Introduction to biology big data and omics

Silvia Bottini, PhD

Junior professor chair INRAE

Institut Sophia Agrobiotech

Silvia.Bottini@inrae.fr

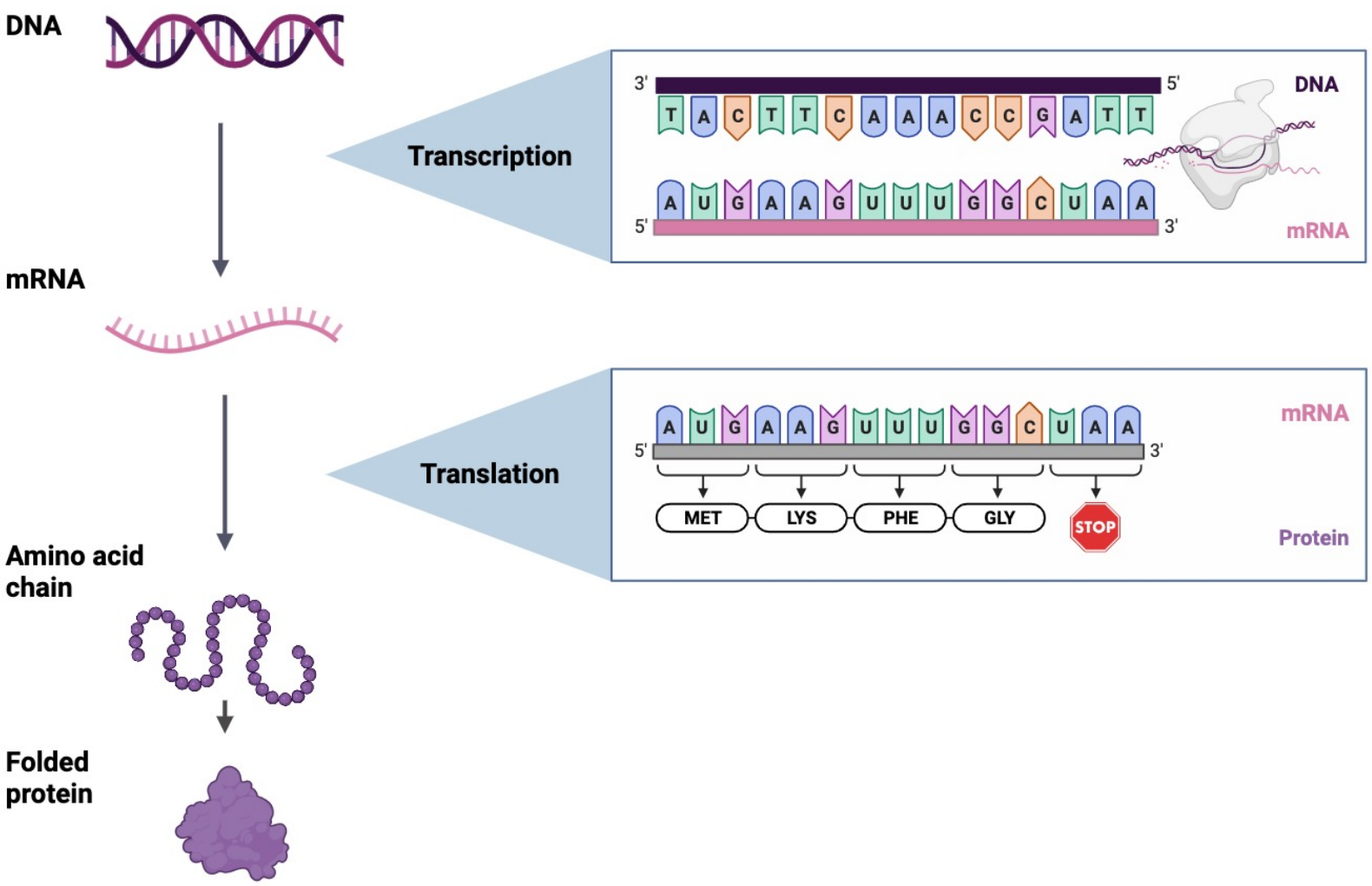
Course Structure

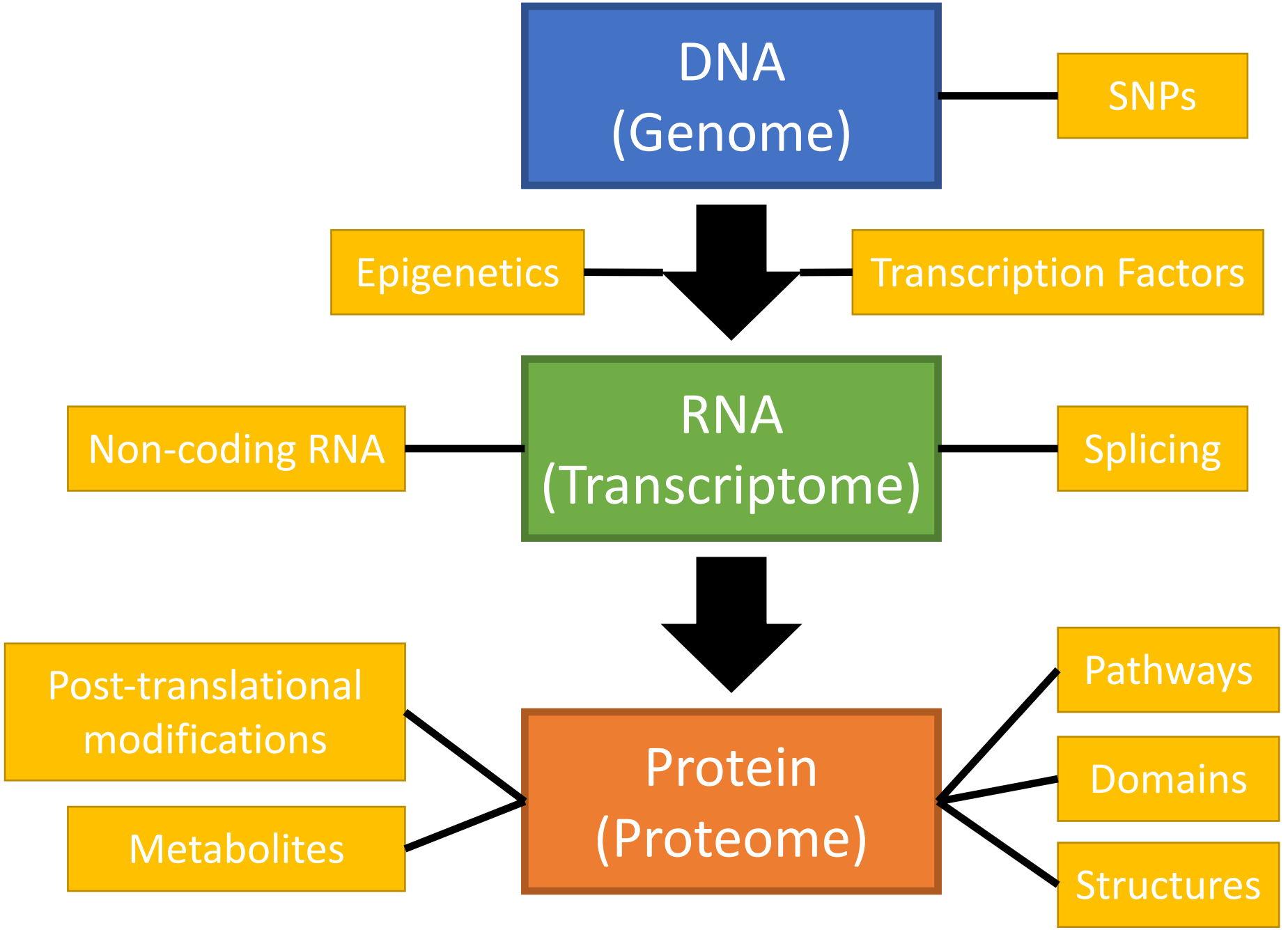
- Central Dogma Data Sources
  - Genomes and Annotations
  - Protein Domains and Structures
  - Reactions, Pathways and Interactions
- Experimental Techniques, Datatypes and Repositories
  - Sequencing and Variants
  - Proteomics and Metabolites
- Omics
  - General concepts
  - Analysis approaches

Course Structure

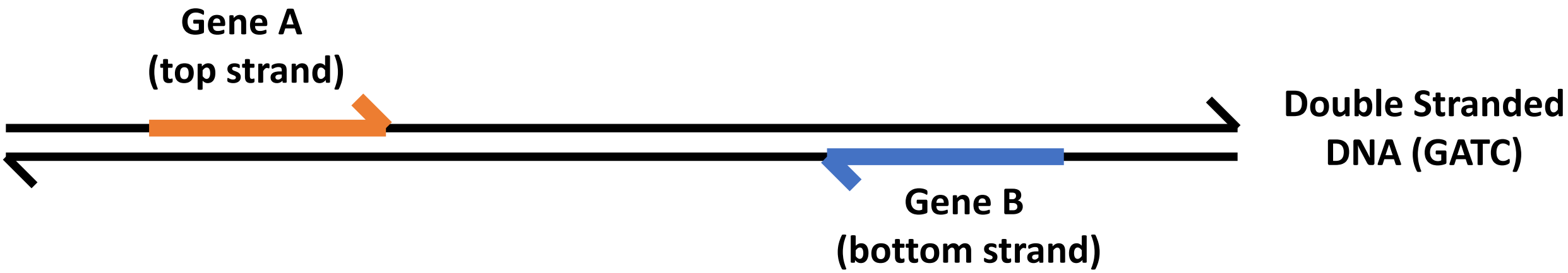
- Central Dogma Data Sources
  - Genomes and Annotations
  - Protein Domains and Structures
  - Reactions, Pathways and Interactions
- Experimental Techniques, Datatypes and Repositories
  - Sequencing and Variants
  - Proteomics and Metabolites
- Omics
  - General concepts
  - Analysis approaches

Central dogma

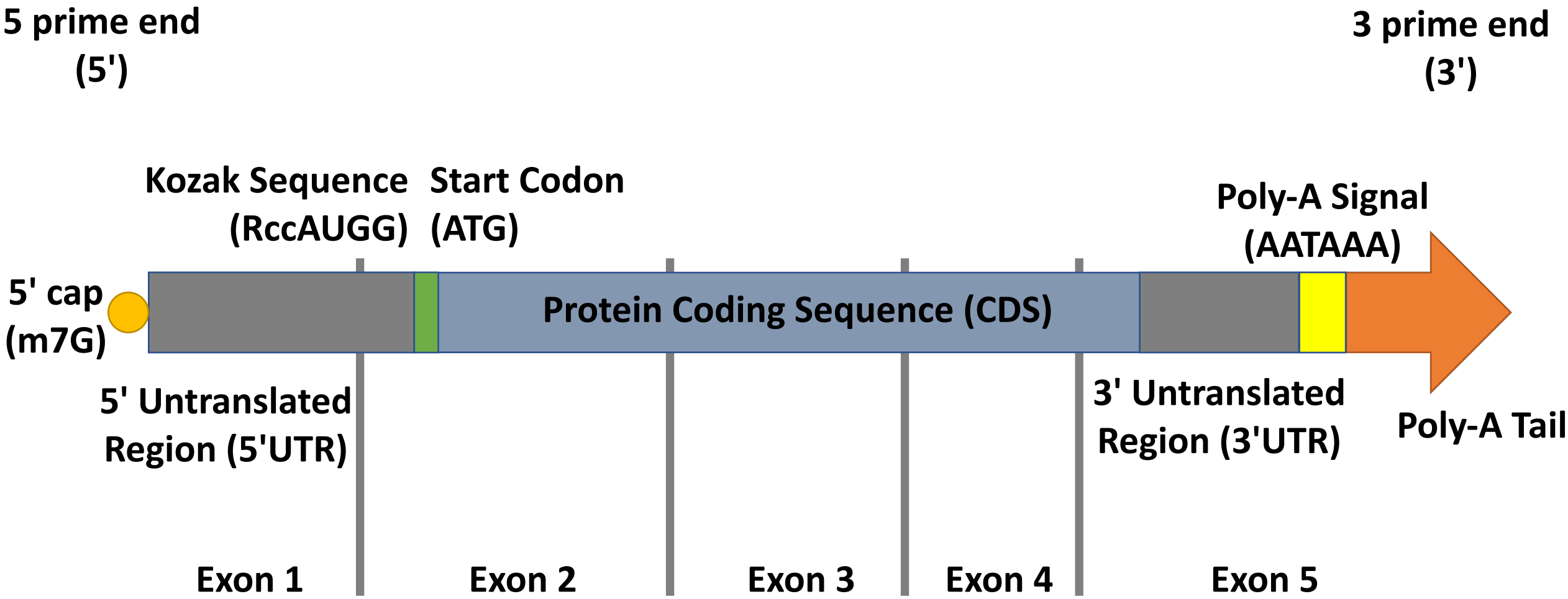



DNA
(Genome)
SNPs
Epigenetics
Transcription Factors
RNA
(Transcriptome)
Non-coding RNA
Splicing
Post-translational
modifications
Metabolites
Protein
(Proteome)
Pathways
Domains
Structures

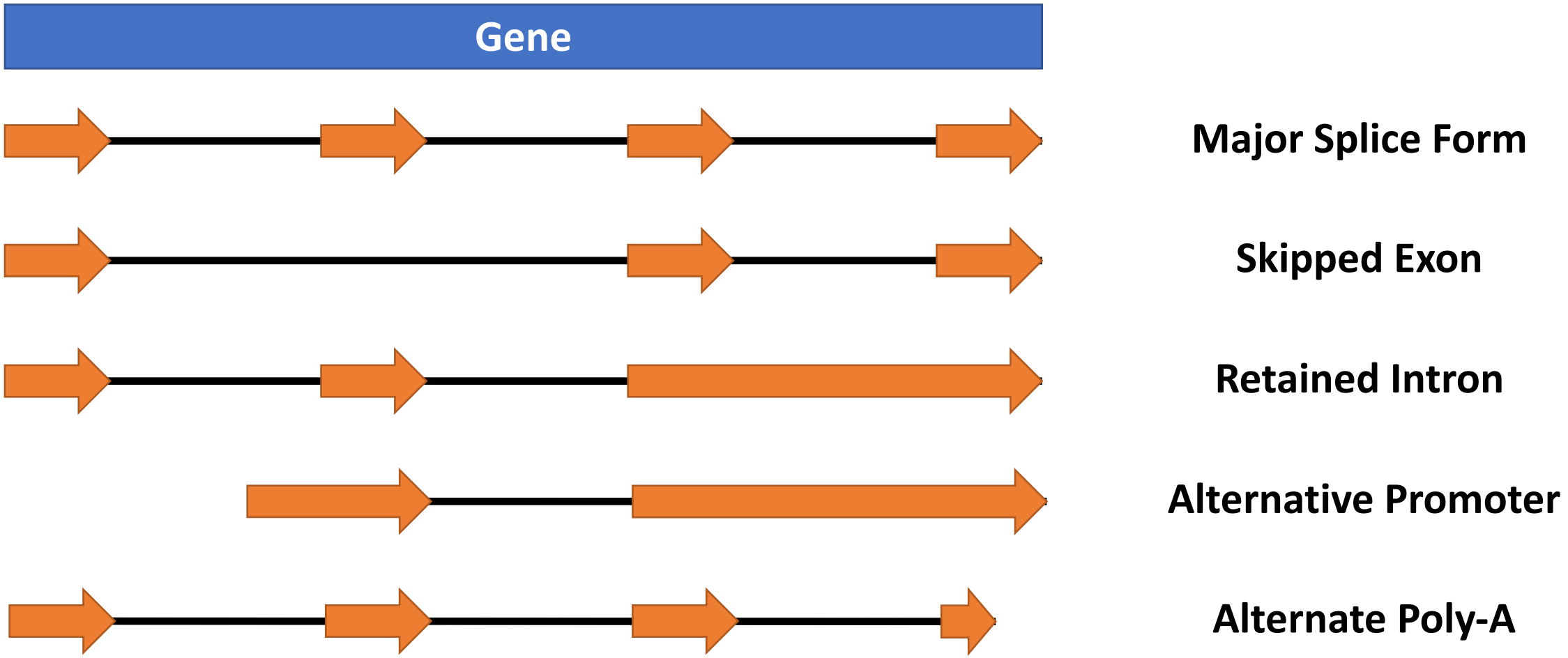
Annotation Structure



Mature Transcript Structure



Alternative Splicing



Genomes and Annotations

- Genome Assemblies
  - Underlying sequence of the organism’s chromosomes
  - Often starts as scaffolds / contigs
  - Eventually assembled into chromosomes (still with holes)
    - Only one chromosome sequence per chromosome
  - Variations (natural or clinical) are stored separately
  - Assembly is refined and improved over time, new releases get new names

Genome Assembly Nomenclature

- Chromosome / Scaffold sequences
  - Originally deposited with ENA / NCBI as sequence records

- Genome Assembly
  - Given an official name by a supervising group (sometimes two!)
  - Fixed coordinates at that point

Current Human Genome

- Assembly Name: GRCh38
- Current Patch: GRCh38.p13
- Managed by: Genome Reference Consortium
- Assembly type: Chromosomal
- Chromosome: Chr1 = CM000663.2 = NC_000001.11
- Genome: GCF_000001405.39 (Assembly Refseq)
  GCA_000001405.28 (Assembly Genbank)

Genome Annotation Sets

- Built on top of a specific assembly
- Combination of prediction tools and real data
- Main annotation is Genes, Transcripts, Coding Sequences
- Many other tracks often added

- Different sites will have different annotations
- Annotations updated more frequently than assemblies

Genome Annotation Details

Genome-Annotation-Data

```
##Genome-Annotation-Data-START##
Annotation Provider::NCBI
Annotation Status::Updated annotation
Annotation Name::Homo sapiens Updated Annotation Release 109.20210226
Annotation Version::109.20210226
Annotation Pipeline::NCBI eukaryotic genome annotation pipeline
Annotation Software Version::8.6
Annotation Method::Best-placed RefSeq; propagated RefSeq model
Features Annotated::Gene; mRNA; CDS; ncRNA
##Genome-Annotation-Data-END##
```

General stats

Total No of Genes	60649	Total No of Transcripts	237012
Protein-coding genes	19955	Protein-coding transcripts	86757
Long non-coding RNA genes	17944	- full length protein-coding	61015
Small non-coding RNA genes	7567	- partial length protein-coding	25742
Pseudogenes	14773	Nonsense mediated decay transcripts	18881
- processed pseudogenes	10667	Long non-coding RNA loci transcripts	48752
- unprocessed pseudogenes	3565		
- unitary pseudogenes	241		
- polymorphic pseudogenes	49		
- pseudogenes	15	Total No of distinct translations	63968
		Genes that have more than one distinct translations	13689
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments	409		
- pseudogenes	236		

Assembly	GRCh38.p14 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.29, Dec 2013
Base Pairs	3,099,750,718
Golden Path Length	3,099,750,718
Assembly provider	Genome Reference Consortium
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Mar 2023
Database version	110.38
Gencode version	GENCODE 44

Gene counts (Primary assembly)

Coding genes	19,831 (excl 650 readthrough)
Non coding genes	25,959
Small non coding genes	4,864
Long non coding genes	18,874 (excl 319 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,239 (excl 1 readthrough)
Gene transcripts	252,894

Viewing Annotated Genomes

- Mostly web based
  - Species specific sites
  - Generic multi-species sites

- Often adds more information
  - Regulation, conservation, repeats
  - Experimental datasets
  - Upload your own

Generic genome viewer sites

Ensembl https://www.ensembl.org

UCSC Browser https://genome.ucsc.edu

WashU Browser https://epigenomegateway.wustl.edu/

Species specific genome viewer sites

tair

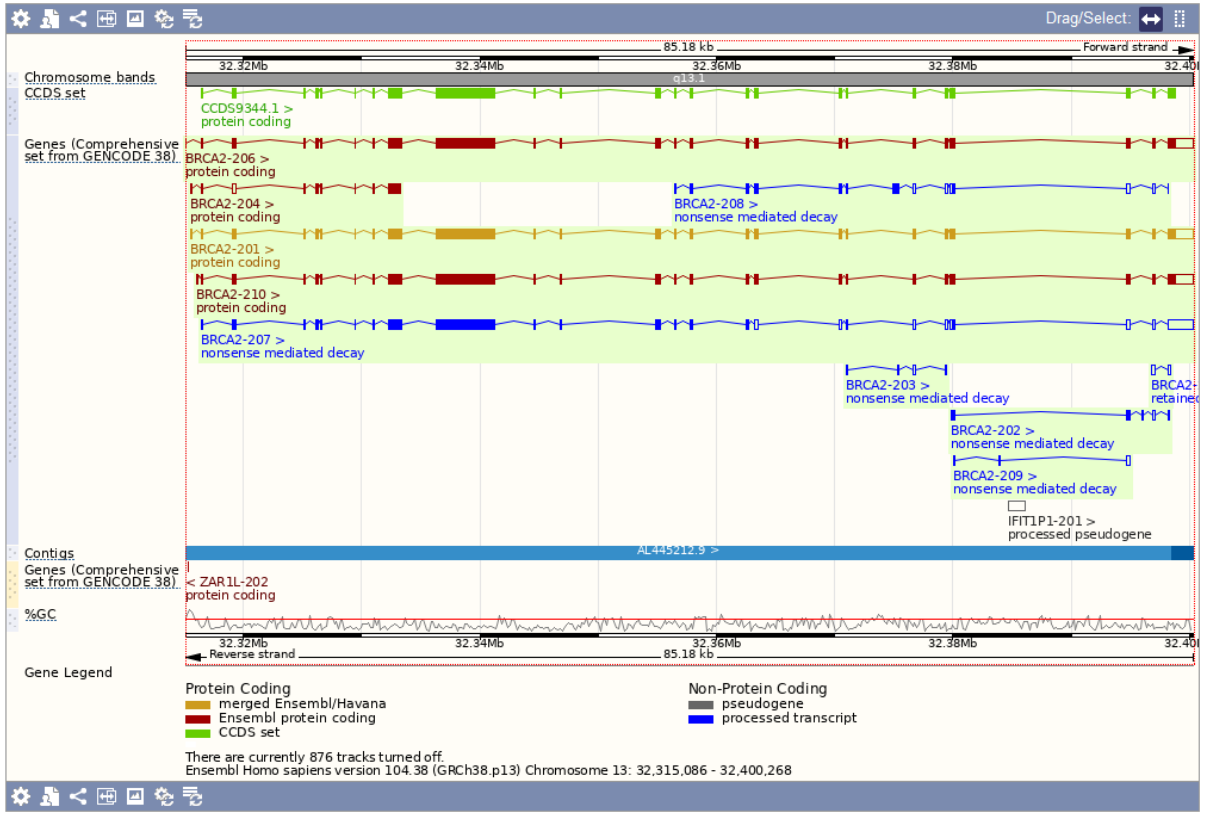
Arabidopsis https://www.arabidopsis.org

Drosophila https://flybase.org/

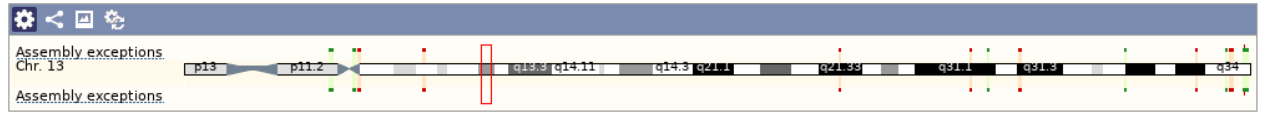
WormBase

Nematode worms https://wormbase.org

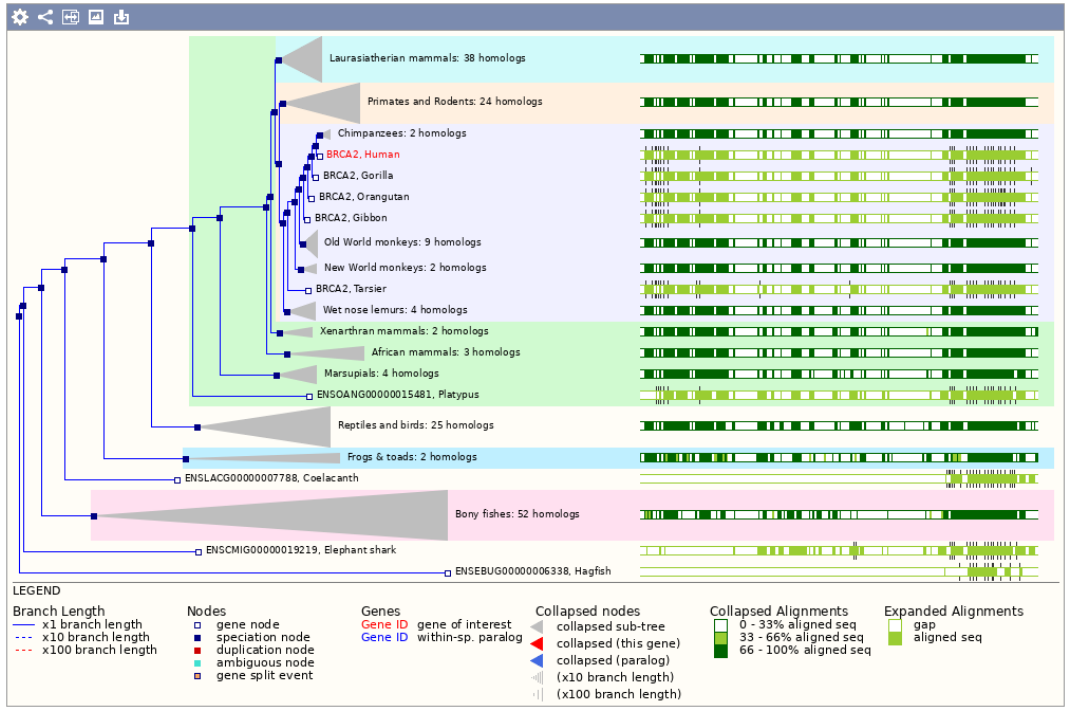
Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt Match
BRCA2-201	ENST00000380152.8	11954	3418aa	Protein coding	CCDS9344	P51587
BRCA2-210	ENST00000680887.1	11880	3418aa	Protein coding	CCDS9344	-
BRCA2-206	ENST00000544455.6	11854	3418aa	Protein coding	CCDS9344	P51587
BRCA2-204	ENST00000530893.6	2011	481aa	Protein coding	-	A0A590UJI7
BRCA2-207	ENST00000614259.2	11763	2649aa	Nonsense mediated decay	-	-



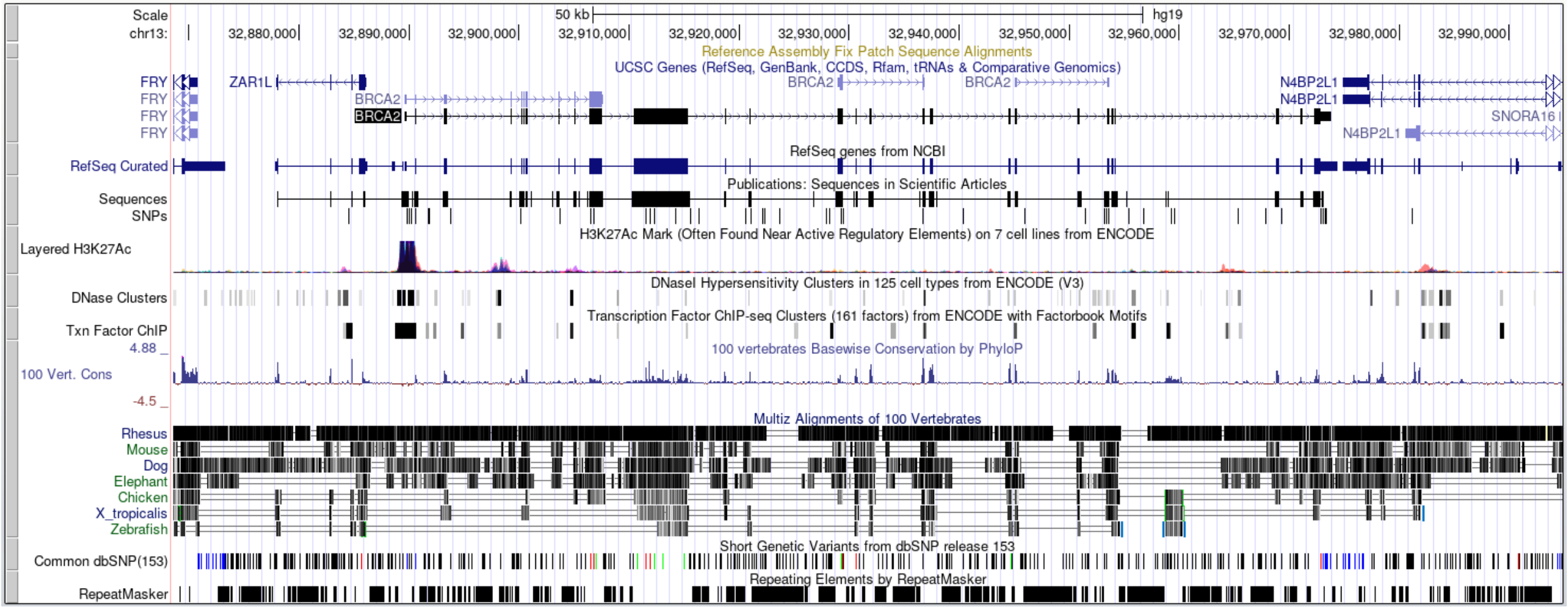
Chromosome 13: 32,315,086-32,400,268



Region in detail



Track Based Displays



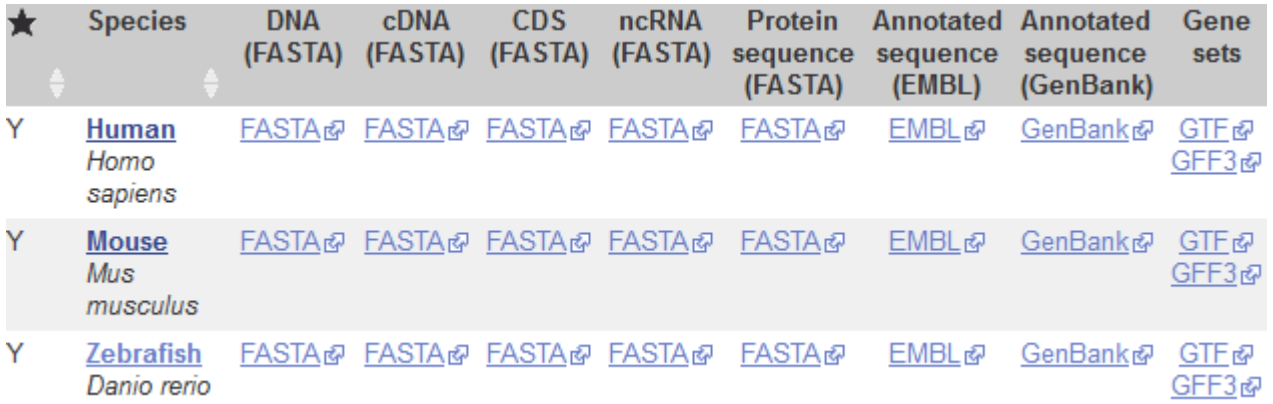
Large Scale Queries

- Large scale querying and export of genomic data

- Annotations, Sequences, Variants etc.
  - Select data type (eg genes)
  - Select genome species
  - Select genes / regions / identifiers
  - Select attributes to export
  - Generate report

Genome File Formats

- Genome Assemblies
  - Chr sequence, FastA format
  - A small header plus DNA bases
  - Also used for RNA / protein



★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets
Y	**Human** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3
Y	**Mouse** *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3
Y	**Zebrafish** *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3

- Gene Annotations
  - GFF or GTF format (both very similar)
  - Hierarchical format linking exons to transcripts to genes

https://www.ensembl.org/info/data/ftp/index.html

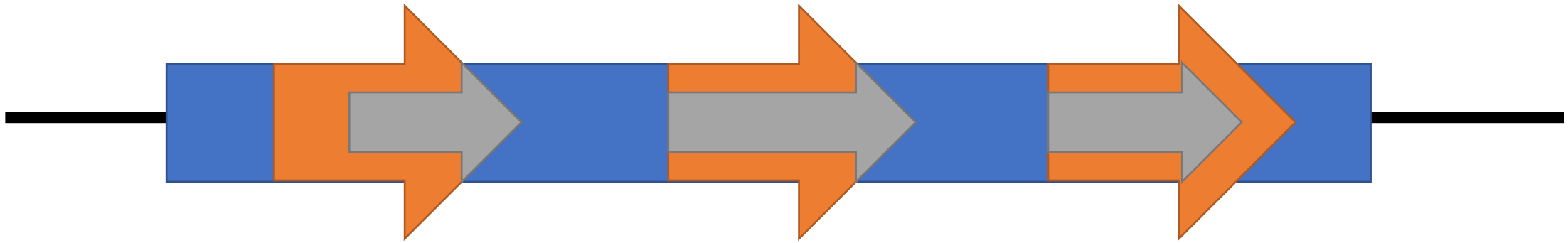
FastA Format Data

>**I** dna:chromosome chromosome:R64-1-1:I:1:230218:1 REF
CCACACCACACCCACACACCCACACACCACACCACACACCACACCCACACACACA
CATCCTAACACTACCCTAACACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTT
ACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAACCATACCACTCCGAAC
CACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
>**II** dna:chromosome chromosome:R64-1-1:II:1:813184:1 REF
AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATGTTCAACCA
AAAGCTACTTACTACCTTTATTTTATGTTTACTTTTTATAGGTTGTCTTTTTATCCCACT
TCTTCGCACTTGTCTCTCGCTACTGCCGTGCAACAAACACTAAATCAAAACAATGAAATA
CTACTACATCAAAACGCATTTTCCCTAGAAAAAAAATTTTCTTACAATATACTATACTAC

IUPAC Ambiguity Codes

IUPAC Code	Meaning
A	A
C	C
G	G
T/U	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	G or A or T or C

Annotation Descriptions



Gene

Exon (combined into transcript)

Coding Exon

GFF (Strictly GFF.2)

- Comprehensive annotation format
- Tab delimited
- Flexible – able to accommodate multi-features

GFF File Fields

1. Chromosome
2. Source
3. Feature Type
4. Start
5. End
6. Score
7. Strand (+/-)
8. Frame (1,2,3)
9. Group/Attributes

```
1 hav gene       11869 14409 . + . ID=gene:ENSG223972;Name=DDX11L1;description=DEAD/H-box 1;gene_id=ENSG223972
1 hav transcript 11869 14409 . + . ID=transcript:ENST456328;Parent=gene:ENSG223972;Name=DDX11L1-002;
1 hav exon       11869 12227 . + . Parent=transcript:ENST456328;exon_id=ENSE2234944;rank=1
1 hav exon       12613 12721 . + . Parent=transcript:ENST456328;exon_id=ENSE3582793;rank=2
1 hav exon       13221 14409 . + . Parent=transcript:ENST456328;exon_id=ENSE2312635;rank=3
```

Positions are 1-indexed, fully open

GTF

- Targeted at gene structure definition
- Variant of GFF with stricter rules about attributes
  - Attributes must use `gene_id` and `transcript_id`
  - Commas mandatory and single space delimited

```
1 havana gene       11869 14409 . + . gene_id "ENSG223972"; gene_name "DDX11L1";

1 havana transcript 11869 14409 . + . gene_id "ENSG223972"; transcript_id "ENST456328"; transcript_name "DDX11L1-202";

1 havana exon       11869 12227 . + . gene_id "ENSG223972"; transcript_id "ENST456328"; exon_number "1"; exon_id
"ENSE2234944";

1 havana exon       12613 12721 . + . gene_id "ENSG223972"; transcript_id "ENST456328"; exon_number "2"; exon_id
"ENSE3582793";

1 havana exon       13221 14409 . + . gene_id "ENSG223972"; transcript_id "ENST456328“; exon_number "3"; exon_id
"ENSE2312635";
```

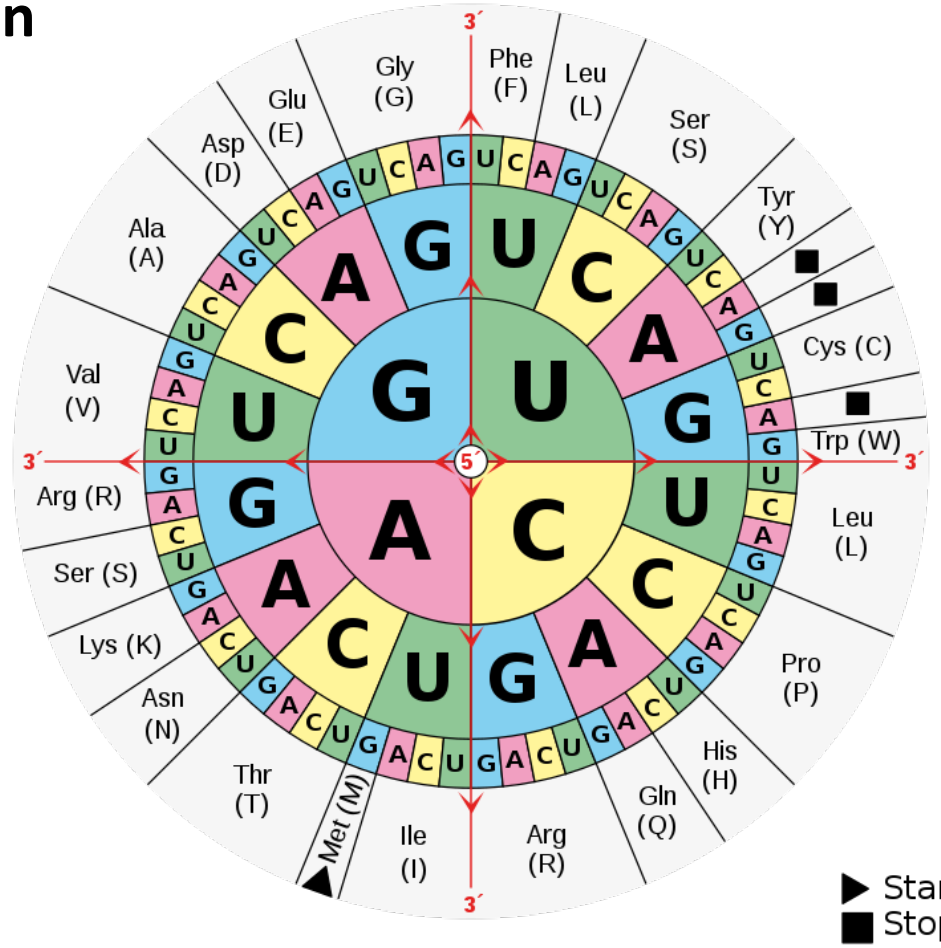
mRNA Translation into Protein

Start Codon **Stop Codon**

```
GACACC ATG AGC ACT GAA ... CTG TGA
UTR    Met Ser Thr Glu     Arg Stp
```

