

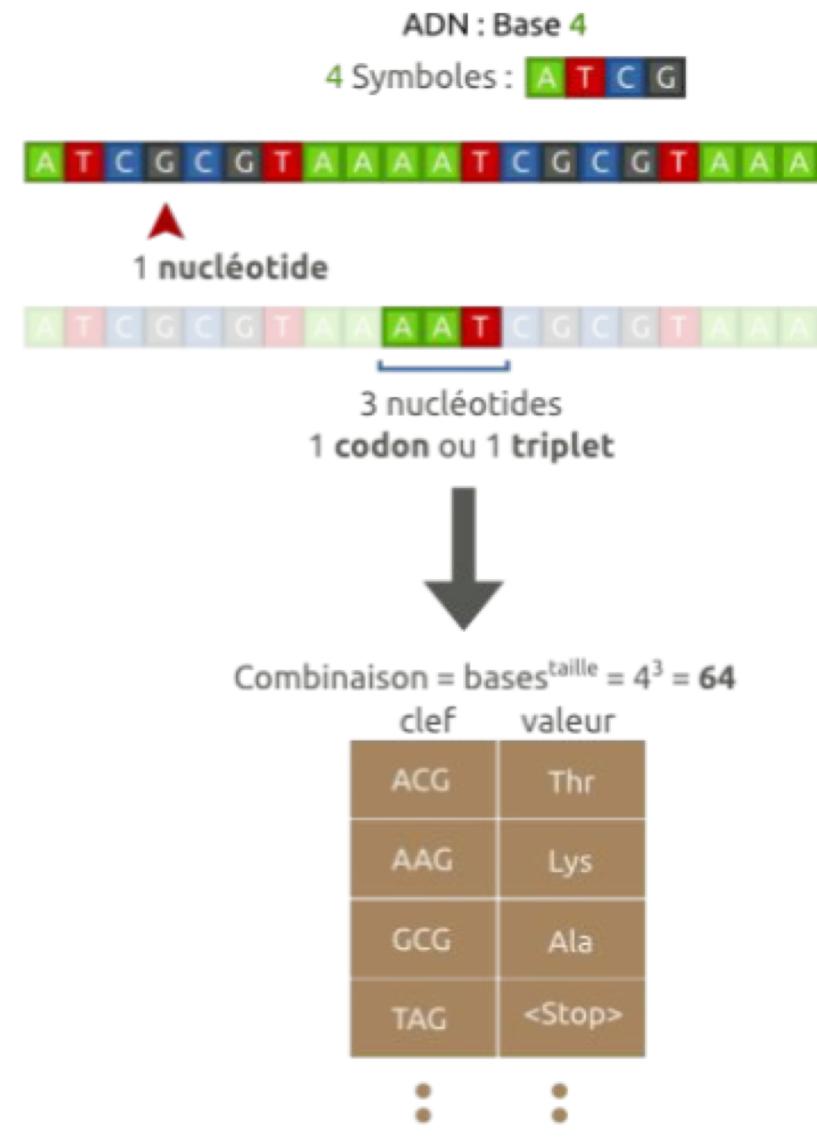
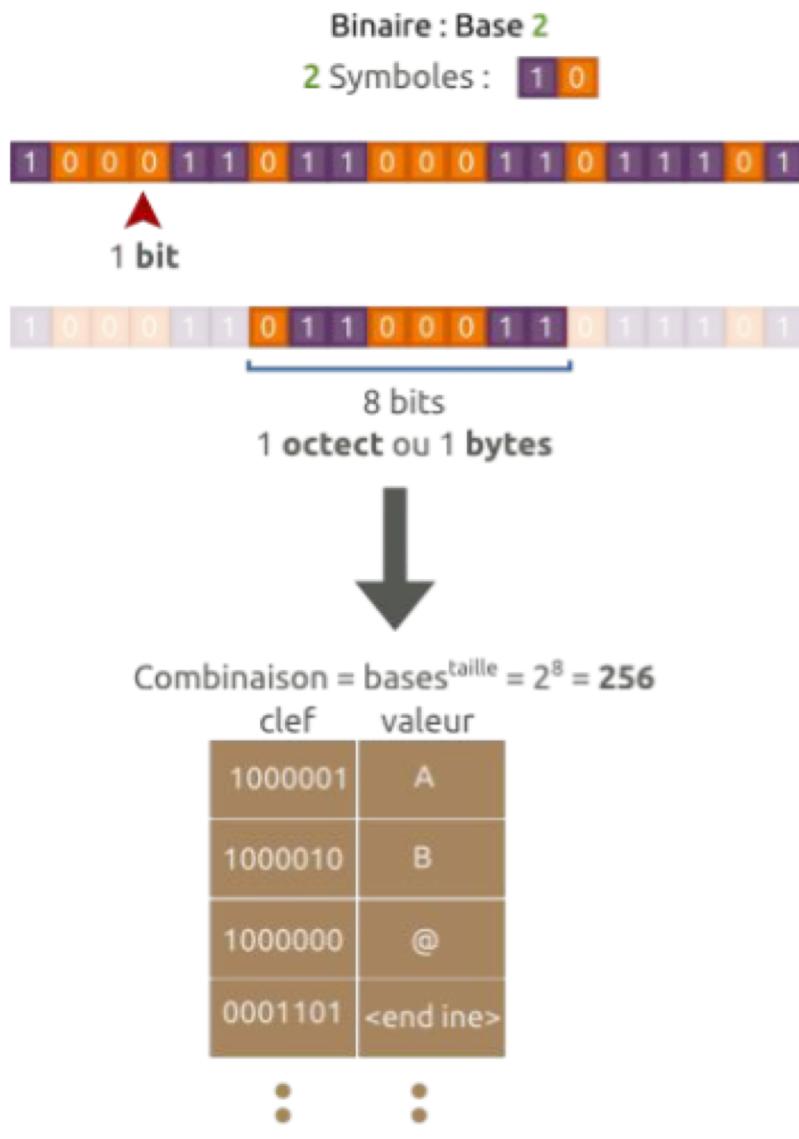
Introduction à la bioinformatique en cancérologie

Daniel Gautheret
2018

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

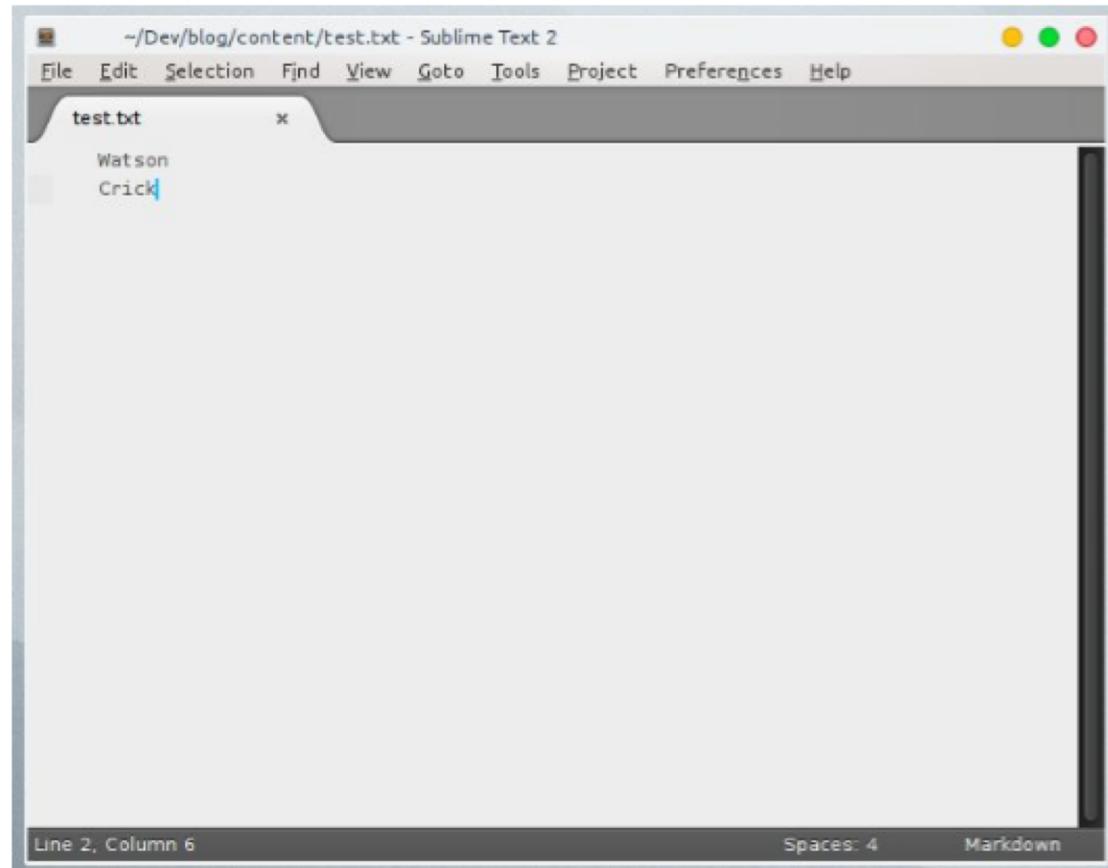


Biologie=science de l'information?

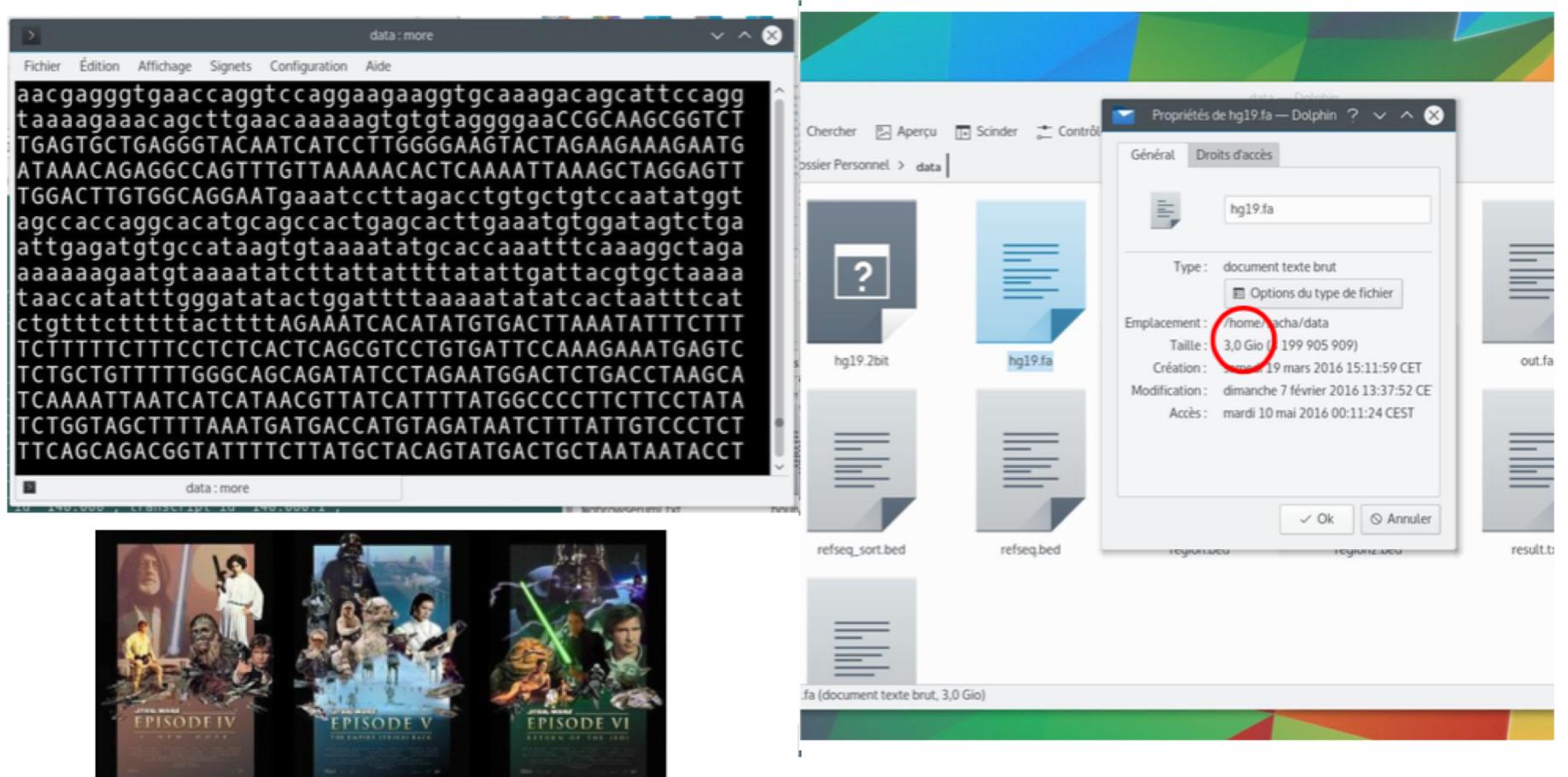


Taille d'un fichier

12 octets
(bytes)=11+1
96 bits (12x8)



