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

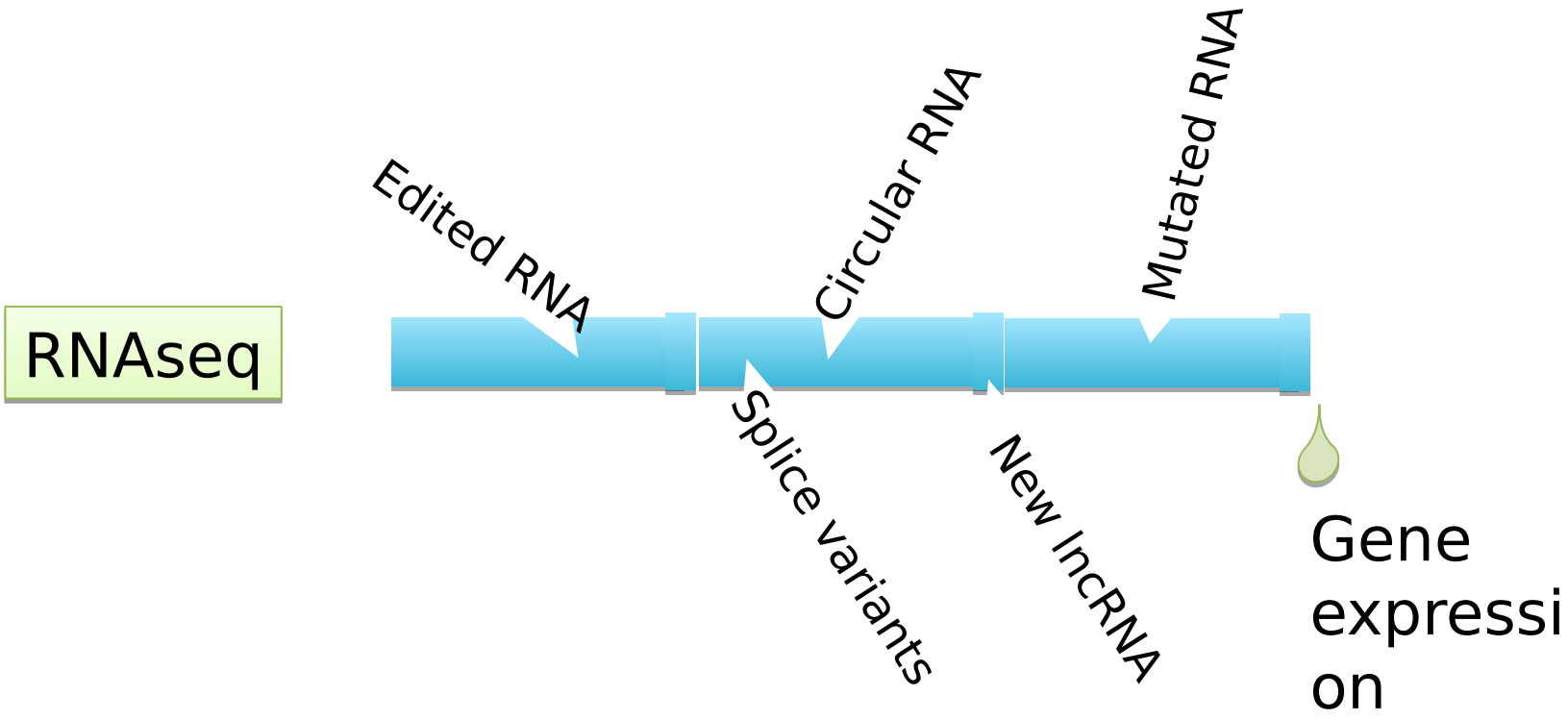
L’ère NGS….

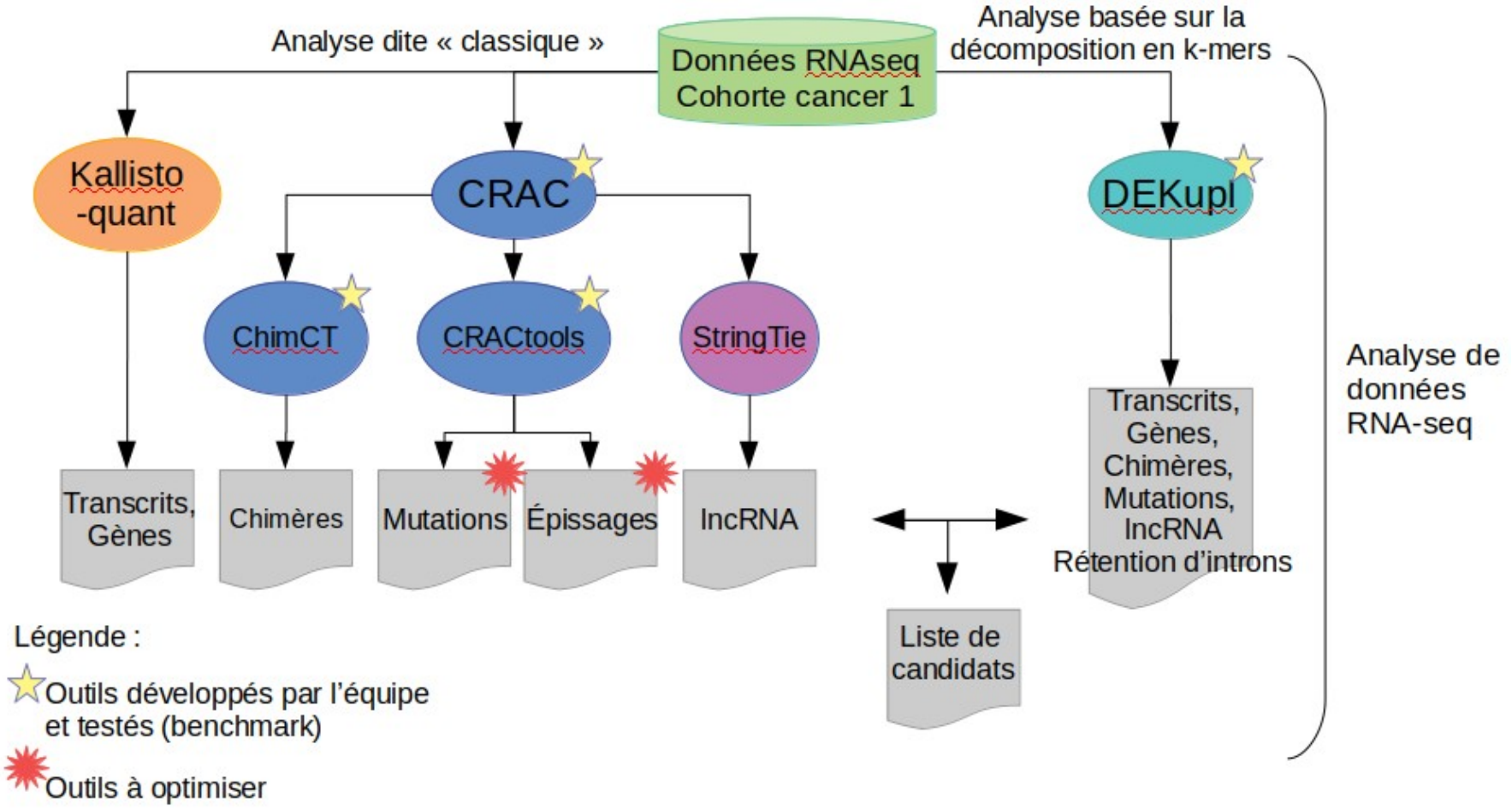
NGS et Transcriptomics

- Introduction: l'ére NGS et état des lieux, principaux séquenceurs (vu en intro)
- Les enjeux de l'analyse, 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
  - ✓ **Méthodes « reference-free», Comment passer à l'échelle**

NGS pipelines are leaky

Bioinformatics challenge





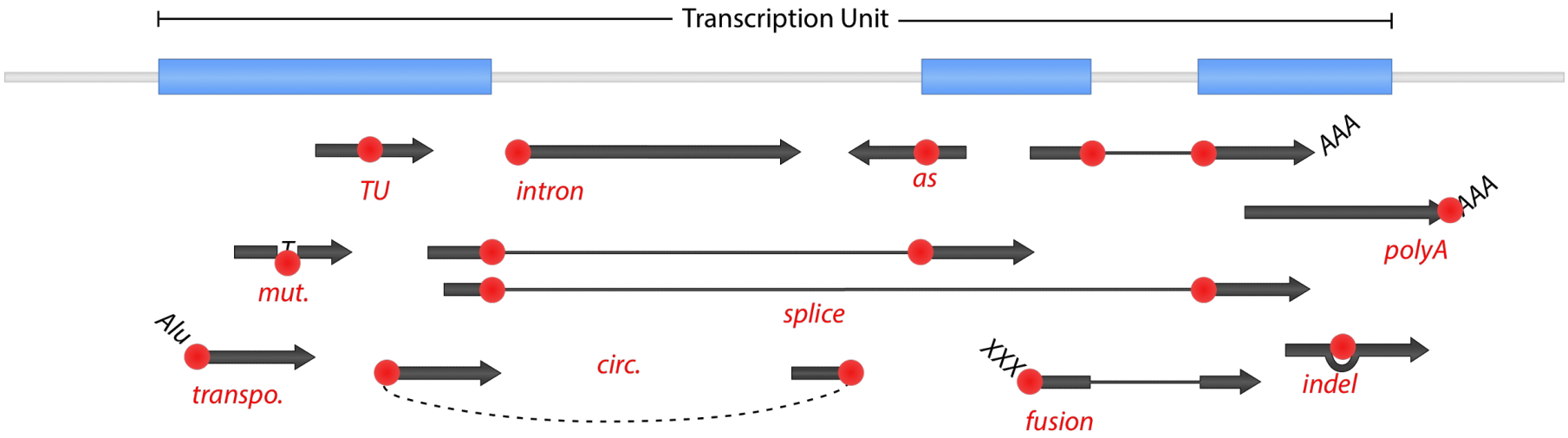
Objectifs of kmer approaches

- High quality analysis of public datasets
- In depth analysis of the whole transcriptome
- Large scale request

- Use of RNAseq resources in biology
- New Biomarkers (outside reference)
- Many applications in biology, human health, pharmaceutics, biomarkers...

k-mer as signature for genomic unit

K ~31



Advantages

- **Without référence**
- **Large pannel events**
- **Rapid counting**
- **Large datasets**

Limits

- **Specificity of *k-mers*** vs complete transcrit
- **Polymorphisms**
- Signal/Noise low ratio

Quantification with kmers

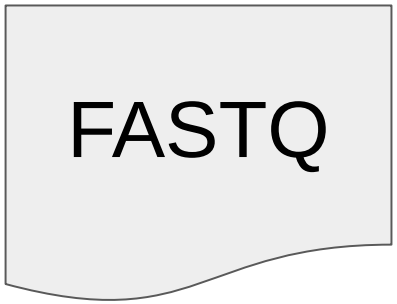
Example k = 4

TGTCAGTGTCGTCGCTAGTAG

TGTCAGTGTCGTCGCTAGTAG

...

