Introduction à la bioinformatique en cancérologie

Daniel Gautheret

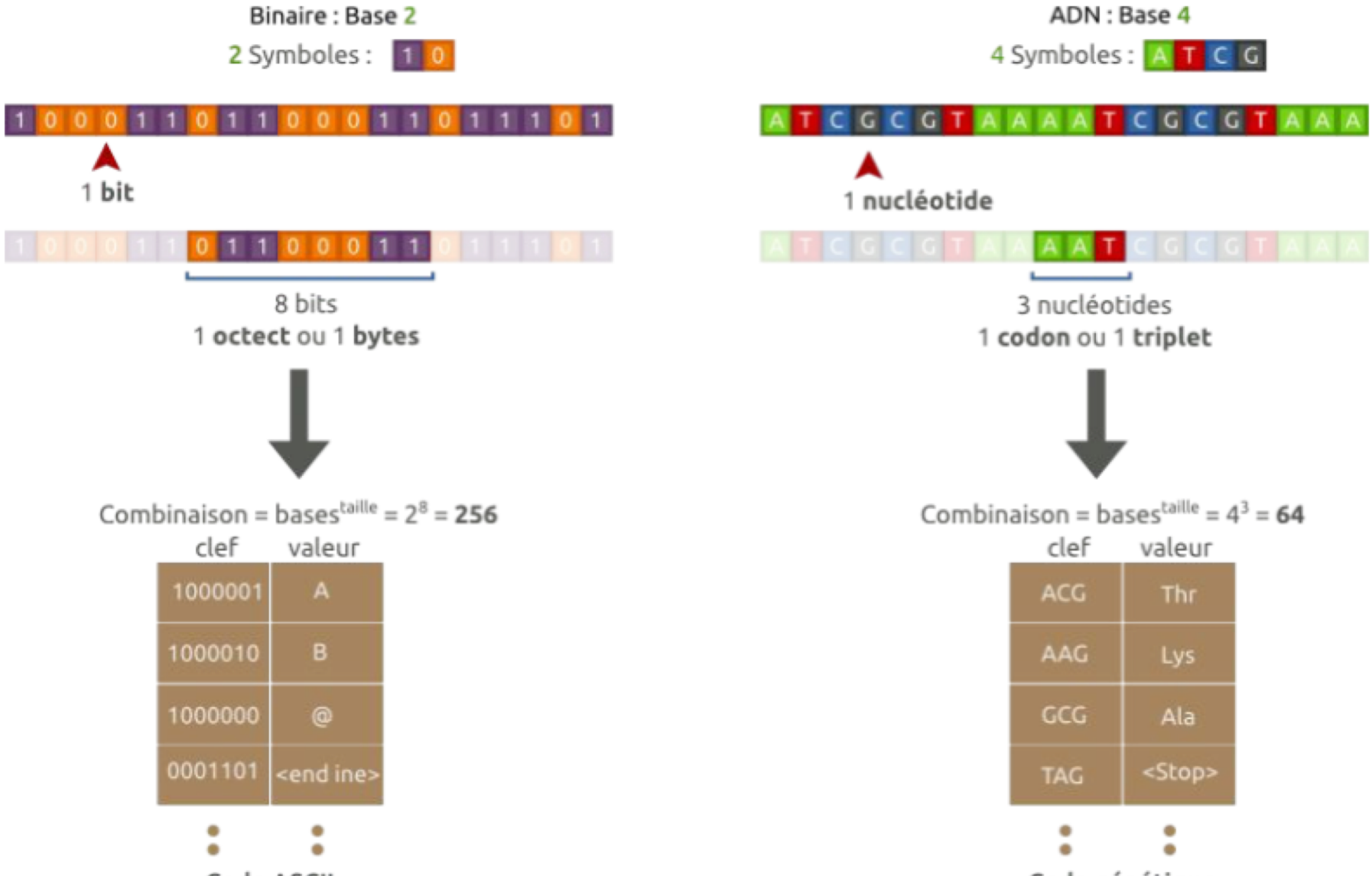
2018

With slides from **Sacha Schutz**
https://github.com/dridk

UNIVERSITÉ PARIS SUD

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

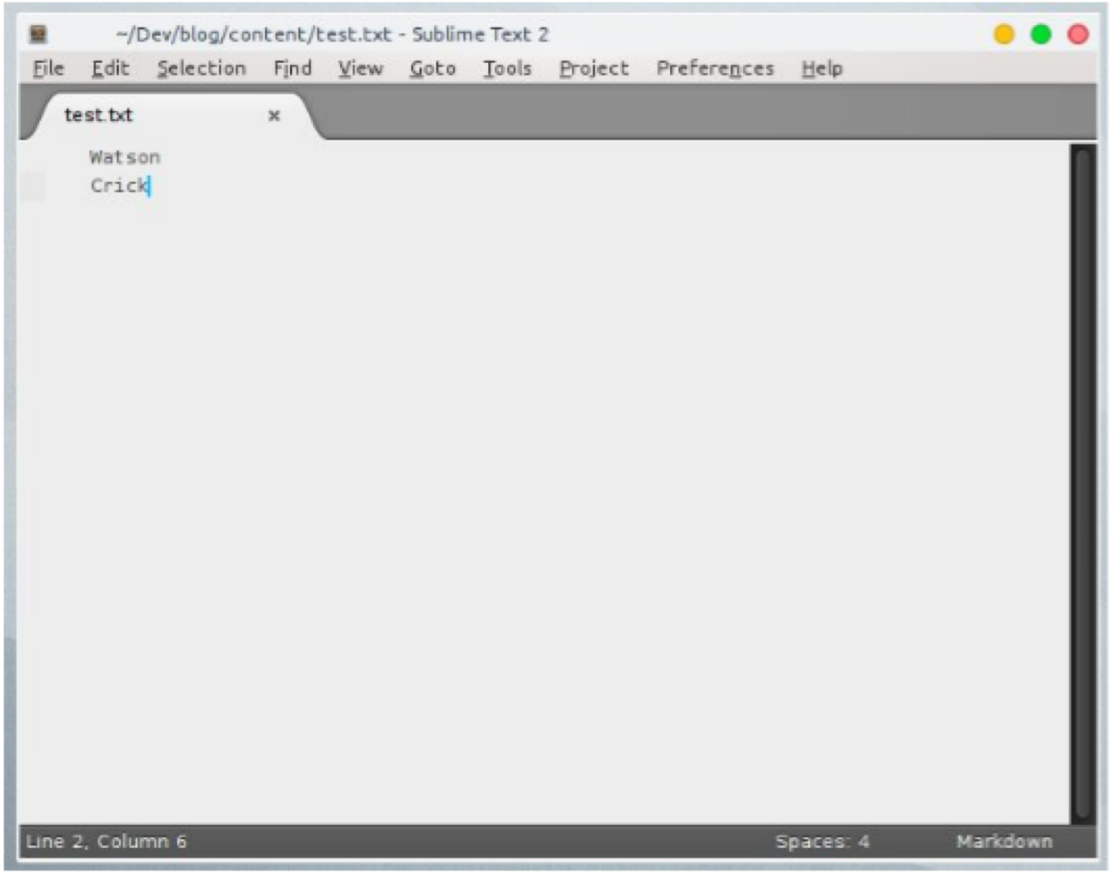
Biologie=science de l'information?



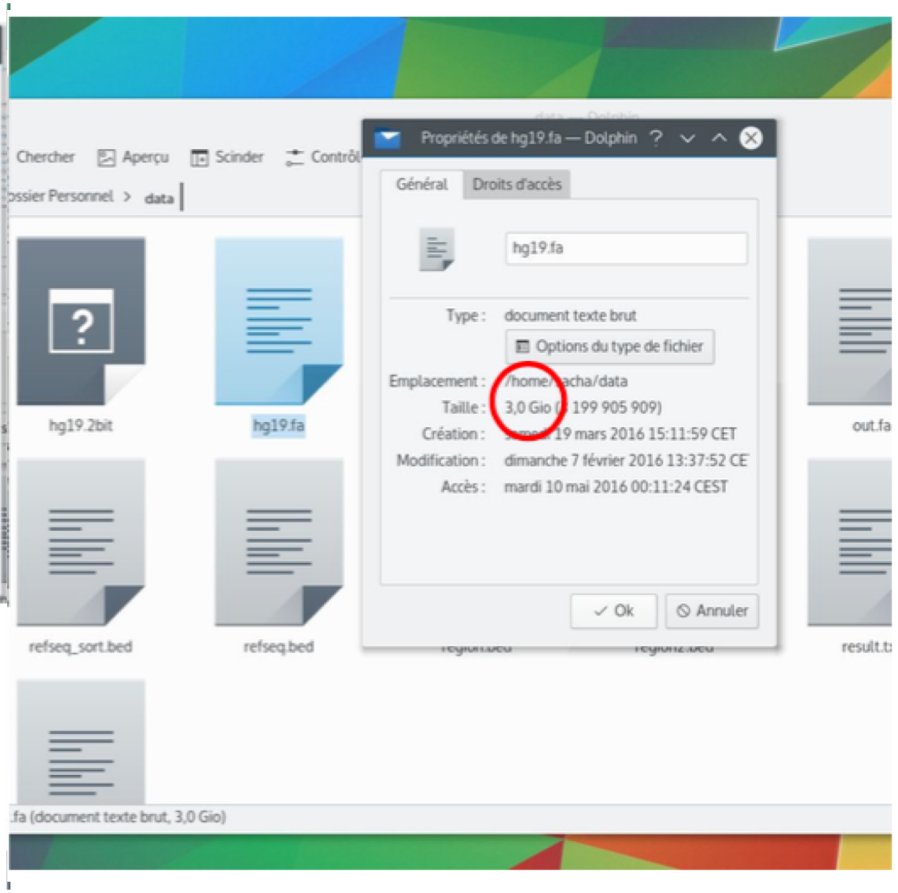
Taille d’un fichier

12 octets (bytes)=11+1

96 bits (12x8)



Taille d’un fichier

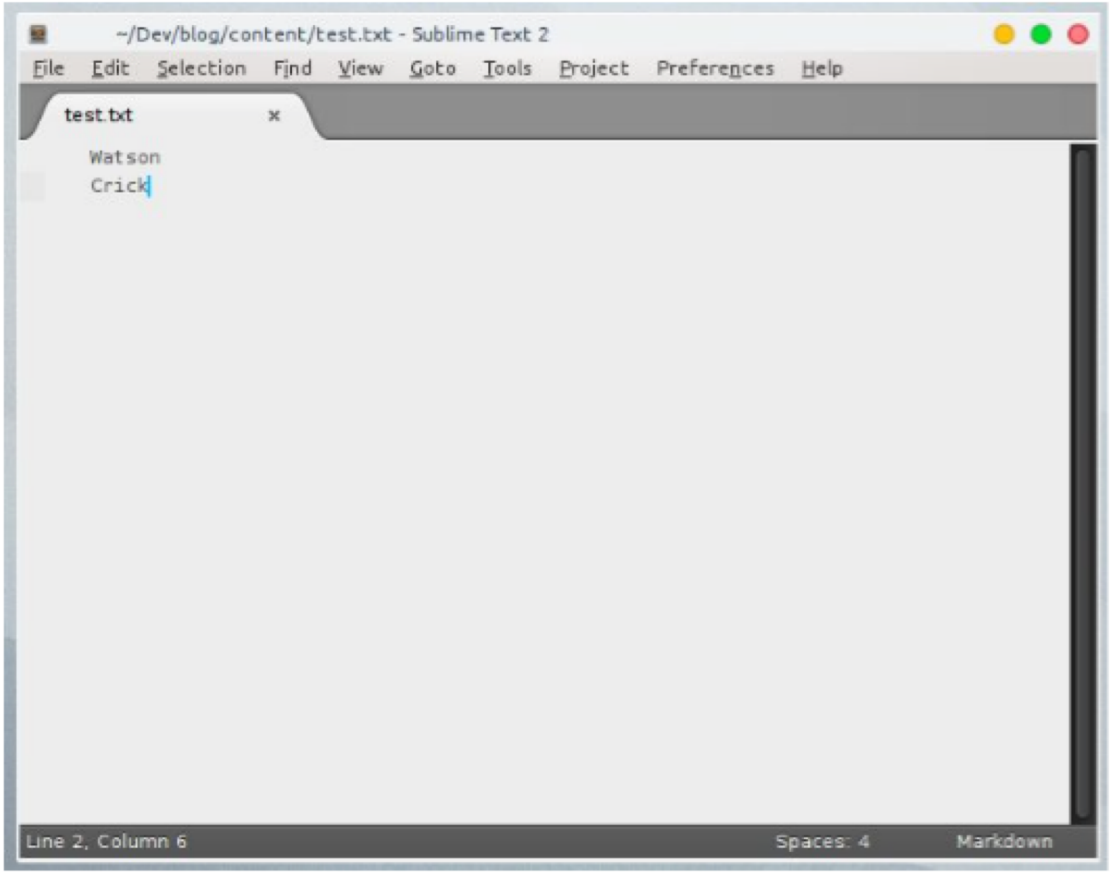


Génome humain version hg19

Taille d’un fichier

Taille du fichier?

- en octets?
- En bits?



