Bienvenue au module #11

Cancer et génomique: des big data aux modèles prédictifs

https://github.com/gustaveroussy/ifsbm-bigdata

université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

2024.1

Planning

	Lundi 16 décembre - Salle: 21 B2M
09:00-10:30	Technologies et données NGS en cancérologie. D. GAUTHERET
10:45-11:15	Exercices: analyser les big data cancer sur CBioPortal. D. GAUTHERET
11:15-12:15	TP Galaxy: Cas d'étude Exome-seq (contrôles qualité, alignements des séquences sur le génome de référence). D. GAUTHERET
13:30-17:00	TP Galaxy : Cas d'étude exome-seq (création d'un workflow, variant-calling, visualisation) D GAUTHERET - Y PRADAT

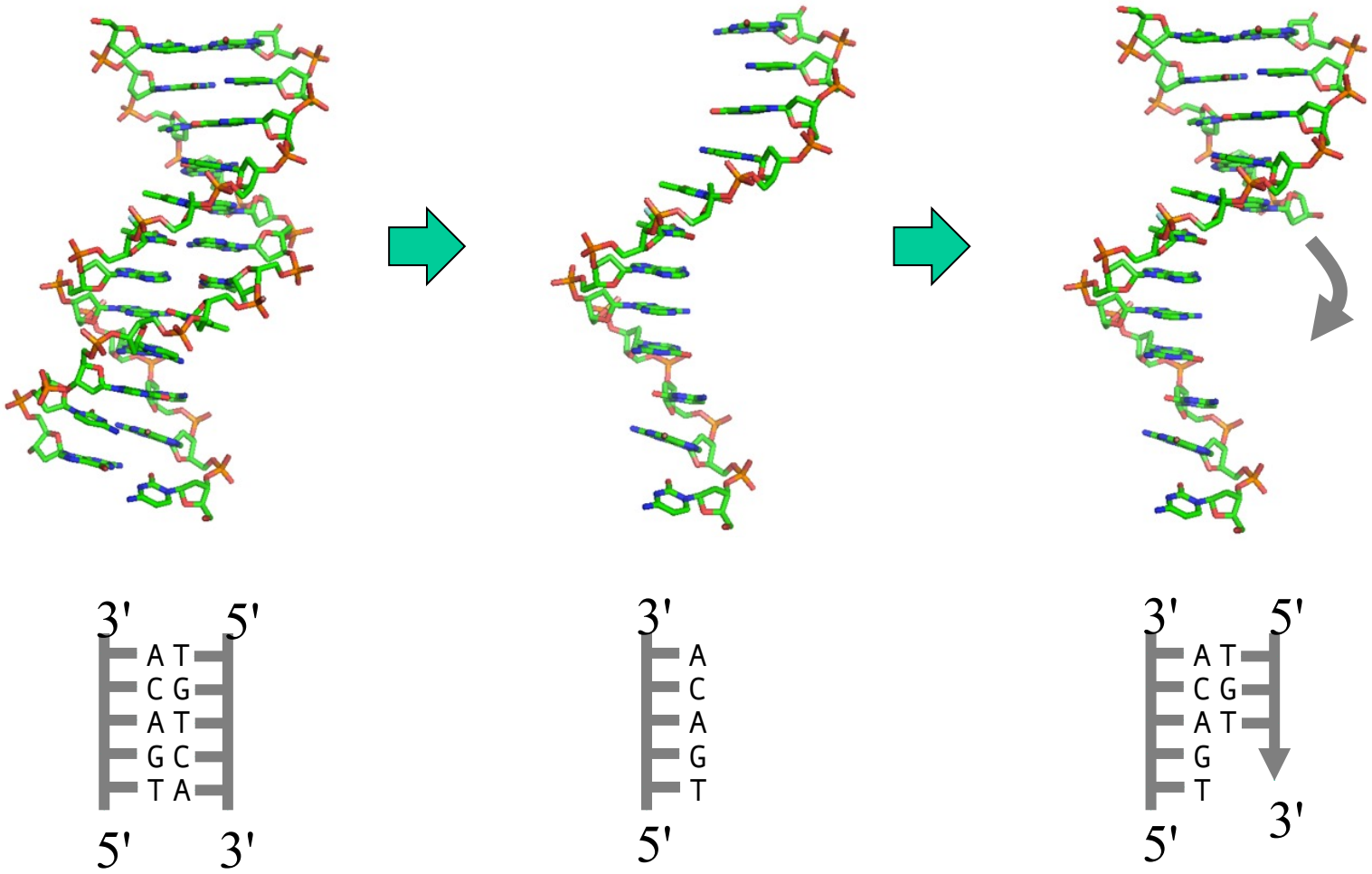
	Mardi 17 décembre - Salle 21 B2M
9:00-10:00	Pourquoi utiliser les methodes d'apprentissage automatique en oncologie personalisée? Julien VIBERT (Gustave Roussy)
10:00-11:00	Methodes de classification supervisée. Yoann PRADAT (Gustave Roussy)
11:30-12:30	Exemple d'un projet de classification médicale: 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).

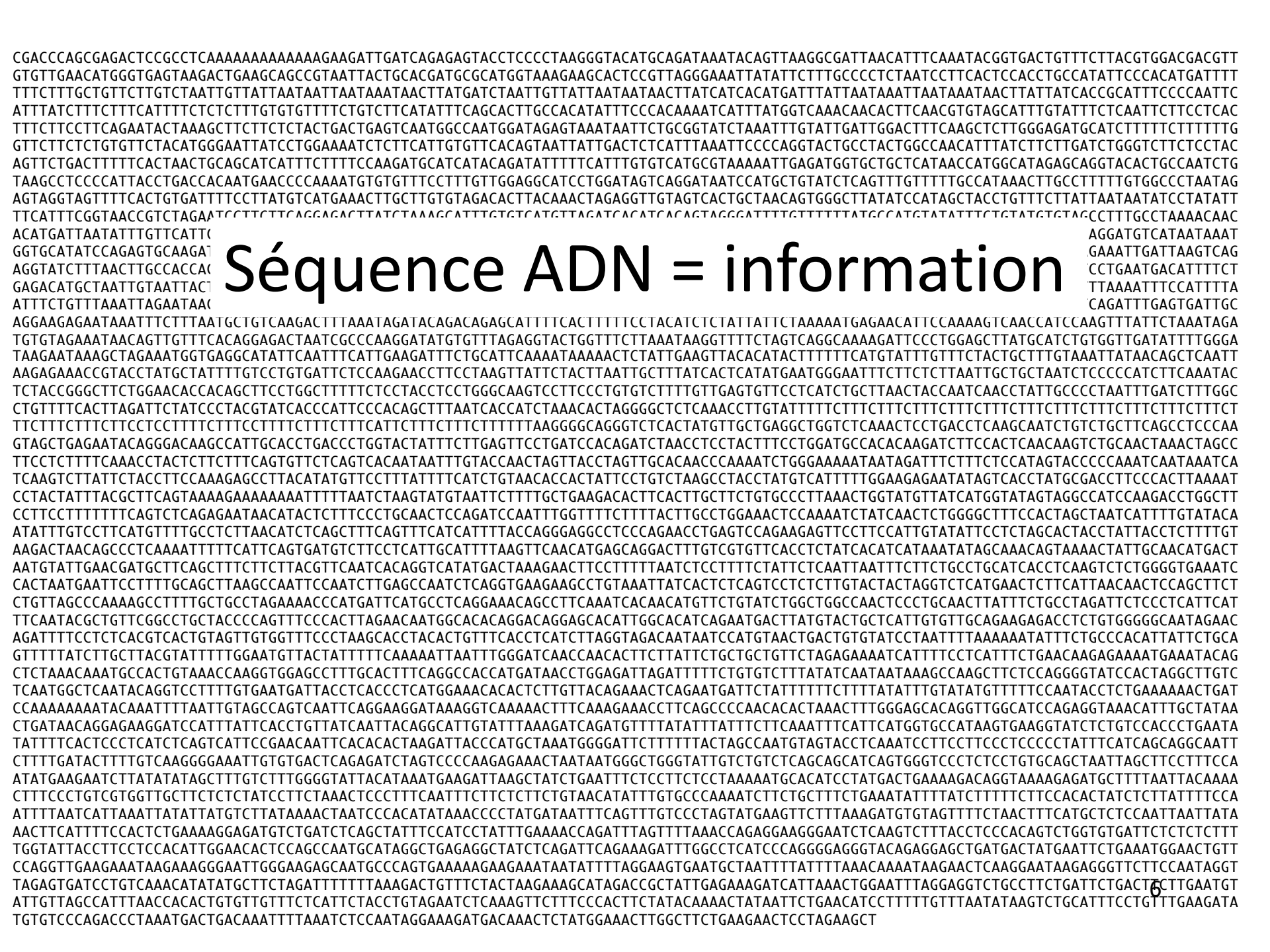
	Mercredi 18 décembre - Salle 21 B2M
9:00-10:00	Méthodes d'analyses génomiques et de modélisation de survie. Yoann PRADAT (Gustave Roussy)
10:00-10:45	Example d'un projet de prédiction de réponse aux immunothérapies: Roger SUN (Gustave Roussy)
11:15-12:00	Exemple d'un projet de prédiction de survie: Elsa 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

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

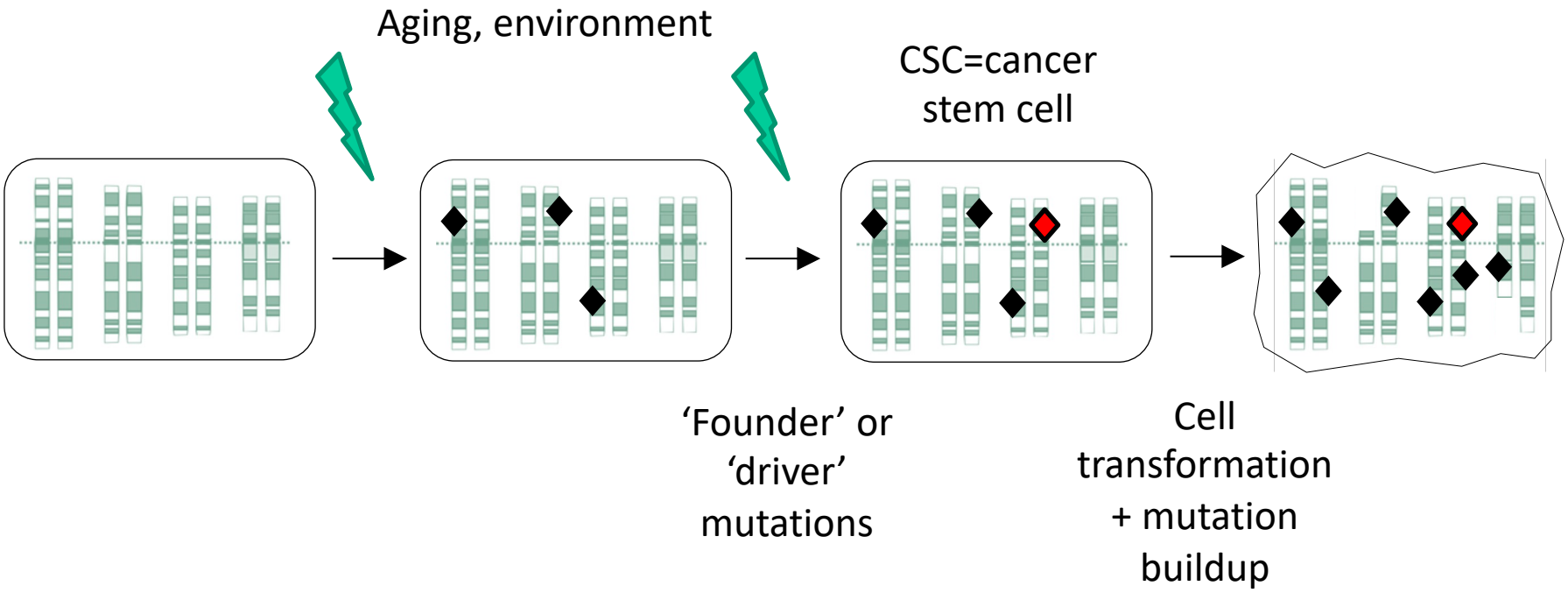
clé	valeur
ACG	Thr
AAG	Lys
GCG	Ala
TAG	stop
ATG	Met
...	...

