



# Genomics in the Cloud

Book Club - Week 15

March 8, 2021


# Agenda

- Chapter 14: Making a Fully Reproducible Paper
- Additional resources
- Open discussion



# Chapter 14: Making a Fully Reproducible Paper

*Genomics in the Cloud* by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.



Our guest  
speaker

Dr. Matthieu  
Miossec

---



# Bringing it all together with a case study

Case study devised as a basis for two GATK/Firecloud (now Terra) workshops. It has since been included in several other workshops in 2019.

ASHG 2018, San Diego, USA



**BROAD INSTITUTE** Create a reproducible paper with FireCloud

T. Miller<sup>1</sup>, K. Noblett<sup>1</sup>, I. Rosenberg<sup>1</sup>, M. J. Miossec<sup>2</sup>, G. A. Van der Auwera<sup>1</sup>  
<sup>1</sup>Data Sciences Platform (DSP), Broad Institute, Cambridge, MA, USA  
<sup>2</sup>Universidad Andrés Bello, Center for Bioinformatics and Integrative Biology, Santiago, Chile

**Abstract**  
The lack of portability and reproducibility of analysis methods limits the effectiveness with which biomedical researchers can benefit from the democratization of genomic analysis. FireCloud is an open-source, freely accessible cloud-based analysis platform developed at the Broad Institute that empowers developers and consumers of analysis methods to overcome these challenges. It bundles data storage, workflow management and interactive Jupyter Notebooks into secure yet readily shareable workspaces. We demonstrated how this system can be used to reproduce someone else's analysis and make your own research more reproducible.

**Reproducing someone else's research: A case study**

**Objective**  
We set out to reproduce the work described by Matthieu Miossec and collaborators in a bioRxiv preprint titled "Deleterious genetic variants in NOTCH1 are a major contributor to the incidence of non-syndromic Tetralogy of Fallot" (ToF). The authors analyzed high-throughput exome sequence data from 867 cases and 1252 controls, identifying 49 deleterious variants within the NOTCH1 gene that appeared associated with this congenital heart disease. Others had previously identified NOTCH1 in families with congenital heart defects, including ToF; however the work by Miossec et al. is the first to scale variant analysis of ToF to nearly a thousand case samples and show that NOTCH1 is a significant contributor to ToF risk.  
**Preprint URL:** <https://www.biorxiv.org/content/early/2018/04/13/300905>

**Overall approach**  
We used the information provided in the preprint and its Supplemental Materials to reconstruct the main phases of the work, distinguishing **Data Input**, **Processing** and **Analysis** as recommended by Kitzes et al. (see section below). For the **Processing** phase, we created a synthetic dataset to get around the lack of appropriate public data to use as input, and applied a variant discovery workflow that we judged equivalent. For the **Analysis** phase, we obtained the original scripts and commands from Dr. Miossec and with his assistance, reimplemented them in two parts: the prediction of variant effects as a workflow in WDL (Workflow Description Language) and the clustering analysis as R code in a Jupyter notebook. We did all the work in the Broad Institute's open-source analysis platform, **FireCloud**.

**DATA INPUTS** **PROCESSING** **ANALYSIS** **SHARING**

ISCB-LA SOIBIO 2018, Viña del Mar, Chile



# Not a hands-on Chapter, but a very instructive one

## Time and Cost

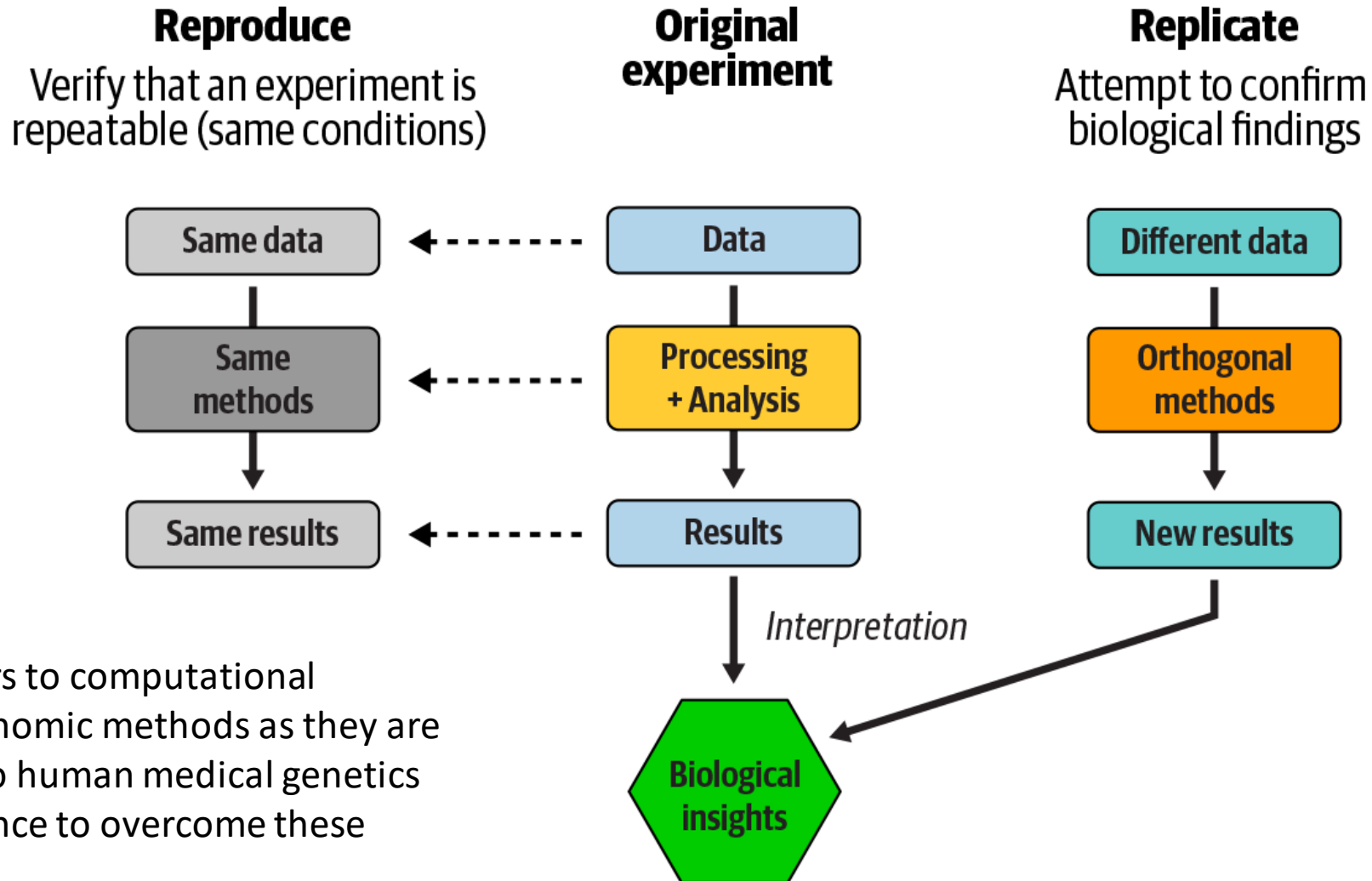
Workflow Name	1 file (range)	100 files	Time to Run 1 file	Time to Run 100 files
1_Collect-1000G-participants	\$1.64 to \$2.90	\$193.75	4.5 hours	12 hours
2_Generate-synthetic-reads	\$2.40 to \$3.44	\$405.67	4.5 hours	12 hours
3_Mutate-reads-with-BAMSurgeon	\$0.02 to \$0.15	\$5.72	.5 hours	2.5 hours
4_Call-single-sample-GVCF-GATK4	\$0.32 to \$0.52	\$39.82	1.75 hours	2.75 hours
5_Joint-call-and-hard-filter		\$10.09		4 hours
P6_redict-variant-effects-GEMINI		\$1.00		4 minutes