TGTCAGTGTCGTCGCTAGTAG



Very fast

Counting *k-mers*

(jellyfish, DSK, ...)

k-mer	abundance
TGAC	3
AATG	34
...	...
TATG	12

Objectifs of kmer approaches

- **High quality analysis of public dataset**
- In depth analysis of the whole transcriptome
- Large scale request

- Use of RNAseq resources in biology
- New Biomarkers (outside reference)
- Many applications in biology, human health, pharmaceutics, biomarkers...

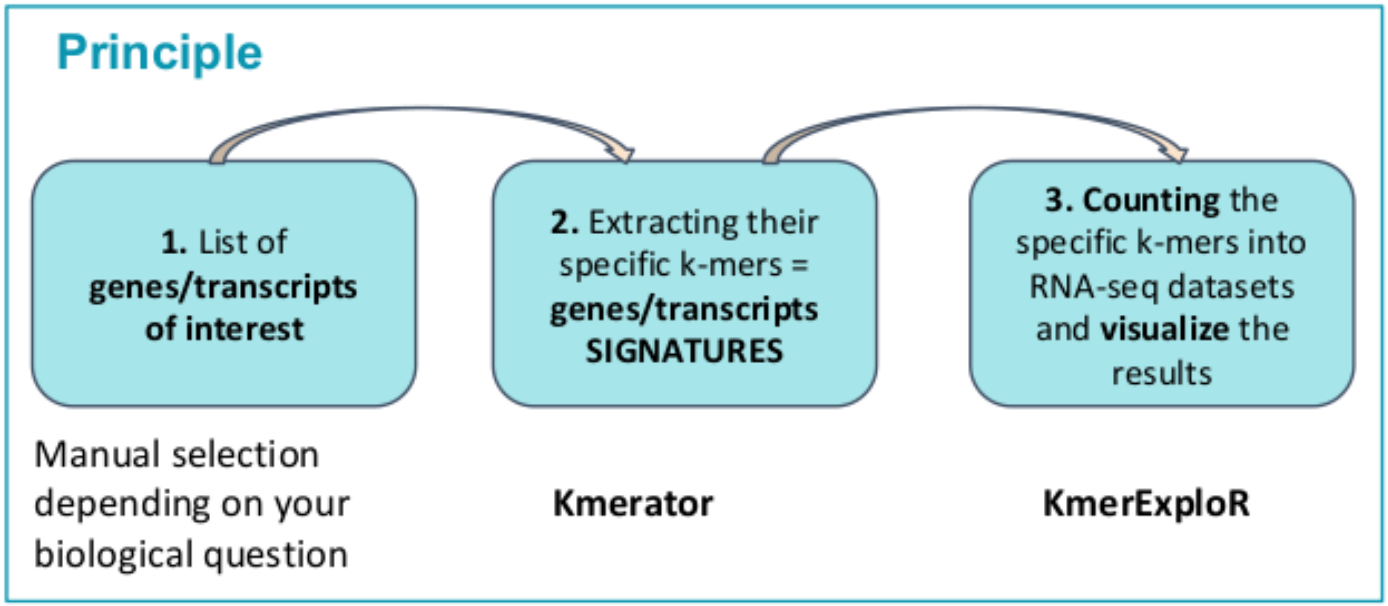
Transcriptome and RNAseq

- Public datasets
- Quality of datasets ?

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 database*

KmeratoR Suite to quickly explore large RNA-Seq datasets

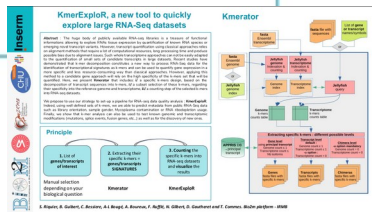


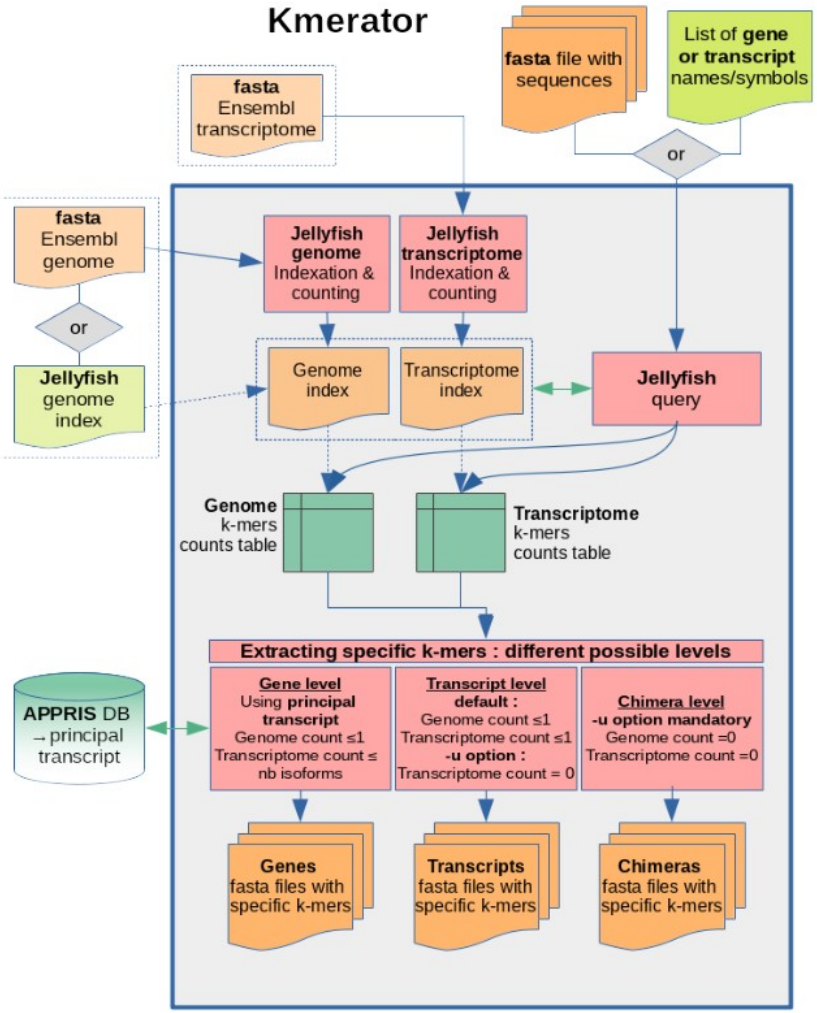
Platform : https://bio2m.montp.inserm.fr/

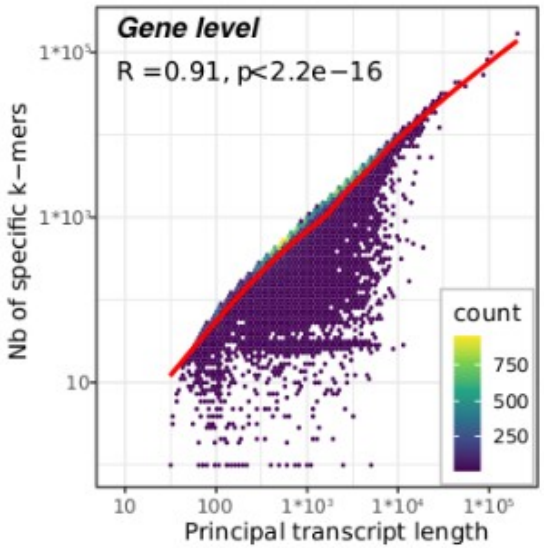
Contacts : benoit.guibert@inserm.fr, sebastien.riquier@inserm.fr, chloe.bessiere@inserm.fr, Therese.Commes@inserm.fr