Encoder une protéine

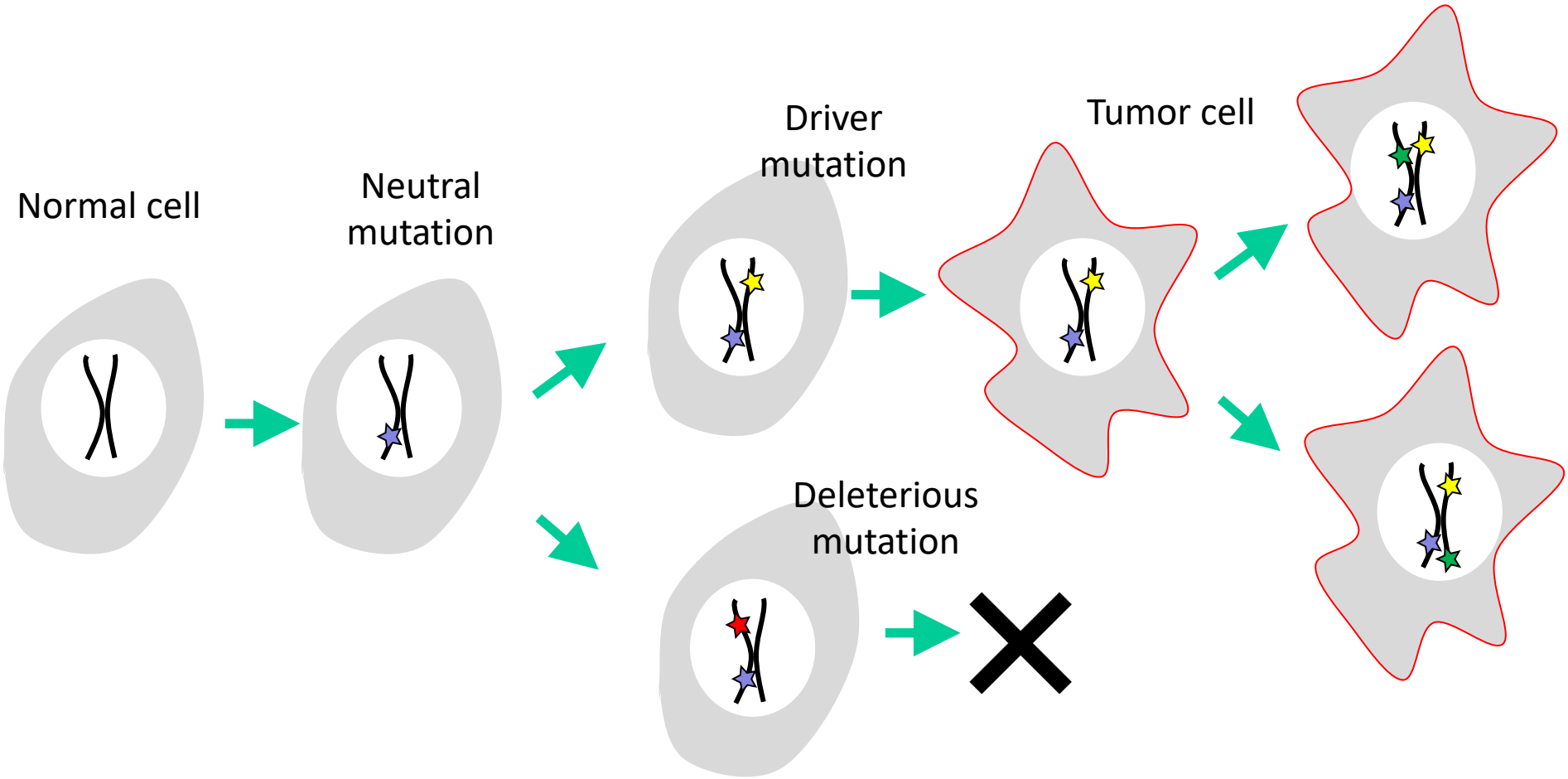
Code génétique

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

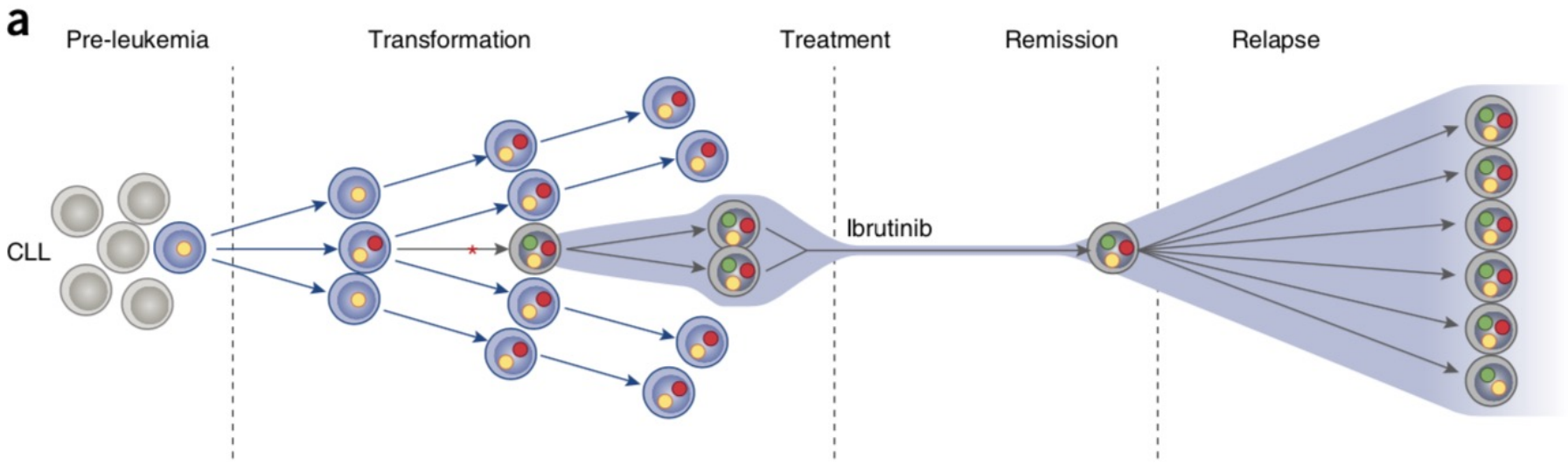
- Le cancer: une maladie du génome



Evolution clonale des cellules cancéreuses

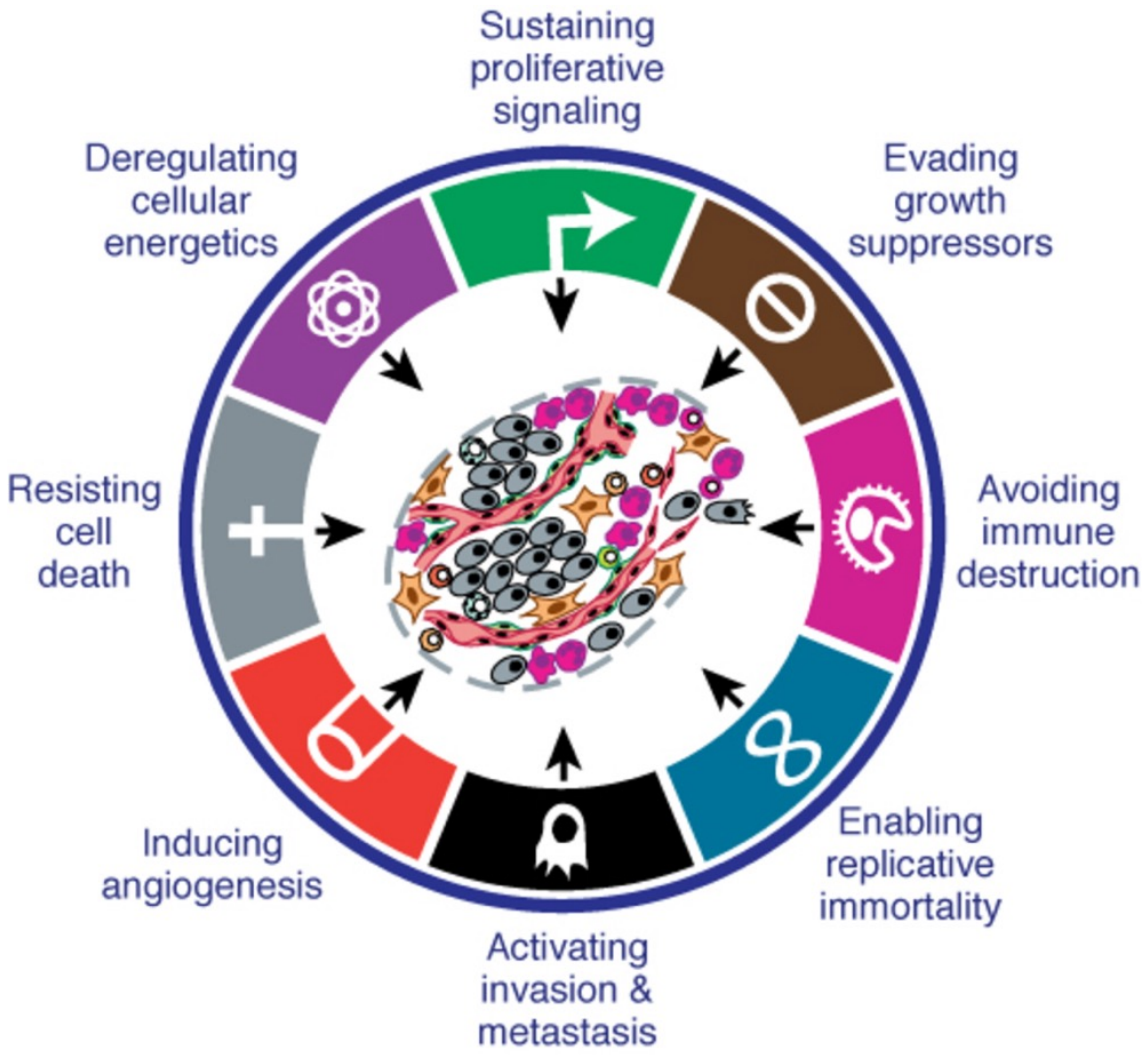


Clonal evolution of cancer



Ferrando & Lopez-Otin, Nature Med. 2017

What is different in a cancer cell?



Hanahan & Weinberg, 2015

Current Chemotherapies

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

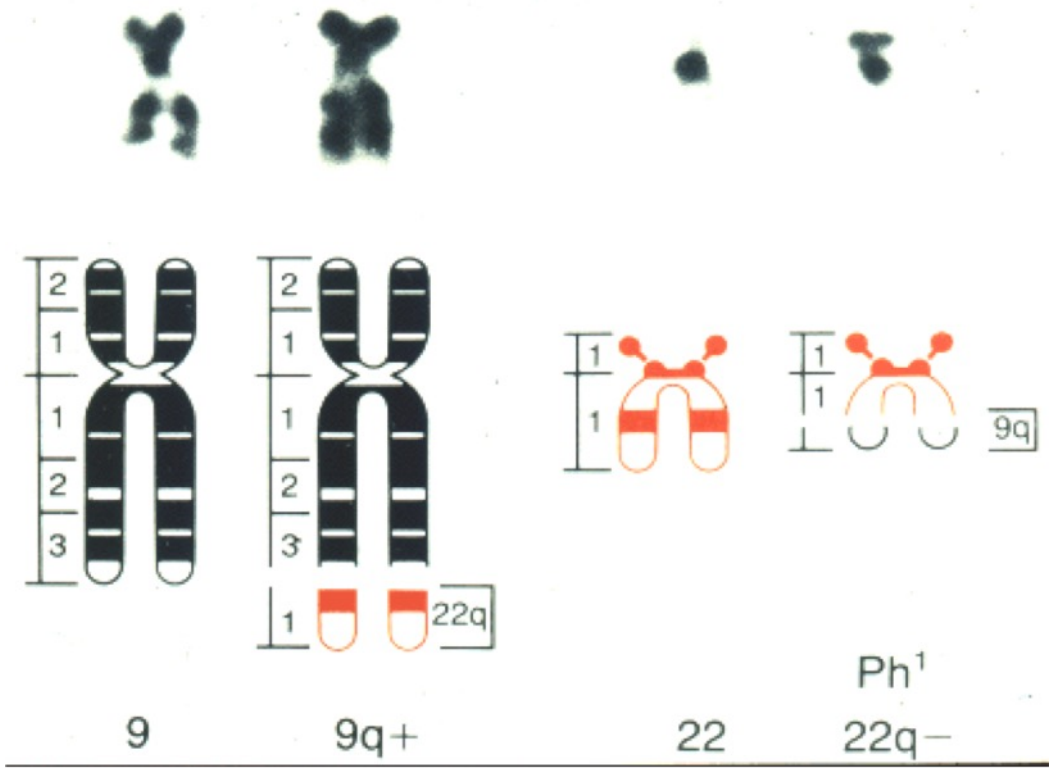
= systemic treatments

Target activities shared by normal cells

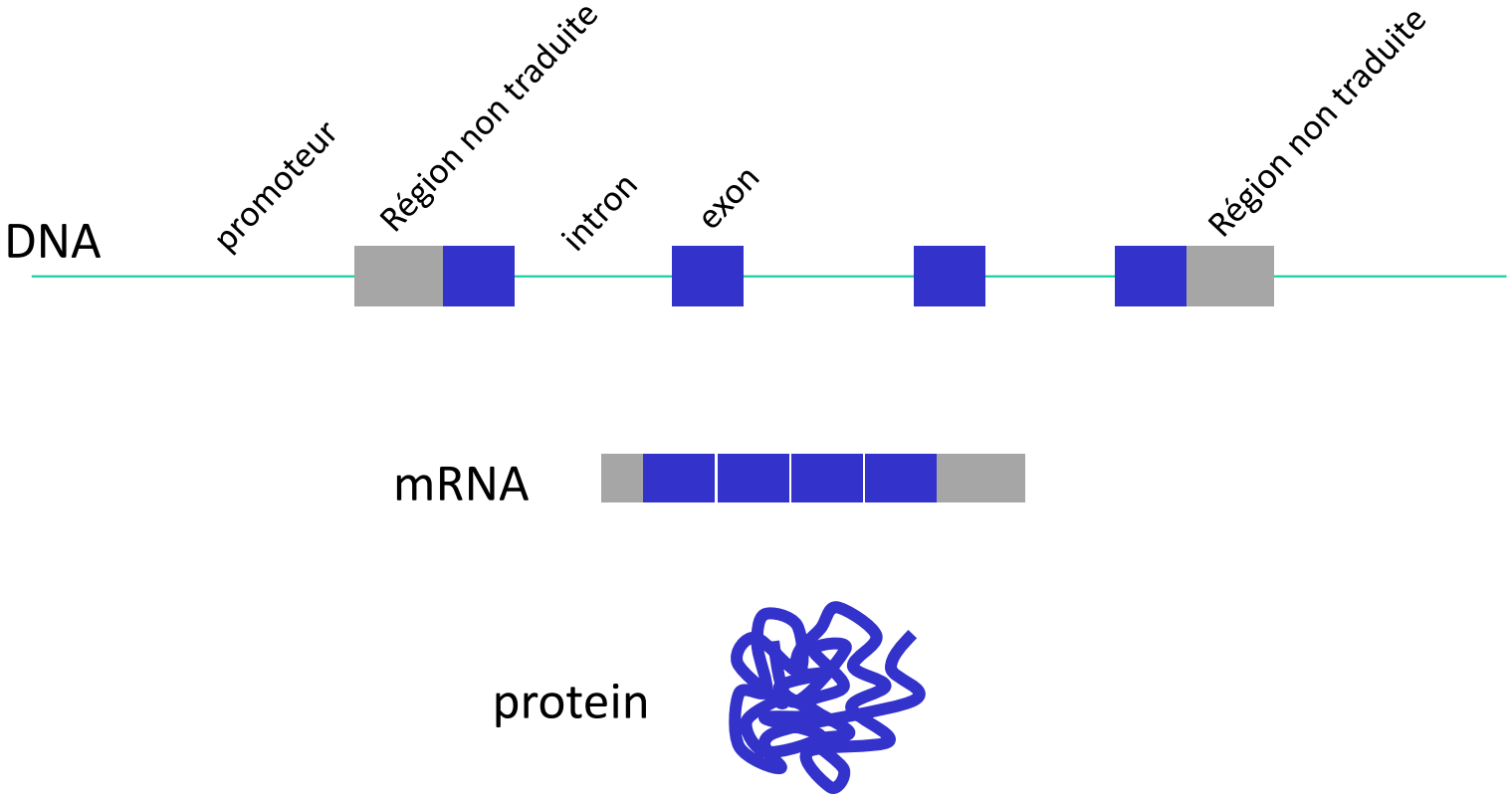
Precision therapies

Le « chromosome de Philadelphie »

