https://ngstc.iutms.umontpellier.fr/formations/m2bs/TP-rnaseq/

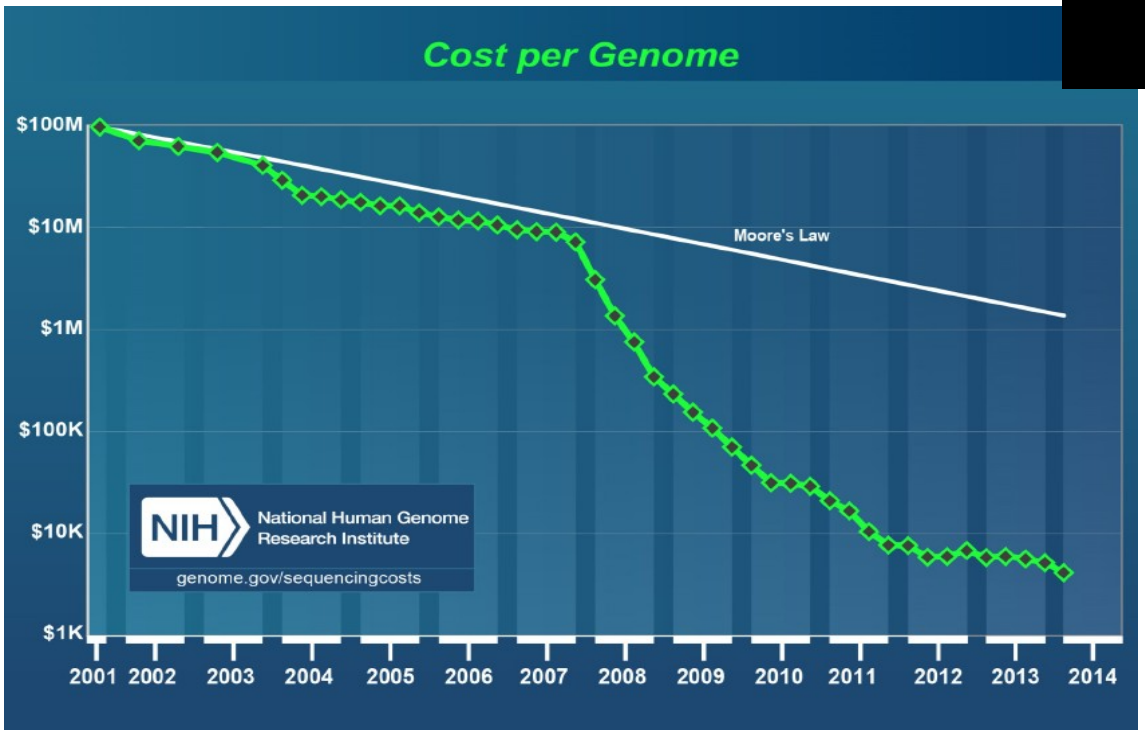
L’ère NGS….

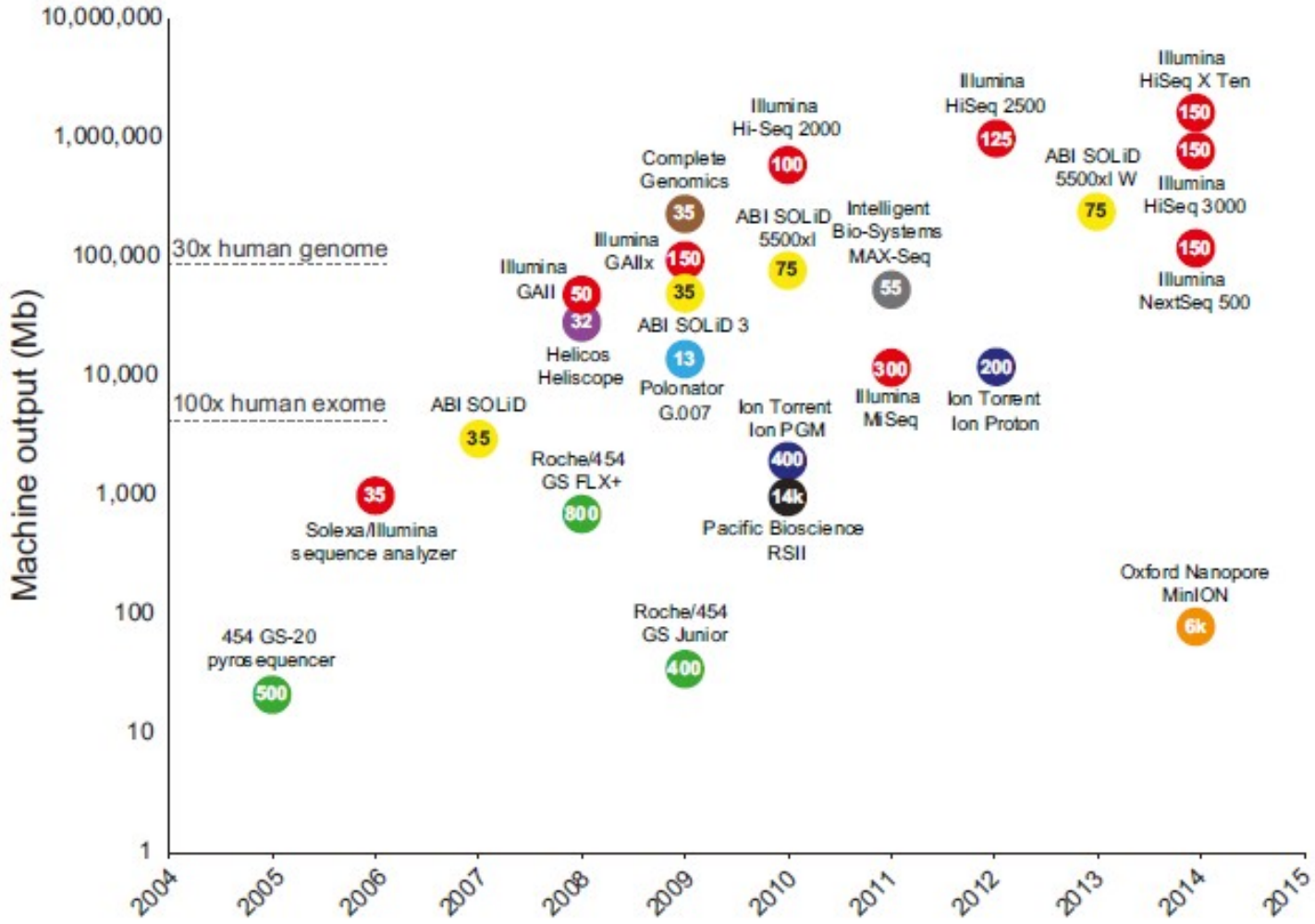
NGS et Transcriptomics

- Introduction: l'ére NGS et état des lieux, principaux séquenceurs (vu en intro)
- Les enjeux de l'analyse, qqls applications en diagnostic (ADN/ARN)
- Transcriptome et RNAseq, les questions biologiques
  - ✓ Les différentes stratégies d'analyse
  - ✓ Le mapping, exemple et limites
  - ✓ Annotations génomiques, transcriptome de référence
  - ✓ Comment passer à l'échelle

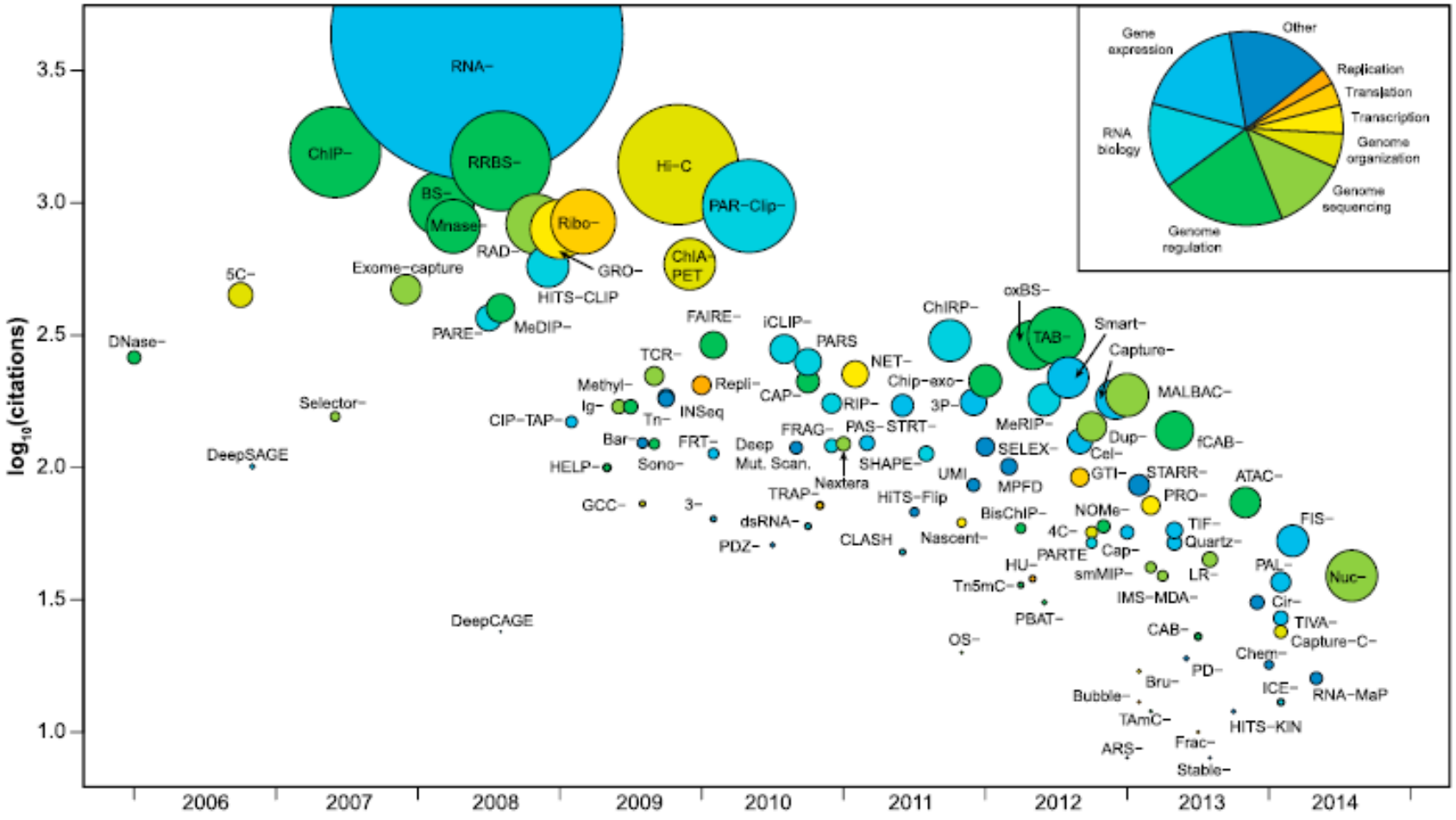
NGS: une nouvelle ère

- Coût (qqls milliers d'euros)
- Temps (Un génome humain en quelques jours)
- Puissance (Milliards bases en quelques heures)

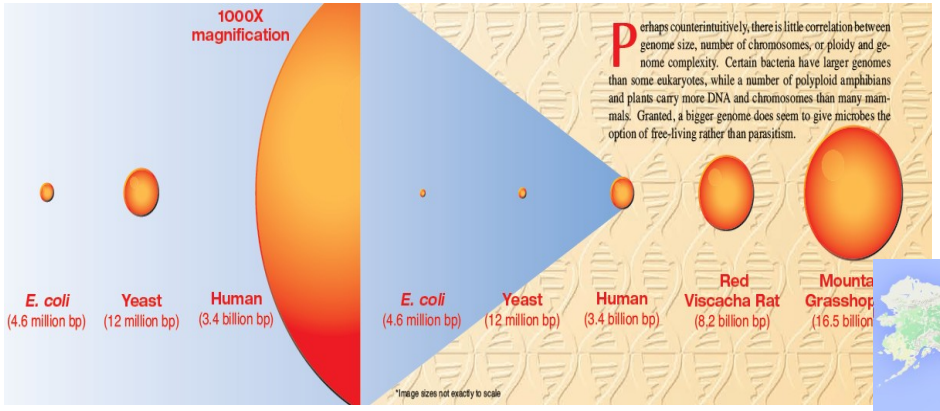




1-Explosion des séquenceurs NGS



2-Explosion des applications -Seq



IGSR and the 1000 Genomes Project

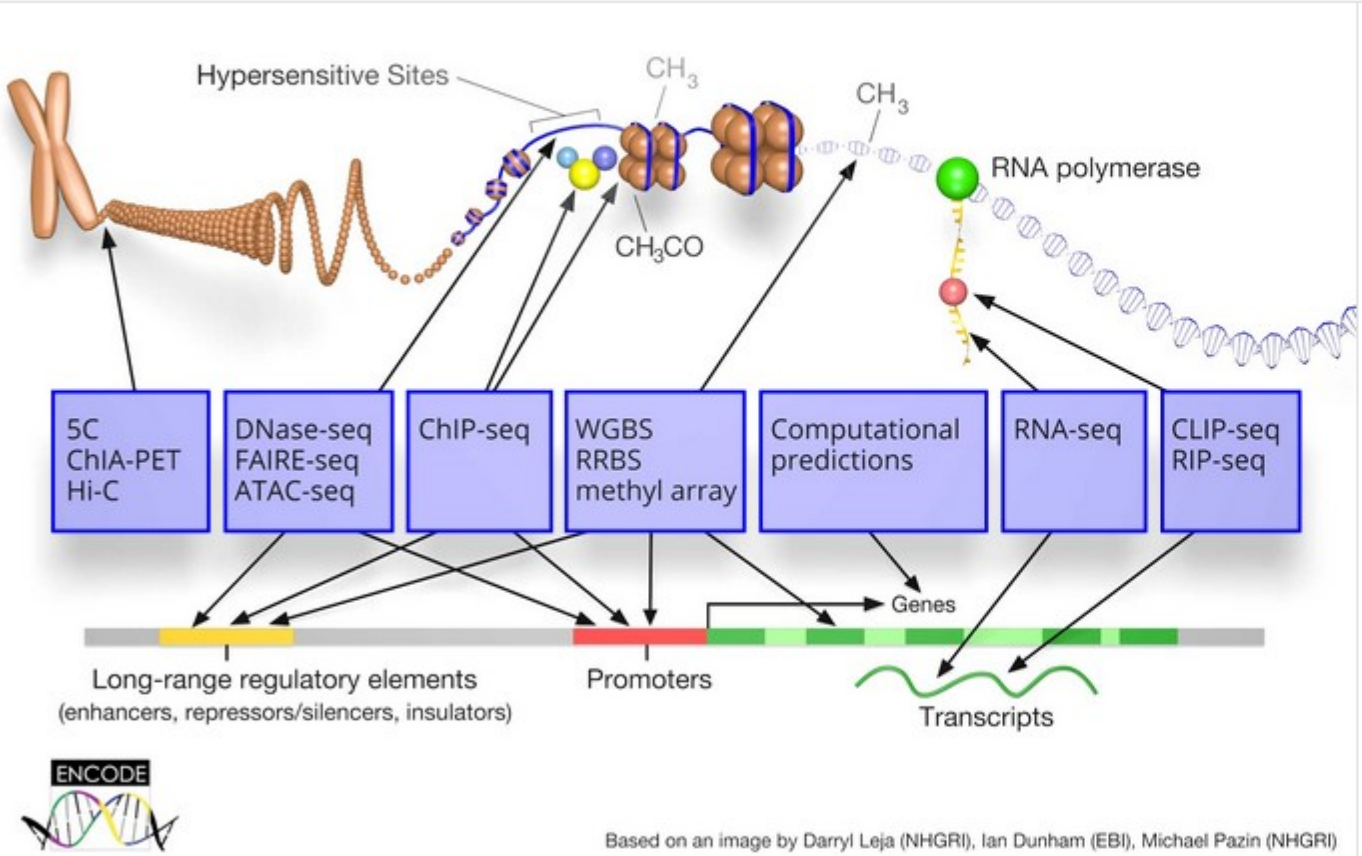
International network of cancer genome projects

The International Cancer Genome Consortium*

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Systematic studies of more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies.

