Medicine

UNIVERSITY OF Southampton

Annotation and filtering

Dr. Reuben J. Pengelly

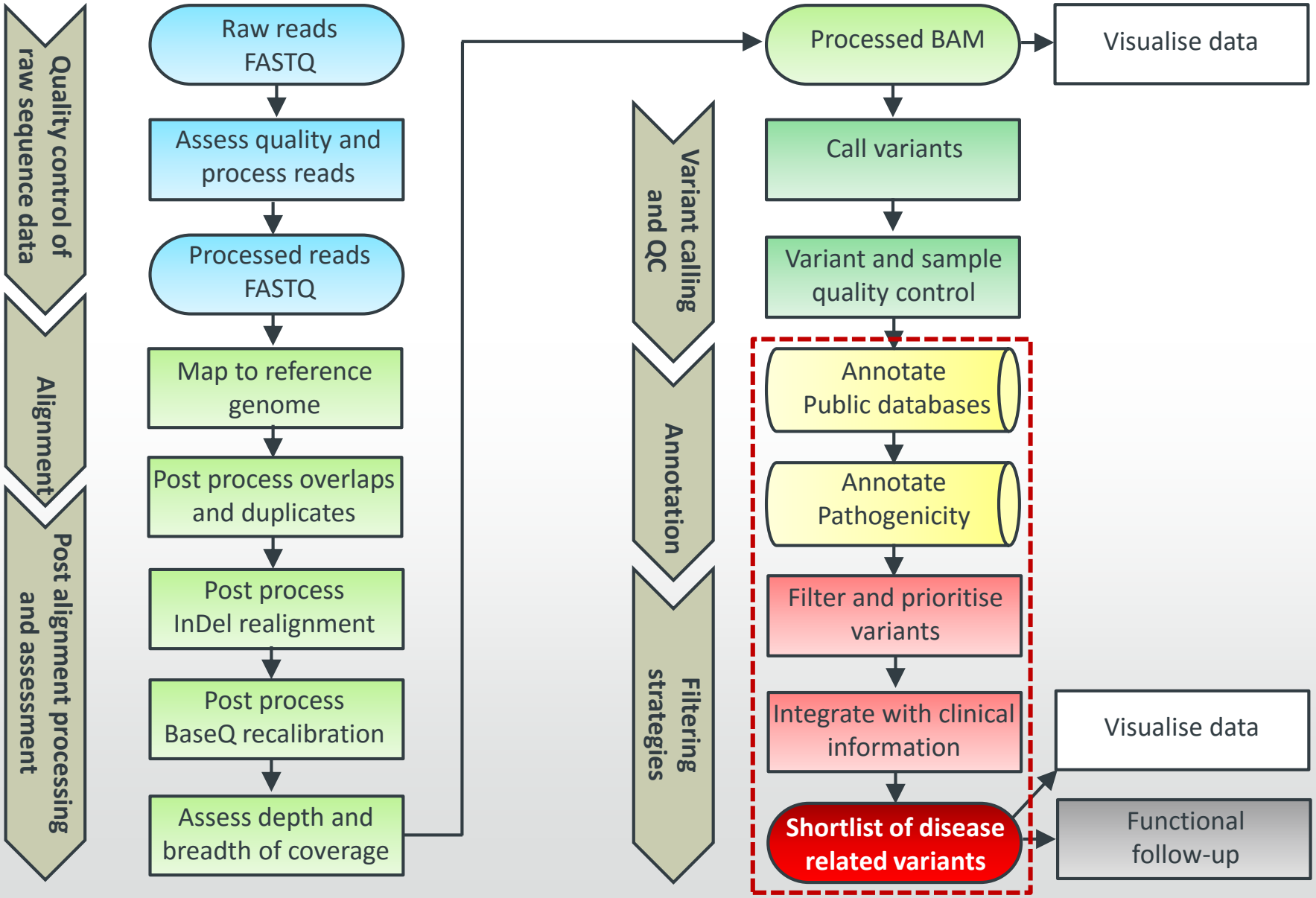
27th February, 2017

Learning outcomes

At the end of this lecture, you should be able to:

1. Describe and evaluate available annotations
2. Justify filtering strategies using genetic knowledge
3. Contrast variant-led and phenotype-led approaches

Analysis workflow



Data to annotate

- Variant functional effect
- Allele frequencies
- Conservation
- Functional effect prediction scores
- Disease databases for disease of interest
  - Genes
  - Variants

Functional effect

- Is variant synonymous, stop-gain, frameshift…
  - Different transcripts may be affected differently
  - Worst affected transcript may not be clinically relevant
  - Strict guidelines (http://varnomen.hgvs.org/)
- Choice of transcript database (RefSeq or ensembl)
- Reduces complex biology to an absolute, sometimes imperfectly

NOD2:NM_001293557:exon3:c.T953C:p.V318A

Variant type	p. description	Remarks
	p.(Arg490Ser)	The protein change is predicted (no experimental proof)
Substitution	p.Arg490Ser/p.R490S p.Trp87Ter / p.Trp78∗/p.W87∗	Both three- (preferred) and one-letter amino acid code may be used; ∗ accepted for one- and three-letter code
Deletion	p.Asp388_Gln393del	No specification of deleted amino acid(s)
Duplication	p.Asp388_Gln393dup	No specification of duplicated amino acid(s)
Insertion	p.Ala228_Val229insTrpPro p.Ala228_Val229insLys∗	Mandatory specification of inserted amino acids
Inversions		Not possible
Frame shift	p.(Arg97fs) p.(Arg97Profs∗23)	Short and long form accepted; long form contains "fsTer" or "fs∗"

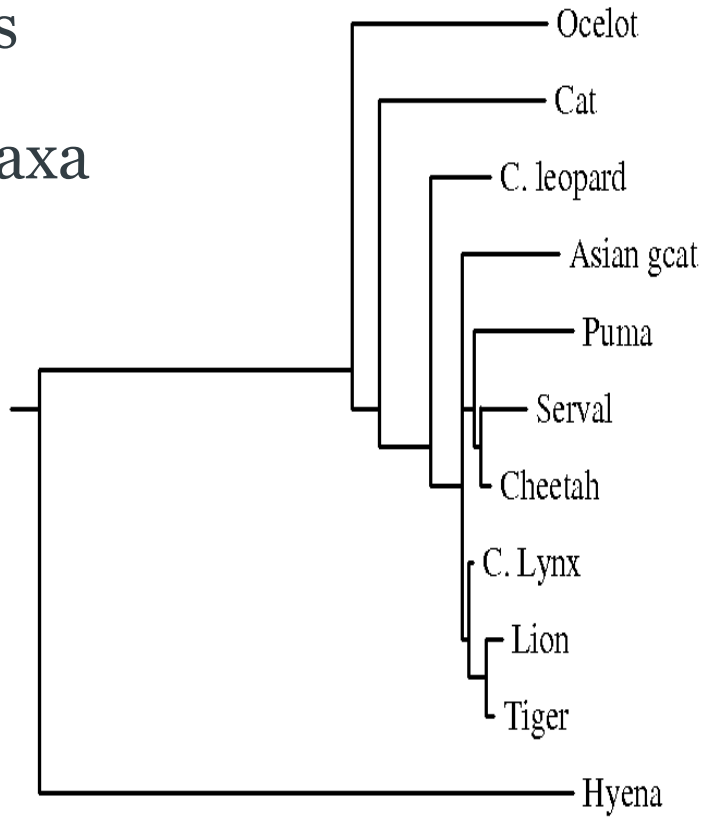
Allele frequencies

- Require baseline of human variance
  - Rare/novel variant ≠ pathogenic
  - Massive collaborations (and infrastructure) required
- 1000 Genomes – 2,504 WGS individuals globally ('healthy')
  - 843 authors; 168 affiliations
- Exome Sequencing Project – 6,515 Americans (EA and AA)
  - 396 authors;
- ExAC – 60,706 WES globally (common disease cohorts)
  - 74 authors + consortium; 52 affiliations

