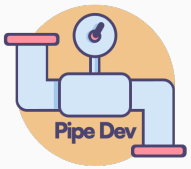


SwissPedHealth Analysis pipelines

Dylan Lawless

May 16, 2023

Content

Bioinformatics

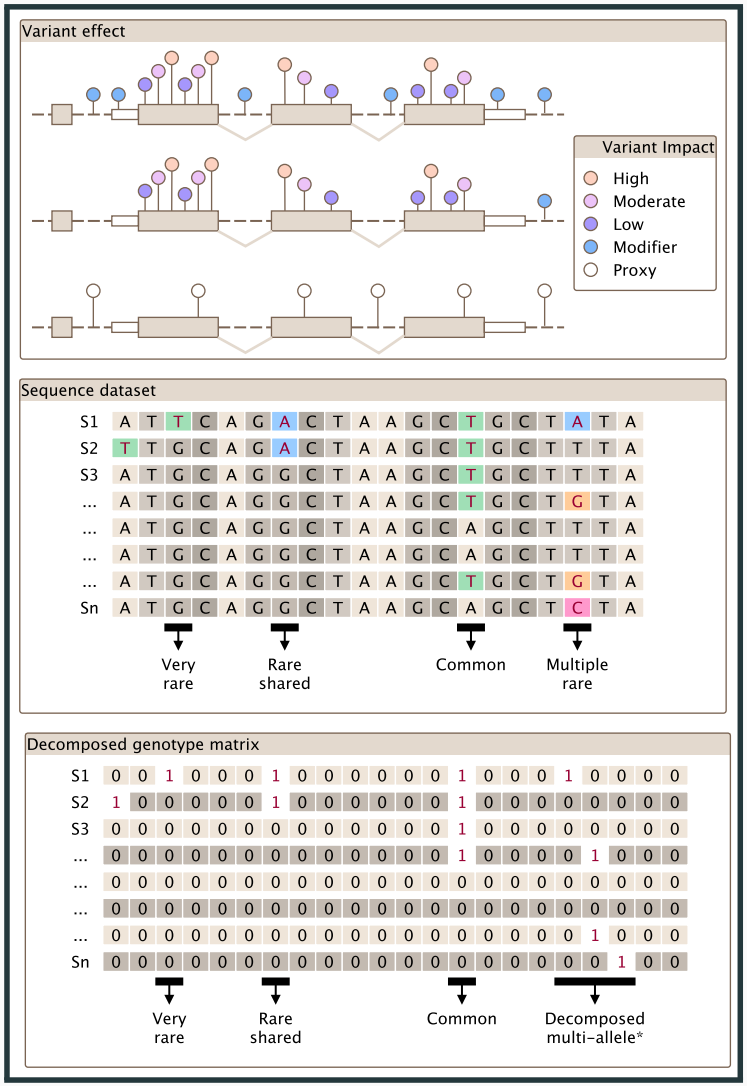
Primary Analysis

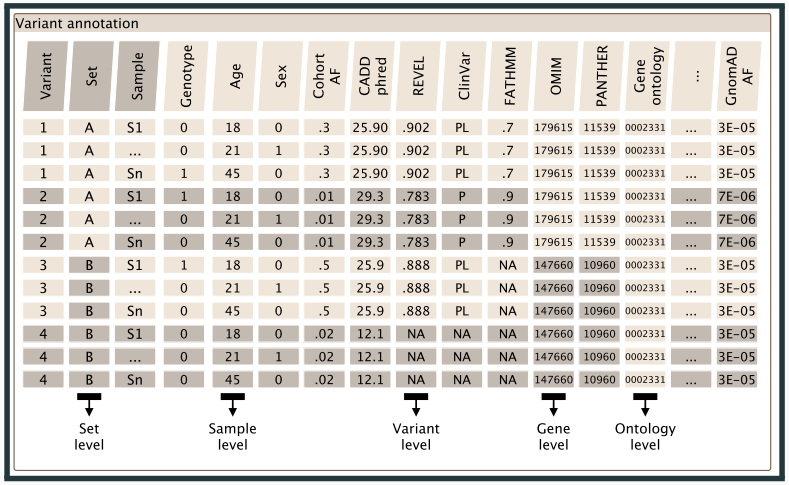
Secondary Analysis

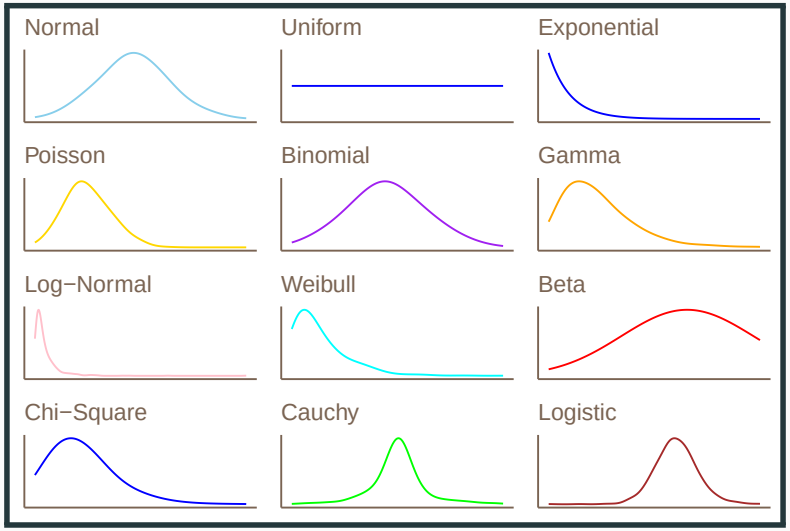
Demo

Bioinformatics

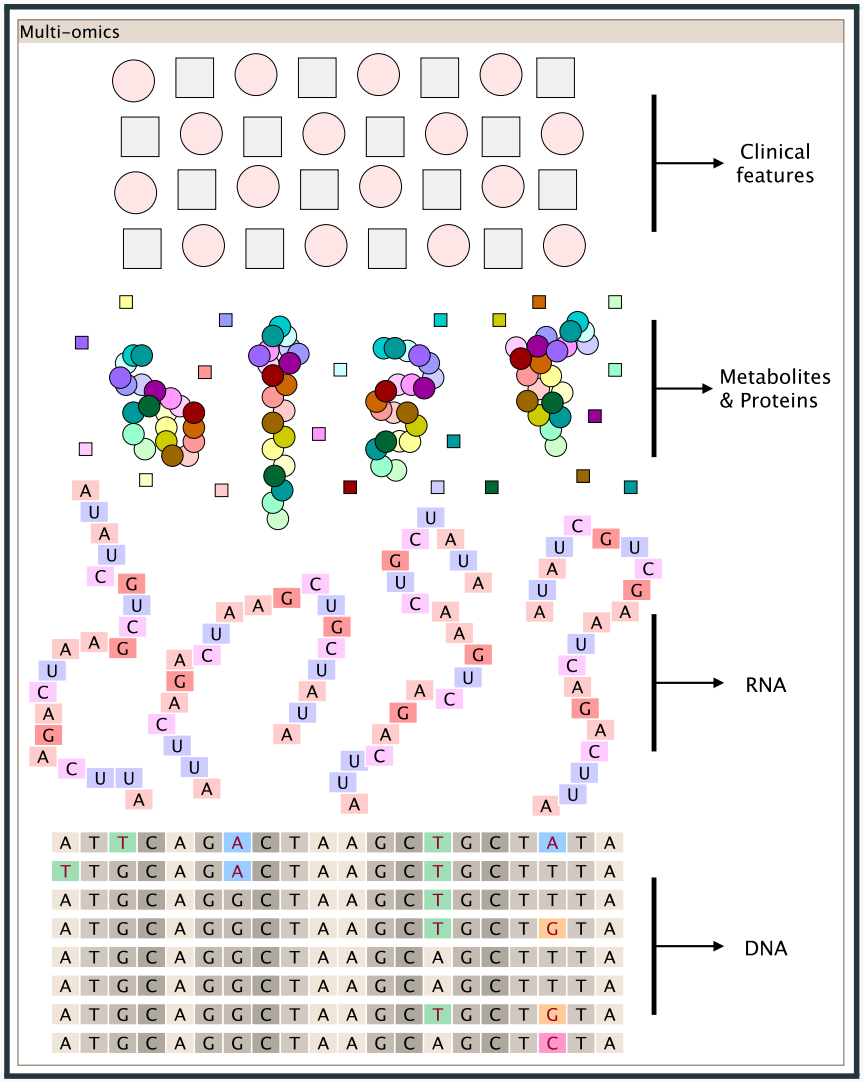
Bioinformatics







Bioinformatics



Demographics

- $Pheno \sim Clin.predictor + age + sex$

Machine learning

- $PredOutcome \sim ClinFeat + age + PC$