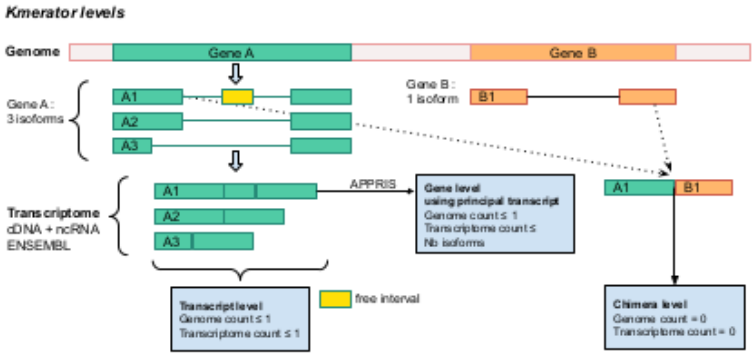
https://github.com/Transipedia/kmerator

https://github.com/Transipedia/kmerexplor

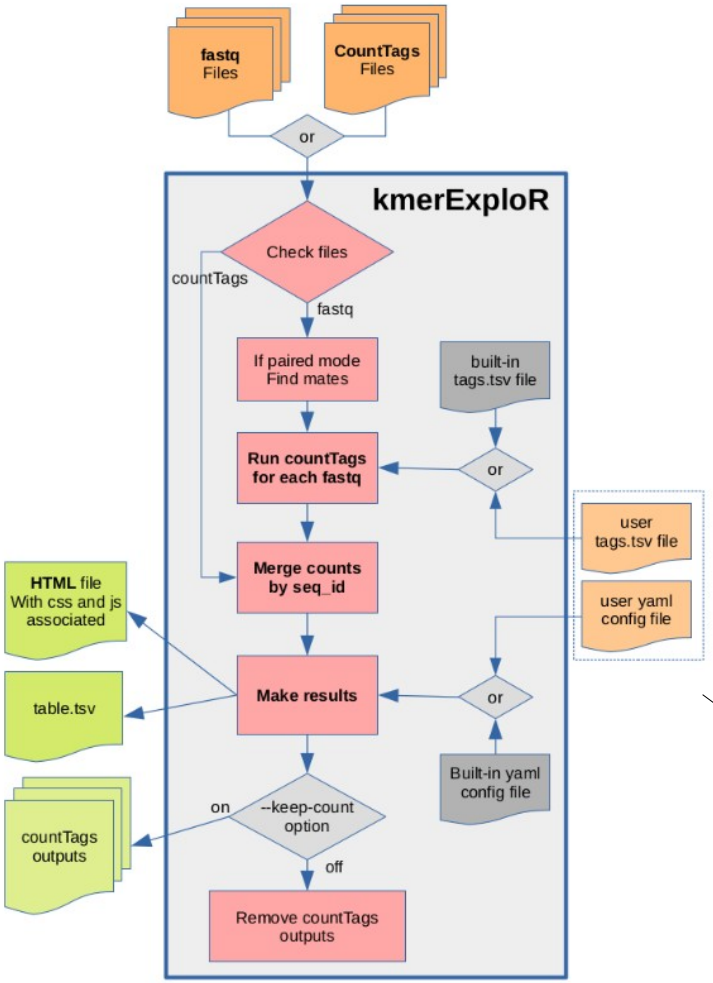






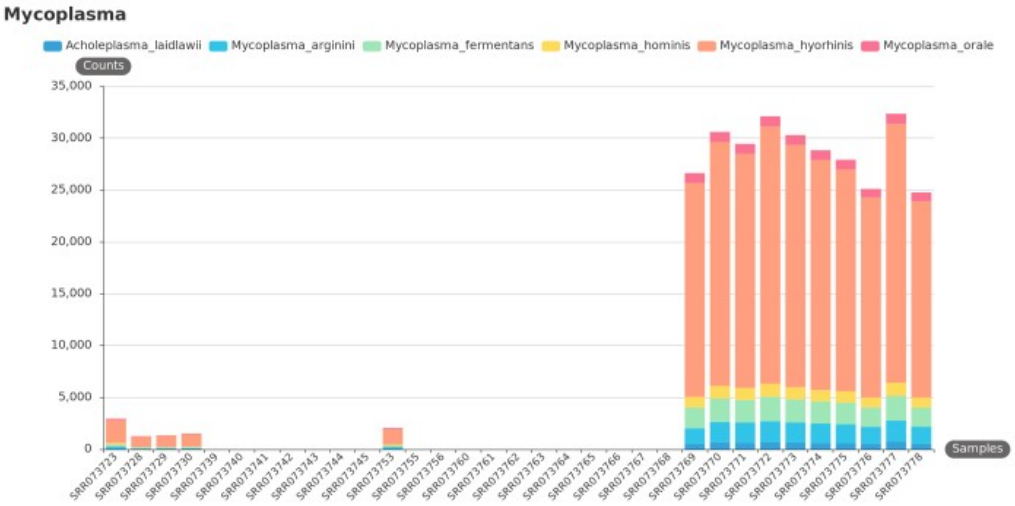


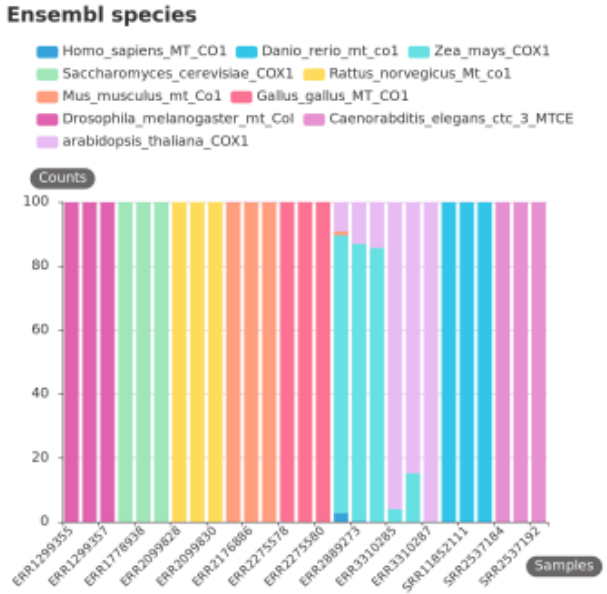
To extract kmers in a specific way for
Gene or Transcript signatures

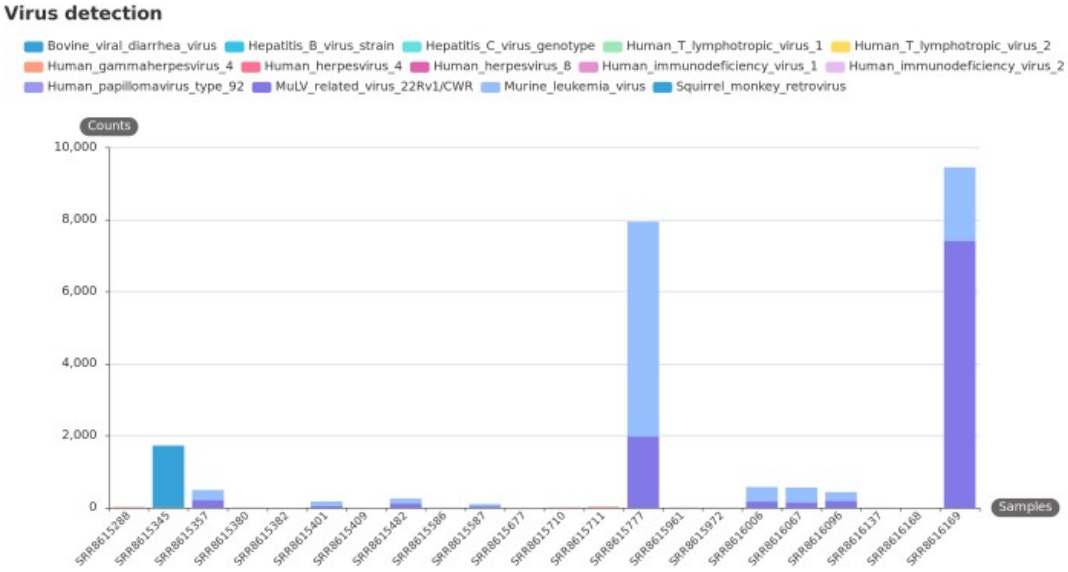


To count kmer predictors for in depth RNAseq analysis

Methodological questions







Most frequent biological questions

Objectifs of kmer approaches

- High quality analysis of public dataset
- In depth analysis of the whole transcriptome
- **Large scale request**

- Use of RNAseq resources in biology
- New Biomarkers (outside reference)
- Many applications in biology, human health, pharmaceutics, biomarkers...

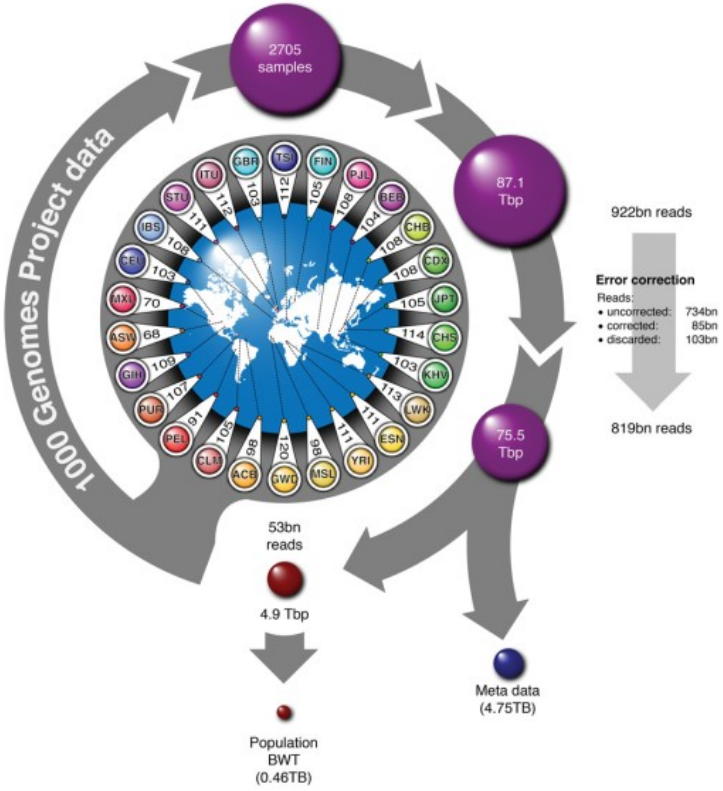
Donner plusieurs exemples d’approches large scale kmers pour genomes et transcriptomes
Redoak…

https://gite.lirmm.fr/doccy/RedOak

Différentes approches (cf Revue Camille Marchet)

Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes

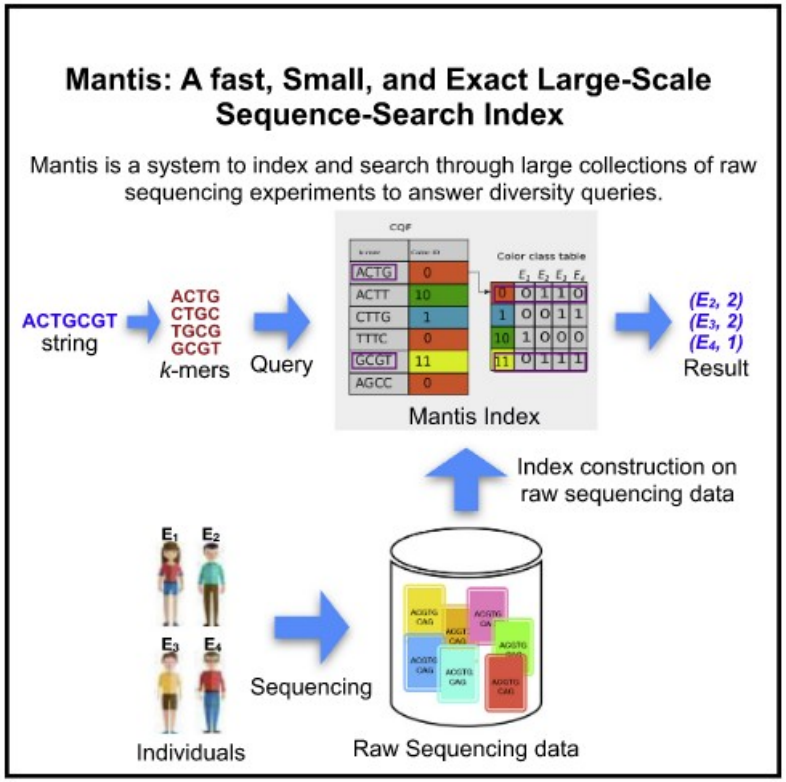
Dirk D. Dolle,[1,6] Zhicheng Liu,[1,2,6] Matthew Cotten,[1] Jared T. Simpson,[3,4] Zamin Iqbal,[5] Richard Durbin,[1] Shane A. McCarthy,[1] and Thomas M. Keane[1,2]