Fichiers texte vs. binaires

Fichier texte

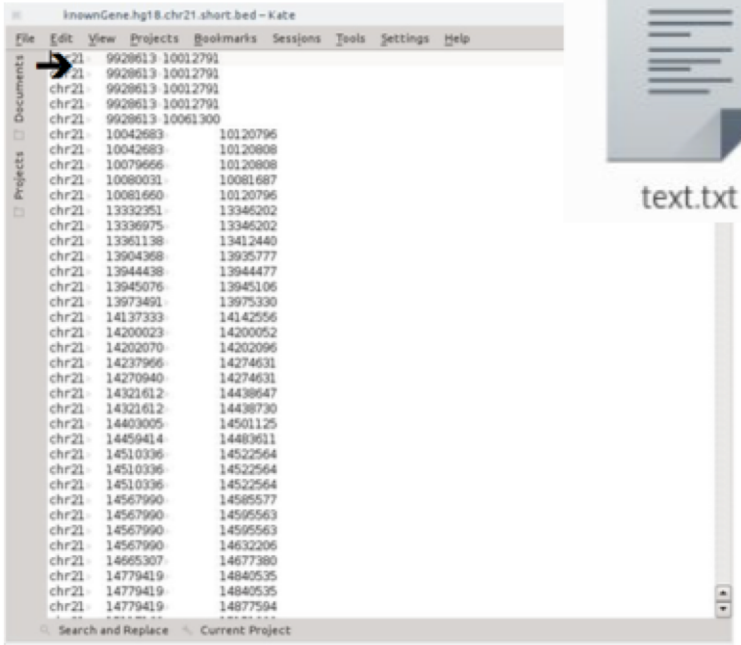
- Unité: 1 octet (byte)
- Lisible par un humain
- Dans un éditeur de texte
- Prend beaucoup d'espace
- Exemple:
  - Fasta, sam, csv, xml, vcf, fastq

Fichier binaire

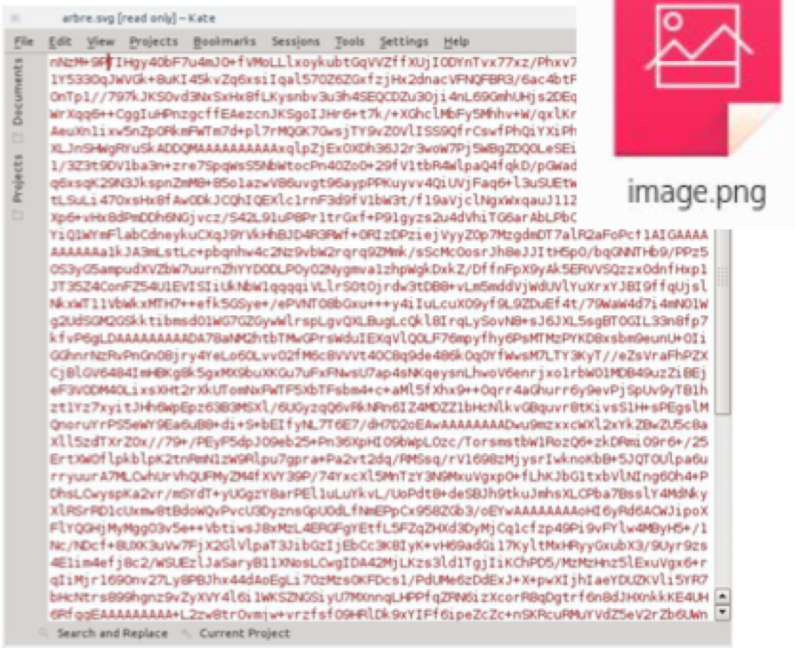
- Unité: 1 bit
- Non lisible par un humain
- Dans un éditeur hexadécimal
- Prend moins d'espace
- Exemple:
  - png, jpg, mp3, raw, bam, xlsx, docx, fastq.gz

Exemples

Fichier texte



Fichier binaire



Espace texte vs. binaire

Exemple de fichier texte pour information Vrai/Faux (V/F):

VFVFFVFF

Total: 8 octets

Exemple de fichier texte pour information Vrai/Faux (1/0):

10100100

Total: 1 octet

Formats de données en bioinformatique

- La majorité des données de bioinfo sont de type texte:
  - FASTA, FASTQ, SAM, VCF, BED, GFF, GTF, TSV, CSV, WML, JSON
- Pour des raisons de performance et d'espace, certains sont en format binaire
  - BAM, VCF.GZ, FASTQ.GZ

Format et spécification

- Le format d'un fichier décrit comment les données sont représentées
- Cette description est fournie dans un document appelé **spécification**.

Même donnée, différents formats

```
users : {
first_name: “James”, last_name:
“Watson”, birthday: “1928-04-06”
}
```

Format **JSON**
https://tools.ietf.org/html/rfc4627

```
<users>
<first_name>James</firstname>
<last_name> Watson</last_name>
<birthday>19280406</birthday>
</users>
```

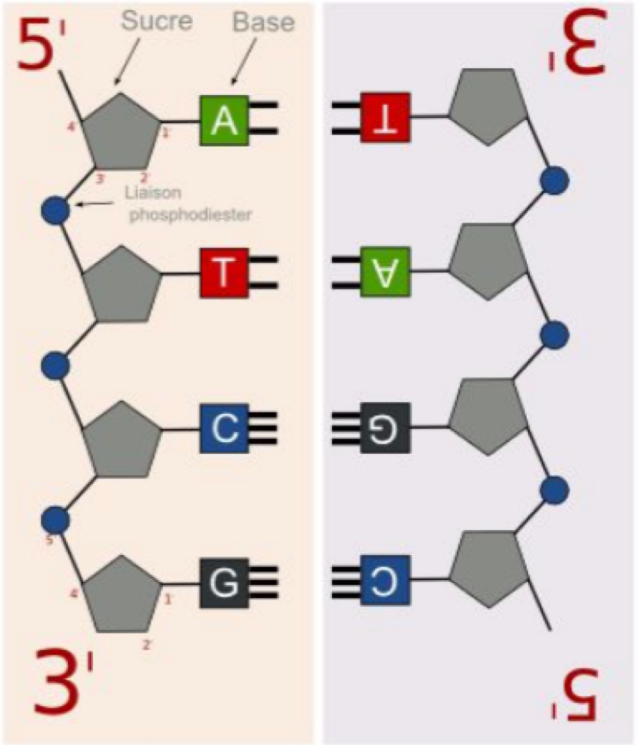
Format **XML**
https://www.w3.org/TR/REC-xml/

Séquences et régions

- En génomique on peut catégoriser les formats en:
  - Formats décrivant des **séquences**
  - Formats décrivant des **régions**

Séquences d'ADN

- Toujours dans le sens 5'->3'
- Sur quel brin?



Format fasta

*.fa , *.fasta

```
>identifiant1 commentaire libre
CAGCATCGATCGTCGGCGATGCATGCGGATGCTAGCTGATCACGATGC
CGCATGCTAGTCAGGCGGGAGGGATATTATTAGCGGCGTATCGGATGA
CAGCATTTACGGCGGGAGTGCTATTATTTATGAGCGGCGAT
>identifiant2 commentaire libre
CAGGCGGAGGTTCTTTATTATATCGGCGGGCGGAGGCGGCGATGCATC
CAGTGCACTGACGTAGTCAGCGATGCATTTTATGACTGACTCAGTTTT
CCCGCTAGCTATGCTATGCTATTCGATCGATTCGTGAGCTGATCTGGC
CAGCTATGCTTTAGTA
```

Format fastq

Descriptif du read (position sur la piste de séquençage, taille,..)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Qualité (probabilité que la base soit correcte) encodé par code ASCII

Qualité dans le format fastq

Qualité=-10 $\log_{10}(P_{erreur})$

```
@SEQ_ID1
GATTTGGGGT
+
42,40,134,47,36 35 40 40 40
```

```
@SEQ_ID1
GATTTGGGGT
+
*(å/$#((((
```

0		32		64	@	96	`	128	Ç	160	á	192	└	224	ó
1	☺	33	!	65	A	97	a	129	ü	161	í	193	┴	225	ß
2	☻	34	"	66	B	98	b	130	é	162	ó	194	┬	226	ô
3	♥	35	#	67	C	99	c	131	â	163	ú	195	├	227	ò
4	♦	36	$	68	D	100	d	132	ä	164	ñ	196	─	228	õ
5	♣	37	%	69	E	101	e	133	à	165	Ñ	197	┼	229	õ
6	♠	38	&	70	F	102	f	134	å	166	ª	198	ã	230	µ
		39	'	71	G	103	g	135	ç	167	º	199	Ã	231	þ
		40	(	72	H	104	h	136	ê	168	¿	200	╚	232	Þ
		41	)	73	I	105	i	137	ë	169	®	201	╔	233	Ú
		42	*	74	J	106	j	138	è	170	¬	202	╩	234	Û
11	♂	43	+	75	K	107	k	139	ï	171	½	203	╦	235	Ù
12	♀	44	,	76	L	108	l	140	î	172	¼	204	╠	236	ý
13		45	-	77	M	109	m	141	ì	173	¡	205	═	237	Ý
14	♫	46	.	78	N	110	n	142	Ä	174	«	206	╬	238	¯
15	☼	47	/	79	O	111	o	143	Å	175	»	207	¤	239	´
16	►	48	0	80	P	112	p	144	É	176	░	208	ð	240	
17	◄	49	1	81	Q	113	q	145	æ	177	▒	209	Ð	241	±
18	↕	50	2	82	R	114	r	146	Æ	178	▓	210	Ê	242	‗
19	‼	51	3	83	S	115	s	147	ô	179	│	211	Ë	243	¾
20	¶	52	4	84	T	116	t	148	ö	180	┤	212	È	244	¶
21	§	53	5	85	U	117	u	149	ò	181	Á	213	ı	245	§
22	▬	54	6	86	V	118	v	150	û	182	Â	214	Í	246	÷
23		55	7	87	W	119	w	151	ù	183	À	215	Î	247	¸
24	↑	56	8	88	X	120	x	152	ÿ	184	©	216	Ï	248	°
25	↓	57	9	89	Y	121	y	153	Ö	185	╣	217	┘	249	¨
26	→	58	:	90	Z	122	z	154	Ü	186	║	218	┌	250	·
27	←	59	;	91	[	123	{	155	ø	187	╗	219	█	251	¹
28	∟	60	<	92	\	124	¦	156	£	188	╝	220	▄	252	³
29	↔	61	=	93	]	125	}	157	Ø	189	¢	221	¦	253	²
30	▲	62	>	94	^	126	~	158	×	190	¥	222	Ì	254	■
31	▼	63	?	95	_	127	⌂	159	ƒ	191	┐	223	▀	255	

Format Genbank

```
LOCUS       L10986                 47233 bp    DNA     linear   INV 21-SEP-2004
DEFINITION  Caenorhabditis elegans cosmid F10E9, complete sequence.
ACCESSION   L10986
VERSION     L10986.2  GI:38638818
KEYWORDS    HTG.
SOURCE      Caenorhabditis elegans
  ORGANISM  Caenorhabditis elegans
            Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida;
            Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis.
REFERENCE   1  (bases 1 to 47233)
  AUTHORS   .
  CONSRTM   WormBase Consortium
  TITLE     Genome sequence of the nematode C. elegans: a platform f
            investigating biology. The C. elegans Sequencing Consort
  JOURNAL   Science 282 (5396), 2012-2018 (1998)
  MEDLINE   99069613
   PUBMED   9851916
FEATURES             Location/Qualifiers
     source          1..47233
                     /organism="Caenorhabditis elegans"
                     /mol_type="genomic DNA"
                     /strain="Bristol N2"
                     /db_xref="taxon:6239"
                     /chromosome="III"
                     /clone="F10E9"
     gene            265..26728
                     /gene="mig-10"
                     /locus_tag="F10E9.6"
     CDS             join(265..338,3266..3515,15194..15317,21507..21
                     21727..21887,23171..23335,24302..24472,24524..24608,
                     25012..25827,26284..26430,26478..26728)
                     /gene="mig-10"
 /translation="MDSCEEECDLEVDSDEEDQLFGEKCISLLSSLLPLSSSTLLSNA
                     INLELDEVERPPPLLNVLEEQQFPKVCANIEEENELEADTEEDIAETADDEESKDPVE
                     KTENFEPSVTMDTYDFPDPYPVQIRARPQVPPKPPIDTVRYSMNNIKESADWQLDELL
                     EELEALETQLNSSNGGDQLLLGVSGIPASSSRENVKSISTLPPPPPALSYHQTPQQPQ
. . .
                     QVYTGIGWEKKYKSPTPWCISIKLTALQMKRSQFIKYICAEDEMTFKKWLVALRIAKN
                     GAELLENYERACQIRRETLGPASSMSAASSSTAISEVPHSLSHHQRTPSVASSIQLSS
                     HMMNNPTHPLSVNVRNQSPASFSVNSCQQSHPSRTSAKLEIQYDEQPTGTIKRAPLDV
                     LRRVSRASTSSPTIPQEESDSDEEFPAPPPVASVMRMPPPVTPPKPCTPLTSKKAPPP
                     PPKRSDTTKLQSASPMAPAKNDLEAALARRREKMATMEC"
. . .
```

```
BASE COUNT     2598 a   2024 c   1888 g   2449 t
ORIGIN
        1 ttctaaaagt cgaaaaacga gcaatttttg atgctagatt ttttgatttg acgaattttt
       61 tcagtttttt ttctttaaaa aaggtttttg accccttaaa gttttccttt cccttccaat
      121 tttttccttc tttcttatac gacttctcaa gtttcaactc taaaacaaag ctacatgtac
      181 atttccggta aactttgtgt ctcagaagat ccatttttctt tttgttacat ttattcaaga
      241 ttgaattcca aaatttcagc caatatggac agttgcgaag aggaatgcga tctggaagtt
      301 gacagtgacg aagaagatca acttttggt gaaaagtggt gagttcttat tgtggtaacc
      361 aaagaaatgt cagtggtccg taaacacttg actcccaaat ggtttctcgt aattacctta
      421 tgcacacttt tcaagtgttt gccgtttgat cttagccaat ttgaaacgtt tagatgttaa
      481 atggaaaatg ggtaaagttt tttattttat agaaaaaagg tttggaaaaa aatcgagtca
      541 ctgaatagtt tgaagaacgg aaaaataaaa ctttccaaaa atcataaaac atttagtgtt
      601 tcgaaaatta tagtgttttt tttgttggta tgttttgaca aaagctaaac catctttatt
      661 gtagttttgt aaaatgttca caaagatgcg ttttttttc aaatttggca ggctatcttt
      721 acattcacat ttggataatt caaatttttc ttatcgctaa caaatttcc tatttttcca
      781 attattcgtt tttataaagc tttggtagta tgttgtgtct atctttagtg gtcatcagtt
. . .
//
```

