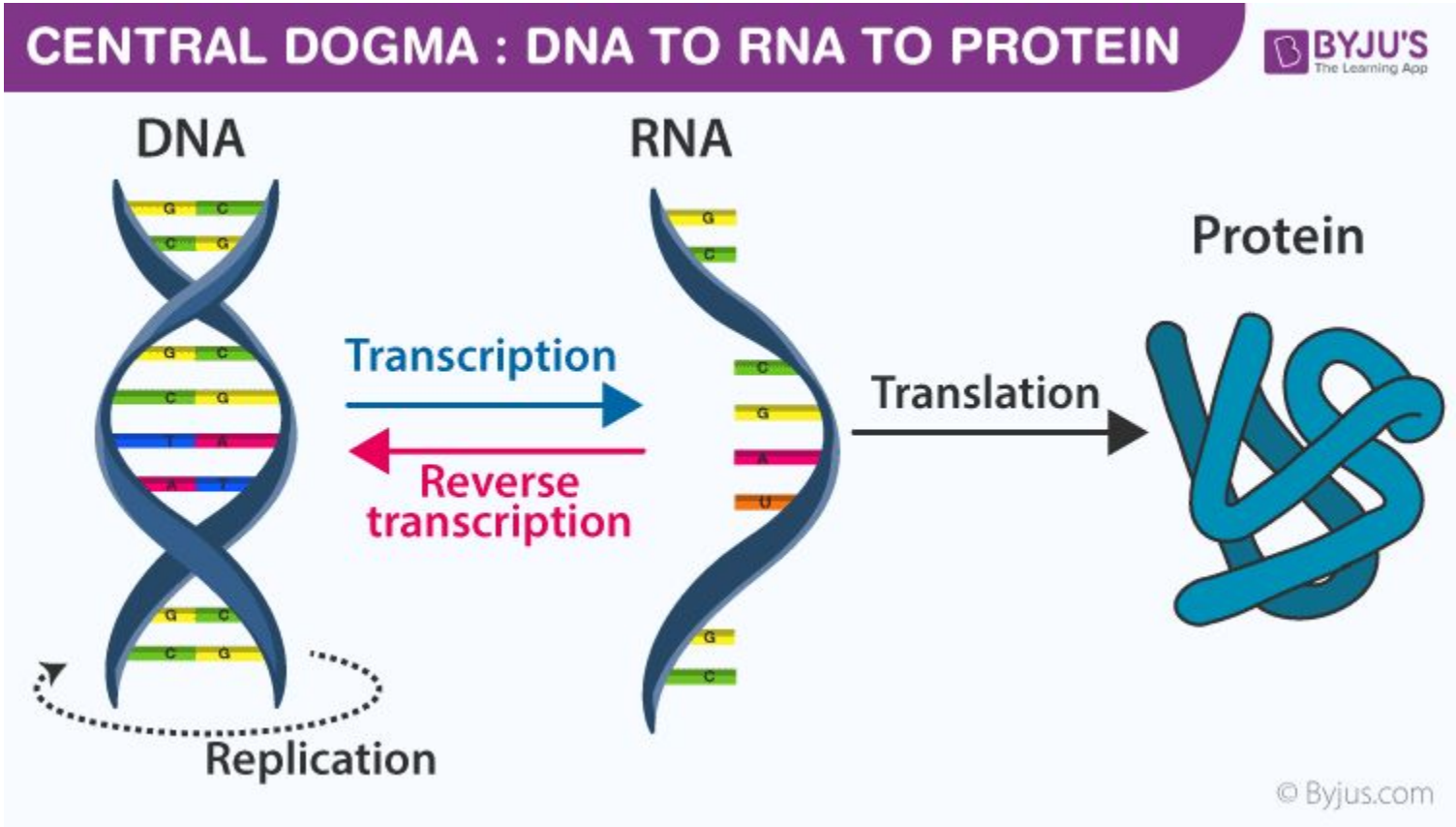
Introduction to DNA & Sequencing

Nicholas Smith and Christian Mertes

Outline

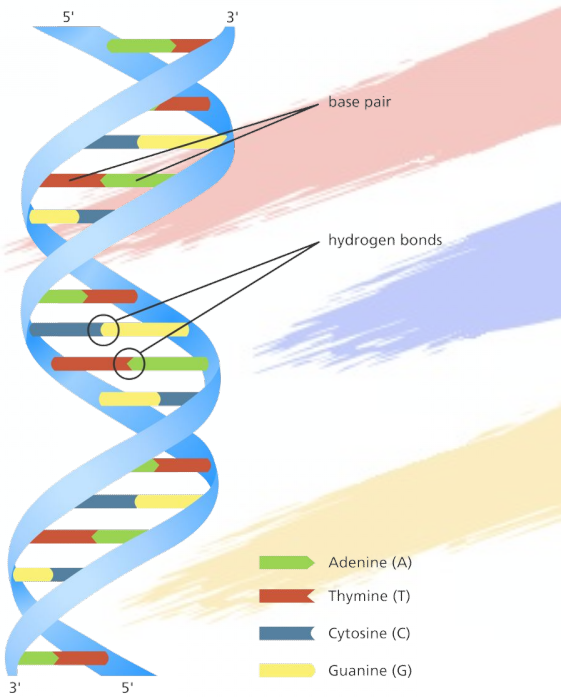
- Introduction to bioinformatics
  - Basic biology
  - DNA sequencing
  - RNA sequencing
- Q&A
- Workflows in GHGA
  - Model summary
  - Data storage and analysis
  - FAIR practices
- Summary
- Q&A #2

Biology


CENTRAL DOGMA : DNA TO RNA TO PROTEIN
BYJU'S
The Learning App
DNA
RNA
Protein
Transcription
Reverse transcription
Translation
Replication
© Byjus.com

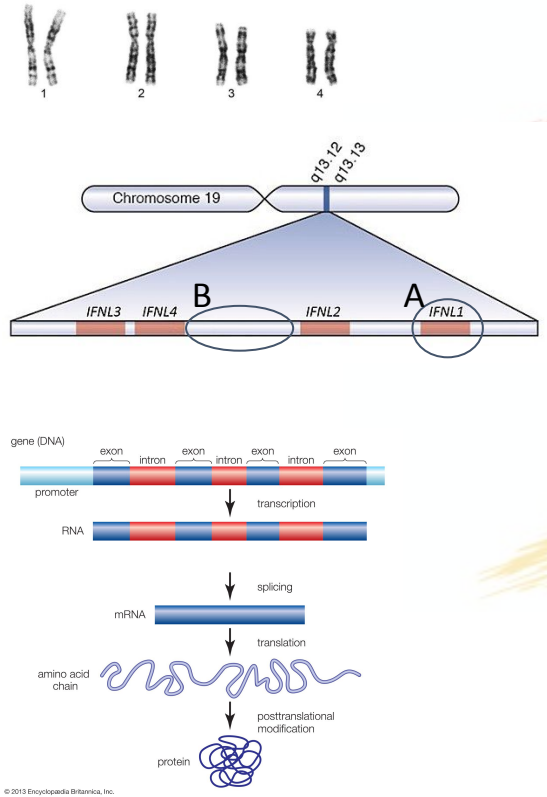
DNA Basics

- DNA has a double helix shape
- The building blocks are 4 nucleobases
  - A, T, C, G
    - A - T pairs
    - C - G pairs
- Each person has 2 chromosomes representing the genetic information inherited from their parents
  - So each person has 2 copies of each gene!
- Human genome is ~3B bases long
  - organized into 23 pairs of chromosomes
- DNA is further organized into 2 umbrella categories
  - genes (1-2 %)
  - non-genes (98-99 %)



Genome organization

- 23 pairs of chromosomes
  - strands of DNA are condensed
- Chromosome broken into genes
  - ~ 20,000 genes in humans
  - 2 copies of each gene and allele
- Genes are organized into exons (coding) and introns (non-coding)
  - Genes have a special start and a stop



Genomic Sequencing

Terms:

- Reads
  - The subsection of DNA

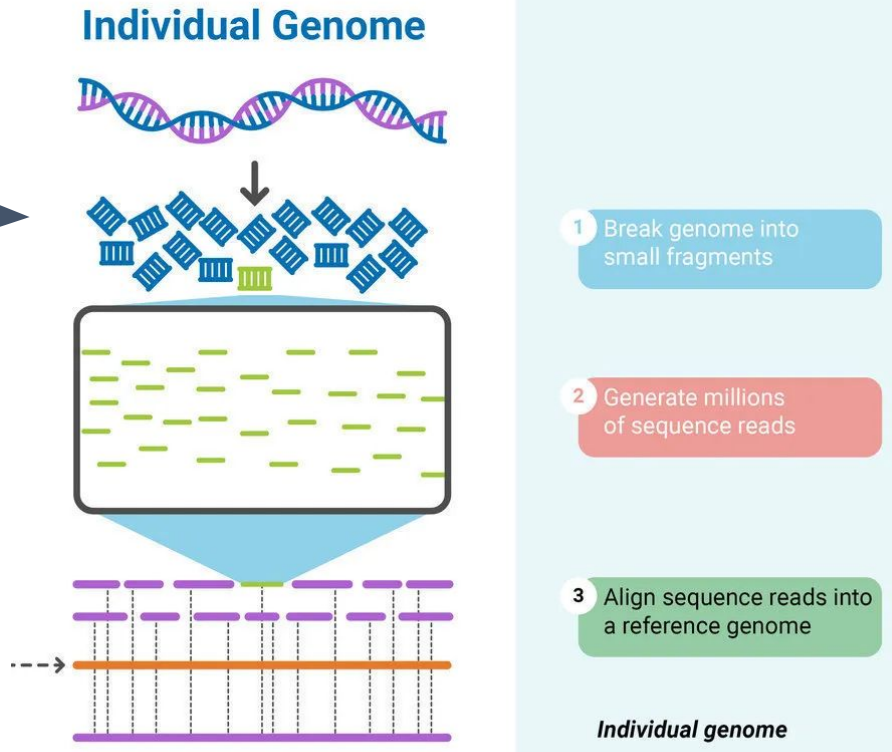


figure courtesy of [sequencing.com]

Genomic Sequencing

Terms:

- Reads
  - The subsection of DNA
- Sequencing depth
  - Amount of overlap for a position