Statistical Genomics

- $DNA \sim Pheno + age + PC$
- $DNA \sim Pheno + RNA + Metab$
- $DNA \sim Metab + PC$

Primary Analysis

Primary - DNA, RNA, metabolomic and proteomic

DNA

- SNV and INDEL
- Structural variant
- Coding and non-coding

RNA

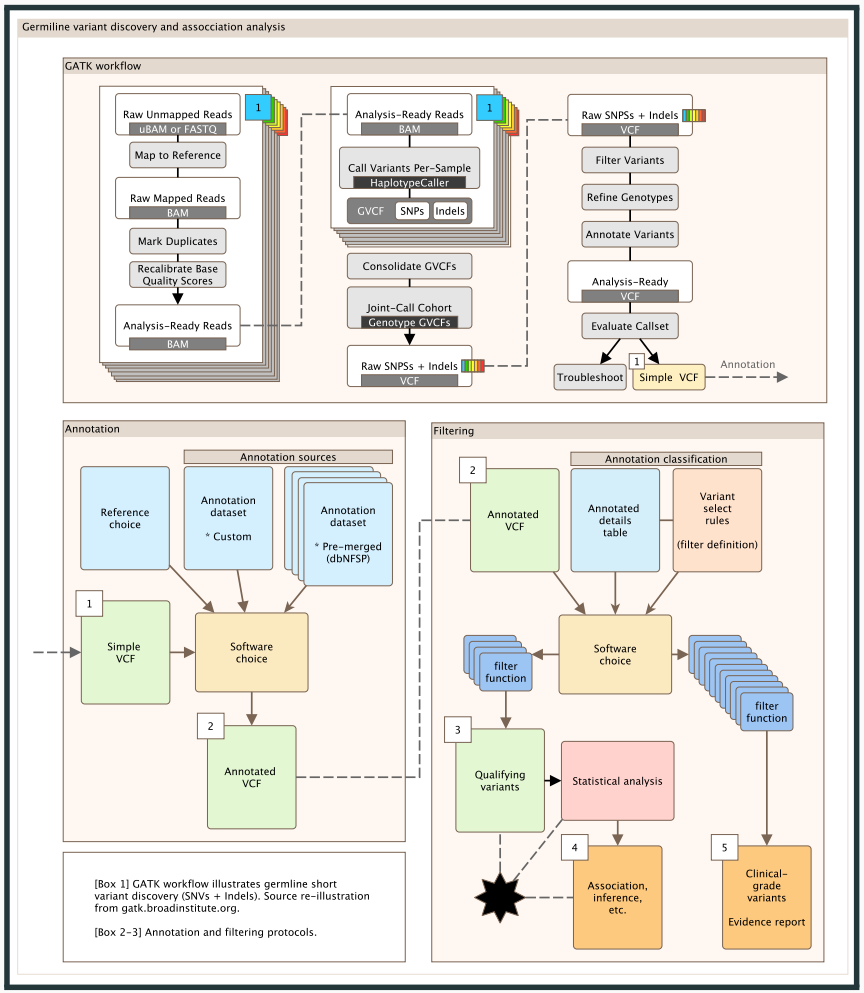
- Quant expression
- Splice
- ASE

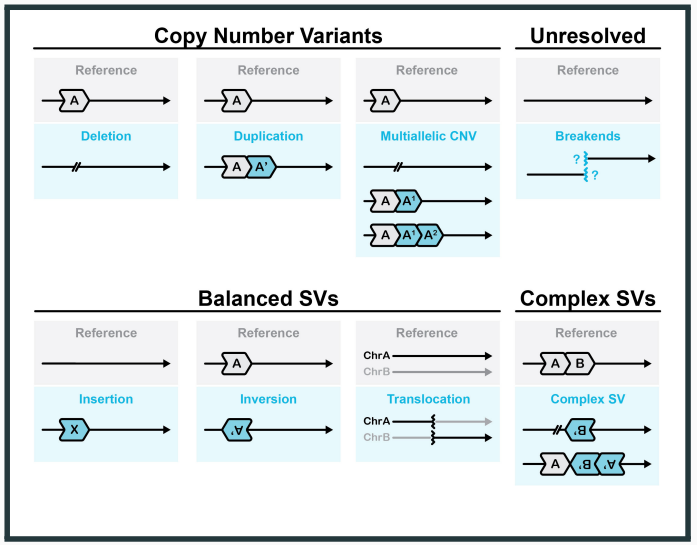
Metabolomic, proteomic, clinical

- Distribution (MetaboAnalystR)
- Visualisation
- QC

Clinical report, ACMG, and best-practice [1, 2, 3, 4]

