SEQOIA
Médecine génomique

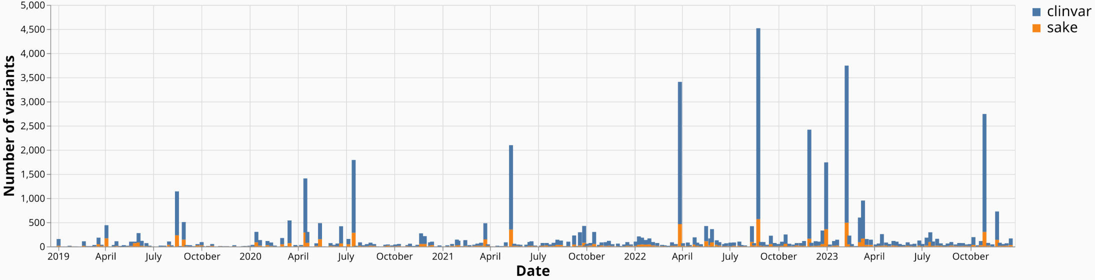
SAKE

How to fish variants in a data lake

Pierre Marijon, Sacha Schutz

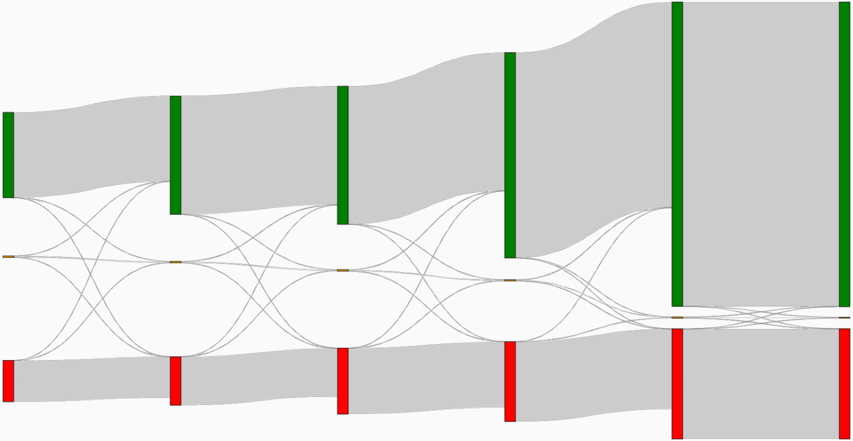
Janvier 9, 2024

Problem: Annotations change over time



ClinVar pathogenic variants found in Seqoia between each release

Problem: Annotations change over time



Clinvar variant classes change per year between 2018 to 2023

Objectif:

Automatic re-analysis the Seqoia data sets

A Big data Challenge

	germinal	somatic	total
#samples	27,615	5,256	32,871
vcf (Gb)	5,888	709	6,189
#unique variants	363,080,825	340,371,185	603,691,147 [1]
#genotypes	144,766,772,625	11,529,461,678	156,296,234,303

[1] ≈ 15 % of commun variants

A Big data Challenge

	germinal	somatic	total
#samples	27,615	5,256	32,871
vcf (Gb)	5,888	709	6,189
#unique variants	363,080,825	340,371,185	603,691,147 [1]
#genotypes	144,766,772,625	11,529,461,678	156,296,234,303
#GnomAD	-	-	759,302,267
#Clinvar	-	-	2,337,929

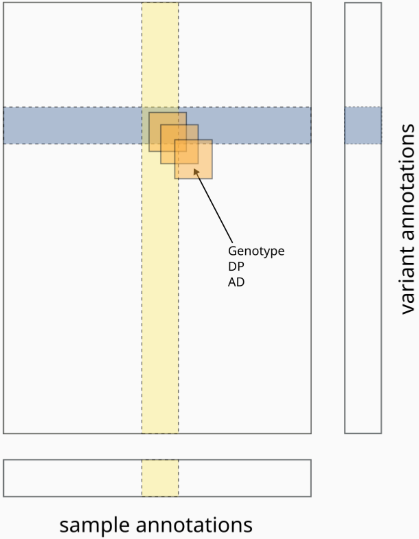
[1] ≈ 15 % of commun variants

A Big data Challenge

	germinal	somatic	total
#samples	27,615	5,256	32,871
vcf (Gb)	5,888	709	6,189
#unique variants	363,080,825	340,371,185	603,691,147 [1]
#genotypes	144,766,772,625	11,529,461,678	156,296,234,303
#GnomAD	-	-	759,302,267
#Clinvar	-	-	2,337,929
#PanelApp genes	-	-	6,031
#Omim genes	-	-	18,138

