SEQOIA
Médecine génomique

SAKE: Seqoia dAta laKe

What to fish in the lake ?

Pierre Marijon
Laura Do Souto Ferreira

April 11, 2025

GCS SeqOIA

A Big Data Challenge

	Germline	Somatic	Total
#Sample	46,839	10,522	58,027
#unique variants	492,284,372	619,758,827	963,515,536
#genotypes	246,931,520,478	23,317,666,538	
#sample with CNV	46,082	3,298	49,380
#CNV	1,445,298,049	2,421,232	1,447,719,281

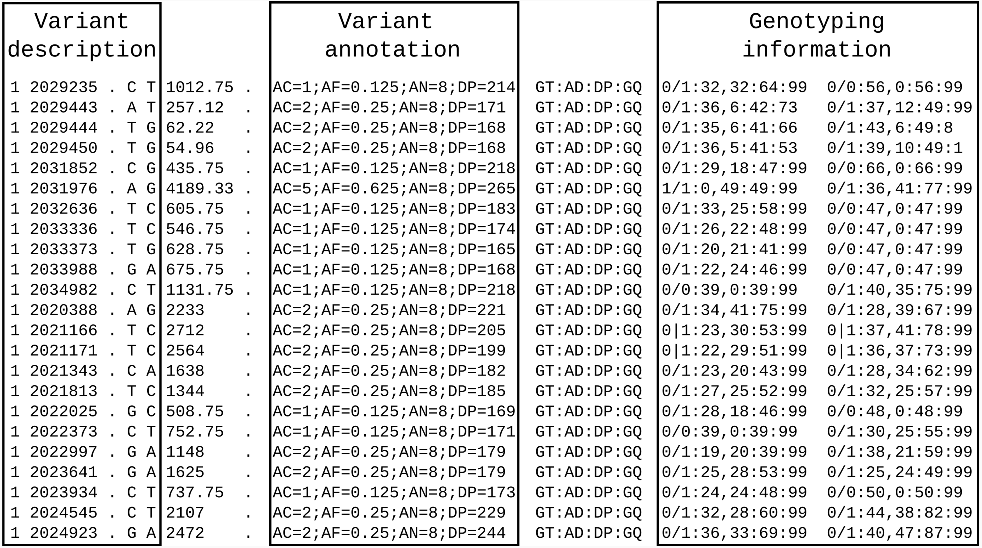
A Big Data Challenge

	Germline	Somatic	Total
#Sample	46,839	10,522	58,027
#unique variants	492,284,372	619,758,827	963,515,536
#genotypes	246,931,520,478	23,317,666,538	
#sample with CNV	46,082	3,298	49,380
#CNV	1,445,298,049	2,421,232	1,447,719,281
SNV size (Tib)	≈488.68	≈84.61	≈573.29
CNV size (Tib)	≈5.97	≈0.15	≈6.13
sake size (Tib)	≈4.21	≈0.34	≈4.81

Update to 03/25

How to organize and request variants ?

What’s in VCF ?