Format Genbank

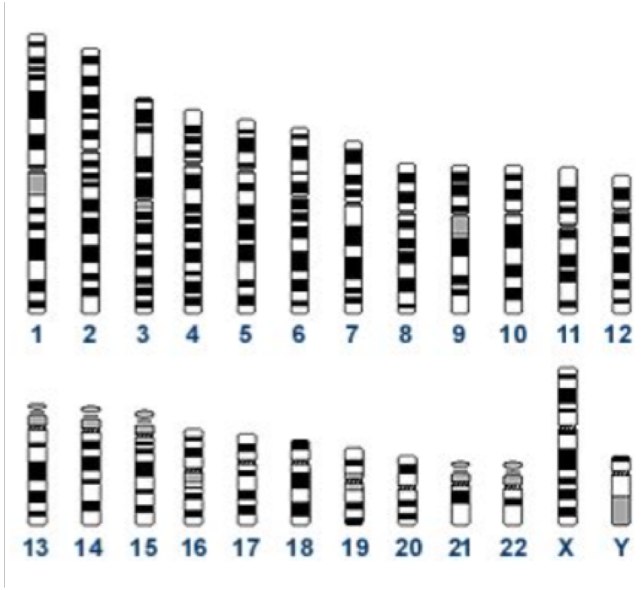
Cherchez sur le site du NCBI le mRNA du gène *GJB2 dans la* section **nucleotides**

- *Sur quel chromosome est le gène?*
- *Combien d'exons comporte le gène?*
- *Quelle est la séquence des 4 premiers nucléotides du 1er intron?*

Les régions

Les coordonnées génomiques permettent de définir une région exacte du génome

<**chromosome**>:<**start**>-<**end**>
`chr7:117465784-117715971`

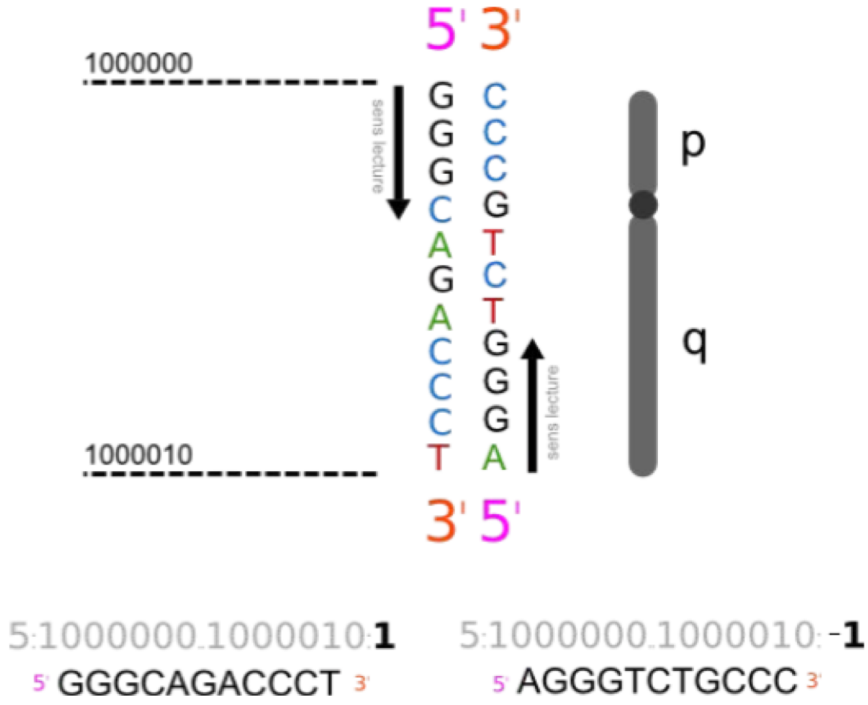


Accéder directement à une région

Par exemple par l'interface programmatique (API) de la banque Ensembl:

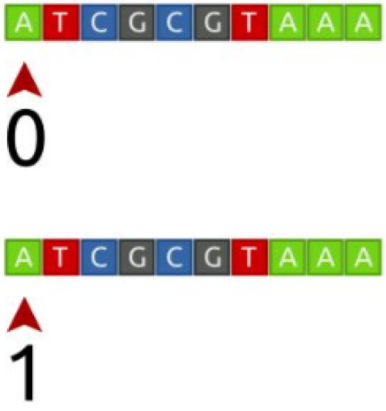
`http://rest.ensembl.org/sequence/region/human/7:117465784..117715971:-1`

Attention aux versions d'assemblage du génome (Hg19, Hg38..)



Formats de fichiers utilisant les régions

- Attention: suivant le format la première base peut être numérotée 0 ou 1



Format/library	Type
BED	0-based
GTF	1-based
GFF	1-based
SAM	1-based
BAM	0-based
VCF	1-based
BCF	0-based
Wiggle	1-based
GenomicRanges	1-based
BLAST	1-based
GenBank/EMBL Feature Table	1-based

Format bed

obligatoire

name score strand Thick start Thick end color

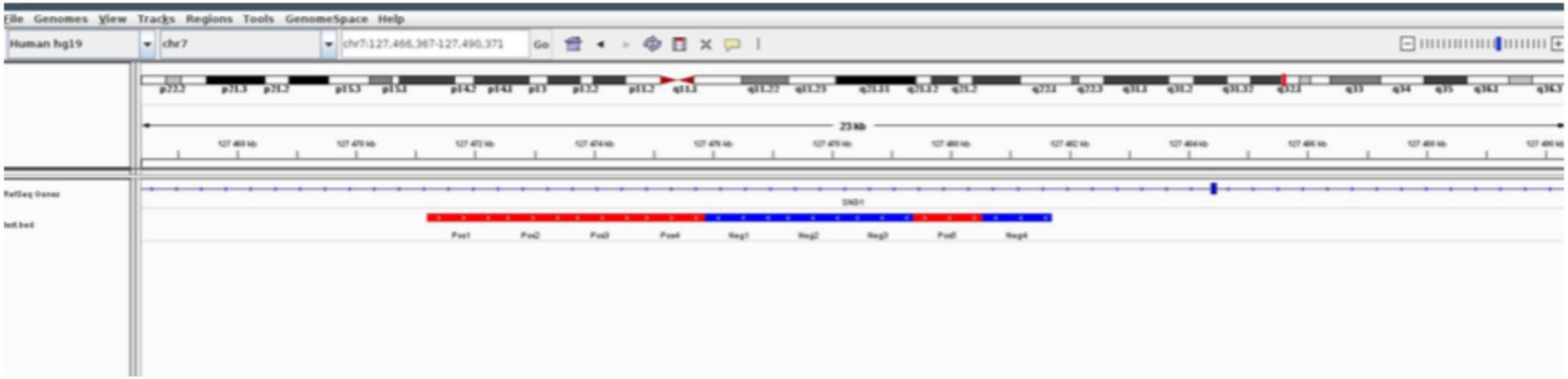
```
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 0,255,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,255
```

Attention

Le premier nucléotide est numéroté 0.

end - start = taille de la séquence

Le format bed est lisible par les navigateurs de génome



IGV genome browser

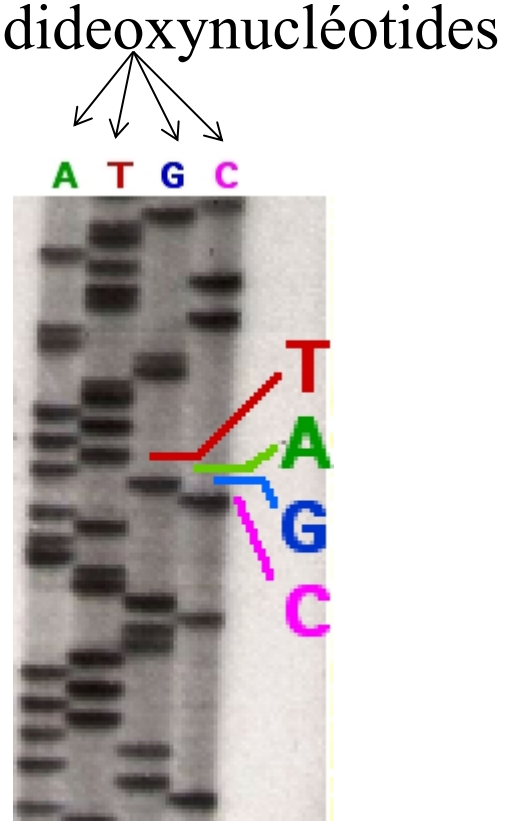
Autres formats de régions

- Nous verrons par la suite:
  - GTF/GFF: annotation de features dans le génome
  - SAM/BAM: alignement de reads de séquence sur le génome
  - VCF: variant calling file

Bioinformatique et NGS

Le séquençage de Sanger (1977)

- Séquençage par terminaison de chaine
  - Synthèse interrompue à un certain type de base.
  - 4 types= 4 réactions
- 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 we could sequence the genome in 30 years. »

Walter Gilbert 1980

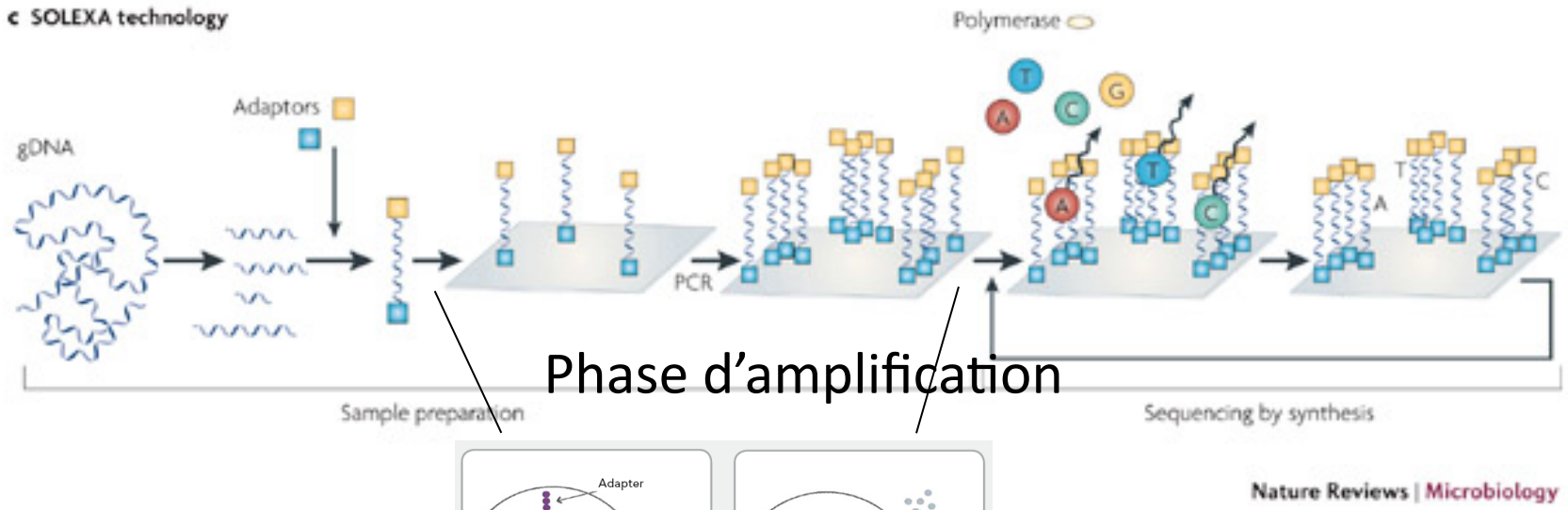
- Project started in 1991 and completed in 2001.
- …using Sanger sequencing

NGS : Next Generation Sequencing (2005-)

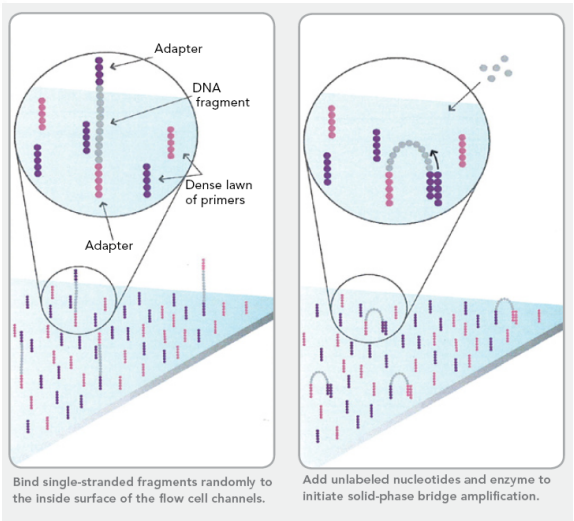
Faster, Cheaper, Deeper

NGS: parallélisation du séquençage

Solexa/Illumina



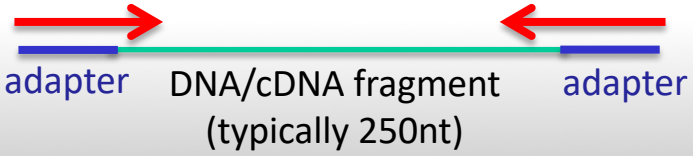
Phase d'amplification



Sequencers & output

	Illumina Hi-Seq 2500	Illumina NextSeq 500	Illumina MySeq	Lifetech Ion torrent PGM	Lifetech Ion proton
Read number	16x300M	1x400M	15M	1M (316chip)	50M
Read size	2x200	2x150	2x250	35-400	35-400