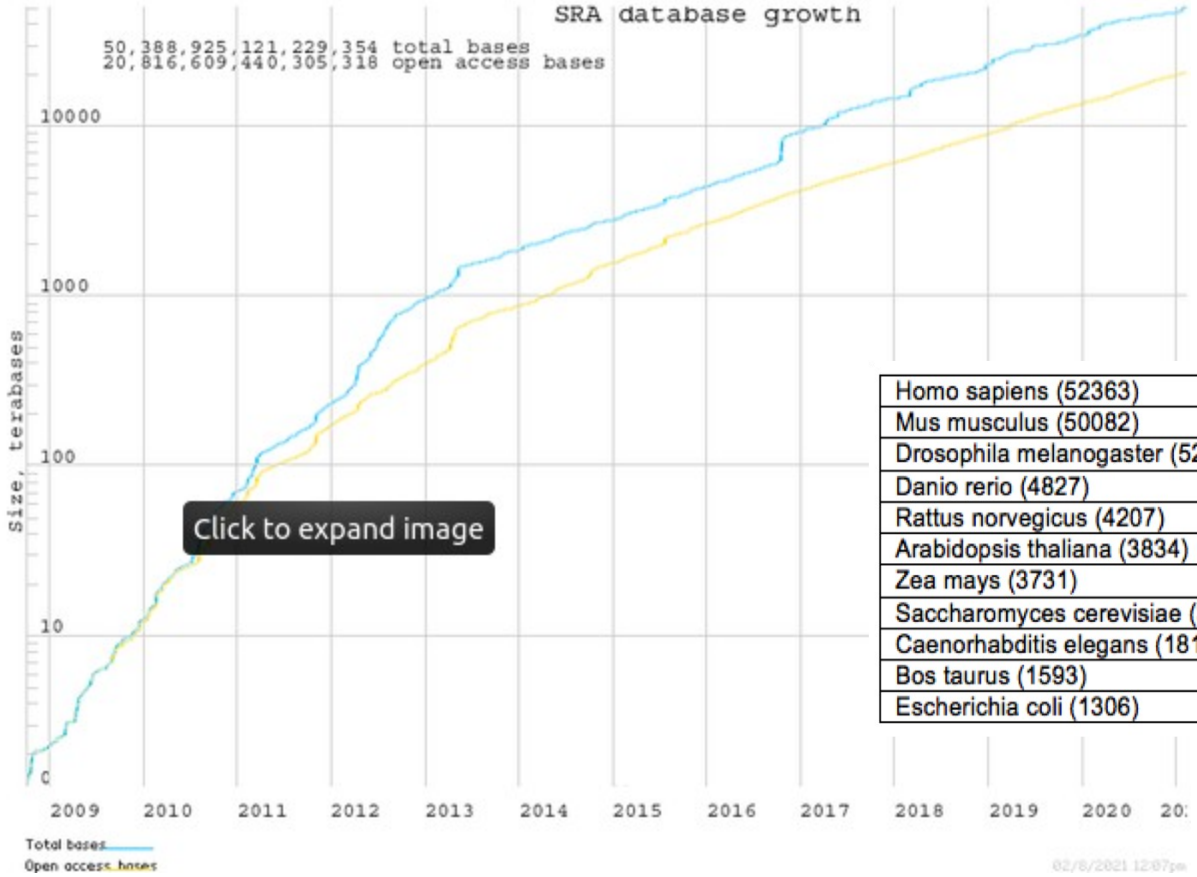
3-Explosion des grands programmes de séquençage

ENCODE: Encyclopedia of DNA Elements



HUMAN | MOUSE | WORM | FLY

3-Explosion des grands programmes de séquençage

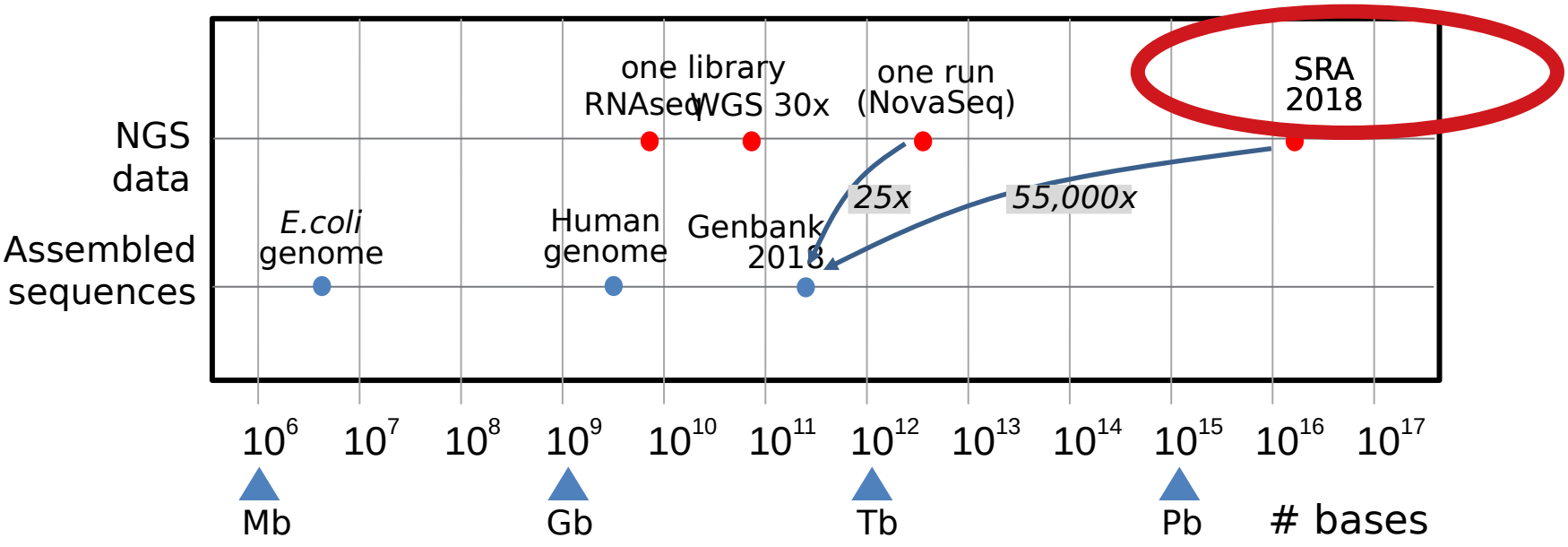


Click to expand image

Homo sapiens (52363)	Oryza sativa (1210)
Mus musculus (50082)	Macaca mulatta (942)
Drosophila melanogaster (5256)	Gallus gallus (904)
Danio rerio (4827)	Zaire ebolavirus (867)
Rattus norvegicus (4207)	human metagenome (861)
Arabidopsis thaliana (3834)	Glycine max (811)
Zea mays (3731)	Mycobacterium tuberculosis (780)
Saccharomyces cerevisiae (3321)	Solanum lycopersicum (761)
Caenorhabditis elegans (1814)	Equus caballus (759)
Bos taurus (1593)	All other taxa (53604)
Escherichia coli (1306)	

Top20 des données RNA-seq dans SRA par espéces

4-Explosion des données publiques



4-Explosion des données publiques

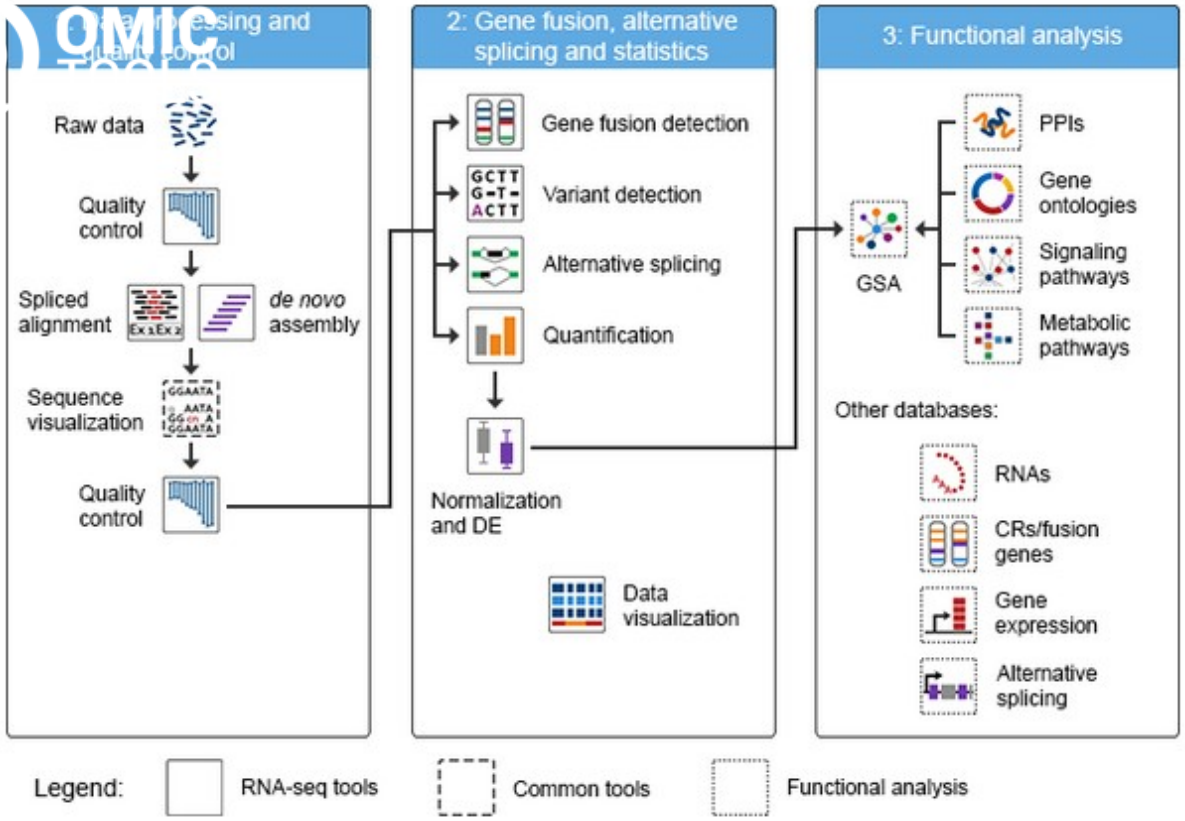
Transcriptome and RNAseq

- Public datasets in SRA

Homo sapiens (52363)	Saccharomyces cerevisiae (3321)	Zaire ebolavirus (867)
Mus musculus (50082)	Caenorhabditis elegans (1814)	human metagenome (861)
Drosophila melanogaster (5256)	Bos taurus (1593)	Glycine max (811)
Danio rerio (4827)	Escherichia coli (1306)	Mycobacterium tuberculosis (780)
Rattus norvegicus (4207)	Oryza sativa (1210)	Solanum lycopersicum (761)
Arabidopsis thaliana (3834)	Macaca mulatta (942)	Equus caballus (759)
Zea mays (3731)	Gallus gallus (904)	All other taxa (53604)

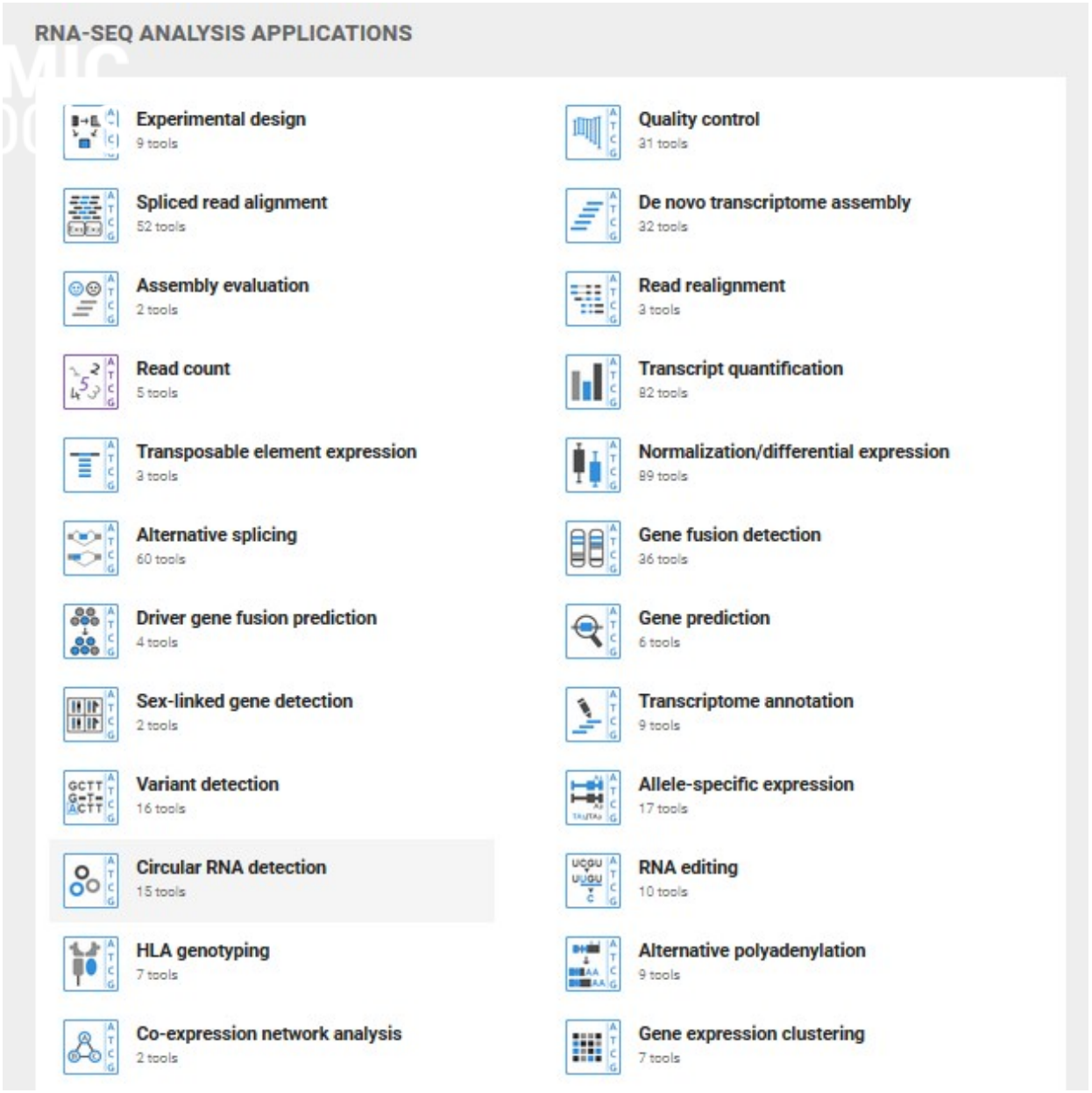
***Figure 1**: The top 20 species ranked by number of RNA-seq libraries (parenthesis) available in the SRA datab*

