

Bienvenue au module #11

Cancer et génomique: des big data aux modèles prédictifs

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

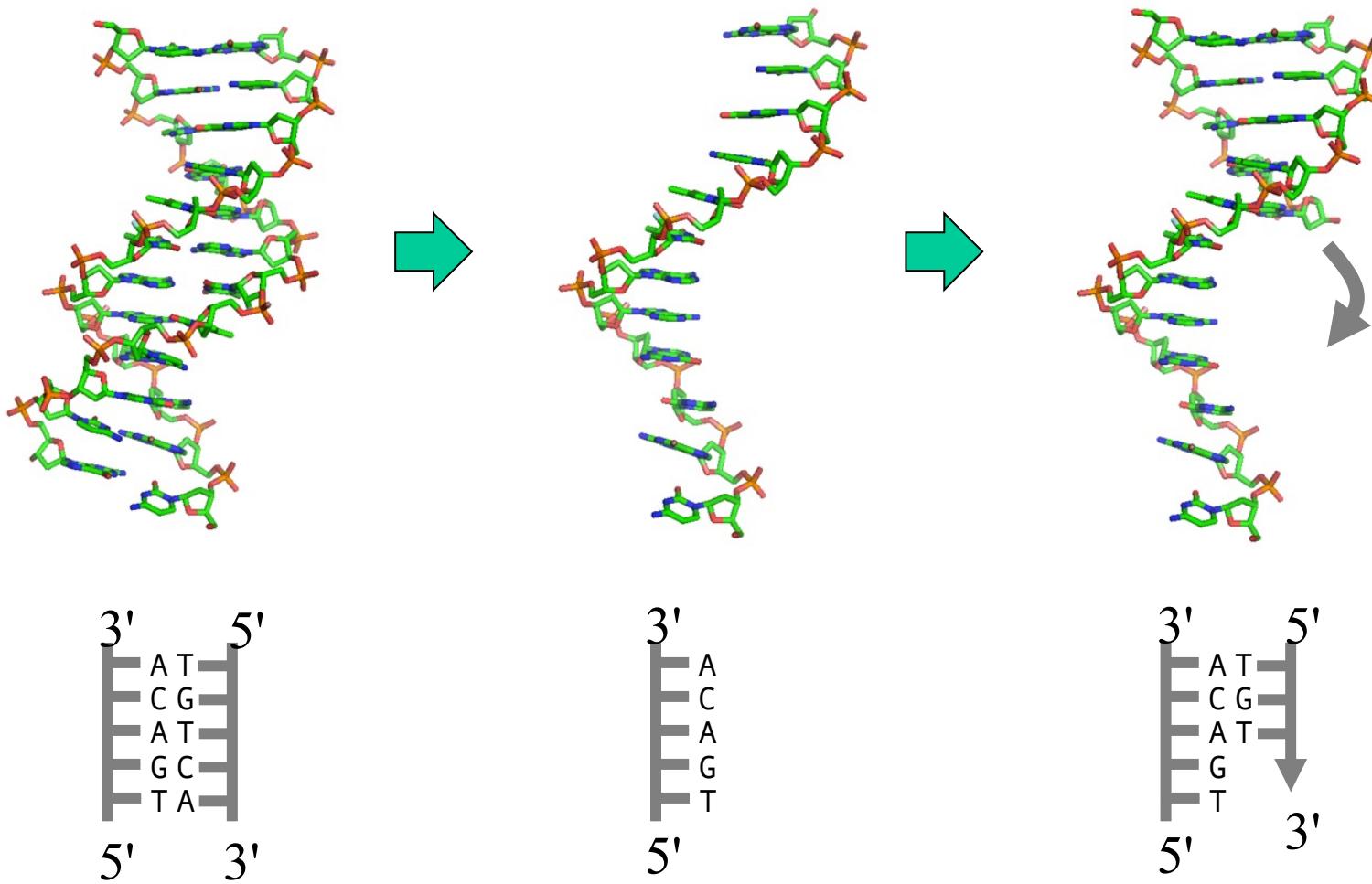
Mardi 23 janvier - Salle 21 B2M	
9:00-10:00	Pourquoi utiliser les méthodes d'apprentissage automatique en oncologie personnalisée? Julien VIBERT (Gustave Roussy)
10:00-11:00	Méthodes de classification supervisée. Yoann PRADAT (Gustave Roussy)
11:30-12:30	Exemple d'un projet de classification médicale: 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

Rappels de biologie moléculaire

Information transmission in the cell



Séquence ADN = information

AGGATGTCATAATAAAT
GAAATTGATTAAGTCAG
CCTGAATGACATTTCT
TTAAAATTCATTTA
CAGATTGAGTGATTGC

Biologie = une science de l'information



Binaire: base 2
2 symboles:

0	1
---	---

0 1 0 0 **1 1 1 0 0 1 0 1** 1 0 1 0 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
---	---	---	---

C T C A T G C A C **A C G T A G A T T**

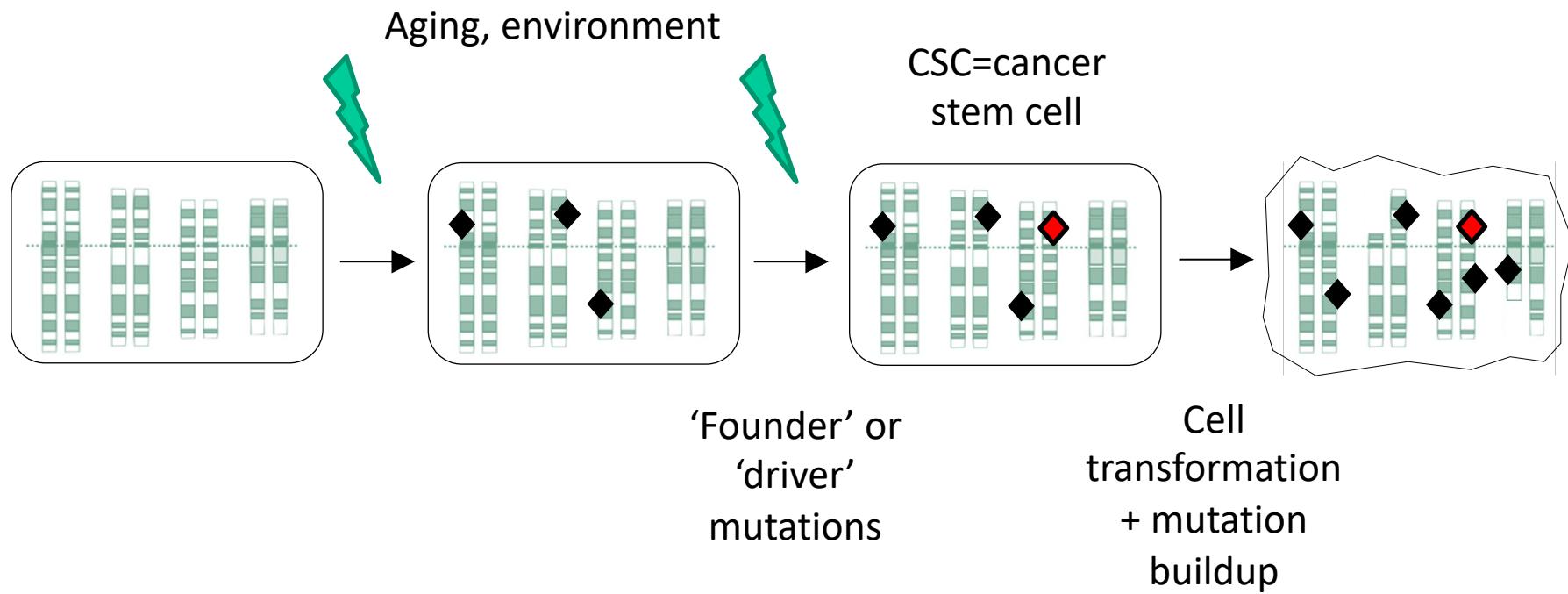
3 nucléotides = 1 codon

Encoder une protéine

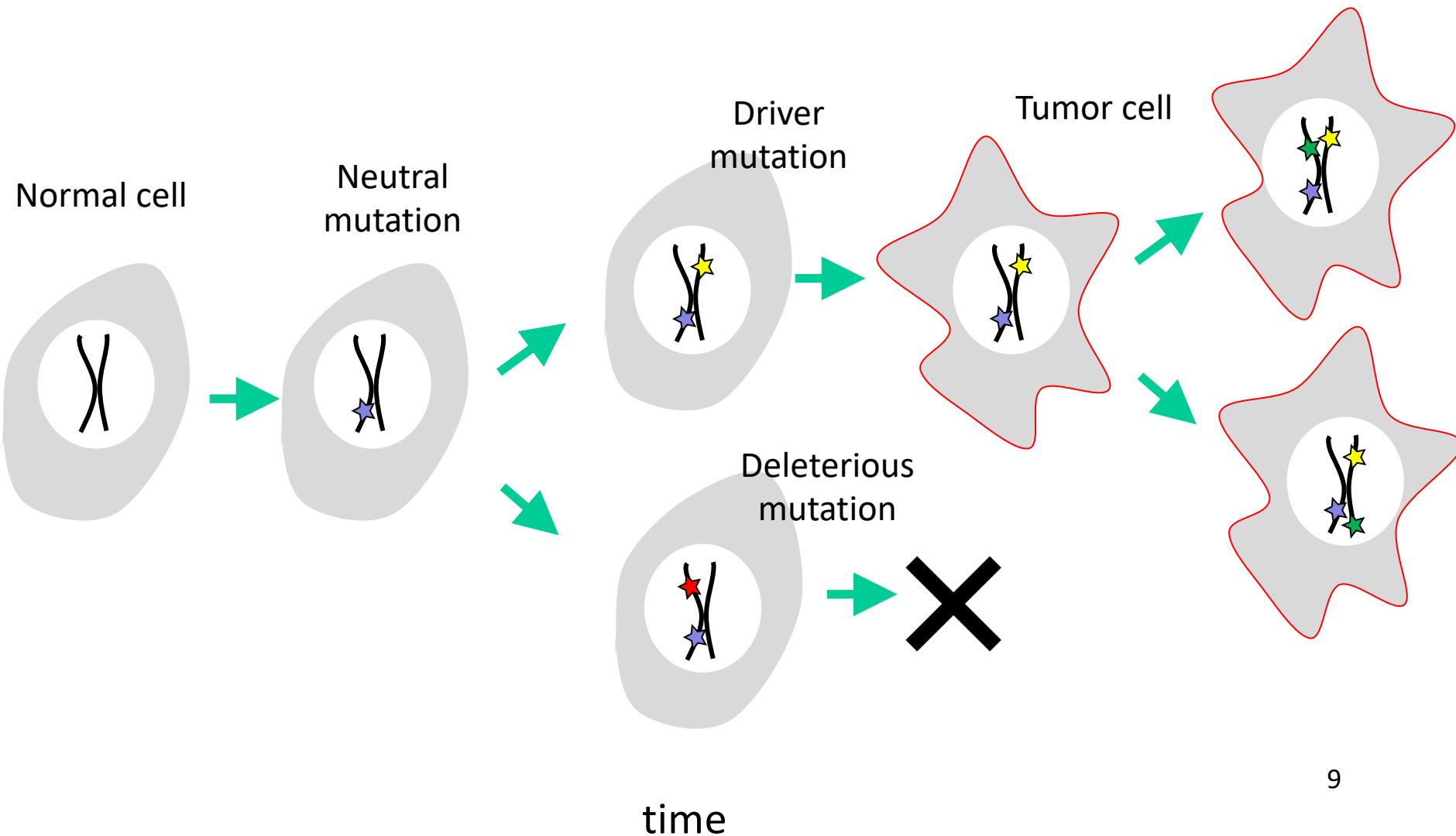
clé	valeur
ACG	Thr
AAG	Lys
GCG	Ala
TAG	stop
ATG	Met
...	...