Cell Systems

Mantis: A Fast, Small, and Exact Large-Scale Sequence-Search Index

Graphical Abstract



Highlights

- Mantis is a tool to search through large collections of raw sequencing experiments
- Mantis index is 20% smaller than the Split-Sequence Bloom Tree (SSBT) search index

Authors

Prashant Pandey,
Fatemeh Almodaresi,
Michael A. Bender, Michael Ferdman,
Rob Johnson, Rob Patro

Correspondence

rob.patro@cs.stonybrook.edu

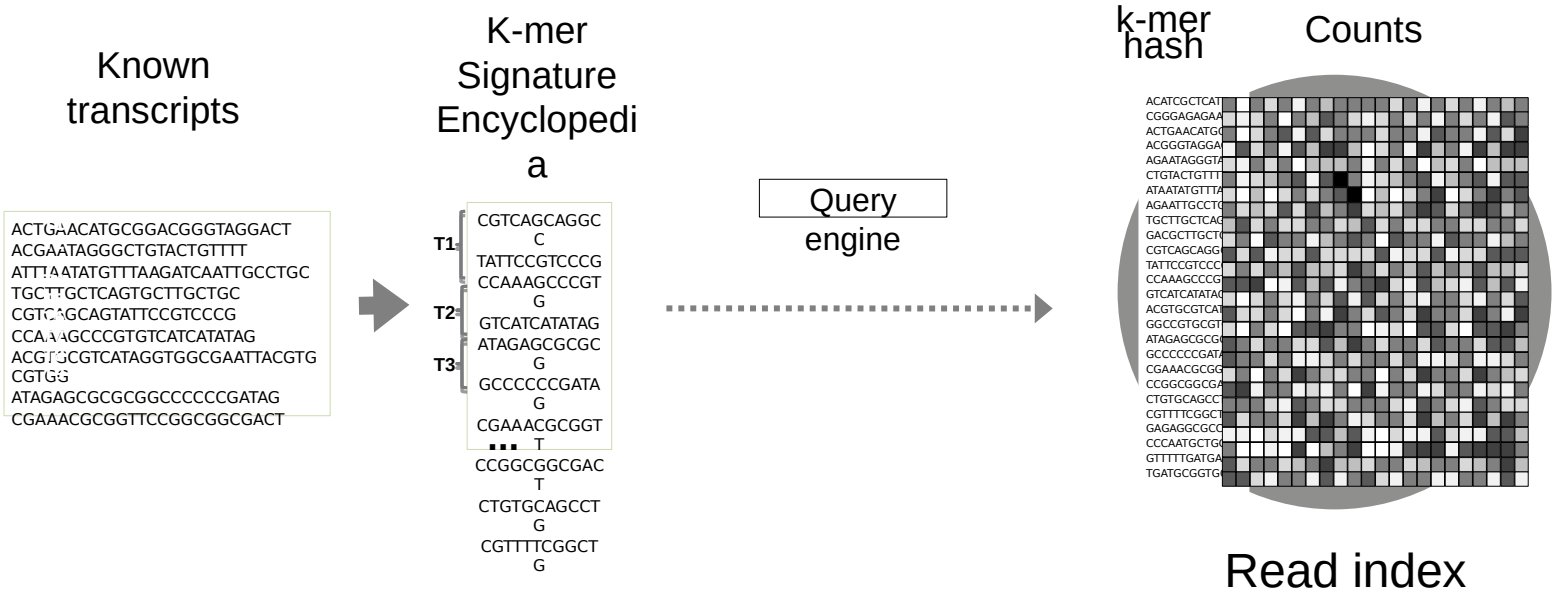
In Brief

Mantis is a system to index and search through large collections of raw sequencing data. The query sequence can be a known or newly assembled gene or any valid nucleotide sequence. Mantis is faster and smaller than existing sequence-search tools and is exact in the sense that it does not report false-positives. To construct the index, Mantis indexes the k-mers (substrings of size k) in the reads of an experiment and then groups k-mers across experiments that exhibit the same patterns of occurrence.

Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;27(6):764–70.

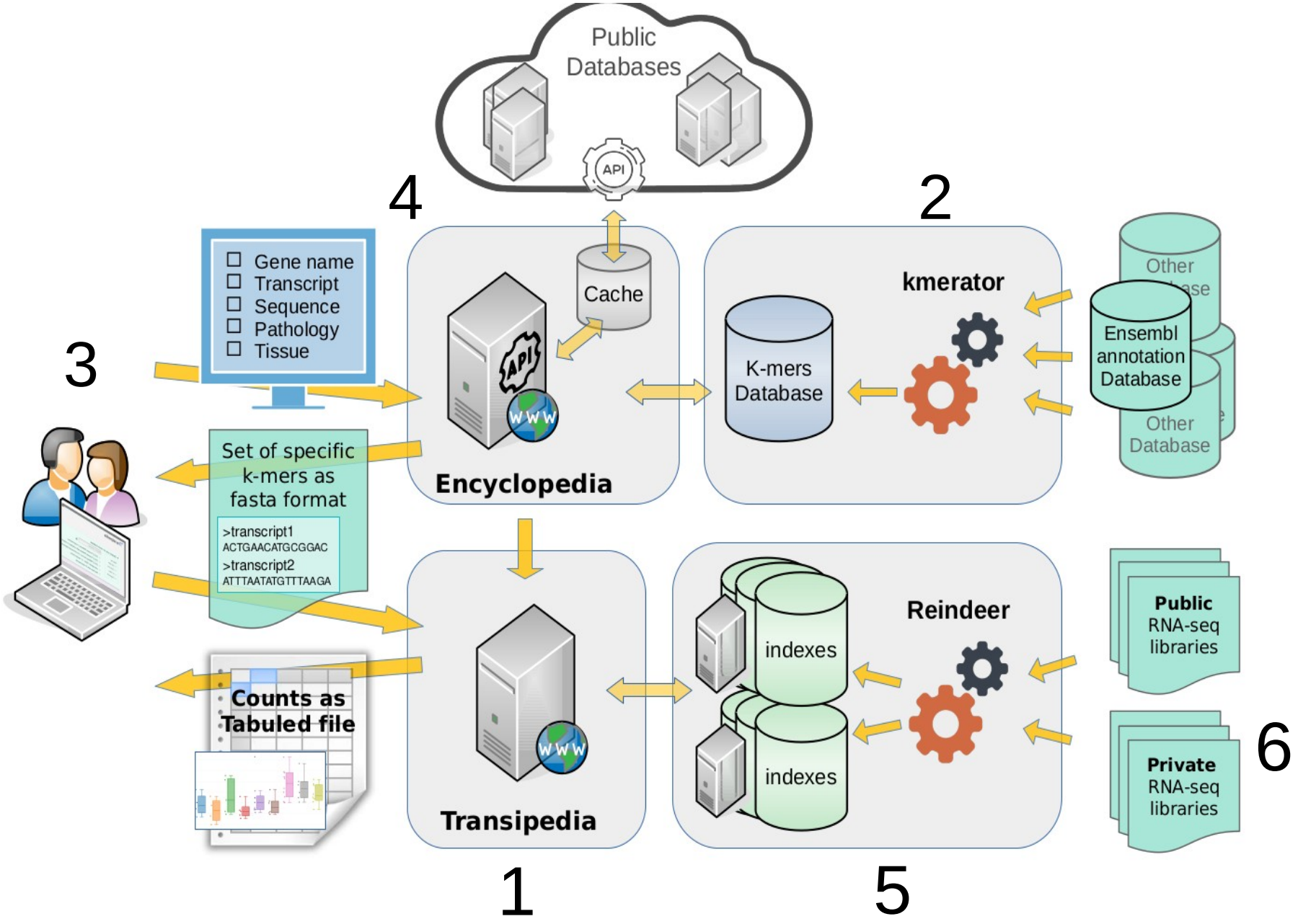
- Kmer index
- Kmer counting
- Kmer request
-
- 3 notions

The TranSiPedia Concept



(1000's of RNA-seq libraries)
(Marchet C & Chikhi R, 2020)

- Discovery of novel transcripts in very large databases
- New disease biomarkers & other medical applications


Public Databases
API
4
2
Gene name
Transcript
Sequence
Pathology
Tissue
3
Cache
kmerator
Other
Ensembl annotation Database
Other Database
K-mers Database
Encyclopedia
Set of specific k-mers as fasta format
>transcript1
ACTGAACATGCGGAC
>transcript2
ATTTAATATGTTTAAGA
Counts as Tabuled file
Transipedia
Reindeer
indexes
indexes
Public RNA-seq libraries
Private RNA-seq libraries
6
1
5

Results

1- Transipedia web site (demo) for request

2- A software for specific kmers design (Kmerator, NARgab)

3- Development of tools for mutations and chimeric RNA (kmers design for mutations and fusion RNA (signatures))

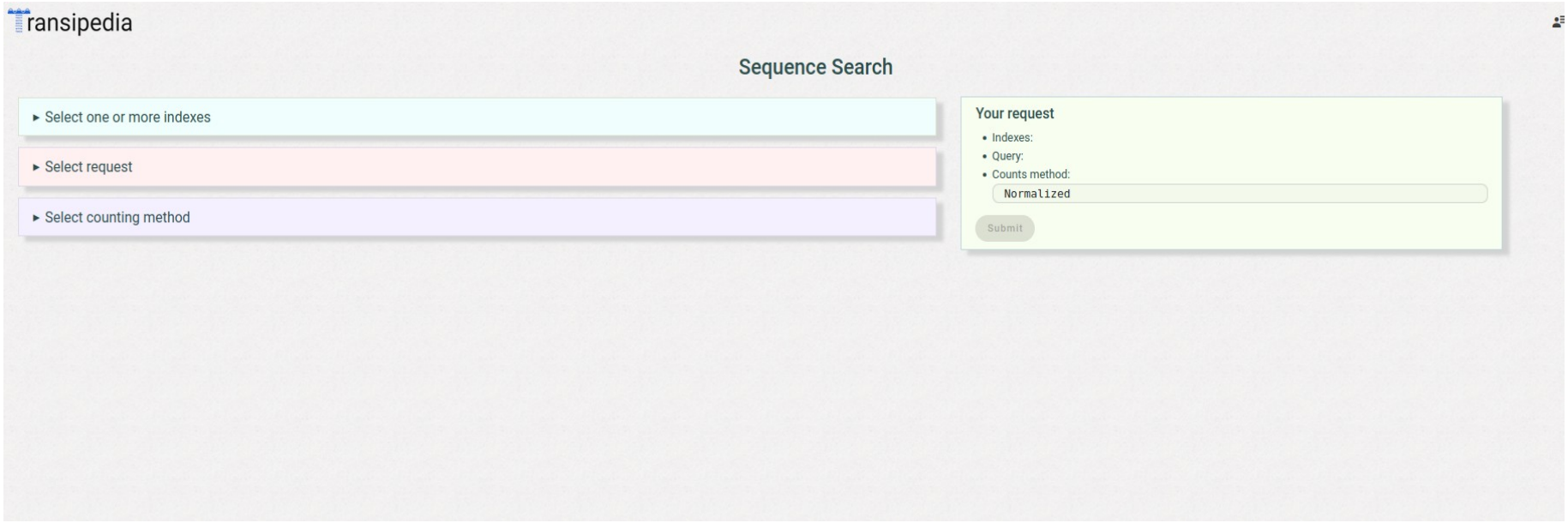
4-Encyclopedia of kmers (prototype BD for transcriptome reference)