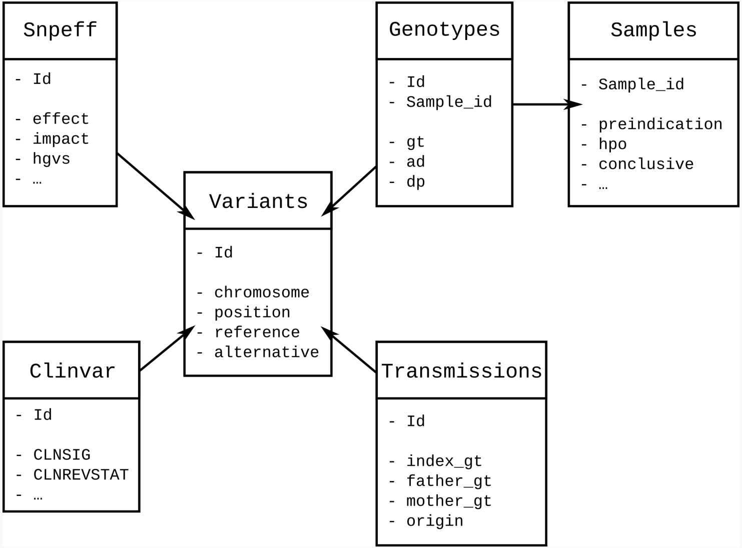
Variant description

Variant annotation

Genotyping information

```
1 2029235 . C T 1012.75 . AC=1;AF=0.125;AN=8;DP=214  GT:AD:DP:GQ 0/1:32,32:64:99  0/0:56,0:56:99
1 2029443 . A T 257.12  . AC=2;AF=0.25;AN=8;DP=171   GT:AD:DP:GQ 0/1:36,6:42:73   0/1:37,12:49:99
1 2029444 . T G 62.22   . AC=2;AF=0.25;AN=8;DP=168   GT:AD:DP:GQ 0/1:35,6:41:66   0/1:43,6:49:8
1 2029450 . T G 54.96   . AC=2;AF=0.25;AN=8;DP=168   GT:AD:DP:GQ 0/1:36,5:41:53   0/1:39,10:49:1
1 2031852 . C G 435.75  . AC=1;AF=0.125;AN=8;DP=218  GT:AD:DP:GQ 0/1:29,18:47:99  0/0:66,0:66:99
1 2031976 . A G 4189.33 . AC=5;AF=0.625;AN=8;DP=265  GT:AD:DP:GQ 1/1:0,49:49:99   0/1:36,41:77:99
1 2032636 . T C 605.75  . AC=1;AF=0.125;AN=8;DP=183  GT:AD:DP:GQ 0/1:33,25:58:99  0/0:47,0:47:99
1 2033336 . T C 546.75  . AC=1;AF=0.125;AN=8;DP=174  GT:AD:DP:GQ 0/1:26,22:48:99  0/0:47,0:47:99
1 2033373 . T G 628.75  . AC=1;AF=0.125;AN=8;DP=165  GT:AD:DP:GQ 0/1:20,21:41:99  0/0:47,0:47:99
1 2033988 . G A 675.75  . AC=1;AF=0.125;AN=8;DP=168  GT:AD:DP:GQ 0/1:22,24:46:99  0/0:47,0:47:99
1 2034982 . C T 1131.75 . AC=1;AF=0.125;AN=8;DP=218  GT:AD:DP:GQ 0/0:39,0:39:99   0/1:40,35:75:99
1 2020388 . A G 2233    . AC=2;AF=0.25;AN=8;DP=221   GT:AD:DP:GQ 0/1:34,41:75:99  0/1:28,39:67:99
1 2021166 . T C 2712    . AC=2;AF=0.25;AN=8;DP=205   GT:AD:DP:GQ 0|1:23,30:53:99  0|1:37,41:78:99
1 2021171 . T C 2564    . AC=2;AF=0.25;AN=8;DP=199   GT:AD:DP:GQ 0|1:22,29:51:99  0|1:36,37:73:99
1 2021343 . C A 1638    . AC=2;AF=0.25;AN=8;DP=182   GT:AD:DP:GQ 0/1:23,20:43:99  0/1:28,34:62:99
1 2021813 . T C 1344    . AC=2;AF=0.25;AN=8;DP=185   GT:AD:DP:GQ 0/1:27,25:52:99  0/1:32,25:57:99
1 2022025 . G C 508.75  . AC=1;AF=0.125;AN=8;DP=169  GT:AD:DP:GQ 0/1:28,18:46:99  0/0:48,0:48:99
1 2022373 . C T 752.75  . AC=1;AF=0.125;AN=8;DP=171  GT:AD:DP:GQ 0/0:39,0:39:99   0/1:30,25:55:99
1 2022997 . G A 1148    . AC=2;AF=0.25;AN=8;DP=179   GT:AD:DP:GQ 0/1:19,20:39:99  0/1:38,21:59:99
1 2023641 . G A 1625    . AC=2;AF=0.25;AN=8;DP=179   GT:AD:DP:GQ 0/1:25,28:53:99  0/1:25,24:49:99
1 2023934 . C T 737.75  . AC=1;AF=0.125;AN=8;DP=173  GT:AD:DP:GQ 0/1:24,24:48:99  0/0:50,0:50:99
1 2024545 . C T 2107    . AC=2;AF=0.25;AN=8;DP=229   GT:AD:DP:GQ 0/1:32,28:60:99  0/1:44,38:82:99
1 2024923 . G A 2472    . AC=2;AF=0.25;AN=8;DP=244   GT:AD:DP:GQ 0/1:36,33:69:99  0/1:40,47:87:99
```

