Graduate School
Bioinformatics Sequence Analysis
Introduction

Barbera van Schaik

Bioinformatics and Biomedical Computing
Epidemiology and Data Science
Amsterdam UMC

b.d.vanschaik@amsterdamumc.nl

March 8, 2021

Related Graduate School courses

- DNA technology
- Unix
- Computing in R
- Practical biostatistics
- Advanced biostatistics
- Bioinformatics
- Bioinformatics Sequence Analysis
- Research Data Management

`https://www.amc.nl/web/leren/graduate-school.htm`

In this course

Bioinformatics Sequence Analysis

You will learn what is behind commonly used **methods for sequence analysis**, how to **analyze datasets** with (reasonably) user-friendly interfaces, and get introduced to **command-line tools** for next generation sequencing (NGS)

Not in this course

1. Sequence assembly
2. Bisulphite sequencing
3. Protein sequence analysis
4. Metagenomics

Bioinformatics Sequence Analysis

1. Introduction to sequence analysis
2. Sequencing techniques
3. Brief introduction Linux and R (self study)
4. NGS pre-processing
5. (Multiple) sequence alignment
6. Case: Neuroblastoma
7. Introduction to R2
8. Exome sequence analysis
9. RNAseq
10. Single cell RNA sequencing

The focus is on human data, but many techniques are also applicable to other organisms

Practical things

Certificate

- Attend all sessions (one day can be skipped, ask for possibility for self-study)
- Active participation

Course material

- Slides and exercises are published on `https://bioinformatics.amc.nl/`

In this hour

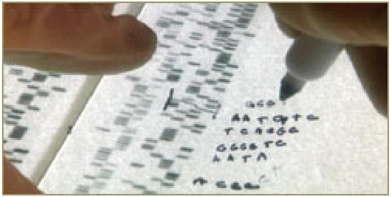
Introduction

You will get an indication about the **scale** of sequence data, how to **handle the data**, where to find **publicly available data and tools**, and what can be done with **NGS**

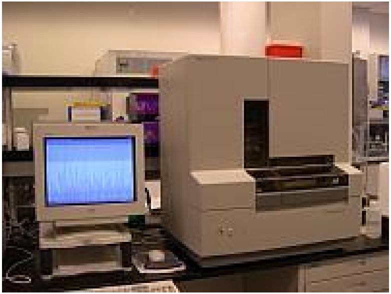
Overview

1. Welcome
2. Scale of sequence data
   - DNA sequencing
   - Genome projects
3. Bioinformatics databases and tools
   - Databases
   - Sequence analysis
4. Handling sequence data
   - Computing
   - Application areas