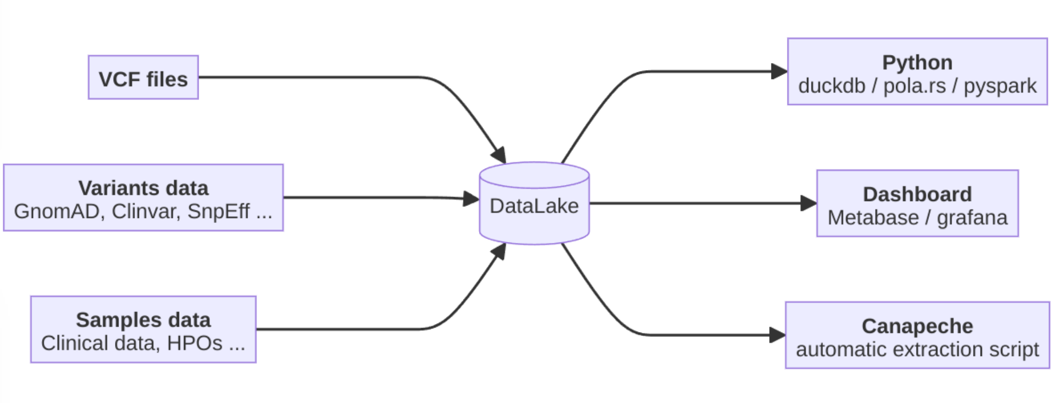
[1] ≈ 15 % of commun variants

Need for a sparse matrix to store everything



Only 0.67 % of all possible positions are fill

Solution: A data lake



VCF files

Variants data
GnomAD, Clinvar, SnpEff ...

Samples data
Clinical data, HPOs ...

DataLake

Python
duckdb / pola.rs / pyspark

Dashboard
Metabase / grafana

Canapeche
automatic extraction script

Advantages over traditional databases

The Seqoia datalake is a collection of **parquet files**:

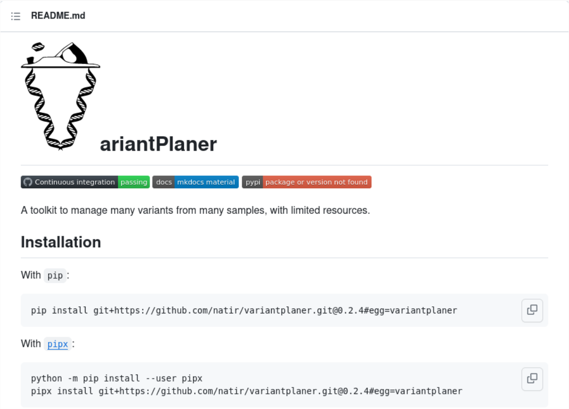
- **Compressed** files
- Column oriented for **fast** analytical process
- **SQL** queryable
- Usable with **Python** or R

Variantplaner

a python package for ingesting VCFs files in a data lake

Python command line tool and library

- Convert vcf into parquet files
- Built for sake but generalizable to other uses
- Open source
- https://github.com/natir/variantplaner


README.md
Continuous integration passing
docs mkdocs material
pypi package or version not found

ariantPlaner

A toolkit to manage many variants from many samples, with limited resources.