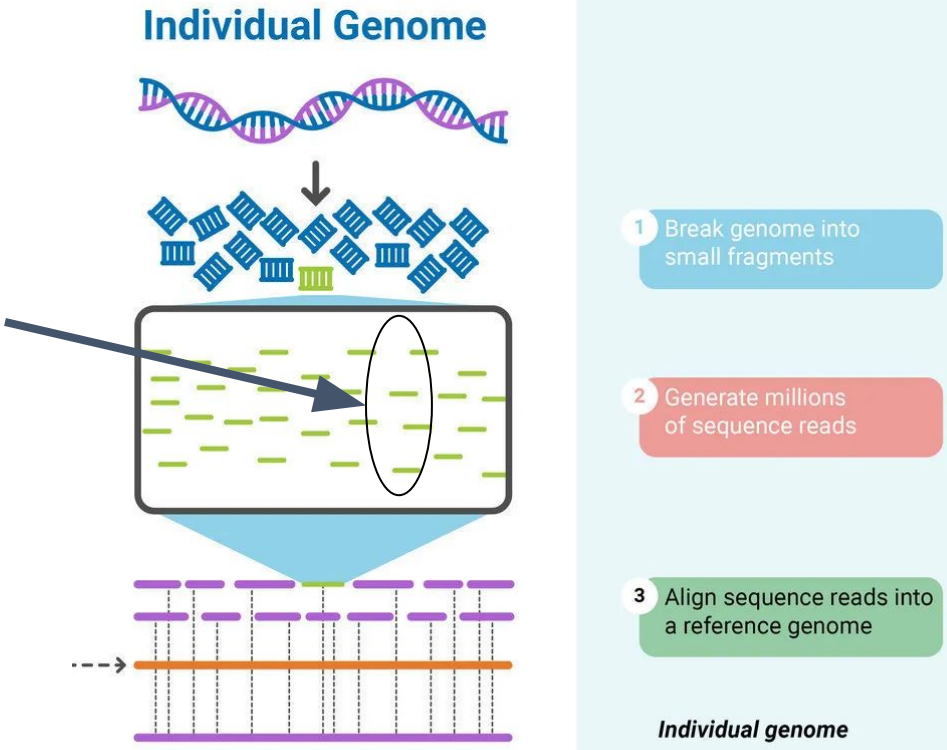


figure courtesy of [sequencing.com]

Genomic Sequencing

Terms:

- Reads
  - The subsection of DNA
- Sequencing depth
  - Amount of overlap for a position
- Reference
  - The gold standard shared human genome

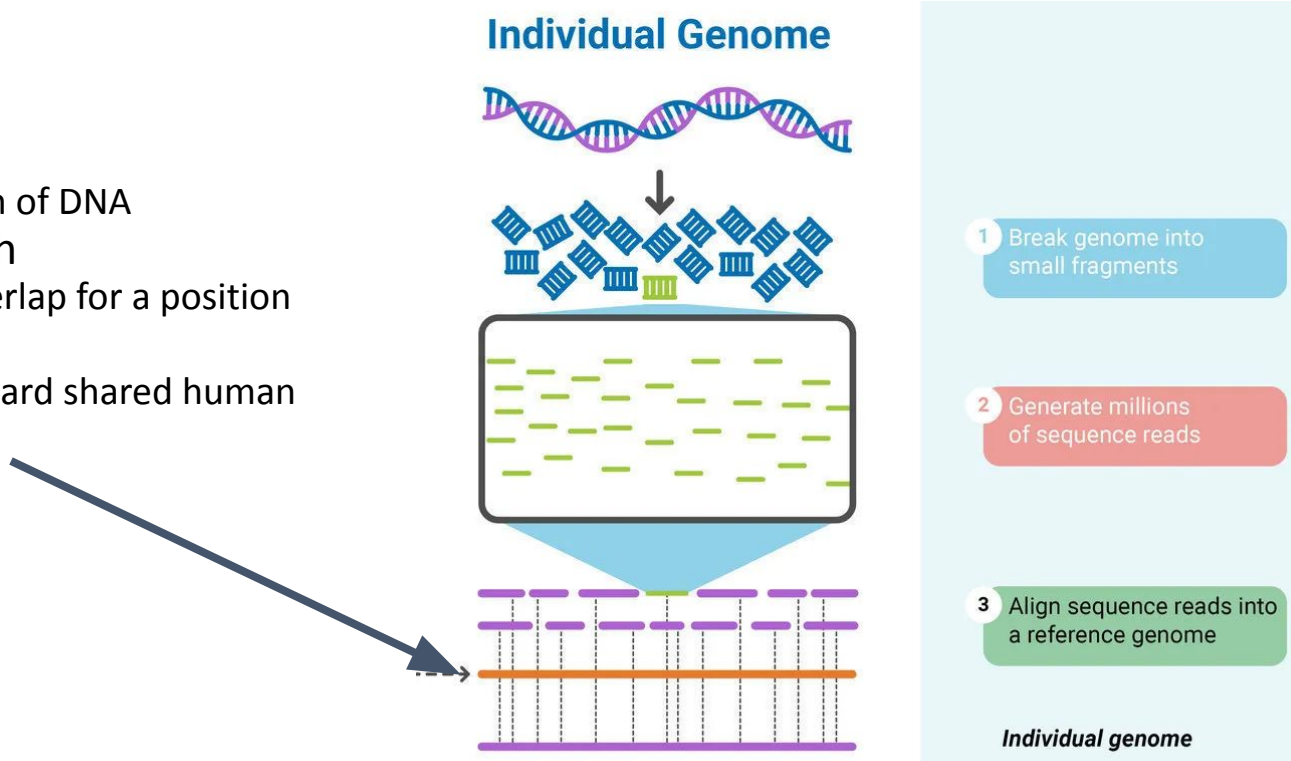


figure courtesy of [sequencing.com]

Genomic Sequencing

Terms:

- Reads
  - The subsection of DNA
- Sequencing depth
  - Amount of overlap for a position
- Reference
  - The gold standard shared human genome
- Alignment
  - Where each read lines up with the reference

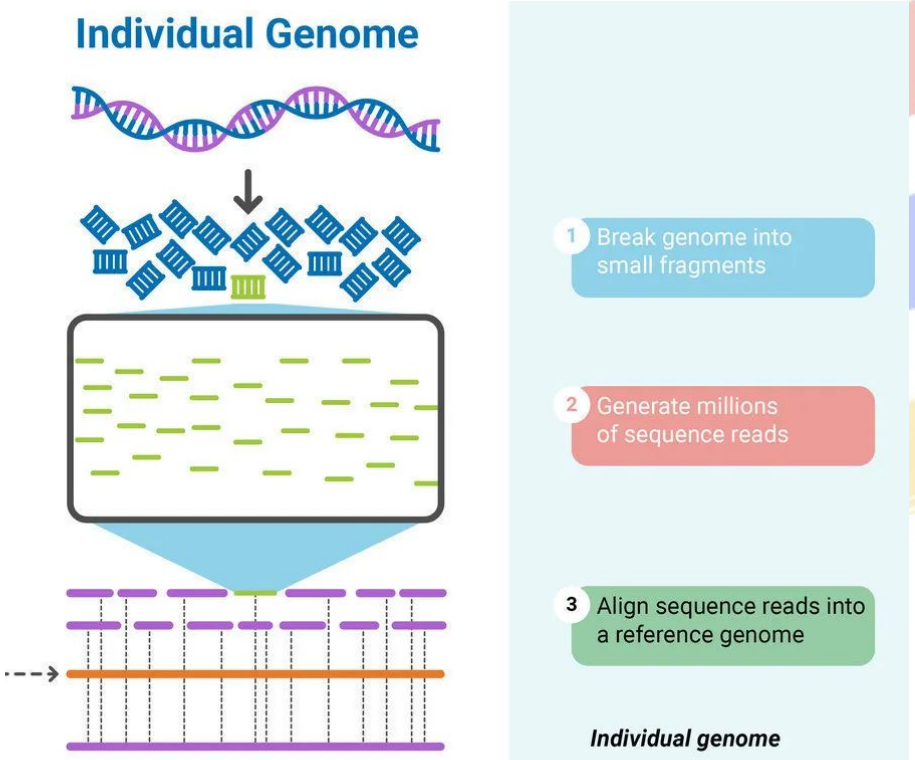
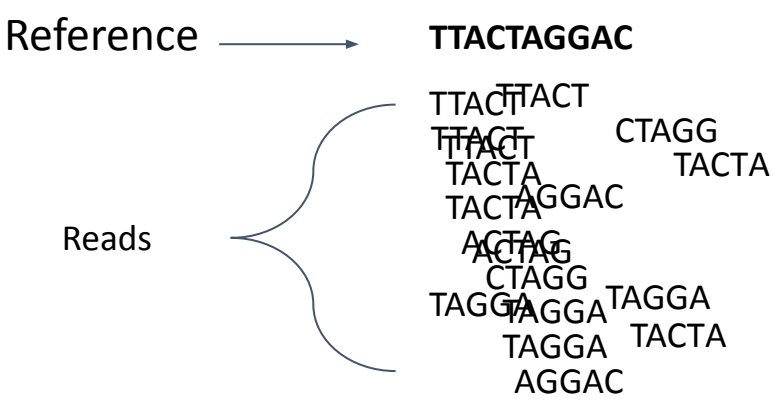
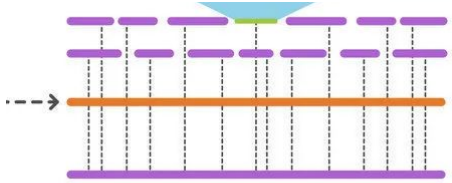


figure courtesy of [sequencing.com]

Alignment





3 Align sequence reads into a reference genome

Individual genome

figure courtesy of sequencing.com

index	1	2	3	4	5	6	7	8	9	10
allele	T	T	A	C	T	A	G	G	A	C
supporting reads	2	4	5	6	8	7	5	4	3	1

Genomic Sequencing

Terms:

- Reads
  - The subsection of DNA
- Sequencing depth
  - Amount of overlap for a position
- Reference
  - The gold standard shared human genome
- Alignment
  - Where each read lines up with the reference
- Variants
  - How the DNA nucleotides differ than the reference

