# Exploring large RNA datasets with k-mers

**Daniel Gautheret**

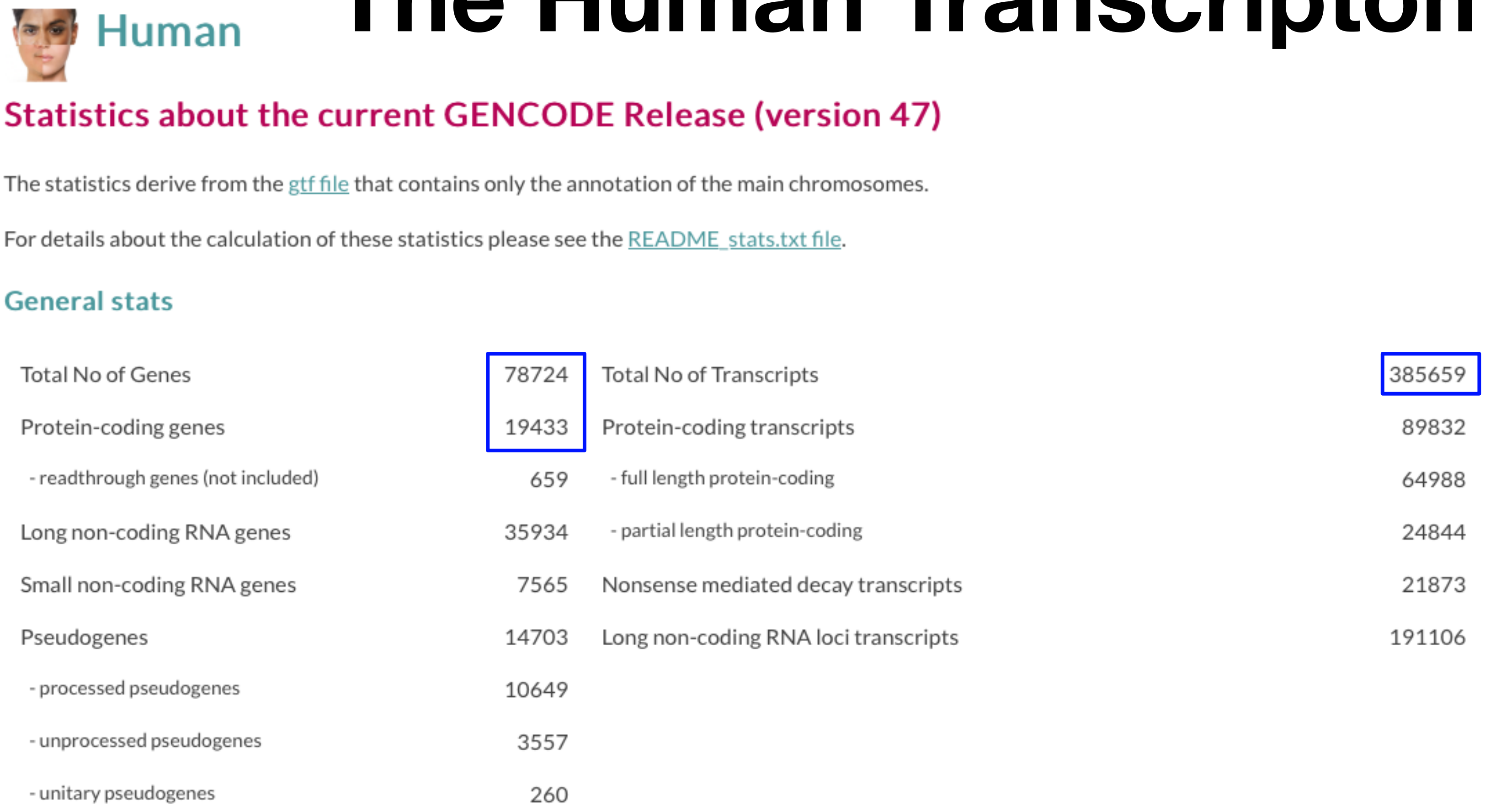
*Master GenE2*

*UE Big Data*

I2BC

université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

# The Human Transcriptome



Human

## Statistics about the current GENCODE Release (version 47)

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 | 78724 | Total No of Transcripts | 385659 |
| Protein-coding genes | 19433 | Protein-coding transcripts | 89832 |
| - readthrough genes (not included) | 659 | - full length protein-coding | 64988 |
| Long non-coding RNA genes | 35934 | - partial length protein-coding | 24844 |
| Small non-coding RNA genes | 7565 | Nonsense mediated decay transcripts | 21873 |
| Pseudogenes | 14703 | Long non-coding RNA loci transcripts | 191106 |
| - processed pseudogenes | 10649 | | |
| - unprocessed pseudogenes | 3557 | | |
| - unitary pseudogenes | 260 | | |

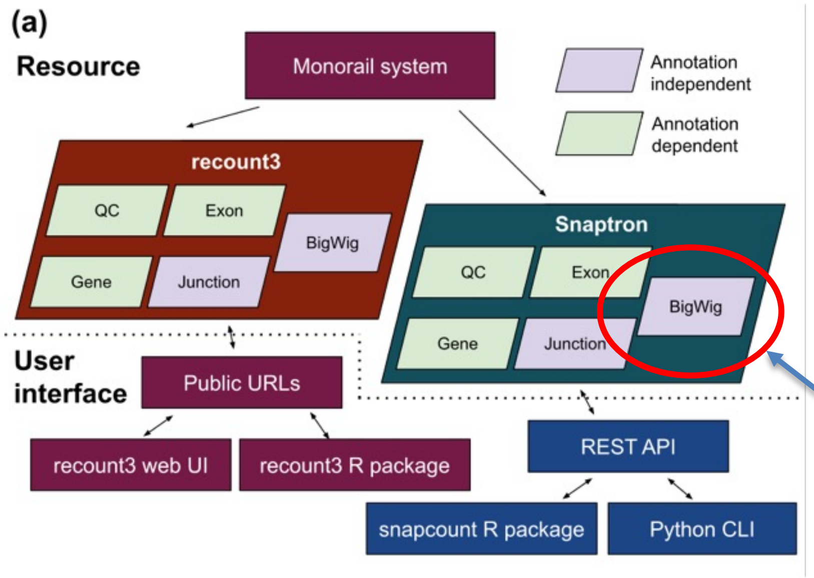
# Diseases reshape our transcriptome

| Pathologic process | Effect on transcriptome | Not in Gencode | Not in Genome |
|---|---|---|---|
| Mutagenic processes (UV, tobacco, aging, repair defects) | SNVs, indels | ✓ | |
| | Fusions | ✓ | ✓ |
| | Microsatellite instability | ✓ | ✓ |
| | SVs, ecDNAs | ✓ | ✓ |
| defaults in gene regulation/ processing | Splice, polyA variants | ✓ | |
| | lncRNAs, circRNAs | ✓ | |
| Epigenetic alterations | Expressed ETs | ✓ | ✓ |
| | Pervasive transcription | ✓ | |
| Infectious agents | Exogenous viral & bacterial RNA | ✓ | ✓ |
| | RNA from neo-integrated virus | ✓ | ✓ |

# What we need

- Capacity to search large sequence datasets (1000’s of samples)
- Non-reference
- Nucleotide-resolution
- Quantitative

