Bienvenue au module #11

Cancer et génomique: des big data aux modèles prédictifs

université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

Planning

	Lundi 22 janvier - Salle: 21 B2M
09:00-10:30	Technologies et données NGS en cancérologie. Daniel GAUTHERET
10:45-12:15	TP Galaxy I: Cas d'étude RNA-seq (contrôles qualité, alignements des séquences sur le génome de référence et quantification de l'expression des gènes). Daniel GAUTHERET
13:30-17:00	TP Galaxy II : Cas d'étude RNA-seq (création d'un workflow, matrice d'expression des gènes et analyse différentielle) D GAUTHERET

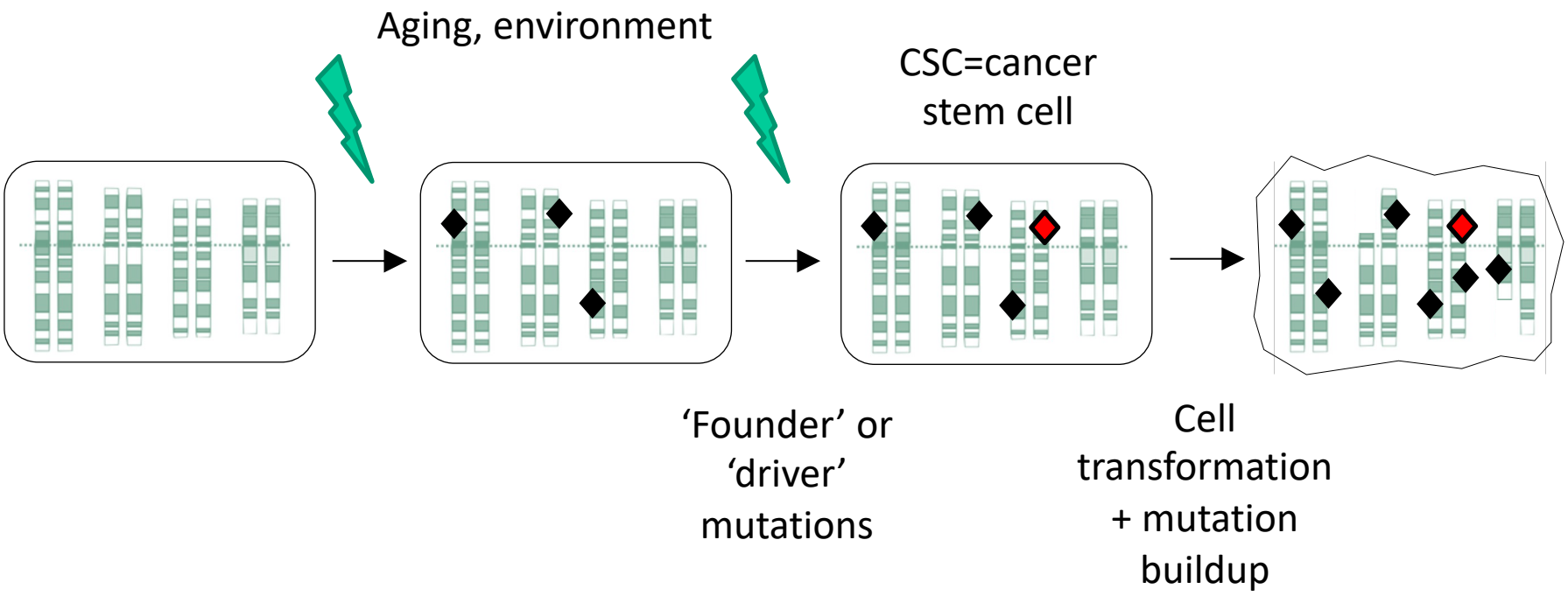
	Mardi 23 janvier - Salle 21 B2M
9:00-10:00	Pourquoi utiliser les methodes d'apprentissage automatique en oncologie personalisée? Julien VIBERT (Gustave Roussy)
10:00-11:00	Methodes de classification supervisée. Yoann PRADAT (Gustave Roussy)
11:30-12:30	Exemple d'un projet de classification médicale: Julien VIBERT (Gustave Roussy)
13:30-17:00	TP: Classification supervisée sur des données d'expression issues de TCGA. Julien VIBERT (Gustave Roussy) et Yoann PRADAT (Gustave Roussy).

	Mercredi 24 janvier - Salle 21 B2M
9:00-10:00	Méthodes d'analyses génomiques et de modélisation de survie. Yoann PRADAT (Gustave Roussy)
10:15-11:00	Example d'un projet de prédiction de réponse aux immunothérapies: Roger SUN (Gustave Roussy)
11:15-12:00	Exemple d'un projet de prédiction de survie: Elsa BERNARD (Gustave Roussy)
13:30-17:00	TP: Factorisation non-négative et modélisation de survie. Yoann PRADAT (Gustave Roussy).

Evaluation:
Protocole de TP commenté

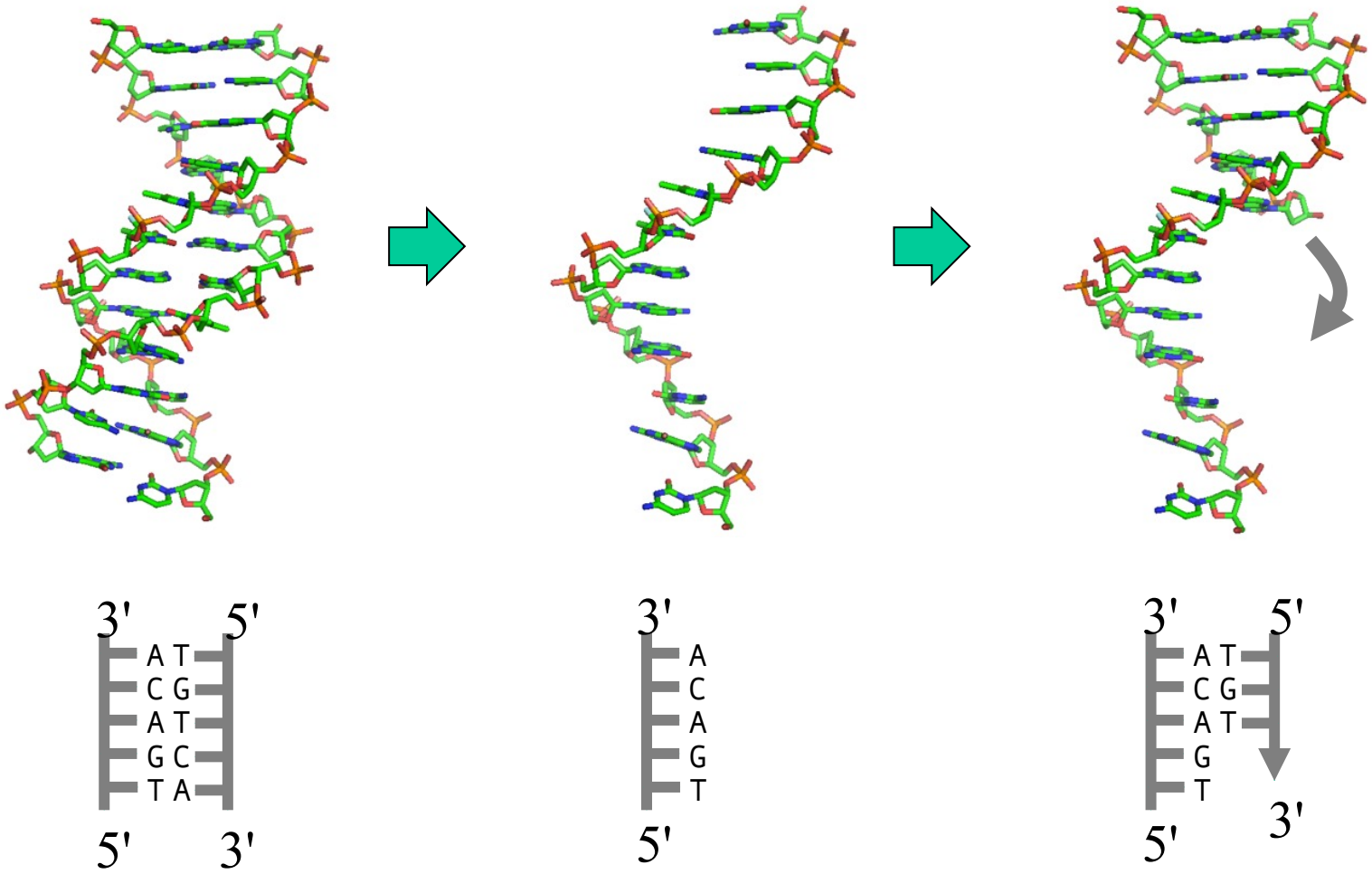
Introduction

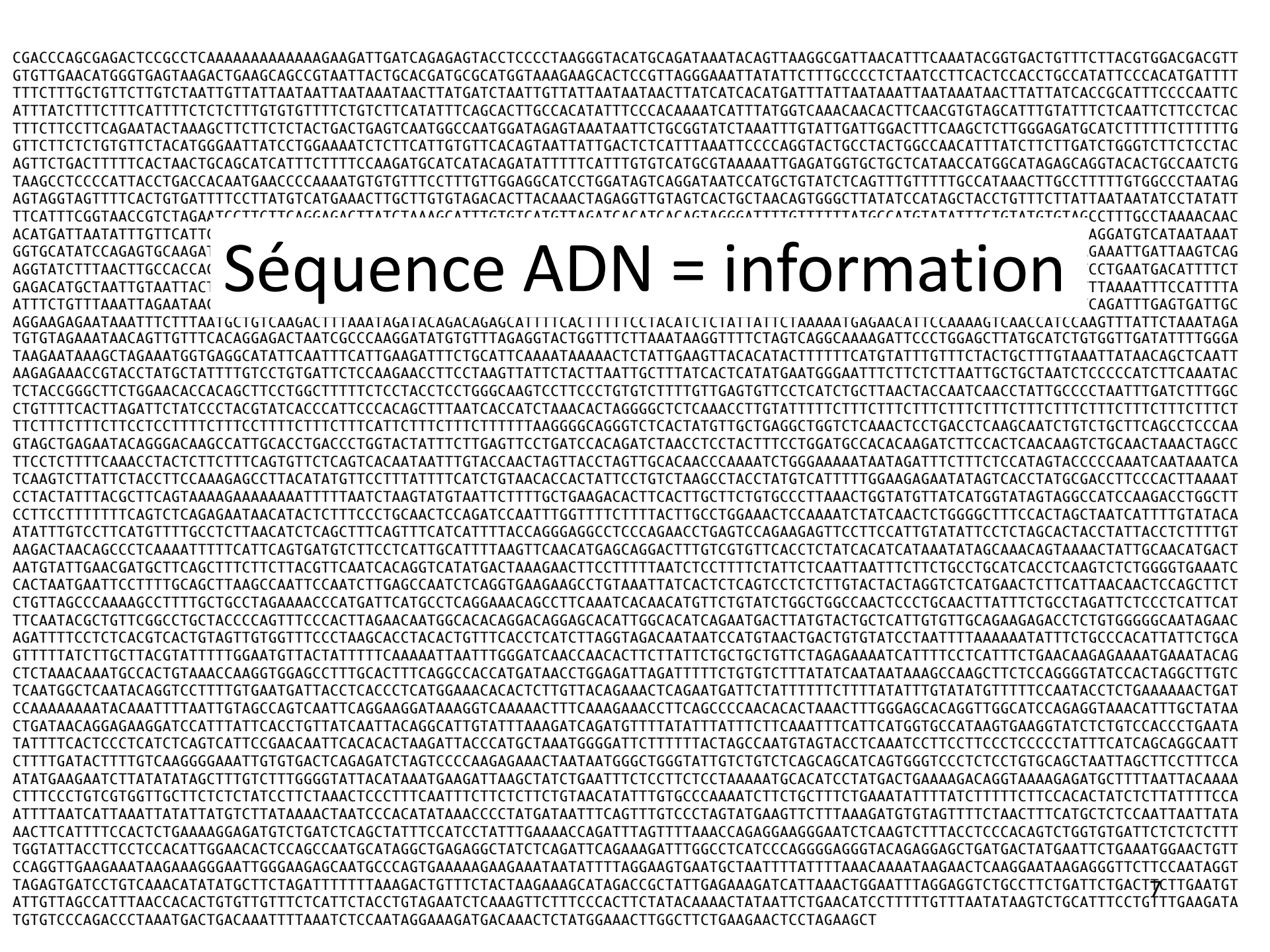
- Le cancer: une maladie du génome



Rappel

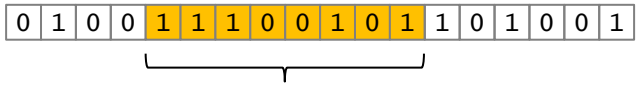
Information transmission in the cell




Séquence ADN = information

Biologie = une science de l'information

Binaire: base 2
2 symboles: 0 1



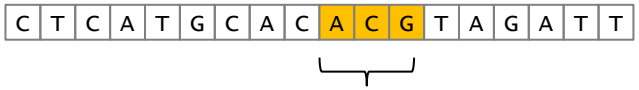
8 bits = 1 octet

Encoder un texte

clé	valeur
01000100	D
01000101	E
01000110	F
01000111	G
01001000	H
...	...

Code ASCII

ADN: base 4
4 symboles: A T G C



3 nucléotides = 1 codon

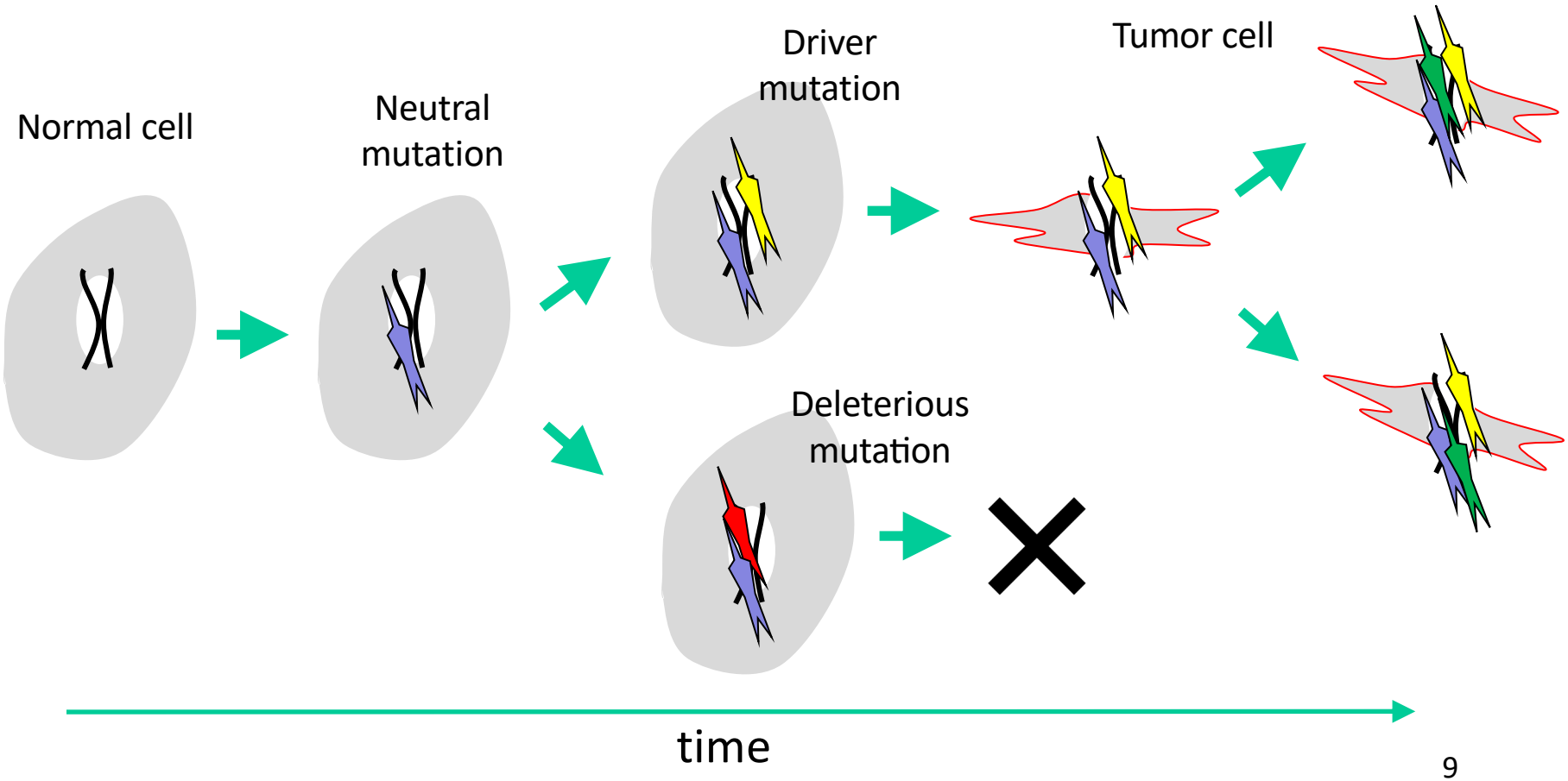
Encoder une protéine

clé	valeur
ACG	Thr
AAG	Lys
GCG	Ala
TAG	stop
ATG	Met
...	...