Code génétique

- Le cancer: une maladie du génome

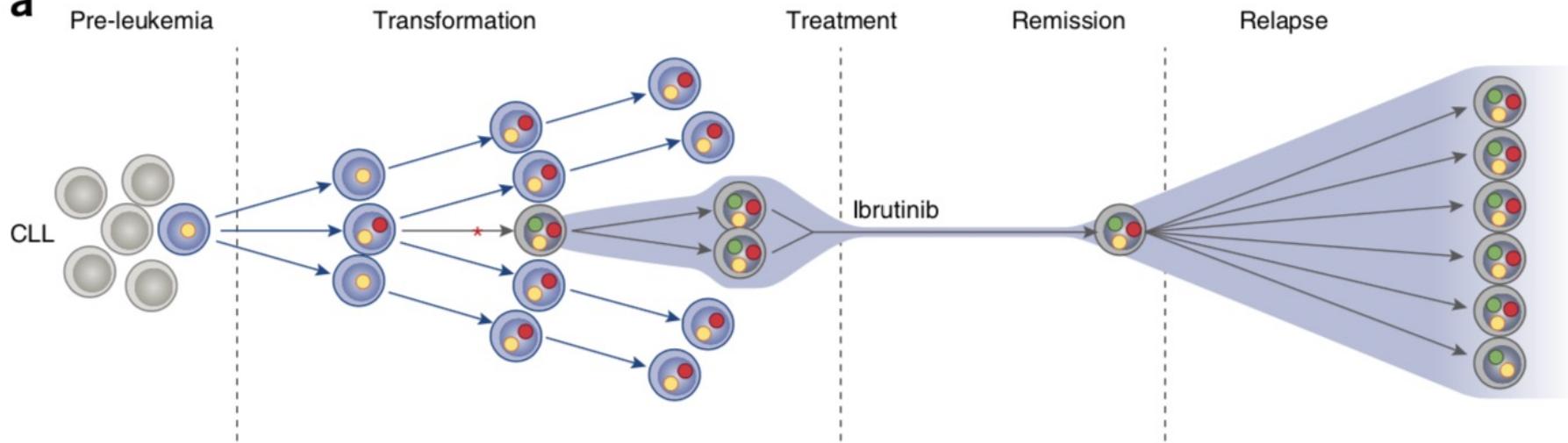


Evolution clonale des cellules cancéreuses



Clonal evolution of cancer

a

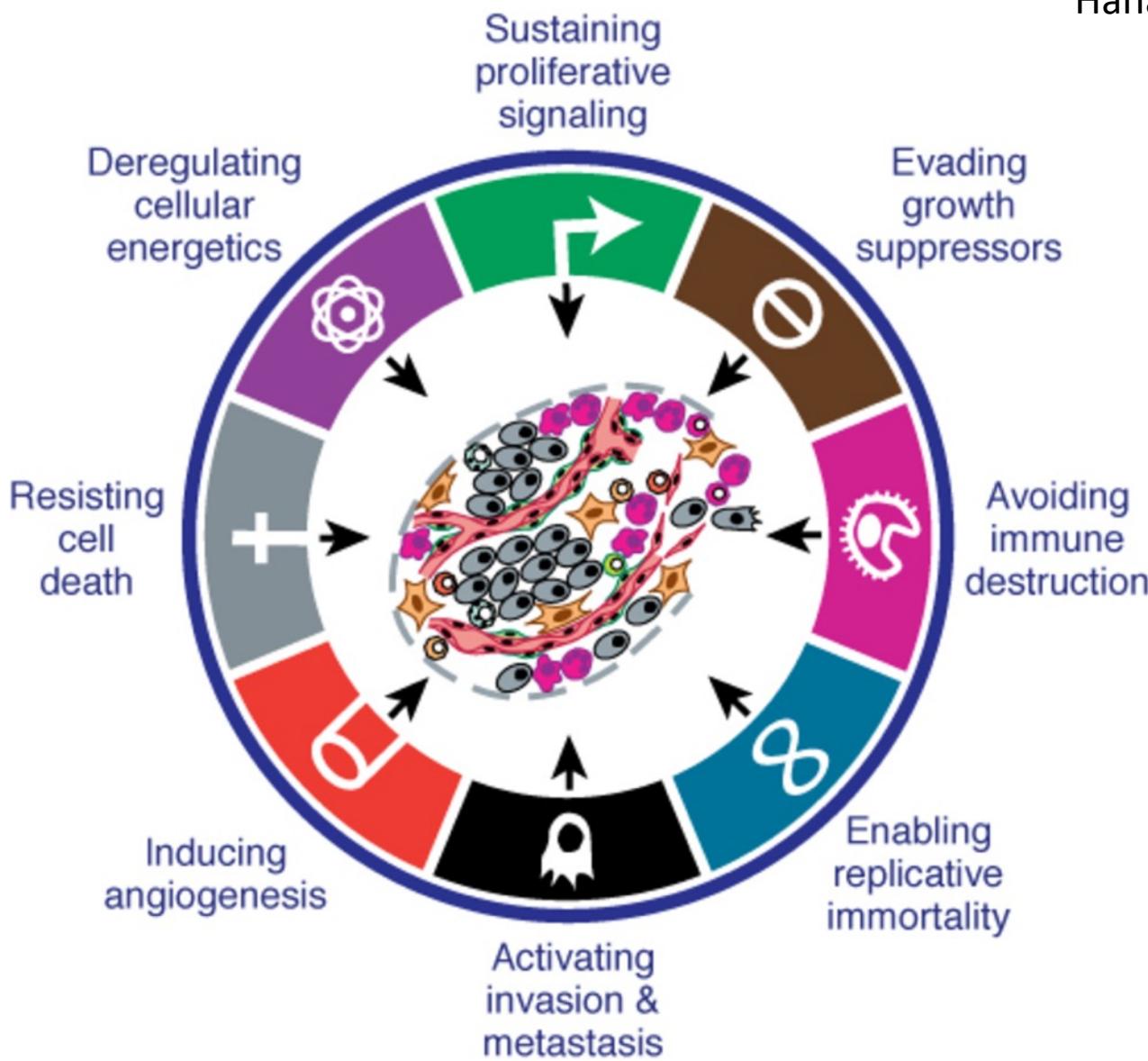


Ferrando & Lopez-Otin, Nature Med. 2017

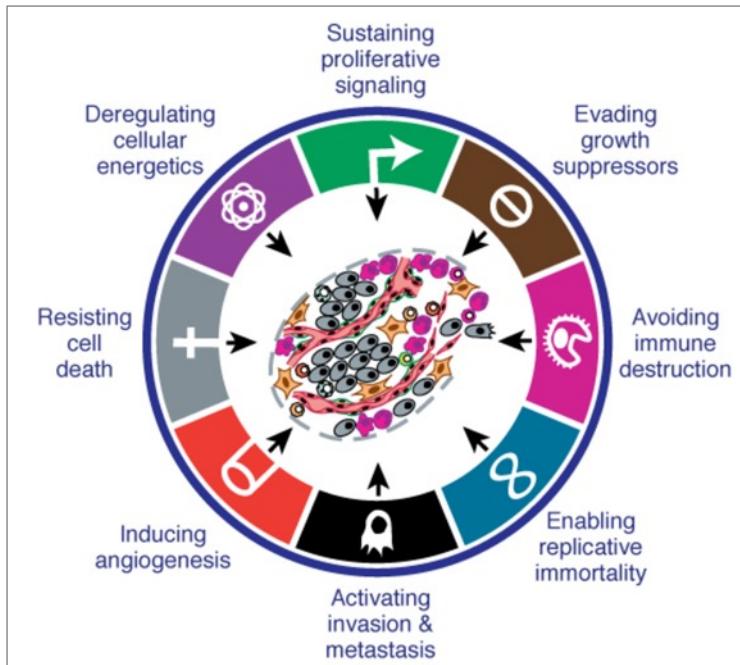
What is different in a cancer cell?

Hallmarks of cancer

Hanahan & Weinberg, 2015



Chemotherapy with nonspecific cancer drugs

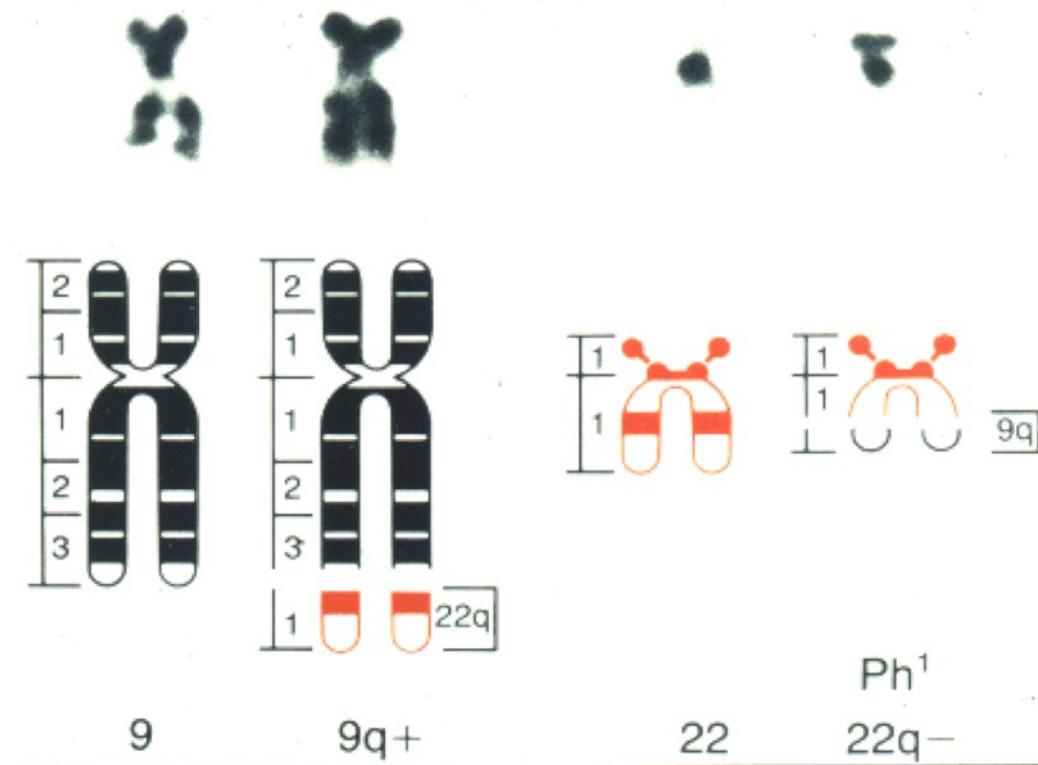


- Taxanes: block cell division
- Alkylating agents (platin salts): kill dividing cells
- Tamoxifene : block hormonal activation

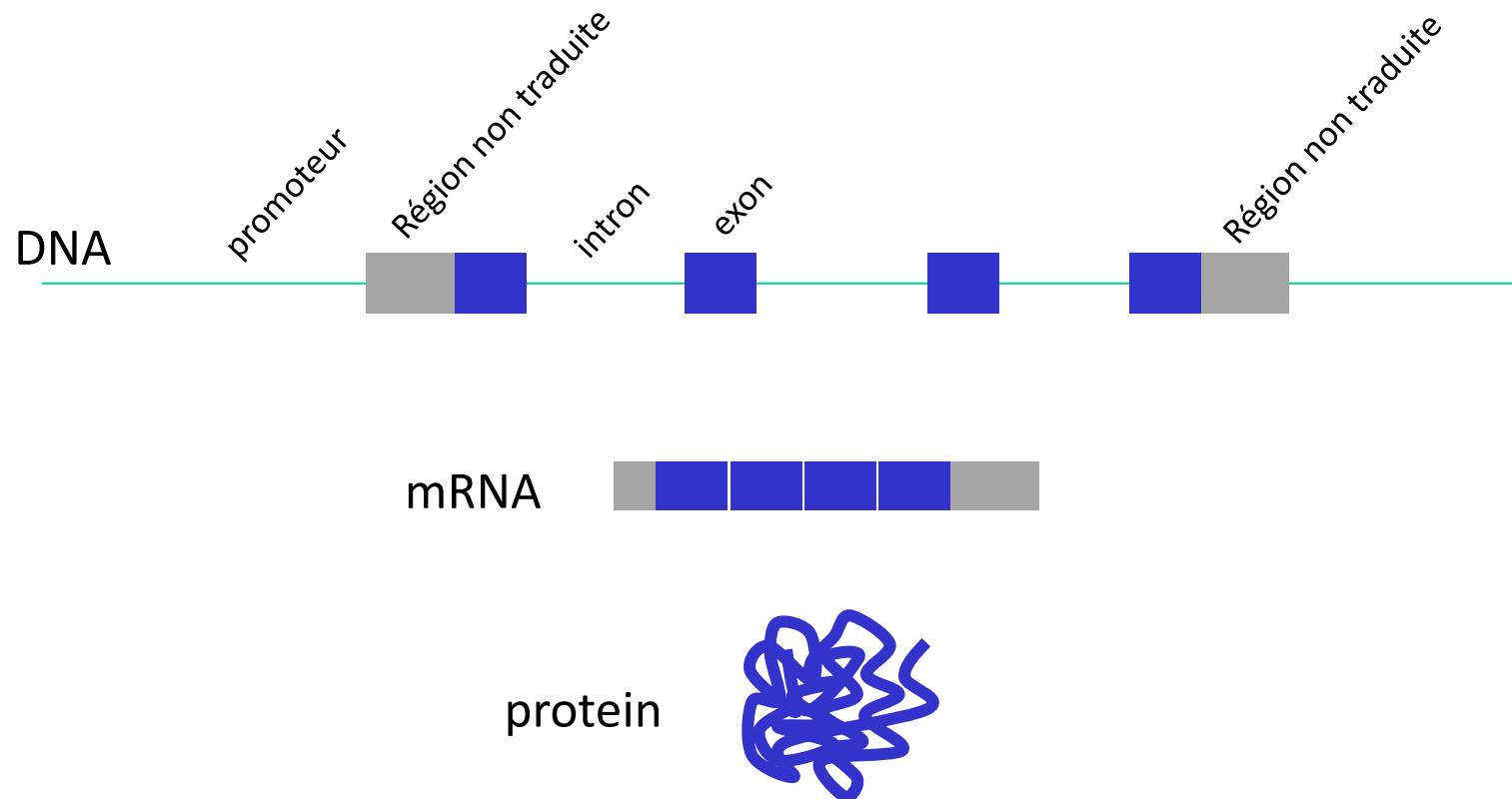
Precision medicine

Le « chromosome de Philadelphie »

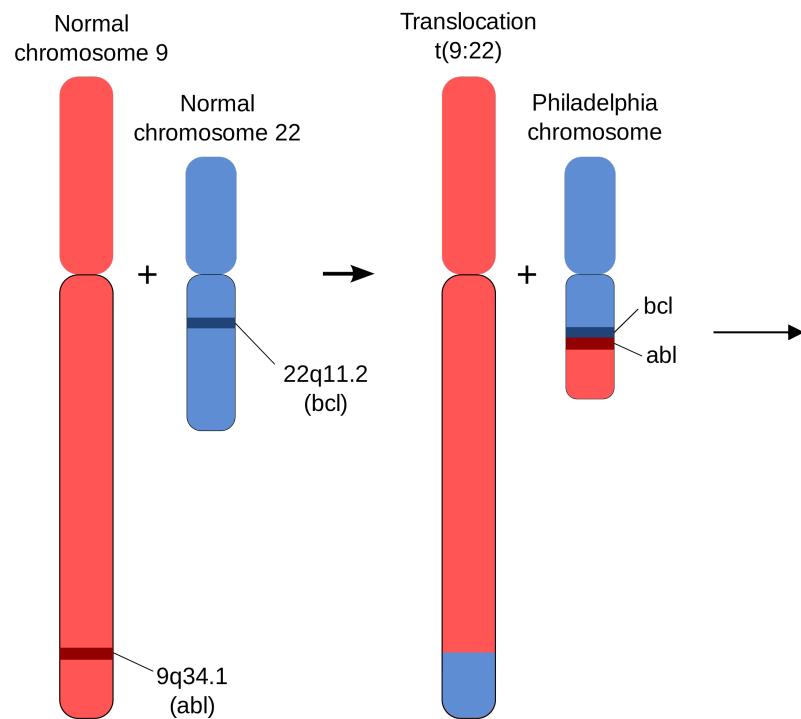
Translocation entre le chromosome 9 et le chromosome 22
dans la leucémie myéloïde chronique (LMC)



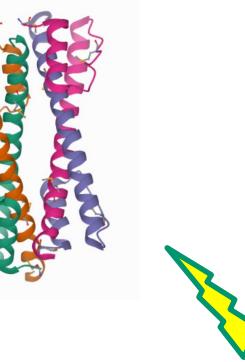
Rappel: structure d'un gène eucaryote (animaux, plantes, levures)



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 myéloïde chronique (LMC).