Taille d'un fichier

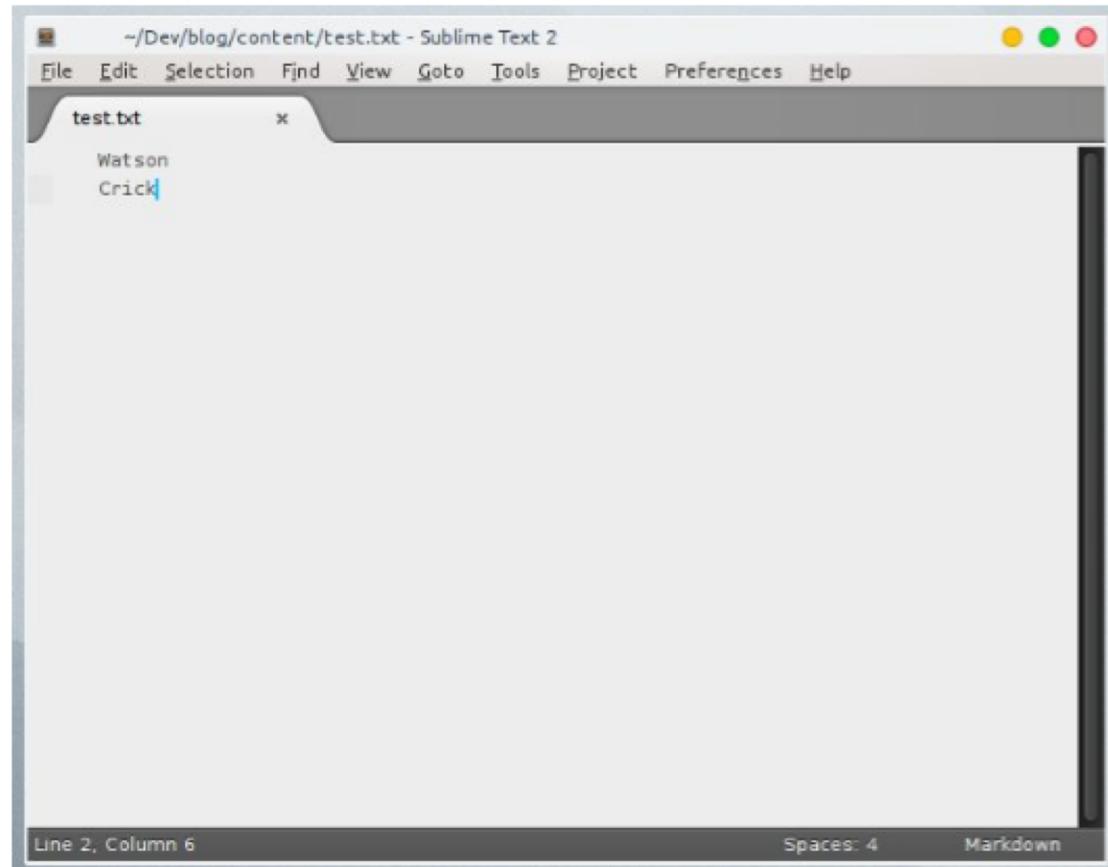


Génome humain version hg19

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

knownGene.hg18.chr21.short.bed – Kate

```

chr21 9926613 10012791
chr21 9926613 10012793
chr21 9926613 10012793
chr21 9926613 10012793
chr21 9926613 10061300
chr21 10042683 10120796
chr21 10042683 10120808
chr21 10079666 10120808
chr21 10080031 10081687
chr21 10083660 10120796
chr21 13332351 13345202
chr21 13366975 13346202
chr21 13361138 13412440
chr21 13904368 13925777
chr21 13944438 13944477
chr21 13945076 13945106
chr21 13973491 13975330
chr21 14137333 14142556
chr21 14200023 14200052
chr21 14202070 14202096
chr21 14237966 14274631
chr21 14270940 14274631
chr21 1432612 14438647
chr21 1432612 14436730
chr21 14403005 14501125
chr21 14459414 14480611
chr21 14510396 14522564
chr21 14510396 14522564
chr21 14510396 14522564
chr21 14567990 14585577
chr21 14567990 14595563
chr21 14567990 14595563
chr21 14567990 14632206
chr21 14665307 14677380
chr21 147779419 14840535
chr21 147779419 14840535
chr21 147779419 14877504

```

Search and Replace Current Project



text.txt

Fichier binaire

arbre.svg [read only] – Kate

```

nHzM+9F1Hgj40bF7u4mJ0+fVMoLLxoykubtGqvVzffxUj10DyntTx7xz/Phxv7
1Y5330qJWVGK+BuKI45kVzq6ks1qal5702620xfzJhx2dnacVFNFQFBK3/Gac4btf
OnTp1//797KJX50vd3In5xkhrlLKysmbv3u3h4SE0CD2u30j14nL69GnhUhjrs2D6q
WrXqq6h+Cgg1UhPnzgcffEEaezcnnJKSgplJHr6+t27/+xGch1MbFySMNhvh+k/gx1Kr
AeuXn1lxw5NzpOfkmPwTm7d+p17rMQQK7GwajTY9v20VLISS9QfrCewfPhq1YXlPh
XLJnSHNgfTrusKADDDQMAAAAAAAAAXqk1p2)Ex00h3Bj2r3ow7p)59BgZDQ0LeSE1
1/332790V1ba3mzne7Spqw59NbwtocPn402z0+29fV1tbR4wLpa4f qkD/pGwad
qfIxsgl29N3okspnZMhB65o1azvB6uvgt96aypPKuywv4QlUpVjFaq+L3uSUEtW
tLSuL1470vXh8kfAv0dKJCOHQExlc1rnFad9fV1baGrt/f19avjclNqnxixauJ12
xp6+vHx8dPnD0h6NGjvcz/542L91uP8Pr1trGxf+P91gyzs2u4dVh1TG6arAbLPbC
Y1Q1MfMFLabcdneykuCxqJ9V1VhHEUD4RGPWt+OR12DP21e+jVyy20p7HtgdrD77aLR2aFpC11A1AAAA
AAAAAA1kJA3MlStLc+pbgqnhw4c2N29vbw2rqrq@ZhM/sScMcOosrJh@eJ1Jt6p0/bgGNNTb9/Pp25
053yGSamudvZBw7tuurZhiYD00LPDy029ygmwa1zhpwgkDxZ/Df1fpPx9yAkSERVfVSQzxx0dnfhkp1
Jt3524ConF254u1LEViS1zUkNbW1qqqq1VLLr50t0jrw3tD89vLm5addvjdudu/lYuXrxYJB19fqrqjUel
NkixT11VbWkxMTH7+-efk5Gsy+/ePVNT08bcuu++y4iLuLcux09yf6L92duE14t/79Waw4d7i4n0IW
g20u5GM205kkbtbe5d0170220yWlrlspLgv0XlBugLcQ18frqlsySoVnB+s3JUL5segtOG1L3n9nfpt7
k.fvP6glDAaaaaAAAADA78at#ChtbTMwGPrswdwIIEkgv1L00LF76epyfThy6PsMTMsPYKD8xsbe@emU-01
GghnrNzRePgn08jry4YfLo60Luv02fM6cBvVvt40C8g9de486k.0q0fTwswM7,TY3kyf//eZsVraFpZK
CjBLGv64841mhBkg8k5pxMg6euXGu7ufxFm9s7u4p4sNkqeysnLhwoV6enrjxo1rbw01MD649uZ186j
+e9V0DM40LxiXhT2rXhUTemNxPTPSxbTsfsm4+c=ahL5fXh9+Qgr4aHhurry9evPjSpUv9TyBLh
zt1Yz7xyt t3HhWpEpz63B3MSX1/60u0yzq6v/FNf61Z4ND221bHcN1kvQ8guvr8tKs vsS1h8pEgsLM
QneruJrPISew9Ea6uB0+di+5+biEl fYNL716E7/d6D2oEawAAAAAAAAdVu9exxxcX12zH2uL5c8a
xL1SzdrXrZbx//79+ /PeYF5dpJ096b25+Pn36xpI098wpL0zC/Torsmsrbw1PezQ6+zkDrri09r6+/25
ErTkwofLpk1pk2tneN1zw9rlpu7gpr+r+Pazvt2dg/r16982M)ysrIwknokB+5jQT00l(pabu
rryuurA7MLCeHuVhQUPHMy2M4fXY39p/74YxcxL5hT2Y3N5Mxuvgxp0+fLhK3bG1txkvLNUng6h4+P
Dhs,OyospKA2vr/+s7D7+yU0gzYbarPE1luLurkvL/UpDdt+o+deSB3hptkuJahsXLCPba7bsLy4M6Nky
xLRsPDIcUkm+etBdwQvPvcU3ydznsOpUodl,fNsdEpCx9582d3/3/EwAAAAAAAACpH6yRd4ACWj1poX
FfYQGHjMyMgg03v5e++VbtwsJx8MzL4EPGfRtEtfL5Fzq2hxd3ByMjCq1cfzp49P19fPylw4M8yH5+/1
Nc/NDcf+BU0X3u/w7fjX2GVLvpA731b2c1jBcEc3k8T1yk+vh69adg17KyltmKhRyygxubX3/9uY92s
4E11m4efj8c2w/SuL2eJlaSaryB11XNesLcwgID442mjLkzs3ld1tqj11kChP05/Mzkh2n51Exuvgx6+r
qIImrj1690nv27Ly6PBjhx44d0eEgl70zMs0PfDcs1/PdUme6zDdExJ+xpwXIjhIaeYD02Kvl15YR7
bkchNtrs896hgnz9vZyXY4l611wKSZN0Gsiyu7M0mknqLHPf qzP9w61zXceP88q0gtrf6n8dJh0nkKE4UH
6rfqgEAAAAAAAALZzw8trOmvj+vrzfsof96RlD9xY1Pf61peZcZc+nSKRcuPMuYvDz5eV2rZb6Jn