Sake struct



Snpeff

- Id

- effect
- impact
- hgvs
- …

Genotypes

- Id
- Sample_id

- gt
- ad
- dp

Samples

- Sample_id

- preindication
- hpo
- conclusive
- …

Variants

- Id

- chromosome
- position
- reference
- alternative

Clinvar

- Id

- CLNSIG
- CLNREVSTAT
- …

Transmissions

- Id

- index_gt
- father_gt
- mother_gt
- origin

Variant Id a 64 bitfields

type	position in concat chromosome	$\mathrm{len}(\mathit{ref})$	$\mathrm{nuc2bit}(\mathit{alt})$
1	$a = \mathrm{ceil}(\log_2(\mathit{genome\_length}))$	b	$c = \mathrm{len}(\mathit{alt}) \times 2$

$$b = 63 - a - c$$

Variant Id a 64 bitfields

$$\overbrace{\boxed{1}}^{\text{type}}\overbrace{\boxed{a = \text{ceil}(\log_2(genome\_length))}}^{\text{position in concat chromosome}}\overbrace{\boxed{b}}^{\text{len}(ref)}\overbrace{\boxed{c = \text{len}(alt) \times 2}}^{\text{nuc2bit}(alt)}$$

$$b = 63 - a - c$$

$$if \ \text{ceil}(\log_2(genome\_length) + \text{len}(ref) + \text{len}(alt) \times 2 > 63 :$$

Variant Id a 64 bitfields

type	position in concat chromosome	$\text{len}(ref)$	$\text{nuc2bit}(alt)$
1	$a = \text{ceil}(\log_2(genome\_length))$	b	$c = \text{len}(alt) \times 2$

$$b = 63 - a - c$$

$$if\ \text{ceil}(\log_2(genome\_length)) + \text{len}(ref) + \text{len}(alt) \times 2 > 63 :$$

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

Variant Id: collision risk

collision risk for k element in N bucket: $1 - \frac{2N!}{(2^{kN}(2^N - k)!)} \approx 1 - \exp\left(-\frac{k^2}{2^{N+1}}\right) \approx \frac{k^2}{2 \cdot N}$

	Germline	Somatic
#short variant	477,322,058 (96.96%)	612,273,467 (98.79%)
#long variants	14,962,314 (3.04%)	7,485,360 (1.20%)
collision risk	$1.21 \cdot 10^{-5}$	$3.04 \cdot 10^{-6}$