Sanger

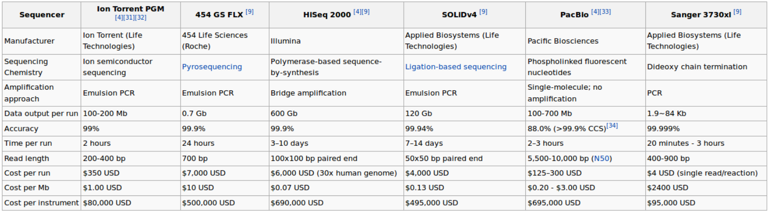


Automated sequencing



Sequencing centers

Next generation sequencing



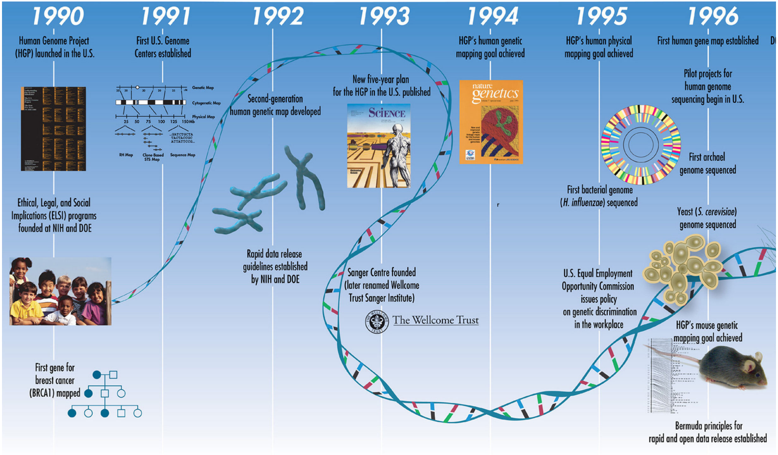
Sequencer	Ion Torrent PGM [4][31][32]	454 GS FLX [9]	HiSeq 2000 [4][9]	SOLiDv4 [9]	PacBio [4][33]	Sanger 3730xl [9]
Manufacturer	Ion Torrent (Life Technologies)	454 Life Sciences (Roche)	Illumina	Applied Biosystems (Life Technologies)	Pacific Biosciences	Applied Biosystems (Life Technologies)
Sequencing Chemistry	Ion semiconductor sequencing	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing	Phospholinked fluorescent nucleotides	Dideoxy chain termination
Amplification approach	Emulsion PCR	Emulsion PCR	Bridge amplification	Emulsion PCR	Single-molecule; no amplification	PCR
Data output per run	100-200 Mb	0.7 Gb	600 Gb	120 Gb	100-700 Mb	1.9~84 Kb
Accuracy	99%	99.9%	99.9%	99.94%	88.0% (>99.9% CCS)[34]	99.999%
Time per run	2 hours	24 hours	3-10 days	7-14 days	2-3 hours	20 minutes - 3 hours
Read length	200-400 bp	700 bp	100x100 bp paired end	50x50 bp paired end	5,500-10,000 bp (N50)	400-900 bp
Cost per run	$350 USD	$7,000 USD	$6,000 USD (30x human genome)	$4,000 USD	$125-300 USD	$4 USD (single read/reaction)
Cost per Mb	$1.00 USD	$10 USD	$0.07 USD	$0.13 USD	$0.20 - $3.00 USD	$2400 USD
Cost per instrument	$80,000 USD	$500,000 USD	$690,000 USD	$495,000 USD	$695,000 USD	$95,000 USD

Table 1. Comparing metrics and performance of next-generation DNA sequencers.[35]

Genome projects

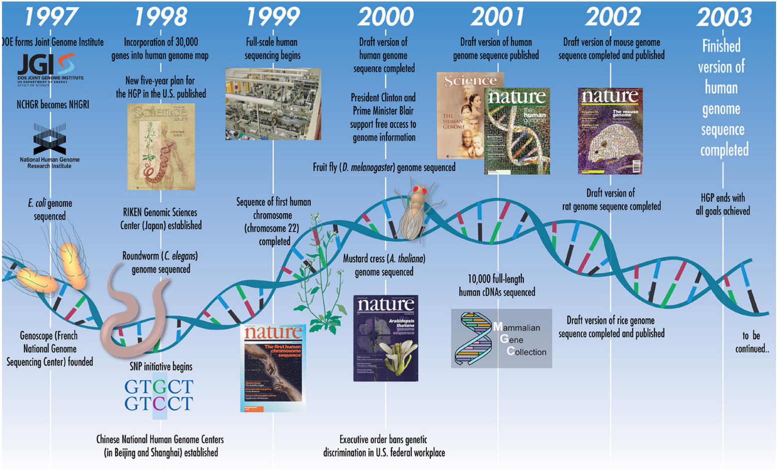
- HGP
- 1000g
- UK10K >100K genomes
- Personal genomes

Human Genome Project


1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
Genetic Map
Cytogenetic Map
Physical Map
RH Map
Clone-Based STS Map
Sequence Map
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the U.S. HGP published
SCIENCE
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
D

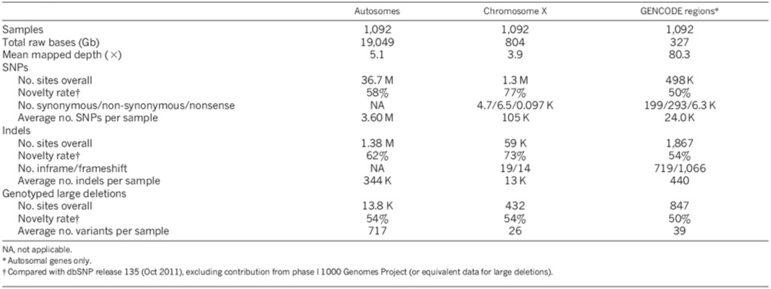
http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml

Human Genome Project


1997
DOE forms Joint Genome Institute
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
Science
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
The first human chromosome sequence
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Arabidopsis thaliana genome sequence
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
THE HUMAN GENOME
nature
the human genome
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
The mouse genome
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued..

http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml

1000 genomes project



	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K
Novelty rate†	58%	77%	50%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K	199/293/6.3 K
Average no. SNPs per sample	3.60 M	105 K	24.0 K
Indels			
No. sites overall	1.38 M	59 K	1,867
Novelty rate†	62%	73%	54%
No. inframe/frameshift	NA	19/14	719/1,066
Average no. indels per sample	344 K	13 K	440
Genotyped large deletions			
No. sites overall	13.8 K	432	847
Novelty rate†	54%	54%	50%
Average no. variants per sample	717	26	39

NA, not applicable.

