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

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.

dideoxynucléotides

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

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

Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
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

Gene expression signatures

Golub 1999:
microarray-based
signatire of Leukemia

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

Genomics
england

About Us | 100,000 Genomes Project

Médecine France génomique 2025

mise à jour : 19.07.17

A+ A-

Le plan « Médecine France génomique 2025 », se concrétise. En effet, le ministère chargé de la santé a lancé en décembre 2016 un appel à projets national amorçant le financement des 2 premières plateformes génomiques à visée diagnostique et de suivi thérapeutique, sur les 12 attendues dans les 5 ans. Ces équipements d'excellence illustrent le soutien constant des pouvoirs publics vis-à-vis de l'innovation médicale, en l'occurrence du séquençage à très haut débit du génome humain qui fonde la médecine génomique, dite aussi « personnalisée ».

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

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 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: Next step

- 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

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)

Memorial Sloan-Kettering
Cancer Center, USA

cBioPortal for Cancer Genomics

- 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

Annexes

- Formats de fichiers NGS

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
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

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	╞	230	µ
		39	'	71	G	103	g	135	ç	167	º	199	╟	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	╨	240	-
17	◄	49	1	81	Q	113	q	145	æ	177	▒	209	╤	241	±
18	↕	50	2	82	R	114	r	146	Æ	178	▓	210	╥	242	‗
19	‼	51	3	83	S	115	s	147	ô	179	│	211	╙	243	¾
20	¶	52	4	84	T	116	t	148	ö	180	┤	212	╘	244	¶
21	§	53	5	85	U	117	u	149	ò	181	Á	213	╒	245	§
22	▬	54	6	86	V	118	v	150	û	182	Â	214	╓	246	÷
23		55	7	87	W	119	w	151	ù	183	À	215	╫	247	¸
24	↑	56	8	88	X	120	x	152	ÿ	184	©	216	╪	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	▐	254	■
31	▼	63	?	95	_	127	⌂	159	ƒ	191	┐	223	▀	255	

Fichiers de régions

Coordonnées génomiques indiquant une région du génome

`<chromosome>:<start>-<end>`

`chr7:117465784-117715971`

Formats de régions

- BED
- GTF/GFF: annotation de features dans le génome
- SAM/BAM: alignement de reads de séquence sur le génome
- VCF: variant calling file

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

end - start = taille de la séquence