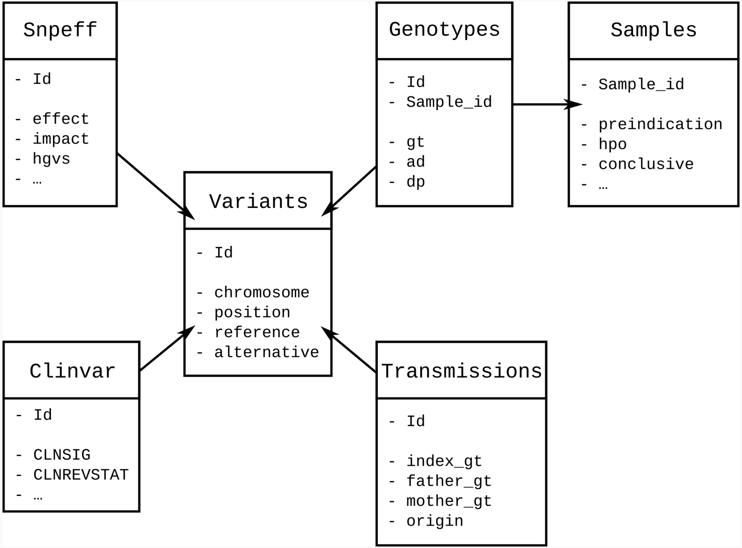
Variant Id: collision risk

collision risk for k element in N bucket: $1 - \frac{2N!}{(2^{kN}(2^N - k)!)} \approx 1 - \exp\left(-\frac{k^2}{2^{N+1}}\right) \approx \frac{k^2}{2 \cdot N}$

	Germline	Somatic
#short variant	477,322,058 (96.96%)	612,273,467 (98.79%)
#long variants	14,962,314 (3.04%)	7,485,360 (1.20%)
collision risk	$1.21 \cdot 10^{-5}$	$3.04 \cdot 10^{-6}$
#star variants	6,716,896 (1.36%)	1,944,446 (0.31%)
collision risk	$3.68 \cdot 10^{-6}$	1.6610^{-6}

```
#CHROM  POS       ID  REF         ALT  GT
10      41905990  .   CAATTAATGGA C    0/1
10      41905993  .   T           *    0/1
10      41905993  .   T           G    0/1
```

File organisation



Snpeff

- Id
- effect
- impact
- hgvs
- …

Genotypes

- Id
- Sample_id
- gt
- ad
- dp

Samples

- Sample_id
- preindication
- hpo
- conclusive
- …

Variants

- Id
- chromosome
- position
- reference
- alternative

Clinvar

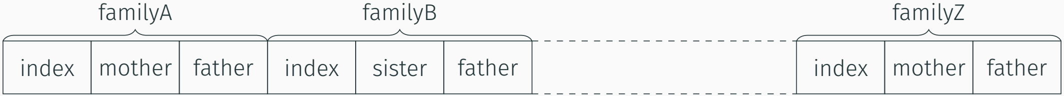
- Id
- CLNSIG
- CLNREVSTAT
- …

Transmissions

- Id
- index_gt
- father_gt
- mother_gt
- origin

Genotypes organisation

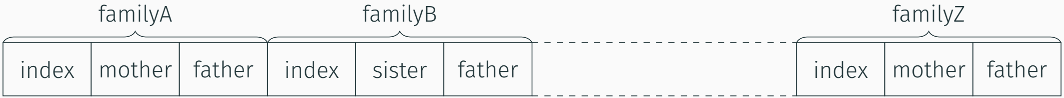
What variants does a patient carry?



familyA			familyB			familyZ		
index	mother	father	index	sister	father	index	mother	father

Genotypes organisation

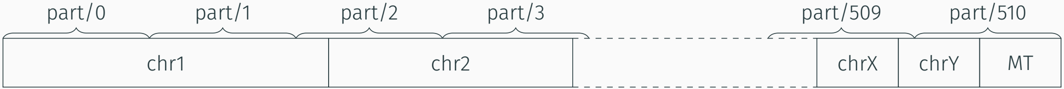
What variants does a patient carry?



familyA familyB familyZ

index mother father index sister father index mother father

Which patients carry one variant?

