Introduction to DNA sequencing data analysis

Sina Majidian

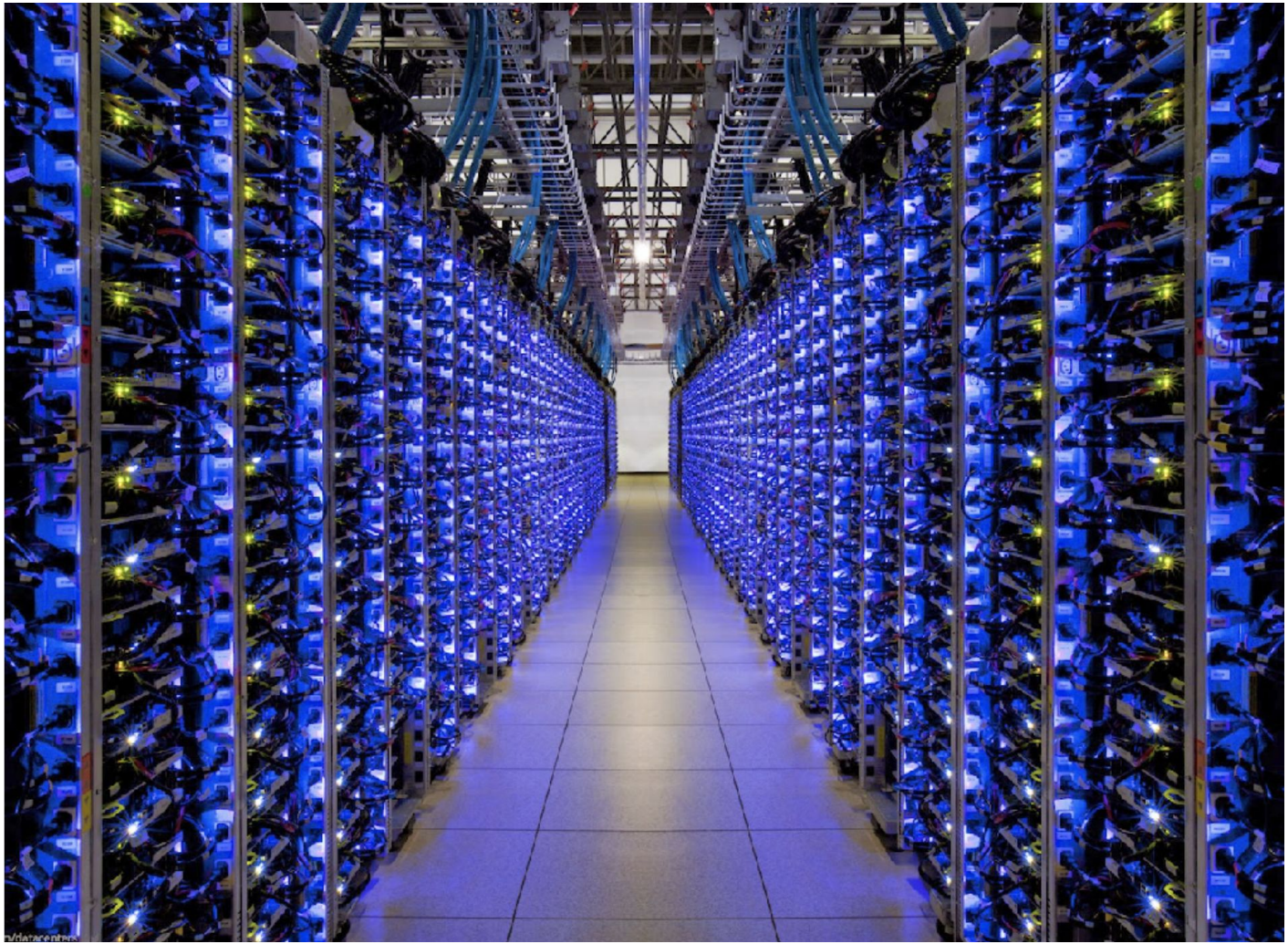
Postdoctoral Fellow at the University of Lausanne.

SIB Member, Swiss Institute of Bioinformatics.

25 Nov. 2021

UNIL | Université de Lausanne

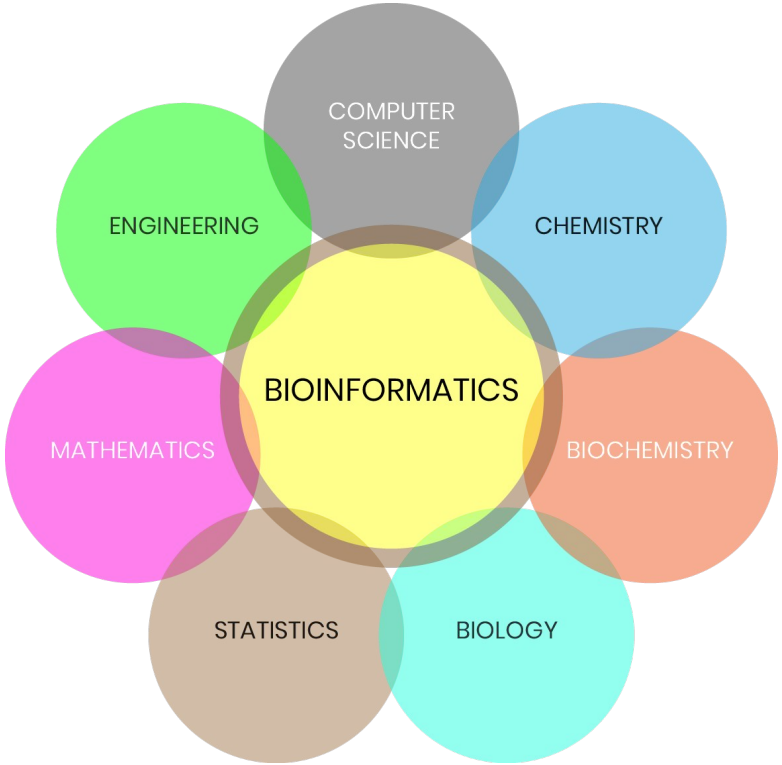
Swiss Institute of Bioinformatics



Learning objectives

- How machines sequence DNA?
- How to store data?
- How to download data?
- How to visualize and manipulate data?

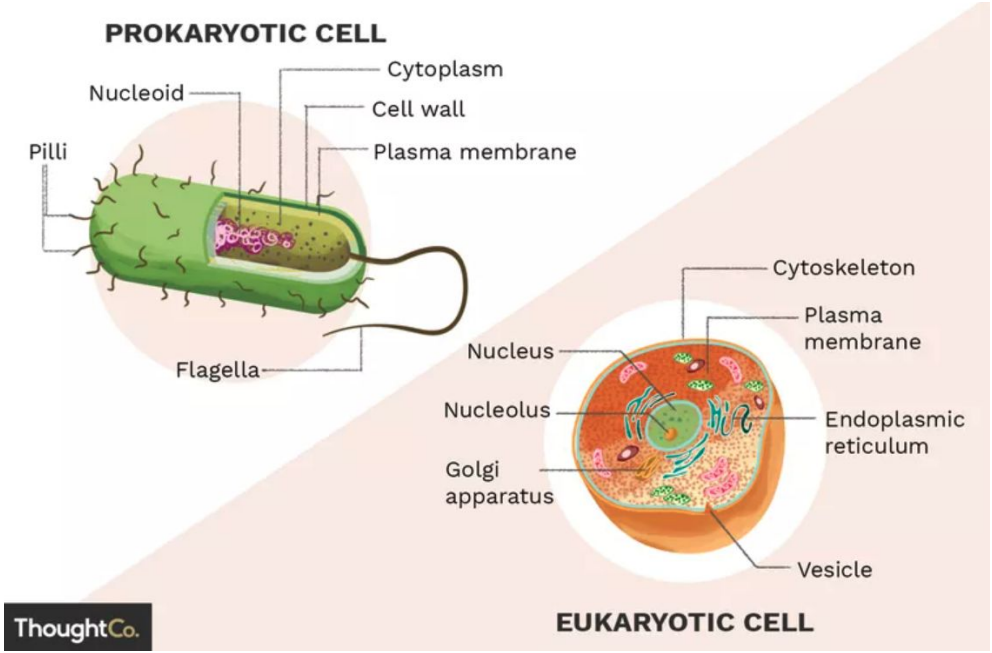
If you are participating with phone, we need computer for hands-on exercises!



Your background?

Cell

The basic unit from which a living organism is made.



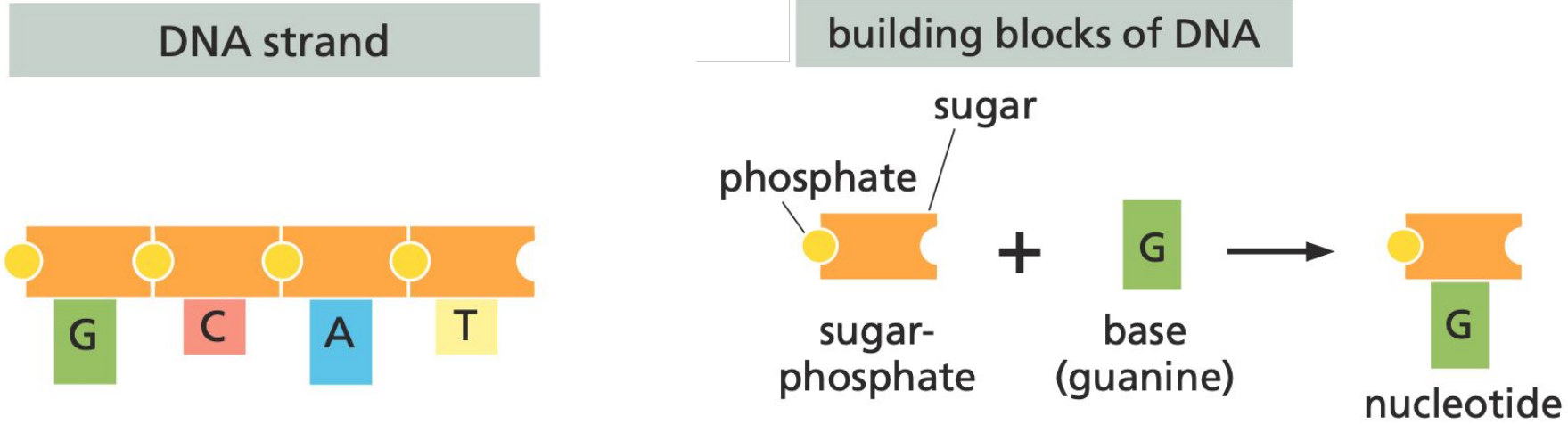
Genome

“The complete set of genes or genetic material present in a cell or organism.”

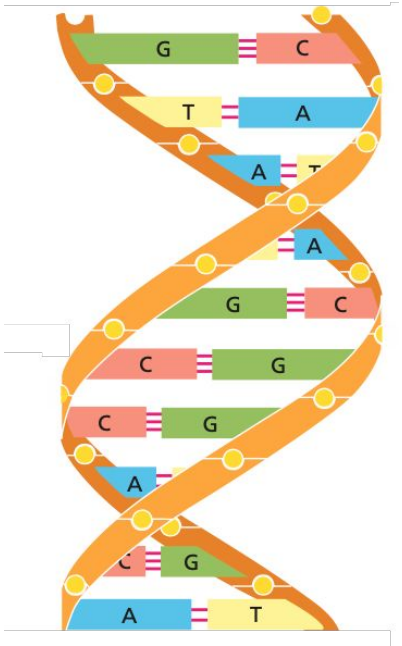
Oxford dictionaries

- containing the information needed to maintain organism’s living.
- is made of DNA (or RNA in some viruses).

DNA: the code of life



DNA double helix

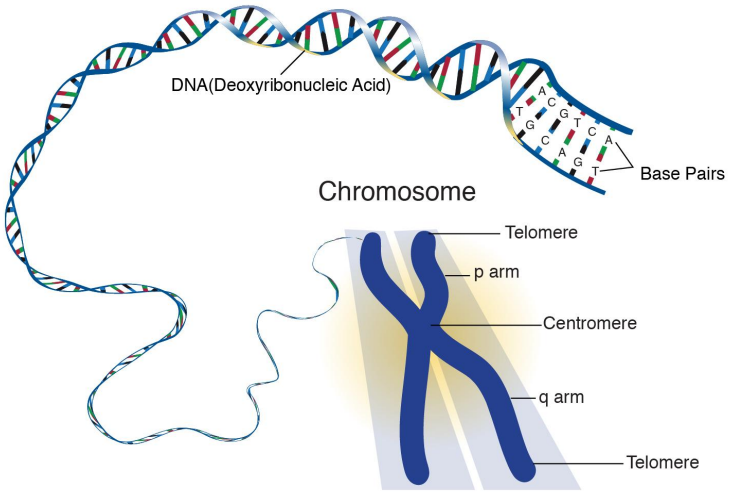


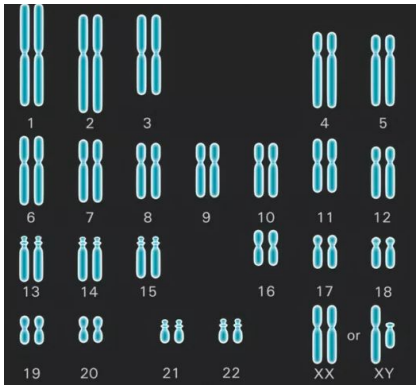
Genome size

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Chromosome

The DNA molecule is packaged into thread-like structures called chromosomes.





genome.gov/genetics-glossary/Chromosome
thoughtco.com

Google ucsc genome browser

All Images Videos News Books More Tools

About 2'920'000 results (1.08 seconds)

https://genome.ucsc.edu

UCSC Genome Browser Home

The **UCSC Genome Browser** is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UCSC Genomics Institute.

Browser Gateway
As with the previous GRCh37 assembly, the Genome ...

Human BLAT Search
Genome: Search all, Assembly: Query type: Sort output: Output ...

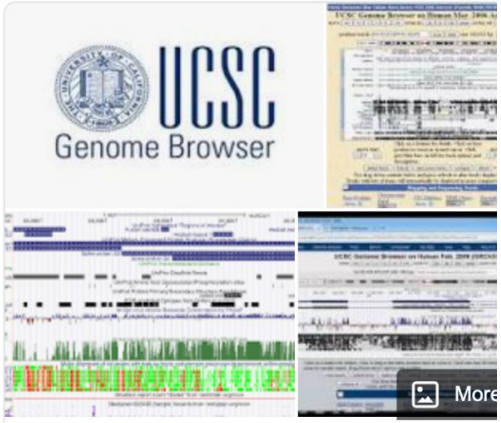
Genome Browser
Use drop-down controls below and press refresh to alter tracks ...

hg19/GRCh37
Use drop-down controls below and press refresh to alter tracks ...

In-Silico PCR
In-Silico PCR searches a sequence database with a pair ...

Table Browser
Use this tool to retrieve and export data from the Genome Browser ...

More results from ucsc.edu »



UCSC Genome Browser

The UCSC Genome Browser is an online and downloadable genome browser hosted by the University of California, Santa Cruz. Wikipedia

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute UCSC **Genome Browser**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Our tools

- **Genome Browser** interactively visualize genomic data
- **COVID-19 Research** use the SARS-CoV-2 genome browser and explore coronavirus datasets
- **BLAT** rapidly align sequences to the genome
- **Table Browser** download data from the Genome Browser database
- **Variant Annotation Integrator** get functional effect predictions for variant calls
- **Data Integrator** combine data sources from the Genome Browser database
- **Genome Browser in a Box (GBiB)** run the Genome Browser on your laptop or server
- **In-Silico PCR** rapidly align PCR primer pairs to the genome
- **LiftOver** convert genome coordinates between assemblies
- **Track Hubs** import and view external data tracks
- **REST API** returns data in JSON format

More tools...

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly assembled genome was released on the web at http://genome.ucsc.edu, along with the initial prototype of a graphical viewing tool, the UCSC Genome Browser. In the ensuing years, the website has grown to include a broad collection of vertebrate and model organism assemblies and annotations, along with a large suite of tools for viewing, analyzing and downloading data. Learn more about our history on the **UCSC Genome Browser Project History** page and by watching this **video**.

What's new

Nov. 23, 2021 - **New blog post about GenArk hubs**

Nov. 18, 2021 - **New clinical rare disease track - Orphadata**

Nov. 17, 2021 - **New Single-Cell track group and data for hg38**

More news...

Subscribe

*The UCSC Genome Browser is developed and maintained by the **Genome Bioinformatics Group**, a cross-departmental team within the **UCSC Genomics Institute**.*

You might want to navigate to your nearest mirror - genome-euro.ucsc.edu

- User settings (sessions and custom tracks) **will differ** between sites. Read more.
- Take me to genome-euro.ucsc.edu
- Let me stay here genome.ucsc.edu

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute — UCSC **Genome Browser Gateway**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Browse/Select Species

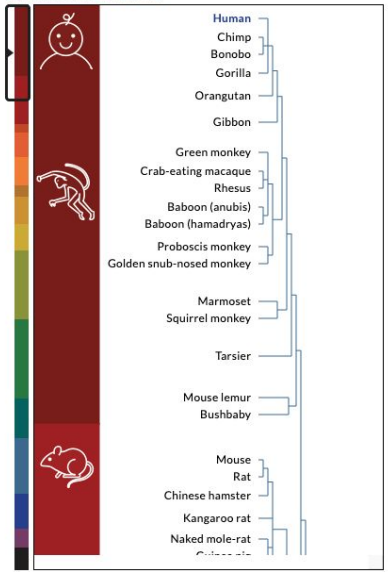
POPULAR SPECIES

Human | Mouse | Rat | Zebrafish | Fruitfly | Worm | Yeast

Enter species, common name or assembly ID

Can't find a genome assembly?

REPRESENTED SPECIES



Find Position

Human Assembly: Dec. 2013 (GRCh38/hg38)

GO

Position/Search Term

Enter position, gene symbol or search terms

Current position: chrX:15,560,138-15,602,945

Human Genome Browser - hg38 assembly

view sequences

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p13 (GCA_000001405.28)
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 13
Assembly accession: GCA_000001405.28
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: GCF_000001405.39 (GRCh38.p13, GCA_000001405.28)
BioProject ID: PRJNA31257

Homo sapiens
(Graphic courtesy of CBSE)

Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- **By gene name:** Type a gene name into the "search term" box, choose your gene from the drop-down list, then press "submit" to go directly to the assembly location associated with that gene. More information.
- **By track type:** Click the "track search" button to find Genome Browser tracks that match specific selection criteria. More information.

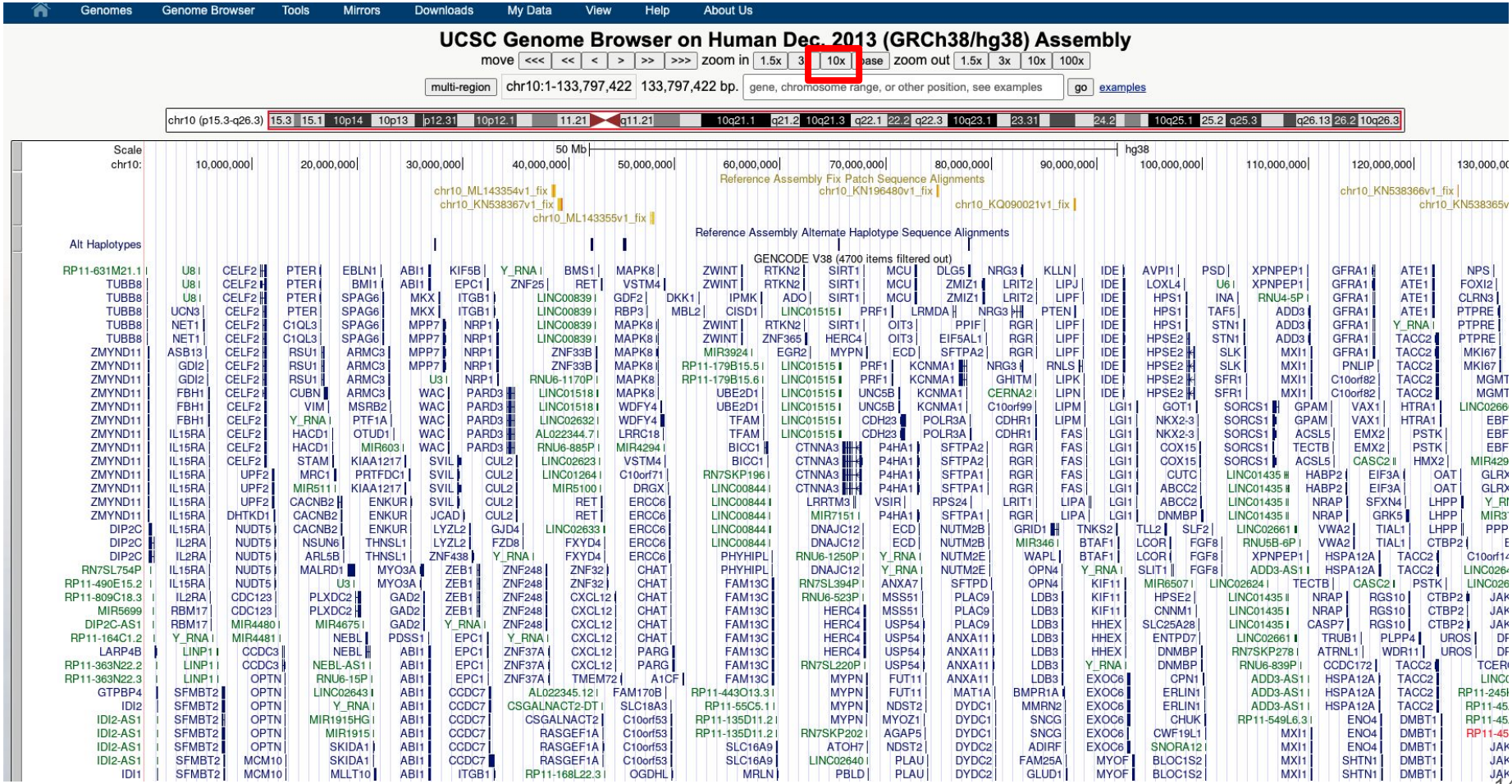
Download sequence and annotation data:

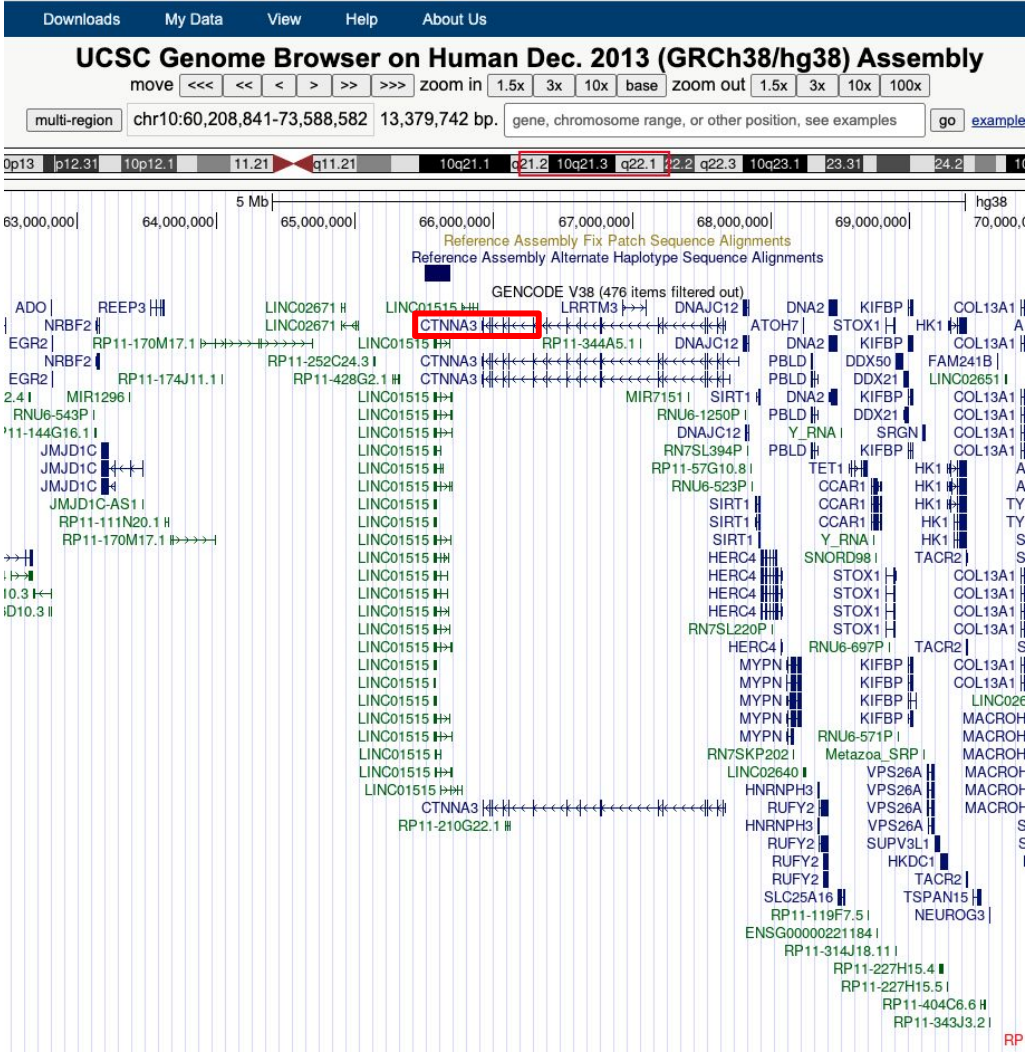
- Using rsync (recommended)
- Using HTTP
- Using FTP
- Data use conditions and restrictions
- Acknowledgments

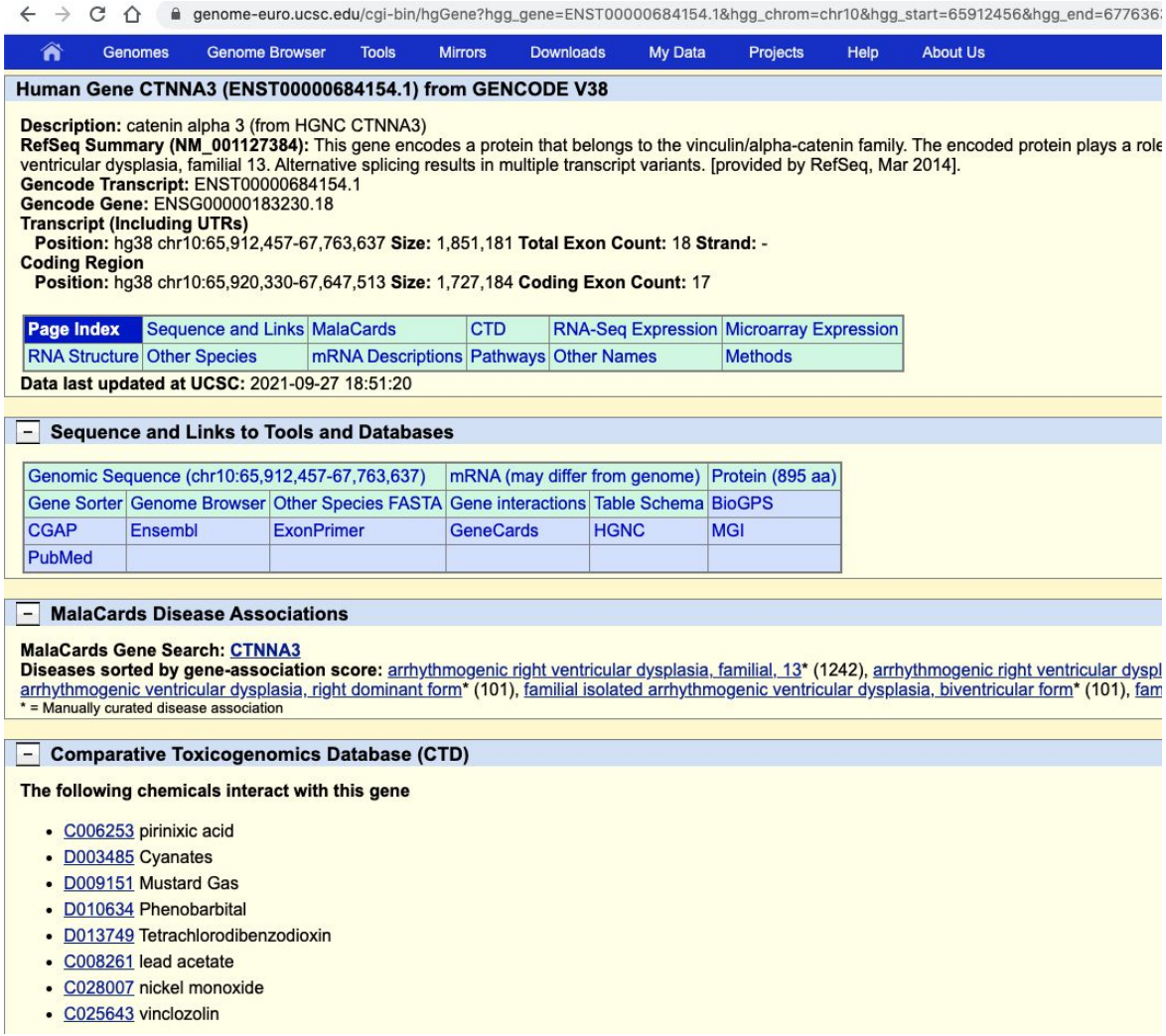
Assembly Details

The GRCh38 assembly is the first major revision of the human genome released in more than four years. As with the previous GRCh37 assembly, the Genome Reference Consortium (GRC) is now the primary source for human genome assembly data submitted to GenBank. Beginning with this release, the UCSC Genome Browser version numbers for the human assemblies now match those of the GRC to minimize version confusion. Hence, the GRCh38 assembly is referred to as "hg38" in the Genome Browser datasets and documentation. For a glossary of assembly-related terms, see the GRC Assembly Terminology page.