- Most species use the same code
- Some have minor differences

https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

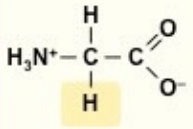


Often internal

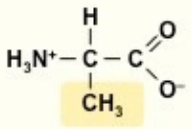
Often surface

Often Binding or catalytic sites

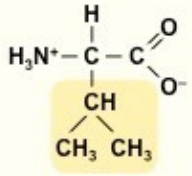
NON-POLAR



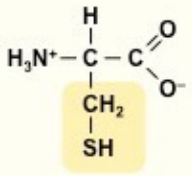
Glycine
(Gly / G)



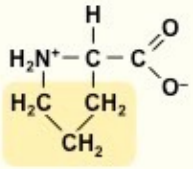
Alanine
(Ala / A)



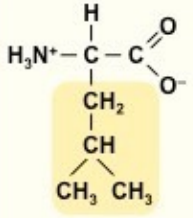
Valine
(Val / V)



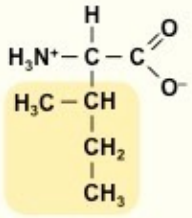
Cysteine
(Cys / C)



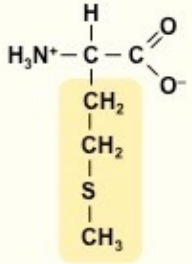
Proline
(Pro / P)



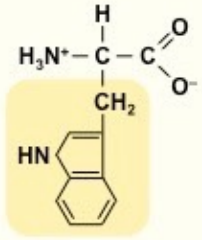
Leucine
(Leu / L)



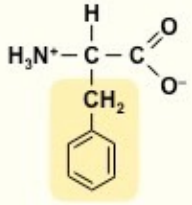
Isoleucine
(Ile / I)



Methionine
(Met / M)

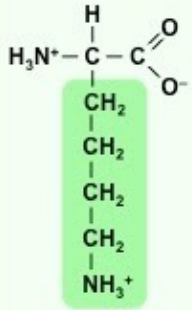


Tryptophan
(Trp / W)

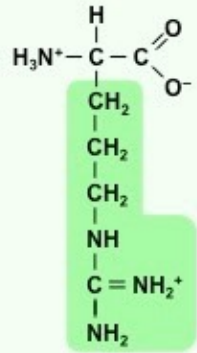


Phenylalanine
(Phe / F)

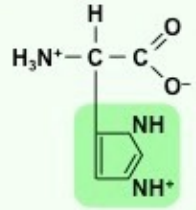
+ CHARGE



Lysine
(Lys / K)

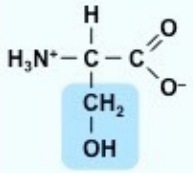


Arginine
(Arg / R)

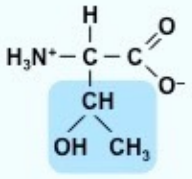


Histidine
(His / H)

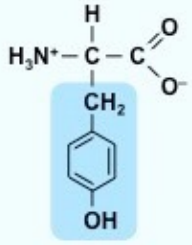
POLAR



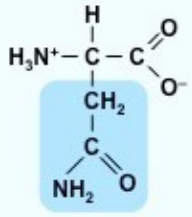
Serine
(Ser / S)



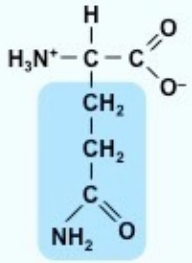
Threonine
(Thr / T)



Tyrosine
(Tyr / Y)

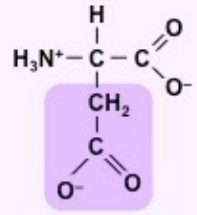


Asparagine
(Asn / N)

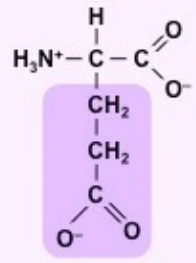


Glutamine
(Gln / Q)

– CHARGE



Aspartic Acid
(Asp / D)



Glutamic Acid
(Glu / E)

Protein Domain Information

- A single protein can have more than one functional unit
  - Proteins are annotated with functional ‘domains’
  - A domain is normally linked with a globular folded structure

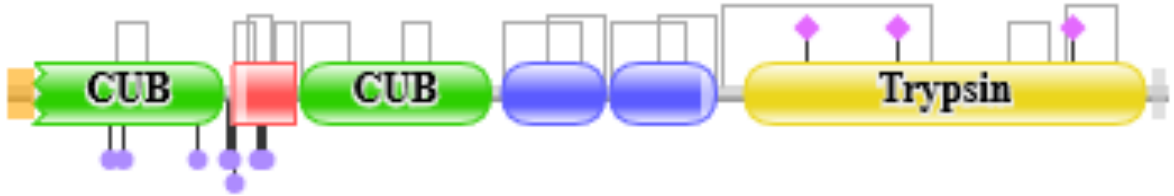
- Domain structures are re-used to provide modular functionality across multiple proteins.
  - Often linked to exon structures or splice variation

- It can be useful to know the key functional amino acids
  - Binding pockets
  - Active sites

Protein Domain Databases

http://smart.embl-heidelberg.de/

https://www.ebi.ac.uk/interpro/



Tryp_SPc domain

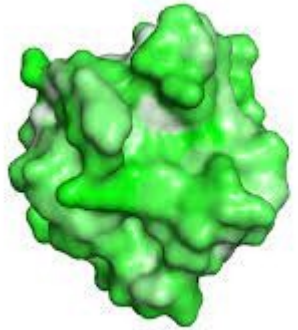
This is a SMART **Tryp_SPc** domain (full annotation).

Position:	**437** to **675**
E-value:	**4.3565597296112e-75** (HMMER2)

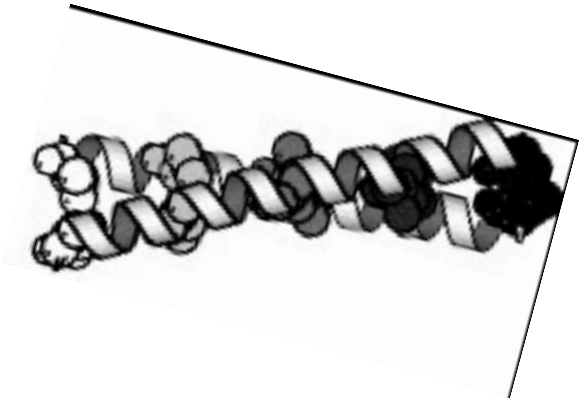
SMART ACC:	SM000020
Definition:	Trypsin-like serine protease
Description:	Many of these are synthesised as inactive precursor zymogens that are cleaved during limited proteolysis to generate their active forms. A few, however, are active as single chain molecules, and others are inactive due to substitutions of the catalytic triad residues.

Types of domain

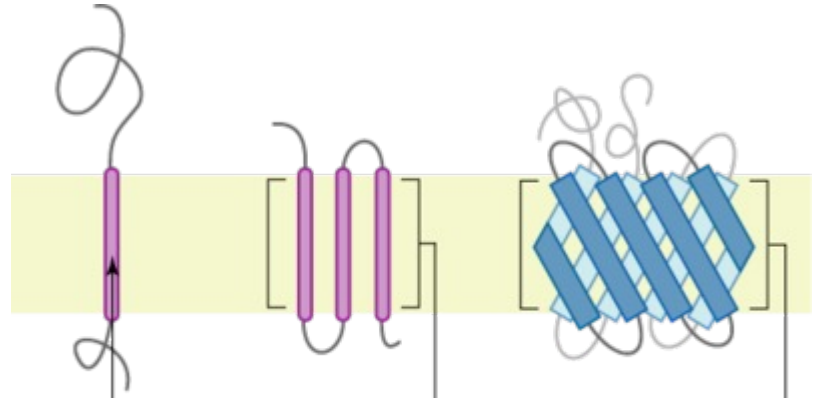
- Globular
  - Forms a concerted 3D structure
  - Most catalytic and some binding domains



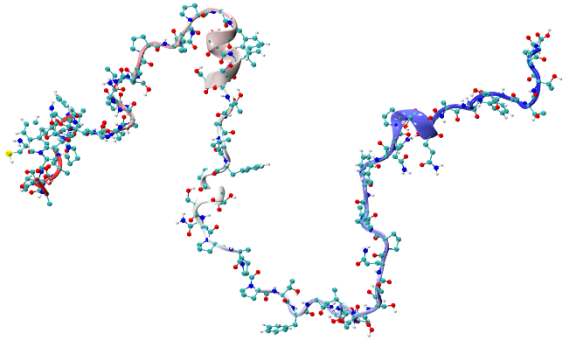
- Semi-ordered
  - Coiled coil
  - Many binding domains



- Transmembrane
  - Threaded through a membrane
  - Transmembrane regions, then internal and external segments



- Disordered / Low Complexity
  - Linker regions
  - Intrinsically disordered proteins



Key Residue Databases

PS00107 **PROTEIN_KINASE_ATP** *Protein kinases ATP-binding region signature* :

376 - 398: [confidence level: (0)] `LGSGLFGVVRlGkwraqyk..........VAIK`

PS00109 **PROTEIN_KINASE_TYR** *Tyrosine protein kinases specific active-site signature* :

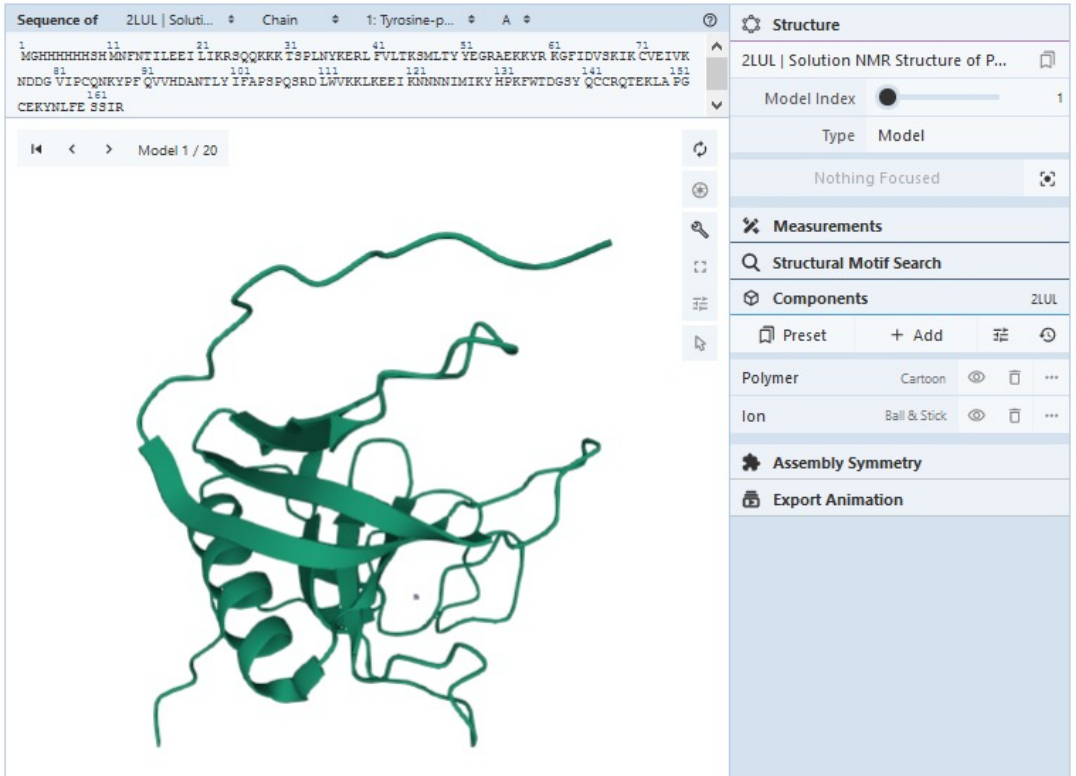
485 - 497: [confidence level: (0)] `FIHrDLAARNCLV`

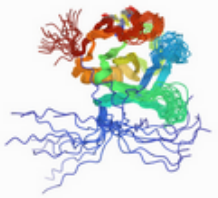
Predicted feature:

ACT_SITE	489	Proton acceptor	[condition: none]

Protein Structure Databases

RCSB PDB
PROTEIN DATA BANK





3D View

2LUL

Download File | View File

Solution NMR Structure of PH Domain of Tyrosine-protein kinase Tec from Homo sapiens, Northeast Structural Genomics Consortium (NESG) Target HR3504C

Liu, G., Xiao, R., Janjua, H., Hamilton, K., Shastry, R., Kohan, E., Acton, T.B., Everett, J.K., Lee, H., Pederson, K., Huang, Y.J., Montelione, G.T., Northeast Structural Genomics Consortium (NESG)

To be published

Released	2012-08-15
Method	SOLUTION NMR
Organisms	Homo sapiens
Macromolecule	Tyrosine-protein kinase Tec (protein)
Unique Ligands	ZN

3D View

1GL5

Download File | View File

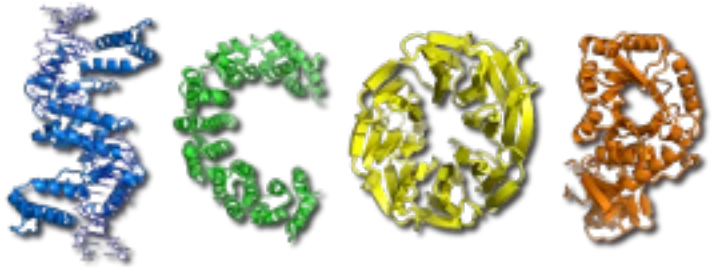
NMR structure of the SH3 domain from the Tec protein tyrosine kinase

Mulhern, T.D., Pursglove, S.E., Booker, G.W.

(2002) J Biol Chem **277**: 755-762

Released	2001-11-28
Method	SOLUTION NMR
Organisms	Mus musculus
Macromolecule	TYROSINE-PROTEIN KINASE TEC (protein)

Protein Structure Classification Databases

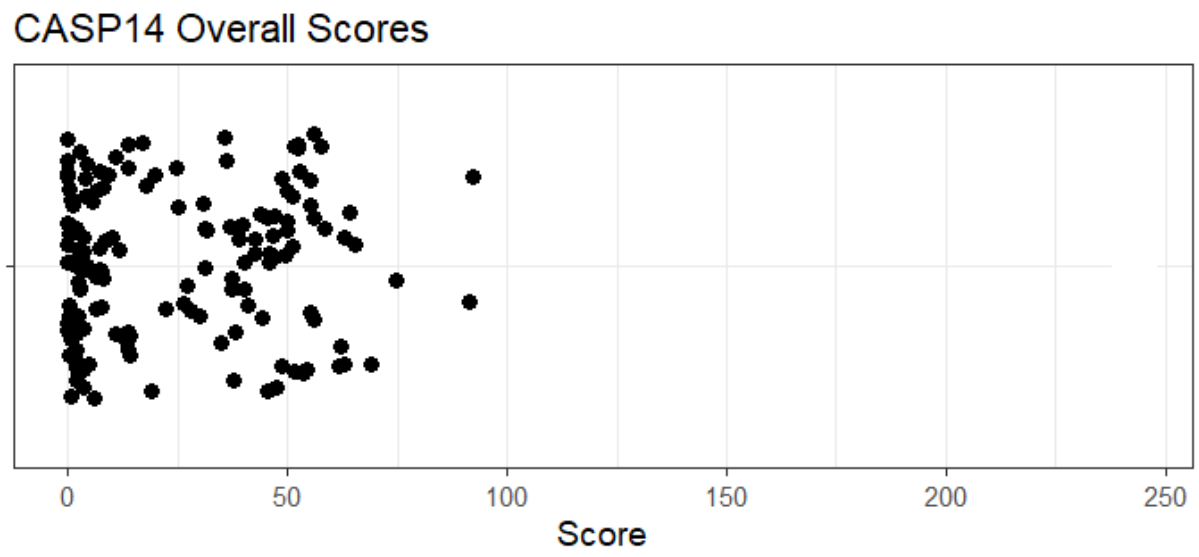


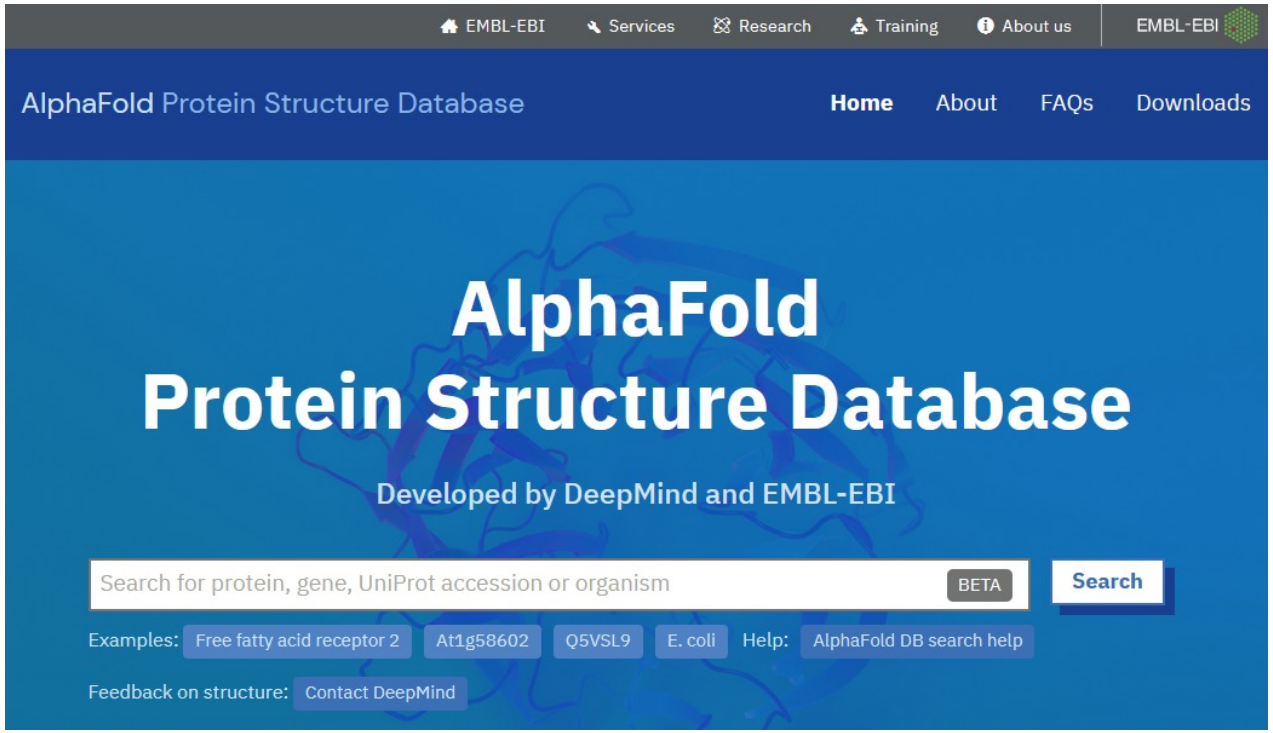
https://scop.mrc-lmb.cam.ac.uk/

https://www.cathdb.info/

Predicted Structure Database

Mar 2022, only 7914/22818 protein coding genes have an experimental 3D structure available





ARTIFICIAL INTELLIGENCE

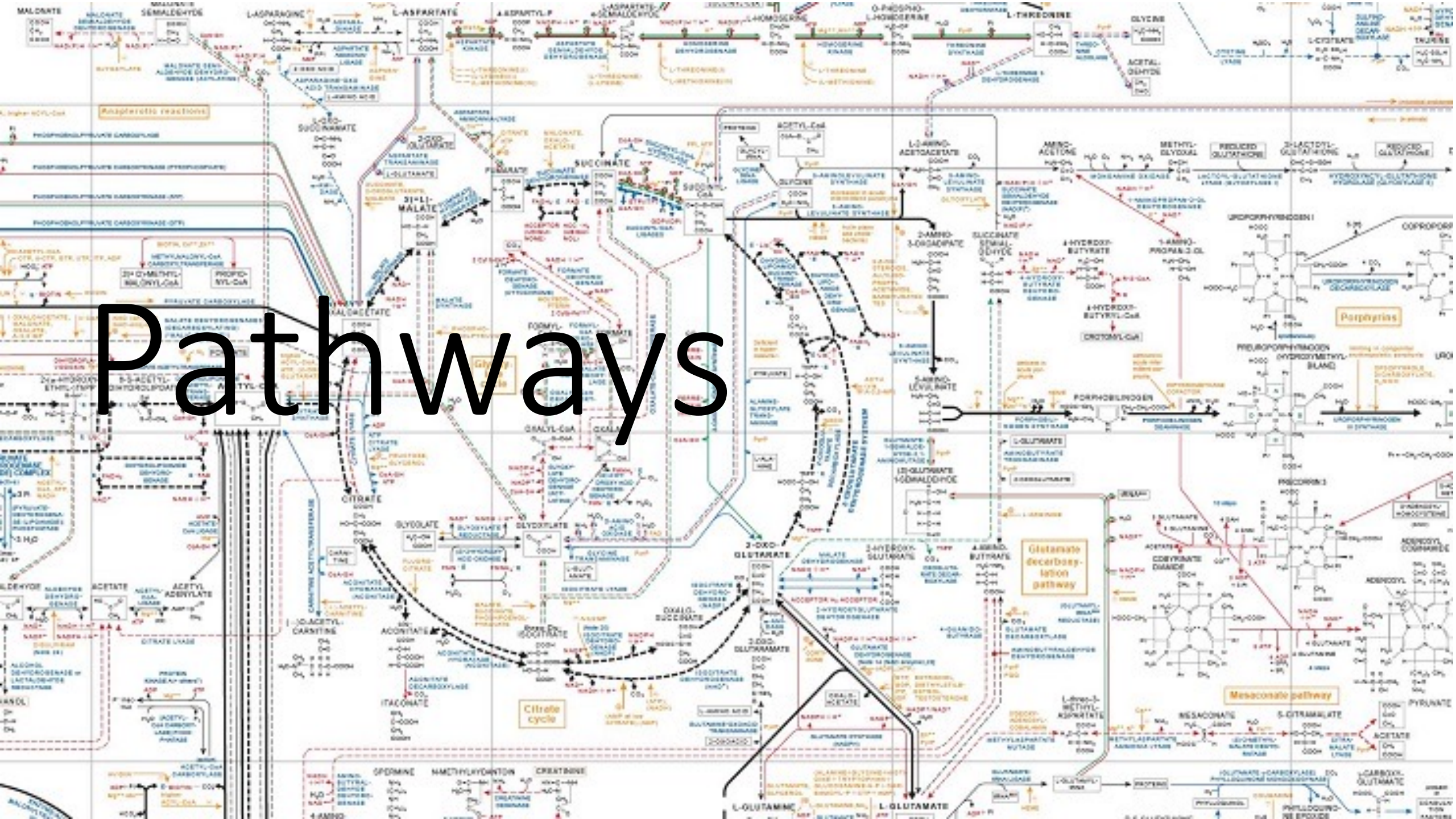
DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology

AlphaFold can predict the shape of proteins to within the width of an atom. The breakthrough will help scientists design drugs and understand disease.

By Will Douglas Heaven

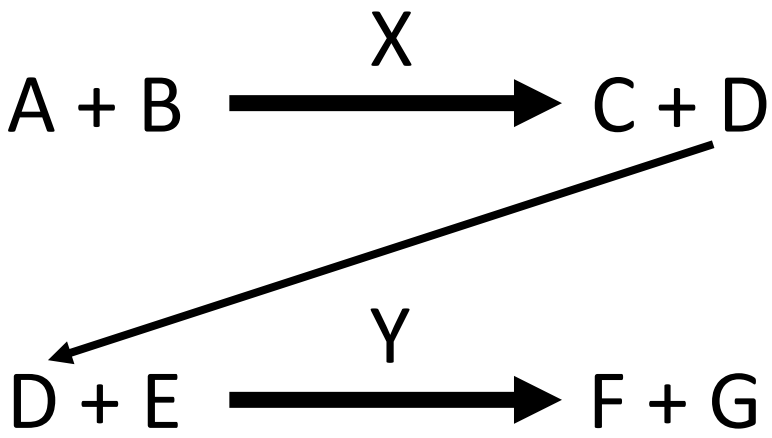
November 30, 2020

https://alphafold.ebi.ac.uk/


Pathways
Citrate cycle
Glutamate decarboxylation pathway
Mesaconate pathway
Porphyrins

Hierarchy of Reaction Annotations

- Components (Reactants / Products)
- Proteins (Enzymes)
- Reactions
- Pathways
- Processes

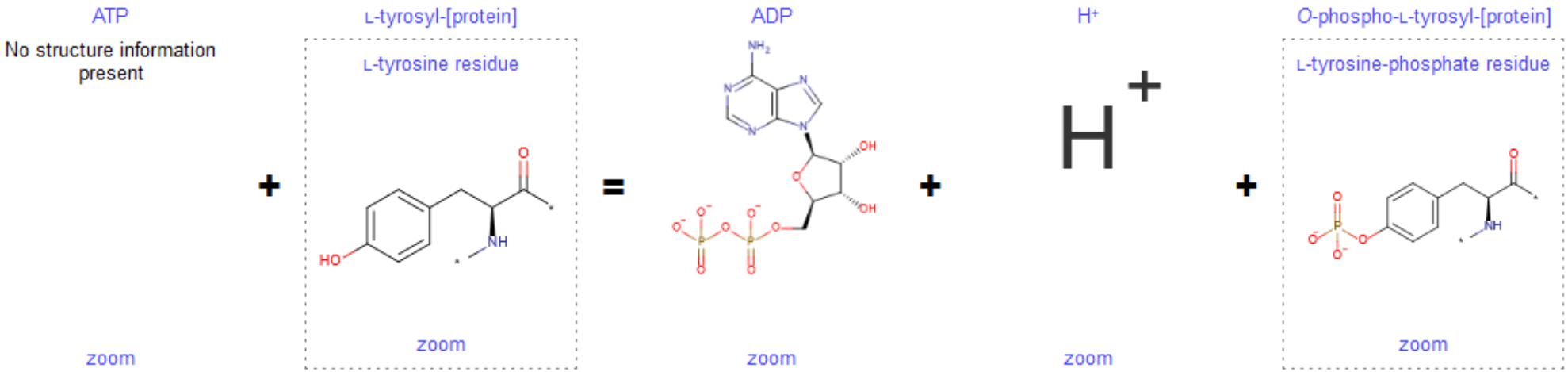


Reactions

Rhea

RHEA:10596

Enzymes	82,966 proteins (UniProtKB)
Enzyme classes	EC 2.7.10.1 Receptor protein-tyrosine kinase EC 2.7.10.2 Non-specific protein-tyrosine kinase EC 2.7.12.1 Dual-specificity kinase EC 2.7.12.2 Mitogen-activated protein kinase kinase



Enzyme Databases

- Enzymes are described by an Enzyme Commission (EC) number
  - EC `2.7.1.10` is phosphoglucokinase
  - Hierarchical structure

- Main Enzyme databases
  - Expasy Enzyme

Expasy

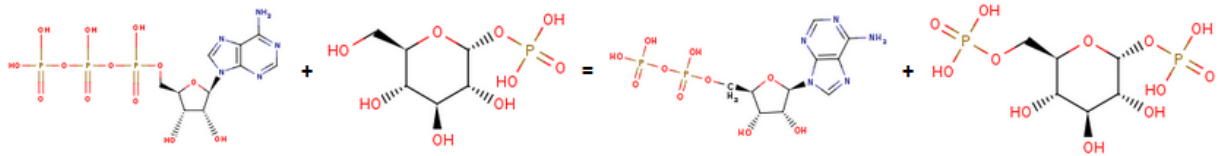
BRENDA
The Comprehensive Enzyme Information System

EC Tree

└ 2 Transferases
  └ 2.7 Transferring phosphorus-containing groups
    └ 2.7.1 Phosphotransferases with an alcohol group as acceptor
      └ EC 2.7.1.10 phosphoglucokinase

Reaction Schemes hide

ATP	+	alpha-D-glucose 1-phosphate	=	ADP	+	alpha-D-glucose 1,6-bisphosphate

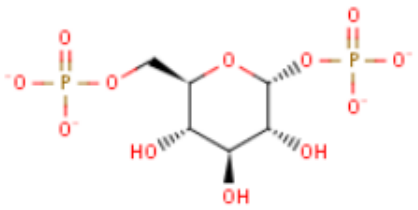


Chemical entities of biological interest

A database of "small" molecules with biological relevance
Natural or synthetic products which intervene in the processes of living organisms

CHEBI:58392 - α-D-glucose 1,6-bisphosphate(4−)

Main | ChEBI Ontology | Automatic Xrefs | Reactions | Pathways | Models



ChEBI Name	**α-D-glucose 1,6-bisphosphate(4−)**
ChEBI ID	**CHEBI:58392**
ChEBI ASCII Name	alpha-D-glucose 1,6-bisphosphate(4-)
Definition	A quadruply-charged organophosphate oxoanion arising from deprotonation of the phosphate OH groups of α-D-glucose 1,6-bisphosphate; major species at pH 7.3.
Stars	★★★ This entity has been manually annotated by the ChEBI Team.
Supplier Information	No supplier information found for this compound.
Download	Molfile XML SDF

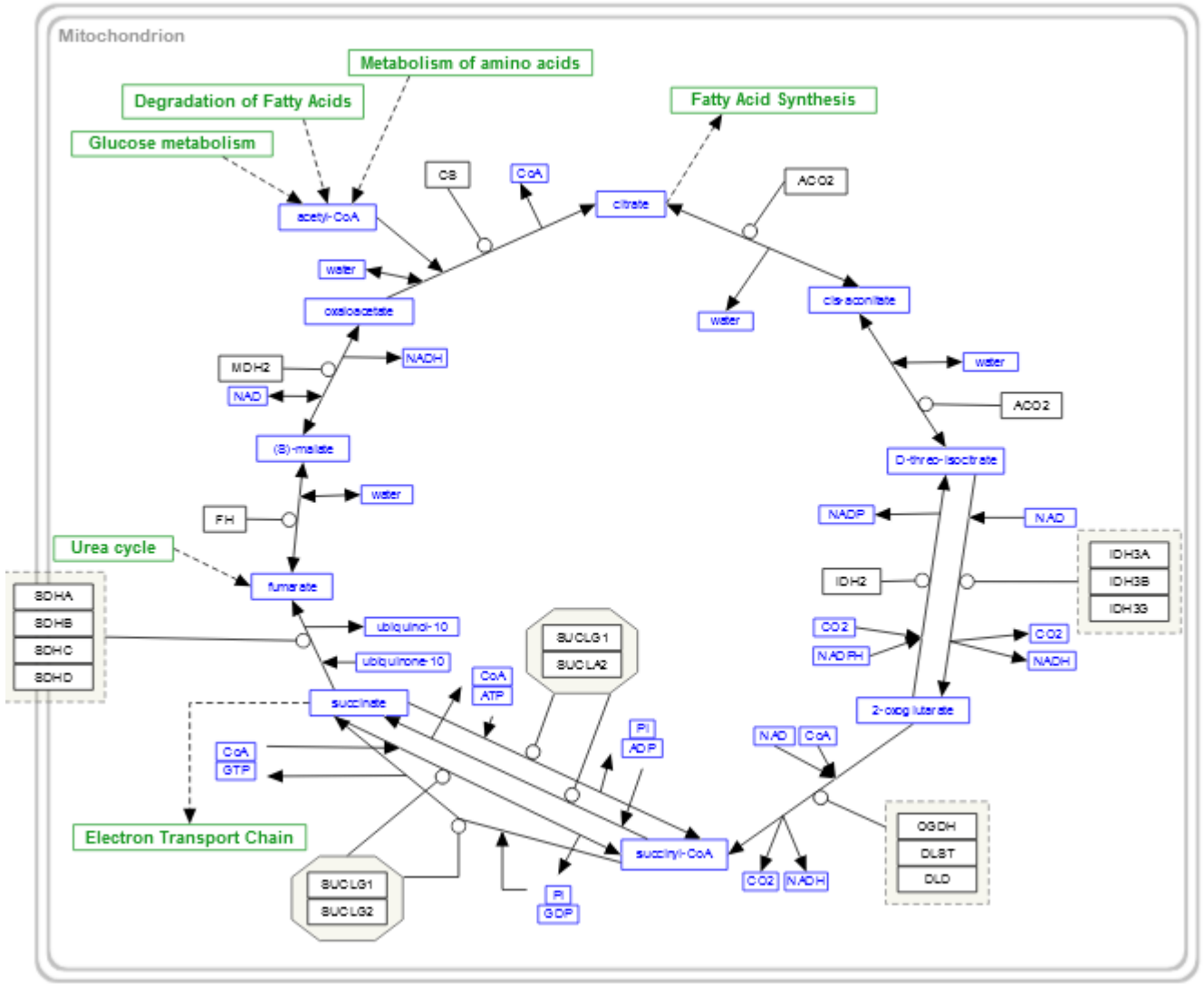
- Find compounds which contain this structure
- Find compounds which resemble this structure
- Take structure to the Advanced Search

Pathways

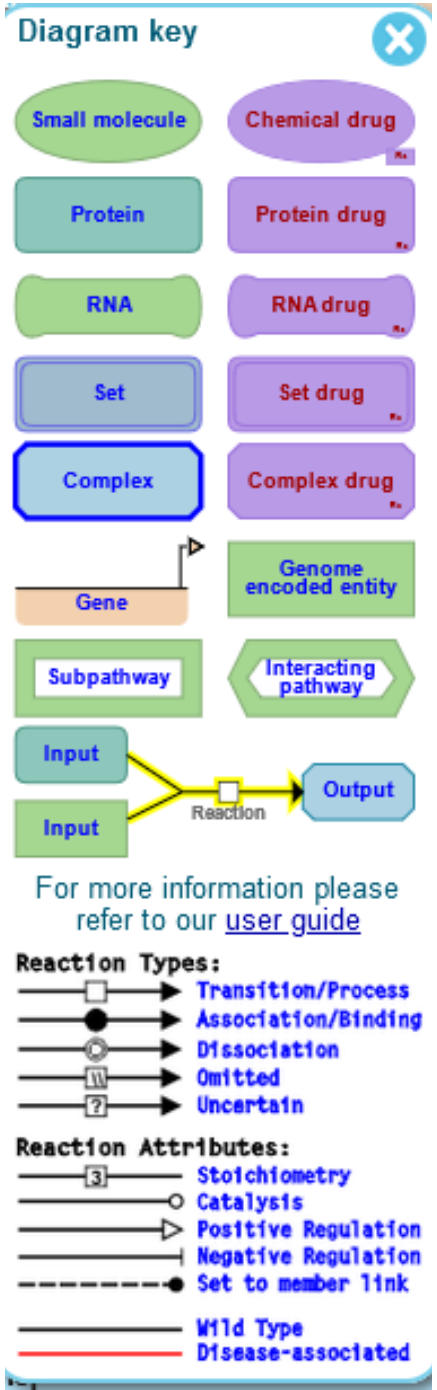
WikiPathways
Pathways for the People

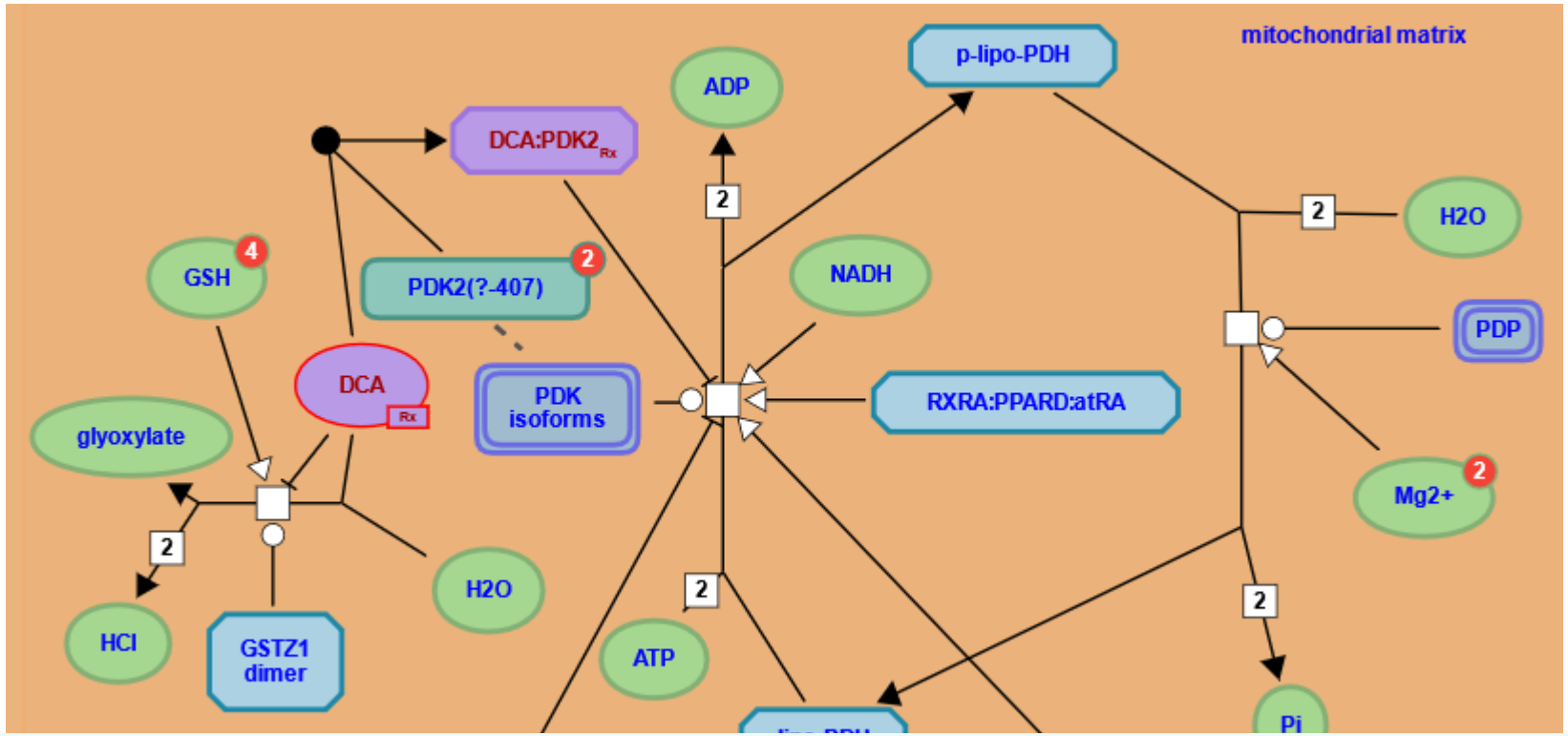
reactome

Kyoto Encyclopedia of Genes and Genomes
KEGG



Reactome





ME1:Mg2+ tetramer oxidatively decarboxylates MAL to PYR | Id: R-HSA-9012036.1 | Species: Homo sapiens