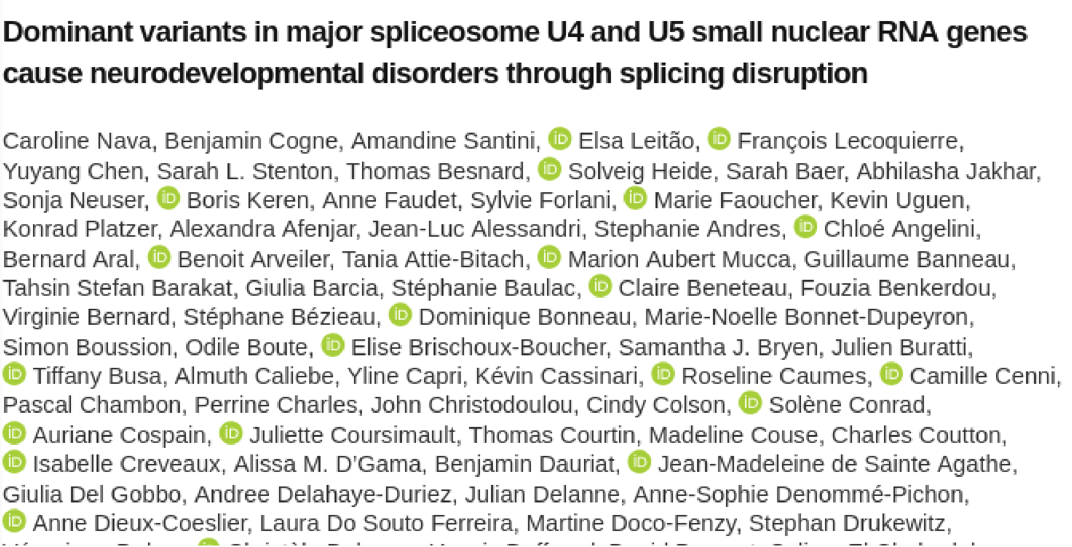


part/0 part/1 part/2 part/3 part/509 part/510

chr1 chr2 chrX chrY MT

Discovery of new patient

We found a new disease-gene link



Dominant variants in major spliceosome U4 and U5 small nuclear RNA genes cause neurodevelopmental disorders through splicing disruption

Caroline Nava, Benjamin Cogne, Amandine Santini, Elsa Leitão, François Lecoquierre, Yuyang Chen, Sarah L. Stenton, Thomas Besnard, Solveig Heide, Sarah Baer, Abhilasha Jakhar, Sonja Neuser, Boris Keren, Anne Faudet, Sylvie Forlani, Marie Faoucher, Kevin Uguen, Konrad Platzer, Alexandra Afenjar, Jean-Luc Alessandri, Stephanie Andres, Chloé Angelini, Bernard Aral, Benoit Arveiler, Tania Attie-Bitach, Marion Aubert Mucca, Guillaume Banneau, Tahsin Stefan Barakat, Giulia Barcia, Stéphanie Baulac, Claire Beneteau, Fouzia Benkerdou, Virginie Bernard, Stéphane Bézieau, Dominique Bonneau, Marie-Noelle Bonnet-Dupeyron, Simon Boussion, Odile Boute, Elise Brischoux-Boucher, Samantha J. Bryen, Julien Buratti, Tiffany Busa, Almuth Caliebe, Yline Capri, Kévin Cassinari, Roseline Caumes, Camille Cenni, Pascal Chambon, Perrine Charles, John Christodoulou, Cindy Colson, Solène Conrad, Auriane Cospain, Juliette Coursimault, Thomas Courtin, Madeline Couse, Charles Coutton, Isabelle Creveaux, Alissa M. D'Gama, Benjamin Dauriat, Jean-Madeleine de Sainte Agathe, Giulia Del Gobbo, Andree Delahaye-Duriez, Julian Delanne, Anne-Sophie Denommé-Pichon, Anne Dieux-Coeslier, Laura Do Souto Ferreira, Martine Doco-Fenzy, Stephan Drukewitz,