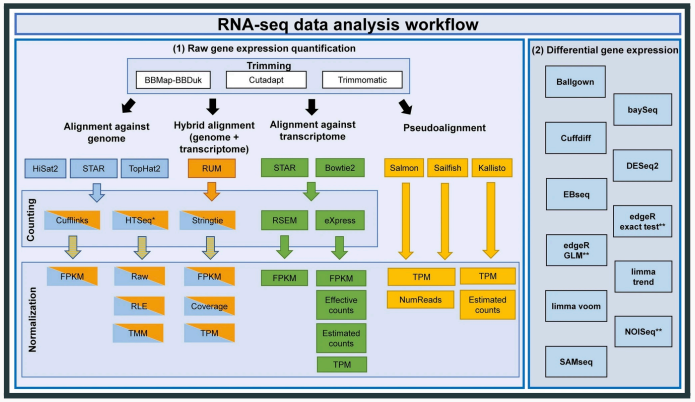
Primary DNA

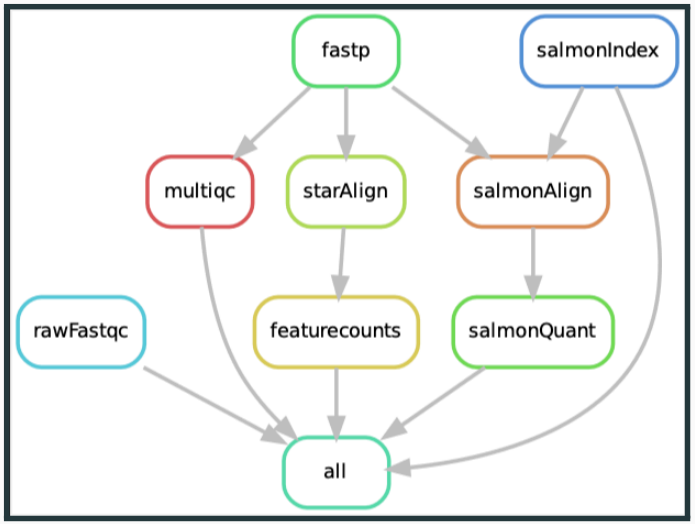


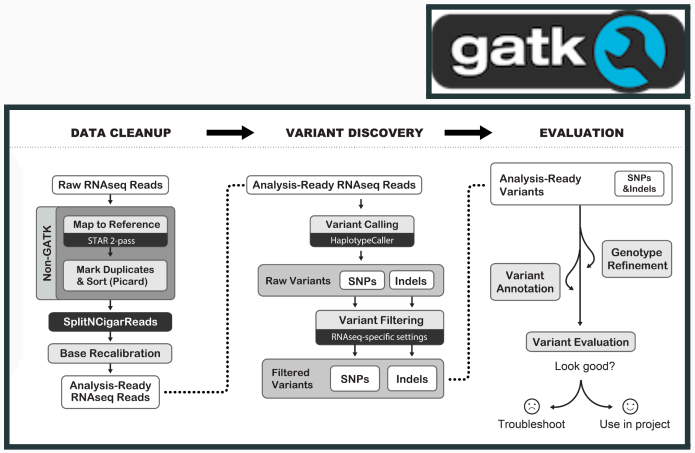


- SNV and INDEL (GATK, VEP) [5]
- Structural variant (GATK, smoove, indexcov)
- Coding / non-coding [6, 7]

Primary RNA







- Bulk RNA-seq pipieline[a]: QC, DEG (DESeq2, edgeR)
- Splicing (SUPPA, rMATS, Leafcutter, SVA, FRASER, and SPOT)

[a]Shweta Pipaliya, PhD.

Secondary Analysis

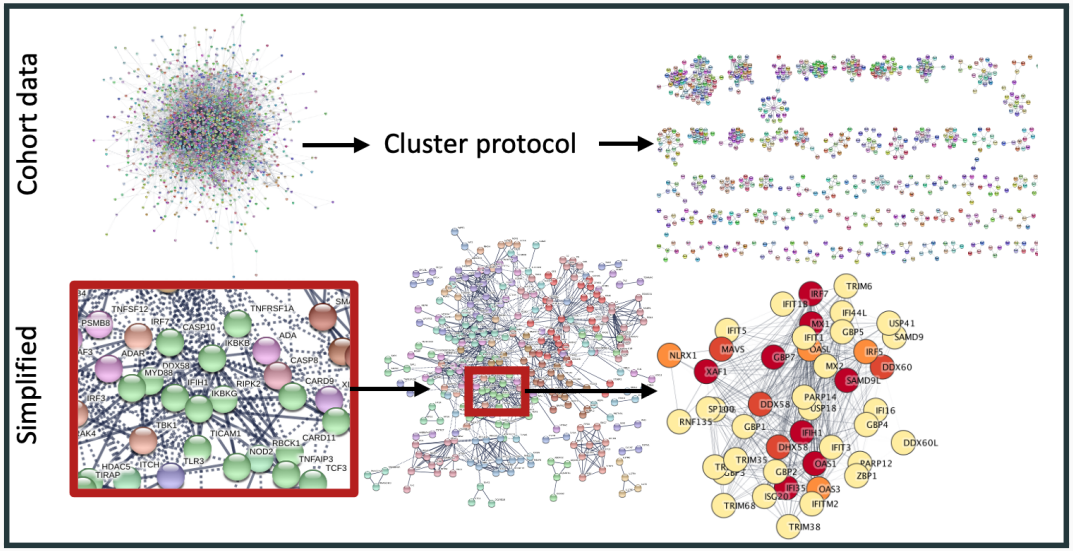
Secondary - Variant, Gene, VSAT

- **DNA** Single variant
- **DNA** Gene burden
- **DNA** Variant set / Protein pathway
- **RNA** DEG, splicing, GSEA, protein pathway
- **Joint** with metabolomic / proteomic