Type: chr10


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move
zoom in
zoom out
multi-region
chr10:1-133,797,422
133,797,422 bp.
gene, chromosome range, or other position, see examples
go
examples
chr10 (p15.3-q26.3)
Scale
chr10:
50 Mb
hg38
Reference Assembly Fix Patch Sequence Alignments
chr10_ML143354v1_fix
chr10_KN538367v1_fix
chr10_ML143355v1_fix
chr10_KN196480v1_fix
chr10_KQ090021v1_fix
chr10_KN538366v1_fix
chr10_KN538365v
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE V38 (4700 items filtered out)

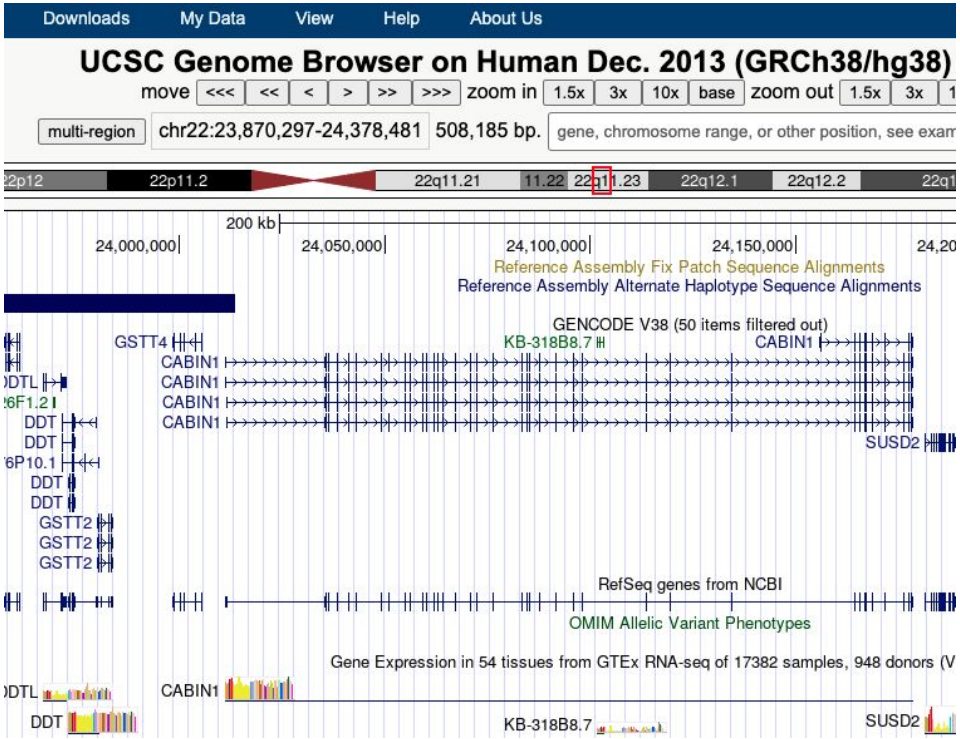

Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move
zoom in
zoom out
multi-region
chr10:60,208,841-73,588,582
13,379,742 bp.
gene, chromosome range, or other position, see examples
go
example
5 Mb
hg38
63,000,000
64,000,000
65,000,000
66,000,000
67,000,000
68,000,000
69,000,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V38 (476 items filtered out)
CTNNA3
RP11-210G22.1


genome-euro.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000684154.1&hgg_chrom=chr10&hgg_start=65912456&hgg_end=677636
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Projects
Help
About Us
Human Gene CTNNA3 (ENST00000684154.1) from GENCODE V38
Description: catenin alpha 3 (from HGNC CTNNA3)
RefSeq Summary (NM_001127384): This gene encodes a protein that belongs to the vinculin/alpha-catenin family. The encoded protein plays a role
ventricular dysplasia, familial 13. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Mar 2014].
Gencode Transcript: ENST00000684154.1
Gencode Gene: ENSG00000183230.18
Transcript (Including UTRs)
Position: hg38 chr10:65,912,457-67,763,637 Size: 1,851,181 Total Exon Count: 18 Strand: -
Coding Region
Position: hg38 chr10:65,920,330-67,647,513 Size: 1,727,184 Coding Exon Count: 17
Page Index
Sequence and Links
MalaCards
CTD
RNA-Seq Expression
Microarray Expression
RNA Structure
Other Species
mRNA Descriptions
Pathways
Other Names
Methods
Data last updated at UCSC: 2021-09-27 18:51:20
Sequence and Links to Tools and Databases
Genomic Sequence (chr10:65,912,457-67,763,637)
mRNA (may differ from genome)
Protein (895 aa)
Gene Sorter
Genome Browser
Other Species FASTA
Gene interactions
Table Schema
BioGPS
CGAP
Ensembl
ExonPrimer
GeneCards
HGNC
MGI
PubMed
MalaCards Disease Associations
MalaCards Gene Search: CTNNA3
Diseases sorted by gene-association score: arrhythmogenic right ventricular dysplasia, familial, 13* (1242), arrhythmogenic right ventricular dysp
arrhythmogenic ventricular dysplasia, right dominant form* (101), familial isolated arrhythmogenic ventricular dysplasia, biventricular form* (101), fam
* = Manually curated disease association
Comparative Toxicogenomics Database (CTD)
The following chemicals interact with this gene
C006253 pirinixic acid
D003485 Cyanates
D009151 Mustard Gas
D010634 Phenobarbital
D013749 Tetrachlorodibenzodioxin
C008261 lead acetate
C028007 nickel monoxide
C025643 vinclozolin


Phenotype and Literature
refresh
OMIM Alleles
dense
CADD
hide
Cancer Gene Expr
hide
ClinGen
hide
Deprecated ClinGen CNVs
hide
ClinVar Variants
hide
Coriell CNVs
hide
COSMIC Regions
hide
Development Delay
hide
Gene Interactions
hide
GeneReviews
hide
GWAS Catalog
hide
HGMD Variants
hide
LOVD Variants
hide
OMIM Cyto Loci
hide
OMIM Genes
hide
New Orphanet
hide
New REVEL Scores
hide
SNPedia
hide
TCGA Pan-Cancer
hide
UniProt Variants
hide
Variants in Papers
hide
COVID-19
refresh
COVID GWAS v4
hide
COVID GWAS v3
hide
Rare Harmful Vars
hide
Single Cell RNA-seq
refresh
New Colon Wang
hide
New Ileum Wang
hide
New Rectum Wang
hide
New Blood (PBMC) Hao
hide
New Cortex Velmeshev
hide
New Fetal Gene Atlas
hide
New Heart Cell Atlas
hide
New Kidney Stewart
hide
New Liver MacParland
hide
New Lung Travaglini
hide
New Muscle De Micheli
hide
New Pancreas Baron
hide
New Placenta Vento-Tormo
hide
New Skin Sole-Boldo
hide
mRNA and EST
refresh
Human ESTs
hide
Human mRNAs
hide
Other ESTs
hide
Other mRNAs
hide
SIB Alt-Splicing
hide
Spliced ESTs
hide
Expression
refresh
GTEx Gene V8
pack
GTEx RNA-Seq Coverage
hide
New Affy Archive
hide
EPDnew Promoters
hide
GNF Atlas 2
hide
GTEx Gene
hide
GTEx Transcript
hide
GWIPS-viz Riboseq
hide
miRNA Tissue Atlas
hide
Regulation
refresh
ENCODE cCREs
dense
ENCODE Regulation
show
CpG Islands
hide
GeneHancer
hide
GTEx cis-eQTLs
hide
Hi-C and Micro-C
hide
New JASPAR Transcription Factors
hide
ORegAnno
hide
RefSeq Func Elems
hide
Comparative Genomics
refresh
Conservation
New Cactus 241-way
Cons 30 Primates
Primate Chain/Net
Placental Chain/Net
Vertebrate Chain/Net

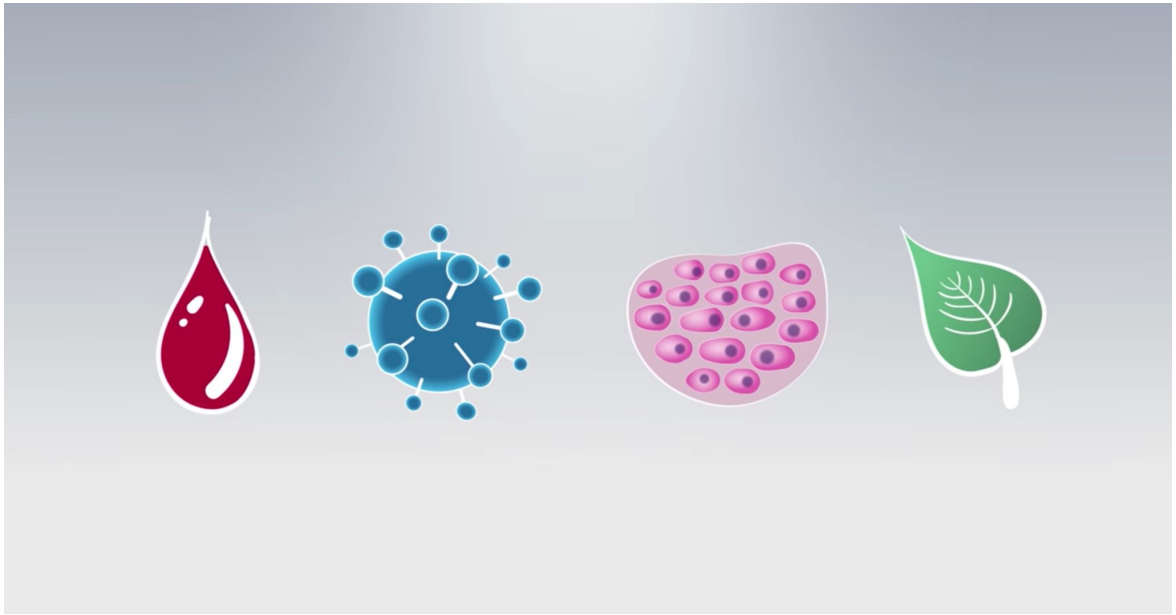
Exercise

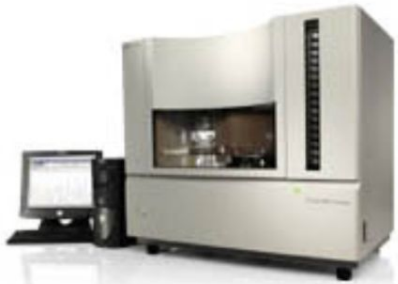
Mention a gene located at chr 22 position 24,100,000!


Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38)
move
zoom in
zoom out
multi-region
chr22:23,870,297-24,378,481
508,185 bp.
22p12
22p11.2
22q11.21
22q11.22
22q11.23
22q12.1
22q12.2
200 kb
24,000,000
24,050,000
24,100,000
24,150,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V38 (50 items filtered out)
GSTT4
KB-318B8.7
CABIN1
SUSD2
GSTT2
RefSeq genes from NCBI
OMIM Allelic Variant Phenotypes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors

Index

DNA sequencing





Sanger DNA sequencing

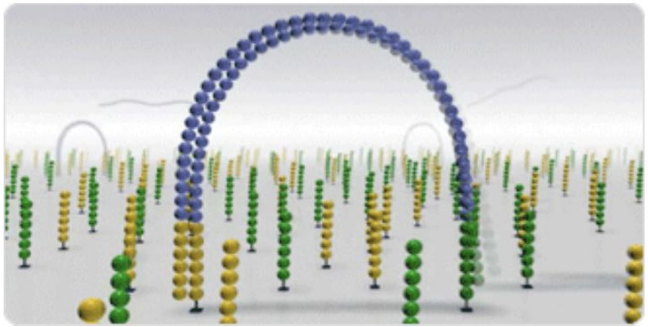
1977-1990s

2nd-generation DNA sequencing

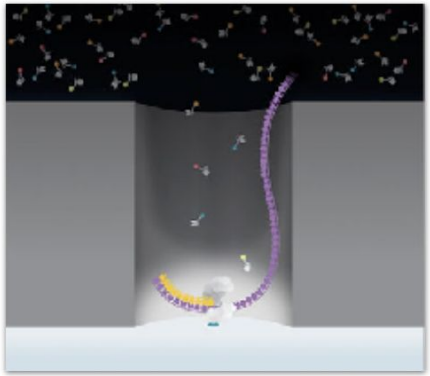
Since ~2007

3rd-generation & single-molecule DNA sequencing

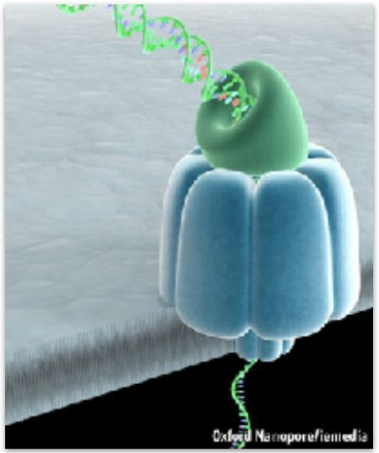
Since ~2010



Synthesis / ligation



SMRT cell

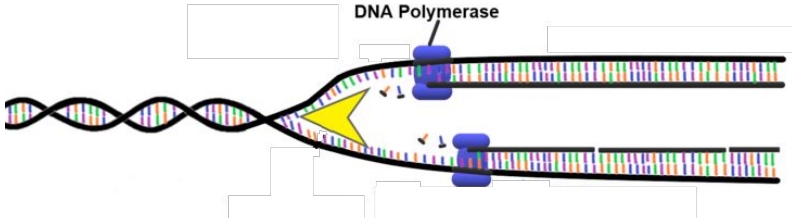


Nanopore

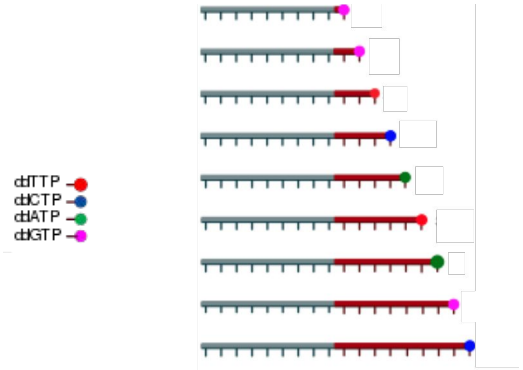
Sanger sequencing

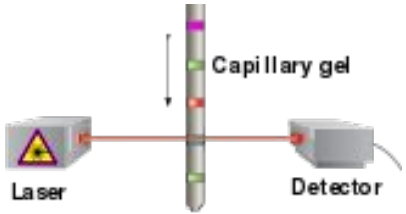
- sequencing by synthesis
- based on the chain termination procedure
- output sequences of different lengths

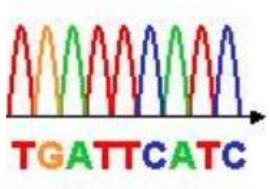
DNA replication



dATP,dCTP,dTTP,dGTP

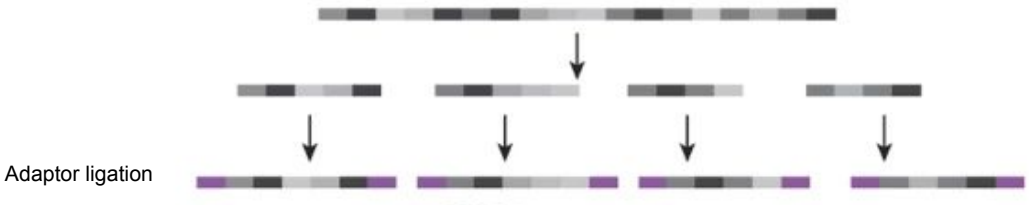


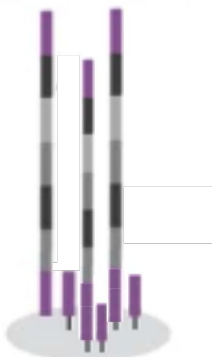


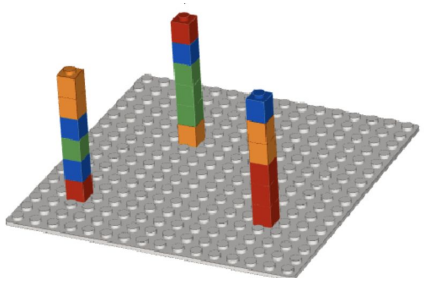


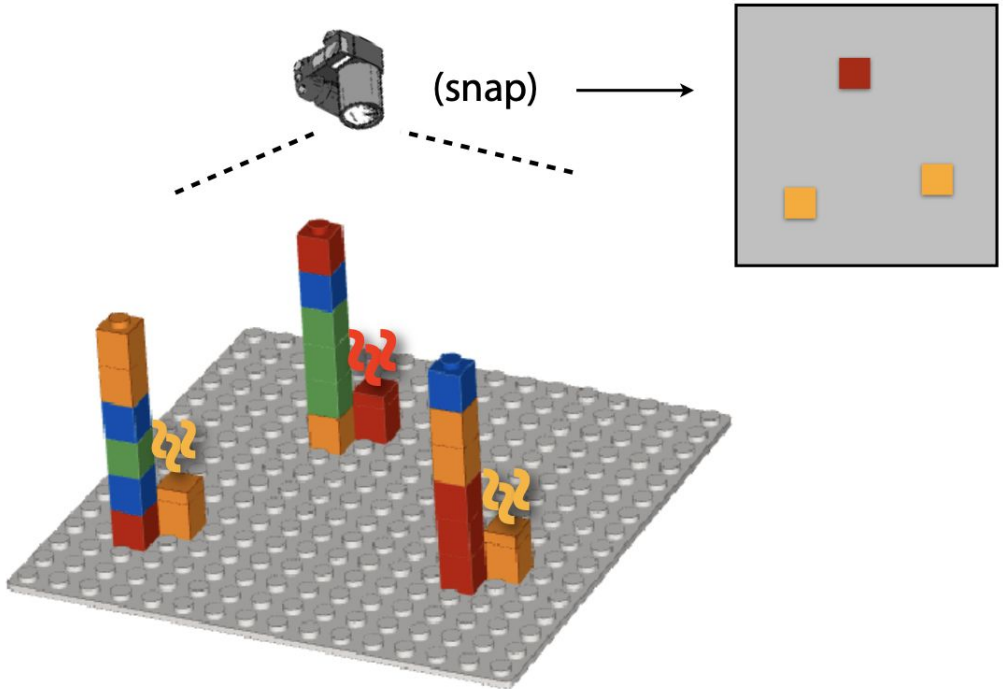
commons.wikimedia.org/wiki/File:DNA_Polymerase_DNA_Replication.png
en.wikipedia.org/wiki/Sanger_sequencing

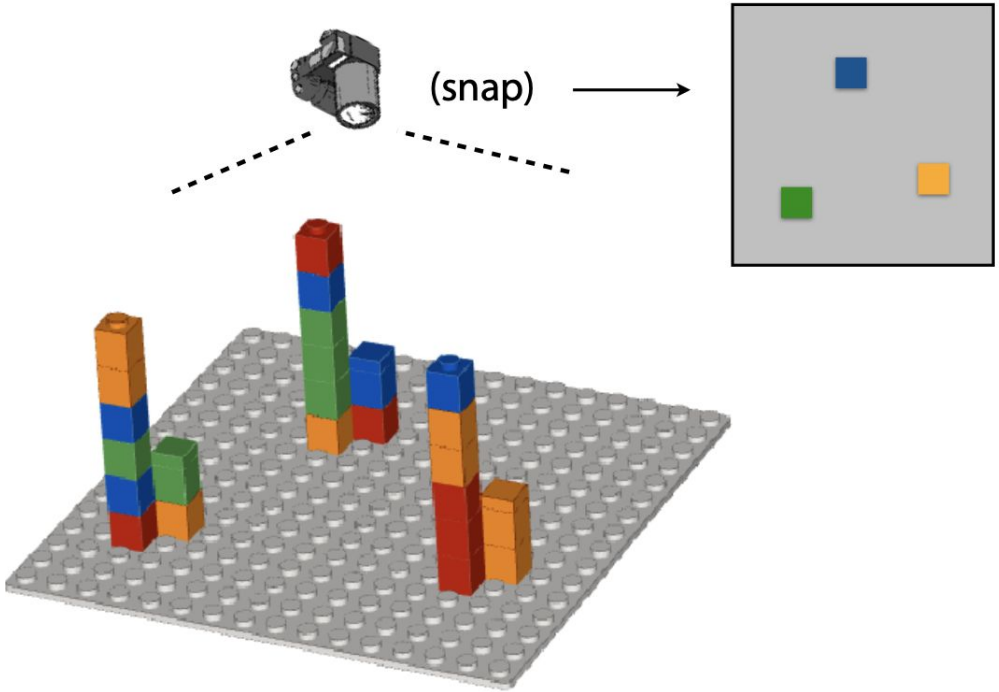
Illumina sequencing

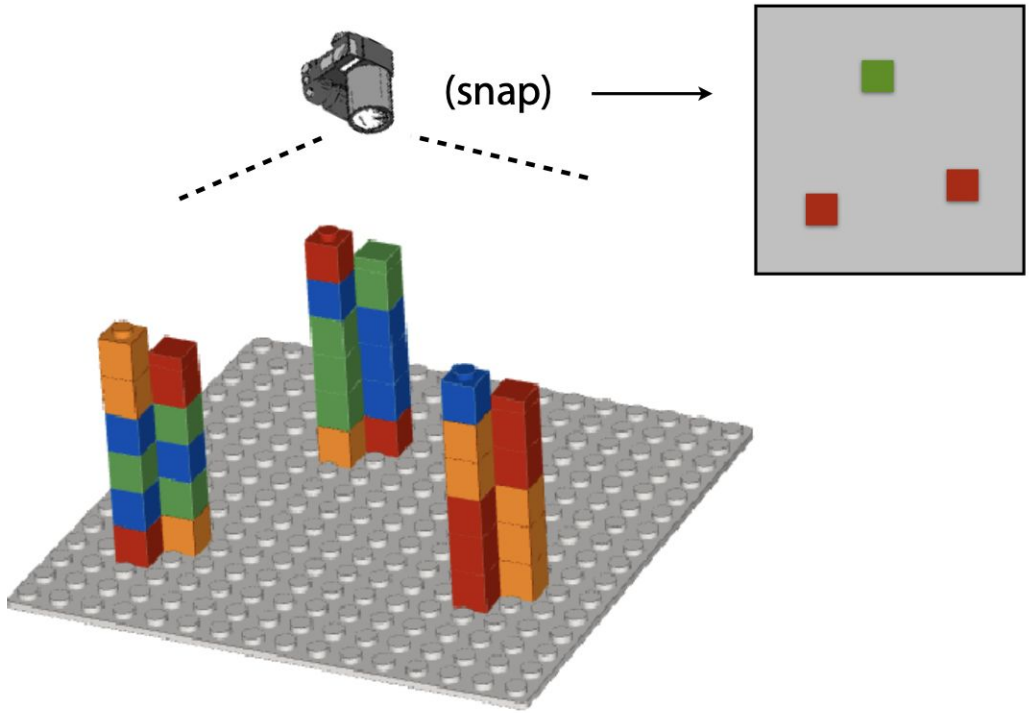


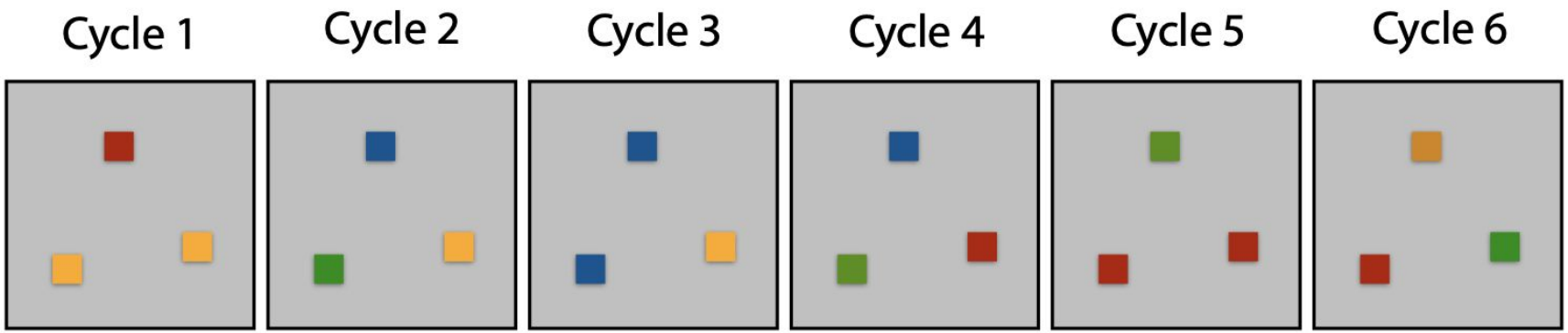


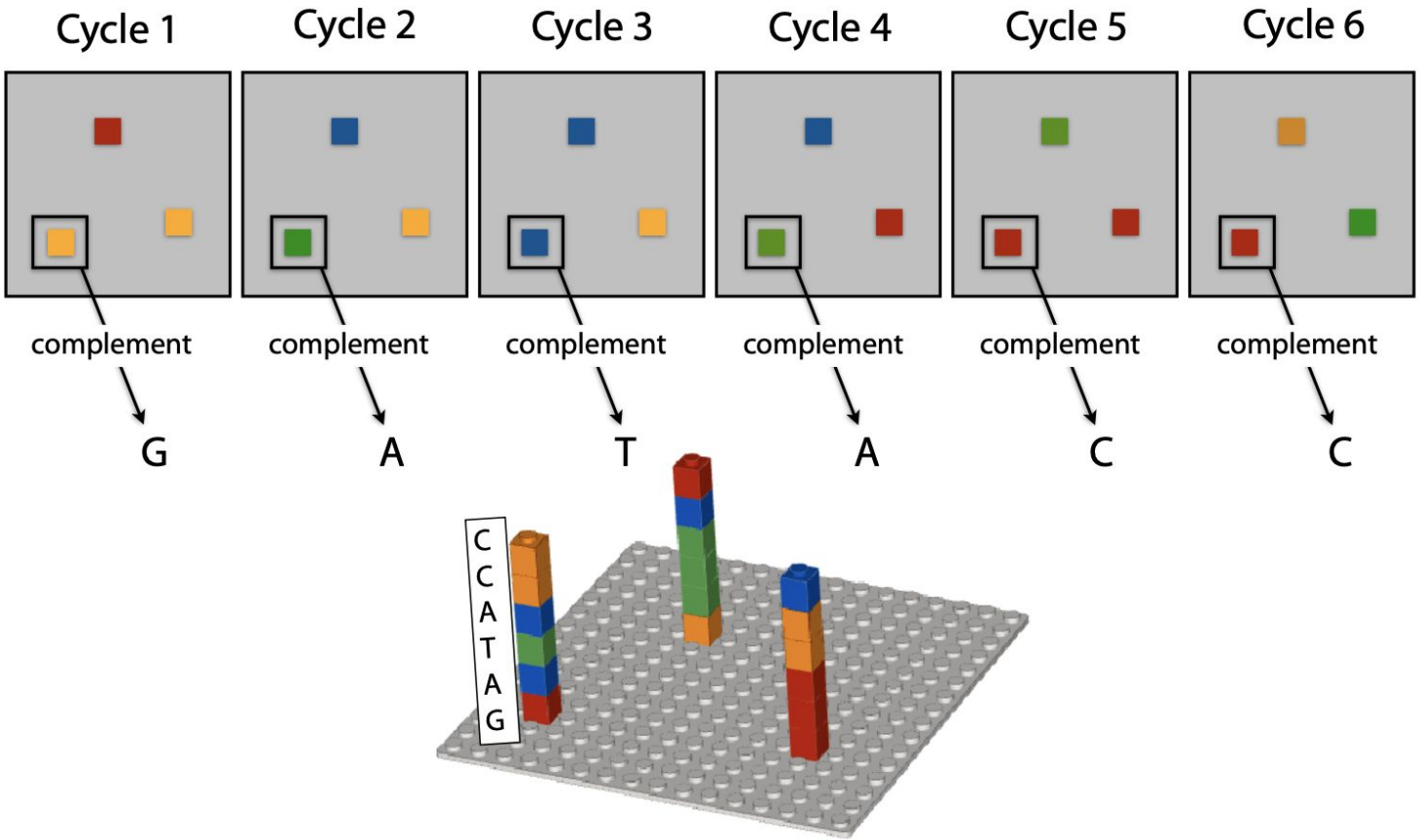



(snap)

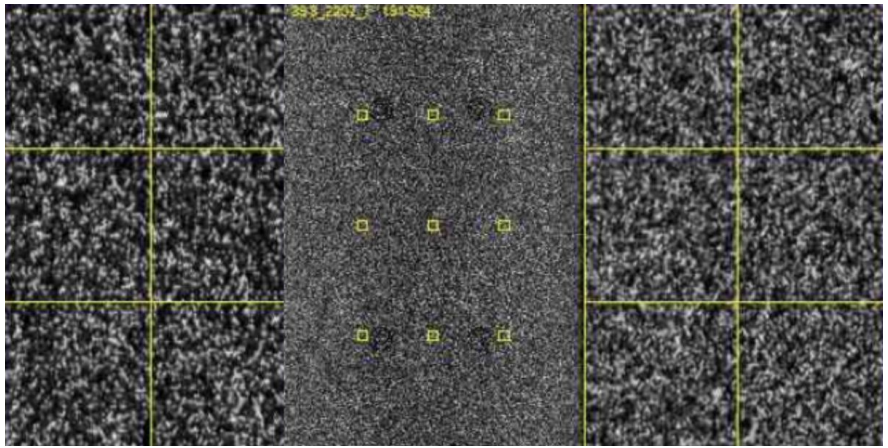

(snap)


(snap)


Cycle 1
Cycle 2
Cycle 3
Cycle 4
Cycle 5
Cycle 6

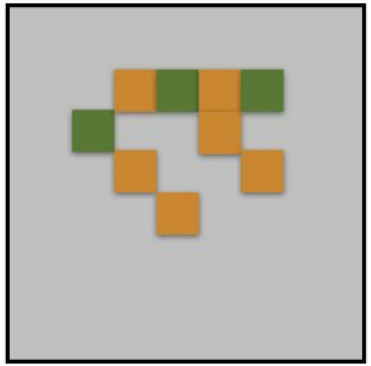

Cycle 1
Cycle 2
Cycle 3
Cycle 4
Cycle 5
Cycle 6
complement
complement
complement
complement
complement
complement
G
A
T
A
C
C
C C A T A G

Actual Illumina HiSeq 3000



https://www.youtube.com/watch?v=CZeN-IgjYCo

dnatech.genomecenter.ucdavis.edu/2015/05/07/first-hiseq-3000-data-download/



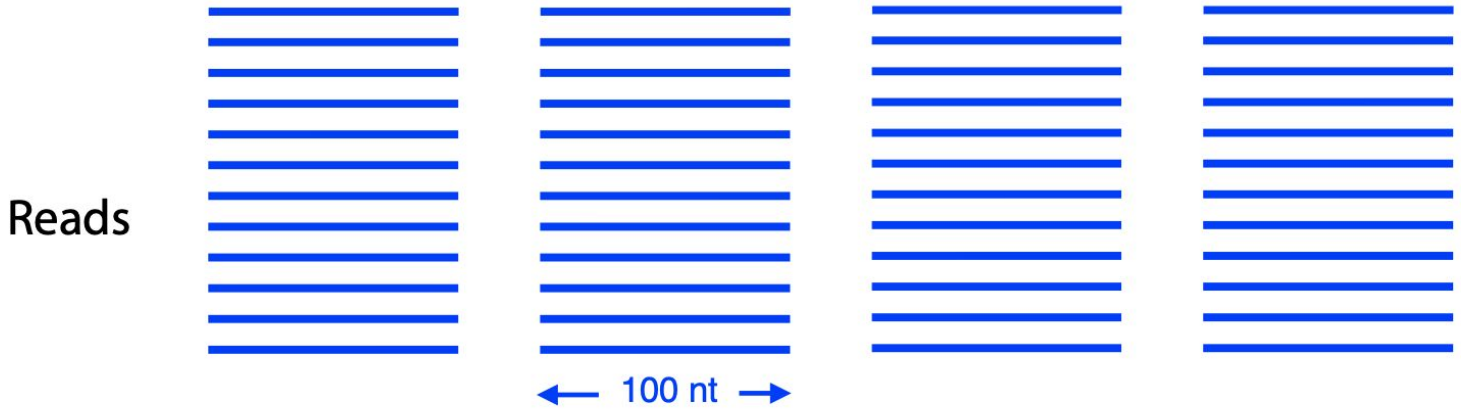
Call: orange (C)

