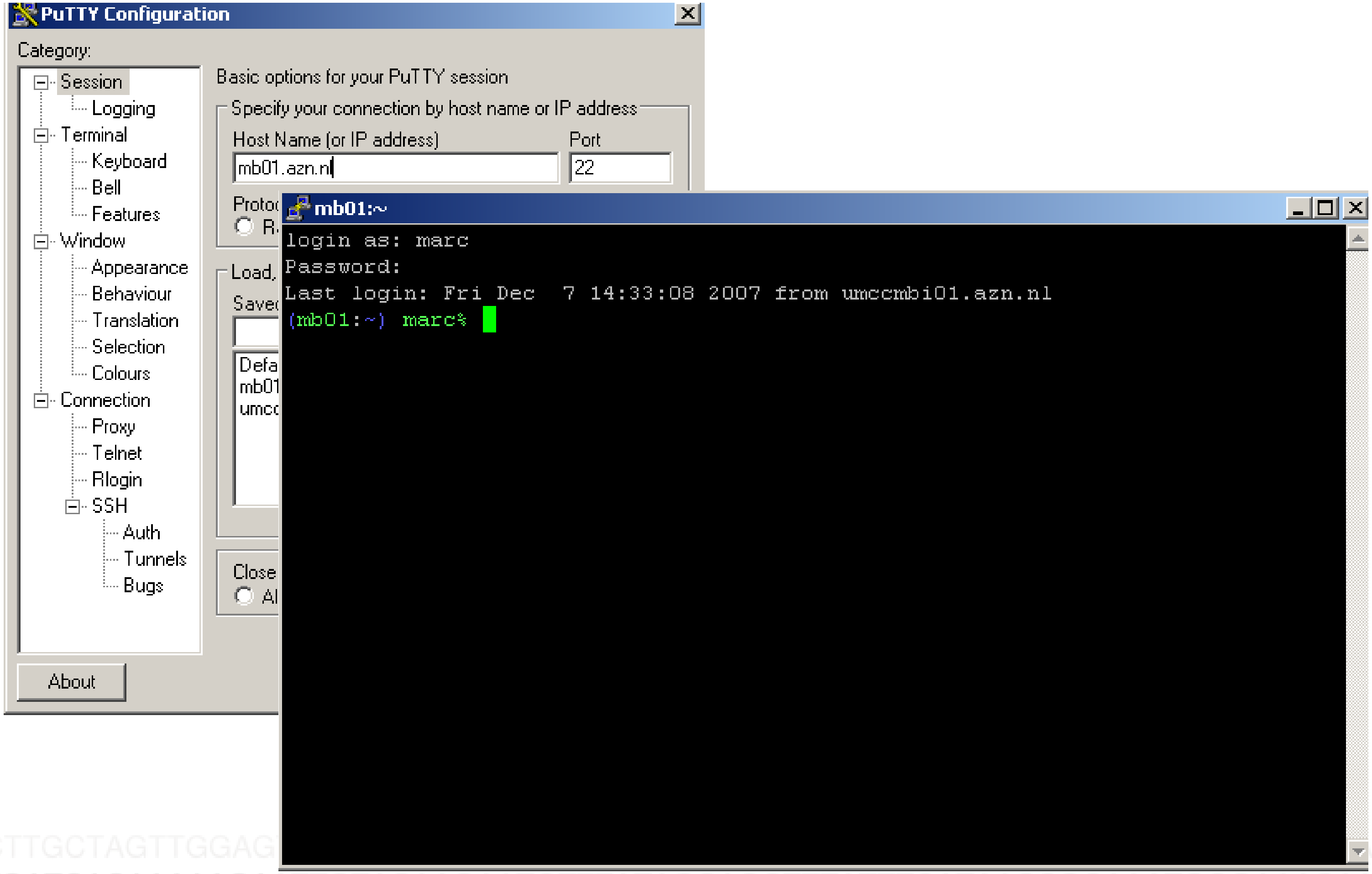


SSH (Secure Shell)

- Connect to server
- Clients available for every OS
- Perform analysis remotely
- Server can have lots of memory and CPU power