ProteoMCLustR

Proteome clustering with
Markov cluster algorithm (MCL) in R
for high performance computing platforms

ProteoMCLustR



ProteoMCLustR

Input:

$N_i,\ i = 1, \ldots, n$:
Nodes (genes) in the STRING database

$E_{ij},\ i, j = 1, \ldots, n$:
Edges (interactions) between nodes in STRING database

S : Score threshold for edges

I : Iteration limit

$L_{\text{min}}, L_{\text{max}}$: Size limits for clusters

e, r : Expansion & inflation parameters for MCL algorithm

Algorithm:

1. Preprocess $(N_i, E_{ij}, S) \rightarrow (N_i', E_{ij}')$

2. ChooseInflation $(N_i', E_{ij}', L_{\text{min}}, L_{\text{max}}) \rightarrow$ inflation

3. RunMCL $(N_i', E_{ij}', I, L_{\text{min}}, L_{\text{max}}, \text{inflation}, e, r)$

3.1. Initialize $M_{ij}^{(0)} = \dfrac{E_{ij}'}{\sum_{k=1}^{n} E_{ik}'}$

3.2. Iterate until convergence:

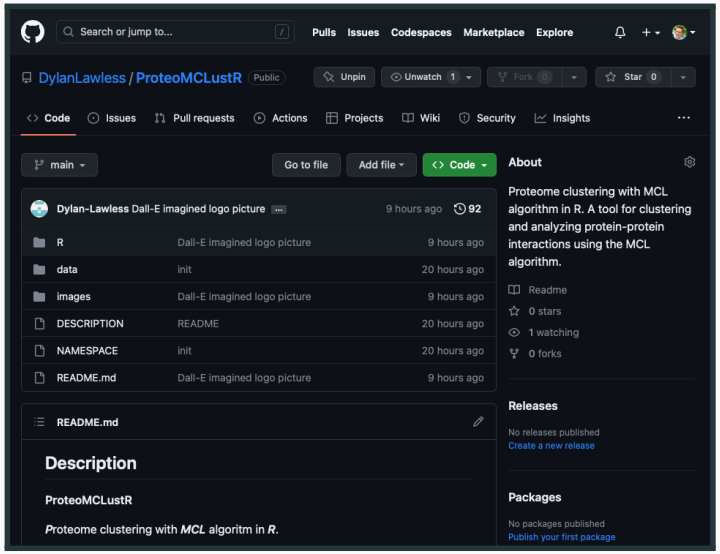
3.2.1. Expansion $M^{(k)} = (M^{(k-1)})^e$

3.2.2. Inflation $M_{ij}^{(k)} = \dfrac{(M_{ij}^{(k-1)})^r}{\sum_{k=1}^{n} (M_{ik}^{(k-1)})^r}$

3.3. Extract clusters from converged matrix $M^{(\text{final})}$

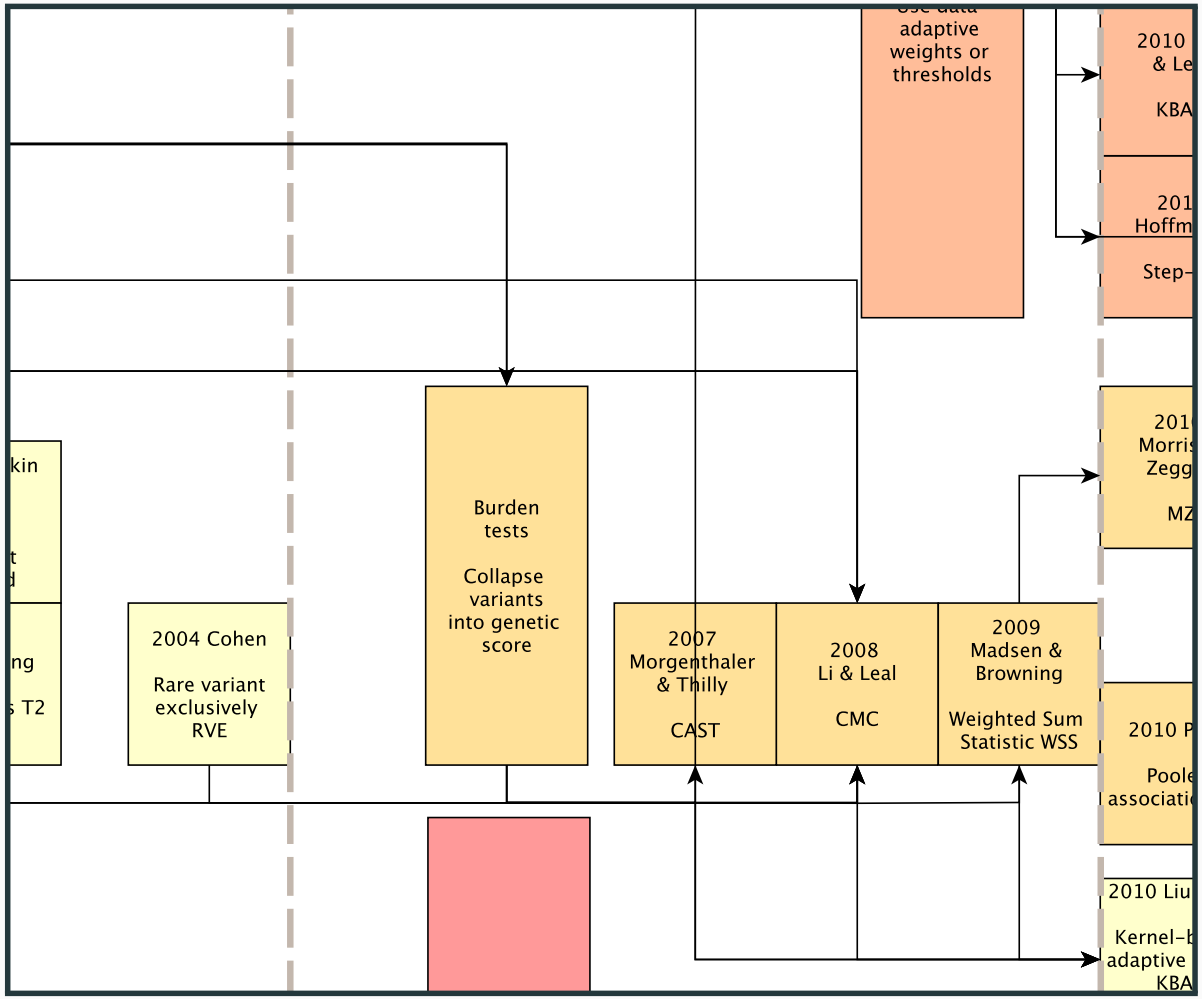
Output:

Clusters: Set of optimized node (gene) clusters



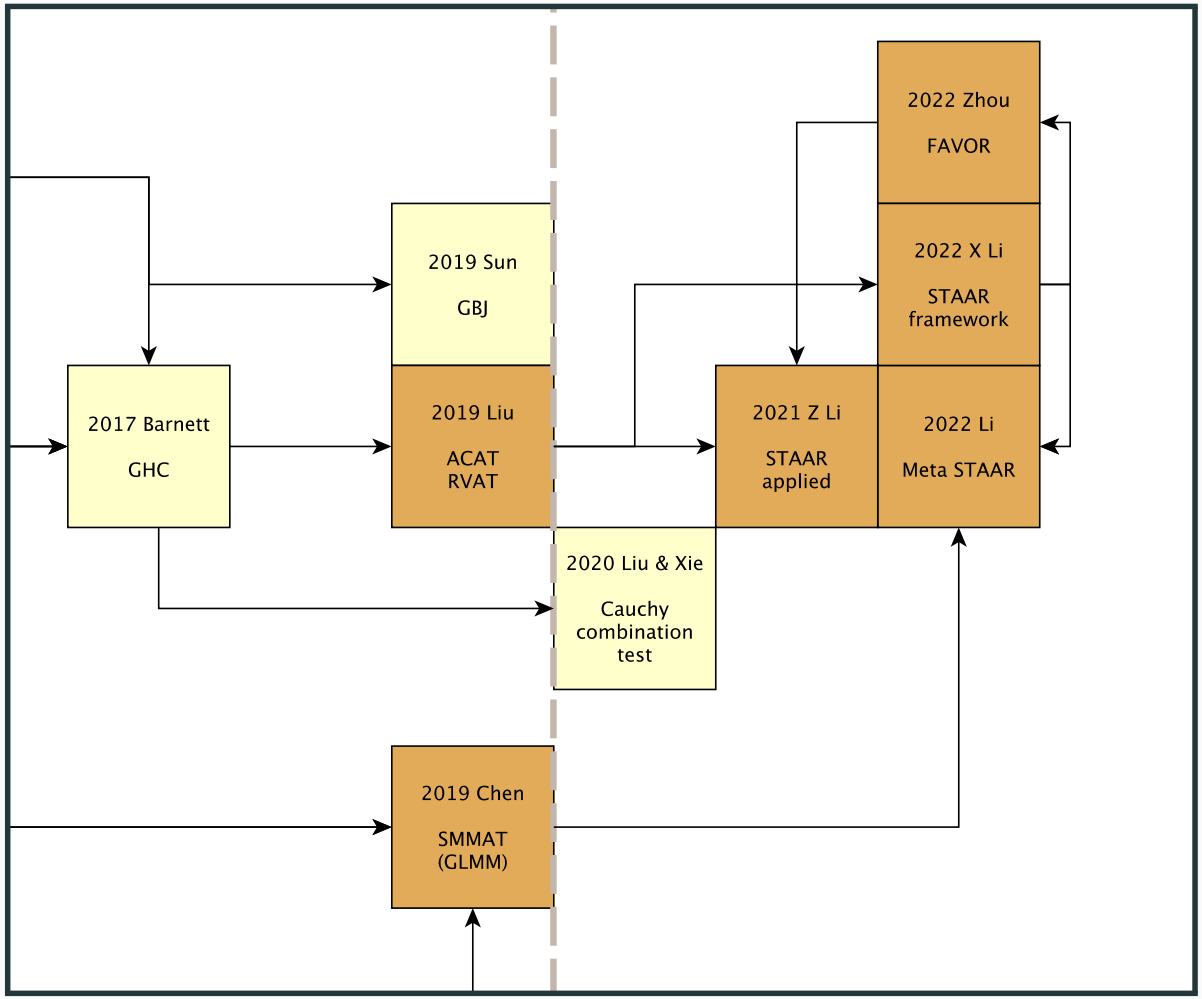
Statistics

Variant analysis map



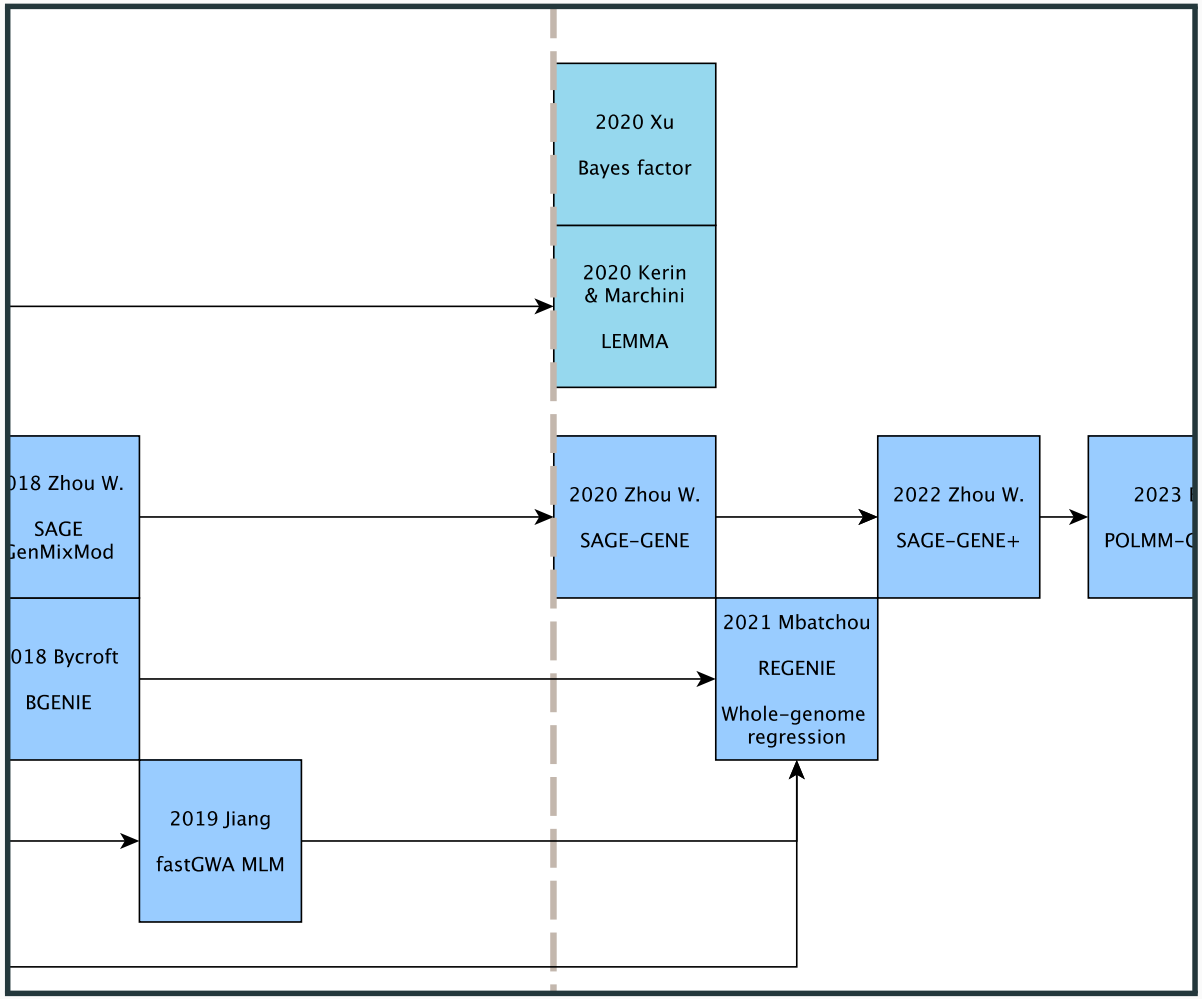
Statistics

Variant analysis map



Statistics

Variant analysis map



Demo

Demo - Case Report

Pertinent findings: Sample - SPH00123

Genomics analysis report: Sample_id - SPH00123

Germline variant ENST00000646337.2:c.888G>A (p.Phe319SerTer15) *AVPR2* was identified as high confidence disease-causing.

1. Based on evidence from all known relevant databases, this variant was interpreted as a disease causing *true positive*.
2. None of our *critical databases* had *missing information* about this variant, thereby reducing the likelihood of a *false positive*.