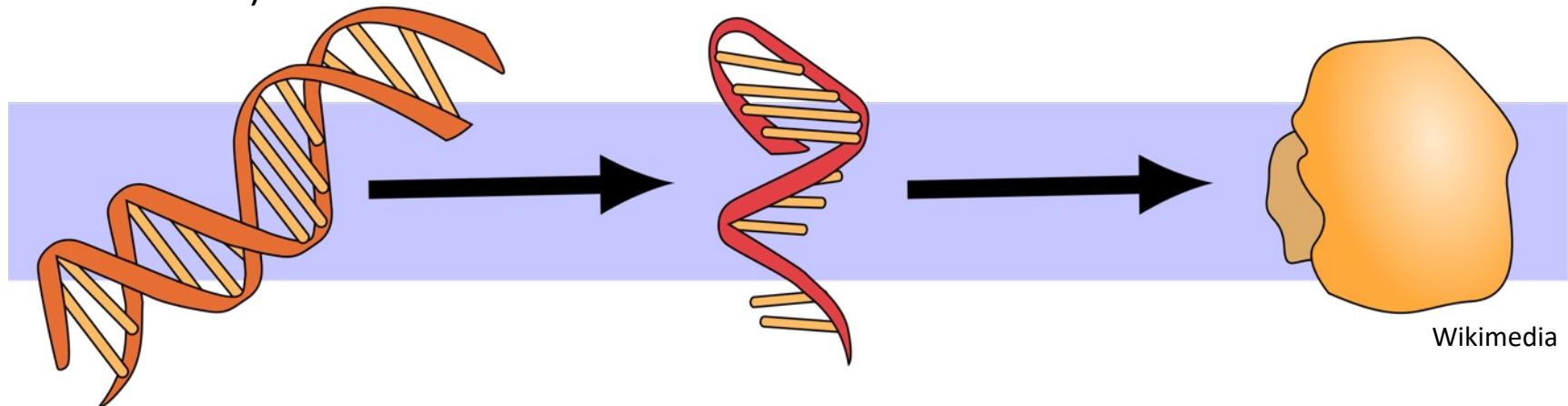
Protéine de fusion
BCR-ABL



Glivec
(inhibiteur de
tyrosine kinase)

Types of tumor-specific alterations

- Epigenetic changes (methylation, histones marks)



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

20 000



Cancer Drivers



500



Actionable



50

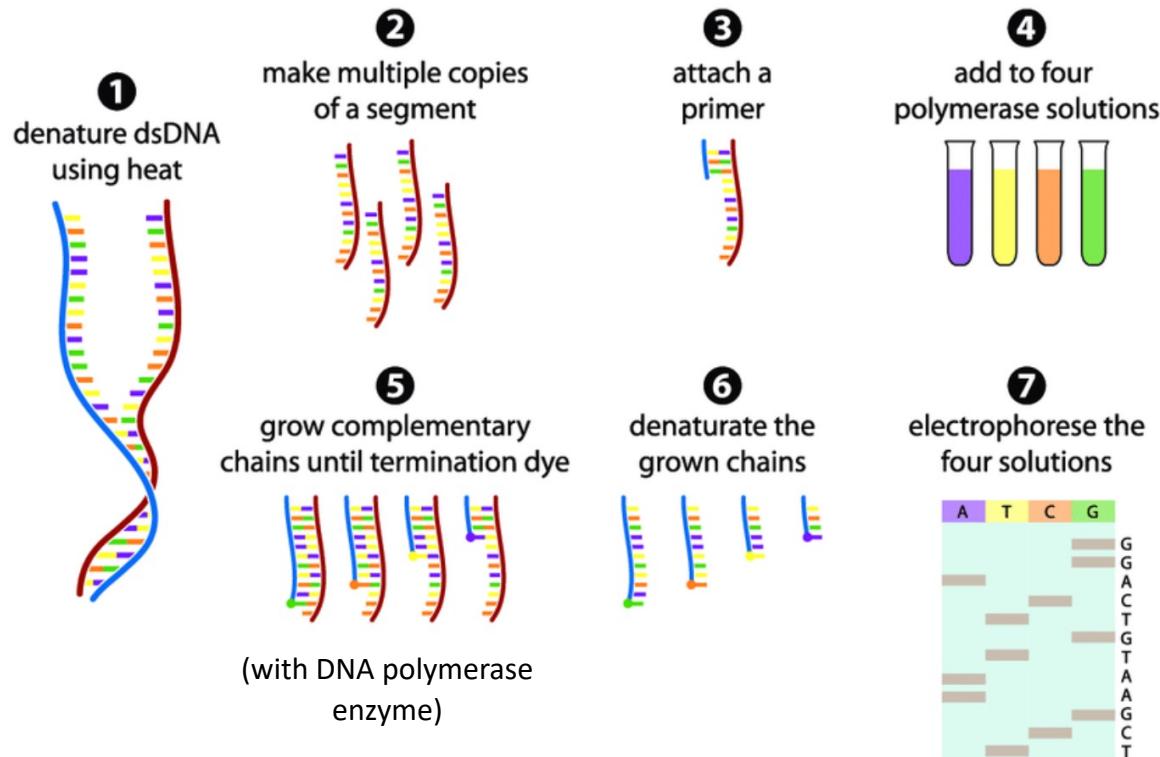


RNA and DNA sequencing:

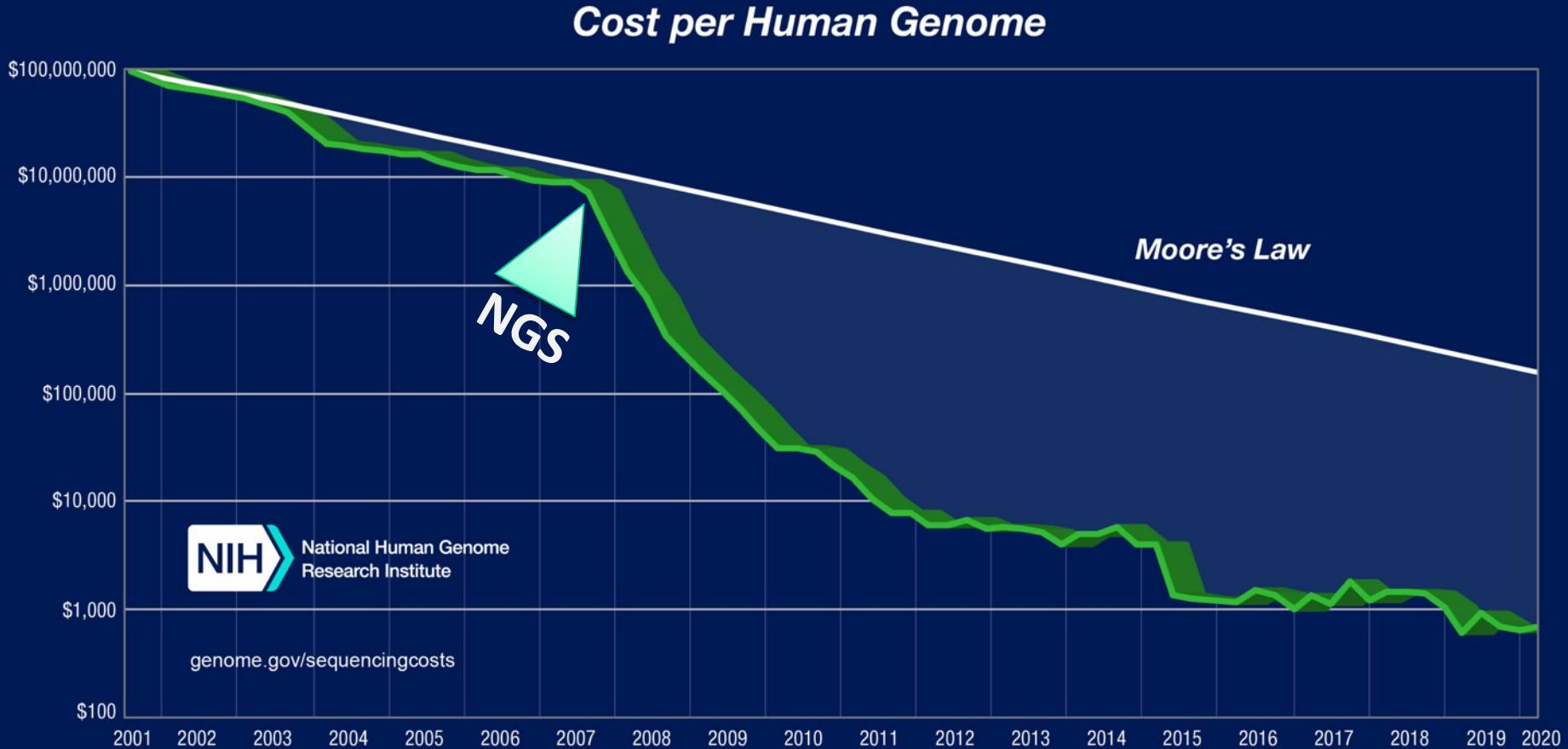
The main omics technologies in oncology

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)



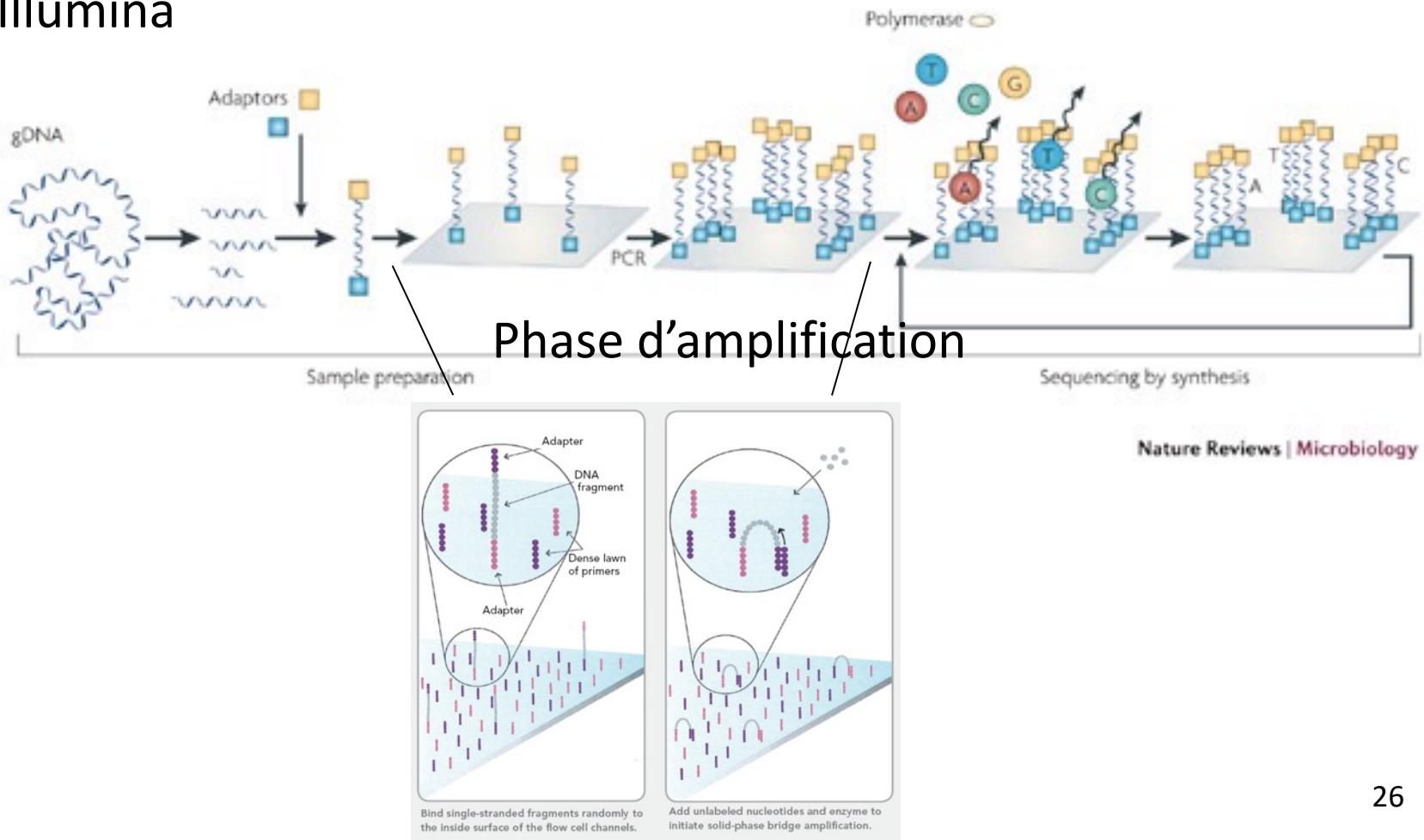
The NGS revolution



NGS :
Next Generation Sequencing
(2005-)