Summation

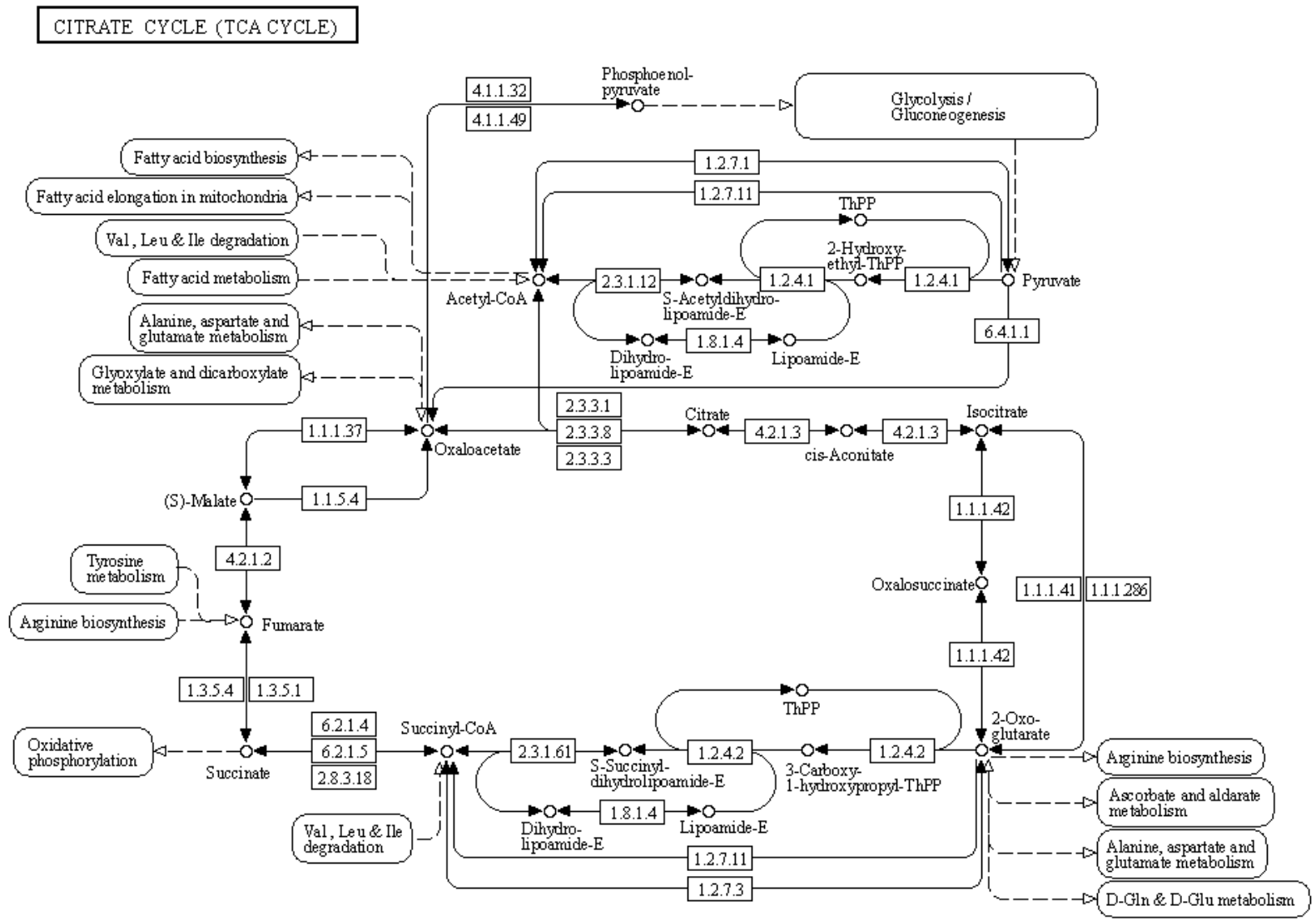
One hallmark of cancer is altered cellular metabolism. Malic enzymes (MEs) are a family of homotetrameric enzymes that catalyse the reversible oxidative decarboxylation of L-malate to pyruvate, with a simultaneous reduction of NAD(P)+ to NAD(P)H. As MEs generate NADPH and NADH, they may play roles in energy production and reductive biosynthesis. Humans possess three ME isoforms; ME1 is cytosolic and utilises NADP+, ME3 is mitochondrial and can utilise NADP+ and ME2 is mitochondrial and can utililse either NAD+ or NADP+ (Chang & Tong 2003, Murugan & Hung 2012).

NADP-dependent malic enzyme (ME1, aka c-NADP-ME) is a cytosolic enzyme that oxidatively decarboxylates (s)-malate (MAL) to pyruvate (PYR) and CO2 using NADP+ as cofactor (Zelewski & Swierczynski 1991). ME1 exists as a dimer of dimers (Murugan & Hung 2012, Hsieh et al. 2014) and a divalent metal such as Mg2+ is essential for catalysis (Chang & Tong 2003).

▶ **Background literature references...**

KEGG databases

Category	Entry point
Systems information	KEGG PATHWAY
	KEGG BRITE
	KEGG MODULE KEGG RModule
Genomic information	KEGG ORTHOLOGY KEGG Annotation
	KEGG GENES KEGG SeqData
	KEGG GENOME KEGG Virus
Chemical information	KEGG COMPOUND
	KEGG GLYCAN
	KEGG REACTION
	KEGG Enzyme
Health information	KEGG NETWORK
	KEGG DISEASE
	KEGG DRUG



Functional Gene Sets

GENEONTOLOGY
Unifying Biology

Molecular Function

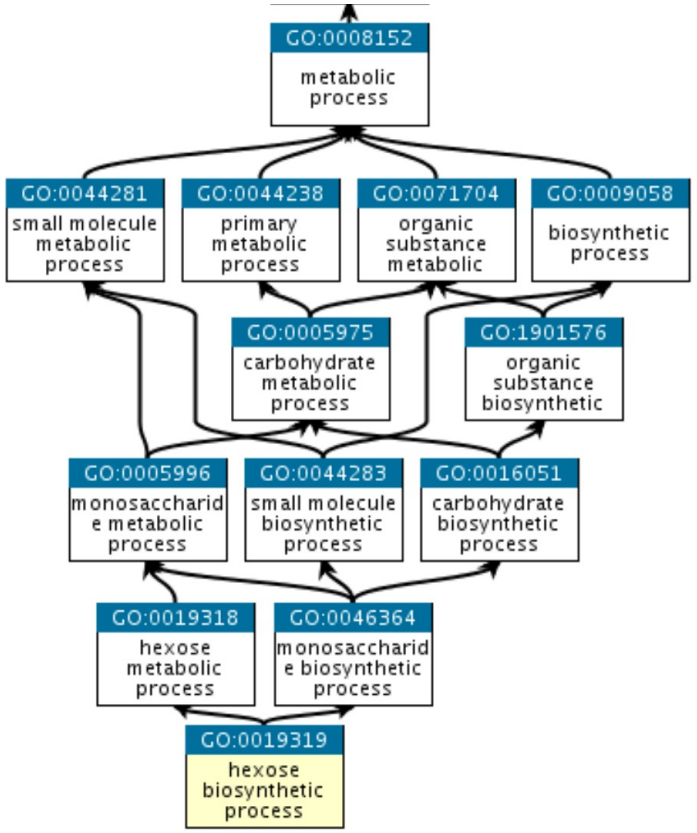
Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as “catalysis” or “transport”. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*. To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word “activity” (a *protein kinase* would have the GO molecular function *protein kinase activity*).

Cellular Component

The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.

Biological Process

The larger processes, or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport* . Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

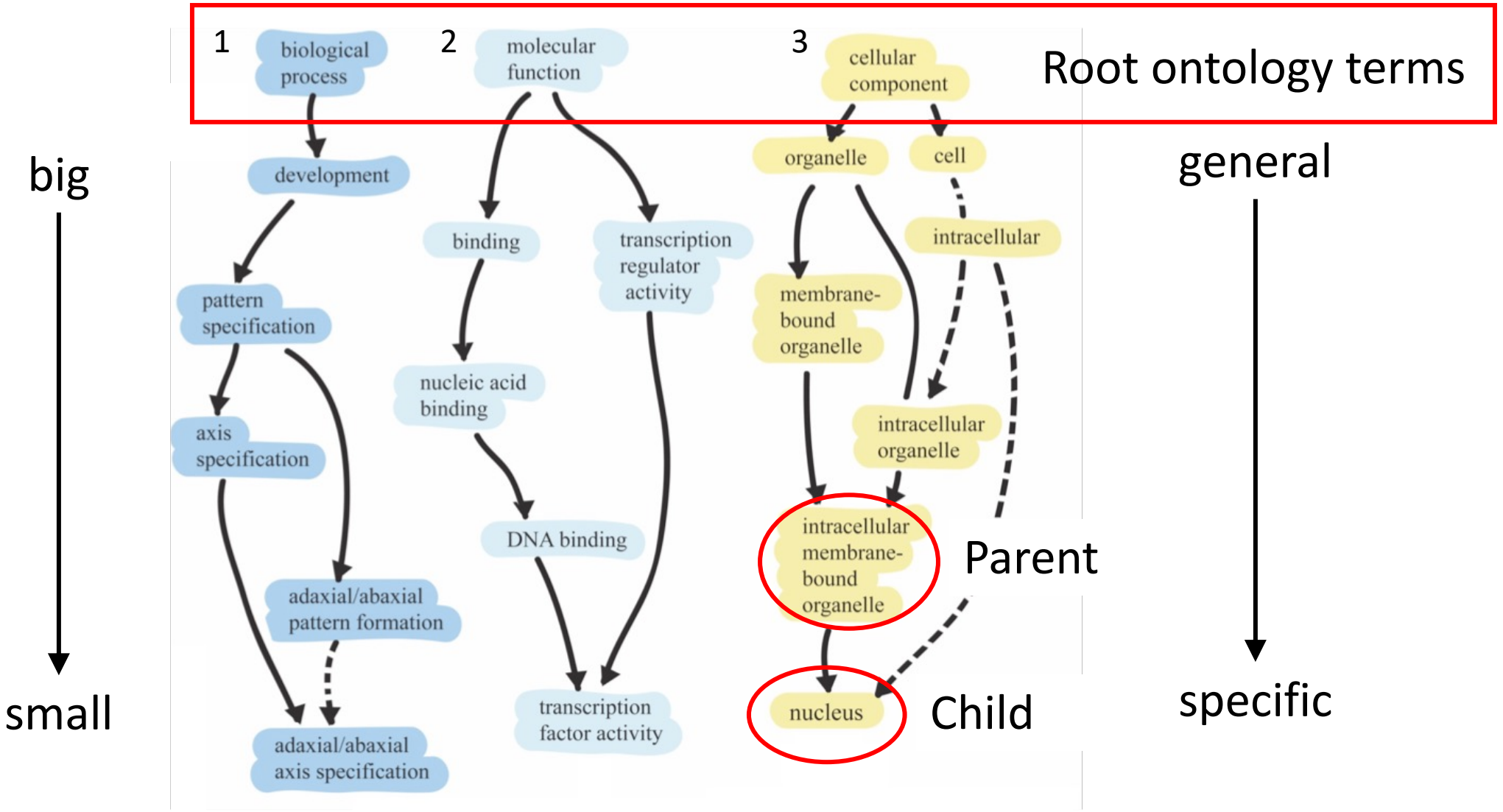


Term Information

Accession GO:0019319
Name hexose biosynthetic process
Ontology biological_process
Synonyms hexose anabolism, hexose biosynthesis, hexose formation, hexose synthesis
Alternate IDs None
Definition The chemical reactions and pathways resulting in the formation of hexose, any monosaccharide with a chain of six carbon atoms in the molecule. *Source:* ISBN:0198506732

Data health

Gene/product	Gene/product name	Organism
Sds	serine dehydratase	Mus musculus
G6pc	glucose-6-phosphatase, catalytic	Mus musculus
Gnpda1	glucosamine-6-phosphate deaminase 1	Mus musculus
Nr3c1	nuclear receptor subfamily 3, group C, member 1	Mus musculus
Gpt	glutamic pyruvic transaminase, soluble	Mus musculus
Ranbp2	RAN binding protein 2	Mus musculus
Ptpn2	protein tyrosine phosphatase, non-receptor type 2	Mus musculus
Stk11	serine/threonine kinase 11	Mus musculus
Gm10768	predicted gene 10768	Mus musculus
Fbp1	fructose bisphosphatase 1	Mus musculus


1
biological process
development
pattern specification
axis specification
adaxial/abaxial pattern formation
adaxial/abaxial axis specification
2
molecular function
binding
nucleic acid binding
DNA binding
transcription regulator activity
transcription factor activity
3
cellular component
organelle
cell
intracellular
membrane-bound organelle
intracellular organelle
intracellular membrane-bound organelle
nucleus
Root ontology terms
big
small
general
specific
Parent
Child

Genes assigned to ontology terms

Nanog homeobox [Source:HGNC Symbol;Acc:HGNC:20857]

- Cellular Component
  - GO:0005634 nucleus
  - GO:0005654 nucleoplasm
  - GO:0005730 nucleolus
- Molecular Function
  - GO:0003677 DNA binding
  - GO:0003700 transcription factor activity, sequence-specific DNA binding
  - GO:0003714 transcription corepressor activity
  - GO:0005515 protein binding
  - GO:0043565 sequence-specific DNA binding
- Biological Process
  - GO:0001714 endodermal cell fate specification
  - GO:0006351 transcription, DNA-templated
  - GO:0006355 regulation of transcription, DNA-templated
  - GO:0007275 multicellular organism development
  - GO:0008283 cell proliferation
  - GO:0019827 stem cell population maintenance
  - GO:0030154 cell differentiation
  - GO:0035019 somatic stem cell population maintenance
  - GO:0045595 regulation of cell differentiation
  - GO:0045944 positive regulation of transcription from RNA polymerase II promoter
  - GO:1903507 negative regulation of nucleic acid-templated transcription

Regulation and Interactions

- The regulation of genes is as important as their structure or function
- Several sources of useful information
  - Regulatory binding proteins, mostly transcription factors
  - Interactions with other proteins to form complexes
  - Composition of known complexes

Transcription Factor Information

JASPAR 2020

Detailed information of matrix profile **MA0605.2**

Home > Matrix > MA0605.2

Profile summary

Add

Name:	ATF3
Matrix ID:	MA0605.2
Class:	Basic leucine zipper factors (bZIP)
Family:	Jun-related factors
Collection:	CORE
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	HT-SELEX
Validation:	12815047
Uniprot ID:	P18847
Source:	28473536
Comment:	

Sequence logo

Download SVG

Frequency matrix

JASPAR | TRANSFAC | MEME | RAW PFM | Reverse comp.

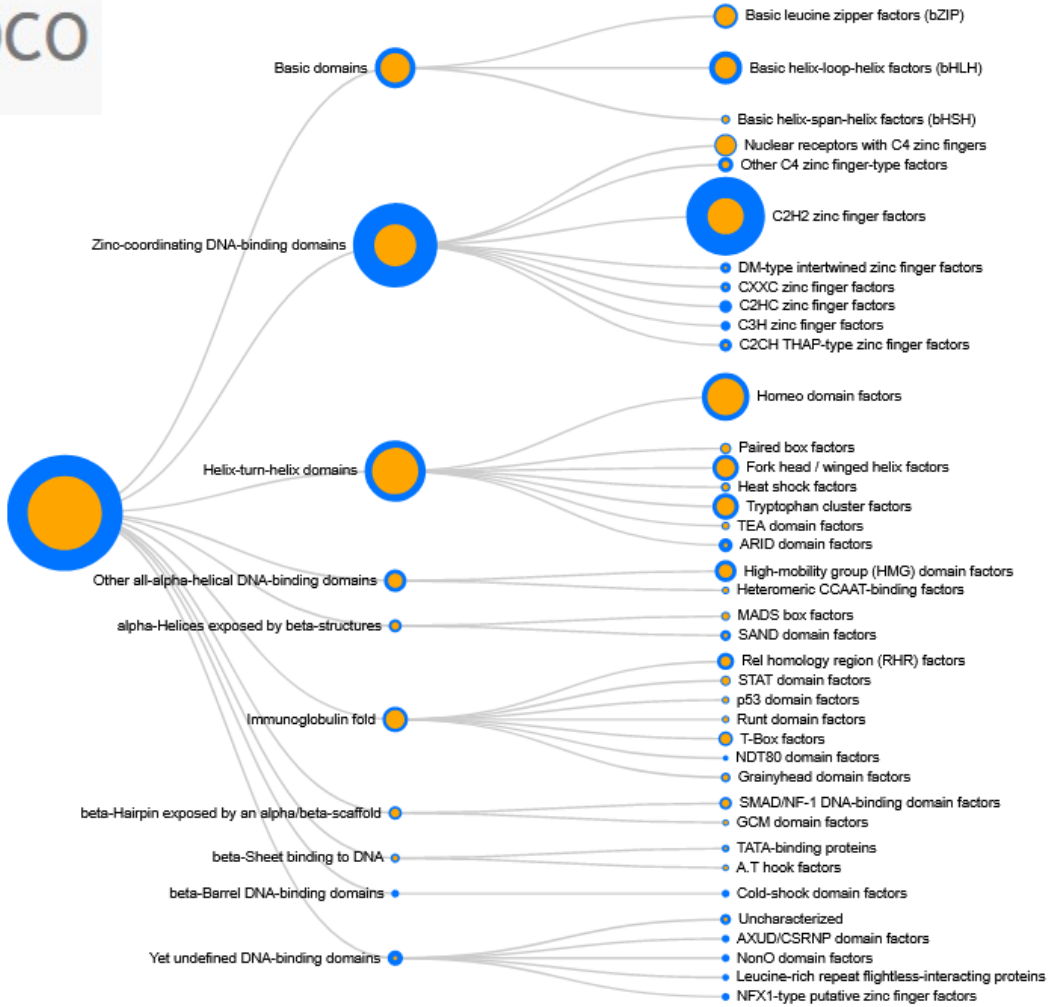
A [	8505	24741	0	0	40546	0	891	0	1520	40546	0	6919	]
C [	8220	1354	0	0	0	40546	0	0	40546	0	15737	16820	]
G [	16894	15805	1	40546	0	0	40546	0	0	0	1094	8242	]
T [	6926	0	40546	1366	0	556	0	40546	0	0	24808	8564	]

Transcription Factor Information

HOCOMOCO

Model info

Transcription factor	BCL6B (GeneCards)
Model	BCL6B_HUMAN.H11MO.0.D
Model type	Mononucleotide PWM
LOGO	
LOGO (reverse complement)	
Data source	HT-SELEX
Model release	HOCOMOCOv10
Model length	11
Quality	D
Motif rank	0
Consensus	nYGCTTTCTAG



Genes Regulated by a Transcription Factor

- Difficult to predict - lots of false positives
  - Swiss Regulon
  - 

  - ReMap2022
  - GTRD
    Gene Transcription Regulation Database

BRCA2

Entrez	Description	Chrom.	Strand	Promoter (Start - Stop)	TSS
675	breast cancer 2, early onset	chr13	+	32884616 - 32890116	32889616

Transcription Factor Binding Sites

Download all TFBS in the BRCA2 promoter

Show 10 entries Search:

Motif	Source	Strand	Start	Stop	PValue	Match Sequence	Overlap w/ Footprints
Pax4_MA0068.1	JASPAR	+	32888990	32889019	0.0E+00	AAAAAAAAAAGCAAAAGATACTACCAAGCC	30
V_GC_01_M00255	TRANSFAC	-	32889167	32889180	0.0E+00	AGTGGGCGGGGCTG	14
V_LDSPOLYA_B_M00317	TRANSFAC	-	32889437	32889452	0.0E+00	AGTGTGTGTTCTCTTC	16
V_SOX2_Q6_M01272	TRANSFAC	-	32889284	32889299	0.0E+00	AATACCTTTGTTCTGA	16
V_SP1_Q4_01_M00932	TRANSFAC	-	32889989	32890001	0.0E+00	AAGGGGCGGGGCT	13
V_STAT5A_01_M00457	TRANSFAC	+	32890101	32890115	0.0E+00	AATTTCTTGGAAACA	15
V_STAT5A_Q6_M01890	TRANSFAC	+	32890100	32890112	0.0E+00	AAATTTCTTGGAA	13
V_STAT5B_01_M00459	TRANSFAC	+	32890101	32890115	0.0E+00	AATTTCTTGGAAACA	15
V_STAT_Q6_M00777	TRANSFAC	+	32890099	32890111	0.0E+00	GAAATTTCTTGGA	13
SP1_C2H2_DBD_monomeric_11_1	SELEX	+	32889169	32889179	1.0E-05	GCCCCGCCCAC	11

Showing 1 to 10 of 196 entries Previous Next

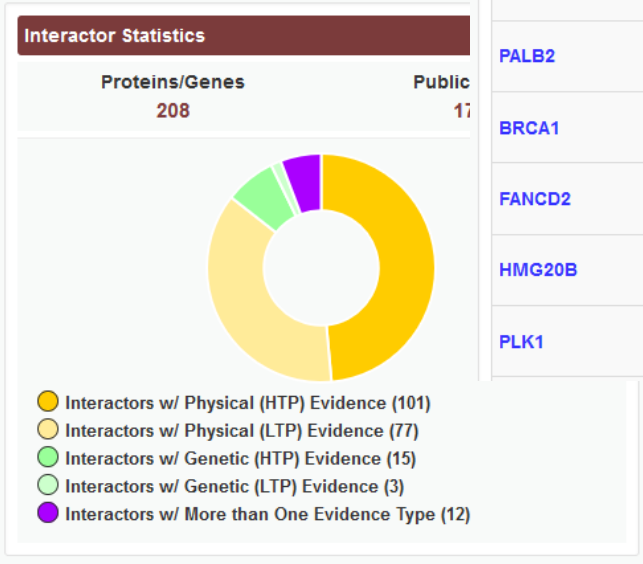
Gene Interactions

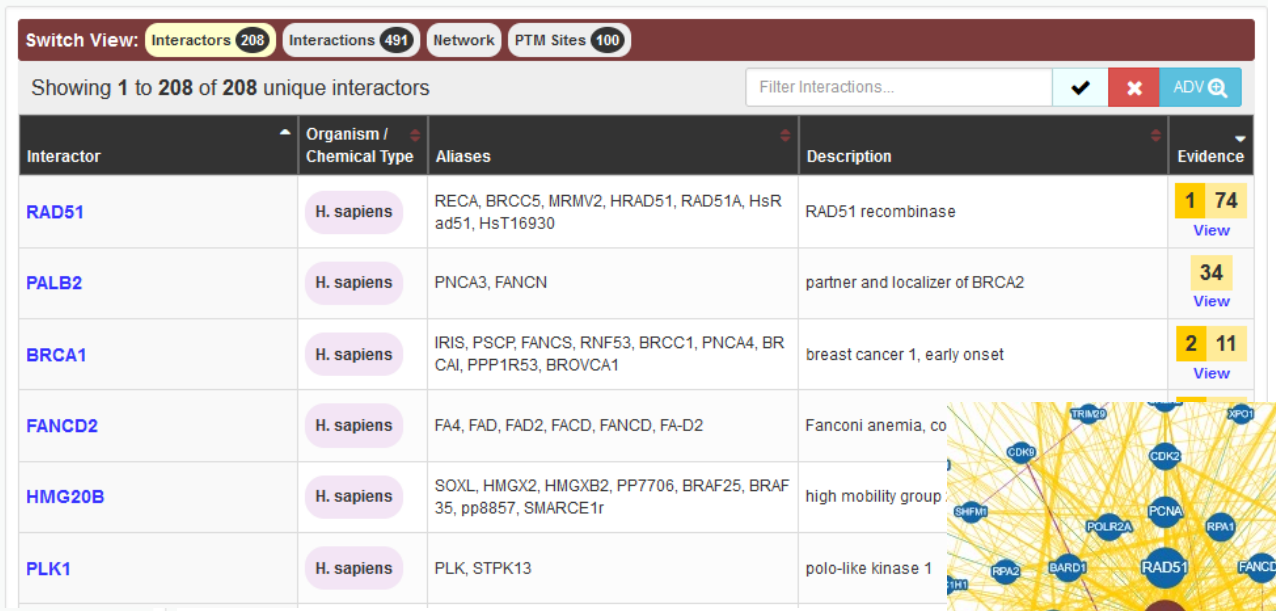
- Many genes form stable or transitory interactions with others
- Knowing the genes that interact helps understand biology

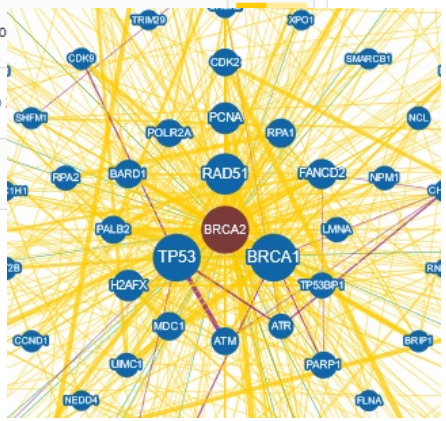
BioGRID 4.4

IntAct

STRING







Types of Interaction

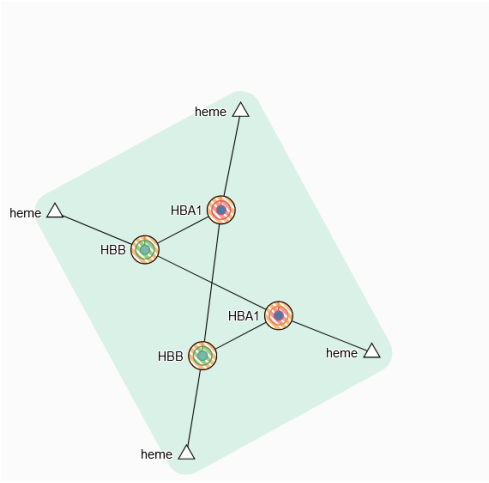
- Physical
  - Two proteins directly interact, either stably or transiently

- Genetic
  - One gene influences another, normally after modification
    - Co-expression
    - Knockout compensation

Complex Prediction

- Many proteins interact with several others, but at different times
- Complexes suggest that multiple proteins directly associate
  - Can't always be clearly predicted from pairwise interactions
  - Other experimental methods are required

Complex Portal



Legend	Description	Stoichiometry
●	**protein - HBA1 (unspecified role)** P69905 Hemoglobin subunit alpha	2
●	**protein - HBB (unspecified role)** P68871 Hemoglobin subunit beta	2
△	**small molecule - heme (cofactor)** CHEBI:30413 heme	4

Course Structure

- Central Dogma Data Sources
  - Genomes and Annotations
  - Protein Domains and Structures
  - Reactions, Pathways and Interactions
- Experimental Techniques, Datatypes and Repositories
  - Sequencing and Variants
  - Proteomics and Metabolites
- Omics
  - General concepts
  - Analysis approaches

Sequence Variants

Reference	GATCTTA**G**CTGA
Variant	GATCTTA**C**CTGA

- Germline variants
  - Happen in sperm or eggs
  - Completely inherited into the next generation
  - Can cause genetic disease
- Somatic variants
  - Happen in other tissues
  - Partially penetrant
  - Common cause of cancer

Types of Variant

Ref GATCTTA**G**CTGA

Var GATCTTA**C**CTGA

Substitution
Single Nucleotide Polymorphism
SNP

Ref GATCTTA**G..**CTGA

Var GATCTTA**CAA**CTGA

Ref GATCTTA**GCT**GA

Var GATCTTA**C..**GA

Insertion

Deletion

InDel

Functional Variant Consequences

- Within Coding Region
  - Silent (codon changes, but same translation)
  - Missense (change translation from one amino acid to another)
  - Nonsense (change translation from one amino acid to STOP)
  - Frameshift (InDel changing the translation frame)

- Outside CDS
  - Breaks or adds splice junction
  - Changes functional binding site

Structural variants

- Chromosomal copy number change
  - Gain or loss of a chromosome
  - Leads to serious genetic disease

- Segmental Deletion / Duplication
  - Large parts of chromosomes deleted, duplicated, inverted, translocated
  - 1kb to 3Mbp
  - Affects many genes, can lead to gene fusions

Databases of Variants

- Common genomic variants
  - Measured across a large population
  - Shows natural variation
  - Not necessarily linked to disease
  - Used for studying populations and families

- Functional variants
  - Variants with an associated phenotype
  - Often disease related but can be any measurable phenotype

Variant Databases

- Single Variants
  - dbSNP (https://www.ncbi.nlm.nih.gov/snp/)
    - Full reference for any reported SNPs, mix of functional and non-functional
  - HGMD (http://www.hgmd.cf.ac.uk)
    - Human genetic disease focussed database
  - COSMIC (https://cancer.sanger.ac.uk/cosmic)
    - Mutations observed in Cancer
    - Also has details of mutations in immortalised cell lines

- Larger Regions
  - dbVar (https://www.ncbi.nlm.nih.gov/dbvar/)
    - Counterpart to dbSNP for larger variants
  - ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)
    - Larger variants with clinical relevance
  - OMIM (https://www.ncbi.nlm.nih.gov/omim)
    - A more wide ranging collection of the phenotypic variation linked to genes

Variant Terminology

- Minor Allele Frequency (MAF)
  - How prevalent the variant is in the population
- Impact scores (SIFT / PolyPhen etc)
  - A quantitative value assessing the likely biological impact of a variant

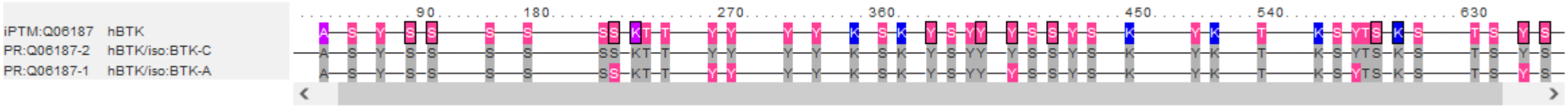
Post translational Modifications

- Many proteins are modified after they have been translated
  - Phosphorylation
  - Glycosylation
  - Ubiquitination
  - Nitrosylation
  - Methylation
  - Acetylation
  - Lipidation
  - Proteolysis

iPTMnet

Both PTMs observed on a protein and proteins modified by a query gene.

Site	PTM Type	PTM Enzyme	Score
All	All		2 selected
S21	Phosphorylation		★★★☆
S115	Phosphorylation		★★★☆
S180	Phosphorylation	P05771 (PRKCB)	★★★★
Y223	Phosphorylation	Q06187 (BTK) , Q08881 (ITK) , P07948 (LYN) , P00519 (ABL1) , A0A173G4P4 (Abl fusion) , P42680 (TEC)	★★★★
Y551	Phosphorylation	P07948 (LYN) , Q06187 (BTK) , P12931 (SRC) , P43405 (SYK)	★★★★



Combined Gene/Protein Centric Datasources

https://www.uniprot.org/

https://www.genecards.org

https://www.ncbi.nlm.nih.gov/gene/

https://www.wikigenes.org/

UniProt

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequences (1+)
- Similar proteins
- Cross-references
- Entry information
- Miscellaneous

wikigenes collaborative publishing

Disease Relevance
High Impact publication summaries
Biological Context
Anatomical Context
Chemical Compound Associations
Physical Interactions
Enzymatic Interactions
Regulatory relationships
Analytical, diagnostic and therapeutic context
References

NCBI Gene

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Expression
- Bibliography
- Variation
- Pathways from PubChem
- Interactions
- General gene information
  - Markers, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

GeneCards® THE HUMAN GENE DATABASE

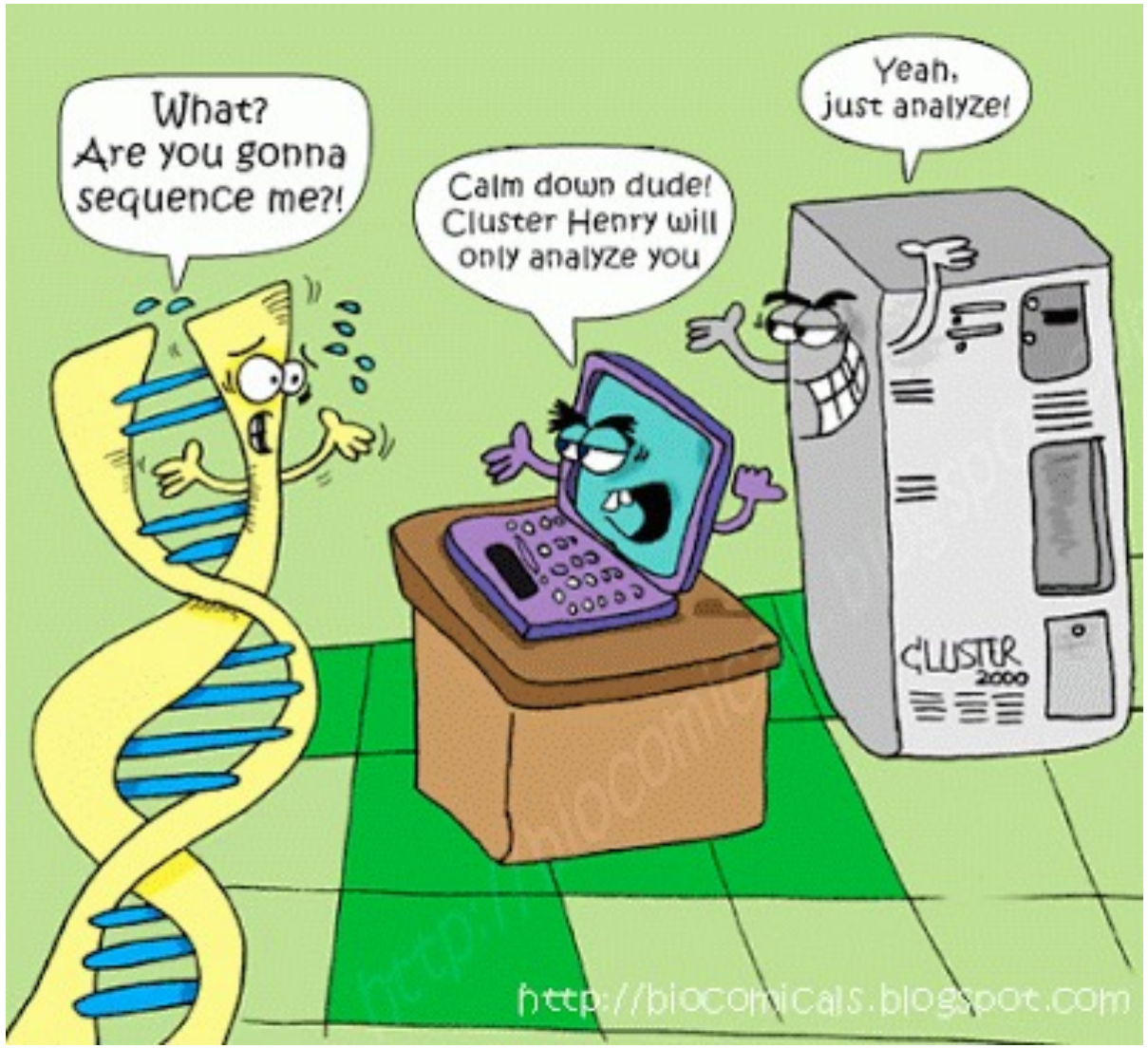
Jump to section									
	Aliases	Disorders	Domains	Drugs	Expression	Function	Genomics	Localization	Orthologs
	Paralogs	Pathways	Products	Proteins	Publications	Sources	Summaries	Transcripts	Variants
Research Products	Antibodies	Assays	Proteins	Inhib. RNA	CRISPR	Exp. Assays	miRNA	Drugs	Animal Models
	Cell Lines	Clones	Primers	Genotyping					