```

Search and Replace Current Project



image.png

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

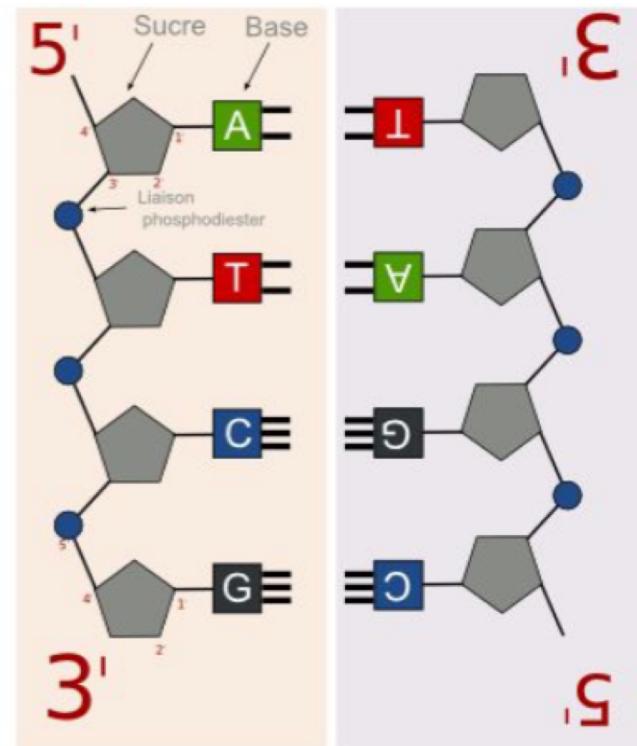
<https://www.w3.org/TR/REC-xml/>

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
CGCATGCTAGTCAGGCAGGGAGGGATATTATTAGCGGCGTATCGGATGA
CAGCATTACGGCGGGAGTGCTATTATTATGAGCGGCGAT
>identifiant2 commentaire libre
CAGGCAGGTTCTTATTATCAGCGGGCGGAGGCAGGCGATGCATC
CAGTGCAGTGCAGTAGTCAGCGATGCATTTATGACTGACTCAGTTT
CCCGCTAGCTATGCTATTGATCGATTGTGAGCTGATCTGGC
CAGCTATGCTTAGTA
```

Format fastq

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

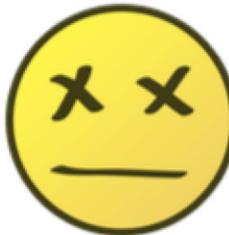
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGCCGCTGCCATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

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

Qualité dans le format fastq

Qualité=-10 log₁₀(P_{erreur})

@SEQ_ID1
GATTGGGGT
+
42,40,134,47,36 35 40 40 40



@SEQ_ID1
GATTTGGGGT
+
*(å/\$#((((



Ø	32	@	64	Ø	96	'	128	Ç	ü	á	192	Ł	224
1	33	!	65	A	97	a	129	é	í	í	193	ł	225
2	34	"	66	B	98	b	130	é	ó	ó	194	ł	226
3	35	#	67	C	99	c	131	á	ú	ú	195	ł	227
4	36	¤	68	D	100	d	132	á	ñ	ñ	196	ł	228
5	37	%	69	E	101	e	133	á	ä	ä	197	ł	229
6	38	&	70	F	102	f	134	é	æ	æ	198	ł	230
	39	'	71	G	103	g	135	ç	œ	œ	199	ł	231
	40	<	72	H	104	h	136	é	ɛ	ɛ	200	ł	232
	41	>	73	I	105	i	137	é	ɛ	ɛ	201	ł	233
	42	*	74	J	106	j	138	é	ɔ	ɔ	202	ł	234
11	€	43	+	75	K	107	k	139	í	í	203	ł	235
12	£	44	-	76	L	108	l	140	í	í	204	ł	236
13		45	-	77	M	109	m	141	í	í	205	ł	237
14	₽	46	.	78	N	110	n	142	á	á	206	ł	238
15	*	47	/	79	O	111	o	143	á	á	207	ł	239
16	▶	48	0	80	P	112	p	144	é	é	208	ł	240
17	◀	49	1	81	Q	113	q	145	é	é	209	ł	241
18	‡	50	2	82	R	114	r	146	é	é	210	ł	242
19	::	51	3	83	S	115	s	147	ô	ô	211	ł	243
20	¶	52	4	84	T	116	t	148	ö	ö	212	ł	244
21	₩	53	5	85	U	117	u	149	ò	ò	213	ł	245
22	-	54	6	86	V	118	v	150	û	û	214	ł	246
23		55	7	87	W	119	w	151	ù	ù	215	ł	247
24	↑	56	8	88	X	120	x	152	ü	ü	216	ł	248
25	↓	57	9	89	Y	121	y	153	ö	ö	217	ł	249
26	→	58	:	90	Z	122	z	154	ø	ø	218	ł	250
27	←	59	:	91	[123	<	155	ø	ø	219	ł	251
28	↶	60	<	92	`	124	:	156	ø	ø	220	ł	252
29	↷	61	>	93]`	125	:	157	ø	ø	221	ł	253
30	▲	62	>	94	^	126	~	158	ø	ø	222	ł	254