# Best human transcriptome index to date: Recount3



⇨ 763,000 samples (human+mouse)

gene-, exon-, or junction-level search

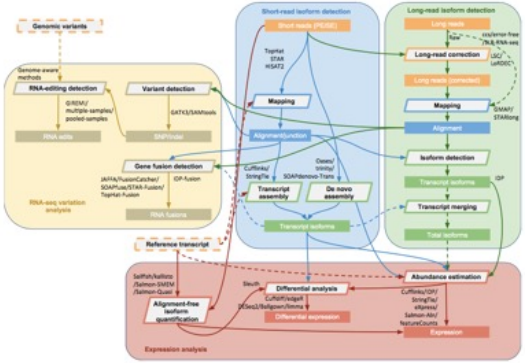
Still reference-based

Has coverage of intergenic regions, but
Mapping-dependent: NO repeats, fusions, unmapped events
No direct query of arbitrary sequence

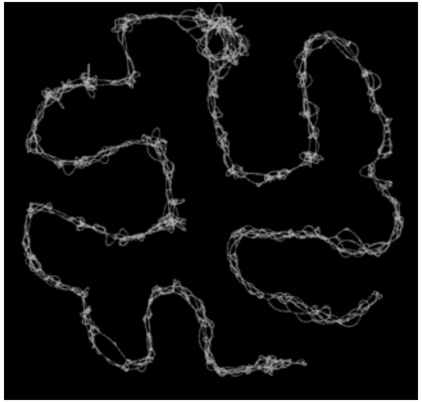
# How to avoid the unique human reference?

## Integrated pipelines

fusions+lncRNA+splice+repeats+
circRNA+virus+de novo assembly...



## Graph genomes

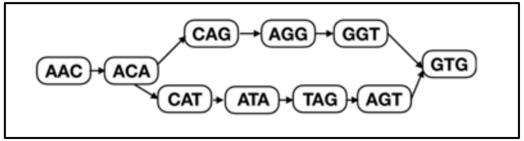


## Real reference-free

```
GTAGAGCTGT
GTA
  TAG
    AGA
      GAG
        AGC
          GCT
            CTG
              TGT
```

k-mers

de Bruijn graph

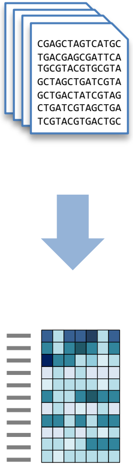


# Exploring RNA diversity with k-mers

ATCAGACTAAA
ATC
TCA
CAG
AGA
GAC
ACT
CTA
TAA
AAA

k-mers = successive subsequences of length *k*

RNA-seq dataset



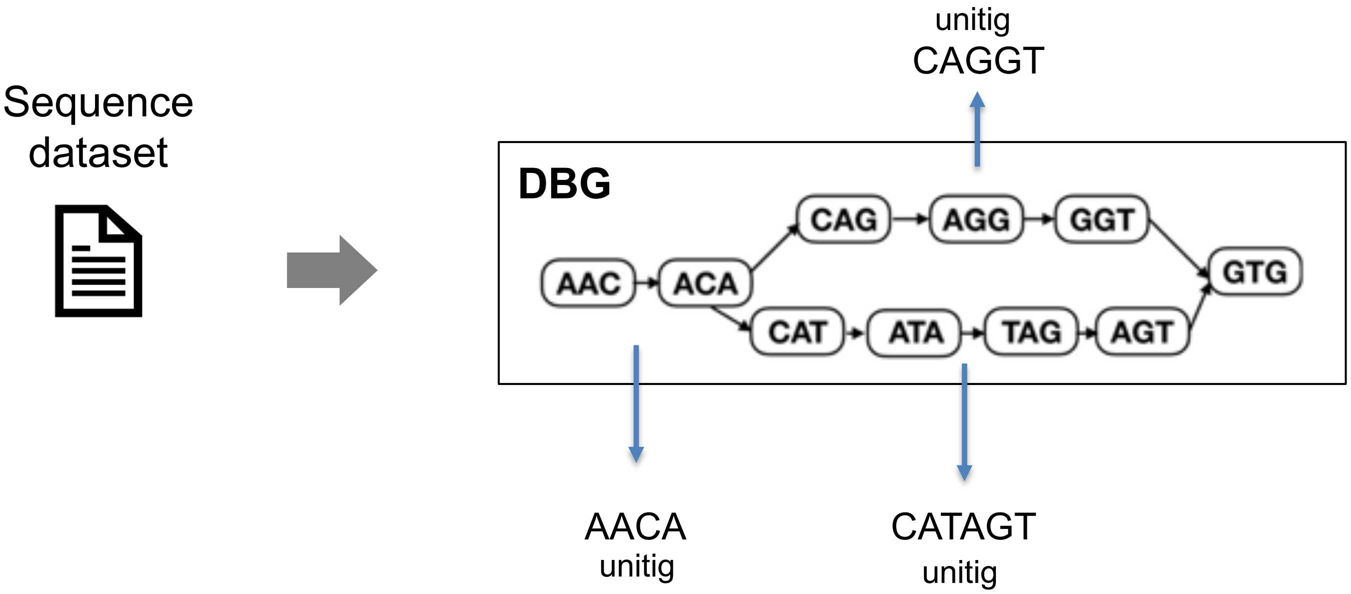
K-mer count index

Reference-free

Scalable

Sustainable

# De Bruijn Graph (DBG)



Sequence dataset

unitig
CAGGT

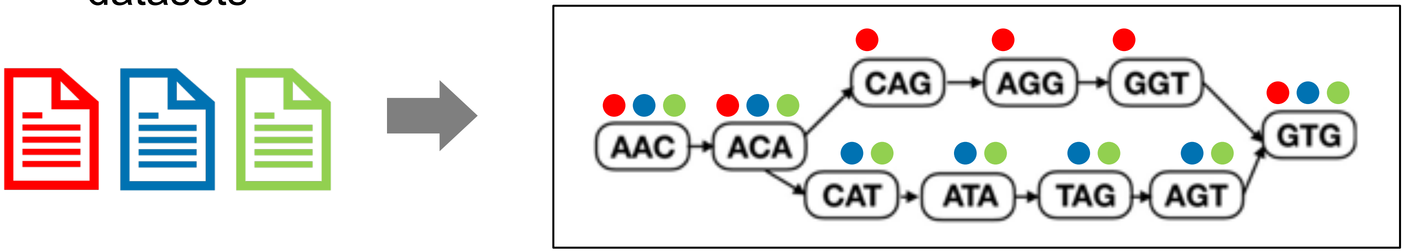
AACA
unitig

CATAGT
unitig

Tools: BCALM, CuttleFish

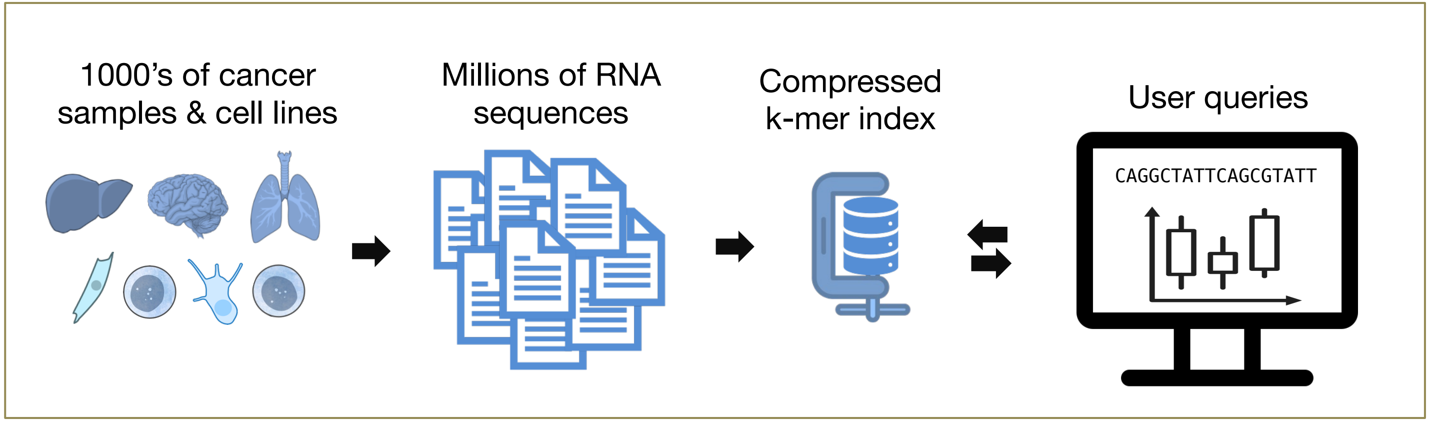
# Colored DBG

Sequence datasets



Colors can represent presence/absence or actual counts

# The Transpedia Project


1000’s of cancer samples & cell lines
Millions of RNA sequences
Compressed k-mer index
User queries
CAGGCTATTCAGCGTATT