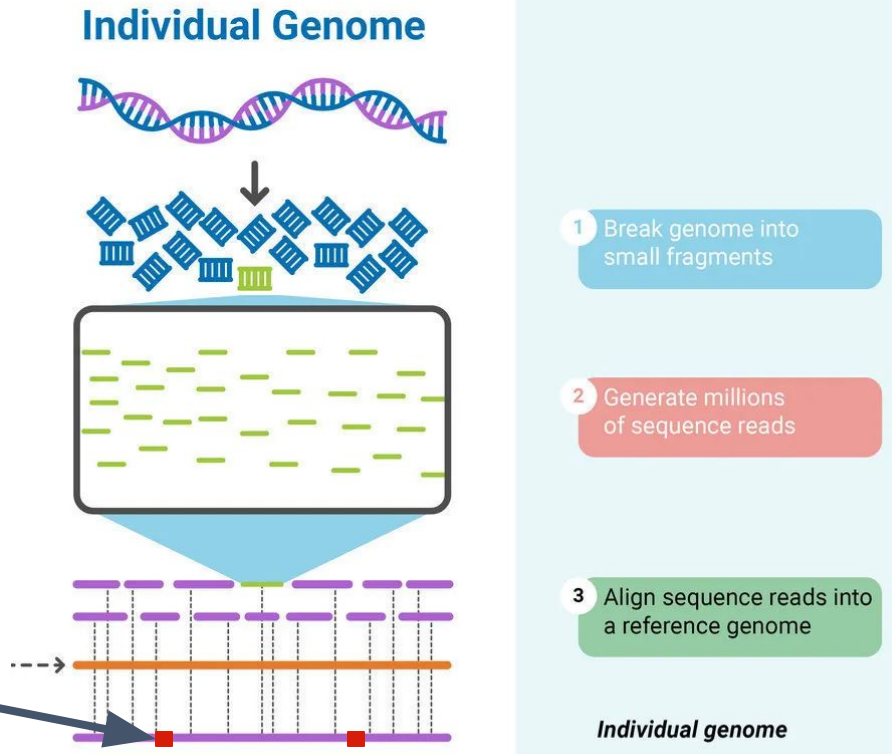
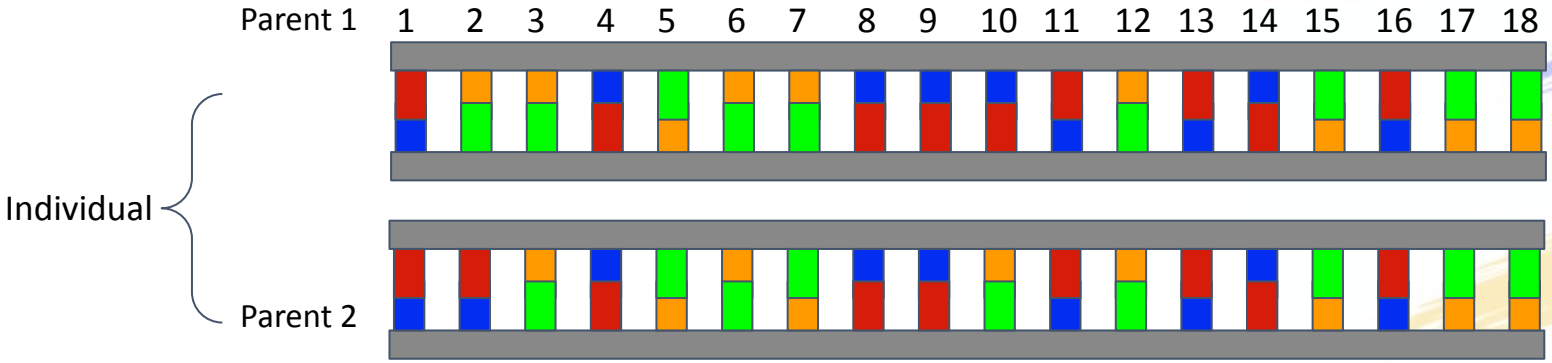


figure courtesy of [sequencing.com]

Simplification!

- How bioinformaticians talk
  - view the world as a computer scientist, often simplifying complex representations and biology



Simplification!

Parent 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5’	T	G	G	A	C	G	T	A	A	A	T	G	T	A	C	A	C	C
3’	A	C	C	T	G	C	A	T	T	T	A	C	A	T	G	T	G	G

Individual

Parent 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5’	T	T	G	A	C	G	C	A	A	C	T	G	T	A	C	A	C	C
3’	A	A	C	T	G	C	G	T	T	G	A	C	A	T	G	T	G	G

Simplification!

- Because we know the nucleotide pairings (A-T and G-C) we don't lose any information by excluding the 3' sequence from our simplification

Parent 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5'	T	G	G	A	C	G	T	A	A	A	T	G	T	A	C	A	C	C

Individual

Parent 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5'	T	T	G	A	C	G	C	A	A	C	T	G	T	A	C	A	C	C

Simplification!

- Most of our DNA is shared, only small parts of it differ from person to person
  - ~99.9% similar from person to person

Individual

Parent 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5’	T	G	G	A	C	G	T	A	A	A	T	G	T	A	C	A	C	C

Parent 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5’	T	T	G	A	C	G	C	A	A	C	T	G	T	A	C	A	C	C

Reference

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5’	T	T	C	A	C	G	C	A	A	A	T	G	T	A	C	A	C	C

Simplification!

- Compare our individual to the Reference (a well studied high quality sample)

Individual

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Parent 1	5’	T	G	G	A	C	G	T	A	A	A	T	G	T	A	C	A	C	C
Parent 2	5’	T	T	G	A	C	G	C	A	A	C	T	G	T	A	C	A	C	C

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Reference	5’	T	T	C	A	C	G	C	A	A	A	T	G	T	A	C	A	C	C

Simplification!

- Since “everyone” uses the same reference we can just list out how or individual changes are different compared to the reference

Position	Reference Allele	Alternative Allele	Genotype
2	T	G	0/1
3	C	G	1/1
7	C	C	0/1
10	A	T	0/1

Variant Call Format

VCF format - example

```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  GEN
chr1    82154   rs4477212   a   .   .   .   GT  0/0
chr1    752566  rs3094315   g   A   .   .   GT  1/1
chr1    752721  rs3131972   A   G   .   .   GT  1/1
chr1    798959  rs11240777  g   .   .   .   GT  0/0
chr1    800007  rs6681049   T   C   .   .   GT  1/1
chr1    838555  rs4970383   c   .   .   .   GT  0/0
chr1    846808  rs4475691   C   .   .   .   GT  0/0
chr1    854250  rs7537756   A   .   .   .   GT  0/0
chr1    861808  rs13302982  A   G   .   .   GT  1/1
chr1    873558  rs1110052   G   T   .   .   GT  1/1
chr1    882033  rs2272756   G   A   .   .   GT  0/1
chr1    888659  rs3748597   T   C   .   .   GT  1/1
chr1    891945  rs13303106  A   G   .   .   GT  0/1
```

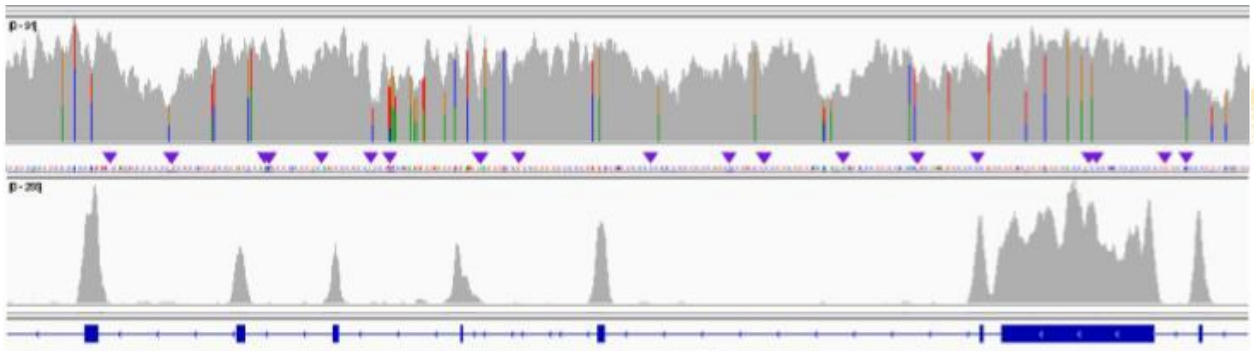
Genomics Data

- Whole Genome Sequencing (WGS)
  - Covers the “entire” genome
  - Becoming the primary choice of sequencing technology
- Whole Exome Sequencing (WES)
  - Covers the coding sequences of the genome
  - Older and cheaper way of sequencing the coding regions of the genome

WGS

WES

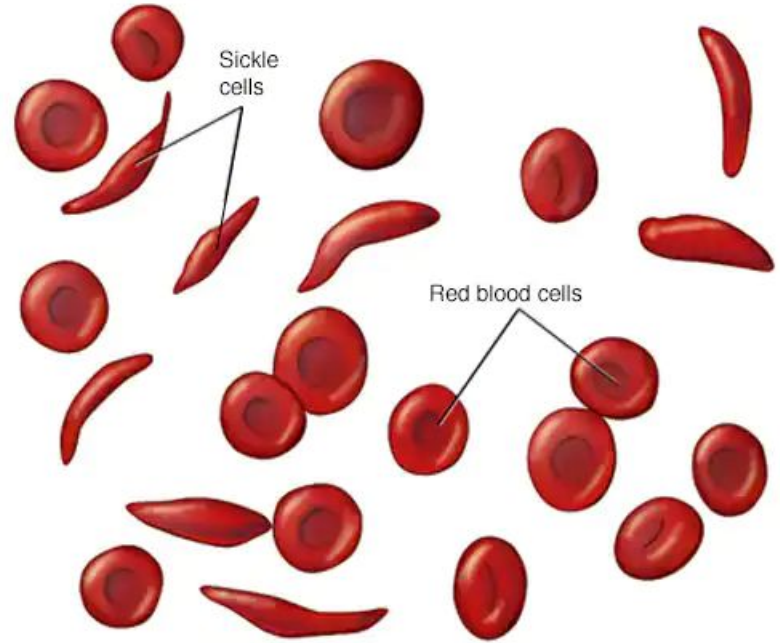
Reference



Case study: Sickle Cell Anemia

Sickle cell anemia [illegible] des of acute illness and p[illegible]

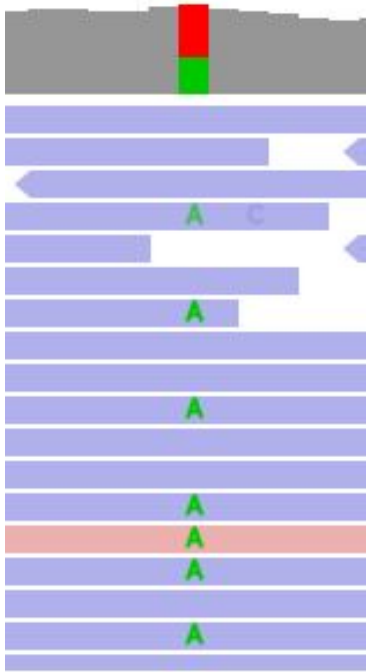
- Most commor[illegible]
- T -> A at posit[illegible]
- causes the pr[illegible]
  - Glutamic A[illegible]

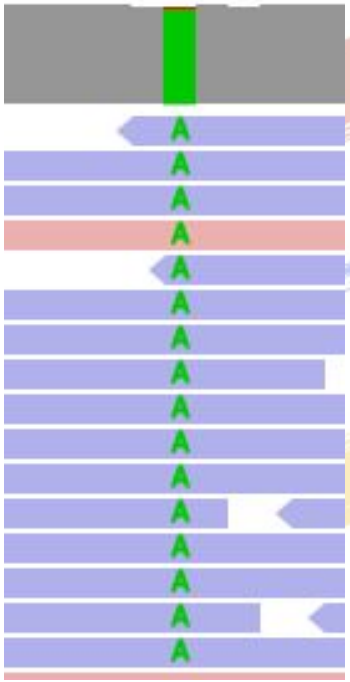


https://www.mayoclinic.org/diseases-conditions/sickle-cell-anemia/symptoms-causes/syc-20355876