Estimate *p*, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

Illumina error rate 0.01%



Your genome

100,000,000 nt

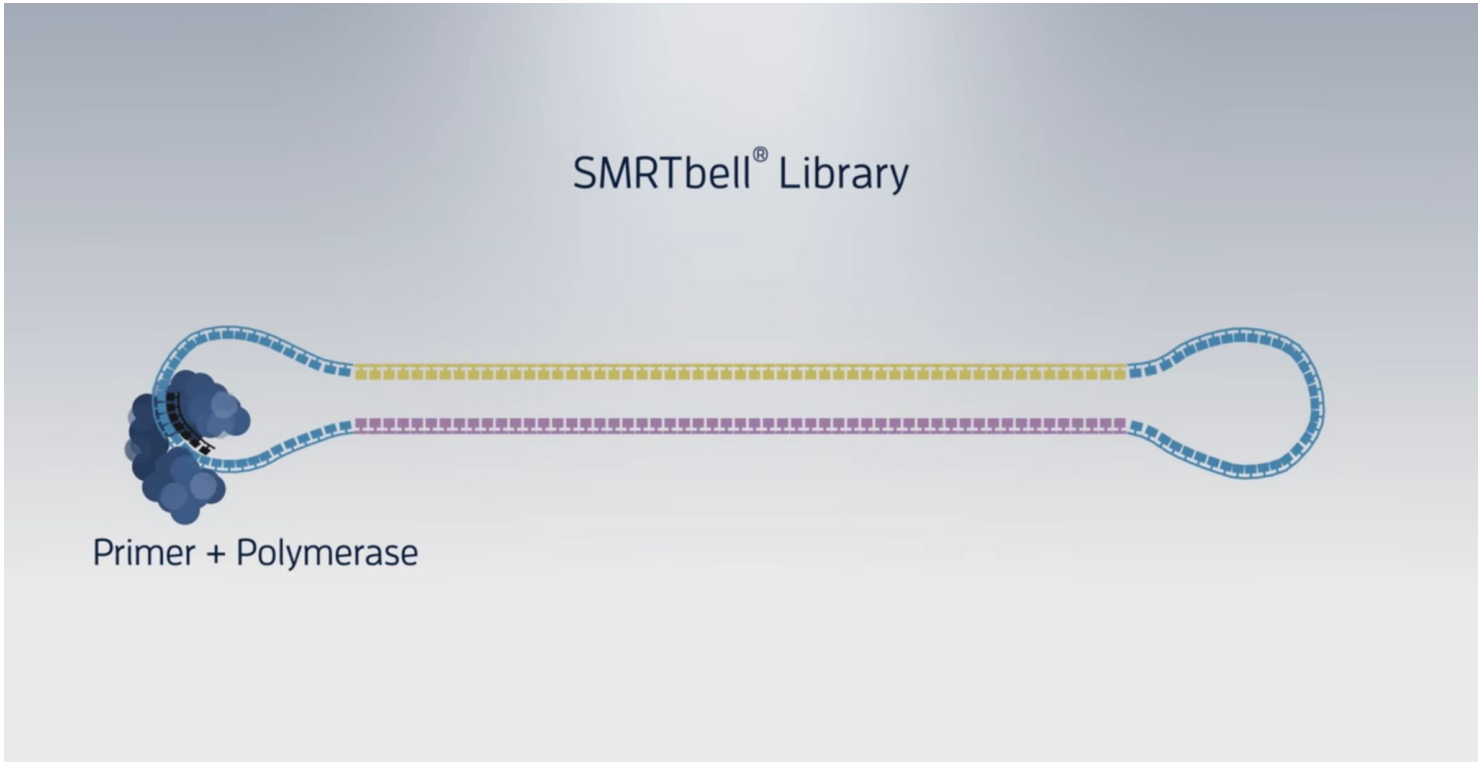
Index

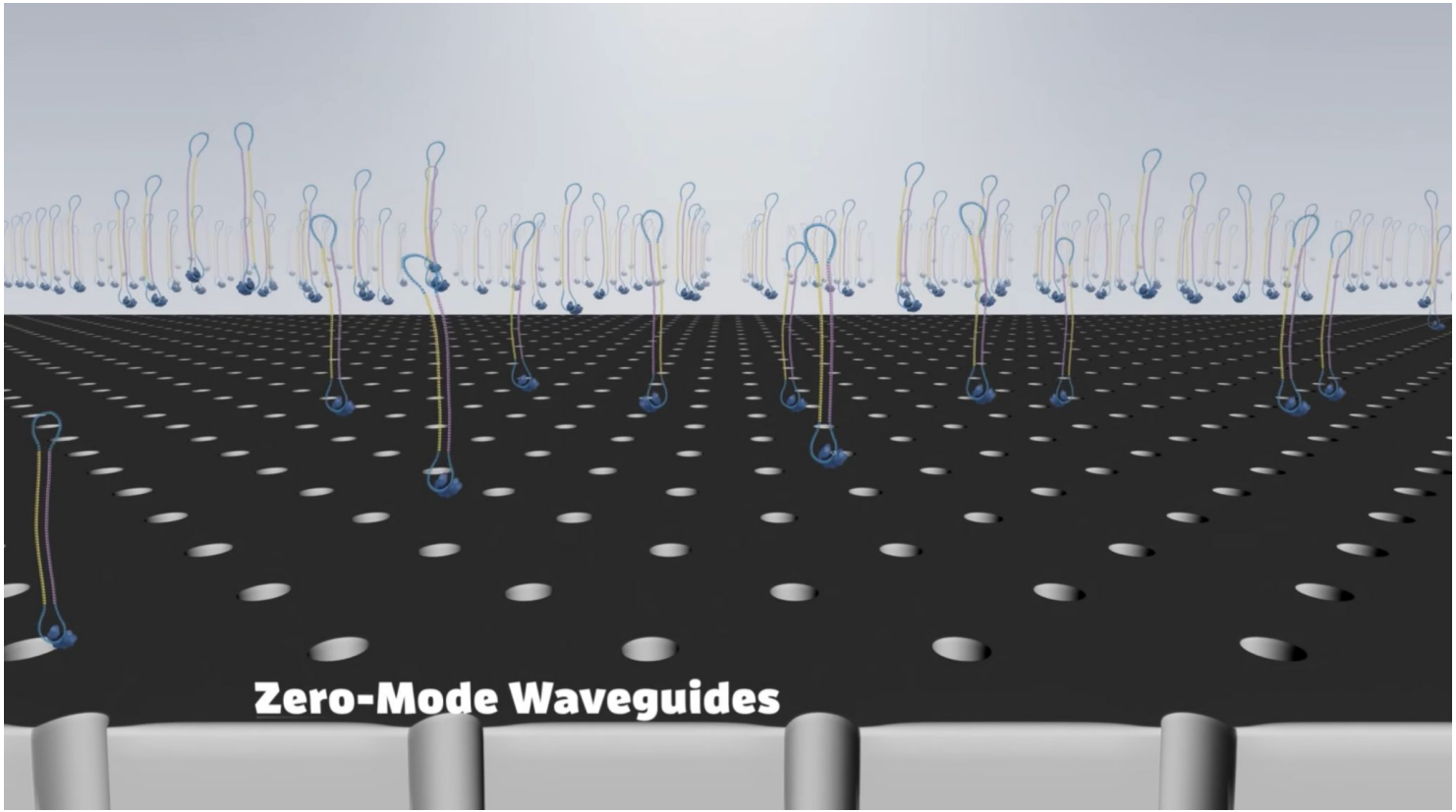
PACBIO

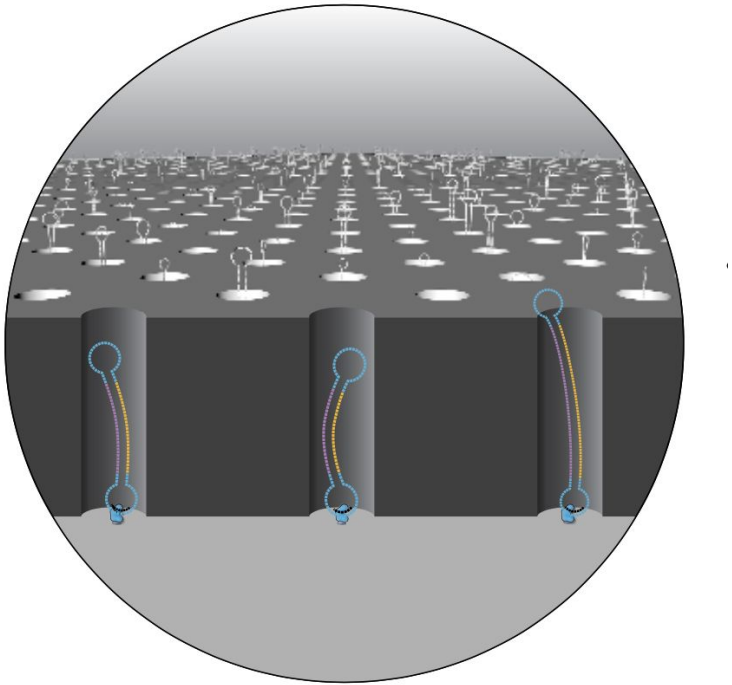
Single Molecule, Real-Time (SMRT®) Sequencing

pacb.com/smrt-science/smrt-sequencing

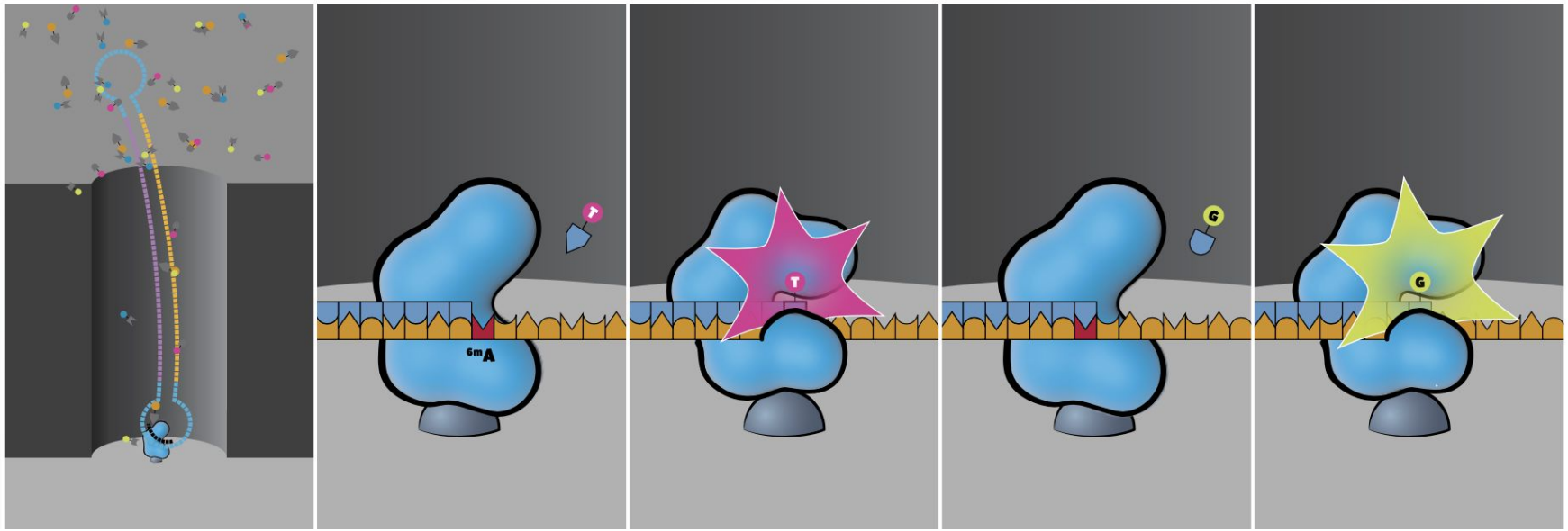
https://www.youtube.com/watch?v=_lD8JyAbwEo

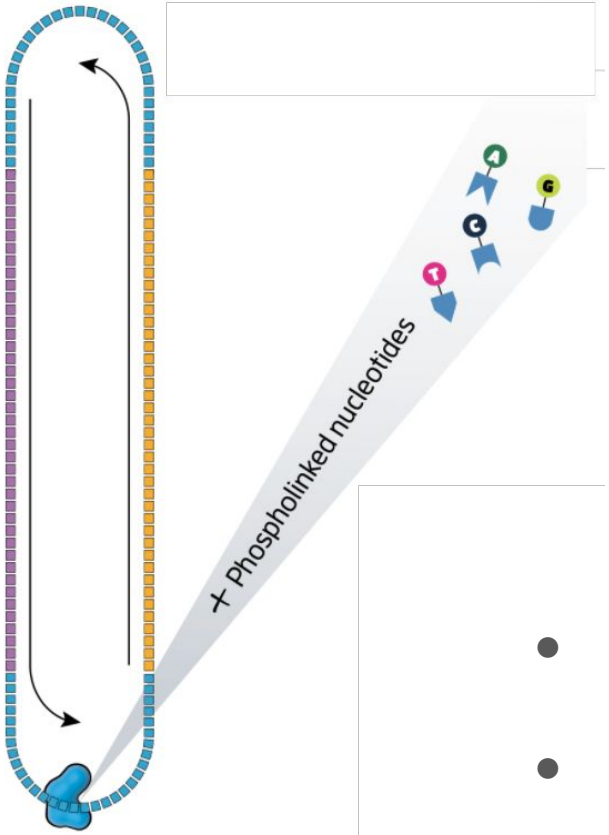

SMRTbell® Library
Primer + Polymerase

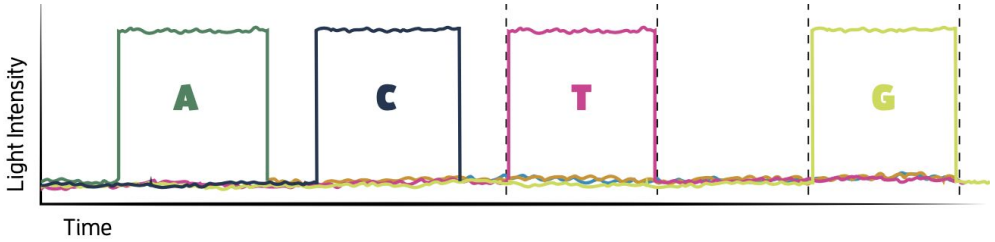

Zero-Mode Waveguides



SMRT Cells contain millions of zero-mode waveguides (ZMWs)

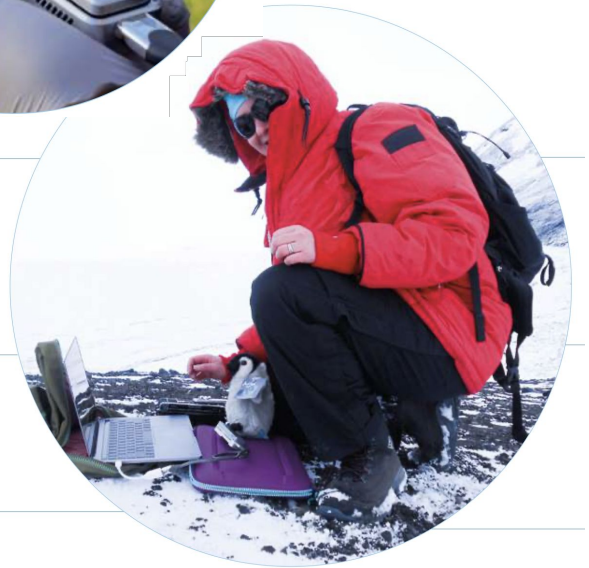

T
6mA
T
G
G





- Circular Consensus sequencing (CCS) or High Fidelity (HiFi)
- ~12,000 bp

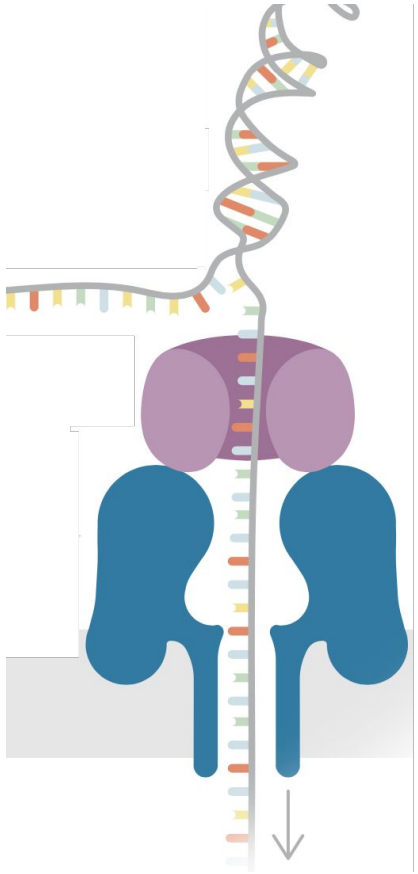
ONT

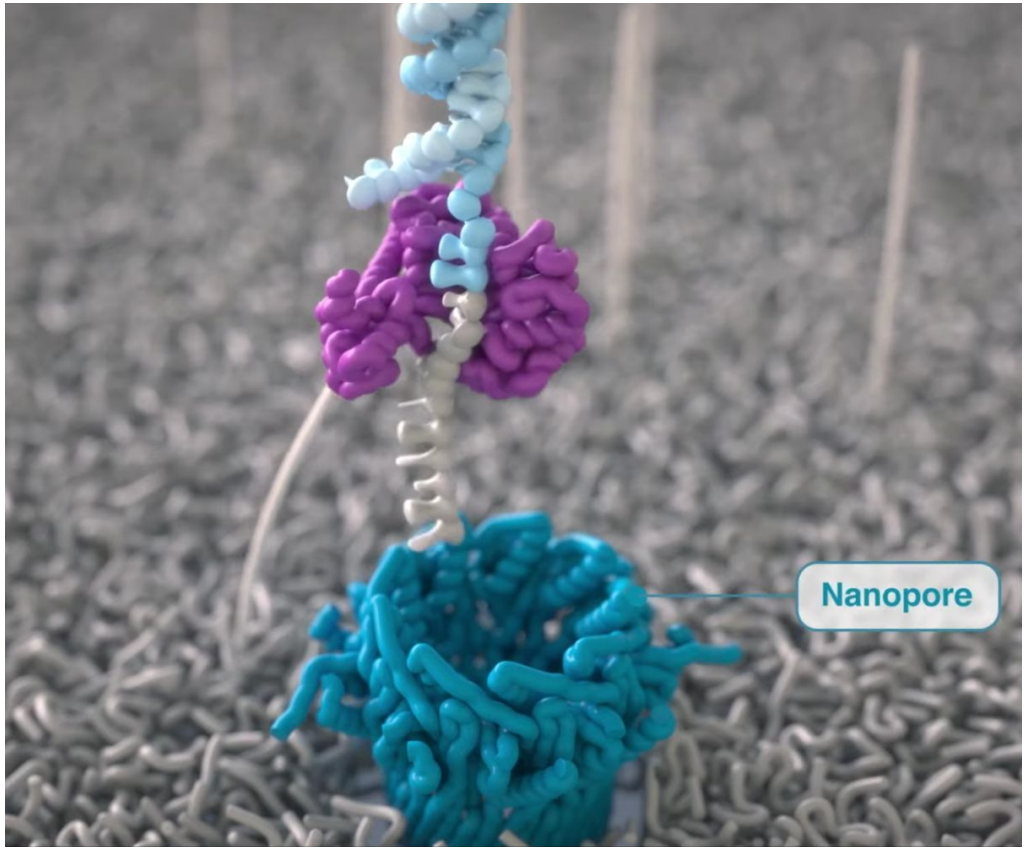


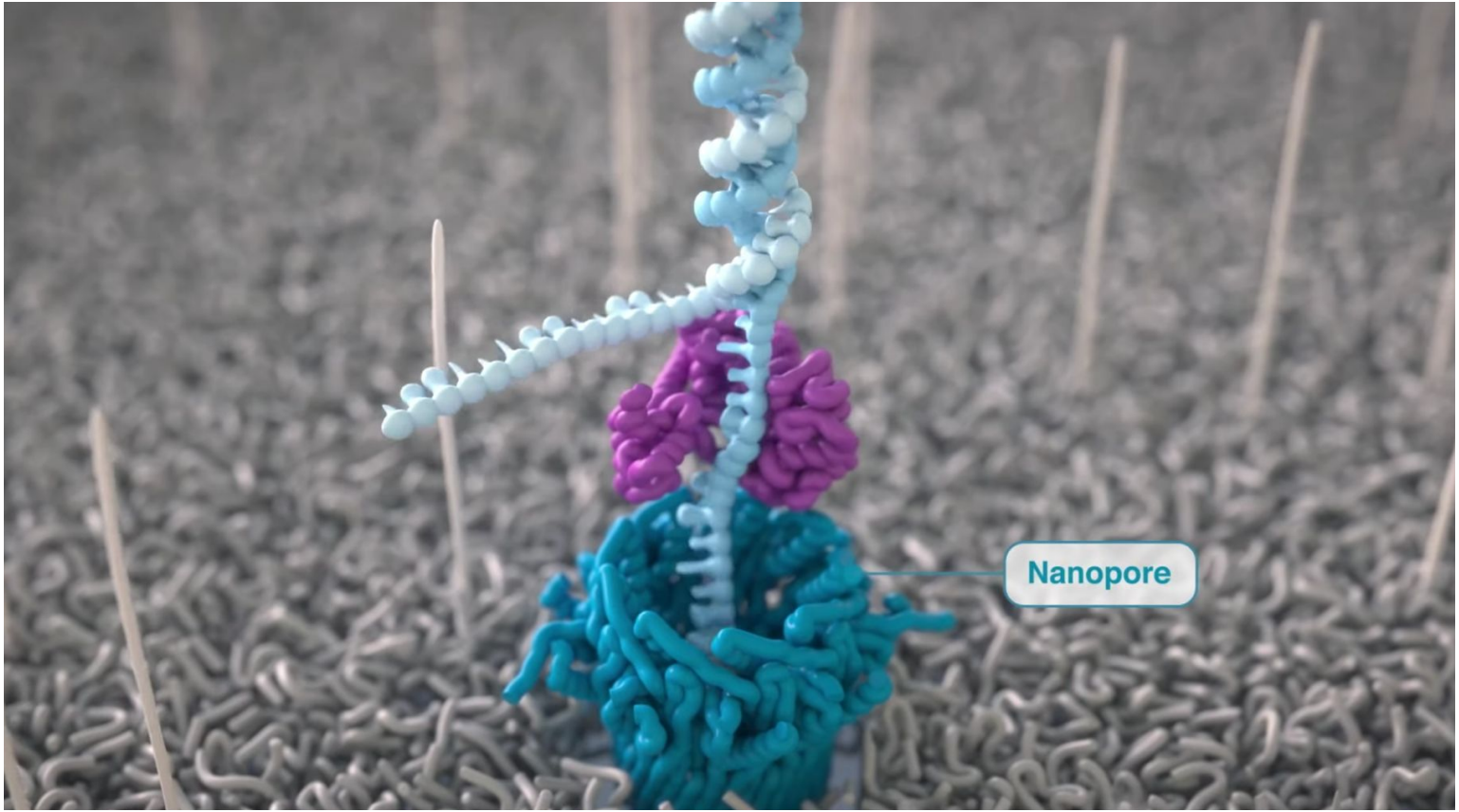
100,000- 4,000,000 bp

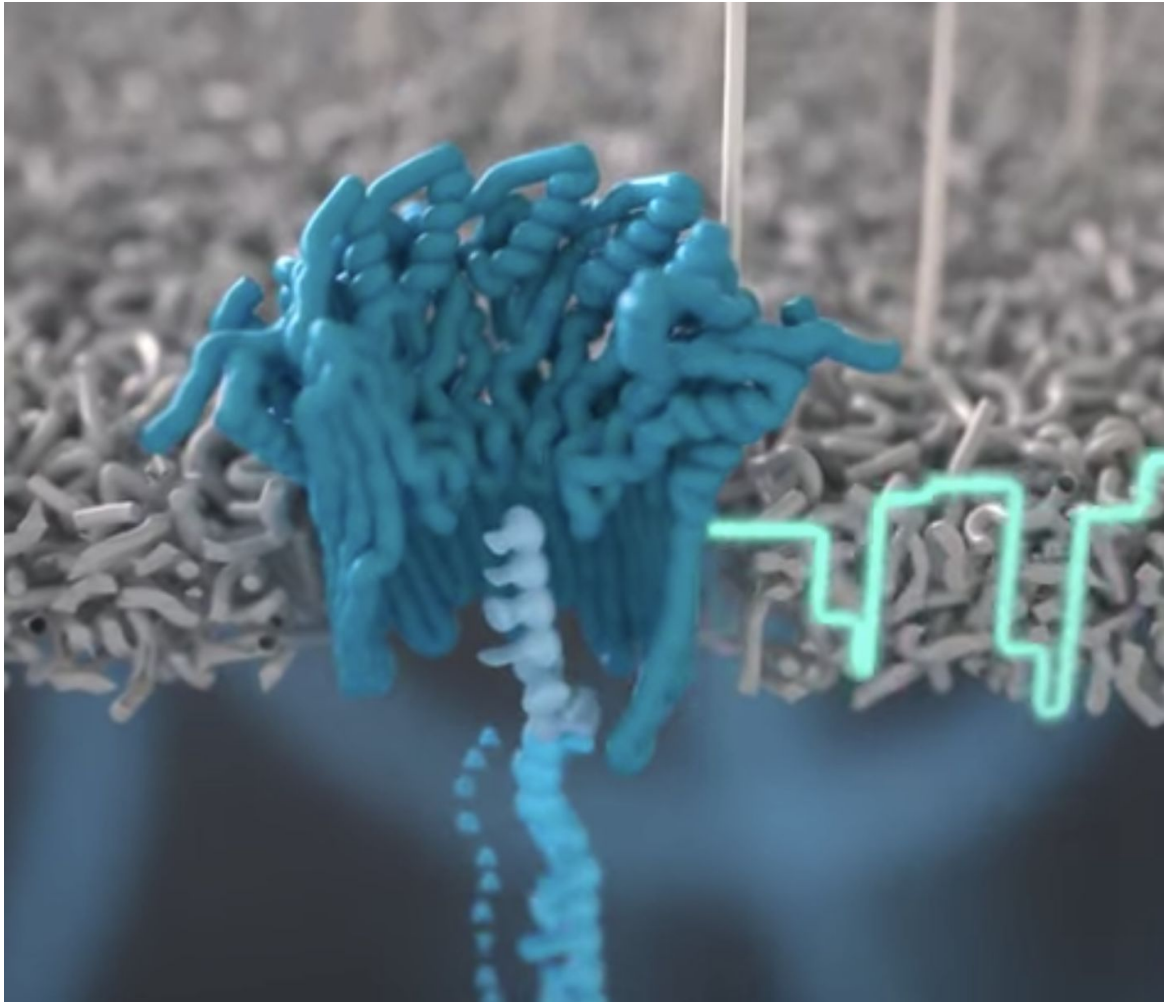
https://www.youtube.com/watch?v=qzusVw4Dp8w

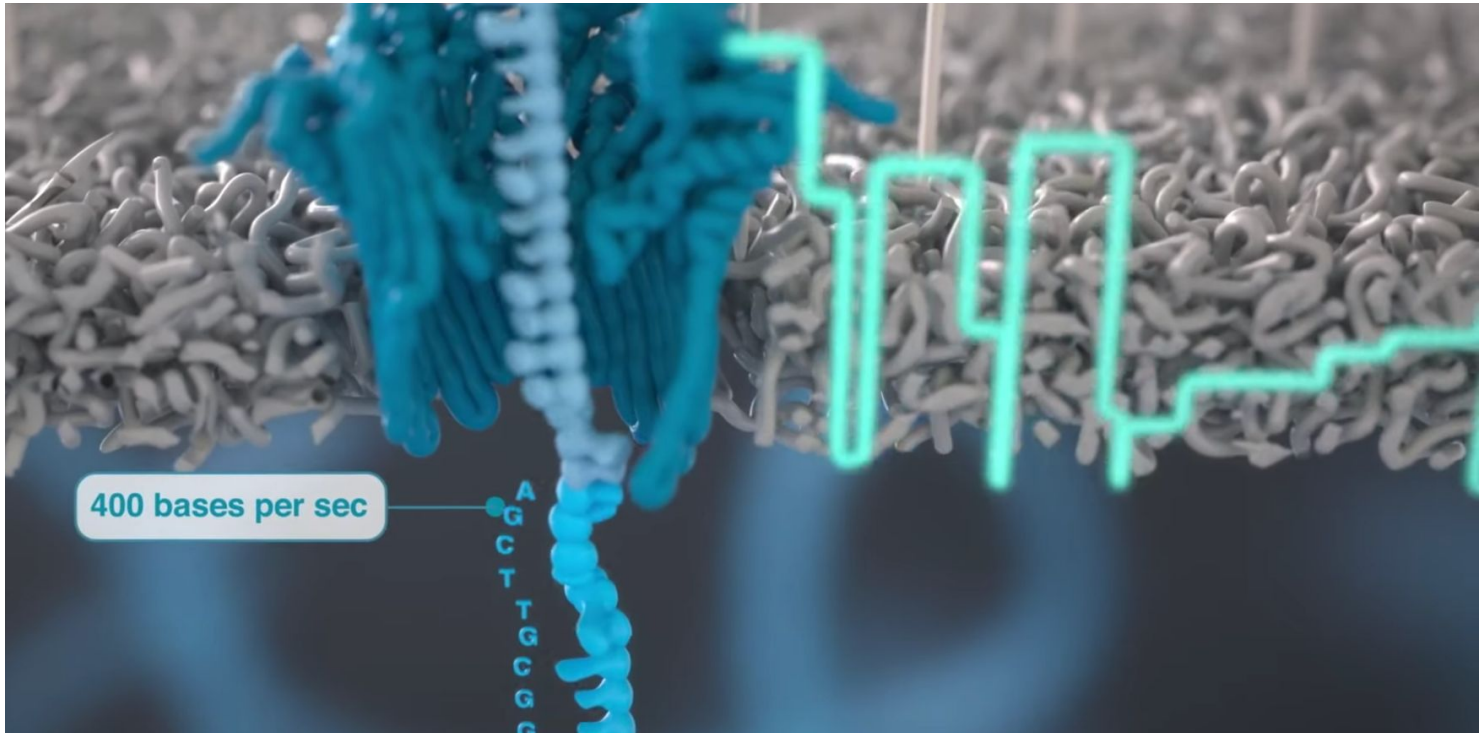
- DNA or RNA fragments pass through a nano-scale hole.
- The fluctuations in current during translocation are used.
- An enzyme motor controls the translocation of the DNA.