For Windows: Putty

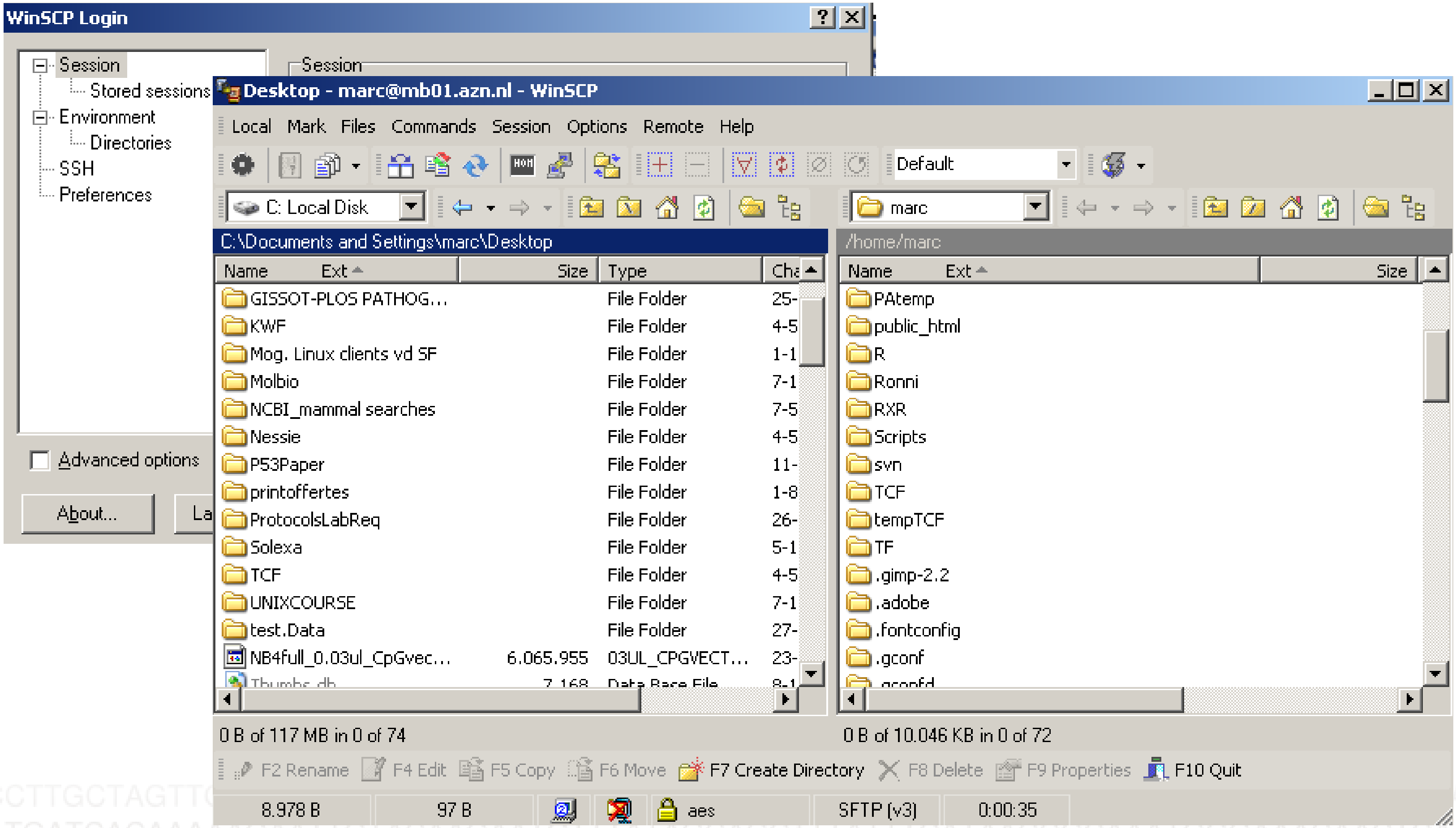


Host Name (or IP address): mb01.azn.nl

Port: 22

```
login as: marc
Password:
Last login: Fri Dec  7 14:33:08 2007 from umccmbi01.azn.nl
(mb01:~) marc%
```

For Windows: WinSCP



C:\Documents and Settings\marc\Desktop

Name	Ext	Size	Type	Cha
GISSOT-PLOS PATHOG...			File Folder	25-
KWF			File Folder	4-5
Mog. Linux clients vd SF			File Folder	1-1
Molbio			File Folder	7-1
NCBI_mammal searches			File Folder	7-5
Nessie			File Folder	4-5
P53Paper			File Folder	11-
printoffertes			File Folder	1-8
ProtocolsLabReq			File Folder	26-
Solexa			File Folder	5-1
TCF			File Folder	4-5
UNIXCOURSE			File Folder	7-1
test.Data			File Folder	27-
NB4full_0.03ul_CpGvec...		6.065.955	03UL_CPGVECT...	23-
Thumbs.db		7.168	Data Base File	8-1

0 B of 117 MB in 0 of 74

/home/marc

Name	Ext	Size
PAtemp		
public_html		
R		
Ronni		
RXR		
Scripts		
svn		
TCF		
tempTCF		
TF		
.gimp-2.2		
.adobe		
.fontconfig		
.gconf		
.gconfd		

0 B of 10.046 KB in 0 of 72

8.978 B | 97 B | aes | SFTP (v3) | 0:00:35

Server IP addresses

- 23.20.162.10
- 54.80.42.82
- 23.20.67.155
- 54.198.43.176
- 54.81.15.78
- 54.242.170.57
- 54.80.155.251