TGCA | **10696** GTEX | **8555** ENCODE | **657** GEUVADIS | **465**



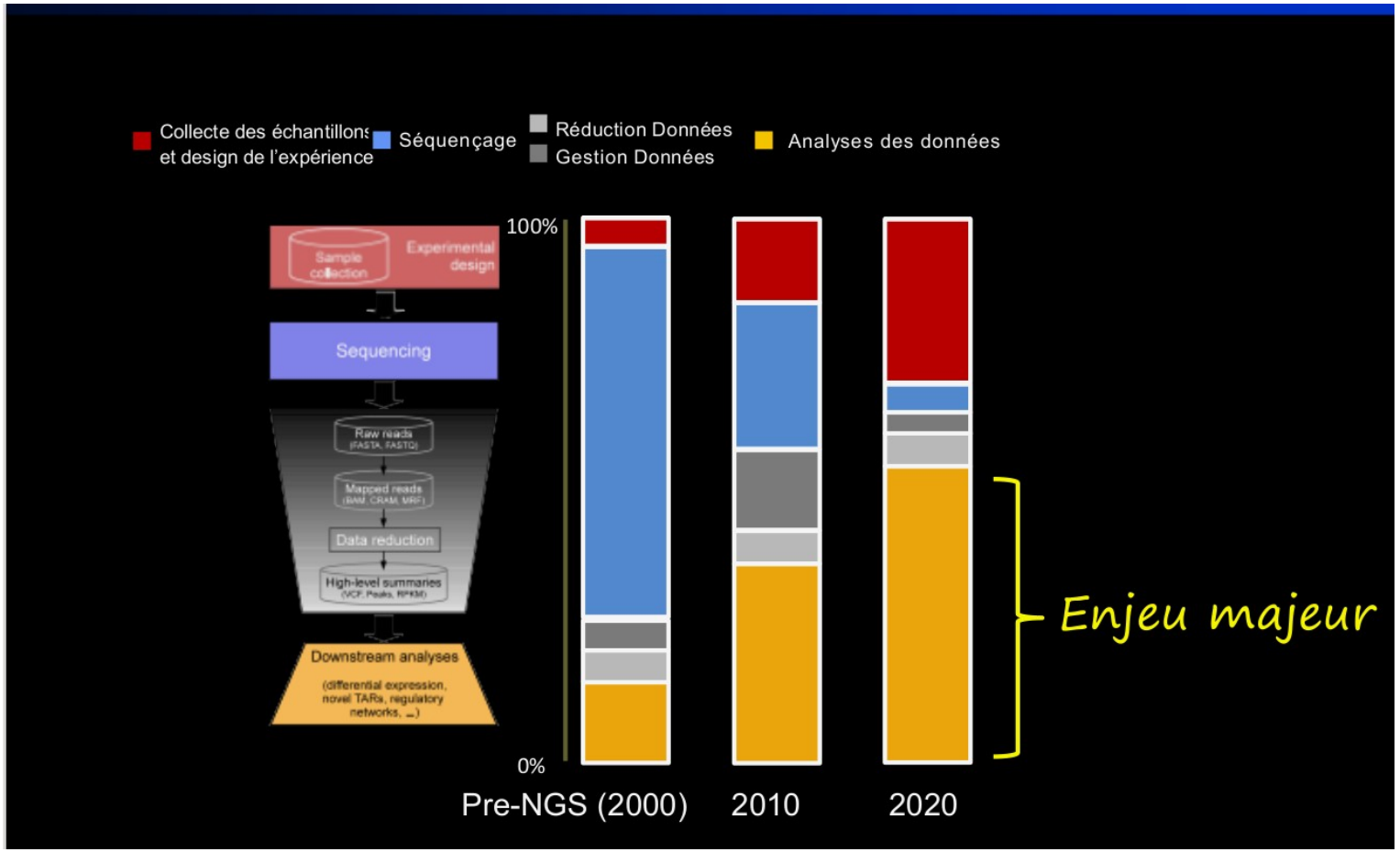
https://omictools.com/

5-Profusion d’Outils Bioinformatiques

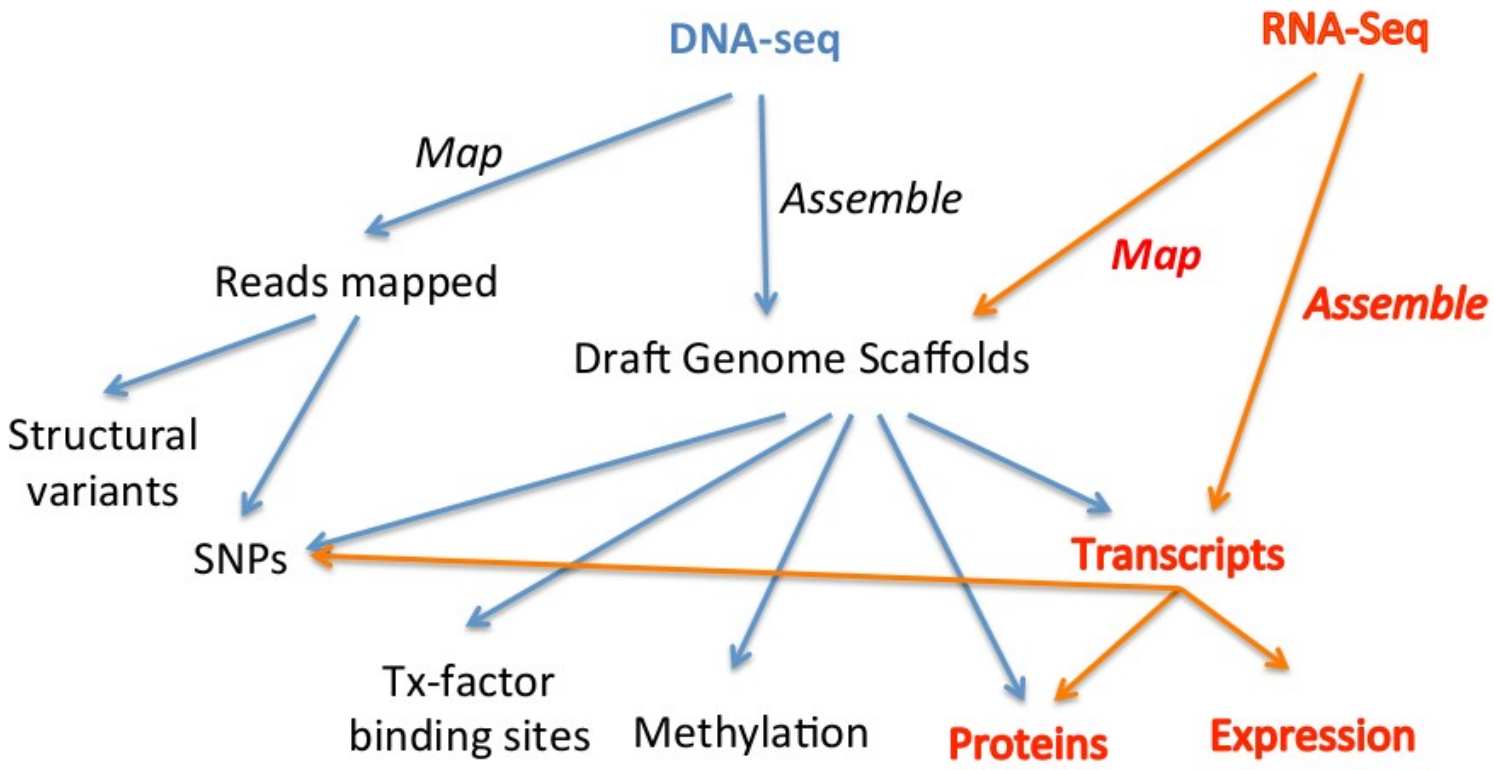


https://omictools.com/whole-exome-sequencing-category

Quelques enjeux de l’analyse des données



Applications



Genome et Transcriptome / Whole DNAseq and RNAseq
Les points communs de l’analyse
Les spécificités

Les champs d’application exemples en diagnostic humain
Quelques limitations

Principales Applications en diagnostic moléculaire

ADN (Panel de gènes ciblés, Exomes, Whole genome ...)
Mutations (Substitutions, Indels, Tandem repeats..)
Anomalies chromosomiques (translocations, Inversions, gènes de fusion)
Variants structuraux à large échelle (CNV)
ADN circulant

ARN (RNA-seq)
Expression des gènes codants
Long non-codingRNA (lncRNA), microRNA (miRNA), ARN circulaires (circRNA)...
Variants d'épissage ...
ARN de fusion...
Indels, mutations.....

ADN&Epigenome, Méthylation, Hi-C (prochaine étape)
Métagenomique (Microbiologie, Virologie, Microbiome..)

DNA-RNA-Applications

- Panel

ADN/ARN

- 50-500Mo
- labTop

- Exome
- RNAseq

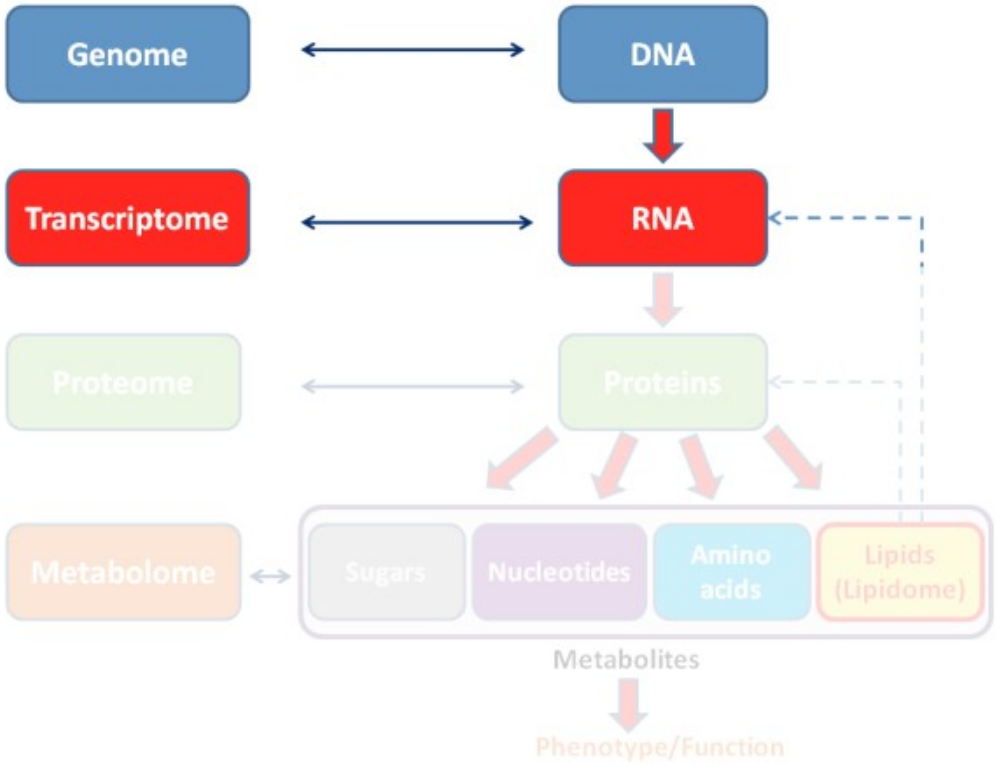
- 1-4Go
- server

- Whole genome

- 40-400Go
- server/ cloud

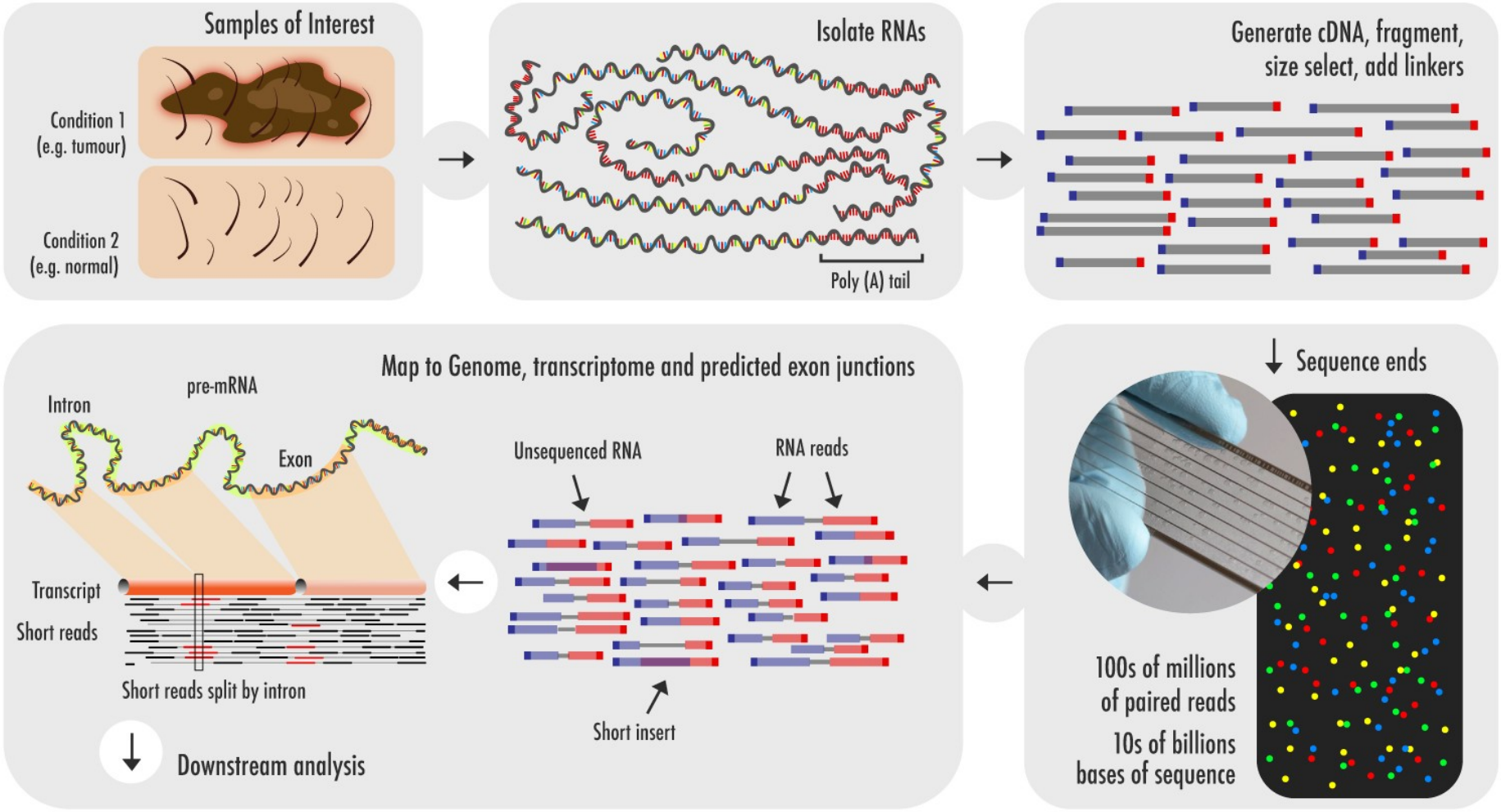
Transcript**omics**

Transcriptome = complete set of all RNA molecules ("transcripts") produced from a genome **OR** specific subset of transcripts present in a particular cell type or under specific growth conditions



Transcriptomics involves large-scale analysis of RNAs to follow when, where, and under what conditions genes are expressed.