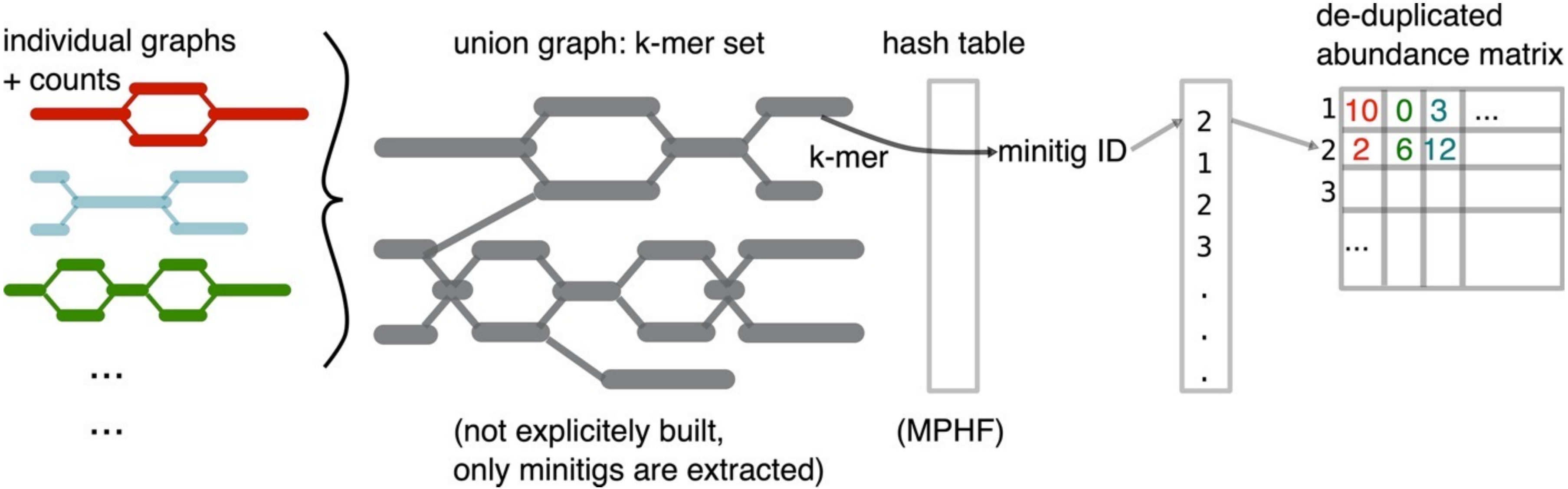
T. Commes (Univ Montpellier)
R. Chikhi (I. Pasteur)
C. Marchet, M. Salson, A. Limasset (CNRS Lille)
D. Gautheret, M. Gallopin (Paris-Saclay)

# Reindeer

With: C. Marchet, M. Salson, R. Chikhi

Bioinformatics 2021

N fastq files



|   |    |   |    |     |
|---|----|---|----|-----|
| 1 | 10 | 0 | 3  | ... |
| 2 | 2  | 6 | 12 |     |
| 3 |    |   |    |     |
|   | ...|   |    |     |

(BCALM)

# Reindeer indexes (on-disk)

| Dataset | #Samples | Fastq.gz size (Gb) | Index size (Gb) | RAM (Gb) | Load time (h:m:s) |
|---|---|---|---|---|---|
| SEQC/MAQC | 16 | 51 | 2.4 | 3.1 | 00:00:33 |
| GSE62852-AML | 40 | 252 | 16 | 10.8 | 00:02:17 |
| GTEx (part) | 1119 | 6100 | 312 | 42.2 | 00:08:58 |
| CCLE | 1019 | 8900 | 236 | 22.3 | 00:05:57 |

15 - 40 times smaller
(depends on dataset diversity)

# Query times
## (index of 1019 samples)

(hash table in memory, counts on disk)

| Query type | # Queries | Query time (sec) |
|---|---|---|
| 31-mers | 1000 | 1.0 |
| | 10000 | 2.0 |
| | 100000 | 16.0 |
| | 500000 | 89.0 |
| | 1000000 | 179.0 |
| full-length mRNAs | 1 | 0.3 |
| | 100 | 12.6 |
| | 1000 | 81 |

550 k-mer/sec

# Transipedia.org

Method | Open access | Published: 10 October 2024

## Transipedia.org: k-mer-based exploration of large RNA sequencing datasets and application to cancer data

Chloé Bessière, Haoliang Xue, Benoit Guibert, Anthony Boureux, Florence Rufflé, Julien Viot, Rayan Chikhi, Mikaël Salson, Camille Marchet, Thérèse Commes & Daniel Gautheret

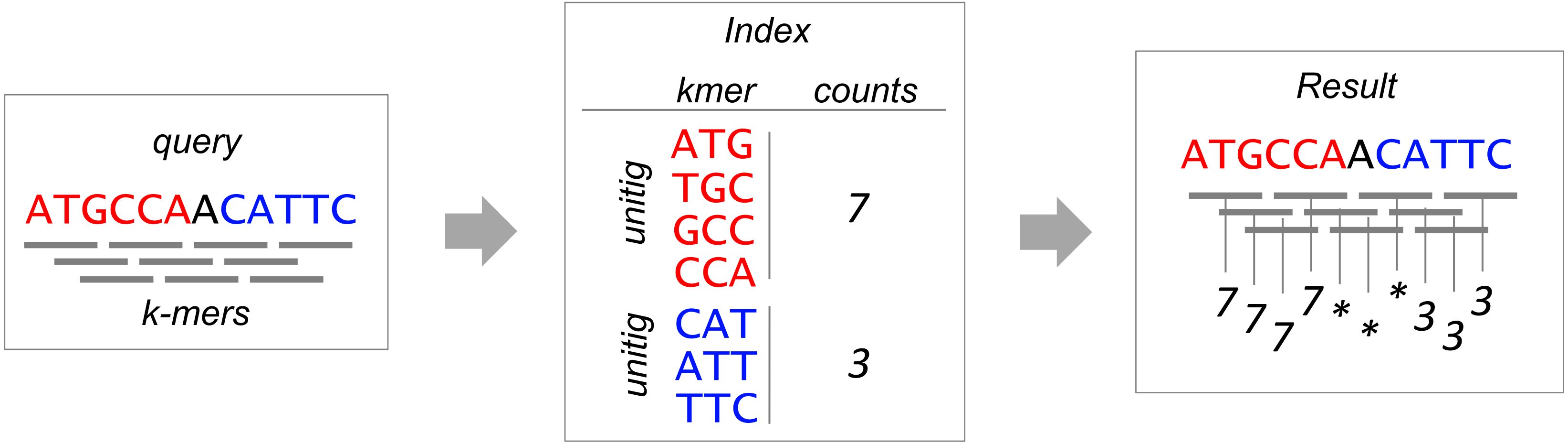
*Genome Biology* **25**, Article number: 266 (2024) | Cite this article

# Assessing Reindeer Accuracy

Quantify

- Gene expression
- Mutations
- Fusion RNAs
- Repeats
- Splice junctions
- ... Anything ...