Single vs. **paired end** sequencing



long-read technologies

Instrument	Durée	Millions de Reads/run	Bases/read	Gb/Run
Applied Biosystems 3730	2h	0,000096	650	0,00006
454 GS Jr. Titanium	10h	0,1	400	0,1
454 FLX Titanium	10h	1	400	0,4
454 FLX+	20h	1	650	0,7
Illumina GA IIx v5 SE	2j	640	36	23
Illumina GA IIx v5 PE	14j	640	*288*	184,3
Illumina MiSeq v2 Nano	17h	1	*300*	0,3
Illumina MiSeq v2 Micro	19h	4	*300*	1,2
Illumina MiSeq v3	20h	22	*150*	3,3
Illumina MiSeq v3	55h	22	*600*	13,2
Illumina NextSeq 500 Mid	15h	130	*150*	19,5
Illumina NextSeq 500 High	18h	400	*150*	60
Illumina HiSeq 2500 Rapid run	27h	300	*200*	60
Illumina HiSeq 2500 v3	11j	1500	*200*	300
Illumina HiSeq X (2 flow cells)	3j	6000	*300*	1800
Ion Torrent – PGM 314 chip	2,3h	0,475	200	0,1
Ion Torrent – PGM 314 chip	3,7h	0,475	400	0,2
Ion Torrent – PGM 316 chip	3h	2,5	200	0,5
Ion Torrent – PGM 316 chip	4,9h	2,5	400	1
Ion Torrent – PGM 318 chip	4,4h	4,75	200	1
Ion Torrent – PGM 318 chip	7,3h	4,75	400	1,9
Ion Torrent - Proton I	4h	70	175	12,3
Ion Torrent - Proton II	5h	280	175	49
Ion Torrent - Proton III	6h	500	175	87,5
Life Technologies SOLiD 5500xl	8j	1410	110	155
Pacific Biosciences RS II	2h	0,03	3000	0,1
Oxford Nanopore MinION	≤6h	0,1	9000	0,9

[ANPGM GT1] Recommandations informatiques pour les analyses NGS, 2016

Sequencing Output per Flow Cell

	NovaSeq 6000 System		
Flow Cell Type	S1*	S2	S4
2 × 50 bp	134-167 Gb	280–333 Gb	NA†
2 × 100 bp	266-333 Gb	560–667 Gb	NA†
2 × 150 bp	400-500 Gb	850–1000 Gb	2400-3000 Gb

=3Tb
(en 44h)

2017: Illumina NovaSeq

Les grandes applications des NGS

- DNA-seq (variants génomiques)
- 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

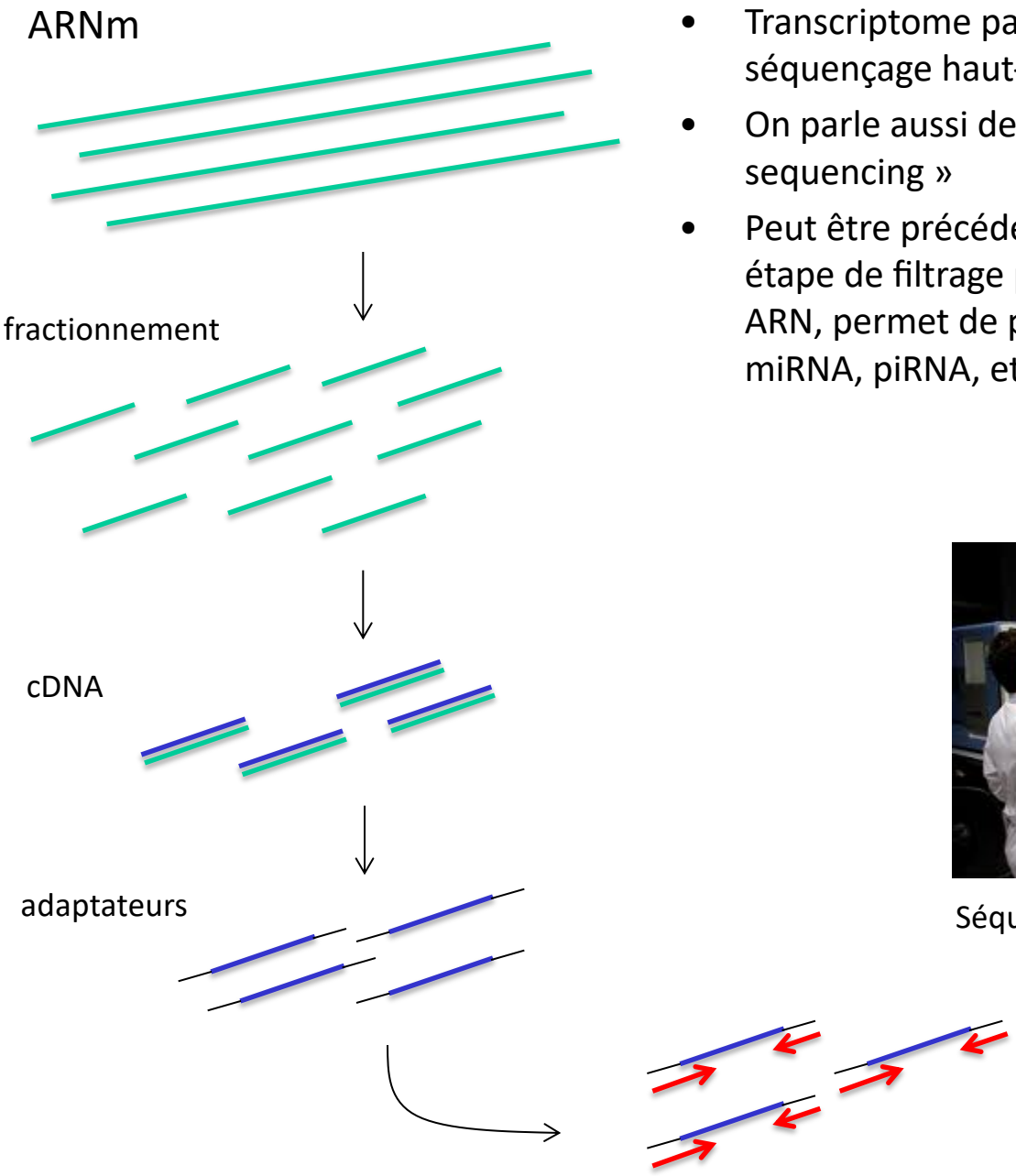
- En cancérologie, 2 grandes applications
  - Génétique constitutionnelle (recherche de prédisposition)
  - Génétique somatique (diagnostic, médecine de précision)

Séquencer quoi?

- Panel de gènes
  - Une série d'exons d'intérêt (gènes de cancer= 100kb)
- Exome (WES)
  - Tous les exons du génome (30 Mb)
- Whole genome (WGS)
  - Le génome complet (3 Gb)

RNA-seq

RNA-Seq



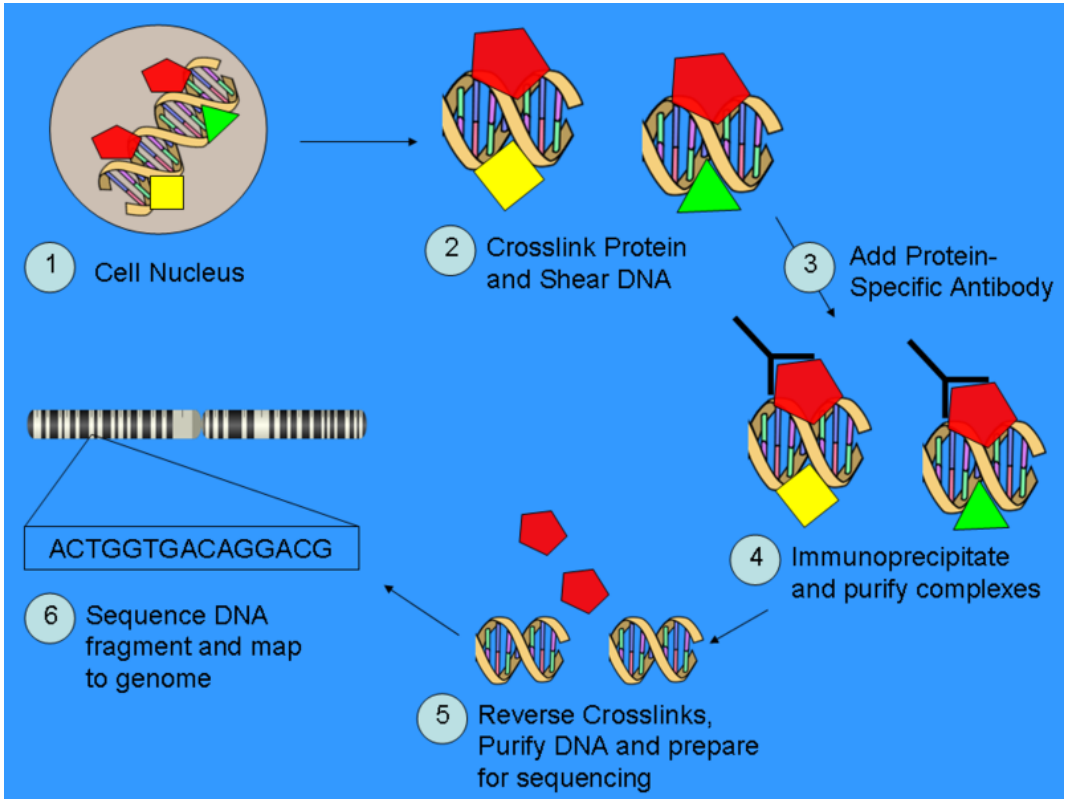
- Transcriptome par séquençage haut-débit.
- On parle aussi de « deep sequencing »
- Peut être précédé d'une étape de filtrage pour petits ARN, permet de pêcher les miRNA, piRNA, etc.