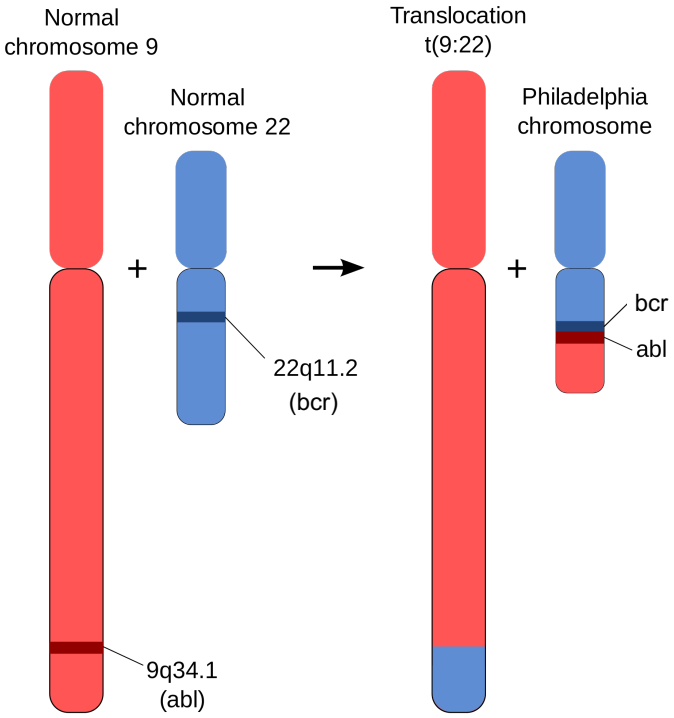
Translocation entre le chromosome 9 et le chromosome 22 dans la leucémie myéloïde chronique (LMC)



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).

