Bienvenue aux UE
Cancer et Génomique

UE11: Big data moléculaire et son traitement
UE12: Big data et modèles prédictifs

UNIVERSITÉ PARIS SUD
FACULTÉ DE MÉDECINE

IFSBM
INSTITUT DE FORMATION SUPÉRIEURE BIOMÉDICALE

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

Planning de la semaine

	Lundi 21 janvier 2019 - Salle 21, RdC
09:00-10:30	Technologies et données omiques en cancérologie. Daniel GAUTHERET
10:40-12:30	Concevoir un plan d'expérience et traiter les données primaires RNA-seq Thibault DAYRIS, ingénieur bioinformaticien
13:30-17:00	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). Gaëlle LELANDAIS

	Mardi 22 janvier 2019 - Salle 21, RdC
09:00-12:30	TP Galaxy II : Cas d'étude RNA-seq (création d'un workflow, matrice d'expression des gènes et analyse différentielle) Gaelle LELANDAIS
13:30-15:20	Problématique de la détection de variants somatiques par séquençage d'exome. D. GAUTHERET
15:30-17:00	L'analyse des altérations de nombre de copies par microarrays et NGS. Bastien JOB, Ingénieur Bioinformaticien, Gustave Roussy

	Mercredi 23 janvier 2018 - Salle 21, RdC
09:00-12:30	TP Galaxy (fin). Analyse exome. Visualisation de résultats avec IGV. D. GAUTHERET

	Mercredi 23 janvier 2019, Salle 21, RdC
13:30-17:00	Premiers pas avec le langage R. D. GAUTHERET

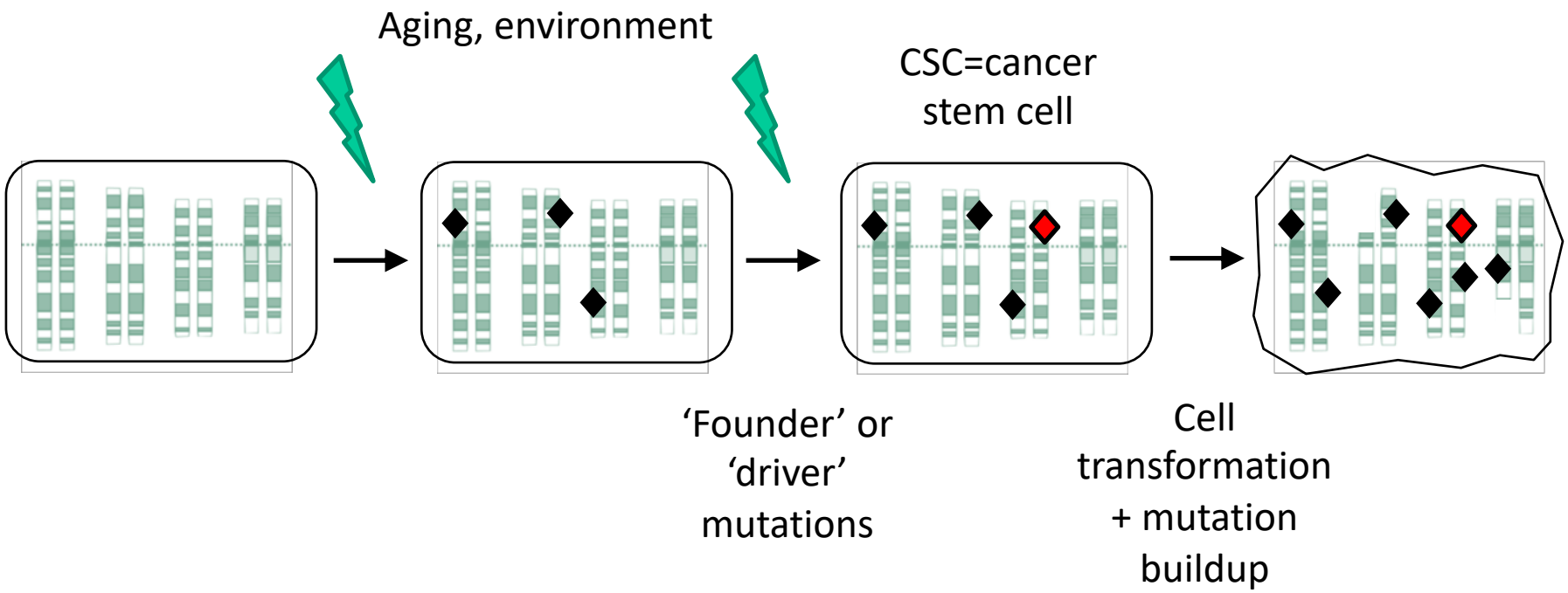
	Jeudi 24 janvier 2019, Salle 253, 2e et.
9:00-11:30	Pourquoi utiliser les methodes d'apprentissage automatique en oncologie personalisée? Loic VERLINGUE
11:30-12:30	Apprentissage automatique pour l'optimisation du dévellopement thérapeutique en oncologie: projet RESOLVED2. Guillaume BEINSE (MD, Msc)
13:30-17:00	TP: Travailler avec des données d'expression issues de TCGA. Identifier une signature predictive d'une variable clinique? Loic VERLINGUE

	Vendredi 25 janvier 2019 - Salle 253, 2e et.
9:00-12:30	Partir à la chasse de nouveaux marqueurs, cibles et médicaments contre le cancer avec des armes de petits ARN. François MAJOR, Pr. chaire d'Alembert, U. Paris-Sud
13:30-17:00	TP. Analyse de réseau avec miRbooking. Identification d'oncomiRs et de miRNA suppresseurs de tumeurs. François MAJOR

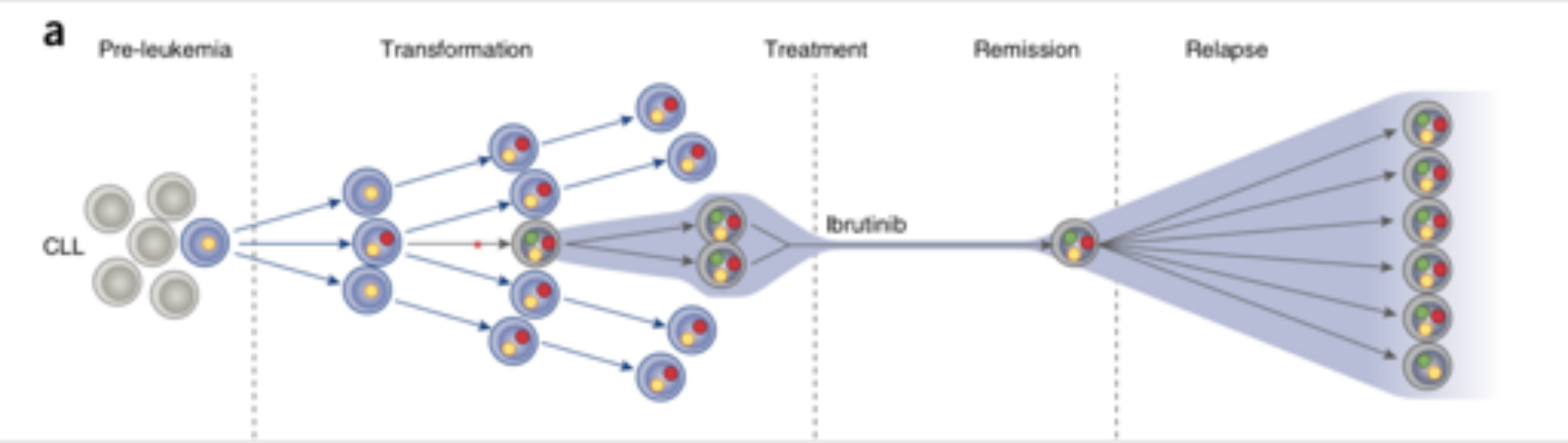
Evaluation:
Vos protocoles de TP
UE 11: Galaxy
UE 12: R

Introduction

- Le cancer: une maladie du génome

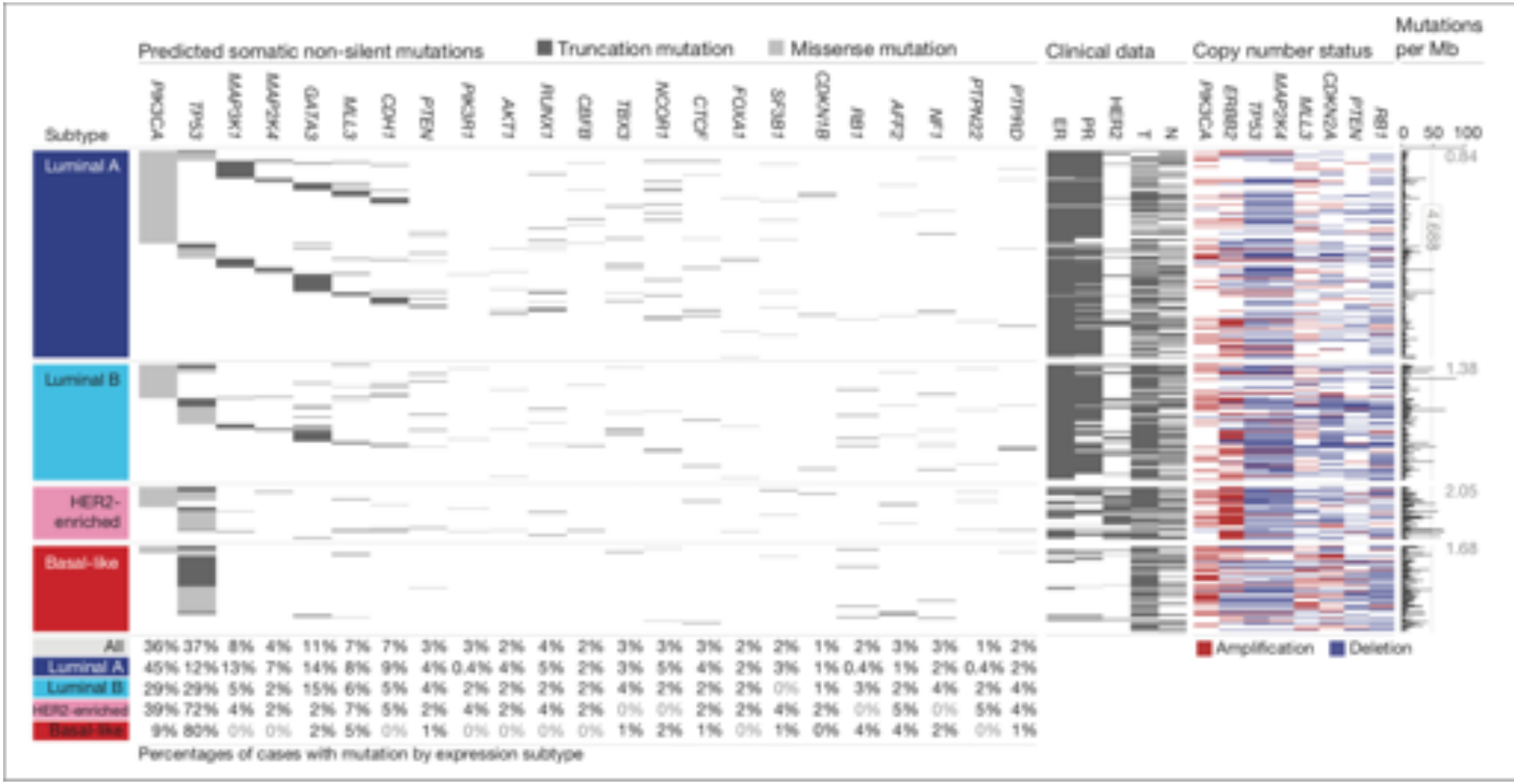


L'évolution clonale du cancer



Ferrando & Lopez-Otin, Nature Med. 2017

Big data génomique: études multi-patients



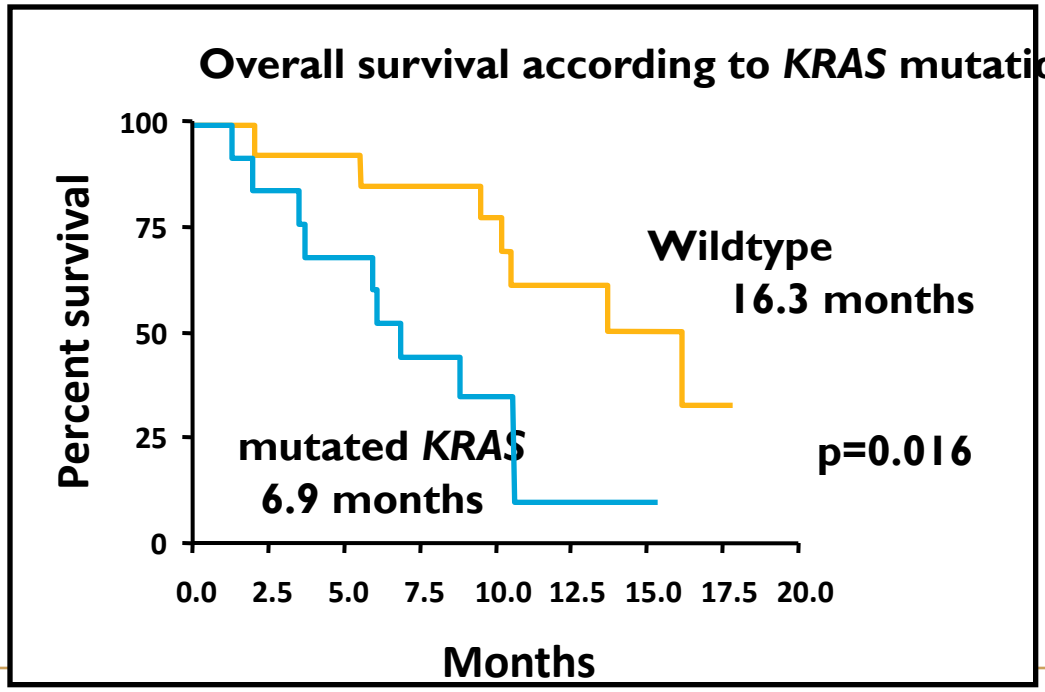
Mutations observées chez 825 patients de cancer du sein
TCGA consortium, Nature, 2012