* Autosomal genes only.

† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

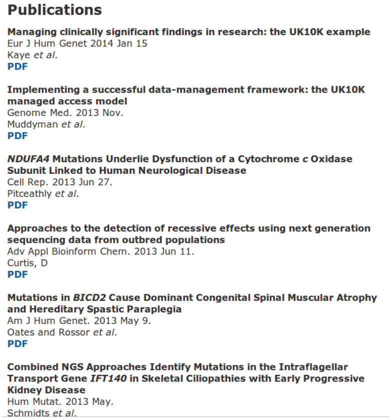
http://www.1000genomes.org/

UK10K

4000 genomes

6000 exomes

http://www.uk10k.org/



Publications

Managing clinically significant findings in research: the UK10K example
Eur J Hum Genet 2014 Jan 15
Kaye *et al.*
PDF

Implementing a successful data-management framework: the UK10K managed access model
Genome Med. 2013 Nov.
Muddyman *et al.*
PDF

***NDUFA4* Mutations Underlie Dysfunction of a Cytochrome *c* Oxidase Subunit Linked to Human Neurological Disease**
Cell Rep. 2013 Jun 27.
Pitceathly *et al.*
PDF

Approaches to the detection of recessive effects using next generation sequencing data from outbred populations
Adv Appl Bioinform Chem. 2013 Jun 11.
Curtis, D
PDF

Mutations in *BICD2* Cause Dominant Congenital Spinal Muscular Atrophy and Hereditary Spastic Paraplegia
Am J Hum Genet. 2013 May 9.
Oates and Rossor *et al.*
PDF

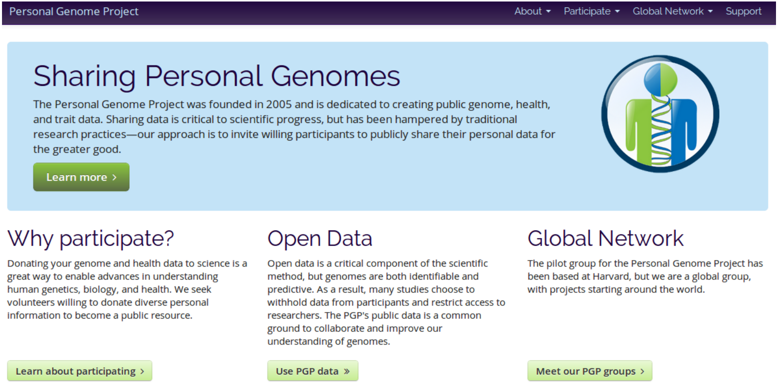
Combined NGS Approaches Identify Mutations in the Intraflagellar Transport Gene *IFT140* in Skeletal Ciliopathies with Early Progressive Kidney Disease
Hum Mutat. 2013 May.
Schmidts *et al.*

The 100K genomes project

Genomics england

The project will focus on patients with a rare disease and their families and patients with cancer. The first samples for sequencing are being taken from patients living in England with discussions taking place with Scotland, Wales and Northern Ireland about potential future involvement.
http://www.genomicsengland.co.uk/

Personal genomes



Sharing Personal Genomes

The Personal Genome Project was founded in 2005 and is dedicated to creating public genome, health, and trait data. Sharing data is critical to scientific progress, but has been hampered by traditional research practices—our approach is to invite willing participants to publicly share their personal data for the greater good.

Why participate?

Donating your genome and health data to science is a great way to enable advances in understanding human genetics, biology, and health. We seek volunteers willing to donate diverse personal information to become a public resource.

Open Data

Open data is a critical component of the scientific method, but genomes are both identifiable and predictive. As a result, many studies choose to withhold data from participants and restrict access to researchers. The PGP's public data is a common ground to collaborate and improve our understanding of genomes.

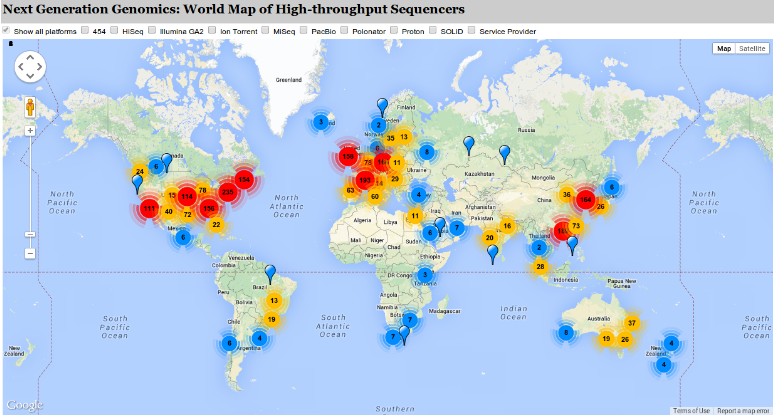
Global Network

The pilot group for the Personal Genome Project has been based at Harvard, but we are a global group, with projects starting around the world.