Big Data Generation

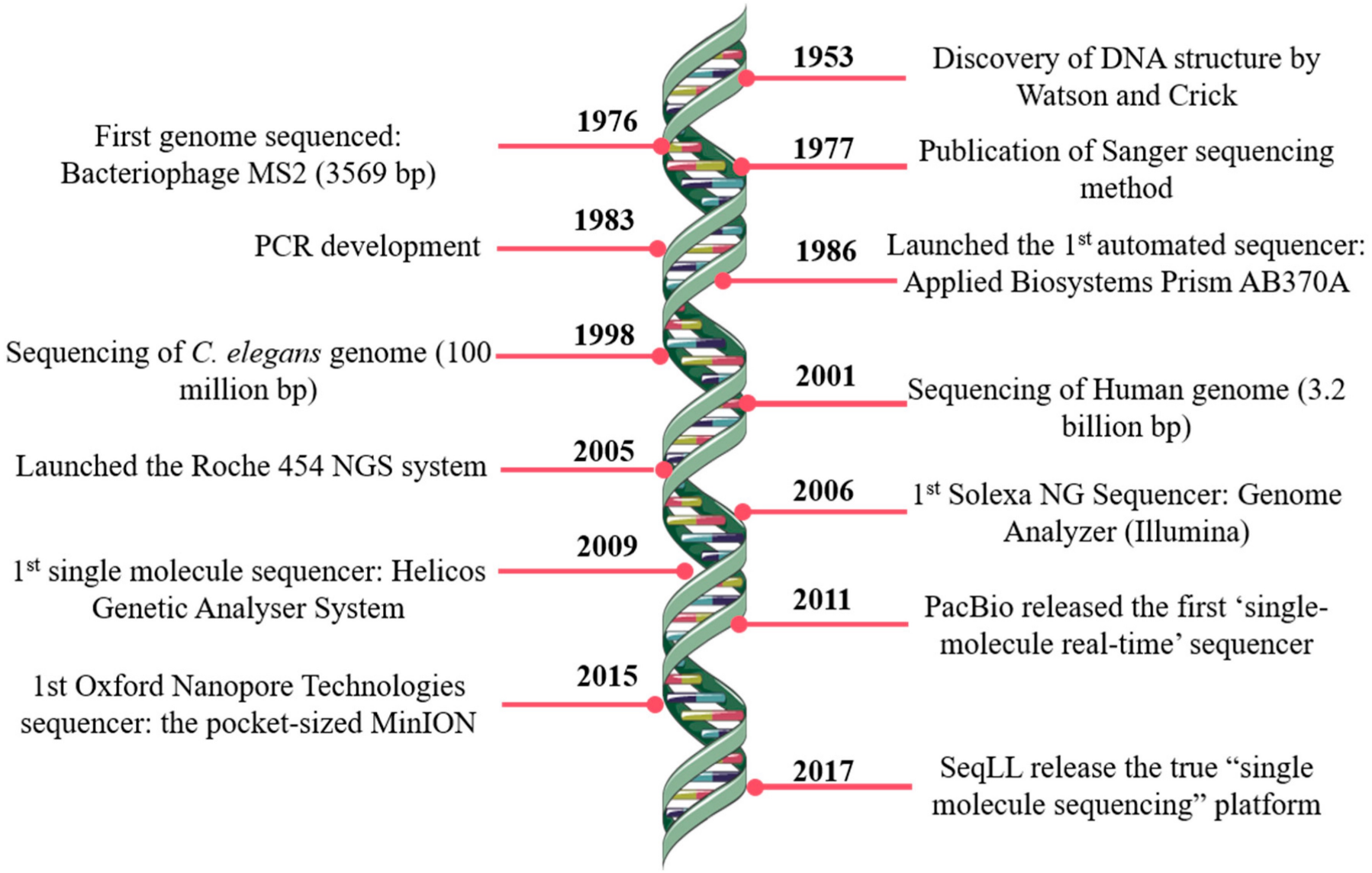
- High throughput sequencing
  - Genomics, Transcriptomics, Epigenetics
- Multi-channel Flow Cytometry
  - Cell surface proteomics
- Mass Spectrometry
  - Proteomics, Metabolomics
- Biological Imaging
  - Cell / Tissue structure, Proteomics, Metabolomics

Sequencing

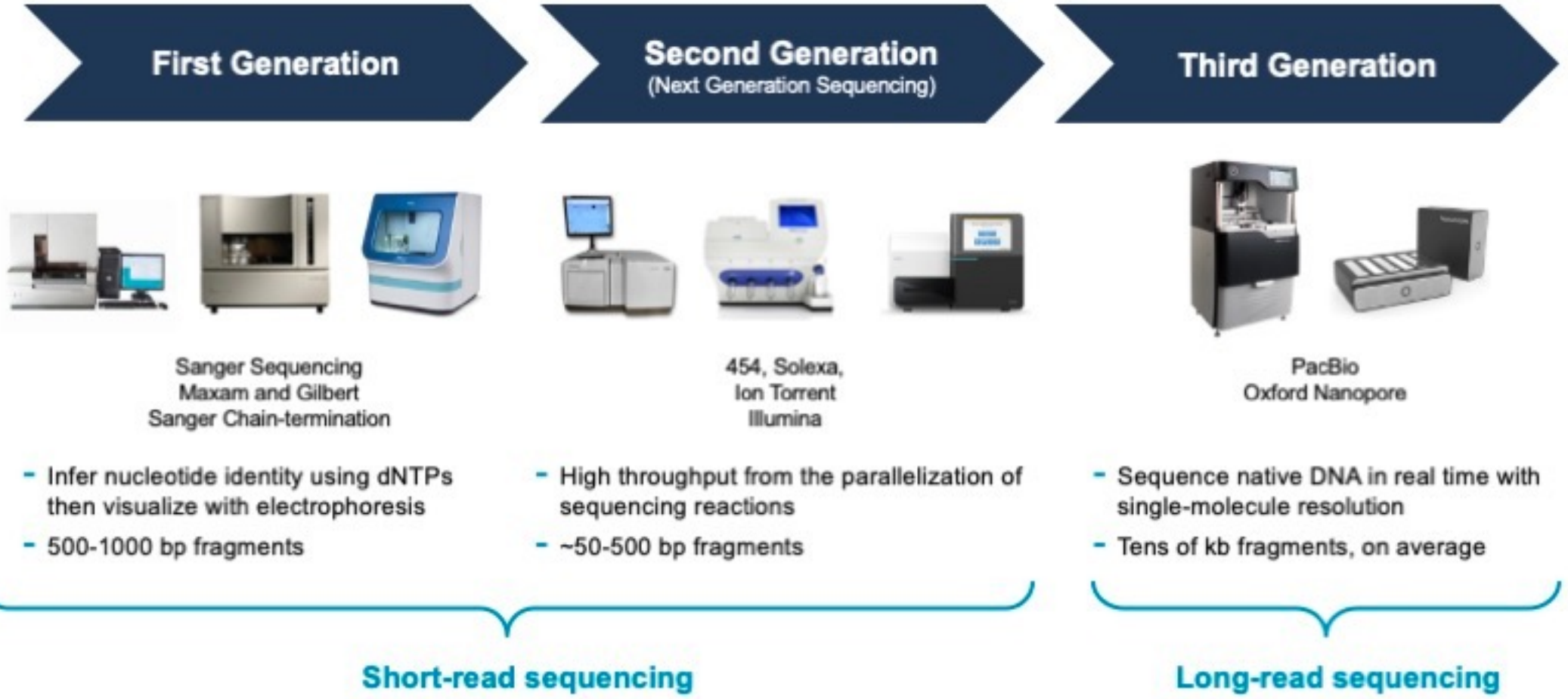
DNA sequencing is a process of determining the order of nucleotides within a DNA molecule.



History of DNA sequencing

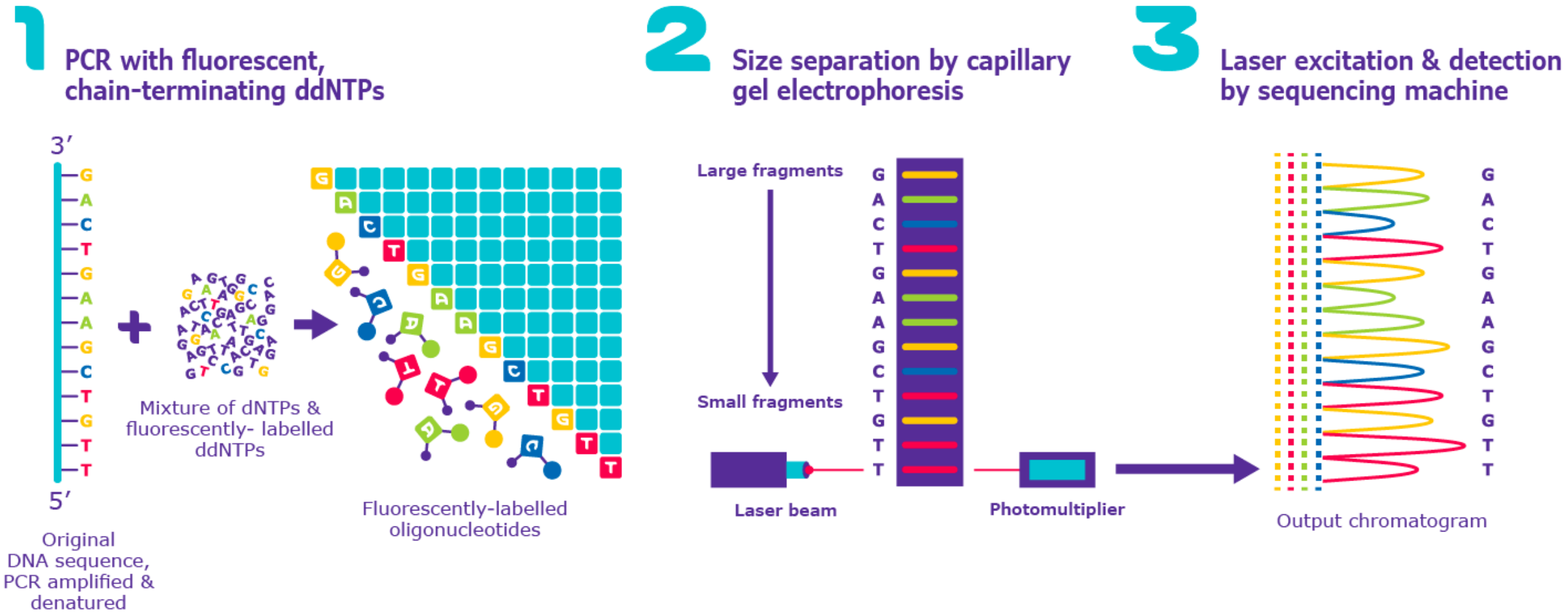


DNA sequencing generations

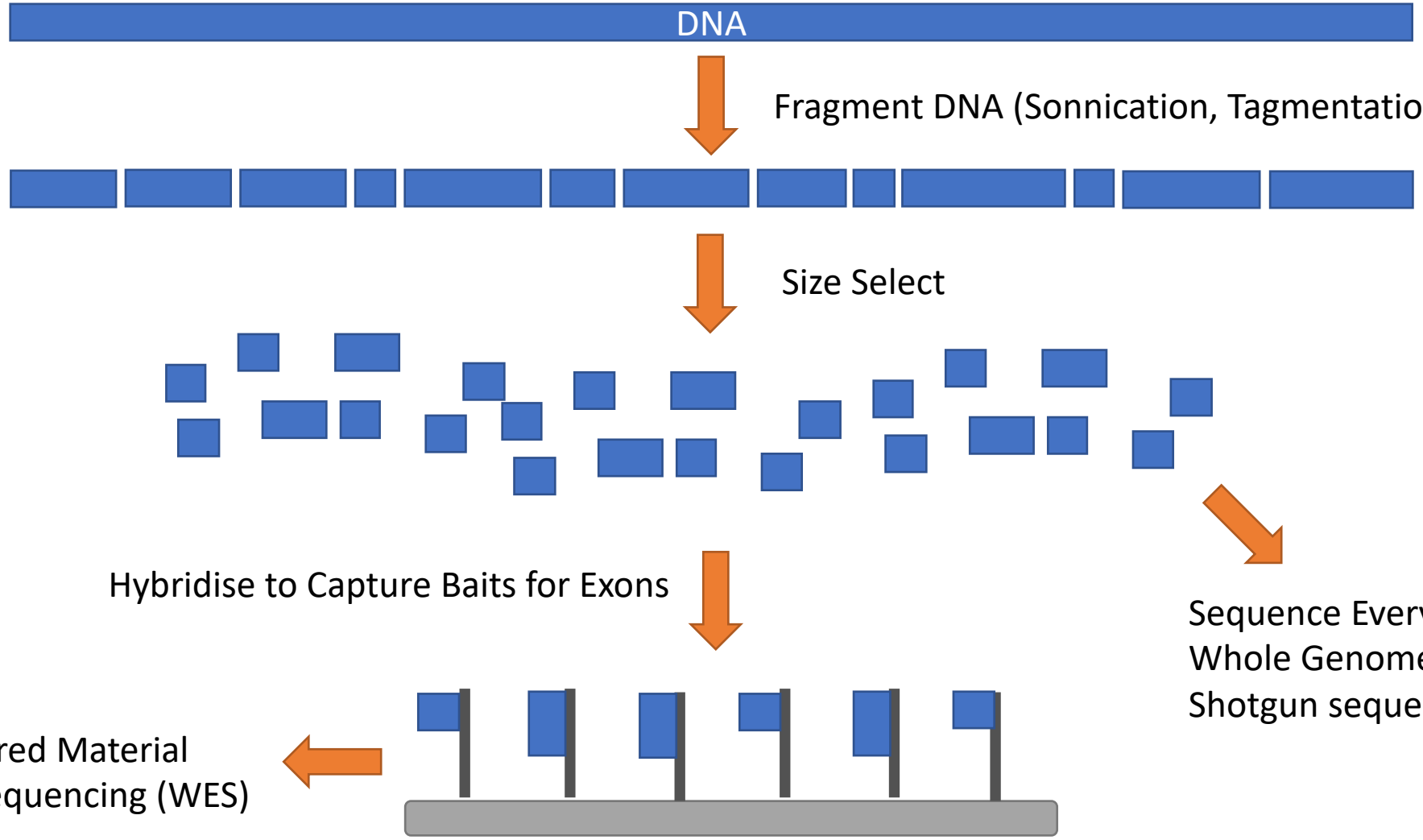


Sanger sequencing

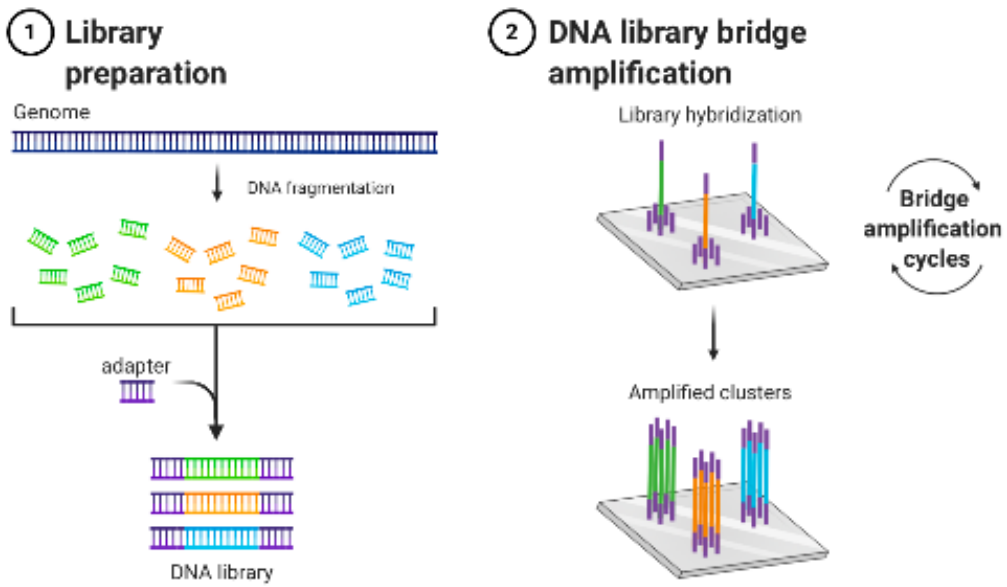
Sanger sequencing is a DNA sequencing method in which target DNA is denatured and annealed to an oligonucleotide primer, which is then extended by DNA polymerase using a mixture of deoxynucleotide triphosphates (normal dNTPs) and chain terminating dideoxynucleotide triphosphates (ddNTPs).



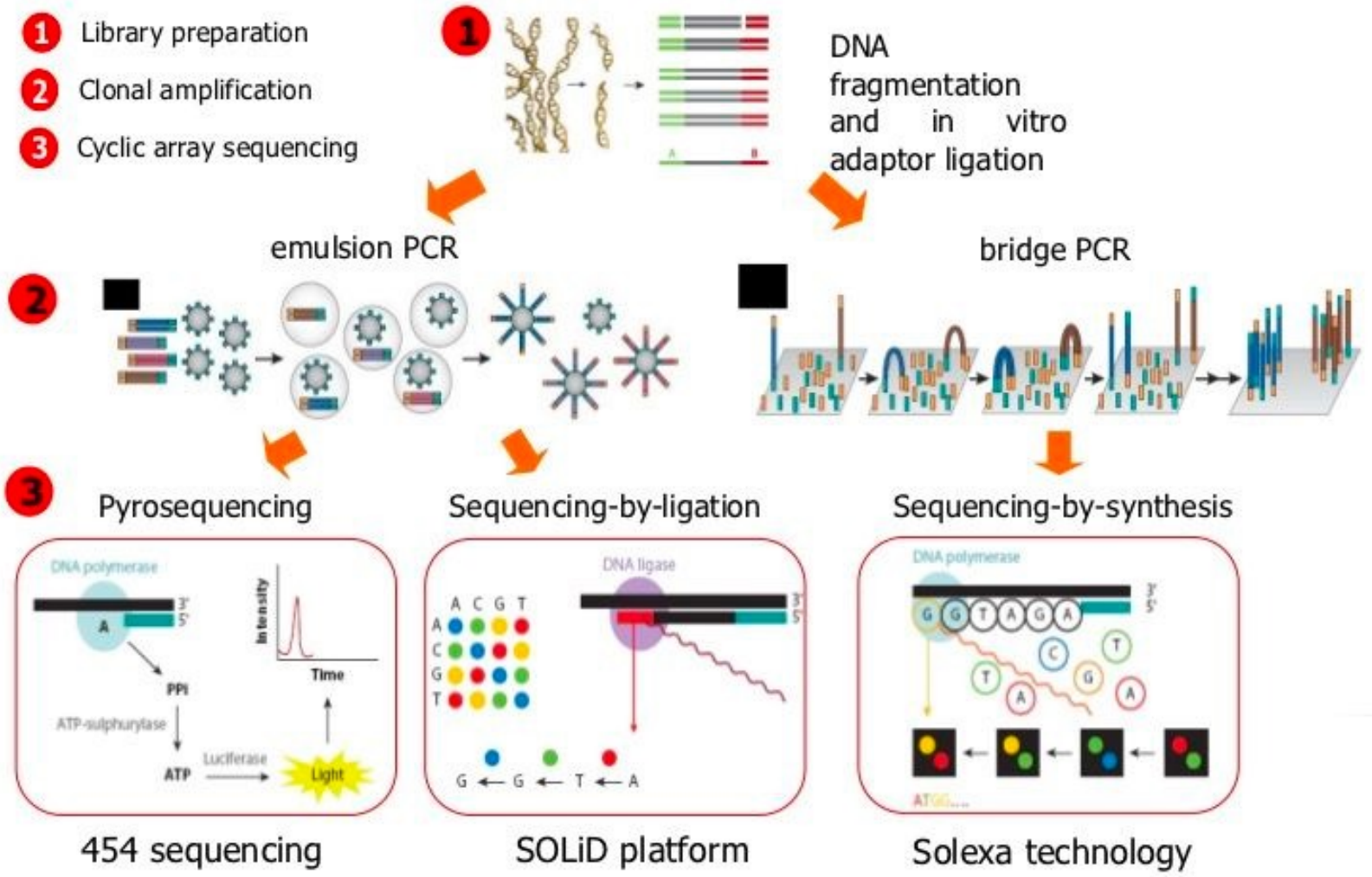
Genome Sequencing



NGS


1 Library preparation
Genome
DNA fragmentation
adapter
DNA library
2 DNA library bridge amplification
Library hybridization
Bridge amplification cycles
Amplified clusters
3 DNA library sequencing
Fluorescently labeled nucleotides
Sequencing cycles
Data collection
4 Alignment and data analysis
Contigs (overlapping regions)
Reads cluster 1
Reads cluster 2
Reads cluster 3
Assembled sequence

Different NGS platforms

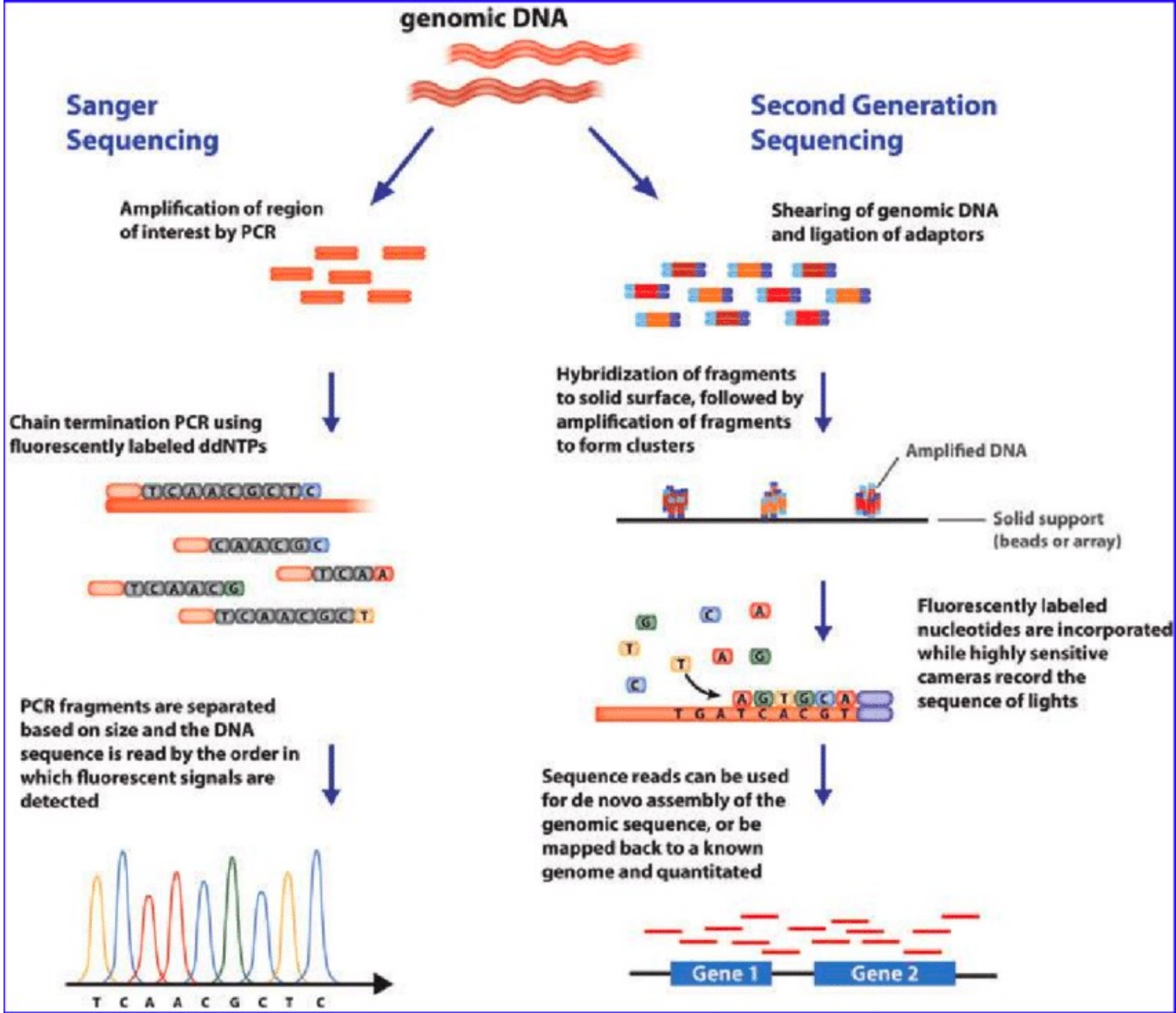


Pyrosequencing is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions.

the sequence extension reaction is not carried out by polymerases but rather by DNA ligase and either one-base-encoded probes or two-base-encoded probes. In its simplest form, a fluorescently labelled probe hybridizes to its complementary sequence adjacent to the primed template

This approach uses reversible terminator-bound dNTPs in a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage.

Protocols comparison



NGS advantages

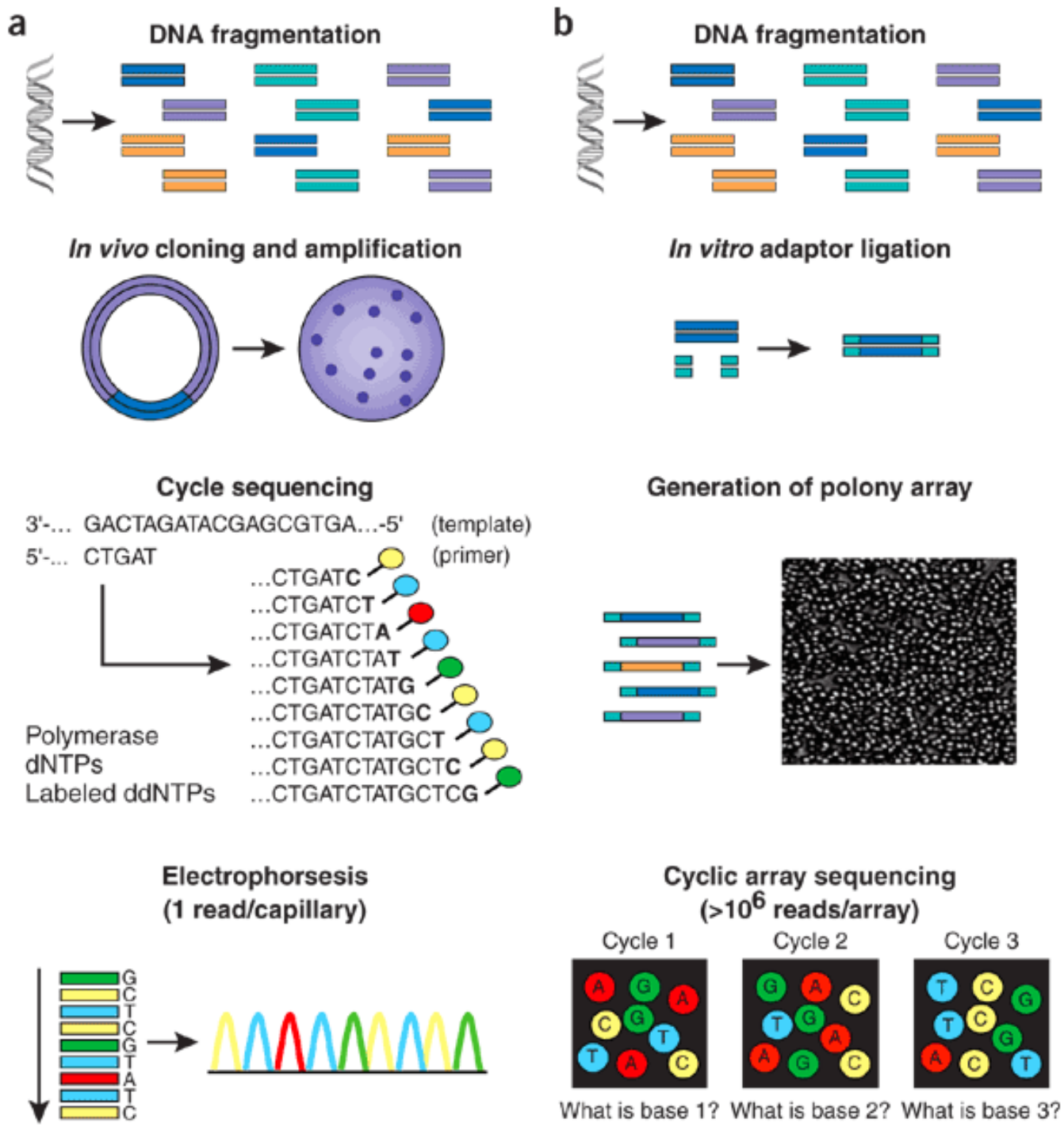
No *in vivo* cloning, transformation, colony picking

High degree of parallelism

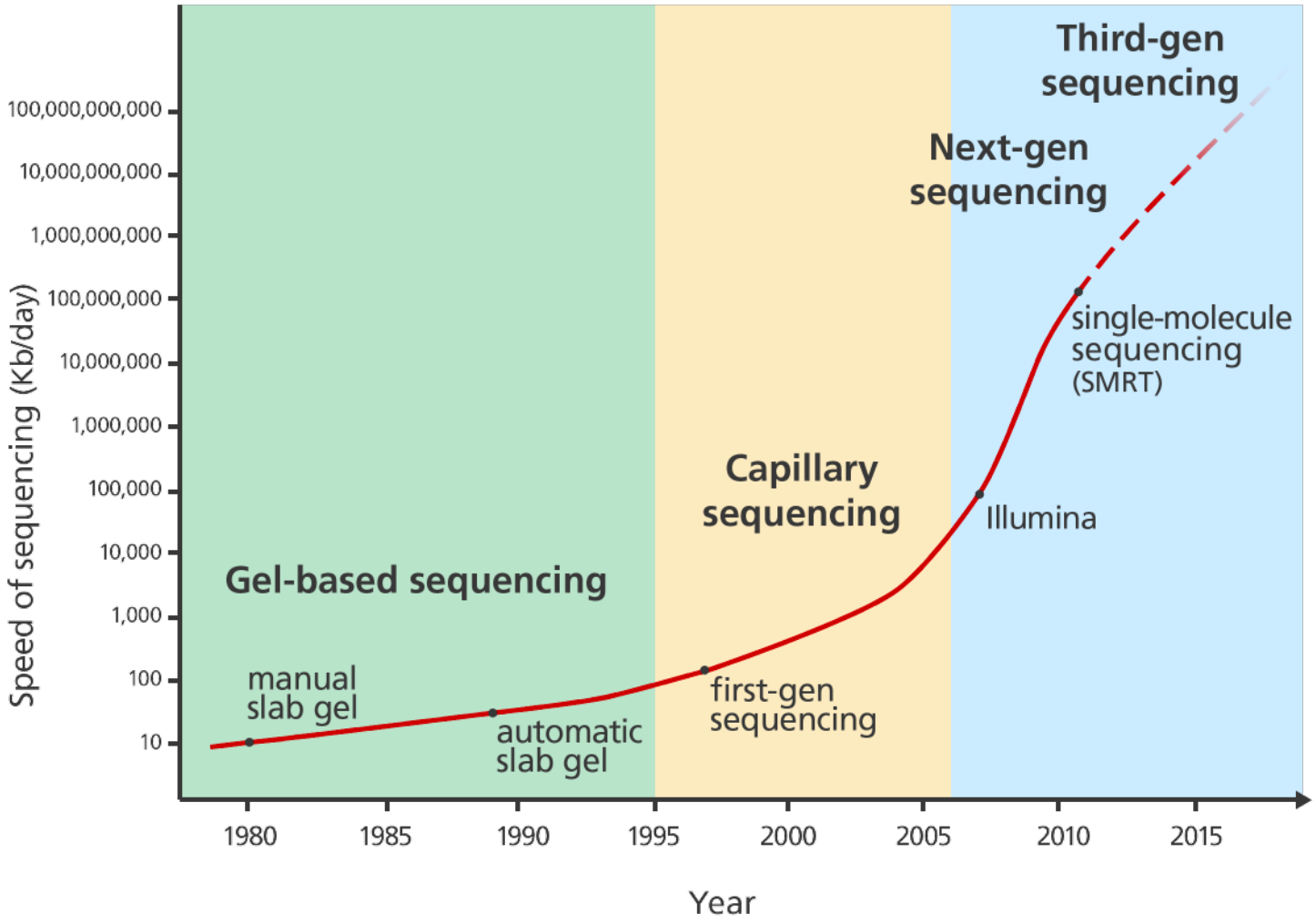
Low Reagent Cost

Reduced Sample Size

Faster



NGS advantages



Next generation sequencing

Illumina MiniSeq

Illumina MiSeq

Illumina NextSeq

Illumina HiSeq

Ion PGM

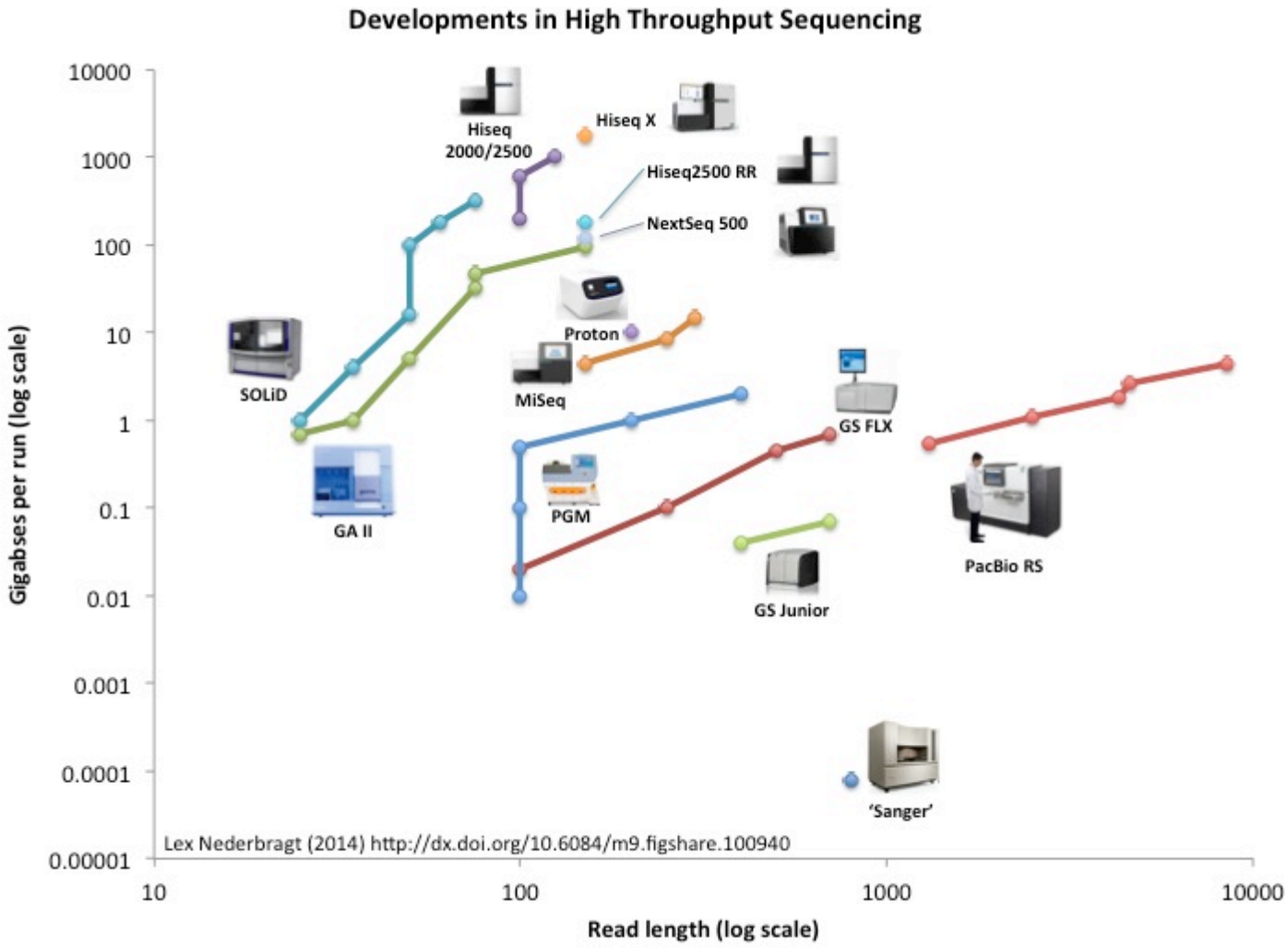
Ion Proton

Ion S5

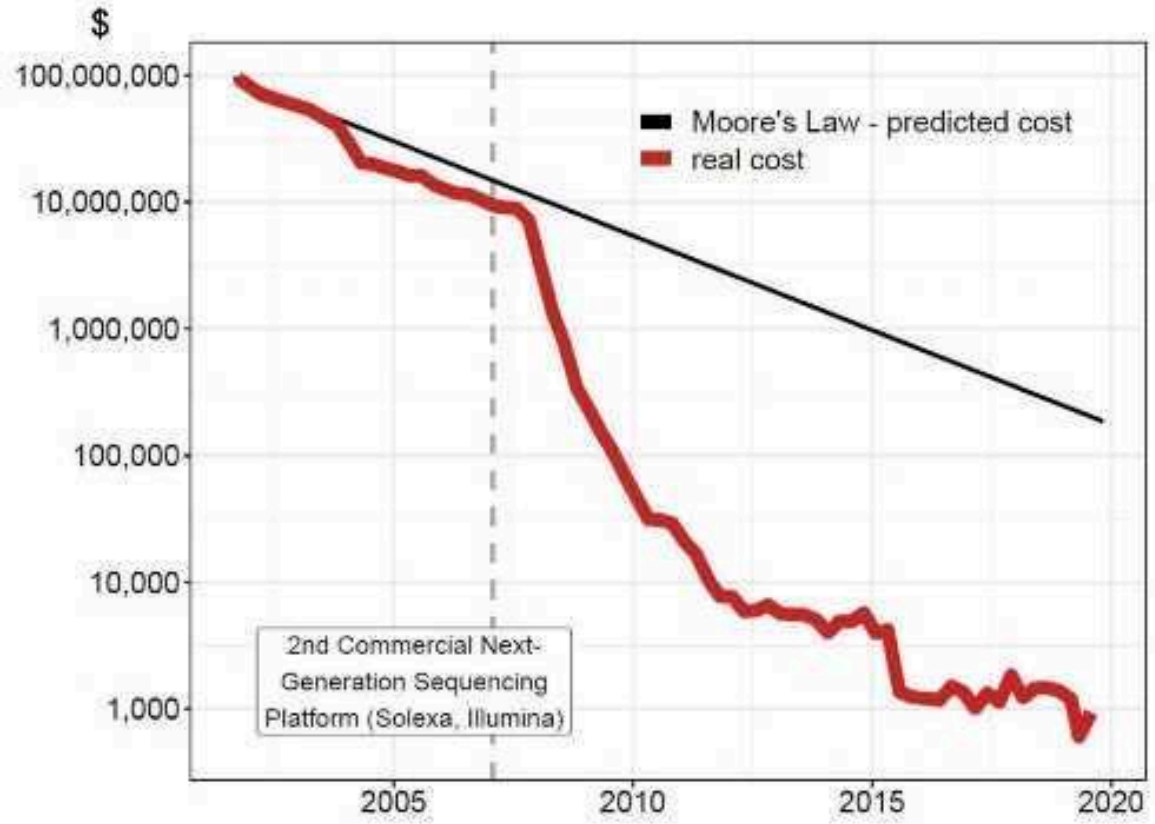
PacBio RS II

PacBio Sequel

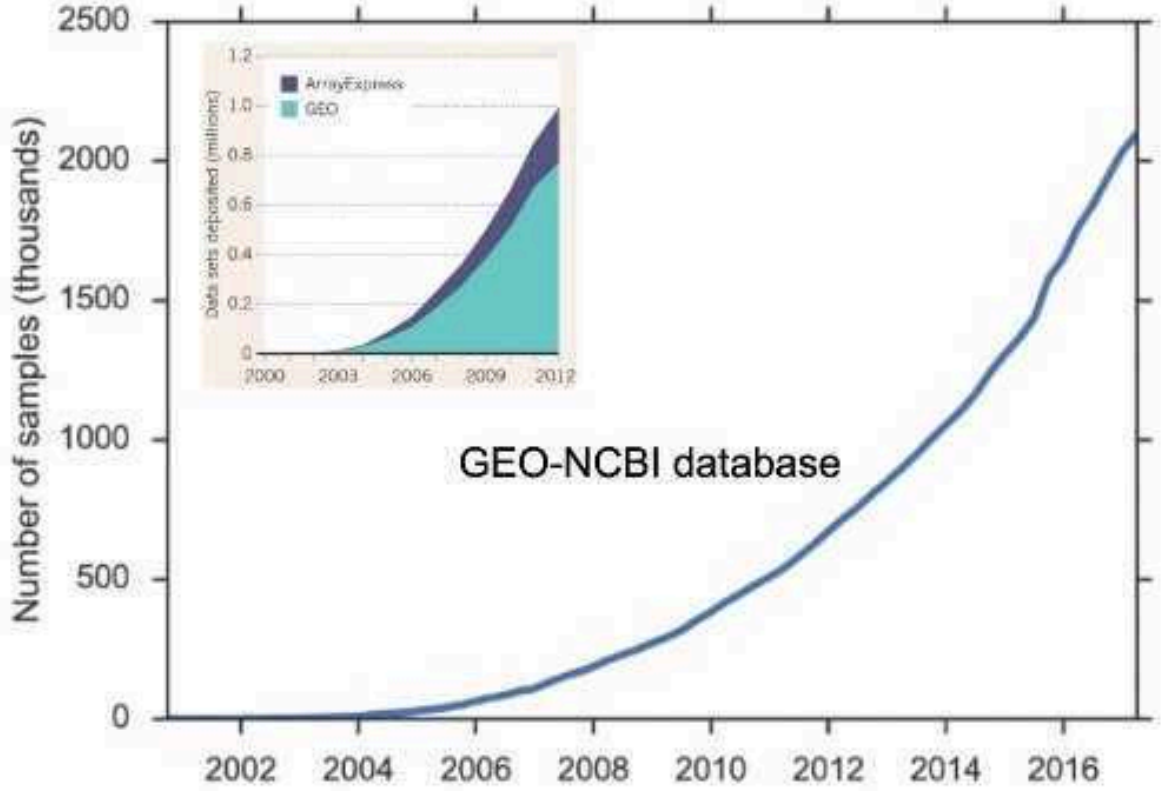
ONT MinION



Sequencing cost



Growth of omics data

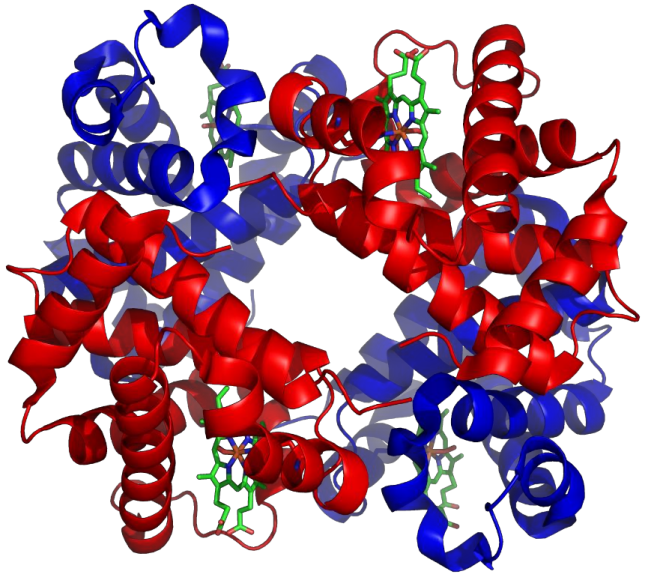


Mass spectrometry

- General purpose method to measure the accurate masses of small molecules

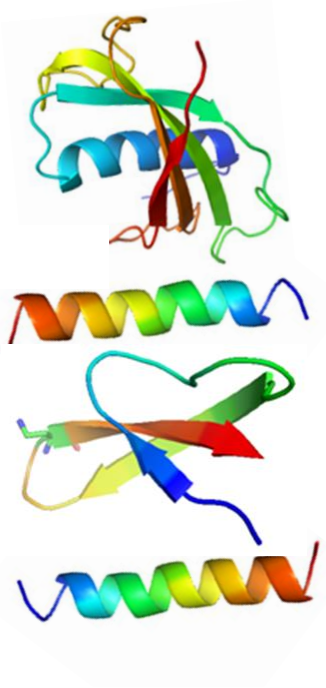
- Can be used to identify
  - Proteins (plus modifications)
  - Metabolites
    - Sugars
    - Nucleotides
    - Amino Acids
    - Lipids