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

Illumina



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

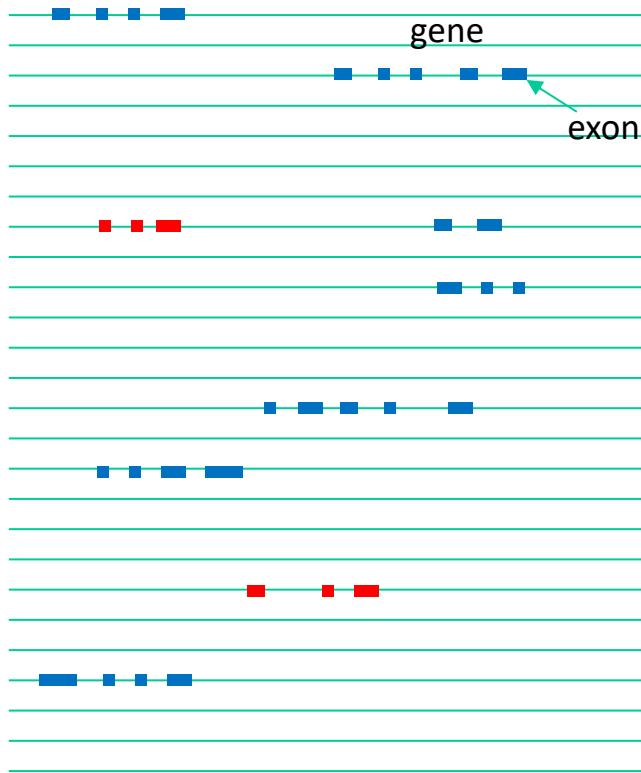
Les grandes applications des NGS

- 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)

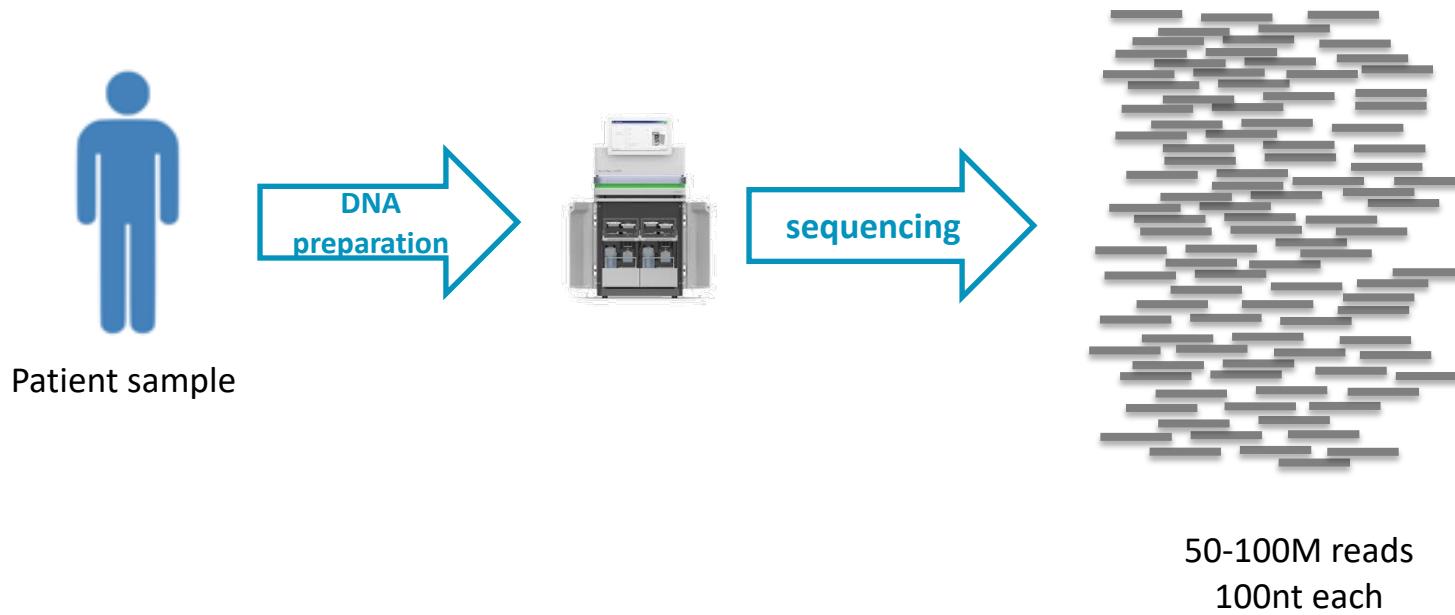
DNA sequencing types



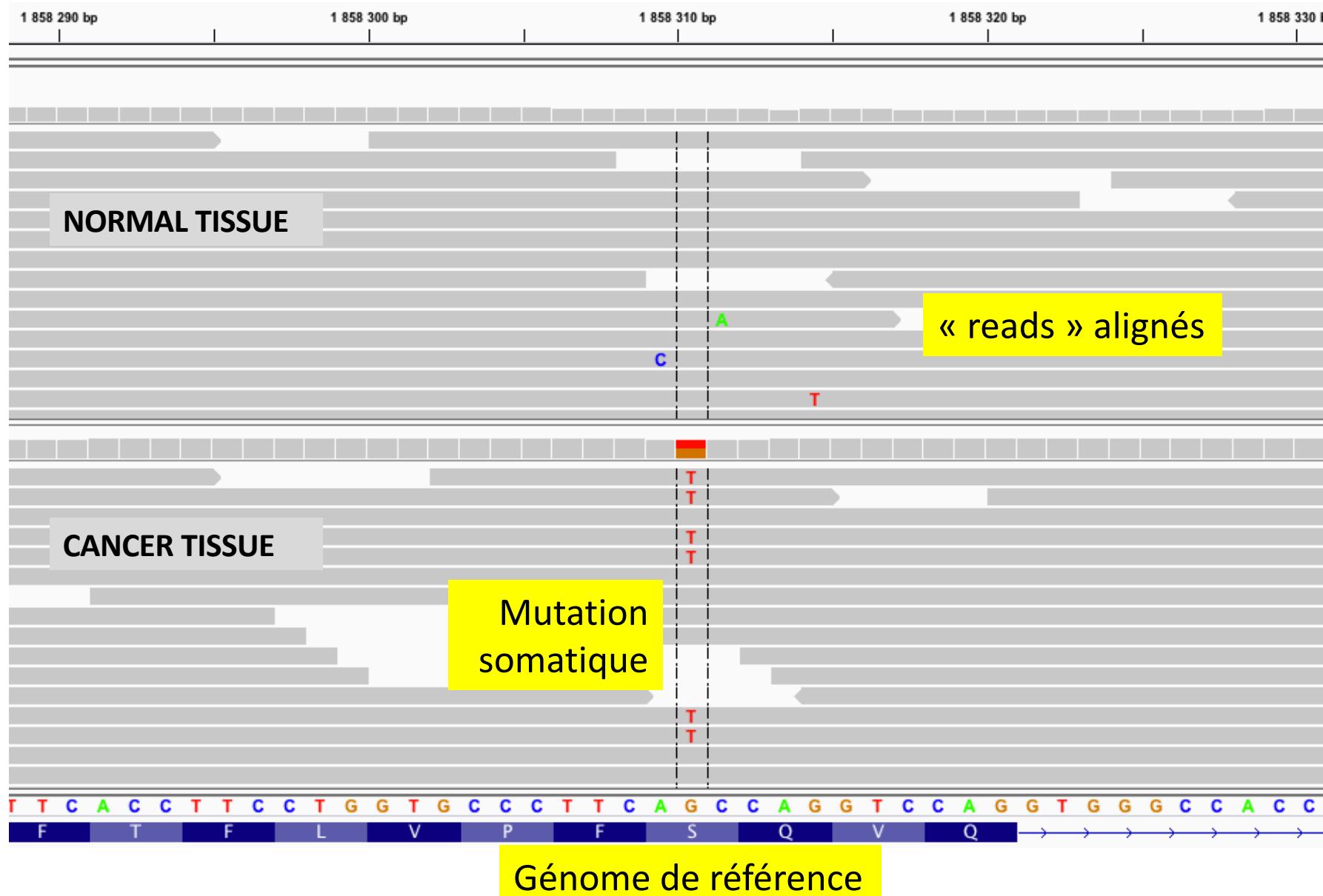
Chromosome sequence

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

Protocole DNA-seq



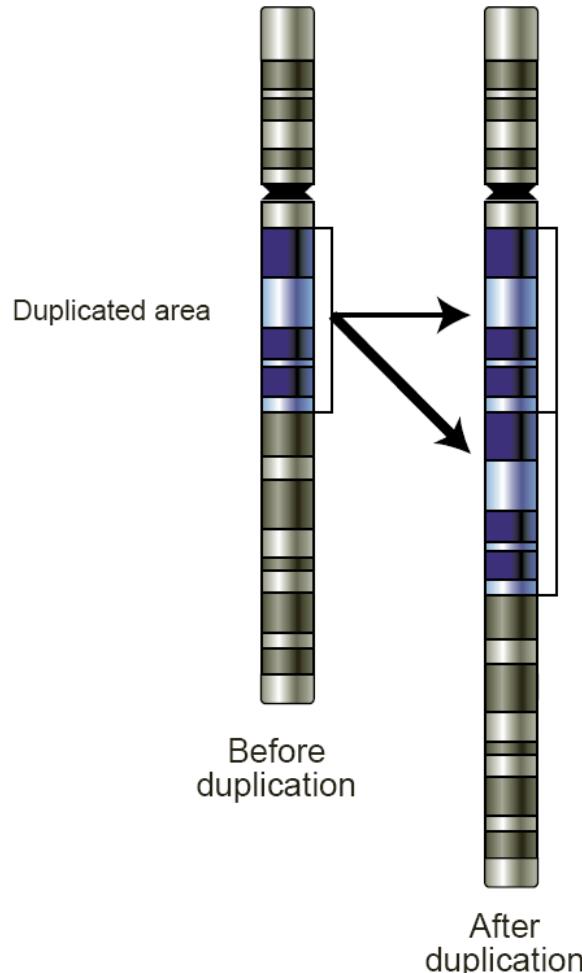
Événements recherchés



Événements recherchés par DNA-seq

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

Copy number variations



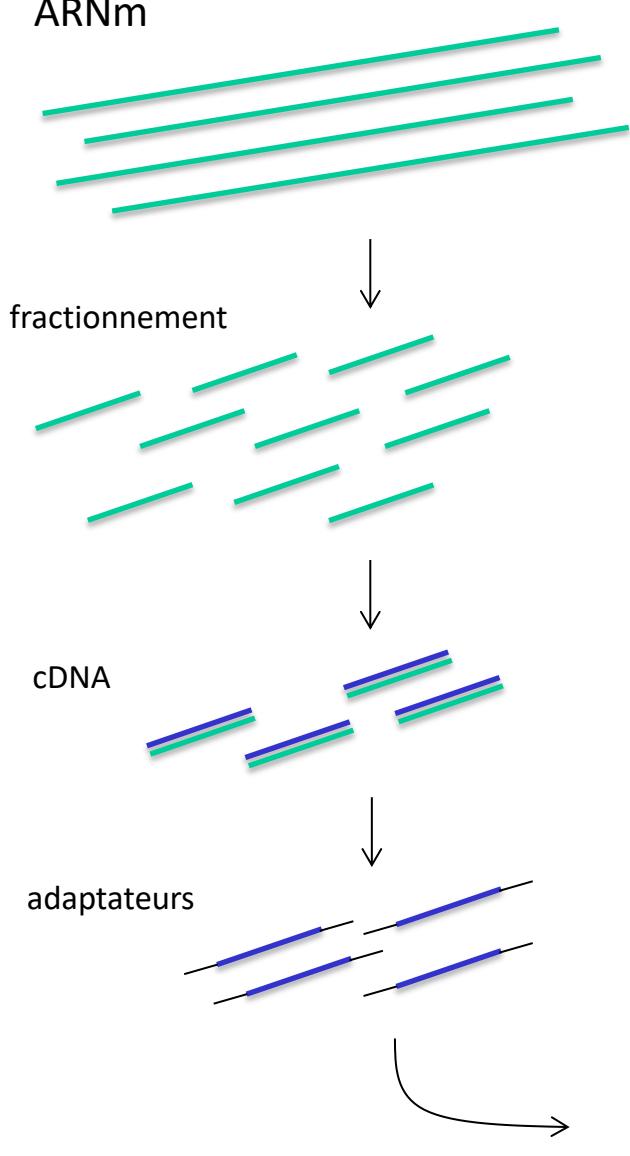
- Caused by recombination errors during DNA repair
- Can be loss or gain
- Oncogenic events= loss of a tumor suppressor or gain of a oncogene