# Reindeer query logic



*query*

ATGCCAACATTC

*k-mers*

*Index*

| | *kmer* | *counts* |
|---|---|---|
| *unitig* | ATG<br>TGC<br>GCC<br>CCA | *7* |
| *unitig* | CAT<br>ATT<br>TTC | *3* |

*Result*

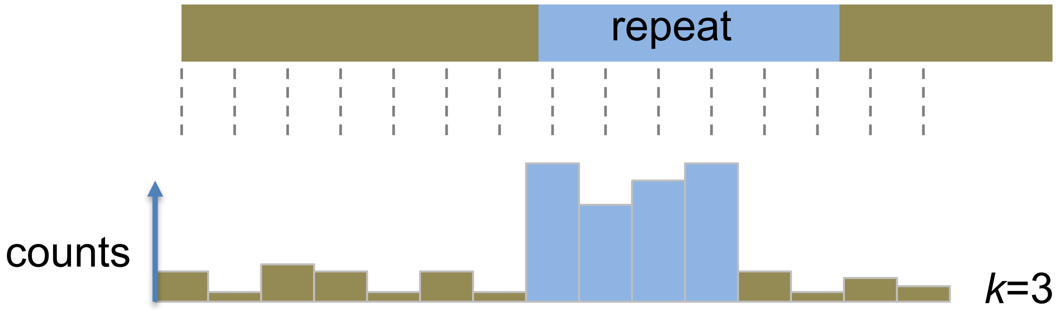
ATGCCAACATTC

*7 7 7 7 * * * 3 3 3*

- Different ways of processing counts
- Warning! some k-mers may belong to a repeat and have unrealistic counts

# Query design

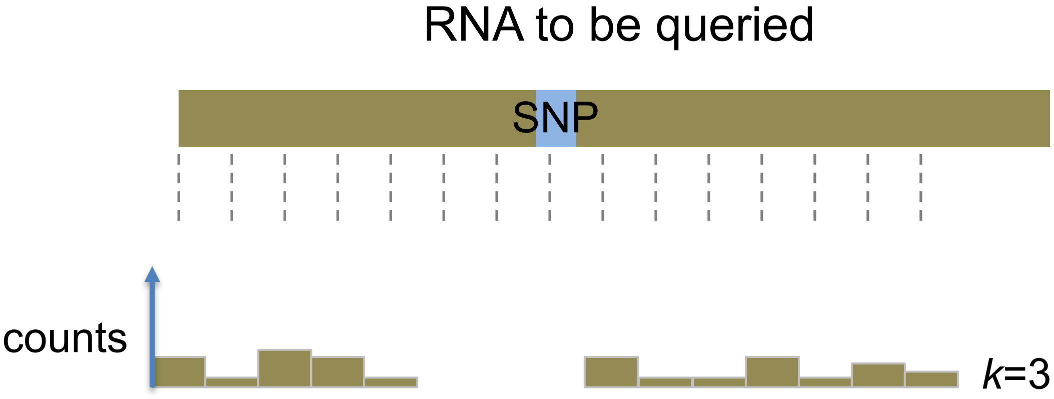
RNA to be queried



*k*-mers in repeat have higher counts

Repeats, low complexity regions, exon borders must be deleted from queries

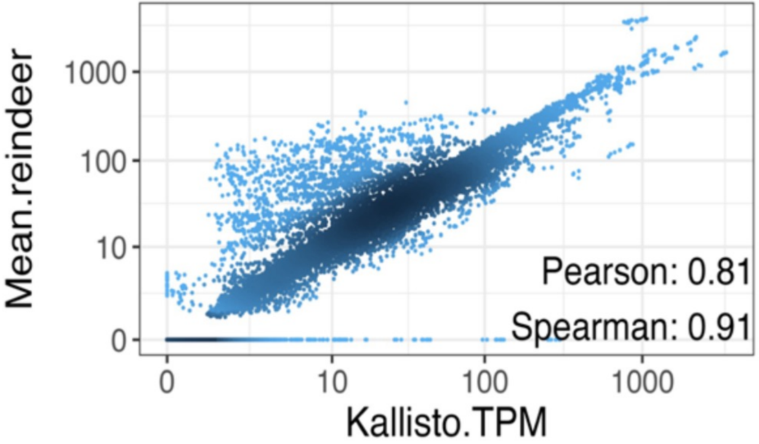
# Output design



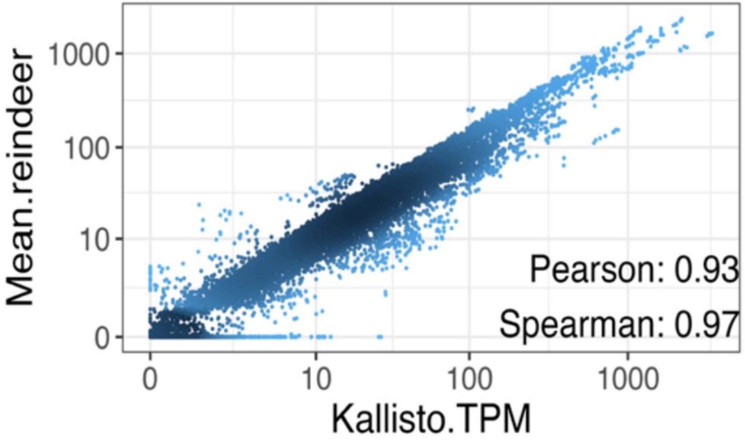
*k* *k*-mers have zero counts

There must be tolerance for missed *k*-mers in output

# Gene expression quantification



Raw



Multimap *k*-mers removed

# Accuracy of mutation and fusion detection

CCLE Dataset
(N=1019)

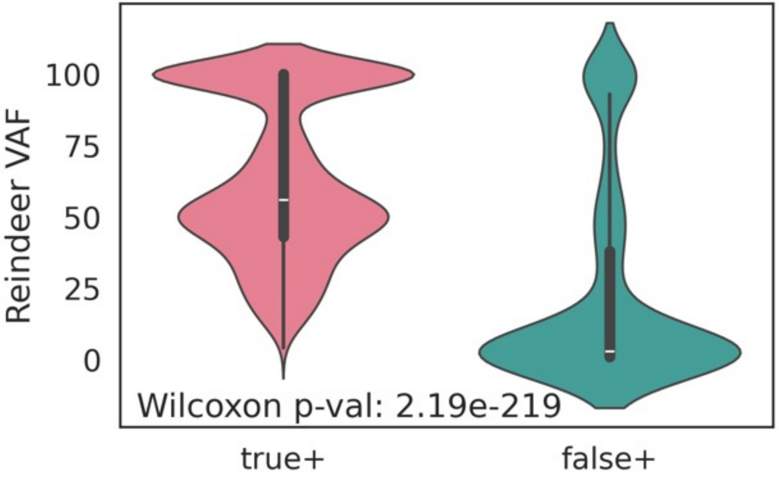
Ground truth=
Depmap calls

| | # probes | | #positive kmers >= 3 |
|---|---|---|---|
| Cosmic Hotspot mutations | 914 | True + | 1676 |
| | | False + | 255 |
| | | False - | 87 |
| | | Precision | 0.868 |
| | | Recall | 0.951 |
| Cosmic fusions | 59 | True + | 98 |
| | | False + | 3 |
| | | False - | 2 |
| | | Precision | 0.970 |
| | | Recall | 0.980 |