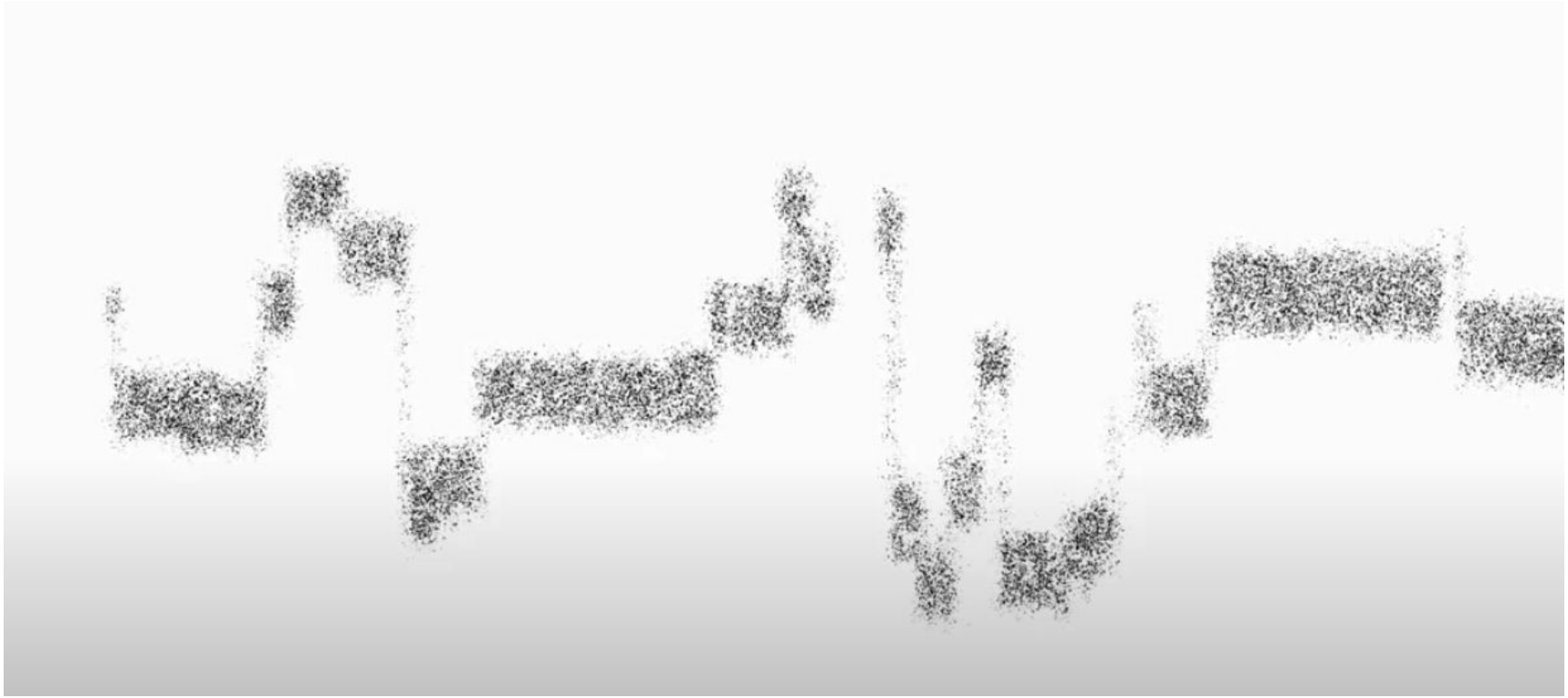



Nanopore

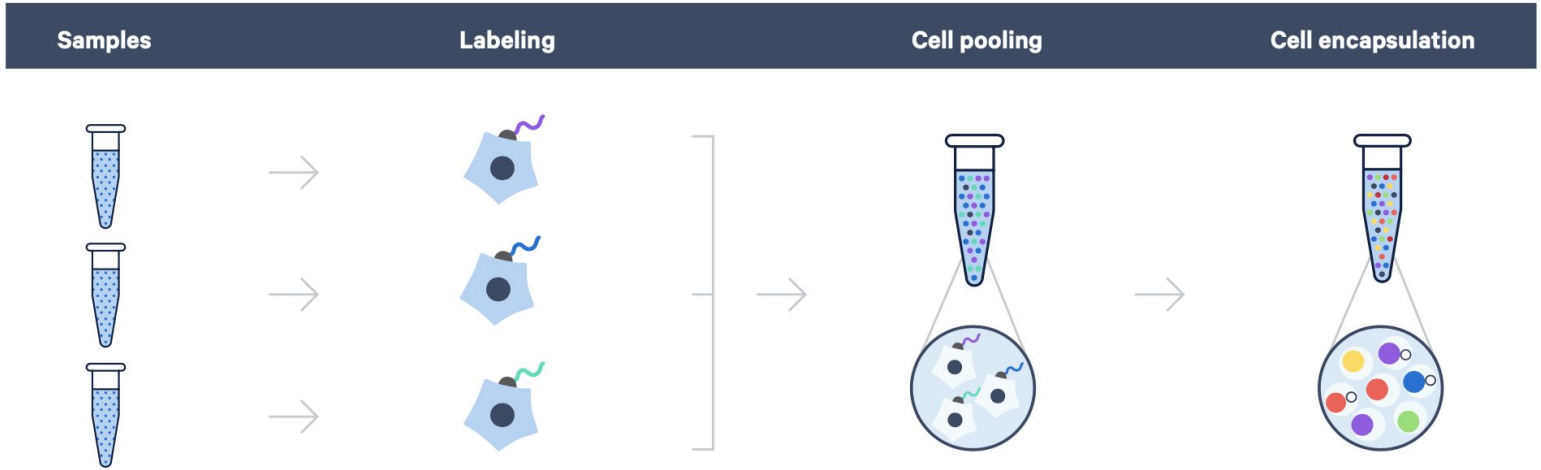

Nanopore




400 bases per sec
A
G
C
T
T G C G

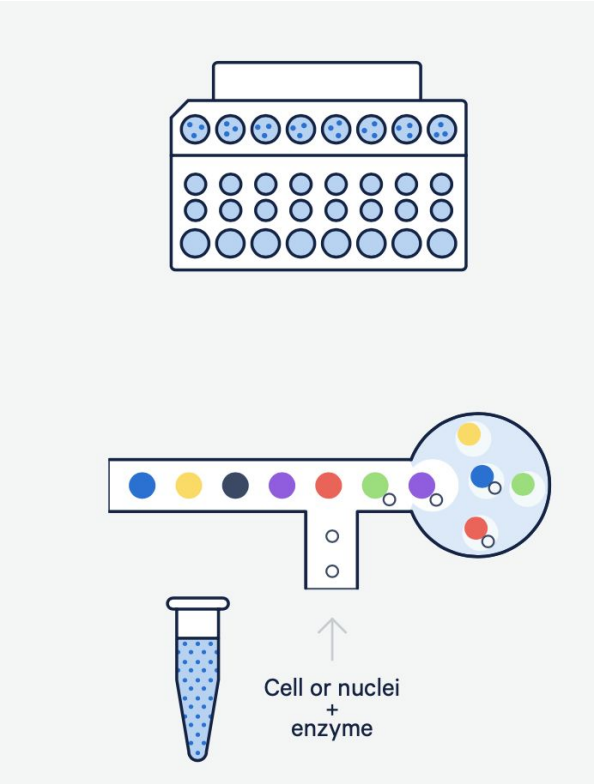


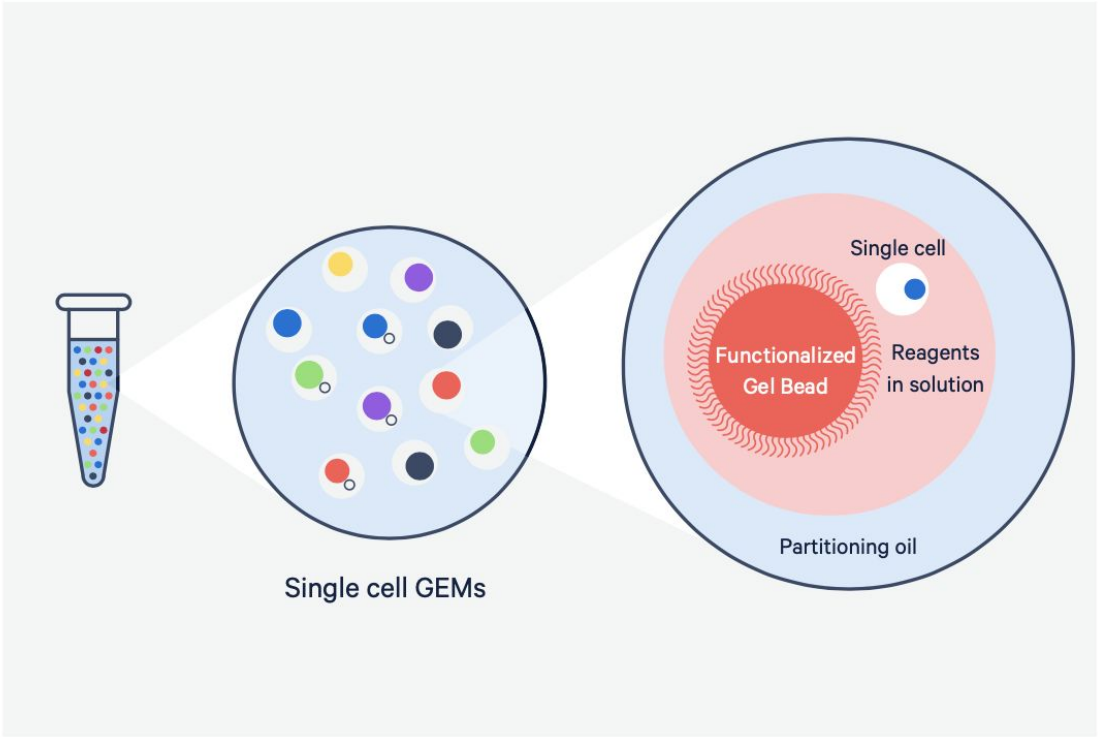
Single cell sequencing

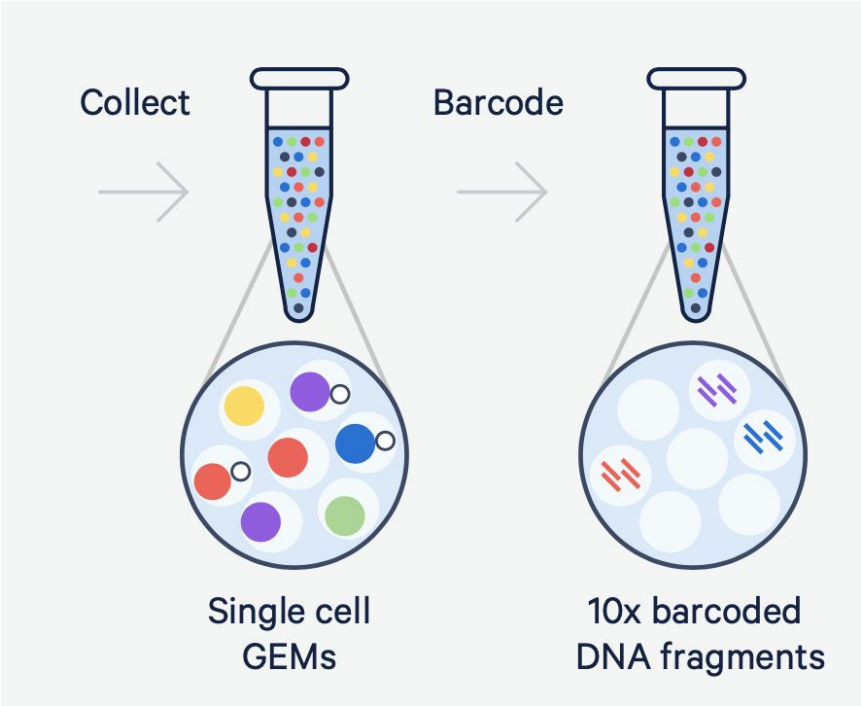


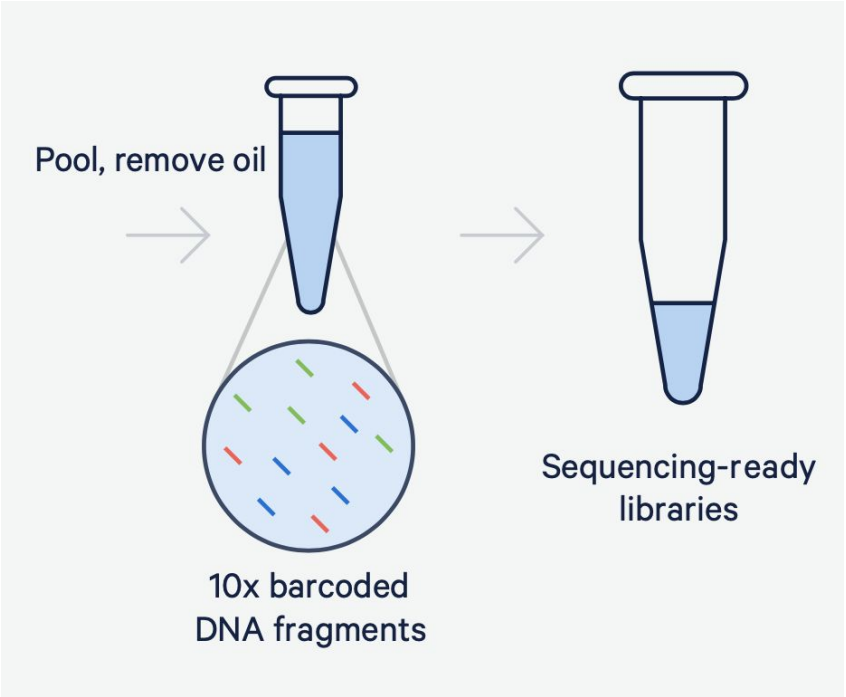
- Gel Beads, cells or nuclei, enzymes, and partitioning oil are loaded onto a chip.

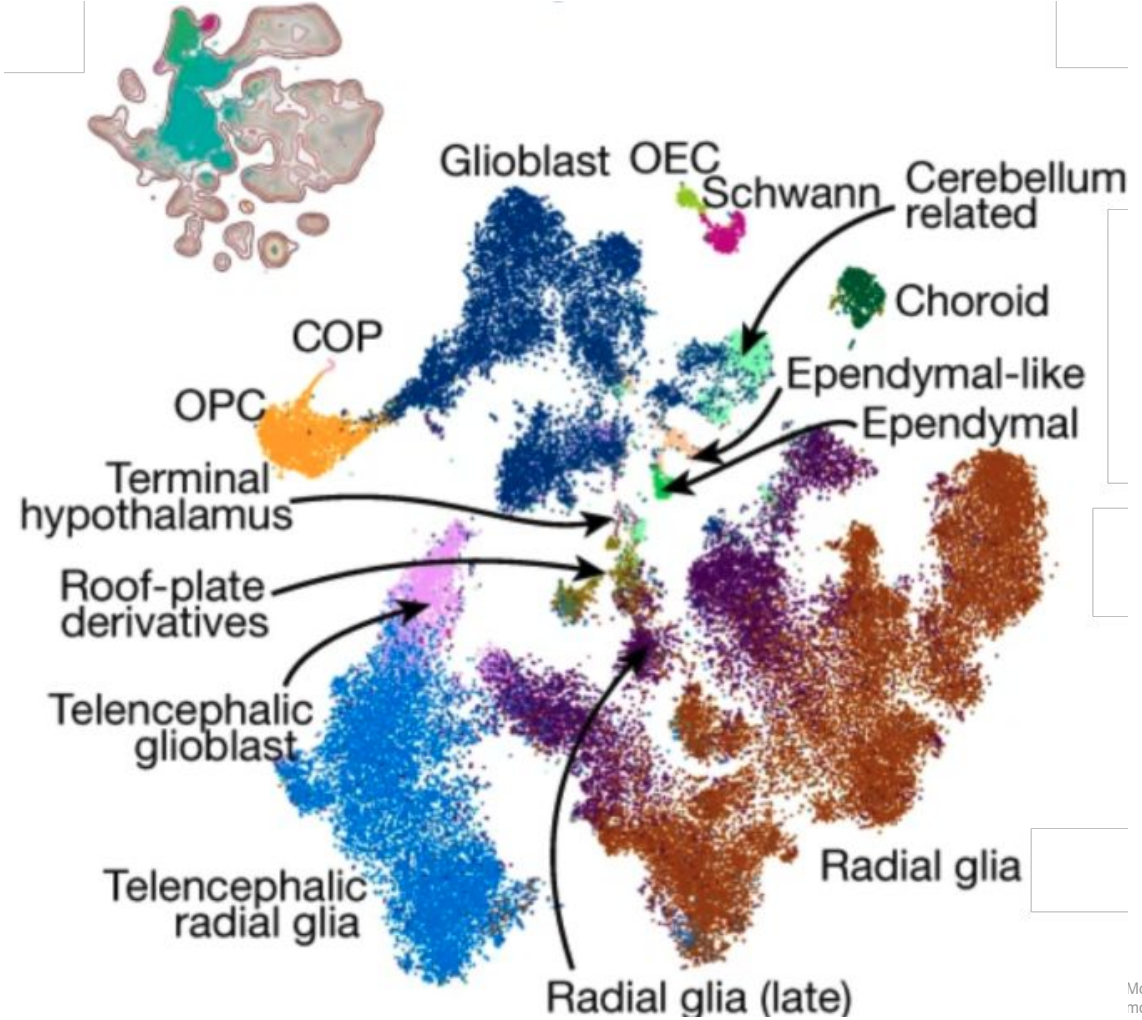
- Barcoded gel beads are mixed with the cells or nuclei, enzymes, and partitioning oil.




Single cell
Functionalized
Gel Bead
Reagents
in solution
Partitioning oil
Single cell GEMs


Collect
Barcode
Single cell
GEMs
10x barcoded
DNA fragments


Pool, remove oil
10x barcoded
DNA fragments
Sequencing-ready
libraries



Molecular architecture of the developing mouse brain, Nature 2021.

Other methods

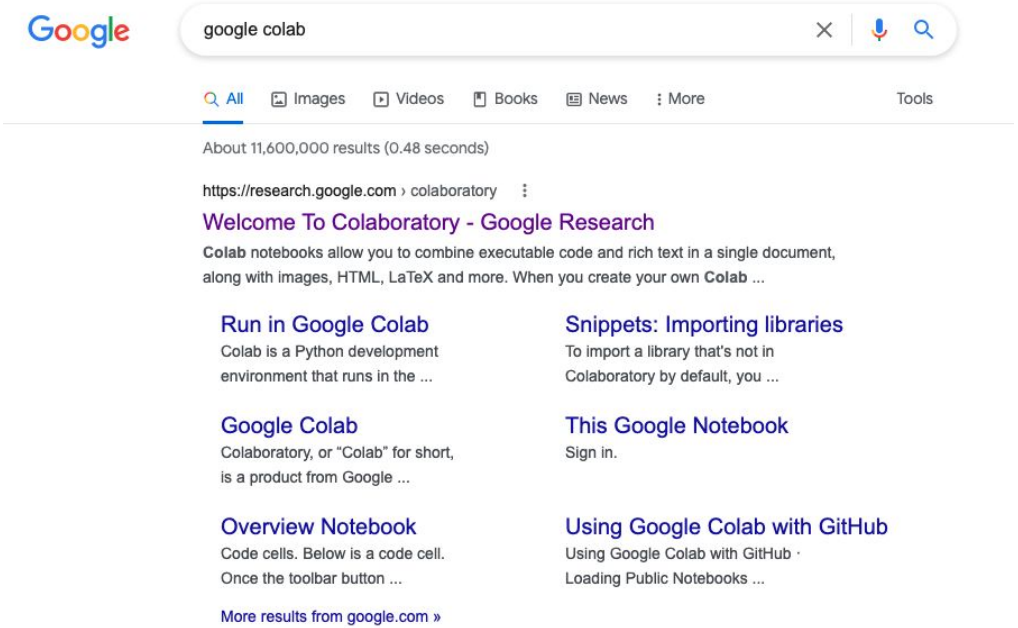
- SOLiD
- Ion Torrent semiconductor sequencing
- 454 Life Sciences (pyrosequencing)

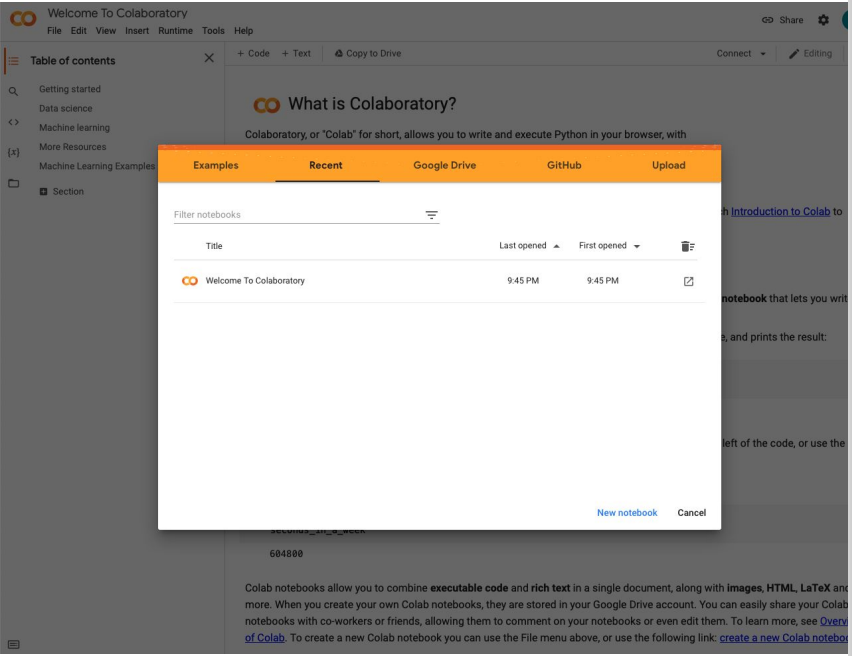
- Hi-C
- Optical mapping, BioNano
- Linked read, 10X genomics

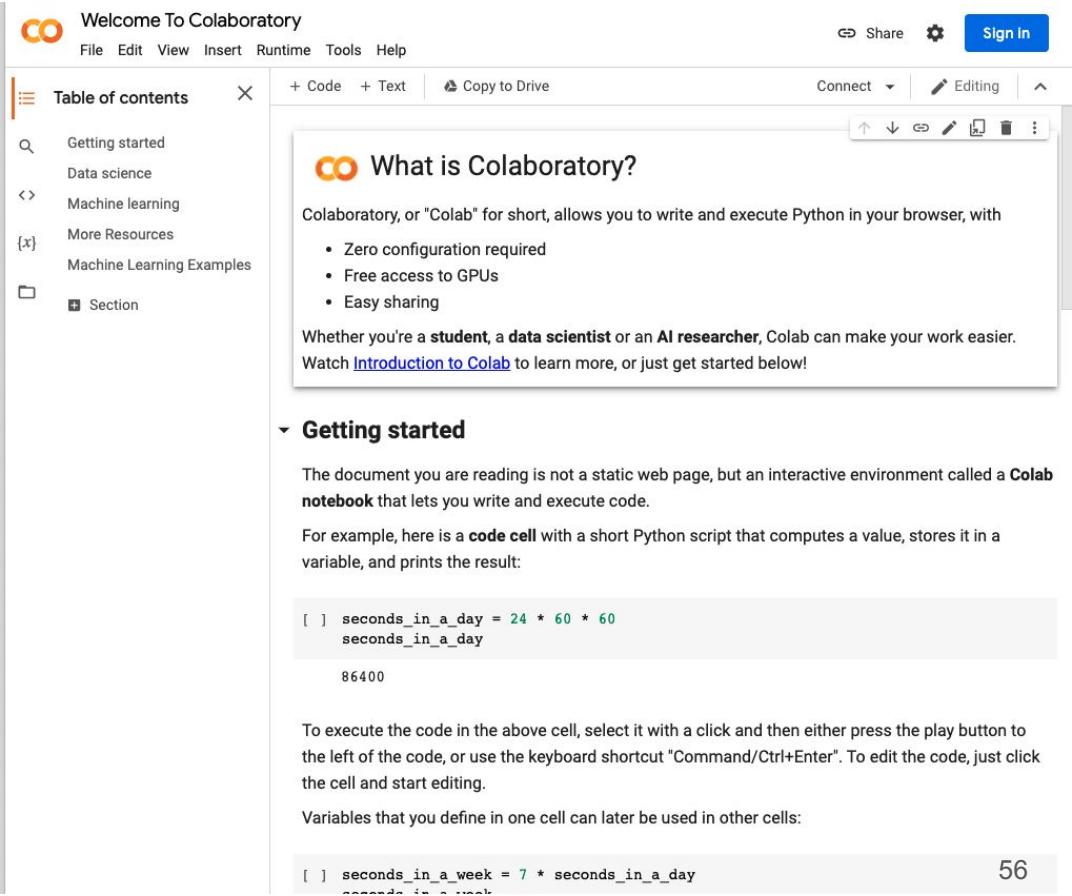
To save time for the next section

Create a gmail account

(if you don’t have)

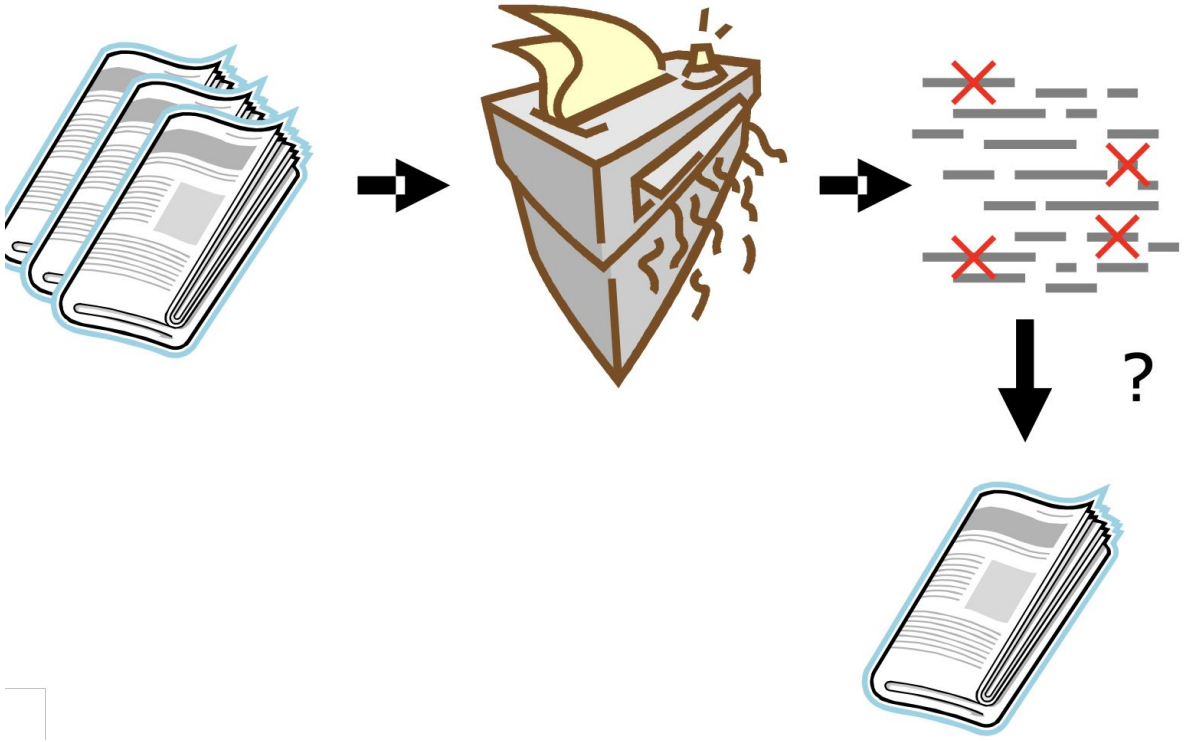



Welcome To Colaboratory
File Edit View Insert Runtime Tools Help
Share
Table of contents
+ Code + Text
Copy to Drive
Connect
Editing
Getting started
Data science
Machine learning
More Resources
Machine Learning Examples
Section
What is Colaboratory?
Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with
Examples
Recent
Google Drive
GitHub
Upload
Filter notebooks
Title
Last opened
First opened
Welcome To Colaboratory
9:45 PM
9:45 PM
New notebook
Cancel
604800
Colab notebooks allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX and
more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab
notebooks with co-workers or friends, allowing them to comment on your notebooks or even edit them. To learn more, see Overv
of Colab. To create a new Colab notebook you can use the File menu above, or use the following link: create a new Colab notebo


Welcome To Colaboratory
File Edit View Insert Runtime Tools Help
Share
Sign In
Table of contents
+ Code + Text
Copy to Drive
Connect
Editing
Getting started
Data science
Machine learning
More Resources
Machine Learning Examples
Section
What is Colaboratory?
Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with
Zero configuration required
Free access to GPUs
Easy sharing
Whether you're a student, a data scientist or an AI researcher, Colab can make your work easier.
Watch Introduction to Colab to learn more, or just get started below!
Getting started
The document you are reading is not a static web page, but an interactive environment called a Colab notebook that lets you write and execute code.
For example, here is a code cell with a short Python script that computes a value, stores it in a variable, and prints the result:
[ ] seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
86400
To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter". To edit the code, just click the cell and start editing.
Variables that you define in one cell can later be used in other cells:
[ ] seconds_in_a_week = 7 * seconds_in_a_day

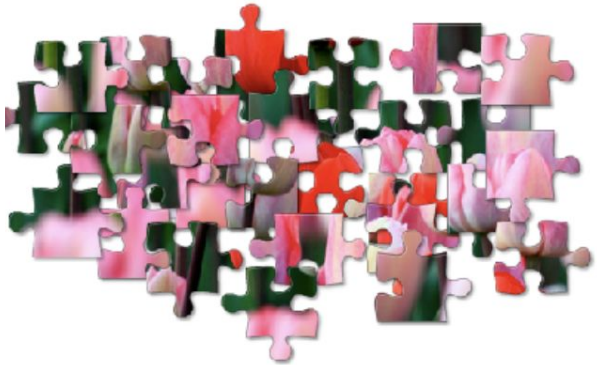
Index

De novo Genome Assembly



Recovery of shredded newspaper

De novo= from the beginning



Overlap-layout-consensus approach

TACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTG

TGCGTGGGATCTCGGCGAAATTCTTTGCCGCA

GTGATGGTATGCGCACCTTGCGTGGGATCTC

CCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGGTAGGTGATGGTATGCG

Overlap-layout-consensus approach

TACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTG

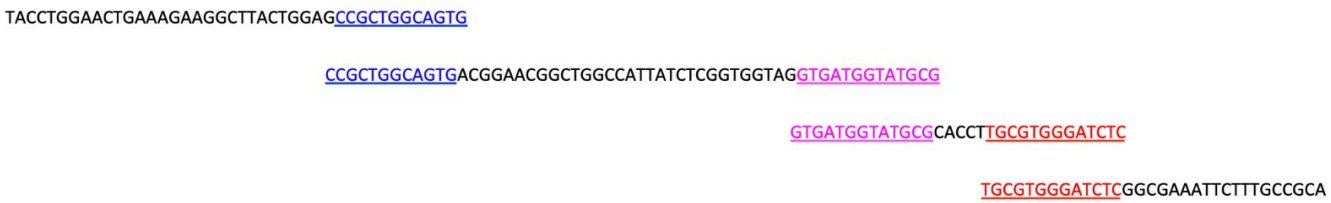
TGCGTGGGATCTCGGCGAAATTCTTTGCCGCA

GTGATGGTATGCGCACCTTGCGTGGGATCTC

CCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGGTAGGTGATGGTATGCG

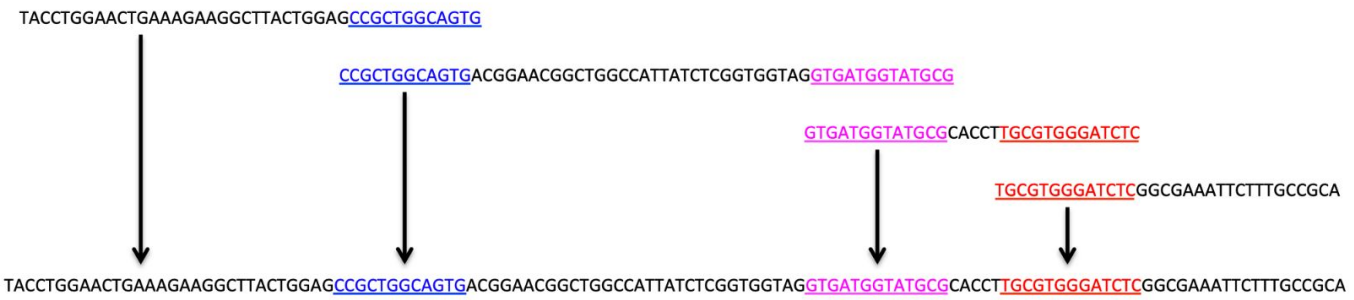
All versus all sequence similarity comparison

Overlap-layout-consensus approach



Alignment of sequences with high sequence identity

Overlap-layout-consensus approach

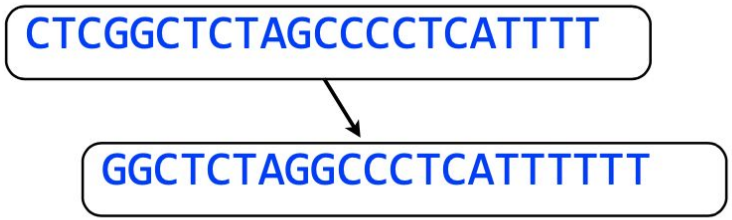


Determination of consensus sequence

Overlap graph

Each node is a read

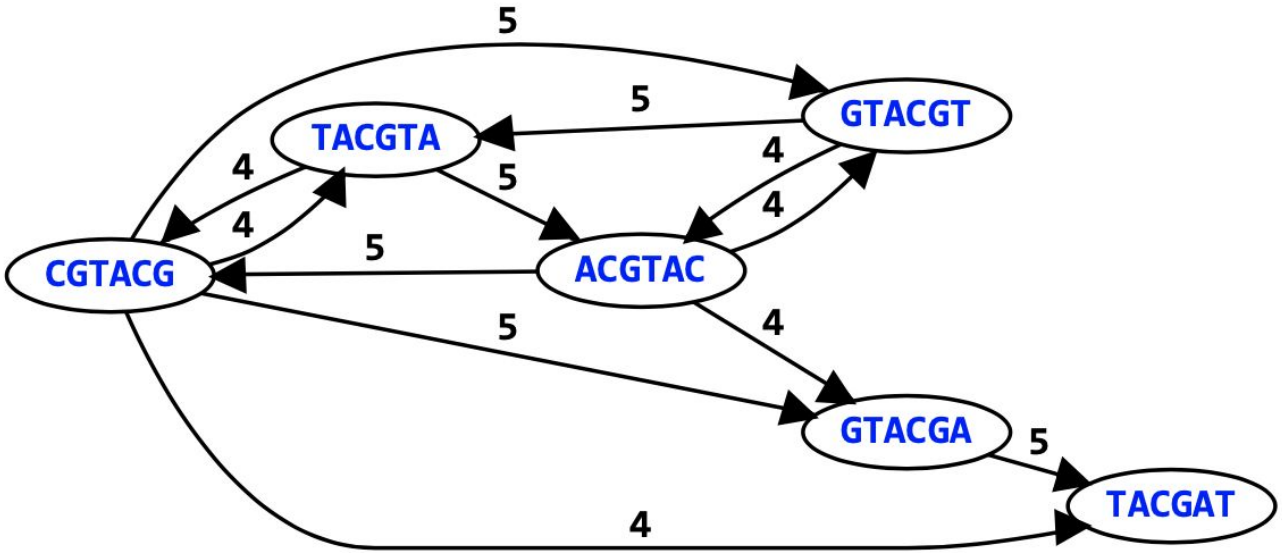
Draw edge A -> B when suffix of A overlaps prefix of B



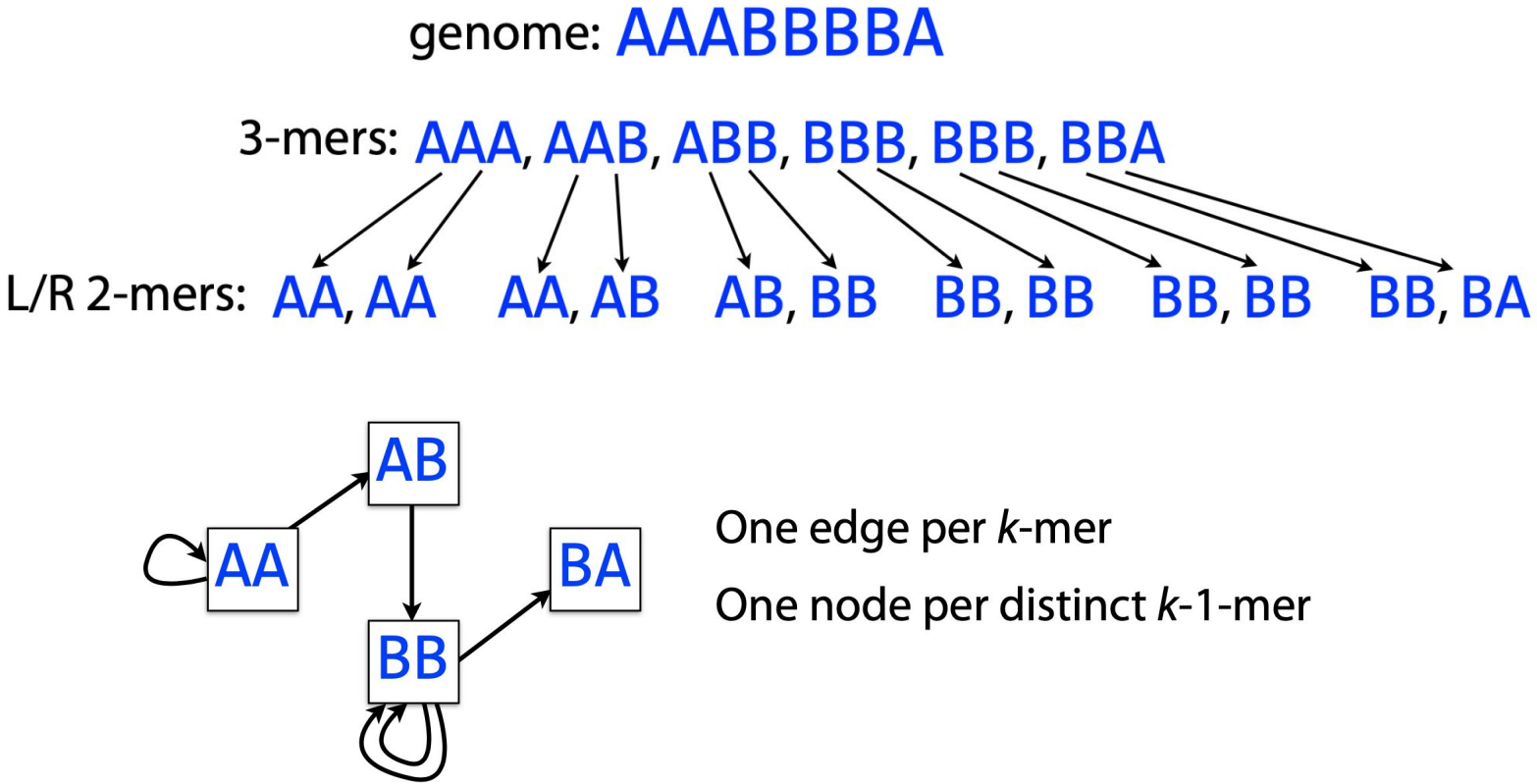
Overlap graph

Nodes: all 6-mers from GTACGTACGAT

Edges: overlaps of length ≥4



De Bruijn graph



CURRENT PROTOCOLS in Bioinformatics
A Wiley Brand

UNIT

Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies

Daniel R. Zerbino

First published: 01 September 2010 | https://doi.org/10.1002/0471250953.bi1105s31 | Citations: 216

Genome Research
Method

Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation

Sergey Koren,[1,5] Brian P. Walenz,[1,5] Konstantin Berlin,[2] Jason R. Miller,[3] Nicholas H. Bergman,[4] and Adam M. Phillippy[1]

[1]*Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA;* [2]*Invincea Incorporated, Fairfax, Virginia 22030, USA;* [3]*J. Craig Venter Institute, Rockville, Maryland 20850, USA;* [4]*National Biodefense Analysis and Countermeasures Center, Frederick, Maryland 21702, USA*

Long-read single-molecule sequencing has revolutionized de novo genome assembly and enabled the automated reconstruction of reference-quality genomes. However, given the relatively high error rates of such technologies, efficient and accurate assembly of large repeats and closely related haplotypes remains challenging. We address these issues with Canu, a successor of Celera Assembler that is specifically designed for noisy single-molecule sequences. Canu introduces support for nano-

github.com/marbl/canu

marbl / canu Public

Watch 55 Star 497 Fork 161

Code Issues 19 Pull requests 2 Actions Projects 1 Wiki Security Insights

master 20 branches 19 tags Go to file Add file Code

brianwalenz Guard against OG == nullptr when called from layoutReads. 08b0347 on 18 Oct 10,284 commits

.github	Add an issue template.	5 years ago
documentation	Update quick-start.rst	4 months ago
scripts	Bump version number.	4 months ago
src	Guard against OG == nullptr when called from layoutReads.	last month
.gitmodules	Use seqrequester submodule for making sqStoreDumpMetaData h...	2 years ago
.readthedocs.yml	Create .readthedocs.yml	2 years ago
README.license.GPL	Move old licensing/ files to root.	6 years ago
README.licenses	Add PacBio licenses.	5 years ago
README.md	Complain about .zip files missing submodules.	15 months ago

README.md

Canu

Canu is a fork of the Celera Assembler, designed for high-noise single-molecule sequencing (such as the PacBio RS II/Sequel or Oxford Nanopore MinION).