31	▼	63	?	95	-	127	△	159	f	ł	223	ł	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 for investigating biology. The *C. elegans* Sequencing Consortium
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="MDSCEECDLEVDSDDEEDQLFGEK CISLLSSLLPLSSSTLLSNA
INLELDEVERPPPLLNVLEBQQFPKVCANIEENELEADTEEDIAETADDEESKDPE
KTENFEPSVTMDTYDFPDYPVQIRARPQVPPKPIDTVRYSMNNIKESADWQLDELL
EELEALETQLNSNGGDQLLLGVSGIPASSSRENVKSISTLPPPPPALSYHQT PQQPQ
...
QVYTGIGWEKKYKSPTPWCI SIKLTALQMRSQFIKYICAEDEMTFKKWLVALRIAKN
GAELLEN YERACQIRRETLGPASSMSAASSSTAISEVPHSLSHHQRTPSVASSIQLSS
HMMNNPHTPLSVNVRNQSPASFSVNSCQQSHPSRTSAKLEI QYDEQPTGTIKRAPLDV
LRRVSRASTSSPTIPQEESDSDEEFPA PAPPVAVSVMRMPVVTPPKCPLTSKKAPPP
PKRSDTTKLQSASPMAPKNDLEAALARREKMATMEC"
...

BASE COUNT 2598 a 2024 c 1888 g 2449 t
ORIGIN
1 ttctaaaagt cgaaaaacga gcaattttt atgctagatt ttttgattt acgaattttt
61 tcgtttttt ttctttaaaa aaggttttt accccttaaa gtttccttt cccttccat
121 ttttccttc ttctttatac gacttctaa gtttcaactc taaaacaag ctacatgtac
181 atttccggta aacttttgtt ctcagaatg ccattttt tttgttacat ttattcaaga
241 ttgaattcca aaatttcagc caatatggc agttgcgaag aggaatgcga tctggaagtt
301 gagactgacg aagaagatca actttttgtt gaaaatgtt gaggttcttat tggtaacc
361 aaagaaatgtt cagtggccg taaacacttgc actcccaat ggttctcg aattaccta
421 tgcaacat tcaagtgtt gccgtttagt cttagccaat ttgaaacgtt tagatgttaa
481 atggaaaatg ggttaaagttt ttatattttt agaaaaaaagg tttggaaaaaa aatcgagtca
541 ctgaatagt tgaagaacgg aaaataaaa ctttccaaaa atcataaaac atttagtgg
601 tcgaaaattt tagtggttt tttgttggta tggttgcaca aaagctaaac catctttatt
661 gtatgtttt aaaaatgttca caaagatgcg ttttttttca aatttggca ggctatcttt
721 acattcacat ttggataat caaatttttcc ttatcgctaa caaattttcc tattttccca
781 attattcgtt ttataaaagc ttggtagta tggtgtct atctttatgtt gtcatcgatt
...
...

Format Genbank

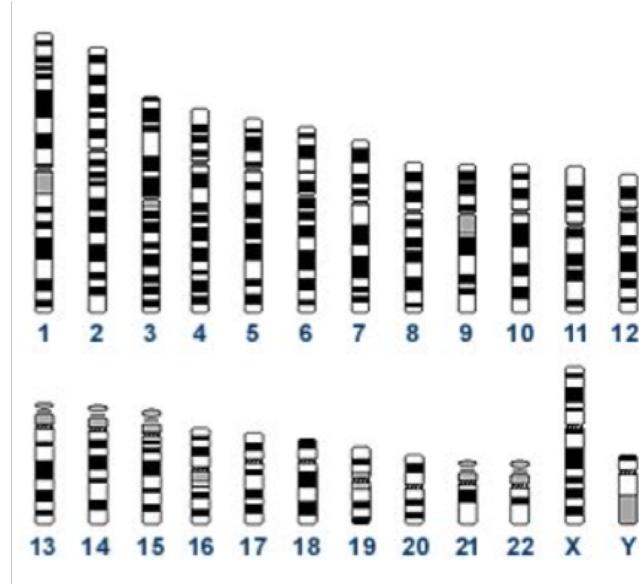
Cherchez sur le site du NCBI le mRNA du gène *GJB2*
dans la section nucleotides

- *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

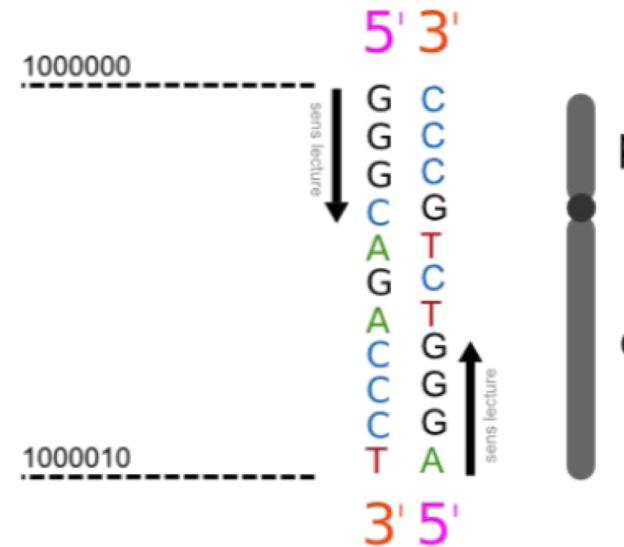


Accéder directement à une région

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..)



5:1000000..1000010:**1**
5' GGGCAGACCCT 3'