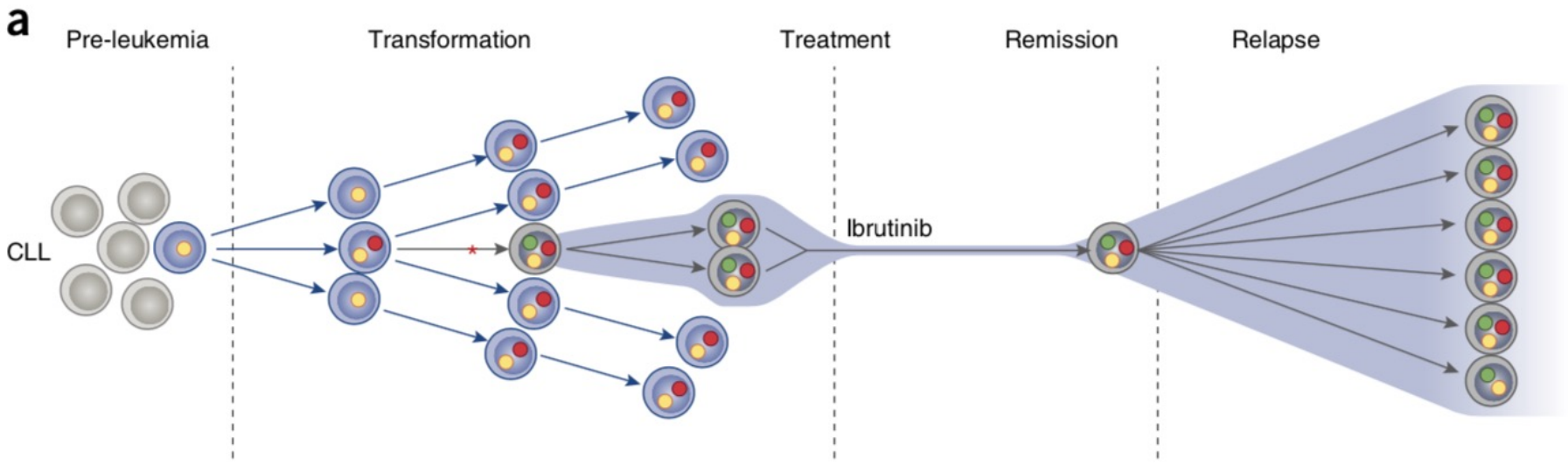
Code génétique

D'après **Sacha Schutz** https://github.com/dridk

Le cas du cancer: des mutations somatiques apparaissant par évolution clonale



Clonal evolution of cancer

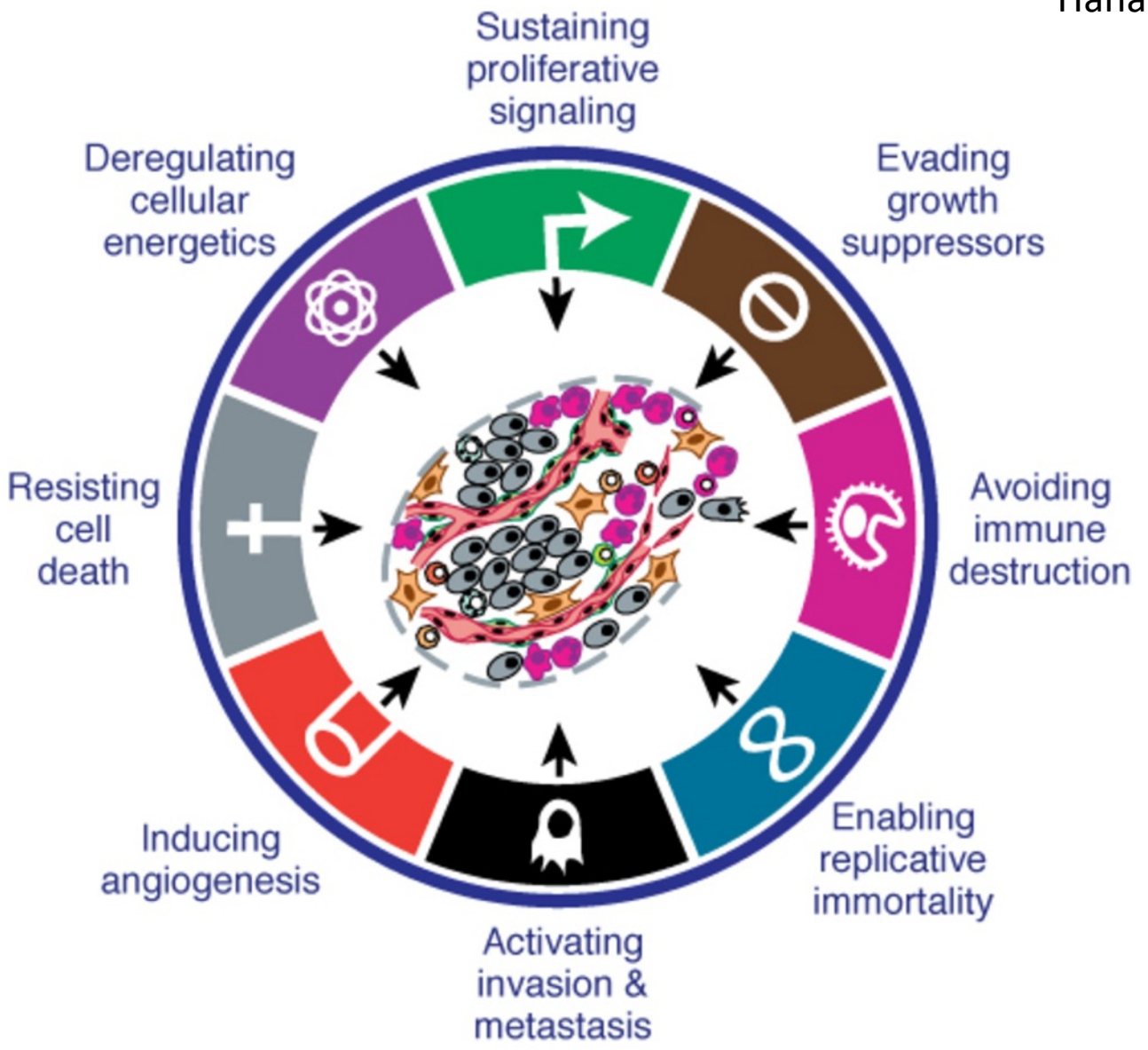


Ferrando & Lopez-Otin, Nature Med. 2017

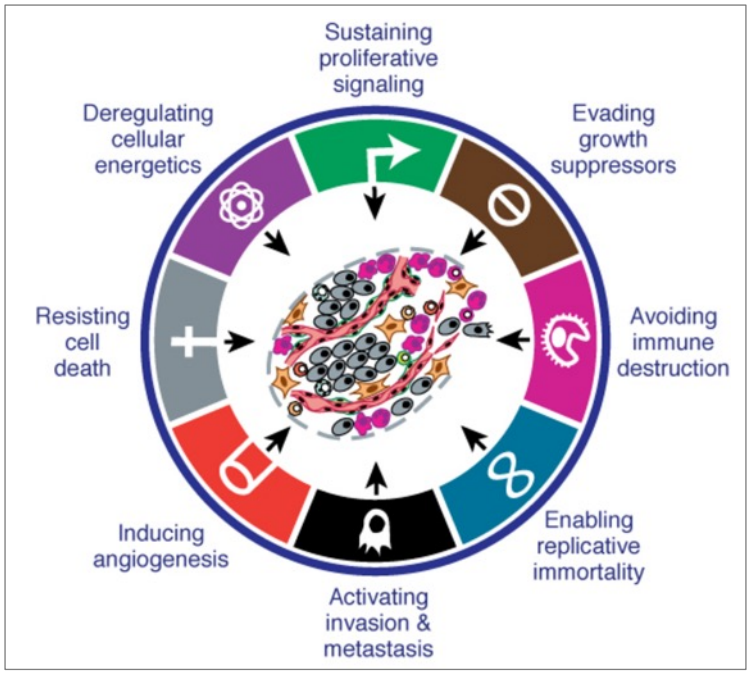
What is different in a cancer cell?

Hallmarks of cancer

Hanahan & Weinberg, 2015



Chemotherapy with nonspecific cancer drugs

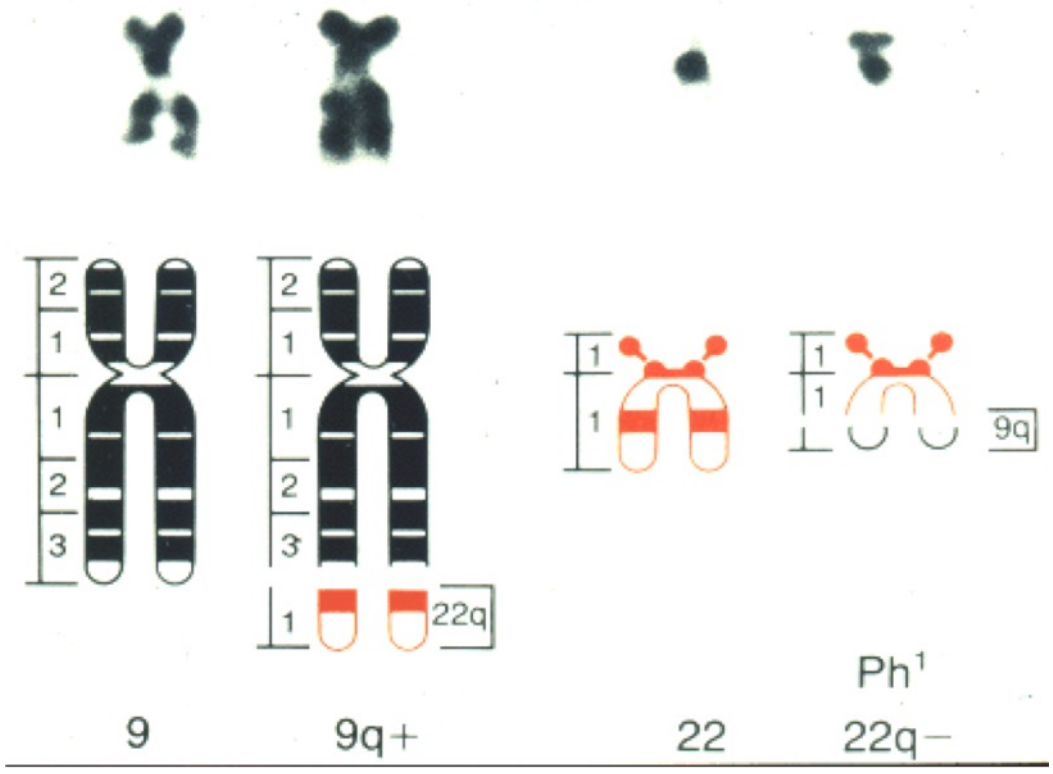


- Taxanes: block cell division
- Alkylating agents (platin salts): kill dividing cells
- Tamoxifene : block hormonal activation

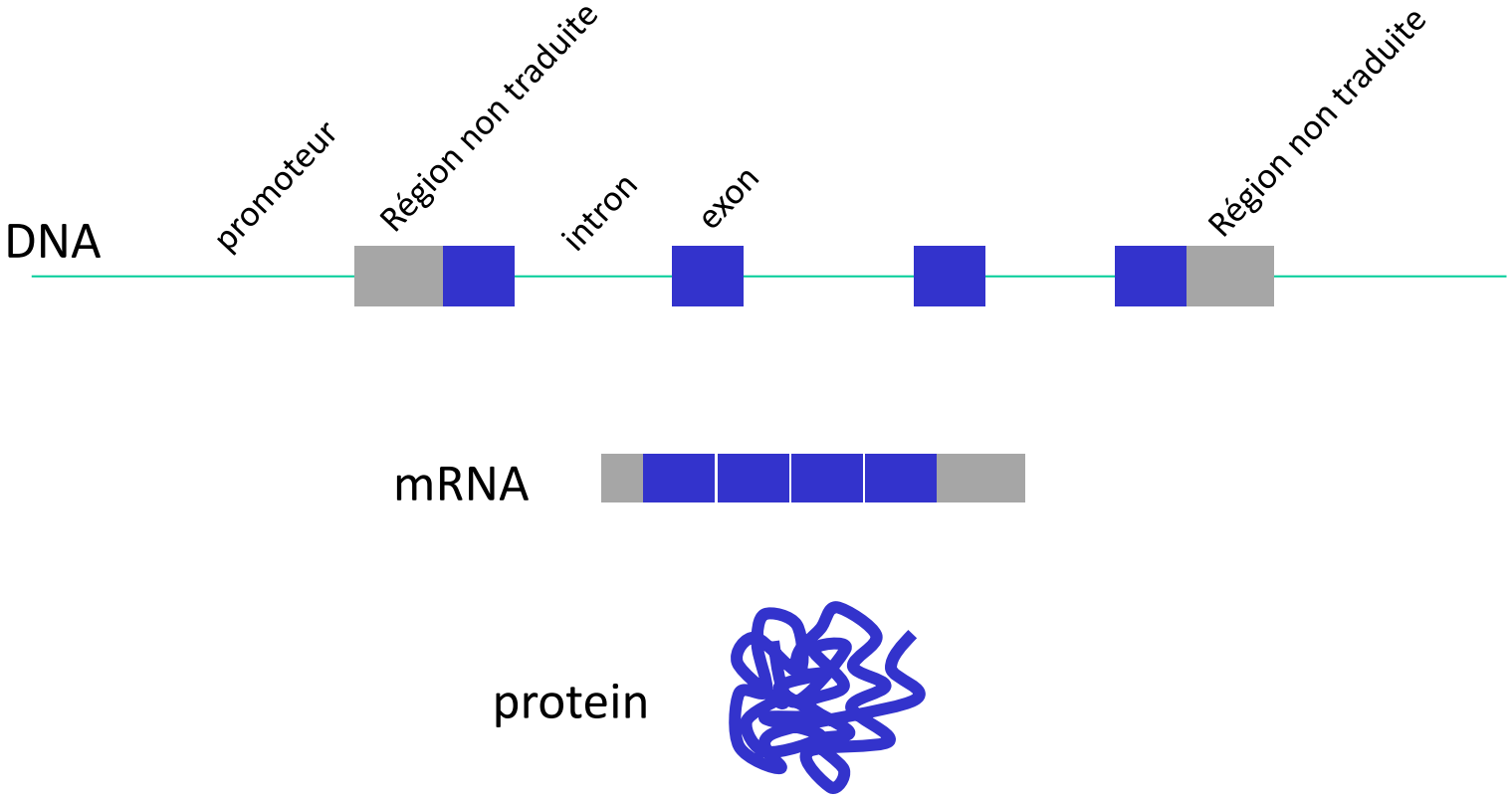
Precision medicine

Le « chromosome de Philadelphie »