Canu is a hierarchical assembly pipeline which runs in four steps:

About

A single molecule sequence assembler for genomes large and small.

canu.readthedocs.io/

Readme

Releases 15

Canu v2.2 Latest
on 27 Aug

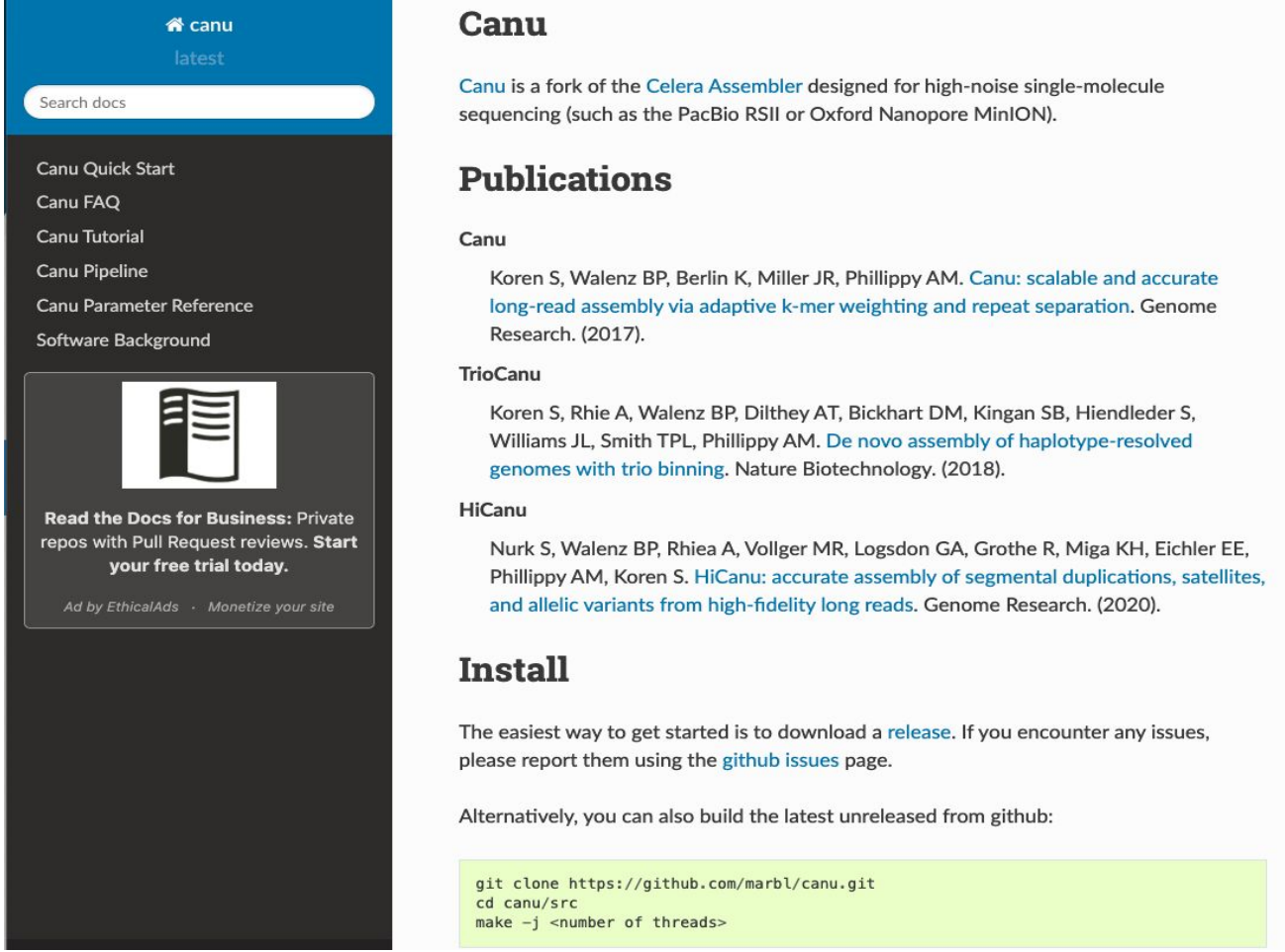
+ 14 releases

Packages

No packages published

Contributors 20

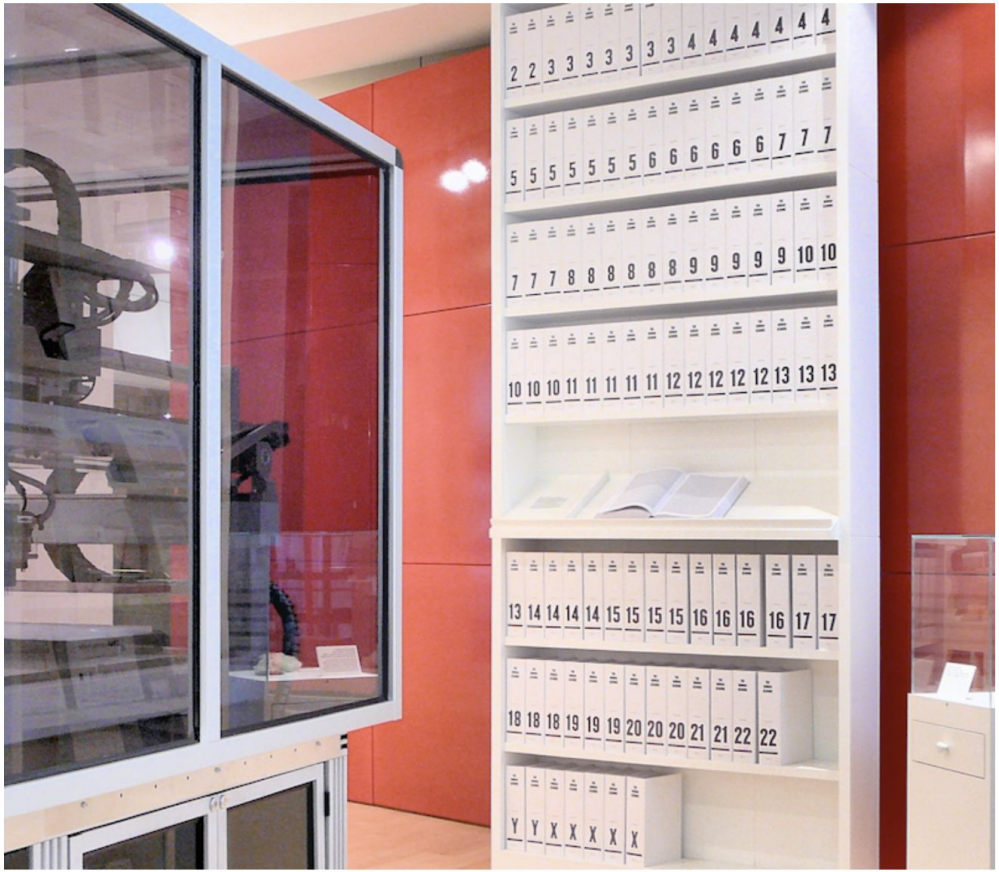
+ 9 contributors



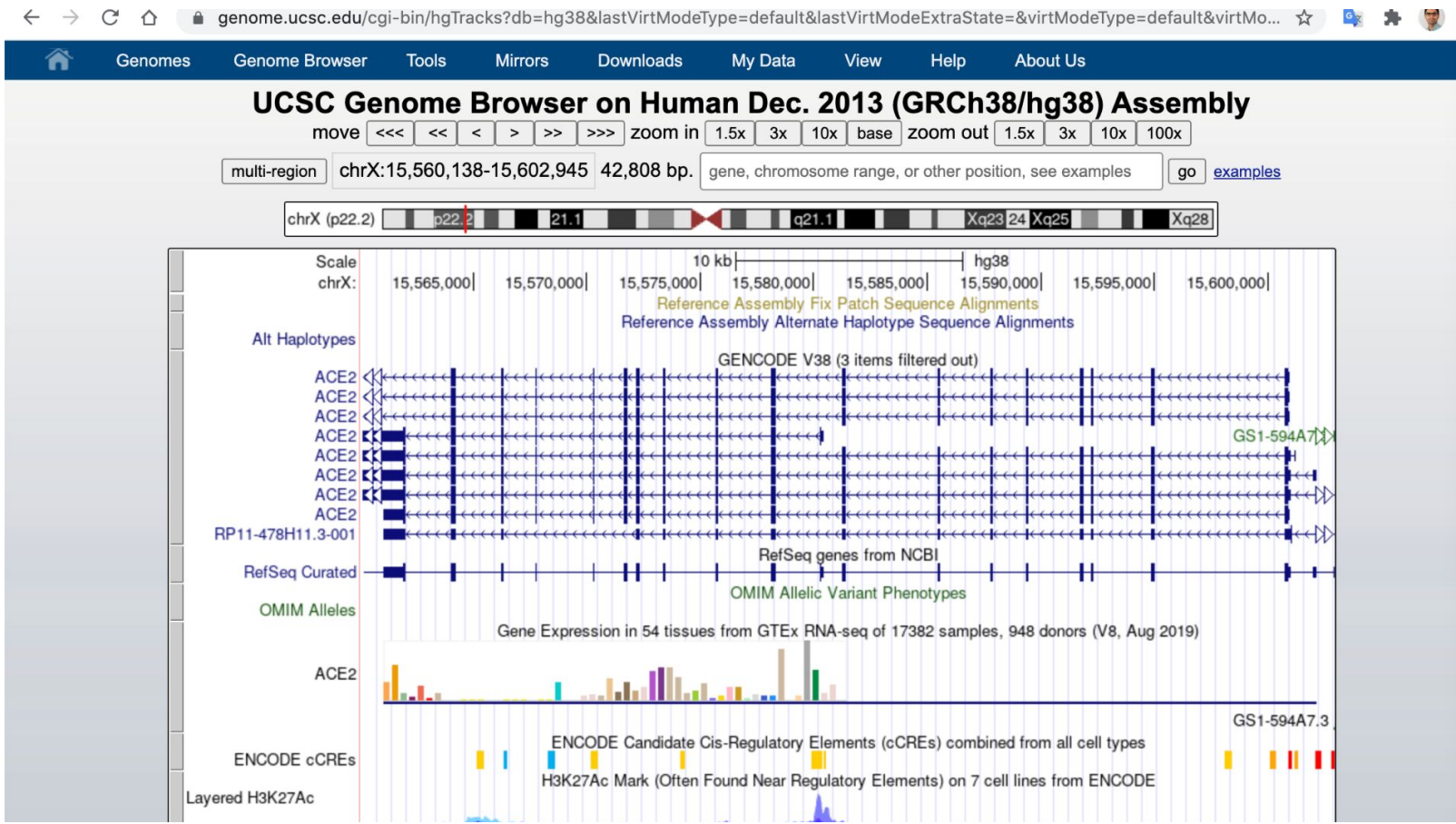
https://canu.readthedocs.io/en/latest/

Reference Genome

Using illumina+pacbio+HiC

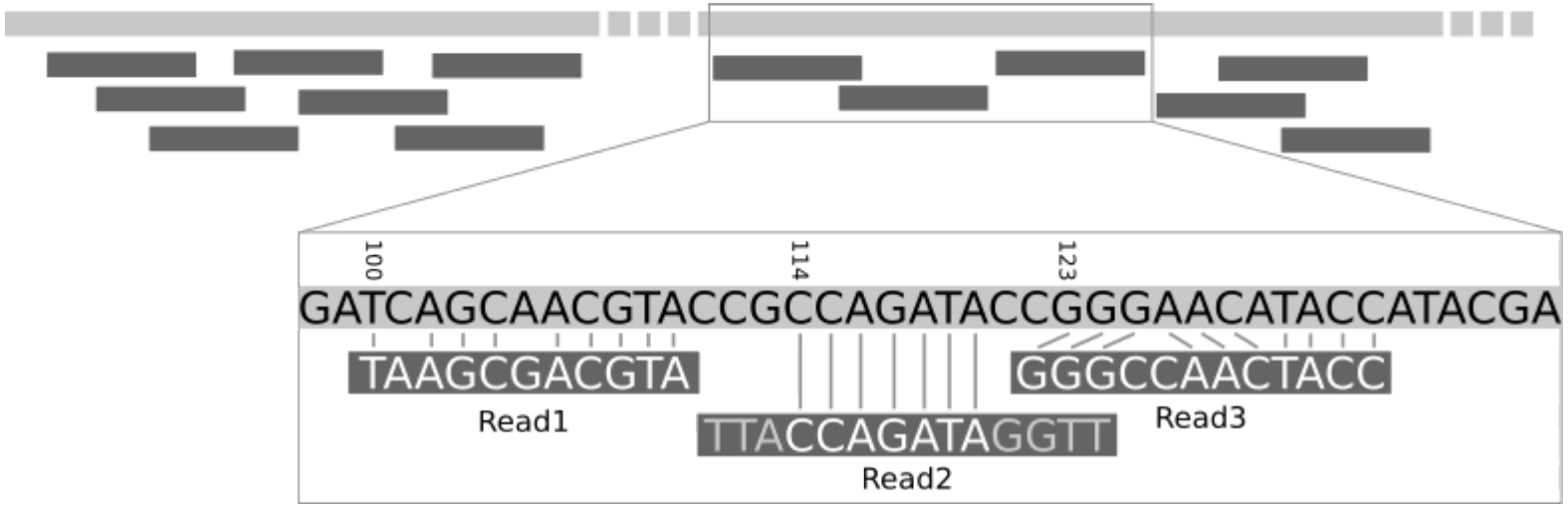


Annotation


genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMo...
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
chrX:15,560,138-15,602,945
42,808 bp.
gene, chromosome range, or other position, see examples
go
examples
chrX (p22.2)
Scale
chrX:
10 kb
hg38
15,565,000
15,570,000
15,575,000
15,580,000
15,585,000
15,590,000
15,595,000
15,600,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE V38 (3 items filtered out)
ACE2
GS1-594A7
RP11-478H11.3-001
RefSeq Curated
RefSeq genes from NCBI
OMIM Alleles
OMIM Allelic Variant Phenotypes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
GS1-594A7.3
ENCODE cCREs
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

Index

Read alignment/mapping


100
114
123
GATCAGCAACGTACCGCCAGATACCGGGAACATACCATACGA
TAAGCGACGTA
Read1
TTACCAGATAGGTT
Read2
GGGCCAACTACC
Read3

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals.

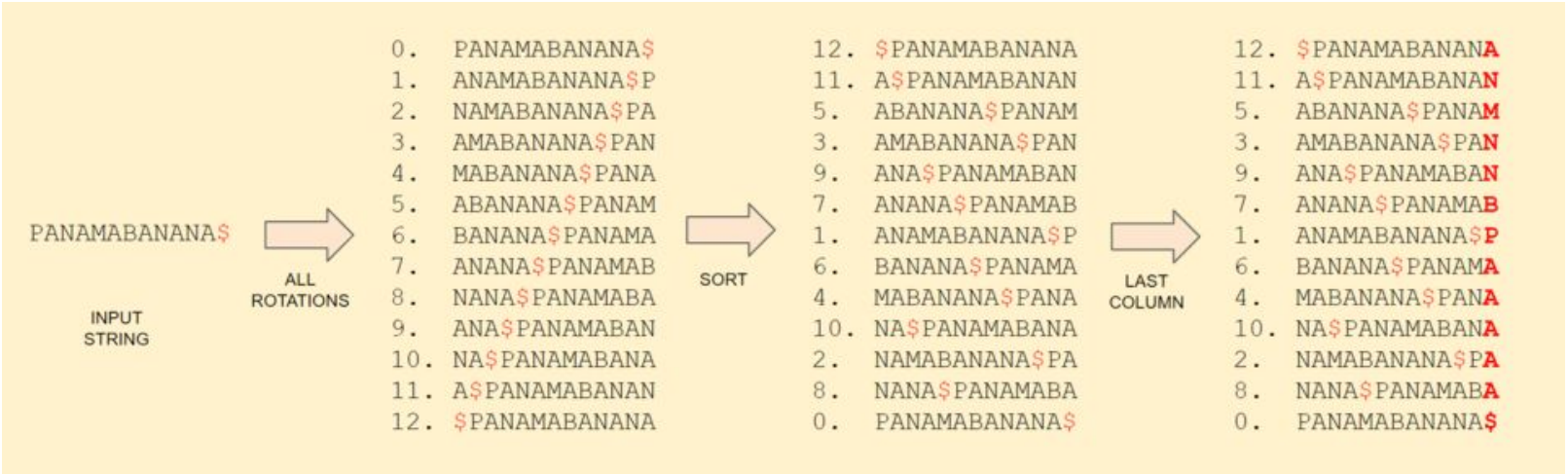
Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short of scanning the whole genome when few reads are aligned. The second category of software, including SOAPv1 (Li *et al.*, 2008b), PASS (Campagna *et al.*, 2009), MOM (Eaves and Gao, 2009), ProbeMatch (Jung Kim *et al.*, 2009), NovoAlign (http://www.novocraft.com), ReSEQ (http://code.google.com/p/re-seq), Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) and BFAST (http://genome.ucla.edu/bfast), hash the genome. These programs can be easily parallelized with multi-threading, but they usually require large memory to build an index for the human genome. In addition, the iterative strategy frequently introduced by these software may make their speed sensitive to the sequencing error rate. The third category includes slider (Malhis *et al.*, 2009) which does alignment by merge-sorting the reference subsequences and read sequences.

BWA: http://bio-bwa.sourceforge.net/

minimap2: https://github.com/lh3/minimap2

Burrows–Wheeler transform

A Block-sorting Lossless Compression Algorithm. Used in bzip



PANAMABANANA” -> ANMNNBPAAAAA$ -> ANM2NBP5A$

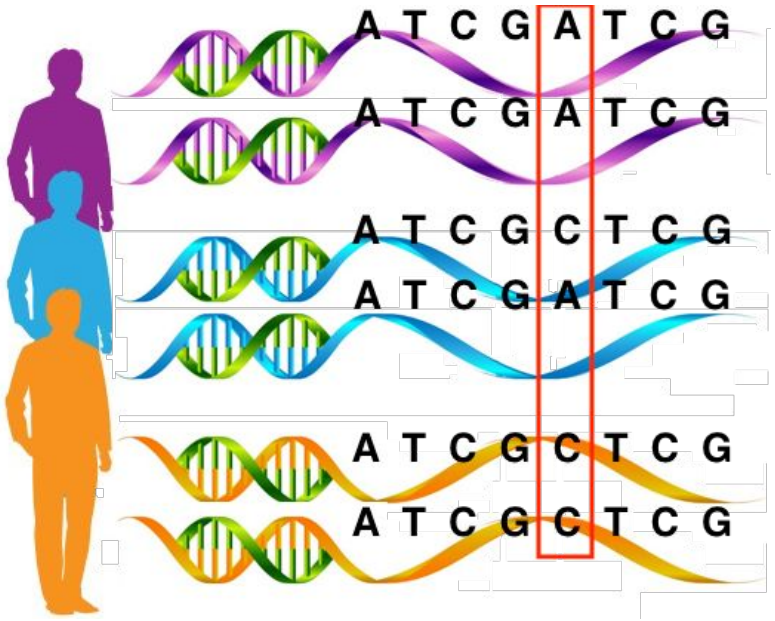
mr-easy.github.io/2019-12-12-burrows-wheeler-transform/

Index

Genetic variation

the difference in DNA among individuals.

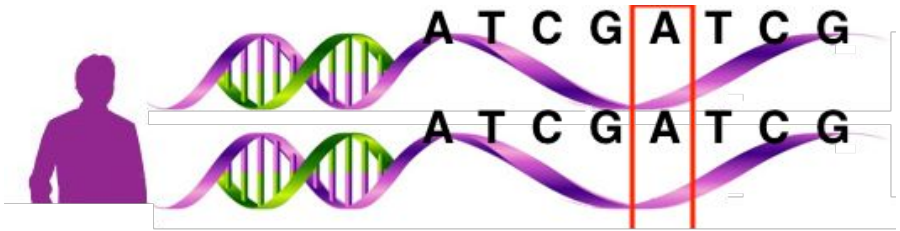
polymorphism



Adapted from nutrigeneticsspecialists.com/single-post/2017/03/27/what-is-a-snp

Aligned reads person1

Ref	ATCG**A**TCG
Read1	ATCG**A**
Read2	ATCG**A**
Read3	G**A**TCG
Read4	G**A**TCG



Aligned reads of person2

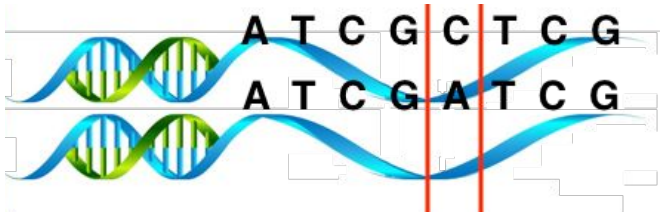
Ref ATCG**A**TCG

Read1 ATCG**C**

Read2 ATCG**A**

Read3 G**C**TCG

Read4 G**A**TCG



Variant calling & Genotyping

Individual 1

Ref ATCG**A**TCG

Read1 ATCG**A**

Read2 ATCG**A**

Read3 G**A**TCG

Read4 G**A**TCG

position 5: A/A

Individual 2

Ref ATCG**A**TCG

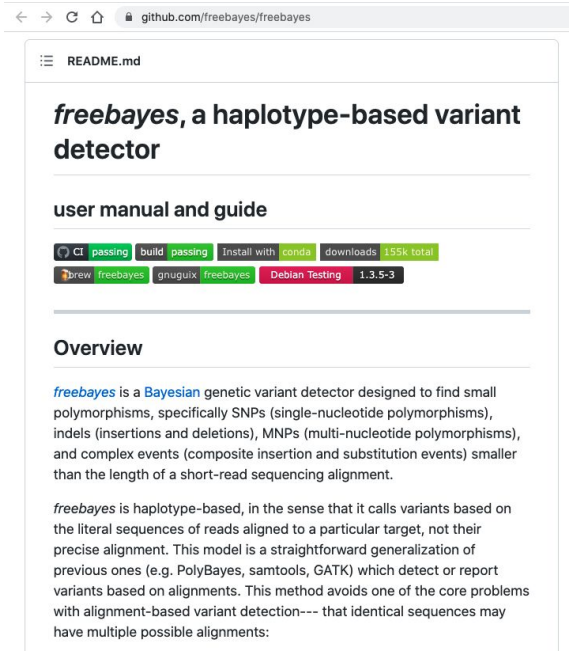
Read1 ATCG**C**

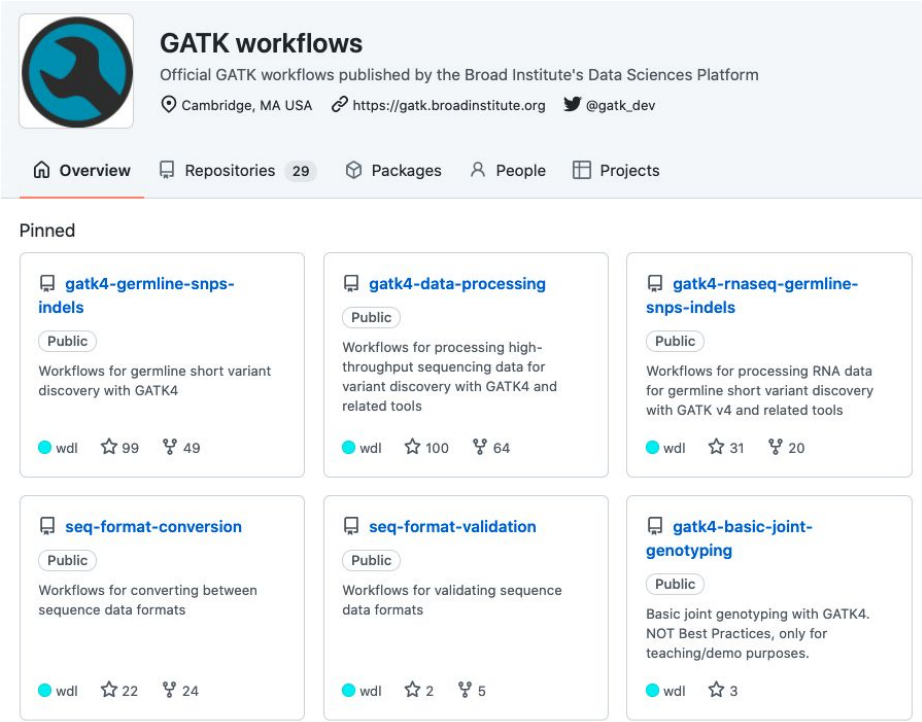
Read2 ATCG**A**

Read3 G**C**TCG

Read4 G**A**TCG

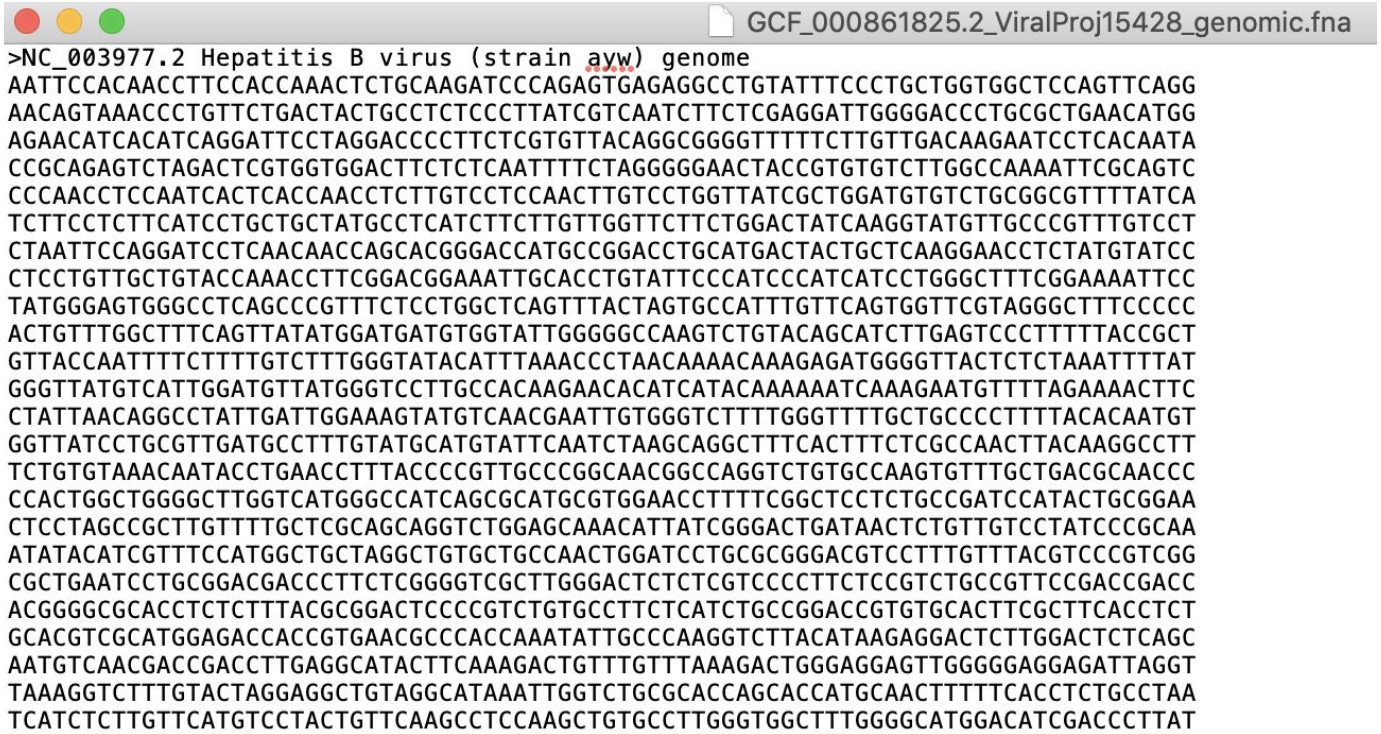
position 5: A/C


github.com/freebayes/freebayes
README.md
freebayes, a haplotype-based variant detector
user manual and guide
CI passing
build passing
Install with conda
downloads 155k total
brew freebayes
gnuguix freebayes
Debian Testing 1.3.5-3
Overview
freebayes is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment.
freebayes is haplotype-based, in the sense that it calls variants based on the literal sequences of reads aligned to a particular target, not their precise alignment. This model is a straightforward generalization of previous ones (e.g. PolyBayes, samtools, GATK) which detect or report variants based on alignments. This method avoids one of the core problems with alignment-based variant detection--- that identical sequences may have multiple possible alignments:


GATK workflows
Official GATK workflows published by the Broad Institute's Data Sciences Platform
Cambridge, MA USA
https://gatk.broadinstitute.org
@gatk_dev
Overview
Repositories 29
Packages
People
Projects
Pinned
gatk4-germline-snps-indels
Public
Workflows for germline short variant discovery with GATK4
wdl 99 49
gatk4-data-processing
Public
Workflows for processing high-throughput sequencing data for variant discovery with GATK4 and related tools
wdl 100 64
gatk4-rnaseq-germline-snps-indels
Public
Workflows for processing RNA data for germline short variant discovery with GATK v4 and related tools
wdl 31 20
seq-format-conversion
Public
Workflows for converting between sequence data formats
wdl 22 24
seq-format-validation
Public
Workflows for validating sequence data formats
wdl 2 5
gatk4-basic-joint-genotyping
Public
Basic joint genotyping with GATK4. NOT Best Practices, only for teaching/demo purposes.
wdl 3

Index

fasta Format



fasta Format

- DNA/RNA/Protein sequence

- Header line, begins with >
- Each line 50/60 characters

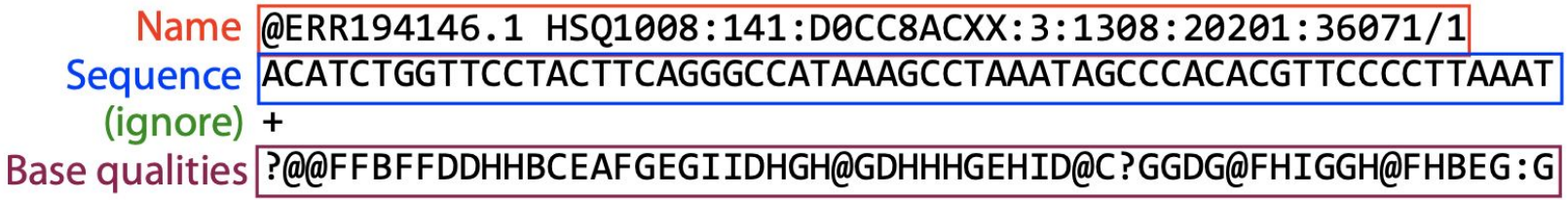
- Extension: .fa .fna .fasta

fastq format

sra_data.fastq.txt

```
@SRR16356243.194720.1 194720 length=151
GGCAGCAAAGTTTTATTGTAAAATAAGAGATCGATATAAAAATGGGATATAAAAAGGGAGAAGGAGGGGAAGGGTGGGGTGAAAATGCAG
ATGTGCTTGCAGAATGTAAAAGATGTTGACCCTTCCAGCTGGACGTGGTGGCTCCAATTGT
+SRR16356243.194720.1 194720 length=151
AAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF/
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFF/
@SRR16356243.194720.2 194720 length=151
CAATTGGAGCCACCACGTCCAGCTGGAAGGGTCAACATCTTTTACATTCTGCAAGCACATCTGCATTTTCACCCCACCCTTCCCCTCCTT
CTCCCTTTTTATATCCCATTTTTATATCGATCTCTTATTTTACAATAAAACTTTGCTGGCA
+SRR16356243.194720.2 194720 length=151
FFFFFFFAFFFFFF/FFFFFAFFFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF/
FFFFFFFFFAFFAFAAAFFFFFFFFFFF=AFF/FFFFFFFFFF/FFFAFF=FFF/=F/FAFFFFFFFF6FF/FF
@SRR16356243.247542.1 247542 length=151
GTTTGACTTAGTATGAATGTGGTTACGTTGGAAGCAAATGTGTCTTCACTTTATCATGAAAAAGTCTGCAAGTGCTCTGCGACGTCCCAG
GGAAATGATCCTACCCTCAATCTTCAGCTCAAAGGGAACCTTTGCTCTTTTTCAGTGACCA
+SRR16356243.247542.1 247542 length=151
FAFF/FA=FAFF/FFFFF/FFAFA///F/A/FF/FFFFAAA///FF/FFFF///A////FFFFFA==//FF/F6FFA//A/FFF//
=F///F/AFA///F/=////F/=/FAA/AF//A/F/FF/AFFF//FFF/F//FFFFFF/F/////
@SRR16356243.247542.2 247542 length=151
ATTGGACACTGAAAAAGAGAAAAGGTTCCCTGTGAGCTGAAGAGTGAGTTTAGGATCATTTCCAAGGGACGTCGCAGAGCACTTGCAGAC
TTTATCATGAAATAGTTAAGACACATTTGATTCCAACGAAACCACATTCAAACTAAGTCAA
+SRR16356243.247542.2 247542 length=151
//F/F/FF/=/FAFFFAFF/F/FF///F//FF/FFFA//FA/F//A/F//FF//AFAF//F//AF6/////FAFFFFAA/F//FF///
FF//F/F/FF/A/F/FF//AA/A/FFA/FA//FFFFF/A/AA///F/A/FFFF/F/FF///=A
```