Christophe Massard

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

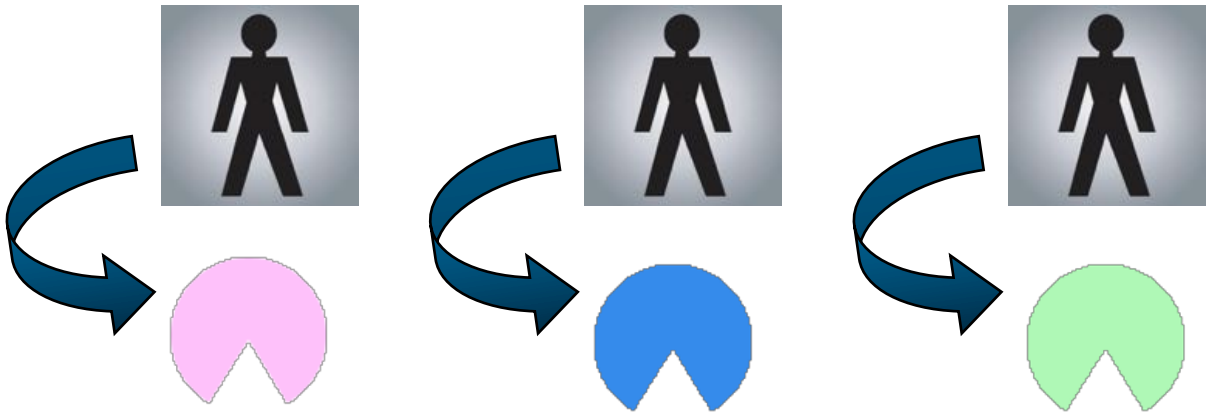
p=0.0003

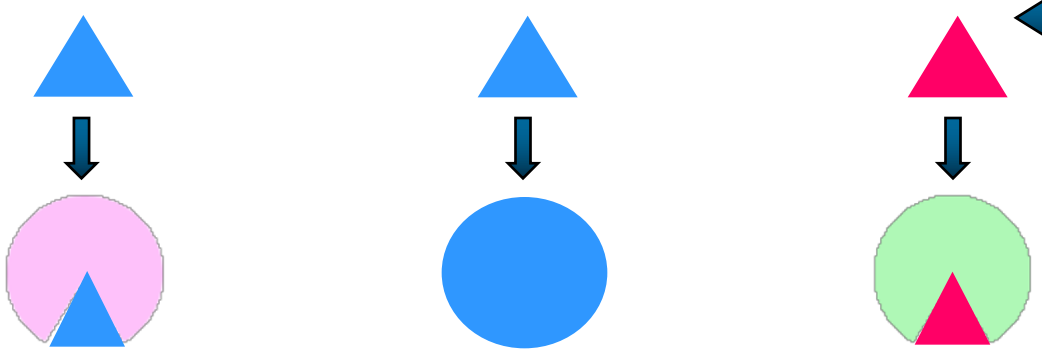


Lièvre A et al. Cancer Res 2006.

* Response according to RECIST criteria

Selecting the right therapy

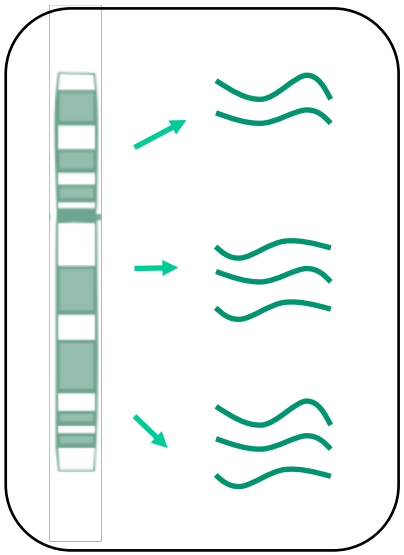




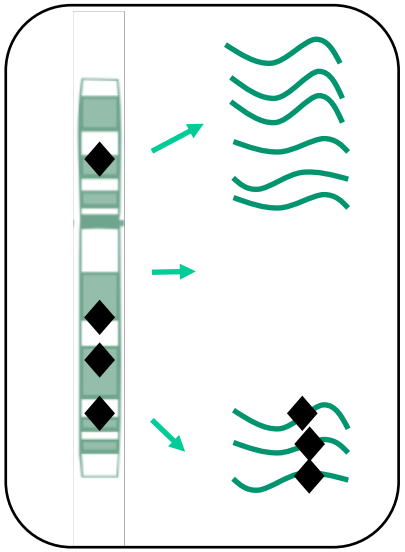
Wrong match | **The right drug for the tumor type** | **Wrong match**

C. Massard, Modified from D Weaver, On-Q-ity

L'ARN: le premier phénotype



Normal cell

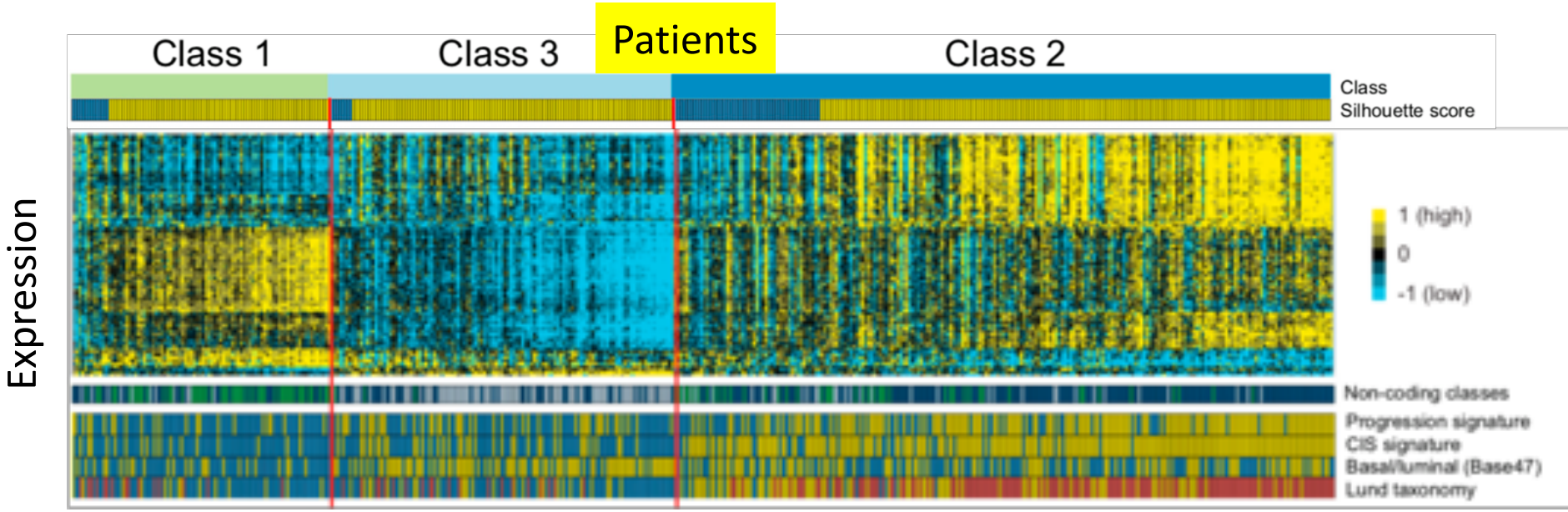


Cancer cell

Altered expression

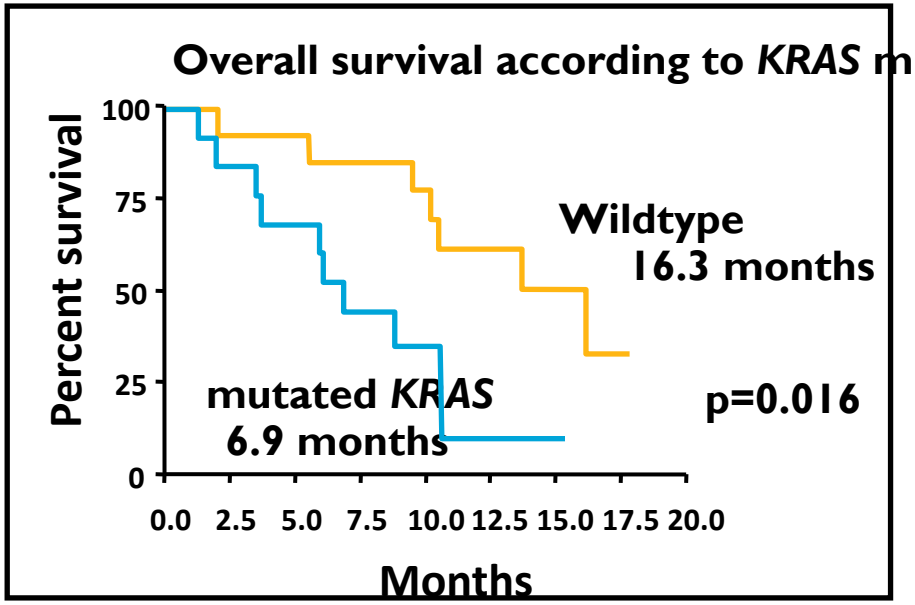
Aberrant transcripts

Big data transcriptomique: études multi-patients



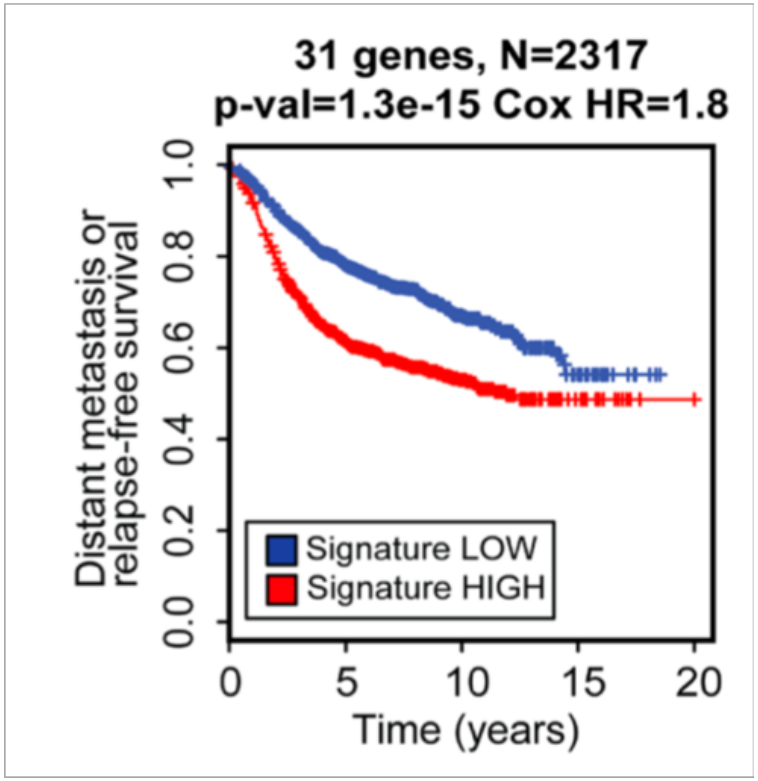
A 117-gene signature for urothelial carcinoma
Hedegaard et al., Cancer Cell, 2016

Signatures prédictives



DNA-based

Expression-based



Les principales technologies omiques en cancérologie

Daniel Gautheret

2019

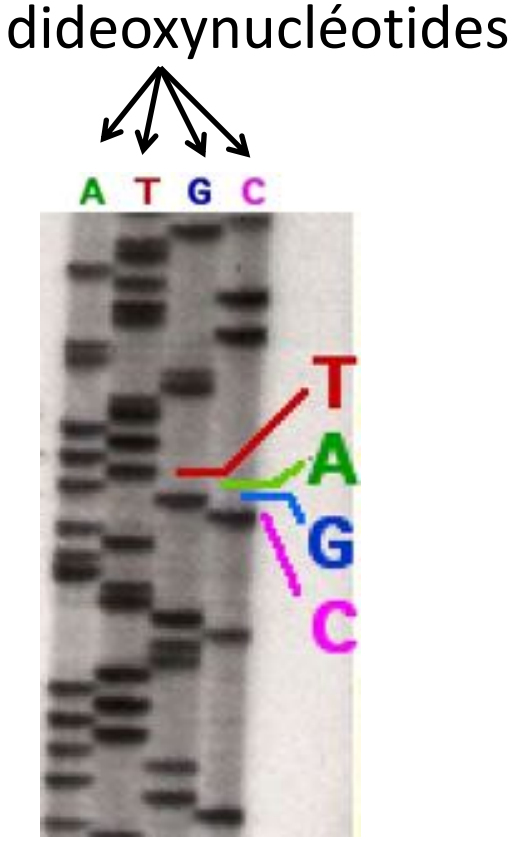
UNIVERSITÉ PARIS SUD
FACULTÉ DE MÉDECINE

IFSBM
INSTITUT DE FORMATION SUPÉRIEURE BIOMÉDICALE

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

Le séquençage de Sanger (1977)

- Séquençage par terminaison de chaine
  - Utilisation de dideoxynucléotides pour interrompre la synthèse à un certain type de base.
  - 4 réactions + marquage radioactif
- Amélioré en 1987 par l'introduction de marqueurs flurorescents (1 seule réaction) et l'automatisation.