Cf cours
Bastien Job

RNA-seq

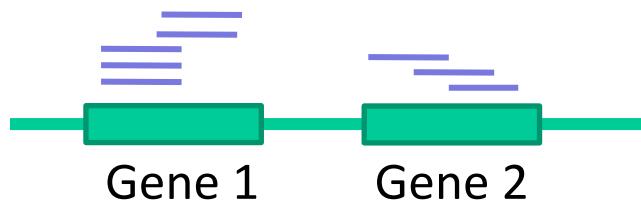
- Pour l'étude du transcriptome
 - Mesure de l'expression de tous les gènes, simultanément

RNA-Seq

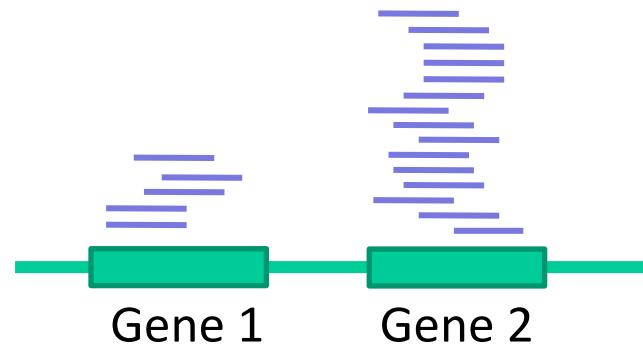


Séquençage

Mesure d'expression par RNA-seq

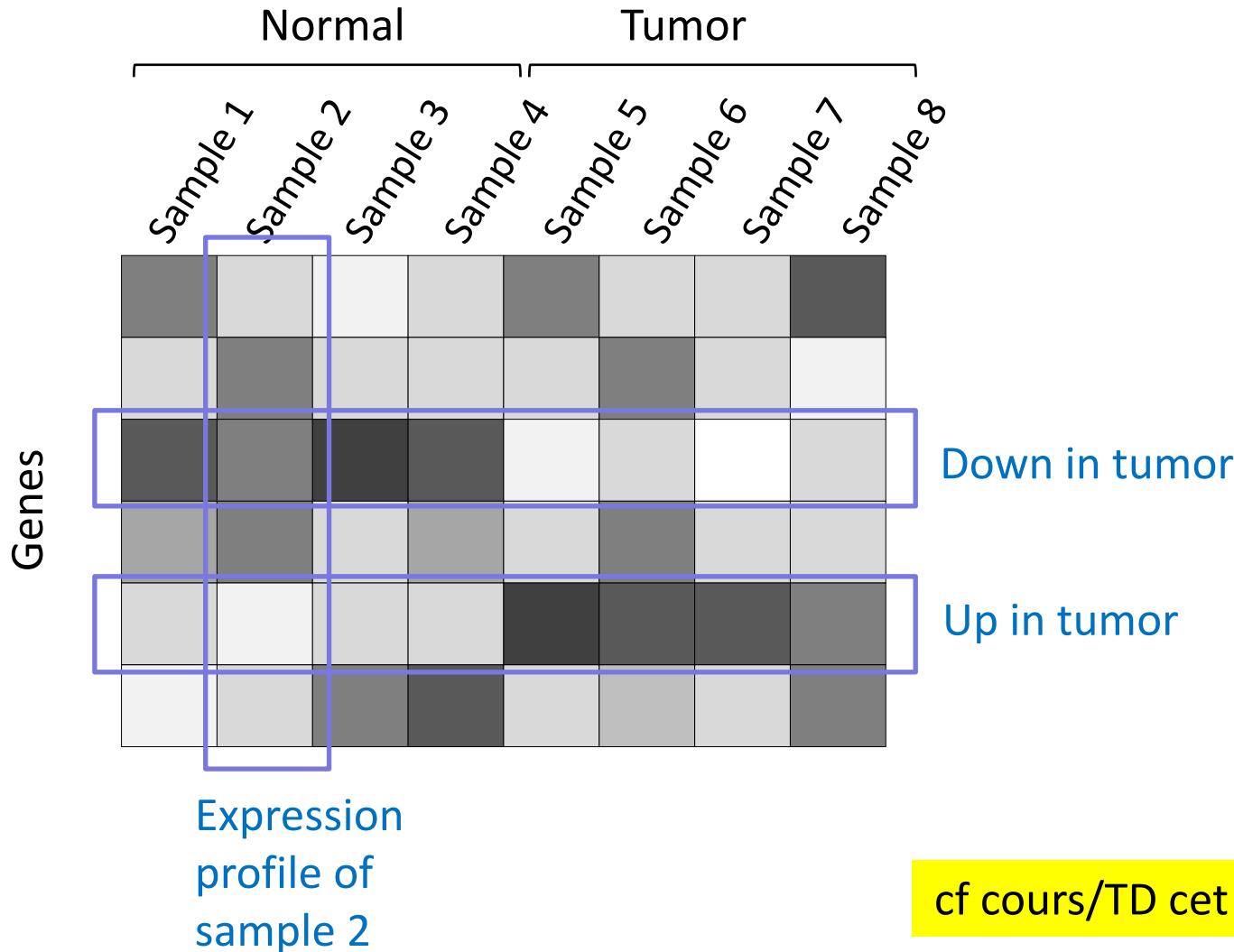


Sample 1



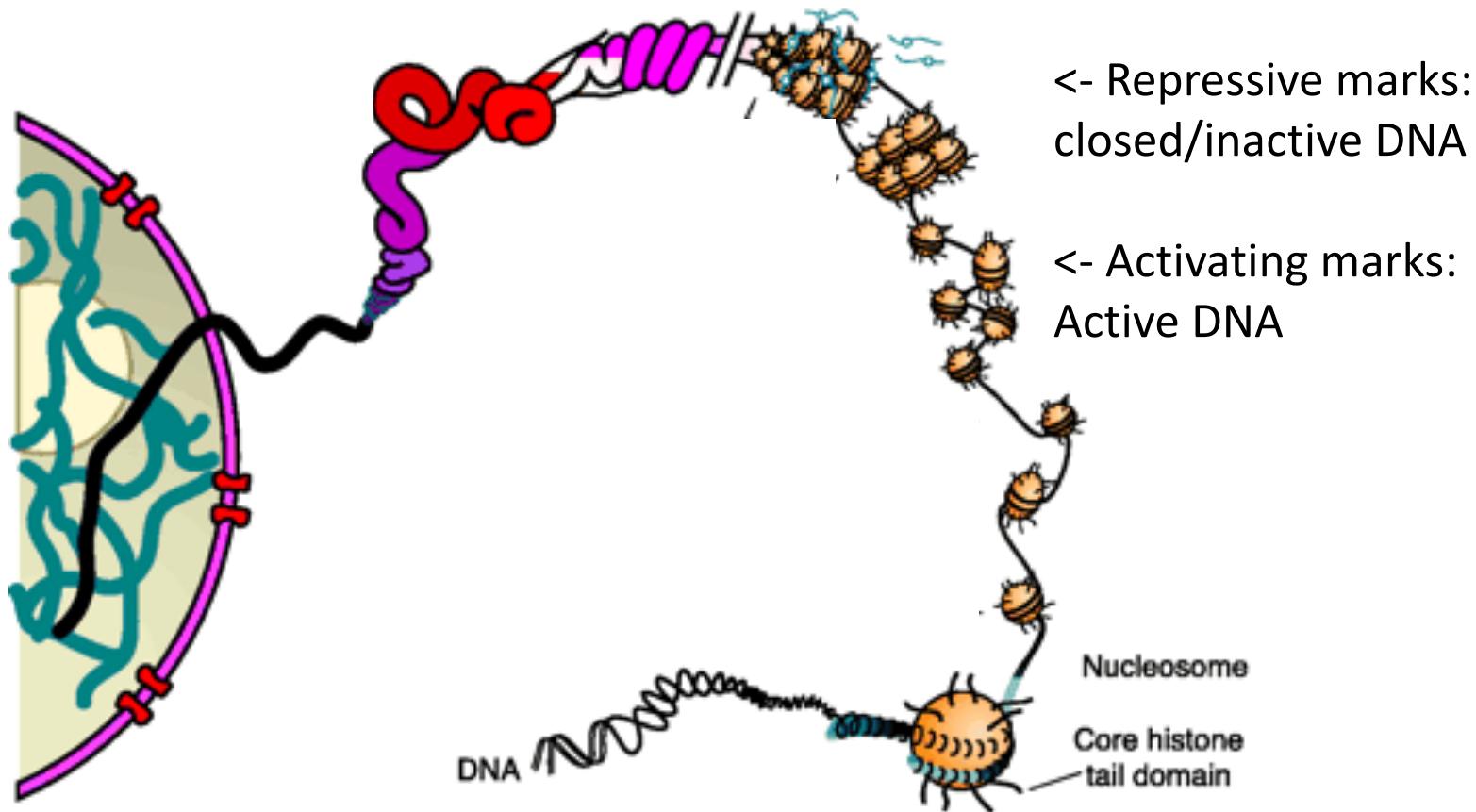
Sample 2

Differential expression analysis & expression profiling

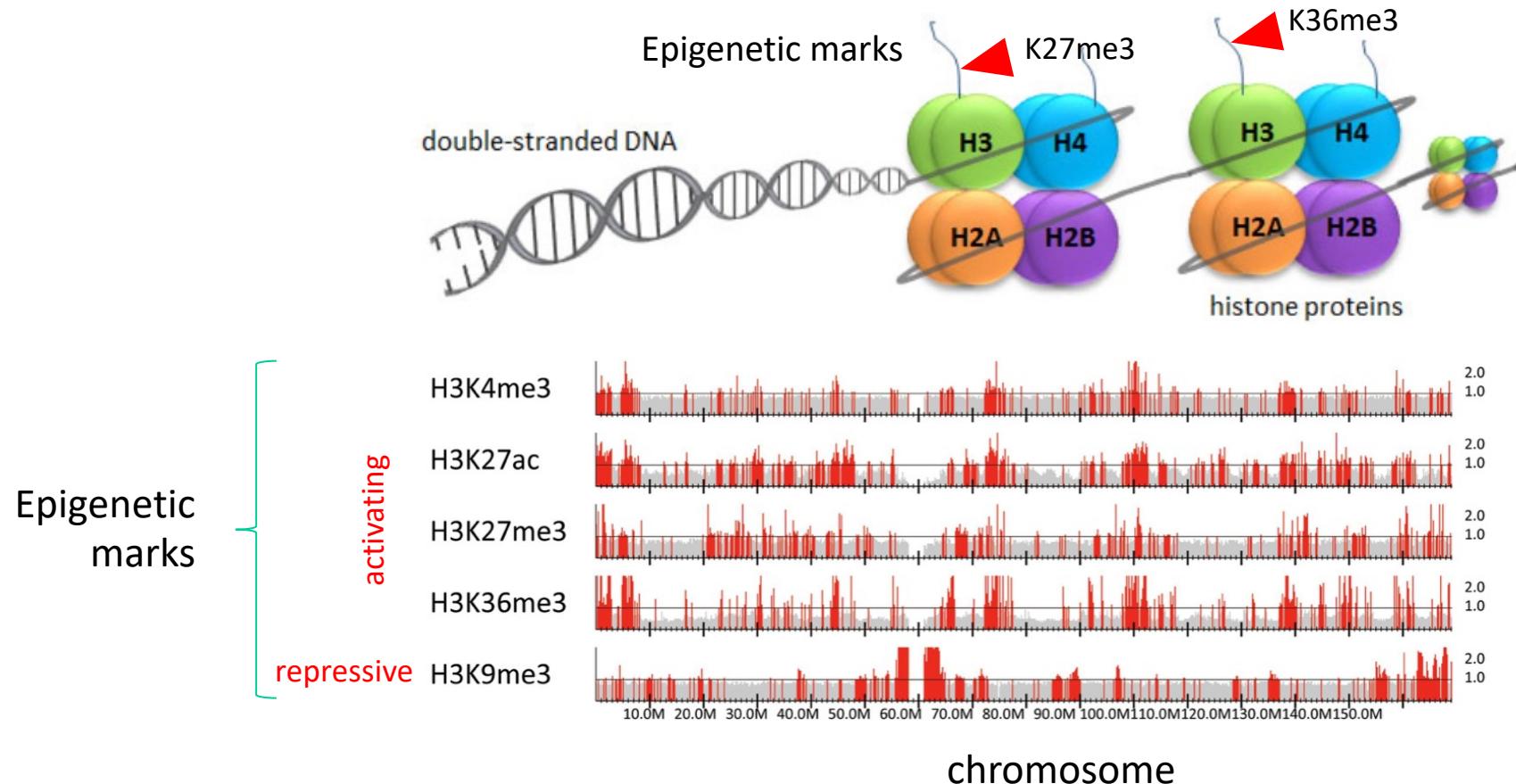


ChIP-seq in oncology

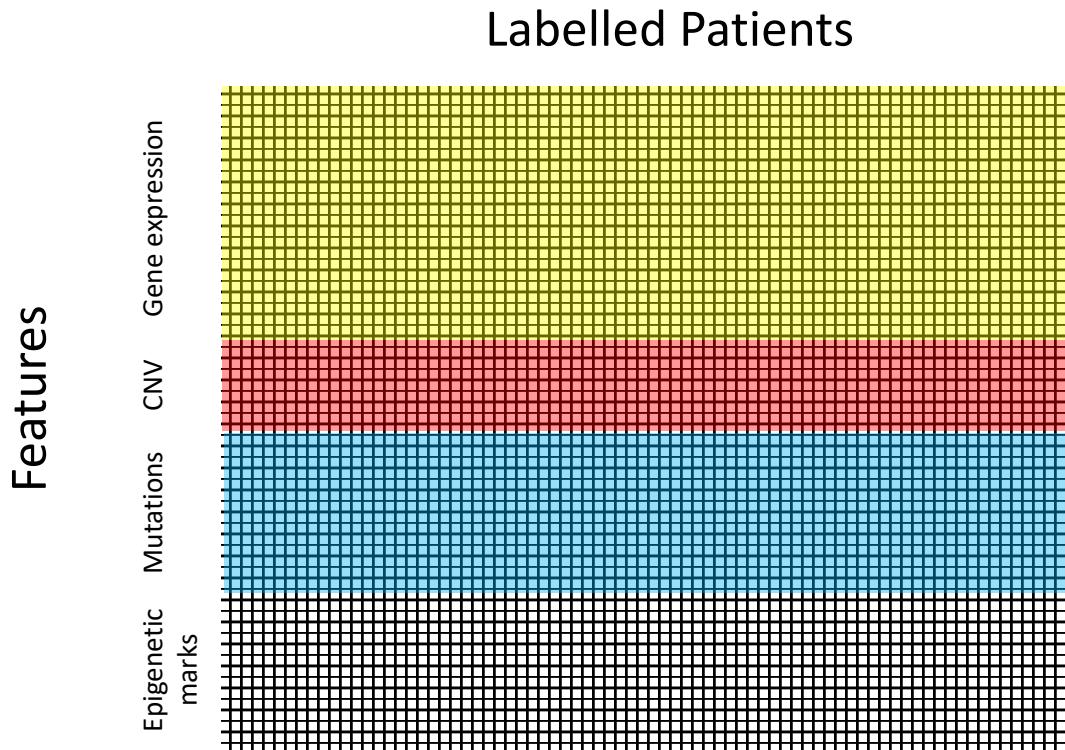
ChIP-seq = Chromatin ImmunoPrecipitation & Sequencing