3. No alternative candidate variants were ignored due to a lack of evidence, thereby reducing the likelihood of a *false positive*.
4. All genomic positions where variants are known to produce similar phenotypes were checked and were not found to contain such variants, thereby reducing the likelihood of *false negatives*.
5. All other genome-wide *VUS* were interpreted as being unrelated to disease, *true negatives*.

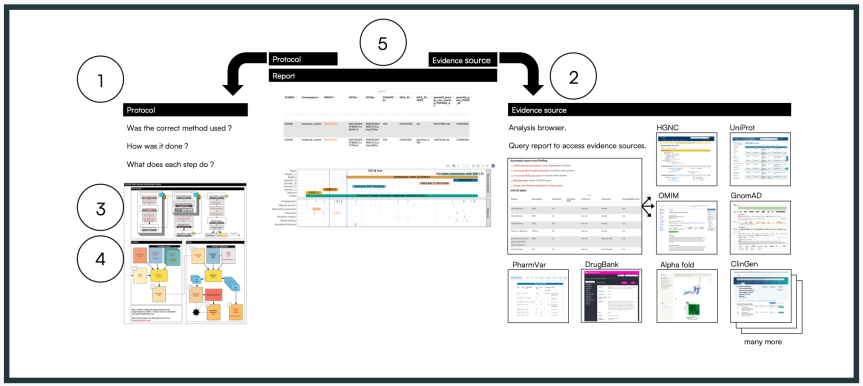
Read more

true positive
critical databases
missing information
false positive

false negatives

VUS
true negatives

Was this analysis performed adequately? See here
Were the evidence sources used up-to-date and reliable? See here
Next steps See here



Demo - Cohort Report

Pertinent findings: Top 25
Based on: 490 cases
Controls: Yes (n=500)