Wikipedia

The Human Genome Project

« I expect that within a few years, our technology will be able to sequence one megabase/technician-year. At that rate 100 technicians could sequence the genome in 30 years. »

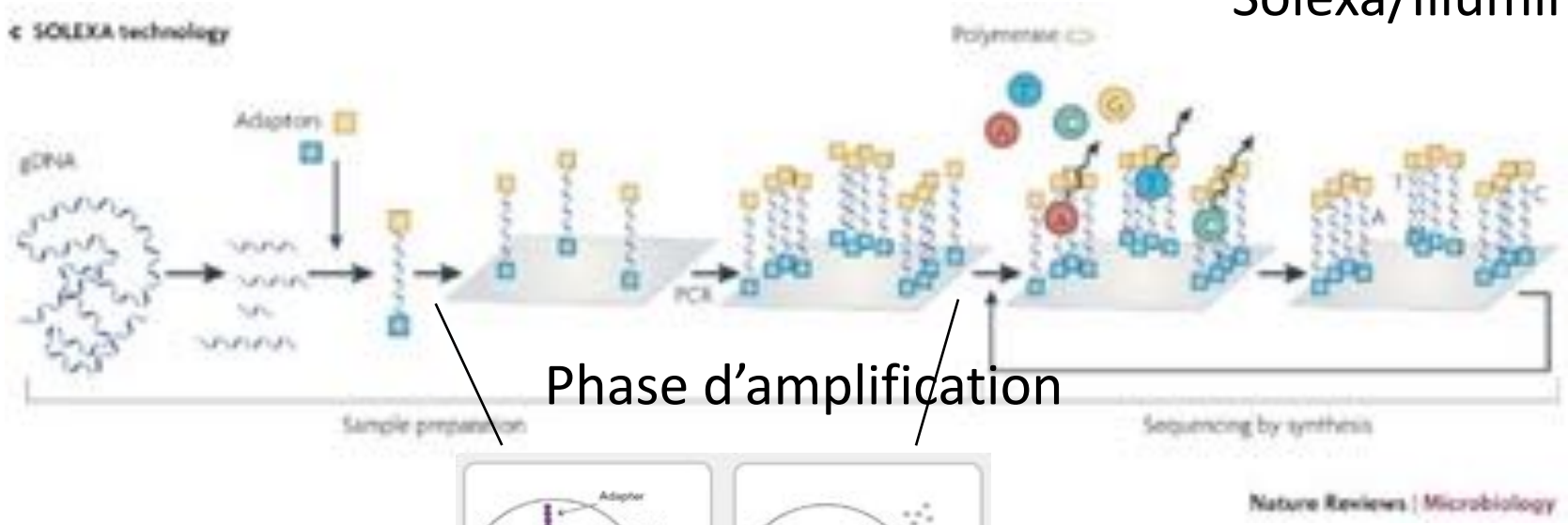
Walter Gilbert 1980

- Project started in 1991 and completed in 2001.
- …using Sanger sequencing

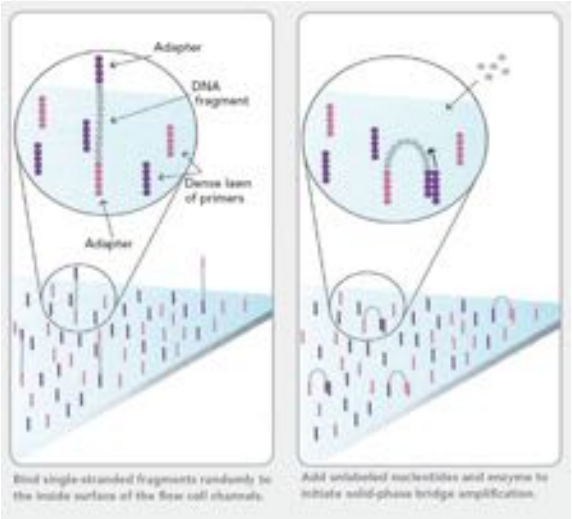
NGS :
Next Generation Sequencing
(2005-)

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

Solexa/Illumina



Phase d'amplification



Next Generation Sequencing

Nanopore MinIon

50Mb

Lifetech Ion torrent PGM

400 Mb

Illumina MySeq

4 Gb

Lifetech Ion proton

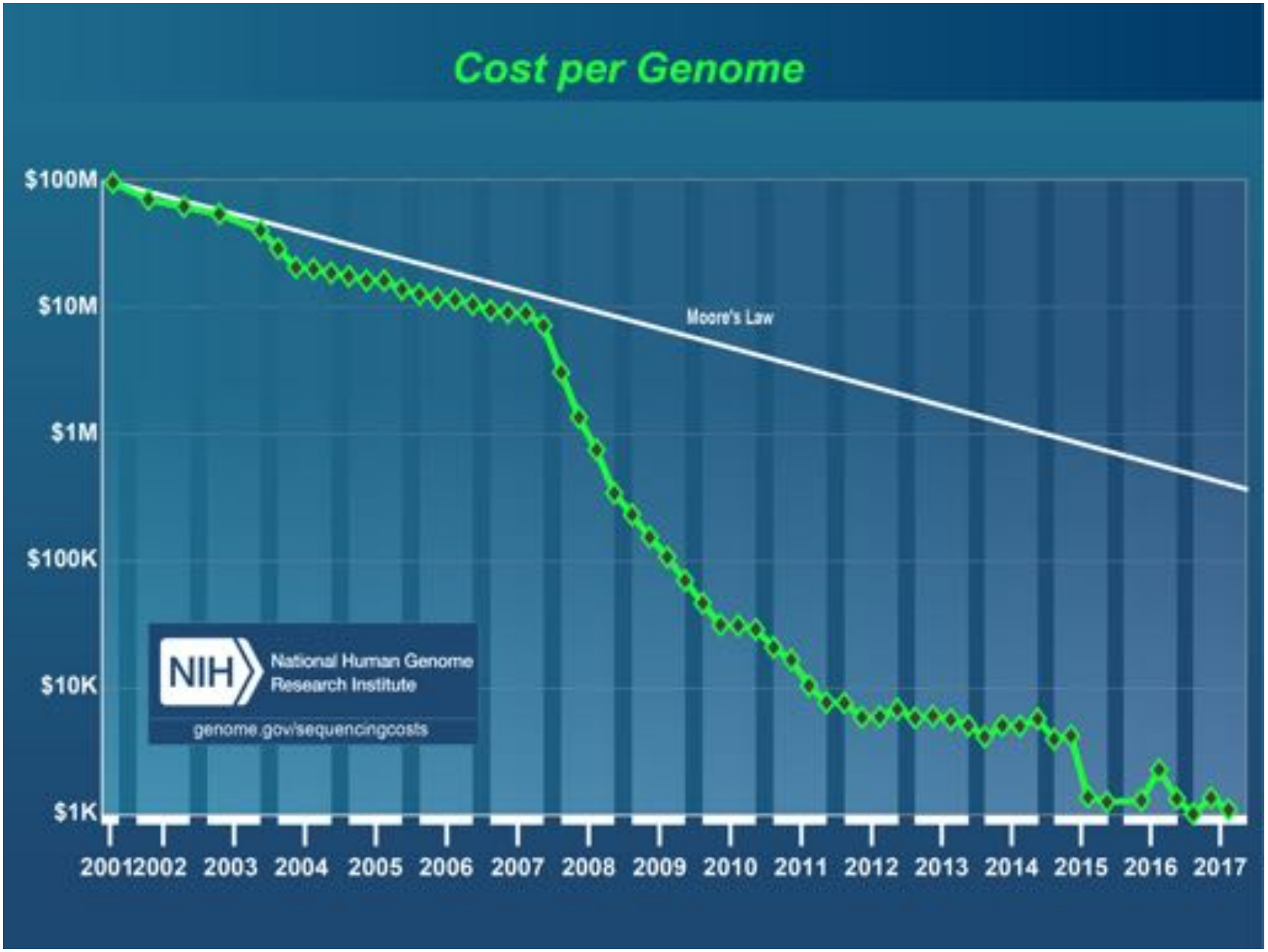
20 Gb

Illumina Hi-Seq 2000

300 Gb

Illumina NovaSeq

3Tb


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
20012002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

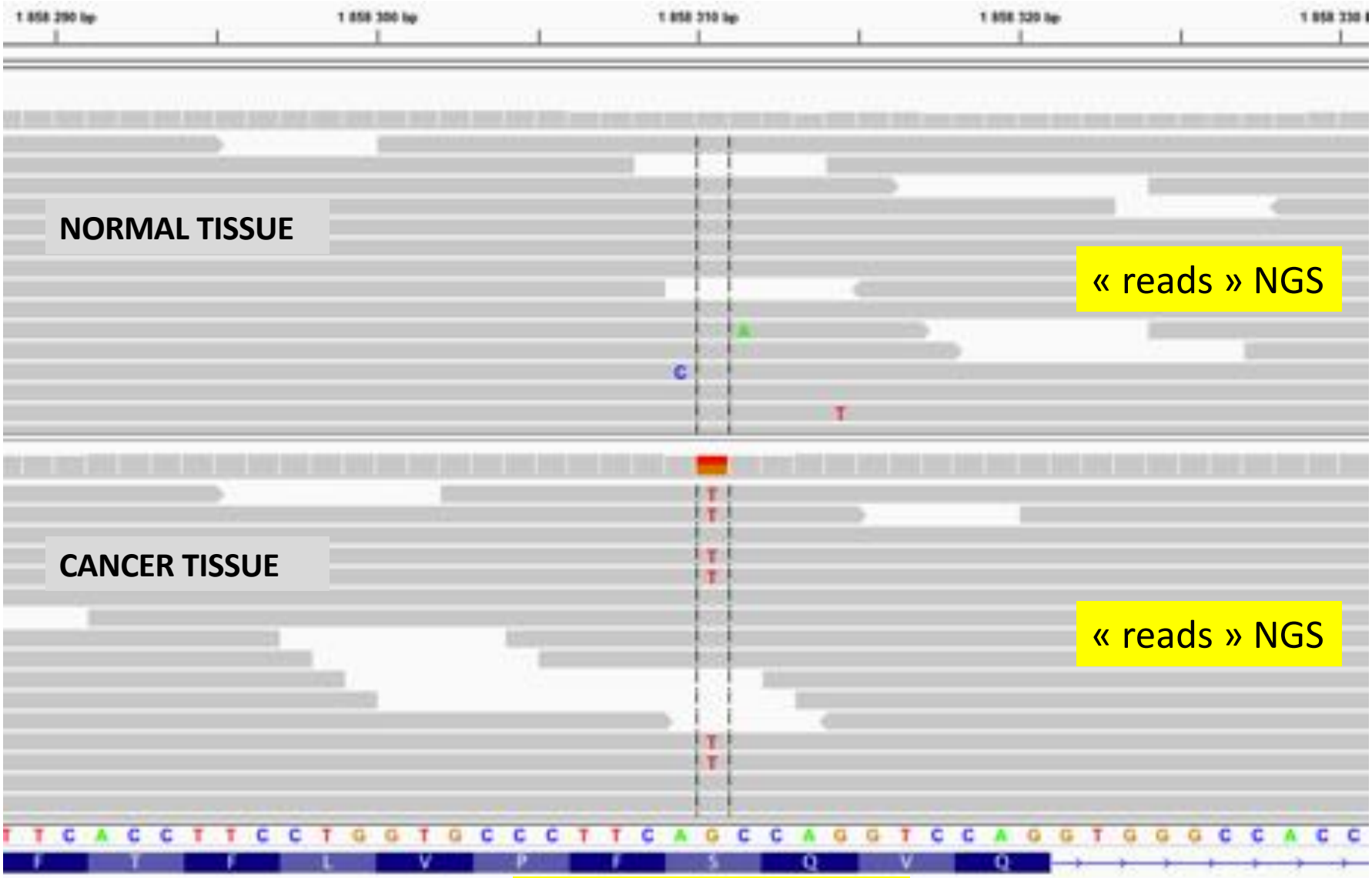
Les grandes applications des NGS

- 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)