Proteins



Digest

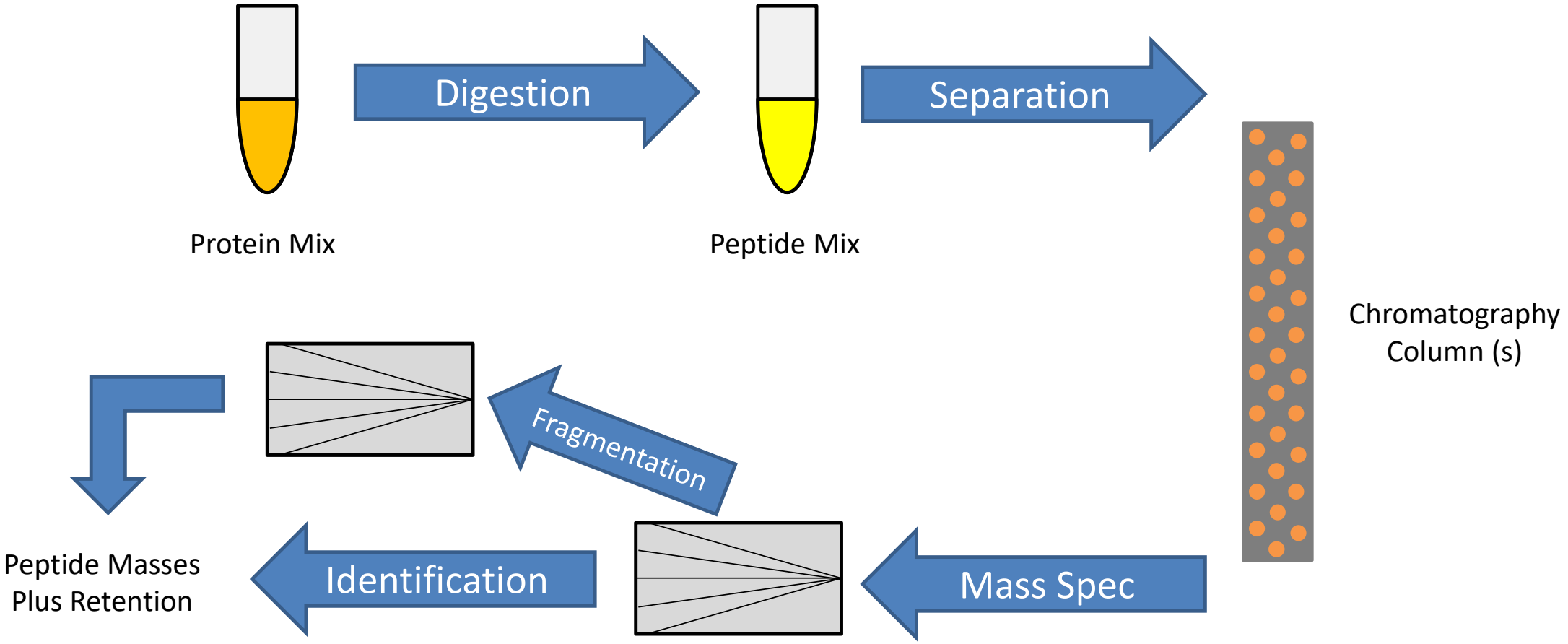
Peptides



Separate

A peptide

Mass-spec protocol



The mass spectometer

Mass spectrometry is an analytical tool useful for measuring the **mass-to-charge ratio** (*m/z*) of one or more molecules present in a sample. These measurements can often be used to calculate the **exact molecular weight** of the sample components as well. Typically, mass spectrometers can be used to identify unknown compounds via molecular weight determination, to quantify known compounds, and to determine structure and chemical properties of molecules.

1. The Ionization Source

2. The Mass Analyzer

3. Ion Detection System

1. The Ionization Source

Molecules are converted to **gas-phase ions** so that they can be moved about and manipulated by external electric and magnetic fields.

a technique called nanoelectrospray ionization

This method allows for creating positively or negatively charged ions, depending on the experimental requirements.

Nanoelectrospray ionization can directly couple the outlet of a small-scale chromatography column directly to the inlet of a mass spectrometer.

2. The Mass Analyzer

Once ionized, the ions are sorted and separated according to **mass-to-charge** (*m/z*) ratios.

There are a number of mass analyzers currently available, each of which has trade-offs relating to speed of operation, resolution of separation, and other operational requirements.

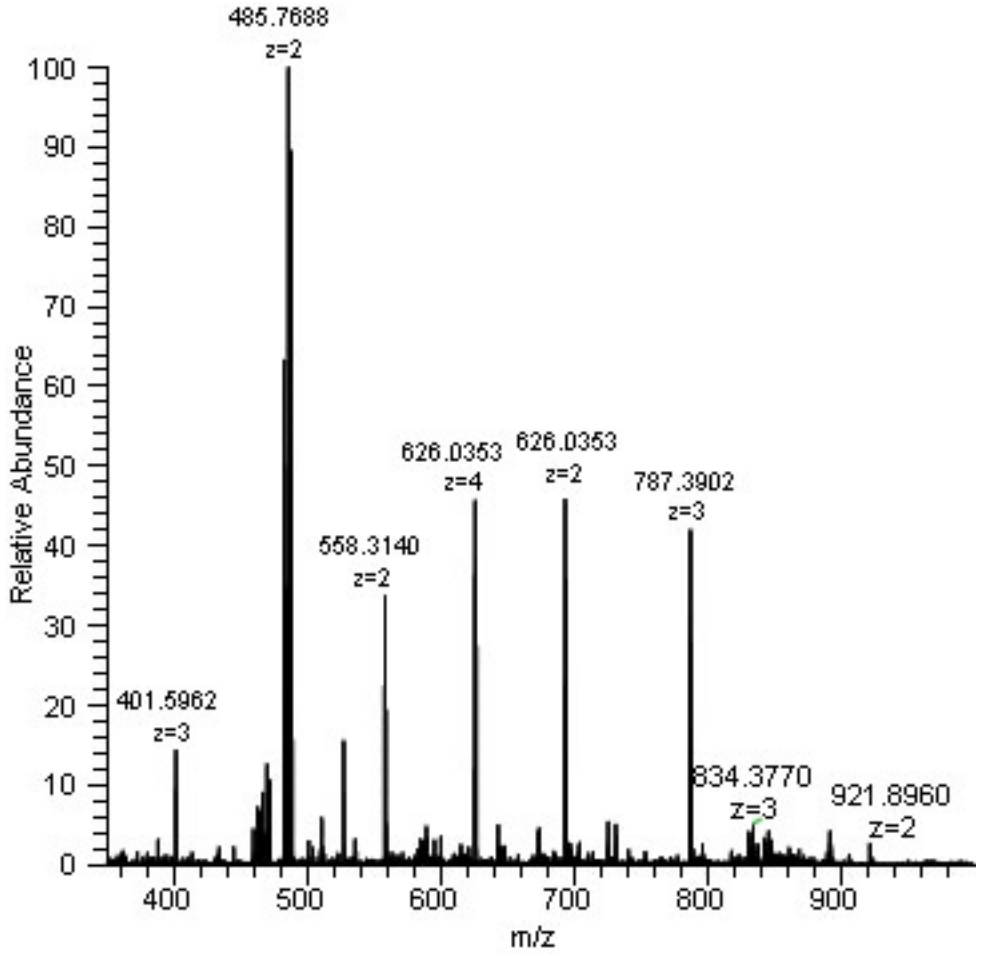
The mass analyzer often works in concert with the ion detection system.

3. Ion Detection System

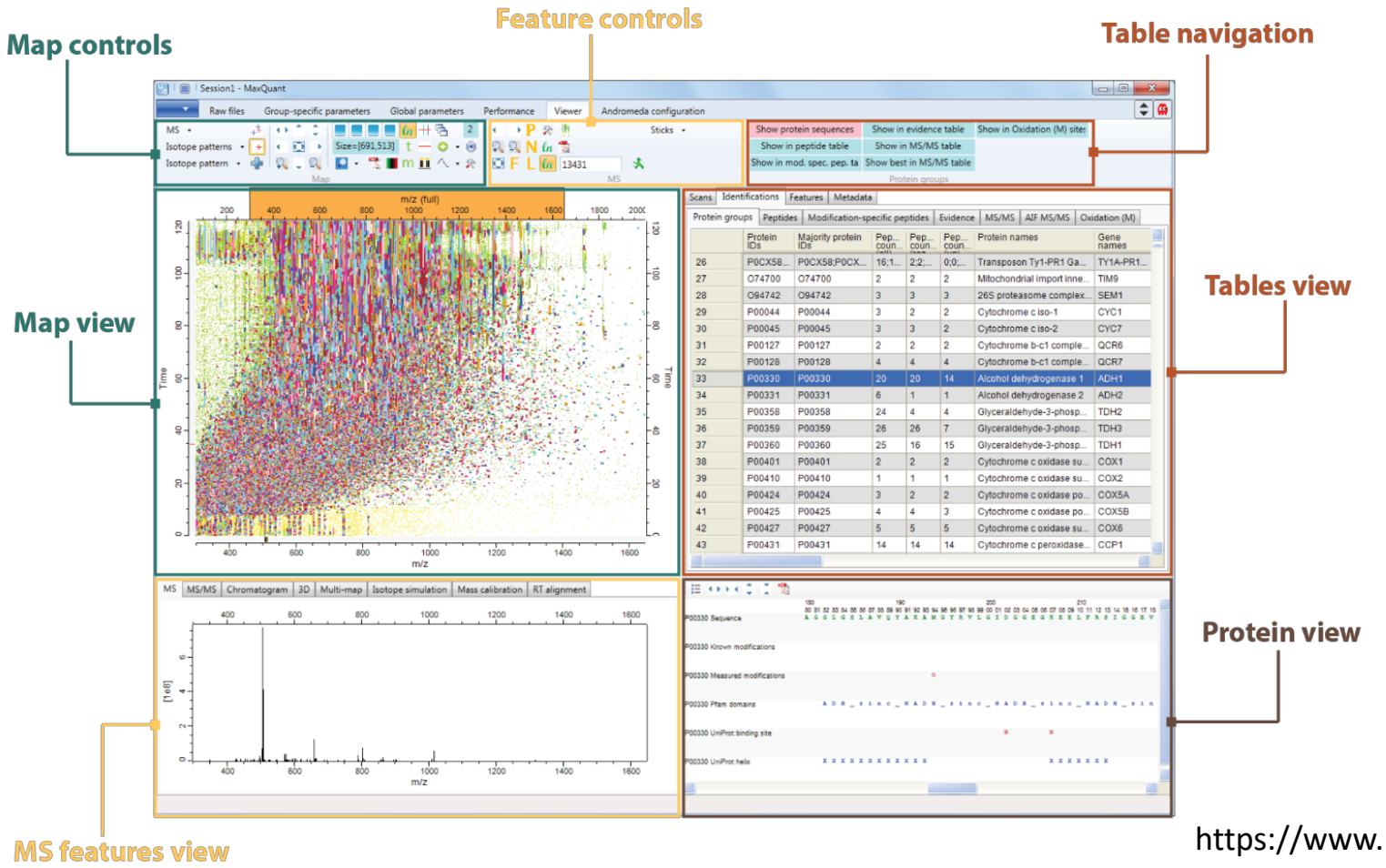
The separated ions are then measured and sent to a data system where the *m/z* ratios are stored together along with their relative abundance.

A **mass spectrum** is simply the *m/z* ratios of the ions present in a sample plotted against their intensities.

Each peak in a mass spectrum shows a component of unique *m/z* in the sample, and heights of the peaks connote the relative abundance of the various components in the sample.

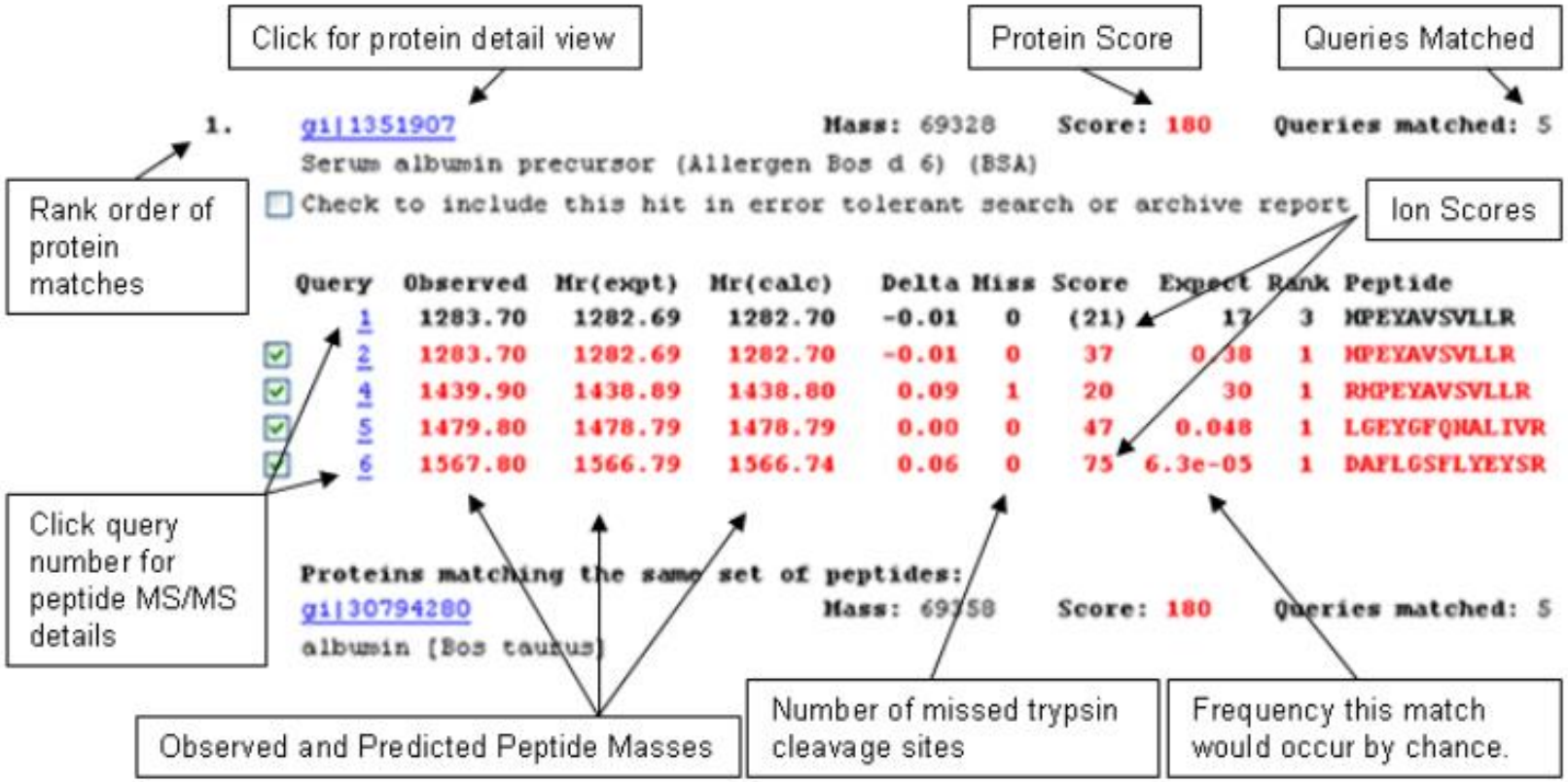


Mass-spec results



https://www.maxquant.org/

Mass-spec results

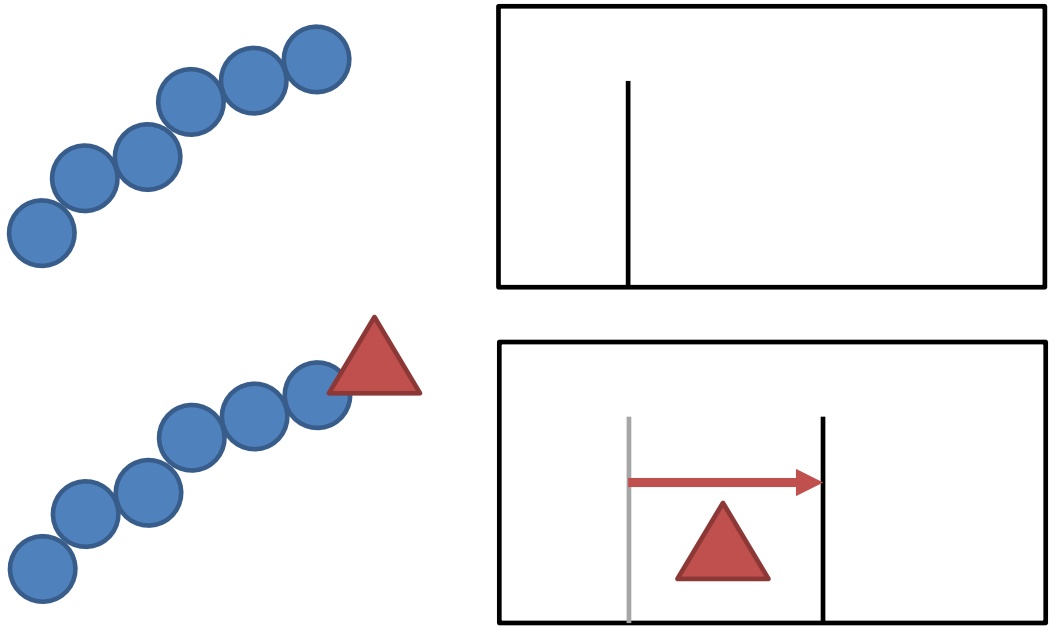


http://www.ohri.ca/proteomics/

Post-translational modifications

- When doing tandem mass spectrometry you can also identify modified peptides

- Phosphorylation
- Acetylation
- Methylation
- Palmitylation
- Acylation
- Ubiquitination
- etc.



Data Repositories

- For many techniques deposition of data in a suitable repository is a condition of publication
- Repositories are more developed and complete for some techniques than others
- Still a growing area

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Gene expression data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database
Proteomics data	PRIDE
*Earth, space & environmental sciences	Recommended Repositories

FAIR Data Principles

- Designed to make data as useful as possible to future researchers
  - Findable
    - Unique accession code
    - Rich metadata
  - Accessible
    - Automated query and download API
  - Interoperable
    - Use of open formats
    - Standard Ontologies for descriptions
  - Reusable
    - Clear licensing
    - Annotated to common community standards

Public Sequencing Databases

- GEO (NCBI)
- Array Express (EBI)
  - Databases for quantitated sequencing data. Provide experimental annotation and metadata and processed quantitated data

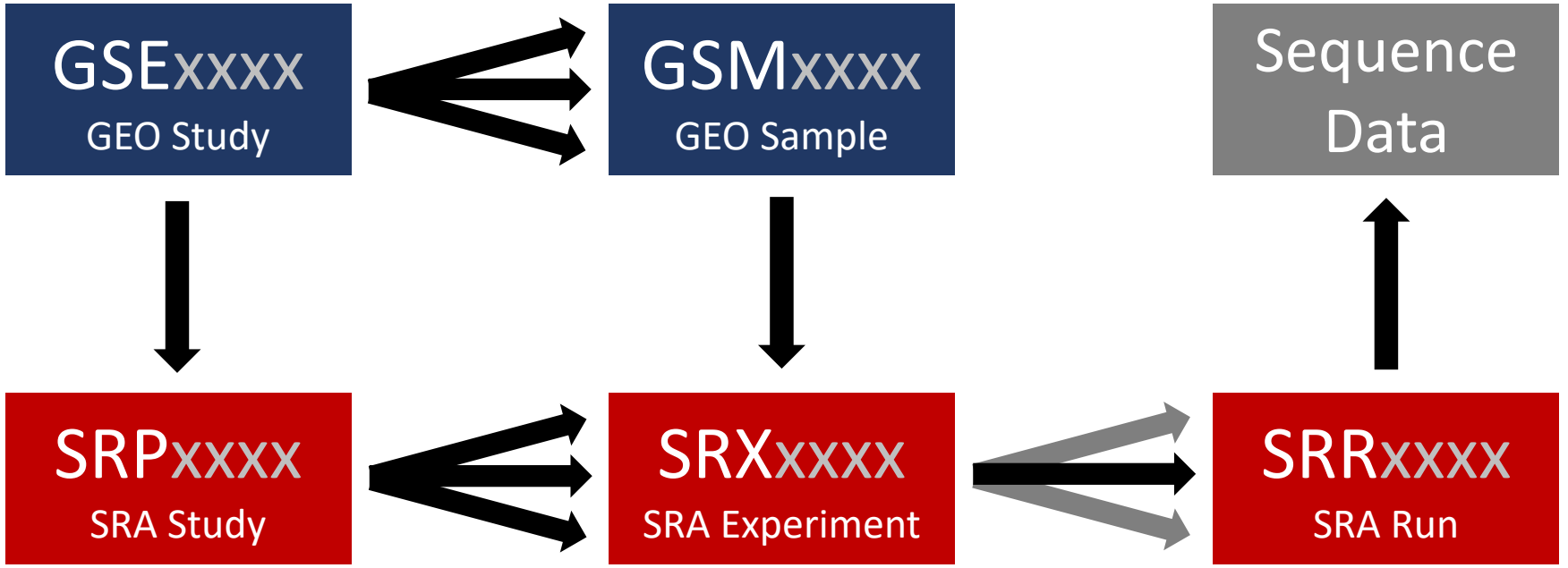
- SRA (NCBI)
- ENA (EBI)
  - Provide raw sequencing data as fastq files

Accession Codes

Transcription-induced formation of extrachromosomal DNA during yeast ageing

Ryan M. Hull[1¤a], Michelle King[1], Grazia Pizza[1¤b], Felix Krueger[2], Xabier Vergara[1¤c], Jonathan Houseley[1]*

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. All sequencing files are available from the GEO database (accession number GSE135542).



Series GSE135542

Query DataSets for GSE135542

Status	Public on Oct 18, 2019
Title	Transcription-induced formation of extrachromosomal DNA during yeast ageing
Organism	Saccharomyces cerevisiae
Overall design	Aged cell samples analysed in pairs of -/+ Cu, for both wt and various mutants. 3 replicates of the 3xCUP1 experiment are included
Contributor(s)	Hull R, King M, Houseley J
Platforms (1)	GPL17342 Illumina HiSeq 2500 (Saccharomyces cerevisiae)
Samples (30)	GSM4015617 3xCUP1_24hr_1_REC-seq
	GSM4015618 3xCUP1_24hr_2_REC-seq
	GSM4015619 3xCUP1_24hr_300uM_Cu_1_REC-seq

⊞ More...

Relations

BioProject	PRJNA559191
SRA	SRP217740

Supplementary file	Size	Download	File type/resource
GSE135542_3xCUP1_processed_data_report.txt.gz	1.3 Mb	(ftp)(http)	TXT
GSE135542_cu_and_gal_processed_data_report.txt.gz	12.4 Mb	(ftp)(http)	TXT
GSE135542_mutants_processed_data_report.txt.gz	1.4 Mb	(ftp)(http)	TXT

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

SRA Run Selector

		BioSample	Bases	Bytes	Experiment	GEO_Accession	Sample Name	source_name	strain
1	SRR9924096	SAMN12529574	1.01 G	344.69 Mb	SRX6673092	GSM4015624	GSM4015624	Cells aged 24 hours in SD media	MEP mus81
2	SRR9924097	SAMN12529572	994.71 M	338.32 Mb	SRX6673093	GSM4015625	GSM4015625	Cells aged 24 hours in SD media	MEP mus81
3	SRR9924098	SAMN12529570	838.88 M	294.83 Mb	SRX6673094	GSM4015626	GSM4015626	Cells aged 24 hours in SD media	MEP mus81
4	SRR9924099	SAMN12529569	631.87 M	250.37 Mb	SRX6673095	GSM4015627	GSM4015627	Cells aged 48 hours in YPD media	MEP [Pgal1-3HA cup1]/CUP1
5	SRR9924100	SAMN12529567	1.11 G	407.87 Mb	SRX6673096	GSM4015628	GSM4015628	Cells aged 48 hours in YPD media	MEP [Pgal1-3HA cup1]/CUP1
6	SRR9924101	SAMN12529565	903.88 M	343.05 Mb	SRX6673097	GSM4015629	GSM4015629	Cells aged 48 hours in YPGal media	MEP [Pgal1-3HA cup1]/CUP1
7	SRR9924102	SAMN12529564	1.45 G	529.28 Mb	SRX6673098	GSM4015630	GSM4015630	Cells aged 48 hours in YPGal media	MEP [Pgal1-3HA cup1]/CUP1
8	SRR9924103	SAMN12529561	646.51 M	227.76 Mb	SRX6673099	GSM4015631	GSM4015631	Cells aged 24 hours in SD media	MEP sae2
9	SRR9924104	SAMN12529560	1.05 G	357.64 Mb	SRX6673100	GSM4015632	GSM4015632	Cells aged 24 hours in SD media	MEP sae2
10	SRR9924105	SAMN12529558	829.11 M	281.75 Mb	SRX6673101	GSM4015633	GSM4015633	Cells aged 24 hours in SD media	MEP sae2
11	SRR9924106	SAMN12529557	823.58 M	284.92 Mb	SRX6673102	GSM4015634	GSM4015634	Cells aged 24 hours in SD media	MEP sae2
12	SRR9924107	SAMN12529555	1.03 G	354.90 Mb	SRX6673103	GSM4015635	GSM4015635	Cells aged 24 hours in SD media	MEP spt3
13	SRR9924108	SAMN12529553	994.63 M	344.93 Mb	SRX6673104	GSM4015636	GSM4015636	Cells aged 24 hours in SD media	MEP spt3
14	SRR9924109	SAMN12529608	551.82 M	242.67 Mb	SRX6673105	GSM4015637	GSM4015637	Cells aged 24 hours in SD media	MEP
15	SRR9924110	SAMN12529606	961.46 M	360.10 Mb	SRX6673106	GSM4015638	GSM4015638	Cells aged 24 hours in SD media	MEP
16	SRR9924111	SAMN12529604	1.33 G	454.02 Mb	SRX6673107	GSM4015639	GSM4015639	Cells aged 24 hours in SD media	MEP
17	SRR9924112	SAMN12529602	1.06 G	395.94 Mb	SRX6673108	GSM4015640	GSM4015640	Cells aged 24 hours in SD media	MEP
18	SRR9924113	SAMN12529601	563.99 M	258.47 Mb	SRX6673109	GSM4015641	GSM4015641	Cells aged 24 hours in SD media	MEP
19	SRR9924114	SAMN12529599	886.36 M	336.01 Mb	SRX6673110	GSM4015642	GSM4015642	Cells aged 24 hours in SD media	MEP
20	SRR9924115	SAMN12529598	1.30 G	446.44 Mb	SRX6673111	GSM4015643	GSM4015643	Cells aged 24 hours in SD media	MEP
21	SRR9924116	SAMN12529596	1.33 G	483.90 Mb	SRX6673112	GSM4015644	GSM4015644	Cells aged 24 hours in SD media	MEP
22	SRR9924117	SAMN12529594	1.13 G	389.68 Mb	SRX6673113	GSM4015645	GSM4015645	Cells aged 24 hours in SD media	MEP
23	SRR9924119	SAMN12529593	921.09 M	324.51 Mb	SRX6673114	GSM4015646	GSM4015646	Cells aged 24 hours in SD media	MEP

ENA
European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

Project: PRJNA559191

Extrachromosomal circular DNA (eccDNA) facilitates adaptive evolution by allowing rapid and extensive gene copy number variation, and is implicated in the pathology of cancer and ageing. Here, we demonstrate that yeast aged under environmental copper accumulate high levels of eccDNA containing the copper resistance gene CUP1. Transcription of CUP1 causes CUP1 eccDNA accumulation, which occurs in the absence of phenotypic selection. We have developed a sensitive and quantitative eccDNA sequencing pipeline that reveals CUP1 eccDNA accumulation on copper exposure to be exquisitely site specific, with no other detectable changes across the eccDNA complement. eccDNA forms de novo from the CUP1 locus through processing of DNA double-strand breaks (DSBs) by Sae2 / Mre11 and Mus81, and genome-wide analyses show that other protein coding eccDNA species in aged yeast share a similar biogenesis pathway. Although abundant we find that CUP1 eccDNA does not replicate efficiently, and high copy numbers in aged cells arise through frequent formation events combined with asymmetric DNA segregation. The transcriptional stimulation of CUP1 eccDNA formation shows that age-linked genetic change varies with transcription pattern, resulting in gene copy number profiles tailored by environment. Overall design: Aged cell samples analysed in pairs of -/+ Cu, for both wt and various mutants. 3 replicates of the 3xCUP1 experiment are included

Show More

Organism:	Saccharomyces cerevisiae (baker's yeast)
Secondary Study Accession:	SRP217740
Study Title:	Transcription-induced formation of extrachromosomal DNA during yeast ageing
Center Name:	Bioinformatics, The Babraham Institute
Study Name:	Transcription-induced formation of extrachromosomal DNA during yeast ageing

Read Files

Show Column Selection

Download report: JSON TSV

Download Files as ZIP | Download selected files

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Download All FASTQ FTP	Sub
PRJNA559191	SAMN12529574	SRX6673092	SRR9924096	4932	Saccharomyces cerevisiae	SRR992409...fastq.gz SRR992409...fastq.gz	

SRA Explorer

This tool aims to make datasets within the Sequence Read Archive more accessible.

Search for: SRP217740[All Fields]

Max Results 100 **Start At Record** 0

Need inspiration? Try GSE30567, SRP043510, PRJEB8073, ERP009109 or human liver miRNA.

Select relevant datasets and click *add to collection*. When you're finished, view all saved datasets with the button in the top right of the page, where you can copy the SRA URLs.

Showing 30 results.