Sickle Cell Anemia

- Left
  - Heterozygous
  - T -> A
- Right
  - Homozygous
  - T -> A
- Knowing the root cause improves patient understanding and treatment options





https://gnomad.broadinstitute.org/variant/11-5248232-T-A?dataset=gnomad_r2_1

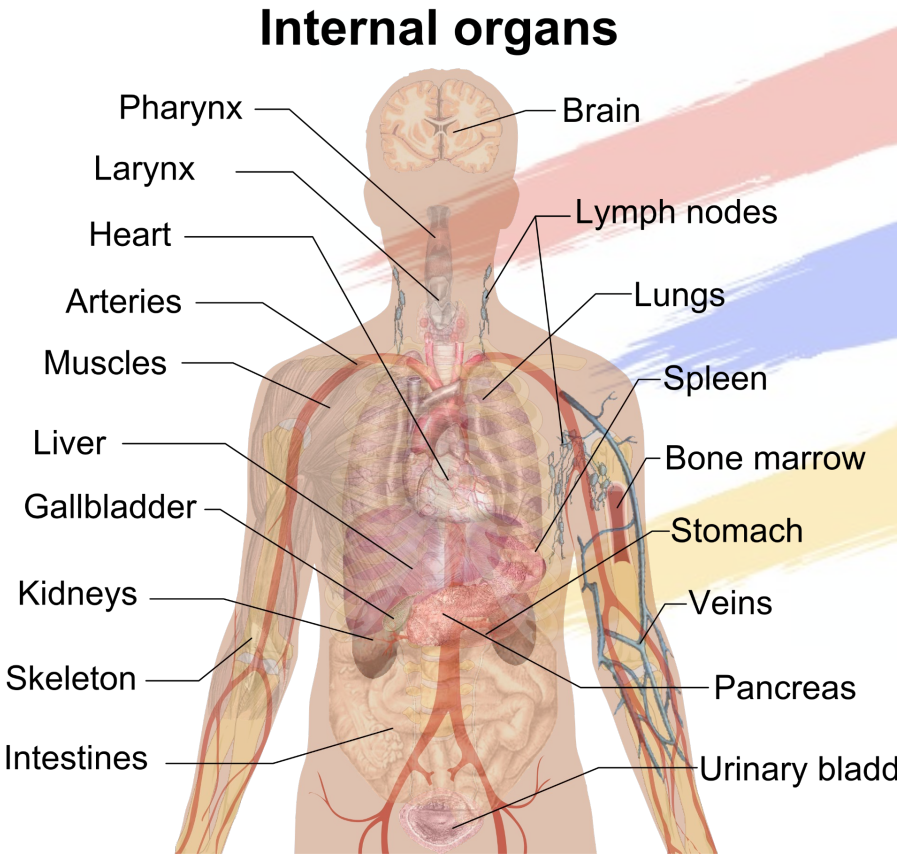
Genomics Application

- Identifying known pathogenic variants
- Rare Disease
  - 300M worldwide patients (Nguengang Wakap, S., Lambert, D.M., Olry, A. et al.)
  - Provide the underlying cause of a genetic disease
- Gene wide association studies (GWAS)
  - General variants across large populations
  - Common disease and genetic contribution
  - Frequency patterns within cohorts of similar phenotypes
- Cancer
  - Comparisons between cancer and healthy tissue
  - Comparative studies
  - hereditary influence on cancer risk (BRCA)

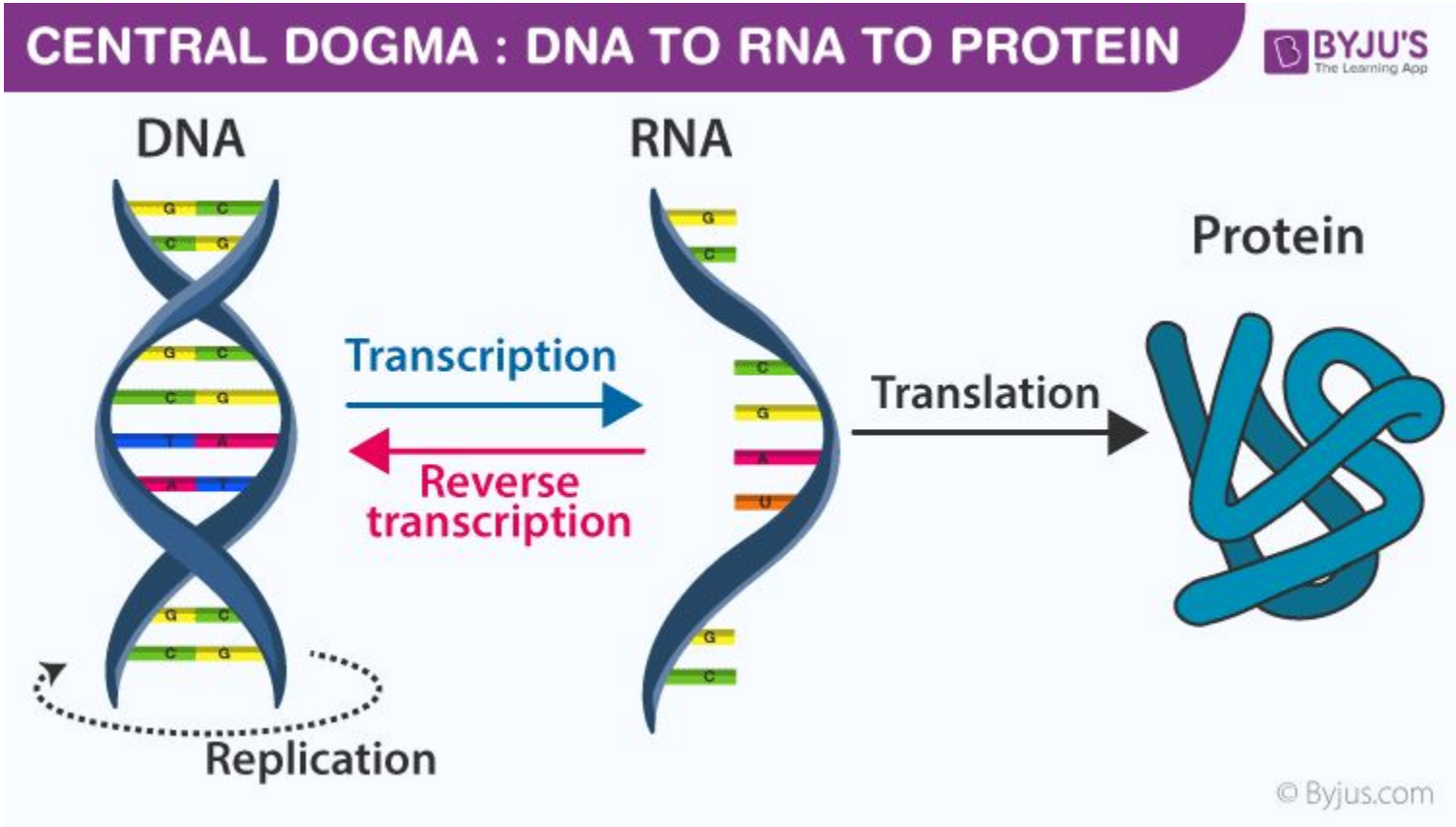
RNA Sequening

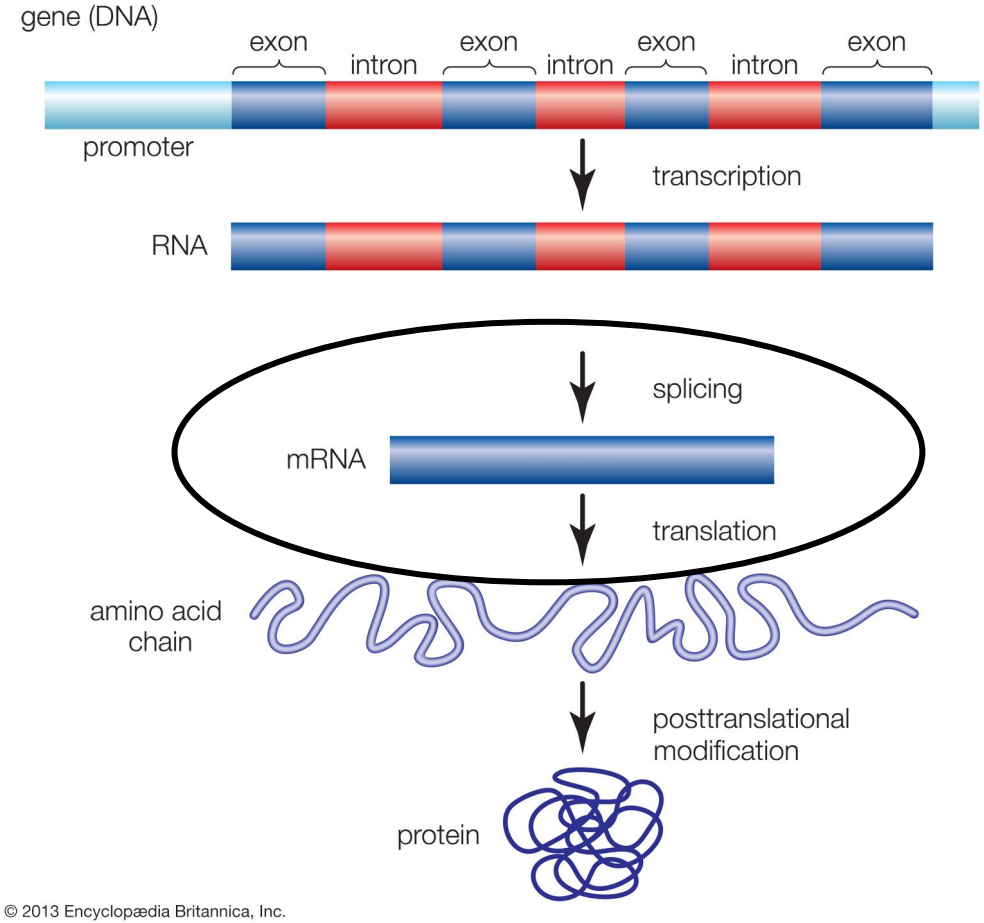
RNA Sequencing

- Different parts of the body do different things
  - but every cell has the same DNA