Evénéments identifiés par DNA-seq



Evénéments identifiés par DNA-seq

Cf cours/TP exome
(B Job, D. Gautheret)

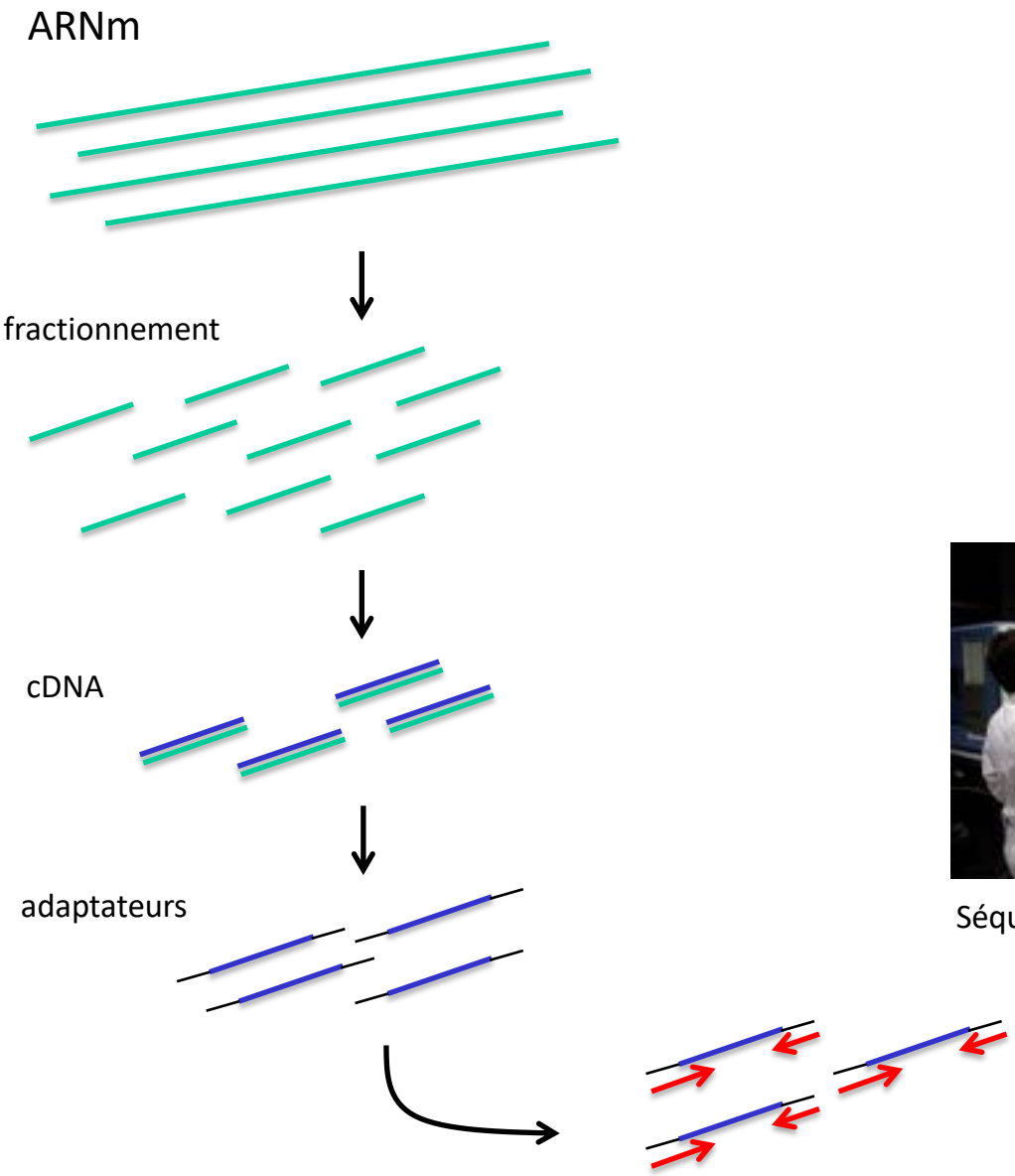
- Mutations ponctuelles
- Réarrangements
- CNV
- Amplification de microsatellites
- Profils mutationnels

RNA-seq

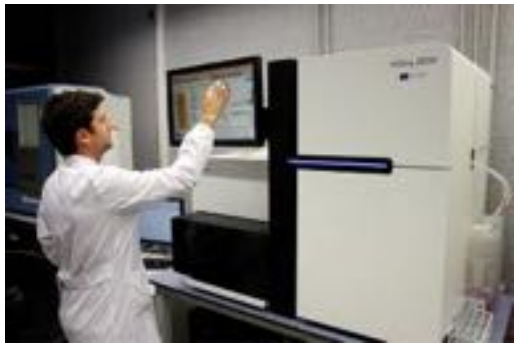
- Transcriptome par NGS = « deep sequencing »

Cf cours/TP RNA-seq
(T Dayris, G Lelandais)

RNA-Seq

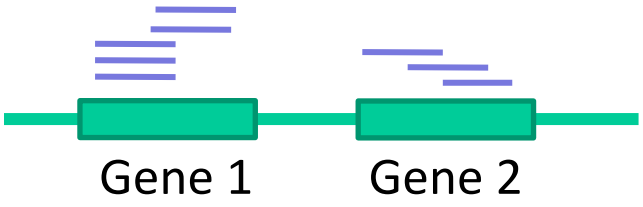


Cf cours DG Lundi pm
TD mardi am

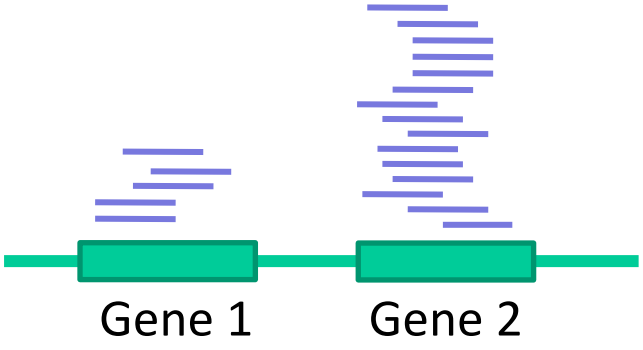


Séquençage

Mesures d'expression par RNA-seq



Sample 1



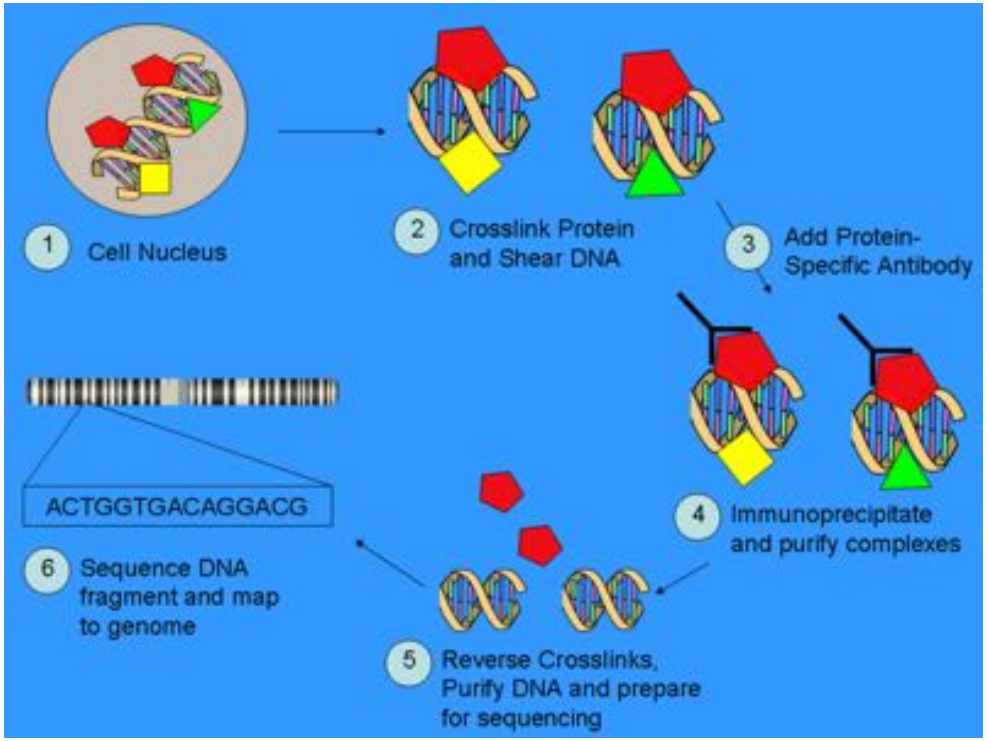
Sample 2

ChIP-Seq

- Chromatin ImmunoPrecipitation & Sequencing

ChIP-Seq

- Permet d'identifier les sites de liaison de protéines (histones, facteurs de transcription, represseurs, enhancers, etc.) sur l'ADN génomique