5:1000000..1000010:**-1**
5' AGGGTCTGCC 3'

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
	1-based
SAM	1-based
BAM	0-based
	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

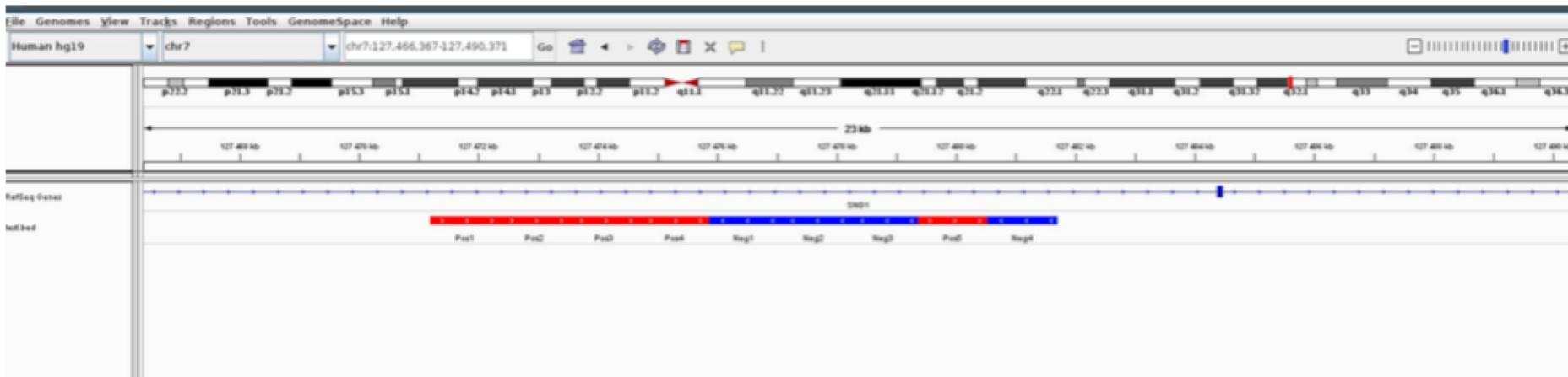
Attention

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

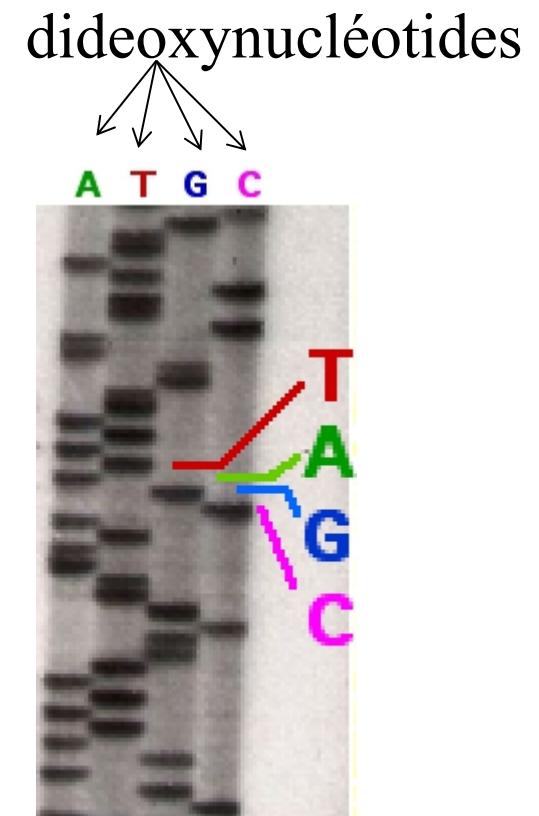
Autres formats de régions

- 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 chaîne
 - Synthèse interrompue à un certain type de base.
 - 4 types = 4 réactions
- Amélioré en 1987 par l'introduction de marqueurs fluorescents (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

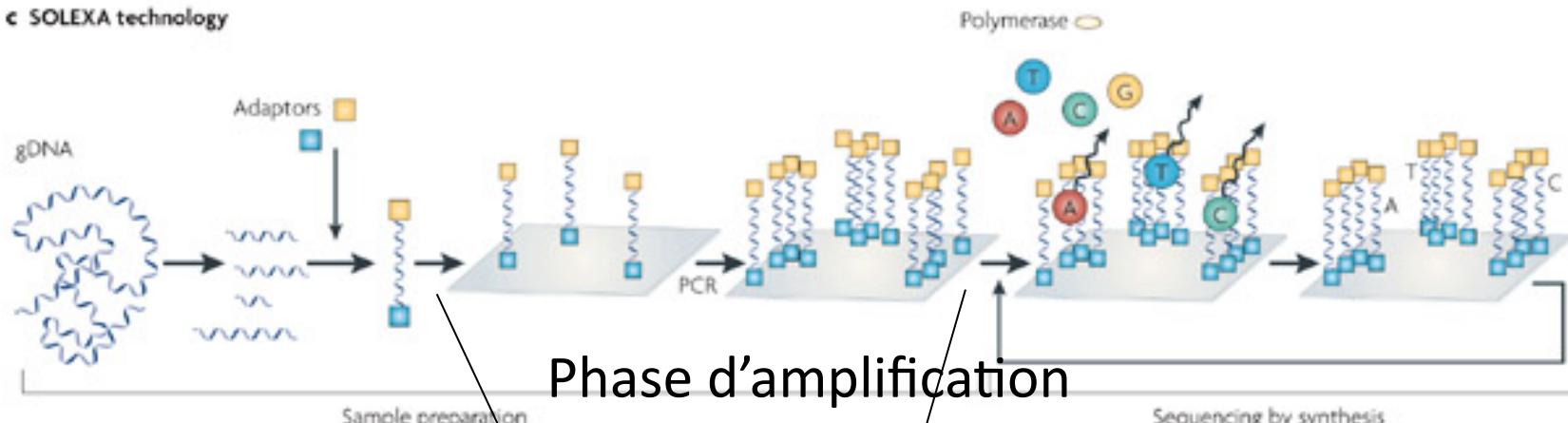
NGS : Next Generation Sequencing (2005-)

Faster, Cheaper, Deeper

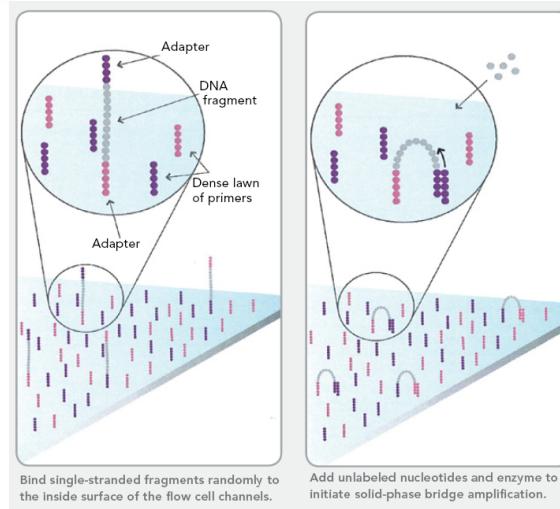
NGS: parallélisation du séquençage

Solexa/Illumina

c SOLEXA technology



Phase d'amplification



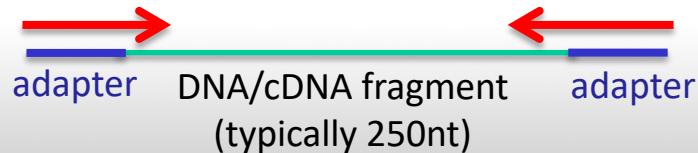
Nature Reviews | Microbiology

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

Sequencing Output per Flow Cell

	NovaSeq 6000 System		
Flow Cell Type	S1*	S2	S4
2 x 50 bp	134-167 Gb	280-333 Gb	NAT†
2 x 100 bp	266-333 Gb	560-667 Gb	NAT†
2 x 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

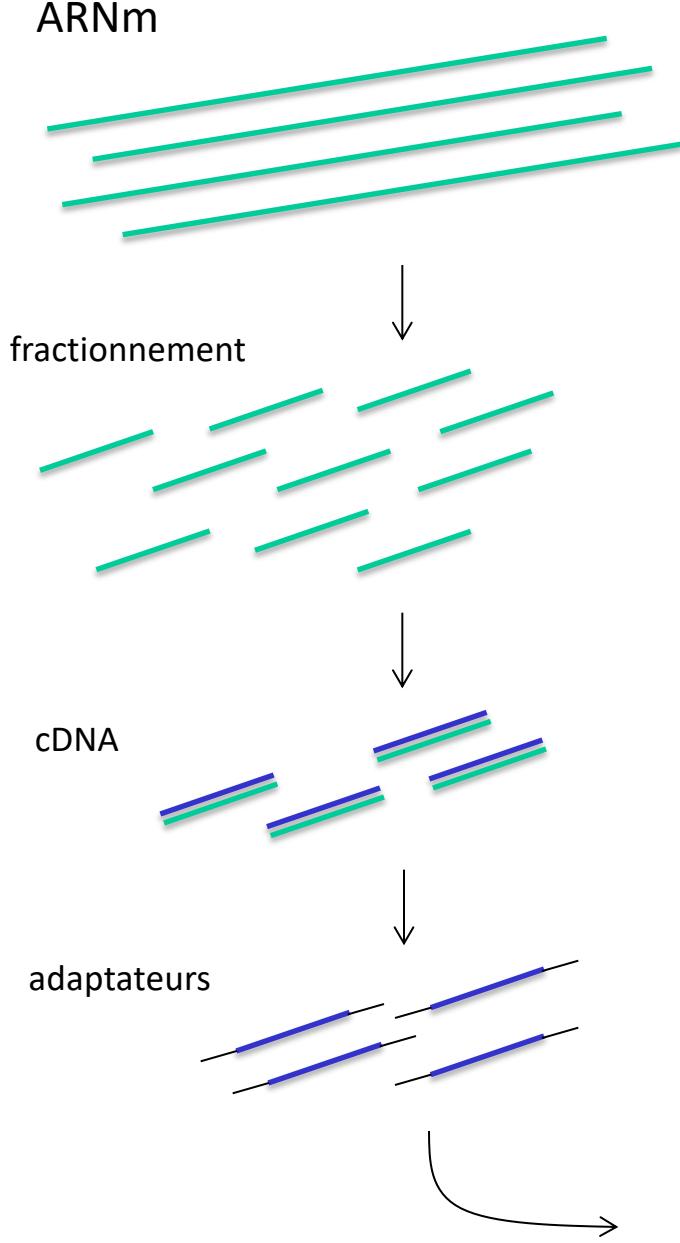
- En cancérologie, 2 grandes applications
 - Génétique constitutionnelle (recherche de prédisposition)
 - Génétique somatique (diagnostic, médecine de précision)

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.

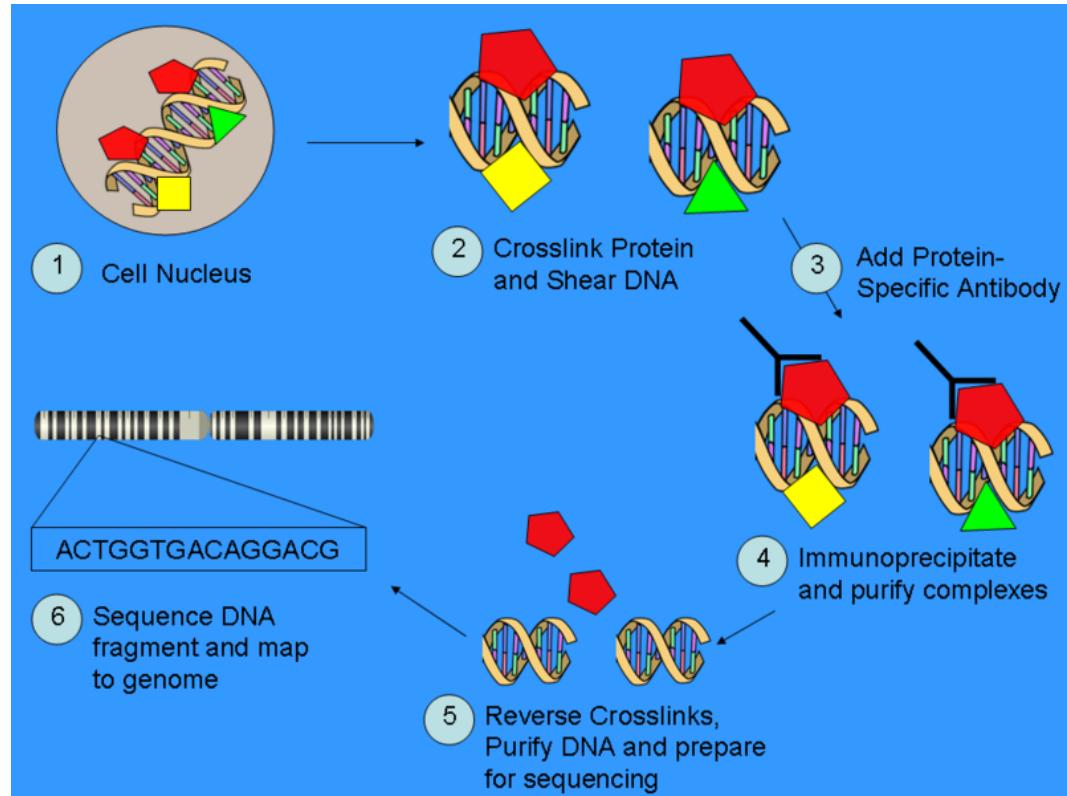


Séquençage

ChiP-Seq

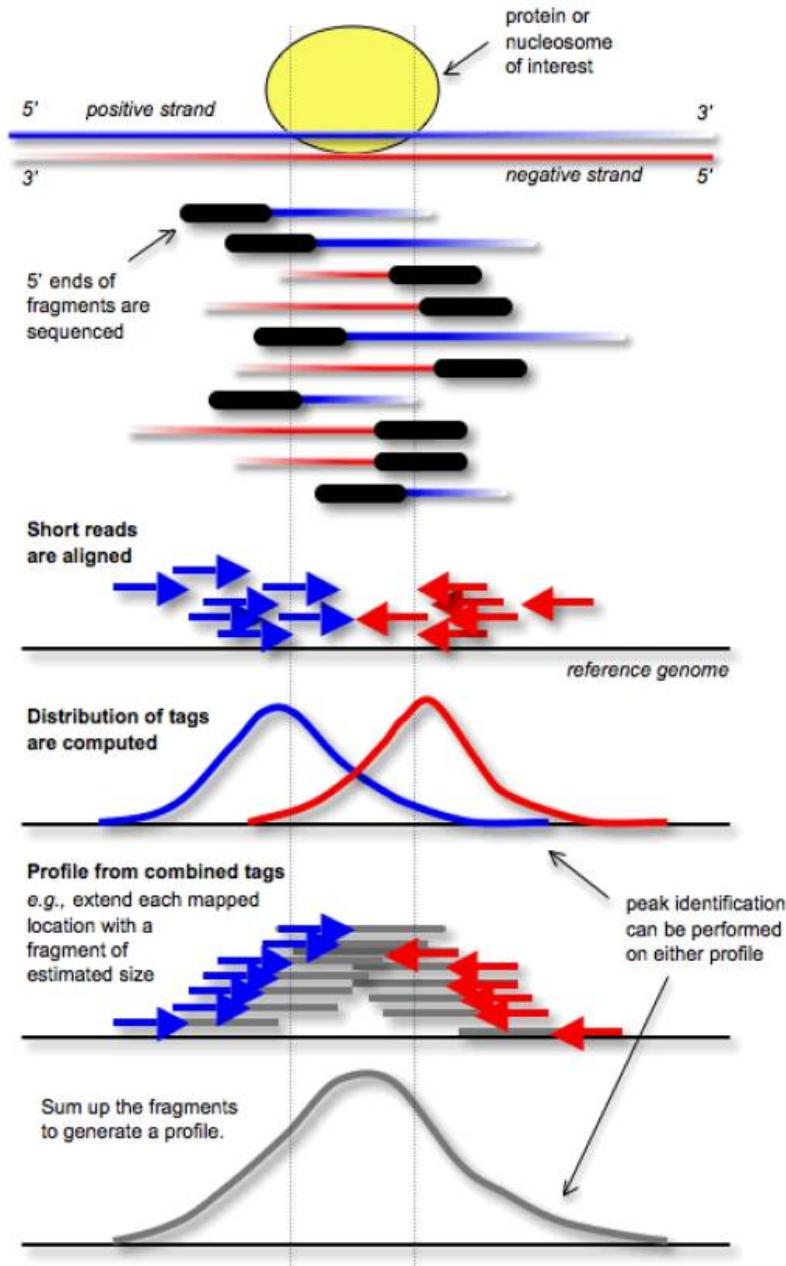
ChIP-Seq

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



Wikipedia

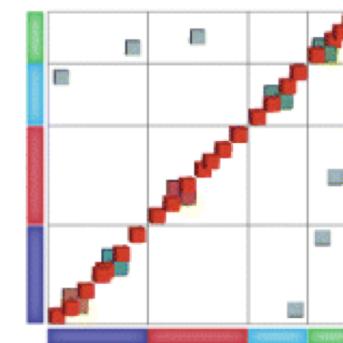
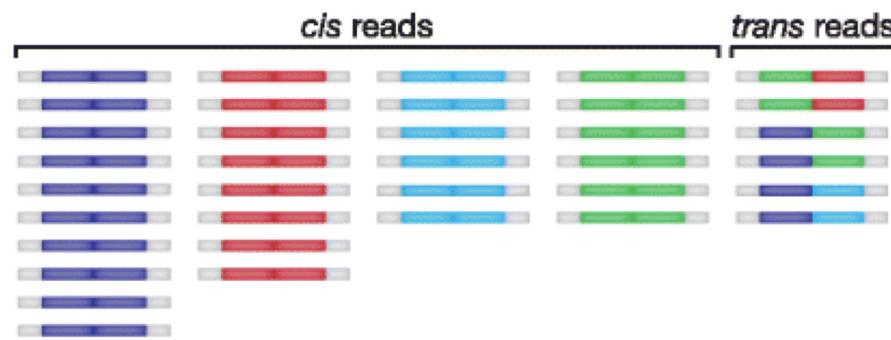
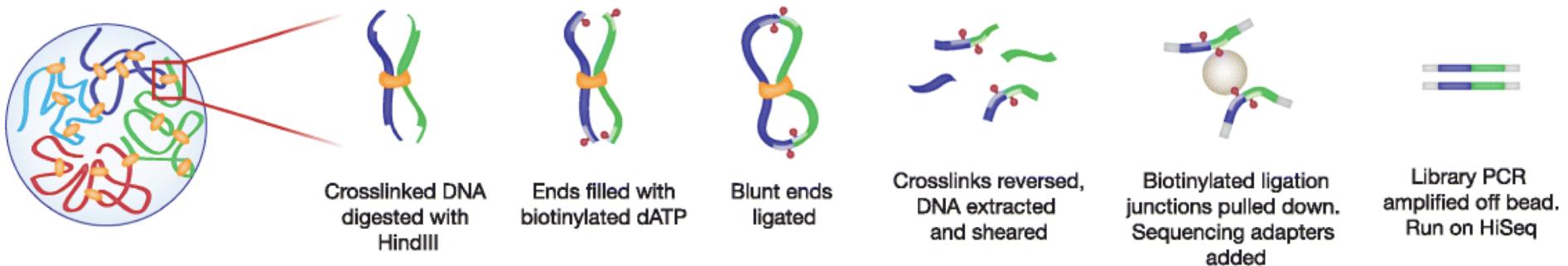
Detection de pics



Hi-C

- Chromosome conformation capture

Hi-C



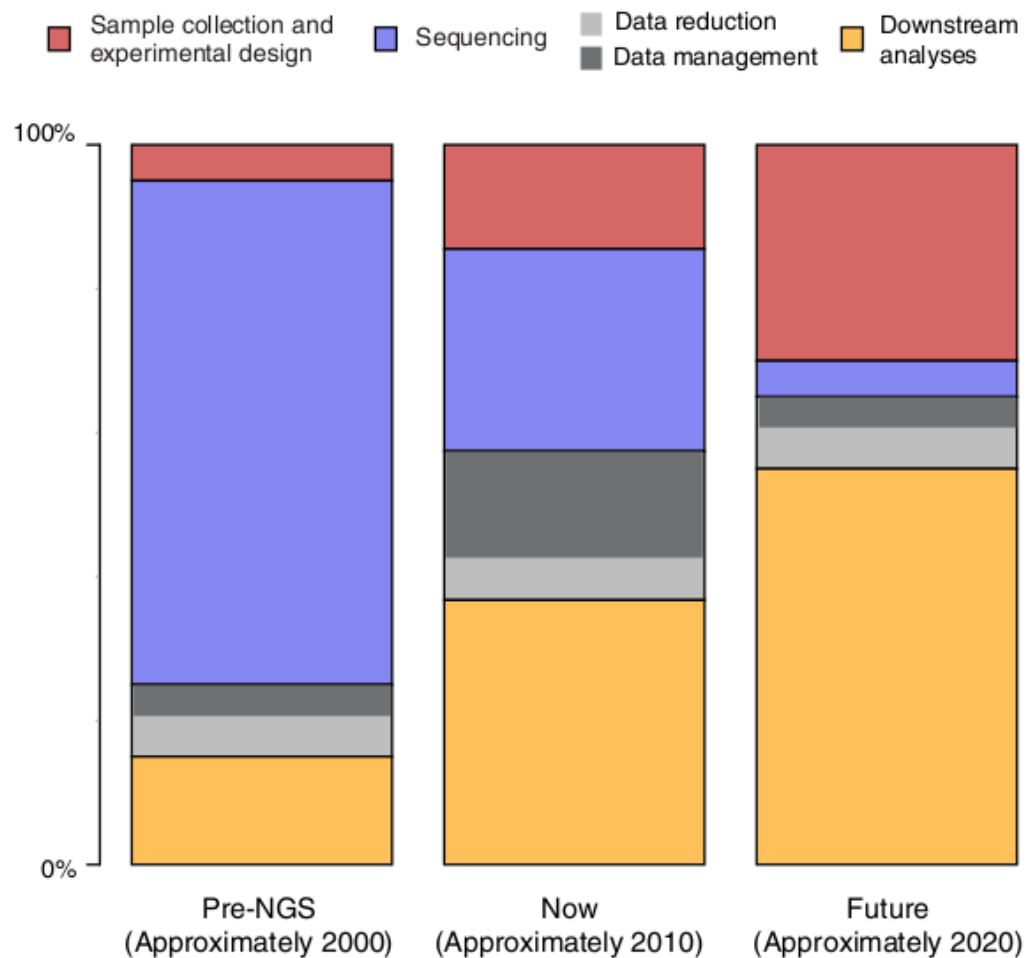
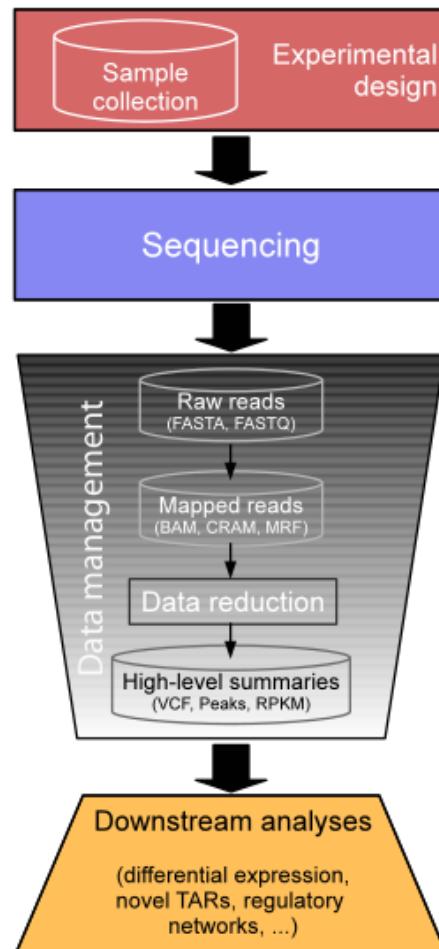
Vers une utilisation systématique du séquençage en médecine

- Déjà tous les patients ont un exome complet séquencé 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