**Note that the size range** for a single 1000 genome vcf file is 669.5 MB to 843.5 MB. The total size of the 100-sample cohort is 72.12 GB.





# Reproducibility of analysis vs replicability of study findings



“Evaluate the barriers to computational reproducibility of genomic methods as they are commonly applied to human medical genetics and teach the audience to overcome these barriers.”

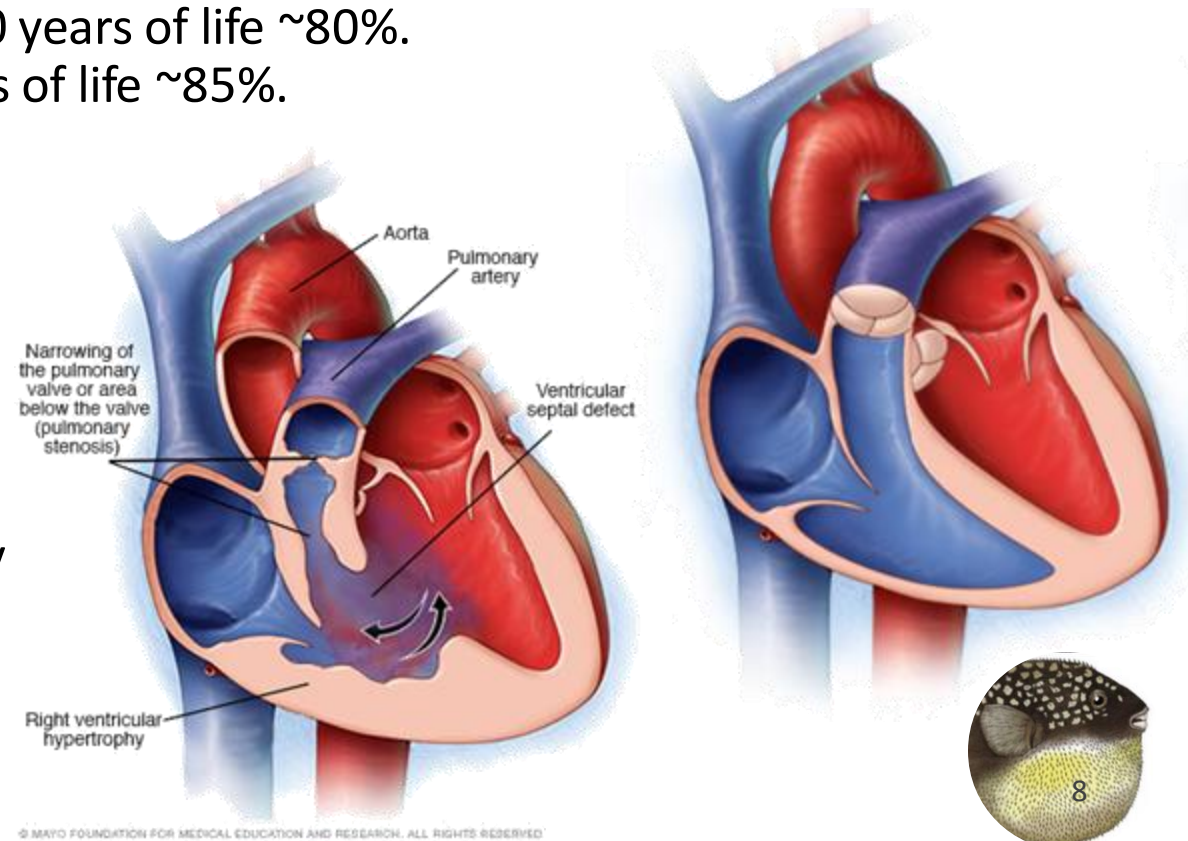


## Whole Exome Sequencing Reveals the Major Genetic Contributors to Nonsyndromic Tetralogy of Fallot

Donna J. Page, Matthieu J. Miossec, Simon G. Williams, Richard M. Monaghan, Elisavet Fotiou, Heather J. Cordell, Louise Sutcliffe, Ana Topf, Mathieu Bourgey, Guillaume Bourque, Robert Eveleigh, Sally L. Dunwoodie, David S. Winlaw, Shoumo Bhattacharya, Jeroen Breckpot, Koenraad Devriendt, Marc Gewillig, J. David Brook, Kerry J. Setchfield, Frances A. Bu'Lock, John O'Sullivan, Graham Stuart, Connie R. Bezzina, Barbara J.M. Mulder, Alex V. Postma, James R. Bentham, Martin Baron, Sanjeev S. Bhaskar, Graeme C. Black, William G. Newman, Kathryn E. Hentges, G. Mark Lathrop, Mauro Santibanez-Koref, Bernard D. Keavney **Show less Authors** ^