Identifies epigenetic marks on chromosomes



ChIP-seq in oncology

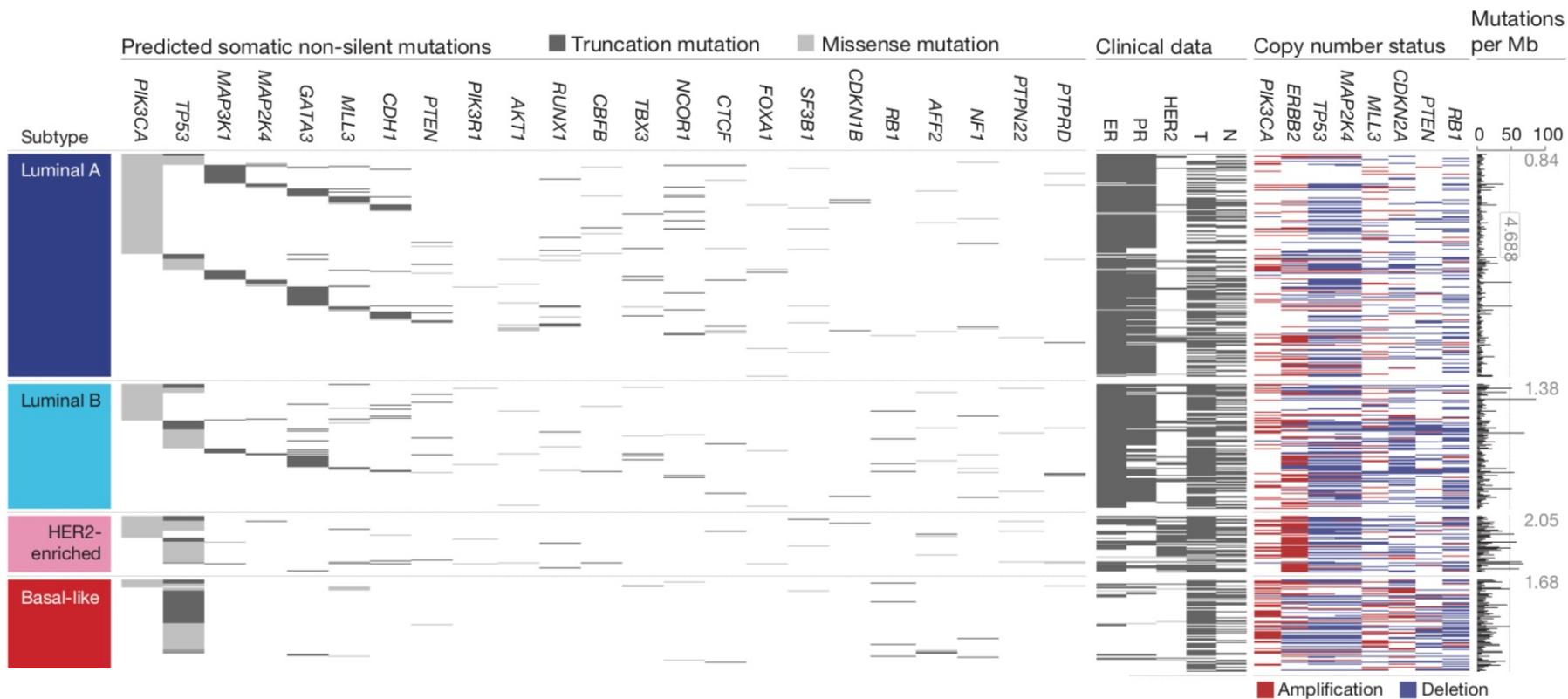


Big data in oncology



Omics cohorts in cancer TCGA project: 10,000 patients

825 Patients

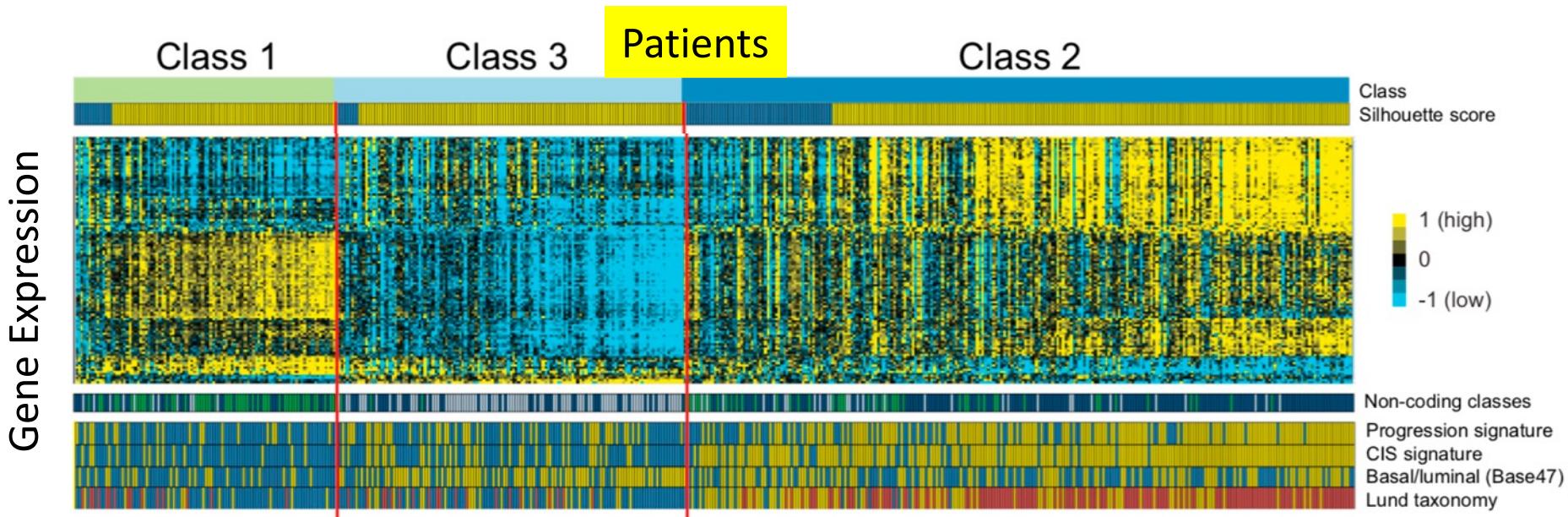


TCGA Breast cancer cohort (2012): 825 patients.
Stratified by cancer type

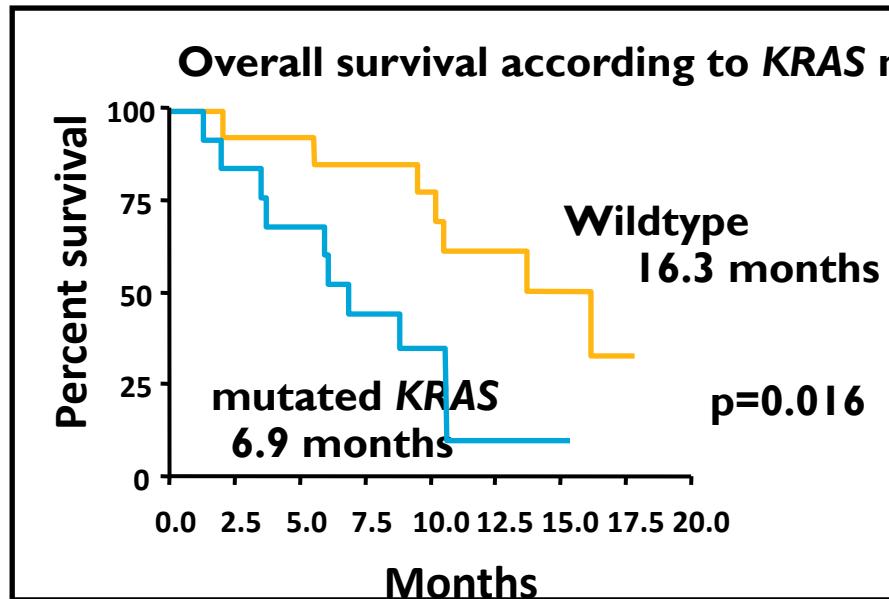
TCGA consortium, Nature, 2012

Subtype discovery (unsupervised classification)

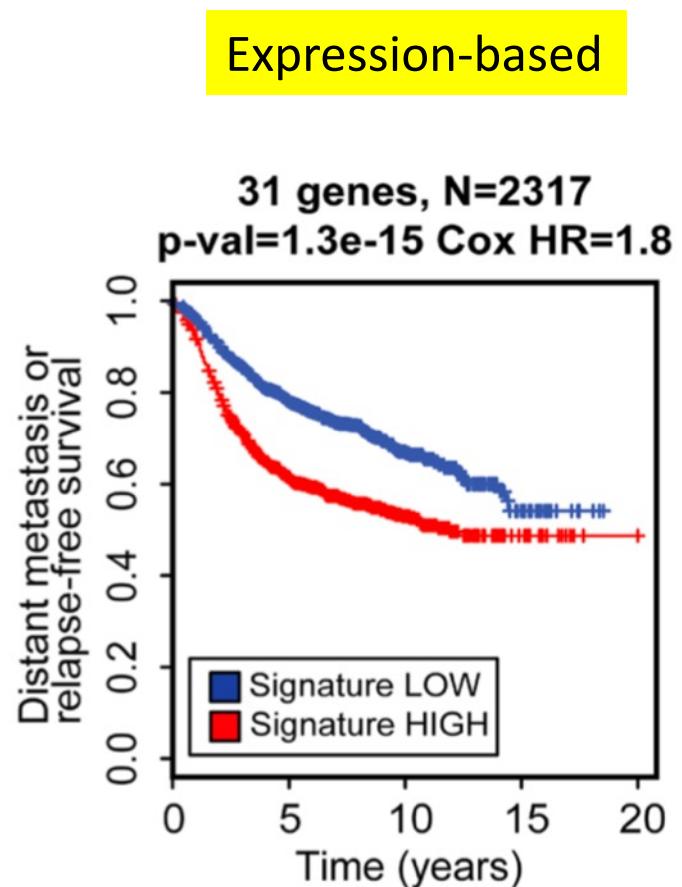
Gene expression profiling in 460 Urothelial carcinoma:
A 117-gene signature.