100,000 genomes plus medical records

`http://www.personalgenomes.org/`

Sequencers around the world

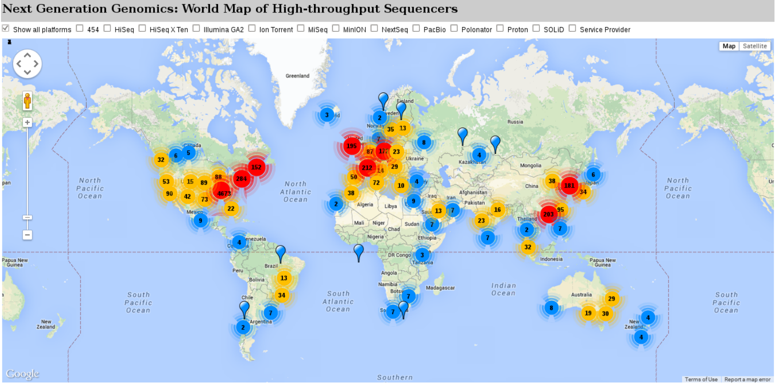


Next Generation Genomics: World Map of High-throughput Sequencers

☑ Show all platforms ☐ 454 ☐ HiSeq ☐ Illumina GA2 ☐ Ion Torrent ☐ MiSeq ☐ PacBio ☐ Polonator ☐ Proton ☐ SOLiD ☐ Service Provider

http://omicsmaps.com/

Sequencers around the world 2015

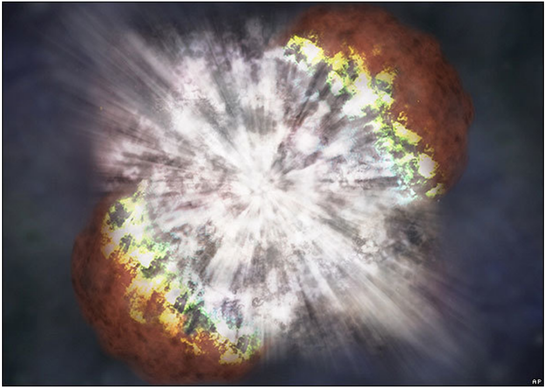


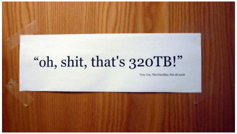
Next Generation Genomics: World Map of High-throughput Sequencers

☑ Show all platforms ☐ 454 ☐ HiSeq ☐ HiSeq X Ten ☐ Illumina GA2 ☐ Ion Torrent ☐ MiSeq ☐ MinION ☐ NextSeq ☐ PacBio ☐ Polonator ☐ Proton ☐ SOLiD ☐ Service Provider

`http://omicsmaps.com/`

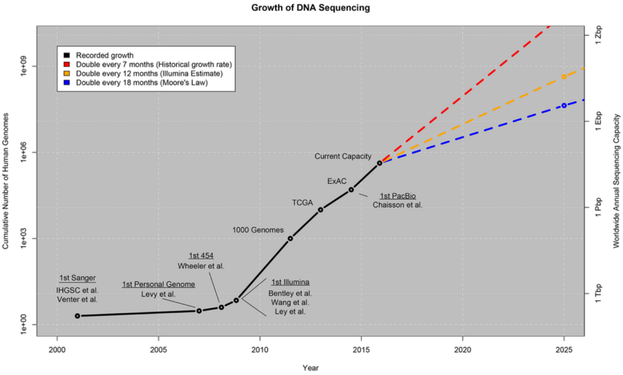
Big data




“oh, shit, that's 320TB!”