1000 Genomes Project Consortium 2015. Nature **526**:68-74. doi:10.1038/nature15393
Fu et al. 2013. Nature **493**:216-220. doi:10.1038/nature11690
Exome Aggregation Consortium et al. 2015. bioRxiv doi:10.1101/030338

Conservation

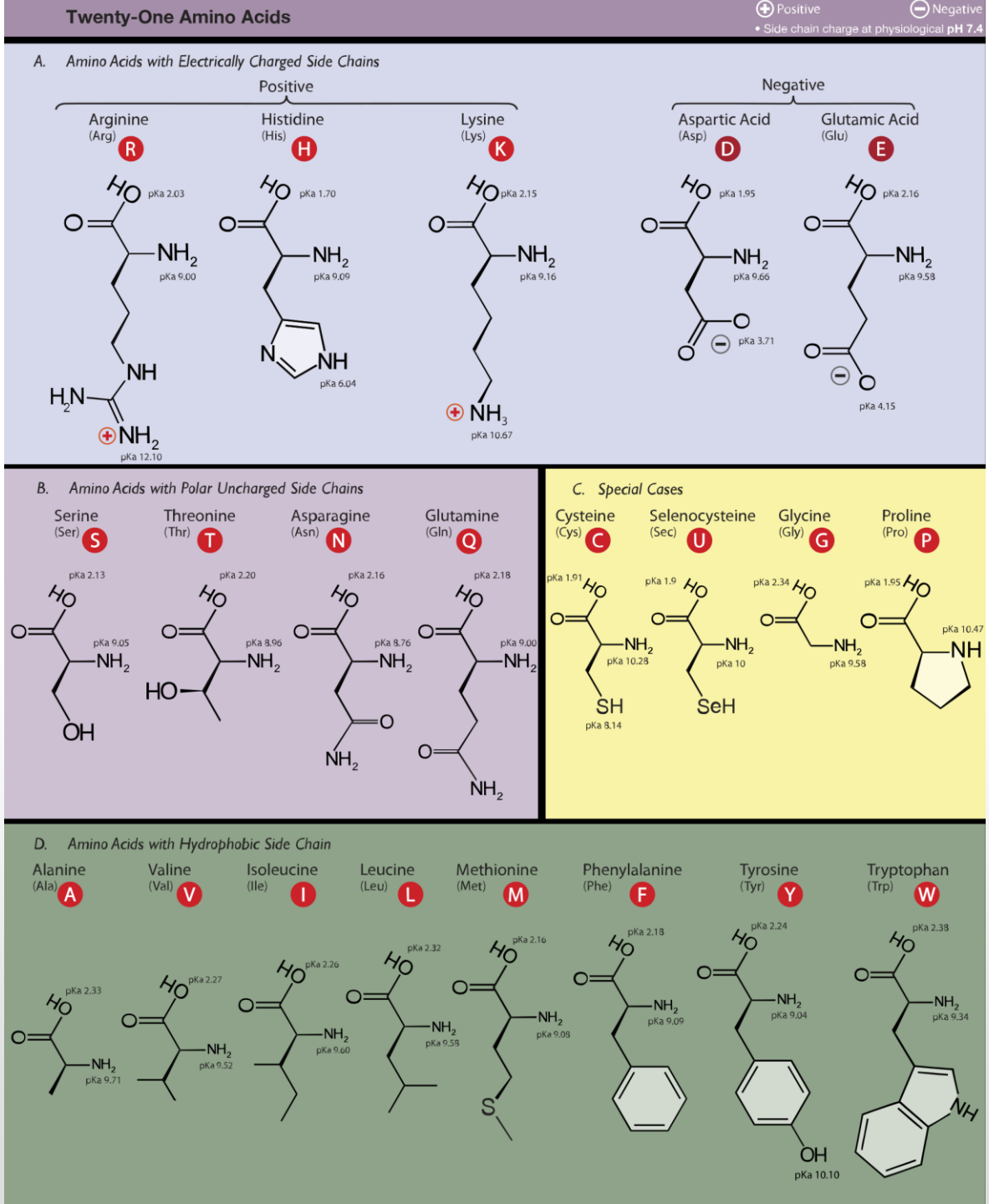
- Measure how much a base varies
- Based on sequence data across taxa
- Requires genome alignments
  - Non-trivial
- PhyloP is common score
- GERP++ more complex



- Highly conserved ➔ more damaging if changed?