RNA Sequencing





Courtesy of Encyclopædia Britannica

RNA sequencing steps

Biology

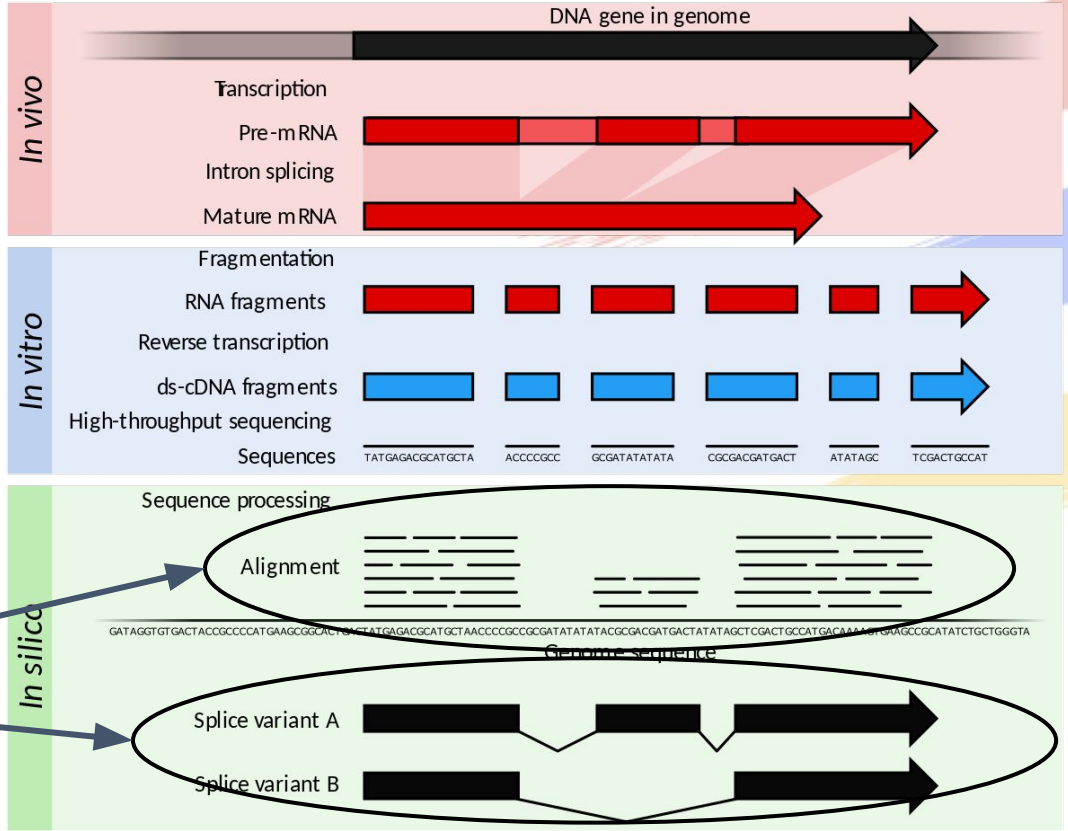
- RNA processing

Sequencing

- similar to DNA seq
- technically less consistent

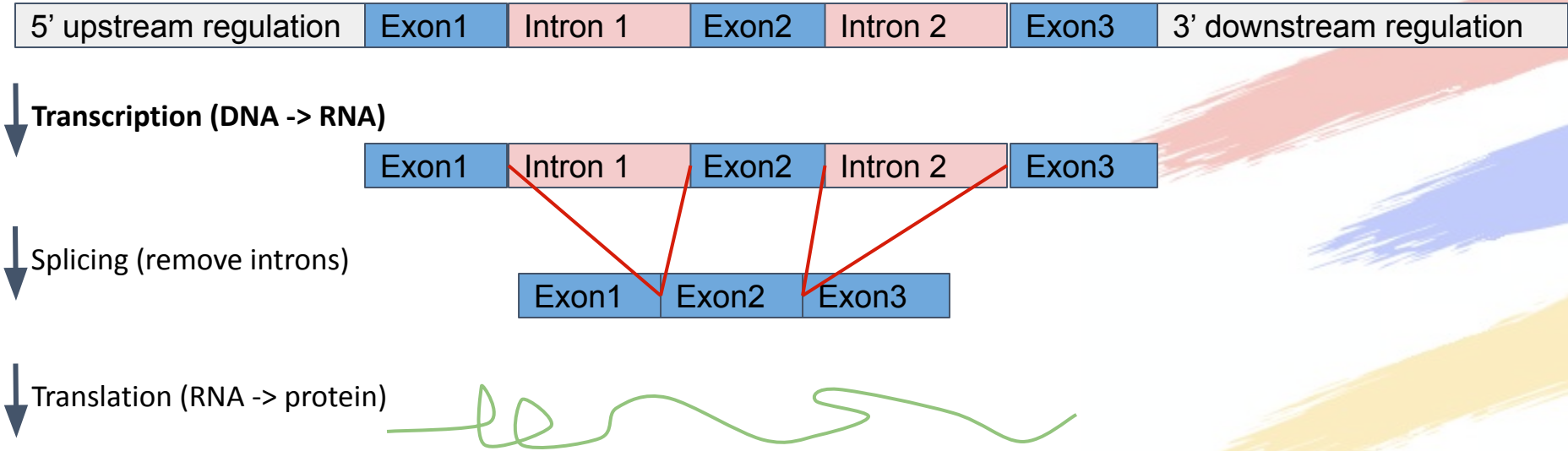
Reads and alignment

- expression counts
- splicing behavior



Expression made easy

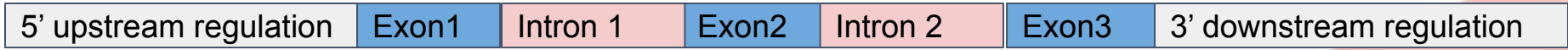
DNA template



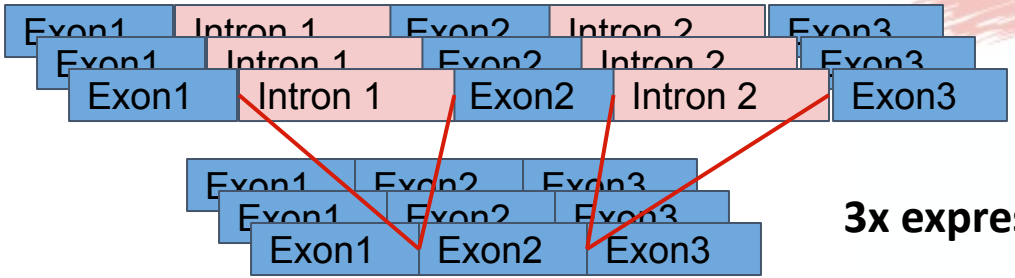
Expression refers to how much of a gene is transcribed and ultimately translated into a protein

Expression made easy

DNA template



Transcription (DNA -> RNA)



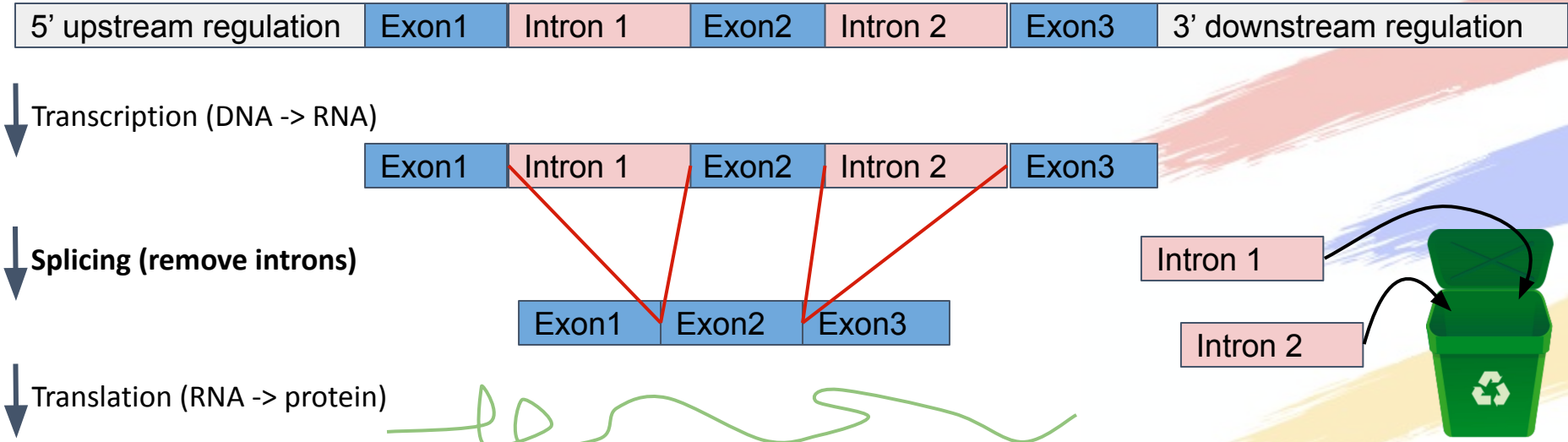
Splicing (remove introns)

3x expression!

Translation (RNA -> protein)

Splicing made easy

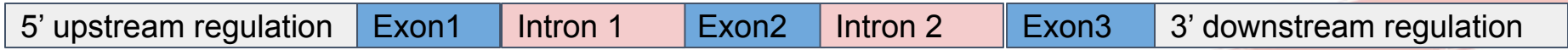
DNA template



Splicing is the process of removing parts of the RNA so that a functioning protein can be built

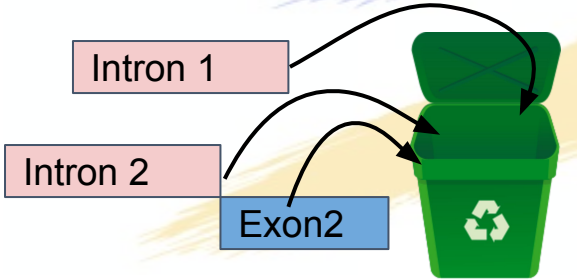
Splicing made easy

DNA template



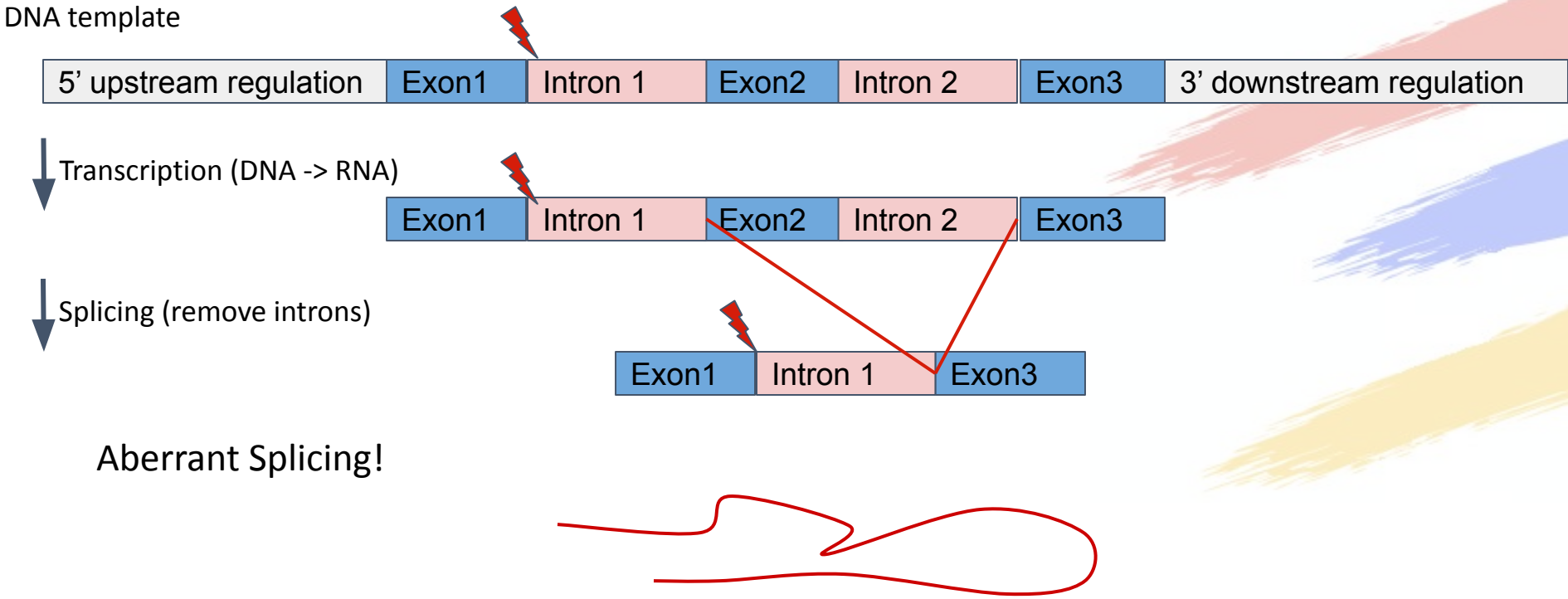
Transcription (DNA -> RNA)