Search

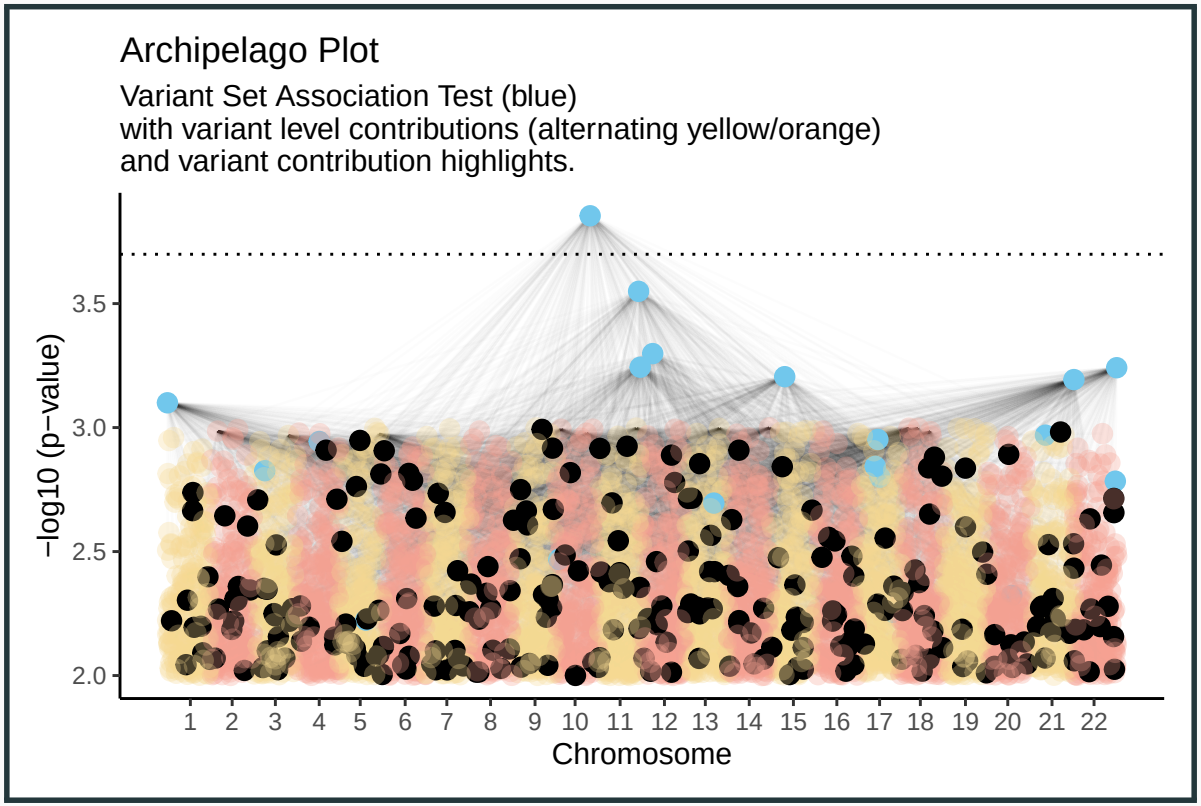
Common.name	Gene.Name	Name	cDNA_position	Codons	CDS_position	Protein_position	Amino_acids	Location	Consequence	IMPACT	EXON	INTRON	HGVSc	HGVSp
Trimetazidine	ACAA1	3-ketoacyl-CoA thiolase, peroxisomal	1138-?	-	1046-?	349-?	-	3:38125813-38125833	splice_donor_variant,splice_donor_5th_base_variant,coding_sequence_variant,intron_variant	HIGH	10/12	10/11	ENST00000333167.13:c.1046_1053+13del	-
Aurothioglucose	ADCY2	Adenylate cyclase type 2	1110-1111	-/CT	951-952	317-318	-/X	5:7695833-7695834	frameshift_variant	HIGH	6/25	-	ENST00000338316.9:c.954_955dup	ENSP00000342952.4:p.Phe319SerfsTer15
Conivaptan	AVPR2	Vasopressin V2 receptor	1159	tgG/tgA	888	296	W/*	X:153906394	stop_gained	HIGH	3/4	-	ENST00000646375.2:c.888G>A	ENSP00000496396.1:p.Trp296Ter
Human immunoglobulin G	C4B	Complement C4-B	-	-	-	-	-	6:32016105	splice_donor_variant	HIGH	-	5/40	ENST00000435363.7:c.626+1G>A	-
Isopropyl alcohol	DDX39B	Spliceosome RNA helicase DDX39B	236-237	gGA/gGAGA	50-51	17	G/GX	6:31540482-31540483	frameshift_variant	HIGH	2/11	-	ENST00000396172.6:c.50_51dup	ENSP00000379475.1:p.Glu17GlyfsTer103
Zinc	DSP	Desmoplakin	3483	Cag/Tag	3238	1080	Q/*	6:7579428	stop_gained	HIGH	23/24	-	ENST00000379802.8:c.3238C>T	ENSP00000369129.3:p.Gln1080Ter
Etanercept	FCGR2C	Low affinity immunoglobulin gamma Fc region receptor II-c	960	Taa/Caa	862	288	*/Q	1:161599693	stop_lost	HIGH	7/7	-	ENST00000466542.6:c.862T>C	ENSP00000426627.1:p.Ter288GlnextTer?
Hyaluronic acid	HAPLN4	Hyaluronan and proteoglycan link protein 4	1075-1076	ccg/cCTTTGAAGGATGAATAAGAGTTCAACAGGCAAACAGTcg	1003-1004	335	P/PLKDE*EFNRQTVX	19:19258022-19258023	stop_gained,frameshift_variant	HIGH	5/5	-	ENST00000291481.8:c.1003_1004insCTTTGAAGGATGAATAAGAGTTCAACAGGCAAACAGT	ENSP00000291481.5:p.Arg336LeufsTer5
Oprelvekin	IL11RA	Interleukin-11 receptor subunit alpha	-	-	-	-	-	9:34658685	splice_donor_variant	HIGH	-	8/12	ENST00000441545.7:c.810+2G>T	-
Efalizumab	ITGAL	Integrin alpha-L	1092	taT/taA	996	332	Y/*	16:30484253	stop_gained	HIGH	9/31	-	ENST00000356798.11:c.996T>A	ENSP00000349252.5:p.Tyr332Ter