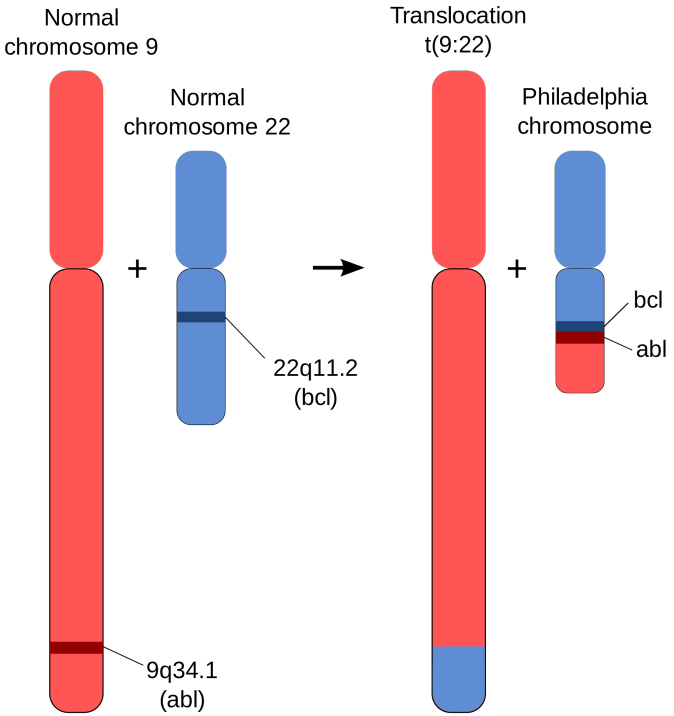
Translocation entre le chromosome 9 et le chromosome 22 dans la leucémie myéloïde chronique (LMC)



Rappel: structure d'un gène eucaryote (animaux, plantes, levures)



Conséquence de la translocation: un oncogène sur chromosome 9 fusionne avec un gène du chromosome 22 : la protéine de fusion BCR-ABL possède une activité tyrosine kinase qui active le cycle cellulaire de manière constitutive et provoque une leucémie myéloide chronique (LMC).

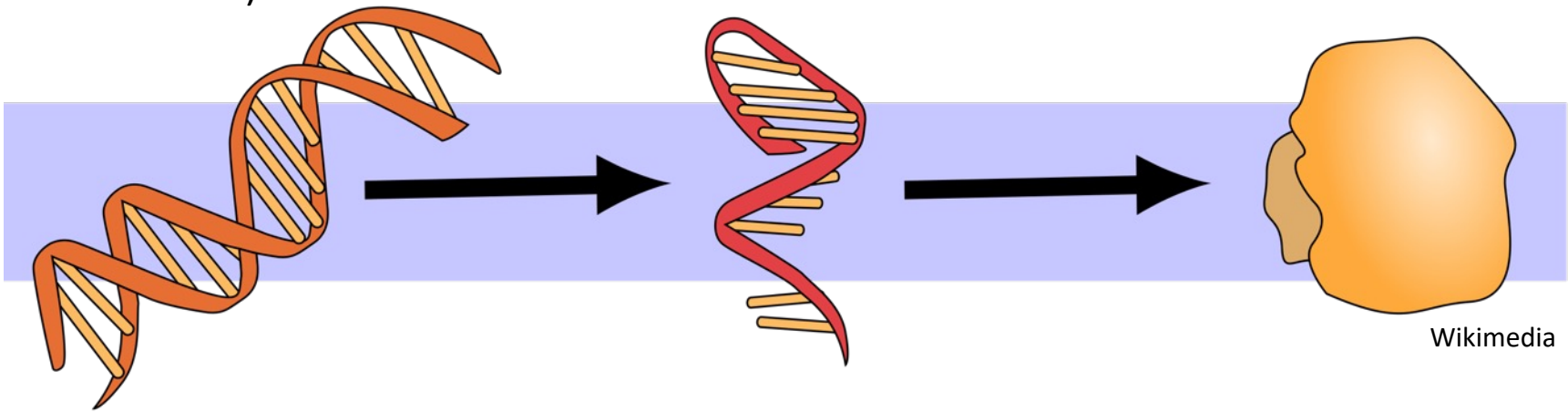


Protéine de fusion
BCR-ABL

Glivec
(inhibiteur de
tyrosine kinase)

Types of tumor-specific alterations

- Epigenetic changes (methylation, histones marks)



Wikimedia

DNA

- Mutations, deletions
- Copy number variations
- Translocations,
- Transposon insertions

RNA

- Change in expression
- Change in processing (splicing)
- Base modifications
- Activation of noncoding genes

Protein

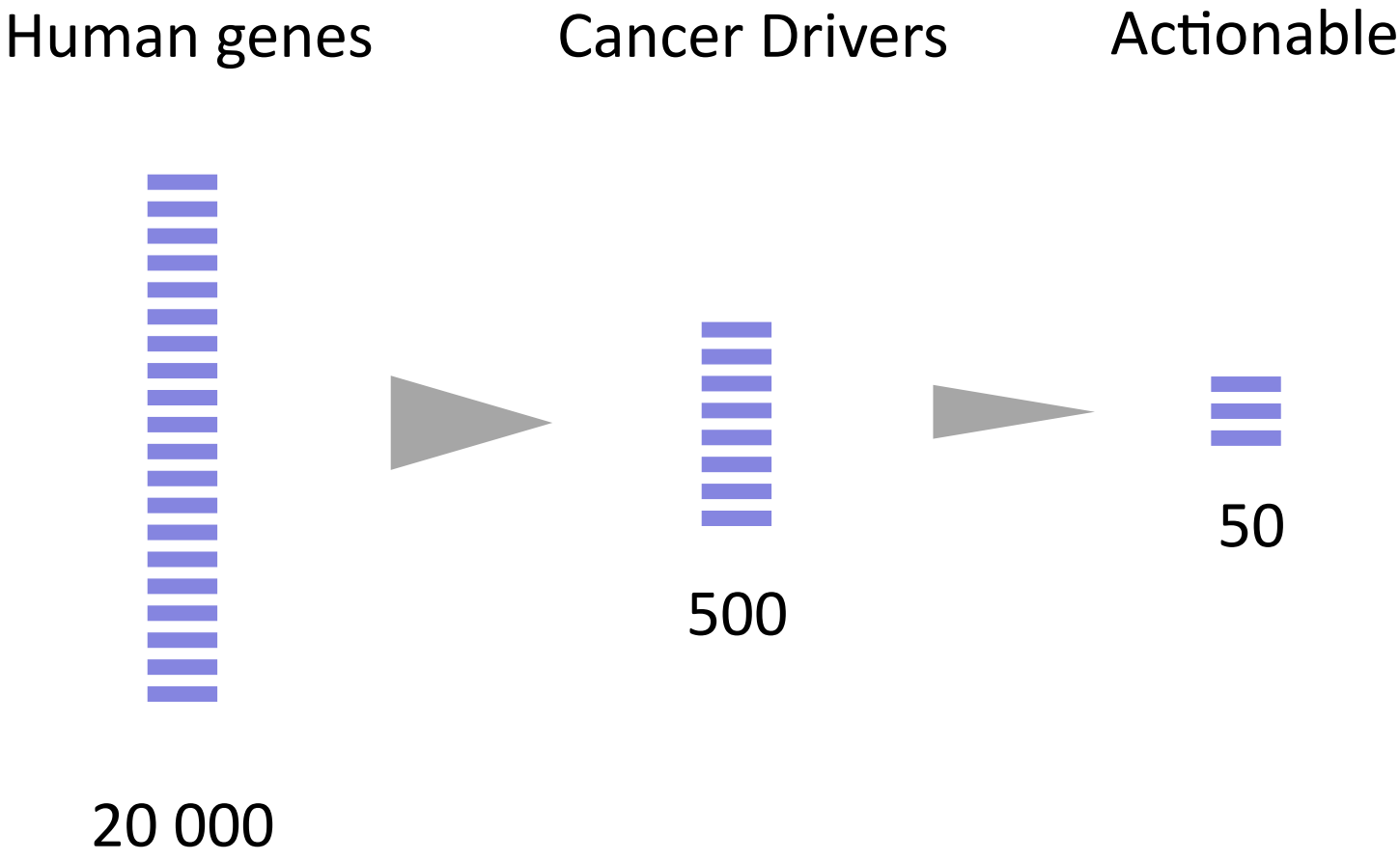
- Change in quantity
- AA modifications

Cancer driver genes

- Genes whose alterations can be oncogenic
- ~500 known drivers in our 20,000 protein-coding genes
  - Tumor suppressors ◀ Impairing mutations
  - Oncogenes ◀ Activating mutations

Actionable genes

- Genes that can be targeted by a known drug
- Ex:
  - BRCA2: PARP inhibitors
  - ERBB2: anti-HER2 antibody
  - KRAS: RAS inhibitor
- < 50 genes

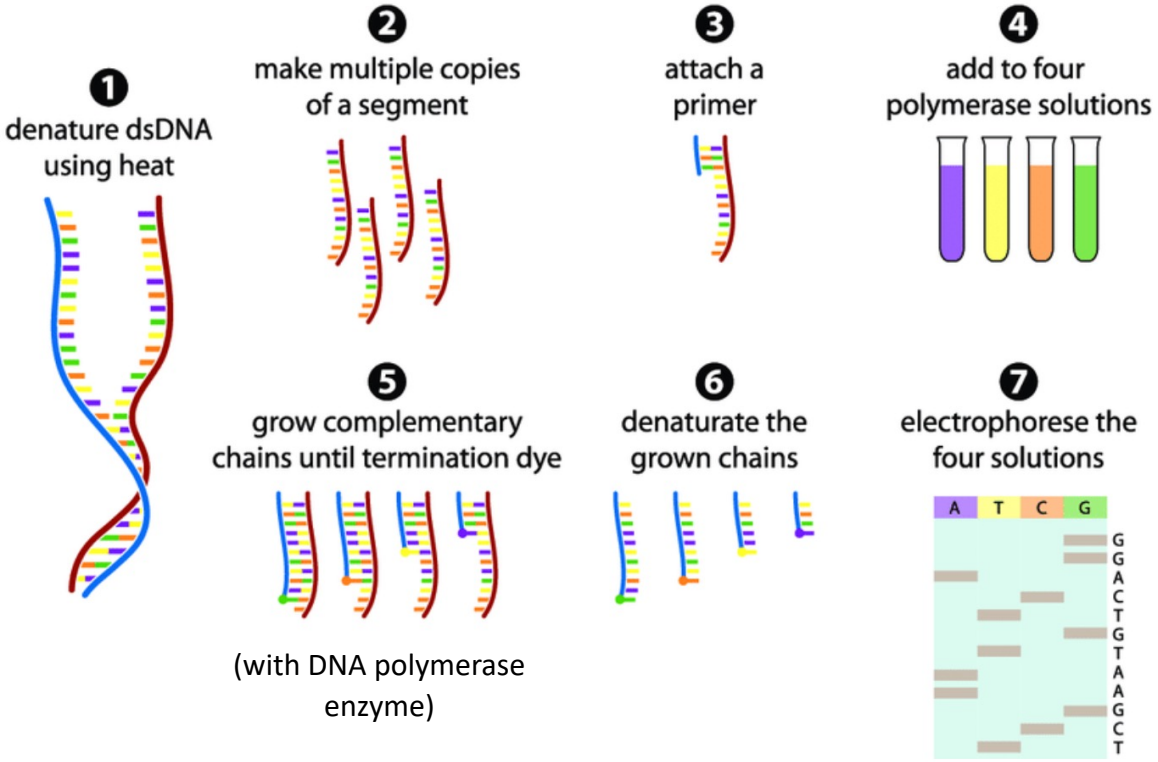

Human genes
Cancer Drivers
Actionable
20 000
500
50

RNA and DNA sequencing:

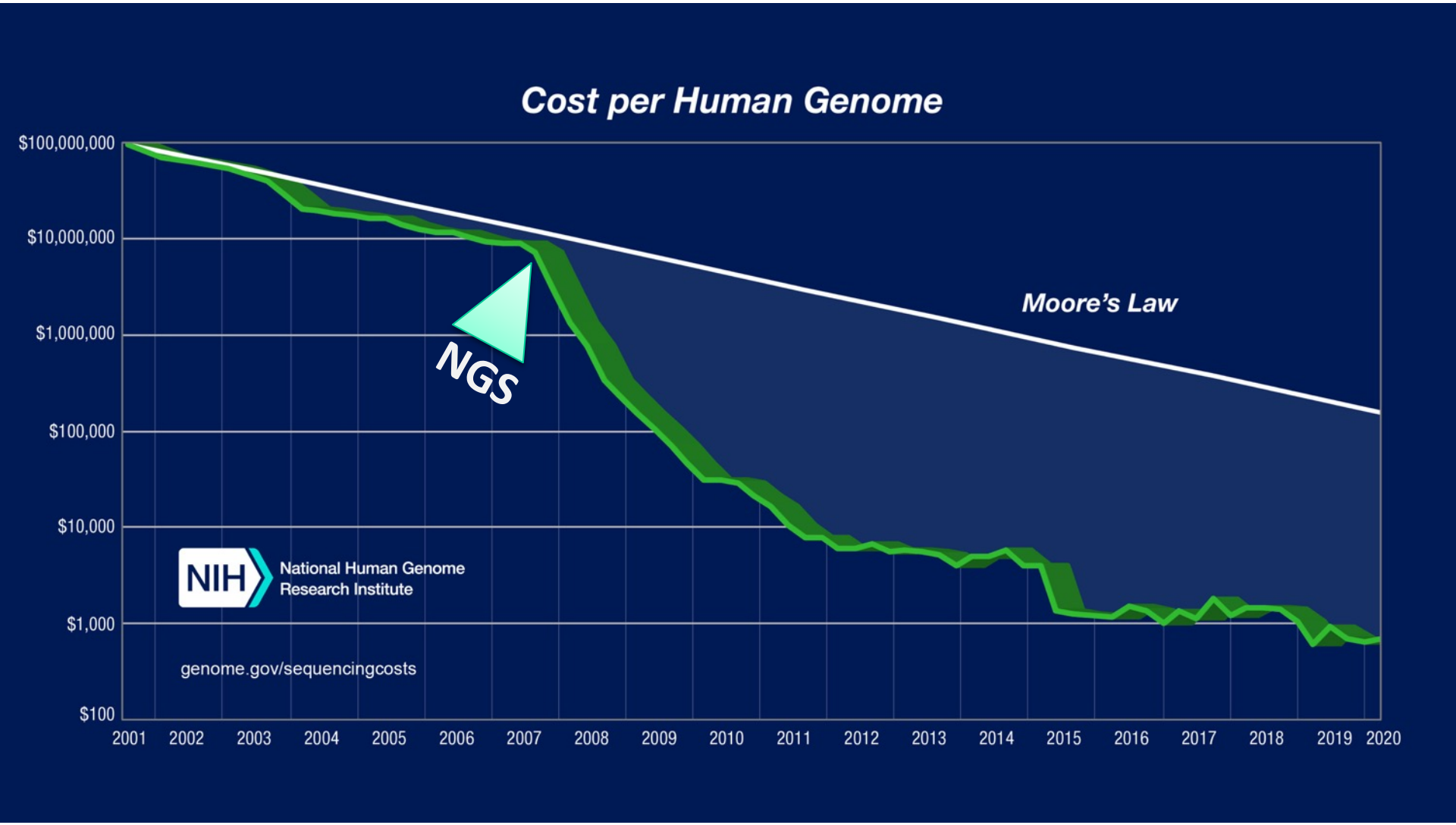
The main omics technologies in oncology

Sanger sequencing (1977)