Conclusion: few misses despite SNPs

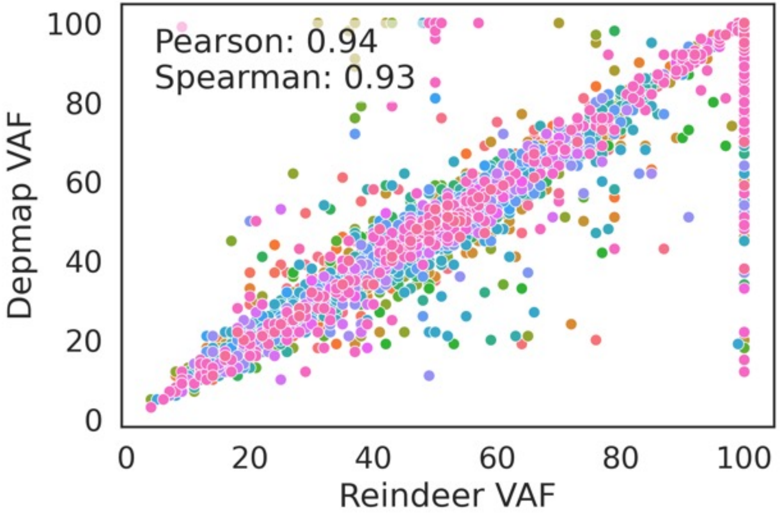
# Allele frequencies

All mutations



+ 70% FP found in WES

All mutations



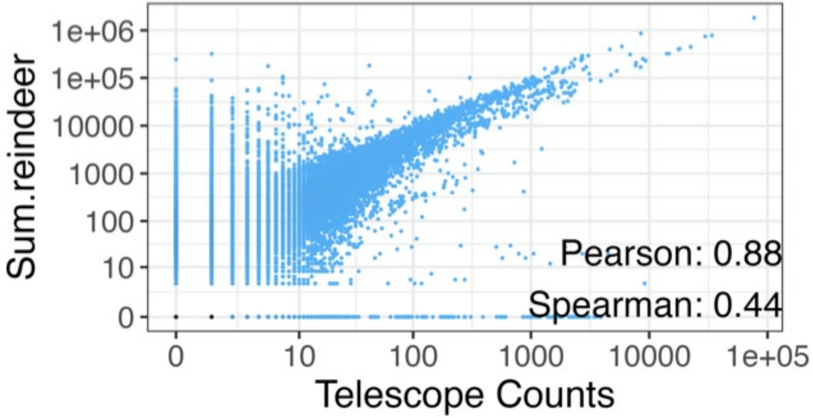
# Repeats

Locus-level

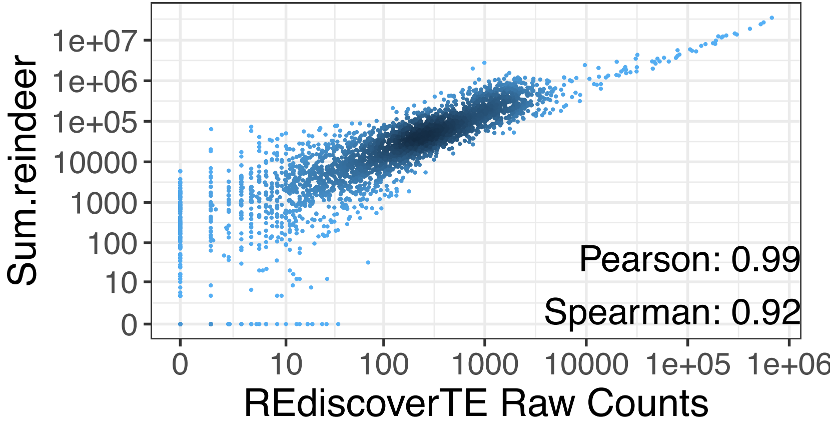
Comparison with two tools of reference

Family-level

1000 random ERV

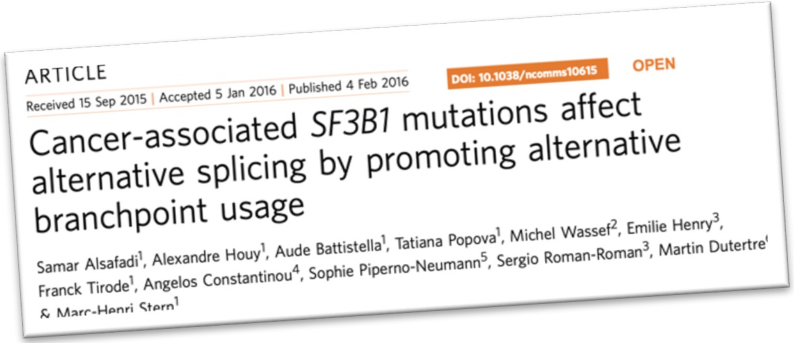


50 ERV families



# Two cancer applications

# Find new splice junction signatures



ARTICLE

Received 15 Sep 2015 | Accepted 5 Jan 2016 | Published 4 Feb 2016

DOI: 10.1038/ncomms10615

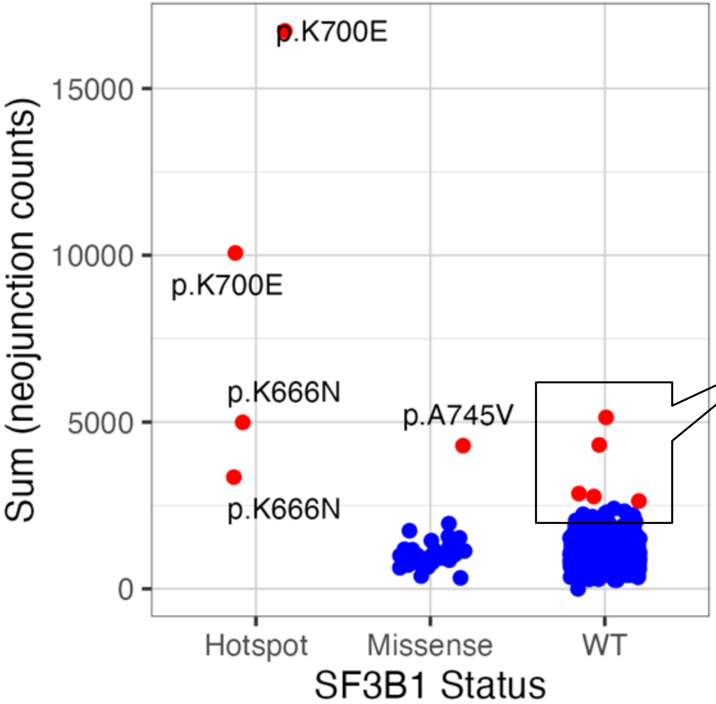
OPEN

Cancer-associated *SF3B1* mutations affect alternative splicing by promoting alternative branchpoint usage

Samar Alsafadi[1], Alexandre Houy[1], Aude Battistella[1], Tatiana Popova[1], Michel Wassef[2], Emilie Henry[3], Franck Tirode[1], Angelos Constantinou[4], Sophie Piperno-Neumann[5], Sergio Roman-Roman[3], Martin Dutertre
& Marc-Henri Stern[1]