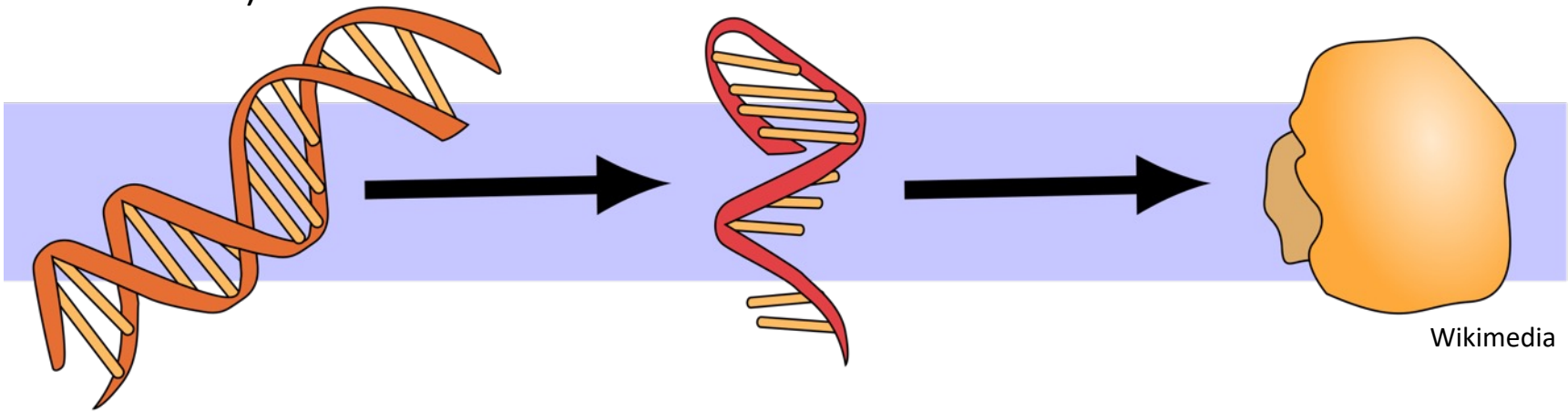


Protéine de fusion
BCR-ABL

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

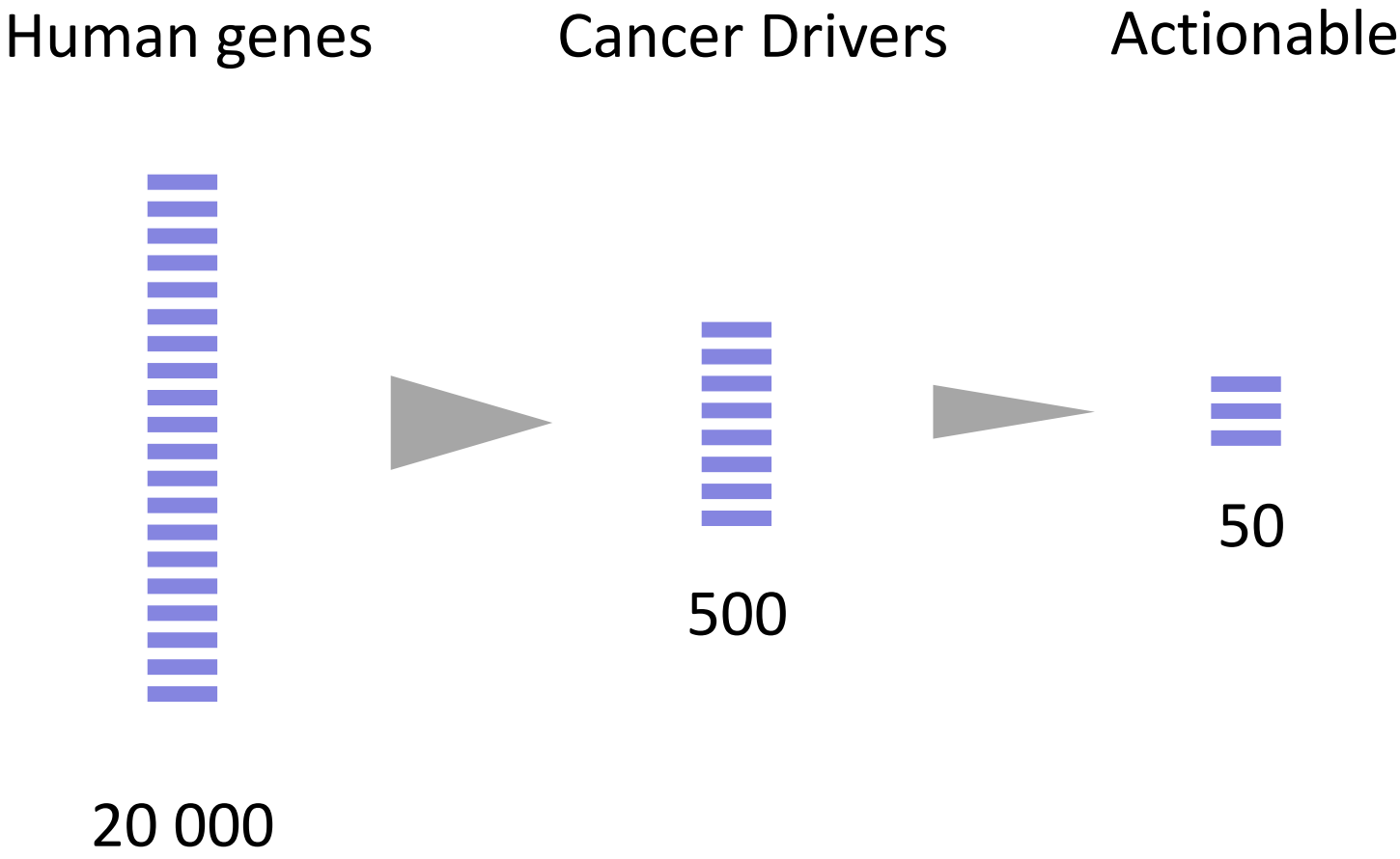
- Change in quantity
- AA modifications

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

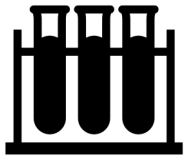
Actionable genes

- 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

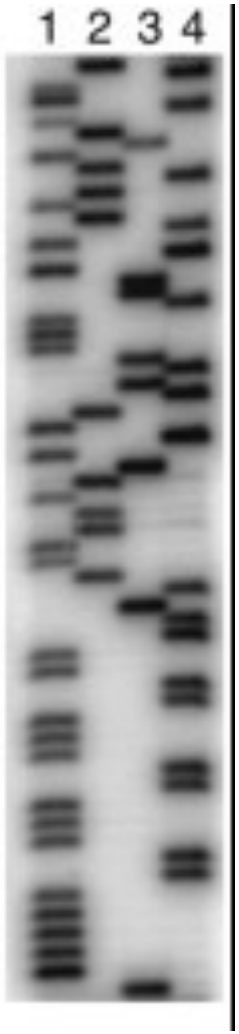
Sequencing technologies

Purified DNA



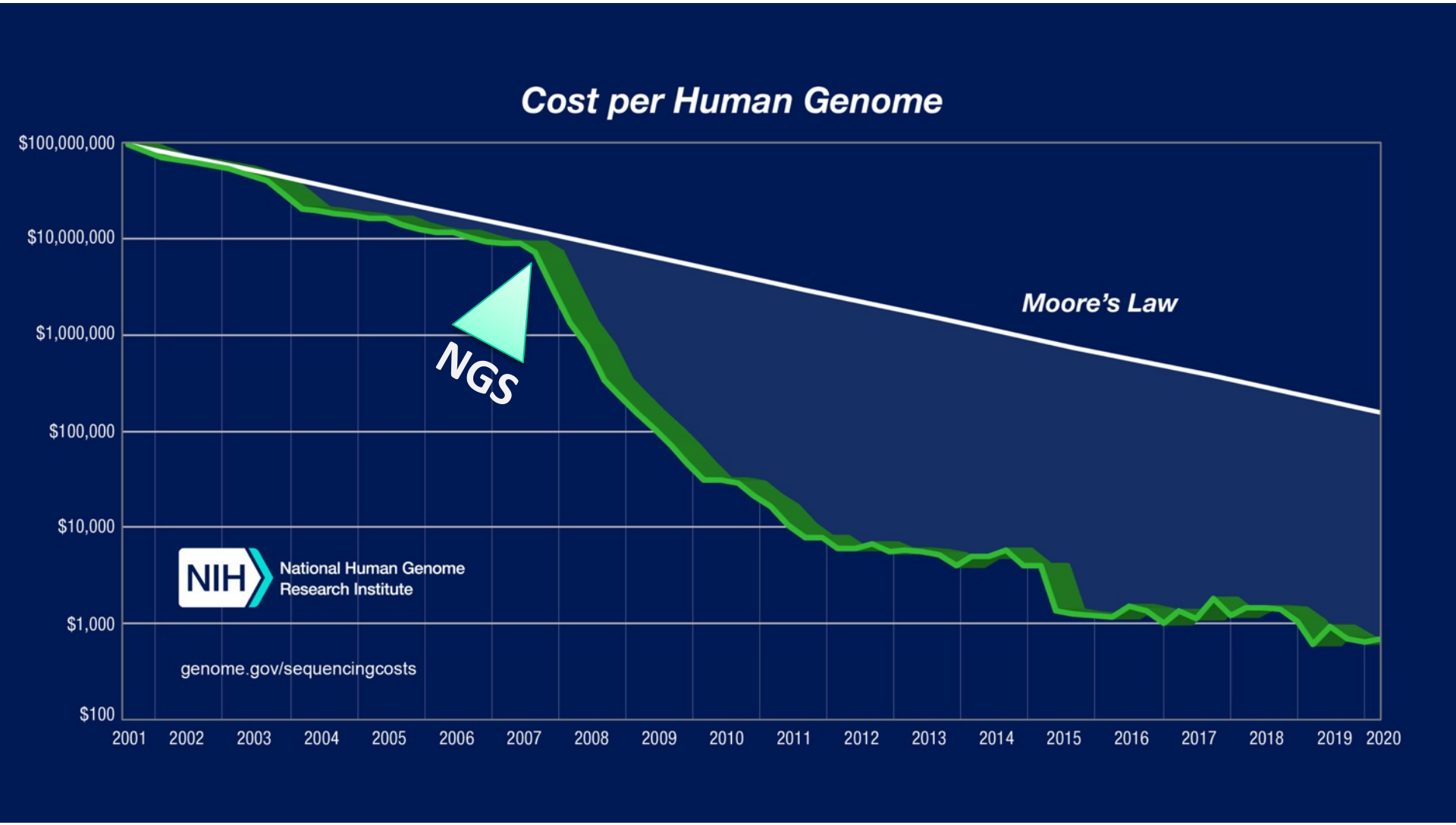
Sequence file

Sanger sequencing



~800 bases / day / operator

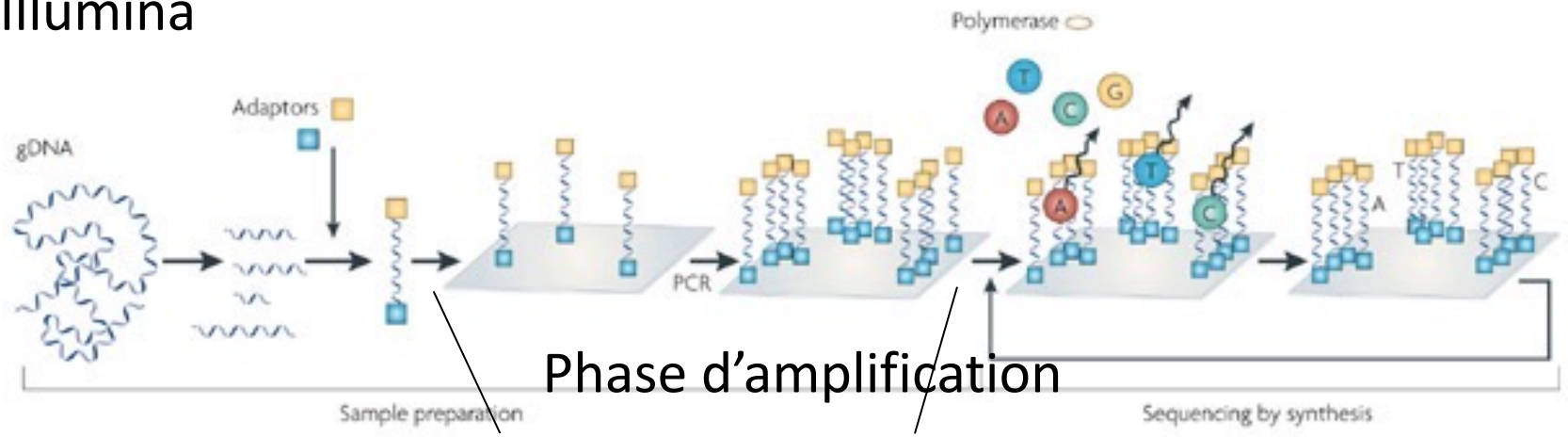
The NGS revolution



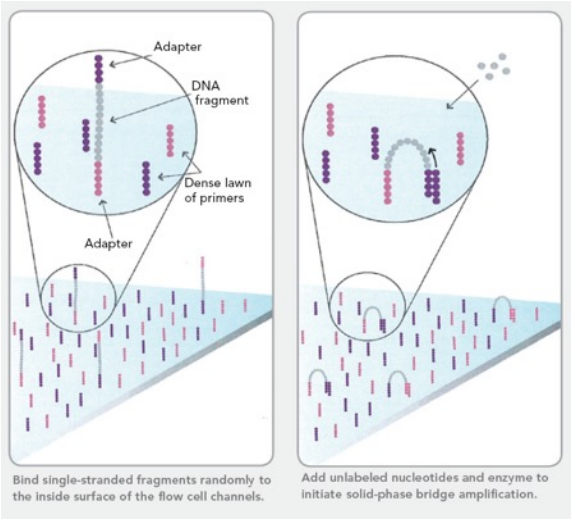
NGS :
Next Generation Sequencing
(2005-)

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

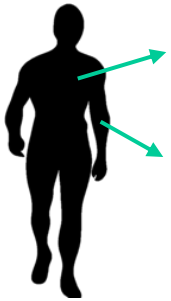
- En cancérologie, 2 grandes applications
  - Génétique constitutionnelle (recherche de prédisposition)
  - Génétique somatique (diagnostic, médecine de précision)

Clinical sequencing

Constitutional mutations (risk factors)

→ Blood sample → Sequence gene panel → Look for germline alteration (BRCA mutation)

Somatic mutations (Precision medicine)



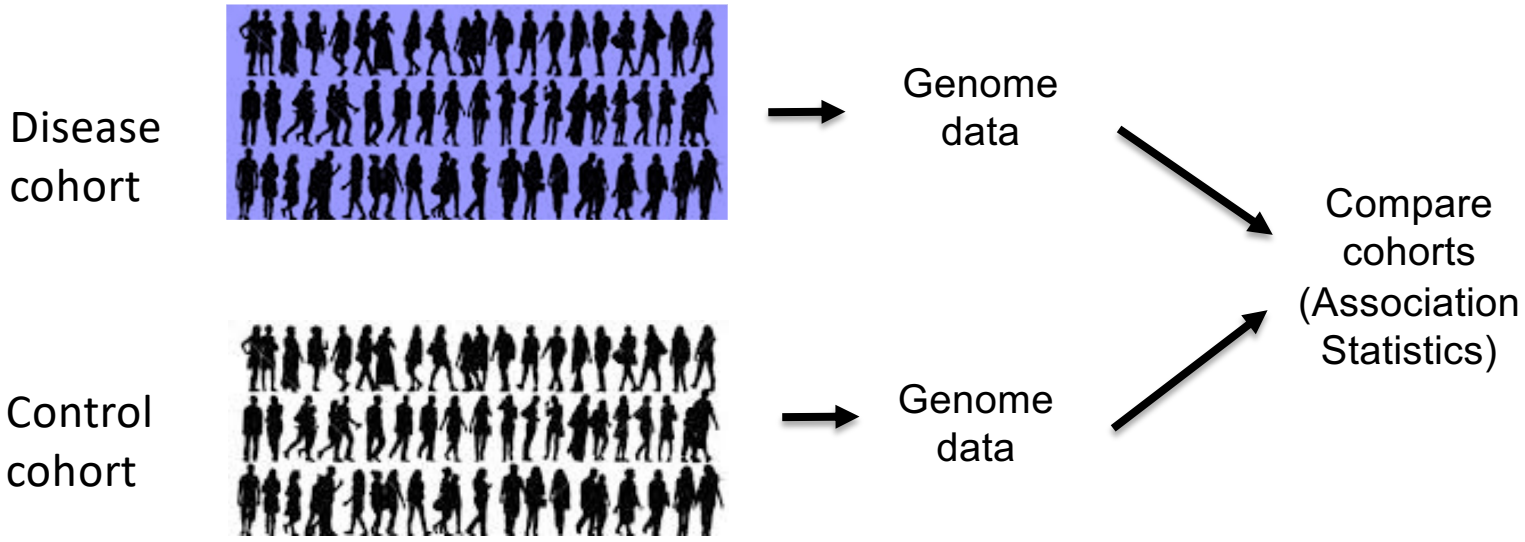
Normal tissue (blood)

Tumor tissue

→ Sequence gene panel → Look for somatic alterations (EGFR mutation)

Research sequencing (constitutional)

UK 100K genomes
AllofUS USA (1M)
France Med Gen 2025



Genome data=
2000's: GWAS=SNP arrays
2010's:WES
2020's: WGS

Research sequencing (cancer/somatic)