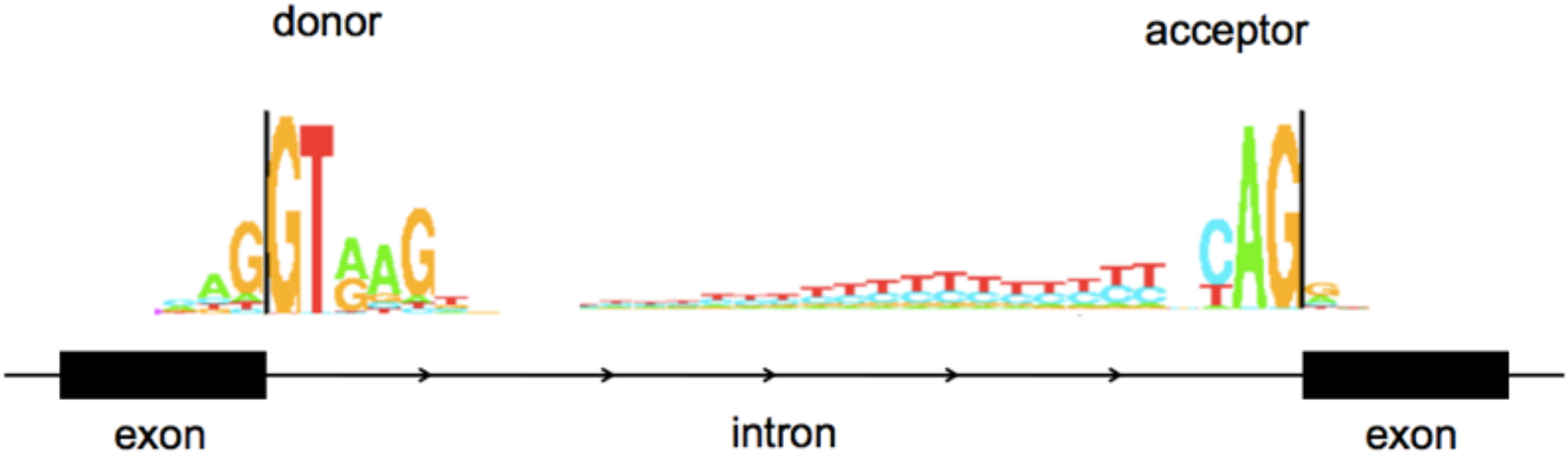
Functional effect scores

- Based on the difference a mutation may make to protein function
- Amino acid changes – Grantham scores
- Plus conservation information – SIFT scores
- Structure based – PolyPhen-2
  - Machine learning based on structural/biochemical features
- Consensus – CADD


Twenty-One Amino Acids
Positive
Negative
• Side chain charge at physiological pH 7.4
A. Amino Acids with Electrically Charged Side Chains
Positive
Arginine (Arg) R
Histidine (His) H
Lysine (Lys) K
Negative
Aspartic Acid (Asp) D
Glutamic Acid (Glu) E
B. Amino Acids with Polar Uncharged Side Chains
Serine (Ser) S
Threonine (Thr) T
Asparagine (Asn) N
Glutamine (Gln) Q
C. Special Cases
Cysteine (Cys) C
Selenocysteine (Sec) U
Glycine (Gly) G
Proline (Pro) P
D. Amino Acids with Hydrophobic Side Chain
Alanine (Ala) A
Valine (Val) V
Isoleucine (Ile) I
Leucine (Leu) L
Methionine (Met) M
Phenylalanine (Phe) F
Tyrosine (Tyr) Y
Tryptophan (Trp) W

Splicing

- Splicing mutations can have similar impact to frameshifts
- Complex process which is hard to predict
- Conservation of motifs at splice sites