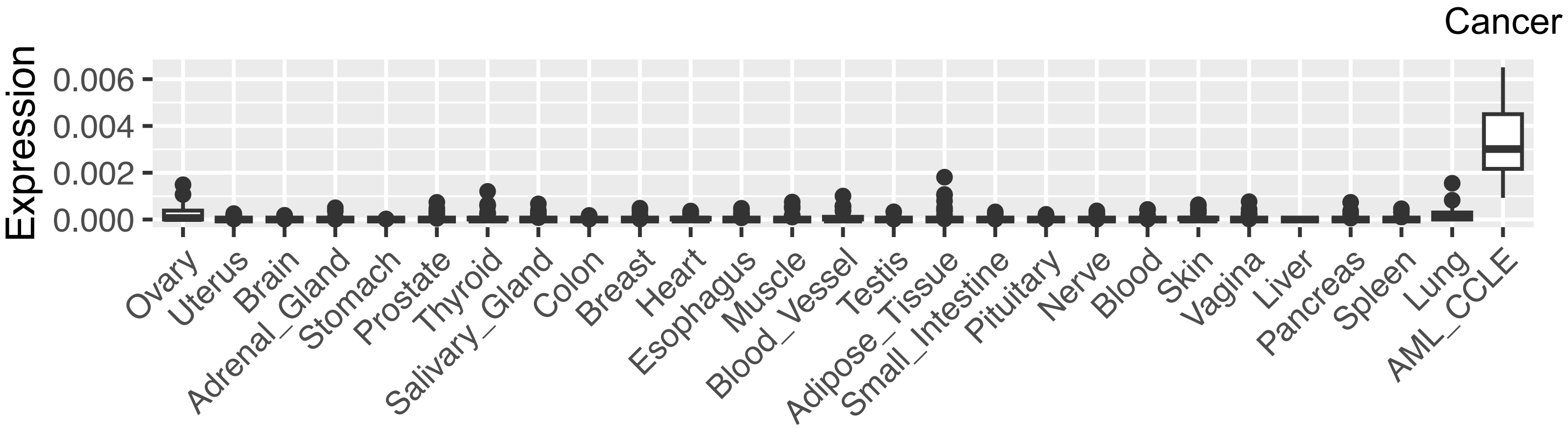
849 SF3B1-induced neojunctions

Quantify in 1019 CCLE cell lines



New genes with same splice defects

# Evaluate neoantigen candidates

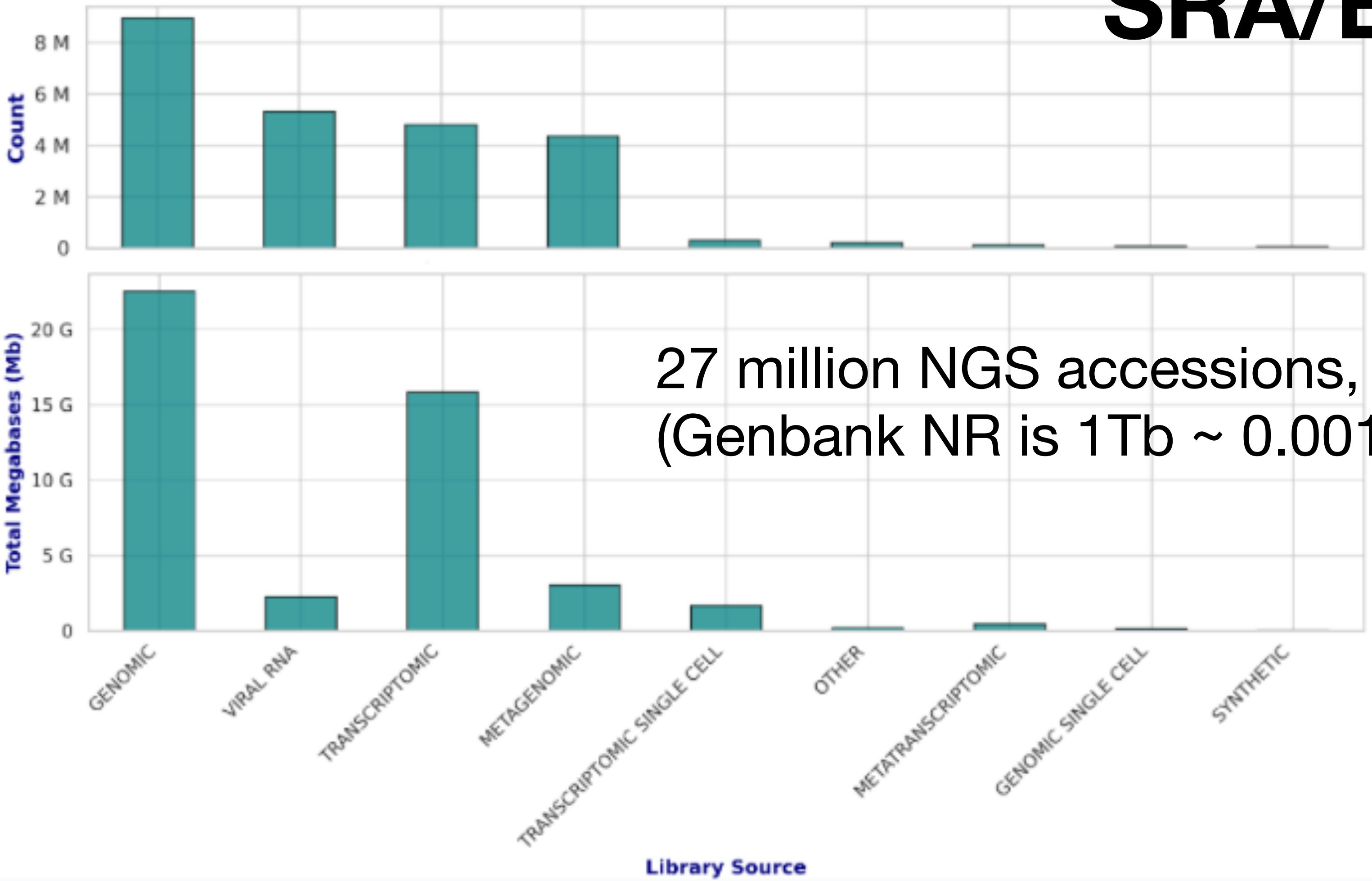


1280 normal tissues

# Next

- OK, 1000 transcriptome is nice
- But what about 1 million?

# SRA/ENA



27 million NGS accessions, ~30Pb
(Genbank NR is 1Tb ~ 0.001 Pb)