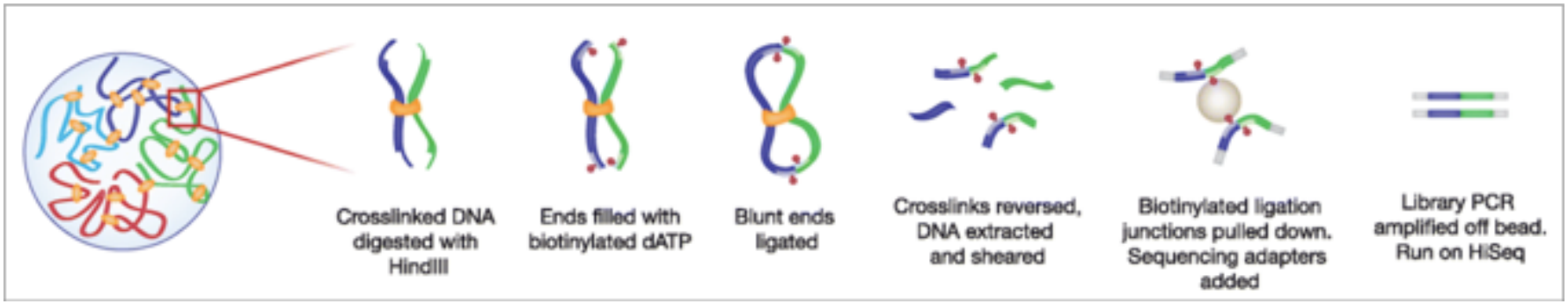


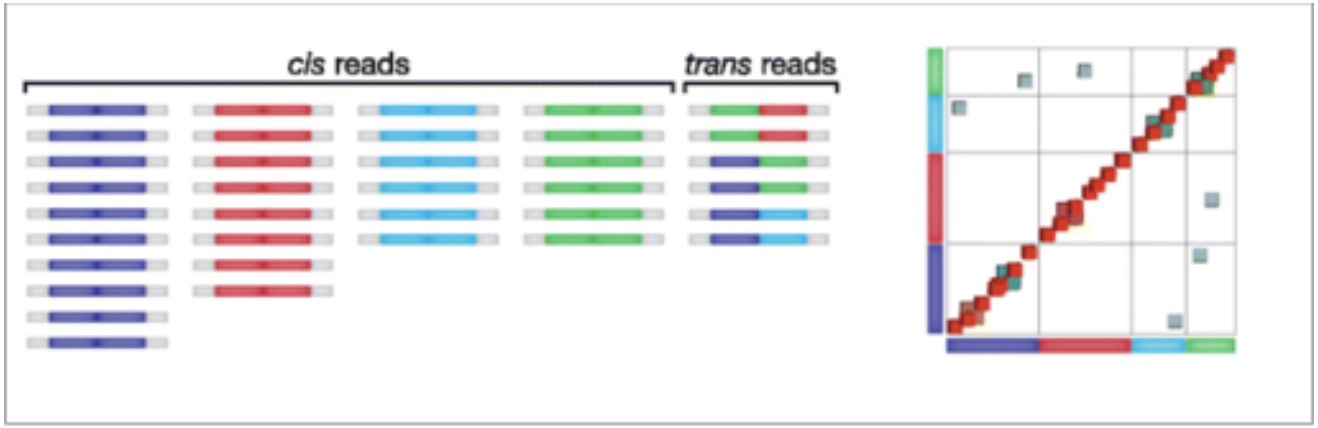
Wikipedia

Hi-C

- Chromosome conformation capture

Hi-C





Harewood et al. Genome Biol 2017

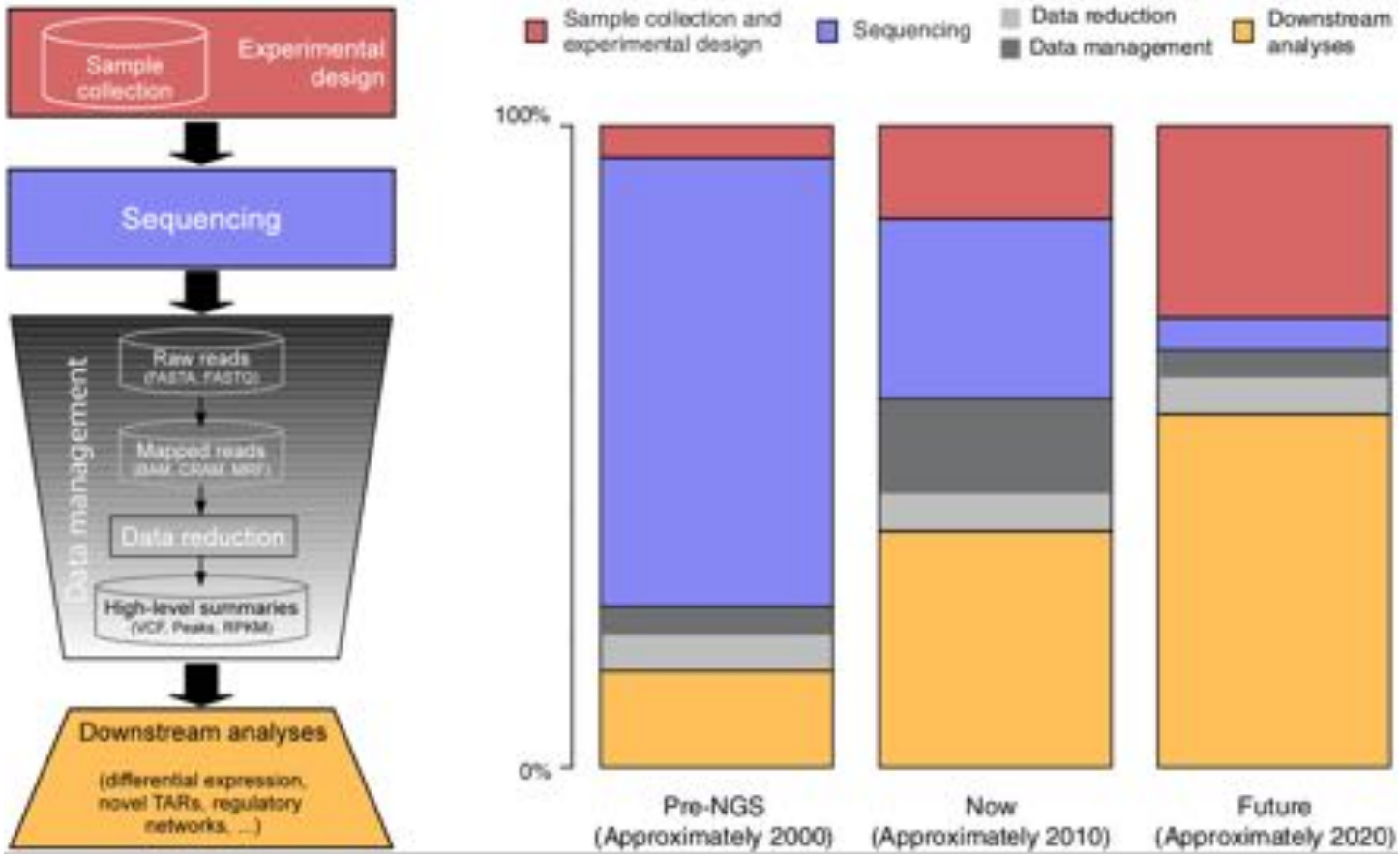
Vers une utilisation systématique du séquençage en médecine

Les données NGS

Volume des données NGS

- Un exome humain (N+T) avec fichiers de mapping et analyse: 70 Go
  - (prévoir ~5 fois le volume des fastq.gz)
- Données génomiques produites annuellement dans un hôpital universitaire: >500 To
- La banque TCGA complete: 1 Po

Costs of sequencing *vs* analysis



Sboner et al. 2011. Genome Biology, **12**:125

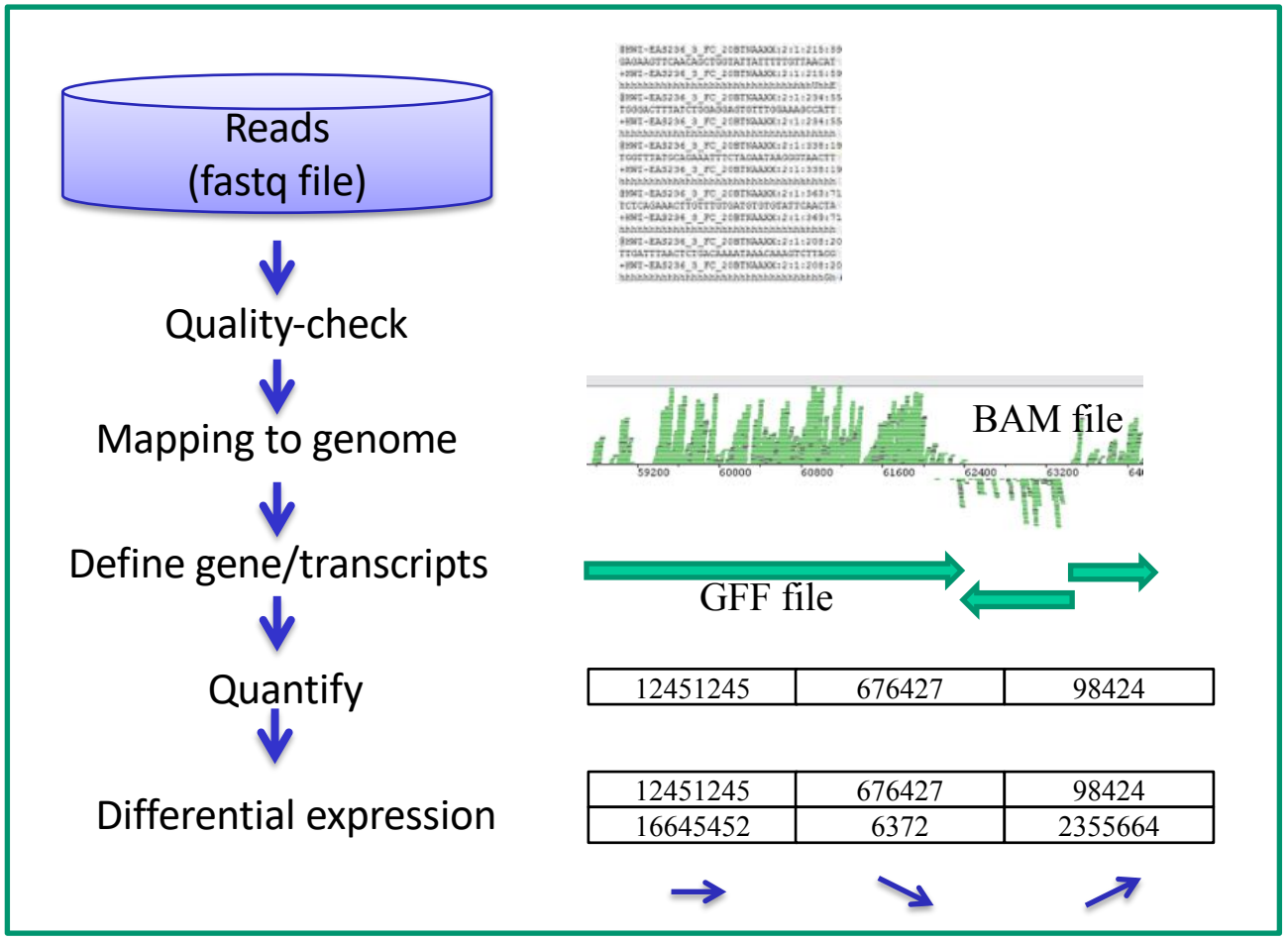
Les outils

« Pipelines » & « workflows »

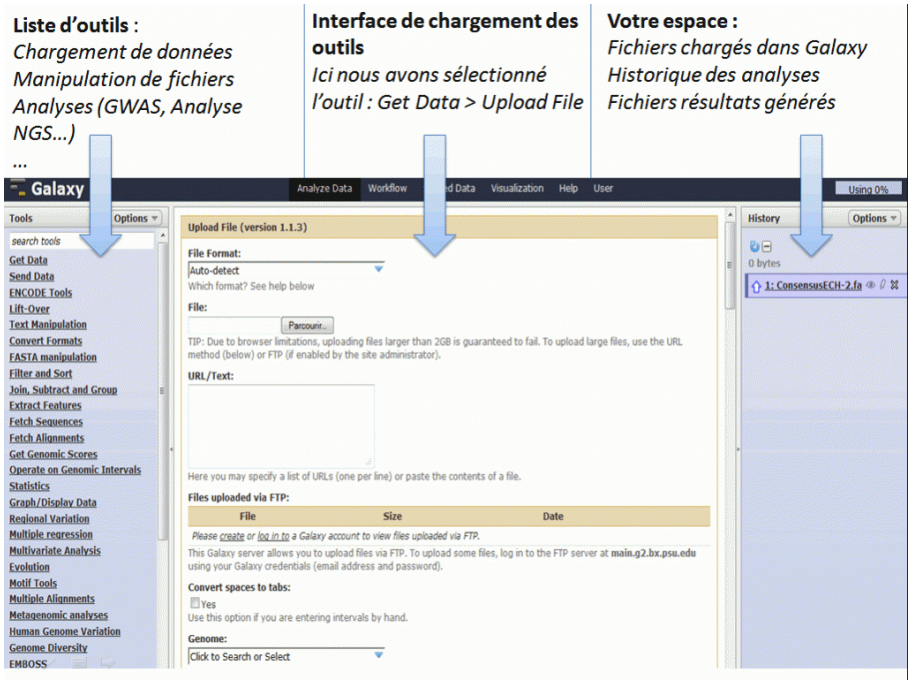
« Bricks » from Unix open source programs

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

NCI, NHGRI, USA

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

TCGA

- launched by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in 2006
- 33 tumor types
- 11,000 patients
- whole-genome sequencing (WGS) for 1,000 tumors

TCGA data types and levels

Controlled access

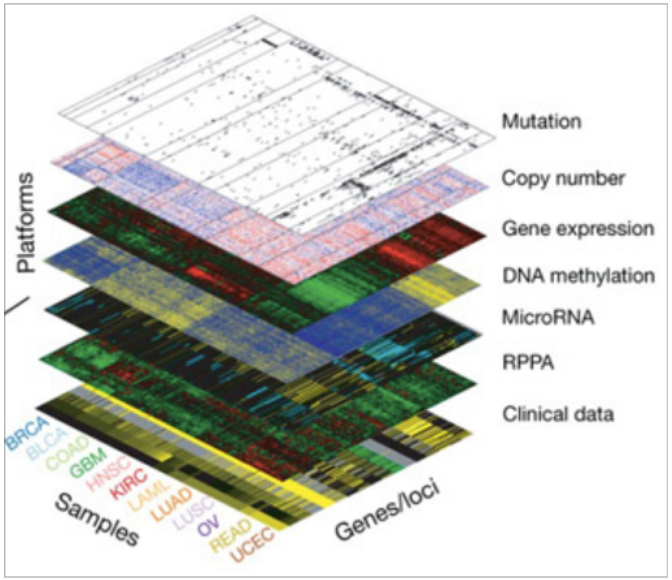
Free access

- 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

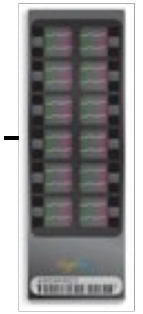


Main data types

- DNA sequencing
  - Whole exome or whole genome DNA se
  - Platform: Illumina HiSeq
- mRNA sequencing / miRNA sequencing
  - PolyA+ RNA / small RNA expression from RNA-seq
  - Platform: Illumina HiSeq or similar
- Array-based expression
  - mRNA expression levels (1 or 2 colors)
  - Illumina or Agilent DNA microarrays