Found variant

```
sake_db = sake_request.Sake()

all_variants = list()
for (chrom, start, end) in regions:
    all_variants.append(sake_db.get_interval(
        chrom, start, end
    ))

variants = concat(all_variants)
```

run time	20s
#rows	9,506

Annotations

```
annotated = sake_db.add_annotation(variants, clinvar)
annotated = sake_db.add_annotation(annotated, gnomad)
annotated = sake_db.add_annotation(variants, snpeff)
```

	clinvar	gnomad	snpeff
run time	6s	115s	104s
#rows	9,506	9,506	51,351

Genotyping

```
annotated = sake_db.add_id_part(annotated)

all_genotyped = list()
for id_part, group in annotated.group_by("id_part"):
    part_genotype = read_genotype(id_part)
    all_genotyped.append(group.join(part_genotype))

genotyped = concat(all_genotyped)
```

	one part	all part
run time	11s	24 min
#rows		2,925,453

Recurrence

```
recurrence = genotyped.group_by("id").aggregate(
    sake_AC = polars.col("gt").sum()
)

all_data = genotyped.join(recurrence)
```

run time	11s
#rows	2,925,453

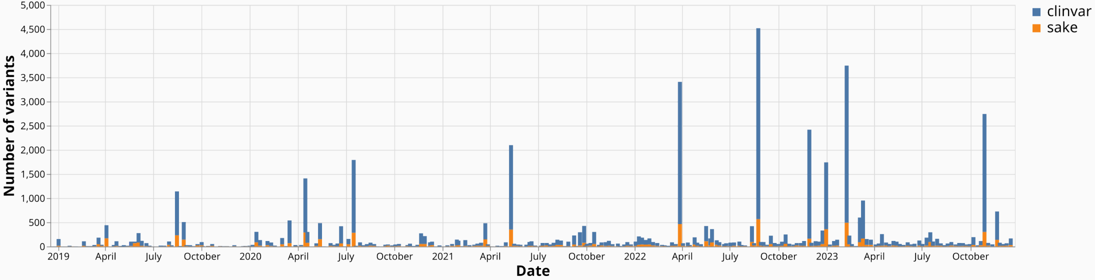
Finialize

```
sample_info = sake_db.add_sample_info(all_data)
homozygote = sample_info.filtre(
    gt == 2 && affected == True
)
heterozygote = sample_info.filtre(
    gt == 1 && affected == True
)
```

	add sample info	homozygote	heterozygote
run time	2.7s	0.3s	0.3s
#rows	2,925,453	382,737	772,645

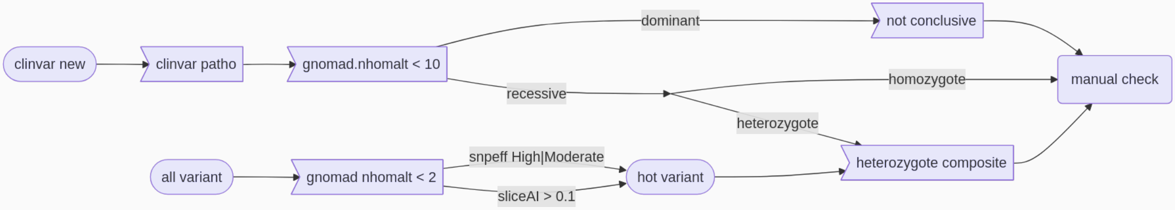
Continuous Reanalysis

We need continuous reanalysis ?



ClinVar pathogenic variants found in Seqoia between each release

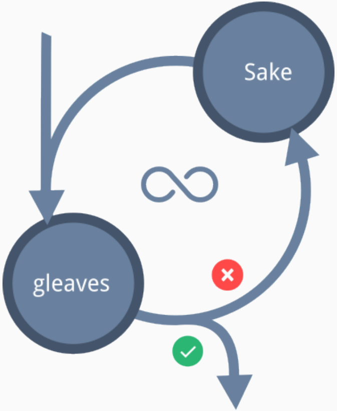
Strategy


clinvar new
clinvar patho
gnomad.nhomalt < 10
dominant
not conclusive
recessive
homozygote
heterozygote
manual check
all variant
gnomad nhomalt < 2
snpeff High|Moderate
sliceAI > 0.1
hot variant
heterozygote composite