Tony Cox, The Guardian, Feb 28 2008

DNA sequencing rate

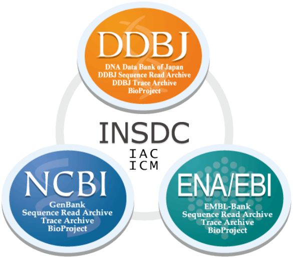


Stephens et al. (2015) PLoS One

GenBank, EMBL and DDBJ

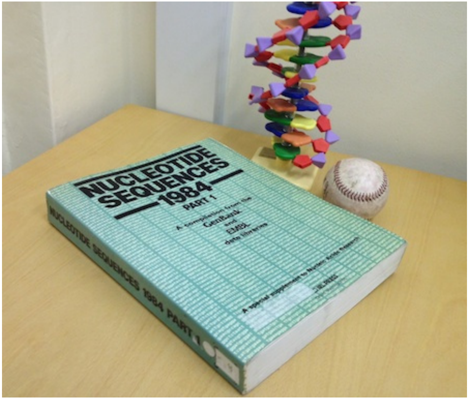
International Nucleotide Sequence Database Collaboration

Daily exchange of sequence data



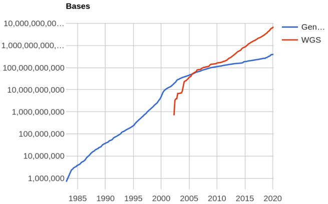
```
https://www.ncbi.nlm.nih.gov/
https://www.ebi.ac.uk/
http://www.ddbj.nig.ac.jp/
```

Nucleotide sequence databases


NUCLEOTIDE SEQUENCES 1984
PART 1
A compilation from the GenBank and EMBL data libraries
NUCLEOTIDE SEQUENCES 1984 PART 1

From: `http://www.davelunt.net/`

GenBank



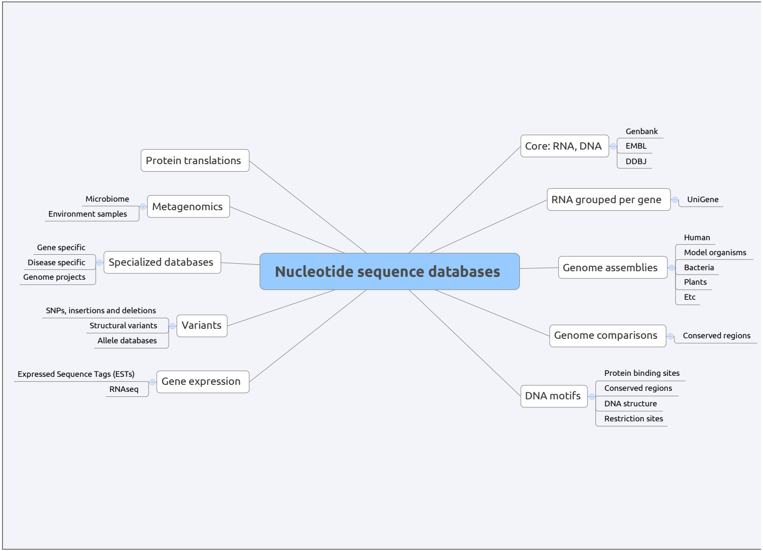
https://www.ncbi.nlm.nih.gov/genbank/statistics/

GenBank has doubled approximately every 18 months

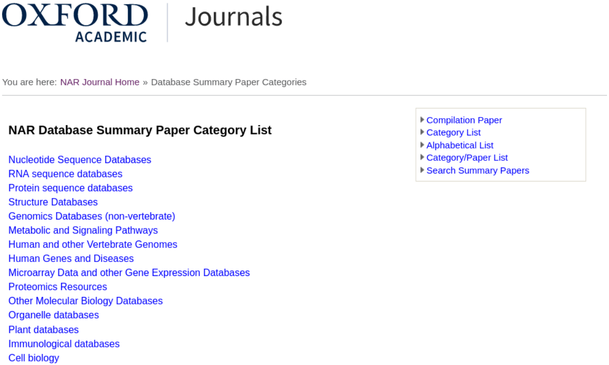
Release 236 (Feb 2020)

has 399,376,854,872 base pairs from 216,214,215 sequences. In addition, there are 1,206,720,688 WGS records containing 6,968,991,265,752 base pairs of sequence data.

Core databases and derivatives



Where to start?



