DE LA RECHERCHE À L'INDUSTRIE

cea

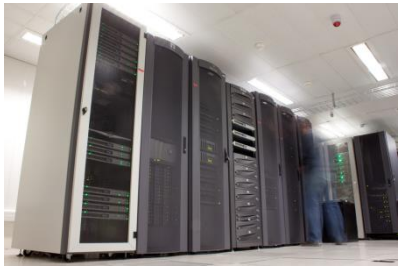
www.cea.fr

HPC EN GÉNOMIQUE

GENOSCOPE

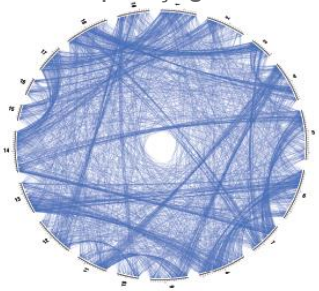
CNG CENTRE NATIONAL DE GENOTYPAGE

Système et réseau



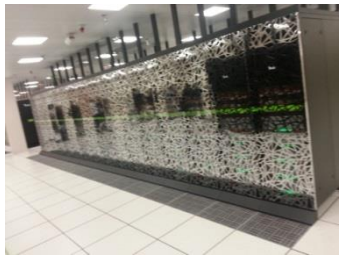
Développement et gestion de production

R&D bioinformatique et séquençage



Vitis vinifera

FRANCE GÉNOMIQUE



Claude SCARPELLI

DRF / INSTITUT DE GÉNOMIQUE / GENOSCOPE

LABORATOIRE D'INFORMATIQUE SCIENTIFIQUE

HISTORIQUE ADMINISTRATIF

1997
- Création du GIP Centre National de Séquençage (Genoscope)

1998
- Création du GIP Centre National de Génotypage (CNG)

2002
- Fusion des 2 GIP dans le GIP CNRG (Consortium National de Recherche en Génomique)

2007
- Intégration des activités du Genoscope et du CNG dans l'Institut de Génomique, 8ème institut de DSV (CEA)

DRF / IG – INSTITUT DE GÉNOMIQUE

- Production et analyse de grandes quantités de données de séquences de génomes / metagénomes (humain, plantes, microorganismes…)
- Développement d'outils bioinformatiques pour l'analyse, l'annotation et l'exploration de la séquence des génomes / metagénomes
- Exploration de la diversité biologique et biochimique du monde bactérien ; applications en chimie, bio remédiation, environnement
- Détermination des variations génétiques entre individus chez l'humain
- Par des approches de séquençage «génome entier» ou ciblé, d'études d'association, d'expression, d'épigénétique…
- Recherche des causes génétiques des maladies humaines, au travers de projets propres et en collaboration
- Interaction gènes / environnement
- Médecine de précision (personnalisée, stratifiée…)

- Activité de service à la communauté, au travers d'appels à projets sélectionnés par un conseil scientifique. Financement ANR, PIA France Génomique et LabEx GENMED
- Projets collaboratifs (financements à coûts partagés, ANR, INRA, consortia,…)
- Projets propres : fertilisation croisée recherche / production

LES UTILISATIONS DU HPC EN GÉNOMIQUE

Recherche fondamentale

- Fonctionnement des génomes («genome biology»),
- Epigénomique, transcriptomique (aussi santé).

Biomédecine et santé

- Médecine et pathologie moléculaire, diagnostic et médecine de précision (cancer)
- microbiomes humains (tube digestif, voies respiratoires, peau…),
- pathologies infectieuses

Agronomie

- Animaux d'élevage,
- Plantes

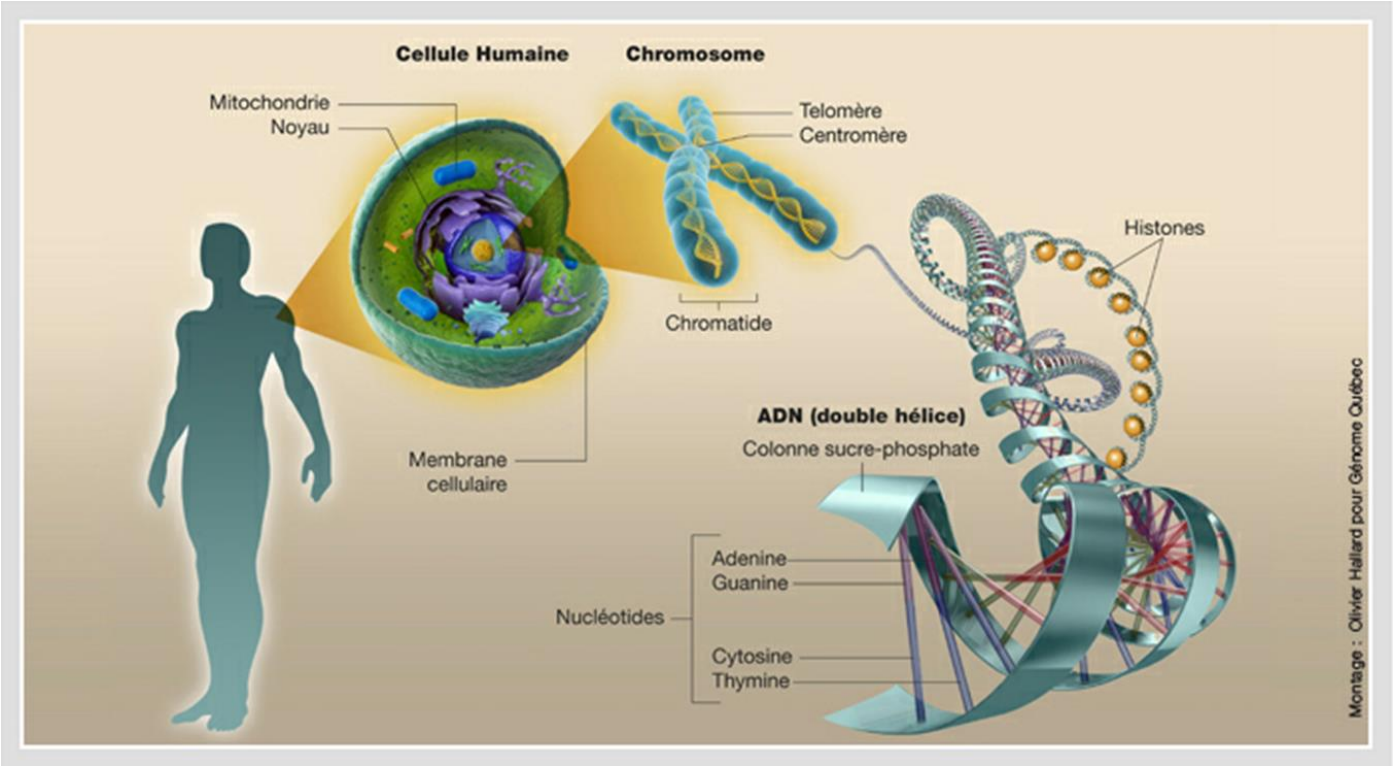
Biodiversité, environnement

- Inventaire des espèces, évolution, variabilité / polymorphisme,
- Inventaire des fonctions, familles de gènes,
- Metagénomique des milieux naturels (sols, eaux…).