Splicing (remove introns)

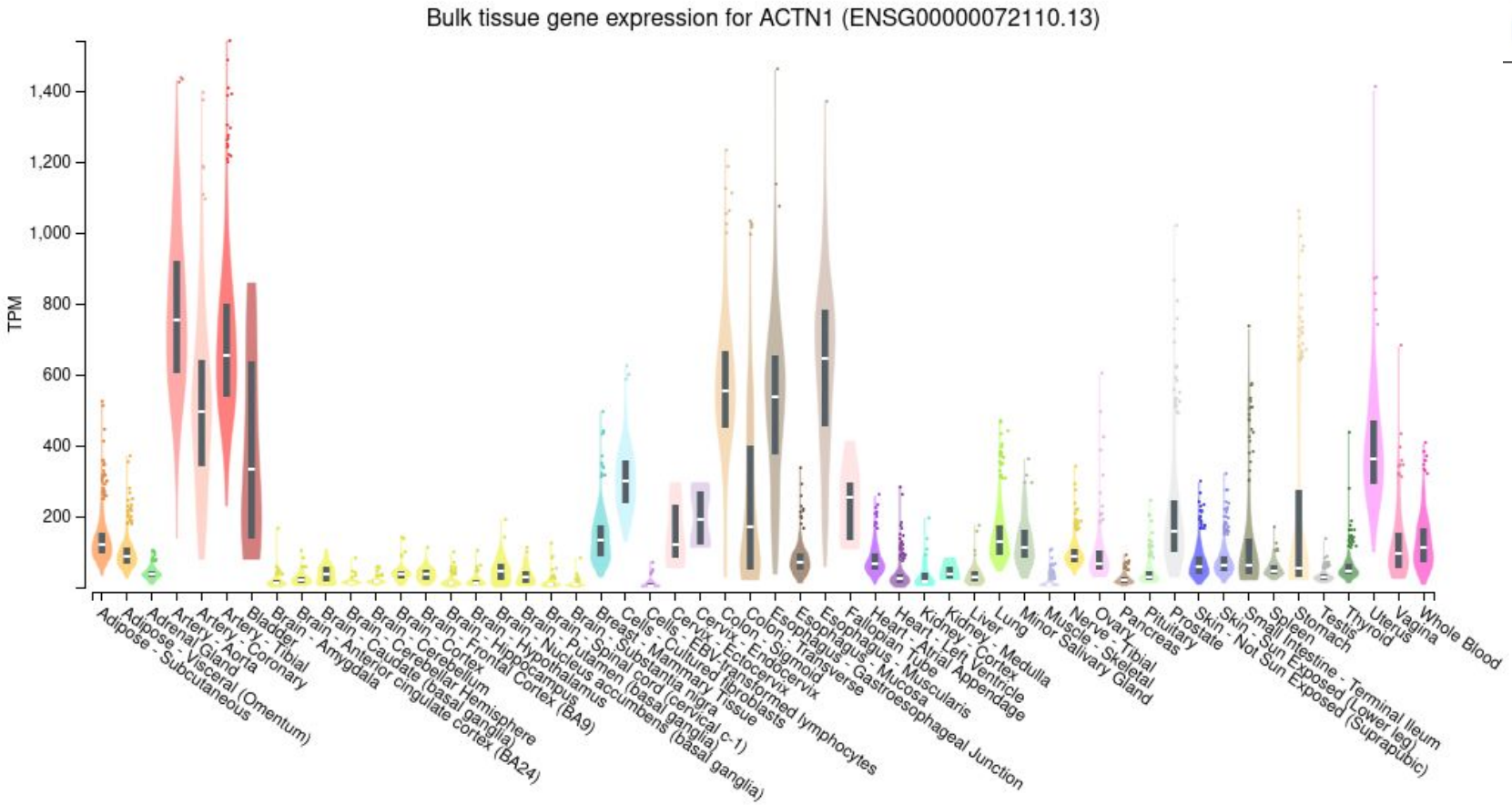


Alternative Splicing

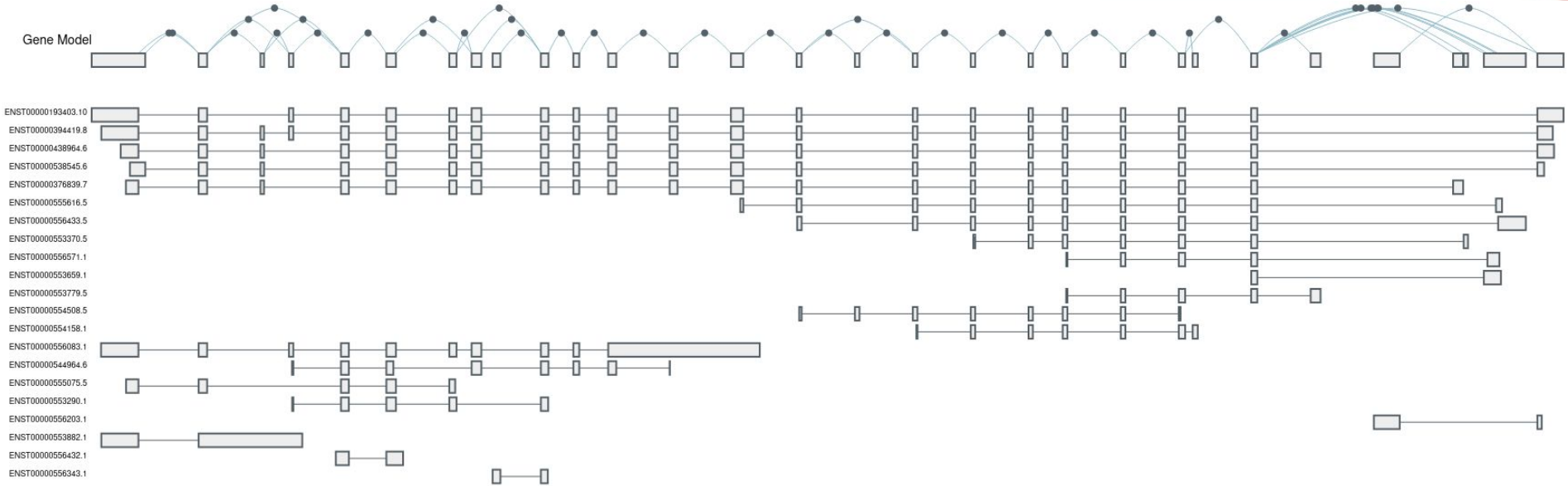
Splicing made easy



Aberrant Splicing!

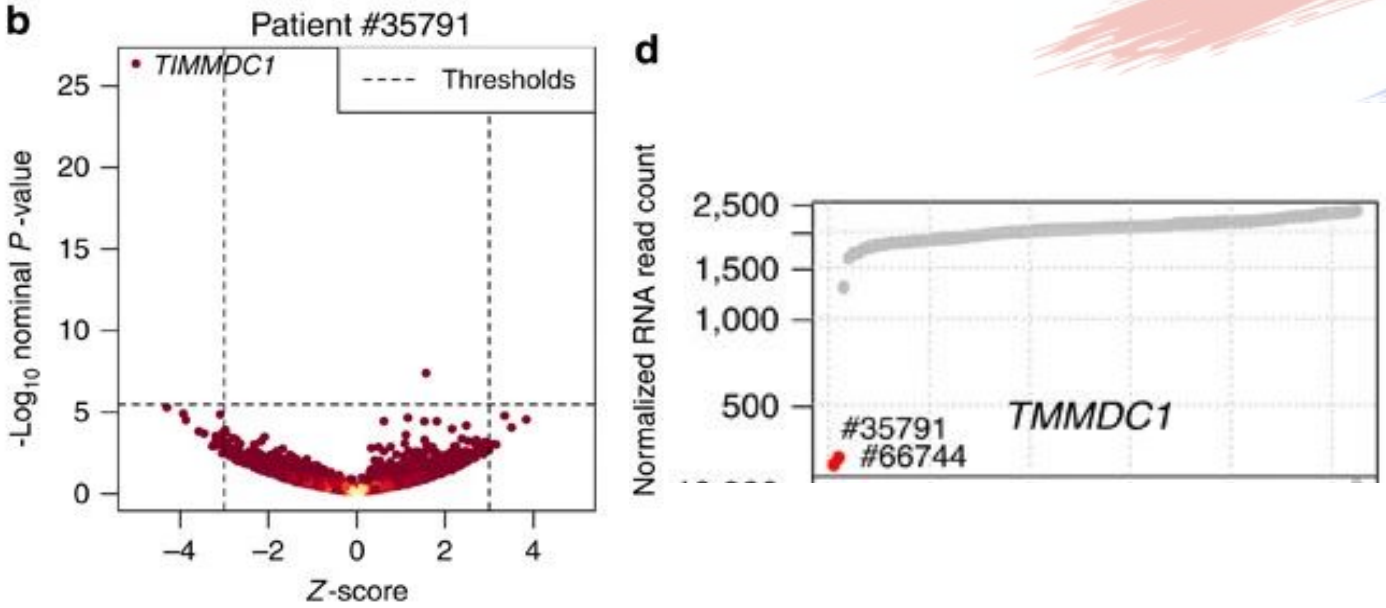


The Genotype-Tissue Expression (GTEx) Project. GTEx Portal on 6/23/22. ACTN1


Gene Model
ENST00000193403.10
ENST00000394419.8
ENST00000438964.6
ENST00000538545.6
ENST00000376839.7
ENST00000555616.5
ENST00000556433.5
ENST00000553370.5
ENST00000556571.1
ENST00000553659.1
ENST00000553779.5
ENST00000554508.5
ENST00000554158.1
ENST00000556083.1
ENST00000544964.6
ENST00000555075.5
ENST00000553290.1
ENST00000556203.1
ENST00000553882.1
ENST00000556432.1
ENST00000556343.1