5-Reindeer index (see ref)

6-Request Applications

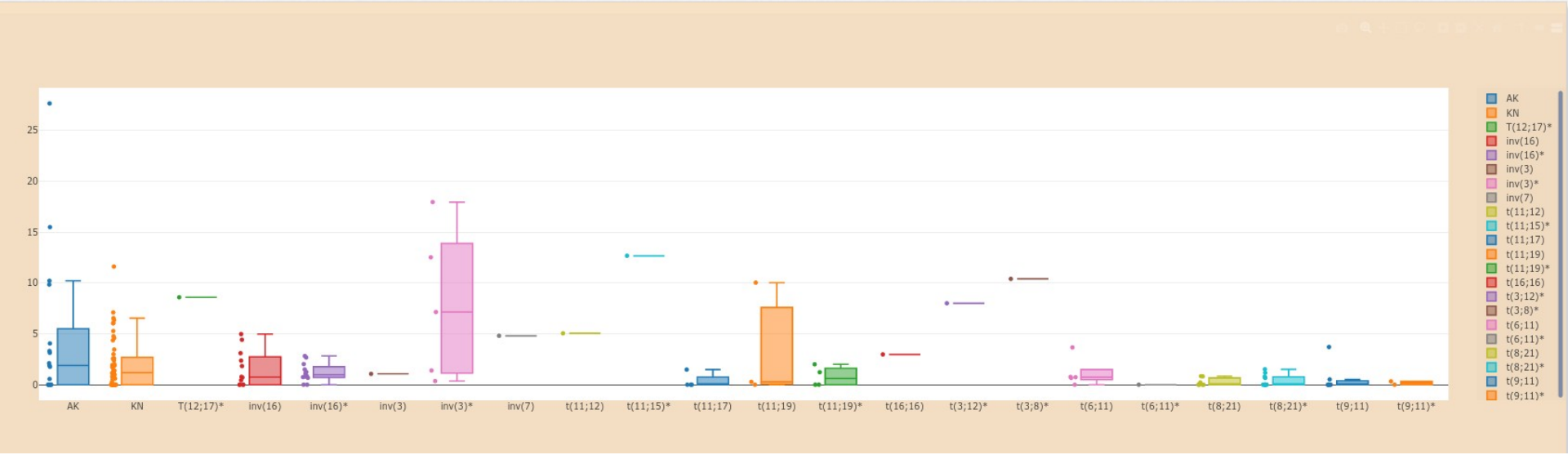
Transipedia website

https://transipedia.montp.inserm.fr/



st

Leucegene with metadata (148 samples)



In a few seconds

seq_name	SRR8615300	SRR8615547	SRR8618305	SRR8616169	SRR8616157	SRR8615641	SRR8615579
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer14	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer15	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer16	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer17	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer18	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer19	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer20	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer21	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer22	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer23	0	0	0	0	0	0	0
Chimera-ABR___PDCD6-ABR___PDCD6_112.kmer24	0	0	0	0	0	0	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer7	0	13	91	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer8	0	13	91	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer9	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer10	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer11	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer12	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer13	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer14	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer15	0	13	95	45	0	11	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer16	0	0	0	0	0	0	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer17	0	0	0	0	0	0	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer18	0	0	0	0	0	0	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer19	0	0	0	0	0	0	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer20	0	0	0	0	0	0	0
Chimera-ADNP2___OR5H15-ADNP2___OR5H15_140.kmer21	0	0	0	0	0	0	0

Acute Myeloid Leukemia (AML)

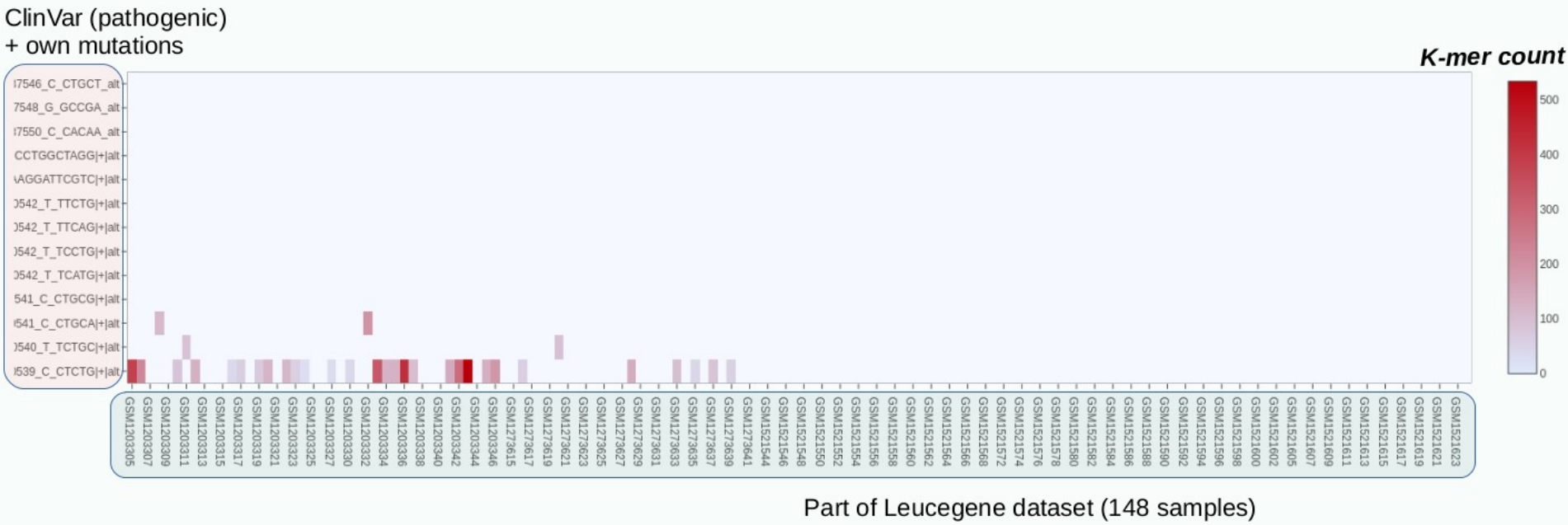
- Disease primarily of older adults
- There are different subtypes of AML (e.g European LeukemiaNet classification)

Based on genetics

Risk category	Genetic abnormality
Favorable	t(8;21)(q22;q22.1) RUNX1-RUNX1T1; inv(16)(p13.1q22) or t(16;16)(p13.1;q22) CBFB-MYH11; **Mutated NPM1** without FLT3-ITD or with FLT3-ITDlow†; Biallelic mutated CEBPA
Intermediate	**Mutated NPM1** and FLT3-ITDhigh†; Wild-type NPM1 without FLT3-ITD or with FLT3-ITDlow† (without adverse-risk genetic lesions); t(9;11)(p21.3;q23.3); MLLT3-KMT2A‡; Cytogenetic abnormalities not classified as favorable or adverse
Adverse	t(6;9)(p23;q34.1); DEK-NUP214; t(v;11q23.3); KMT2A rearranged; t(9;22)(q34.1;q11.2); BCR-ABL1; inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2,MECOM(EVI1); −5 or del(5q); −7; −17/abn(17p); Complex karyotype,§ monosomal karyotype‖; **Wild-type NPM1** and FLT3-ITDhigh†; Mutated RUNX1¶; Mutated ASXL1¶; Mutated TP53#

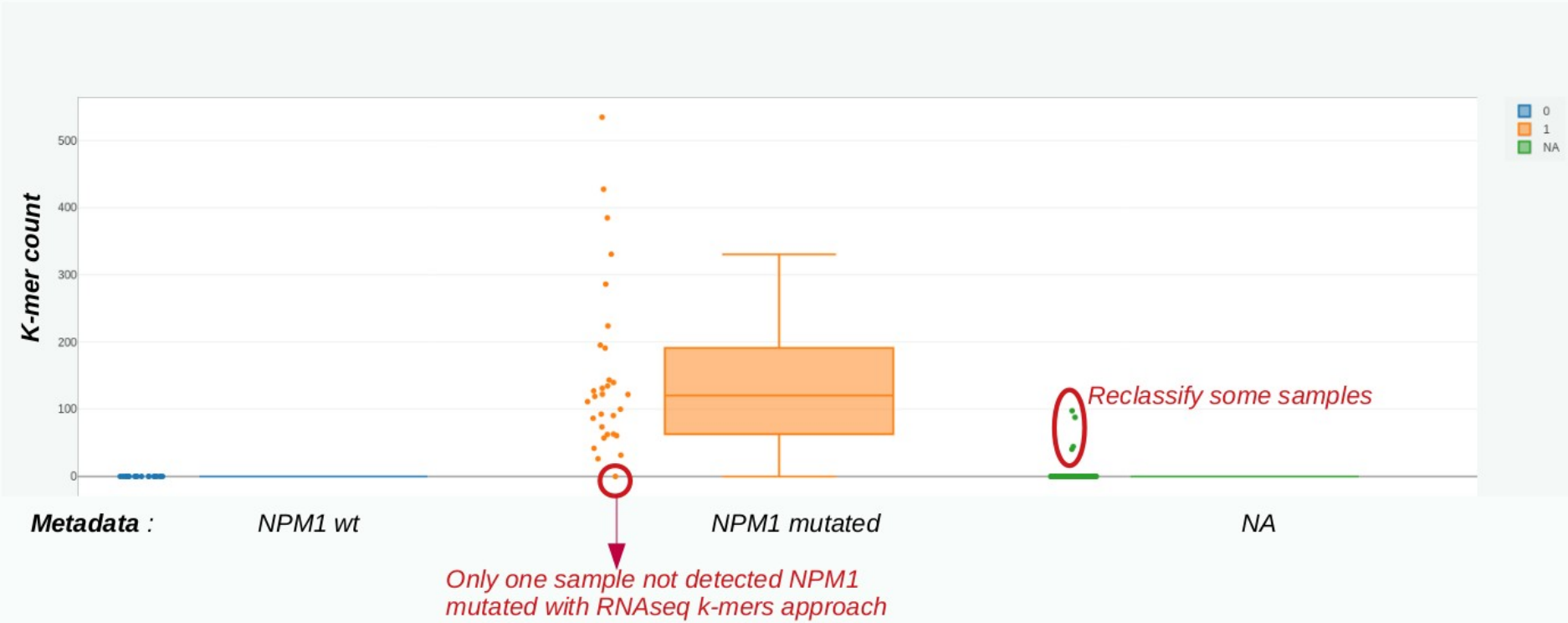
- Prognosis of AML patients is influenced by patient-associated factors (age, conditions) and by disease-related factors (white-cell count, leukemic-cell changes)