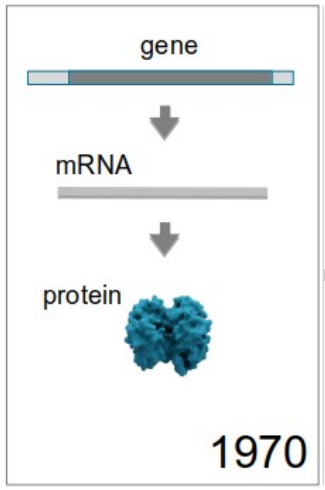
Transcriptome and RNAseq

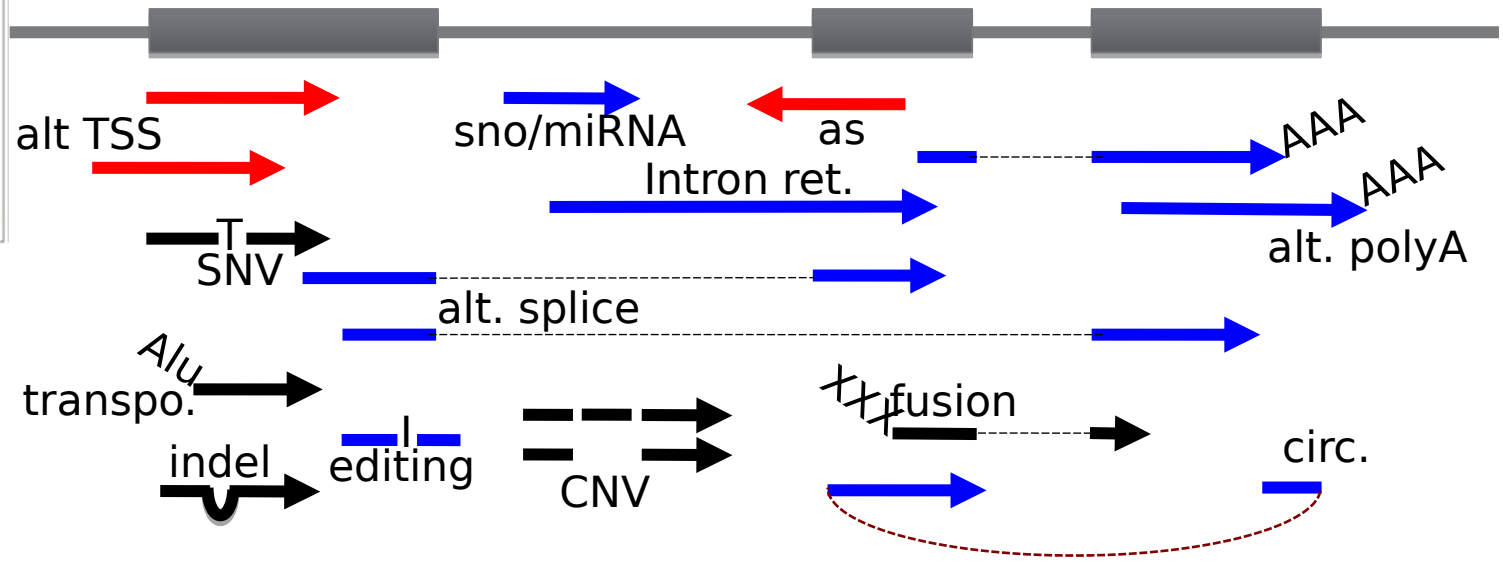


A simple method

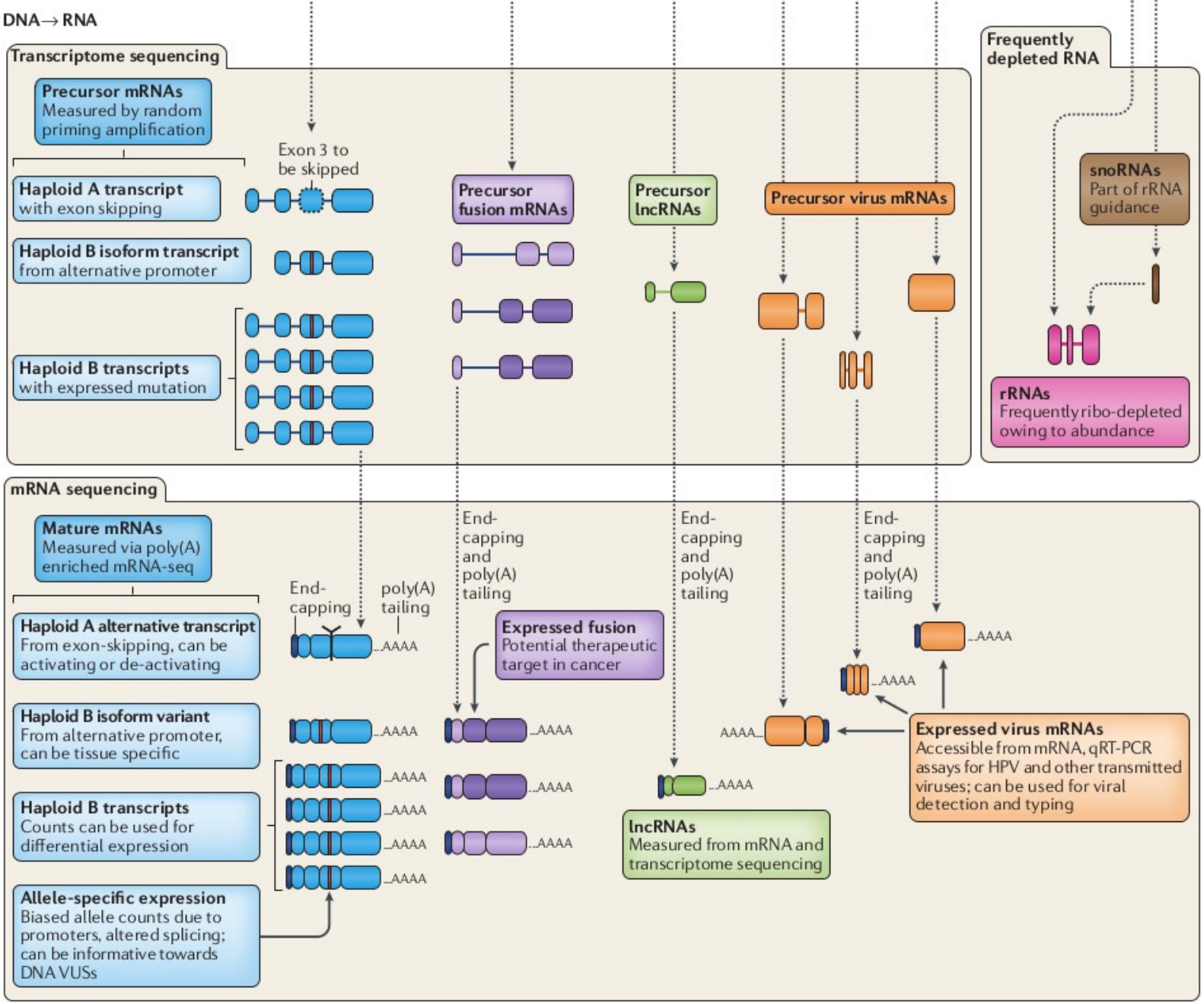
The Real Transcriptome

is combinatorial

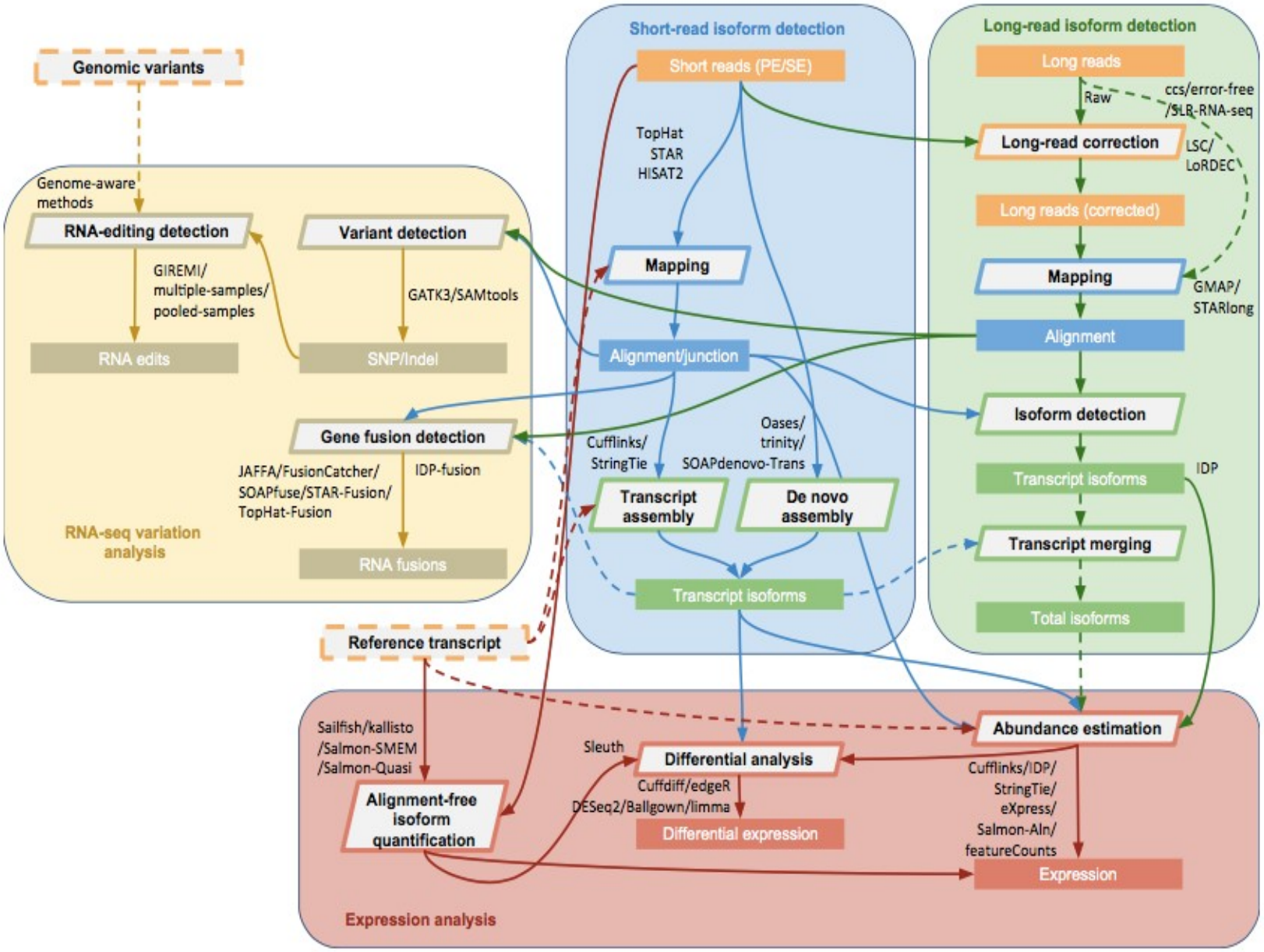




Genetic X **Transcriptional** X **Post-transcriptional variation**

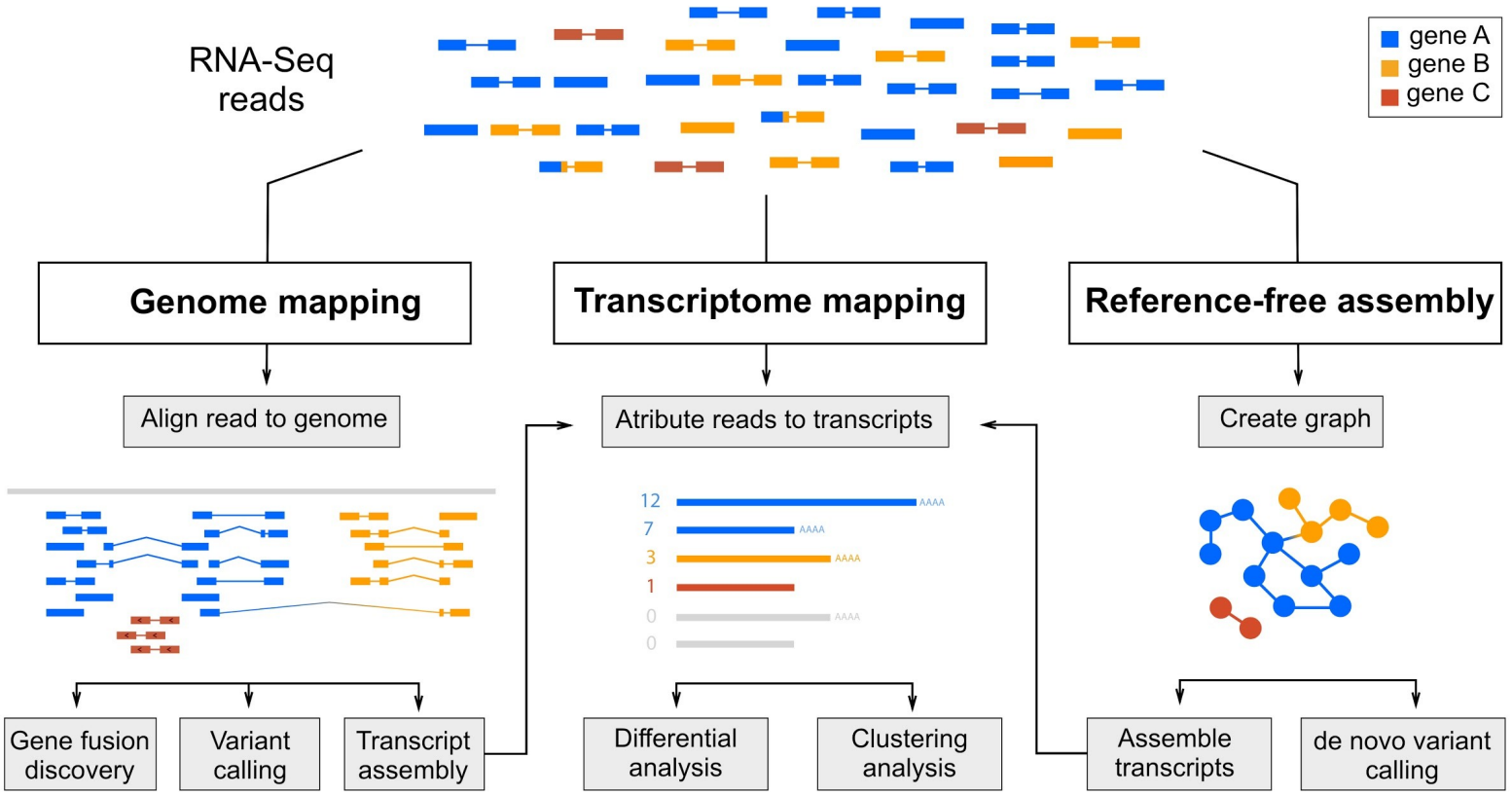


Translating RNA sequencing into clinical diagnostics: opportunities and challenges
Byron et al. 2016



.RNAseq analysis (nat genetics)

Transcriptome and RNA-seq



A complex analysis

Enjeux de l’analyse

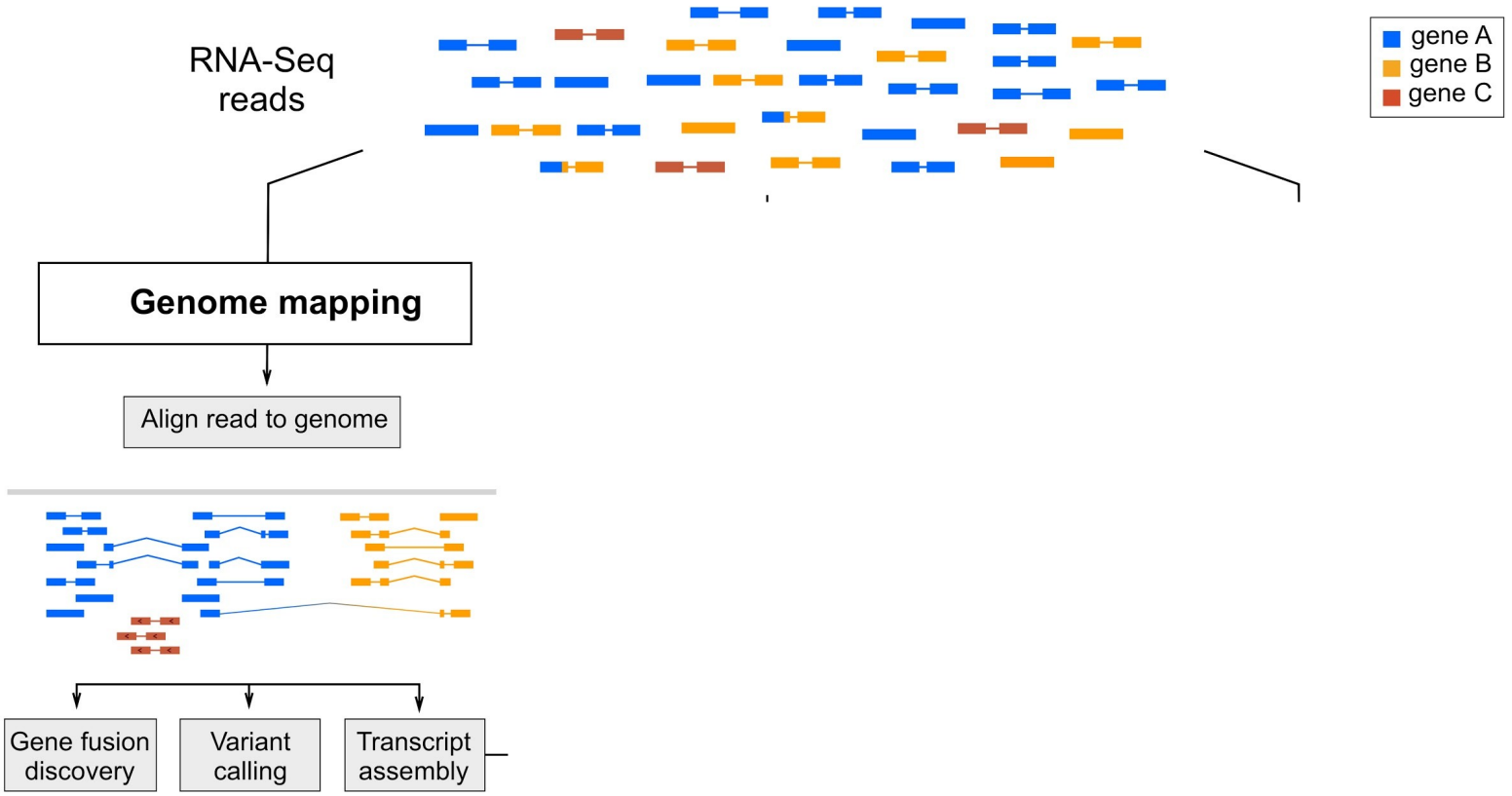
De nombreux softwares :