# What is the paper about?

- Non-syndromic Tetralogy of Fallot
  - Affects ~3 out of 10 000 births.
    - Without surgery, mortality rate in first 10 years of life ~80%.  
With surgery, survival in the first 30 years of life ~85%.
  - Patients with Tetralogy of Fallot typically present 4 cardiac malformations:
    - A large ventricular septal defect
    - Displacement of the aorta over the septal defect
    - Narrowing of the pulmonary valve
    - Progressive right ventricular hypertrophy



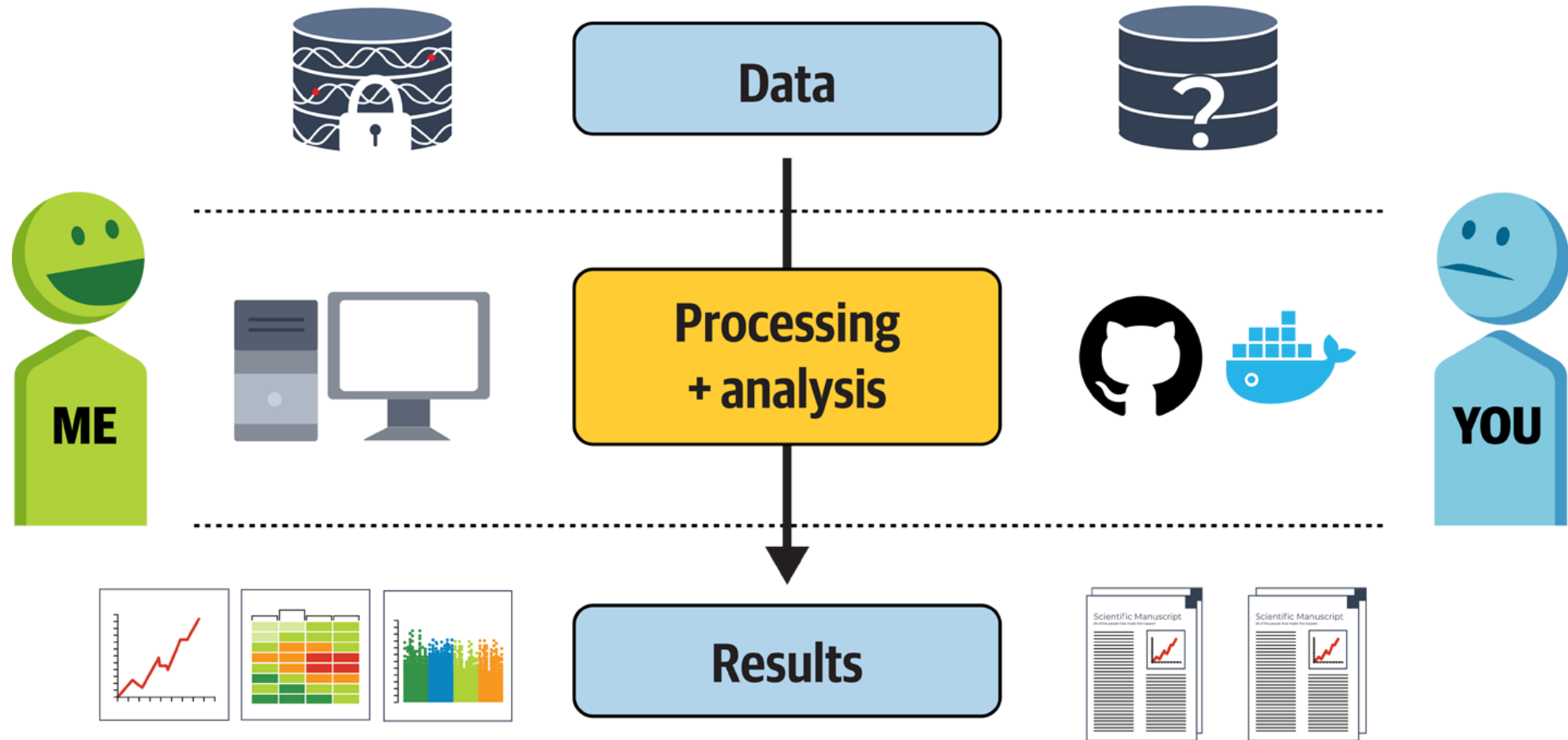
<sup>1</sup>Apitz (2009) 'Tetralogy of Fallot', Lancet, 374(9699), pp.1462-71.

<sup>2</sup>Baillard and Anderson (2009) 'Tetralogy of Fallot', Orphanet Journal of Rare Diseases, 4(1), p.2.