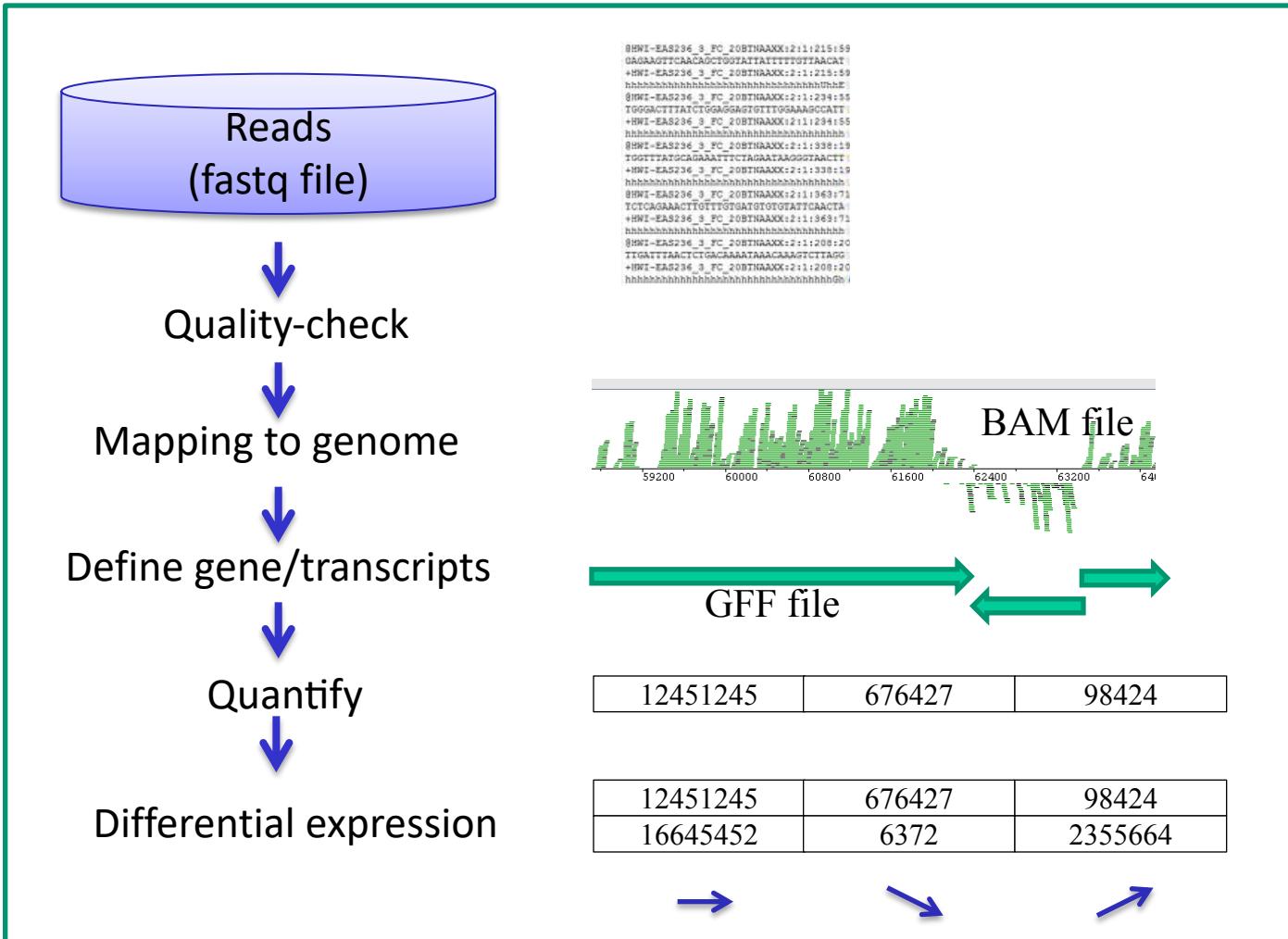
Les outils

« Pipelines » & « workflows »

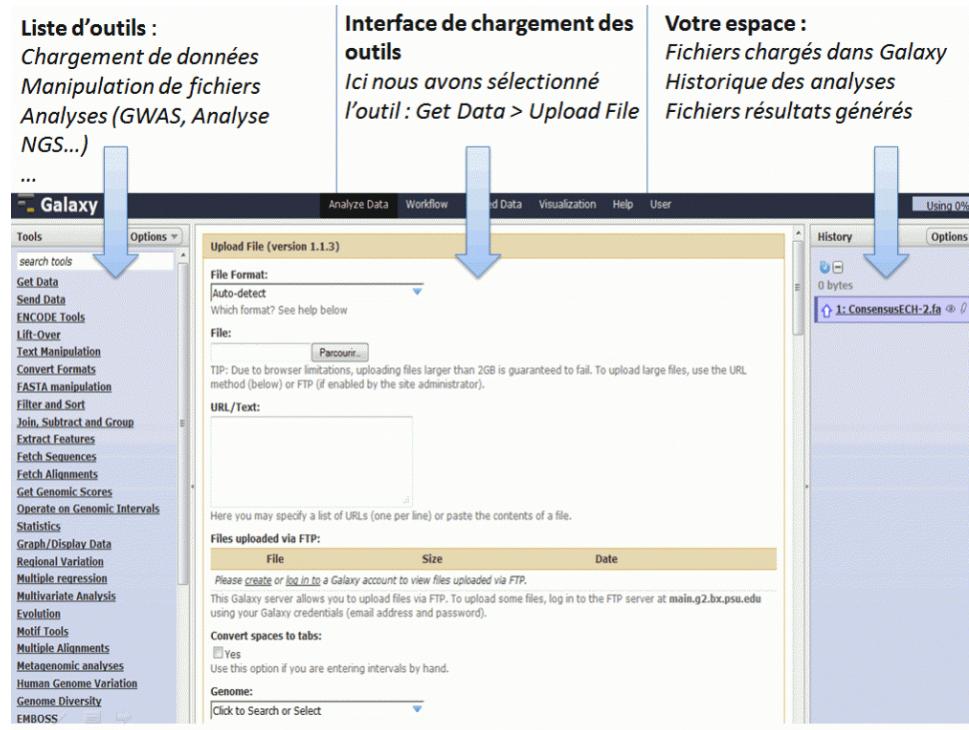
Example: an RNA-seq pipeline

« 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

The Cancer Genome Atlas



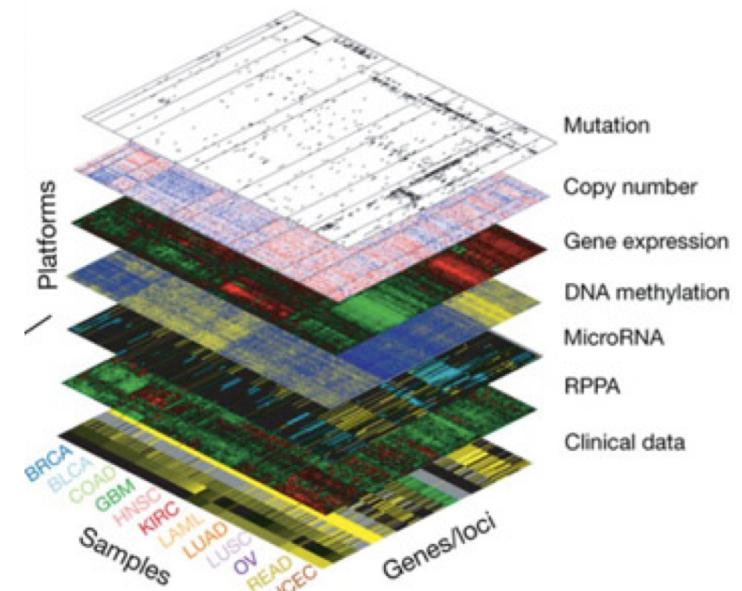
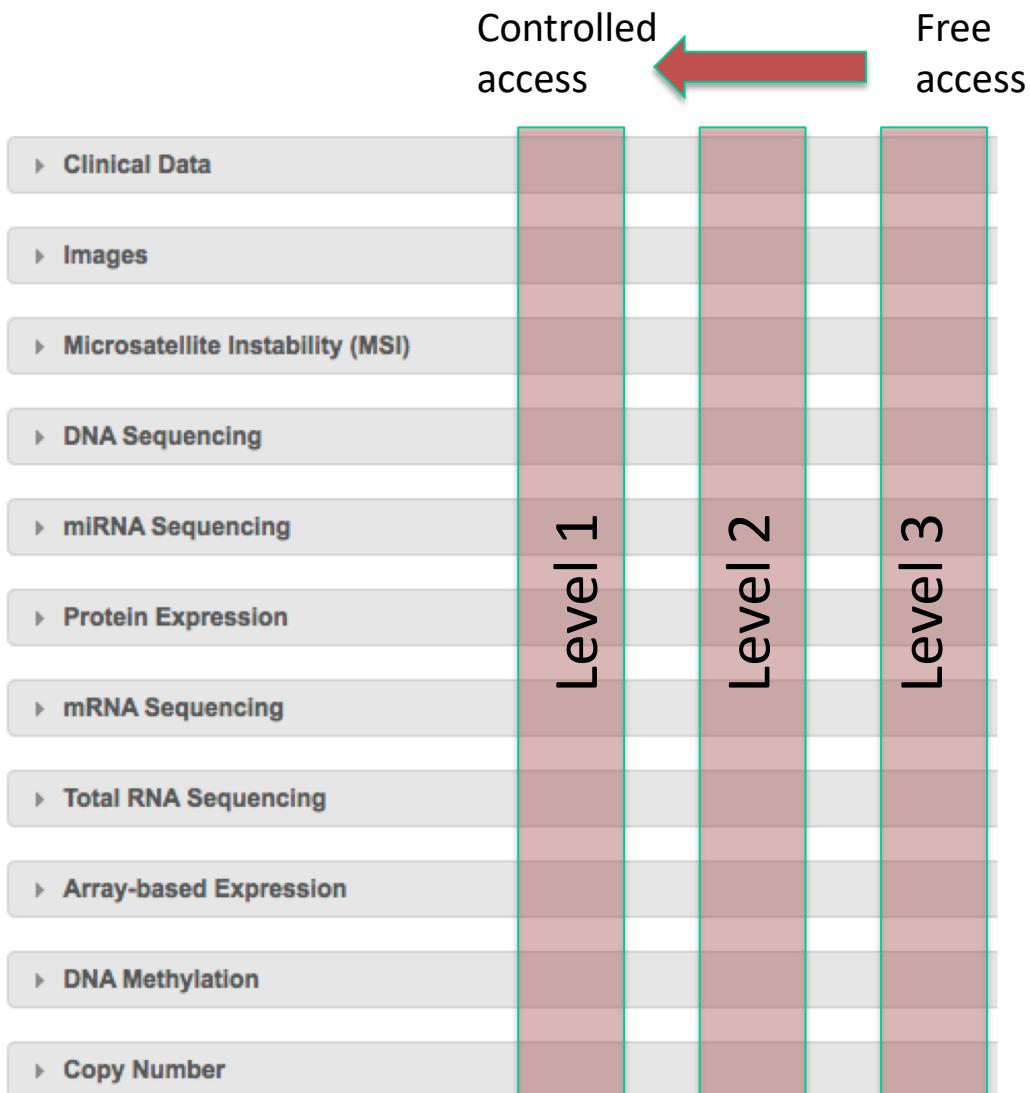
*Understanding genomics
to improve cancer care*

NCI, NHGRI, USA

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)

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:

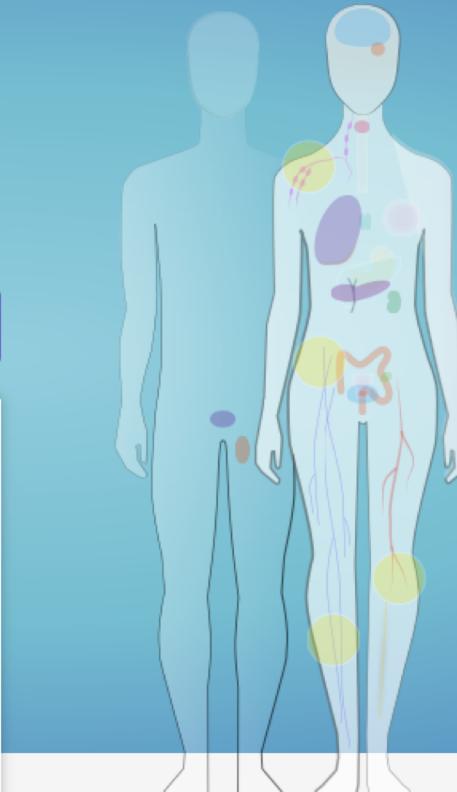
Projects Exploration Repository

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

Primary Site	Cases
Adrenal Gland	~400
Bile Duct	~10
Bladder	~450
Blood	~500
Bone	~450
Bone Marrow	~300
Brain	~1200
Breast	~1200
Cervix	~400
Colorectal	~550
Esophagus	~100
Eye	~50
Head and Neck	~500
Kidney	~1500
Liver	~500
Lung	~1200
Lymph Nodes	~50
Nervous System	~1200
Ovary	~550
Pancreas	~100
Pleura	~50
Prostate	~500
Skin	~500
Soft Tissue	~300
Stomach	~500
Testis	~50
Thymus	~50
Thyroid	~500
Uterus	~500

The GDC Data portal

NATIONAL CANCER INSTITUTE
GDC Data Portal

Home Projects Data Analysis Quick Search Login Cart 0 GDC Apps

Cases Files < Hide Filters Add a Case/Biospecimen Filter

Start searching by selecting a facet or try the Advanced Search [Advanced](#)

Summary Cases (14,551) Files (274,724) [Browse Annotations](#)

Add all files to the Cart Download Manifest

FILES 274,724 CASES 14,551 FILE SIZE 470.57 TB

File Counts by Project 39 Projects

File Counts by Access Level 2 Access Levels

File Counts by Data Format 7 Data Formats

File Counts by Primary Site

File Counts by Data Type

File Counts by Experimental Str...

Search for Case Id

Search for Submitter Id

Primary Site

- Kidney 1,681
- Brain 1,133
- Nervous System 1,127
- Breast 1,098
- Lung 1,089

24 More...

Cancer Program

- TCGA 11,315
- TARGET 3,236

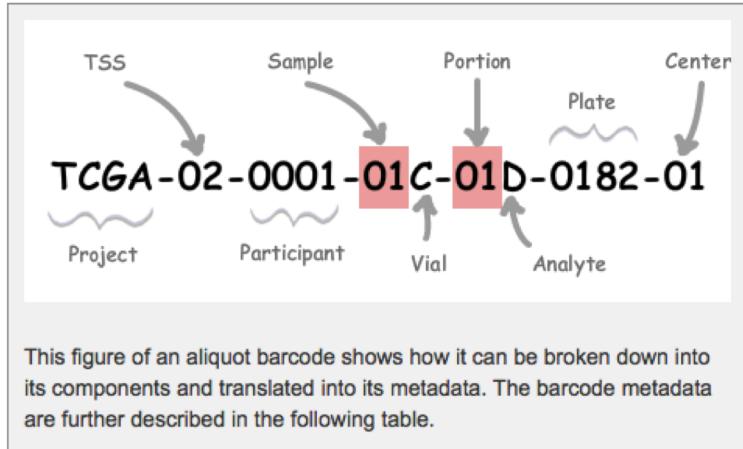
Project

- TARGET-NBL 1,127
- TCGA-BRCA 1,098
- TARGET-AML 988
- TARGET-WT 652
- TCGA-GBM 617

34 More...

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

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 — 179x47