GENE : BASIC UNIT OF HEREDITY

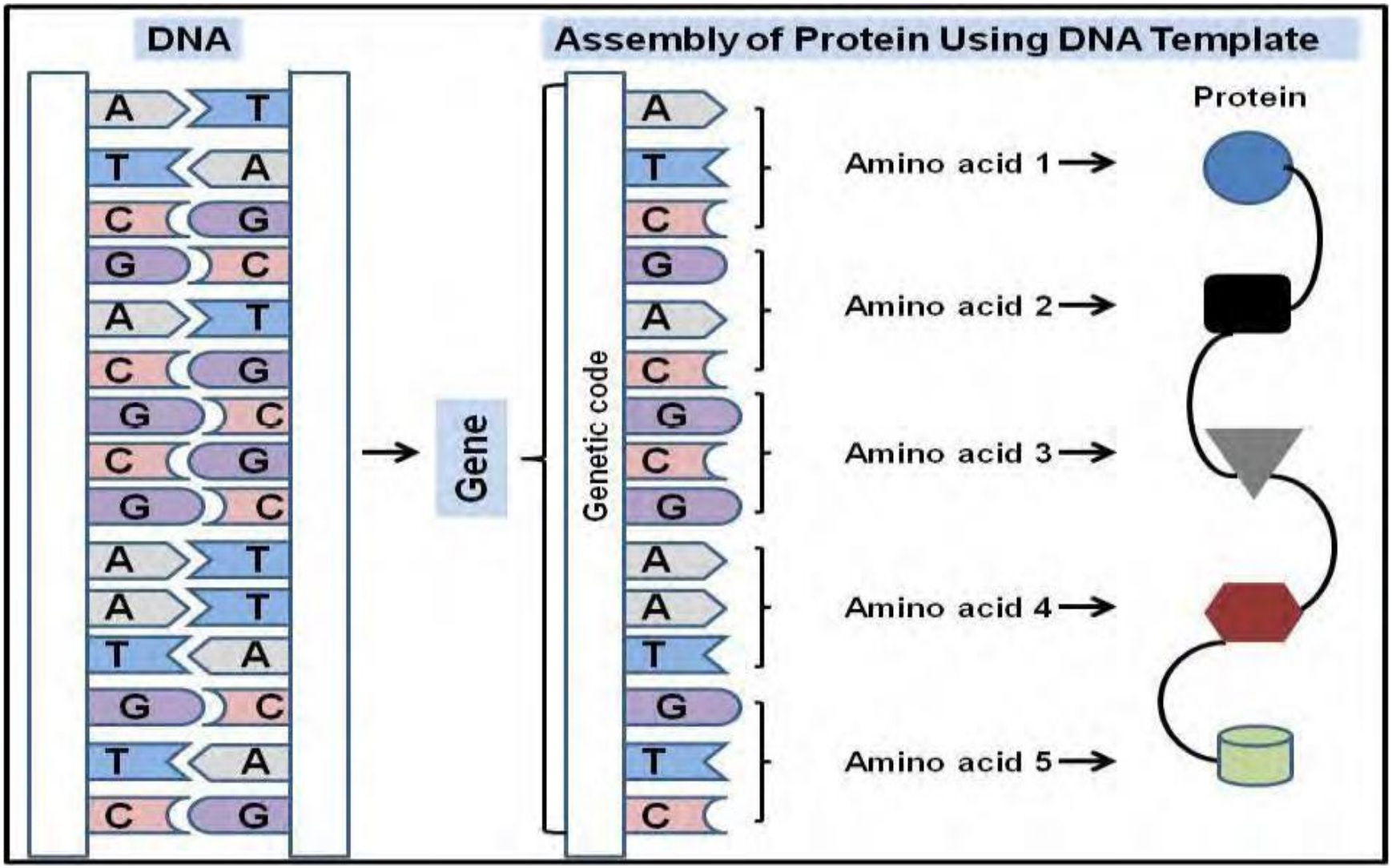
Linear sequence of DNA (deoxyribonucleic acid)

Genes → proteins → cellular functions

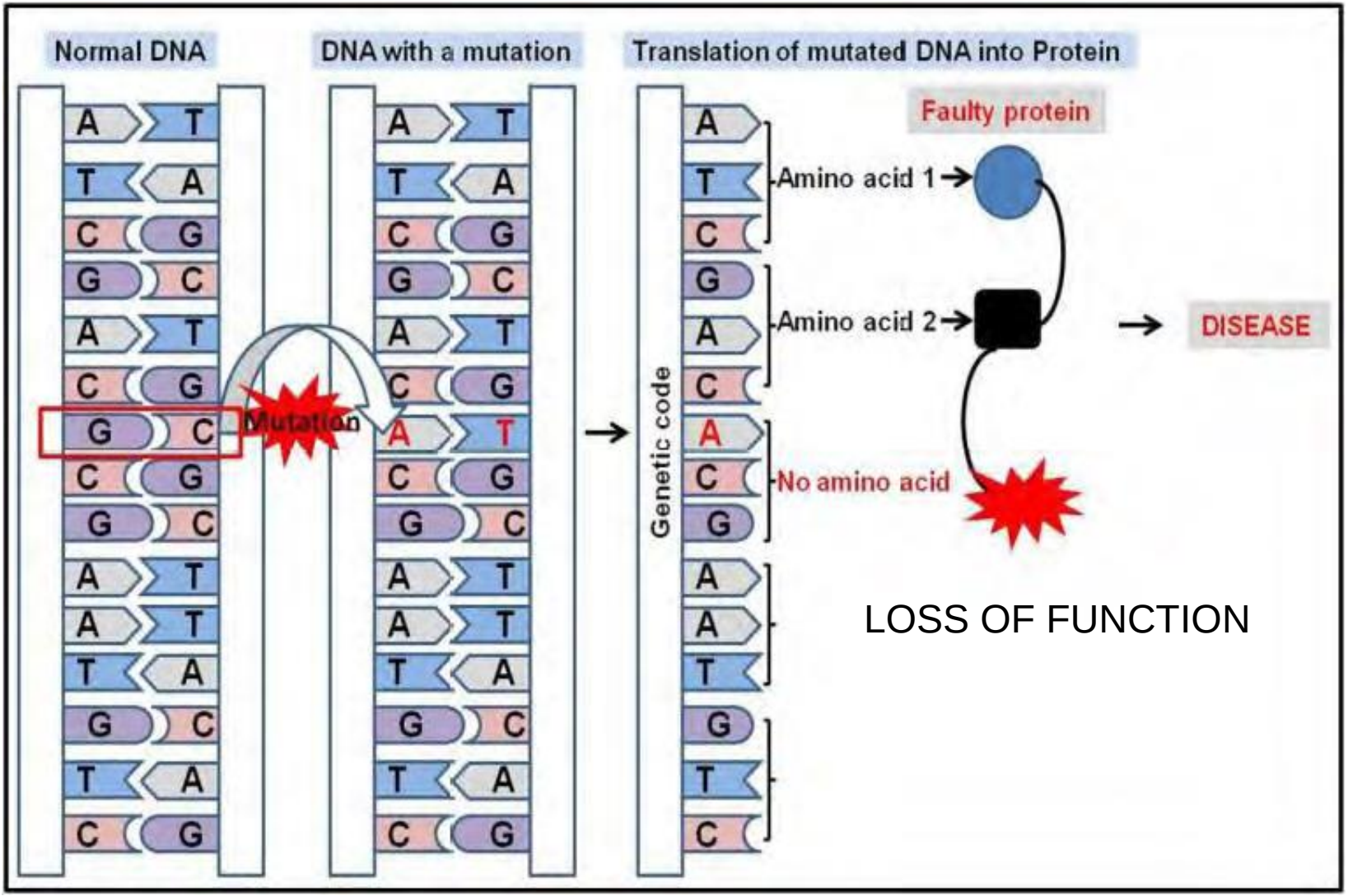


- ~ 20, 000 genes
- The entire collection of DNA is called the G**enome**

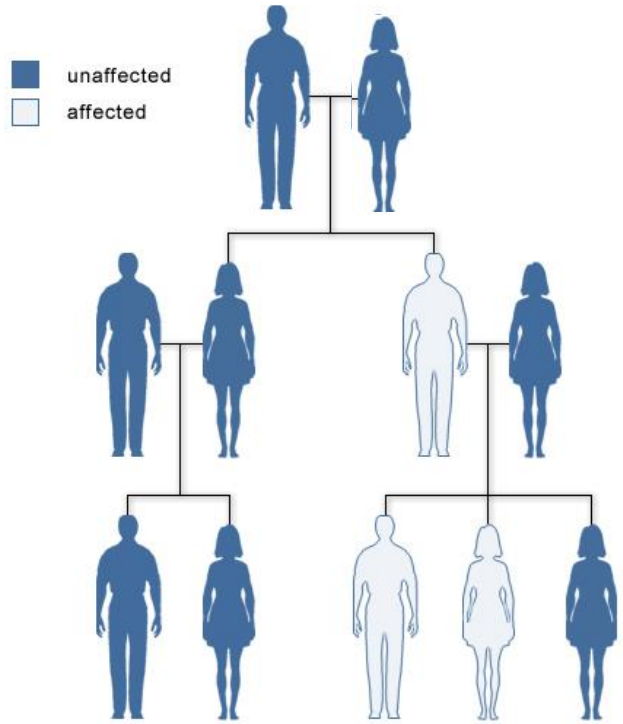
DNA, GENETIC CODE AND MAKING OF PROTEIN FOR NORMAL FUNCTION



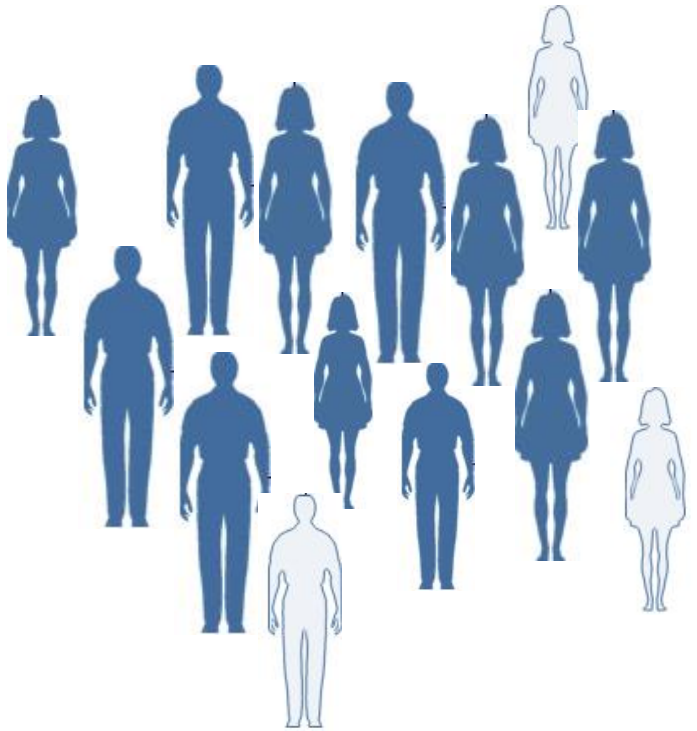
GENETIC CHANGE THAT CAN PREDISPOSE TO PATHOLOGY: SINGLE GENE



GENETIC STUDIES



U.S. National Library of Medicine



Single Gene Disorders

- **Caused by mutations in single genes**
- **Inheritance pattern (dominant, recessive, X-linked)**
- **New mutation or *de novo***