TCGA
PCAWG

Cancer cohort

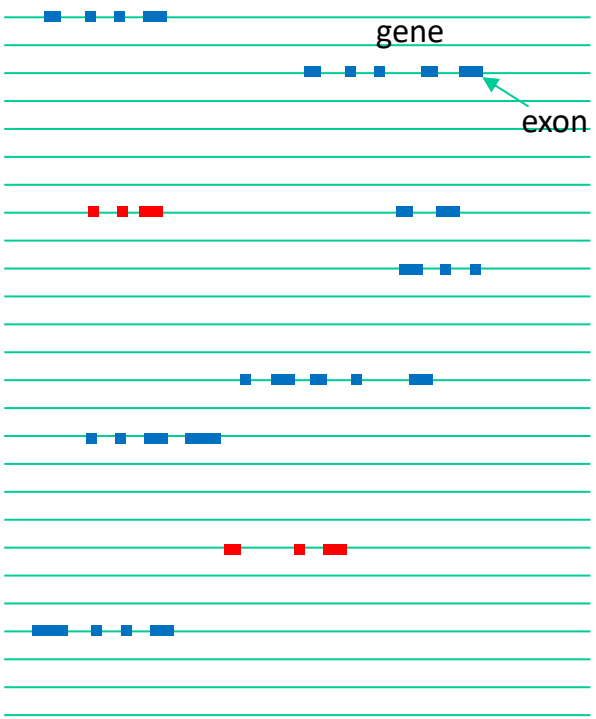
Tumor Sample → Sequence

Normal Sample → Sequence

Sequence → Compare

Genome data=
2010's:WES
2020's: WGS

DNA sequencing types



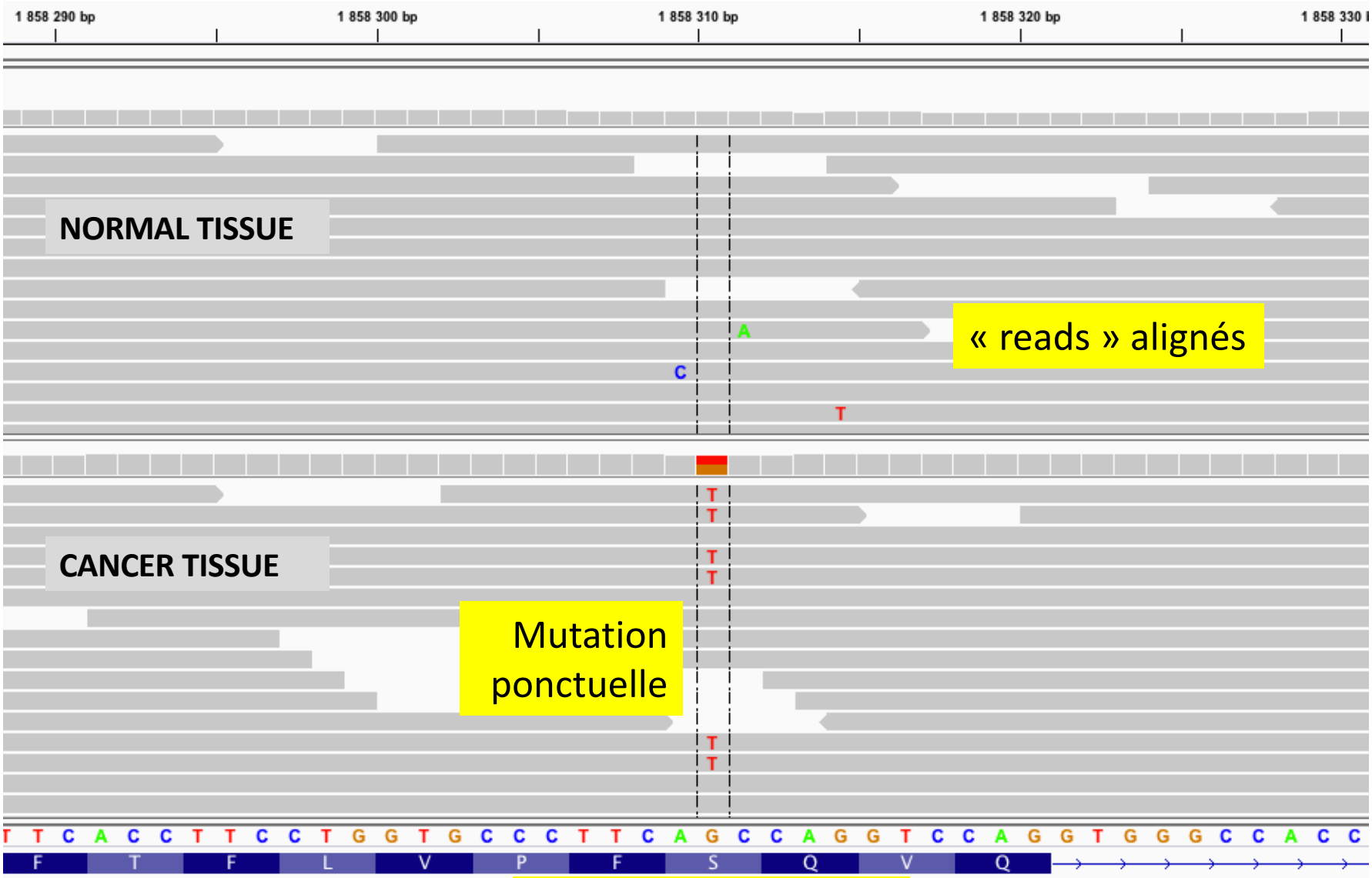
Chromosome sequence

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

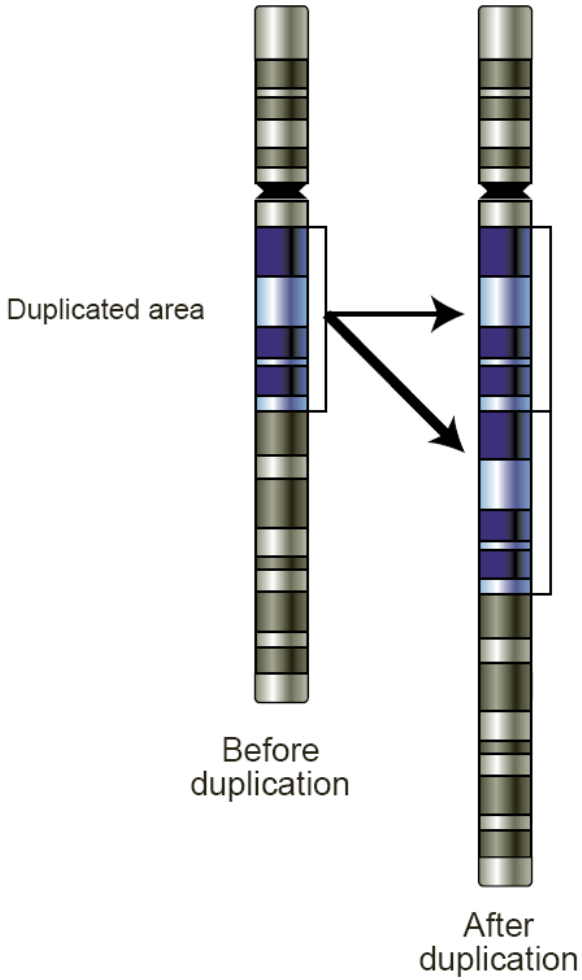
Evénéments recherchés par DNA-seq

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

Mutation ponctuelle



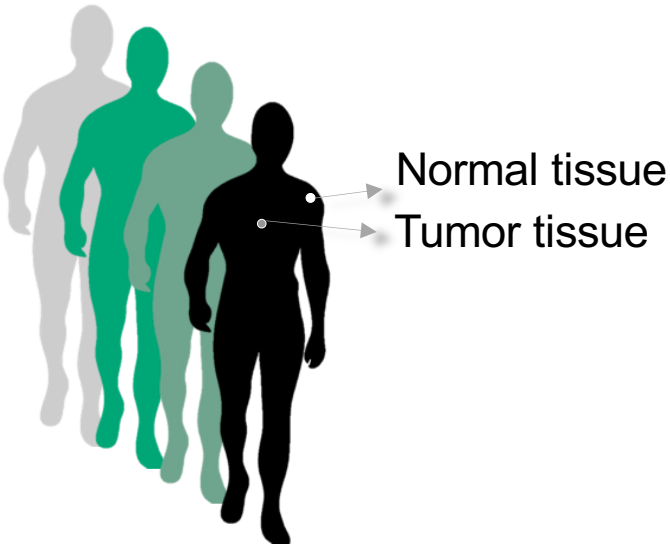
Copy number variations



- 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

Mutations somatiques sur cohortes de patients



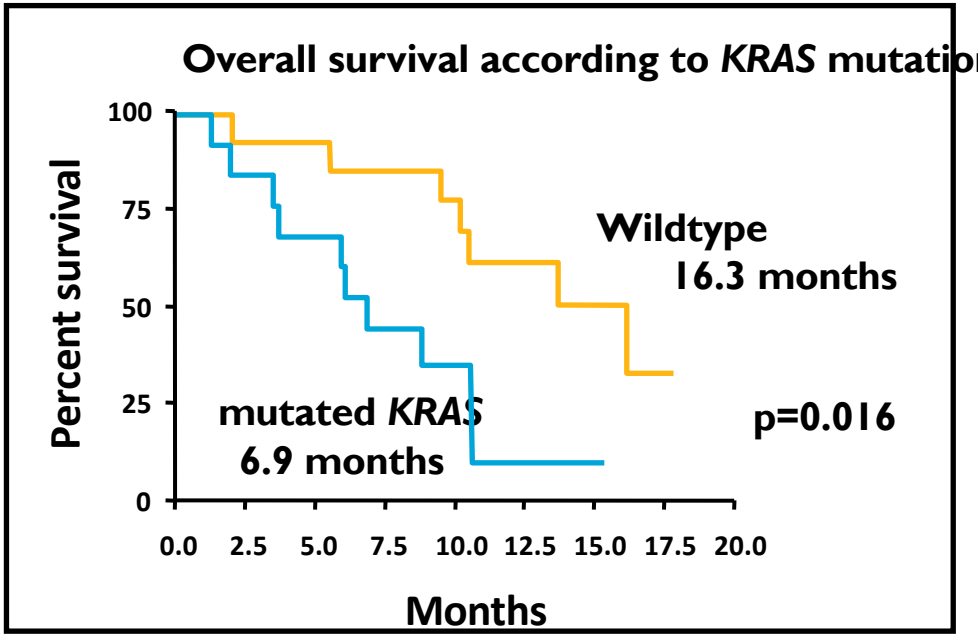
Discovery of outcome-related mutations

KRAS Mutation and Anti-EGFR therapy in advanced colorectal cancer

KRAS Status	Responders*	Non responders*	Total
KRAS mutation (%)	0 (0)	13 (100)	13
Wildtype (%)	11 (65)	6 (35)	17

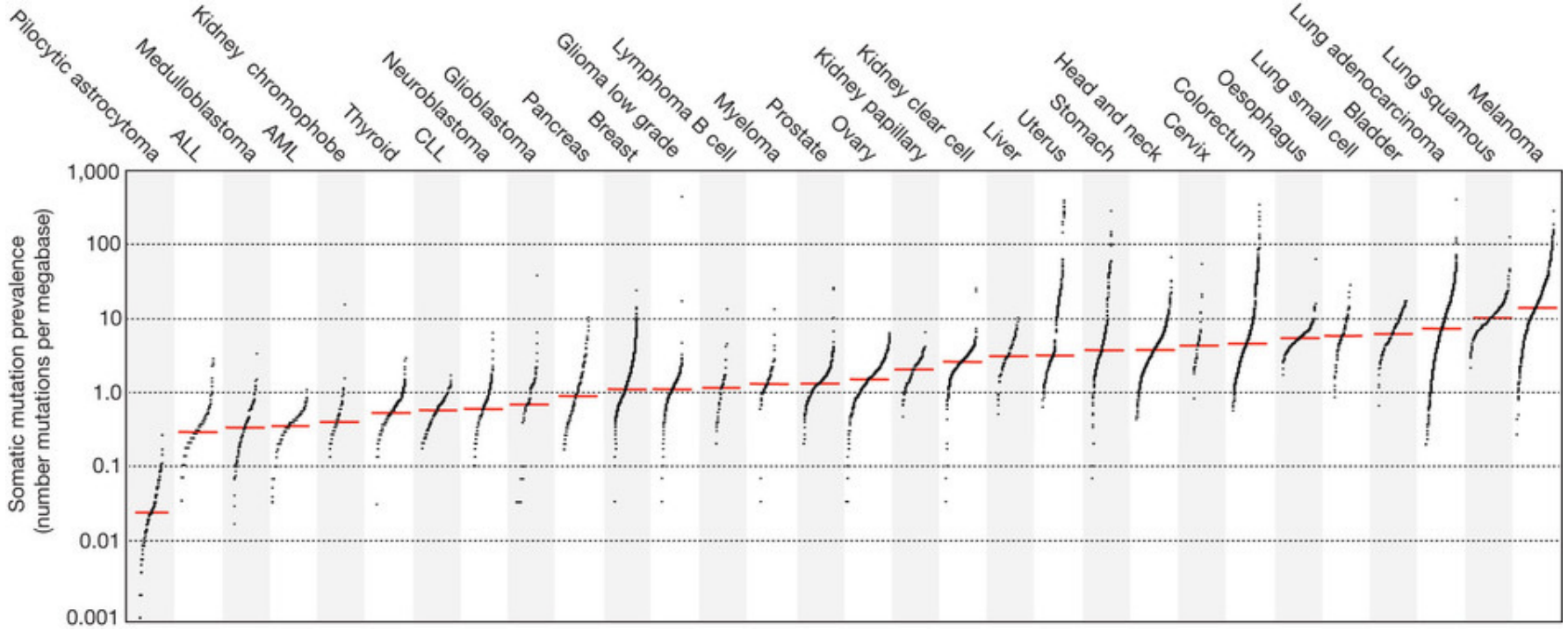
p=0.0003

Courbe de Kaplan-Meier



Lièvre A et al. Cancer Res 2006.

Fréquences des mutations par type de tumeur (10,000 tumeurs)

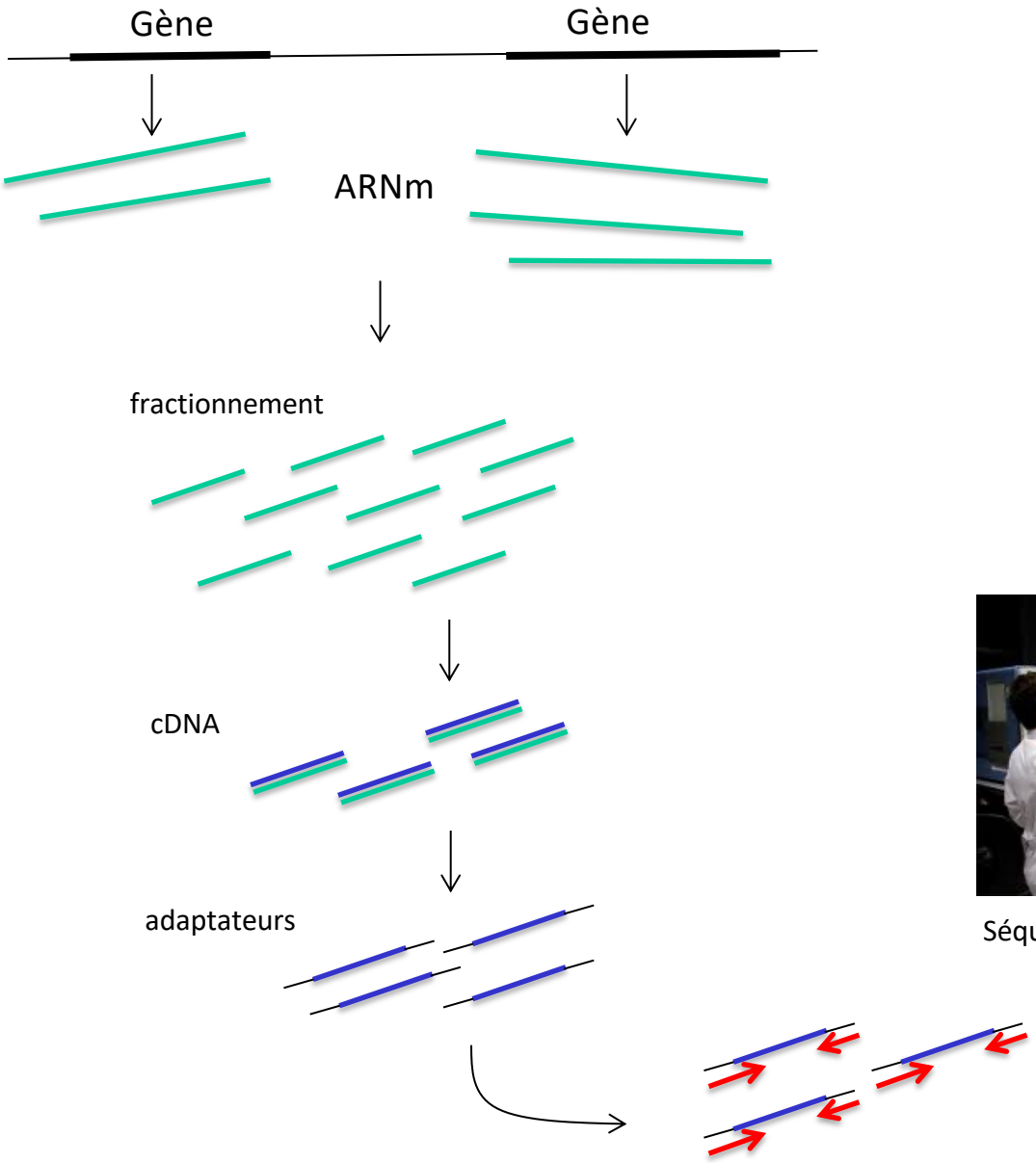


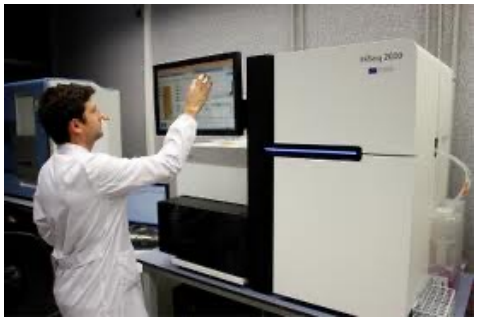
L. Alexandrov…M Stratton, Nature 2013.