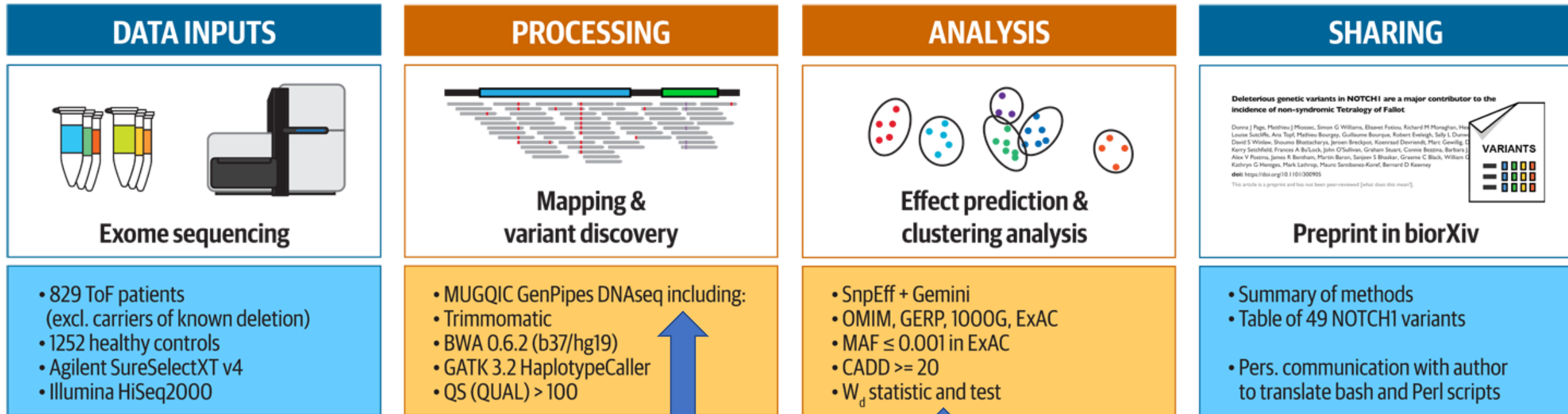
<sup>3</sup>Starr (2010) 'Tetralogy of Fallot: Yesterday and Today', World Journal of Surgery, 34(4), pp. 658-668.



# Typical information asymmetry



# Summary of tetralogy of Fallot pre-print

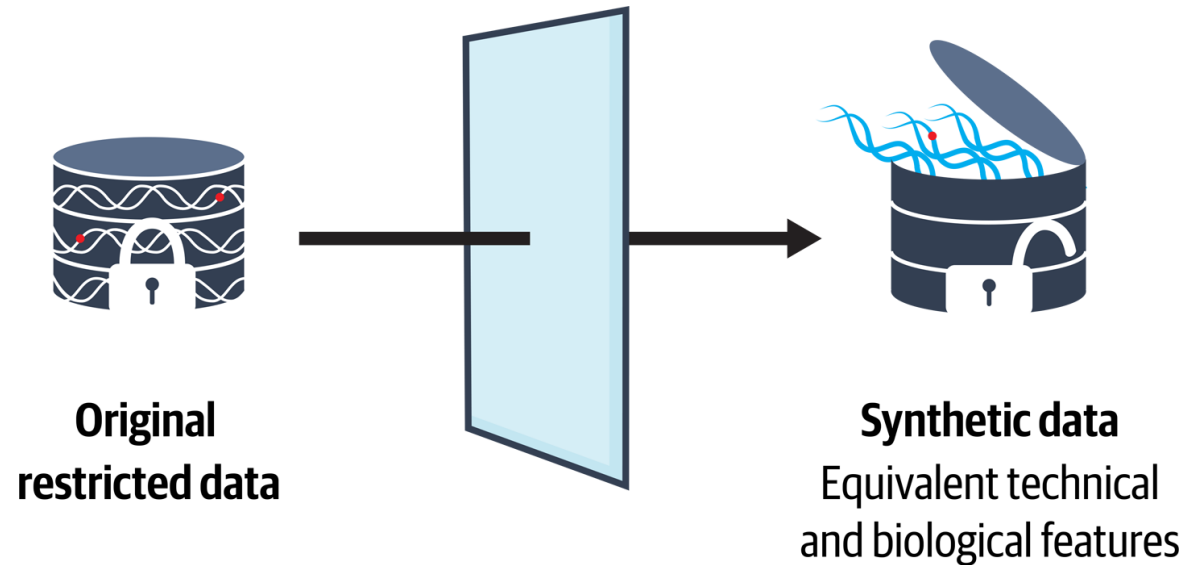


Too vague to reproduce exactly

Going to need some help with that one!



# Designing a Reproducible implementation



- Data input: not directly available.
- Next best thing: synthetic data with equivalent features.

- Case study inevitably an imperfect reproduction of original study.
  - “Evaluate the computational barriers to reproducibility of genomics methods...”
- But result should be fully reproducible itself.
  - Showcasing an example of reproducibility.



# Synthetic data generation overview

Open-access callset  
(1000 Genomes)



Extract variants

Individual VCFs

CHROM	POS	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA12877	NA12878	NA12882
20	61080	-	C	455.13		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1241;F=				
20	61138	-	C	155.18		AC=1;AF=0.107;AN=4;BaseQRankSum=-7.35;C1LogRRankSum=0.00;DP=1				
20	61795	-	G	384.18		AC=2;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61710	-	C	126.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61244	-	A	823.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61701	-	C	176.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	64223	-	A	159.44		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61096	-	G	155.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61088	-	A	1817.13		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61176	-	A	813.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	66728	-	C	2284.18		AC=2;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	67388	-	T	7102.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	68049	-	C	4265.18		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61064	-	G	1162.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61060	-	C	1198.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	61066	-	G	1042.12		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				
20	70404	-	CTCT	823.89		AC=1;AF=0.107;AN=4;BaseQRankSum=-3.55;C1LogRRankSum=0.00;DP=1				

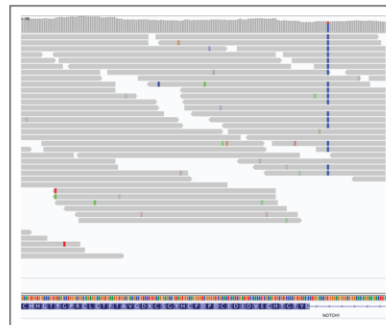
Generate reads

NEAT genReads

+ Exome intervals



Open-access synthetic reads  
with original variants



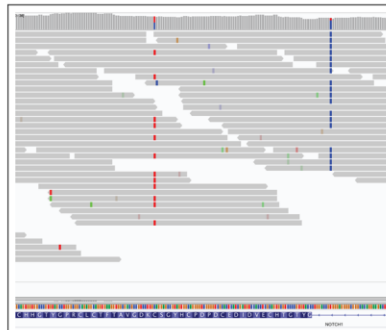
Spike in variants

BAM Surgeon

+ NOTCH1 variant



Open-access synthetic reads  
with *NOTCH1* variant(s)



BED file

Neutral variants  
to avoid batch  
effects.