```
<?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

- PCAWG¹: a collaboration with ICGC² 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



- 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

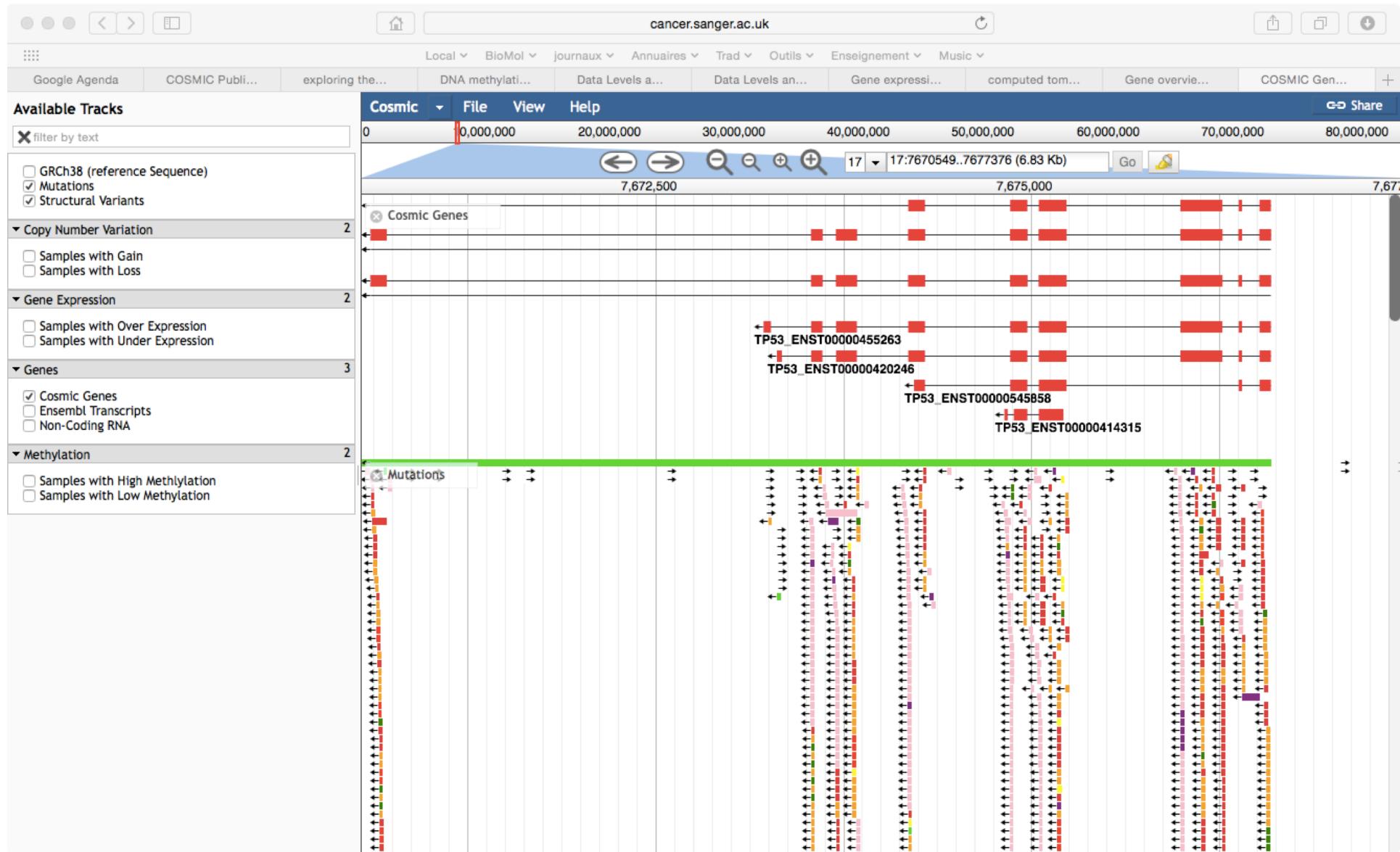
COSMIC Curation

- 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. »

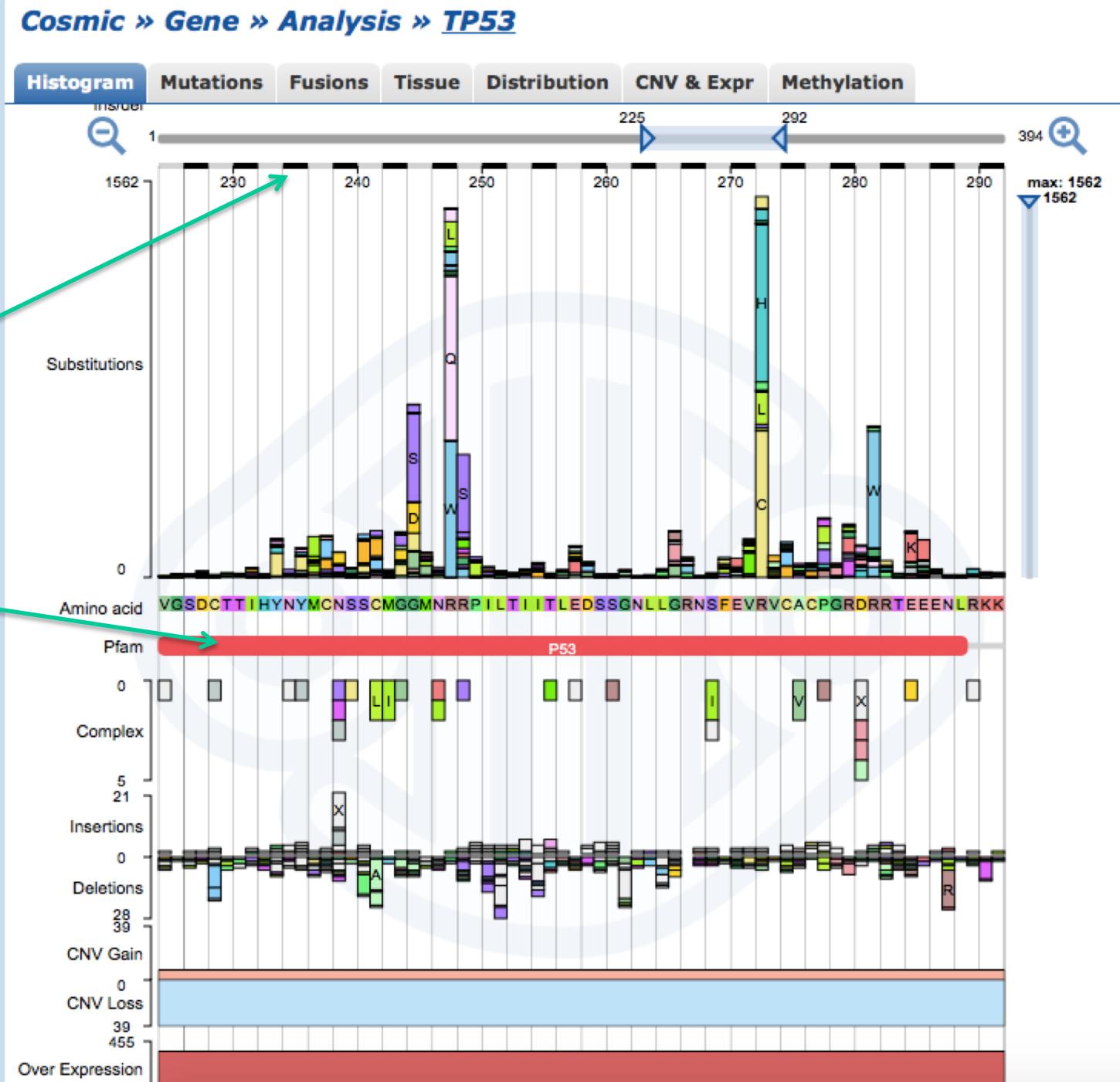
COSMIC genome browser



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: ?

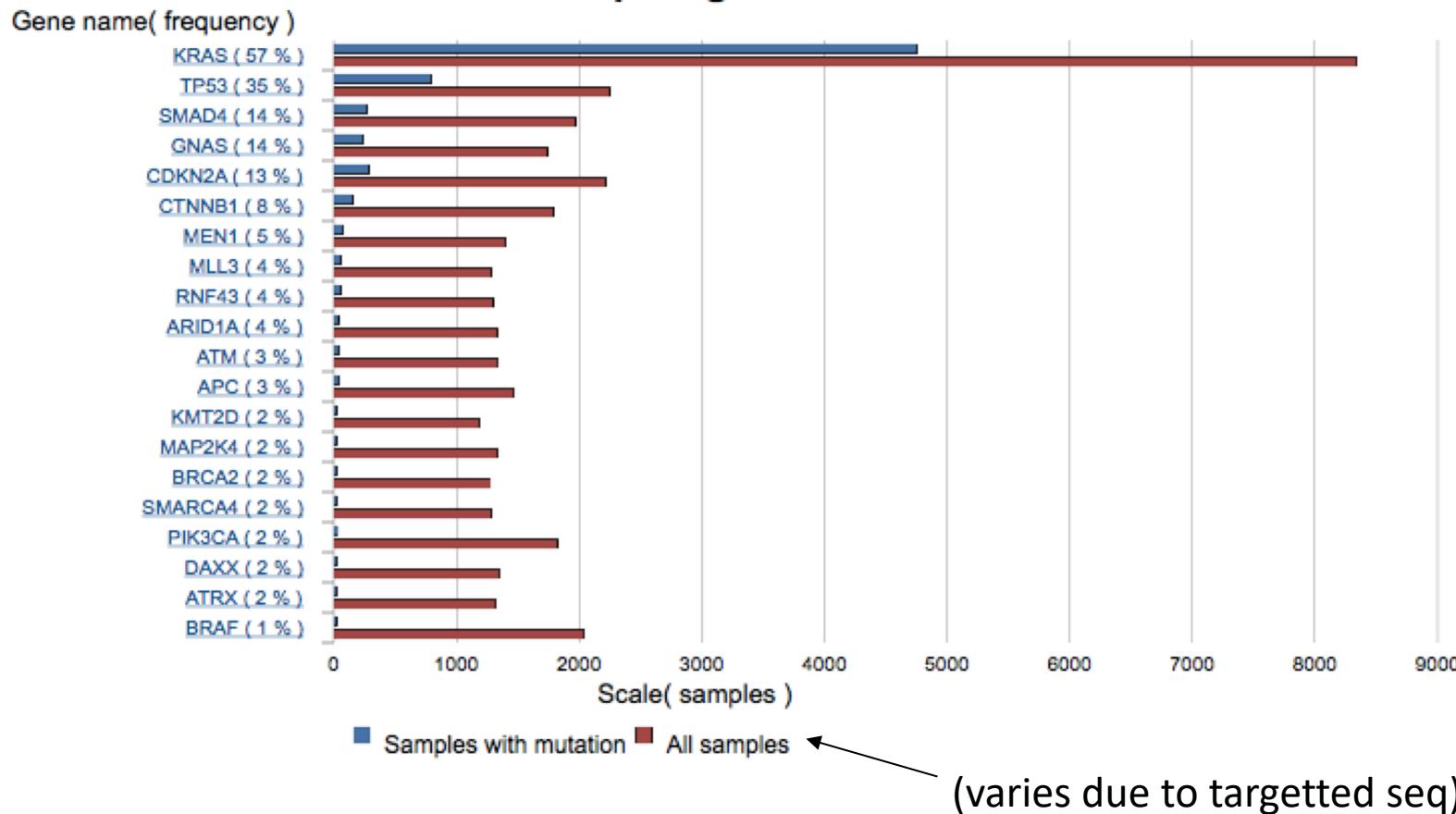
Tissue	Point Mutations		Copy Number Variation		Gene Expression		Methylation	
	% Mutated	Tested	Variant %	Tested	% Regulated	Tested	% Diff. Methylated	Tested
Adrenal gland	508	-	-	79	-	-	-	
Autonomic ganglia	586	-	-	-	-	-	-	
Biliary tract	872	-	-	-	-	-	-	
Bone	955	83	-	-	-	-	-	
Breast	11869	966	1032	707	-	-	-	
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

Cosmic » Cancer Browser » Pancreas

Top genes Mutation Matrix Fusion Genes with Mutations Genes without Mutations Methylation CNV & Expr Table CM