Filter results: Enter search term | All Fields

Add 0 to collection

Title	Accession	Instrument	Total Bases (Mb)	Date Created
GSM4015617: 3xCUP1_24hr_1_REC-seq; Saccharomyces cerevisiae; OTHER	SRR9924120	Illumina HiSeq 2500	8685	21 Oct 2019
GSM4015618: 3xCUP1_24hr_2_REC-seq; Saccharomyces cerevisiae; OTHER	SRR9924121	Illumina HiSeq 2500	10212	21 Oct 2019
GSM4015619: 3xCUP1_24hr_300uM_Cu_1_REC-seq; Saccharomyces cerevisiae; OTHER	SRR9924122	Illumina HiSeq 2500	9693	21 Oct 2019
GSM4015620: 3xCUP1_24hr_300uM_Cu_2_REC-seq; Saccharomyces cerevisiae; OTHER	SRR9924123	Illumina HiSeq 2500	9602	21 Oct 2019

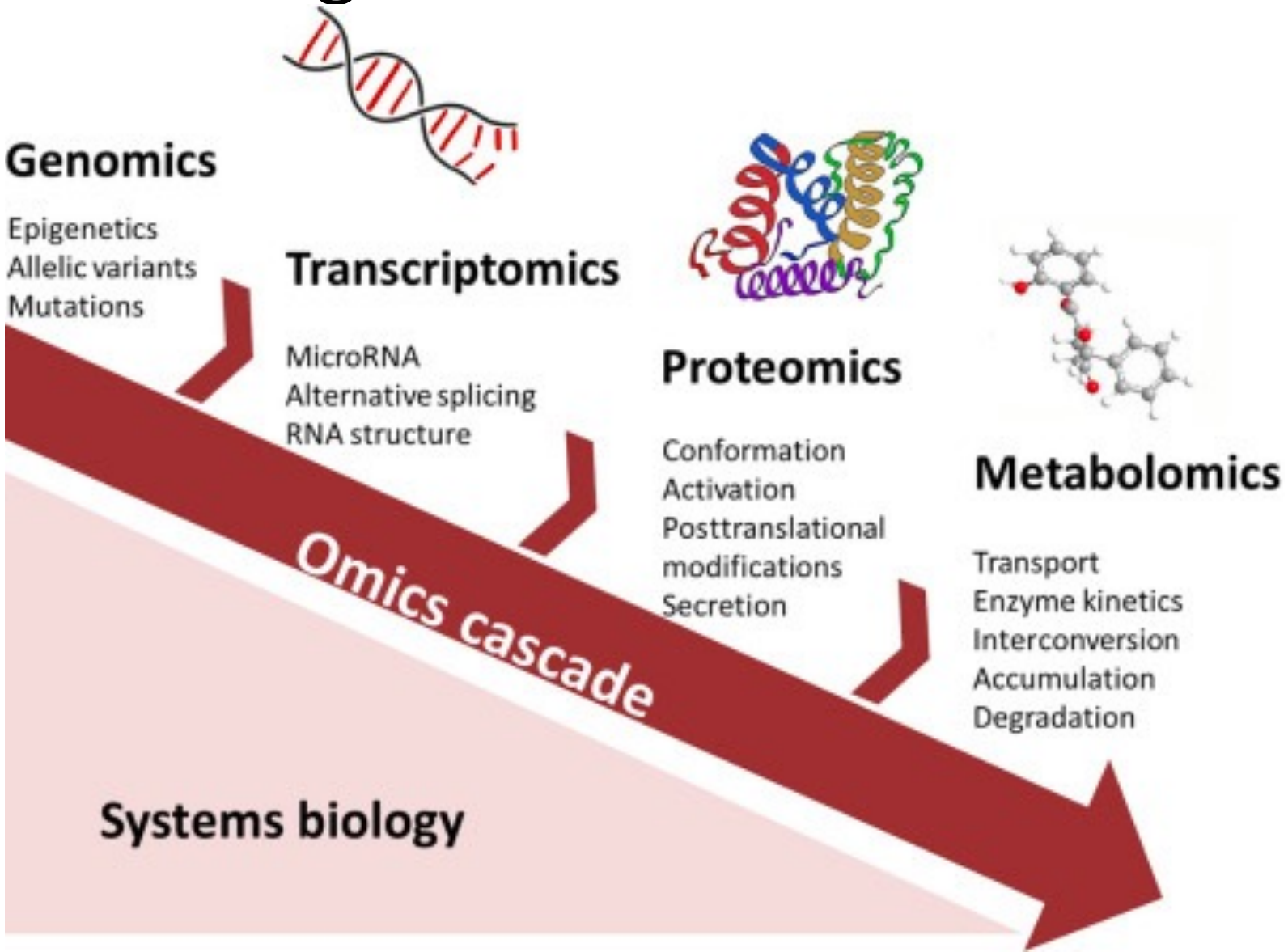
SRA Downloader

```
sradownloader SRR9924120
```

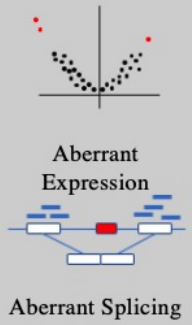
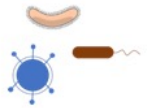
Course Structure

- Central Dogma Data Sources
  - Genomes and Annotations
  - Protein Domains and Structures
  - Reactions, Pathways and Interactions
- Experimental Techniques, Datatypes and Repositories
  - Sequencing and Variants
  - Proteomics and Metabolites
- Omics
  - General concepts
  - Analysis approaches

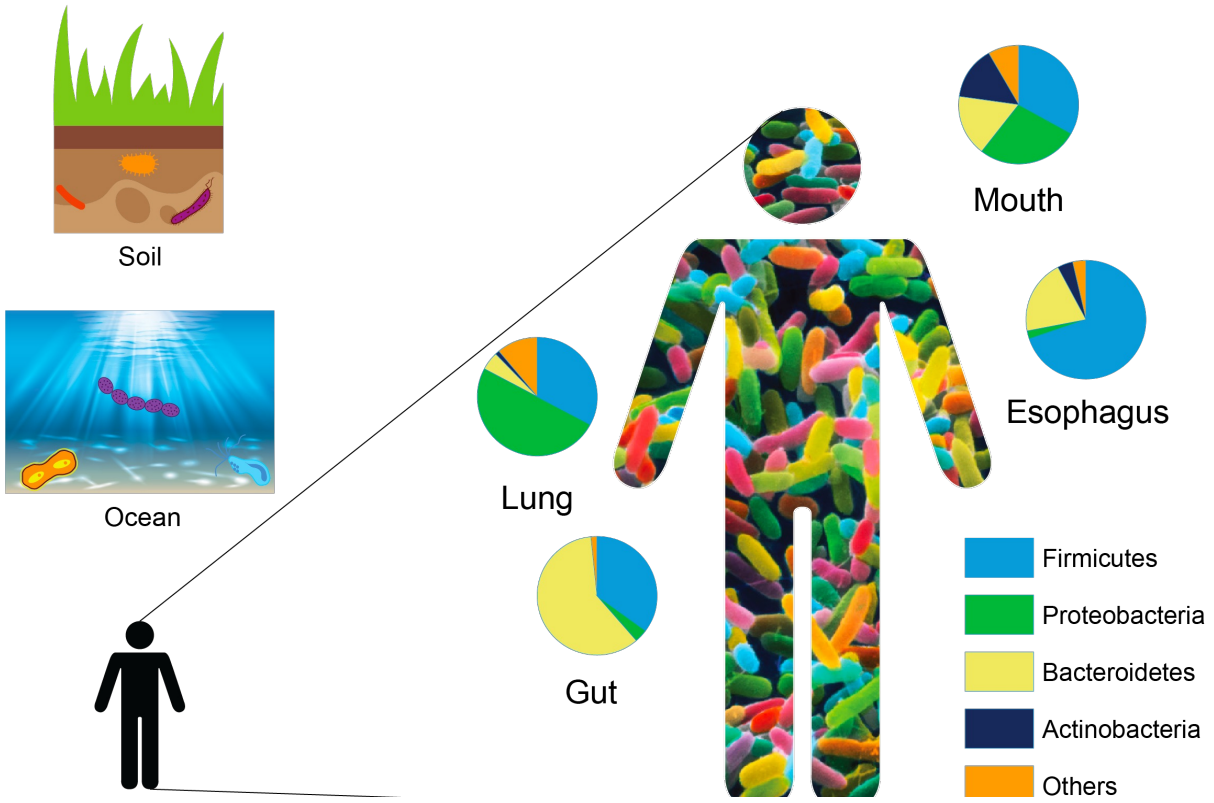
Omics technologies



What are OMICS for?

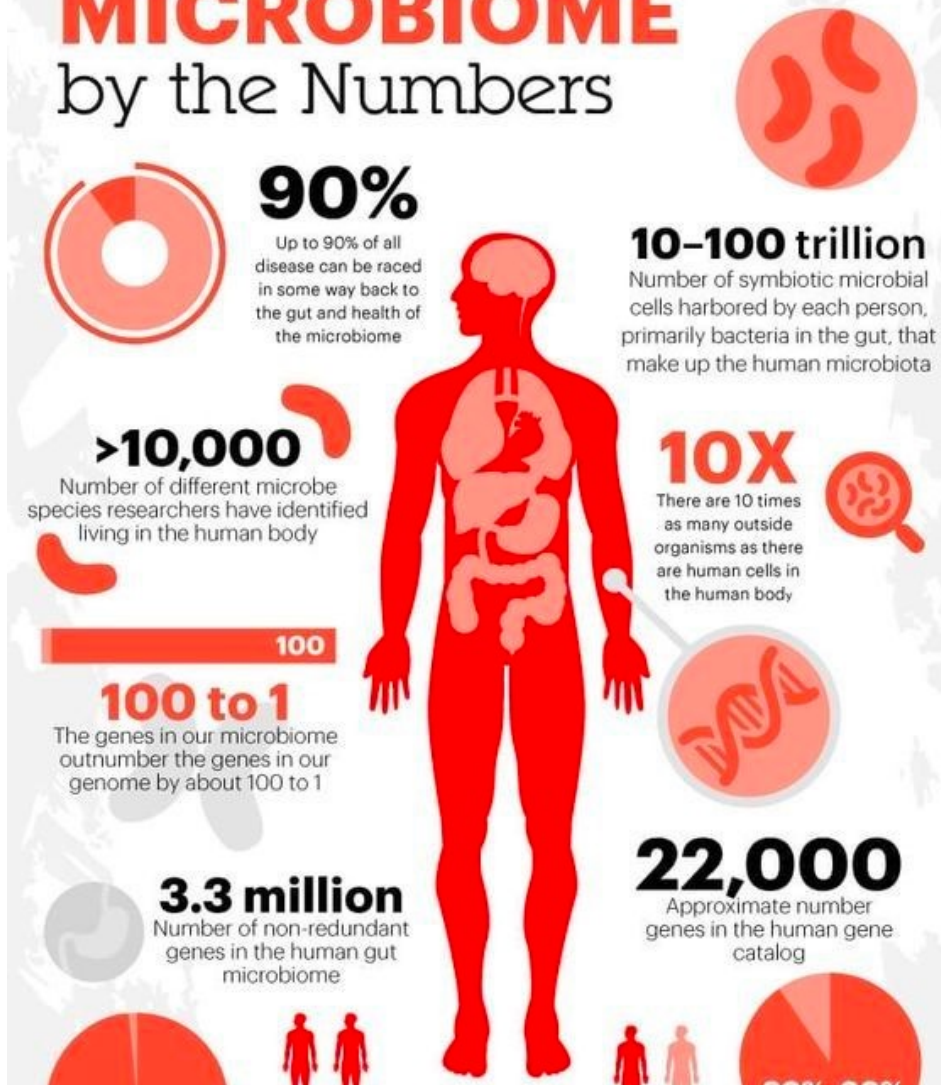
Omics technology		Molecule	Target	Technique/s	Purposes
Genomics		DNA	Genome	NGS WES - WGS	Identification of genetic variants
Epigenomics		DNA methylation, histone acetylation	Epigenome	NGS (ChIP-seq, MeDIP-seq, BS-seq et d'autre variantes	Determination of epigenetic changes in DNA that regulate gene expression
Transcriptomics		RNA	Transcriptome	NGS RNA-seq SmallRNA-seq	Characterization expression levels of genes and identification of non-coding transcripts; alternative splicing events;
Microbiomics		RNA16s	Microbiome	NGS (16S ribosomal abundance)	Identification of microorganisms populating the skin, mucosal surfaces and the gut
Proteomics/ Metabolomics		Proteins / Metabolites	Proteome / metabolome	Mass spectrometry	Characterization of the abundance of proteins / metabolites

What is microbiomics?



‘Microbiomics’ is a fast-growing field in which all the microorganisms of a given community (a ‘microbiota’) are investigated together. This could be the microbiota of an environmental sample (e.g. soil or water), a particular body site (e.g. the gut or the mouth) or from a particular organism (e.g. farm or zoo animals).

The Importance of the MICROBIOME by the Numbers



90%

Up to 90% of all disease can be raced in some way back to the gut and health of the microbiome

10–100 trillion

Number of symbiotic microbial cells harbored by each person, primarily bacteria in the gut, that make up the human microbiota

>10,000

Number of different microbe species researchers have identified living in the human body

10X

There are 10 times as many outside organisms as there are human cells in the human body

100

100 to 1

The genes in our microbiome outnumber the genes in our genome by about 100 to 1

3.3 million

Number of non-redundant genes in the human gut microbiome

22,000

Approximate number genes in the human gene catalog

99.9%

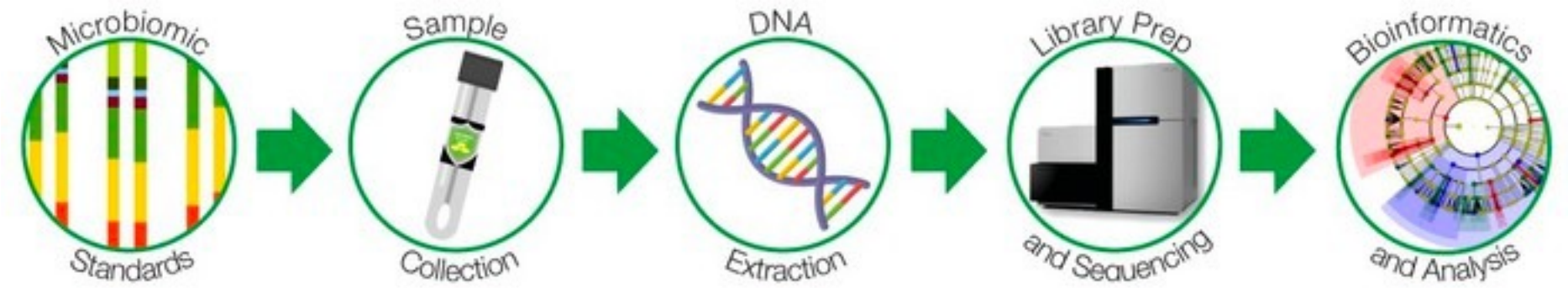
Percentage individual humans are identical to one another in terms of host genome

80%–90%

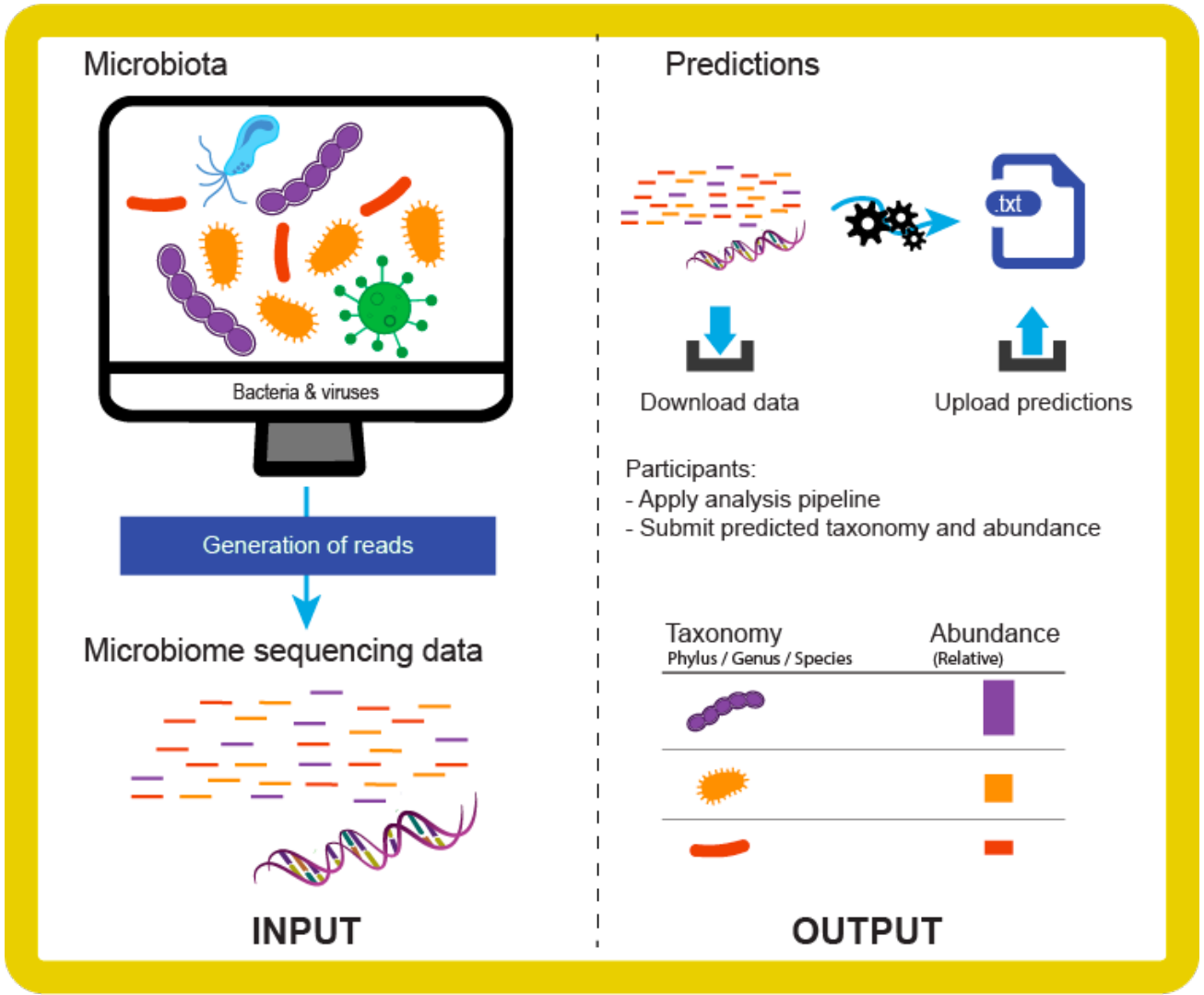
Percentage individual humans are different from one another in terms of the microbiome

Microbiota profiling

At the simplest level, microbiomics looks to investigate the make-up of a particular microbial community and how this might change over time, or with particular pressure. This is typically done using 16S profiling. DNA is extracted from the target sample (soil / water / faeces) and the 16S RNA gene amplified using the polymerase chain reaction (PCR). The resulting amplified DNA fragments are sequenced, and can be matched against sequence databases to achieve identifications of the organisms present. This can give a ‘snapshot’ of the communities present in a sample at any given time, and can be used to compare the communities in different samples, or follow changes in the communities of a particular sample site over time.



Microbiota profiling



Applications

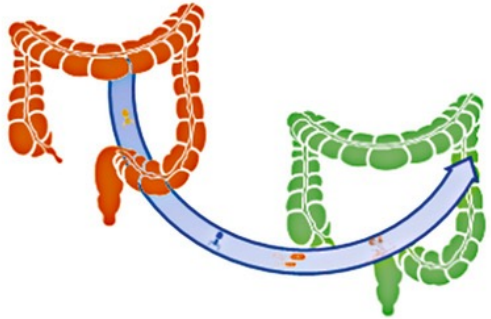
BIOMARKERS
Sequencing Stool Samples

DRUG DISCOVERY
Microbiome-Derived Compound
or microbiome-target drugs

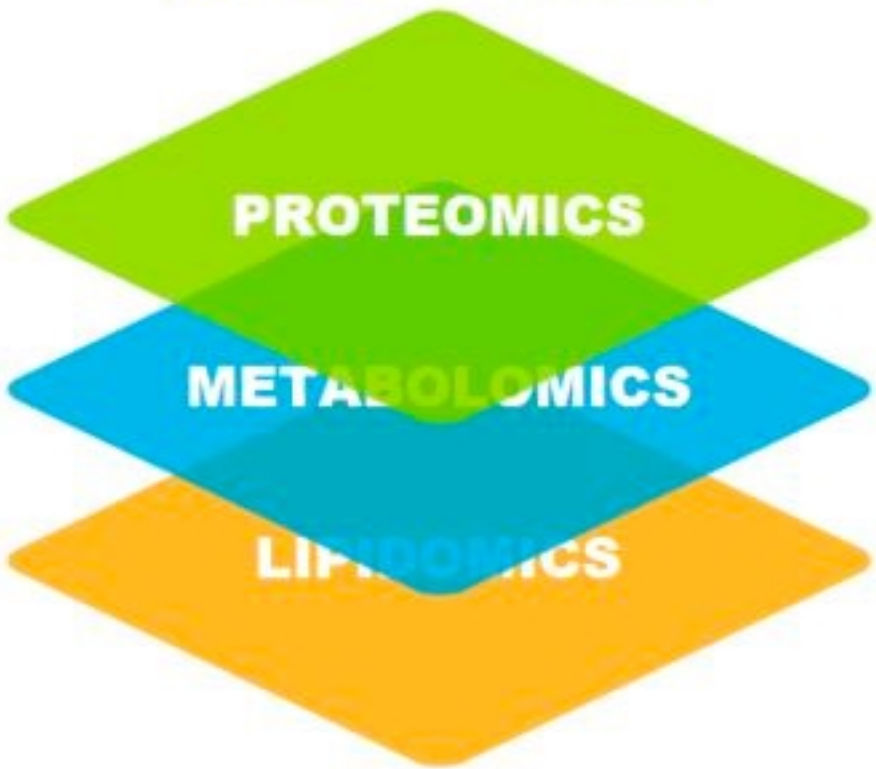
THERAPY OPTIMIZATION
Immunoterapeutics effect

PRECISION MEDICINE
Computational Biologist

MICROBIOMA THERAPY
Fecal Microbial Transplantation



What is proteomics?



1. Proteome

- Protein identification and quantitation
- Post-translational modification

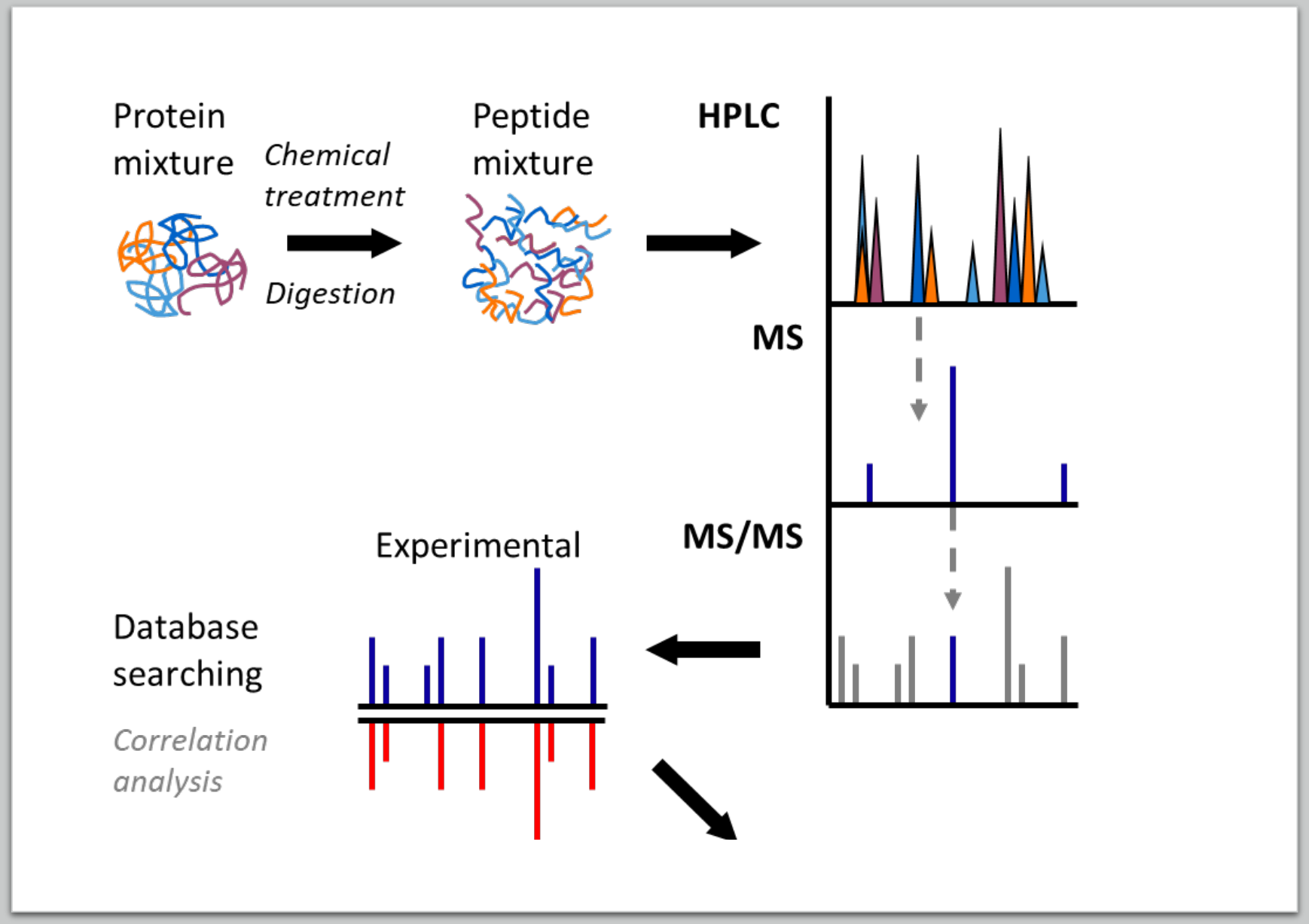
2. Metabolome

- Primary metabolites
- Targeted & untargeted platforms

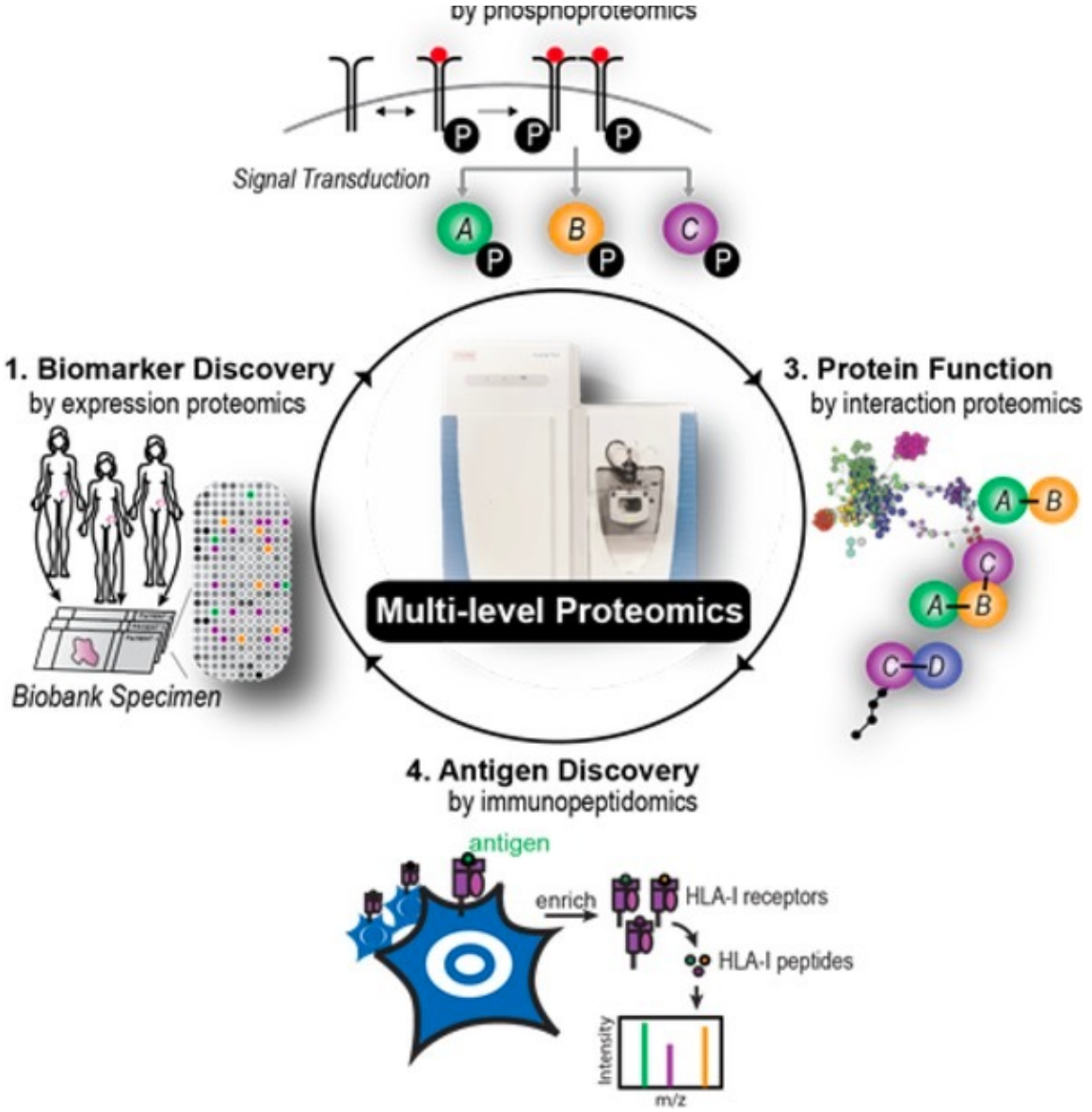
3. Lipidome

- Large-scale analysis of phospholipids
- Lipid mediators analysis

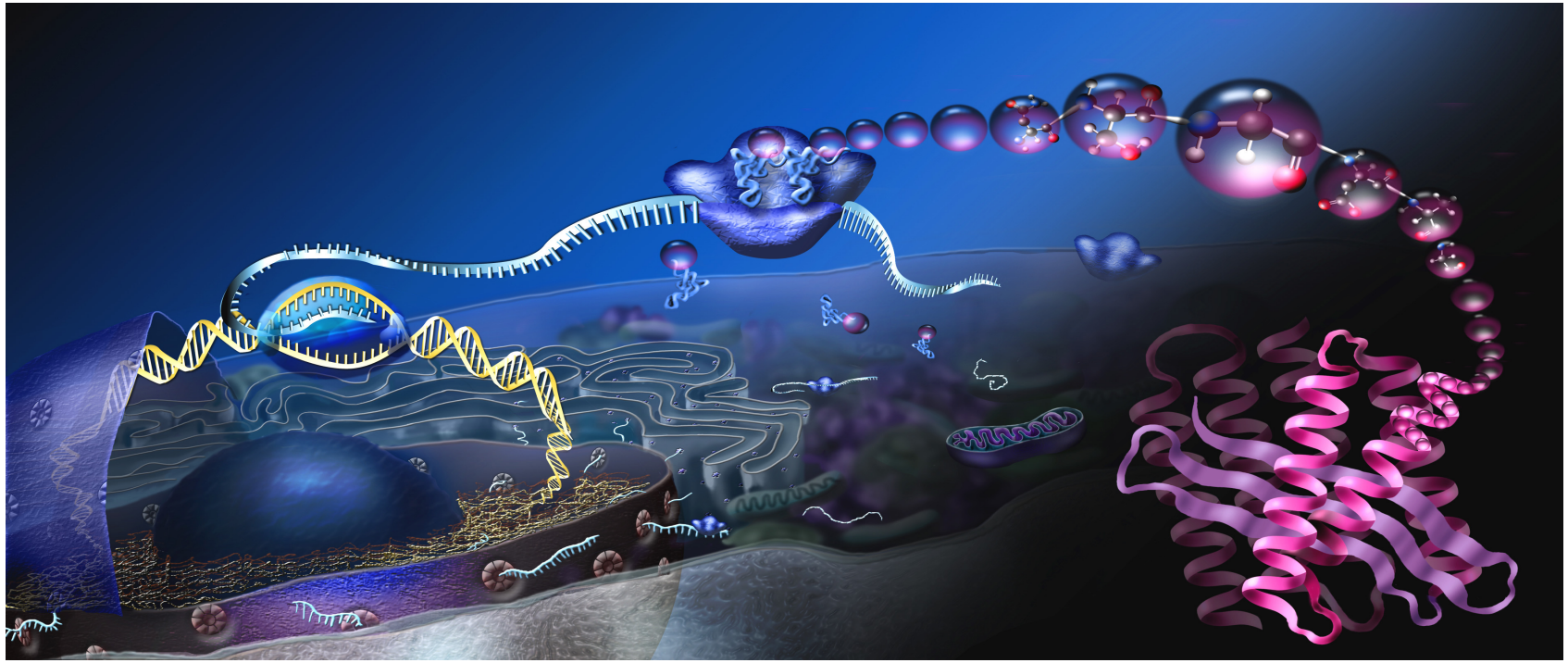
How do we study proteomics?



What can you do with proteomics?

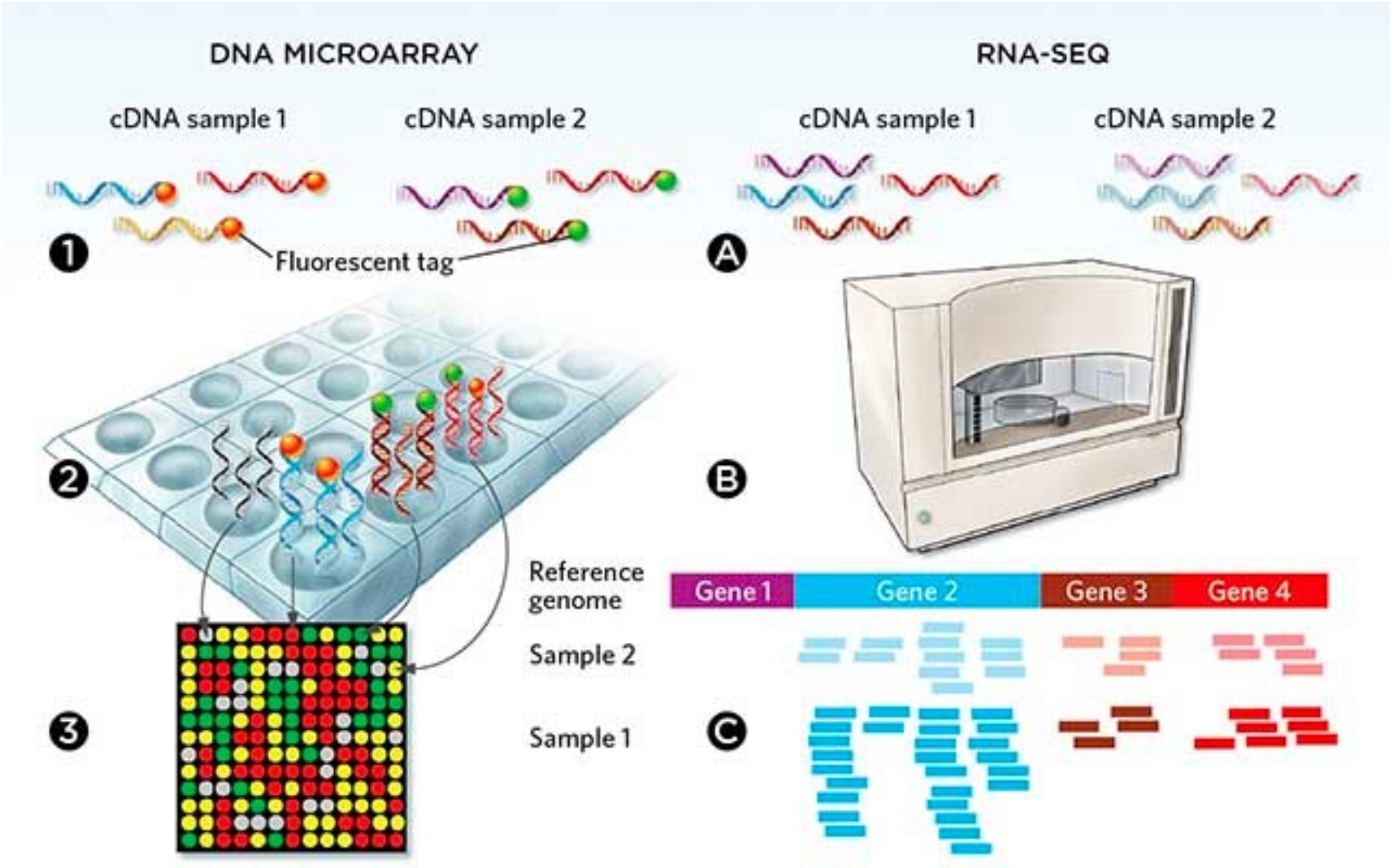


What is transcriptomics?

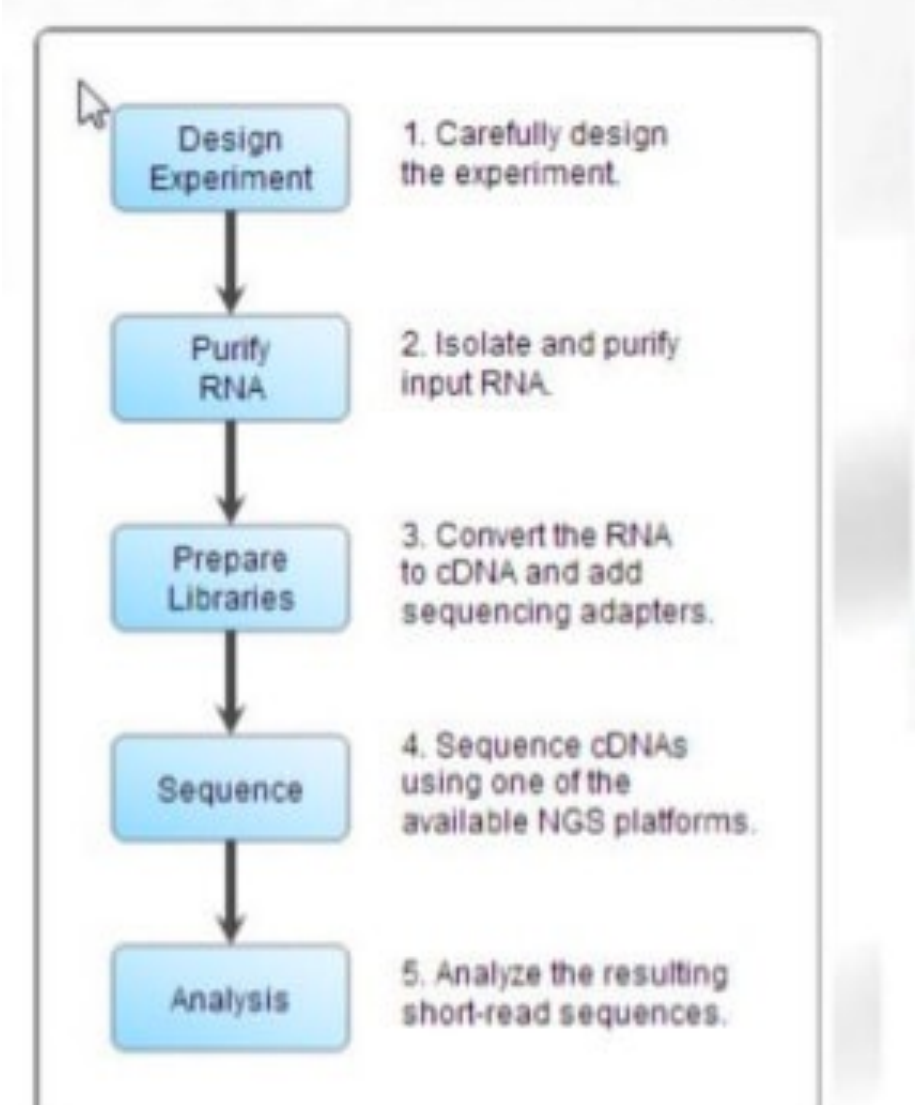


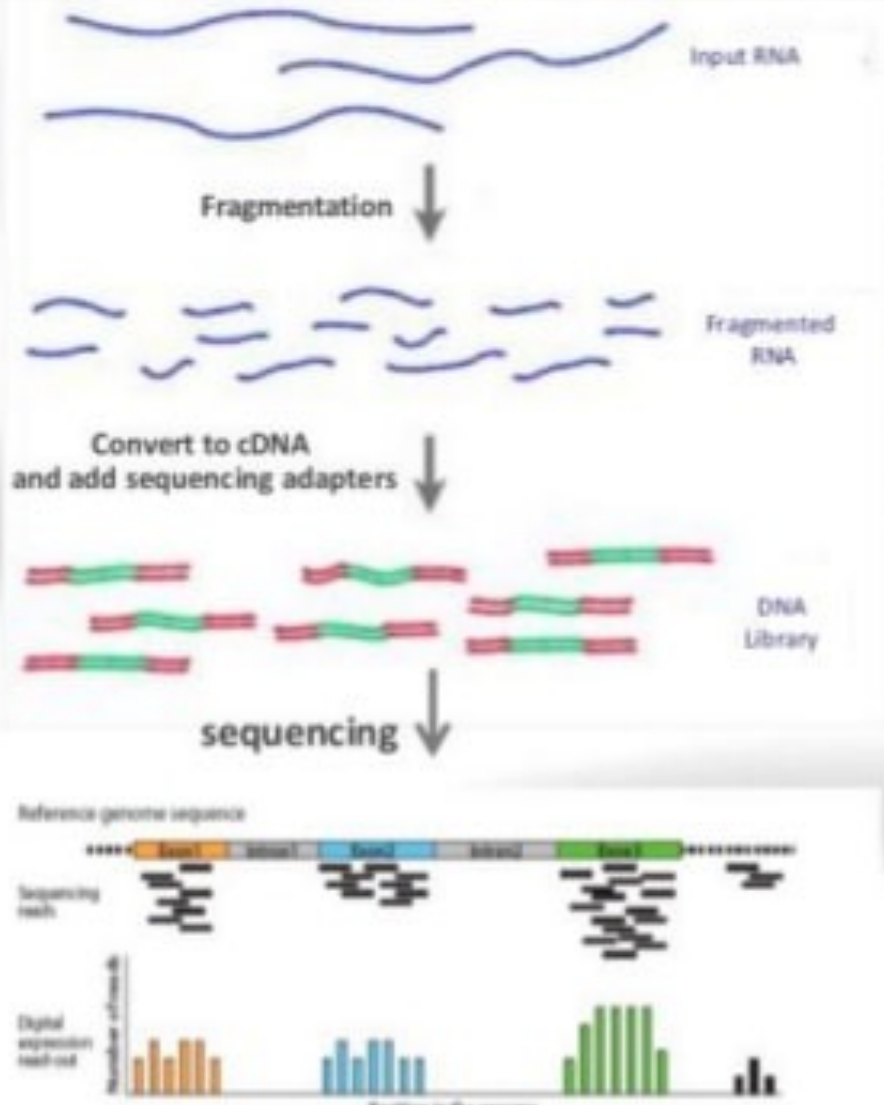
Transcriptomics is essentially the study of the transcriptome. Every complete set of RNA fragment transcribed from DNA in any organism is part of their transcriptome. The comparison among transcriptomes allows further understanding of differentially expressed genes throughout different cell populations.