1–10 of 25 rows Show 10

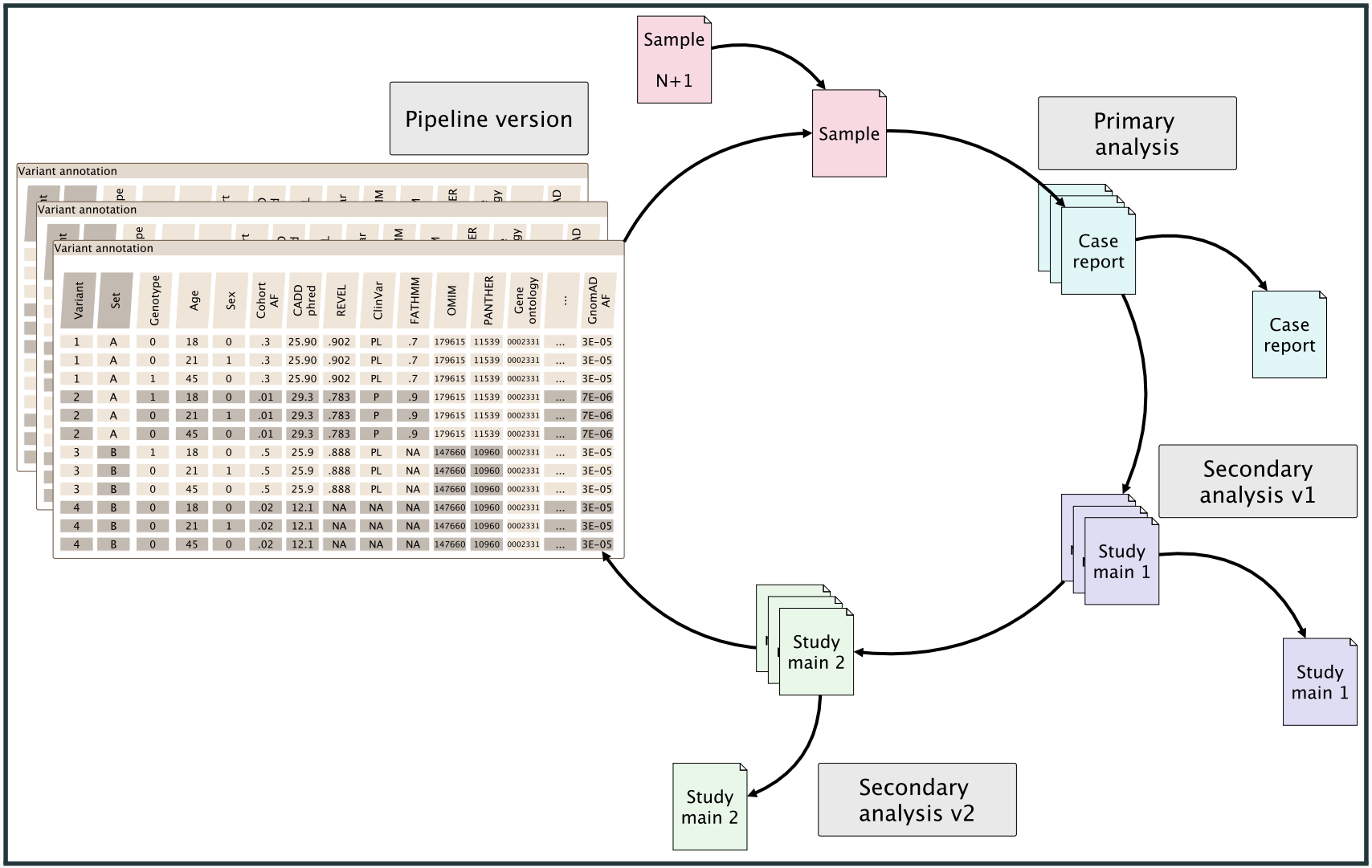
Previous 1 2 3 Next

Demo - Secondary Report

Pertinent findings: Enriched protein pathway
Based on: 490 cases
Controls: Yes (n=500)

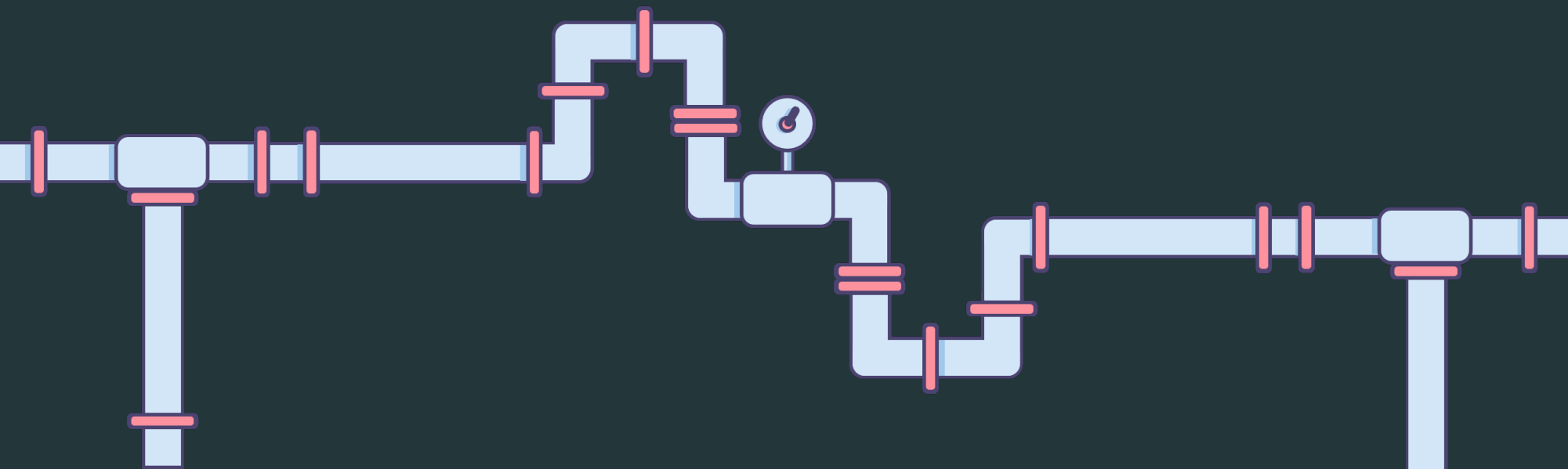


Summary



Variant	Set	Genotype	Age	Sex	Cohort AF	CADD phred	REVEL	ClinVar	FATHMM	OMIM	PANTHER	Gene ontology	...	GnomAD AF
1	A	0	18	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	0	21	1	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	1	45	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
2	A	1	18	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	0	21	1	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	0	45	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
3	B	1	18	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	0	21	1	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	0	45	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
4	B	0	18	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	0	21	1	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	0	45	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05

SwissPedHealth Analysis pipelines



References i

Sue Richards, Nazneen Aziz, Sherri Bale, et al. “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology”. In: *Genetics in medicine* 17.5 (2015), pp. 405—423.

Marilyn M Li, Michael Datto, Eric J Duncavage, et al. “Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists”. In: *The Journal of molecular diagnostics* 19.1 (2017), pp. 4—23.

Erin Rooney Riggs, Erica F Andersen, Athena M Cherry, et al. “Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMGe and the Clinical Genome Resource (ClinGen)”. In: *Genetics in Medicine* 22.2 (2020), pp. 245—257.

Brent S Pedersen, Joe M Brown, Harriet Dashnow, et al. “Effective variant filtering and expected candidate variant yield in studies of rare human disease”. In: *NPJ Genomic Medicine* 6.1 (2021), pp. 1—8.

Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, et al. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline”. In: *Current Protocols in Bioinformatics* 43.1 (2013), pp. 11.10.1—11.10.33. DOI: 10.1002/0471250953.bi1110s43. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471250953.bi1110s43. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43.

References ii

Zilin Li, Xihao Li, Hufeng Zhou, et al. “A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies”. en. In: *Nature Methods* 19.12 (Dec. 2022), pp. 1599—1611. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-022-01640-x. URL: https://www.nature.com/articles/s41592-022-01640-x (visited on 05/03/2023).

Xihao Li, Zilin Li, Hufeng Zhou, et al. “Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale”. en. In: *Nature Genetics* 52.9 (Sept. 2020), pp. 969—983. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-020-0676-4. URL: https://www.nature.com/articles/s41588-020-0676-4 (visited on 05/03/2023).

Gundula Povysil, Slavé Petrovski, Joseph Hostyk, et al. “Rare-variant collapsing analyses for complex traits: guidelines and applications”. In: *Nature Reviews Genetics* 20.12 (2019), pp. 747—759. DOI: 10.1038/s41576-019-0177-4. URL: https://doi.org/10.1038/s41576-019-0177-4.