RNA-seq

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

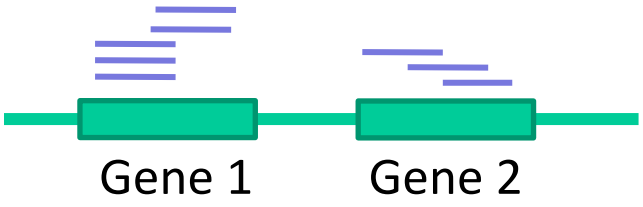
RNA-Seq



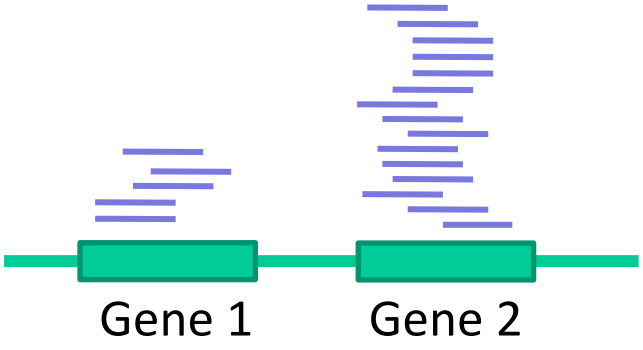


Séquençage

Mesure d'expression par RNA-seq

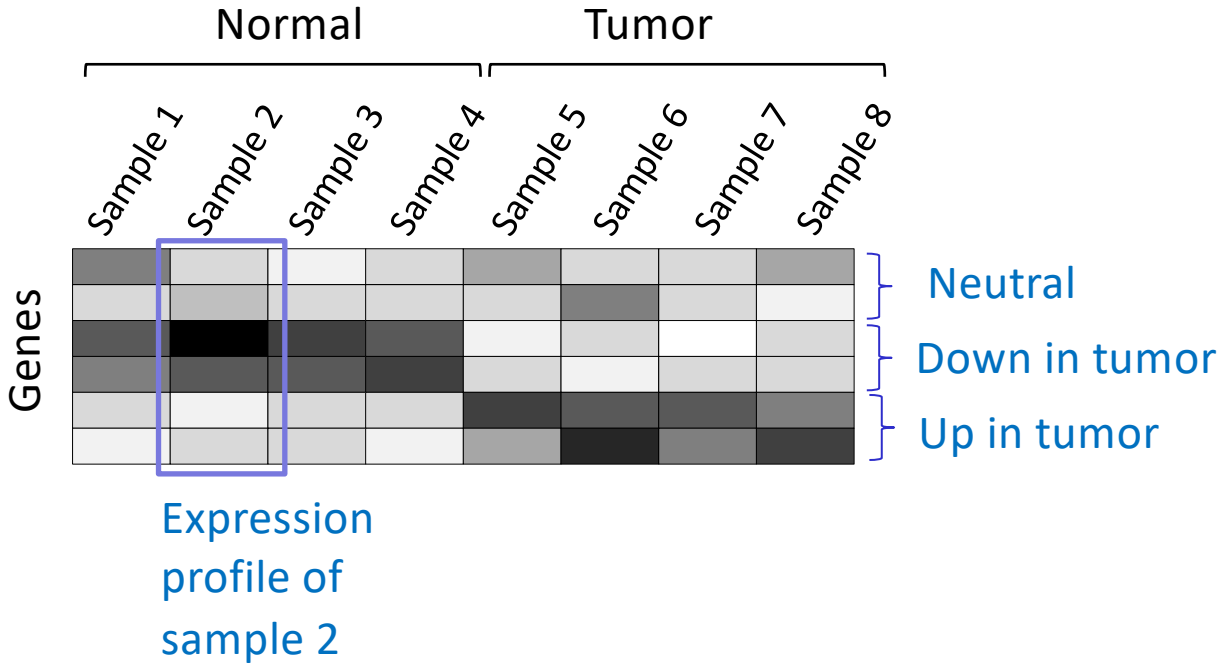


Sample 1

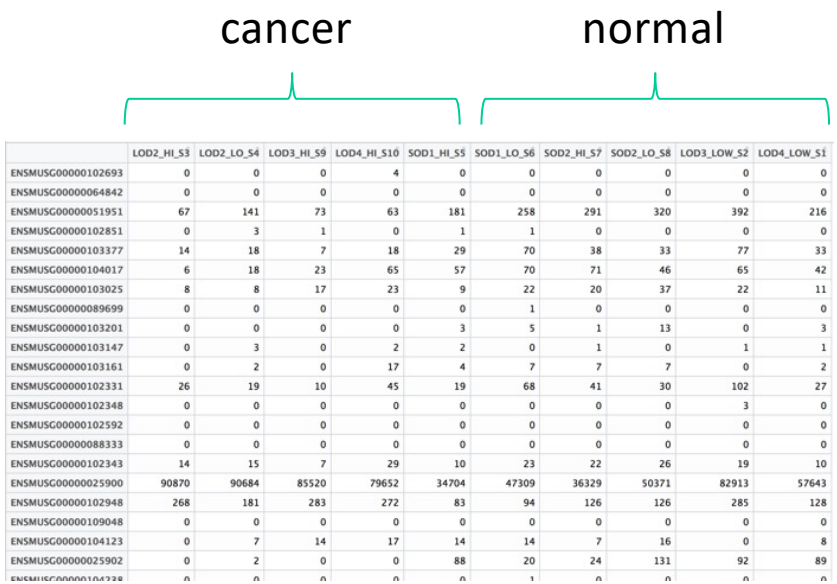


Sample 2

Differential expression analysis



Modèles prédictifs avec données transcriptome



	LOD2_HI_S3	LOD2_LO_S4	LOD3_HI_S9	LOD4_HI_S10	SOD1_HI_S5	SOD1_LO_S6	SOD2_HI_S7	SOD2_LO_S8	LOD3_LOW_S2	LOD4_LOW_S1
ENSMUSG00000102693	0	0	0	4	0	0	0	0	0	0
ENSMUSG00000064842	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000051951	67	141	73	63	181	258	291	320	392	216
ENSMUSG00000102851	0	3	1	0	1	1	0	0	0	0
ENSMUSG00000103377	14	18	7	18	29	70	38	33	77	33
ENSMUSG00000104017	6	18	23	65	57	70	71	46	65	42
ENSMUSG00000103025	8	8	17	23	9	22	20	37	22	11
ENSMUSG00000089699	0	0	0	0	0	1	0	0	0	0
ENSMUSG00000103201	0	0	0	0	3	5	1	13	0	3
ENSMUSG00000103147	0	3	0	2	2	0	1	0	1	1
ENSMUSG00000103161	0	2	0	17	4	7	7	7	0	2
ENSMUSG00000102331	26	19	10	45	19	68	41	30	102	27
ENSMUSG00000102348	0	0	0	0	0	0	0	0	3	0
ENSMUSG00000102592	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000088333	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000102343	14	15	7	29	10	23	22	26	19	10
ENSMUSG00000025900	90870	90684	85520	79652	34704	47309	36329	50371	82913	57643
ENSMUSG00000102948	268	181	283	272	83	94	126	126	285	128
ENSMUSG00000109048	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000104123	0	7	14	17	14	14	7	16	0	8
ENSMUSG00000025902	0	2	0	0	88	20	24	131	92	89

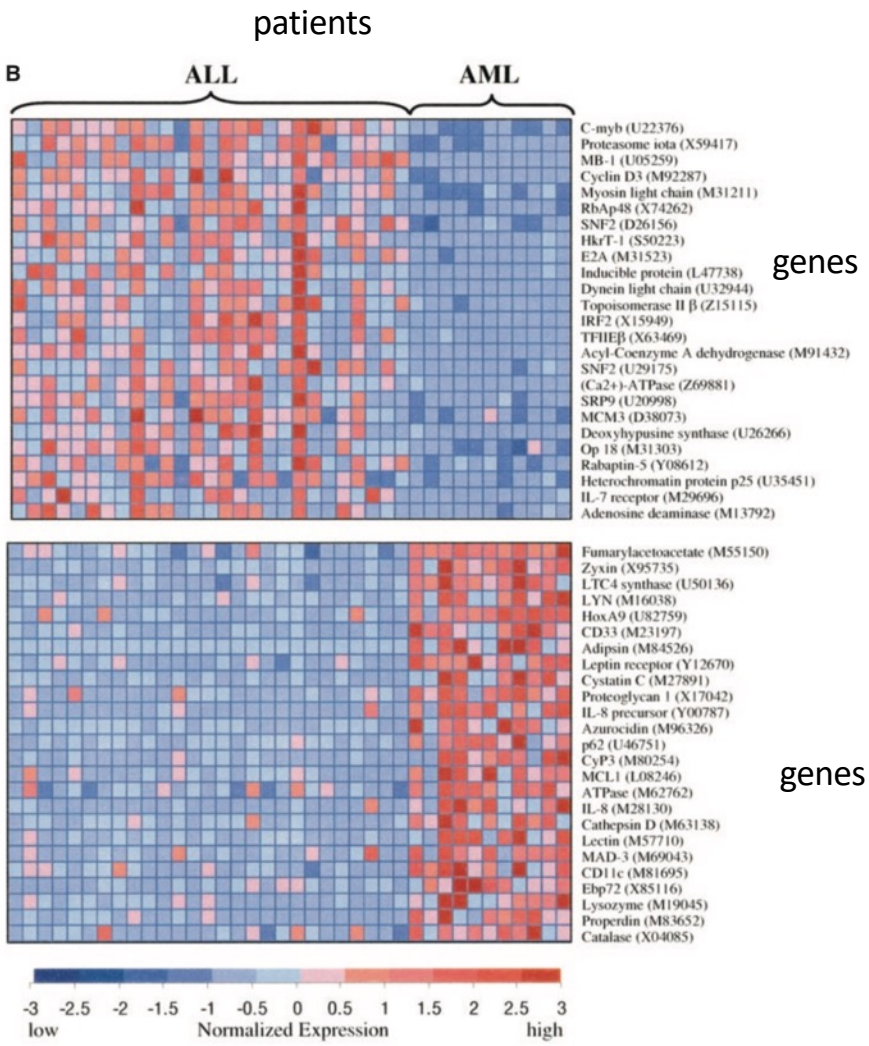
Count matrix

Predict binary status:
- Logistic regression
- Random Forests

Predict survival
- Cox regression

Unsupervised clustering
- Hierarchical
- K-Means
- SOMs

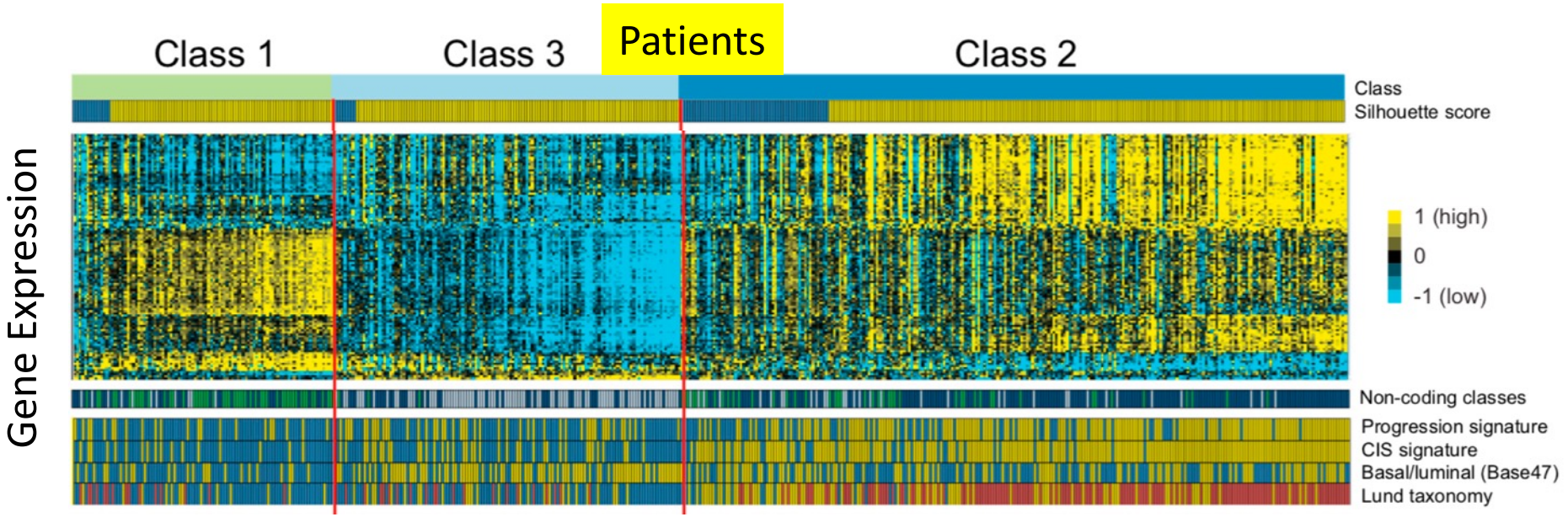
Subtype discovery: *circa* 1999



Todd Golub 1999: microarray-based signature of Leukemia

Subtype discovery (unsupervised classification)

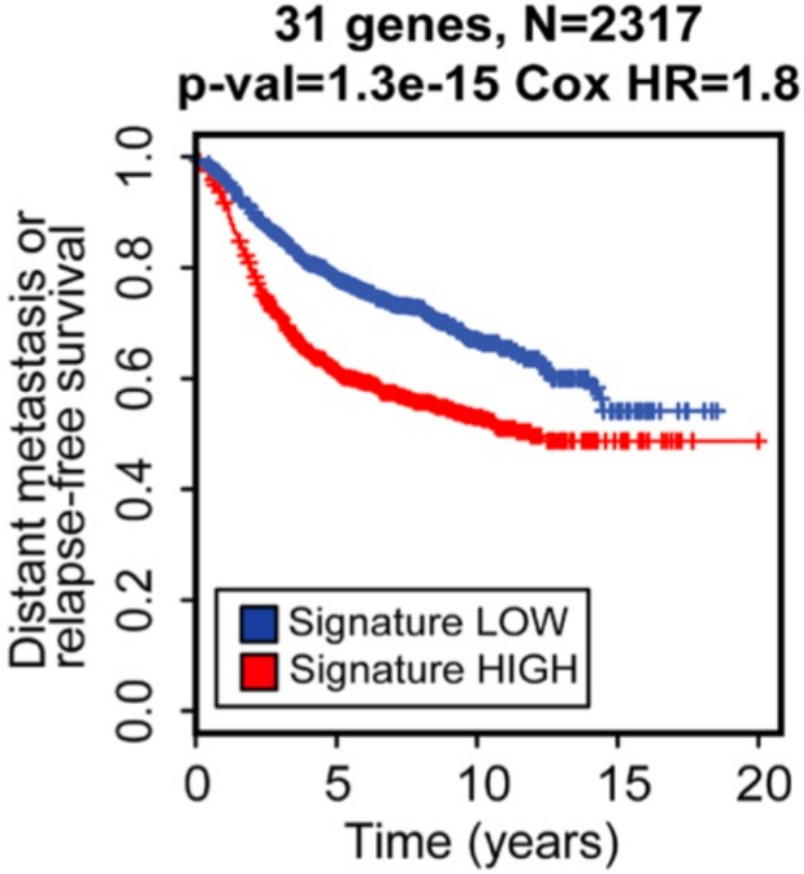
Gene expression profiling in 460 Urothelial carcinoma:
A 117-gene signature.