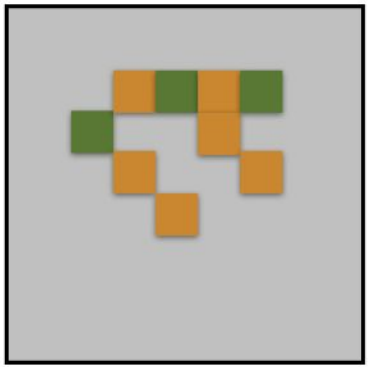
Storing sequencing reads



Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
||||||||||||||||||||||||||||||||
HHHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

Base quality is ASCII-encoded version of $Q = -10 \log_{10} p$



Call: orange (C)

Estimate *p*, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

$Q = -10 \log_{10} 1/3 = 4.77$

$$Q = -10 \cdot \log_{10} p$$

Base quality

Probability that base call is incorrect

$Q = 10 \rightarrow$ 1 in 10 chance call is incorrect

$Q = 20 \rightarrow$ 1 in 100

$Q = 30 \rightarrow$ 1 in 1,000

ASCII table

phred33ToQ(qual): ord(qual)-33

QtoPhred33(Q): chr(int(round(Q)) + 33)

43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z

SAM format

ST-E00223:32:H5J57CCXX:6:2123:15189:52872 97 1 10001 0 4S15M1I54M2I50M25S 4 699063 0
ACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCAACCCTAACCCCCCCCCACCCAACCCCACCCCCCAC
AAFFFKKKKKKKKKKKKKKKKKKKKFKKFKKKKKKFKKAKKKKFFKKKKKFKF<FF7FFFFK7FK,AA,FKFFKF,FKKK7<KKK,,7FFF7F,AAFFK,AKA,,FFF<,A7FKK,,,7A,,,7,,,,,((,(,,(
,,,,,,,(,,A(,,(MC:Z:80S11M1D58M MD:Z:119 RG:Z:15-0017315_1 NM:i:3 MQ:i:47 AS:i:104 XS:i:103
ST-E00223:46:HG7V5CCXX:2:1116:12601:22862 1123 1 10006 0 81M70S = 10106 143
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACACTCACCCTAACCCTACCCCTAACCCTAACCCTATCATTCACTCGAACCCTAACACTACCGCTAG
CGCTAACTCCAGCCCGCACACTATCGCTAACCCTCACGC
AAFFFKKKKKKKKKKKKKKKKFKKA,,A,A,,77<,,A,,7FFK7,,,,AF,,,A7,,,77,,,7AA,,7,FFK,<A,,,,7<<,,,,AA,7,AA,,7,,,7,,,,,,,,,,A7,,,,,,7,7,7(,(,(((,,,7,A,F,7,,,<,,AA,, MC:Z:108S43M
MD:Z:47C2A12A17 RG:Z:15-0017315_1 NM:i:3 MQ:i:2 AS:i:66 XS:i:71
ST-E00223:32:H5J57CCXX:5:2208:10074:43308 99 1 10008 36 101M1I41M7S = 10107 137
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTACCCCTAACCCTAACCCTAACCCTAACCCTCACCCTCACCCTCACCCTAGCCCAAACCCTAACC
CTAACCCTAACCCTAACCCTAACCCTAACCCTACCCCG
AAFFFKFKKKKKKKKKKKKKKKKKFKKKKKKKFKKK7<KA7<F,AF7F,7,A,FA7<KKKKKFKKKK<,,FKAKKAFFA,A7,7,,,7,7,,,,,7F<,,7,,,,7,,A,,,<F7,,,7FA,7,,<F,,,<A,7<AFKK<,<,,7,,,,
,(MC:Z:112S38M MD:Z:49A28A5A5A6A44 RG:Z:15-0017315_1 NM:i:6 MQ:i:36 AS:i:110 XS:i:113
ST-E00223:46:HG7V5CCXX:5:2119:12936:64896 99 1 10013 0 90M61S = 10176 211
TAACCCTAAGCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCATAACCCAAACCCTAACCCTAACCCTAACCCGAACCGTAAGCCAAAACATAAC
CACAACCATAACAATAACCAAAACCTTAACGTTAAACAT
A<AFFKKKK,AAAFKK<FKKKKAK,<A,AFKKKKAFFFKKKKAFKKFAFFFKKA7,FFA,F,F,7F,AFF77AFFAFFKAFKKA,,FFKF,AA,F,,7F,A,,7A,,,7A,,,,,,AFK,,,AA,,,7FKA,A,AFF,<,,,,,
,,77<AA MC:Z:99S48M4S MD:Z:9C49C6T23 RG:Z:15-0017315_1 NM:i:3 MQ:i:0 AS:i:75 XS:i:75
ST-E00223:32:H5J57CCXX:1:1205:17290:54577 99 1 10019 1 92M59S = 10354 414
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAAACCCCCCACCCCAAACCCAAG
CCCAACCCTACCCCCTAACCCCTAACCCCAACCCTGACC
AAFFFKKKKFKKKAKKKKKKKKKKKKKKKKF7KKKKKKKKKKKKKKFKKKKFKKKKKKKAKKKFFKKFFKKKK,<,7<FA,FFFK,,7FFF,,,,,A,((,(((,7,7A,,,,,A,,,777A,AAKK<<,7FFA7,A,,
AA,<,,AFA,7AF MC:Z:72S79M MD:Z:92 RG:Z:15-0017315_1 NM:i:0 MQ:i:20 AS:i:92 XS:i:97

Sequence Alignment Map (SAM)

- for storing reads aligned to reference
- Tab-delimited text
- header section: lines start with @

Read1 163 chr20 7 20 8M2I4M1D3M ..

- read name
- Flag
- chr
- Position of start alignment on the reference sequence
- Mapping quality
- CIGAR

BAM: Binary representation of SAM; compressed version.

VCF (Variant Call Format)

```
##fileformat=VCFv4.3
##fileDate=20090805

#CHROM  POS      ID          REF   ALT    QUAL  FILTER   INFO      FORMAT   NA00001
20      14370    rs6054257   G     A      29    PASS     AF=0.5     GT       0|0
20      17330    .           T     A      3     q10      AF=0.017   GT       0/1
```

Index

OS: operating system

A system software that manages computer hardware, software resources, and provides common services for computer programs.

For computers:

- Windows
- Linux
- macOS

Why linux?

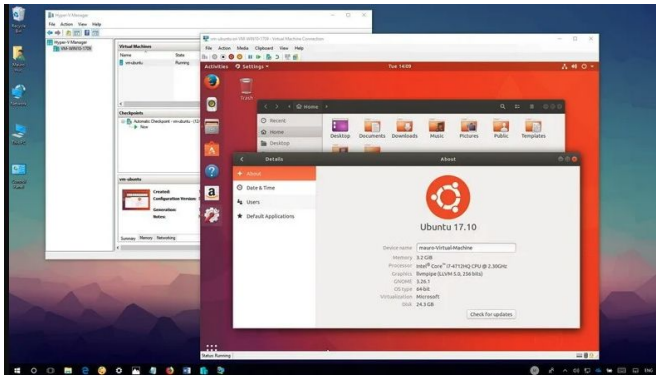
- many bioinformatics software developed only for Linux.
  - Graphical or web user interfaces exists but often struggle to provide flexibility
- free (!)
- performance
- security (virus)

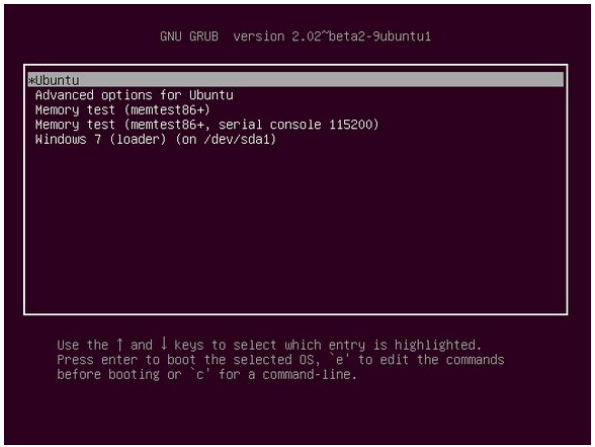
Linux on a computer with windows

1- connect to a Linux machine (server) using PuTTy

2- install on a drive (D/E not C) [danger]

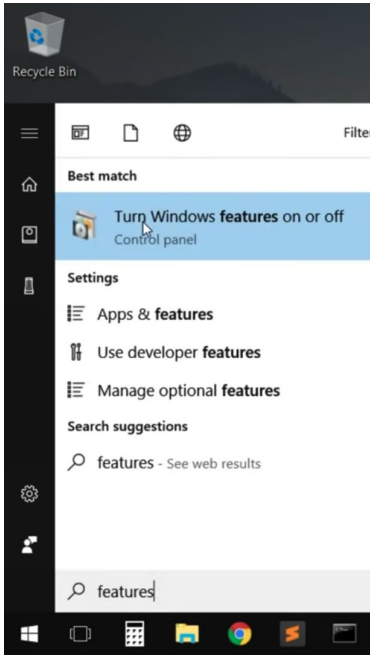
3- use VirtualBox or VMware

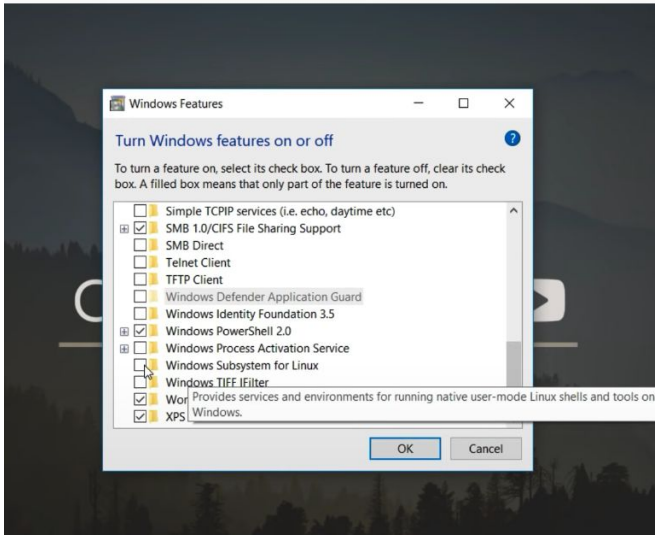




4-install bash on windows

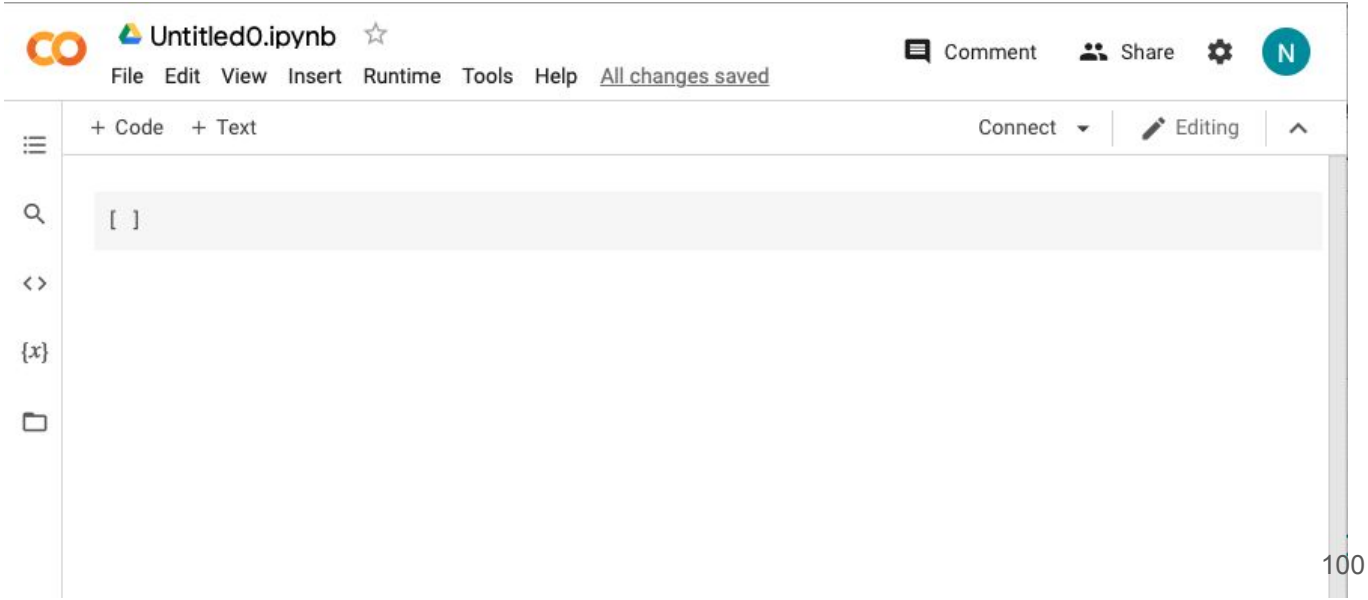
https://www.laptopmag.com/articles/use-bash-shell-windows-10





5- use online Linux terminal

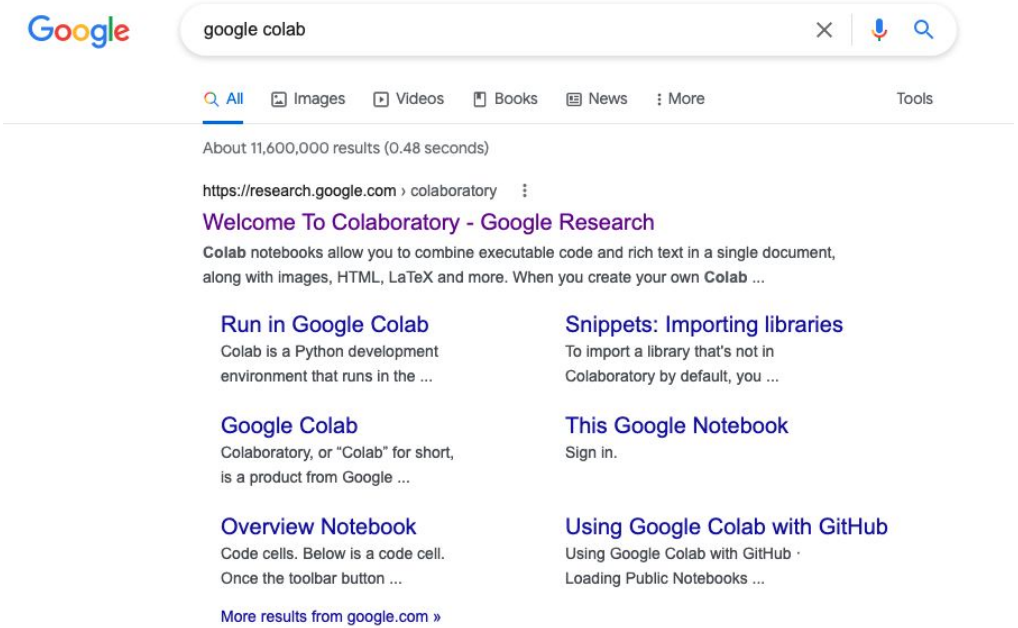
- best is https://research.google.com/colaboratory/
- https://copy.sh/v86/?profile=buildroot
- https://bellard.org/jslinux/vm.html?url=alpine-x86.cfg&mem=192
- http://cb.vu/

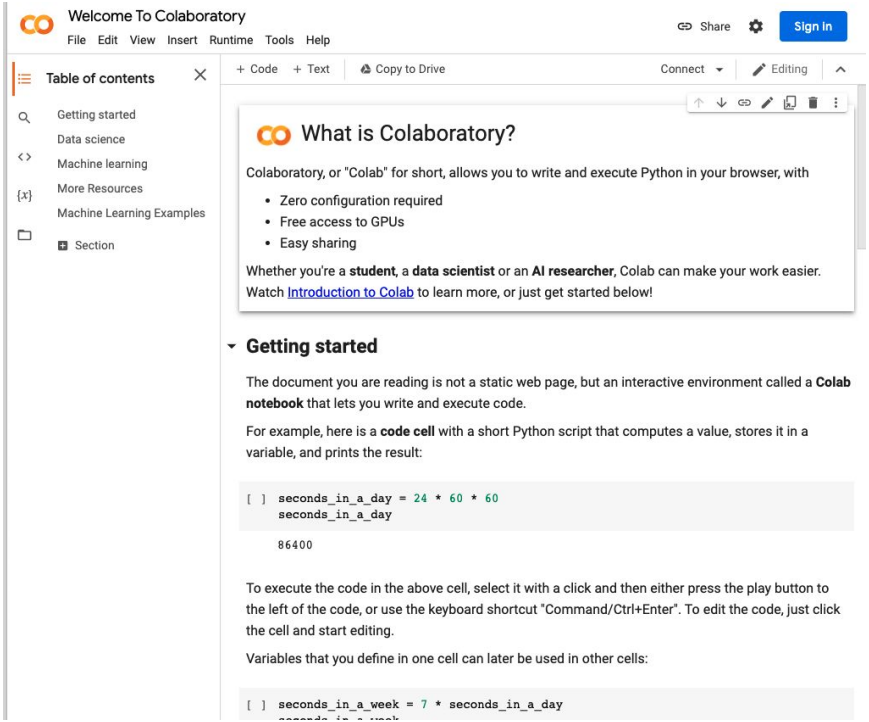


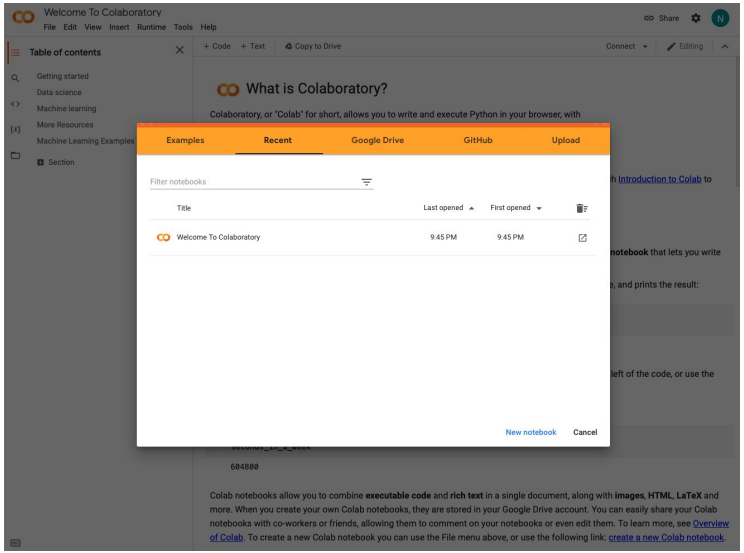
To save time for the next section

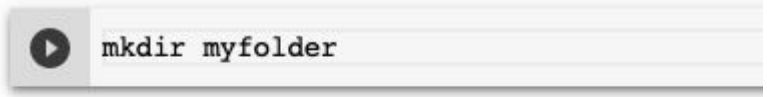
Create a gmail account

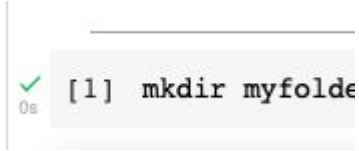
(if you don’t have)

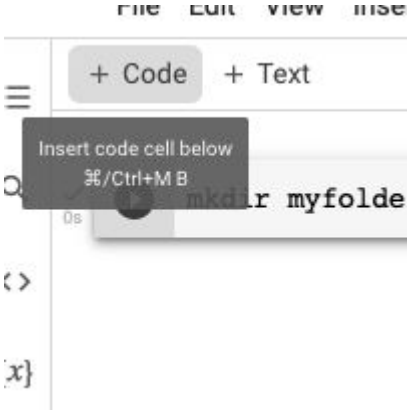



Welcome To Colaboratory
File Edit View Insert Runtime Tools Help
Share
Sign in
Table of contents
+ Code + Text
Copy to Drive
Connect
Editing
Getting started
Data science
Machine learning
More Resources
Machine Learning Examples
Section
What is Colaboratory?
Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with
Zero configuration required
Free access to GPUs
Easy sharing
Whether you're a student, a data scientist or an AI researcher, Colab can make your work easier.
Watch Introduction to Colab to learn more, or just get started below!
Getting started
The document you are reading is not a static web page, but an interactive environment called a Colab notebook that lets you write and execute code.
For example, here is a code cell with a short Python script that computes a value, stores it in a variable, and prints the result:
[ ] seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
86400
To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter". To edit the code, just click the cell and start editing.
Variables that you define in one cell can later be used in other cells:
[ ] seconds_in_a_week = 7 * seconds_in_a_day


Welcome To Colaboratory
Examples
Recent
Google Drive
GitHub
Upload
Filter notebooks
Title
Last opened
First opened
Welcome To Colaboratory
9:45 PM
9:45 PM
New notebook
Cancel
604800
Colab notebooks allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab notebooks with co-workers or friends, allowing them to comment on your notebooks or even edit them. To learn more, see Overview of Colab. To create a new Colab notebook you can use the File menu above, or use the following link: create a new Colab notebook.


mkdir myfolder


[1] mkdir myfolde
0s


+ Code
+ Text
Insert code cell below
⌘/Ctrl+M B
mkdir myfolde
0s

```
[1] mkdir myfolder
```

```
[2] ls
```

```
myfolder/  sample_data/
```

```
[3] cd myfolder
```

```
/content/myfolder
```

```
[4] pwd
```

```
'/content/myfolder'
```

```
cd a
```

```
[Errno 2] No such file or directory: 'a'
/content/myfolder
```

Basic linux

- pwd

Print Working Directory

- ls

List files and directories

- man ls
- man pwd

Colab limitation

use ! before each command

pwd

mkdir yzd

ls yzd

cd yzd

mkdir medicine

mkdir math

ls -alht

cd ..

df -h

```
echo "yazd"

echo "yazd" > file.txt

cat file.txt

cp file file2

cp yazd/* folder1

rm mv zip unzip
```

mkdir

head -n 2

tail -n 2

grep

sed

awk

```
echo ">chr1"  > file
echo "aaaa tttt gggg" > > file

head –n 2 chr.fa
tail –n 5 chr.fa
wc –l chr.fa
sed –n 2,3p chr.fa
sed s/a/mm/ chr.fa
```

tiny dataset

copy the link to fasta file from github

https://github.com/brainstorm/tiny-test-data

```
!fgrep -o G hg19.fa | wc -l
```

$$\text{GC content} = \frac{G + C}{A + T + G + C} \times 100\%$$

Exercise: Calculate number of C

Index

NCBI Resources How To Sign in to NCBI

NCBI
National Center for Biotechnology Information

All Databases Search

UNITE

A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

LEARN MORE

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

Literature	
Bookshelf	2,371
MeSH	26
NLM Catalog	171
PubMed	55,565
PubMed Central	81,053

Genes	
Gene	825
GEO DataSets	1,081
GEO Profiles	54,675
HomoloGene	3
PopSet	1,254

Proteins	
Conserved Domains	30
Identical Protein Groups	98,143
Protein	195,051
Protein Family Models	104
Structure	151

Genomes	
Assembly	63
BioCollections	0
BioProject	238
BioSample	2,438
Genome	1
Nucleotide	179,903

Clinical	
ClinicalTrials.gov	1,277
ClinVar	27
dbGaP	40
dbSNP	0
dbVar	0
GTR	44

PubChem	
BioAssays	3,958
Compounds	5
Pathways	0
Substances	29

NCBI Resources How To

Assembly

Assembly | Hepatitis B virus

Create alert Advanced Browse by organism

Organism group
Viruses (63)
Customize ...

Status clear
Latest (63)
Latest GenBank (63)
Latest RefSeq (1)

Assembly level
Complete genome (53)
Chromosome (9)
Scaffold (1)
Contig (0)

RefSeq category
Reference (0)
Representative (0)

Exclude clear
Exclude partial (0)
Exclude from large multi-isolate project (0)
Exclude anomalous (0)
Customize ...

Annotation status
Has annotation (21)
GenBank has annotation (20)
RefSeq has annotation (1)

Taxonomy check status

Summary 20 per page Sort by Significance Send to:

Download Assemblies

TAXONOMY

Hepatitis B virus

Hepatitis B virus is a species of dsDNA-RT virus in the family Hepadnaviridae (hepatitis B-type viruses).

Taxonomy ID: 10407

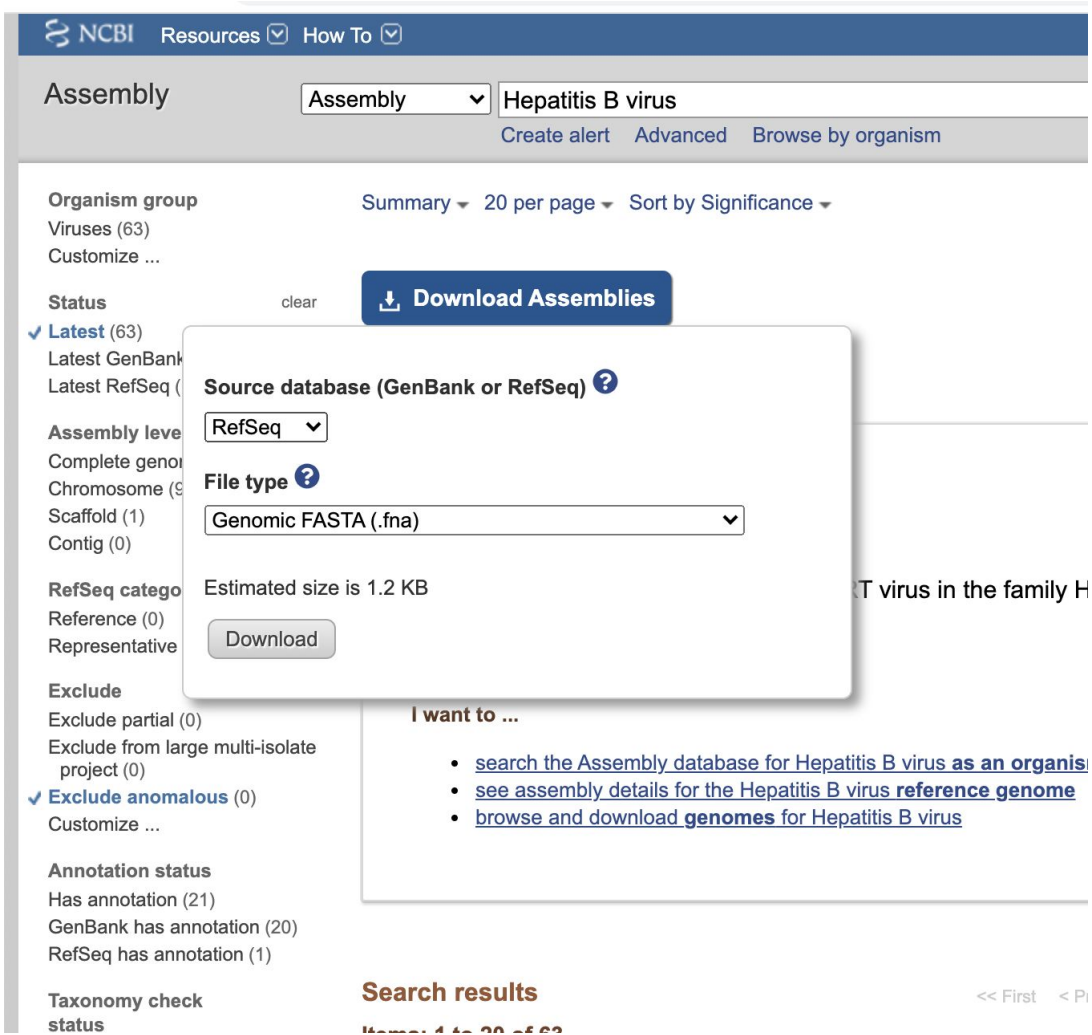
I want to ...