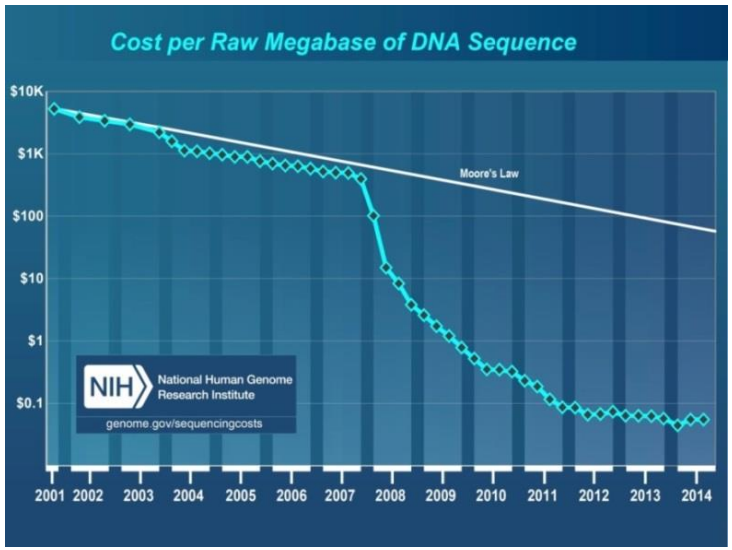
Same gene mutation may produce multiple different types or severity

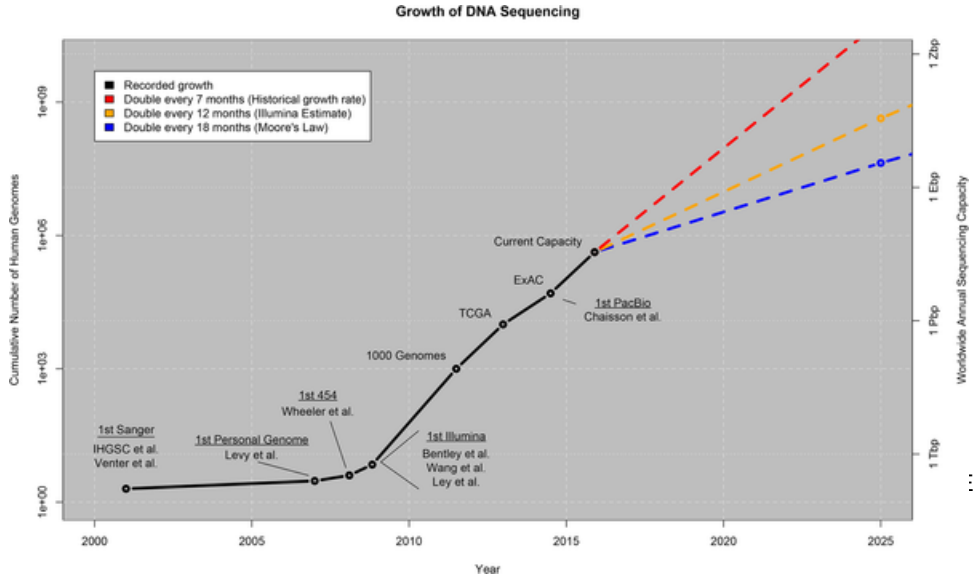
Not all individuals inheriting the affected gene will develop a pathology

More than one gene can cause the same disorder

DÉLUGE DE DONNÉES EN GÉNOMIQUE

- De 100 M$ /génome humain en 2001, on est passé à 10 M$ en 2007, 10 k$ vers 2011,
- On approche 1000 $/génome humain
- Il en a résulté une augmentation spectaculaire (dans des proportions équivalentes à la diminution du coût) du volume des données générées
- Accès à des applications inenvisageables il y a quelques années
- Les données sont le point clé :
  - Bigdata : 3(+)V (Volume, Velocity, Variability,...) + calcul distribué
  - Il faut alimenter les CPU efficacement (NFS à l'IG, Lustre au TGCC, HDFS...)





Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195 http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

DÉLUGE DE DONNÉES EN GÉNOMIQUE

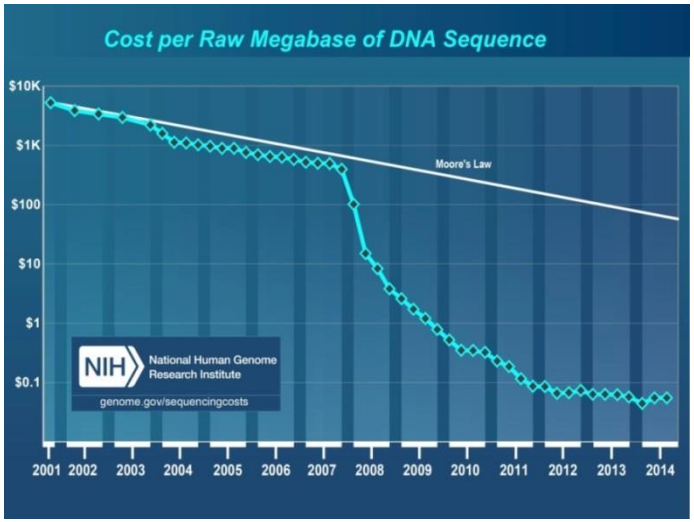
- Technologie «lectures courtes» :

2010
Hiseq 2000
200 Gbase / 10 jours

2013 (x 3)
Hiseq 2500
600 Gbase / 10 jours
200 whole genomes / an

2015 (x 6)
Hiseq 4000
1.2 Tbase / 3 jours
1.000 whole genomes / an

2016 (x …)
X5 (5 Hiseq 4000)
7,5 Tbase / 3 jours
9.000 whole genomes / an
Cout : ¼ HiSeq 2500



http://www.genome.gov/sequencingcosts/

- Technologie «longues lectures», faible encombrement

➢ Oxford Nanopore

2014-2015
10 kb+ / lecture
512 pores
Beta-test mondial
Flowcell format USB, à connecter sur portable
300 Mbase / 2 jours

Q2 2015
Vitesse de passage 30 bps -> 500 bps
20 Gbase / 2 jours

2016
512 pores -> 3.000 pores
Rack de 48 flowcells
120 Gbase / jours / FC
6 Tbase / jours

GÉNOMIQUE & BIG DATA

- Big Data: Astronomical or Genomical?, Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015), PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195, July 7th 2015

- La génomique est au niveau de ces trois autres domaines, identifiés comme des acteurs Big Data

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

ORDRES DE GRANDEUR

Institut de Génomique
2eme centre de séquençage Européen
17 machines Illumina HiSeq2000 / 2500

250 To / an

7 génomes (30X) / jour

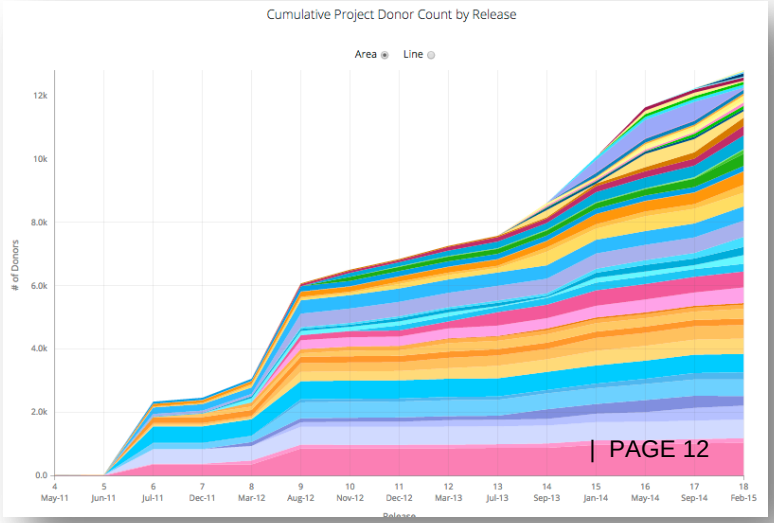
Plateforme Illumina HiSeq X Ten

1,8 Po de données /an
18,000 GH/an
1000 $/génome (30x)