Séquençage

ChiP-Seq

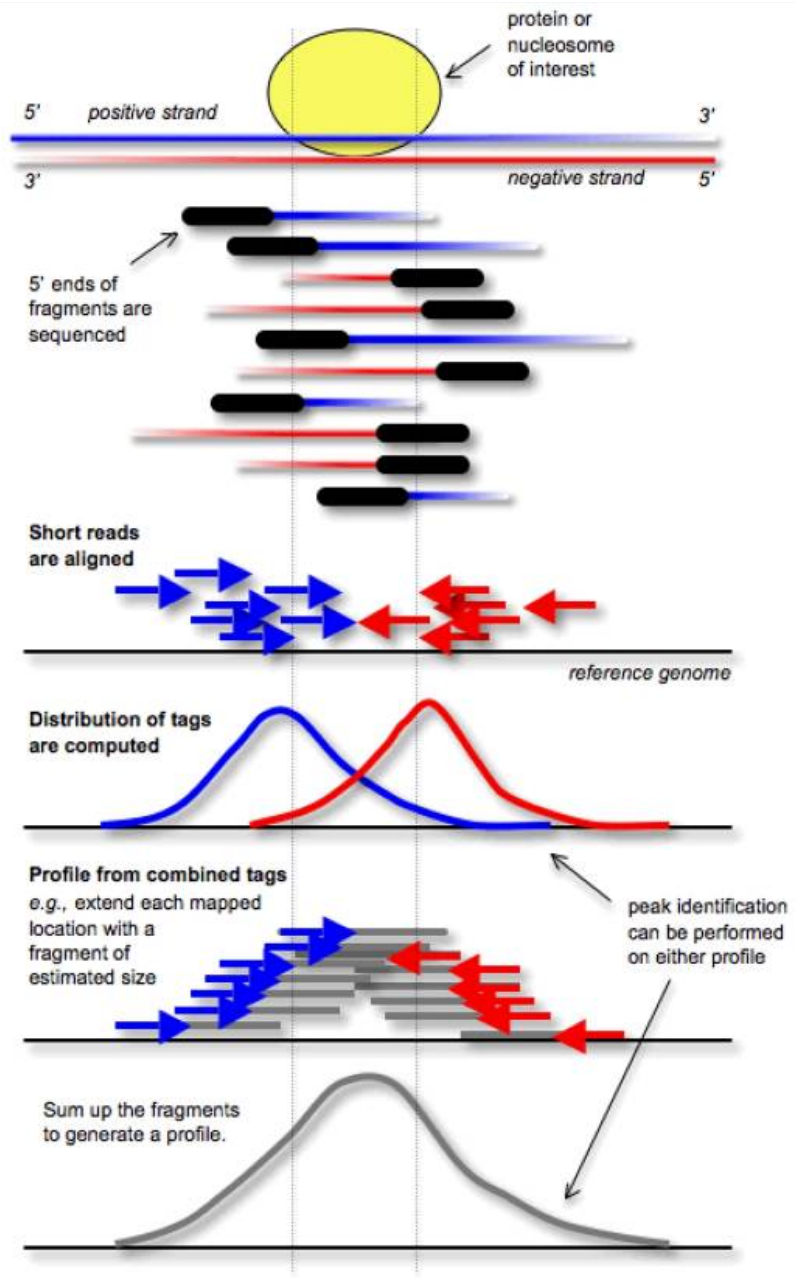
ChIP-Seq

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



Wikipedia

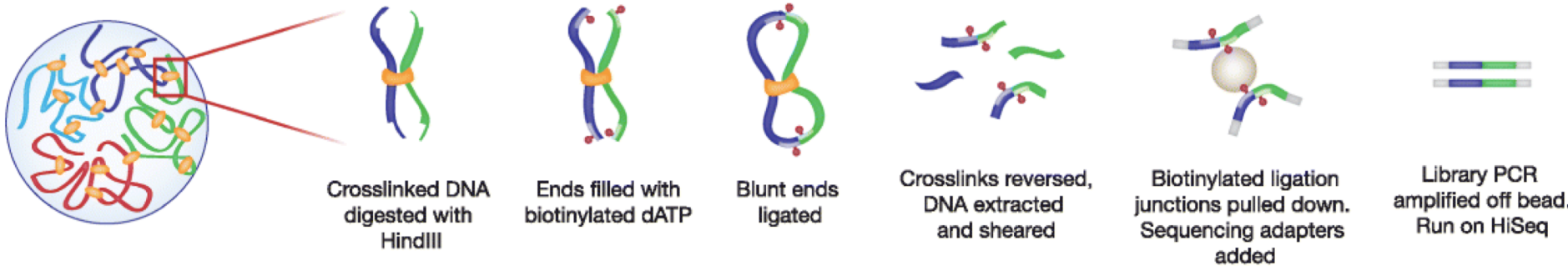
Detection de pics

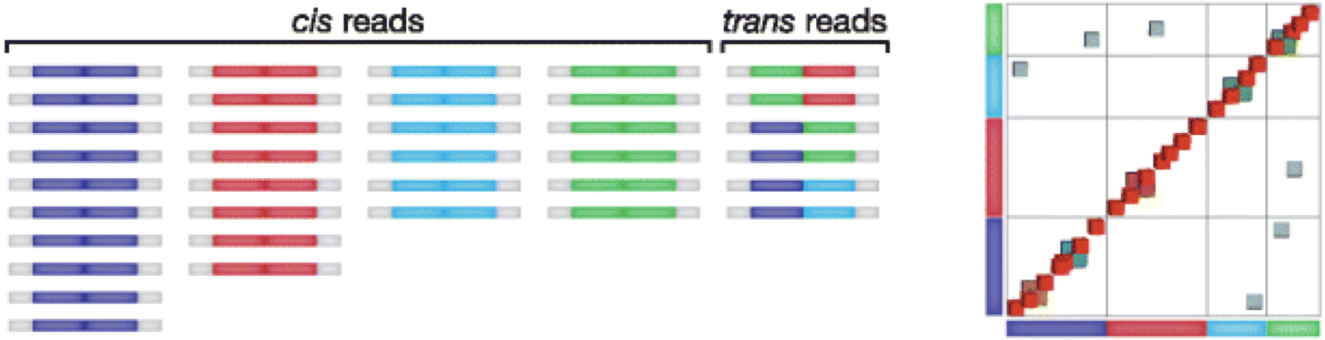


Hi-C

- Chromosome conformation capture

Hi-C





Harewood et al. Genome Biol 2017

Vers une utilisation systématique du séquençage en médecine

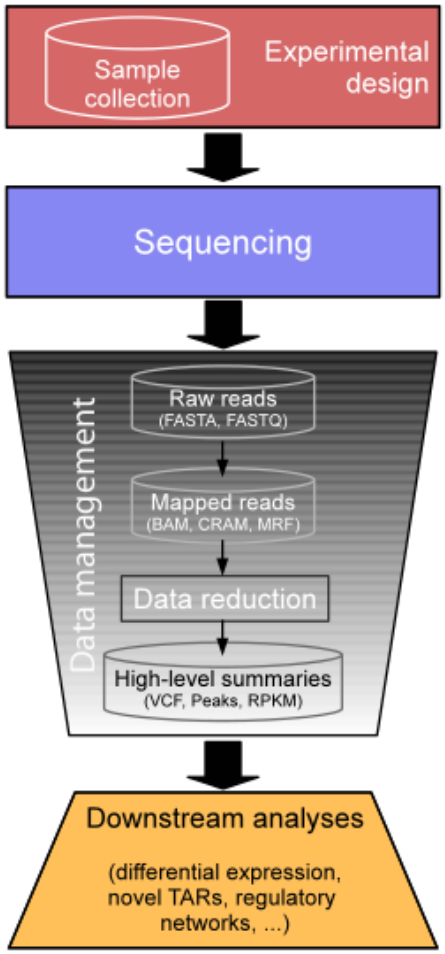
- Déjà tous les patients ont un exome complet séquençé dans certains centres de cancérologie

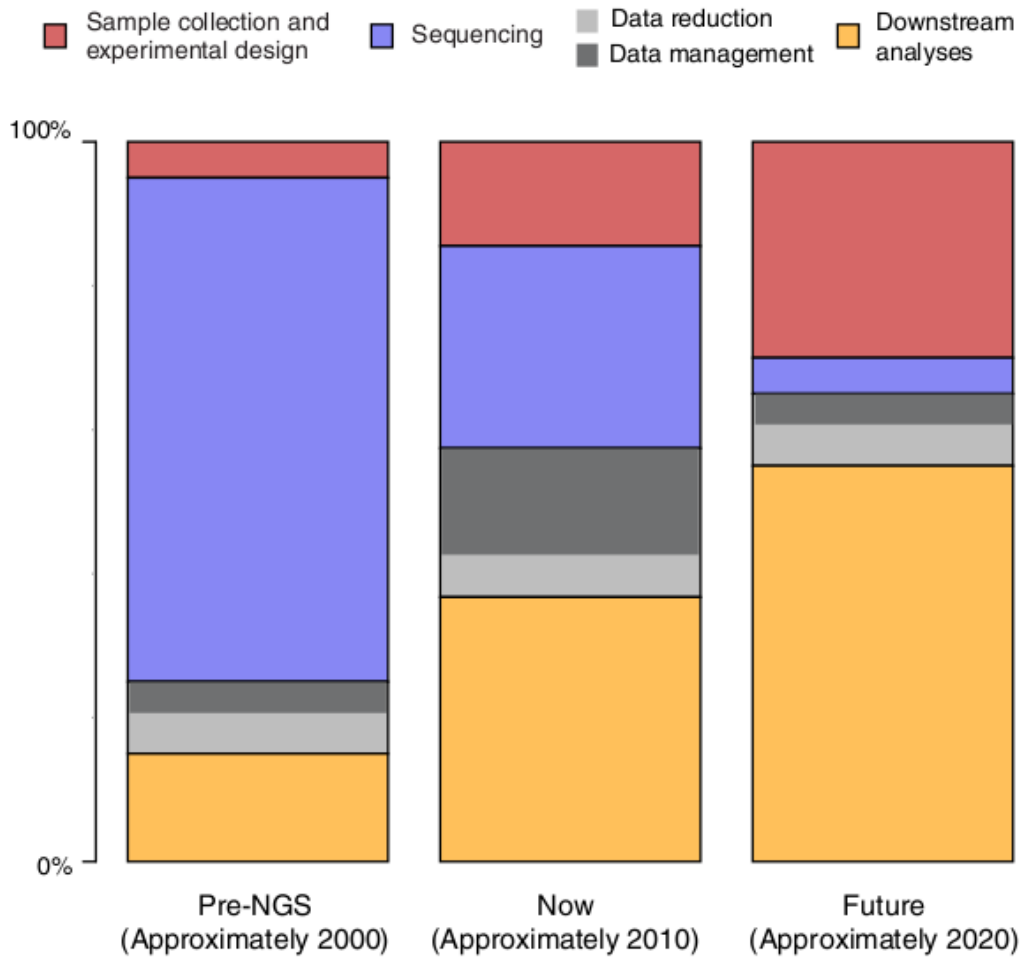
  -> généralisation de la médecine de précision
- Génome humain aujourd'hui < 1000$

Volume des données NGS

- Un exome humain (N+T) avec fichiers de mapping et analyse: 50 Go
  - (prévoir ~5 fois le volume des fastq.gz)
- Données génomiques produites annuellement dans un hôpital universitaire: 100-500 To
- La banque TCGA complete: 1 Po

Evolution of sequencing cost





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

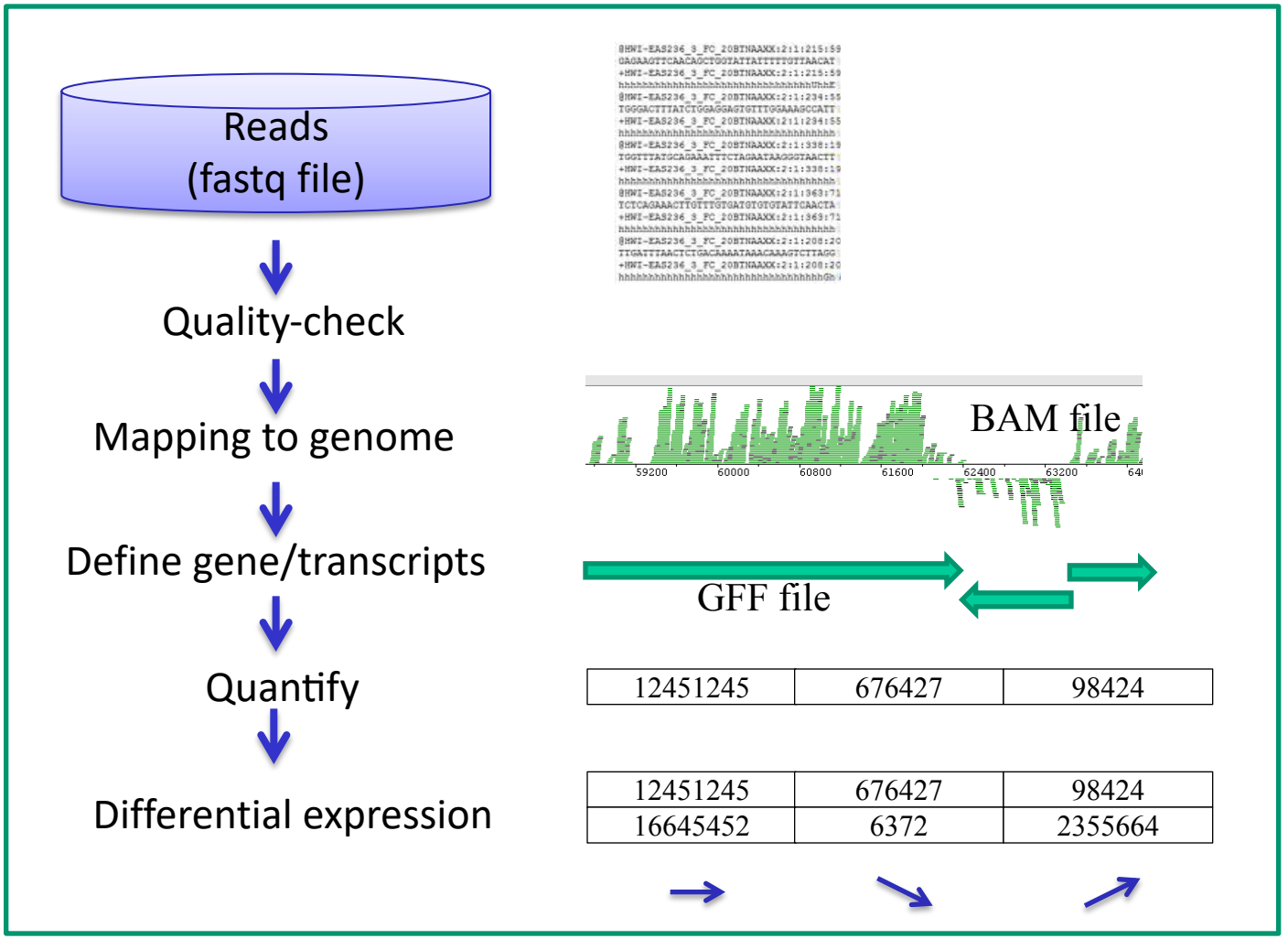
Les outils

« Pipelines » & « workflows »

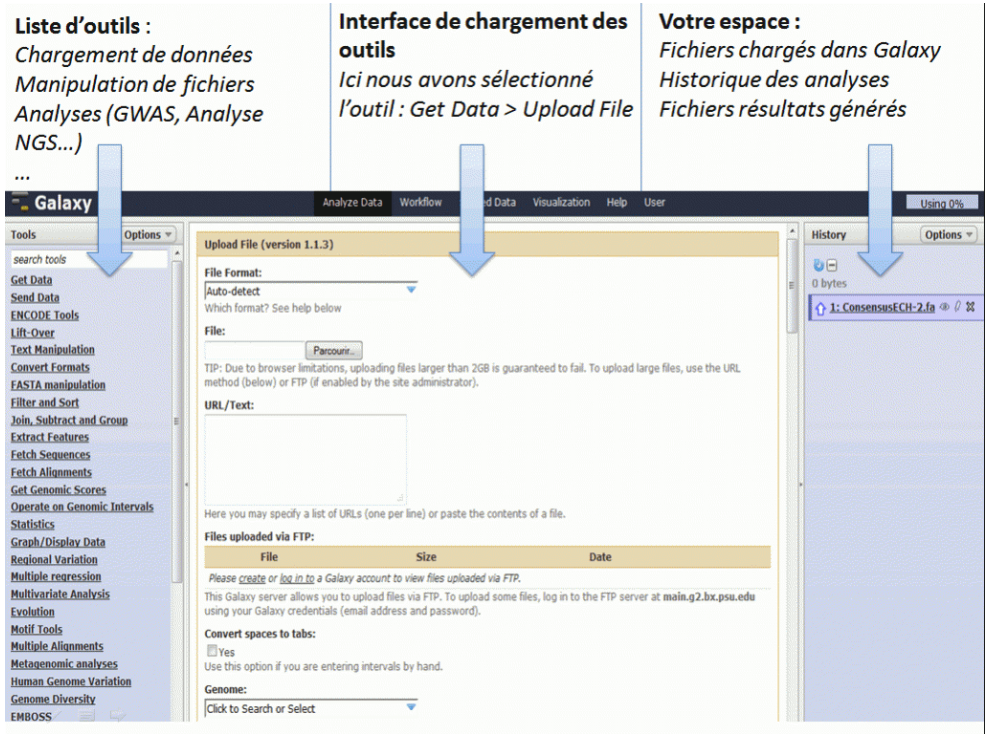
« Bricks » from Unix open source programs