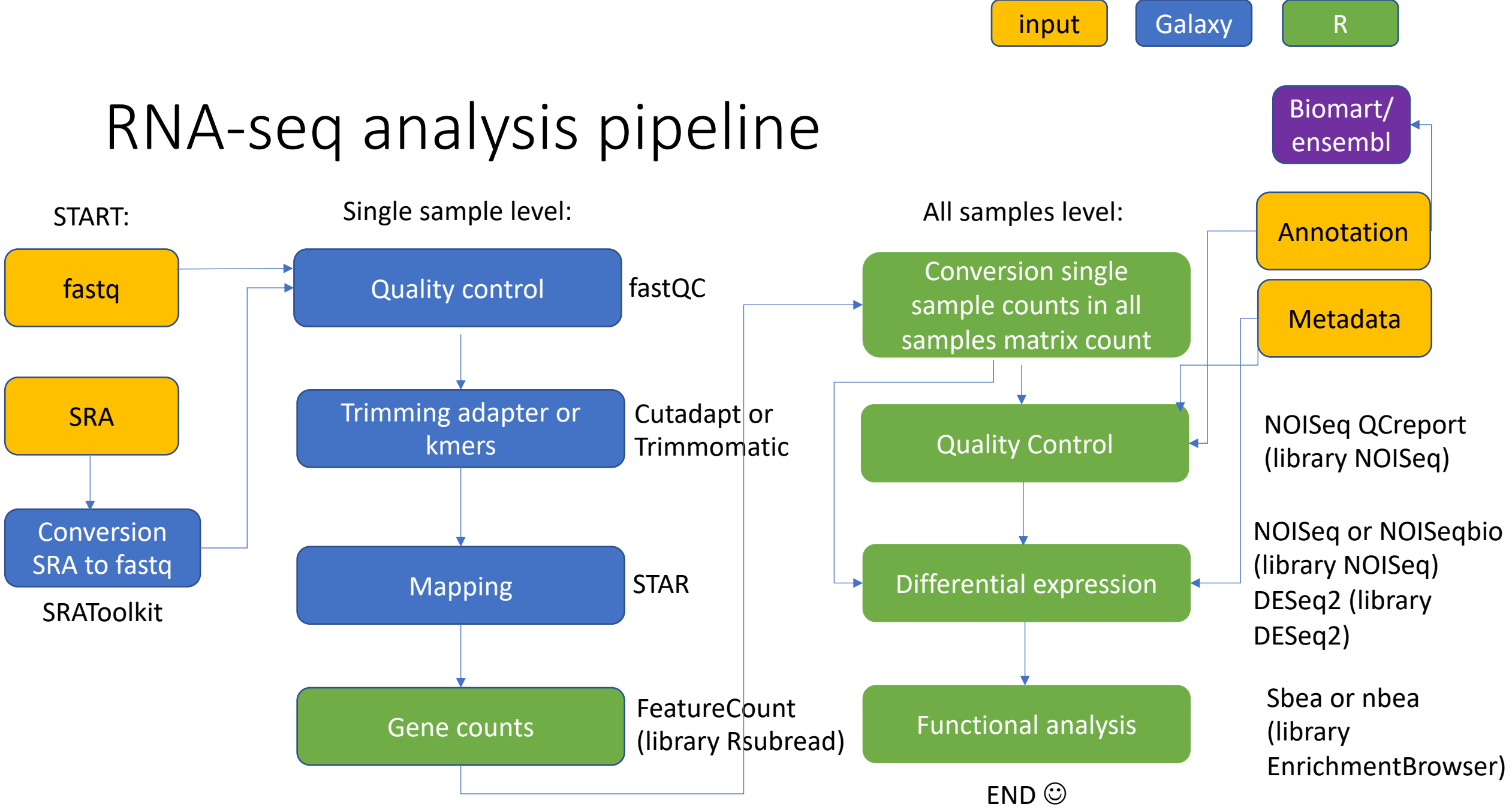
How to interrogate the transcriptome?



RNA-seq experimental workflow

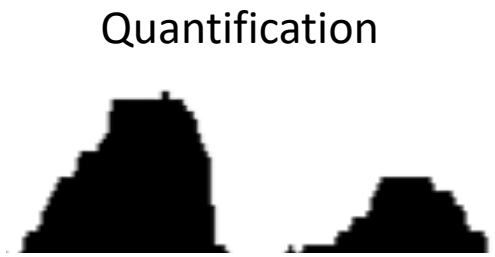




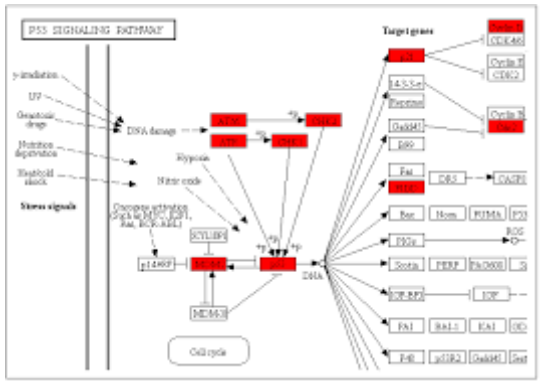

input
Galaxy
R
RNA-seq analysis pipeline
START:
Single sample level:
All samples level:
Biomart/ ensembl
Annotation
fastq
Quality control
fastQC
Conversion single sample counts in all samples matrix count
Metadata
SRA
Trimming adapter or kmers
Cutadapt or Trimmomatic
Quality Control
NOISeq QCreport (library NOISeq)
Conversion SRA to fastq
SRAToolkit
Mapping
STAR
Differential expression
NOISeq or NOISeqbio (library NOISeq) DESeq2 (library DESeq2)
Gene counts
FeatureCount (library Rsubread)
Functional analysis
Sbea or nbea (library EnrichmentBrowser)
END ☺

What can you do with RNA-seq?

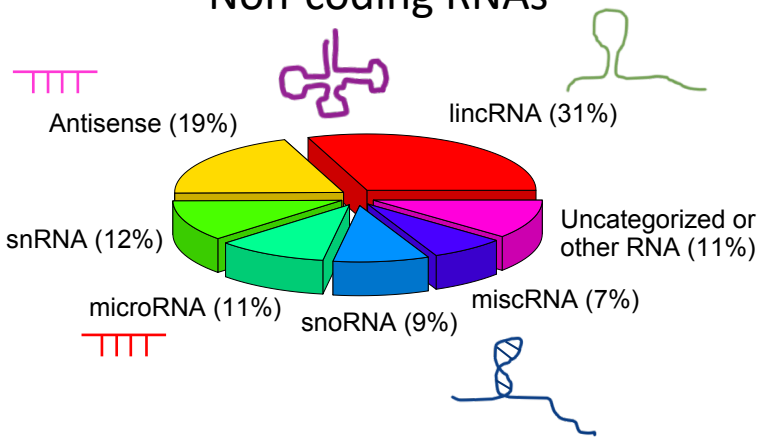
Quantification



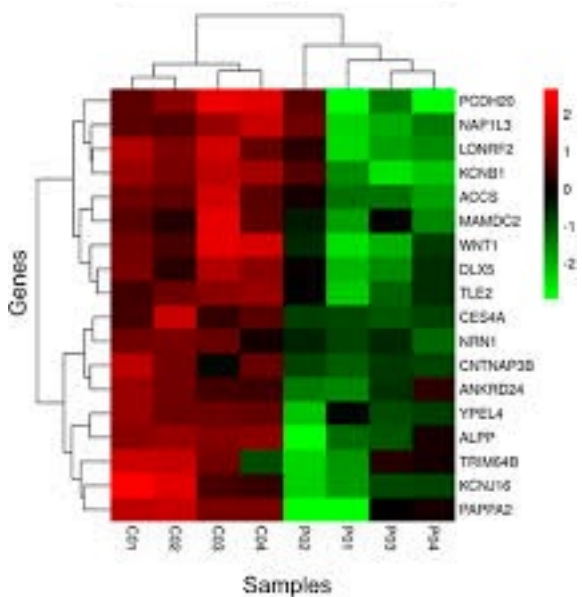
Pathway analysis



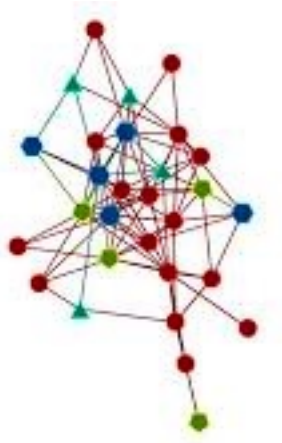
Non-coding RNAs



Differential expression

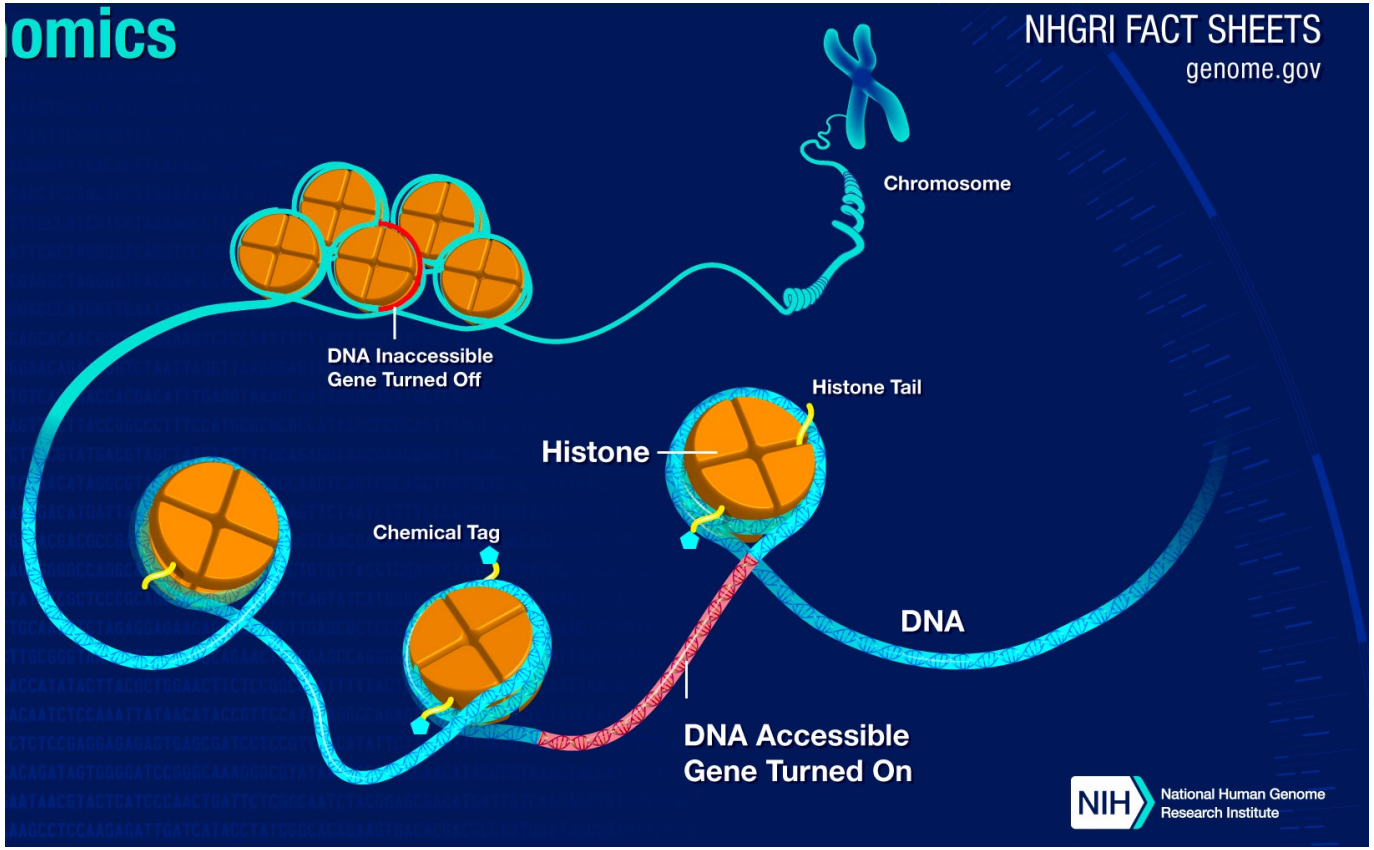


Network inference



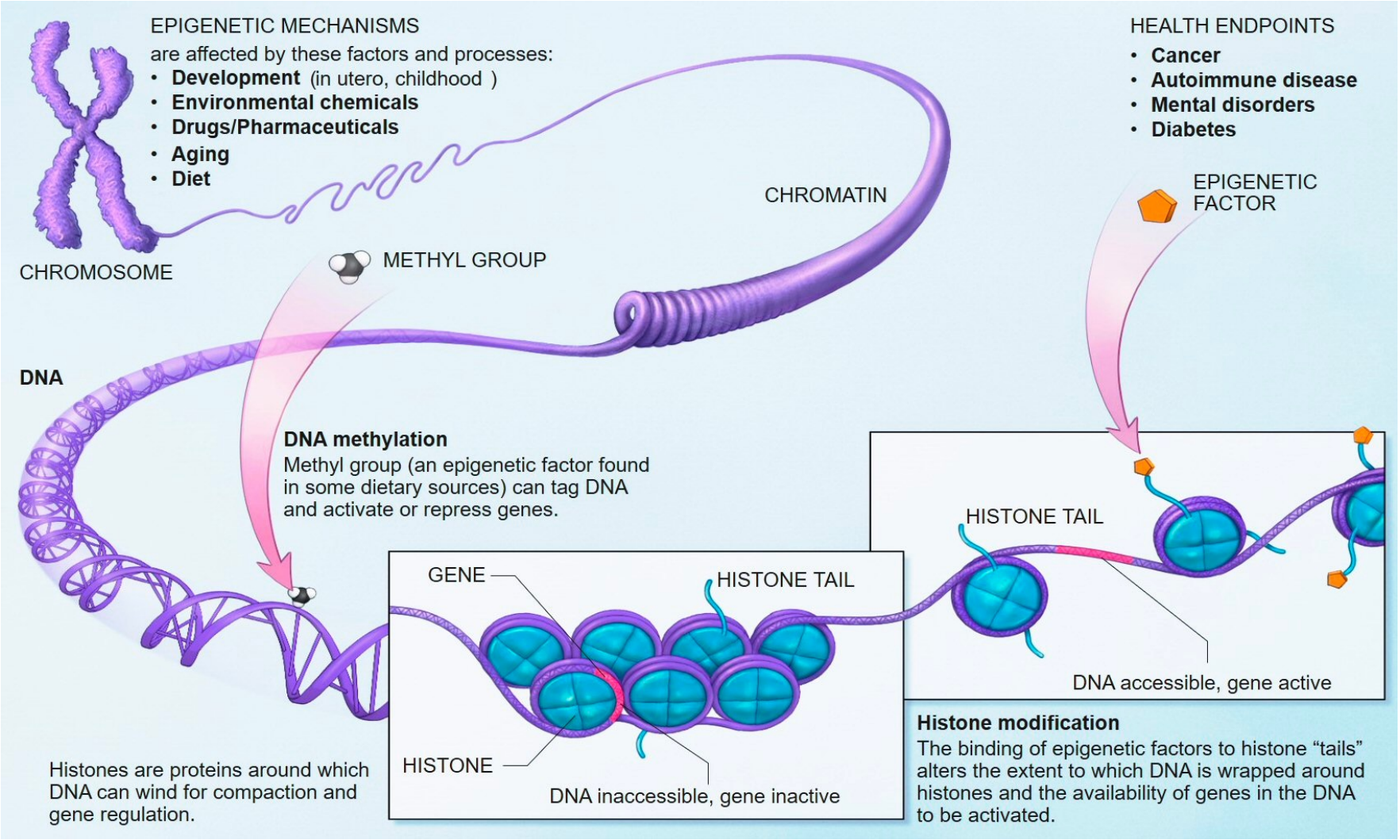
Epigenomics

> The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off, controlling the production of proteins in particular cells. When epigenomic compounds attach to DNA and modify its function, they are said to have "marked" the genome. These marks do not change the sequence of the DNA. Rather, they change the way cells use the DNA's instructions. The marks are sometimes passed on from cell to cell as cells divide. They also can be passed down from one generation to the next.

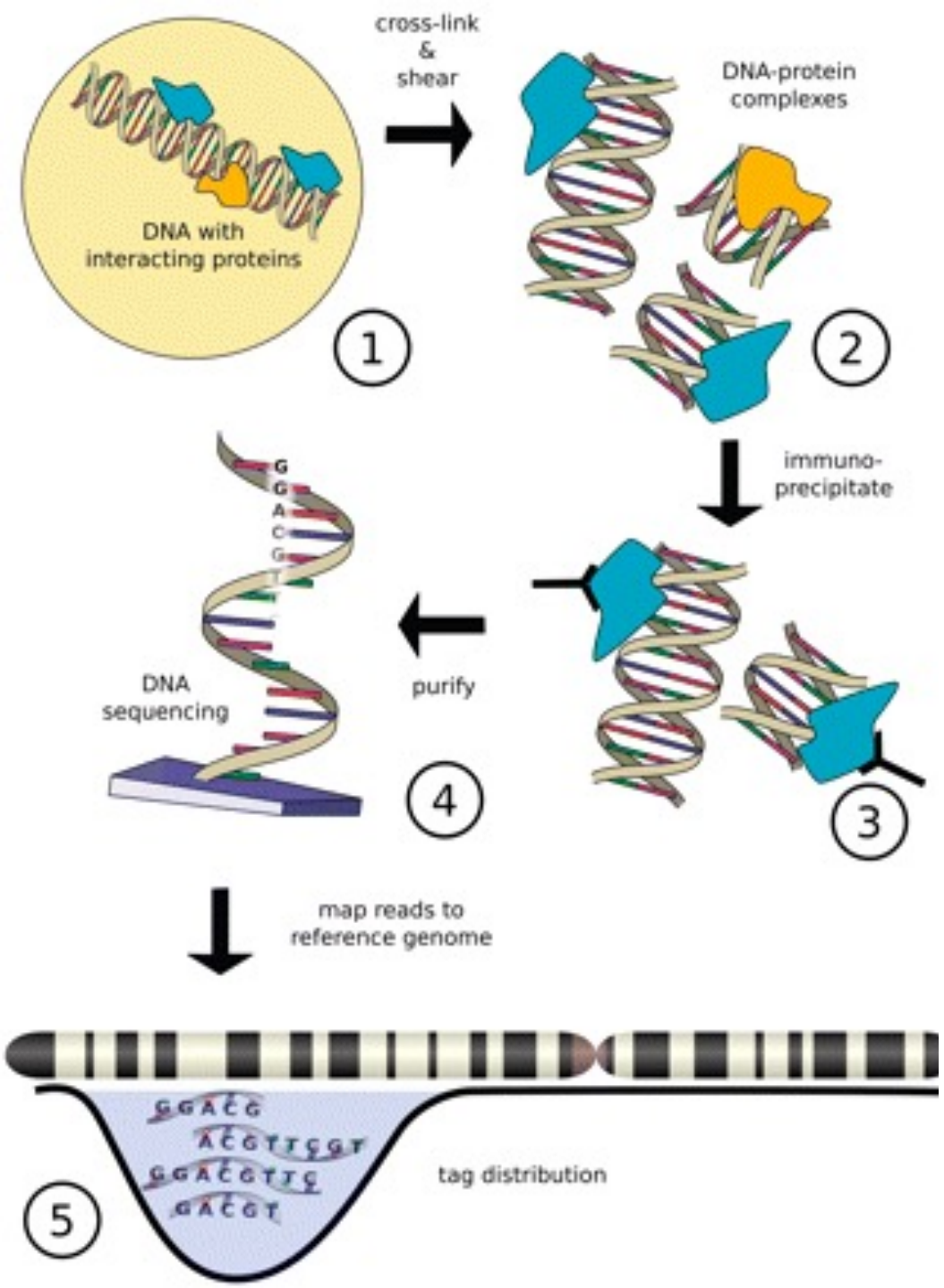


What makes up the epigenome?

The epigenome is the set of chemical modifications to the DNA and DNA-associated proteins in the cell, which alter gene expression, and are heritable (via meiosis and mitosis)

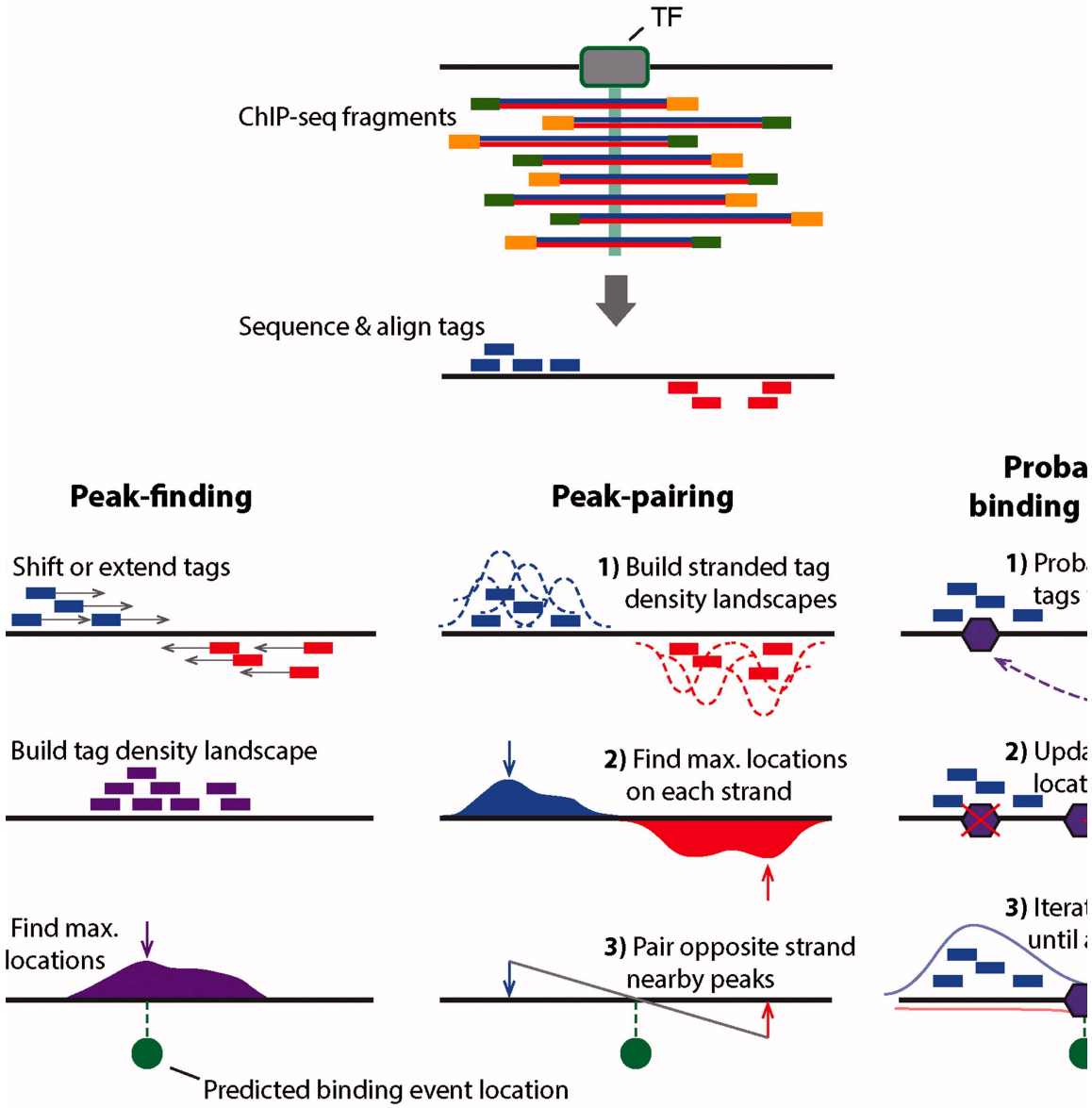


The modifications occur as a natural process of development and tissue differentiation, and can be altered in response to environmental exposures or disease.



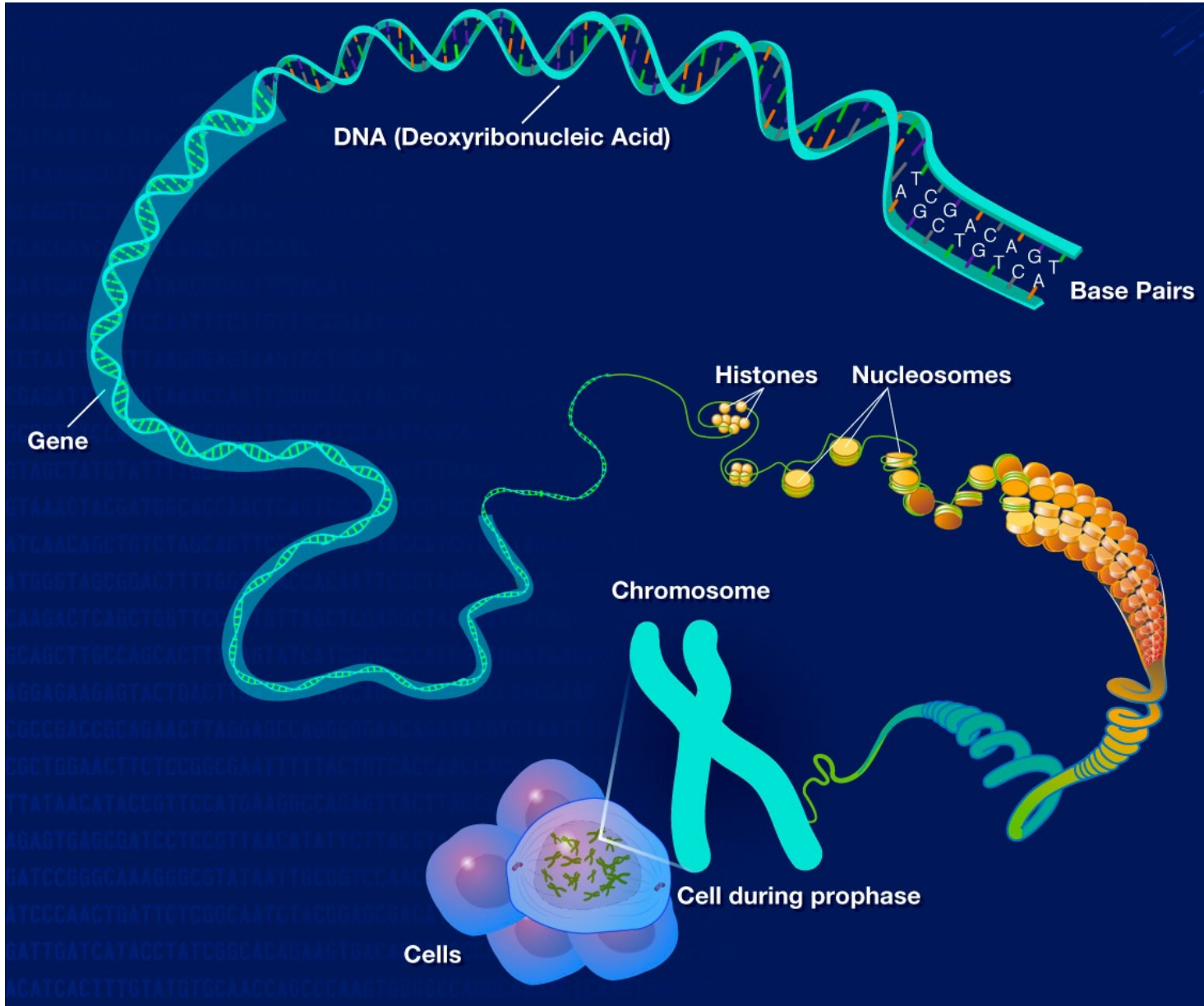
How do we explore the epigenome?

ChIP-seq data analysis

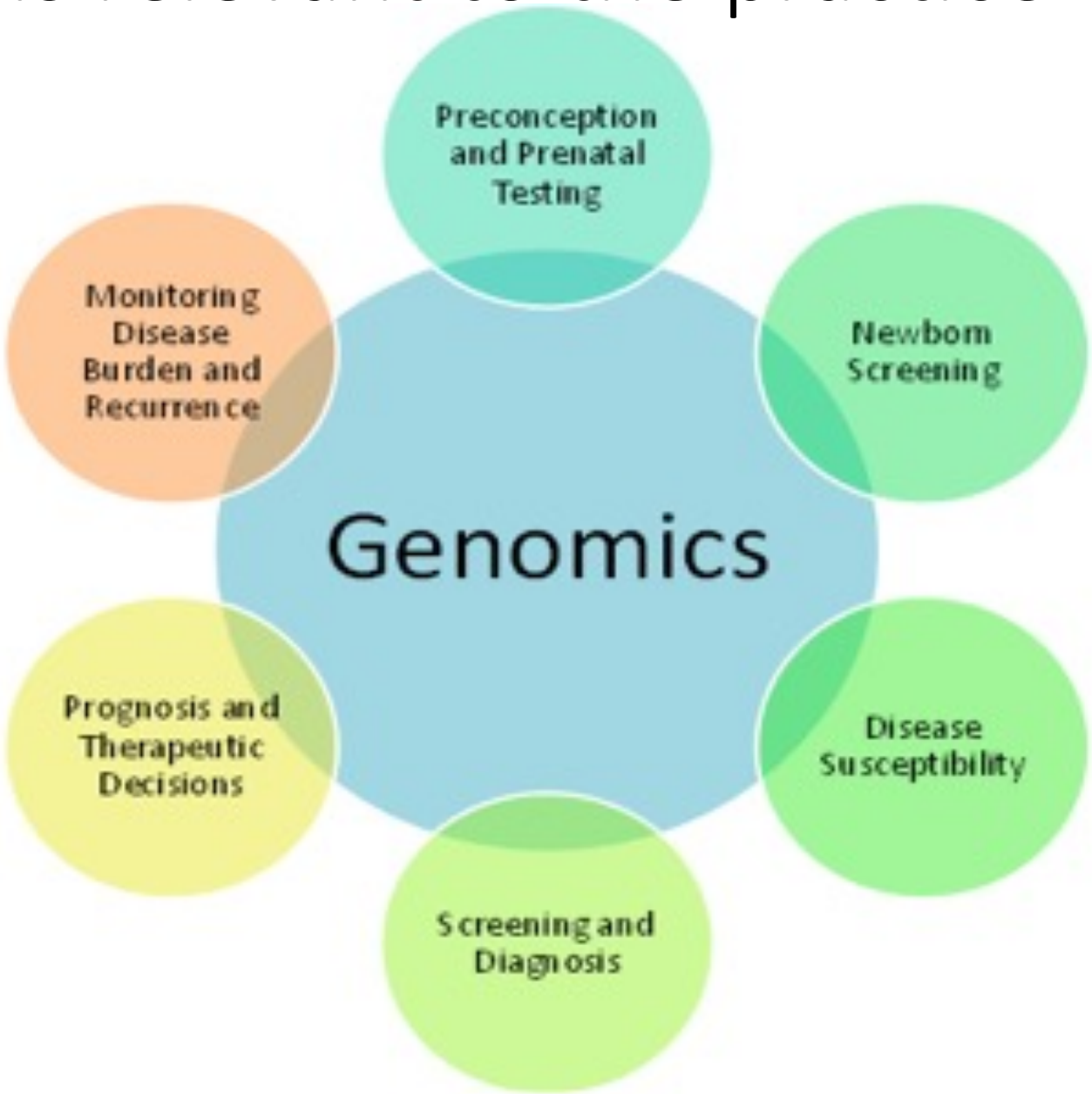


What is genomics?

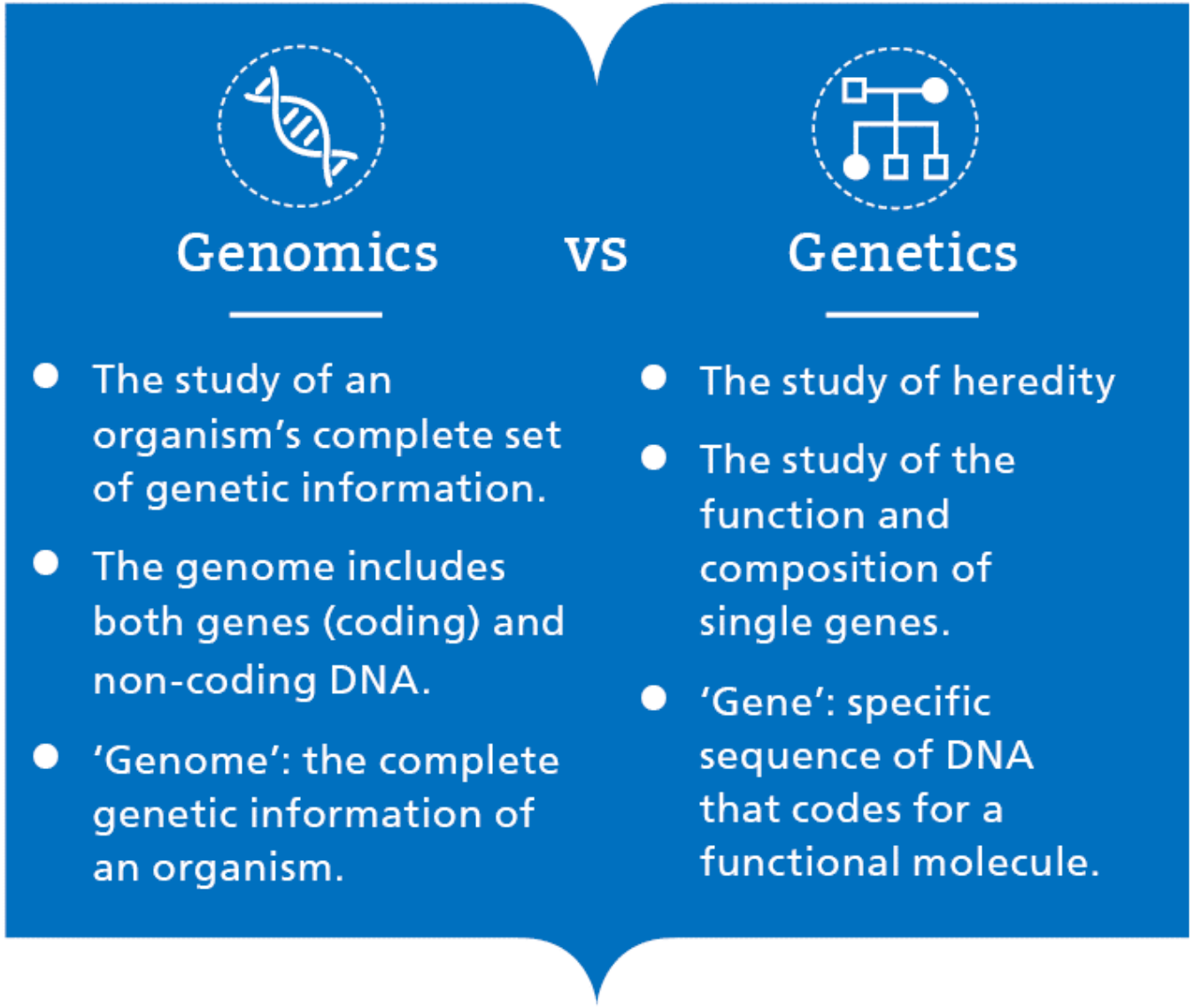
> The genome is the entire set of genetic instructions found in a cell. In humans, the genome consists of 23 pairs of chromosomes, found in the nucleus, as well as a small chromosome found in the cells' mitochondria. Each set of 23 chromosomes contains approximately 3.1 billion bases of DNA sequence (https://www.genome.gov/glossary/index.cfm?id=90). Genomic-based healthcare involves use of this genomic information about an individual as part of their clinical decisions (i.e., for diagnostic or therapeutic decision-making) and consideration of implications of that use.