Projet « 100,000 génomes » (UK)

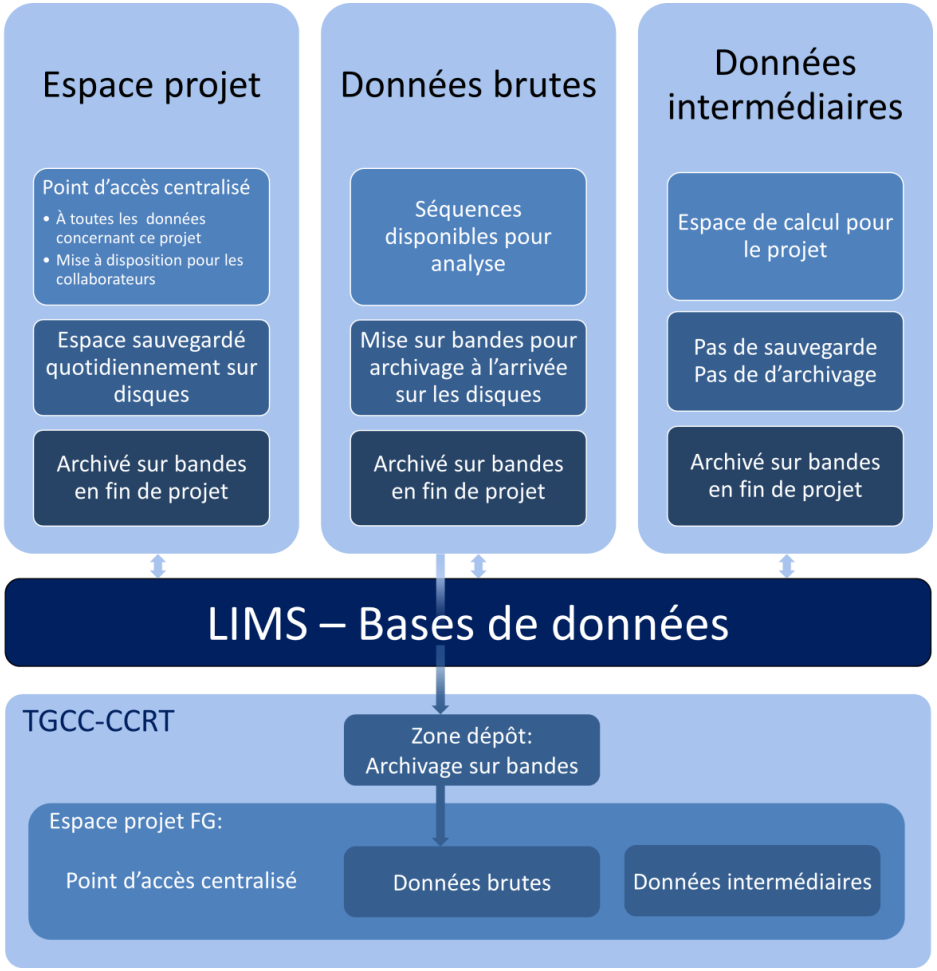
10 Po de données sur 4 ans

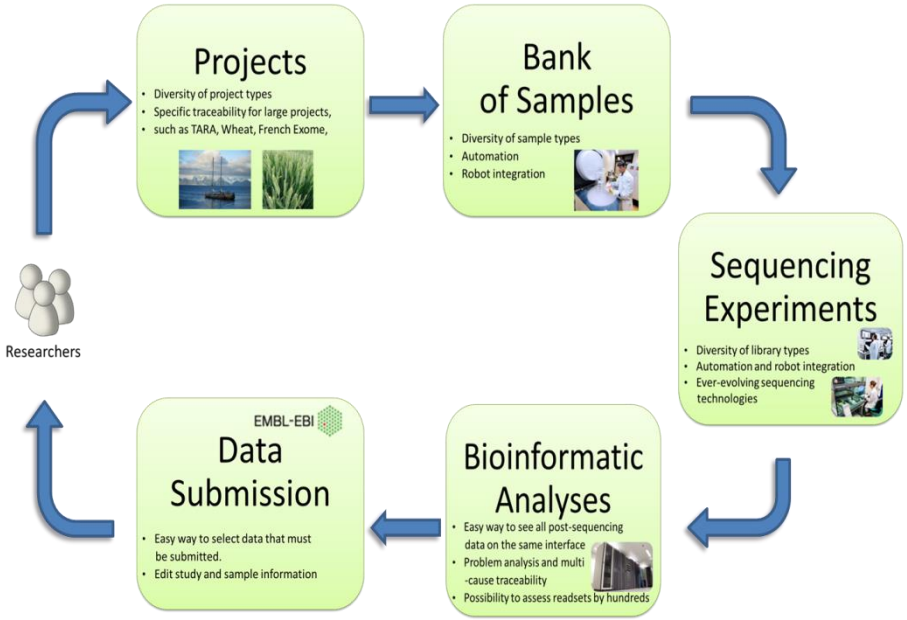
NGS de « troisième génération »?
(longues lectures, molécules uniques)



https://dcc.icgc.org/projects/history

SÉCURITÉ, TRAÇABILITÉ, DISPONIBILITÉ





APPLICATIONS

- Service à la communauté et projets propres

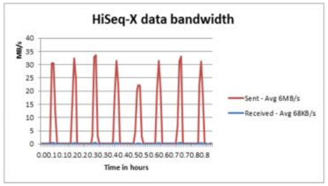
- 2 grandes catégories d'applications
  - Assemblage (parcours de graphe et construction de tables d'index) : large memory, multithread
    - typiquement : quelques semaines, ~16-32 cœurs, espace d'adressage de 100 Go à quelques To.
  - Des applications intrinsèquement parallèles, sans besoin fort de synchronisation («embarrassingly parallel») :
    - Adapter les données à ces applications (pré-traitements, découpage…),
    - Mais ratio mémoire / cœur assez important (8 Go/cœur aujourd'hui).
- généralités :
  - «ecosystème» riche et productif : 1.500 à 2.000 programmes, issus de 200 packages, rythme élevé de mise à jour,
  - besoin de workflows
  - temps de calcul longs (> 24 heures) et imprévisibles, ou du moins très variables
  - Gestion de production, de workflows, IHM (pilotage, reporting),

- évolution vers des besoins d'intégration de données hétérogènes (IHM pour l'exploration)

Partenariat structurant avec la DAM au TGCC

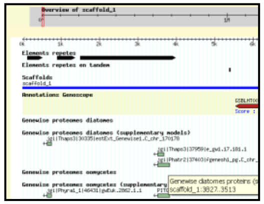
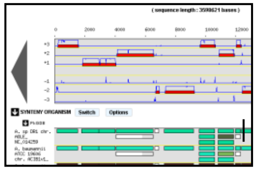
- **Analyses des besoins de la génomique à très haut débit :**
  - **Facilités (électricité, froid, charge au sol)**
  - **Calcul distribué, sécurité, entrées/sorties performantes**
- **Équipe IG dédiée en interface des équipes DAM : prise en compte des besoins spécifiques**
  - **Accumulation de données vivantes**
  - **Parallélisme par les données (cas favorable), assemblage de (meta)génomes, classification (comptage) : machines grande mémoire**
  - **Temps d'exécution peu prédictibles**
  - **Disponibilités des codes, versions multiples**
  - **Mode de travail par groupe**
- **Prochaines étapes : réinvestissement dans la machine Cobalt (engagé), stockage (2017-2018)**
- **Moyen terme : bases de données, analytique («big data»)**


Génération des données (IG)



Stockage et traitement des données (TGCC, IG)