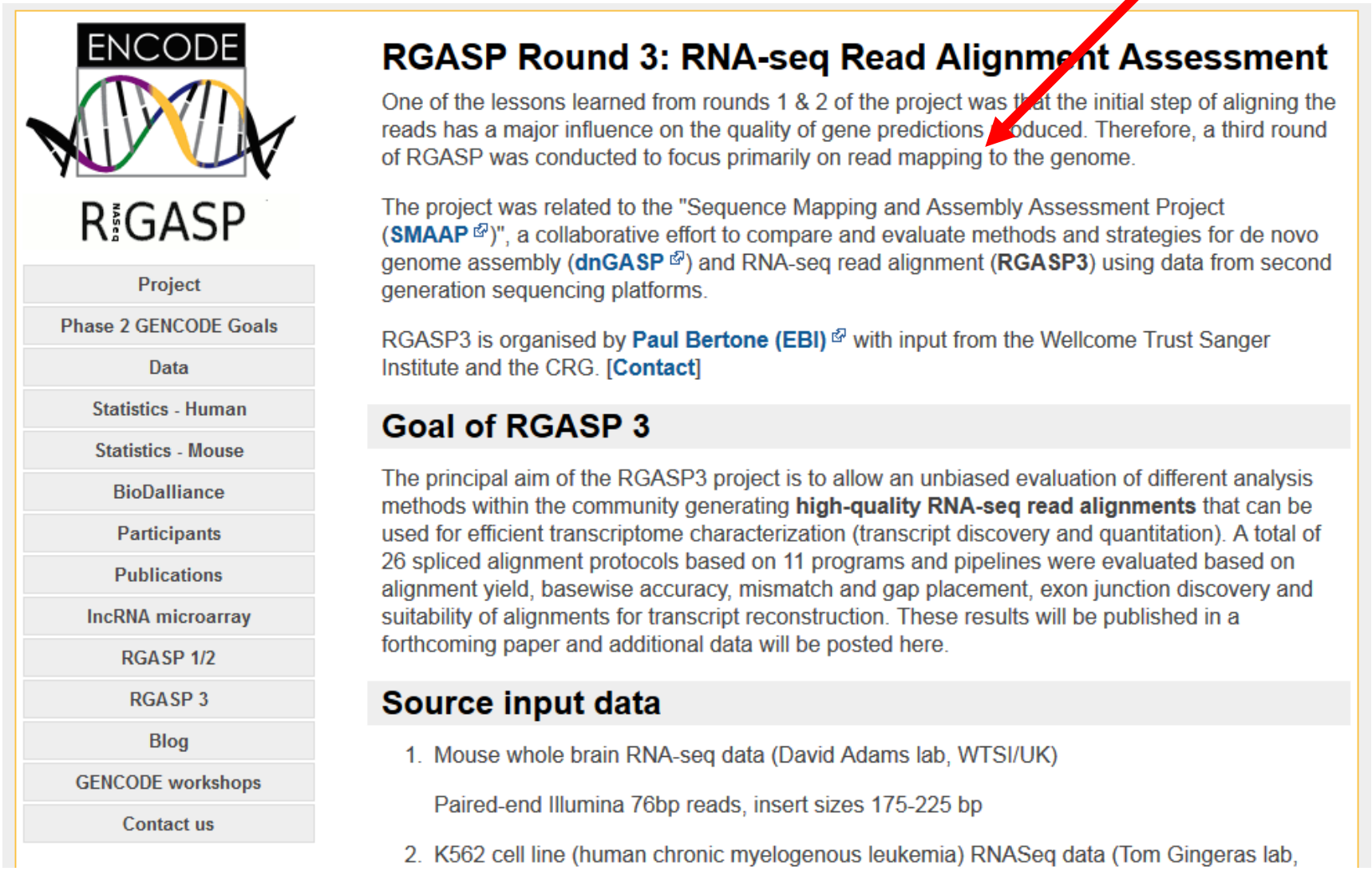
Comment choisir le meilleur et le plus approprié à une question biologique donnée?

Du principe algorithmique à la biologie (et inversément)

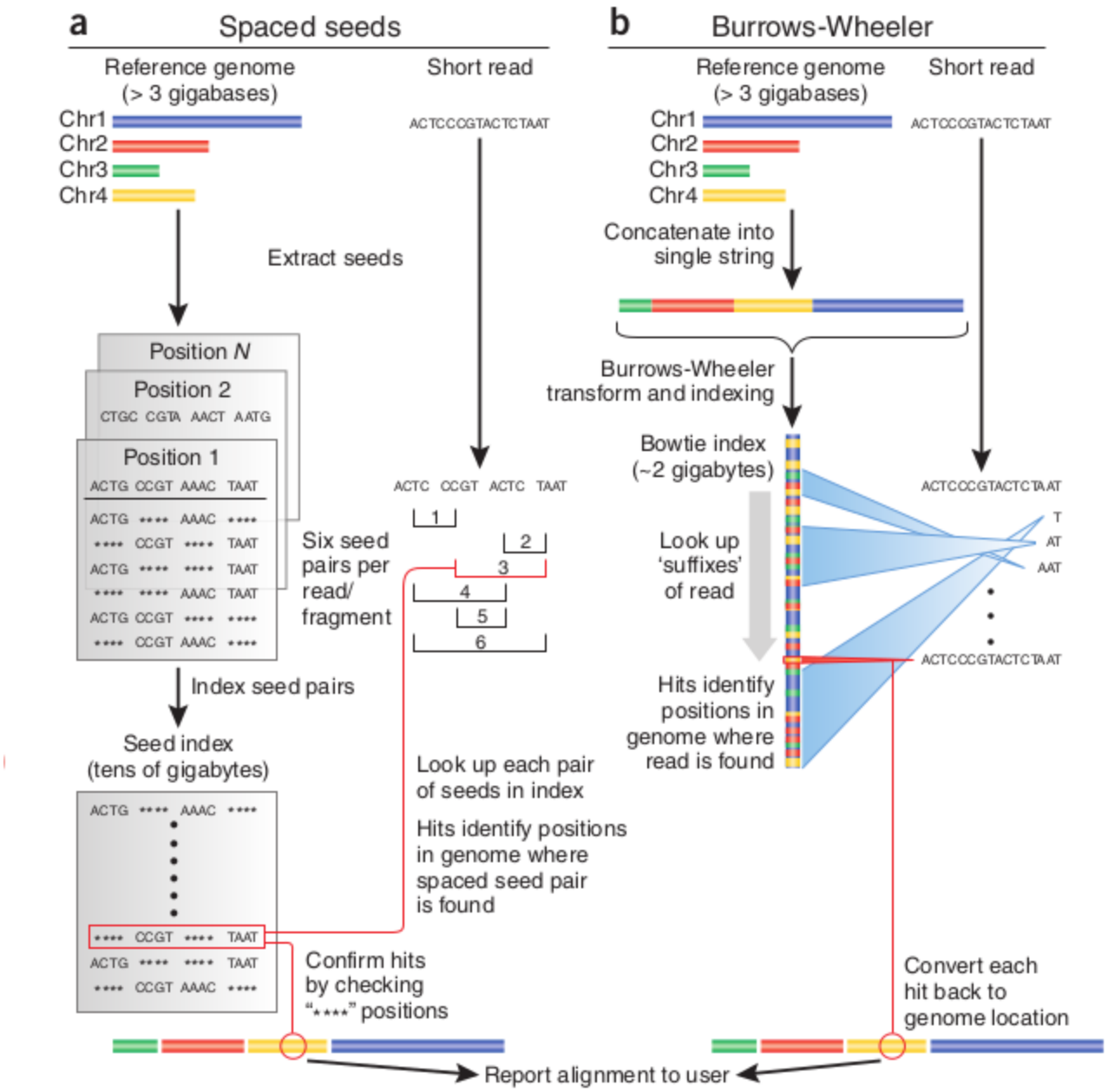
Exemple de mapping, limites..

RNA-seq and genome mapping





L’étape de mapping n’est pas toujours simple…..



nature biotechnology volume 27 number 5 may 2009, Trapnell

Technology dictates algorithms:
recent developments in read alignment
Alser et al. Genome Biology (2021) 22:249

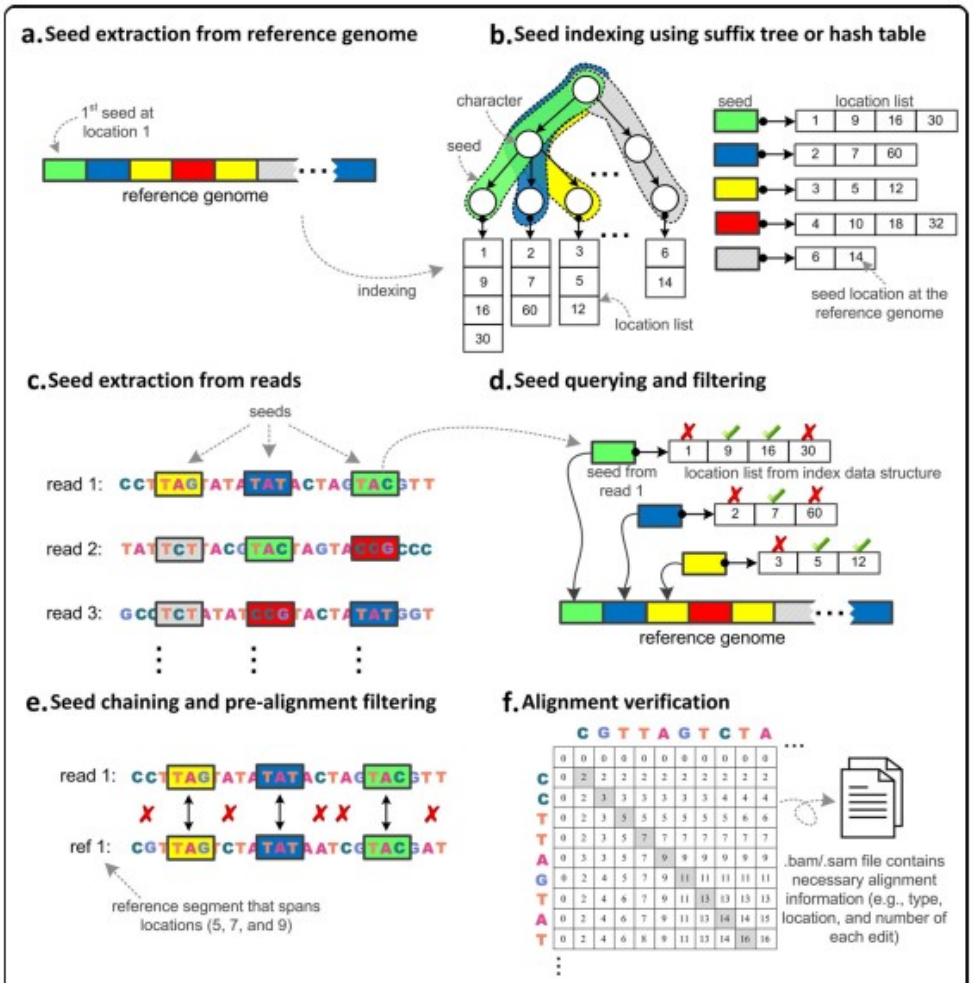
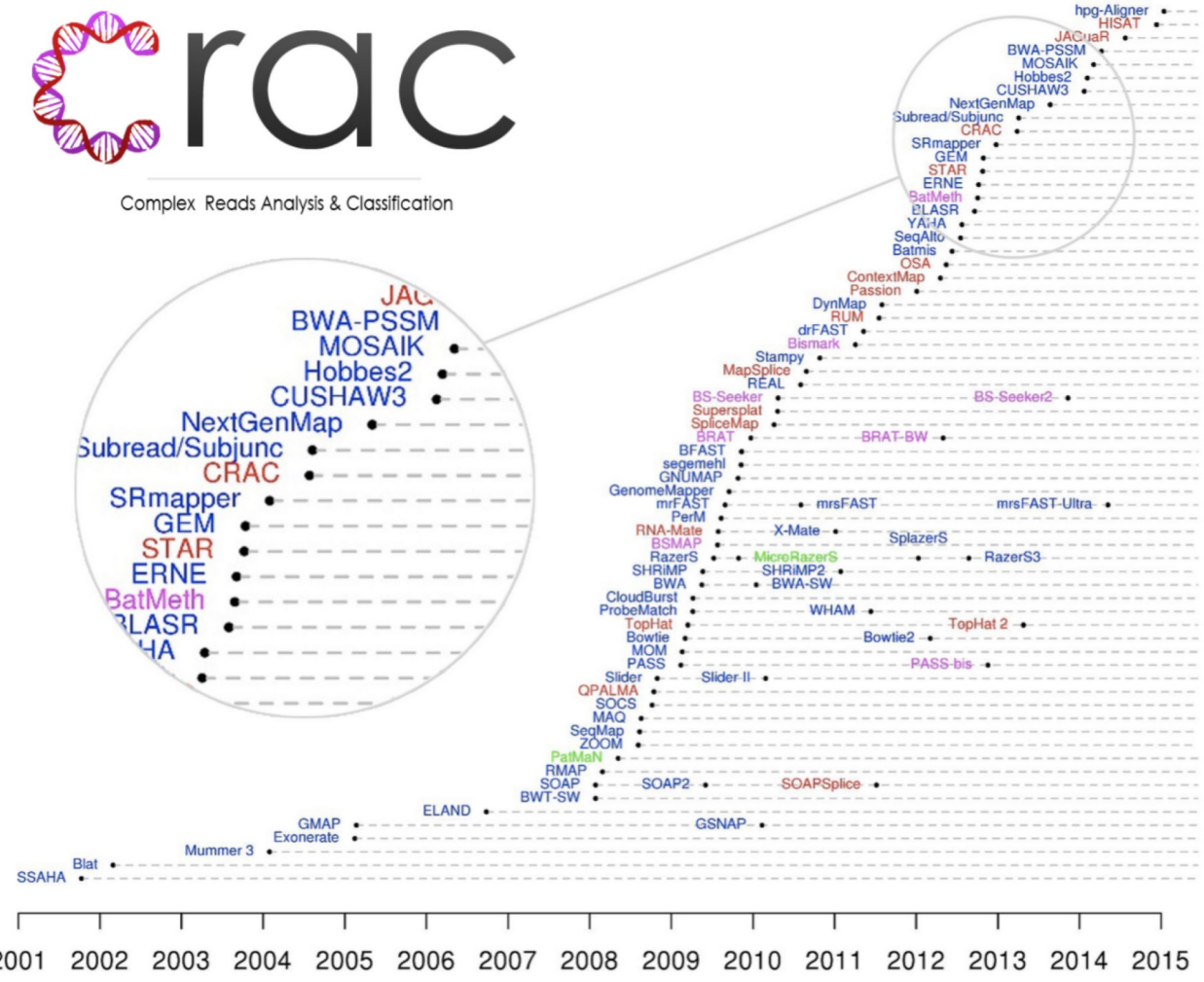