Top 20 genes



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

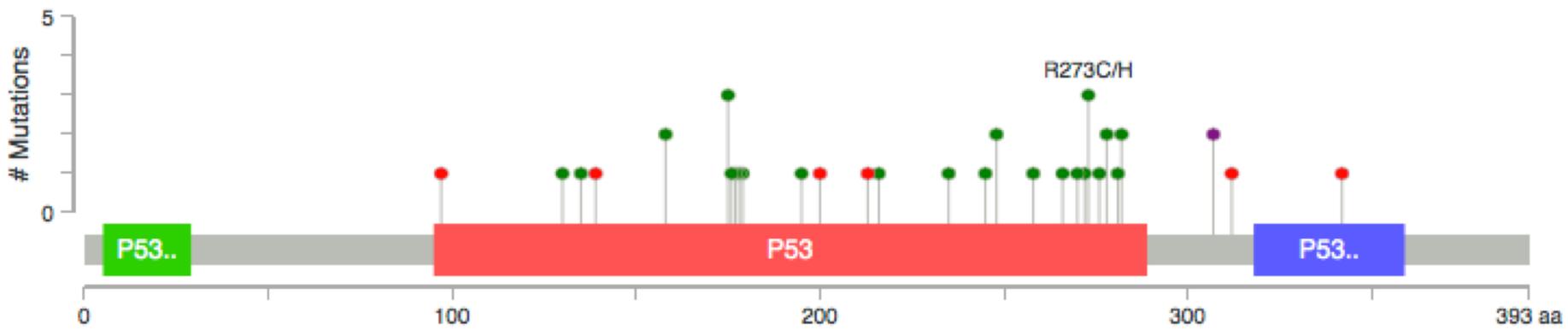


- Integration of Data from 89 cancer genomics studies.
- Focus on analysis tools
 - Mutual exclusivity
 - Gene networks

Mapped mutations on proteins

TP53: [Somatic Mutation Rate: 34.1%]

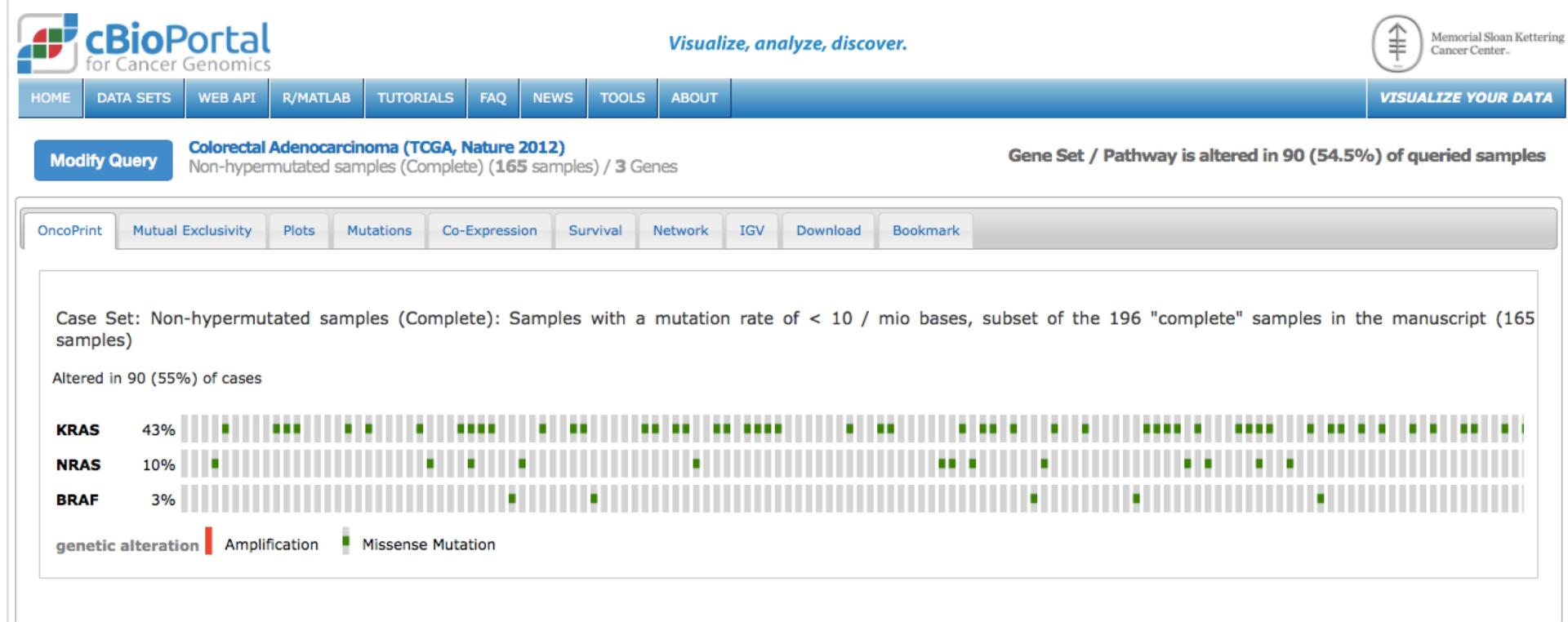
P53_HUMAN [PDF](#) [SVG](#) [Customize](#) [Color Codes](#)



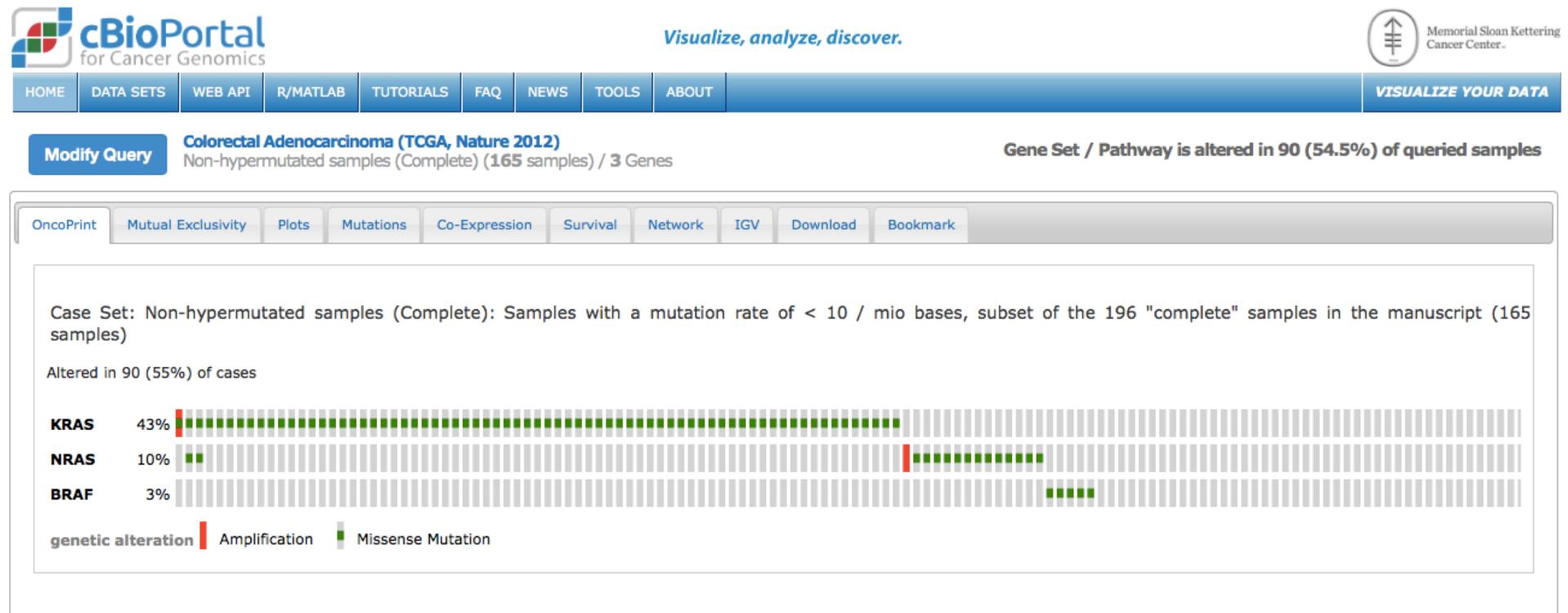
Mutations mapped on TP53 in Glioblastoma dataset (TCGA, Nature 2008)

See also « MutationMapper » tool

« Oncoprint » view

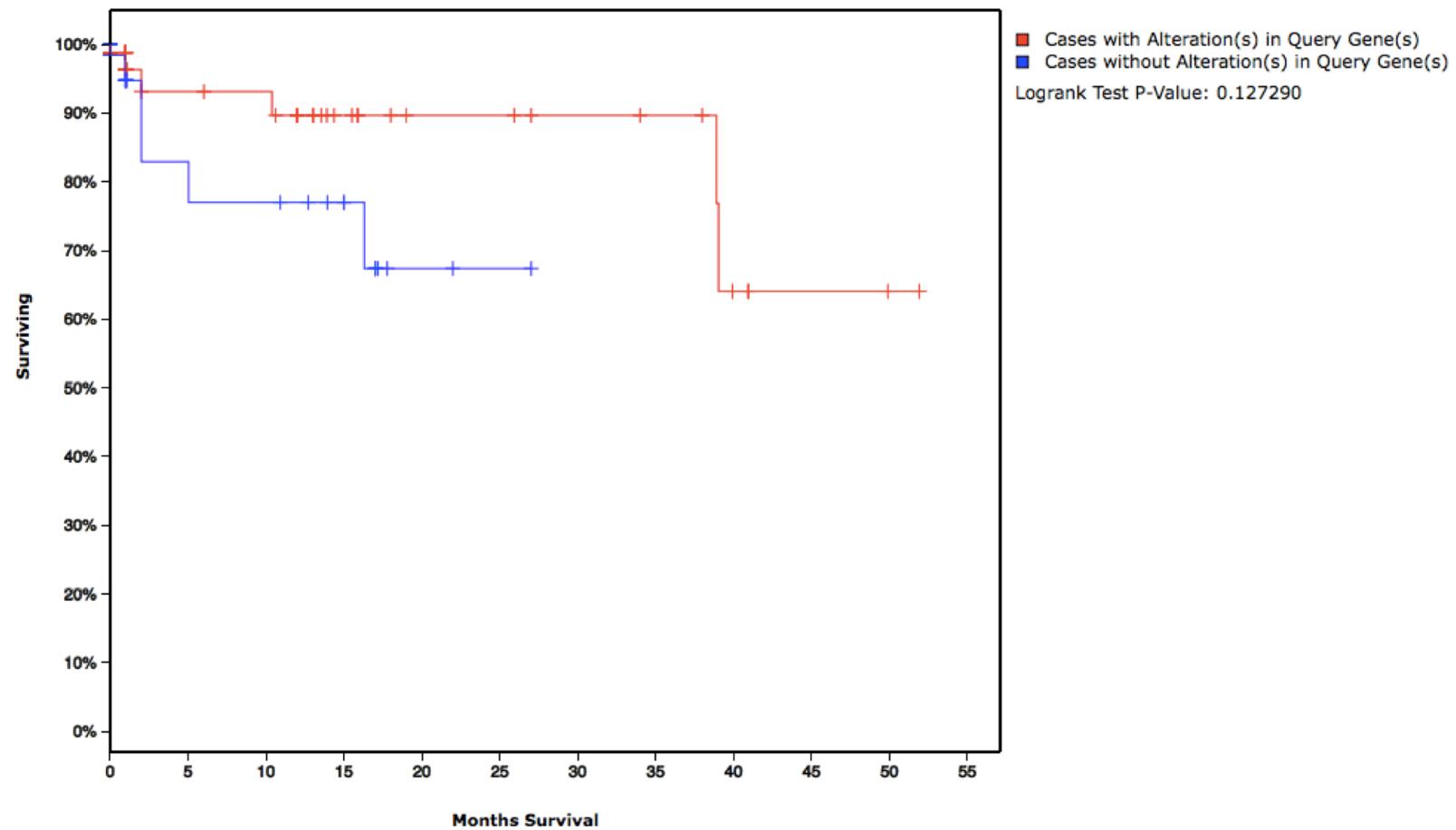


Mutual exclusivity



Kaplan-Meier Curves

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



Programmatic Interfaces

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