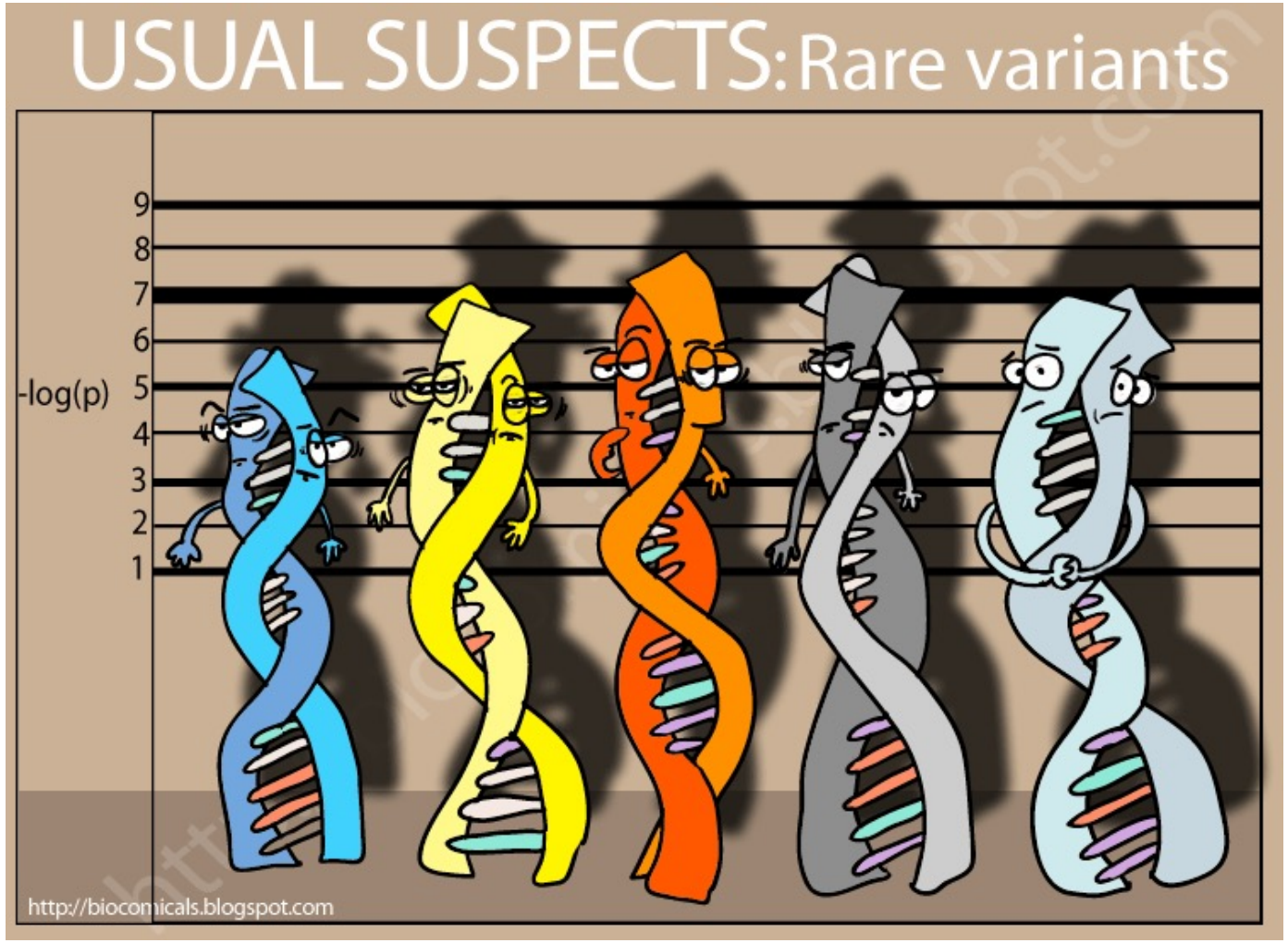
Genomics is relevant to the practice of all nurses.



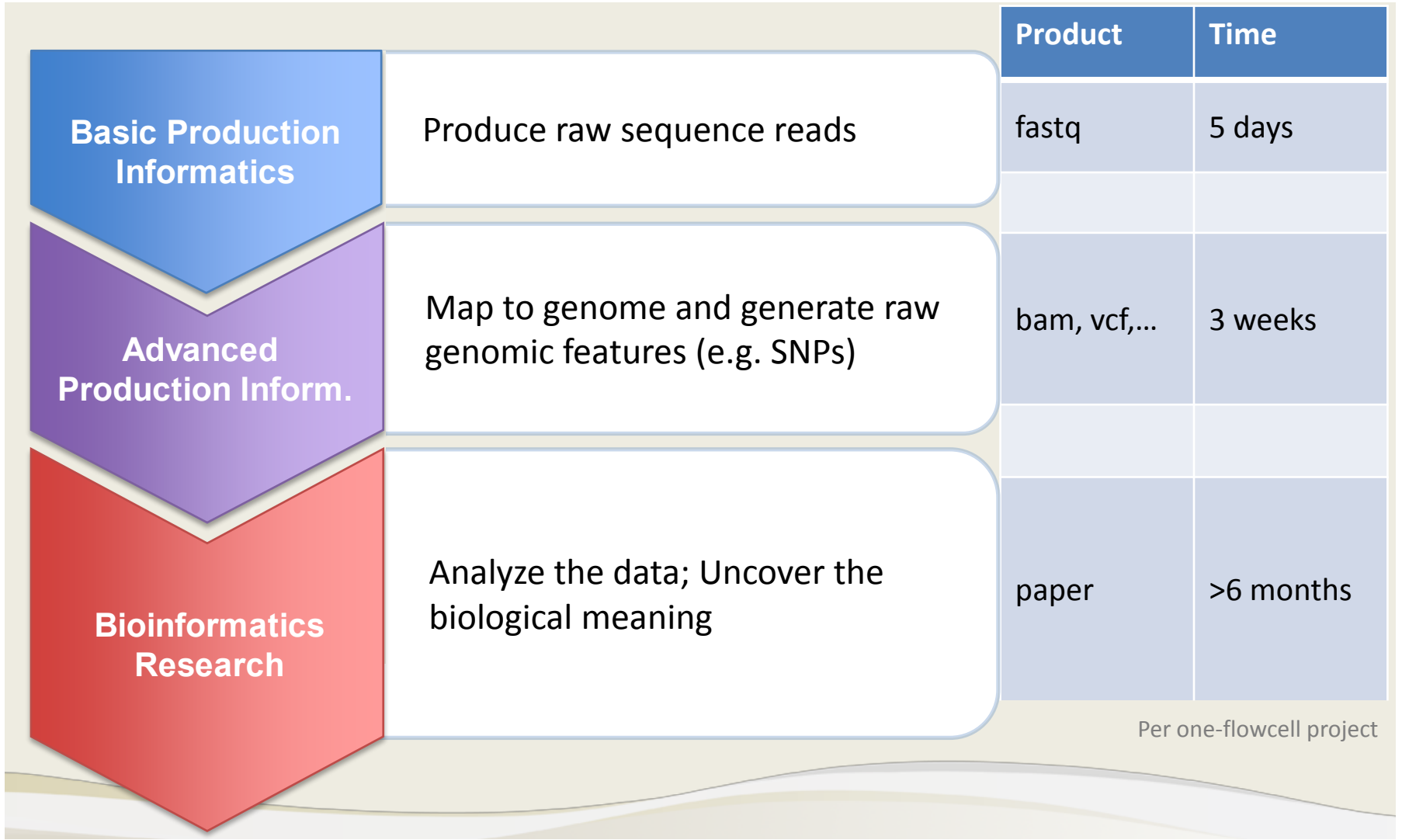
Genomics vs genetics



In practice: look for rare variants



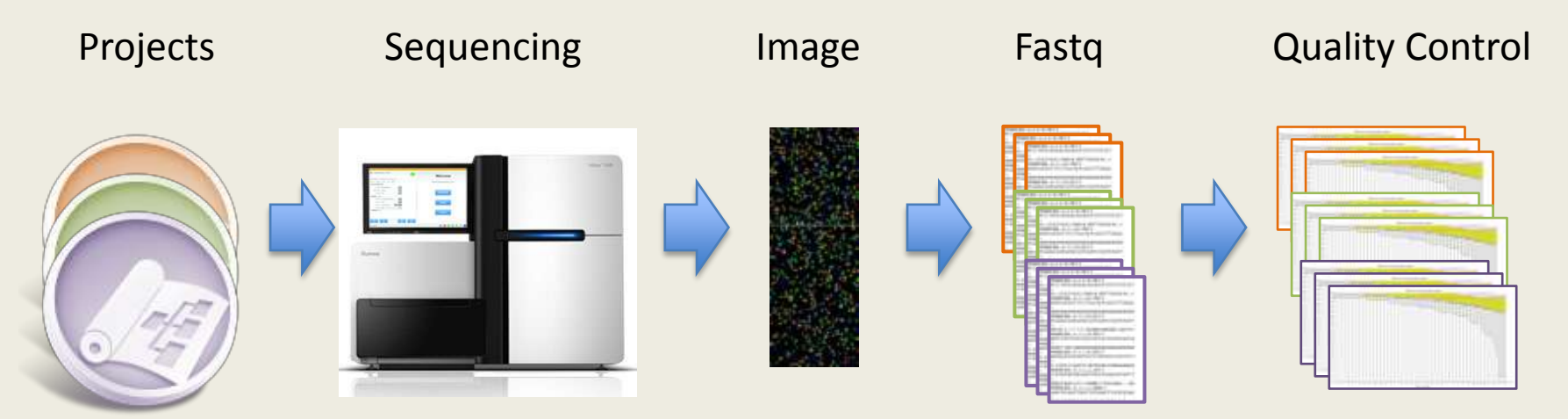
Data production



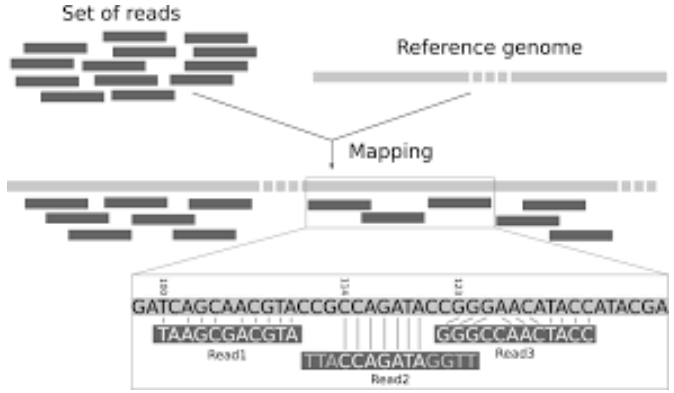
Product	Time
fastq	5 days
bam, vcf,…	3 weeks
paper	>6 months

Per one-flowcell project

Bioinformatic workflow

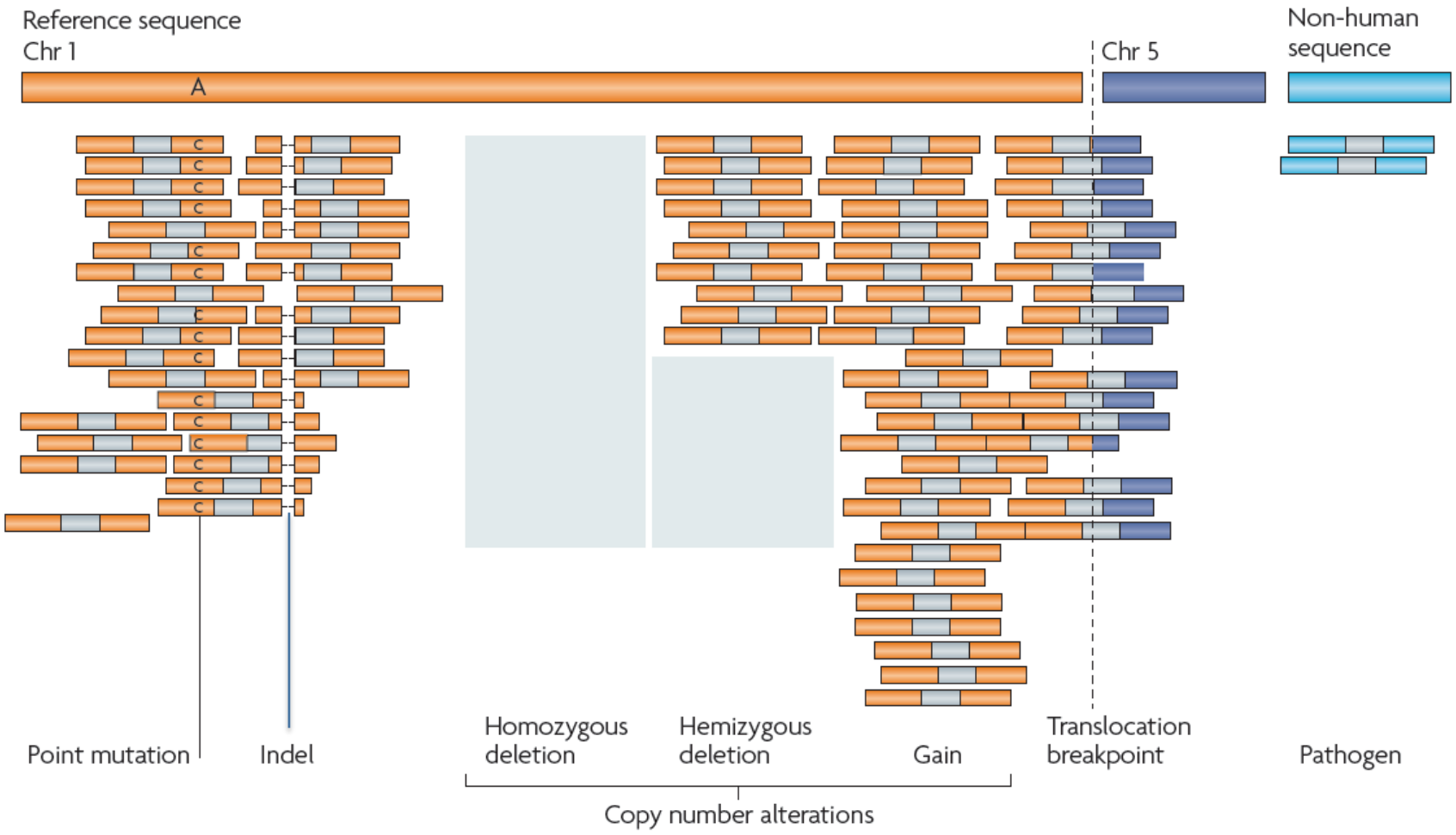


NGS processing



Variant calling, filtering and annotation

Types of variants



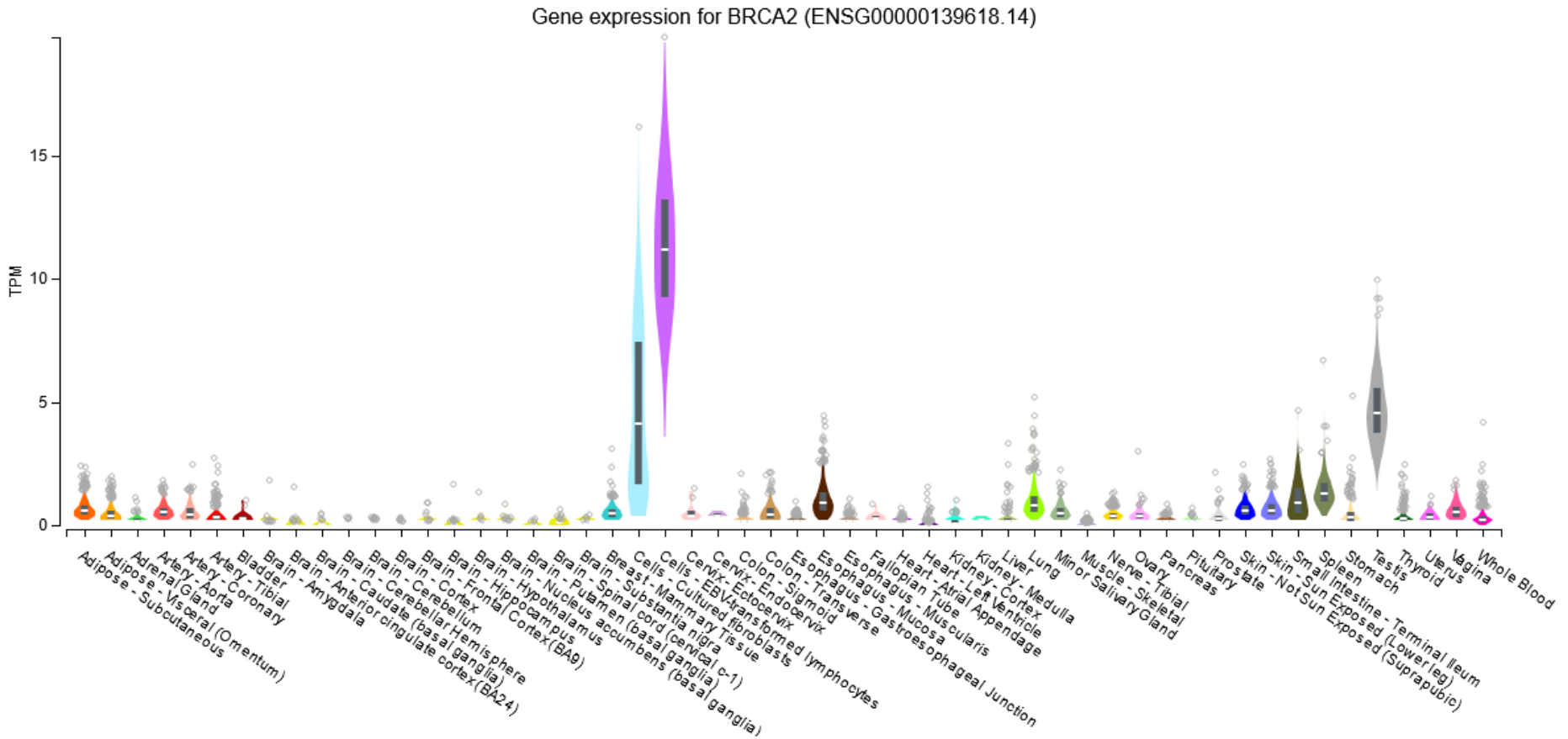
Meyerson, Gabriel and Getz. Nat. Rev. Genet. (2010)

SNV vs SNP

- **SNV** (**S**ingle **N**ucleotide **V**ariant)
  - ➢All nucleotides alteration without implication on the population frequency
- **SNP** (**S**ingle **N**ucleotide **P**olymorphism)
  - ➢Only variants shared by at least 1% of population

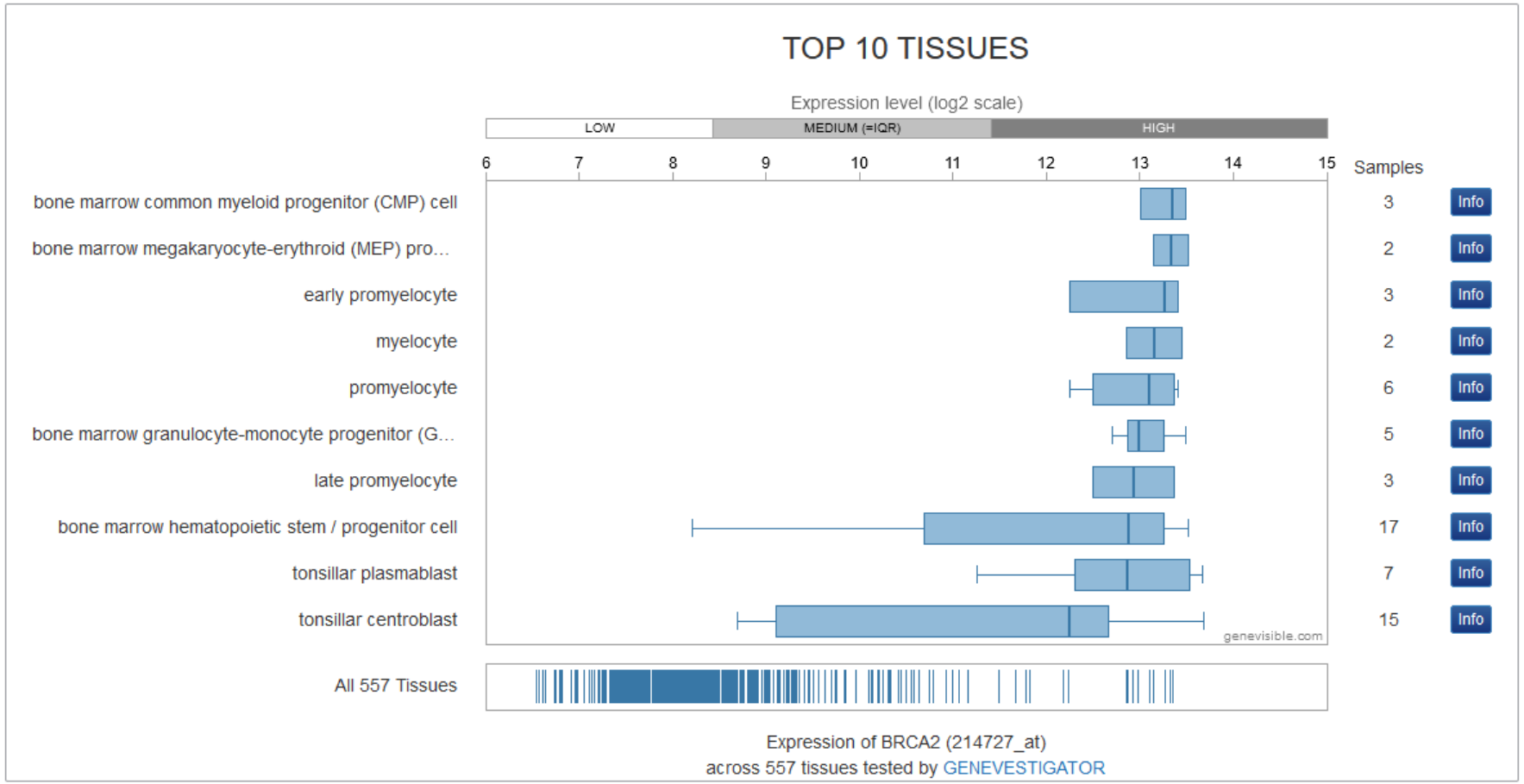
Omics data repositories

Gene Expression Information

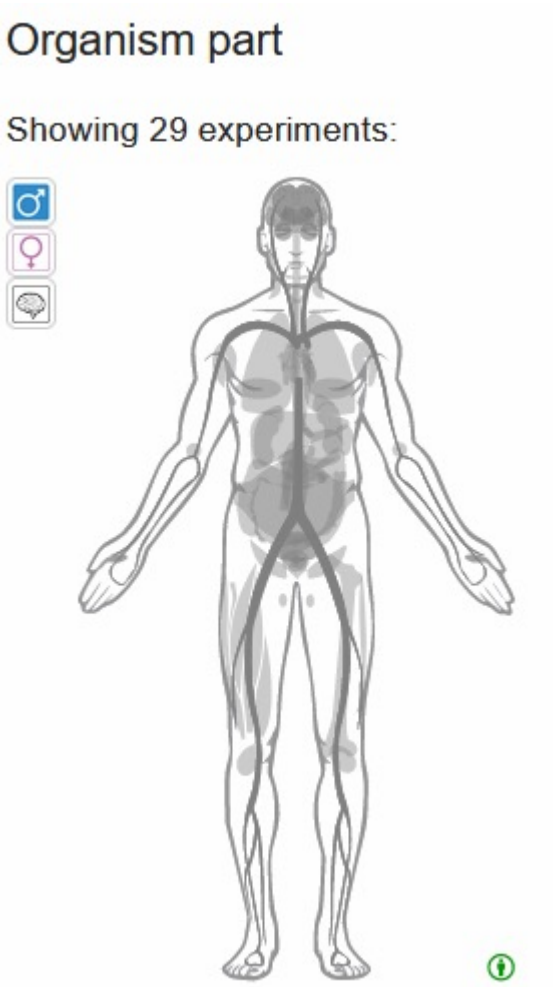


Gene Expression Information

Genevisible

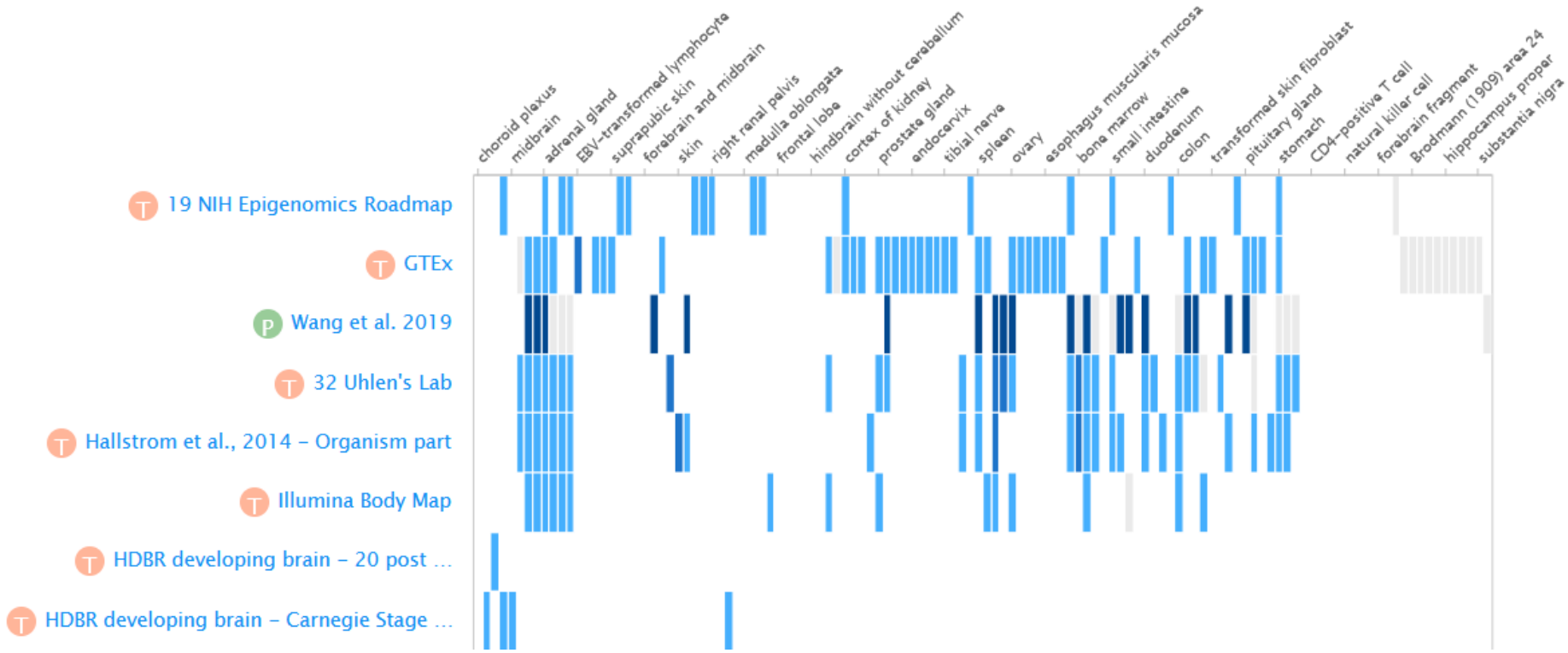


Gene Expression Information

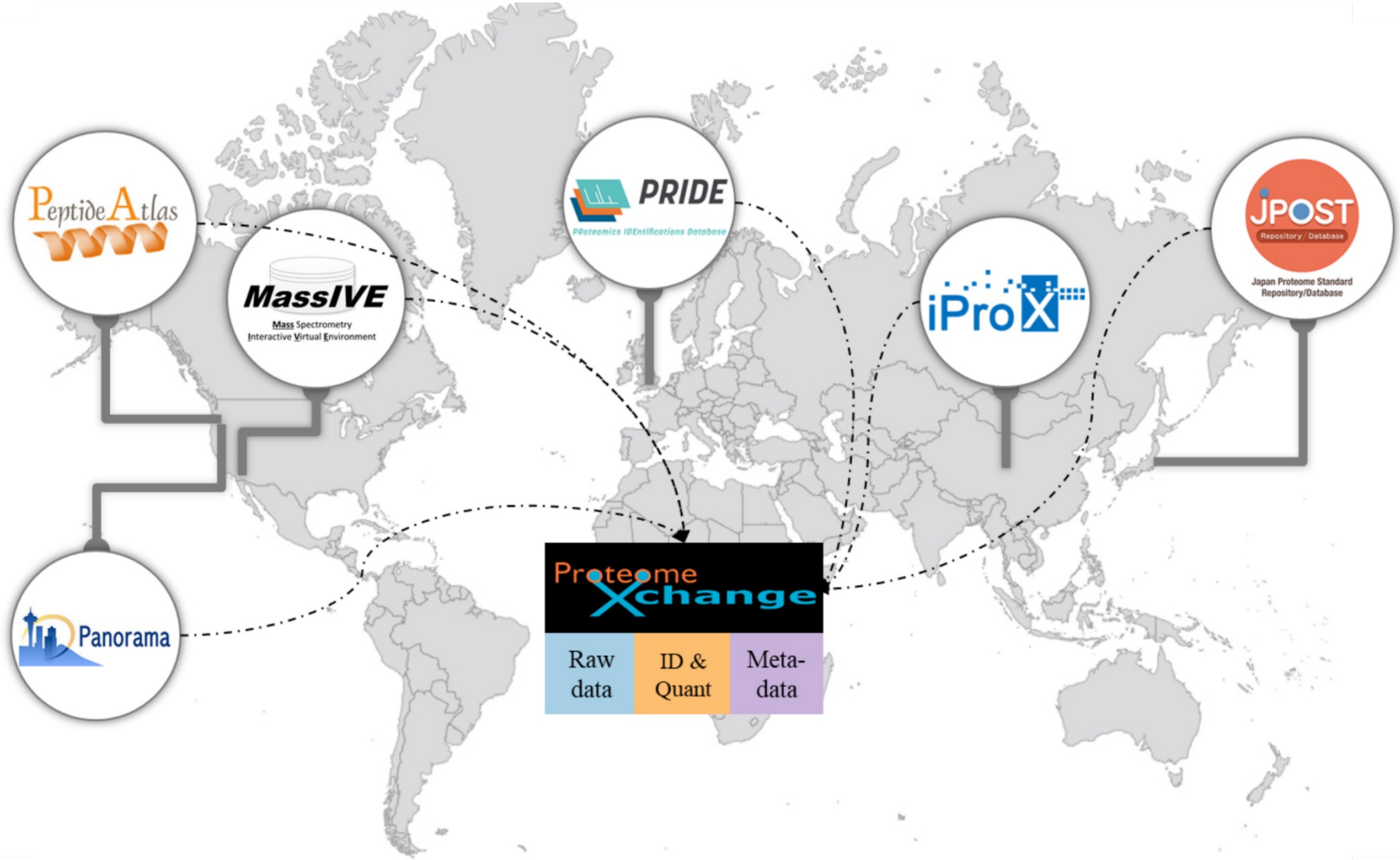


Expression Atlas

Gene expression across species and biological conditions



Data Repositories for Proteomics Mass Spec



Data Repositories for Proteomics Mass Spec

- Varying amounts of experimental annotation
- Good description of processing and preparation
- Raw data files available
  - Mass spec still uses a lot of proprietary vendor file formats
  - Open mzML format is defined but often not used
  - Converters exist but often lose information.

Plasma

Dataset Identifier	Title	Repos	Species	Instrument	Publication	LabHead	Announce Date	Keywords
PXD026962	Multi-omic Profiling of Plasma Identify Biomarkers and Pathogenesis of COVID-19 in Children	iProX	Homo sapiens	QTRAP 6500+	Dataset with its publication pending	Xi Zhou	2021-06-28	Multi-omic Profiling, COVID-19, Children,
PXD026928	Mycoplasma gallisepticum WhiA knockdown and overexpression	PRIDE	Mycoplasma gallisepticum S6	Q Exactive Plus	Dataset with its publication pending	Gleb Fisunov	2021-06-25	mycoplasma, transcription factor, WhiA, overexpression, knockdown,
PXD022361	Recombinant SWATH library for identification of low abundant human plasma proteins	MassIVE	Homo sapiens	TripleTOF 6600	Ahn et al. (2021)	Prof Mark S. Baker	2021-06-24	SWATH, Recombinant Protein, Plasma Proteome,
PXD021581	Prognostic accuracy of Mass Spectrometric Analysis of Plasma in COVID-19	PRIDE	Homo sapiens	LTQ Orbitrap Velos	Dataset with its publication pending	Giuseppe Palmisano	2021-06-21	Sars-cov-2, Covid-19, COVID-19, SARS-CoV-2, Mass spectrometry, Biomarker, Plasma, Prognosis,

PXD026962

PXD026962 is an **original dataset** announced via ProteomeXchange.

Dataset Summary

Title	Multi-omic Profiling of Plasma Identify Biomarkers and Pathogenesis of COVID-19 in Children
Description	Although people of all ages are susceptible to COVID-19, children usually develop less severe disease than adults. Little is known about the pathogenesis of COVID-19 in children. Herein, we conduct the plasma proteomic and metabolomic profiling of a cohort of COVID-19 children patients with mild symptoms, and uncovered that many proteins involved in immune response are significantly up-regulated in a stronger extent than in adults with COVID-19. Interestingly, more molecules involved in protective processes of reducing inflammation are also stimulated to antagonize the deleterious effect in both proteomic and metabolomic levels. By developing a machine learning-based pipeline, we prioritize two set of biomarker combinations, and identify 5 proteins and 5 metabolites as potentially children-specific biomarkers. Further experiments demonstrate these protective metabolites not only inhibit the expression of pro-inflammatory factors, but also suppress the viral replication. Taken together, our study not only discover the protective mechanisms in children with COVID-19, but also shed light on potential therapies targets for treating COVID-19.
HostingRepository	iProX
AnnounceDate	2021-06-28
AnnouncementXML	Submission_2021-06-28_01:26:39.414.xml
DigitalObjectIdentifier	
ReviewLevel	Peer-reviewed dataset
DatasetOrigin	Original dataset
RepositorySupport	Unsupported dataset by repository
PrimarySubmitter	Yang Qiu
SpeciesList	scientific name: Homo sapiens; NCBI TaxID: 9606;

Project Information

Project ID

IPX0002673000

ProteomeXchange ID

PXD026962

Project Title

Multi-omic Profiling of Plasma Identify Biomarkers and Pathogenesis of COVID-19 in Children

XML File

PX_IPX0002673000.xml Download

Download All Files (7.55G)

Aspera Download *recommended

Http Download

Metabolite Mass Spectrometry

- Similar concepts to protein mass spec
- Range of starting material
  - Serum
  - Urine
  - Cerebrospinal fluid
  - Saliva
- Different separations
- Up to 5000 different metabolites to find

Data Repository for Metabolomics Data

MoNA - MassBank of North America

- Reference spectra for biological molecules
  - Used for searching and quantitation

- Experimental datasets of MassSpec Studies
  - Used to answer biological questions
  - Also provides visualisations and tools

TCGA

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

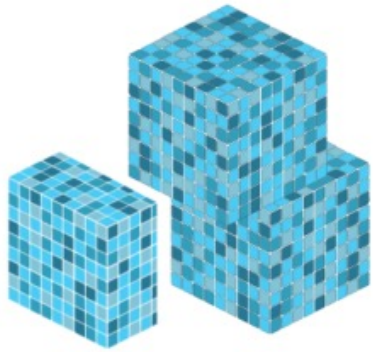
TCGA BY THE NUMBERS

TCGA produced over

2.5

PETABYTES

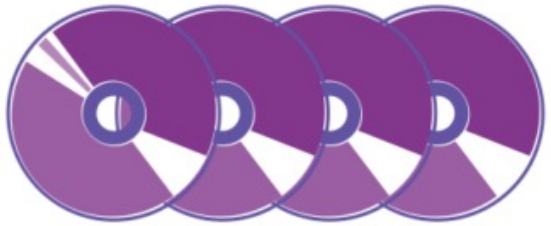
of data



To put this into perspective, **1 petabyte** of data is equal to

212,000

DVDs



TCGA data describes

33

DIFFERENT TUMOR TYPES

...including

10

RARE CANCERS

...based on paired tumor and normal tissue sets collected from

11,000

PATIENTS

...using

7

DIFFERENT DATA TYPES

Table 1. Key portals for accessing publicly available multi-omics datasets

Name	URL	Omic and other data types	Notes
TCGA (Campbell et al., 2020)	https://portal.gdc.cancer.gov/	• Genomics • Epigenomics • Transcriptomics	• Tumor data • Large coverage of tumors
ICGC (Campbell et al., 2020)	https://dcc.icgc.org/	• Genomics • Transcriptomics	• Tumor data • Powerful online analytics tools
CPTAC	https://cptac-data-portal.georgetown.edu/cptacPublic/	• Proteomics	• Tumor data • The largest proteomic data portal
COSMIC Cell Lines (Iorio et al., 2016)	https://cancer.sanger.ac.uk/cell_lines	• Genomics • Epigenomics • Transcriptomics • Drug response • CRISPR-Cas9 screen	• Cancer cell line data • Manually curated • Large coverage of cell lines
DepMap (Broad, 2020)	https://depmap.org/portal/	• Genomics • Epigenomics • Transcriptomics • Proteomics • Drug response • CRISPR-Cas9 screen	• Cancer cell line data • Large coverage of omic types • Powerful online tools
COSMIC (Tate et al., 2019)	https://cancer.sanger.ac.uk/cosmic	• Genomics • Epigenomics • Transcriptomics	• Tumor data • Manually curated • Focus on genomics • Overlap with other portals