OXFORD ACADEMIC | Journals

You are here: NAR Journal Home » Database Summary Paper Categories

NAR Database Summary Paper Category List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

- ▸ Compilation Paper
- ▸ Category List
- ▸ Alphabetical List
- ▸ Category/Paper List
- ▸ Search Summary Papers

`https://www.oxfordjournals.org/nar/database/c/`

Sequence analysis

Sequence alignment

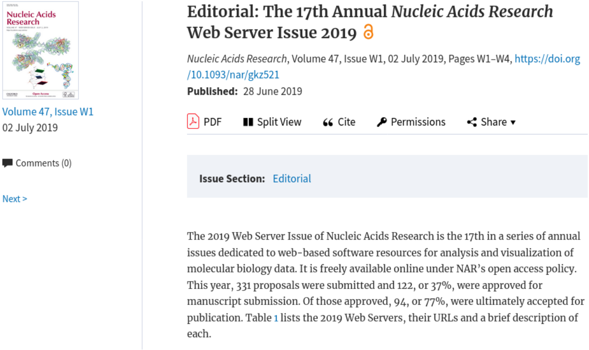
- Needleman-Wunsch
- Smith-Waterman
- BLAST
- BLAT
- ClustalW
- BWA, BFAST, Bowtie, Tophat, etc, etc

Sequence suites/packages

- Emboss package
- CLCbio workbench
- Galaxy
- R Bioconductor

Hundreds of tools to analyse sequence data...

Tools



Volume 47, Issue W1

02 July 2019

Comments (0)

Next >

Editorial: The 17th Annual *Nucleic Acids Research* Web Server Issue 2019

Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W1–W4, https://doi.org/10.1093/nar/gkz521

Published: 28 June 2019

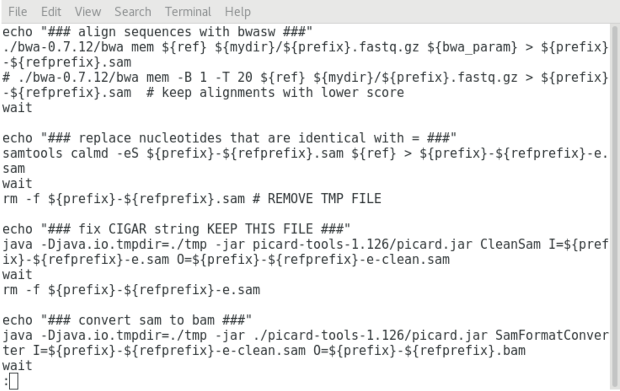
Issue Section: Editorial

The 2019 Web Server Issue of Nucleic Acids Research is the 17th in a series of annual issues dedicated to web-based software resources for analysis and visualization of molecular biology data. It is freely available online under NAR's open access policy. This year, 331 proposals were submitted and 122, or 37%, were approved for manuscript submission. Of those approved, 94, or 77%, were ultimately accepted for publication. Table 1 lists the 2019 Web Servers, their URLs and a brief description of each.

`https://academic.oup.com/nar/article/47/W1/W1/5524725`

Tools

Most tools are only available via the command-line (on linux systems)



```
echo "### align sequences with bwasw ###"
./bwa-0.7.12/bwa mem ${ref} ${mydir}/${prefix}.fastq.gz ${bwa_param} > ${prefix}
-${refprefix}.sam
# ./bwa-0.7.12/bwa mem -B 1 -T 20 ${ref} ${mydir}/${prefix}.fastq.gz > ${prefix}
-${refprefix}.sam  # keep alignments with lower score
wait

echo "### replace nucleotides that are identical with = ###"
samtools calmd -eS ${prefix}-${refprefix}.sam ${ref} > ${prefix}-${refprefix}-e.
sam
wait
rm -f ${prefix}-${refprefix}.sam # REMOVE TMP FILE

echo "### fix CIGAR string KEEP THIS FILE ###"
java -Djava.io.tmpdir=./tmp -jar picard-tools-1.126/picard.jar CleanSam I=${pref
ix}-${refprefix}-e.sam O=${prefix}-${refprefix}-e-clean.sam
wait
rm -f ${prefix}-${refprefix}-e.sam

echo "### convert sam to bam ###"
java -Djava.io.tmpdir=./tmp -jar ./picard-tools-1.126/picard.jar SamFormatConver
ter I=${prefix}-${refprefix}-e-clean.sam O=${prefix}-${refprefix}.bam
wait
:
```