Installation

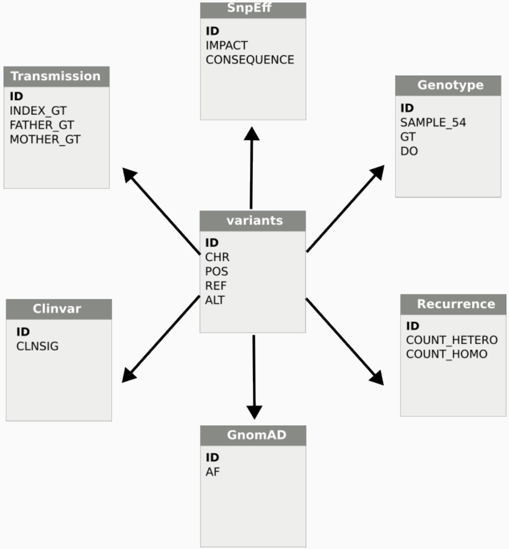
With `pip`:

```
pip install git+https://github.com/natir/variantplaner.git@0.2.4#egg=variantplaner
```

With `pipx`:

```
python -m pip install --user pipx
pipx install git+https://github.com/natir/variantplaner.git@0.2.4#egg=variantplaner
```

`variantplaner`: Star schema



SnpEff
ID
IMPACT
CONSEQUENCE

Transmission
ID
INDEX_GT
FATHER_GT
MOTHER_GT

Genotype
ID
SAMPLE_54
GT
DO

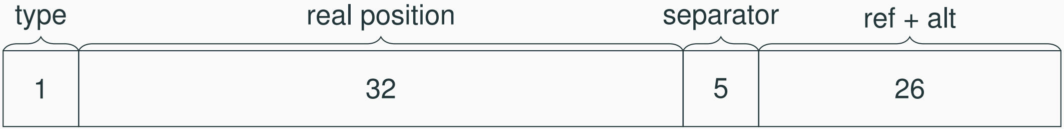
variants
ID
CHR
POS
REF
ALT

Clinvar
ID
CLNSIG

Recurrence
ID
COUNT_HETERO
COUNT_HOMO

GnomAD
ID
AF

variantplaner: variant id



type	real position	separator	ref + alt
1	32	5	26

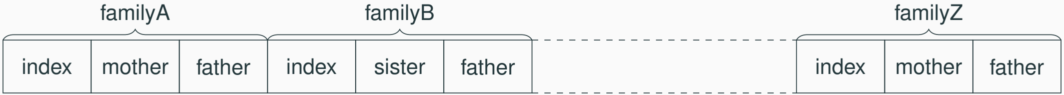
variantplaner: variant id

type	real position	separator	ref + alt
1	32	5	26

type	hash(real position + ref + alt)
1	63

variantplaner: genotype partition

what variants does a patient carry?



familyA

index mother father

familyB

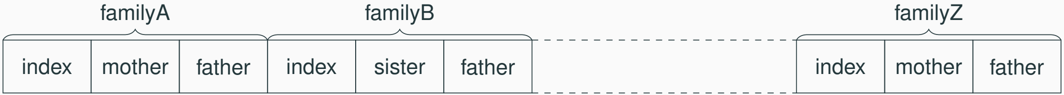
index sister father

familyZ

index mother father

variantplaner: genotype partition

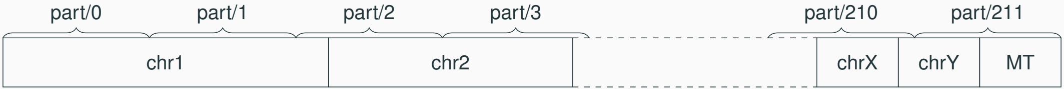
what variants does a patient carry?



familyA familyB familyZ

index mother father index sister father index mother father

which patients carry one variant?