Result

	2019/07/15	2023/05/08	2024/01/07	runtime
#sample	16,006	9,482	19,957	
dominant	67	7	14	15 min
recessive homozygote	785	68	154	40 min
recessive heterozygote				(too long)

Generalisation

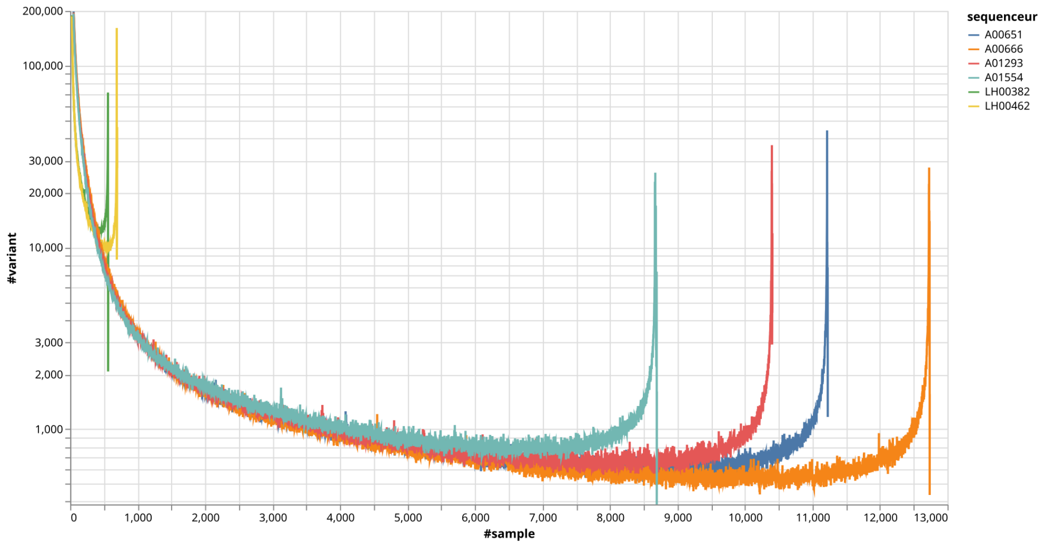


Management of systemic sequencing errors

Sequencers Bias

```
for samples_group in samples.group_by(sequencers):
    for part in 0..512
        compute_recurrence(samples_group, part)
```

Sequencers Bias



Sequencers Bias

```
xplus.AF > 0.3 & xplus.AF > sixk.AF
xplus.AF > 0.3 & xplus.AF > gnomad.AF
```

	sixk	gnomad
#variants	32,312,609	33,796,472
#snpeff HIGH	4,255 (0.013%)	5,176 (0.015%)
#snpeff MODERATE	46,358 (0.14%)	52,473 (0.15%)
#clinvar Patho	145 (0.0004%)	164 (0.0004%)
#clinvar Patho*	3210 (0.009%)	3975 (0.011%)

Utilisation du datalake en RNASeq

Conclusion

Conclusion

Build ”SAKE ”: https://github.com/SeqOIA-IT/variantplaner

Interogate SAKE: https://github.com/SeqOIA-IT/sake_request

Open for PR, bugs, suggestion, etc...

[1] Diagnostic Use Only

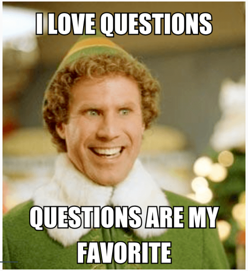
Conclusion

Build "**SAKE** ": https://github.com/SeqOIA-IT/variantplaner

Interogate **SAKE**: https://github.com/SeqOIA-IT/sake_request

Open for PR, bugs, suggestion, etc...

For any DUO[1] request contact: sake@bioinfo.aphp.fr



[1] Diagnostic Use Only