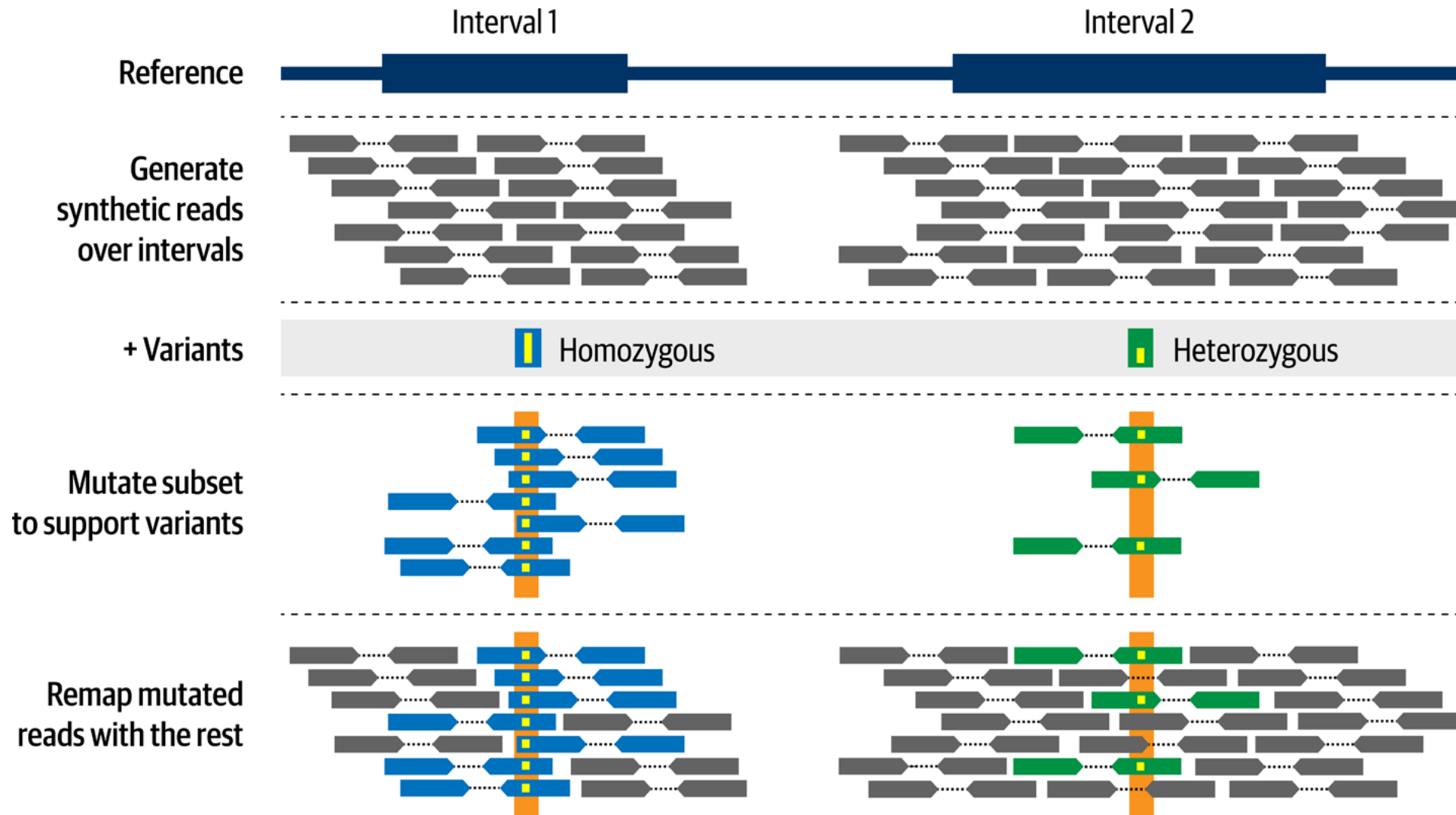
**Deleterious genetic variants in *NOTCH1* are a major contributor to the incidence of non-syndromic Tetralogy of Fallot**

Donna J. Page, Matthieu J. Miossec, Simon G. Williams, Elisavet Fotiou, Richard M. Monaghan, Heather J. Cordell, Louise Sutcliffe, Ana Topf, Mathieu Bourgey, Guillaume Bourque, Robert Eveleigh, Sally L. Dunwoodie, David S. Winlaw, Shoumo Bhattacharya, Jeroen Breckpot, Koenraad Devriendt, Marc Gewillig, David Brook, Kerry Setchfield, Frances A. Bu'Lock, John O'Sullivan, Graham Stuart, Connie Bezzina, Barbara J.M. Mulder, Alex V. Postma, James R. Bentham, Martin Baron, Sanjeev S. Bhaskar, Graeme C. Black, William G. Newman, Kathryn E. Hentges, Mark Lathrop, Mauro Santibanez-Koref, Bernard D. Keavney  
doi: <https://doi.org/10.1101/300905>

	A	B	C	D	E	F	G	H	I	J
1	chrom	start	end	ref	alt	qual	gene	transcript	aa_changis_exo	
2	chr9	139418414	139418415	C	T	698.71	NOTCH1	ENST00000277541 V53M		
3	chr9	139418360	139418361	C	T	803.78	NOTCH1	ENST00000277541 A71T		
4	chr9	139418357	139418358	C	T	889.28	NOTCH1	ENST00000277541 G72R		
5	chr9	139418226	139418228	GC	G	796.28	NOTCH1	ENST00000277541 G115		
6	chr9	139417615	139417616	G	A	2393.22	NOTCH1	ENST00000277541 P143L		
7	chr9	139417602	139417604	GT	G	289.2	NOTCH1	ENST00000277541 N147		
8	chr9	139417591	139417592	T	C	541.13	NOTCH1	ENST00000277541 N151S		
9	chr9	139417498	139417499	C	T	386.14	NOTCH1	ENST00000277541 C182Y		
10	chr9	139417465	139417466	C	G	322.43	NOTCH1	ENST00000277541 G193A		