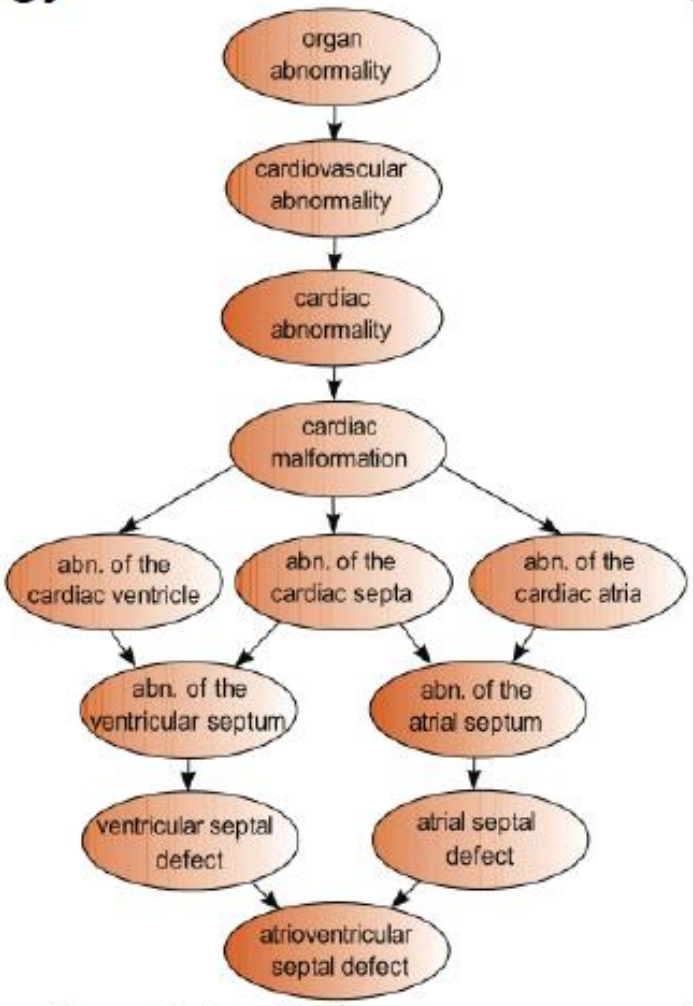
Disease databases

- Can use for gene and variant prioritisation
- Human Gene Mutation Database (HGMD)
  - Team of dedicated curators, industry funded
- Online Mendelian Inheritance in Man (OMIM)
  - Dedicated curators, public grant funded
- Leiden Open Variation Database (LOVD)
  - Infrastructure only, community implemented

Phenotype-led interrogation

- Can use systematic phenotyping for prioritisation
- Human phenotype ontology (HPO) provides a platform
- Link phenotypes to known causal genotypes in a network
- Create gene lists (Phenomizer, Phenotips)
- Prioritise variants (Exomiser, PhenIX)