# Sequence Bloom Trees

Solomon & Kingsford, 2016

Bit vectors

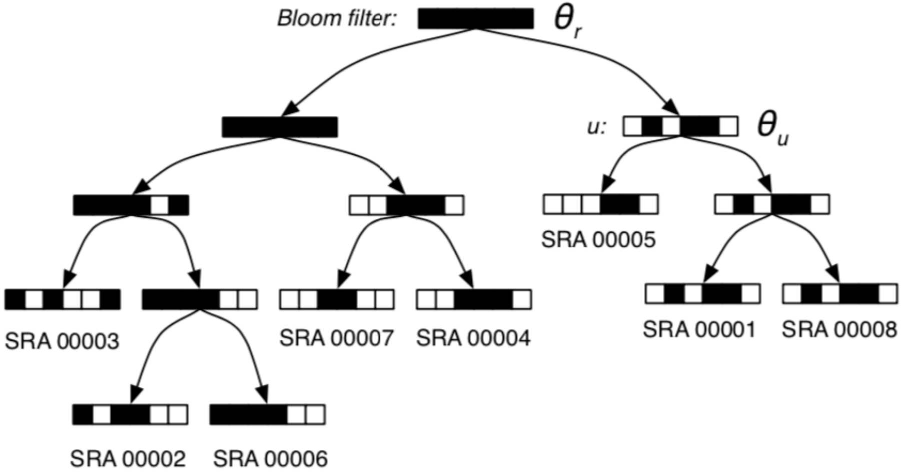
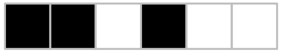
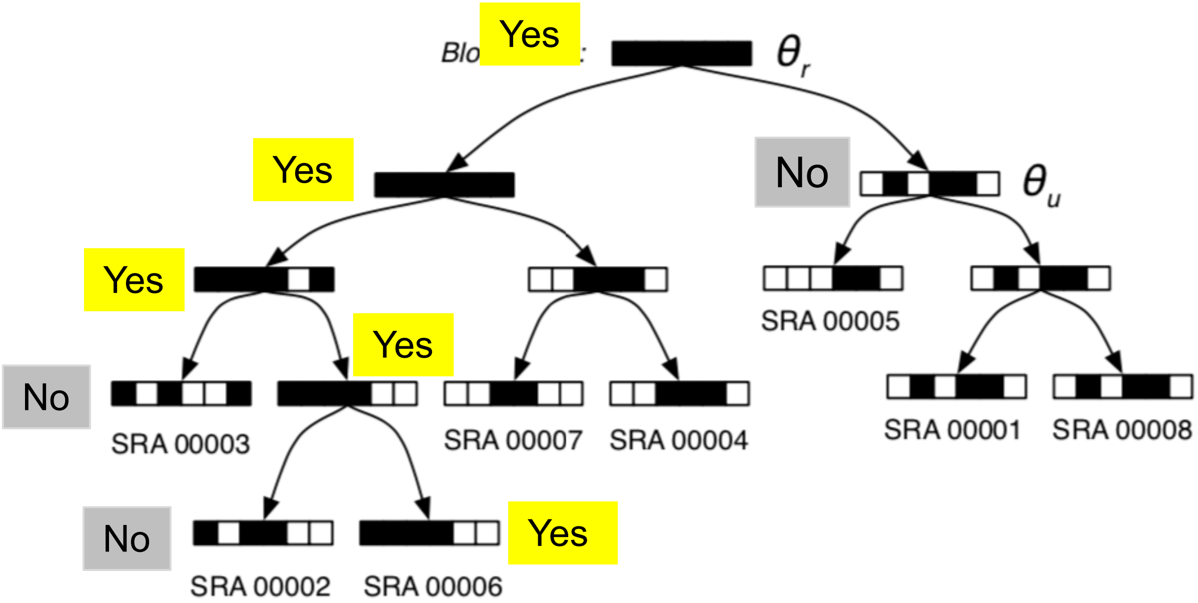


Figure 1: Schematic of a Sequence Bloom Tree. Each node contains a bloom filter containing the kmers present in the sequencing experiments under it.

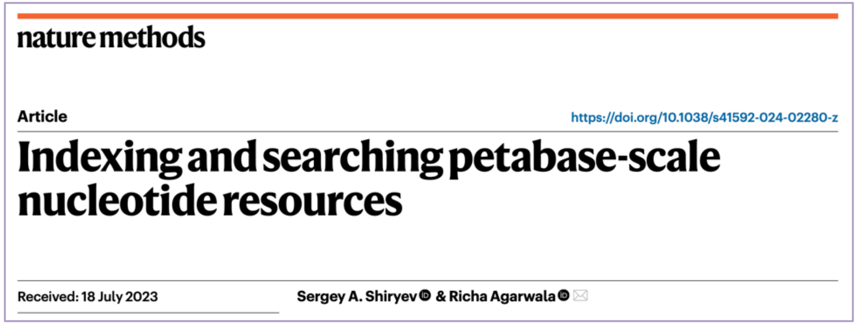
# Example

Sequence X hashes to: 1,2,4





# The future of SRA analysis



nature methods

Article

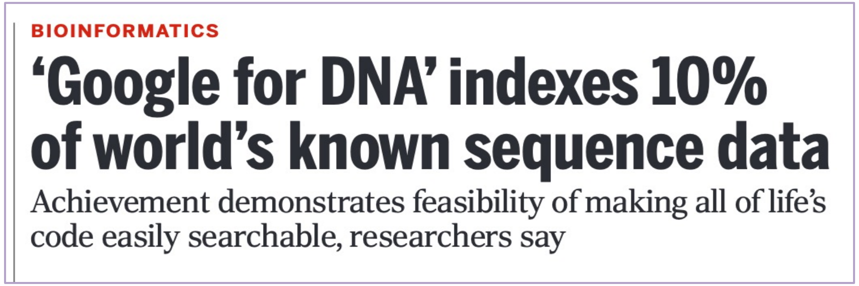
https://doi.org/10.1038/s41592-024-02280-z

Indexing and searching petabase-scale nucleotide resources

Received: 18 July 2023

Sergey A. Shiryev & Richa Agarwala

NCBI Pebblescout: Bloom trees. 3.7 PB. No counts



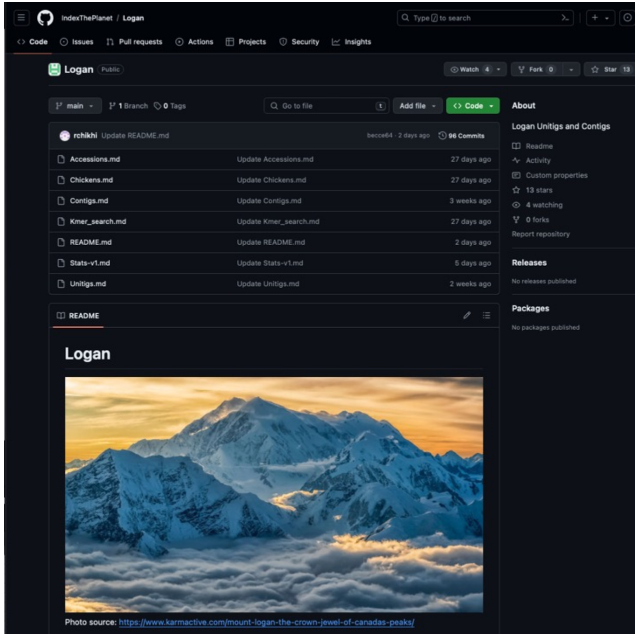
BIOINFORMATICS

‘Google for DNA’ indexes 10% of world’s known sequence data

Achievement demonstrates feasibility of making all of life’s code easily searchable, researchers say

ETHZ Metagraph. Colored « pruned » DBG. 5 PB. (10% of SRA) with counts

# The real future of SRA analysis



rchikhi Update README.md becce64 · 2 days ago 96 Commits

| | | |
|---|---|---|
| Accessions.md | Update Accessions.md | 27 days ago |
| Chickens.md | Update Chickens.md | 27 days ago |
| Contigs.md | Update Contigs.md | 3 weeks ago |
| Kmer_search.md | Update Kmer_search.md | 27 days ago |
| README.md | Update README.md | 2 days ago |
| Stats-v1.md | Update Stats-v1.md | 5 days ago |
| Unitigs.md | Update Unitigs.md | 2 weeks ago |

About

Logan Unitigs and Contigs

Readme<br>
Activity<br>
Custom properties<br>
13 stars<br>
4 watching<br>
0 forks<br>
Report repository