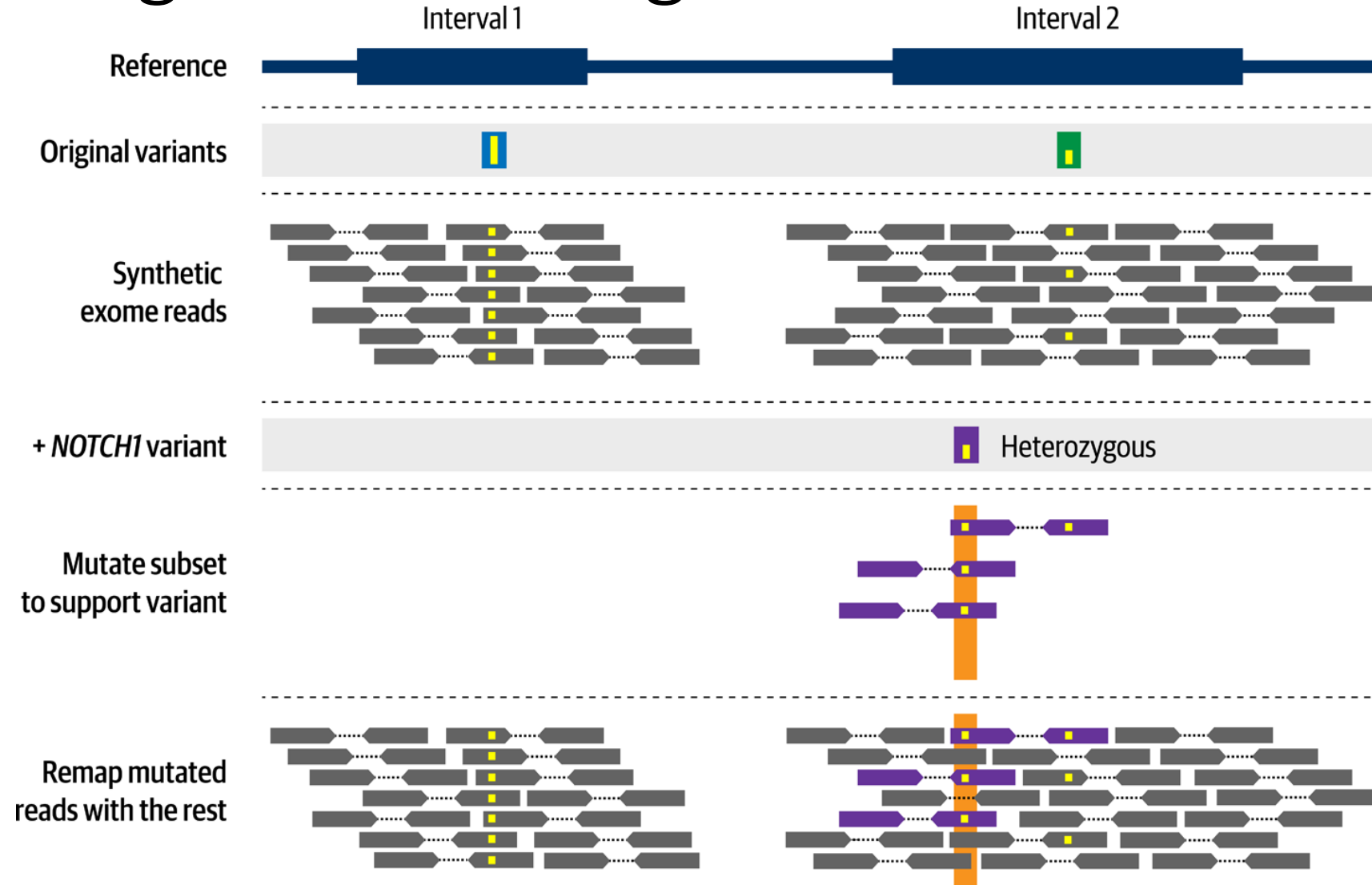


# NEAT-genReads – simulated read data





# BAMSurgeon – making fake exomes



Note: neutral variant added to other exomes to avoid batch effect



# Processing and analysis phases

## PROCESSING



### Mapping & variant discovery

- MUGQIC GenPipes DNaseq including:
- Trimmomatic
- BWA 0.6.2 (b37/hg19)
- GATK 3.2 HaplotypeCaller
- QS (QUAL) > 100

Using GATK4 instead in this reproduction (speed, scalability, etc...)

At this stage we have a large list of deviations from the reference genome (many variants and some errors)