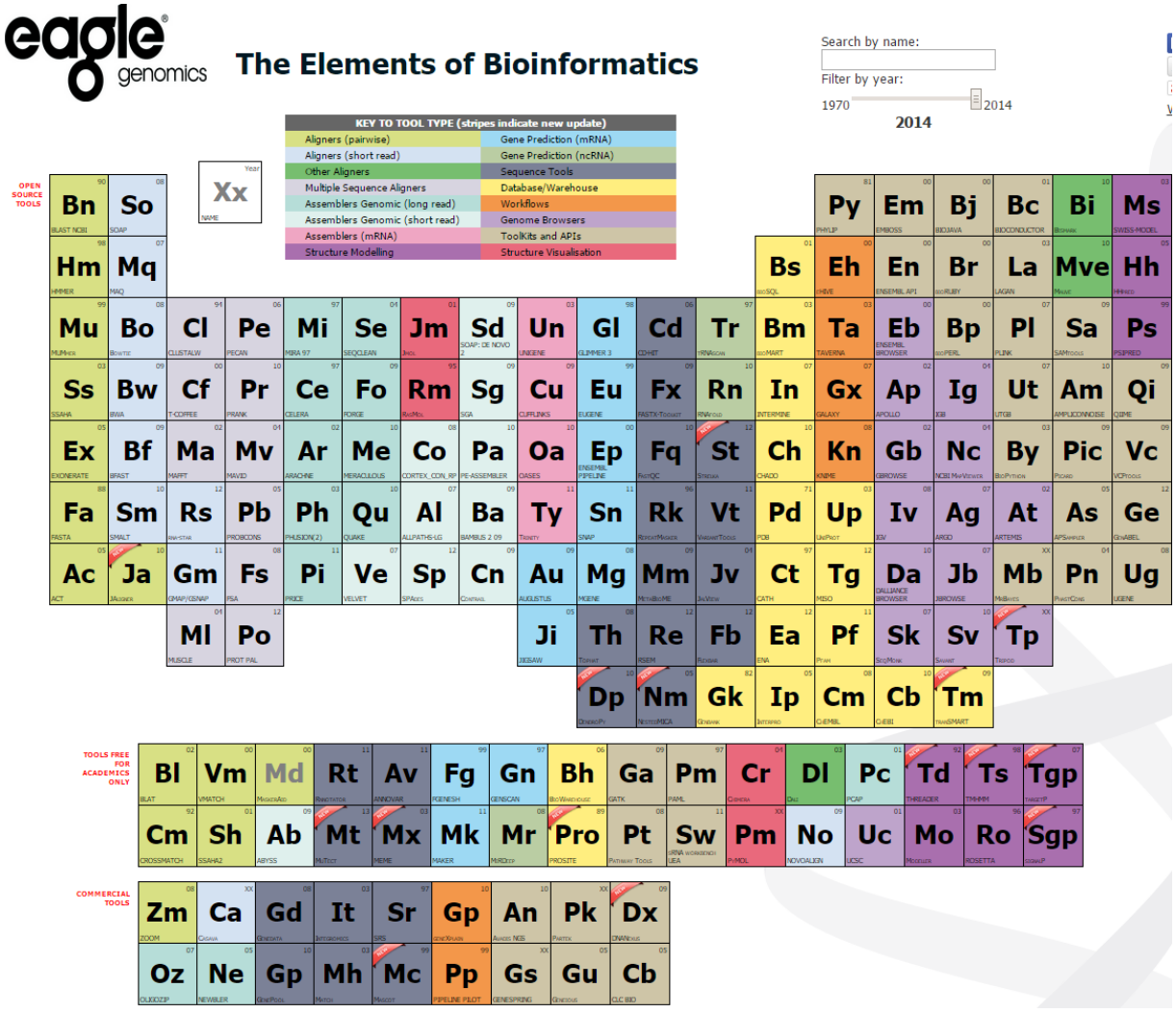
Présentation et exploration des données (IG)

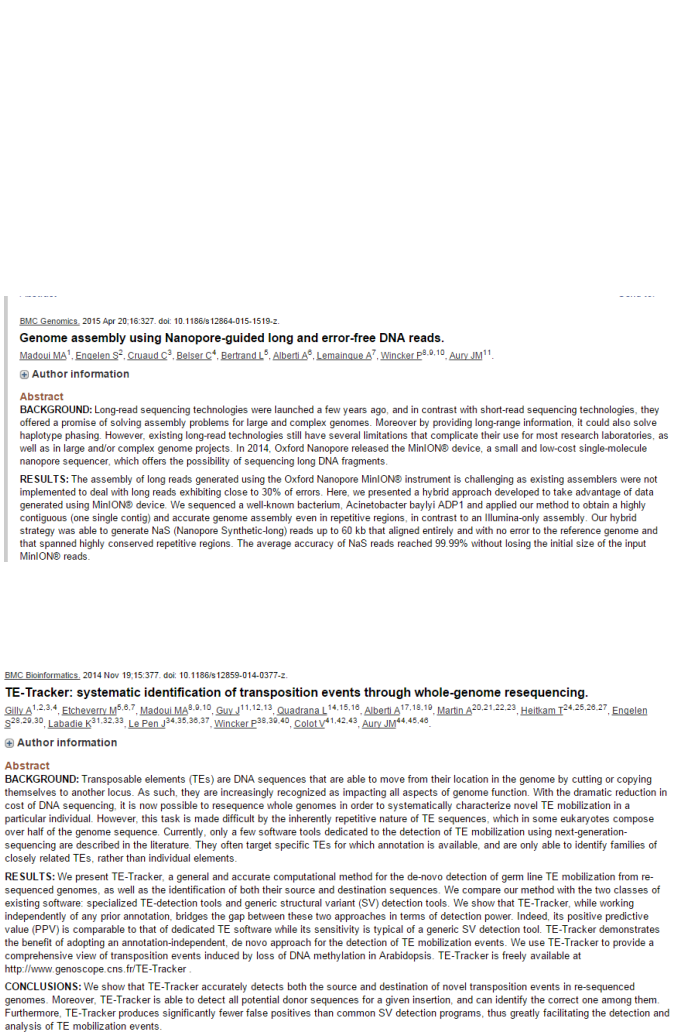




BESOINS ET CONTRAINTES SPÉCIFIQUES POUR LA GÉNOMIQUE

- Nécessité de pouvoir exécuter tous les types de programmes utilisés par la communauté génomique,
  - Durée d'exécution inconnue à priori, et pouvant dépasser la journée
  - «embarrassingly parallel» :
    - nécessité «d'outils supports» (Glost, Pegasus, systèmes «maison»)
  - Empreinte mémoire importante (> 2-4 To).
- Gros volumes de données mis en œuvre :
  - Hiseq 4000 : 2 x 1 To en 3 jours
    - Besoin de stockage vivant et à long terme, pour limiter les mouvements de données durant la vie du projet,
- Support pour le travail par groupe
  - respect de contraintes de confidentialité sur les données (notion d'espace partagé)
  - haut niveau de sécurité (données sensibles),
  - Quota de groupe
- Support pour la gestion des versions de codes, supports de versions multiples
- Publication des données par des portails,
- Activités de production ET de recherche,
- Futur : évolution vers des besoins d'intégration de données hétérogènes (IHM pour l'exploration, l'analytique)

PORTEFEUILLE DES APPLICATIONS





BMC Genomics. 2015 Apr 20;16:327. doi: 10.1186/s12864-015-1519-z.

Genome assembly using Nanopore-guided long and error-free DNA reads.

Madoui MA[1], Engelen S[2], Cruaud C[3], Belser C[4], Bertrand L[5], Alberti A[6], Lemainque A[7], Wincker P[8,9,10], Aury JM[11].

Author information

Abstract

BACKGROUND: Long-read sequencing technologies were launched a few years ago, and in contrast with short-read sequencing technologies, they offered a promise of solving assembly problems for large and complex genomes. Moreover by providing long-range information, it could also solve haplotype phasing. However, existing long-read technologies still have several limitations that complicate their use for most research laboratories, as well as in large and/or complex genome projects. In 2014, Oxford Nanopore released the MinION® device, a small and low-cost single-molecule nanopore sequencer, which offers the possibility of sequencing long DNA fragments.

RESULTS: The assembly of long reads generated using the Oxford Nanopore MinION® instrument is challenging as existing assemblers were not implemented to deal with long reads exhibiting close to 30% of errors. Here, we presented a hybrid approach developed to take advantage of data generated using MinION® device. We sequenced a well-known bacterium, Acinetobacter baylyi ADP1 and applied our method to obtain a highly contiguous (one single contig) and accurate genome assembly even in repetitive regions, in contrast to an Illumina-only assembly. Our hybrid strategy was able to generate NaS (Nanopore Synthetic-long) reads up to 60 kb that aligned entirely and with no error to the reference genome and that spanned highly conserved repetitive regions. The average accuracy of NaS reads reached 99.99% without losing the initial size of the input MinION® reads.

BMC Bioinformatics. 2014 Nov 19;15:377. doi: 10.1186/s12859-014-0377-z.

TE-Tracker: systematic identification of transposition events through whole-genome resequencing.

Gilly A[1,2,3,4], Etcheverry M[5,6,7], Madoui MA[8,9,10], Guy J[11,12,13], Quadrana L[14,15,16], Alberti A[17,18,19], Martin A[20,21,22,23], Heitkam T[24,25,26,27], Engelen S[28,29,30], Labadie K[31,32,33], Le Pen J[34,35,36,37], Wincker P[38,39,40], Colot V[41,42,43], Aury JM[44,45,46].