Open source

Free as in freedom

You can use, change, integrate, and review the code

Open source allows sharing and promotes collaboration

No vendor lock-in

Open source

- Software
- Databases
- Journals
- Standards

https://en.wikipedia.org/wiki/Open_source

- Hardware
- Art
- Money
- Drinks
- Medicine
- Fashion
- Education

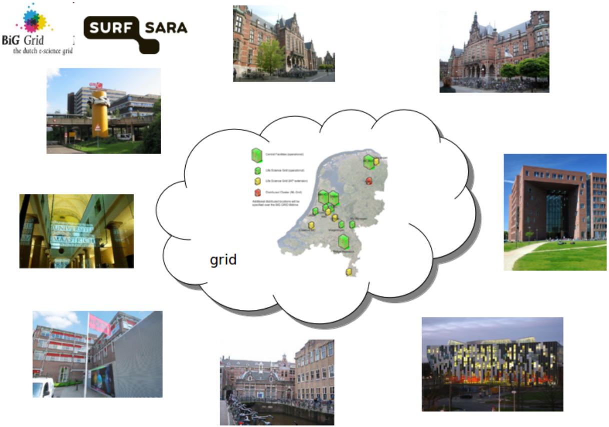
Handling sequence data

PC

Small cluster

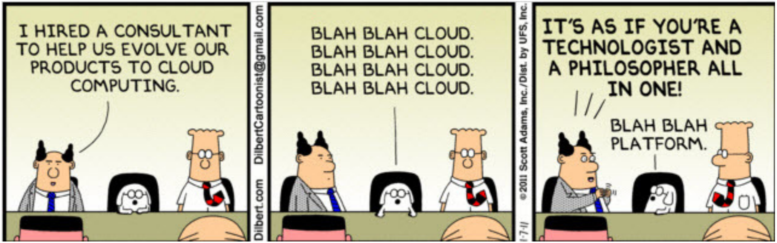
Buy a bigger cluster (centralized model)

Dutch life science grid



http://surfsara.nl/

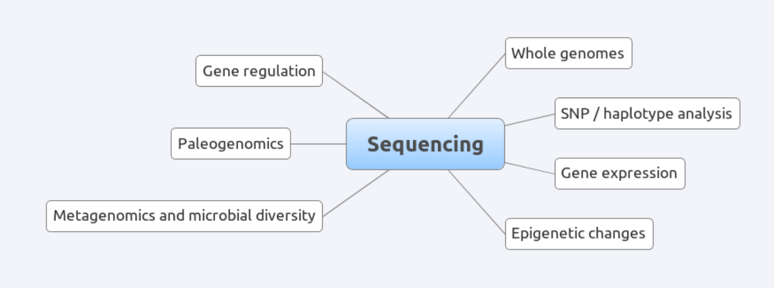
Cloud computing


I HIRED A CONSULTANT TO HELP US EVOLVE OUR PRODUCTS TO CLOUD COMPUTING.
Dilbert.com DilbertCartoonist@gmail.com
BLAH BLAH CLOUD. BLAH BLAH CLOUD. BLAH BLAH CLOUD. BLAH BLAH CLOUD.
1-7-11 ©2011 Scott Adams, Inc./Dist. by UFS, Inc.
IT'S AS IF YOU'RE A TECHNOLOGIST AND A PHILOSOPHER ALL IN ONE!
BLAH BLAH PLATFORM.

HPC cloud at SurfSara

You will use a linux environment that runs on the HPC cloud to get acquainted with command-line tools

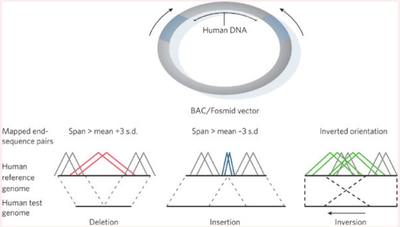
NGS application areas



Whole genomes

- De novo sequencing
- Re-sequencing
- Copy number variations
- Rearrangements
- New insertions/deletions/mutations

Structural variation



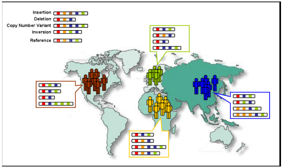
Genomic libraries are constructed from fragmented DNA and subcloned into circular vectors such as BACs or fosmids. The ends of these fragment inserts are directly sequenced from universal vector primers near the subcloning site (arrows) and are termed end-sequence pairs or paired-end sequences. End-sequence pairs are mapped to their best location in the human reference genome sequence assembly. End-sequence pairs that are discordant in terms of length (> 3 s.d. from the mean insert length) and/or orientation when mapped against the reference genome assembly may be indicative of deletions, insertions or inversion, as indicated (red, blue and green, respectively). End-sequence pairs consistent in terms of length and orientation are shown as grey.