Fig. 1 Overview of a read alignment algorithm. **a** The seeds from the reference genome sequence are extracted. **b** Each extracted seed and all its occurrence locations in the reference genome are stored using the data structure of choice (suffix tree and hash table are presented as an example). Common prefixes of the seeds are stored once in the branches of the suffix tree, while the hash table stores each seed individually. **c** The seeds from each read sequence are extracted. **d** The occurrences of each extracted seed in the reference genome are determined by querying the index database. In this example, the three seeds from the first read appear adjacent at locations 5, 7, and 9 in the reference genome. Two of the same seeds appear also adjacent at another two locations (12 and 16). Other non-adjacent locations are filtered out (marked with X) as they may not span a good match with the first read. **e** The adjacent seeds are linked together to form a longer chain of seeds by examining the mismatches between the gaps. Pre-alignment filters can also be applied to quickly decide whether or not the computationally expensive DP calculation is needed. **f** Once the pre-alignment filter accepts the alignment between a read and a region in the reference genome, then DP-based (or non-DP-based) verification algorithms are used to generate the alignment file (in BAM or SAM formats), which contains alignment information such as the exact number of differences, location of each difference, and their type.

problème = **# logiciels** x **# paramètres** x **# d'applications**

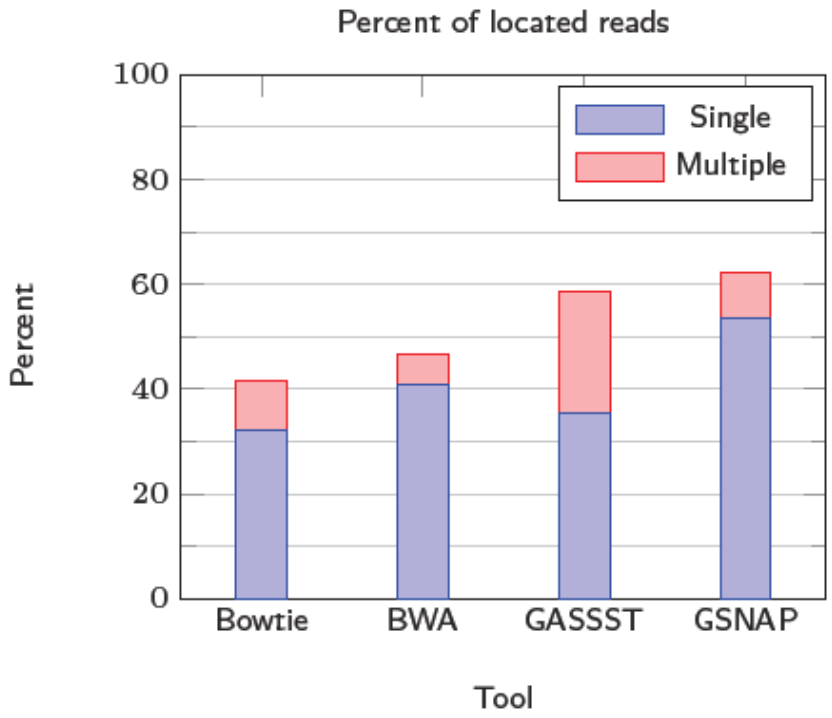


Nuno A. Fonseca Johan Rung Alvis Brazma John C. Marioni, **Tools for mapping high-throughput sequencing data**. Bioinformatics (2012) 28 (24): 3169-3177.

Exemple de problèmes bioinformatiques: 1.le mapping

An example

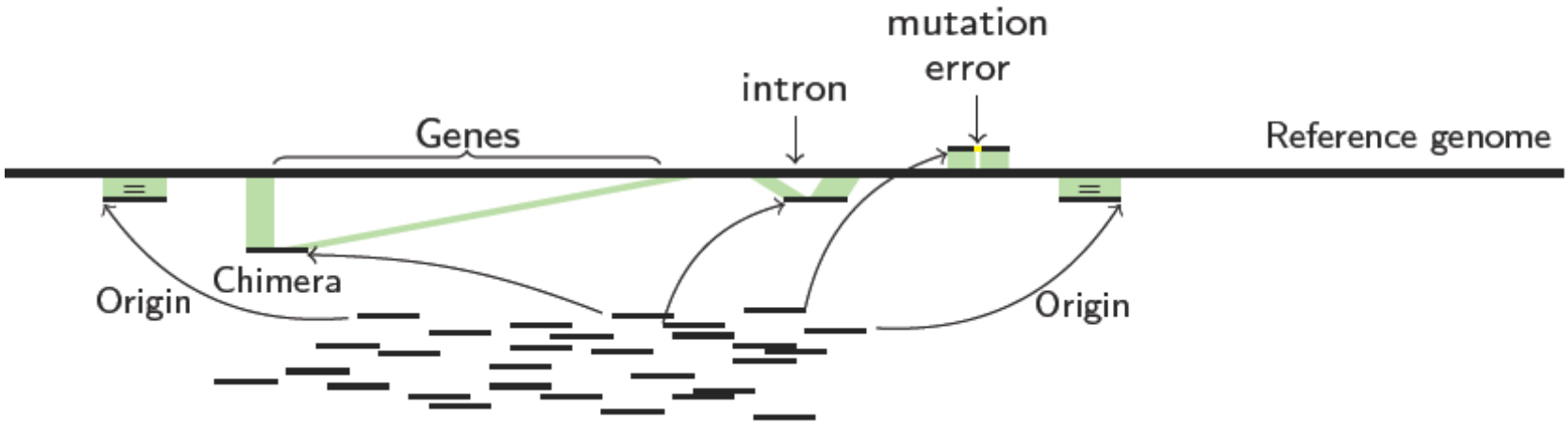
- RNA-Seq data from a K562 library (human cancer cells)
- $\sim$ 70 millions of reads of 75 bp



Inputs

RNA-seq: collection de millions de reads (~100 bp)

Reference genome: $3x10^9$ pb (genome humain)

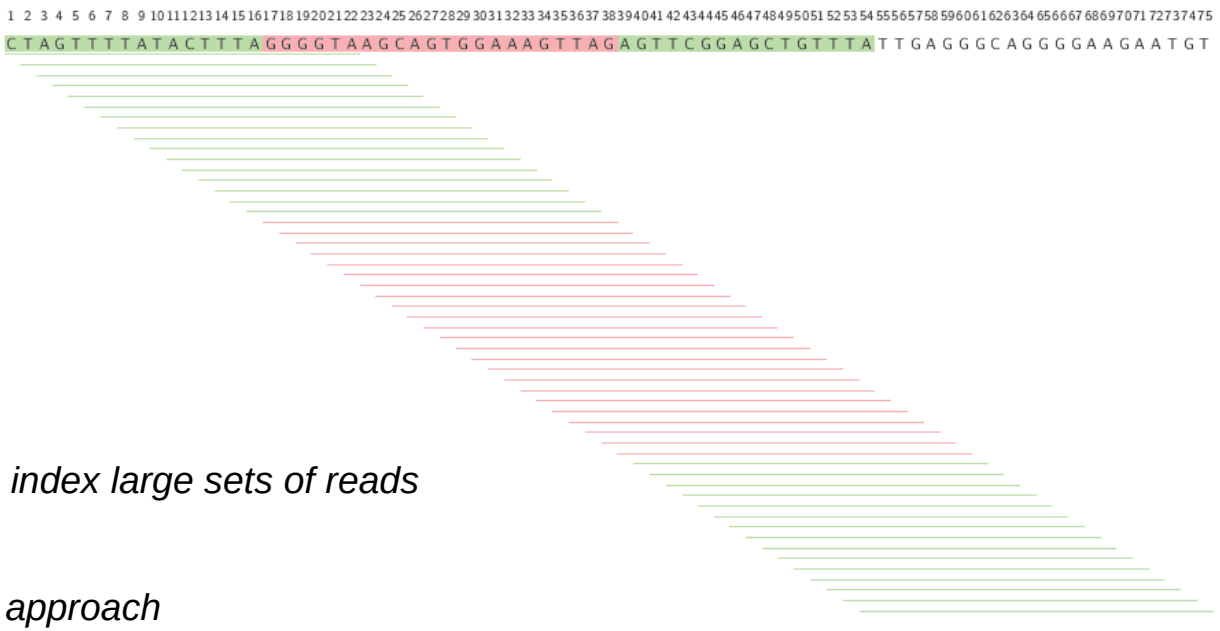


« Etapes de Mapping complexes »