NPM1 mutation profile into Leucegene data



https://transipedia.montp.inserm.fr/

NPM1 mutation profile into Leucegene data (2)



https://transipedia.montp.inserm.fr/

Bioinformatics, 36, 2020, i177–i185
doi: 10.1093/bioinformatics/btaa487
ISMB 2020

OXFORD

REINDEER: efficient indexing of *k*-mer presence and abundance in sequencing datasets

Camille Marchet[1,*], Zamin Iqbal[2], Daniel Gautheret[3], Mikaël Salson[1] and Rayan Chikhi[4]

A prototype with CCLE dataset

nombre d'échantillons	temps prévu	volume généré	volume en entrée
1000	8 h	48 Go	11 To
10000	50 h	480 Go	110 To
100000	500 h	4.8 To	1100 To

https://www.youtube.com/watch?v=CbTaM5zX09U

Tp Bioinfo - KmerExploR

- **Objectif:** analyse qualité des données RNaseq, approches par k-mers,
- recherche d'information biologiques dans les base de données,
- création d'un nouveau module de kmerExploR
-
  - **Lundi 9 Octobre :** CM1 et TD1 (présentation des objectifs et des outils,
  - base de données, questions biologiques, recherche des séquences d'intérêt.....) _
  - TCommes
  - intérêt des Kmers/Transipedia / (demo)
  - sujet sur la question bio

  - **Lundi 16 Octobre: TD2/TP1** (Présentation des outils kmerator et KmerExplor, le code , comment ça fonctionne) _
  - BGuibert , vérification des séquences choisies, accès aux jeux de données RNAseq tests
  - **Lundi 6 Novembre:** TP2 Présentation des résultats, questions .... _BGuibert, TCommes

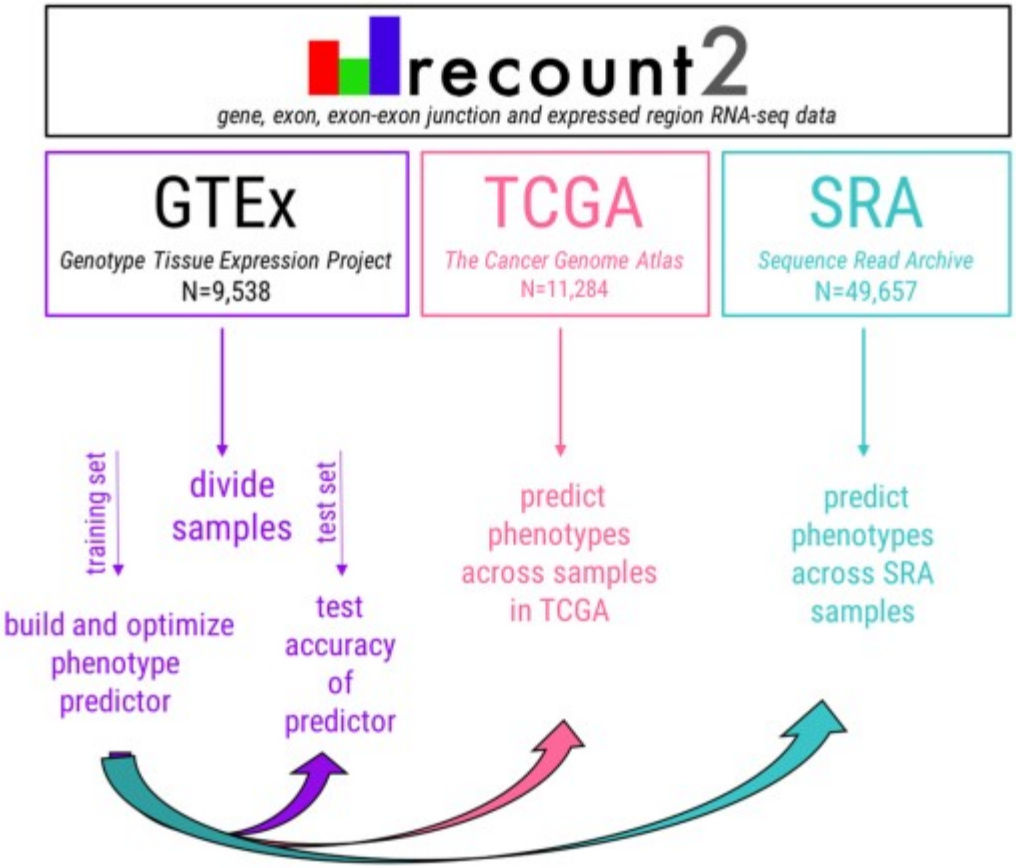


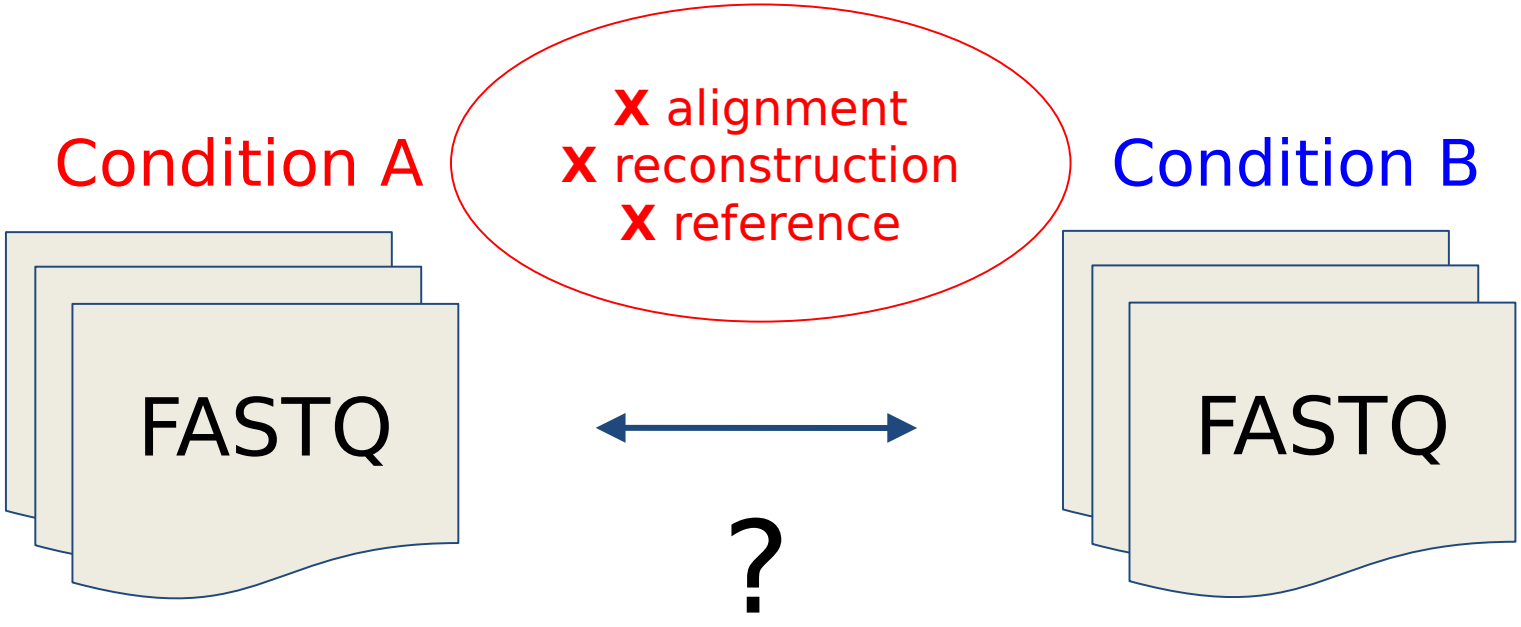
Figure 2. General approach to phenotype prediction. To predict phenotype information, the training data are first randomly divided and the predictor is built. Accuracy is first tested in the training data. Upon achieving sufficient accuracy ($\geq$85%), the predictor is tested in the remaining half of the training data set. Phenotypes can then be predicted across all samples in `recount2`.

Objectifs of kmer approaches

- High quality analysis of public dataset
- **In depth analysis of the whole transcriptome**
- Large scale request

- Use of RNAseq resources in biology
- New Biomarkers (outside reference)
- Many applications in biology, human health, pharmaceutics, biomarkers...