Modèles prédictifs de survie



DNA-based



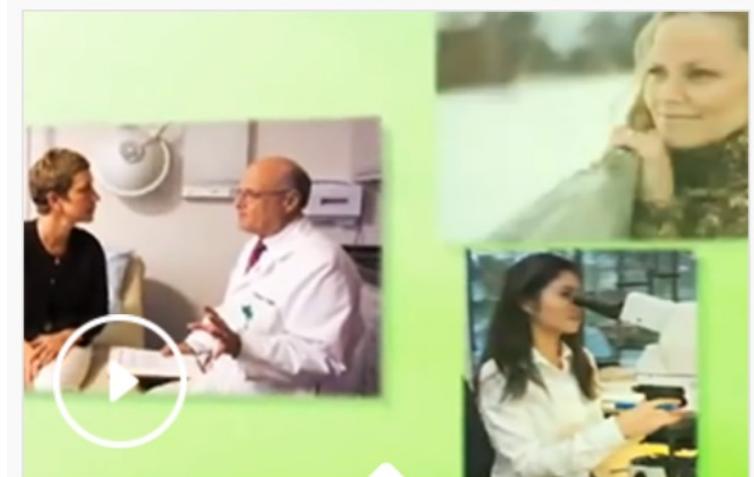
Transcriptional signatures: current applications

Diagnostic

- Tumor vs normal
- Tumor subtyping

Precision medicine

- Response to treatment
- Relapse prediction



Genomic Health et le test Oncotype DX

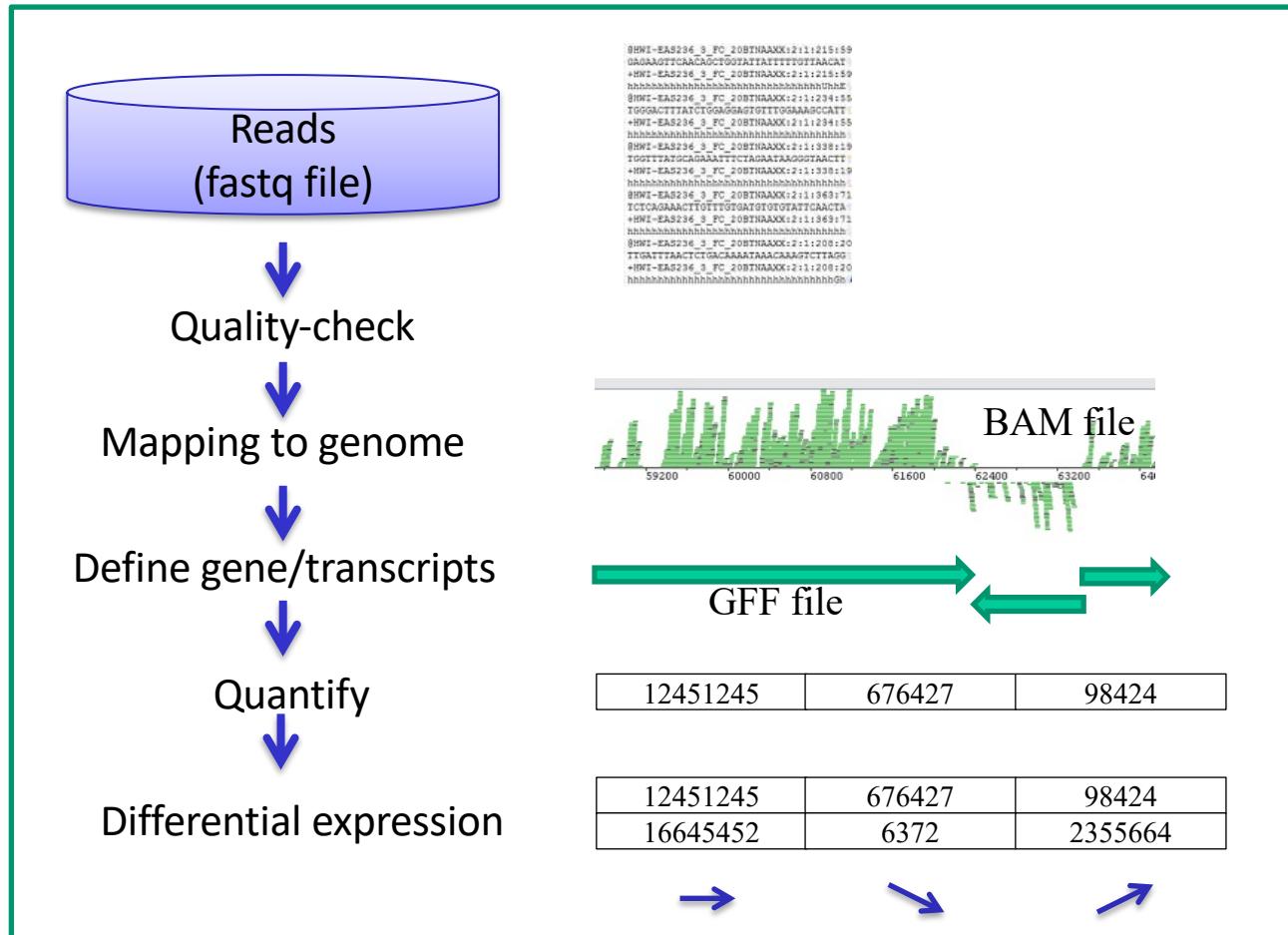
Traitement du Cancer: Comment les tests Oncotype DX permettent de personnaliser les décisions thérapeutiques.

Les outils d'analyse

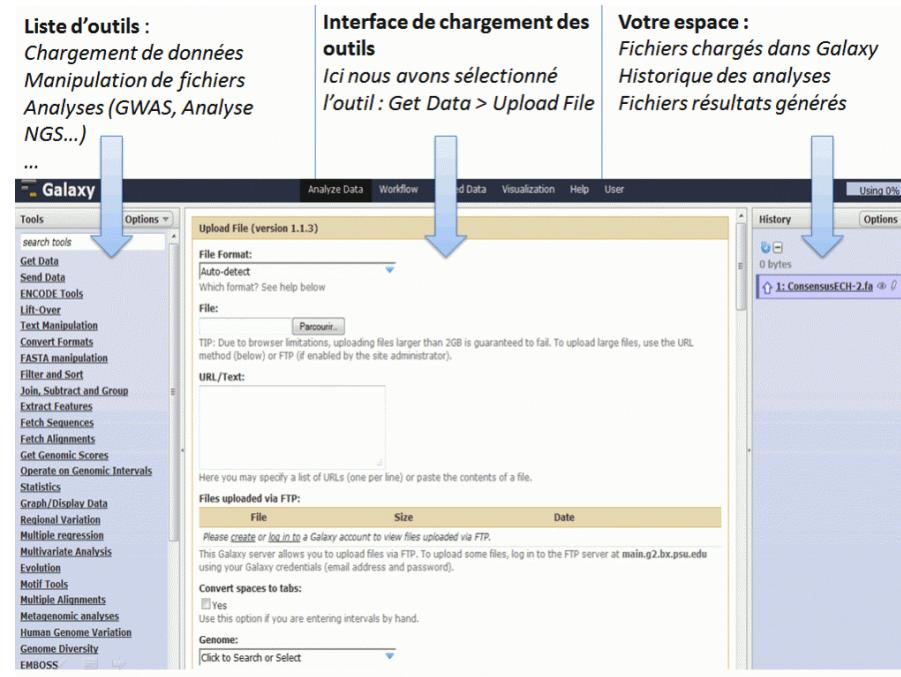
« Pipelines » & « workflows »

« Bricks » from Unix open source programs

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



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

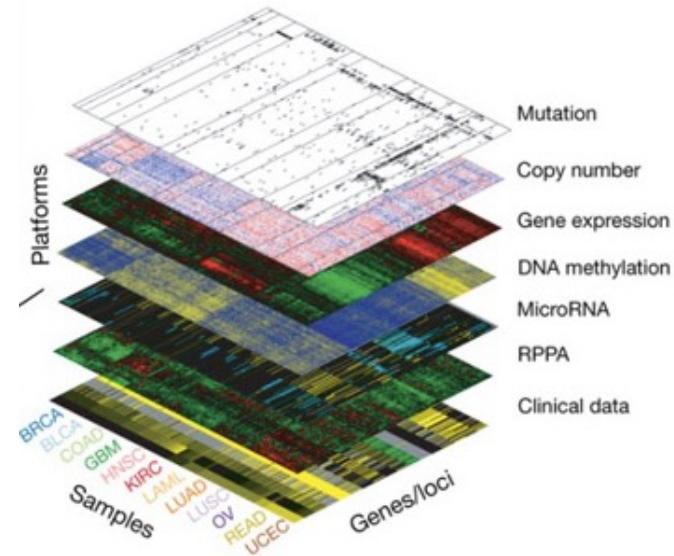
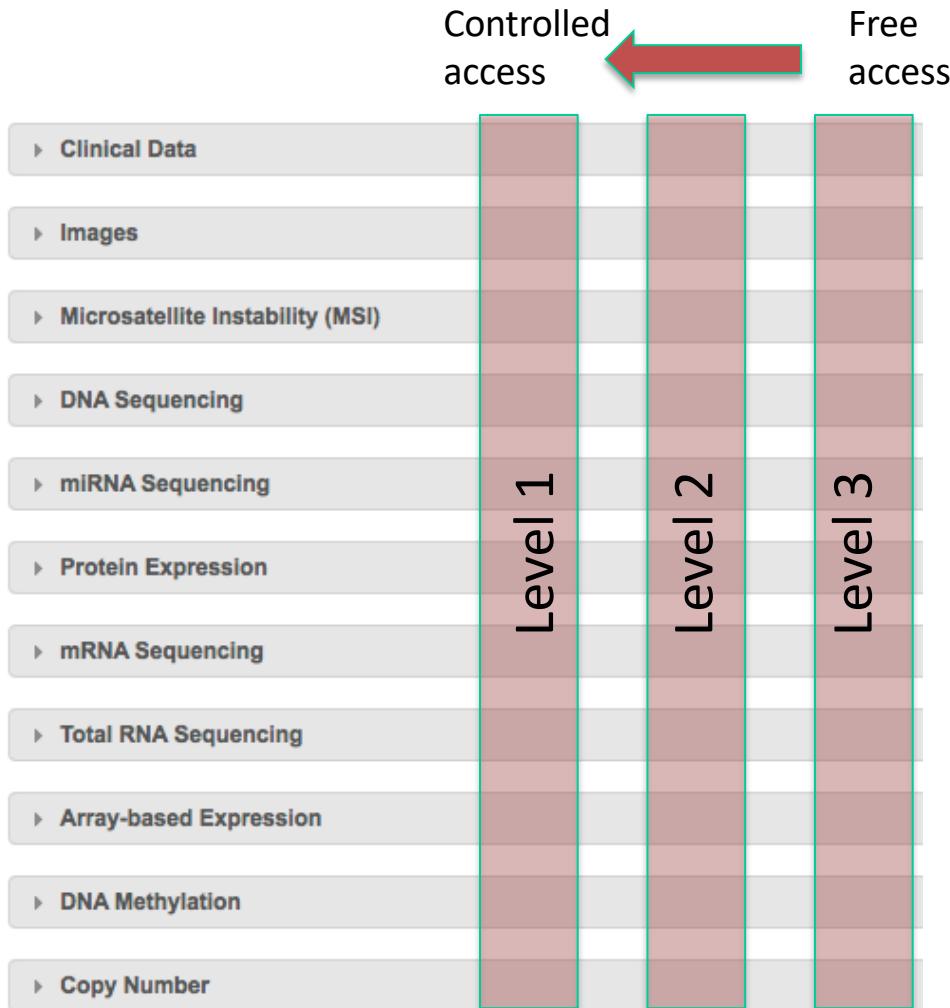
- TCGA: the Cancer Genome Atlas
- COSMIC
- cBioPortal
- CCLE: Cancer Cell Lines Encyclopedia
- GDSC: Genomics of Drug Sensitivity in Cancer
- dbGaP: database of Genotypes and Phenotypes
- GEO: Gene Expression Omnibus
- ArrayExpress



TCGA

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

TCGA data types and levels



TCGA access via the GDC portal (Genomics Data Commons)

NATIONAL CANCER INSTITUTE
GDC Data Portal

Home Projects Exploration Repository

Quick Search Login Cart 0 GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 8.0 - August 22, 2017

PROJECTS	PRIMARY SITES	CASES
39	29	14 551
FILES	GENES	MUTATIONS
274 724	22 144	3 115 606

Cases by Primary Site

Cancer Type	Cases
Adrenal Gland	~400
Bile Duct	~50
Bladder	~400
Blood	1000
Bone	~400
Bone Marrow	~200
Brain	1000
Breast	1000
Cervix	~400
Colorectal	~500
Esophagus	~100
Eye	~50
Head and Neck	~500
Kidney	1500
Liver	~400
Lung	1000
Lymph Nodes	~50
Nervous System	1000
Ovary	~500
Pancreas	~100
Pleura	~50
Prostate	~500
Skin	~500
Soft Tissue	~300
Stomach	~500
Testis	~50
Thymus	~50
Thyroid	~500
Uterus	~500

From TCGA to PCAWG

- PCAWG¹: a collaboration TCGA-ICGC² 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. »