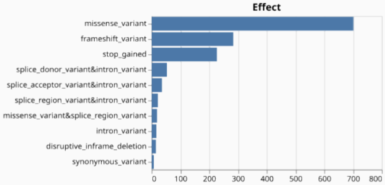
part/0 part/1 part/2 part/3 part/210 part/211

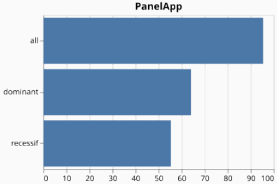
chr1 chr2 chrX chrY MT

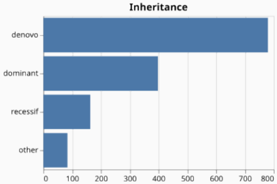
Canapeche

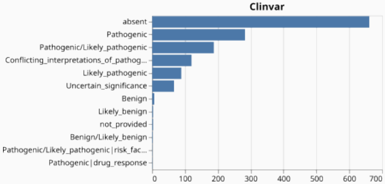
Extract pathogenic variants from the lake

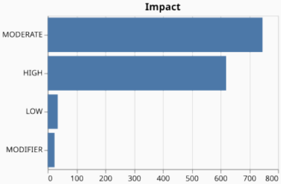
Canapeche: **Characteristic of pathogenic variants**

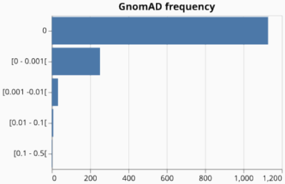






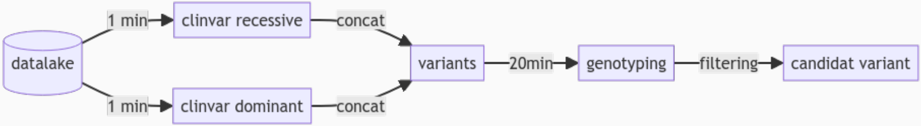






Total: 1374 conclusive pathogen variants

Canapeche: Extraction pipeline



datalake

1 min

clinvar recessive

1 min

clinvar dominant

concat

concat

variants

20min

genotyping

filtering

candidat variant

Compute on:

- Intel Xenon 2.5 GHz × 40
- 190 Gb of ram
- Disk access time: 10 Gbytes

Canapeche: Dominant strategy

Select all heterozygous pathogenic **clinvar** variants inherited from one parent with **GnomAD allele count < 10** present in **dominant genes** according to **panelapp**.

`Canapeche`: Dominant strategy

```
SELECT DISTINCT v.id FROM '{VARIANTS}' v
JOIN '{CLINVAR}' c  ON c.id = v.id
JOIN '{SNPEFF}' s ON s.id = v.id
LEFT JOIN '{GNOMAD}' g ON g.id = v.id
JOIN '{PANNELAPP}' p ON p.gene_symbol = s.gene
WHERE c.CLNSIG LIKE '%patho%'
AND (g.AC[1] < 10 OR g.AC[1] IS NULL)
AND p.inheritance LIKE 'MONOALLELIC'
```

Canapeche: Dominant strategy

Select all homozygous pathogenic **clinvar** variants inherited from both parents with **GnomAD nhomalt < 10** present in **recessives genes** according to **panelapp**.

2876 variants

Canapeche: Results for Clinvar strategies

	Canapeche	Human	Common	Recall	Precision
variant recessive	199	160	79	49.7%	39.7%
variant dominant	2876	390	100	25%	3.4%
variant denovo	758	760	309	40.7%	40.8%

Canapeche: Results for Clinvar strategies

False Positive

- Switch clinvar
- Missing diagnosis
- Mismatched phenotype +++

False negative

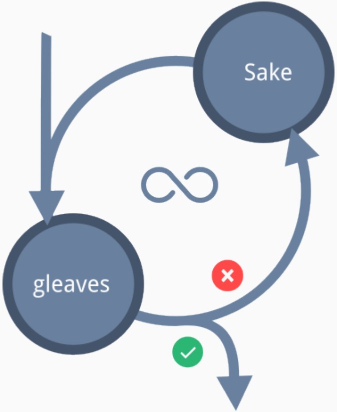
- Switch clinvar
- Other strategies
- Absent from clinvar +++

Canapeche: Results for Clinvar strategies

	Project count
variant recessive	220
variant dominant	3672
variant denovo	838

Conclusion

- Creating a new strategy
- Improve strategy to reach 90% of recall
- Import CNV data into the lake
- Use Large Model Language (LLM) to use clinical data



Example of LLM usage