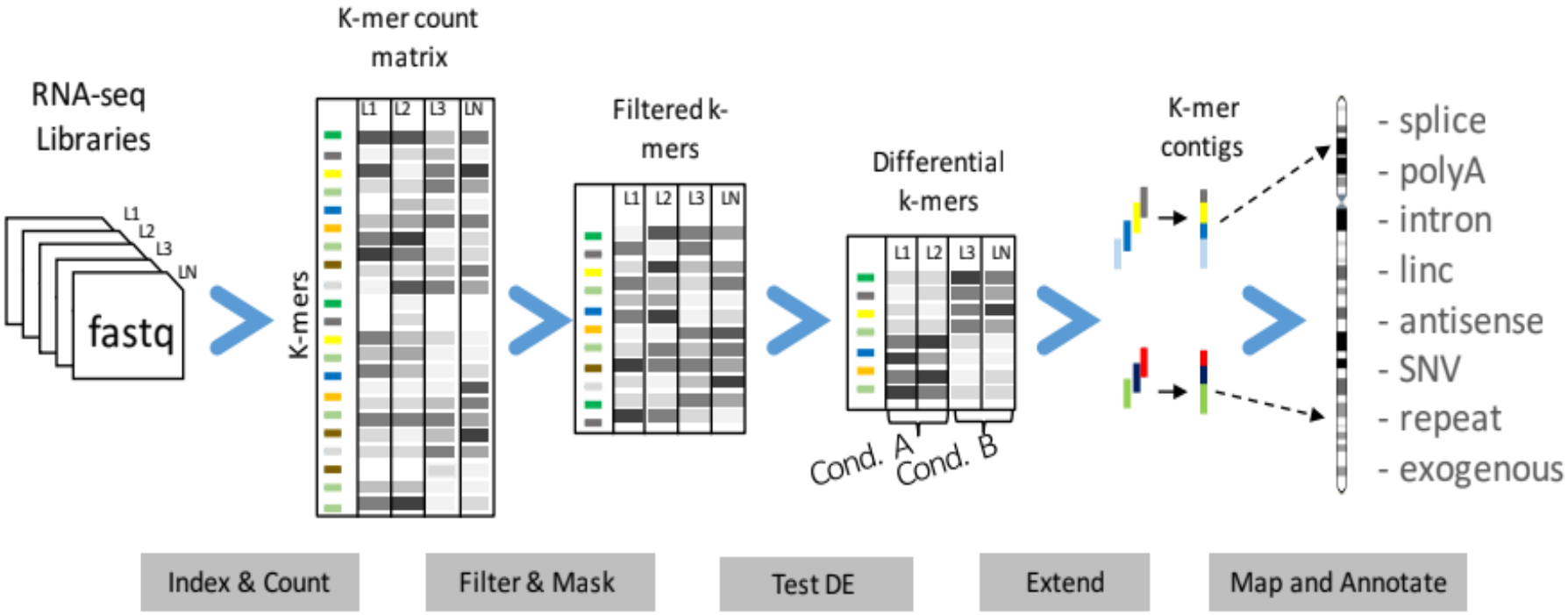
Comparison “*sans a priori*” of RNA-Seq data for new differential transcriptional events



Kmer pipeline

- De-Kupl
- Imoka
- Kmarat
- Gecko..
-

DE-kupl : find functional K-mers through Differential Expression

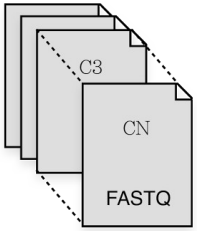


Audoux et al. Genome Biol, 2017

M2-BCD

DE-kupl procedure (New transcript events)

***N* RNA-Seq samples**



k-mers counts (Jellyfish, DSK..), normalisation

	C1	C2	CN
kmer 1	C11	C21	CN1
kmer 2	C12	C22	CN2
kmer 3	C13	C23	CN3
..	..	..	..
kmer N	C1N	C2N	CNN

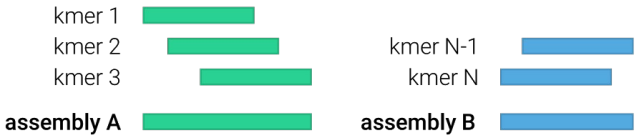
Filter annotated and non–recurrent k-mers

Extract differential *k-mers* (DE)

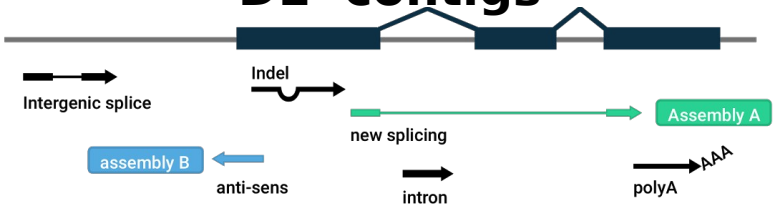
	Conditon A		Condition B		statistics log2FC	p-value
kmer 1						
kmer 2						
kmer 3						
..	..	..	..	..	..	..
kmer N						

Filter non-DE *k-mers*

Assembling DE *k-mers* in contings

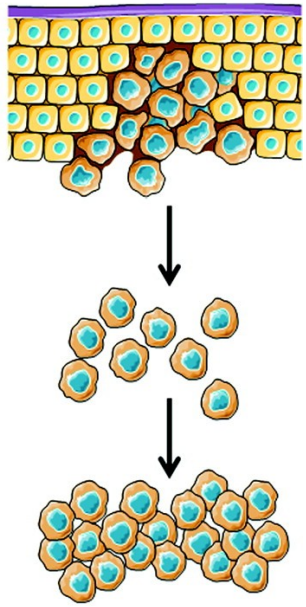


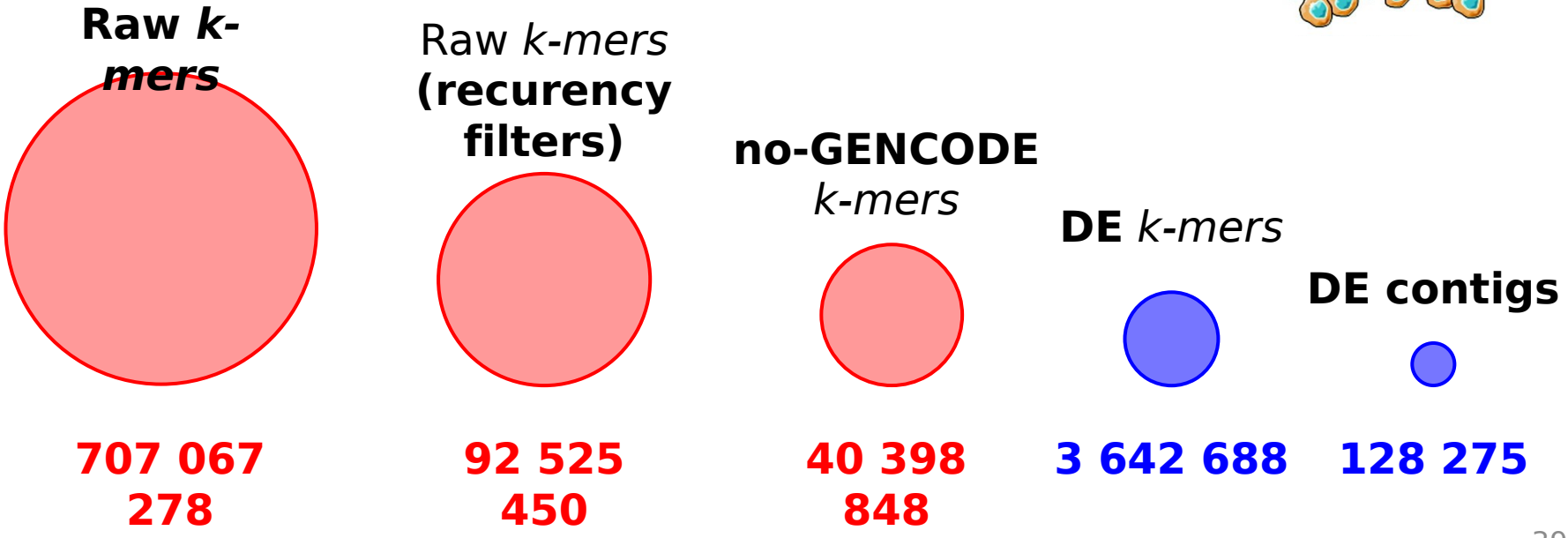
Alignment & annotations of DE contigs



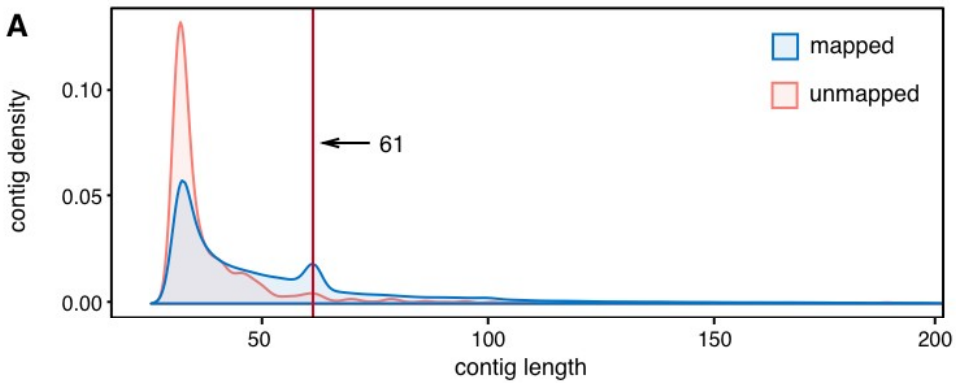
Dataset=Epithelial-Mesenchymal Transition experiment

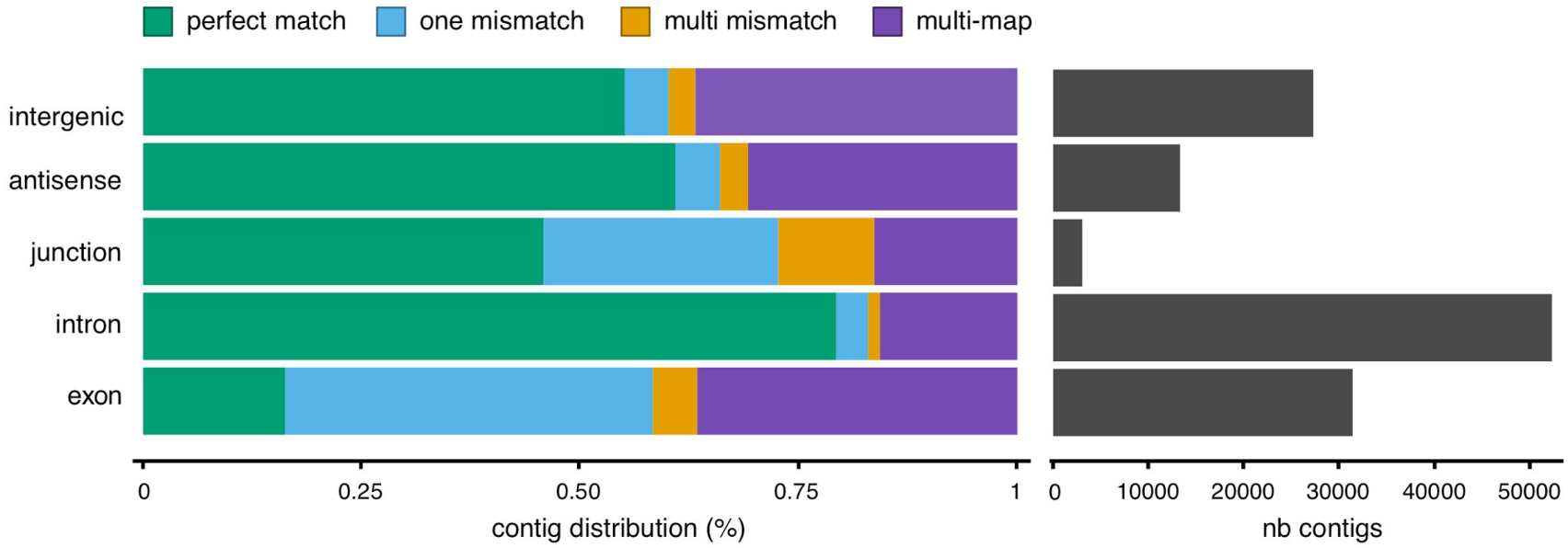
- **12 RNA-Seq cell line EMT** model
- **Performances :** 10 h, (4 cores, 16Gb RAM, 60Gb hard disk).