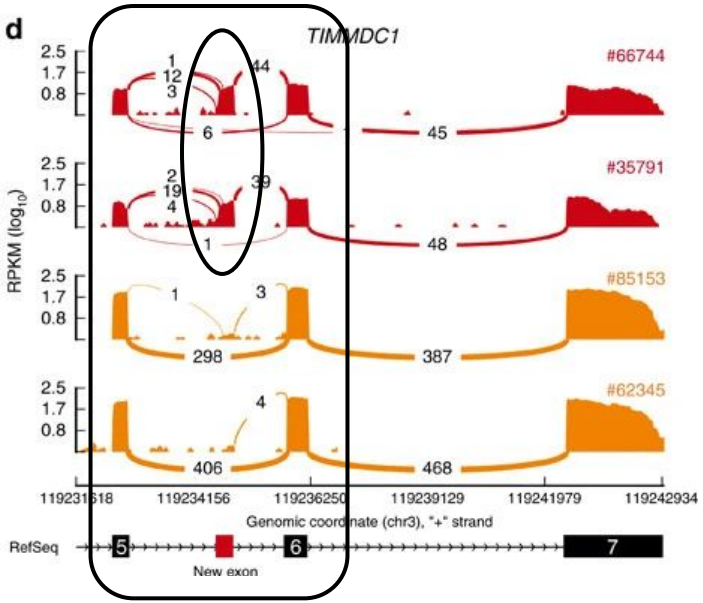
Real life benefit of RNA-seq: TIMMDC1

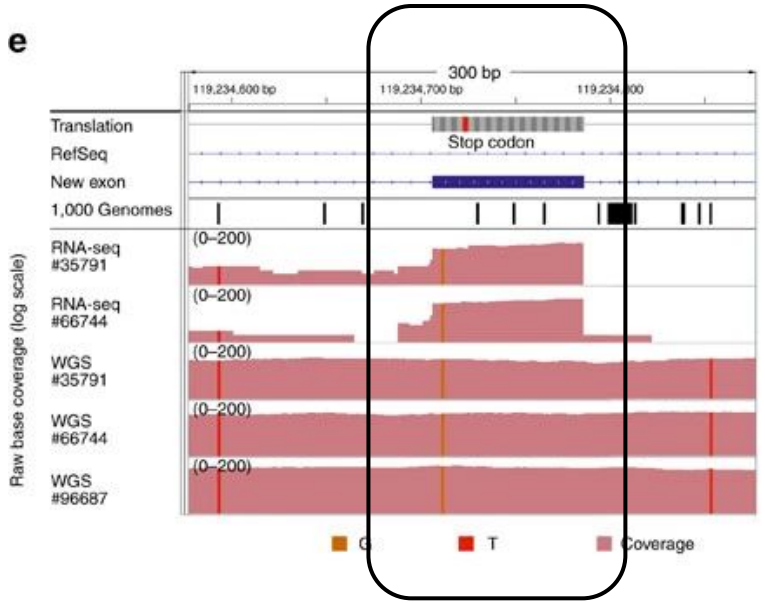
Genetic diagnosis of Mendelian disorders via RNA sequencing -L S. Kremer



Real life benefit of RNA-seq: TIMMDC1

Genetic diagnosis of Mendelian disorders via RNA sequencing -L S. Kremer





Transcriptomics Applications

RNA carries the message from the DNA to the proteins

- Offers a glimpse into a more specialized view. (e.g. different tissues)
  - allows you to infer differences that may affect the protein environment
- Tissue specific
- Downstream analysis applications in:
  - Rare Disease, Cancer, ALS, …

Summary of Sequencing

- Genome level (DNA)
  - coding variants that affect downstream proteins
  - variants that affect gene regulation
  - tumor specific variations

- Transcriptomic level (RNA)
  - expression differences
  - splicing changes

- Understanding DNA and RNA allows us to learn about the causes of human disease and understand the impact our genome has on human health

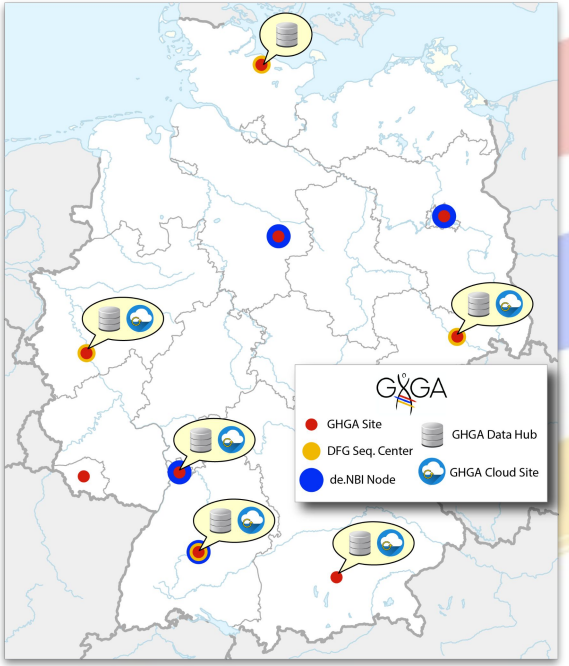
Q&A #1

FAIR Workflows in GHGA

Nicholas Smith and Christian Mertes

GHGA Overview

- Funded since 10/2020 as one of the initial nine first-round **NFDI consortia**
- **Network of six data hubs** co-located with major academic **sequencing centers** (HD, Tü, K, M, DD, Ki)
- Connected to national **cloud infrastructure** (**de.NBI cloud**)
- GHGA Consortium is representing **9 univ., 6 Helmholtz & 5 other research institutions**
    - Broad Interdisciplinary expertise from clinical care, bioinformatics, ethics, legal and IT ⇒ from researchers for researchers



GHGA Contributors and Governance

Board of Directors


O. Kohlbacher
(Univ. Tübingen)


J. Korbel
(EMBL)


O. Stegle
(DKFZ & EMBL)


E. Winkler
(Univ. Heidelberg)

GHGA Team

Co-Applicant Institutions

HMGU

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

University of Cologne

TECHNISCHE UNIVERSITÄT DRESDEN

lrz

TUM

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

UNIVERSITÄT BONN

HEIDELBERGER AKADEMIE DER WISSENSCHAFTEN
Akademie der Wissenschaften des Landes Baden-Württemberg

German Biobank Node
bbmri.de

MAX-PLANCK-GESELLSCHAFT

Universitätsklinikum Tübingen

ikmb
Institut für Klinische Molekularbiologie

CISPA

UNIVERSITÄT DES SAARLANDES

EMBL

DZNE
German Center for Neurodegenerative Diseases within the Helmholtz Association

NCT

MDC
MAX-DELBRÜCK-CENTRUM FÜR MOLEKULARE MEDIZIN IN DER HELMHOLTZ-GEMEINSCHAFT

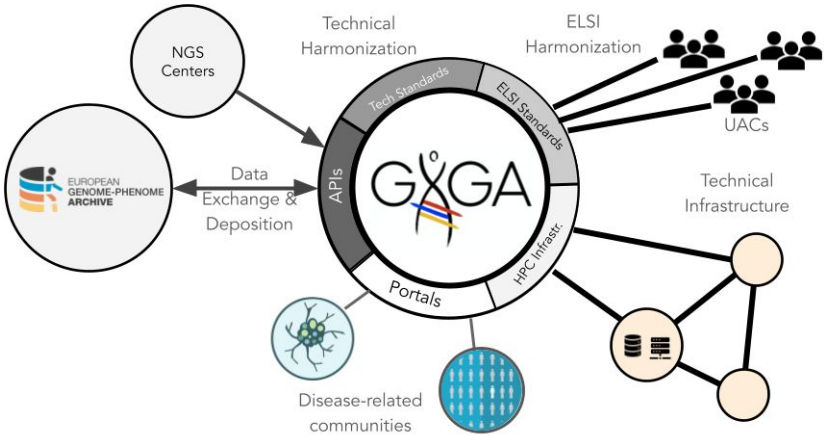
dkfz.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

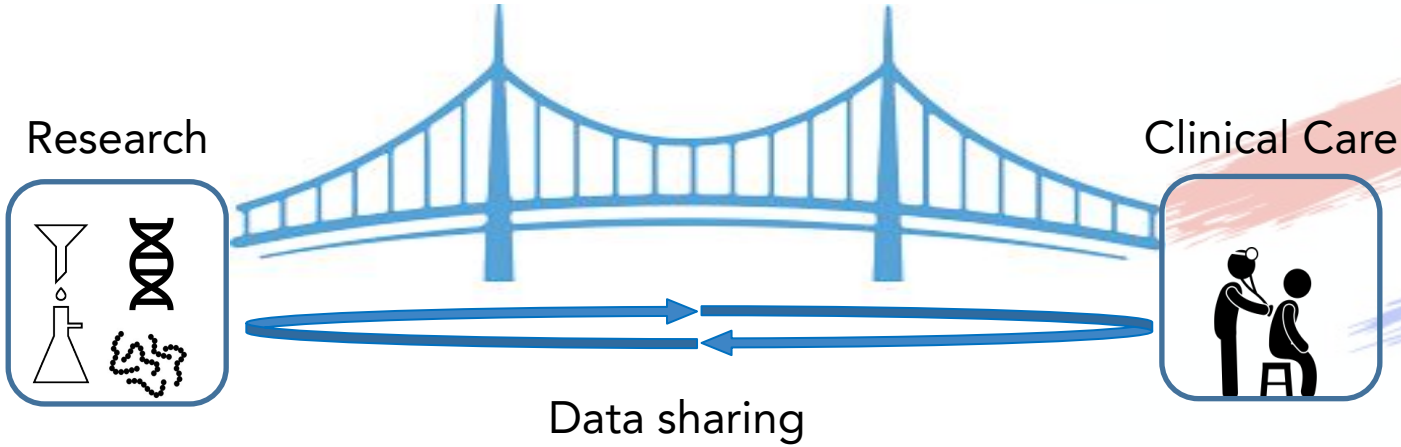
GHGA Goals & Core Objectives

Core Mission: Establish a national infrastructure for human omics data.