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

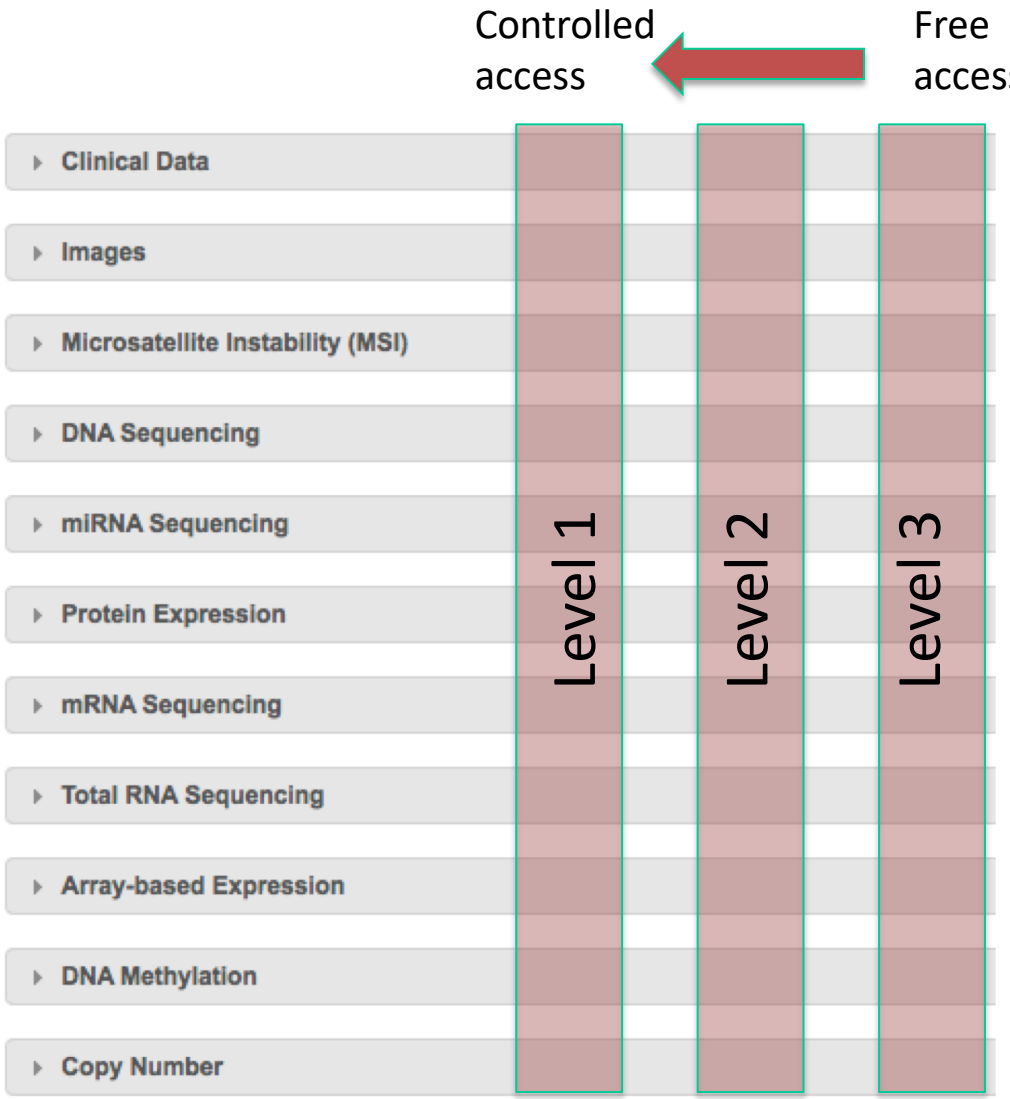
NCI, NHGRI, USA

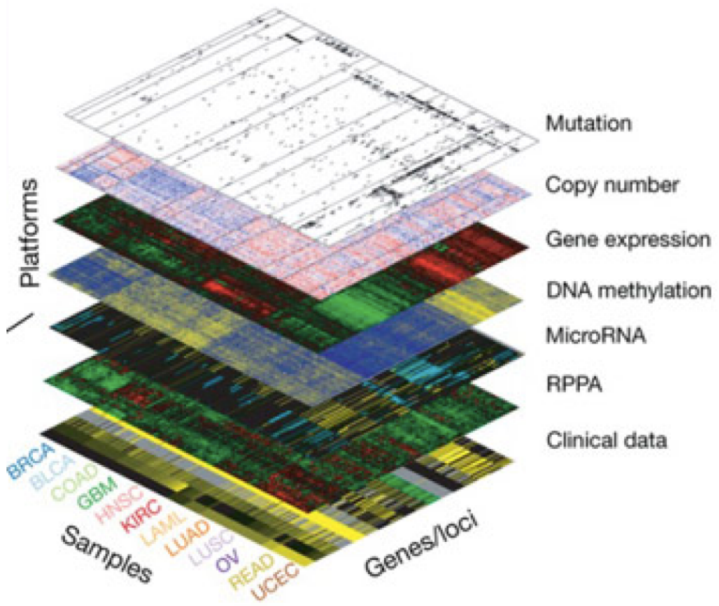
The Cancer Genome Atlas *Understanding genomics to improve cancer care*

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





Main data types

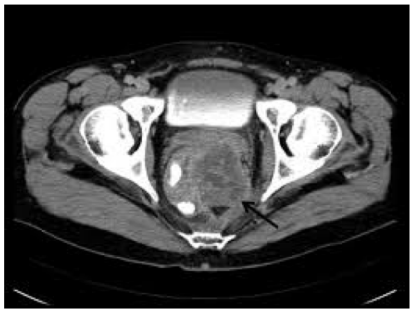
- DNA sequencing
  - Whole exome or whole genome DNA sequence
  - Platform: Illumina HiSeq
- mRNA sequencing / miRNA sequencing
  - PolyA+ RNA / small RNA expression from RNA-seq
  - Platform: Illumina HiSeq or similar
- Array-based expression
  - mRNA expression levels (1 or 2 colors)
  - Illumina or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

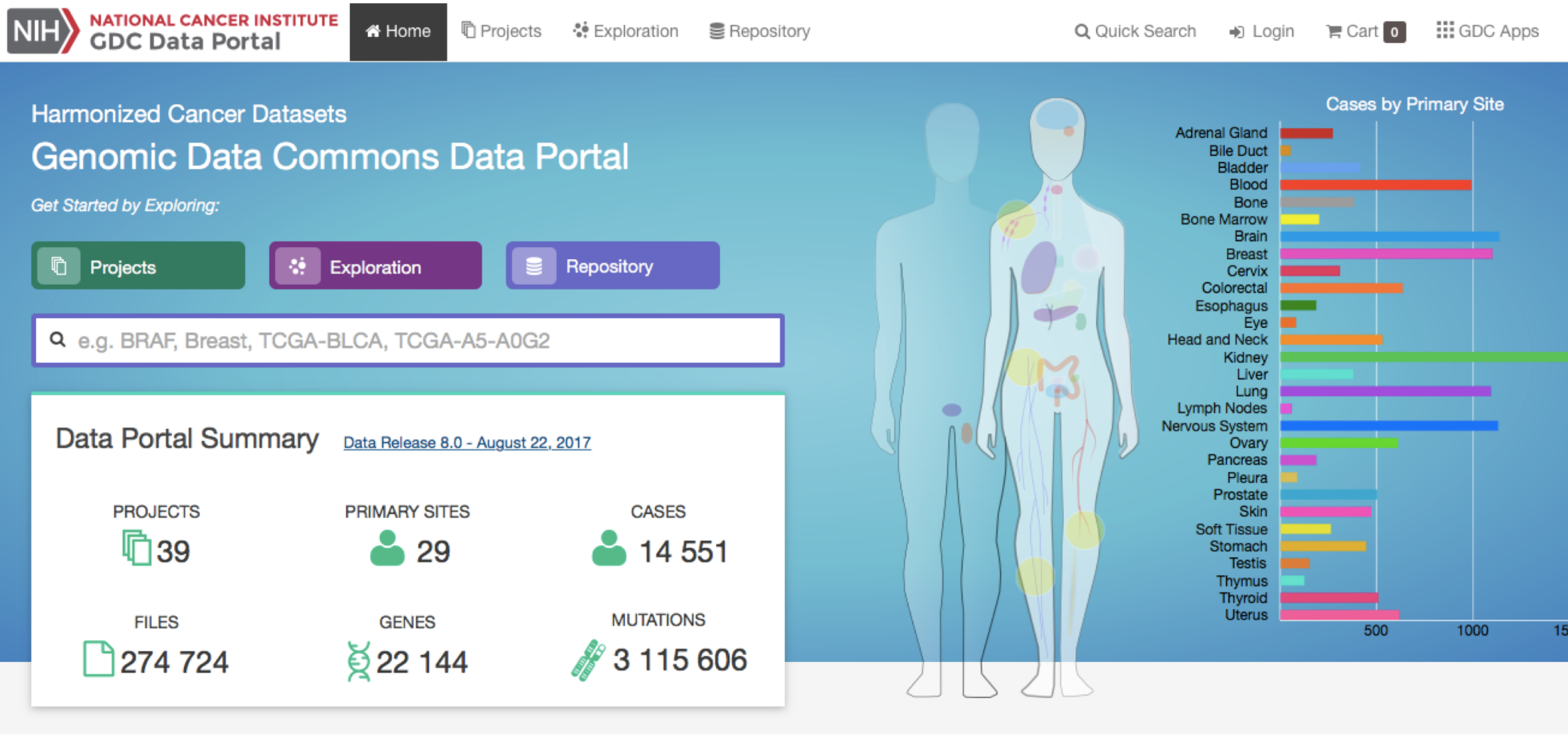
Main data types

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

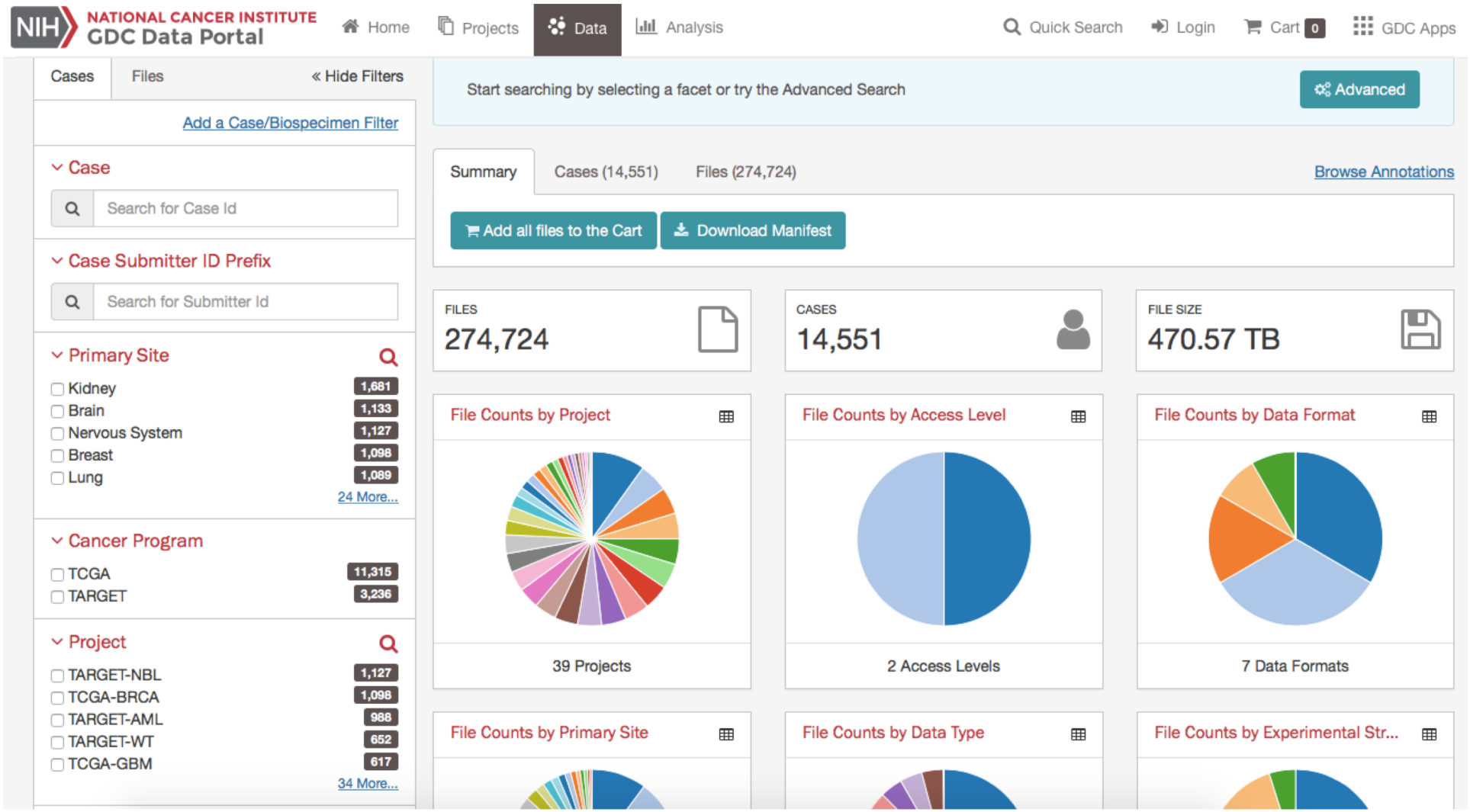


- Image
  - Images of tissue samples
  - CT (computed tomography), DX (digital radiography), CR (computed radiography)
- Clinical data
  - Available clinical information for each participant (demographic, treatment, survival, etc)
  - Biospecimen data: how specimen was processed

TCGA access via the GDC portal (Genomics Data Commons)

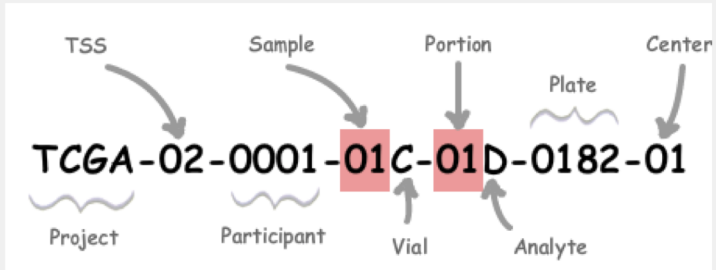


The GDC Data portal



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	



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.