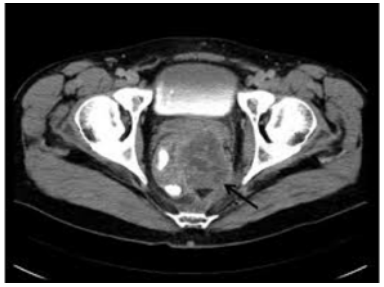
Main data types

- DNA Methylation
  - covalent modification of cytosine bases at the C-5 position, generally within a CpG sequence context
  - platforms: Illumina Methyl arrays
- Protein expression
  - protein expression & concentration
  - Platform: custom antibody array (5ABx1000 samples/slide)
- Copy number
  - Loss and gain of DNA fragments
  - Platform: Agilent CGH array



Main data types

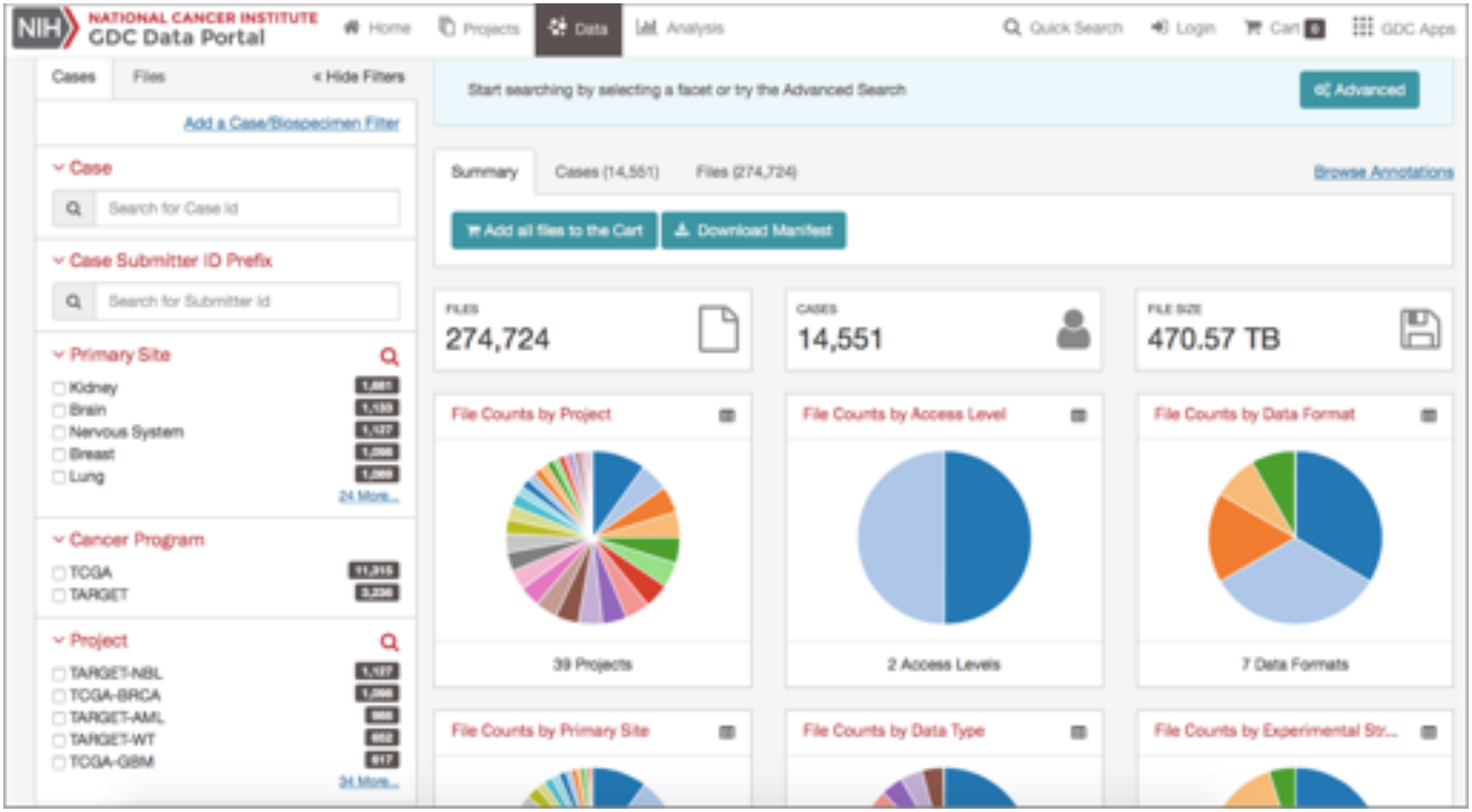
- Microsatellite instability
  - MSI-Mono-Dinucleotide Assay: panel of 4 mononucleotide and 3 dinucleotide repeat loci



- Image
  - Images of tissue samples
  - CT (computed tomography), DX (digital radiography), CR (computed radiography)
- Clinical data
  - Available clinical information for each participant (demographic, treatment, survival, etc)
  - Biospecimen data: how specimen was processed

TCGA access via the GDC portal (Genomics Data Commons)

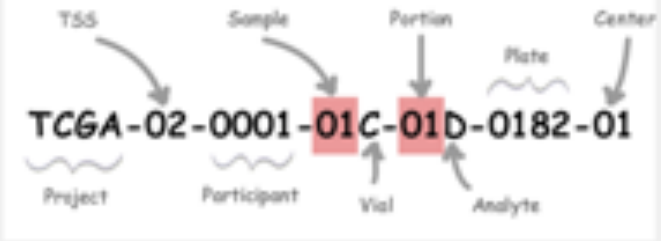
TCGA data access via the GDC Data portal



Example of access levels

	Level 1	Level 2	Level 3
RNA-seq	mRNA sequence for each participant's tumor sample		The calculated expression signal of a particular composite exon of a gene, per sample
DNA-seq	Whole exome sequence for both tumor and normal sample for each participant	Somatic mutation calls for each participant	

1 sample = 1 TCGA barcode



This figure of an aliquot barcode shows how it can be broken down into its components and translated into its metadata. The barcode metadata are further described in the following table.

Label	Identifier for	Value	Value description	Possible values
Project	Project name	TCGA	TCGA project	TCGA
TSS	Tissue source site	02	GBM (brain tumor) sample from MD Anderson	See Code Tables Report
Participant	Study participant	0001	The first participant from MD Anderson for GBM study	Any alpha-numeric value
Sample	Sample type	01	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	01	The first portion of the sample	01-99
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	0182	The 182nd plate	4-digit alphanumeric value
Center	Sequencing or characterization center that will receive the aliquot for analysis	01	The Broad Institute GCC	See Code Tables Report

TCGA **Clinical Data** (patient or sample XML file)

Clinical — emacs — 179×47

```
<?xml version="1.0" encoding="UTF-8"?>
<brca:tcga_bcr>
    <admin:admin>
        <admin:bcr xsd_ver="1.17">Nationwide Children's Hospital</admin:bcr>
        <admin:batch_number xsd_ver="1.17">80.48.0</admin:batch_number>
        <admin:disease_code xsd_ver="2.3">BRCA</admin:disease_code>
        <admin:day_of_dcc_upload xsd_ver="1.17">27</admin:day_of_dcc_upload>
        <admin:month_of_dcc_upload xsd_ver="1.17">2</admin:month_of_dcc_upload>
        <admin:year_of_dcc_upload xsd_ver="1.17">2014</admin:year_of_dcc_upload>
        <admin:patient_withdrawal>
            <admin:withdrawn>false</admin:withdrawn>
        </admin:patient_withdrawal>
    </admin:admin>
    <brca:patient>
        <shared:tumor_tissue_site>Breast</shared:tumor_tissue_site>
        <shared:gender>FEMALE</shared:gender>
        <shared:vital_status>Alive</shared:vital_status>
        <shared:race>WHITE</shared:race>
        <shared:bcr_patient_barcode>TCGA-BH-A0B2</shared:bcr_patient_barcode>
...
        <shared:history_of_neoadjuvant_treatment>No</shared:history_of_neoadjuvant_treatment>
        <shared:informed_consent_verified>YES</shared:informed_consent_verified>
...
        <shared:age_at_initial_pathologic_diagnosis>43</shared:age_at_initial_pathologic_diagnosis>
...
        <shared:histological_type>Infiltrating Ductal Carcinoma</shared:histological_type>
        <brca_shared:breast_carcinoma_progesterone_receptor_status>Positive</brca_shared:breast_carcinoma_progesterone_receptor_status>
...
        <brca_shared:breast_carcinoma_estrogen_receptor_status>Positive</brca_shared:breast_carcinoma_estrogen_receptor_status>
        <brca_shared:lab_proc_her2_neu_immunohistochemistry_receptor_status>Negative</brca_shared:lab_proc_her2_neu_immunohistochemistry_receptor_status>
...
        <brca_nte:new_tumor_events>
            <nte:new_tumor_event_after_initial_treatment>
        </brca_nte:new_tumor_events>
    </brca:patient>
</brca:tcga_bcr>
```

200/300 lines per file

Extract of patient xml clinical file

TCGA is over
PCAWG is on

- PCAWG[1]: a collaboration with 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

- Expert-curated database of cancer somatic mutations & other events
- 2017 (V82):
  - 4.8M coding point mutations
  - 18k Gene fusions
  - 1.2 M CNV
  - 9M gene expression variants
  - 202 cancer genes (tier 1) + 300 fusions

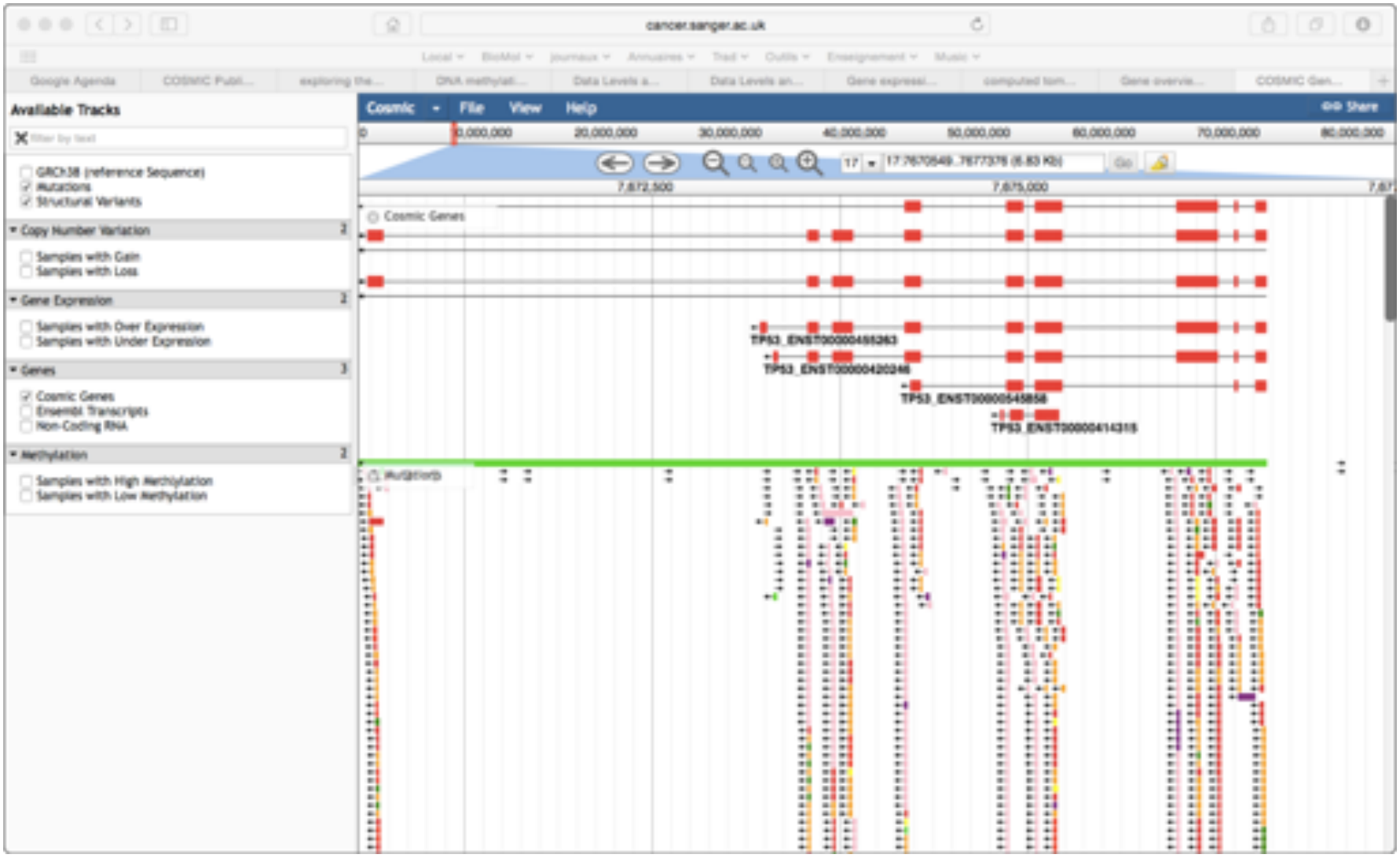
COSMIC Curation

- Manual curation
  - 25000 articles analyzed
- Automated curation
  - 1M 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. »

