



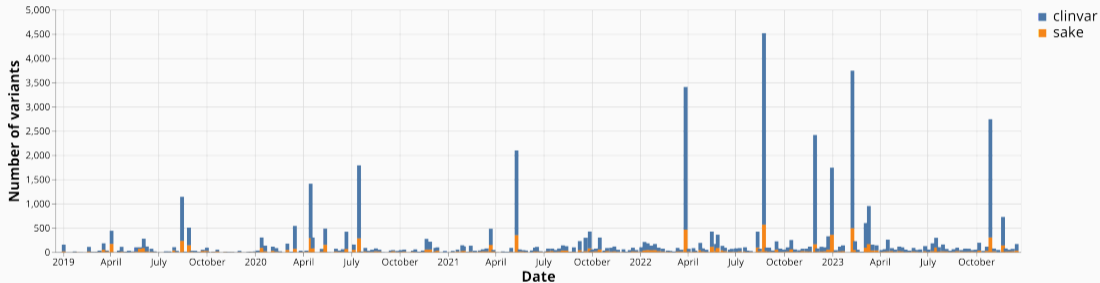
SAKE

How to fish variants in a data lake

Pierre Marijon, Sacha Schutz

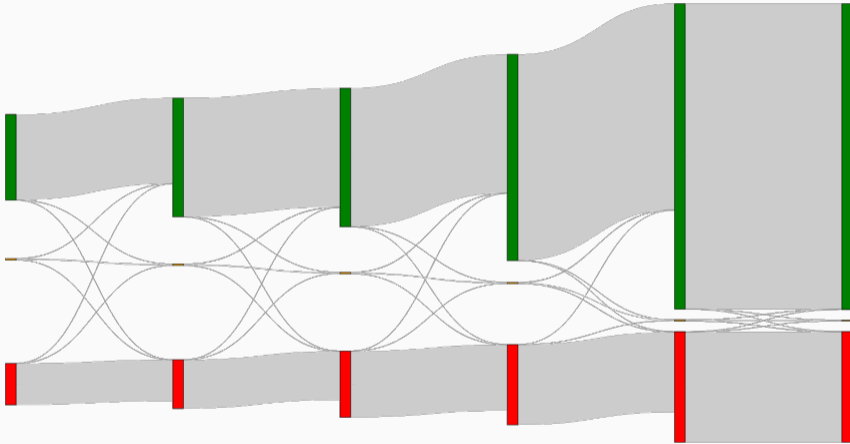
Janvier 9, 2024

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 Sequoia data sets

	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

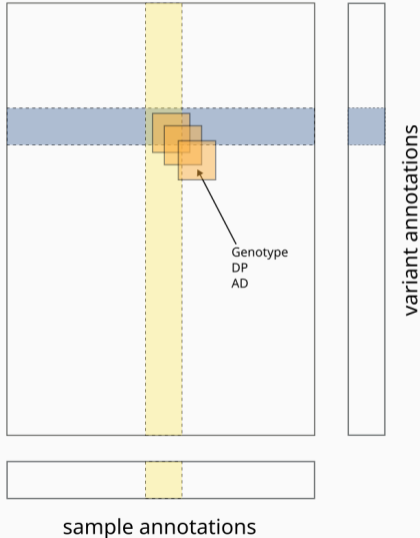
	germinal	somatic	total
#samples	27,615	5,256	32,871
vcf (Gb)	5,888	709	6,189
#unique variants	363,080,825	340,371,185	603,691,147 ¹
#genotypes	144,766,772,625	11,529,461,678	156,296,234,303
#GnomAD	-	-	759,302,267
#Clinvar	-	-	2,337,929

¹ 15 % of commun variants

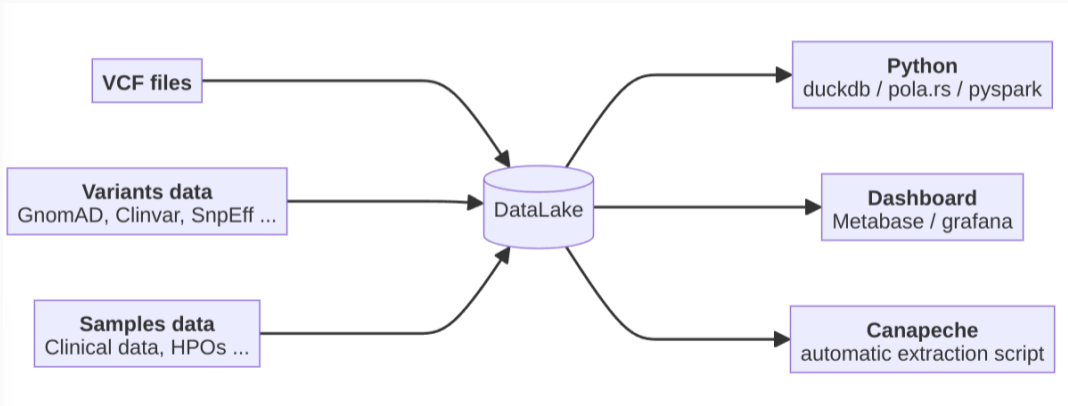
	germinal	somatic	total
#samples	27,615	5,256	32,871
vcf (Gb)	5,888	709	6,189
#unique variants	363,080,825	340,371,185	603,691,147 ¹
#genotypes	144,766,772,625	11,529,461,678	156,296,234,303
#GnomAD	-	-	759,302,267
#Clinvar	-	-	2,337,929
#PanelApp genes	-	-	6,031
#Omim genes	-	-	18,138