CRAC: an integrated approach for RNA-seq analysis

For each read
k factors and their
genomic mapping

CRAC: an approach to classify reads

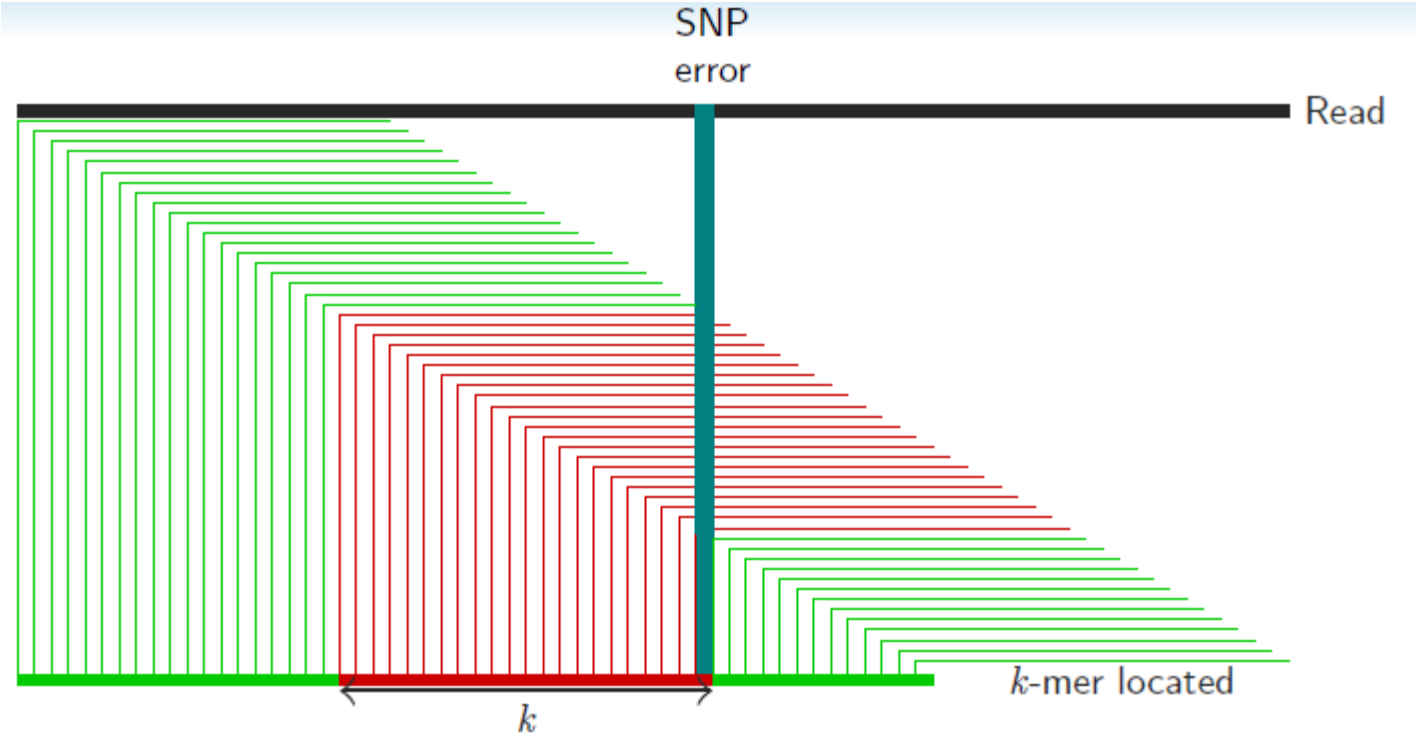


Mapping information
used for SNP
Error and splice…

Based on the Gk arrays structure to index large sets of reads

Based on a k-mer approach

Length of k ~22 nt for th human genome



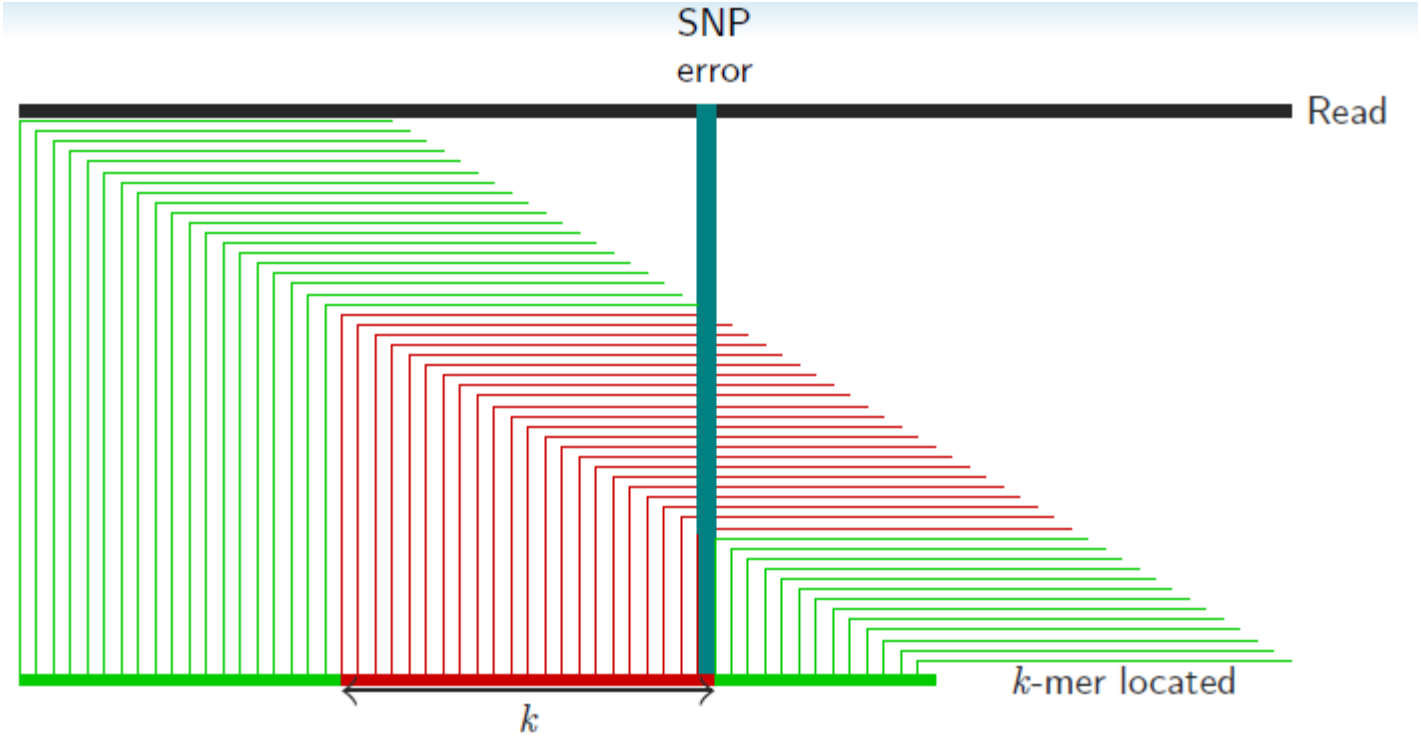
Analysis of the support variation

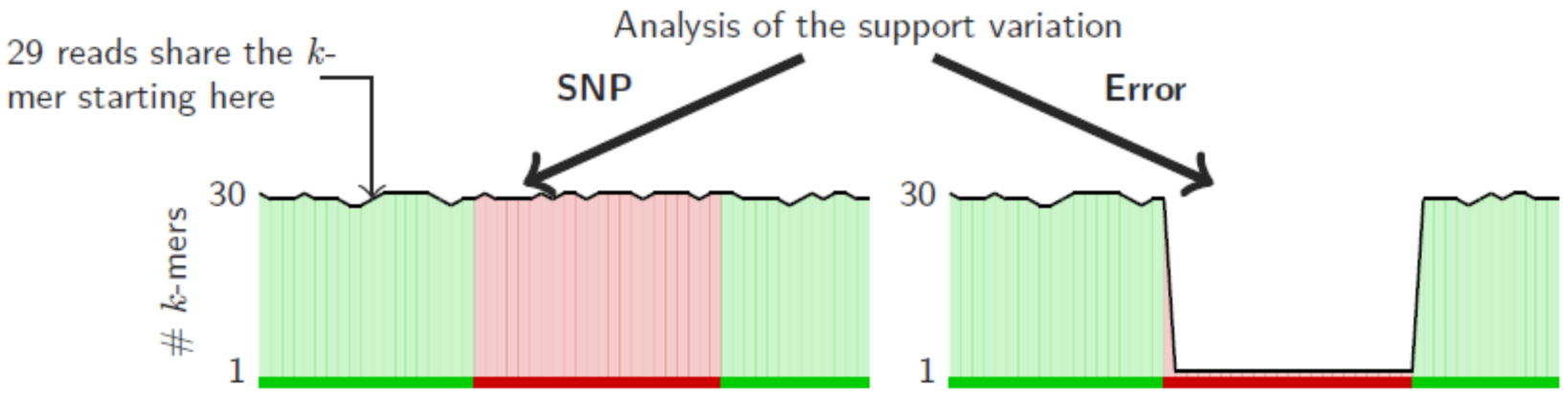
break

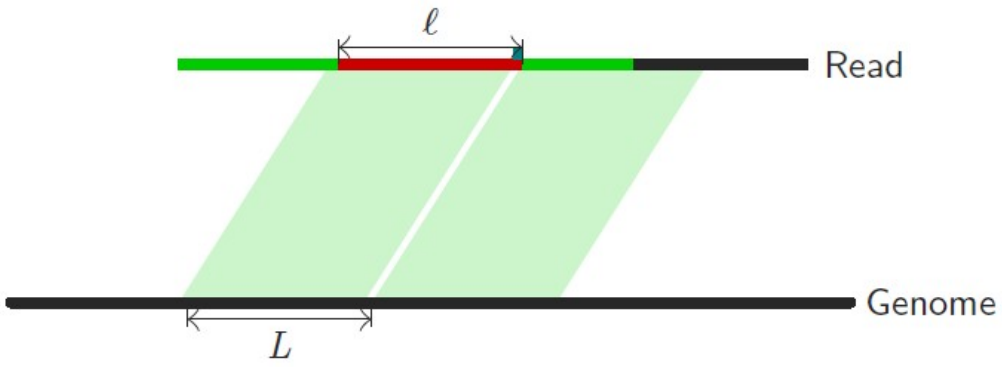
An interruption of the k-mers location on the genome

support

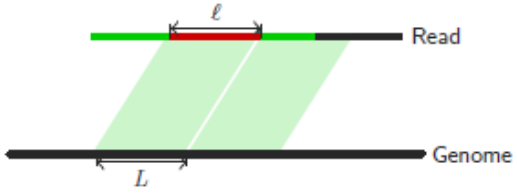
A k-mer support is the number of reads that contain it (local coverage)


SNP
error
Read
k-mer located
k

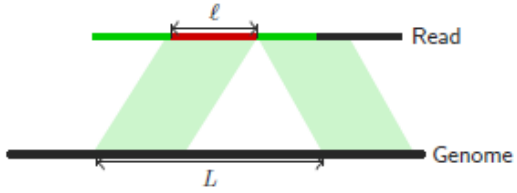

Analysis of the support variation
29 reads share the k-mer starting here
SNP
Error
30
1
k-mers
30
1



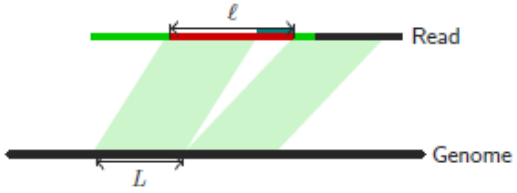
Substitution



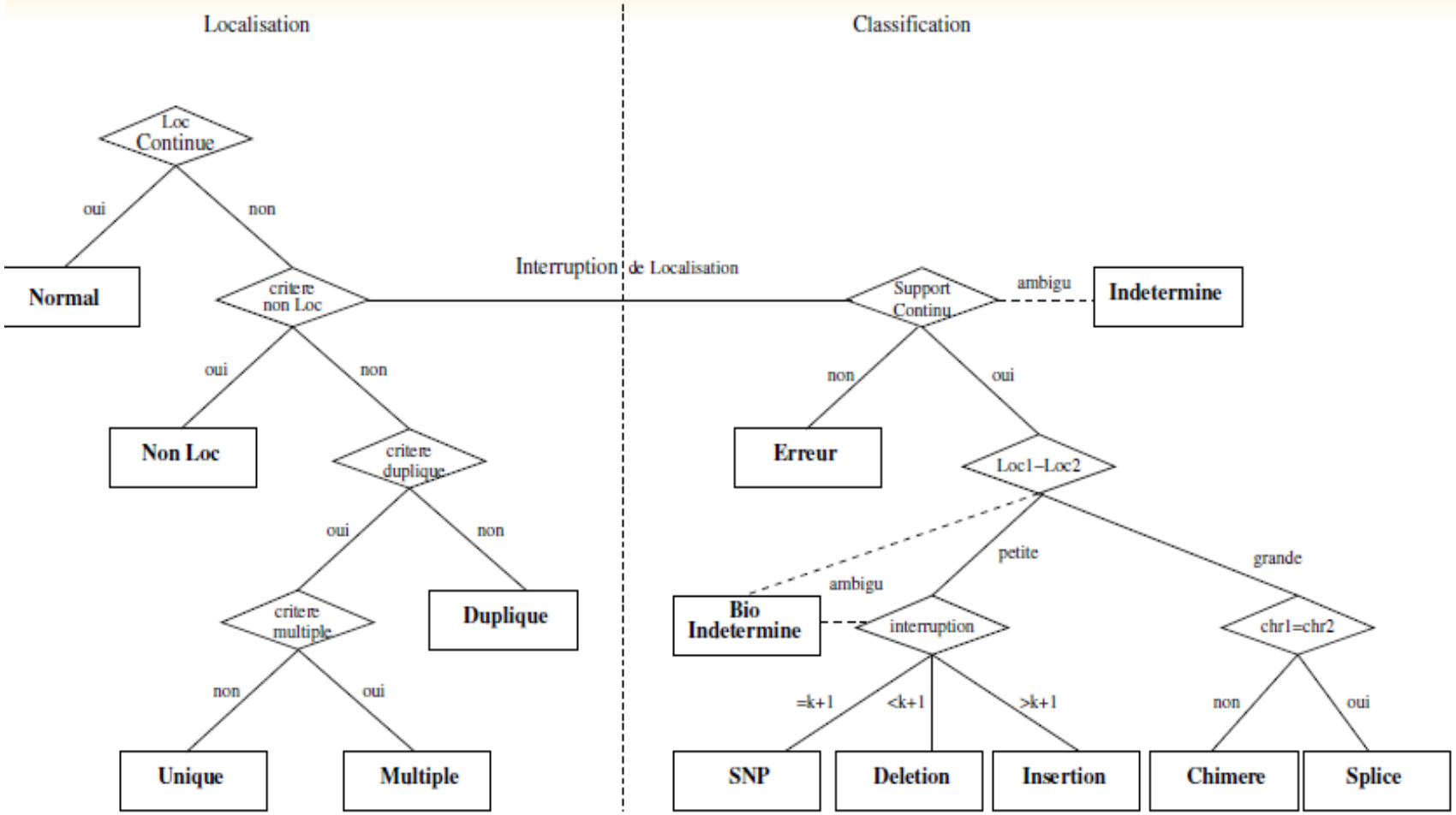
Insertion



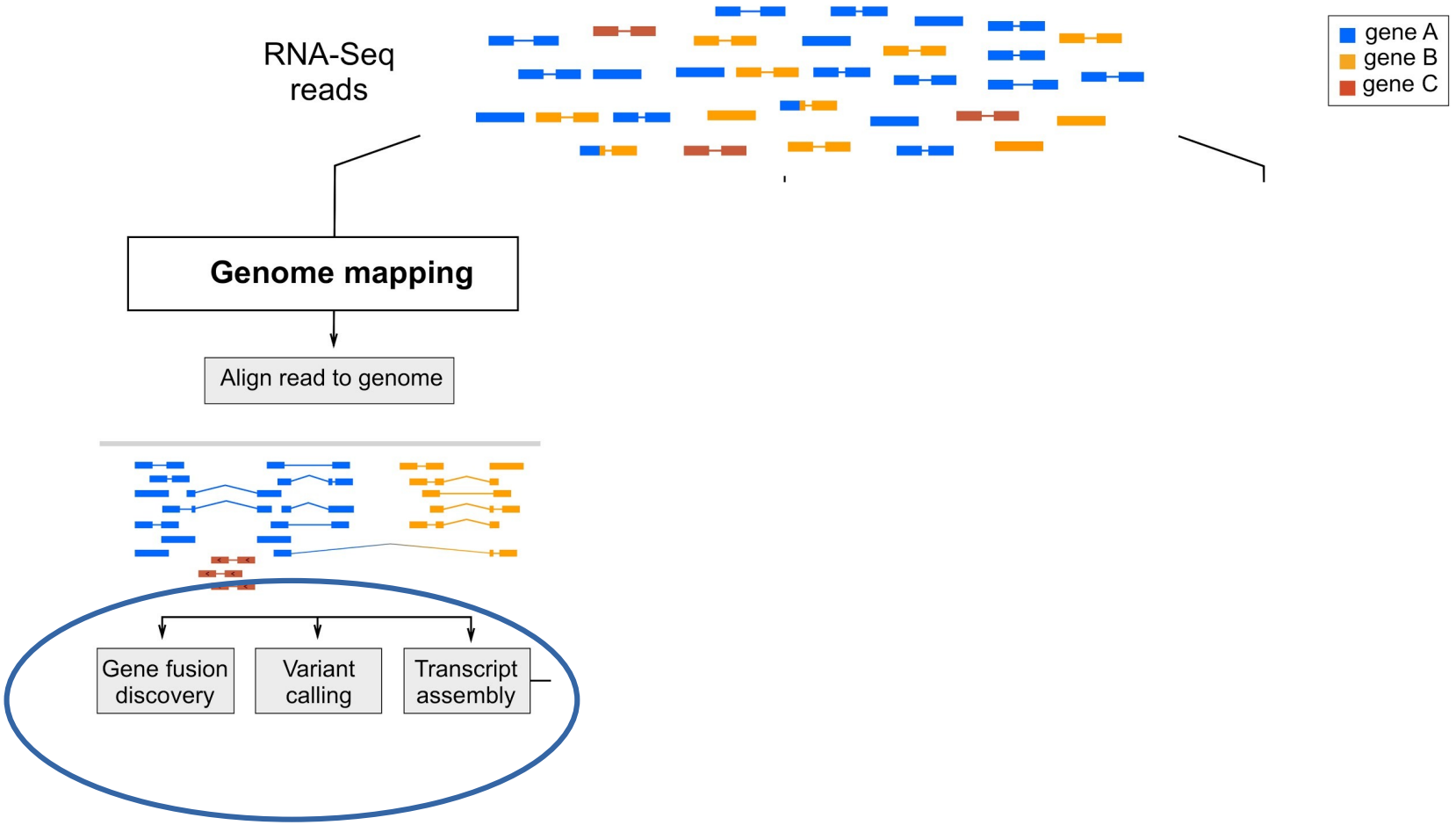
Deletion



Splicing junction


Localisation
Classification
Loc Continue
oui
non
Normal
critere non Loc
Interruption de Localisation
Support Continu
ambigu
Indetermine
oui
non
Non Loc
critere duplique
non
oui
Erreur
Loc1–Loc2
oui
non
petite
grande
ambigu
Bio Indetermine
interruption
chr1=chr2
critere multiple
Duplique
non
oui
=k+1
<k+1
>k+1
non
oui
Unique
Multiple
SNP
Deletion
Insertion
Chimere
Splice

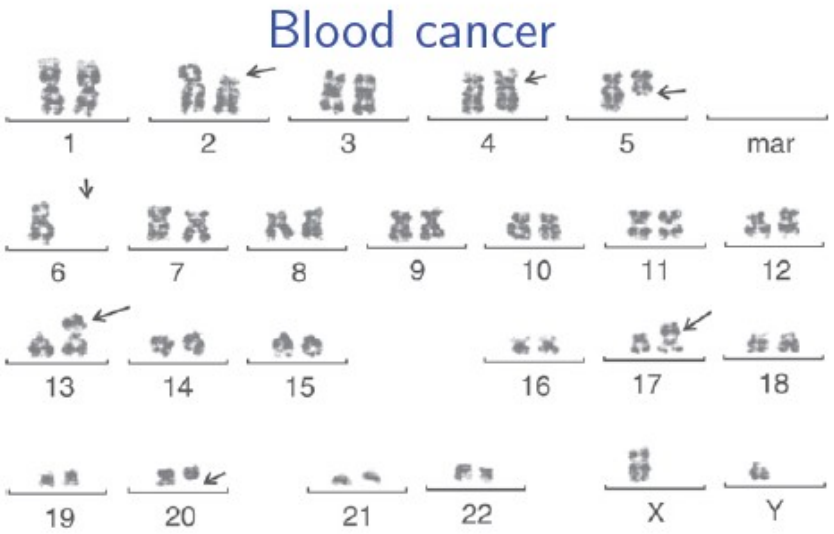
RNA-seq and genome mapping



Trouver des évenements rares

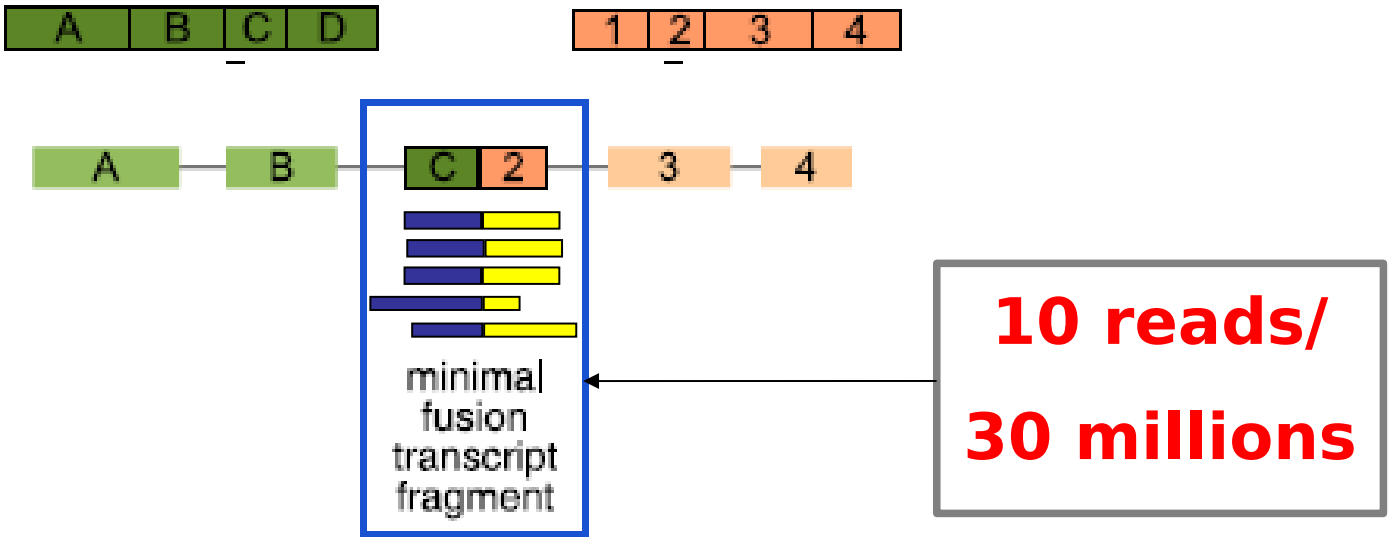
Abnormal chromosome pool in cancer

- Diagnosis of chronic myelogenous leukemia (CML)
- Prognosis in myelodysplastic syndrome