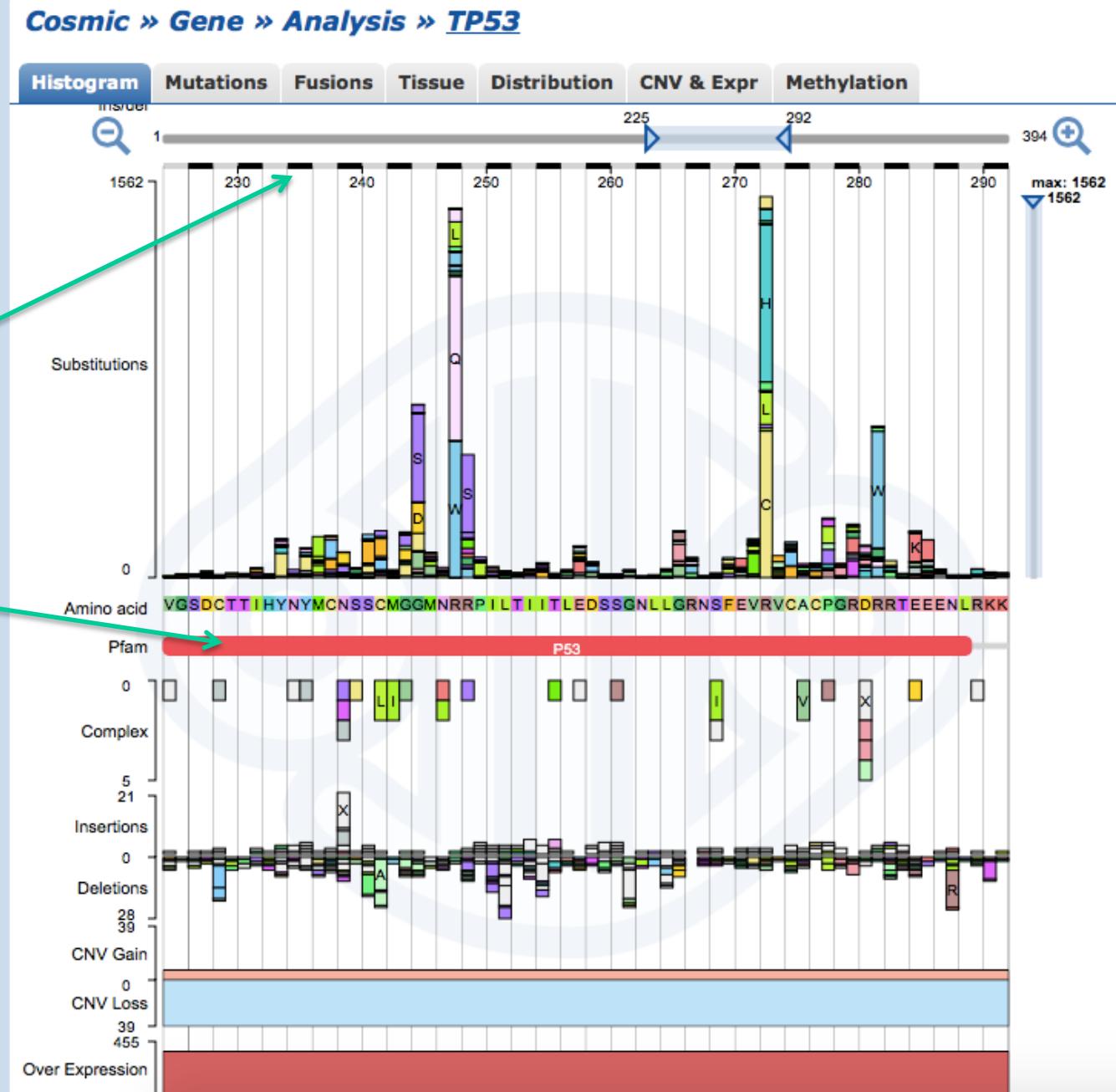


- Expert-curated database of cancer somatic mutations & other events
- 2022 (V97):
 - 23M variants génomiques
 - 19k gene fusions
 - 1.2 M CNV
 - 9M gene expression variants
 - 736 cancer genes

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

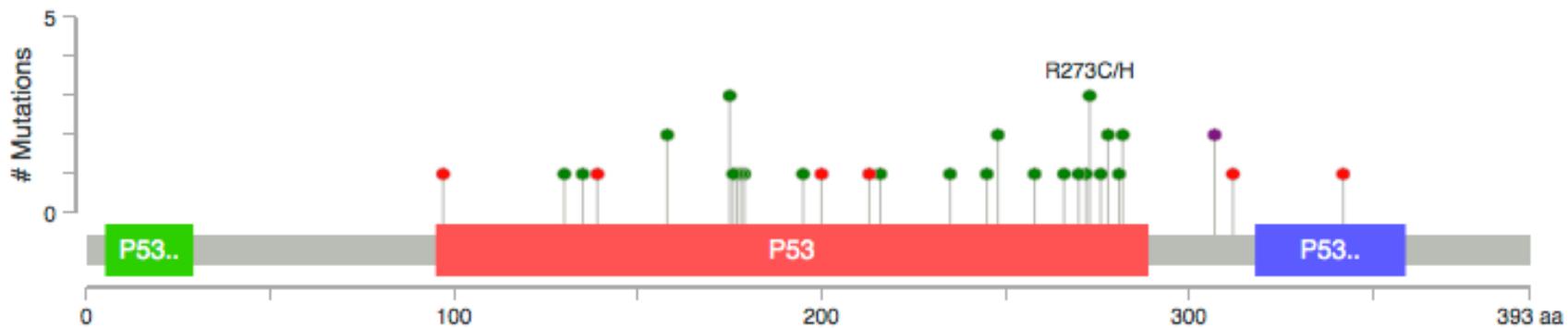


- Integration of Data from >350 cancer genomics studies.
- Focus on analysis tools
 - Mutual exclusivity
 - Gene networks

'Lollipop plots': mutations on proteins

TP53: [Somatic Mutation Rate: 34.1%]

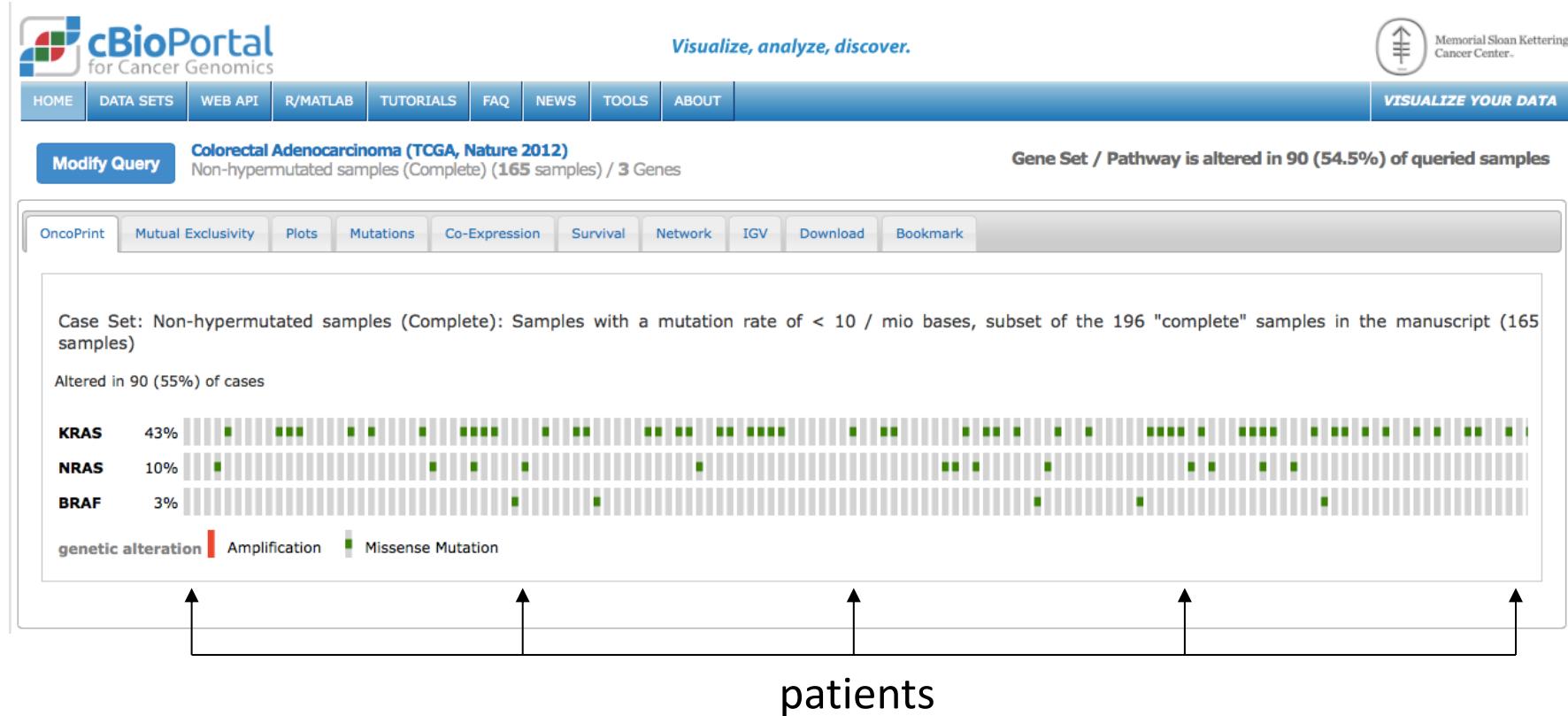
P53_HUMAN PDF SVG Customize Color Codes



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

See also « MutationMapper » tool

« Oncoplot » view



Mutual exclusivity

cBioPortal for Cancer Genomics

Visualize, analyze, discover.

Memorial Sloan Kettering Cancer Center.

HOME DATA SETS WEB API R/MATLAB TUTORIALS FAQ NEWS TOOLS ABOUT VISUALIZE YOUR DATA

Modify Query Colorectal Adenocarcinoma (TCGA, Nature 2012)
Non-hypermutated samples (Complete) (165 samples) / 3 Genes

Gene Set / Pathway is altered in 90 (54.5%) of queried samples

OncoPrint Mutual Exclusivity Plots Mutations Co-Expression Survival Network IGV Download Bookmark

Case Set: Non-hypermutated samples (Complete): Samples with a mutation rate of < 10 / mio bases, subset of the 196 "complete" samples in the manuscript (165 samples)

Altered in 90 (55%) of cases

KRAS 43% 

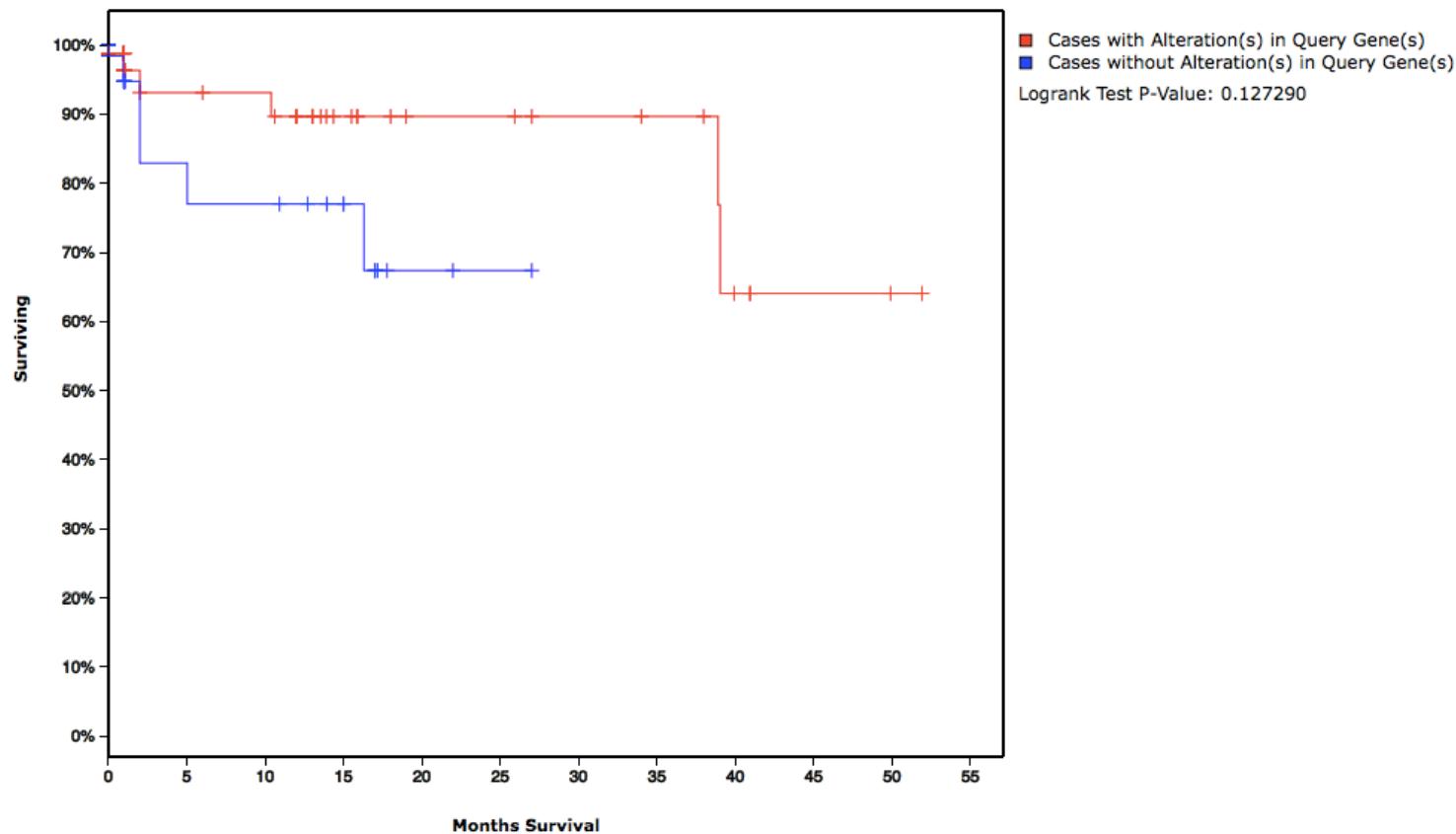
NRAS 10% 

BRAF 3% 

genetic alteration | Amplification | Missense Mutation

Kaplan-Meier Curves

Overall Survival Kaplan-Meier Estimate [SVG](#) [PDF](#)



Thank you

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