Author information

Abstract

BACKGROUND: Transposable elements (TEs) are DNA sequences that are able to move from their location in the genome by cutting or copying themselves to another locus. As such, they are increasingly recognized as impacting all aspects of genome function. With the dramatic reduction in cost of DNA sequencing, it is now possible to resequence whole genomes in order to systematically characterize novel TE mobilization in a particular individual. However, this task is made difficult by the inherently repetitive nature of TE sequences, which in some eukaryotes compose over half of the genome sequence. Currently, only a few software tools dedicated to the detection of TE mobilization using next-generation-sequencing are described in the literature. They often target specific TEs for which annotation is available, and are only able to identify families of closely related TEs, rather than individual elements.

RESULTS: We present TE-Tracker, a general and accurate computational method for the de-novo detection of germ line TE mobilization from re-sequenced genomes, as well as the identification of both their source and destination sequences. We compare our method with the two classes of existing software: specialized TE-detection tools and generic structural variant (SV) detection tools. We show that TE-Tracker, while working independently of any prior annotation, bridges the gap between these two approaches in terms of detection power. Indeed, its positive predictive value (PPV) is comparable to that of dedicated TE software while its sensitivity is typical of a generic SV detection tool. TE-Tracker demonstrates the benefit of adopting an annotation-independent, de novo approach for the detection of TE mobilization events. We use TE-Tracker to provide a comprehensive view of transposition events induced by loss of DNA methylation in Arabidopsis. TE-Tracker is freely available at http://www.genoscope.cns.fr/TE-Tracker .

CONCLUSIONS: We show that TE-Tracker accurately detects both the source and destination of novel transposition events in re-sequenced genomes. Moreover, TE-Tracker is able to detect all potential donor sequences for a given insertion, and can identify the correct one among them. Furthermore, TE-Tracker produces significantly fewer false positives than common SV detection programs, thus greatly facilitating the detection and analysis of TE mobilization events.

MISE À DISPOSITION DES CODES

- «standard» Linux system, Slurm resources manager
- «Modules» software management system, to deal with multiple versions,
- 220+ software installed (genomic field), including many that depends on interpreters (Python, Perl, R…) whose versions may be incompatible

abyss	bowtie	dazz_db	gaze	kraken	ncbi-c++	primer3	sambamba	tabix
allpathslg	bowtie2	delly	gem	kseq	ncbi-tools	prodigal	samblaster	tophat
amos	bpipe	dendropy	geneid	last	ngsplot	pybedtools	samtools	trf
ant	breakdancer	dindel	glibc	lbzip2	parallel	pydnase	saxon	trimgalore
bam-readcount	bsseeker	eigen	graphlan	libgd	pari	pyloh	scala	trimmomatic
bamtools	bwa	eigensoft	graphmap	libmatheval	patchelf	pysam	screed	trinityrnaseq
bcftools	bx-python	elprep	h5py	lumpy	pbcore	pysnptools	seqan	ucsc-tools
bedops	ca-pbcr	emboss	h5utils	macs	pbsuite	pytables	seqtk	varscan
bedtools	cegma	est_genome	happy	maq	pbzip2	python	sicer	vcftools
besst	checkm	exonerate	hmmer	maqview	pegasus	qualimap	skewer	velvet
besst_rna	circos	f-seq	homer	methylkit	perl5-common	r_pore	smalt	wise
binutils	cnvnator	falcon	htseq	mhap	phantompeak	ray	snpeff	wkhtmltopdf
bioconductor	control-freec	fastaq	igv	muParser	picard-tools	reforge	soapdenovo2	xslthl
bioperl-live	cramtools	fastqc	io_lib	multitail	pigz	repeatmasker	spades	xz
biopython	crossmap	fop	isaac	mummer	pindel	rna-seqc	sparsehash	
bismark	cufflinks	freebayes	jcvi	mutect	platypus	rnaseq-mats	ssaha2	+150 packages python
blasr	cutadapt	freedup	jellyfish	nanocorrect	poa	rpy2	star	+214 packages R/Bioc
blast+	daligner	gatk+queue	k-mer-tools	nanook	poretools	rsem	subread	+114 modules Perl
blat	danpos	gatk	khmer	nanopolish	pplacer	ruffus	svdetect	

- «Modules» also used for users directory selection (one user / multiple projects)

```
$ module load extenv/fg                (loads FG extension, ie modules shown above)
$ module load dfldatadir/fgXXXX        (loads FG specific variables, for easier access to per-project data directory)
$ module avail
$ module load samtools                 (loads default version of samtools)
$ module load vcftools/0.1.12          (loads a specific version of samtools)
```

VARSCOPE: AN EXAMPLE OF A RUNNING WORKFLOW

- Varscope: CNG's WGS pipeline, used at production level. Mix of "embarrassingly parallel" (ie map/reduce) paradigm and multi-threads. See the various parallelism level at each steps **1. Varscope: High throughput calibrated process**

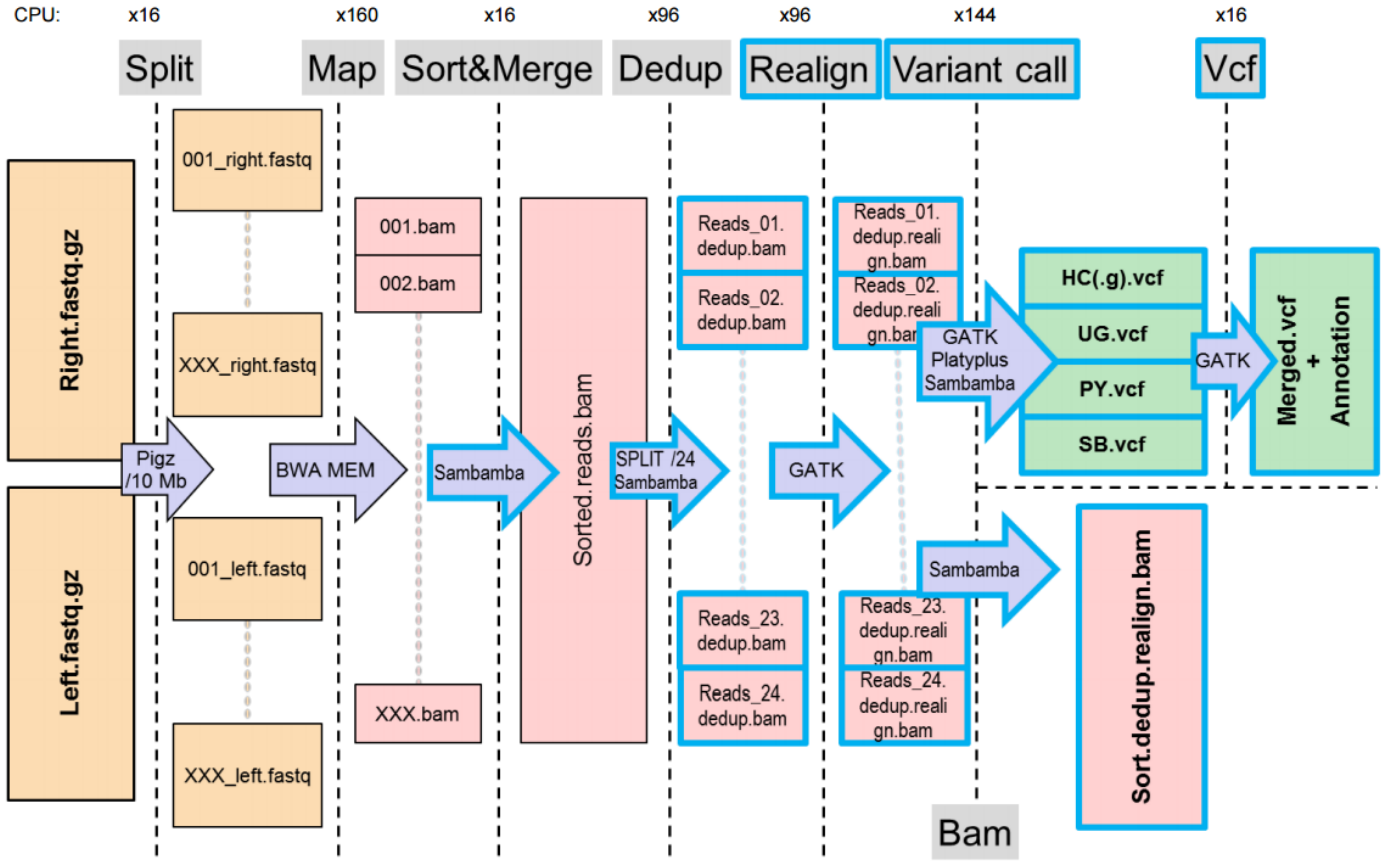
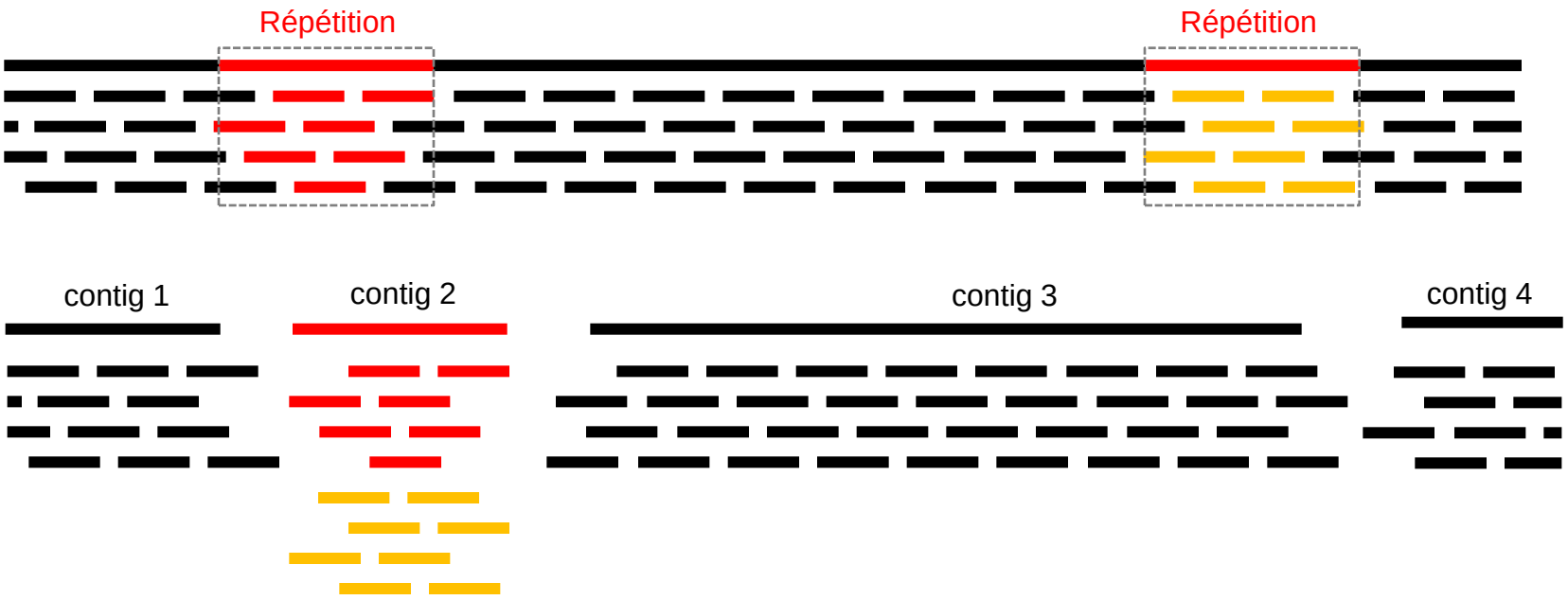