- Chain termination sequencing or "sequencing by synthesis"
  - Termination nucleotides are used to block synthesis at a specific base. (A, T , G or C)



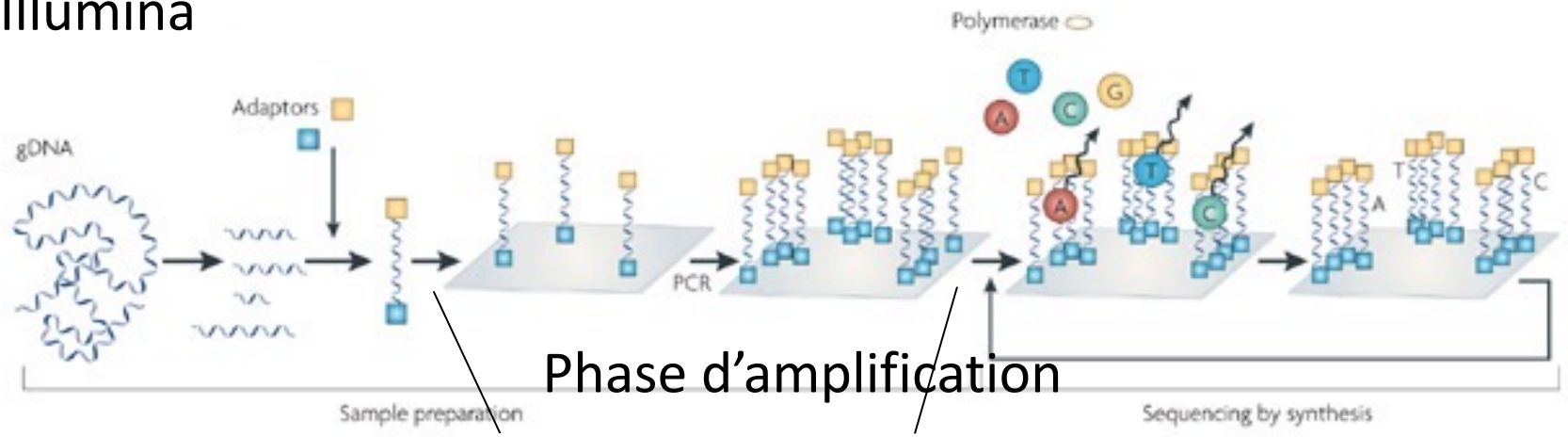
The NGS revolution



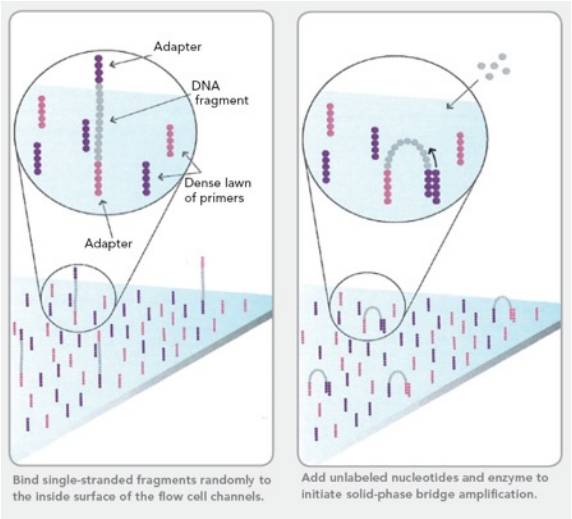
NGS :
Next Generation Sequencing
(2005-)

The most common NGS is « sequencing by synthesis » too!

Illumina



Phase d'amplification



Next Generation Sequencing

Nanopore MinION

7Gb

Illumina MySeq

4-10 Gb

Nanopore GridiON

35Gb

Thermofisher Ion Torrent

50 Gb

Illumina Hi-Seq

500 Gb

Illumina NovaSeq

6Tb

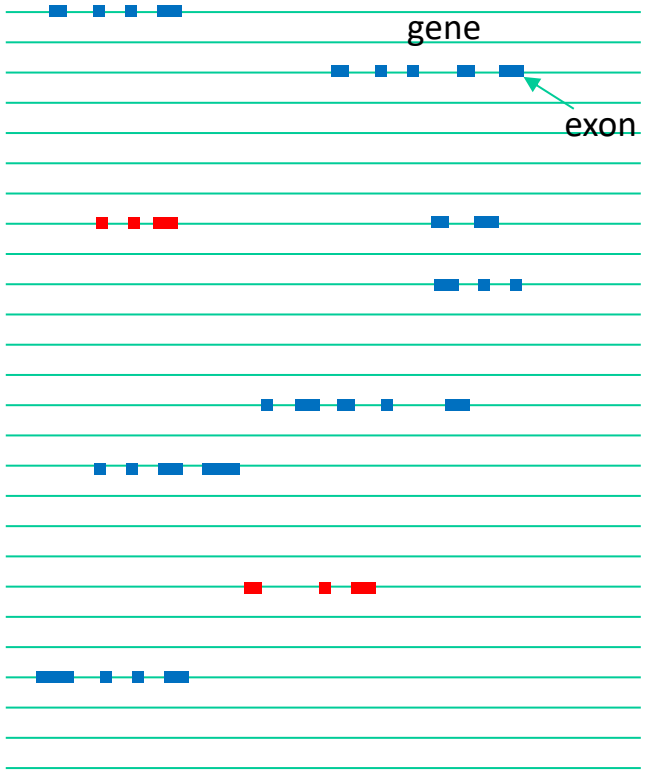
Les grandes applications des NGS

- DNA-seq (variants génomiques, de novo)
- RNA-seq (transcriptome)
- ChiP-Seq (sites de liaisons à l'ADN)
- Autres applications
  - Hi-C, clip-seq, net-seq, ribosome profiling etc.

DNA-seq: Recherche de variants génomiques

- En cancérologie, 2 grandes applications
  - Génétique constitutionnelle (recherche de prédisposition)
  - Génétique somatique (diagnostic, médecine de précision)

DNA sequencing types



Chromosome sequence

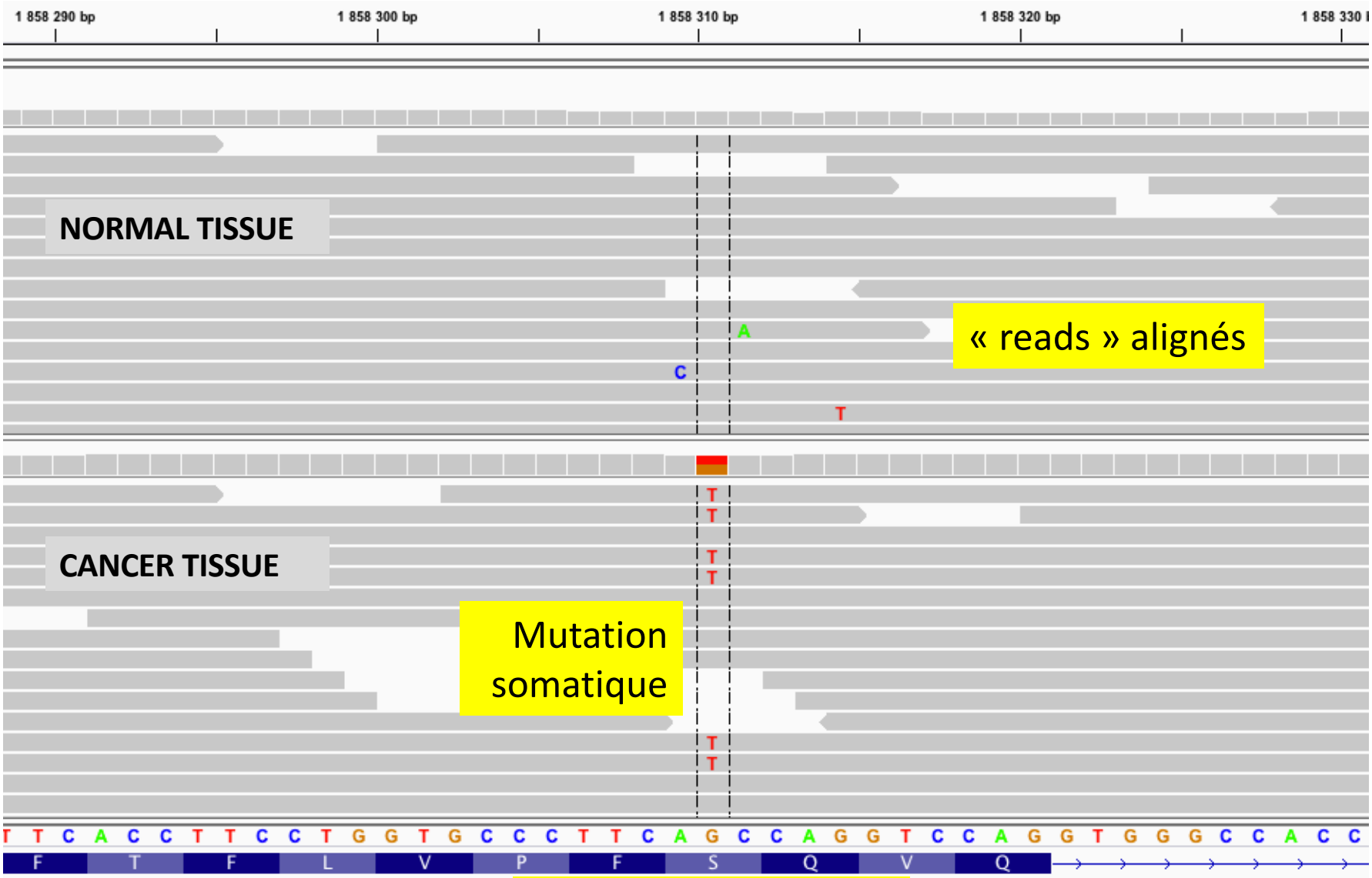
- WGS= Whole genome (3Gb)
- WES: whole exome (50Mb)
- Panel: selected genes (200kb)

Protocole DNA-seq

Patient sample

50-100M reads
100nt each

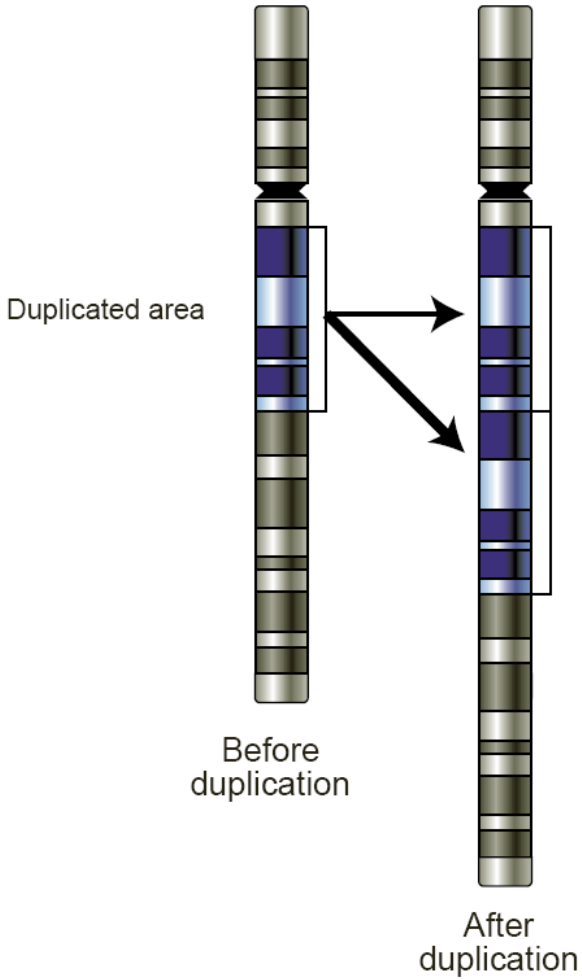
Evénéments recherchés



Evénéments recherchés par DNA-seq

- Variations ponctuelles=SNV
- Réarrangements
- Changement de nombres de copies =CNV
- Amplification de microsatellites
- Profils mutationnels

Copy number variations



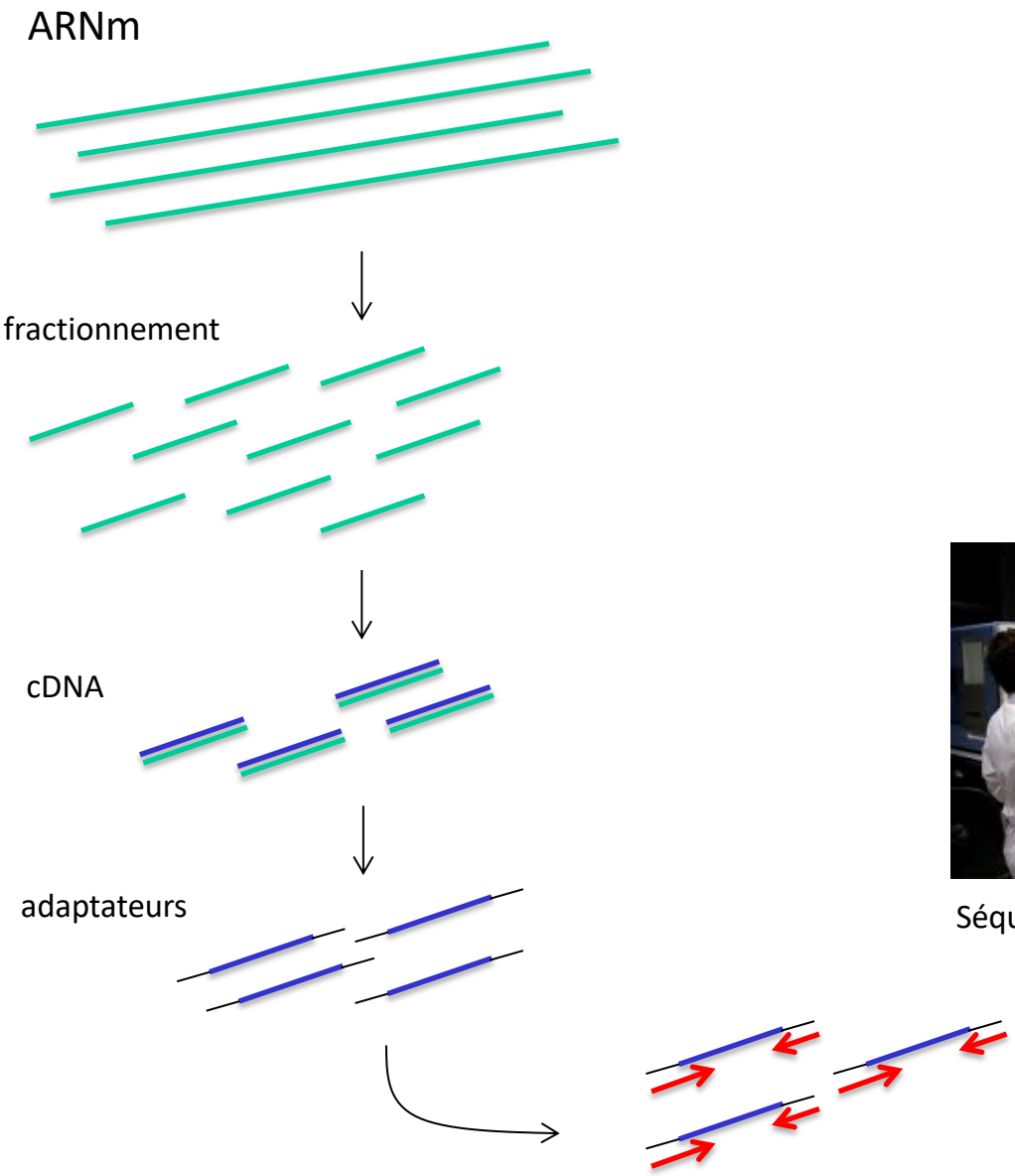
- Caused by recombination errors during DNA repair
- Can be loss or gain
- Oncogenic events= loss of a tumor suppressor or gain of a oncogene