- search the Assembly database for Hepatitis B virus **as an organism**
- see assembly details for the Hepatitis B virus **reference genome**
- browse and download **genomes** for Hepatitis B virus

Search results

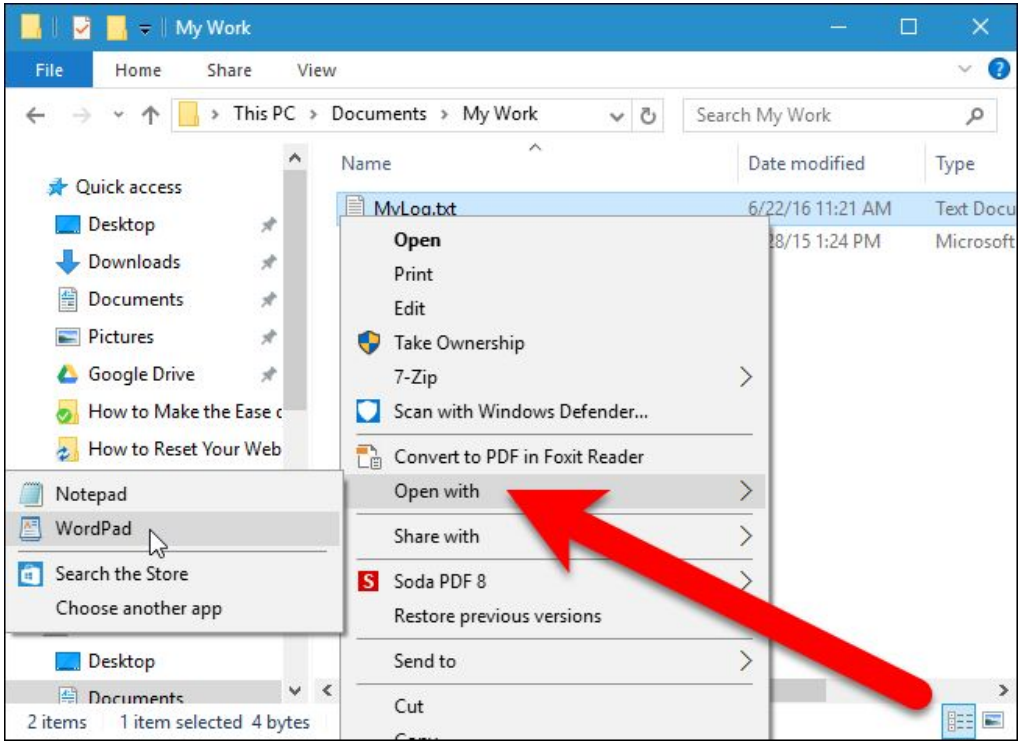
<< First < Prev Page 1 of 4 Next > Last >>

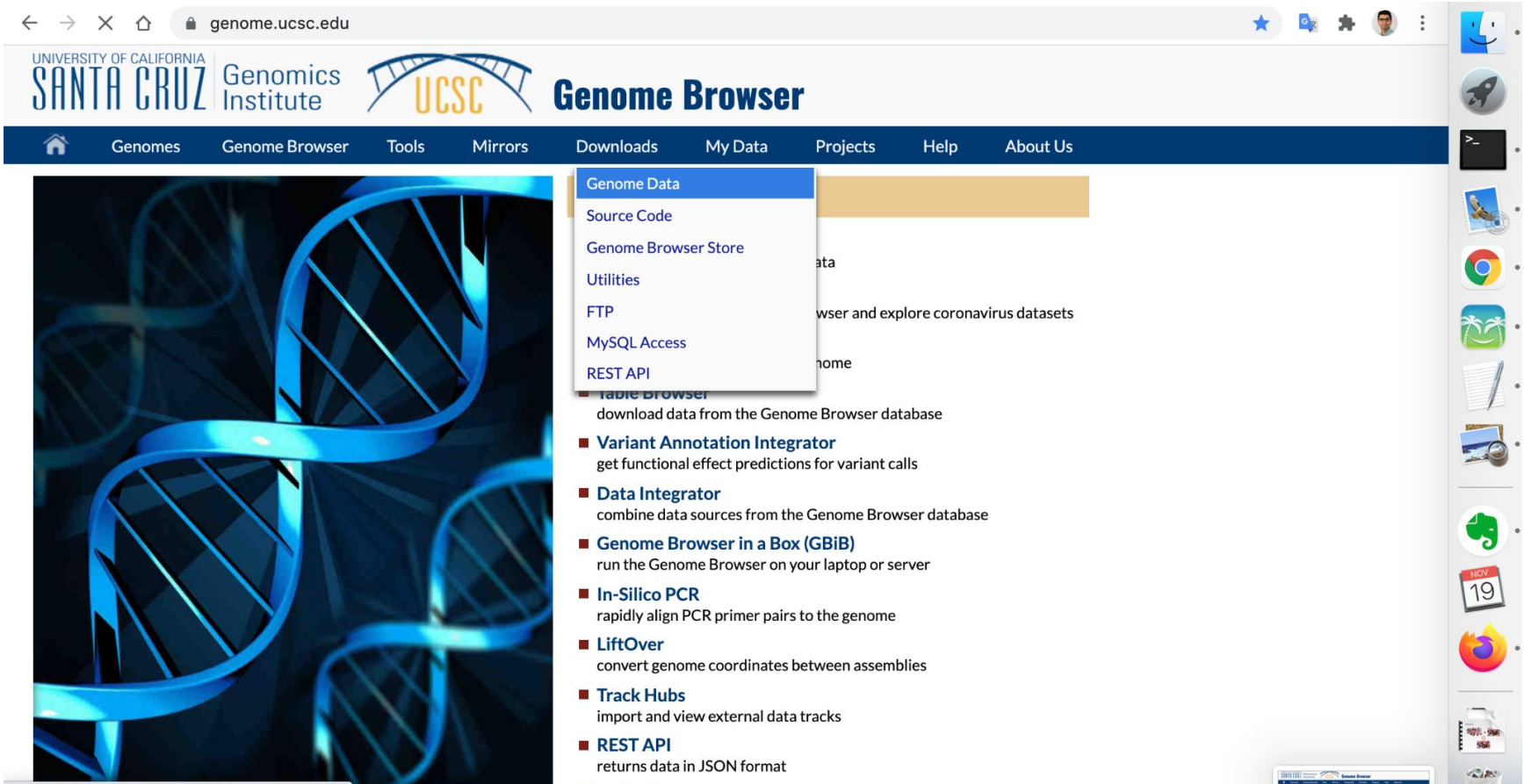
Items: 1 to 20 of 63

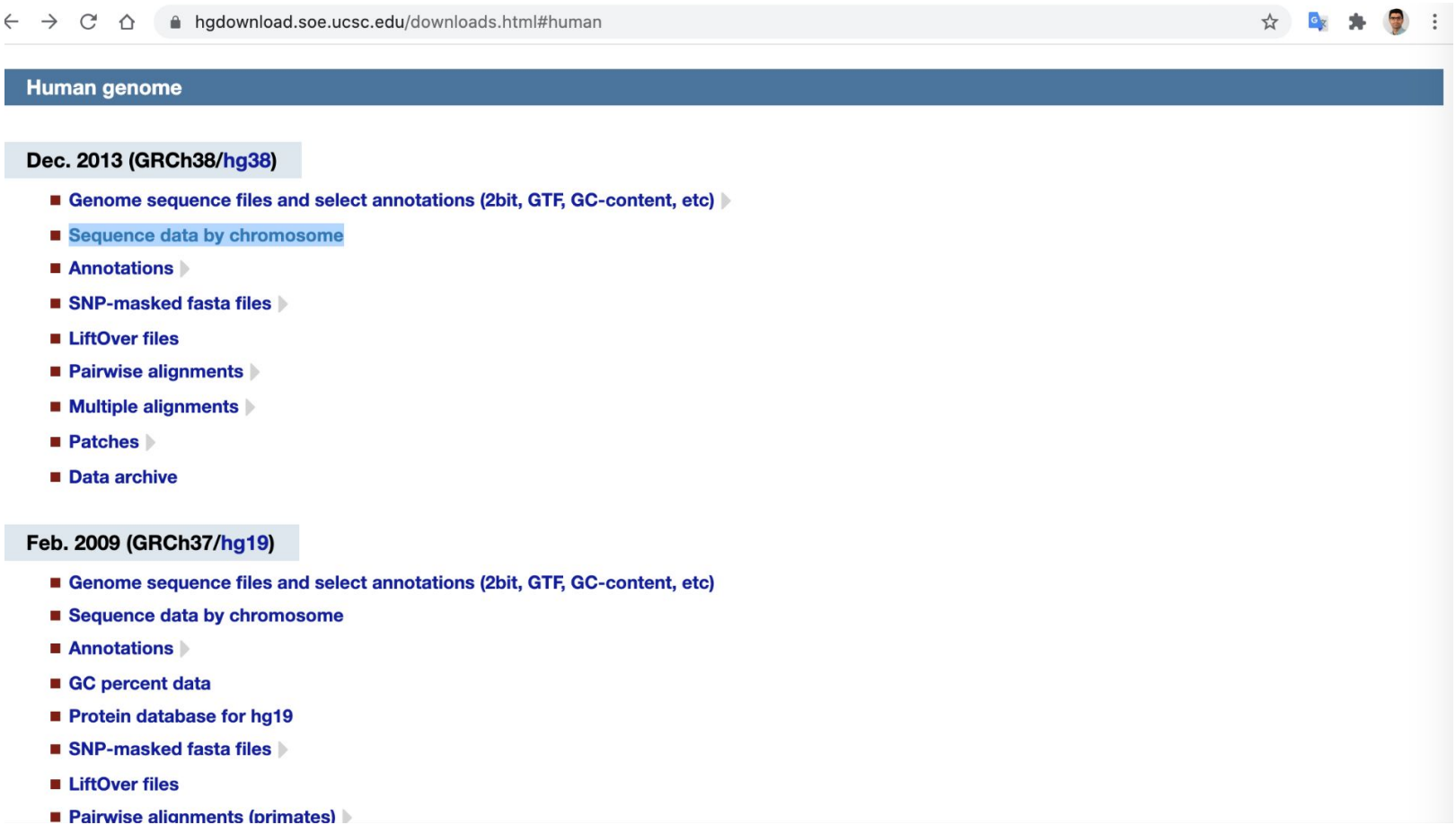

NCBI
Resources
How To
Assembly
Assembly
Hepatitis B virus
Create alert
Advanced
Browse by organism
Organism group
Viruses (63)
Customize ...
Status
clear
Latest (63)
Latest GenBank
Latest RefSeq (
Assembly leve
Complete genor
Chromosome (9
Scaffold (1)
Contig (0)
RefSeq catego
Reference (0)
Representative
Exclude
Exclude partial (0)
Exclude from large multi-isolate project (0)
Exclude anomalous (0)
Customize ...
Annotation status
Has annotation (21)
GenBank has annotation (20)
RefSeq has annotation (1)
Taxonomy check status
Summary
20 per page
Sort by Significance
Download Assemblies
Source database (GenBank or RefSeq)
RefSeq
File type
Genomic FASTA (.fna)
Estimated size is 1.2 KB
Download
T virus in the family H
I want to ...
search the Assembly database for Hepatitis B virus as an organis
see assembly details for the Hepatitis B virus reference genome
browse and download genomes for Hepatitis B virus
Search results
<< First
< Pr

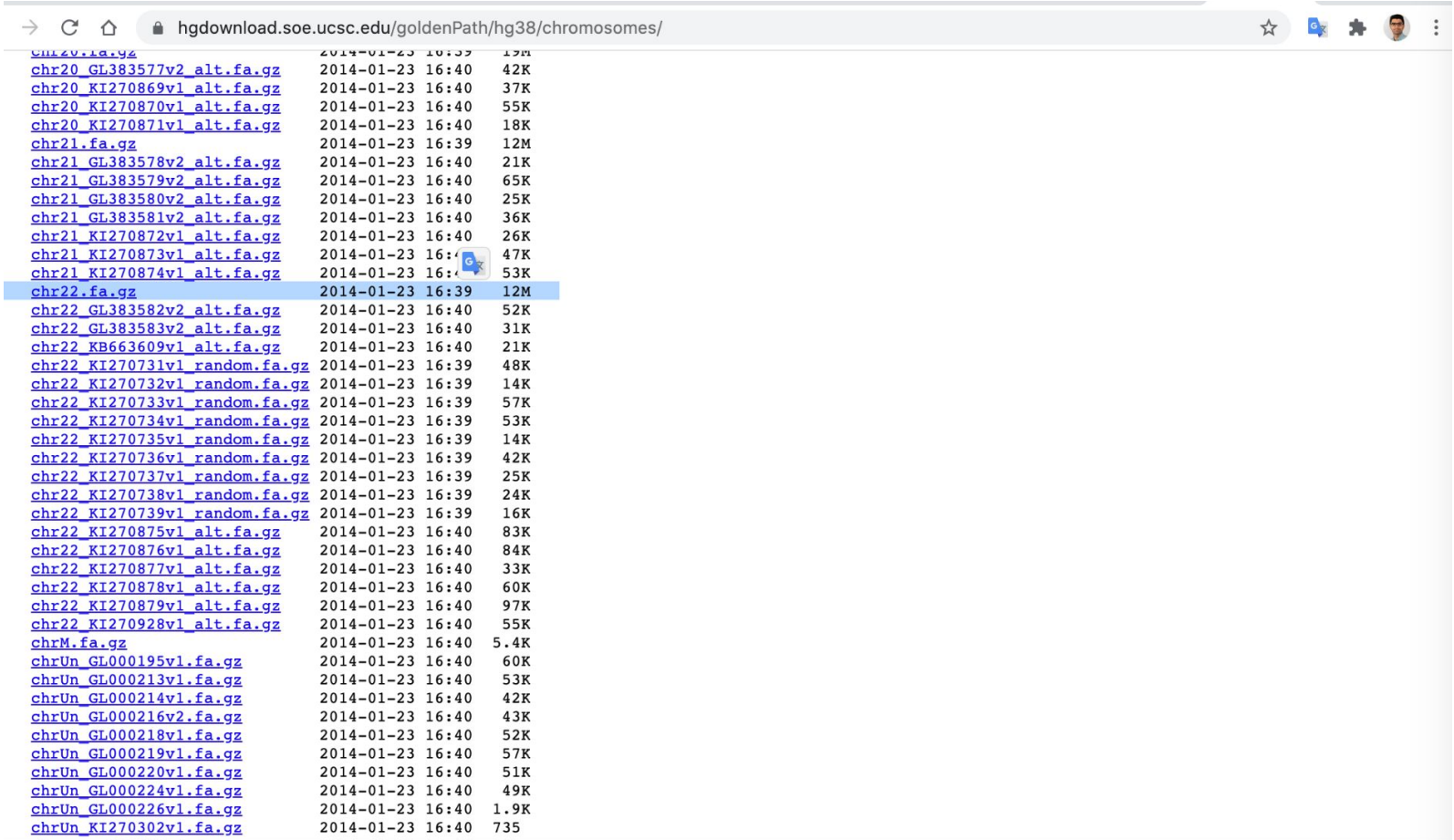
Unzip


VS
VS
7zip


My Work
File
Home
Share
View
This PC › Documents › My Work
Search My Work
Quick access
Desktop
Downloads
Documents
Pictures
Google Drive
How to Make the Ease
How to Reset Your Web
Name
Date modified
Type
MyLog.txt
6/22/16 11:21 AM
Text Docu
8/15 1:24 PM
Microsoft
Open
Print
Edit
Take Ownership
7-Zip
Scan with Windows Defender...
Convert to PDF in Foxit Reader
Open with
Share with
Soda PDF 8
Restore previous versions
Send to
Cut
Notepad
WordPad
Search the Store
Choose another app
Desktop
Documents
2 items
1 item selected 4 bytes

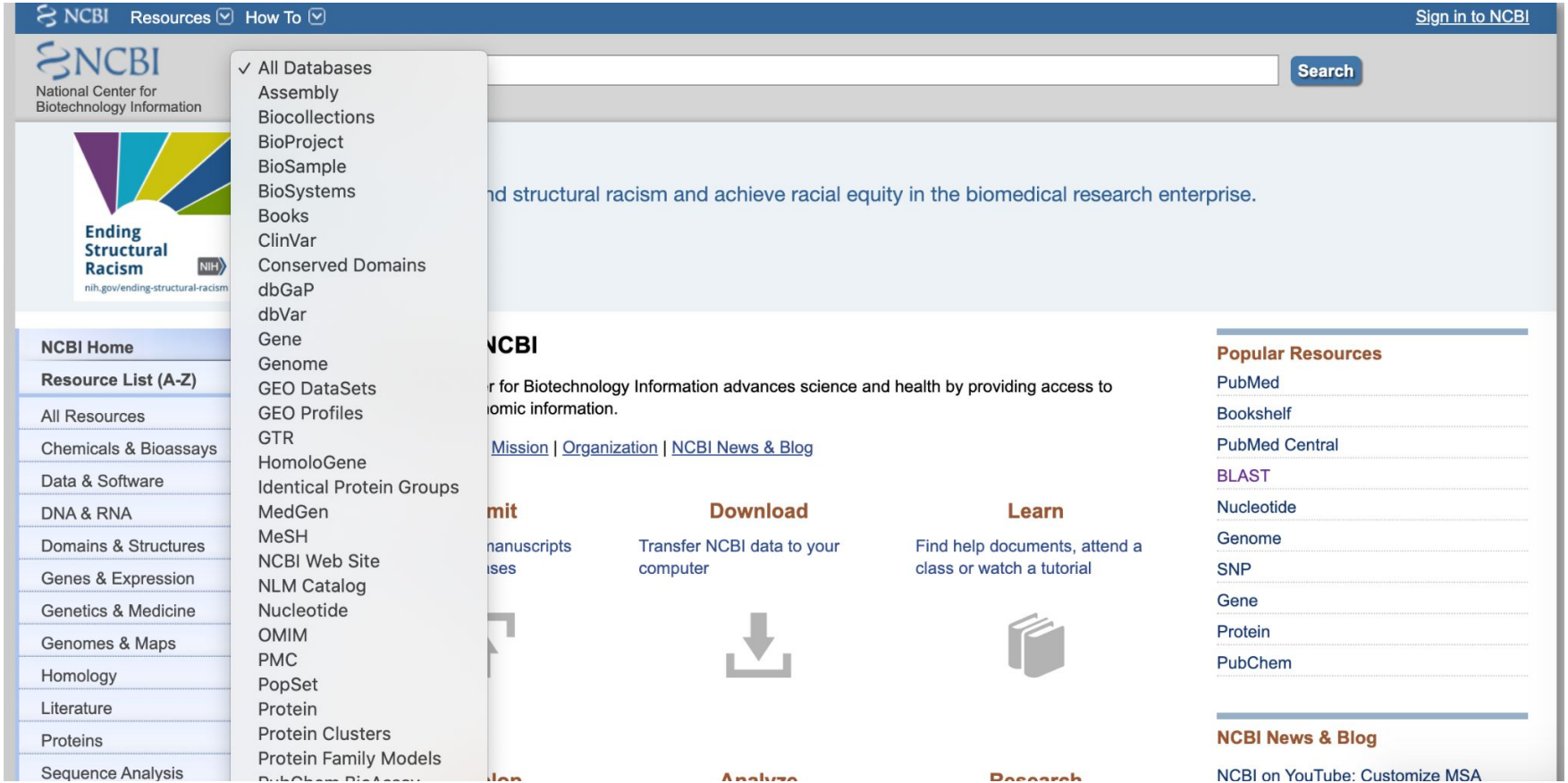

genome.ucsc.edu
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute
UCSC
Genome Browser
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Projects
Help
About Us
Genome Data
Source Code
Genome Browser Store
Utilities
FTP
MySQL Access
REST API
ata
wser and explore coronavirus datasets
nome
download data from the Genome Browser database
Variant Annotation Integrator
get functional effect predictions for variant calls
Data Integrator
combine data sources from the Genome Browser database
Genome Browser in a Box (GBiB)
run the Genome Browser on your laptop or server
In-Silico PCR
rapidly align PCR primer pairs to the genome
LiftOver
convert genome coordinates between assemblies
Track Hubs
import and view external data tracks
REST API
returns data in JSON format

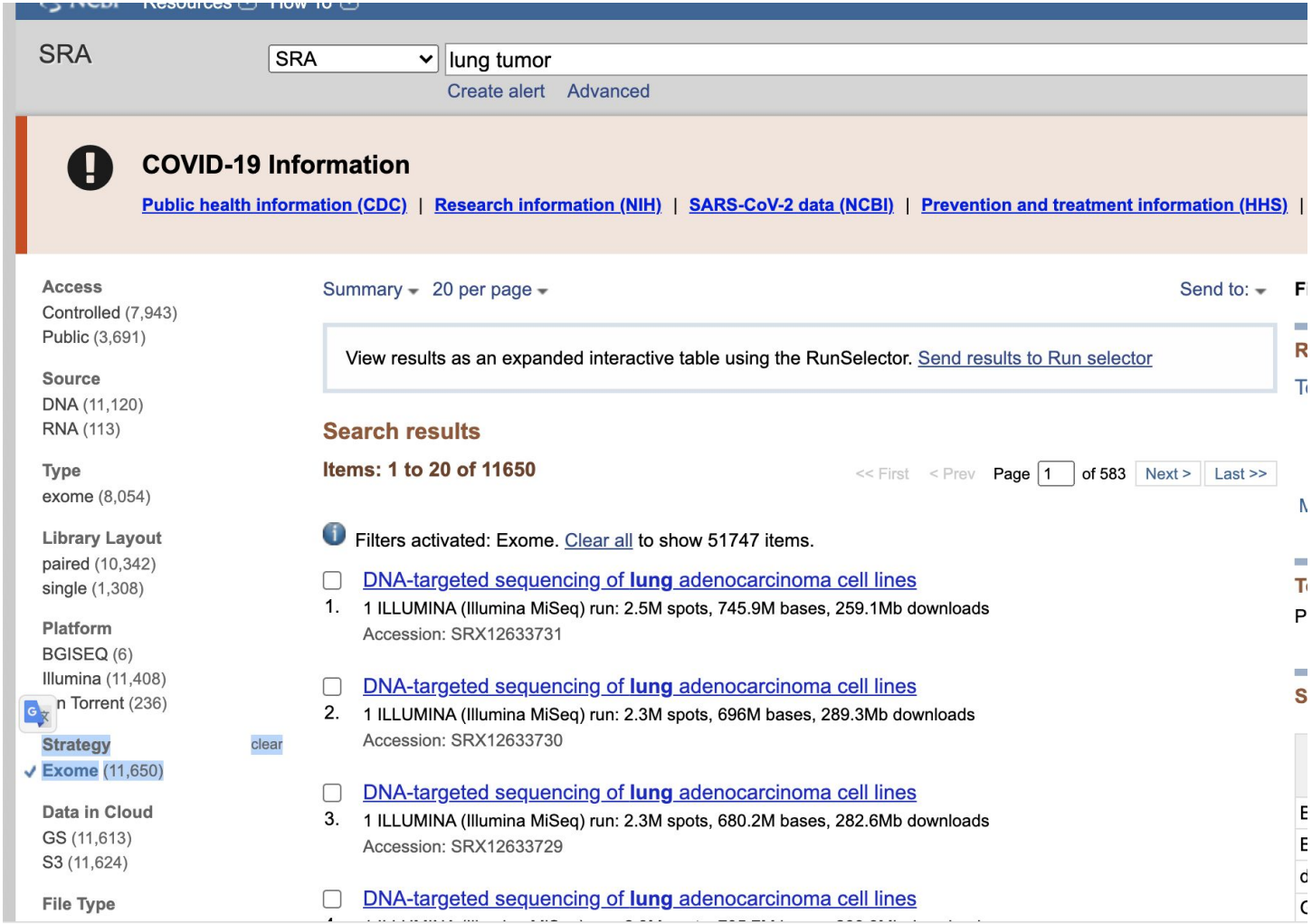

hgdownload.soe.ucsc.edu/downloads.html#human
Human genome
Dec. 2013 (GRCh38/hg38)
Genome sequence files and select annotations (2bit, GTF, GC-content, etc)
Sequence data by chromosome
Annotations
SNP-masked fasta files
LiftOver files
Pairwise alignments
Multiple alignments
Patches
Data archive
Feb. 2009 (GRCh37/hg19)
Genome sequence files and select annotations (2bit, GTF, GC-content, etc)
Sequence data by chromosome
Annotations
GC percent data
Protein database for hg19
SNP-masked fasta files
LiftOver files
Pairwise alignments (primates)

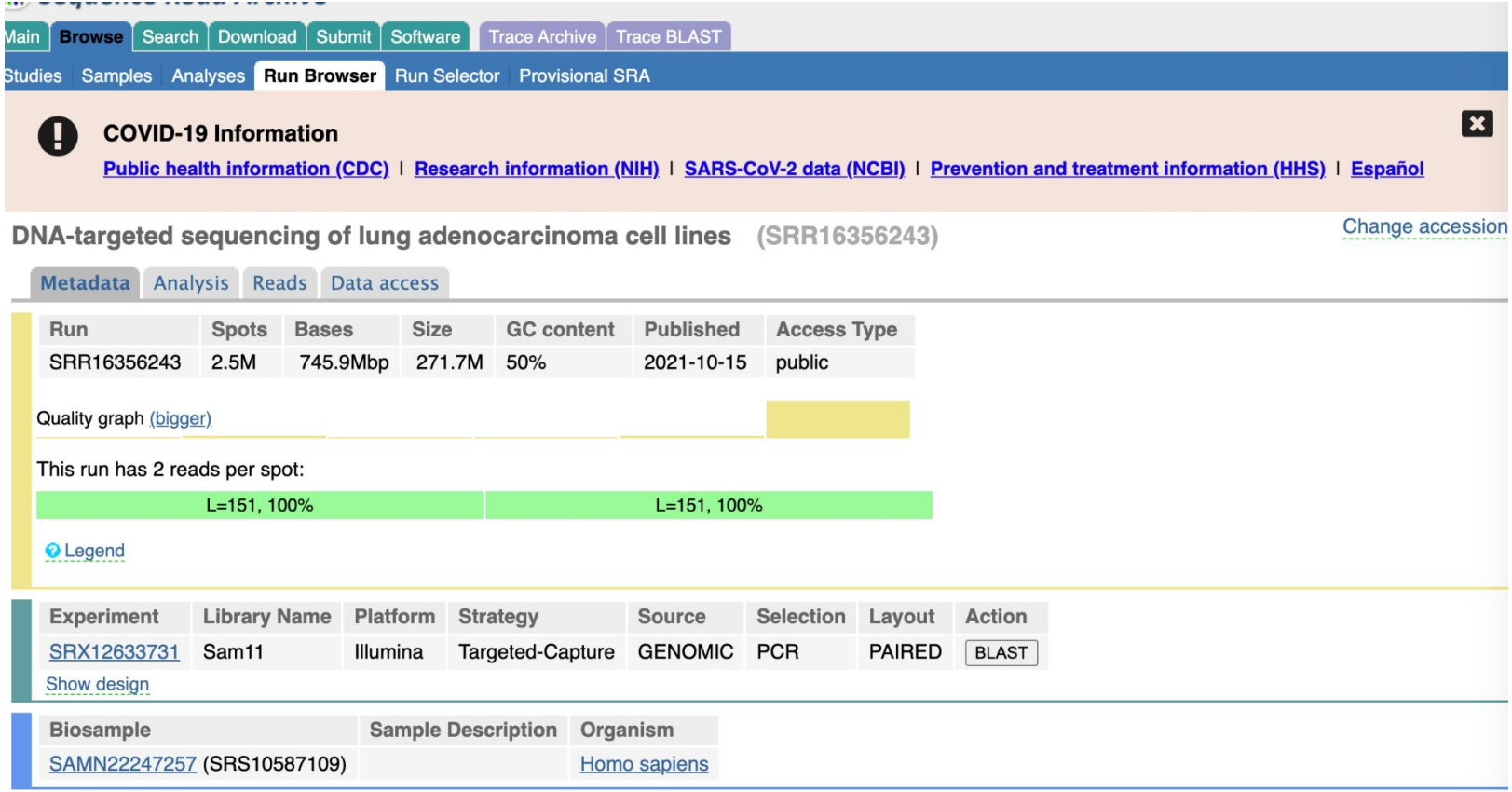

hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/
chr20.fa.gz 2014-01-23 16:39 19M
chr20_GL383577v2_alt.fa.gz 2014-01-23 16:40 42K
chr20_KI270869v1_alt.fa.gz 2014-01-23 16:40 37K
chr20_KI270870v1_alt.fa.gz 2014-01-23 16:40 55K
chr20_KI270871v1_alt.fa.gz 2014-01-23 16:40 18K
chr21.fa.gz 2014-01-23 16:39 12M
chr21_GL383578v2_alt.fa.gz 2014-01-23 16:40 21K
chr21_GL383579v2_alt.fa.gz 2014-01-23 16:40 65K
chr21_GL383580v2_alt.fa.gz 2014-01-23 16:40 25K
chr21_GL383581v2_alt.fa.gz 2014-01-23 16:40 36K
chr21_KI270872v1_alt.fa.gz 2014-01-23 16:40 26K
chr21_KI270873v1_alt.fa.gz 2014-01-23 16:40 47K
chr21_KI270874v1_alt.fa.gz 2014-01-23 16:40 53K
chr22.fa.gz 2014-01-23 16:39 12M
chr22_GL383582v2_alt.fa.gz 2014-01-23 16:40 52K
chr22_GL383583v2_alt.fa.gz 2014-01-23 16:40 31K
chr22_KB663609v1_alt.fa.gz 2014-01-23 16:40 21K
chr22_KI270731v1_random.fa.gz 2014-01-23 16:39 48K
chr22_KI270732v1_random.fa.gz 2014-01-23 16:39 14K
chr22_KI270733v1_random.fa.gz 2014-01-23 16:39 57K
chr22_KI270734v1_random.fa.gz 2014-01-23 16:39 53K
chr22_KI270735v1_random.fa.gz 2014-01-23 16:39 14K
chr22_KI270736v1_random.fa.gz 2014-01-23 16:39 42K
chr22_KI270737v1_random.fa.gz 2014-01-23 16:39 25K
chr22_KI270738v1_random.fa.gz 2014-01-23 16:39 24K
chr22_KI270739v1_random.fa.gz 2014-01-23 16:39 16K
chr22_KI270875v1_alt.fa.gz 2014-01-23 16:40 83K
chr22_KI270876v1_alt.fa.gz 2014-01-23 16:40 84K
chr22_KI270877v1_alt.fa.gz 2014-01-23 16:40 33K
chr22_KI270878v1_alt.fa.gz 2014-01-23 16:40 60K
chr22_KI270879v1_alt.fa.gz 2014-01-23 16:40 97K
chr22_KI270928v1_alt.fa.gz 2014-01-23 16:40 55K
chrM.fa.gz 2014-01-23 16:40 5.4K
chrUn_GL000195v1.fa.gz 2014-01-23 16:40 60K
chrUn_GL000213v1.fa.gz 2014-01-23 16:40 53K
chrUn_GL000214v1.fa.gz 2014-01-23 16:40 42K
chrUn_GL000216v2.fa.gz 2014-01-23 16:40 43K
chrUn_GL000218v1.fa.gz 2014-01-23 16:40 52K
chrUn_GL000219v1.fa.gz 2014-01-23 16:40 57K
chrUn_GL000220v1.fa.gz 2014-01-23 16:40 51K
chrUn_GL000224v1.fa.gz 2014-01-23 16:40 49K
chrUn_GL000226v1.fa.gz 2014-01-23 16:40 1.9K
chrUn_KI270302v1.fa.gz 2014-01-23 16:40 735

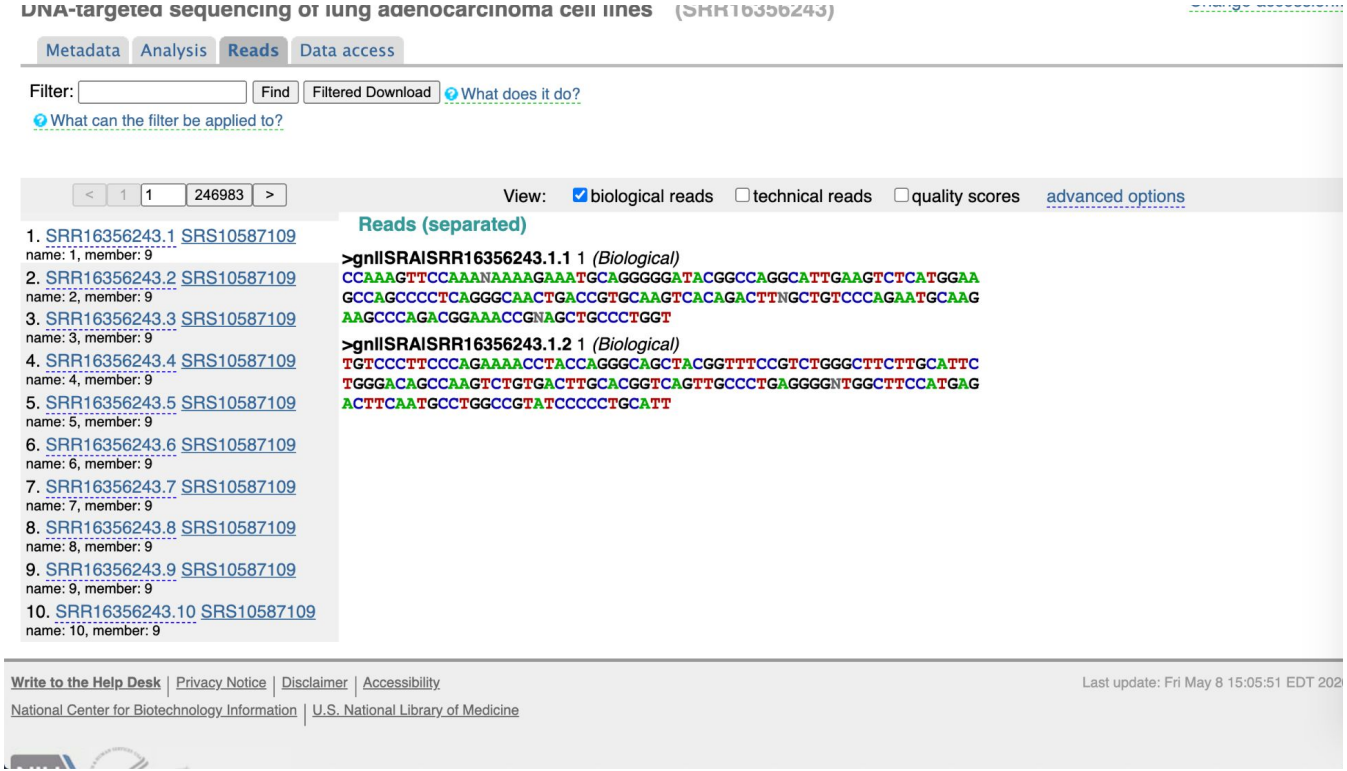
Index

SRA database

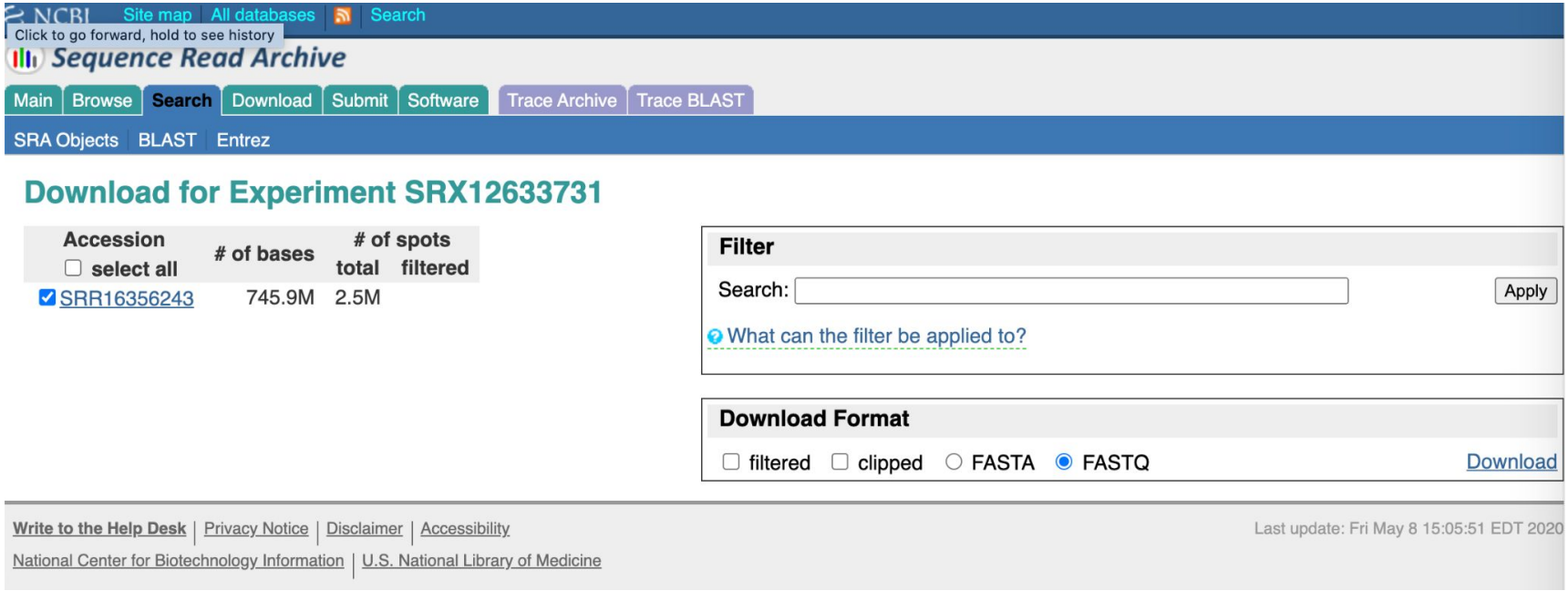



SRA
SRA
lung tumor
Create alert
Advanced
COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) |
Access
Controlled (7,943)
Public (3,691)
Source
DNA (11,120)
RNA (113)
Type
exome (8,054)
Library Layout
paired (10,342)
single (1,308)
Platform
BGISEQ (6)
Illumina (11,408)
n Torrent (236)
Strategy
clear
Exome (11,650)
Data in Cloud
GS (11,613)
S3 (11,624)
File Type
Summary
20 per page
Send to:
View results as an expanded interactive table using the RunSelector. Send results to Run selector
Search results
Items: 1 to 20 of 11650
<< First
< Prev
Page 1 of 583
Next >
Last >>
Filters activated: Exome. Clear all to show 51747 items.
DNA-targeted sequencing of lung adenocarcinoma cell lines
1. 1 ILLUMINA (Illumina MiSeq) run: 2.5M spots, 745.9M bases, 259.1Mb downloads
Accession: SRX12633731
DNA-targeted sequencing of lung adenocarcinoma cell lines
2. 1 ILLUMINA (Illumina MiSeq) run: 2.3M spots, 696M bases, 289.3Mb downloads
Accession: SRX12633730
DNA-targeted sequencing of lung adenocarcinoma cell lines
3. 1 ILLUMINA (Illumina MiSeq) run: 2.3M spots, 680.2M bases, 282.6Mb downloads
Accession: SRX12633729
DNA-targeted sequencing of lung adenocarcinoma cell lines