Fig. 1: Overview of the CNG/CCRT mapping and variant calling pipeline organization. Last developments are highlighted in blue

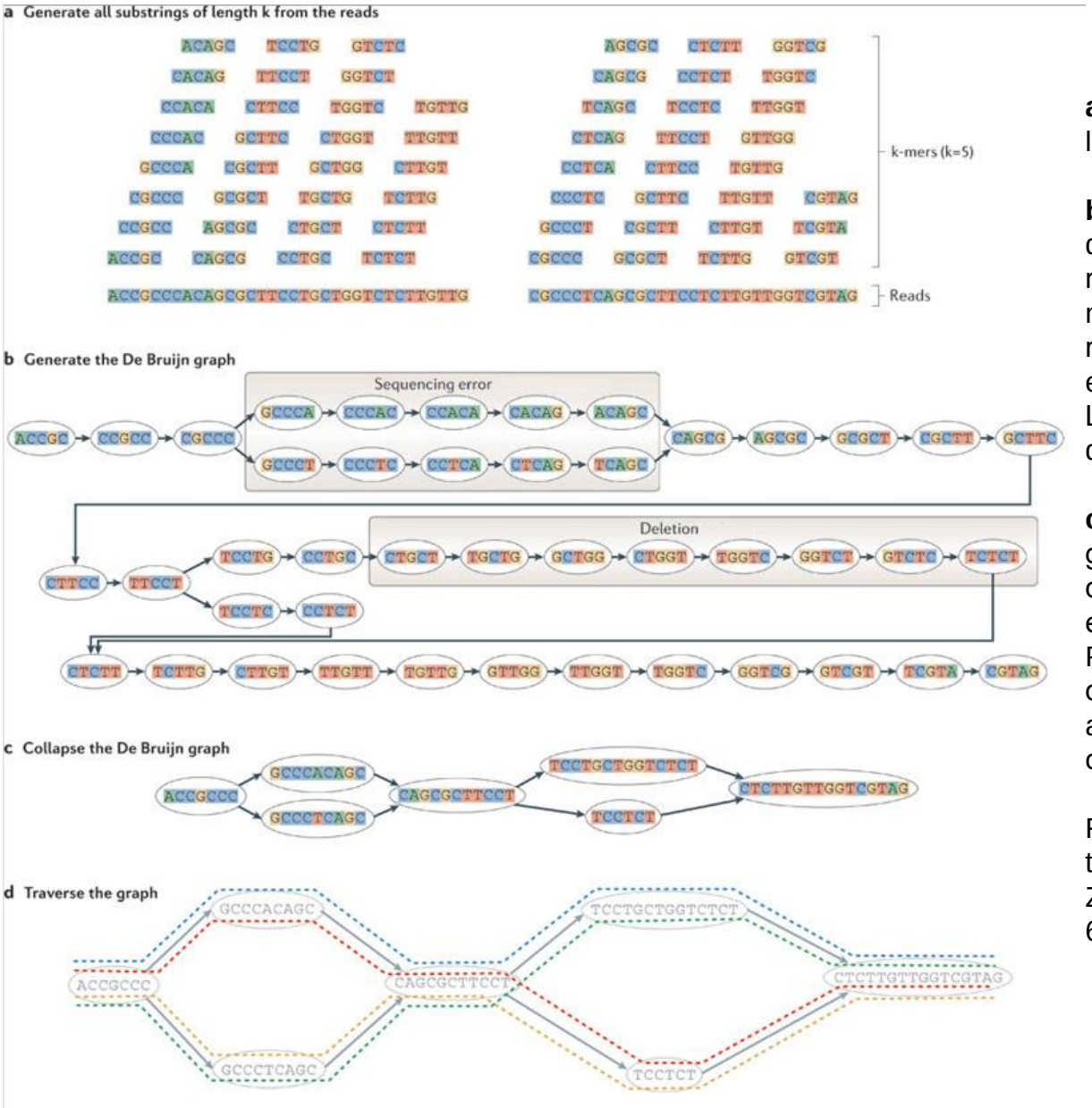
L'assemblage de génomes complexes

- Evaluation des capacités de chaque type de séquenceur, en mettant l'accent sur l'exploitation de ces données pour les applications du Genoscope, par exemple l'assemblage de novo

- L'assemblage d'un génome est une étape cruciale dans un projet d'analyse de génome et nécessite une expertise en biologie (spécificité du génome) mais aussi en informatique (consommation importante de ressources : CPU, mémoire)



Assemblage de génomes : graphes de De Bruijn



a | Les k-mers sont générés à partir des lectures, ici k=5.

b | Chaque k-mer unique représente un nœud dans le graphe de De Bruijn et une paire de nœuds est connectée si, en décalant un k-mer d'un caractère, on obtient deux (k-1)-mers identiques. L'exemple montre une erreur de séquençage (A/T) et une délétion. Les différences d'une base crées des 'bulles' de taille k dans le graphe.

c,d | Les chaines de nœuds adjacent dans le graphe sont collapsées en un unique nœud, quand le 1er nœud a un degré de sortie de 1 et le second nœud un degré d'entrée de 1. Pour finir, le parcours du graphe permet d'extraire la séquence finale, ici à cause des ambiguïtés nous obtiendrons 7 séquences correspondant à chaque nœud du graphe.