Hedegaard et al., Cancer Cell, 2016

Modèles prédictifs de survie

Courbe KM: modèle prédictif multi features (expression) par régression de Cox



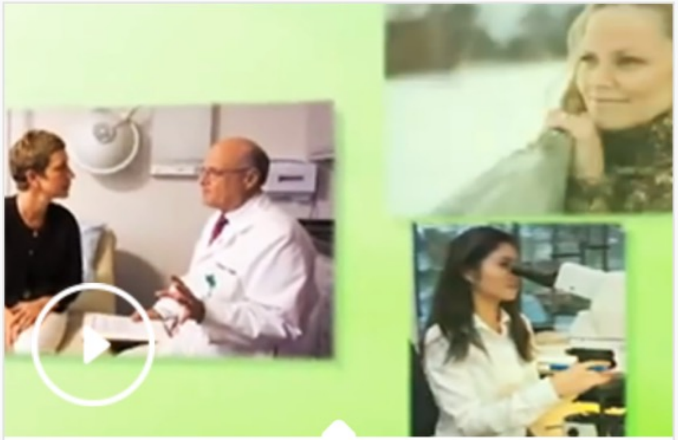
Applications cliniques: prédire la réponse au traitement

Diagnostic

- Tumor vs normal
- Tumor subtyping

Precision medicine

- Response to treatment
- Relapse prediction



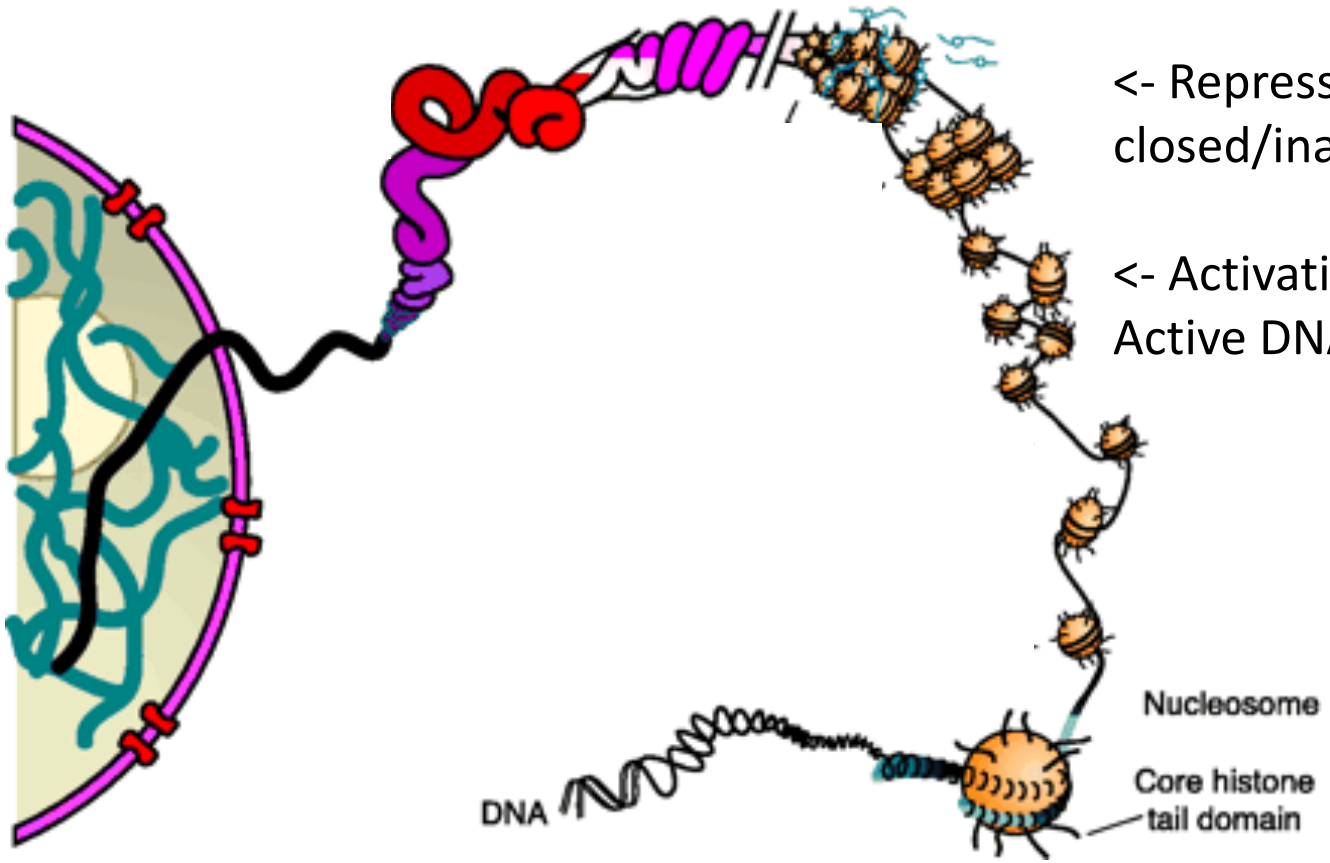
Genomic Health et le test Oncotype DX

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

Epigenetic data

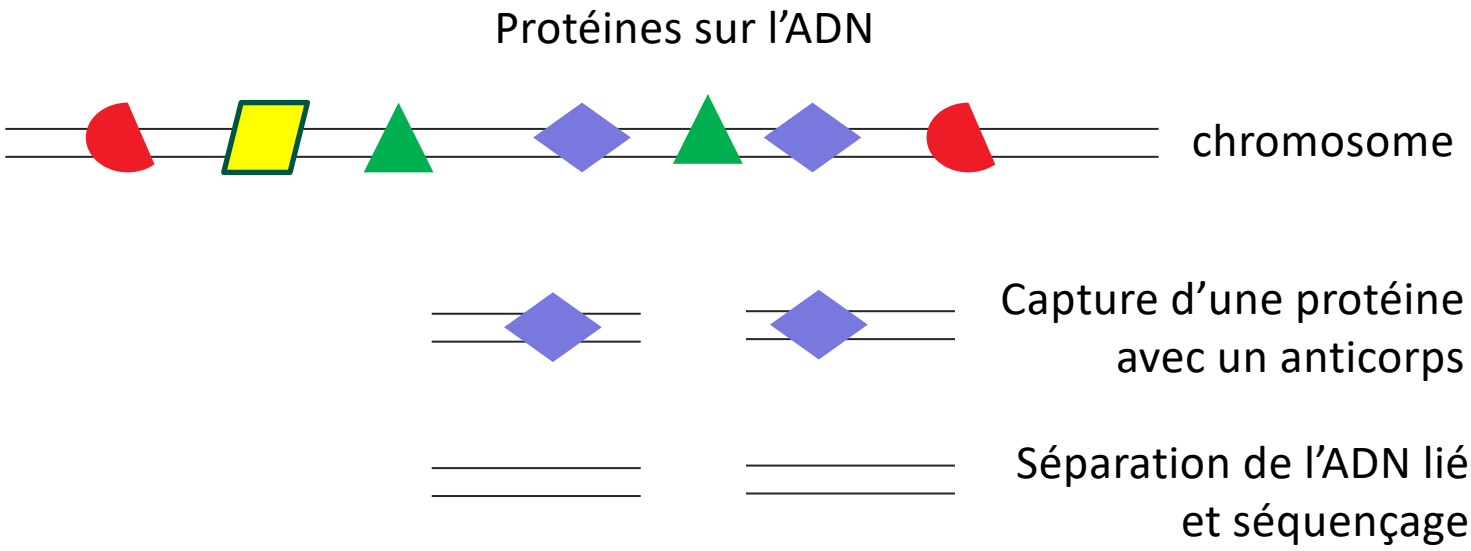
ChIP-seq in oncology

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



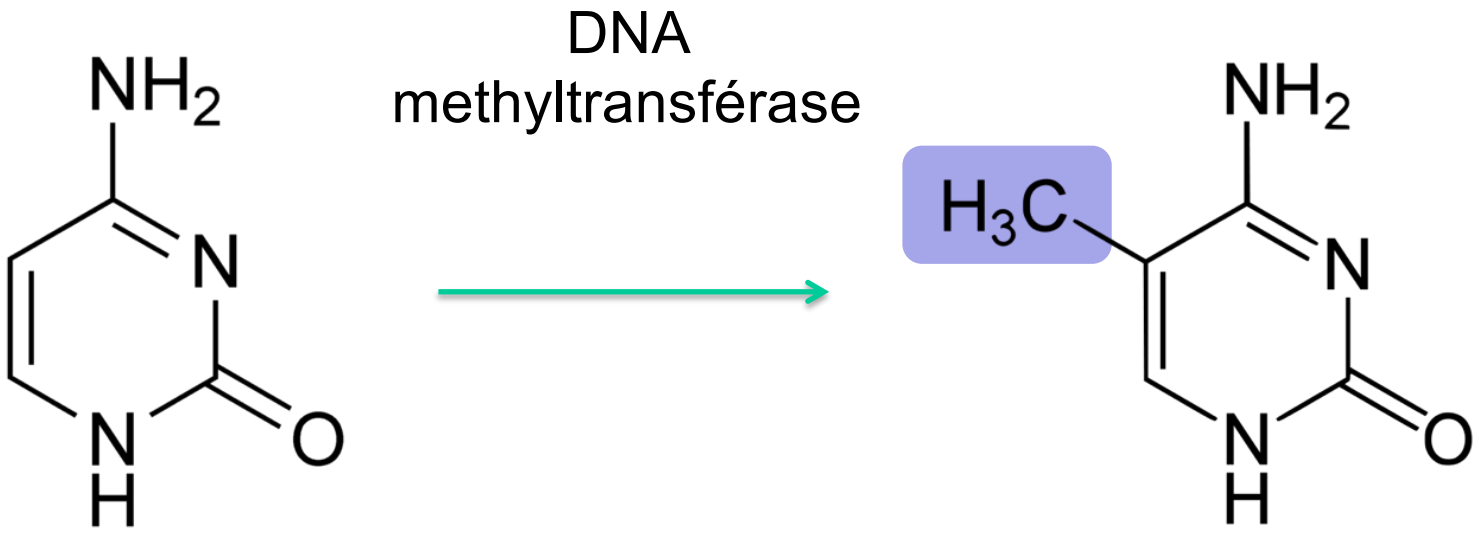
ChIP-Seq

- Permet d'identifier les sites de liaison de protéines sur l'ADN génomique



Methyl-seq (bisulfite-seq)

La méthylation de l'ADN

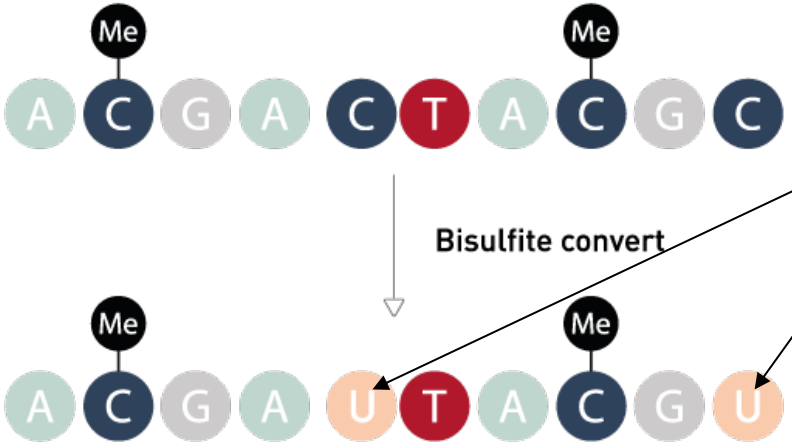


Cytosine

Variations de la méthylation

- L'exposition aux carcinogènes, le régime alimentaire, le vieillissement modifient la méthylation
- Les cancers modifient profondément la méthylation

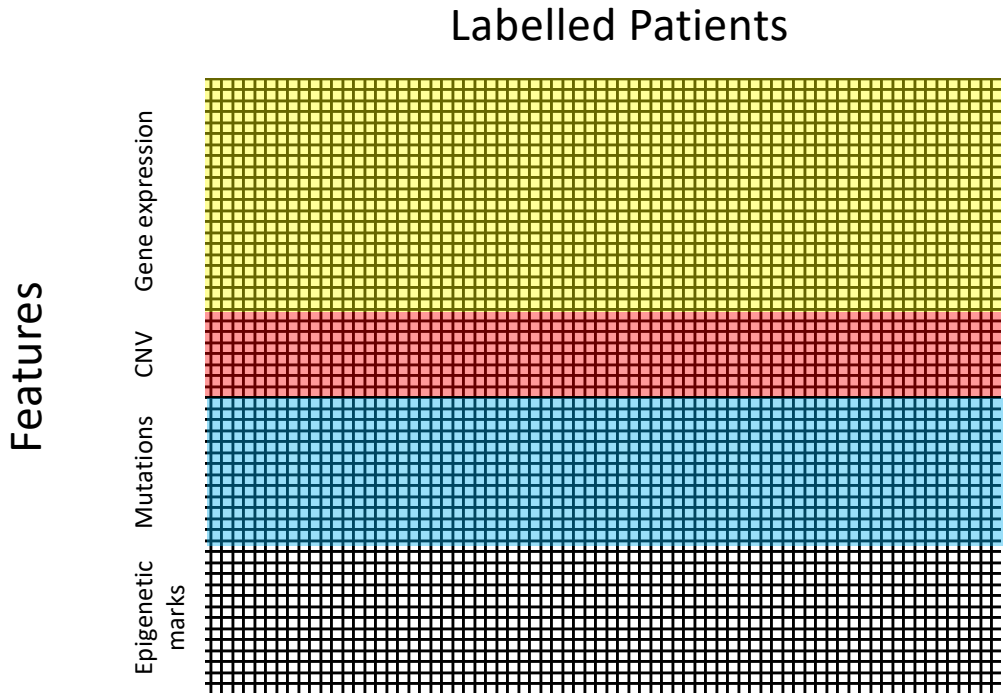
Conversion par bisulfite



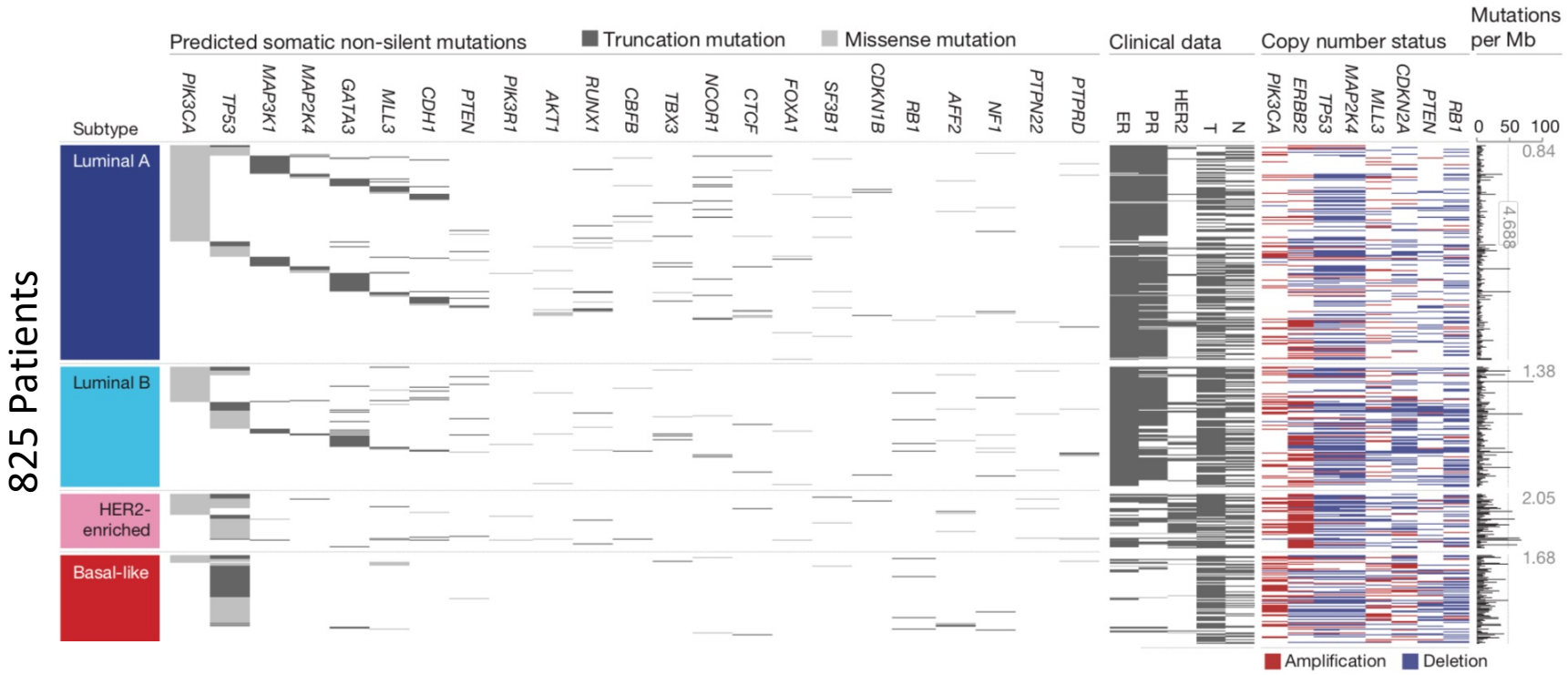
Les C non méthylés sont convertis en U

+ séquençage et alignement des reads:
Les positions avec un T dans le bisulfite et un T dans le témoin sont méthylées