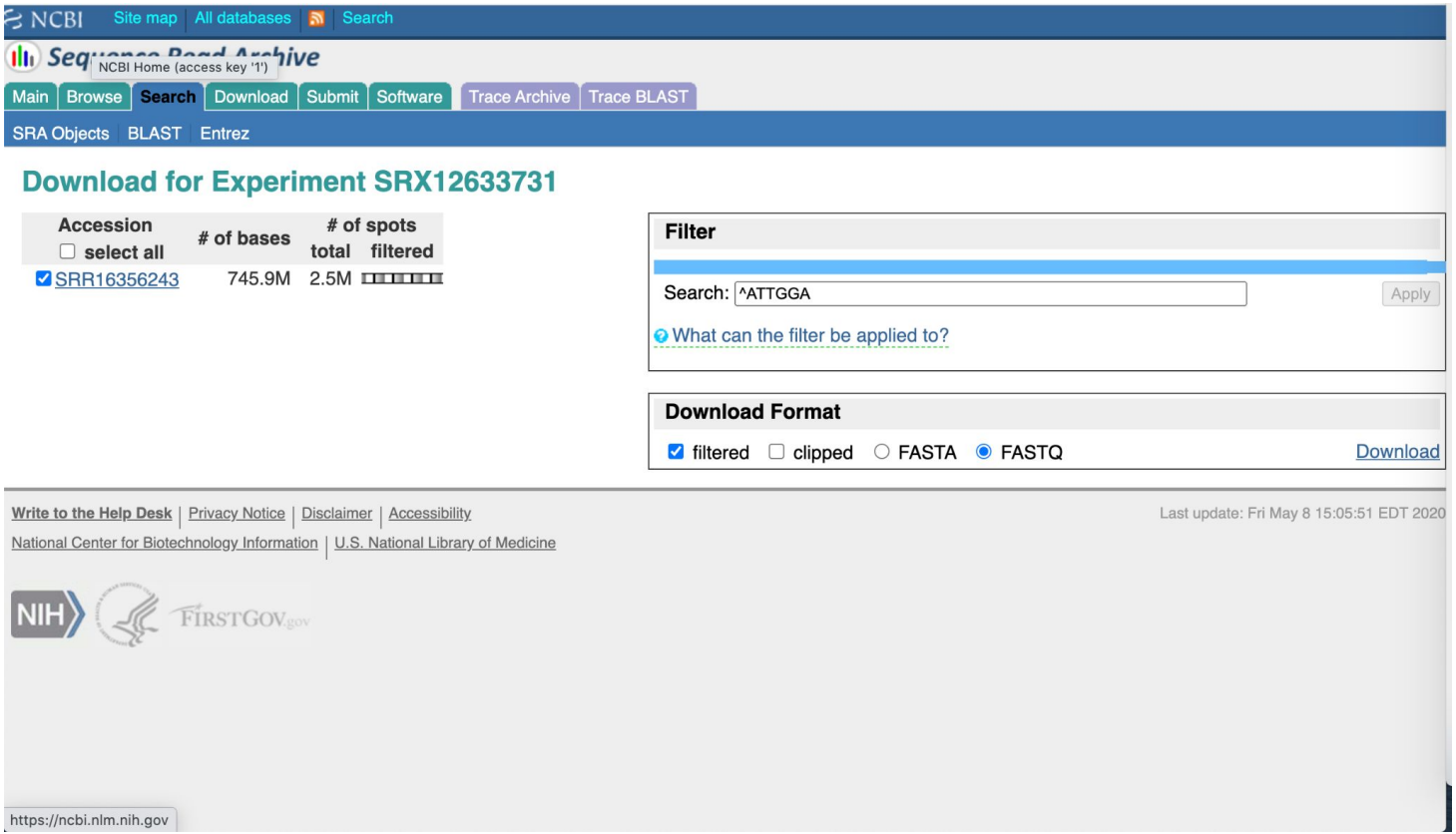

Main
Browse
Search
Download
Submit
Software
Trace Archive
Trace BLAST
Studies
Samples
Analyses
Run Browser
Run Selector
Provisional SRA
COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español
DNA-targeted sequencing of lung adenocarcinoma cell lines (SRR16356243)
Change accession
Metadata
Analysis
Reads
Data access
Run
Spots
Bases
Size
GC content
Published
Access Type
SRR16356243
2.5M
745.9Mbp
271.7M
50%
2021-10-15
public
Quality graph (bigger)
This run has 2 reads per spot:
L=151, 100%
L=151, 100%
Legend
Experiment
Library Name
Platform
Strategy
Source
Selection
Layout
Action
SRX12633731
Sam11
Illumina
Targeted-Capture
GENOMIC
PCR
PAIRED
BLAST
Show design
Biosample
Sample Description
Organism
SAMN22247257 (SRS10587109)
Homo sapiens



https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser&run=SRR16356243


NCBI Site map | All databases | Search
Click to go forward, hold to see history
Sequence Read Archive
Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace BLAST
SRA Objects | BLAST | Entrez
Download for Experiment SRX12633731
Accession
select all
# of bases
# of spots
total
filtered
SRR16356243
745.9M
2.5M
Filter
Search:
Apply
What can the filter be applied to?
Download Format
filtered
clipped
FASTA
FASTQ
Download
Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility
National Center for Biotechnology Information | U.S. National Library of Medicine
Last update: Fri May 8 15:05:51 EDT 2020

^ATTGGA


NCBI
Site map
All databases
Search
NCBI Home (access key '1')
Main
Browse
Search
Download
Submit
Software
Trace Archive
Trace BLAST
SRA Objects
BLAST
Entrez
Download for Experiment SRX12633731
Accession
select all
# of bases
# of spots
total
filtered
SRR16356243
745.9M
2.5M
Filter
Search: ^ATTGGA
Apply
What can the filter be applied to?
Download Format
filtered
clipped
FASTA
FASTQ
Download
Write to the Help Desk
Privacy Notice
Disclaimer
Accessibility
Last update: Fri May 8 15:05:51 EDT 2020
National Center for Biotechnology Information
U.S. National Library of Medicine
NIH
FIRSTGOV.gov
https://ncbi.nlm.nih.gov

sra_data.fastq.txt

```
@SRR16356243.194720.1 194720 length=151
GGCAGCAAAGTTTATTGTAAAATAAGAGATCGATATAAAAATGGGATATAAAAAGGGAGAAGGAGGGGAAGGGTGGGGTGAAAATGCAG
ATGTGCTTGCAGAATGTAAAAGATGTTGACCCTTCCAGCTGGACGTGGTGGCTCCAATTGT
+SRR16356243.194720.1 194720 length=151
AAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF/
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFF/
@SRR16356243.194720.2 194720 length=151
CAATTGGAGCCACCACGTCCAGCTGGAAGGGTCAACATCTTTTACATTCTGCAAGCACATCTGCATTTTCACCCCACCCTTCCCCTCCTT
CTCCCTTTTTATATCCCATTTTTATATCGATCTCTTATTTTACAATAAAACTTTGCTGGCA
+SRR16356243.194720.2 194720 length=151
FFFFFFFAFFFFFF/FFFFFAFFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF/
FFFFFFFFFAFFAFAAAFFFFFFFFFFF=AFF/FFFFFFFFFFF/FFFAFF=FFF/=F/FAFFFFFFFF6FF/FF
@SRR16356243.247542.1 247542 length=151
GTTTGACTTAGTATGAATGTGGTTACGTTGGAAGCAAATGTGTCTTCACTTTATCATGAAAAAGTCTGCAAGTGCTCTGCGACGTCCCAG
GGAAATGATCCTACCCTCAATCTTCAGCTCAAAGGGAACCTTTGCTCTTTTTCAGTGACCA
+SRR16356243.247542.1 247542 length=151
FAFF/FA=FAFF/FFFFF/FFAFA///F/A/FF/FFFFAAA///FF/FFFF///A////FFFFFA==//FF/F6FFA//A/FFF//
=F///F/AFA///F/=////F/=/FAA/AF//A/F/FF/AFFF//FFF/F//FFFFFF/F/////
@SRR16356243.247542.2 247542 length=151
ATTGGACACTGAAAAAGAGAAAAGGTTCCCTGTGAGCTGAAGAGTGAGTTTAGGATCATTTCCAAGGGACGTCGCAGAGCACTTGCAGAC
TTTATCATGAAATAGTTAAGACACATTTGATTCCAACGAAACCACATTCAAACTAAGTCAA
+SRR16356243.247542.2 247542 length=151
//F/F/FF/=/FAFFFAFF/F/FF///F//FF/FFFA//FA/F//A/F//FF//AFAF//F//AF6/////FAFFFFAA/F//FF///
FF//F/F/FF/A/F/FF//AA/A/FFA/FA//FFFFF/A/AA///F/A/FFFF/F/FF///=A
@SRR16356243.315284.1 315284 length=151
AAAAGATCTTATTGGCTTTGGTCTTCAAGTAGCCAAAGGCATGAAATATCTTGCAAGCAAAAAGTTTGTCCACAGAGACTTGGCTGCAAG
AAACTGTATGTAAGTATCAGAATCTCTGTGCCACAATCCAAATTAAGTGACAAGGAGGAAT
+SRR16356243.315284.1 315284 length=151
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFF6FFFFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFF
```

1000 Genome project

VCF files

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

Genetic map

https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html

GIAB

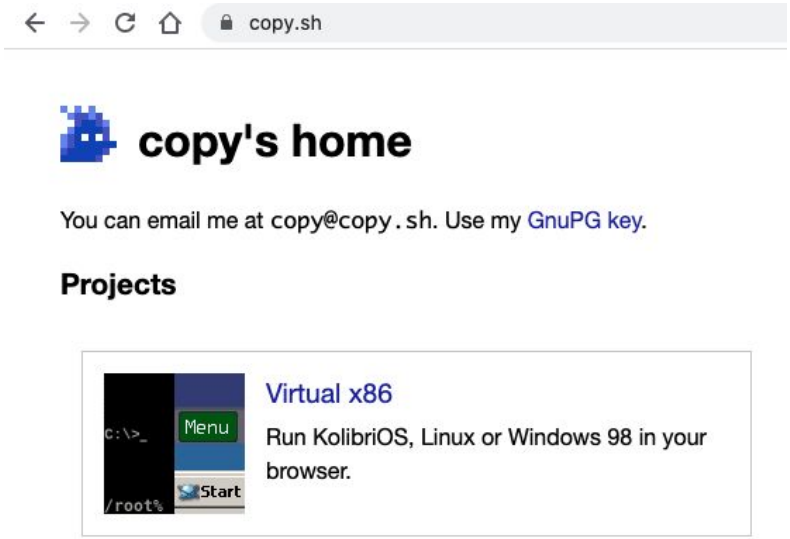
https://github.com/genome-in-a-bottle/giab_data_indexes

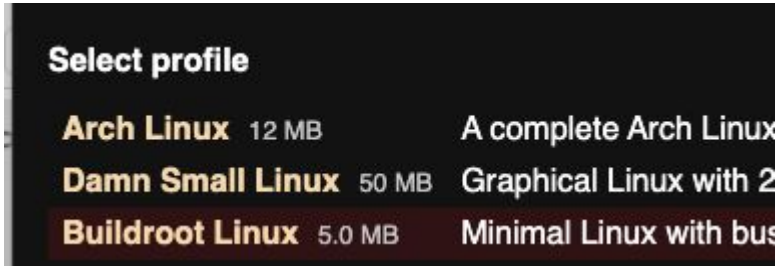
AshkenazimTrio

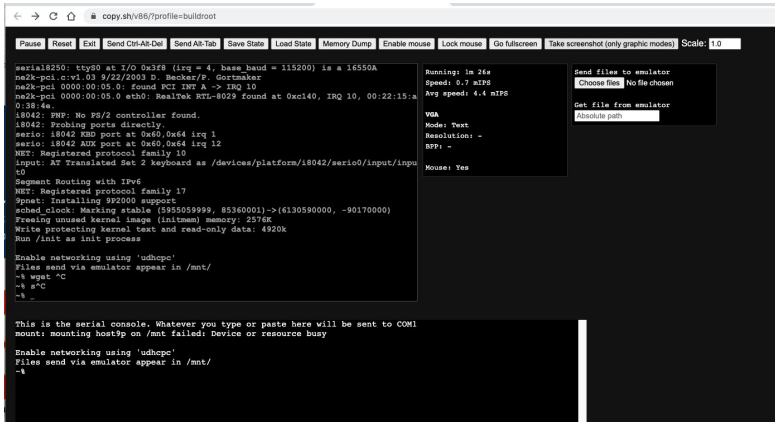
Son:HG002 https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/
Father:HG003 https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/
Mother:HG004 https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/

Sequencing Platform	Sequence	Alignment
Illumina WGS 2x150bp 300X per individual	All HG002 HG003 HG004	novoalign: All HG002 HG003 HG004
Illumina 6KB Matepair	All HG002 HG003 HG004	bwamem:hg19 All HG002 HG003 HG004
Illumina WGS 2X250bp	All HG002 HG003 HG004	isaac:hg19 All HG002 HG003 HG004 novoalign: All HG002 HG003 HG004
Moleculo	All HG002 HG003 HG004	
Illumina Whole Exome	-	bwamem:hg19 All HG002 HG003 HG004
SOLiD 60x for son	All HG002	LifeScope:hg19 All HG002
CompleteGenomics	-	CGAtools:hg19 All HG002 HG003 HG004

Another linux website





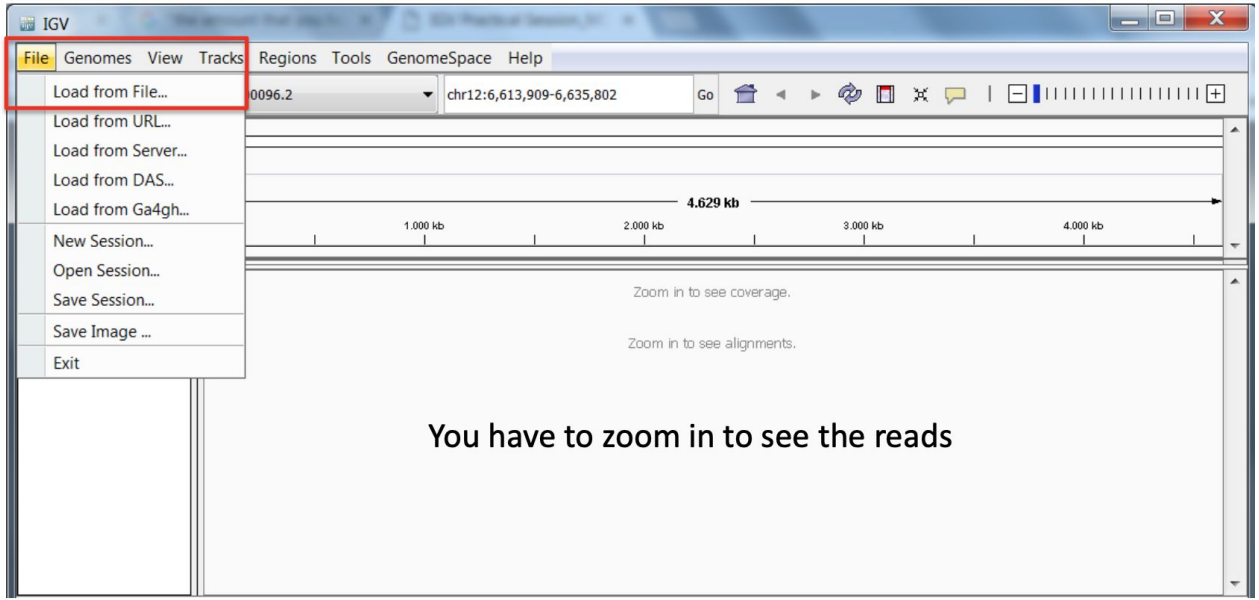


Disable mouse.

No internet

Index

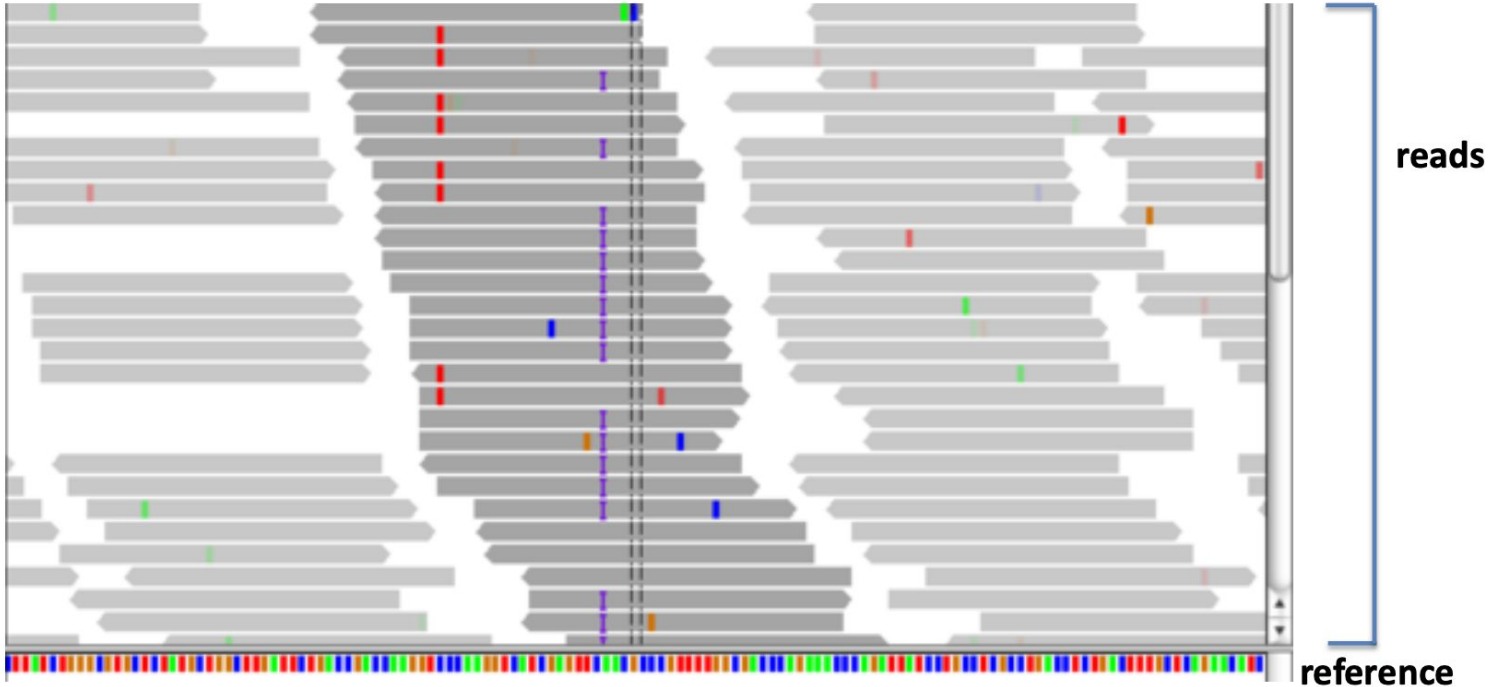
Visualization of mapping results: Loading data into IGV



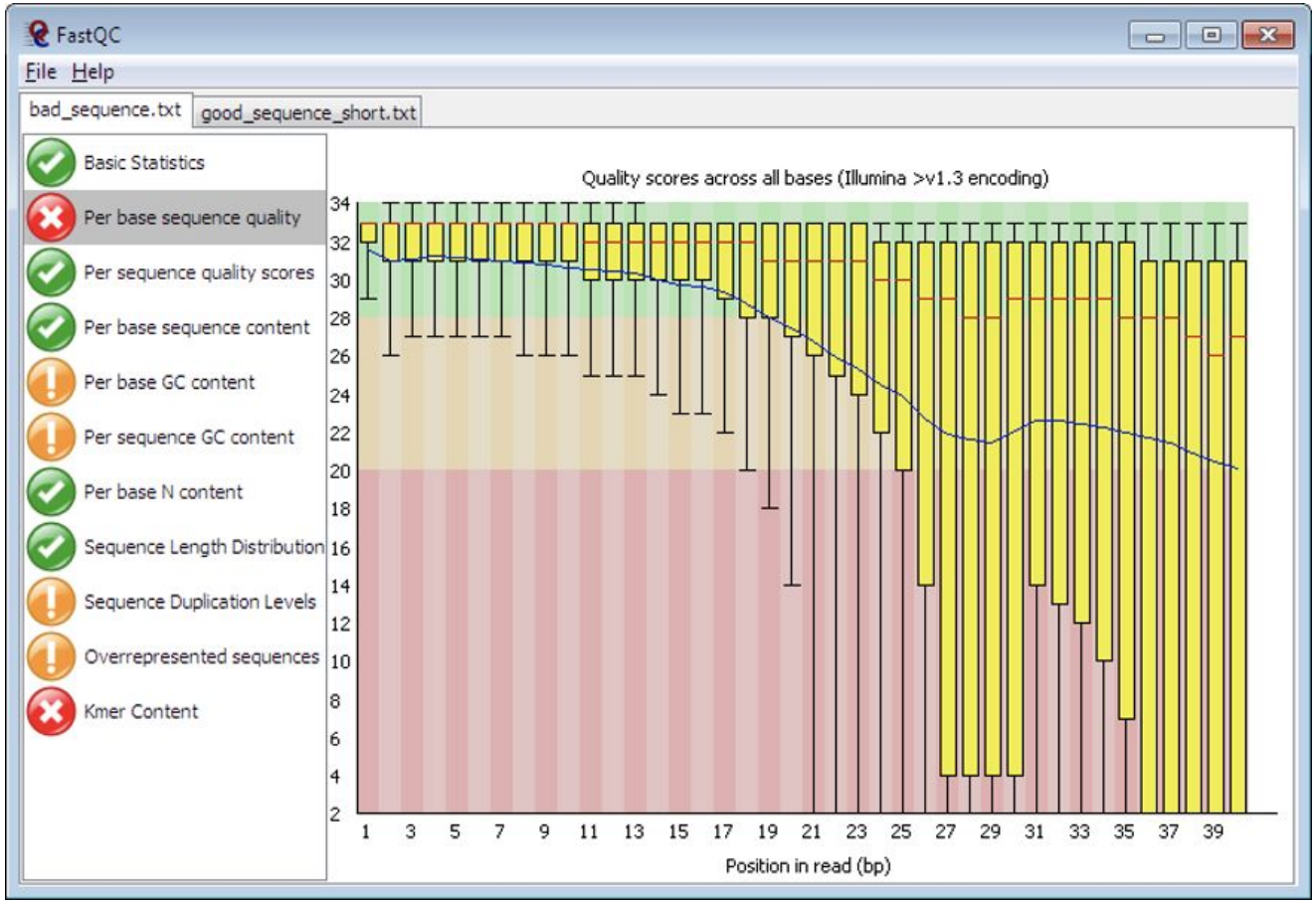
Grey: nuc

Colored: r

Arrows m



FASTQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

Trimming low quality bases from the ends of the reads

G	A	T	T	T	C	T	T	A	G	T	T	G	T	A	T	A	A	T
12	19	17	40	23	30	36	37	12	33	34	26	21	28	34	24	26	12	14

5‘ trimming bases with quality below 20

~~G~~	~~A~~	~~T~~	T	T	C	T	T	A	G	T	T	G	T	A	T	A	A	T
~~12~~	~~19~~	~~17~~	40	23	30	36	37	12	33	34	26	21	28	34	24	26	12	14

3‘ trimming bases with quality below 20

G	A	T	T	T	C	T	T	A	G	T	T	G	T	A	T	A	~~A~~	~~T~~
12	19	17	40	23	30	36	37	12	33	34	26	21	28	34	24	26	~~12~~	~~14~~

Koboldt *Genome Medicine* (2020) 12:91
https://doi.org/10.1186/s13073-020-00791-w

Genome Medicine

REVIEW Open Access

Best practices for variant calling in clinical sequencing

Daniel C. Koboldt[1,2]

Strategy	
Alignment and pre-processing	
Read alignment	BWA-MEM [25], Bowtie 2 [26], minimap2 [27], Novoalign
Marking duplicates	Picard tools [28], Sambamba [29], SAMBLASTER [30]
BAM file creation	Samtools [31], GATK [19]
Sequencing metrics	BEDTools [32], Picard tools [28], QualiMap 2 [33]
Sample quality control	KING [34], VerifyBamID [35]
Variant calling	
Inherited SNVs/indels	FreeBayes [36], GATK HaplotypeCaller [19], Platypus [20], Samtools/BCFtools [37]
Somatic mutations	deepSNV [38], MuSE [39], MuTect2 [40], SomaticSniper [41], Strelka2 [42], VarDict [43], VarScan2 [44]
Copy number variants	cn.MOPS [45], CONTRA [46], CoNVEX [47], ExomeCNV [48], ExomeDepth [49], XHMM [50]
Structural variants	DELLY [51], Lumpy [52], Manta [53], Pindel [54], SVMerge [55]
Gene fusions (RNA-seq)	fusionCatcher [56], fusionMap [57], mapSplice [58], SOAPfuse [59], STAR-Fusion [60], TopHat-Fusion [61]
Variant review/storage	
Visualization and review	Artemis [62], Integrative Genomics Viewer [63]
VCF/BCF file manipulation	BCFtools [37]

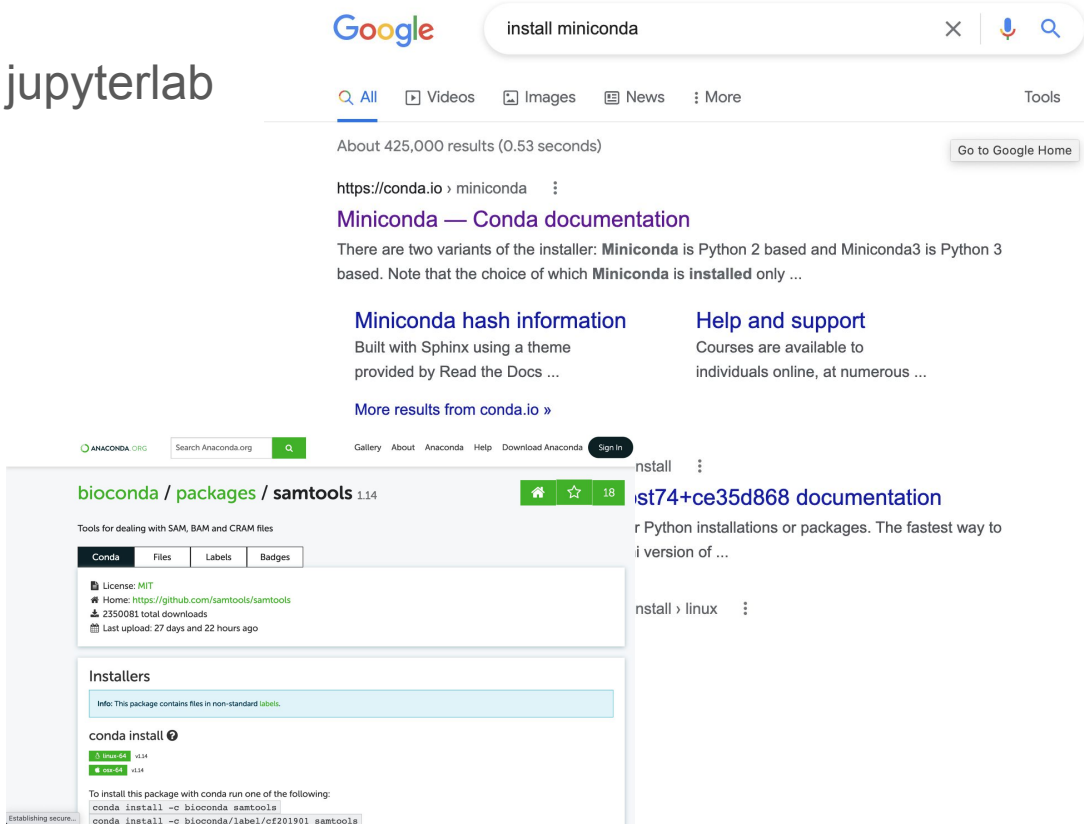
conda

conda install -c conda-forge jupyterlab

conda install samtools

conda install fastqc

conda install jupyter

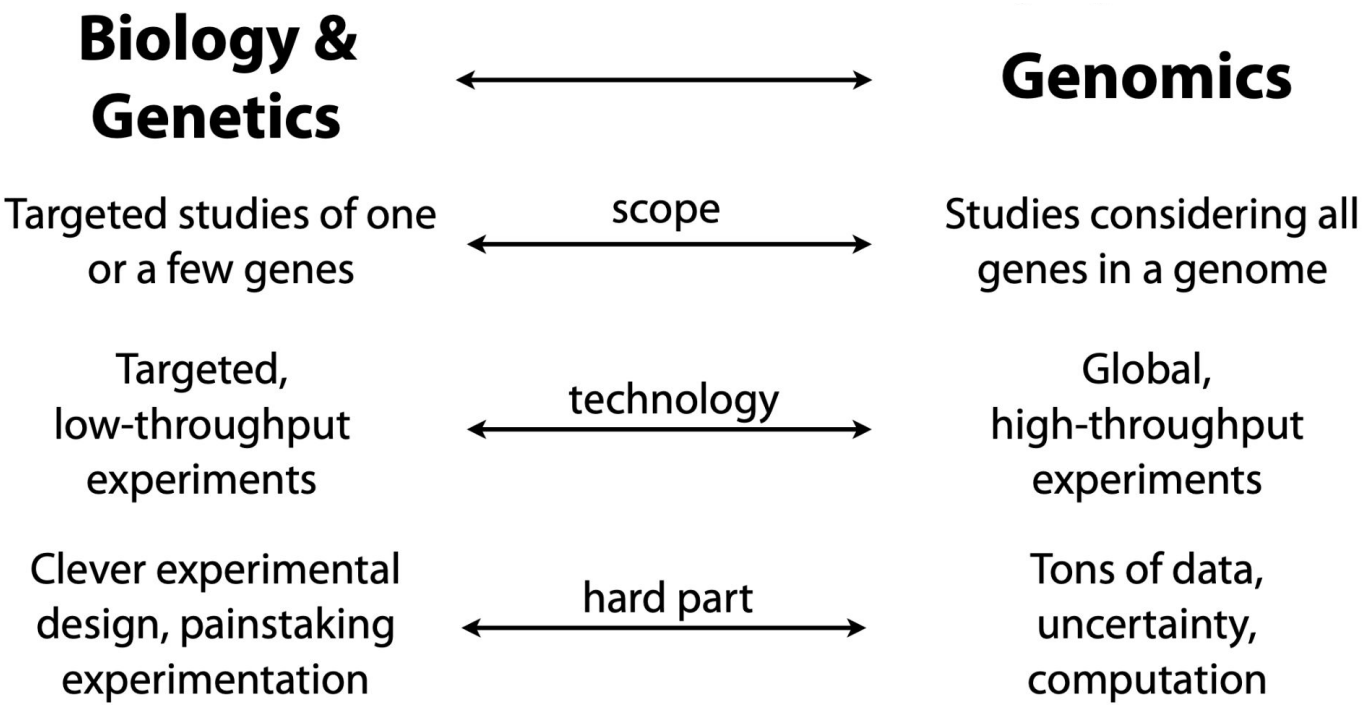


Resources

- Coursera
- https://maktabkhooneh.org/ (دکتر شریفی زارچی)

- Galaxy: Work on public server: https://usegalaxy.org/

- Slides on github
- Algorithms: https://langmead-lab.org/teaching-materials/


Biology & Genetics
Genomics
Targeted studies of one or a few genes
scope
Studies considering all genes in a genome
Targeted, low-throughput experiments
technology
Global, high-throughput experiments
Clever experimental design, painstaking experimentation
hard part
Tons of data, uncertainty, computation

Index

Thanks for your attention!

Sina.Majidian
@gmail.com

twitter

ResearchGate

LinkedIn