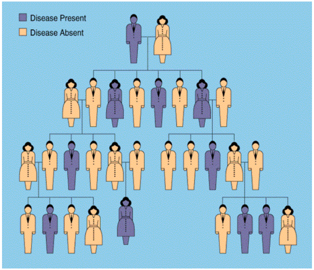
The Human Genome Structural Variation Working Group, Nature 2007

SNP / haplotype analysis

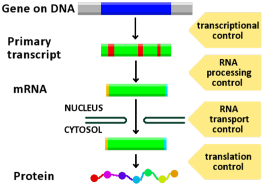
Linkage studies

Forensic research





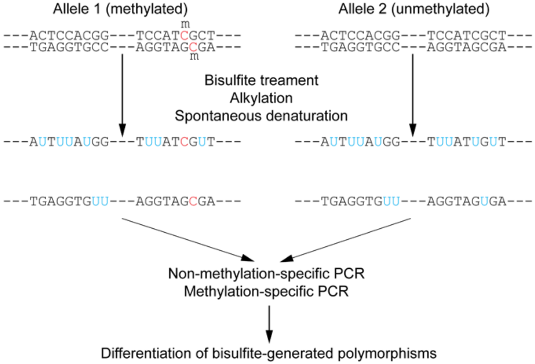
Gene expression



https://en.wikipedia.org/wiki/Regulation_of_gene_expression

- Full-length transcripts
- EST sequencing
- 5’ transcript ends (5’-RATE, CAGE)
- SAGE ditag sequencing
- SAGE-like 3’ end sequencing
- Nebulized fragments
- ncRNA sequencing

Epigenetics



Allele 1 (methylated)

```
            m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
                     m
```

Allele 2 (unmethylated)

```
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
```

Bisulfite treament
Alkylation
Spontaneous denaturation

```
---AUTUUAUGG---TUUATCGUT---        ---AUTUUAUGG---TUUATUGUT---

---TGAGGTGUU---AGGTAGCGA---        ---TGAGGTGUU---AGGTAGUGA---
```

Non-methylation-specific PCR
Methylation-specific PCR

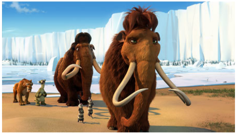
Differentiation of bisulfite-generated polymorphisms

Treatment with sodium bisulfite
Unmethylated cytosines change into uracil
Methylated cytosines are unchanged
Compare sequences with reference sequence

Metagenomics and microbial diversity

Study genomic content in a complex mixture of microorganisms
(bacteria or viruses in some environment)
Identify new species

Paleogenomics



Sequencing of ancient DNA

Mummies

Sabretooth

Mammoth

Neanderthal

Gene regulation

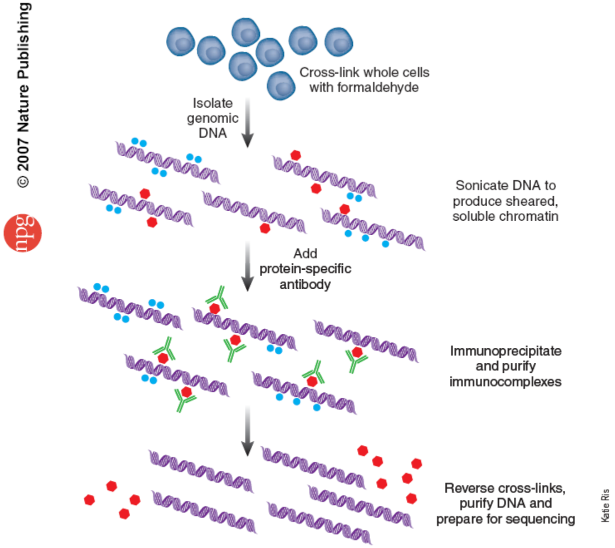
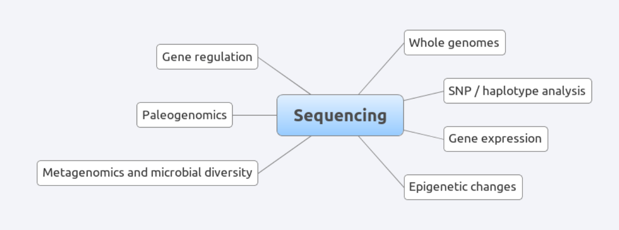

© 2007 Nature Publishing
npg
Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing
Katie Ris

Figure 1 | Workflow of Chip-seq. DNA and proteins are cross-linked and purified; then bound DNA is analyzed by massively parallel short-read sequencing.

Sequence analysis



Usually starts with sequence alignment or sequence assembly

Depending on the application other tools/methods are used or developed

With a click of a button...

.. or perhaps not. You will find out during this course.
Computer exercises sequence analysis:

1. Via web tools
2. Creating pipelines online
3. With command-line tools in a Linux environment

Bioinformatics Sequence Analysis

1. Introduction to sequence analysis
2. Sequencing techniques
3. Brief introduction Linux and R (self study)
4. NGS pre-processing
5. (Multiple) sequence alignment
6. Case: Neuroblastoma
7. Introduction to R2
8. Exome sequence analysis
9. RNAseq
10. Single cell RNA sequencing