Releases

No releases published

Packages

No packages published

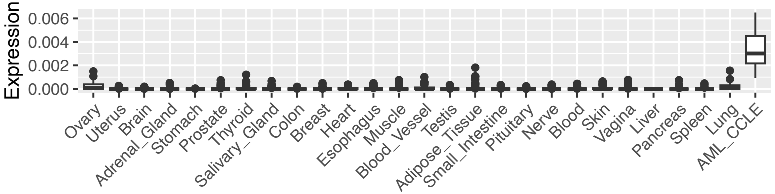
Logan

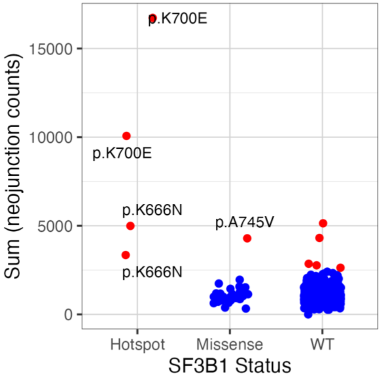
Photo source: https://www.karmactive.com/mount-logan-the-crown-jewel-of-canadas-peaks/

Rayan Chikhi *et al*.: Logan
Index of the full SRA (27M entries)
- 3500 CPU years
- 25PB > 2.1PB Unitigs

Pierre Peterlongo *et al*.: Logan-Search
(Bloom tree-based)
[https://logan-search.org/](https://logan-search.org/)

# Counts are essential for transcriptomics





# Towards 100k human RNA-seq **with counts**

- The ESCALATE project
  - A colored DBG + hash-based index
  - Compress counts
  - Public data

# Mining SRA metadata

- >1million human transcriptome records
- One record: 160 standard + optional fields

```
age
altitude
assembly_quality
assembly_software
base_count
binning_software
bio_material
bisulfite_protocol
broad_scale_environmental_context
broker_name
cage_protocol
cell_line
cell_type
center_name
checklist
chip_ab_provider
chip_protocol
chip_target
collected_by
completeness_score
contamination_score
control_experiment
country
cultivar
culture_collection
datahub
description
dev_stage
disease
dnase_protocol
ecotype
elevation
environment_biome
environment_feature
environment_material
environmental_medium
environmental_sample
experiment_accession
experiment_alias
experiment_target
experiment_title
experimental_factor
experimental_protocol
extraction_protocol
faang_library_selection
first_created
first_public
germline
hi_c_protocol
host
host_body_site
host_genotype
host_gravidity
host_growth_conditions
host_phenotype
host_scientific_name
host_sex
host_status
host_tax_id
identified_by
instrument_model
instrument_platform
investigation_type
isolate
isolation_source
last_updated
library_construction_protocol
library_gen_protocol
library_layout
library_max_fragment_size
library_min_fragment_size
library_name
library_pcr_isolation_protocol
library_prep_date
library_prep_date_format
library_prep_latitude
library_prep_location
library_prep_longitude
library_selection
library_source
library_strategy
local_environmental_context
location
location_end
location_start
marine_region
mating_type
ncbi_reporting_standard
nominal_length
nominal_sdev
pcr_isolation_protocol
project_name
protocol_label
read_count
read_strand
restriction_enzyme
restriction_enzyme_target_sequence
restriction_site
rna_integrity_num
rna_prep_3_protocol
rna_prep_5_protocol
rna_purity_230_ratio
rna_purity_280_ratio
rt_prep_protocol
run_accession
run_alias
salinity
sample_accession
sample_alias
sample_capture_status
sample_collection
sample_description
sample_material
sample_prep_interval
sample_prep_interval_units
sample_storage
sample_storage_processing
sample_title
sampling_campaign
sampling_platform
sampling_site
scientific_name
secondary_project
secondary_sample_accession
secondary_study_accession
sequencing_date
sequencing_date_format
sequencing_location
sequencing_longitude
sequencing_method
sequencing_primer_catalog
sequencing_primer_lot
sequencing_primer_provider
serotype
serovar
sex
specimen_voucher
status
strain
study_accession
study_alias
study_title
sub_species
sub_strain
submission_accession
submission_tool
submitted_host_sex
submitted_md5
submitted_read_type
tag
target_gene
tax_id
taxonomic_classification
taxonomic_identity_marker
temperature
tissue_lib
tissue_type
transposase_protocol
variety
```

age
altitude
assembly_quality
assembly_software
base_count
binning_software
bio_material
bisulfite_protocol
broad_scale_environmental_context
broker_name
cage_protocol
cell_line
cell_type
center_name
checklist
chip_ab_provider
chip_protocol
chip_target
collected_by
completeness_score
contamination_score
control_experiment
country
cultivar
culture_collection
datahub
description
dev_stage
disease
dnase_protocol
ecotype
elevation
environment_biome
environment_feature
environment_material
environmental_medium
environmental_sample
experiment_accession
experiment_alias
experiment_target

experiment_title
experimental_factor
experimental_protocol
extraction_protocol
faang_library_selection
first_created
first_public
germline
hi_c_protocol
host
host_body_site
host_genotype
host_gravidity
host_growth_conditions
host_phenotype
host_scientific
host_sex
host_s
h

y_construction_protocol
brary_gen_protocol
library_layout
library_max_fragment_size
library_min_fragment_size
library_name
library_pcr_isolation_protocol
library_prep_date
library_prep_date_format
library_prep_latitude
library_prep_location
library_prep_longitude
library_selection
library_source

library_strategy
local_environmental_context
location
location_end
location_start
marine_region
mating_type
ncbi_reporting_standa
nominal_length
nominal_sde
pcr_is
pr

get_sequence

y_num
p_3_protocol
a_prep_5_protocol
rna_purity_230_ratio
rna_purity_280_ratio
rt_prep_protocol
run_accession
run_alias
salinity
sample_accession
sample_alias
sample_capture_status
sample_collection
sample_description
sample_material
sample_prep_interval
sample_prep_interval_units
sample_storage
sample_storage_processing
sample_title
sampling_campaign
sampling_platform

sampling_site
scientific_name
secondary_project
secondary_sample_accession
secondary_study_accession
sequencing_date
sequencing_date_format
sequencing_location
sequencing_longitude
sequencing_method
sequencing_primer_catalog
sequencing_primer_lot
sequencing_primer_provider
serotype
serovar
sex
specimen_voucher
status
strain
study_accession
study_alias
study_title
sub_species
sub_strain
submission_accession
submission_tool
submitted_host_sex
submitted_md5
submitted_read_type
tag
target_gene
tax_id
taxonomic_classification
taxonomic_identity_marker
temperature
tissue_lib
tissue_type
transposase_protocol
variety

>80% empty

No controlled vocabulary

# Filling missing fields

- Unstructured information is there
  - Abstract
  - Methods
  - Optional sample info (« healthy sample »)
- Obvious approach: LLMs
  - Now testing Llama with various « bio » models

# Exploring large RNA-seq data: take home lessons

- Public NGS data is huge
  - Not only human RNA: just visit SRA/ENA and find out
- It has been locked to date
  - File retrieval is not practical
  - Mapping strategies do not scale and are not sustainable
- Unlocking is possible with special data structures
- Reuse of existing data has big energy-saving potential

# Research group on RNA Sequence, Structure and Function

## Group members on this project

Daniel Gautheret
Hugues Herrmann, PhD student
Fiona Hak, PhD student
Safa Maddouri, Bioinformatician

## Collaborators

**I2BC**: Melina Gallopin
**Gustave Roussy**: Caroline Robert
**U. Lille**: Camille Marchet, Mikaël Salson
**Pasteur**: Rayan Chikhi
**U. Montpellier**: Thérèse Commes
**Curie**: Antonin Morillon
**Iztech**: Bünyamin Akgül, Cansu Dürer

**Funding**
ANR,
INSERM Mathematics & Informatics for Cancer

https://github.com/Transipedia/