1 sample
= 1 TCGA barcode

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: Next step

- PCAWG[1]: a collaboration with ICGC[2] to analyze whole genome data from 2,000 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
Catalogue of somatic mutations in cancer

- 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

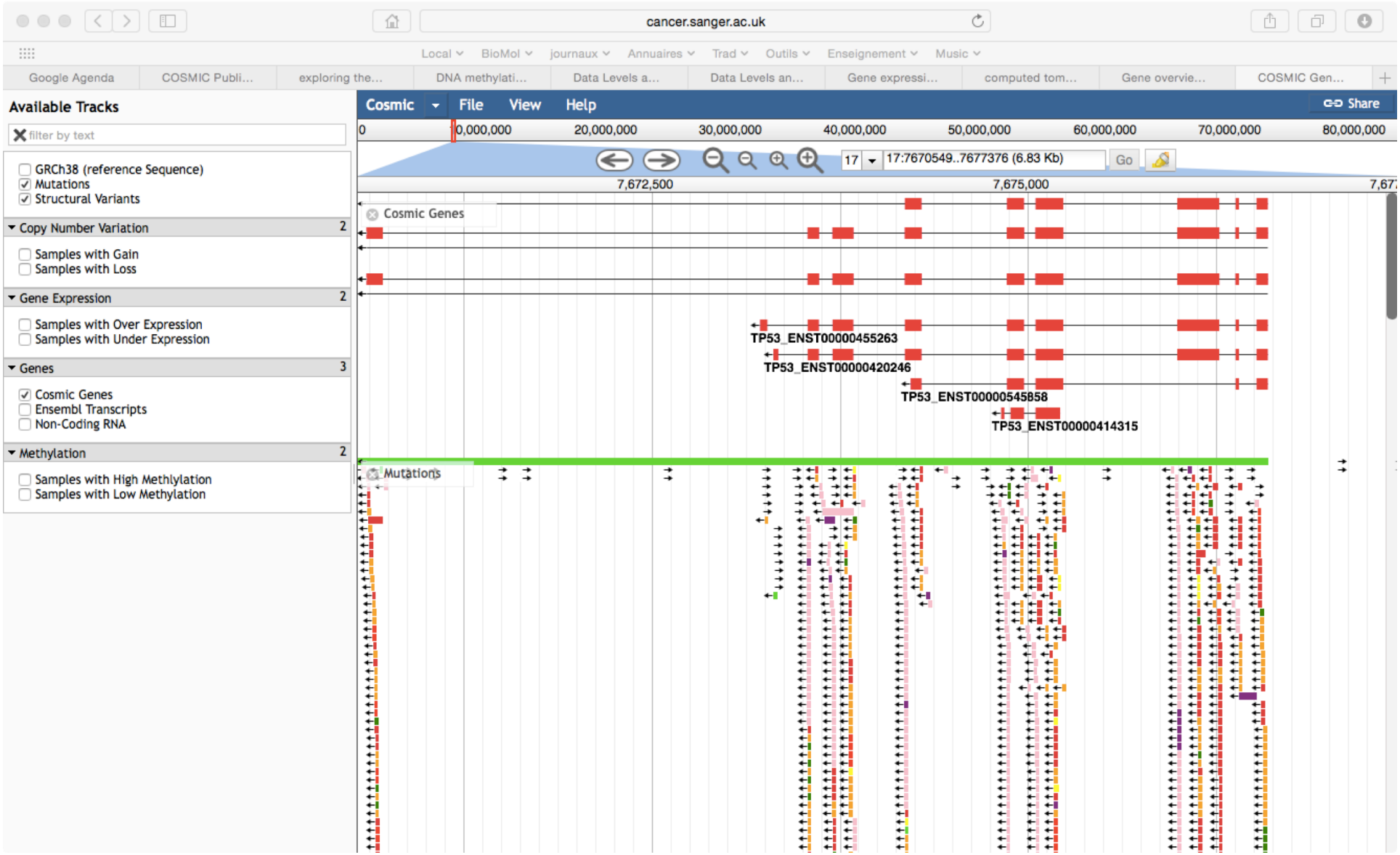
COSMIC Curation

- 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. »

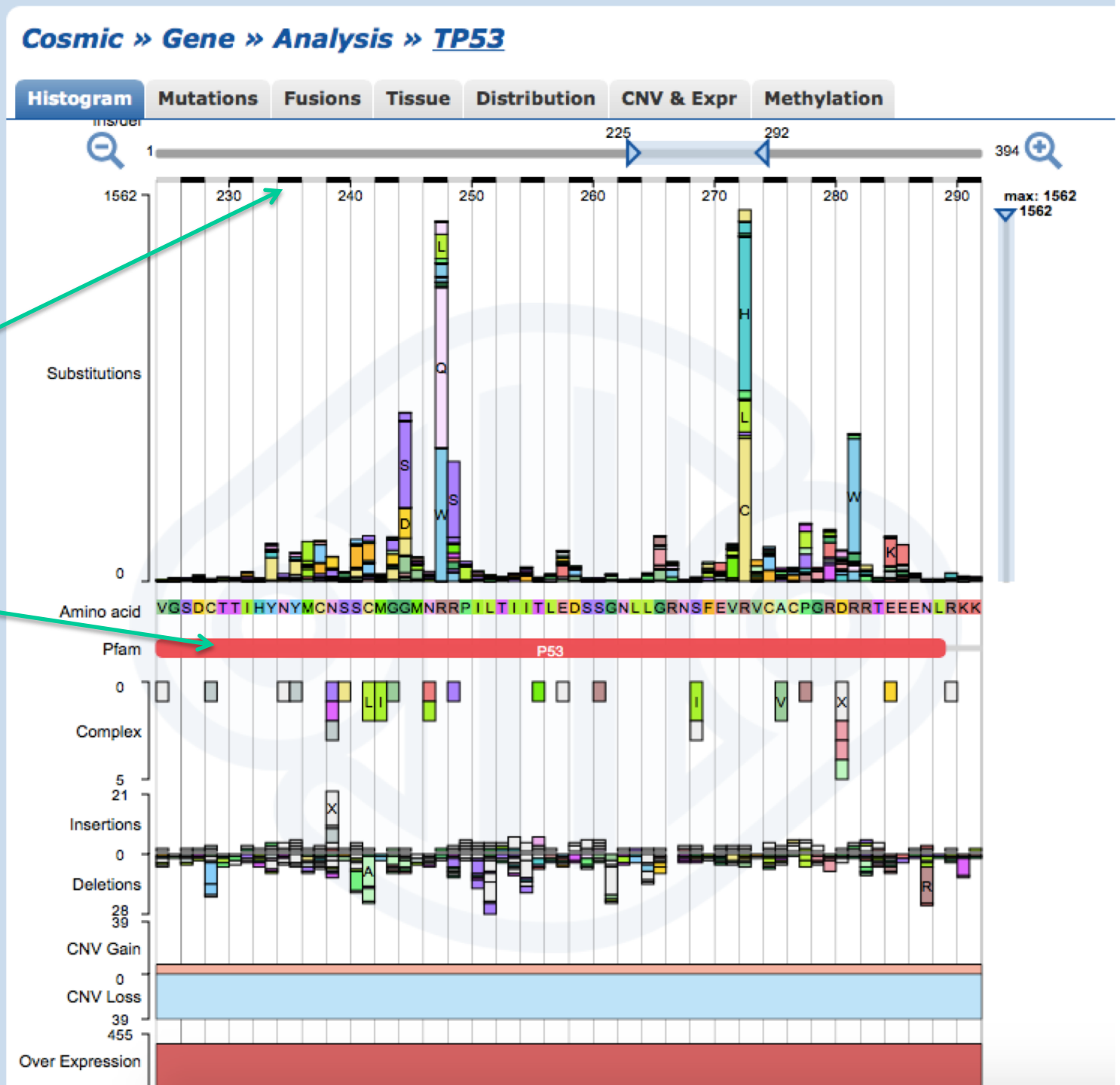
COSMIC genome browser



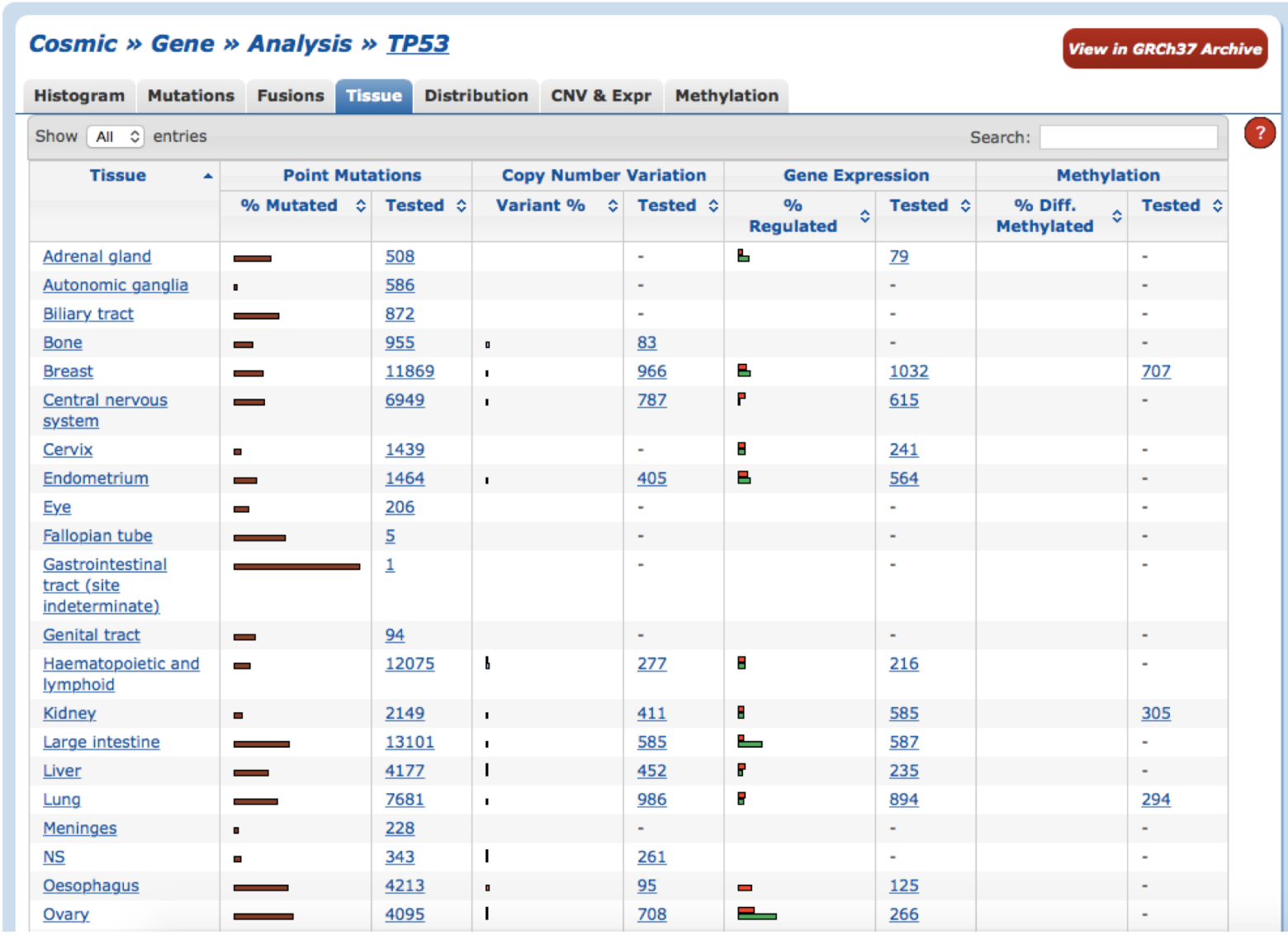
Histogram view

Protein coordinates

Protein domain



Tissue-distribution of mutations



Cosmic » Gene » Analysis » TP53

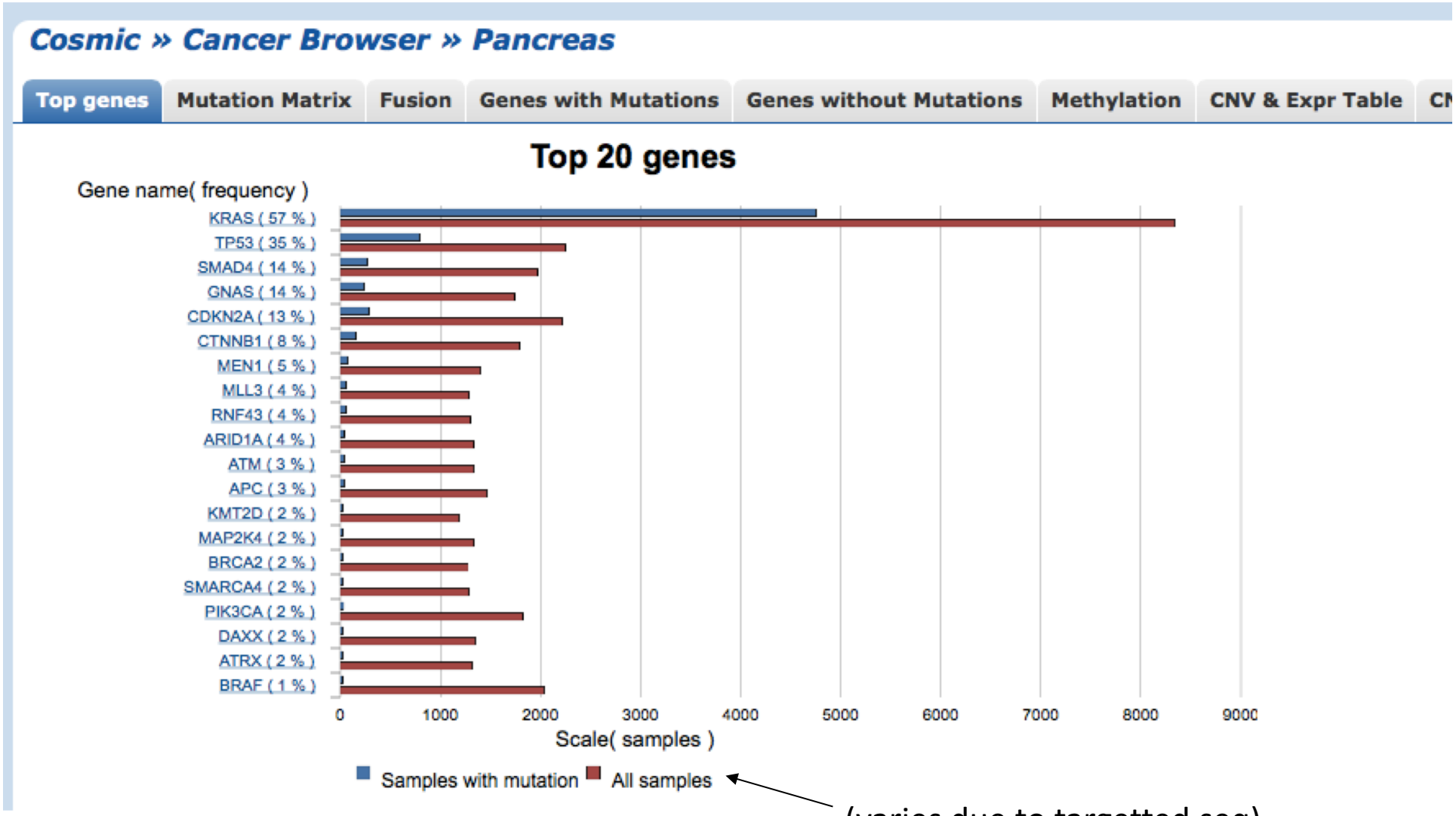
View in GRCh37 Archive

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)