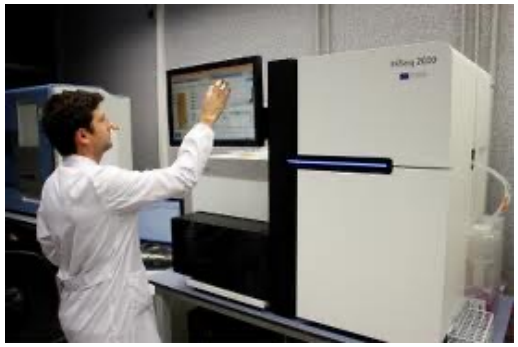
Cf cours Bastien Job

RNA-seq

- Pour l'étude du transcriptome
  - Mesure de l'expression de tous les gènes, simultanément

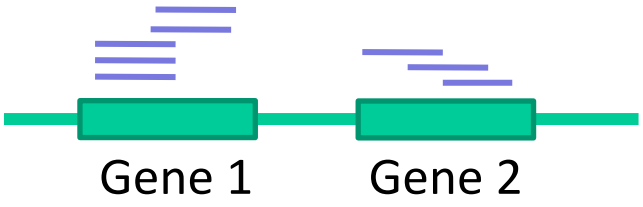
RNA-Seq



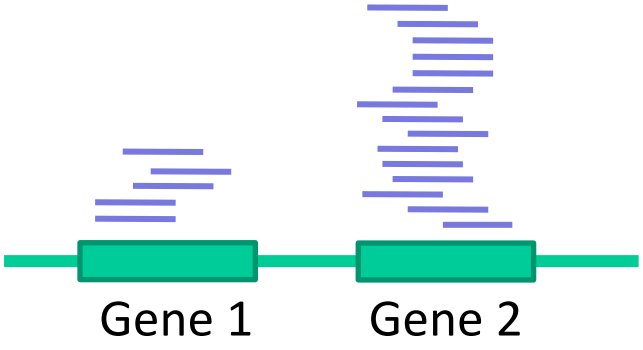


Séquençage

Mesure d'expression par RNA-seq

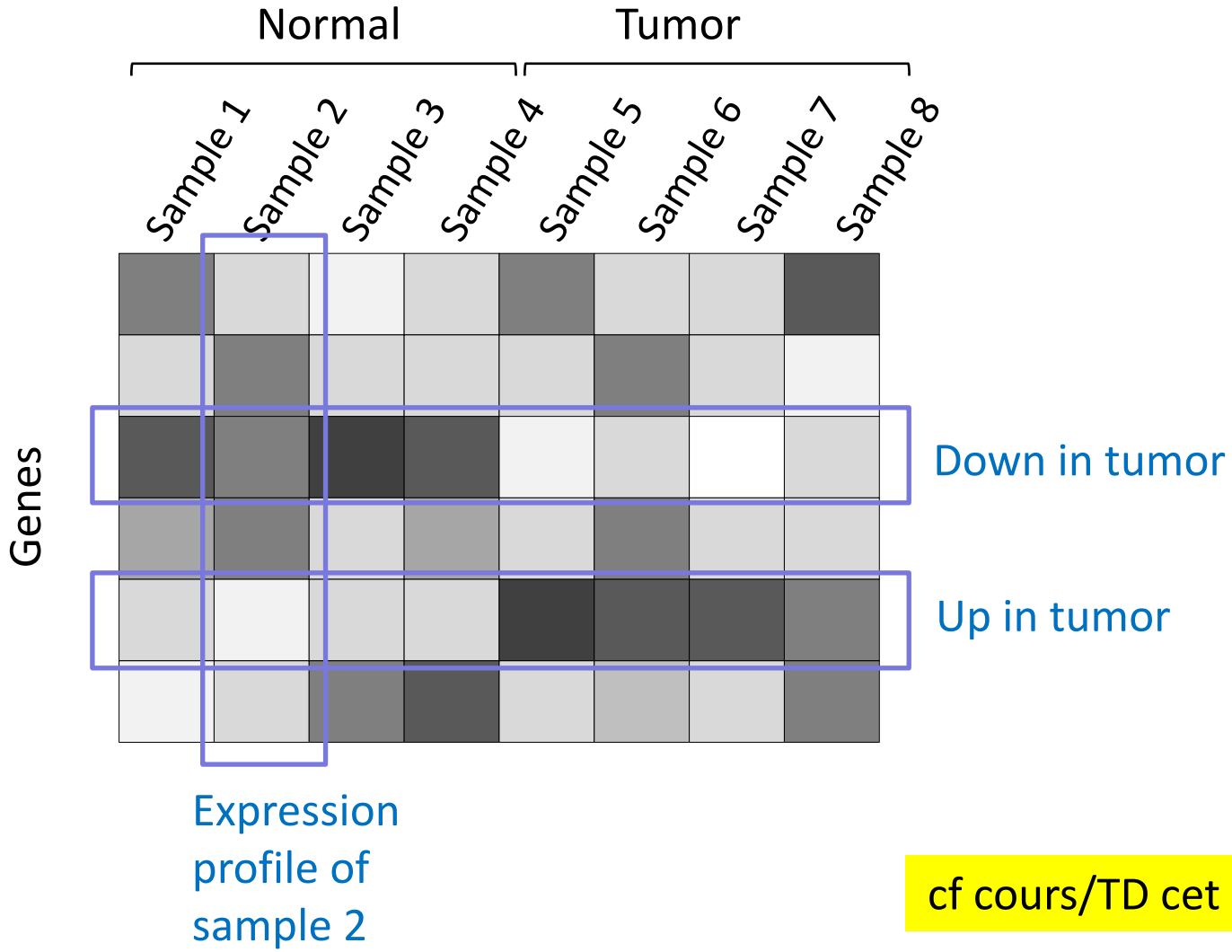


Sample 1



Sample 2

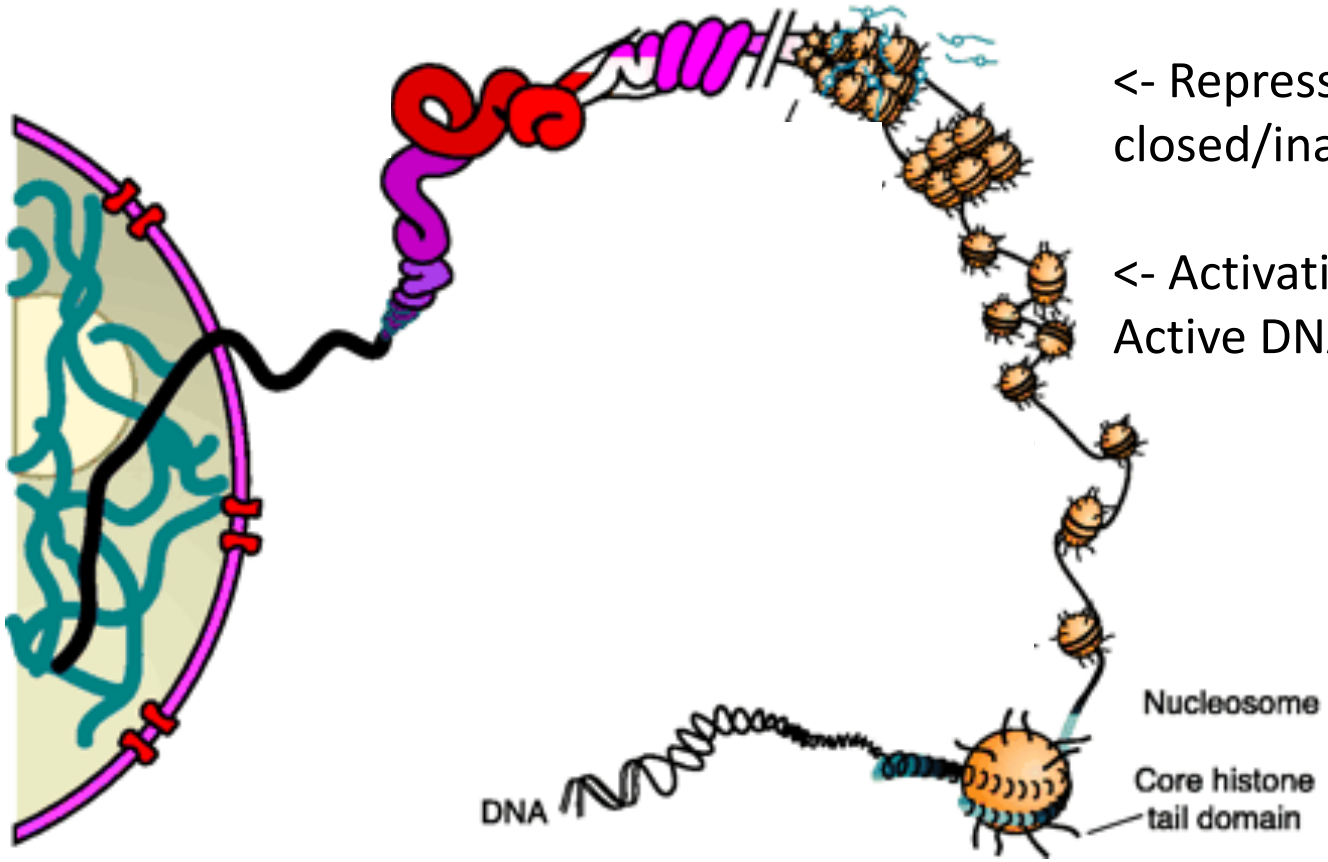
Differential expression analysis & expression profiling



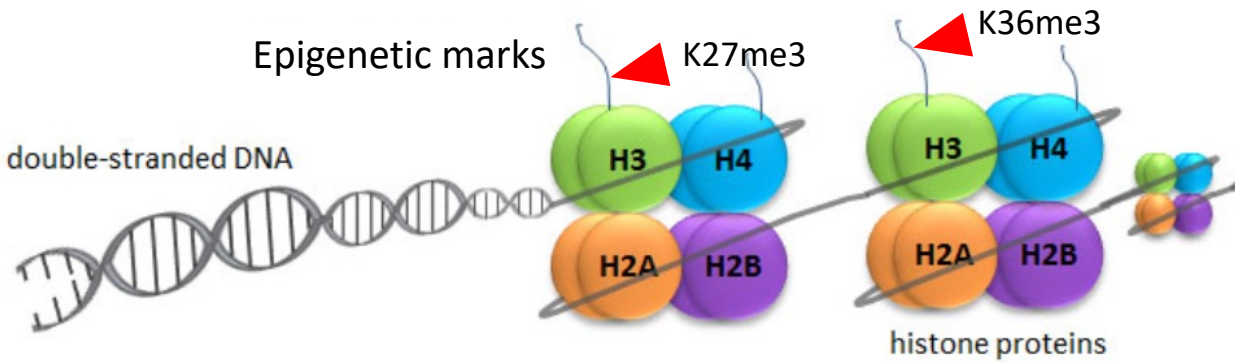
cf cours/TD cet après midi

ChIP-seq in oncology

ChIP-seq = Chromatin ImmunoPrecipitation & Sequencing
Identifies epigenetic marks on chromosomes