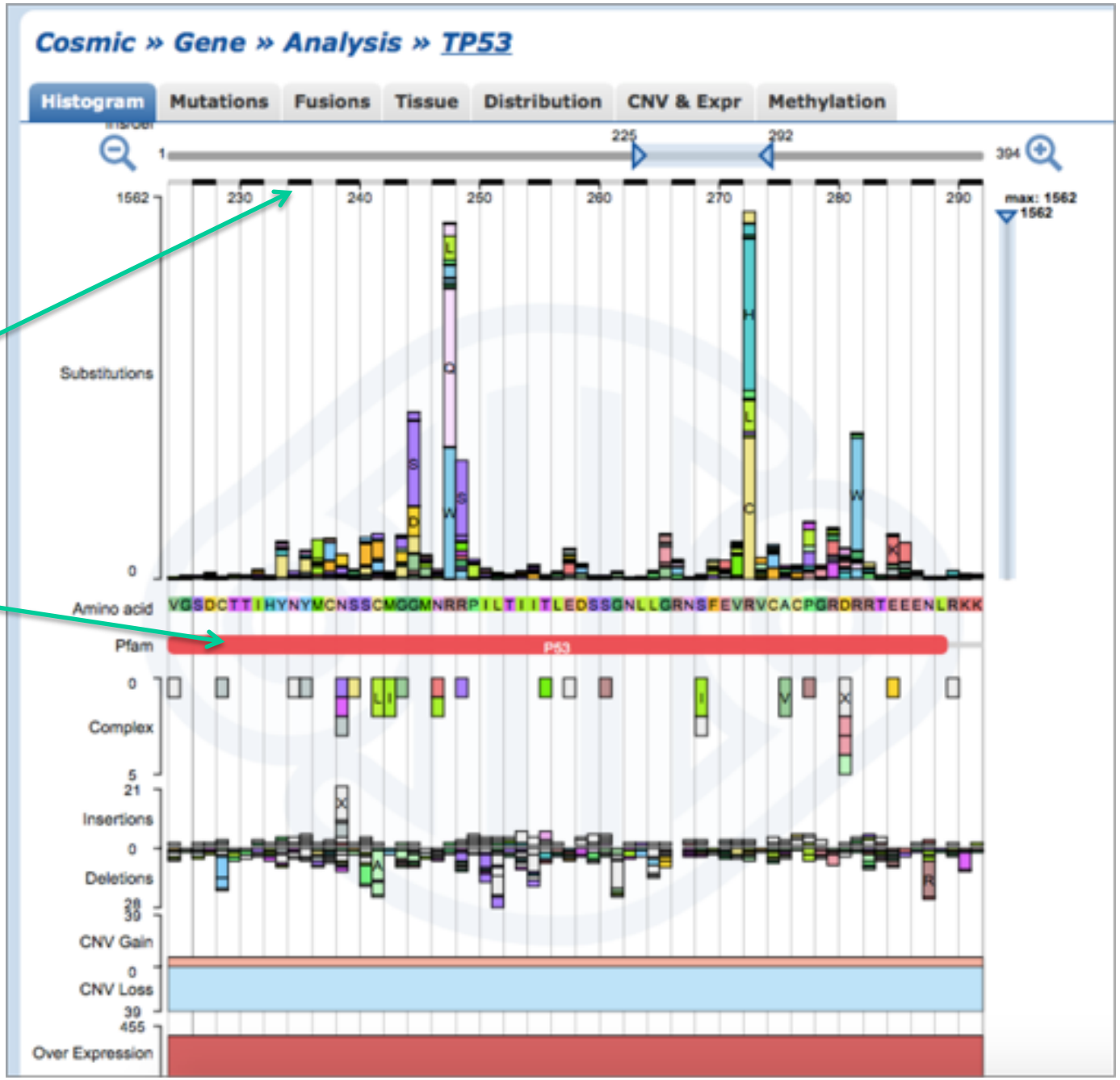
COSMIC genome browser



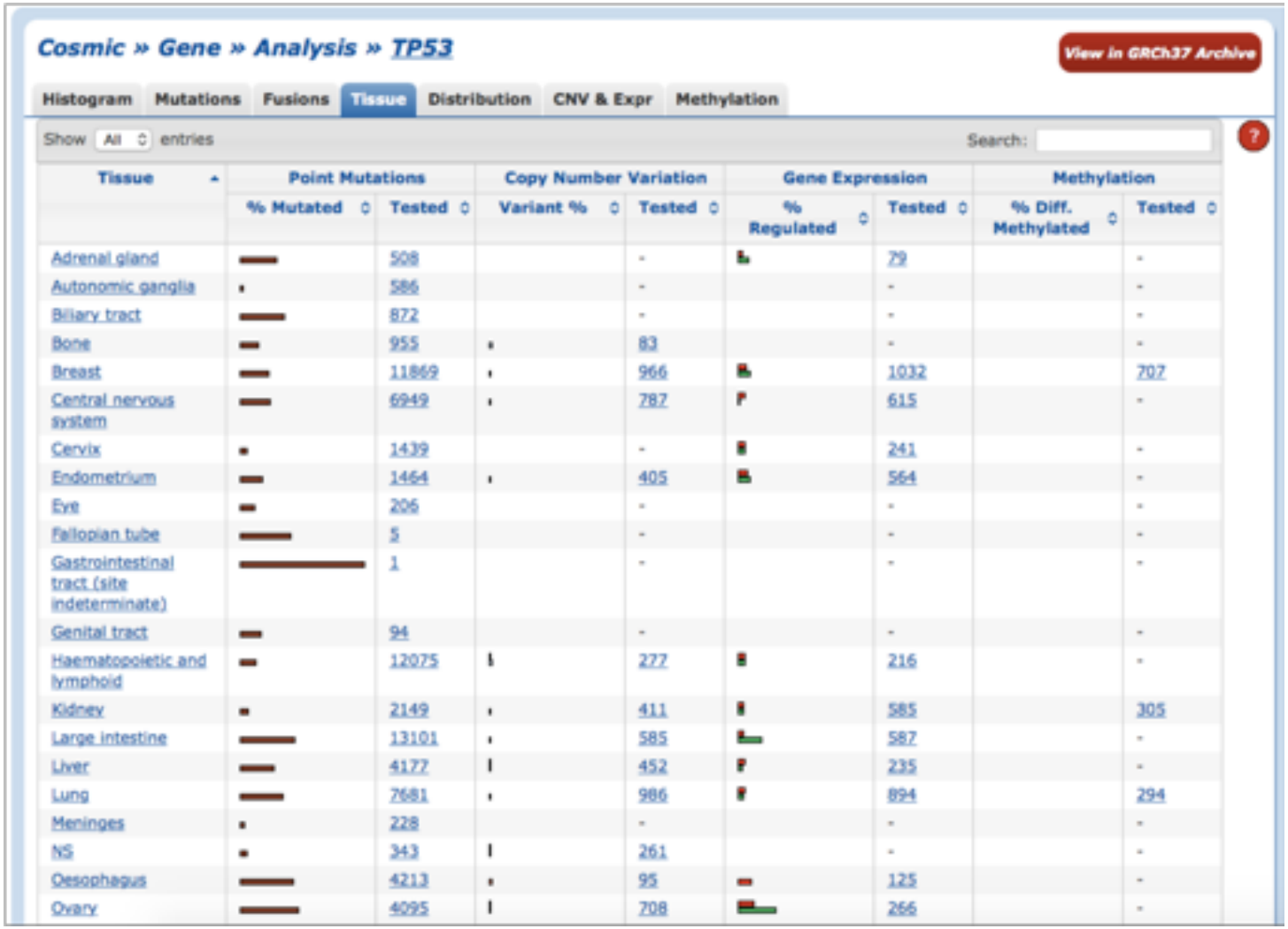
Histogram view

Protein coordinates

Protein domain



Tissue-distribution of mutations



Cosmic » Gene » Analysis » TP53

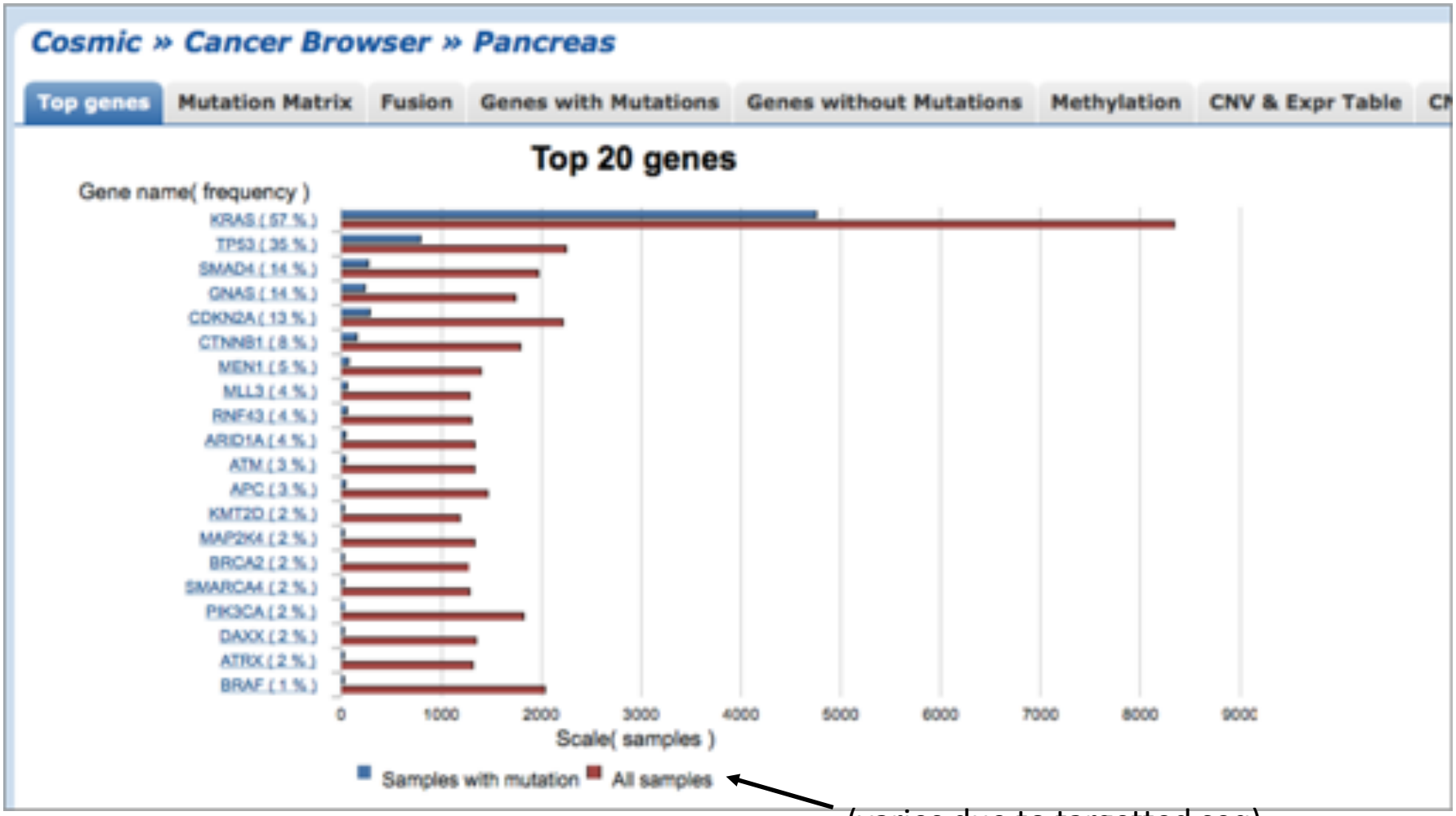
View in GRCh37 Archive

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		-

Cancer browser

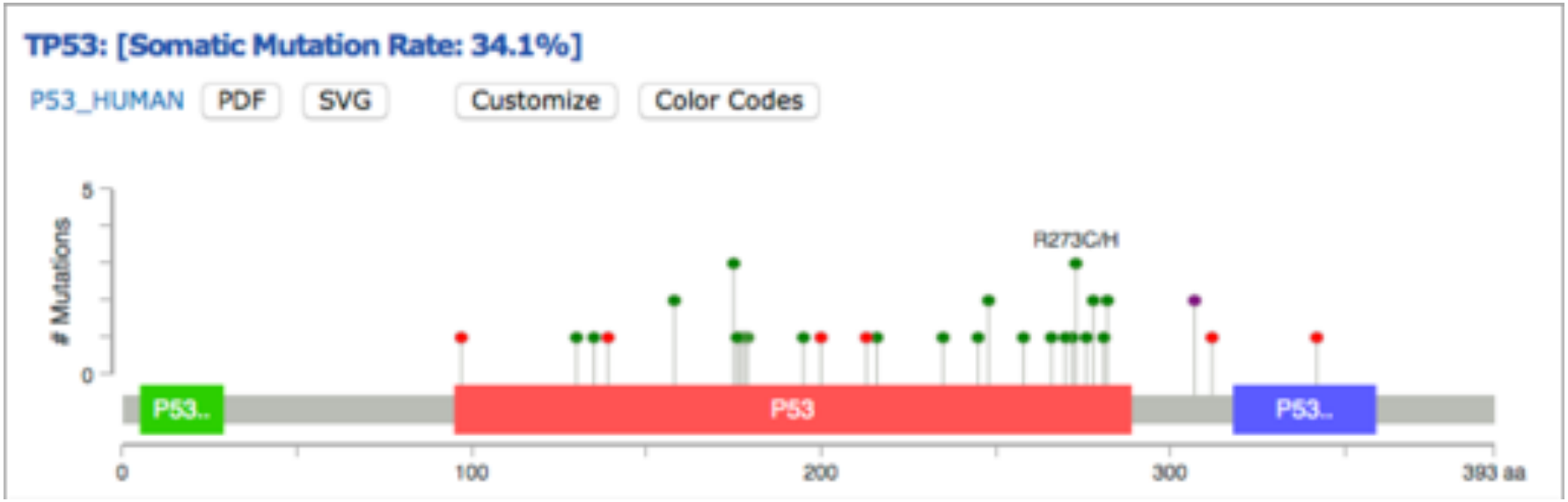


(varies due to targetted seq)

Memorial Sloan-Kettering
Cancer Center, USA

- Integration of Data from 89 cancer genomics studies.
- Focus on analysis tools