¹ 15 % of commun variants

Need for a sparse matrix to store everything



Only 0.67 % of all possible positions are fill



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

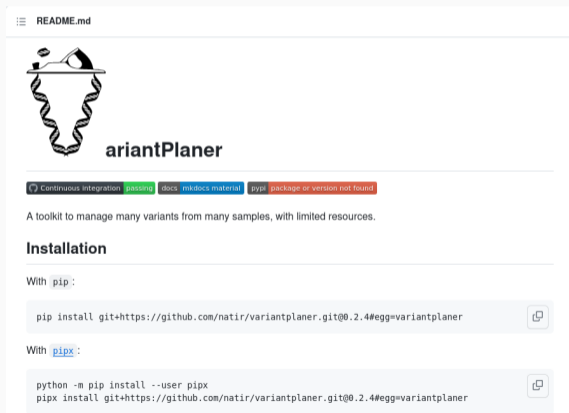
- **Compressed files**
- Column oriented for **fast** analytical process
- **SQL** queryable
- Usable with **Python** or **R**

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/variantplanner>



☰ README.md



ariantPlanner

🔄 Continuous integration passing docs mktdocs material pypl package or version not found

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

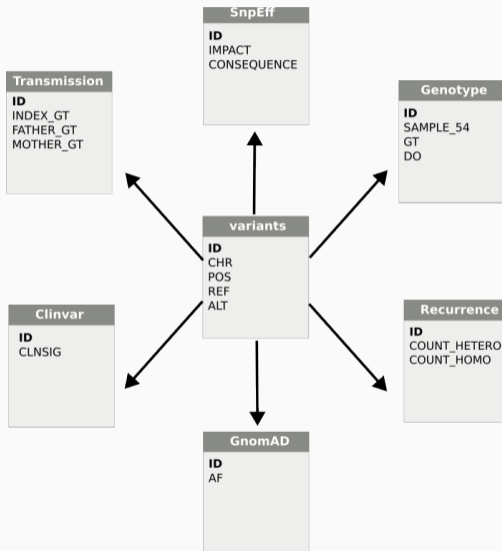
Installation

With [pip](#):

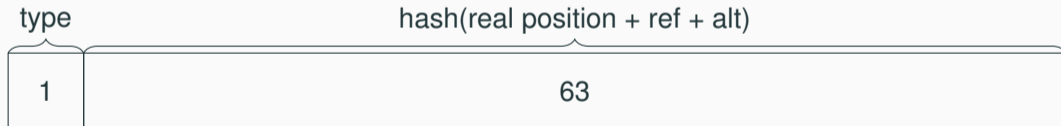
```
pip install git+https://github.com/natir/variantplanner.git@0.2.4#egg=variantplanner
```

With [pipx](#):

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







what variants does a patient carry?



what variants does a patient carry?

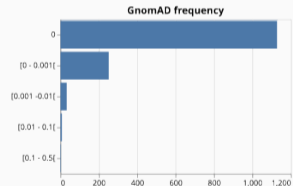
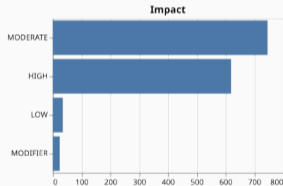
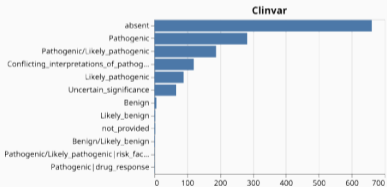
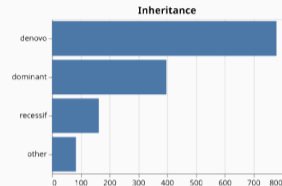
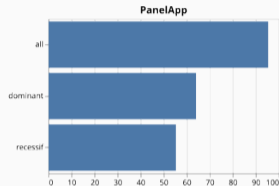
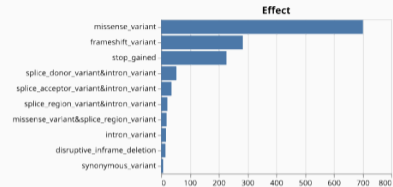


which patients carry one variant?

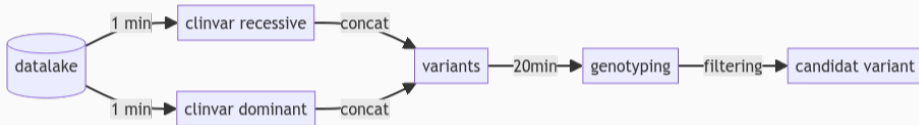


Canapeche

Extract pathogenic variants from the lake



Total: 1374 conclusive pathogen variants



Compute on:

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

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

```
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'
```

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

2876 variants

	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%

False Positive

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

False negative

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

	Project count
variant recessive	220
variant dominant	3672
variant denovo	838

- 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