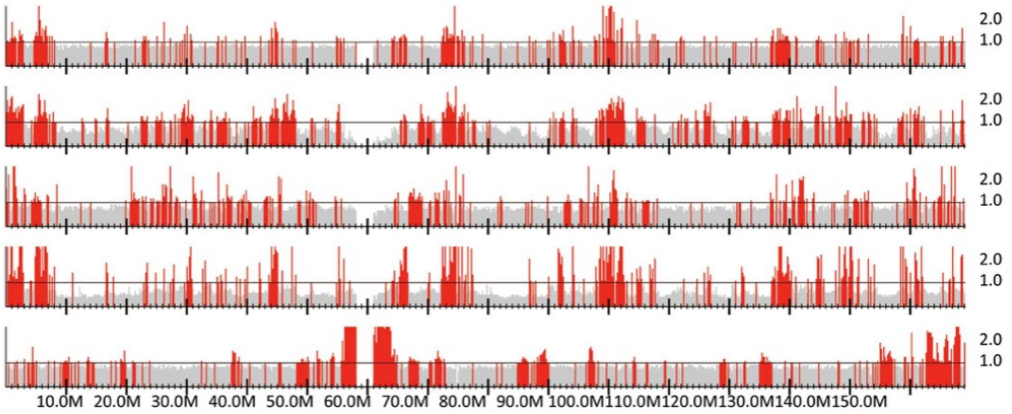
ChIP-seq in oncology



Epigenetic marks

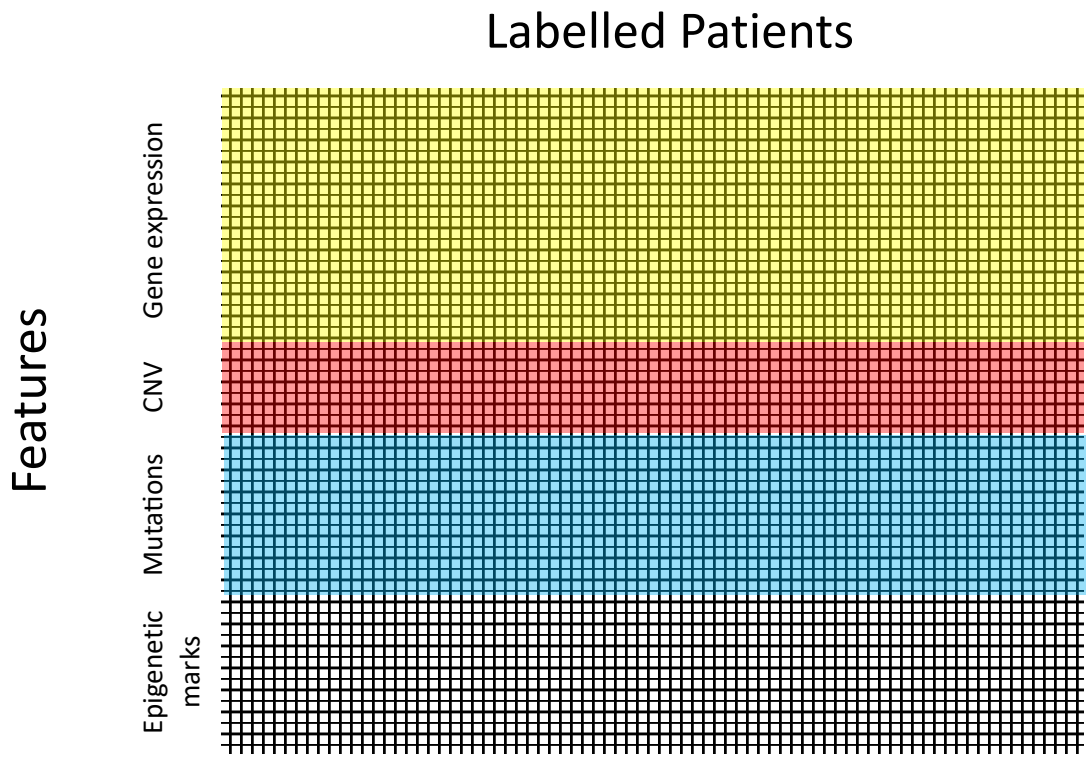
activating

repressive



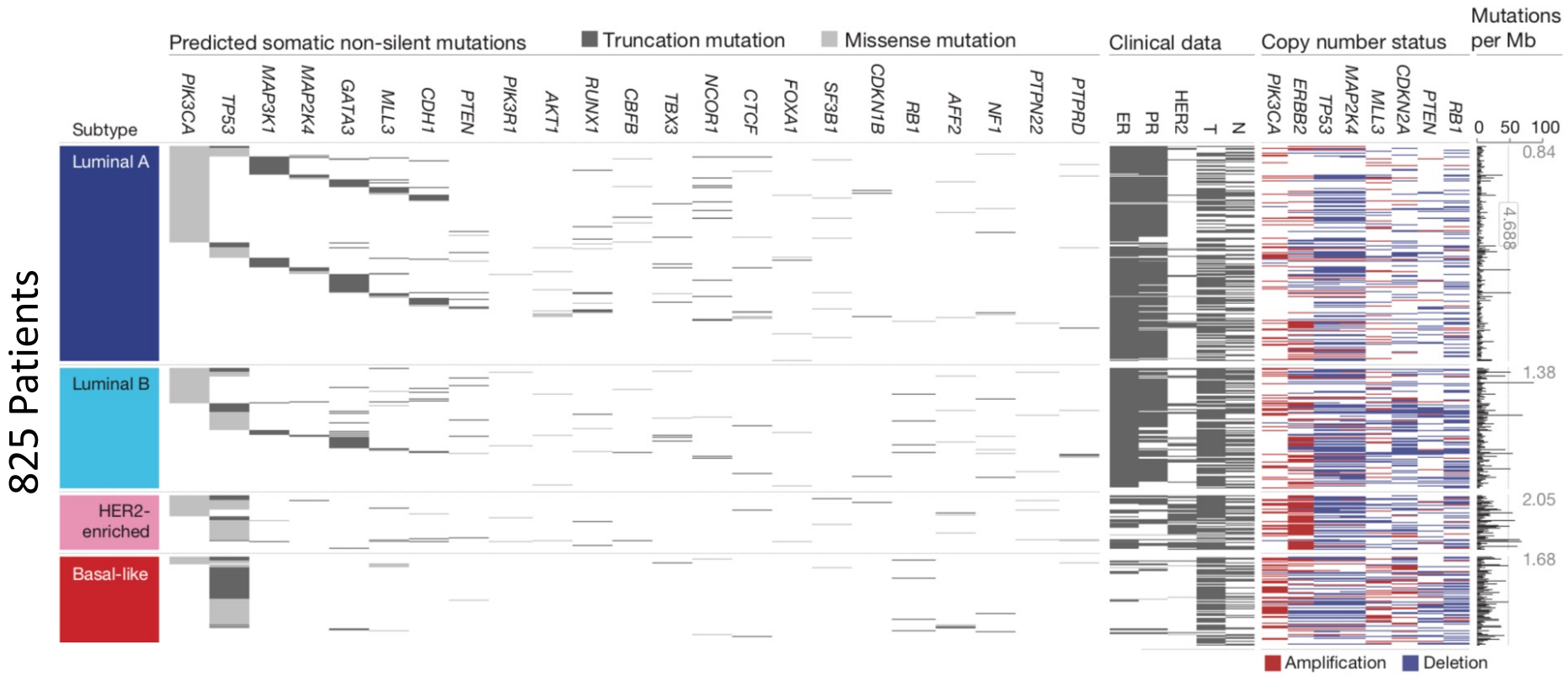
chromosome

Big data in oncology



cf UE12: cours/TD Yoann Pradat & Julien Vibert

Omics cohorts in cancer
TCGA project: 10,000 patients

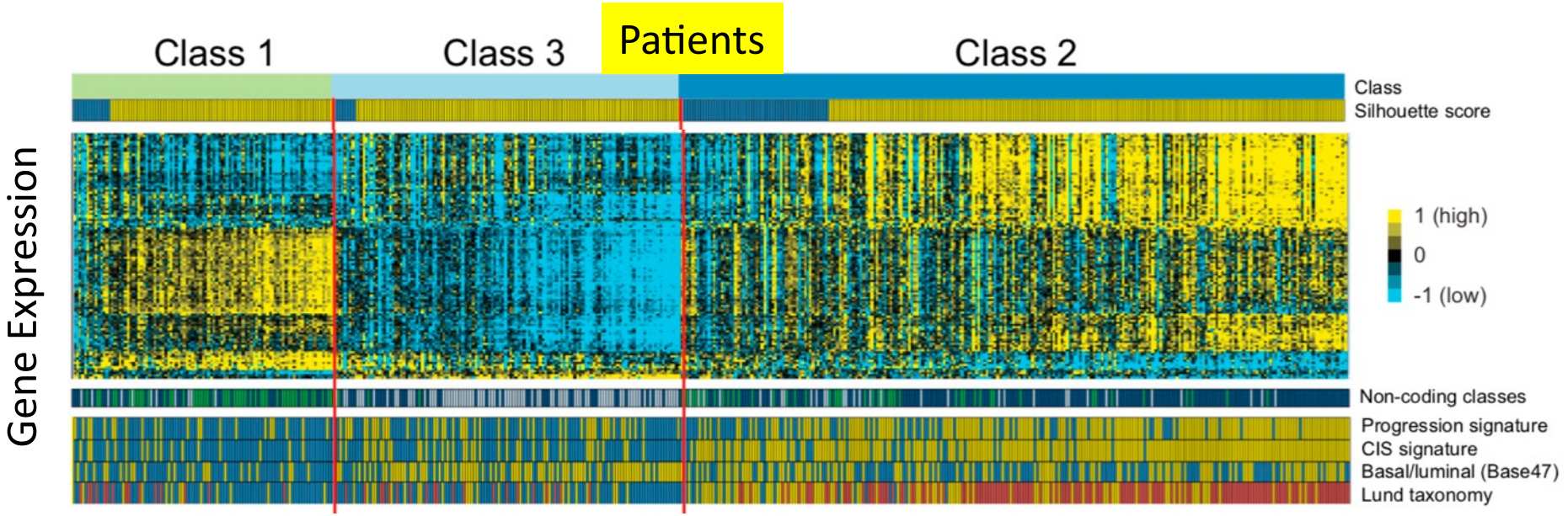


TCGA Breast cancer cohort (2012): 825 patients.
Stratified by cancer type

TCGA consortium, Nature, 2012

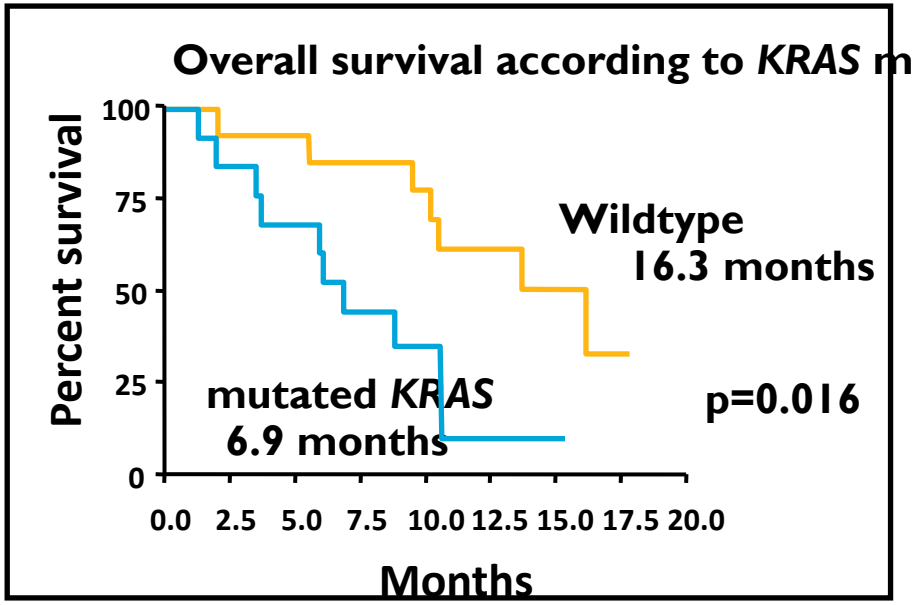
Subtype discovery (unsupervised classification)

Gene expression profiling in 460 Urothelial carcinoma:
A 117-gene signature.



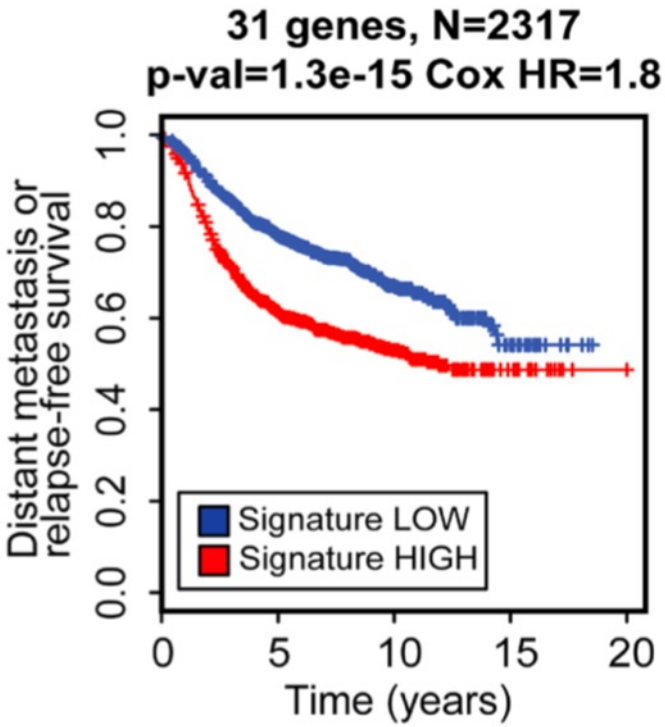
Hedegaard et al., Cancer Cell, 2016

Modèles prédictifs de survie



DNA-based

Expression-based



Transcriptional signatures: current applications

Diagnostic

- Tumor vs normal
- Tumor subtyping

Precision medicine

- Response to treatment
- Relapse prediction

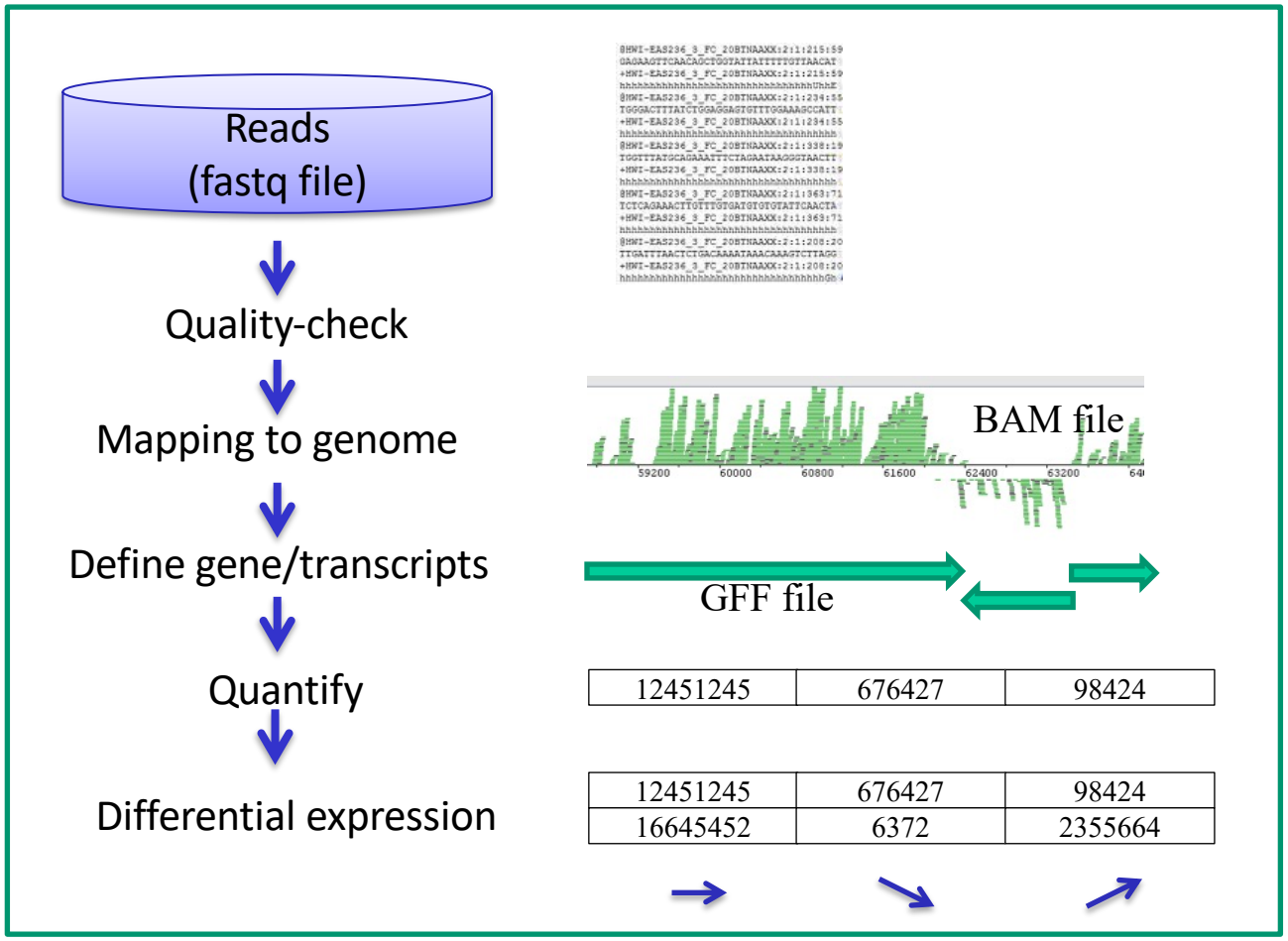
Les outils d'analyse

« Pipelines » & « workflows »