  - Mutual exclusivity
  - Gene networks

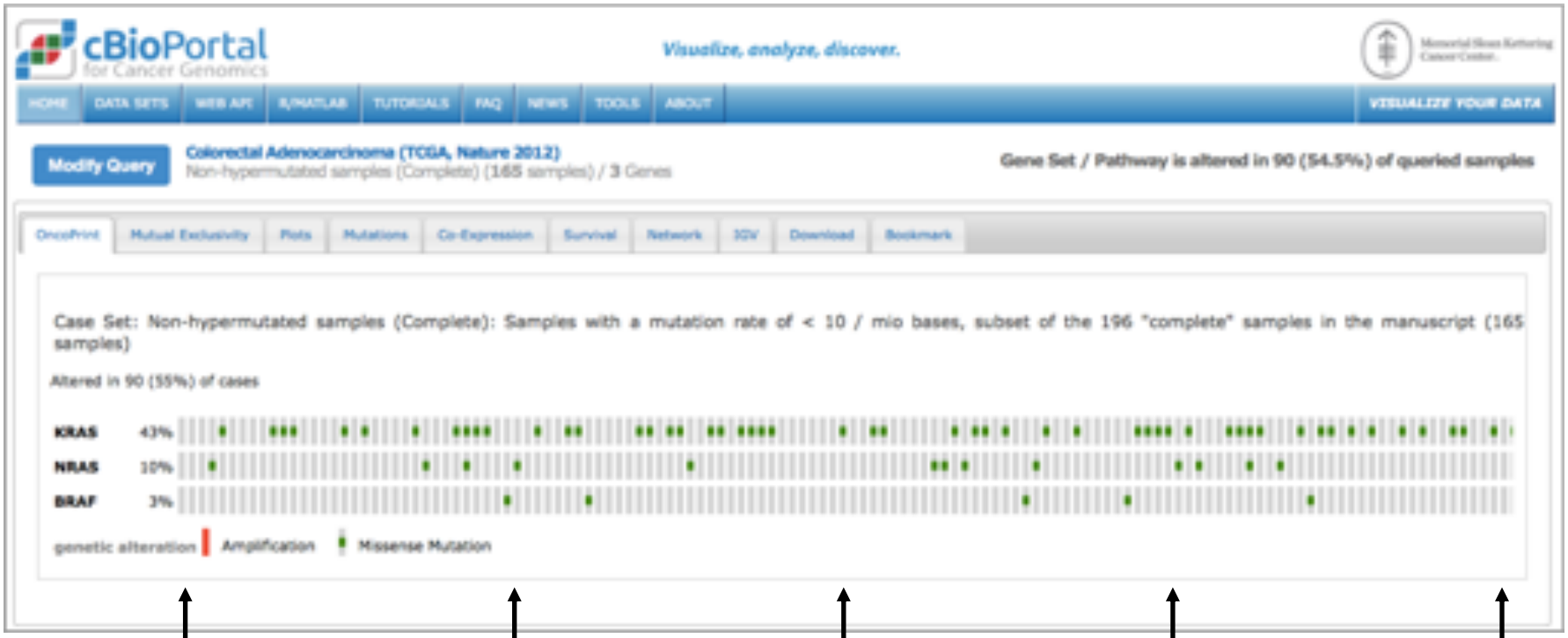
Mapped mutations on proteins



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

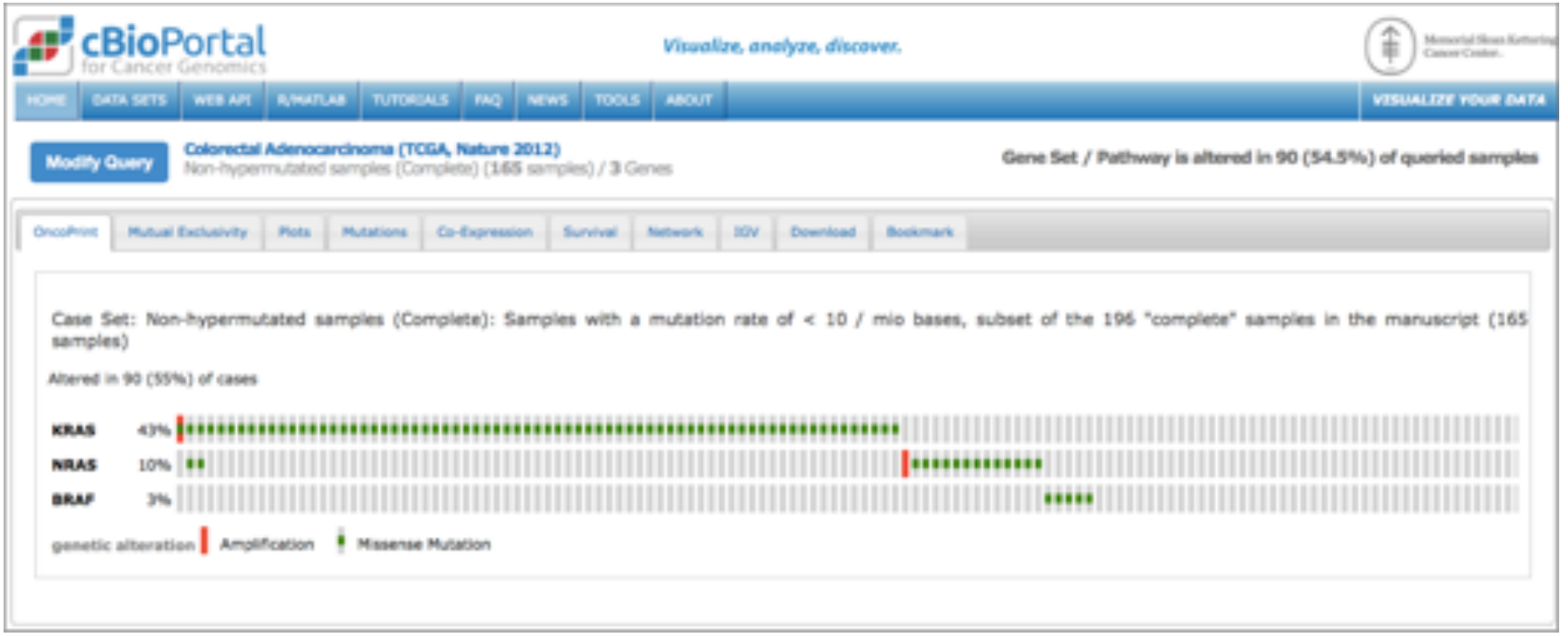
See also « MutationMapper » tool

« Oncoprint » view

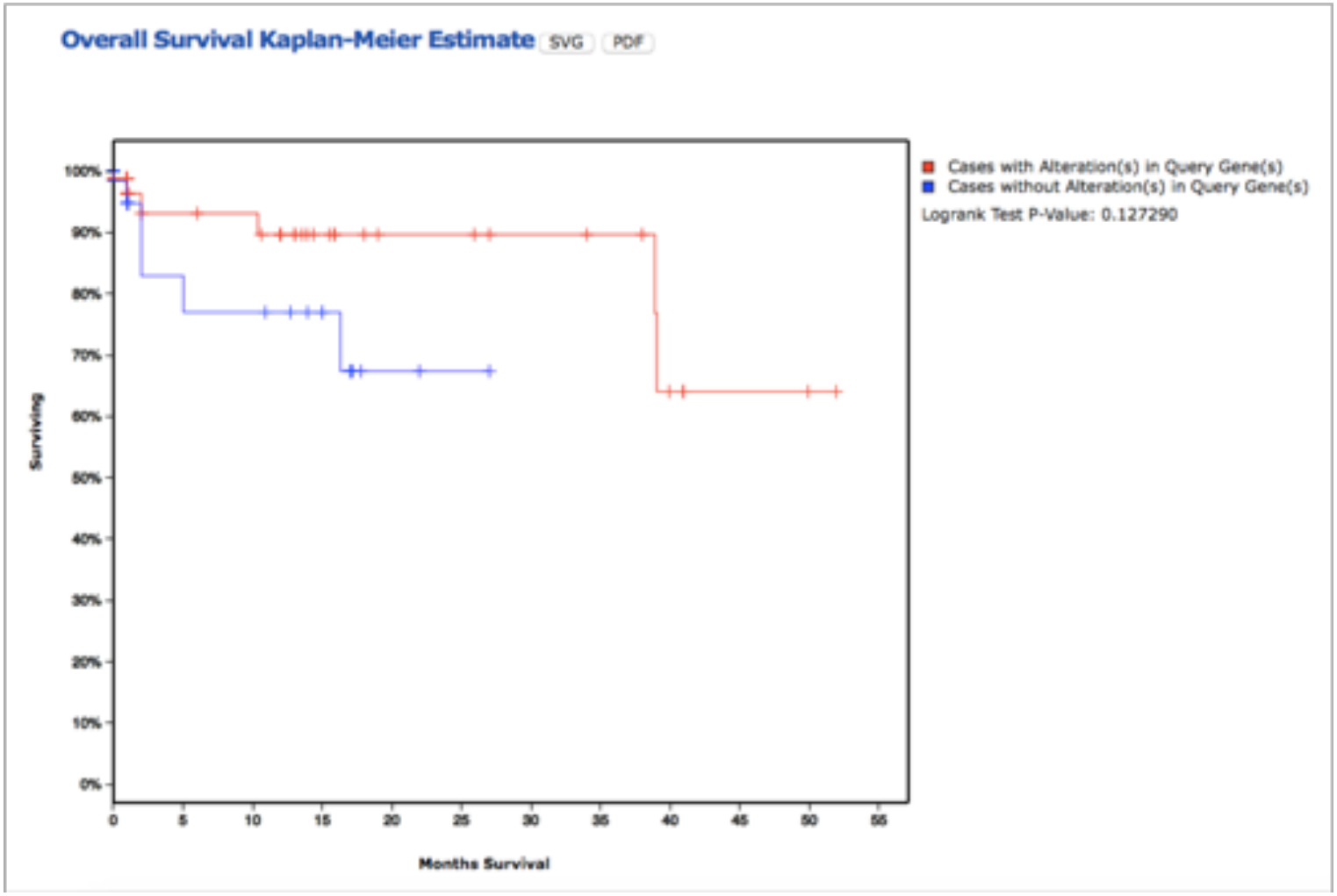


patients

Mutual exclusivity



Kaplan-Meier Curves



Programmatic Interfaces

- Webservice
  - http://www.cbioportal.org/webservice.do?cmd=getCaseLists&cancer_study_id=gbm_tcga
- R library
  - CGDS package (CRAN)
- Matlab Library
  - CGDS toolbox @ MatLab Central