Localisation des contigs DE





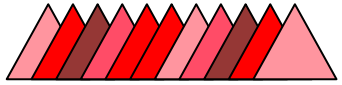
99.2% des 128k contigs DE alignés sur le génome humain à **1 633 loci différents**

Towards hypothesis-free theranostics*

Patients under treatment

Responders

Non responders

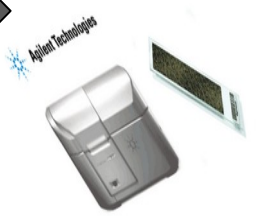


RNA-seq

K-mer matrix

Predictive k-mers

DNA chip



*diagnostic that distinguishes patient or disease type and allows selection of therapy

KAMRAT /IMOKA

A Comparative Analysis of Reference-Free and Conventional Transcriptome Signatures for Prostate Cancer Prognosis

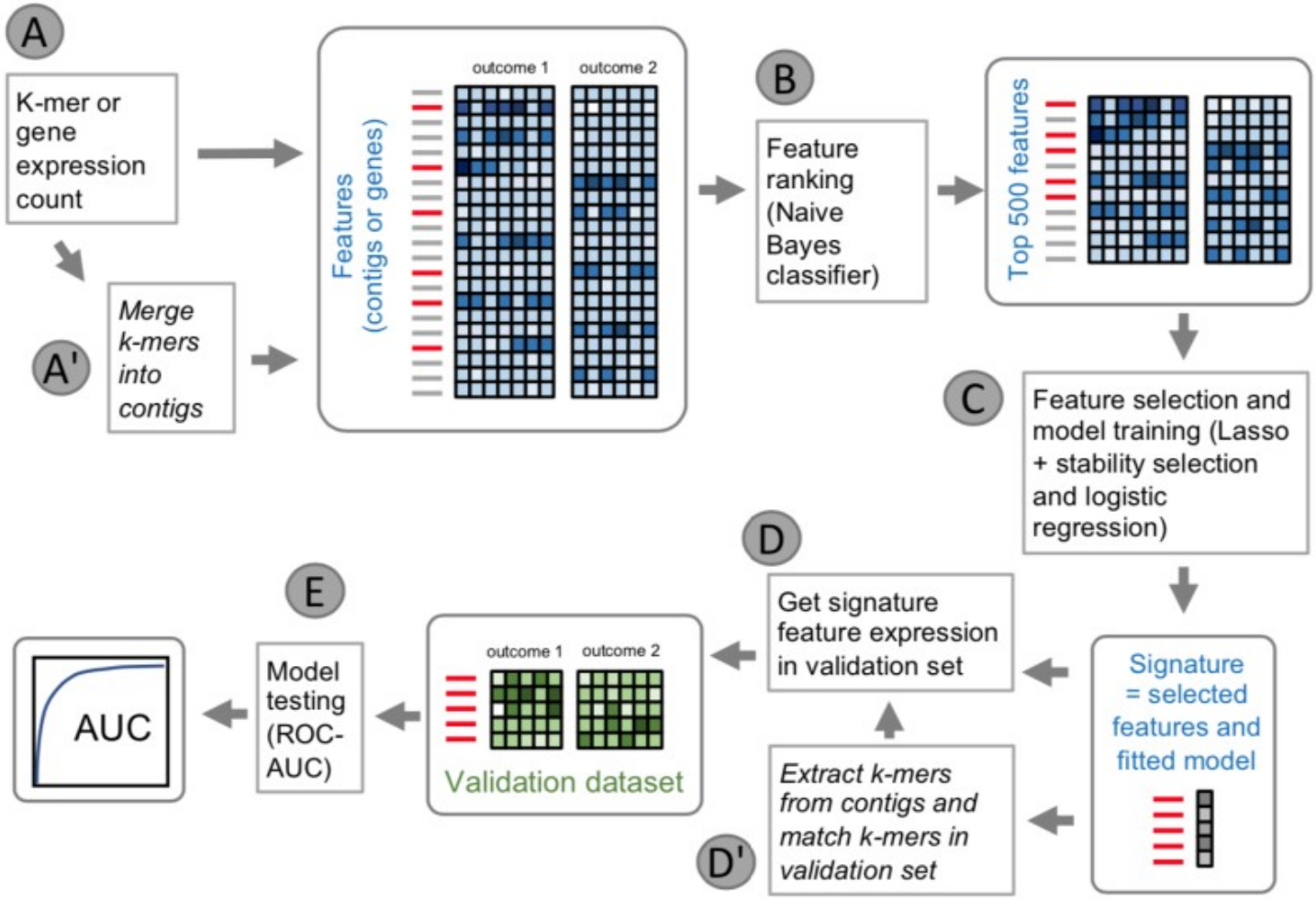
Ha TN Nguyen[1], Haoliang Xue[1], Virginie Firlej[2], Yann Ponty[3], Mélina Gallopin[1], Daniel Gautheret[1*]

September 20, 2020

METHOD Open Access

iMOKA: *k*-mer based software to analyze large collections of sequencing data

Claudio Lorenzi[1], Sylvain Barriere[1], Jean-Philippe Villemin[1], Laureline Dejardin Bretones[1], Alban Mancheron[2] and William Ritchie[1*]



https://github.com/i2bc/PCa-gene-based_vs_gene-free/tree/master/KaMRaT

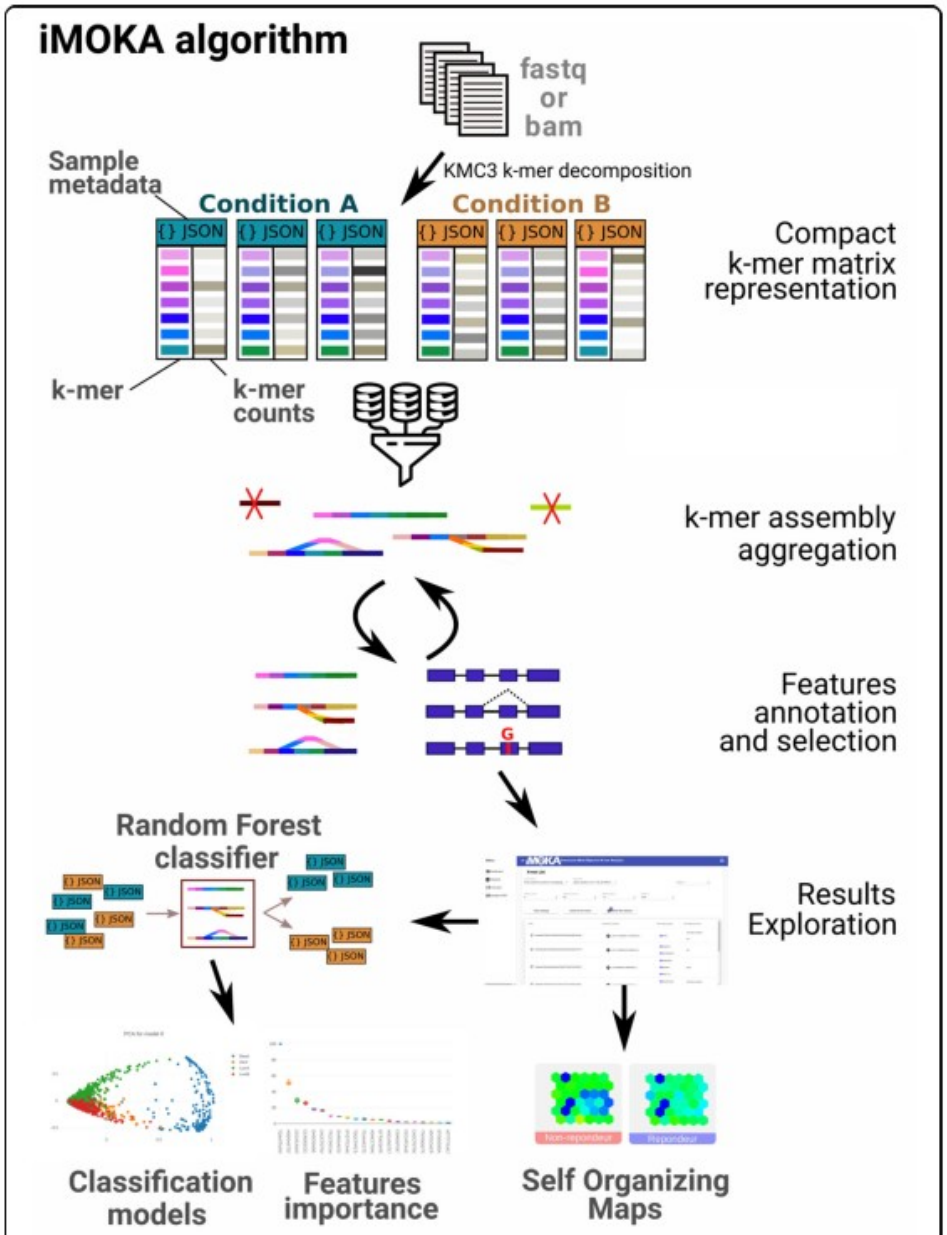


Fig. 1 Overview of the iMOKA algorithm. The software accepts sequencing reads in FASTQ, FASTA, BAM formats, or SRR identifiers. The *k*-mer count in each file is calculated and stored using a dedicated file format. *k*-mers are then filtered using an Entropy boosted Bayes filter with Monte Carlo cross validation to obtain the *k*-mers that are able to classify the input samples. These are combined into graphs and annotated using GMAP or another user-defined aligner. The final list of highly informative *k*-mers can be explored using the graphical interface to create classification models, inspect individual *k*-mers, and detect sample outliers using self-organizing maps