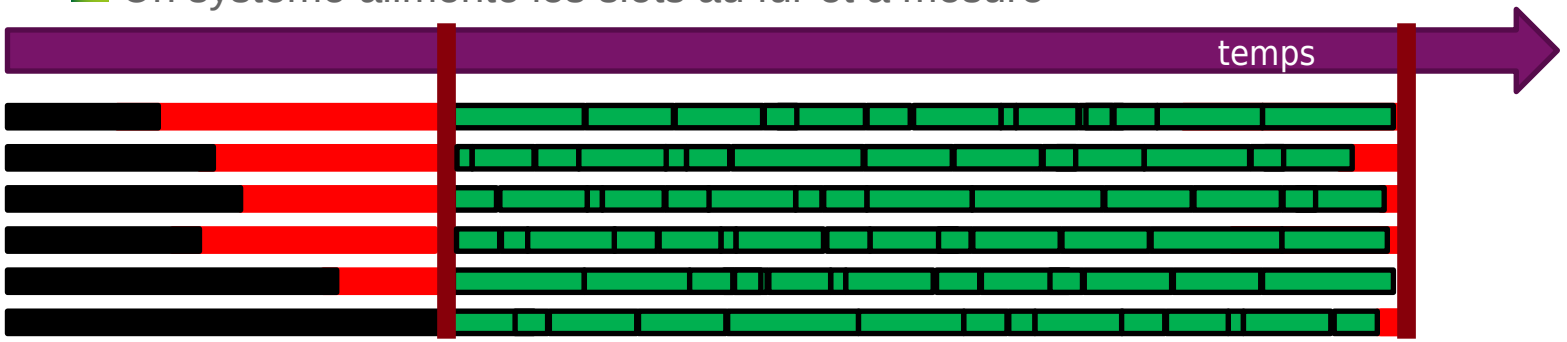
Figure extraite de l'article : Next-generation transcriptome assembly, Jeffrey A. Martin & Zhong Wang, Nature Reviews Genetics 12, 671-682 (2011).

UN EXEMPLE DE PIÈGE...

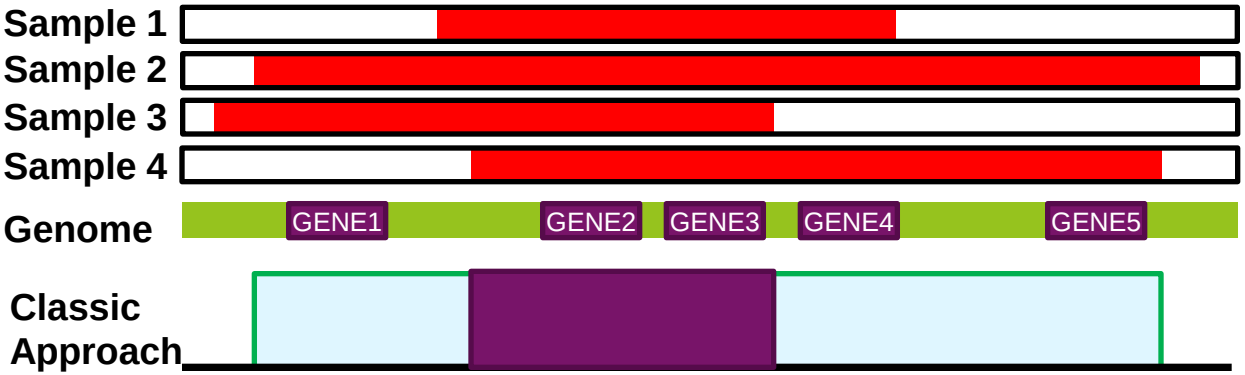
- «embarrassingly parallel» «naïf»
  - Le temps d'exécution des taches unitaires est variable

Temps d'exécution «perdu» (exception : backfill scheduling)

- «feed the slots»
  - La réservation est déterminée par le délai de restitution
  - Un système alimente les slots au fur et à mesure



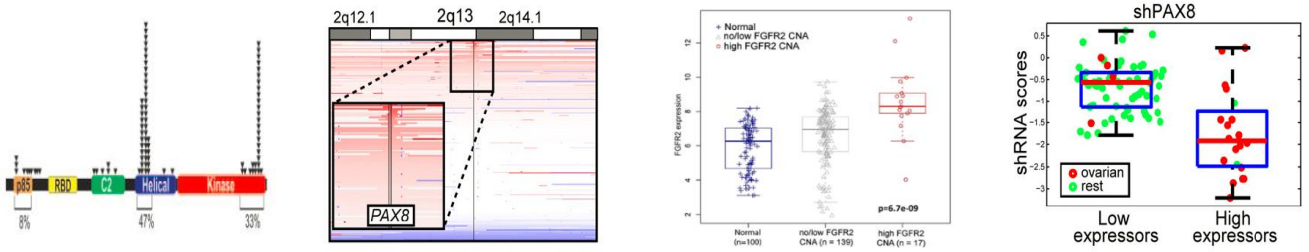
INTEGRATIVE APPROACH



deterministic 0/1 decision

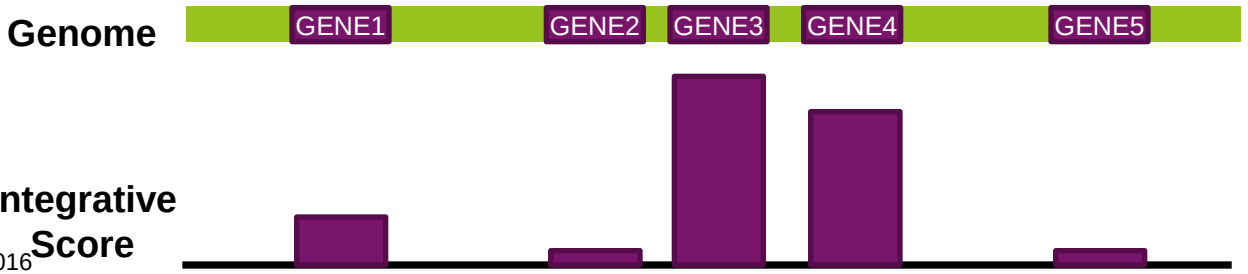
Classic Approach

Features



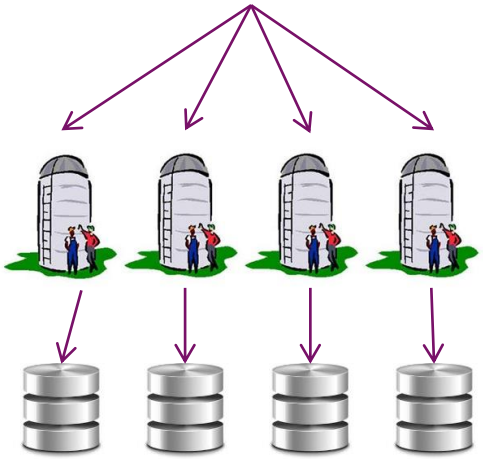
Sequence | Copy Number | Expression | shRNA

Weight and combine

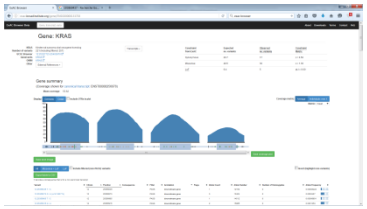


French Exac Genomics Integration

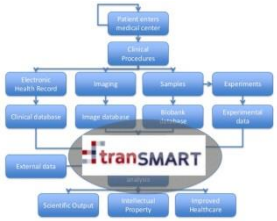
Projets



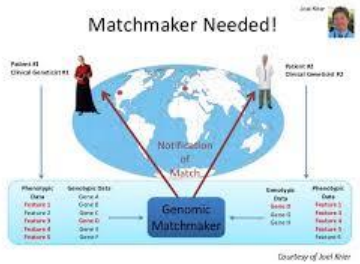
- **Sharing = Aggrégation**
  - Filtration
  - Burden studies

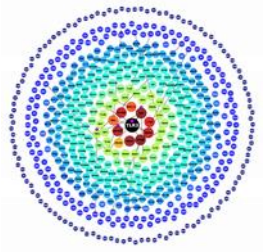


- **Integration :**
  - Genomics <-> Clinics



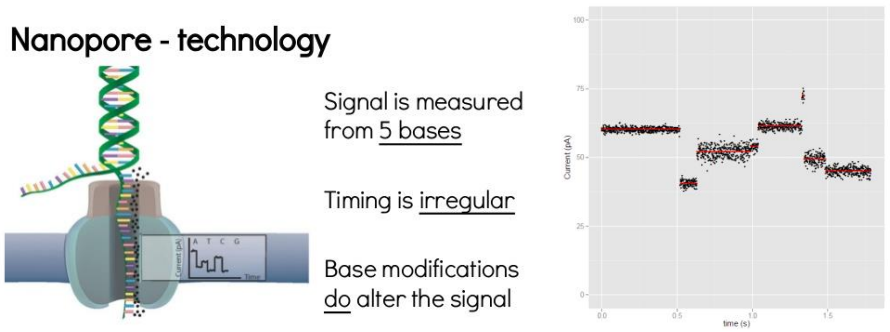
New Services





QUELQUES CHALLENGES PLURI-DISCIPLINAIRES

- Traitement du signal, micro fluidique



- Confidentialité
  - Projet PPGEN (couplage DSV/DRT) : cryptographie homomorphe (DRT/LIST, R. SIRDEY)

- Cloud computing
  - Minimiser les mouvements de données,
  - Big data analytics
  - Machine learning, approche supervisées et non supervisées

DE LA RECHERCHE À L'INDUSTRIE

cea

QUESTIONS ?