Multi-omics


Labelled Patients
Features
Gene expression
CNV
Mutations
Epigenetic marks

Multi-omics of
Breast cancer cohort: 825 patients



Stratified by cancer type

TCGA consortium, Nature, 2012

Cancer Big Data Ressources

The Cancer Genome Atlas

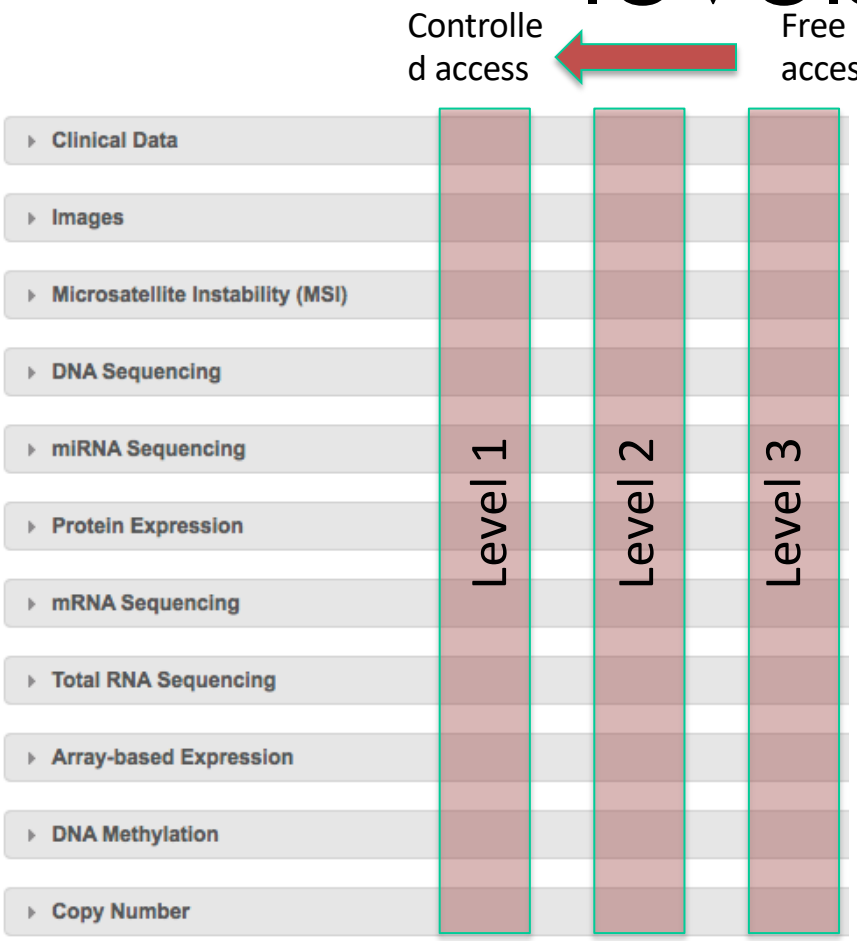
Understanding genomics to improve cancer care

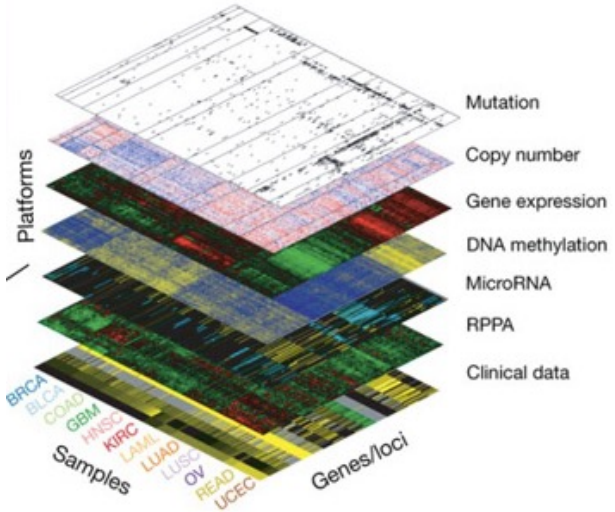
NCI, NHGRI, USA

TCGA/PCAWG

- launched in 2006
- 33 tumor types
- 11,000 patients
- whole-genome sequencing (WGS) for 2,800 tumors

TCGA data types and levels





Cancer Genomics Portals

- cBioPortal
  - For analysis
- COSMIC
  - Cancer gene census
  - Cancer mutation census
- OncoKB
  - For clinicians -> which drug for which mutations?

Memorial Sloan-Kettering
Cancer Center, USA

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

Exercice pratique: cBioPortal

Analyse de mutations:
Les mutations oncogéniques du cancer du poumon

- Sur cBioPortal, choisir le dataset:
  - Metastatic Non-Small Cell Lung Cancer MSK (2621 samples)
- « Query by gene » + enter genes: KRAS, EGFR, TP53
- Analyses à réaliser
  - Quel est le % de patients mutés sur chacun des 3 gènes? (oncoprint)
  - Les 3 mutations sont elles mutuellement exclusives? (mutual exclusivity)
  - Comment les mutations sont-elles distribuées le long du gène? (mutations)
  - Comment la survie des patients diffère-telle selon les mutations? (comparison+survival)

Analyse d'expression:
les gènes de cycle cellulaire dans le cancer du sein

- Choisir le dataset:
  - Breast Invasive Carcinoma (TCGA, PanCancer Atlas) (N=1084)
- Query by gene
- Select genomic profile: mRNA expression (z-score)
- Enter genes: cell cycle control + Query
- Tracks + heatmaps + add genes to heatmap
- Tracks + clinical + subtype + sort subtypes (a_Z)
- Sur le heatmap: peut-on expliquer l'aggressivité supérieure des cancer de type basal, puis luminal B, puis luminal A ?

Programmatic Interfaces to cBioPortal

- Webservice (via URL)
- R library
  - CGDS package (CRAN)
- Matlab Library
  - CGDS toolbox @ MatLab Central

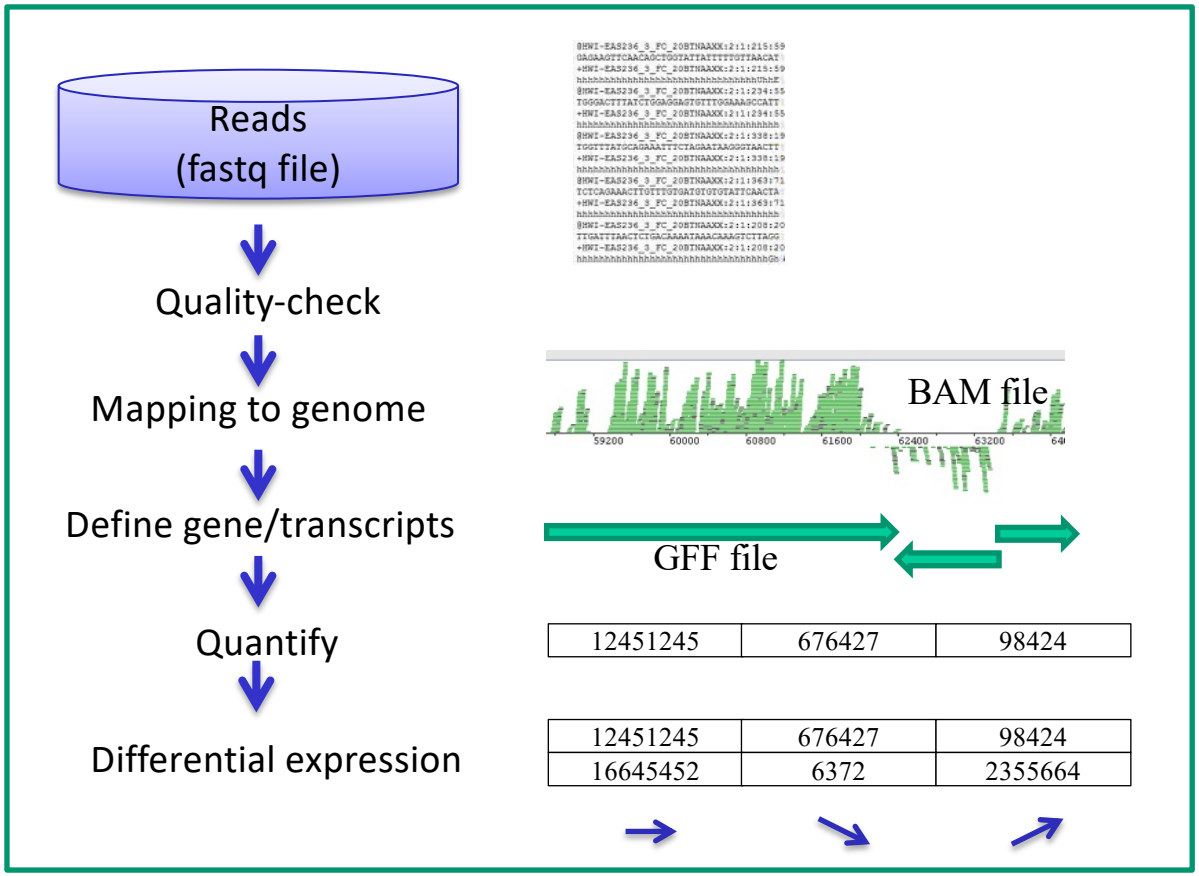
Construire ses propres protocoles d'analyse

« Pipelines » & « workflows »

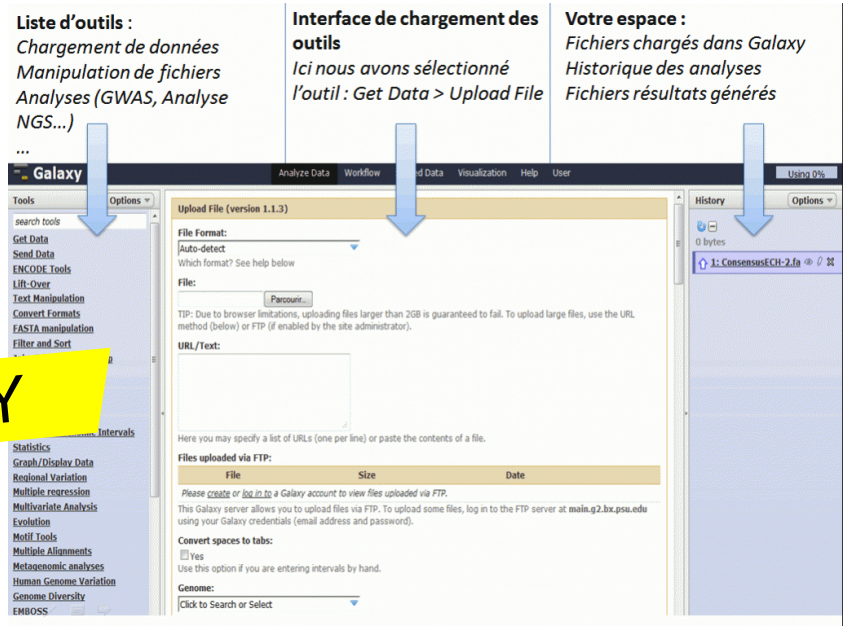
« Bricks » from Unix open source programs

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

Example: an RNA-seq pipeline



Galaxy: user-friendly interface to NGS pipelines

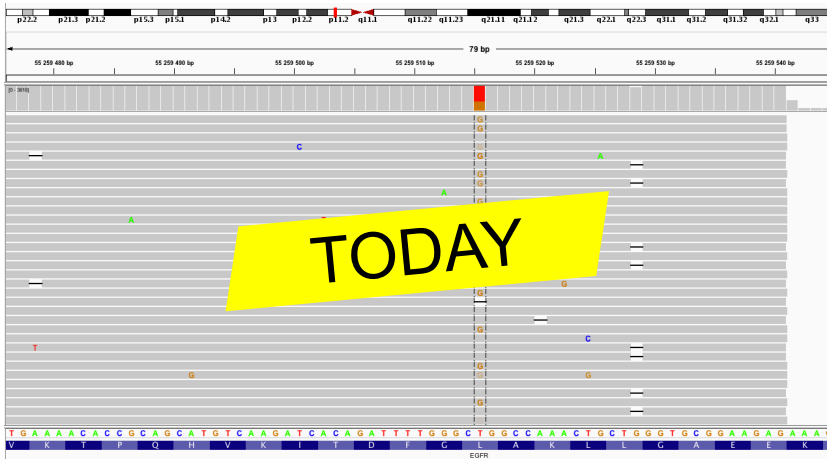


Credit: Biorigami

- 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)

Minimal starter pack for cancer genomics analysis

File formats: Fastq, BAM, VCF, BED, GFF



Visualisation: IGV

Statistics on expression, mutation, methylation tables

NEXT 2 DAYS

Thank you

- Next:
  - Practice: Exome-seq analysis with Galaxy