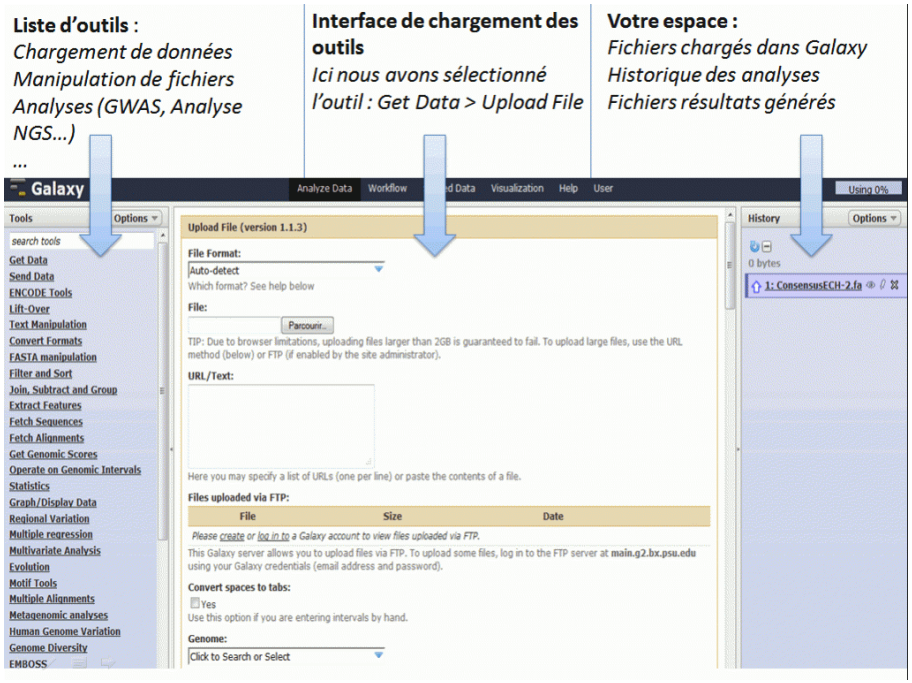
« Bricks » from Unix open source programs

Combined into pipelines (typically a few hours to days to run)

Example: an RNA-seq pipeline



Galaxy: user-friendly interface to NGS pipelines



Credit: Biorigami

- Interest: avoiding Unix command line + traçability
- But: running NGS workflow on real human data often requires a computer cluster (will not run on a single-node Galaxy server)

Les bases de données en génomique du cancer

Cancer Genomics Databases

- TCGA: the Cancer Genome Atlas
- COSMIC
- cBioPortal
- CCLE: Cancer Cell Lines Encyclopedia
- GDSC: Genomics of Drug Sensitivity in Cancer
- dbGaP: database of Genotypes and Phenotypes
- GEO: Gene Expression Omnibus
- ArrayExpress

The Cancer Genome Atlas *Understanding genomics to improve cancer care*

TCGA

- launched in 2006
- 33 tumor types
- 11,000 patients
- whole-genome sequencing (WGS) for 1,000 tumors

TCGA data types and levels

Controlled access

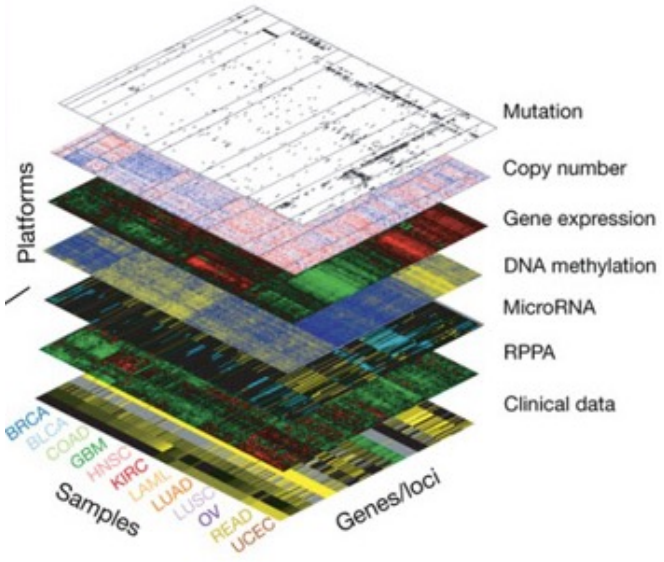
Free access

- Clinical Data
- Images
- Microsatellite Instability (MSI)
- DNA Sequencing
- miRNA Sequencing
- Protein Expression
- mRNA Sequencing
- Total RNA Sequencing
- Array-based Expression
- DNA Methylation
- Copy Number

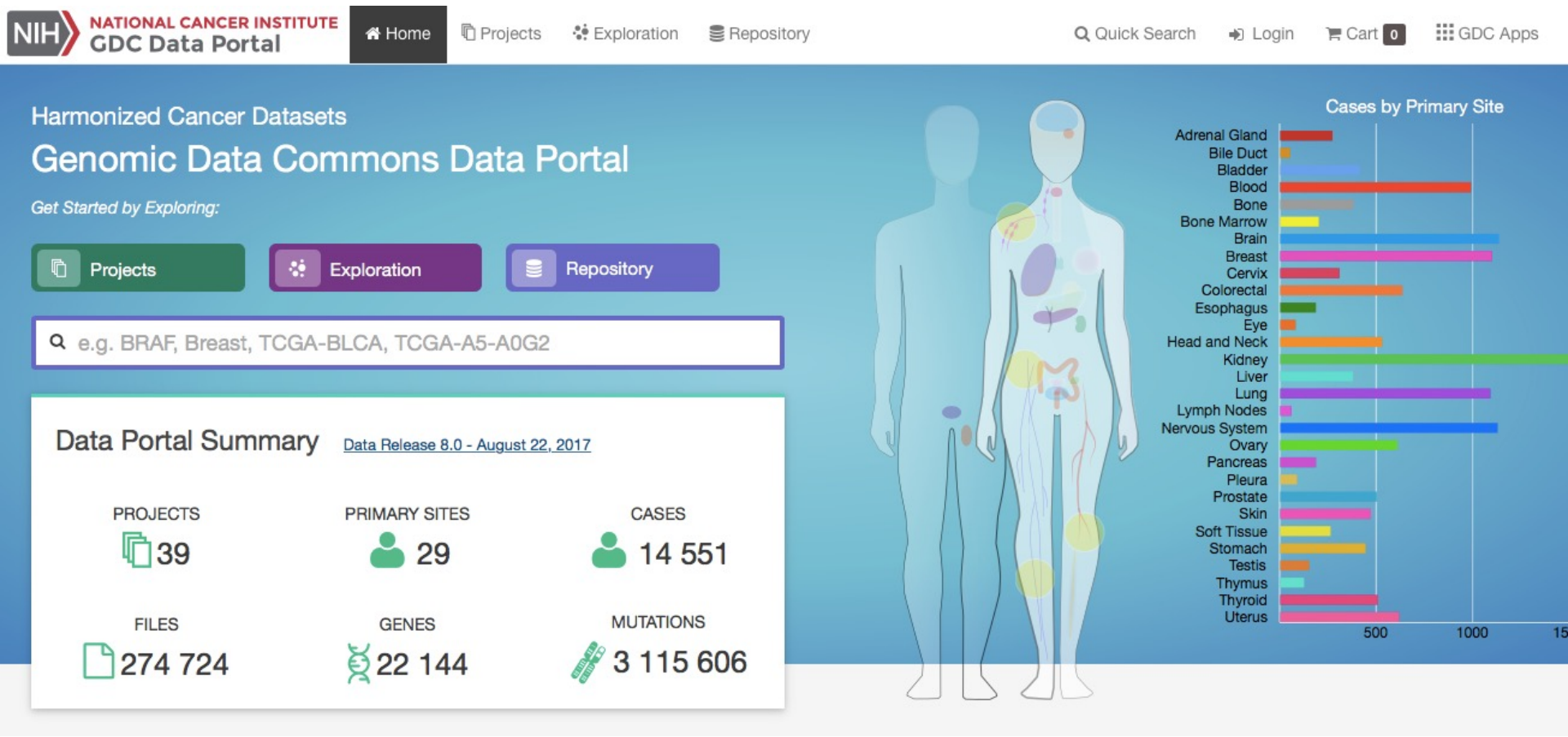
Level 1

Level 2

Level 3



TCGA access via the GDC portal (Genomics Data Commons)



From TCGA to PCAWG

- PCAWG[1]: a collaboration TCGA-ICGC[2] to analyze whole genome data from 2,800 pairs of tumor and normal samples and integrate the results with clinical and other molecular data available on those same cases.

1. PCAWG: Pan-Cancer Analysis of Whole Genomes
2. ICGC: International Cancer Genome Consortium

Sanger Institute, UK

COSMIC Curation

- Manual curation
  - >25000 articles analyzed
- Automated curation
  - 1.4M samples (incl. 31k WGS) (TCGA & ICGC)
  - Annotation pipeline (Variant effect predictor

« Most [mutations] have no effect on the development of disease. We are adapting our curation processes to reduce this noise and highlight high-value information. »

« Samples with over 20 000 point mutations, none of which have been validated are excluded from curation as their noise vastly outweighs their signal. »

COSMIC
Catalogue of somatic mutations in cancer

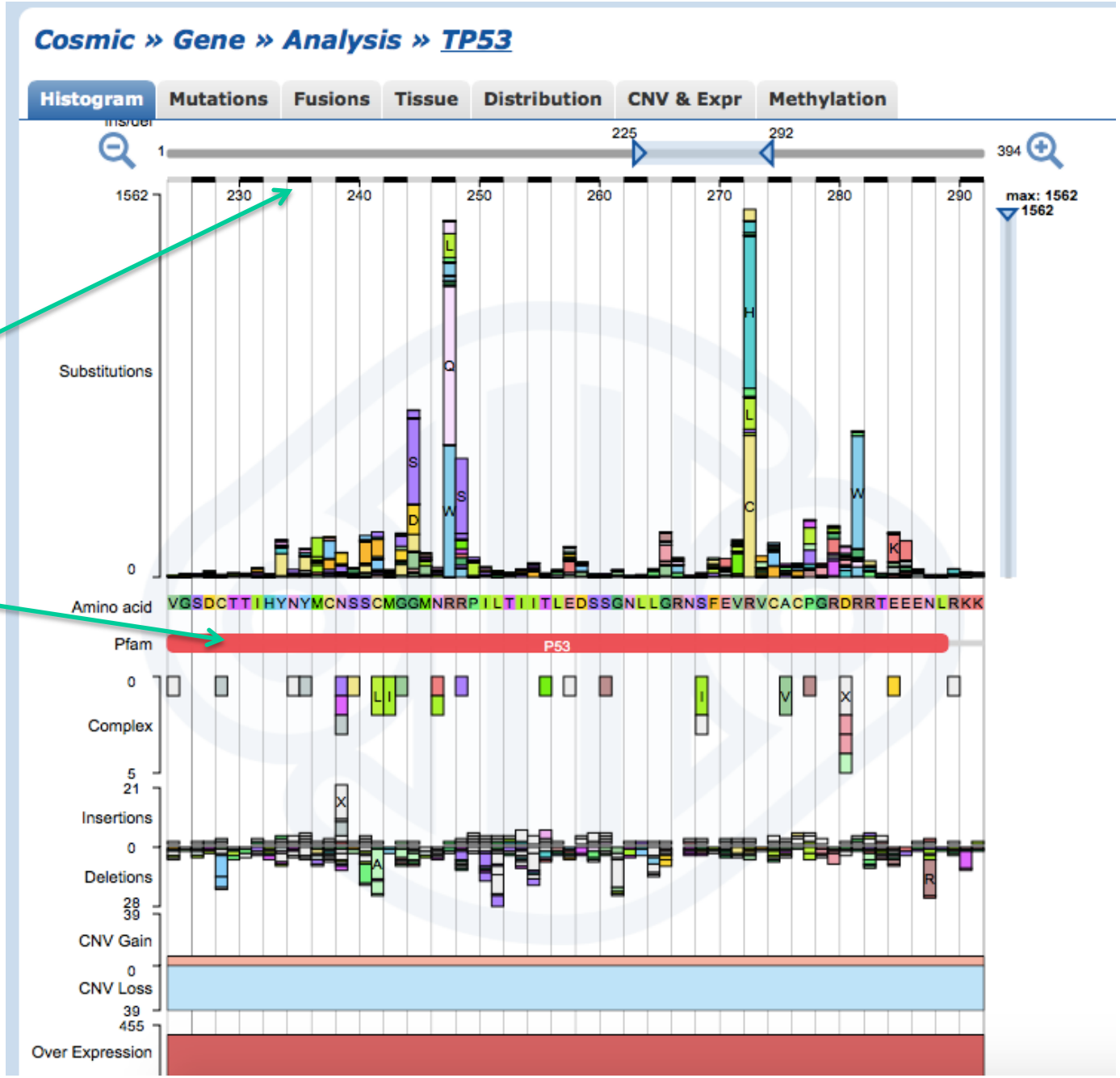
- Expert-curated database of cancer somatic mutations & other events
- 2022 (V97):
  - 23M variants génomiques
  - 19k gene fusions
  - 1.2 M CNV
  - 9M gene expression variants
  - 736 cancer genes

(2022: V95)

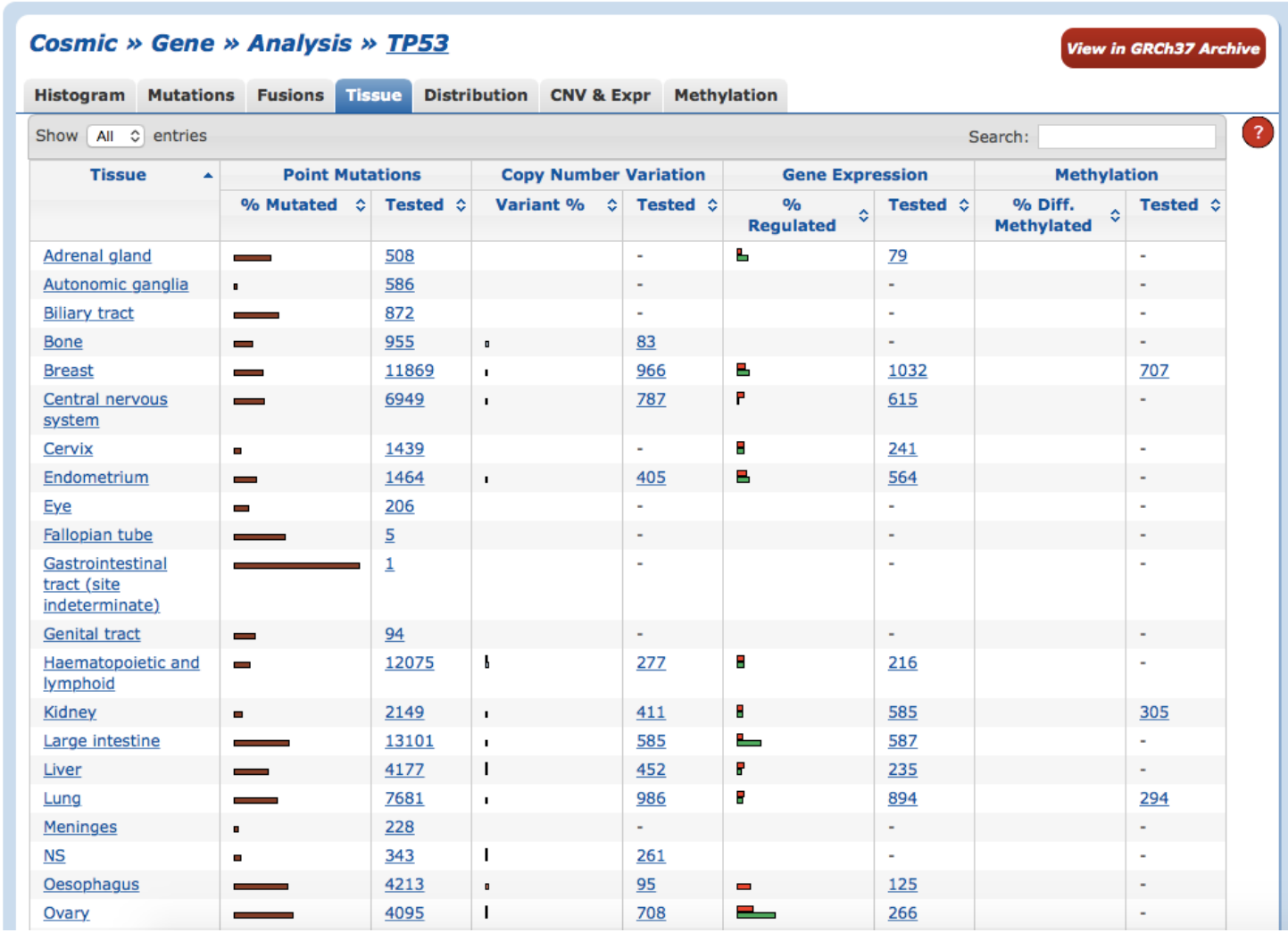
Histogram view

Protein coordinates

Protein domain



Tissue-distribution of mutations



Cosmic » Gene » Analysis » TP53

View in GRCh37 Archive

Histogram | Mutations | Fusions | Tissue | Distribution | CNV & Expr | Methylation