**Which genetic variants should we be looking at more closely?**

## ANALYSIS



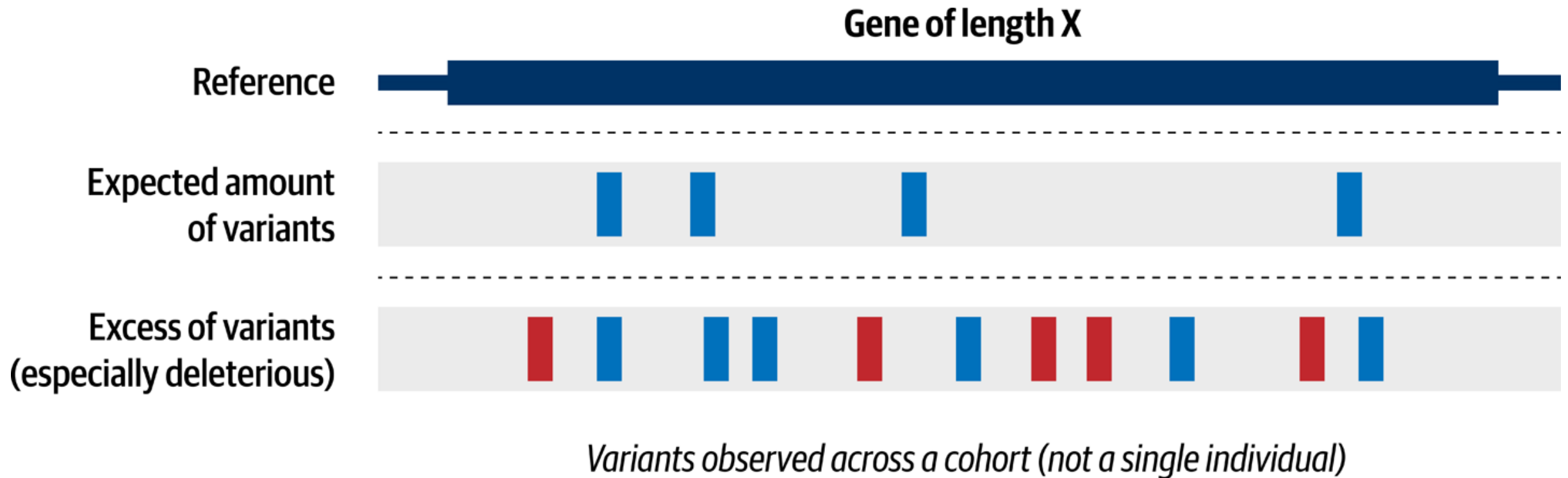
### Effect prediction & clustering analysis

- SnpEff + Gemini
- OMIM, GERP, 1000G, ExAC
- $MAF \leq 0.001$  in ExAC
- CADD  $\geq 20$
- $W_d$  statistic and test

In the Jupyter notebook.



# Comparing variant load across multiple samples



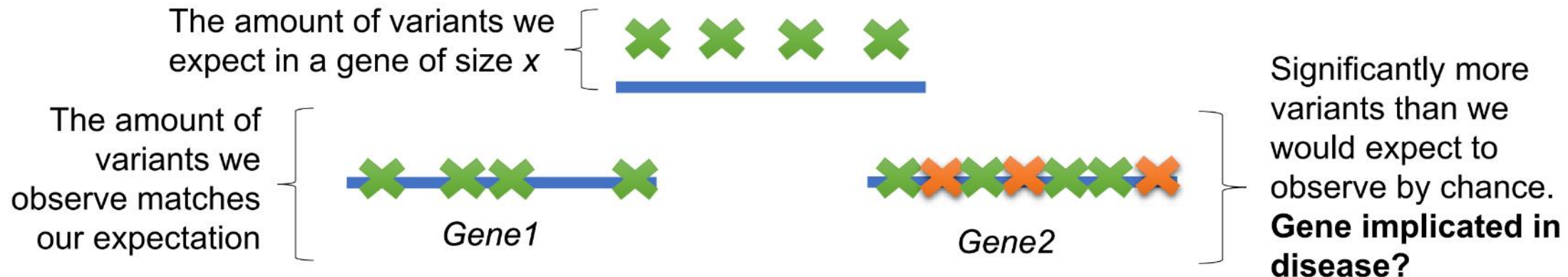
# Variant Load Analysis (Overview)

This is what my version of that last graphic looked like in November of 2018.

We are still left with thousands of variants:

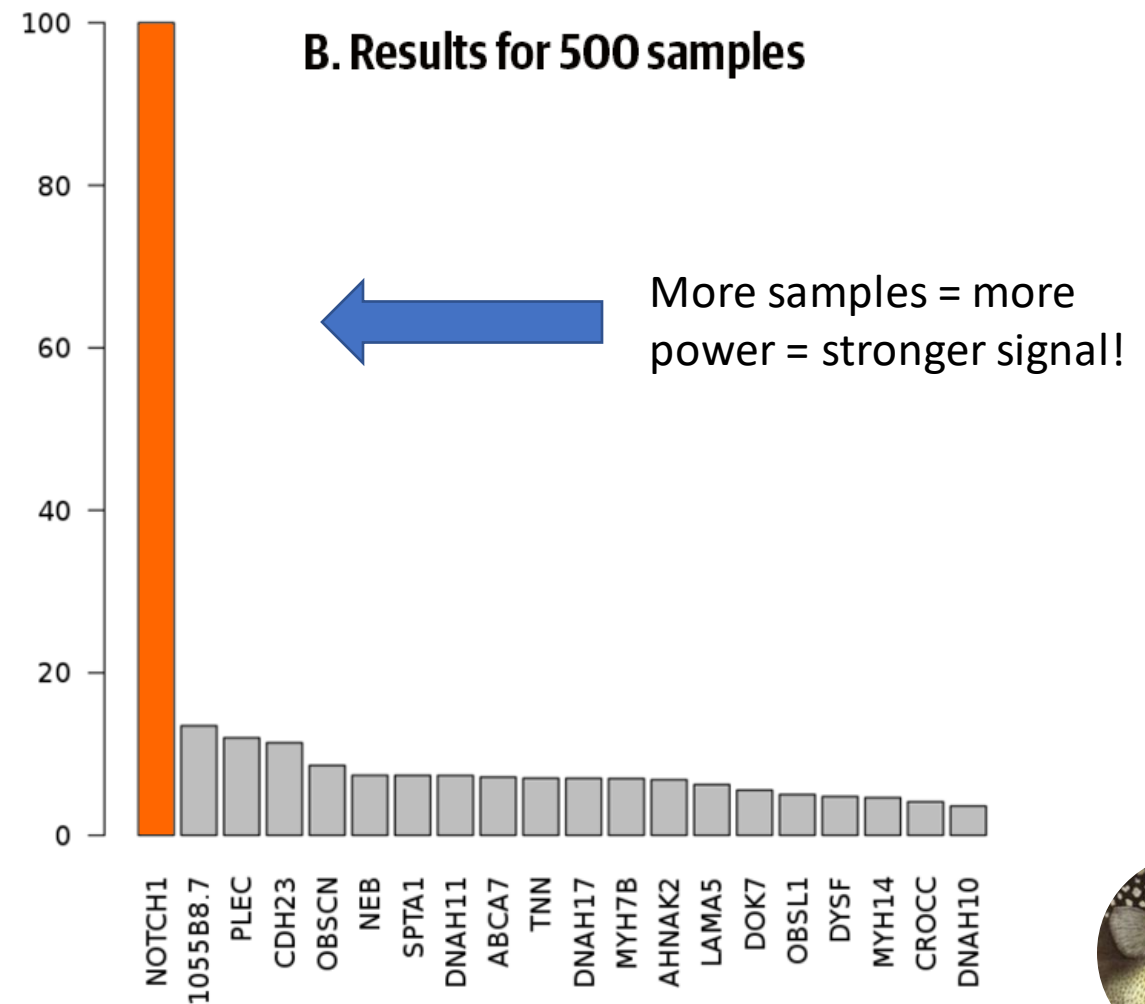
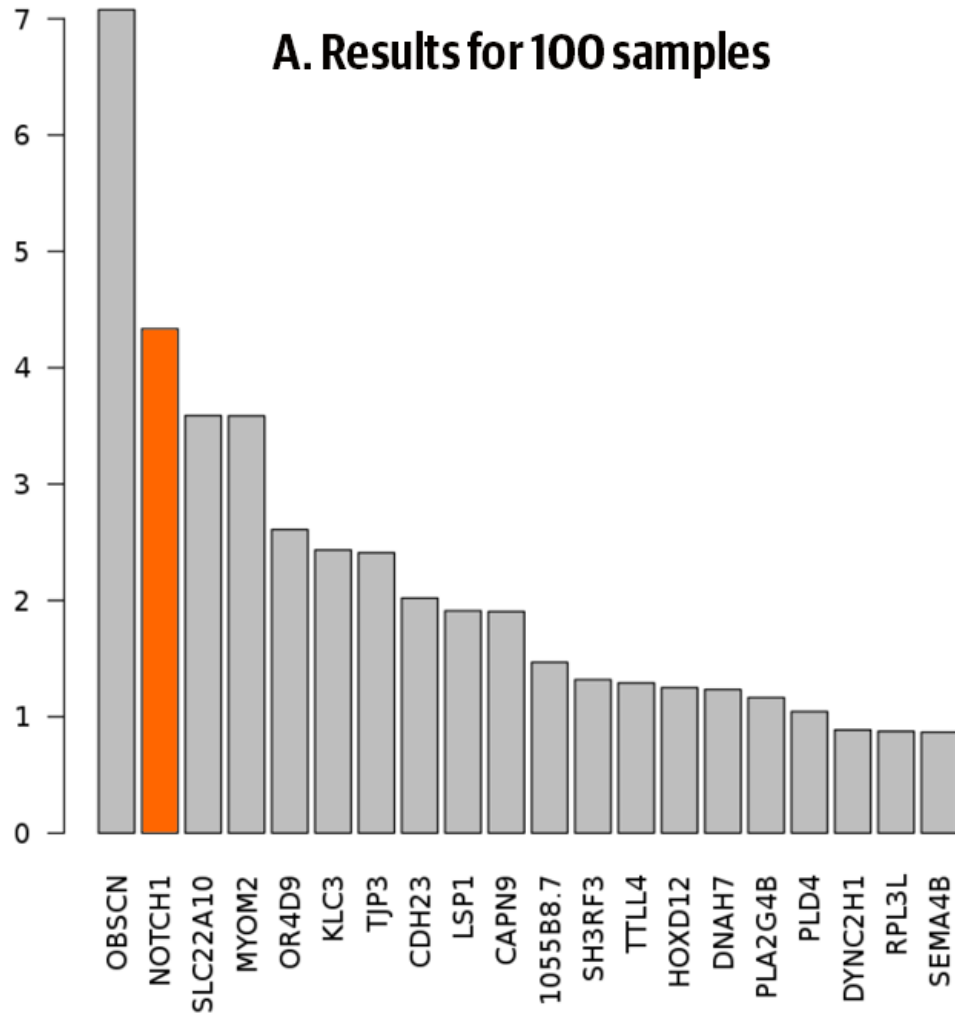
- Not every variant that is rare is of high impact.
- Not every variant that is predicted deleterious actually is.

However, our dataset should be enriched for truly deleterious rare variants. These should aggregate in disease-related genes. We need a method of identifying such genes.



More on this later...

# Clustering test rankings





# The long, Winding Road to FAIRness

- Achieving computational reproducibility and FAIRness in your work:
  - Keep it open source!
  - Version control
    - Automatic versioning in Terra via Github, Dockstore or the internal method repository.
  - Automation and Portability
    - Prefer WDL/CWL over intricate BASH, Python or R scripts.
  - Built-in documentation
    - Jupyter notebooks are great for teaching/learning and therefore also great for documenting the logic of various steps of your analysis. Be explicit.
  - Open Data
    - Refer to open-access or synthetic data that is sufficient to demonstrate your analysis if original data is not available (eg. Protected patient data).



# Additional resources

- Tetralogy of Fallot resources
  - <https://doi.org/10.1161/CIRCRESAHA.118.313250>
  - <https://www.childrenshospital.org/conditions-and-treatments/conditions/t/tetralogy-of-fallot>
- Organizational resources
  - <https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989>
  - <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424>
- Reproducibility best practice resources
  - <http://www.practicereproducibleresearch.org>
  - <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745>
  - <https://repro4everyone.org/pages/explore-topics/>
- FAIR resources
  - <https://www.go-fair.org/fair-principles>



# Additional resources (case study since 2018)

- Posters:
  - <http://broad.io/ASHG2018> (ASHG 2018)
  - <https://f1000research.com/posters/8-1380> (BOSC 2019)
- Workshop in Viña del Mar:
  - <https://broad.io/gatk-1811> (ISCB-LA SOIBIO 2018)
- Recorded presentations:
  - <https://youtu.be/xOzwWNLXdHc> (BroadE, July 2019)
- Books:
  - Genomics in the Cloud (2020)





Thank you for joining us on this journey!

Special thanks to:

- Our presenters
- Our members
- Dr. Geraldine Van der Auwera