- Make human omics data **FAIR**
  - Federated national platform for long-term **data archival**
- **Standardization and Harmonization**:
  - Provide standardized metadata and workflows
  - Standardized processing of data
- **Ethico-legal and data use framework** for
  - Data sharing
  - Protection
  - Analysis
- Establish strong Interfaces with **international initiatives**
  - Connect with EGA, GA4GH, 1+Mio. Genomes, Solve-RD



Human omics data sharing - bench to bedside and back

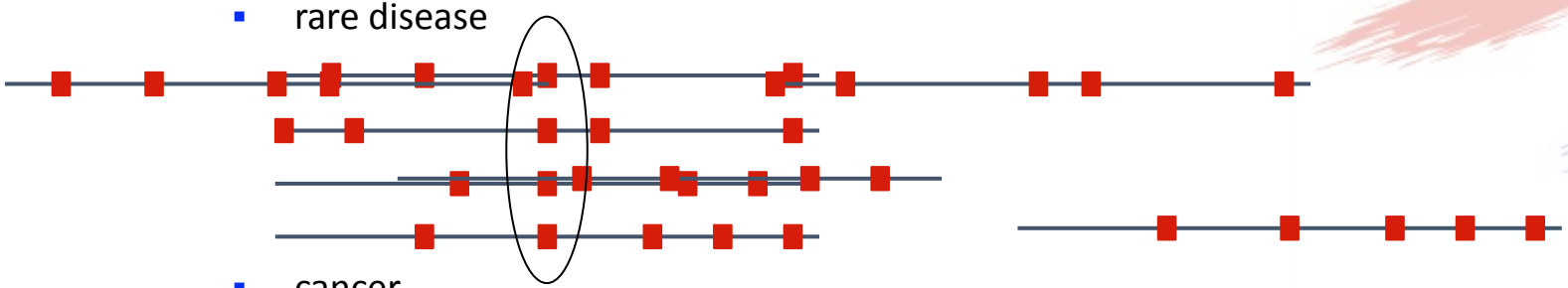


- reuse of omics data in research: biological discovery & replication of findings to show validity.
- Translation of research insights: delivering value in genomic medicine.
- Interactions with diverse communities: computational biologists, molecular geneticists, clinician scientists, physicians, and patients

Archiving Data

What can we do with sequencing data?

- Improves the quality of care and research
  - rare disease



  - cancer
  - common disease
  - covid
- Allows for reanalysis
  - improved healthcare knowledge

Archiving Data

How do we keep data?

- Focus on security, and access control
  - encryption
  - strict definitions of data submitter, data controller, data requester
    - clear cut roles for each

- General Data Protection Regulation (GDPR)
  - requirements surrounding data sharing
  - consent surrounding sensitive health data

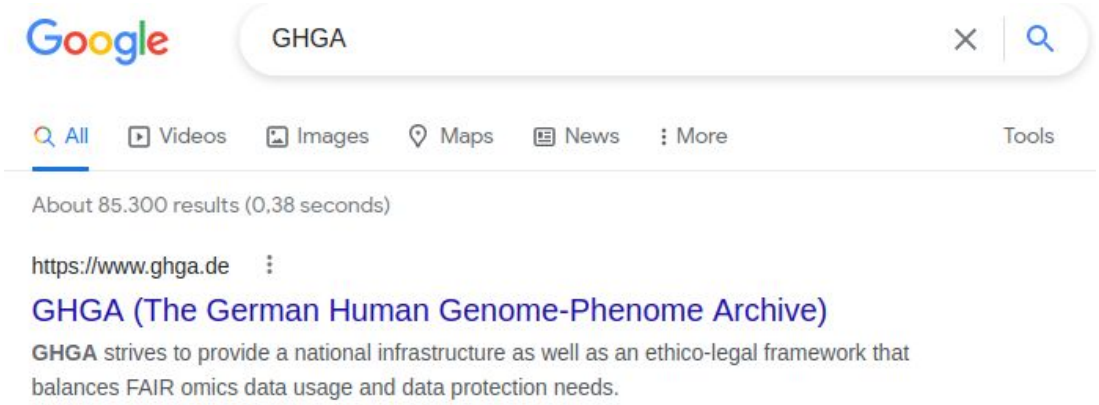
Fair workflows and data

- Findable
- Accessible
- Interoperable
- Reusable

Fair workflows and data

Findable Accessible Interoperable Reusable

- Data, workflows, ontologies, and tools must be findable
- Unique and persistent identifiers



Adapted from the RD-Alliance

FAIR workflows and data

Findable **Accessible** Interoperable Reusable

- Authentication and authorisation must be clear and transparent.
- Usage of international standard protocols and APIs, e.g. GA4GH

Adapted from the RD-Alliance

Fair workflows and data

Findable Accessible **Interoperable** Reusable

- Standardized processing: assembly, variant annotation, …
- Software and workflow applications should be hardware agnostic
- Data read/write/exchange should comply with GA4GH standards

Adapted from the RD-Alliance

Fair workflows and data

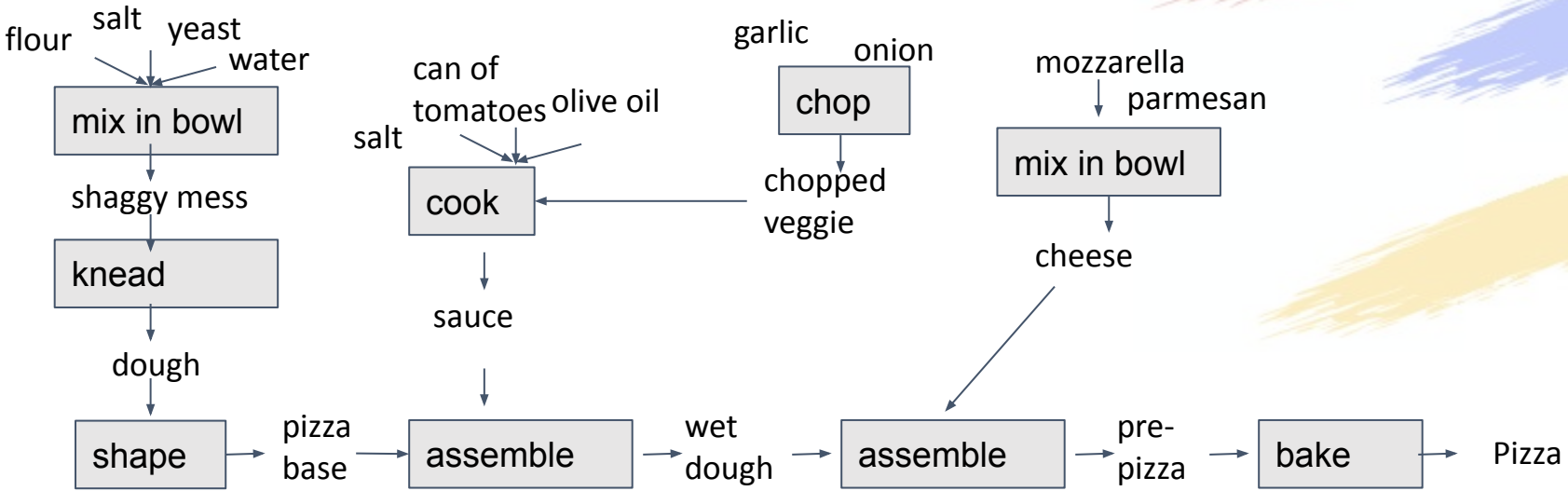
Findable Accessible Interoperable **Reusable**

- Data and software should be able to be reused
- Using the right licenses and consents

Adapted from the RD-Alliance

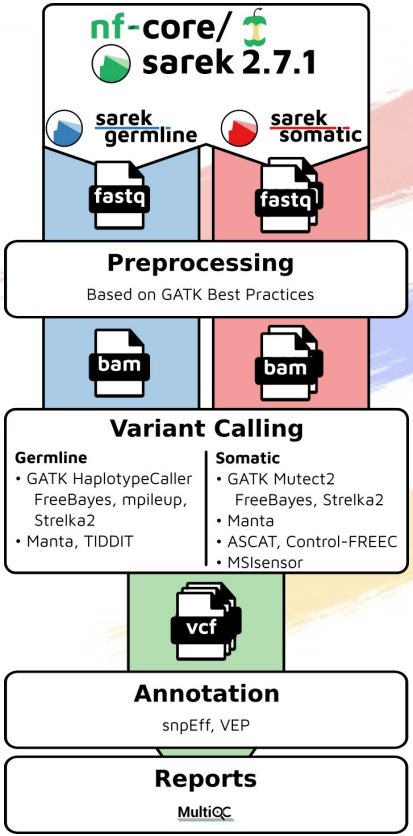
Workflows

- A collection of tools that collect a predescribed input and output packaged to accomplish a larger task
- Understanding the bottlenecks and what can be done in parallel



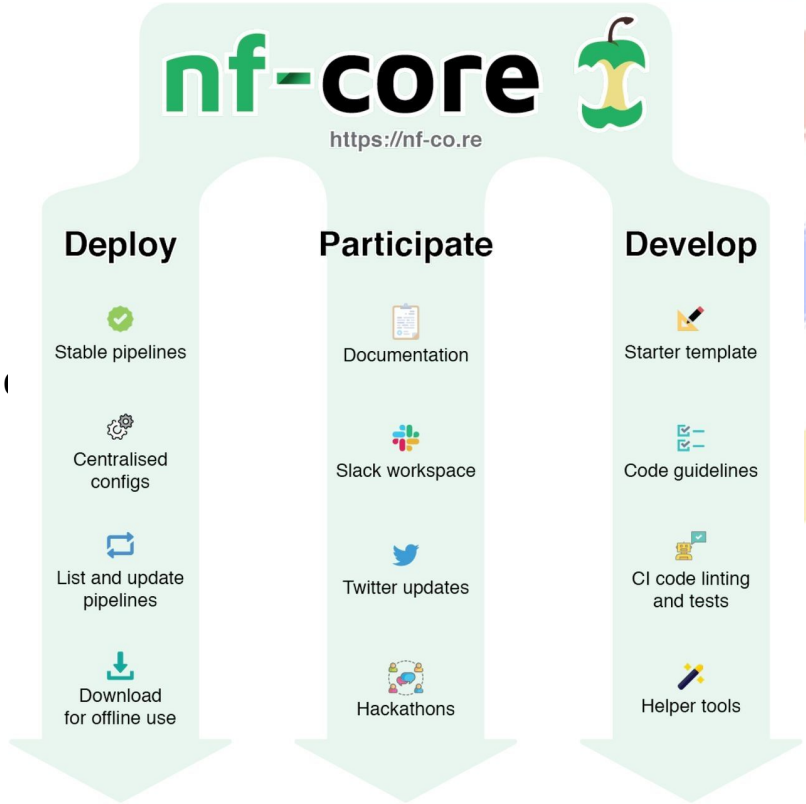
Workflow for short-read DNA sequencing

- GHGA team specifies a fixed set of parameters
- Workflow covers:
  - Read trimming
  - Read alignment
  - Variant calling
  - Structural variant calling
  - Reporting and Annotation



Collaborating with the bioinformatics community

- nf-core community
  - Community of thousands
  - Sharable and reproducible workflows
- Modular design
  - Can reuse specific tools
- Accurate and accepted
  - Software and tools are accepted by th wider bioinformatics community
  - Developing benchmarking for major version release



Workflow Summary and Newsletter

- GHGA is establishing a legal and technical infrastructure
- Useful to use databases to improve diagnostic and clinical care
- Data and workflows needs to be FAIR
- GHGA aims to serve the research community

Survey

Analysis services

Secondary analysis

Matchmaking / Beacon service

Access to raw data

Summary statistics

Standardized Workflows

Intermediate files

https://commons.wikimedia.org/wiki/File:Professor_Thinking_Cartoon.svg