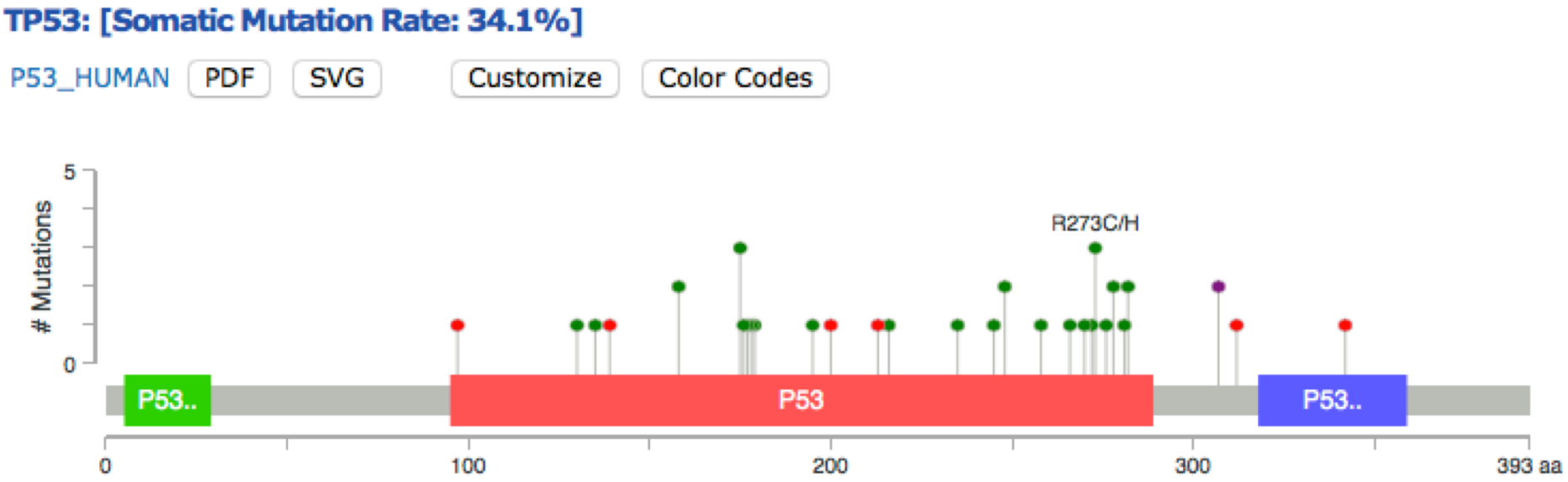
Exercice

- A l'aide du site COSMIC, comparez les profils de mutations de TP53 et BRAF.
- En quoi diffèrent-ils? Pourquoi?

Memorial Sloan-Kettering
Cancer Center, USA

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

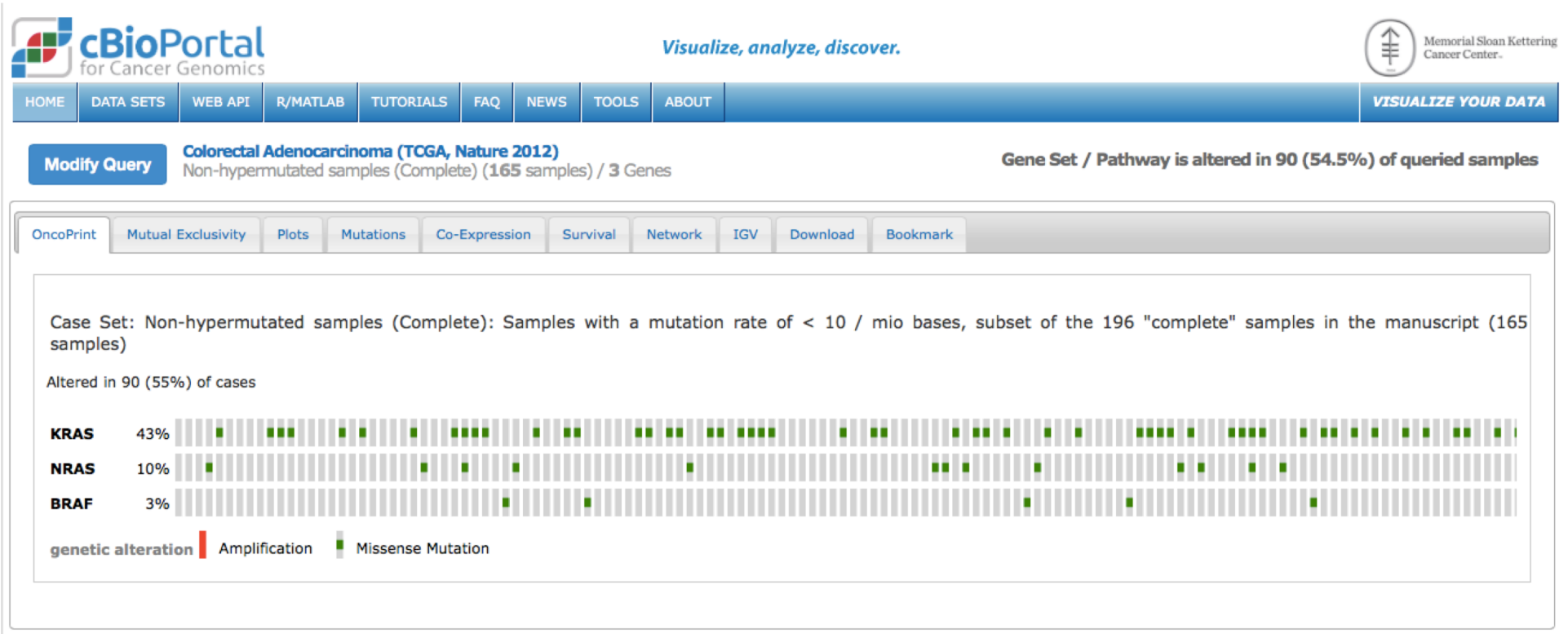
Mapped mutations on proteins



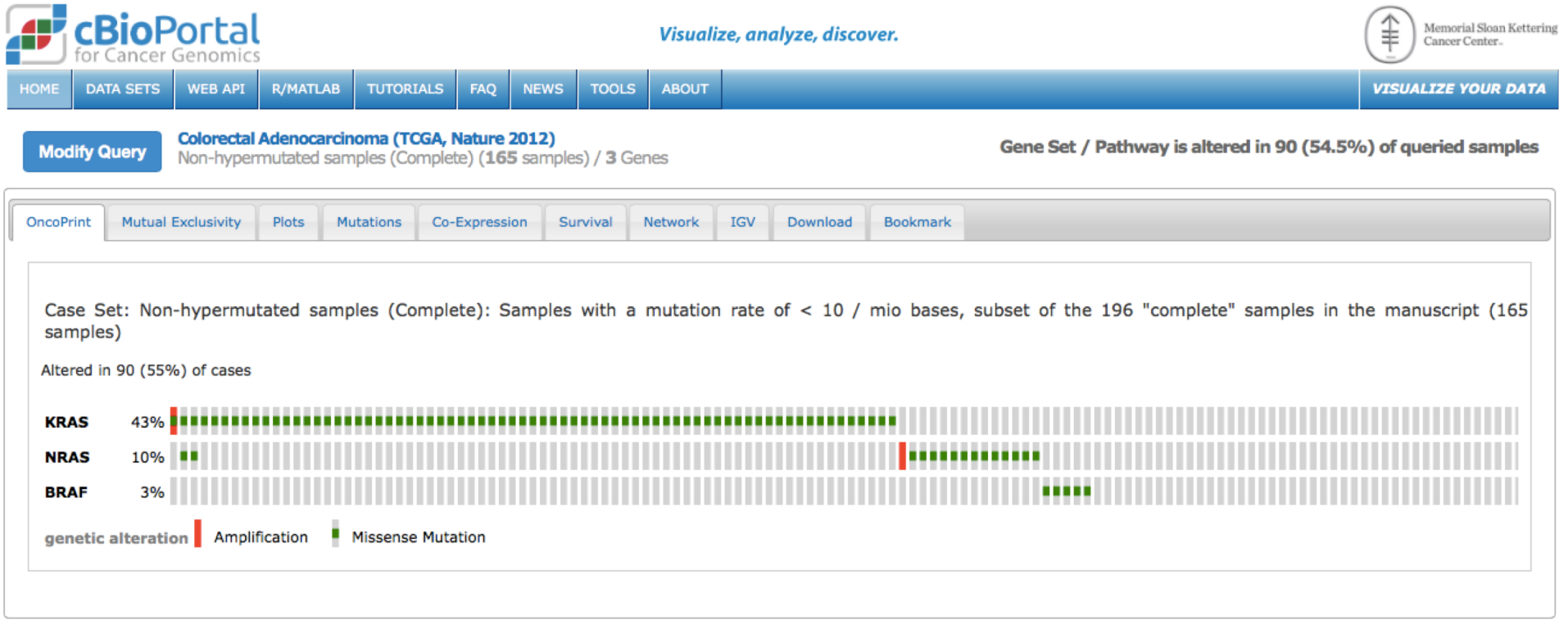
Mutations mapped on TP53 in Glioblastoma dataset (TCGA, Nature 2008)

See also « MutationMapper » tool

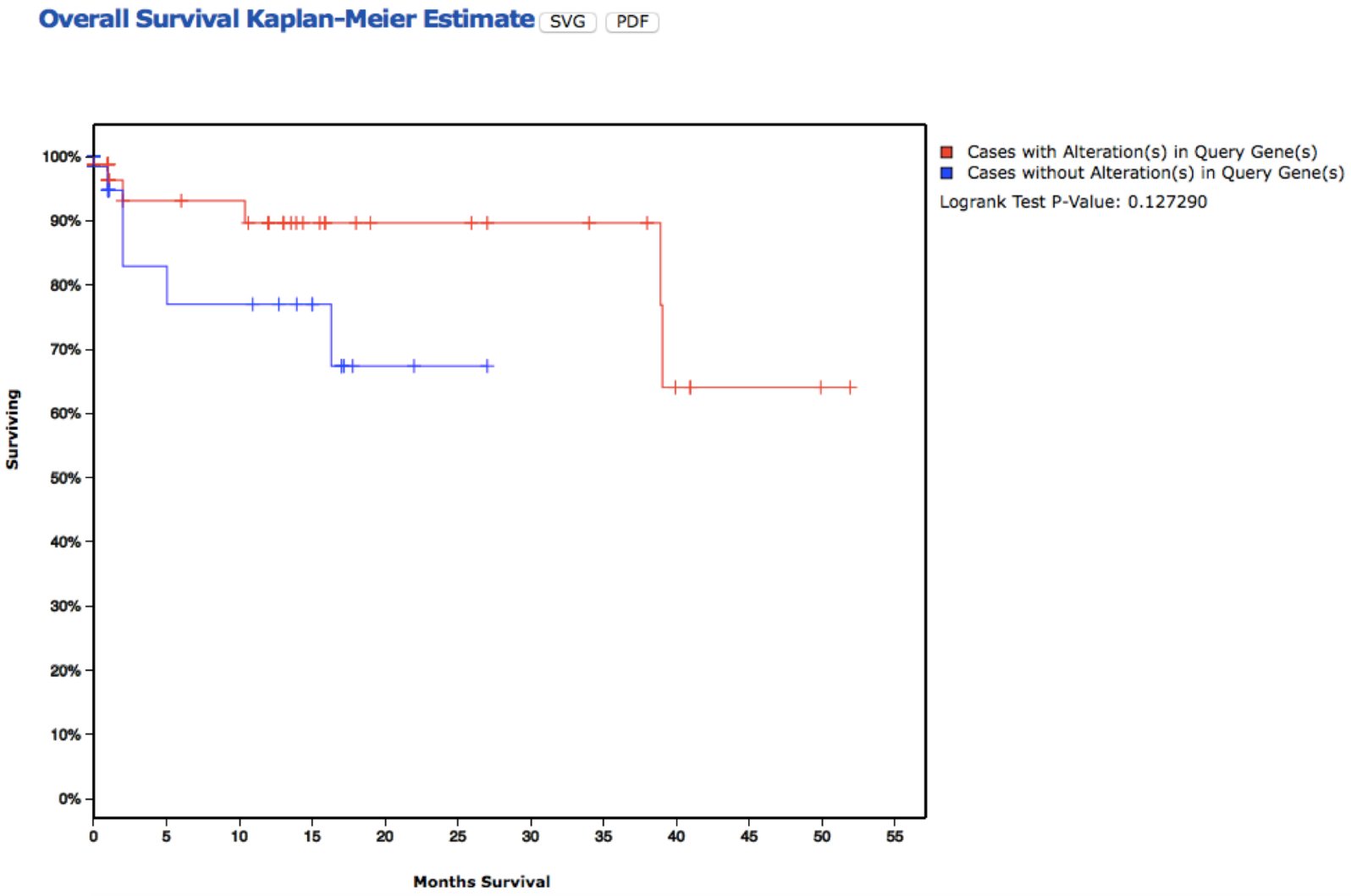
« Oncoprint » view



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