Human Phenotype Ontology

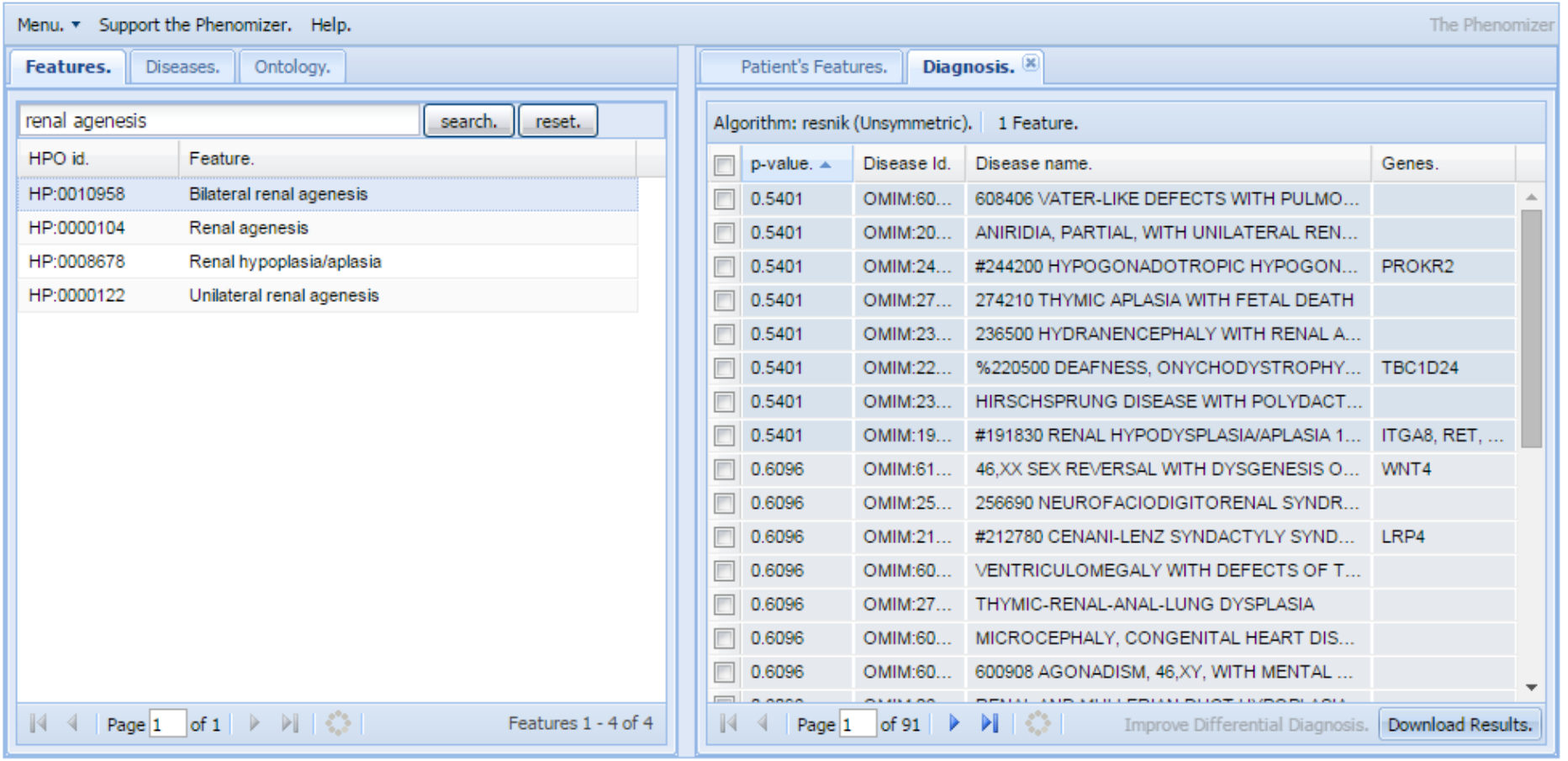


Atrioventricular septal defect

Robinson P, Köhler S, Bauer S, Seelow D, Horn D, Mundlos: The Human Phenotype Ontology: A Tool for annotating and analyzing human hereditary disease, *Am J Hum Genet. 2008 Nov*

Medicine **Phenomizer**

UNIVERSITY OF Southampton



*compbio.charite.de/**phenomizer**/*

Medicine

Phenotype defined target

UNIVERSITY OF Southampton

Narrowing diagnostic scope to phenotype-defined target

Unmasking exome data using phenotypically compatible genes

Advantages	Disadvantages
Individually defined for each patient, in accordance to presenting phenotype	Depend on the completeness of gene-phenotype mappings
Does not depend on the existence of **diagnostic hypothesis**	Depend on the completeness of describing patient phenotype presentation
Is **robust to diagnostic re-classification**	Generally result in larger gene target for analysis
Based on a **continuously updated** resource (HPO)	Moderate increase in the possibility of finding pertinent findings
Can be used a **standardized approach** for panel generation	