Trouver des évenements rares (ARNde fusion)

1/ trouver des séquences rares et spécifiques dans une collection de read (10 /30. 10^{6})



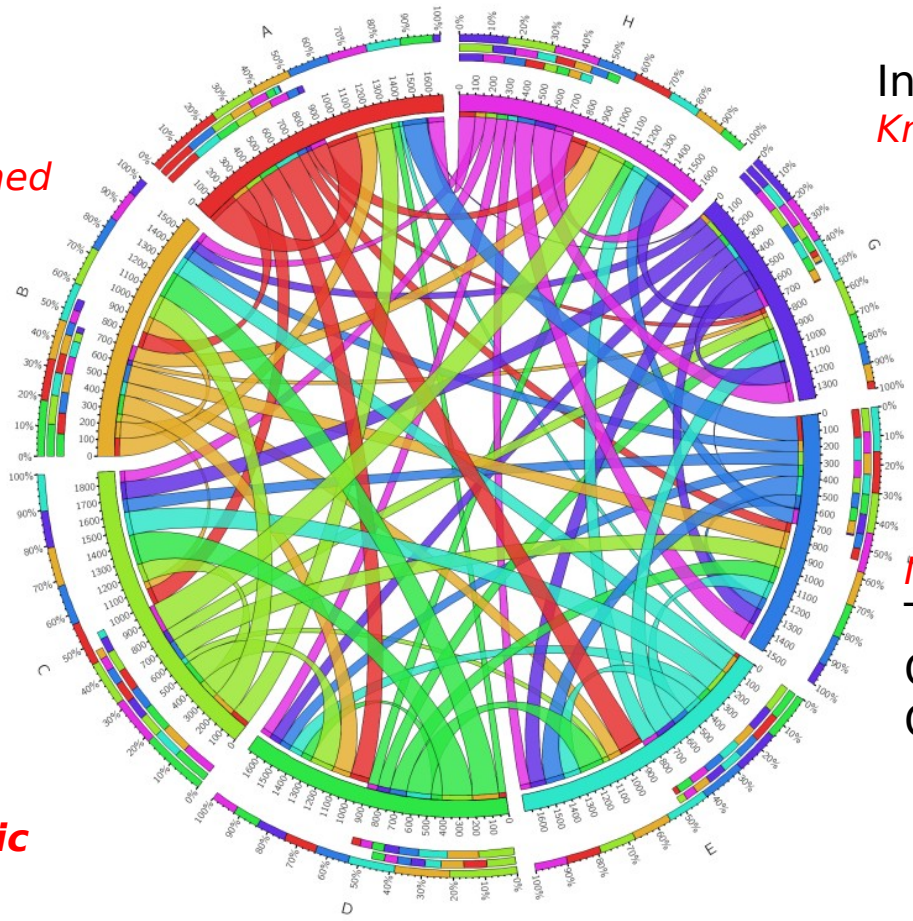
2/ Distinguer les différents évenements biologiques et les différencier des artefacts

The **chimeric** transcriptome in AML

Translocation
New translocation
Variant of well-defined translocation

Tandem repeat
Inverted & direct

Read-through
Tumor and non-tumor specific



Inversion
Known Inversion

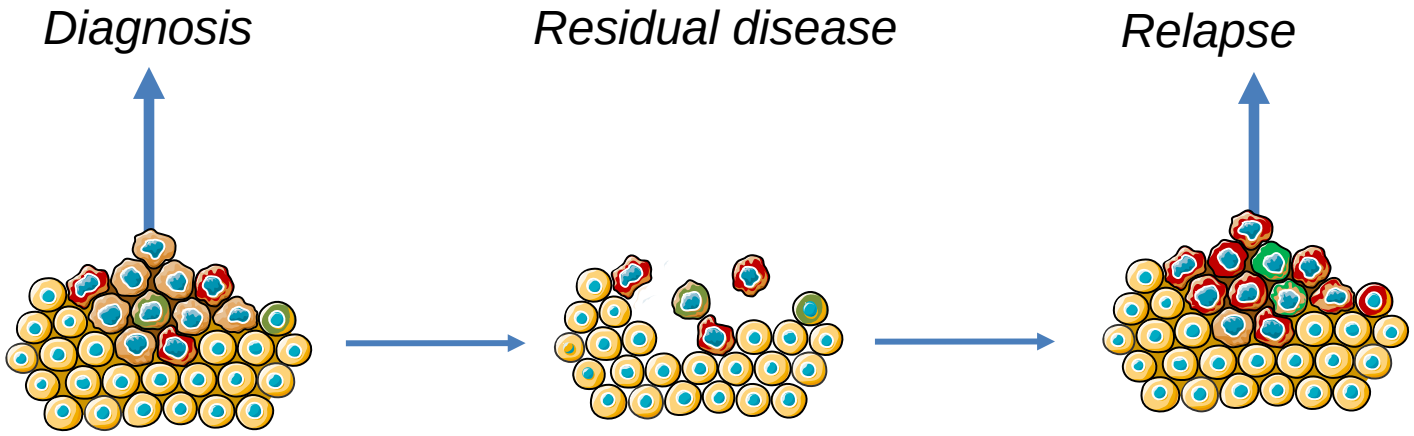
New classes 2,3, 4
Trans-splicing ?
CNV ?
Other

New Chimeric RNA in all classes

17 biological Validations

Biomarqueurs et médecine de précision

Diagnostic, suivi du patient en cancerologie



primitive tumor

Clonal expansion

Tumor status change

Patients classification
Therapeutical proposal

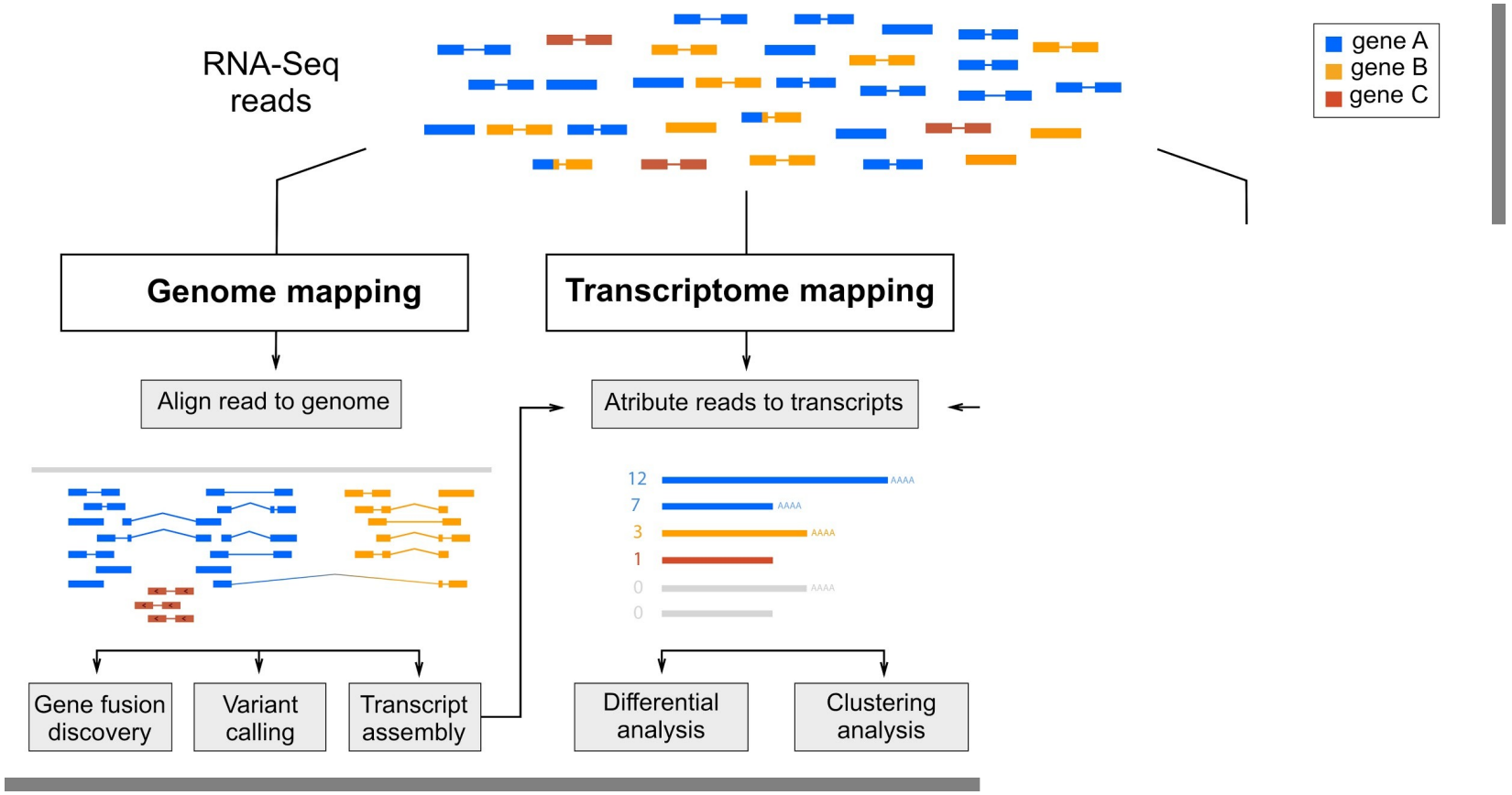
Disease follow-up
Heterogeneity

Treatment resistance
Targeting therapy

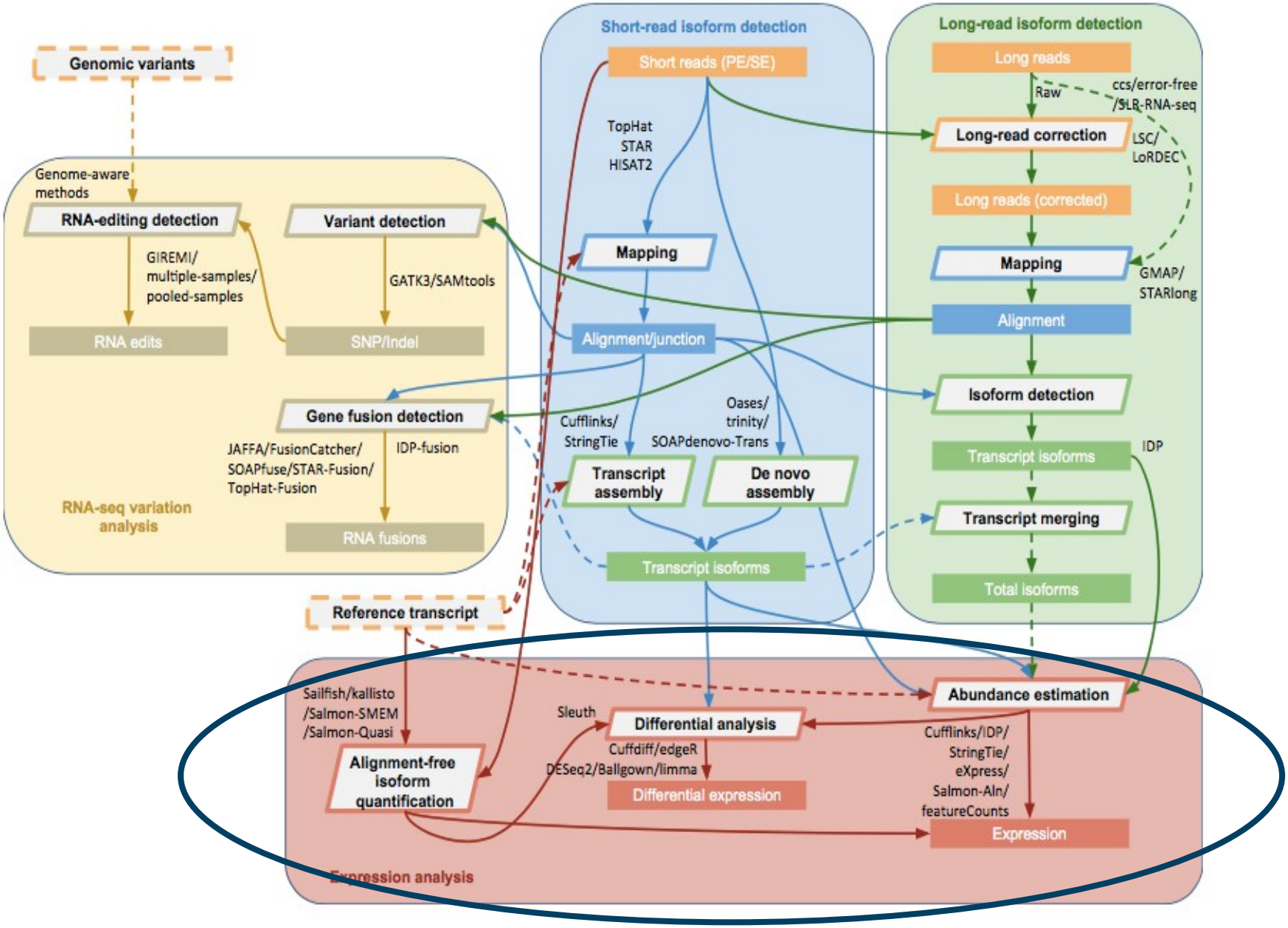
La solution : une combination de Biomarqueurs

(Mutations, Fusion genes, Gene expression, Splicing events, lncRNA, miRNA)

Transcriptome and RNA-seq



Transcriptome mapping for DGE



Differentiel gene expression (DGE) ...Application la plus fréquente...

NGS et Transcriptomics

- Introduction: l'ére NGS et état des lieux, principaux séquenceurs (vu en intro)
- Les enjeux de l'analyse, principales applications en diagnostic (ADN/ARN)
- Transcriptome et RNAseq, les questions biologiques
  - ✓ Les différentes stratégies d'analyse
  - ✓ Le mapping, exemple et limites
  - ✓ **Annotations génomiques, transcriptome de reference**
  - ✓ Comment passer à l'échelle

GENCODE

GENCODE | Data | Stats | Browser | Blog

Statistics about the current Human GENCODE Release (version 25)

* The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

Compare with the previous release (GENCODE 24) »

Version 25 (March 2016 freeze, GRCh38) - Ensembl 86

General stats

Total No of Genes	58037	Total No of Transcripts	198093
Protein-coding genes	19950	Protein-coding transcripts	80087
Long non-coding RNA genes	15767	- full length protein-coding:	54755
Small non-coding RNA genes	7258	- partial length protein-coding:	25332
Pseudogenes	14650	Nonsense mediated decay transcripts	13769
- processed pseudogenes:	10725	Long non-coding RNA loci transcripts	27692
- unprocessed pseudogenes:	3400		
- unitary pseudogenes:	214		
- polymorphic pseudogenes:	51		
- pseudogenes:	21	Total No of distinct translations	60033
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13536
- protein coding segments:	411		
- pseudogenes:	239		

GENCODE

Human | Mouse | How to access data | FAQ | Documentation

Human

Statistics about the current GENCODE Release (version 38)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

General stats

Total No of Genes	60649	Total No of Transcripts	237012
Protein-coding genes	19955	Protein-coding transcripts	86757
Long non-coding RNA genes	17944	- full length protein-coding	61015
Small non-coding RNA genes	7567	- partial length protein-coding	25742
Pseudogenes	14773	Nonsense mediated decay transcripts	18881
- processed pseudogenes	10667	Long non-coding RNA loci transcripts	48752
- unprocessed pseudogenes	3565		
- unitary pseudogenes	241		
- polymorphic pseudogenes	49		
- pseudogenes	15	Total No of distinct translations	63968
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct tra	13689
- protein coding segments	409		
- pseudogenes	236		