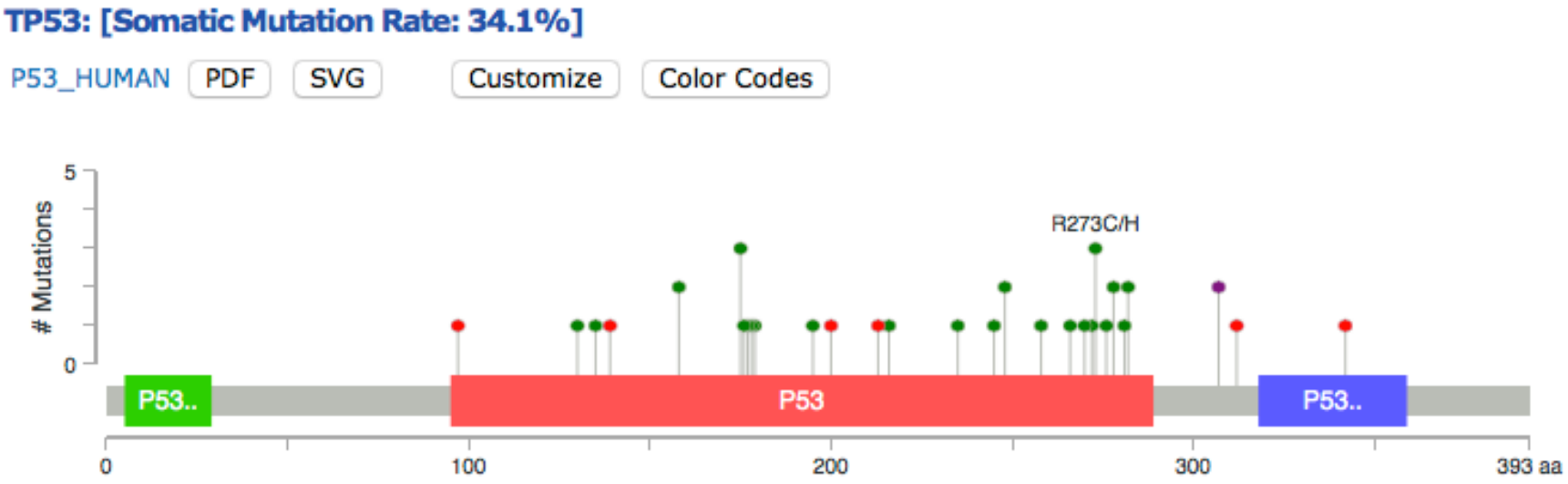
Show All entries Search:

Tissue	Point Mutations		Copy Number Variation		Gene Expression		Methylation	
	% Mutated	Tested	Variant %	Tested	% Regulated	Tested	% Diff. Methylated	Tested
Adrenal gland		508		-		79		-
Autonomic ganglia		586		-		-		-
Biliary tract		872		-		-		-
Bone		955		83		-		-
Breast		11869		966		1032		707
Central nervous system		6949		787		615		-
Cervix		1439		-		241		-
Endometrium		1464		405		564		-
Eye		206		-		-		-
Fallopian tube		5		-		-		-
Gastrointestinal tract (site indeterminate)		1		-		-		-
Genital tract		94		-		-		-
Haematopoietic and lymphoid		12075		277		216		-
Kidney		2149		411		585		305
Large intestine		13101		585		587		-
Liver		4177		452		235		-
Lung		7681		986		894		294
Meninges		228		-		-		-
NS		343		261		-		-
Oesophagus		4213		95		125		-
Ovary		4095		708		266		-

Memorial Sloan-Kettering
Cancer Center, USA

- Integration of Data from >350 cancer genomics studies.
- Focus on analysis tools
  - Mutual exclusivity
  - Gene networks

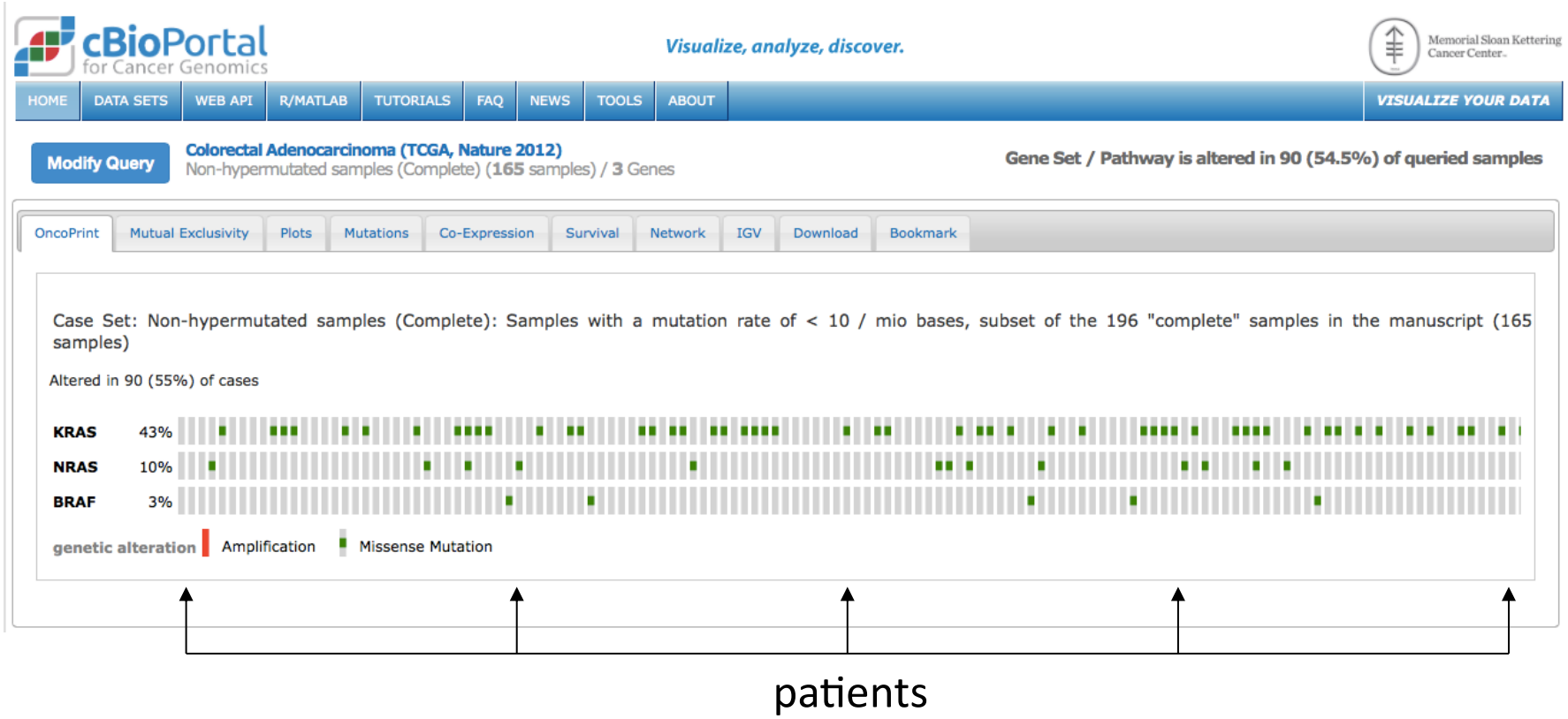
‘Lollipop plots’: mutations on proteins



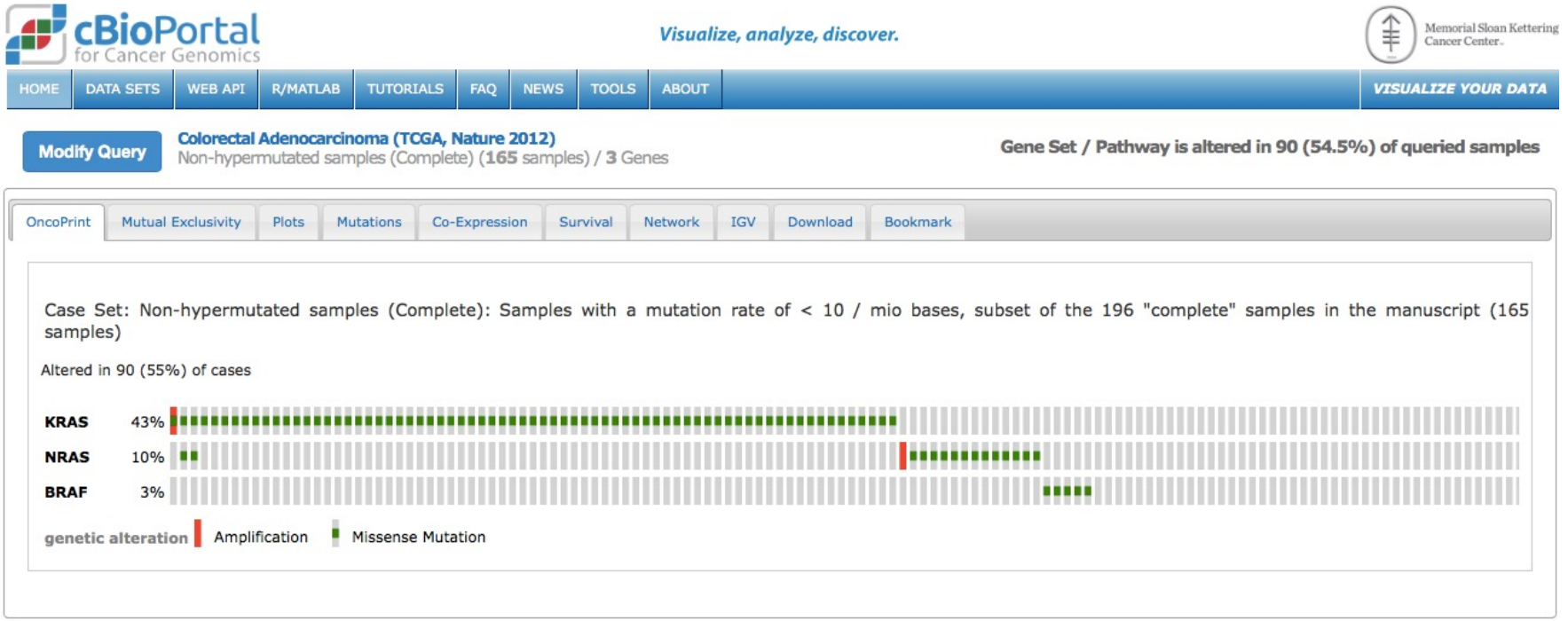
Mutations mapped on TP53 in Glioblastoma dataset (TCGA, Nature 2008)

See also « MutationMapper » tool

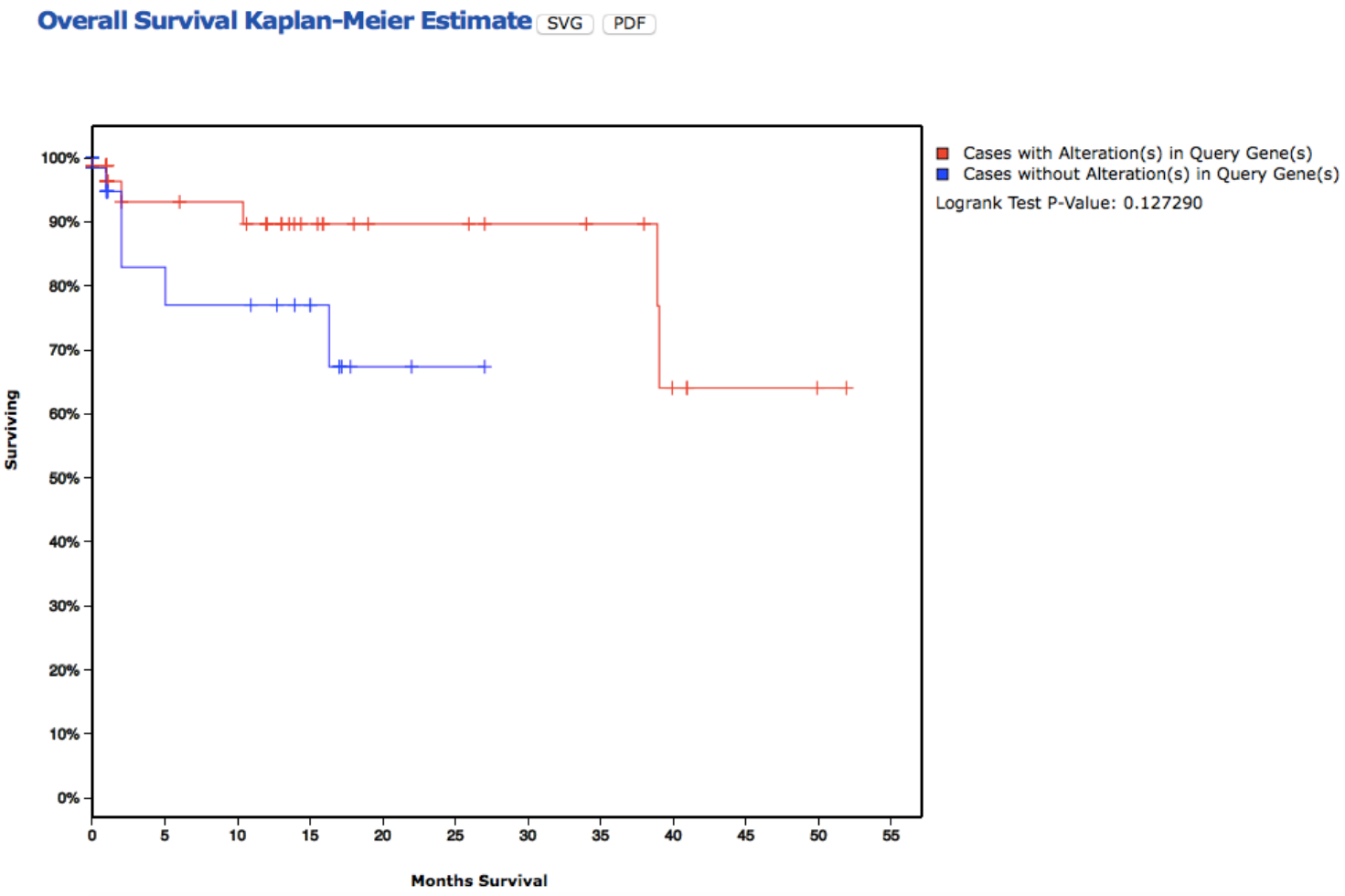
« Oncoplot » view



Mutual exclusivity



Kaplan-Meier Curves



Thank you

- Today:
  - Bastien Job on genome alterations
  - Practice: RNA-seq analysis with Galaxy