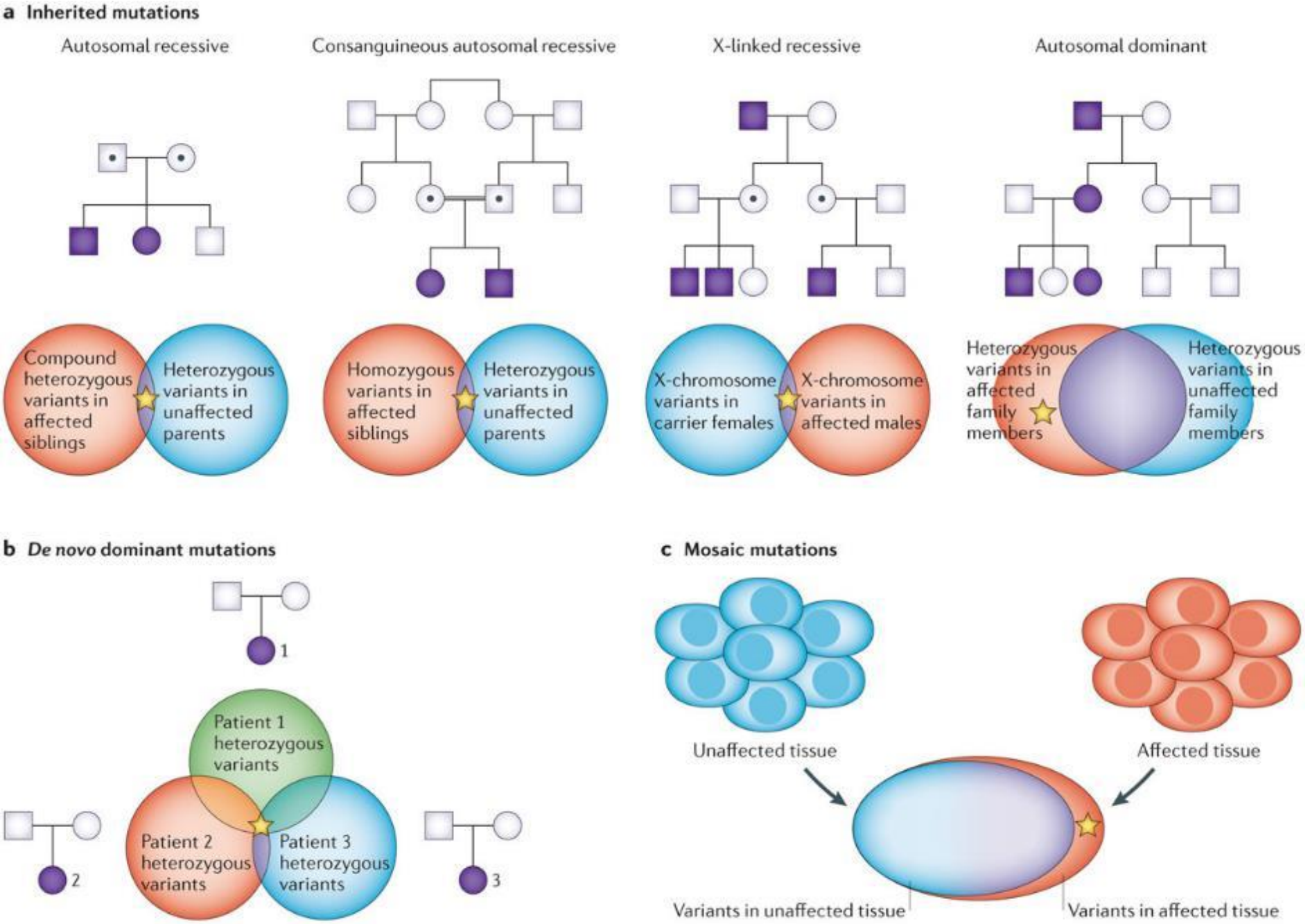
Variant shortlist

- Use Phenotype to restrict the candidates

- However, different resources give different phenotype to gene interactions

Renal Agenesis	
OMIM	RET, **ITGA8**, PAX2
Orphanet	FGF20, **ITGA8**
Congenital anomalies of the kidney and urinary tract (CAKUT) panel	BICC1, BMP4, CHD1L, EYA1, FOXC1, GATA3, GDNF, RET, ROBO2, SIX1, SIX5, SOX17, TFAP2A, TRAP1, UPK3A, WT1
Phenomizer	PROKR2, RET, PAX2, TBC1D24, LRP4,

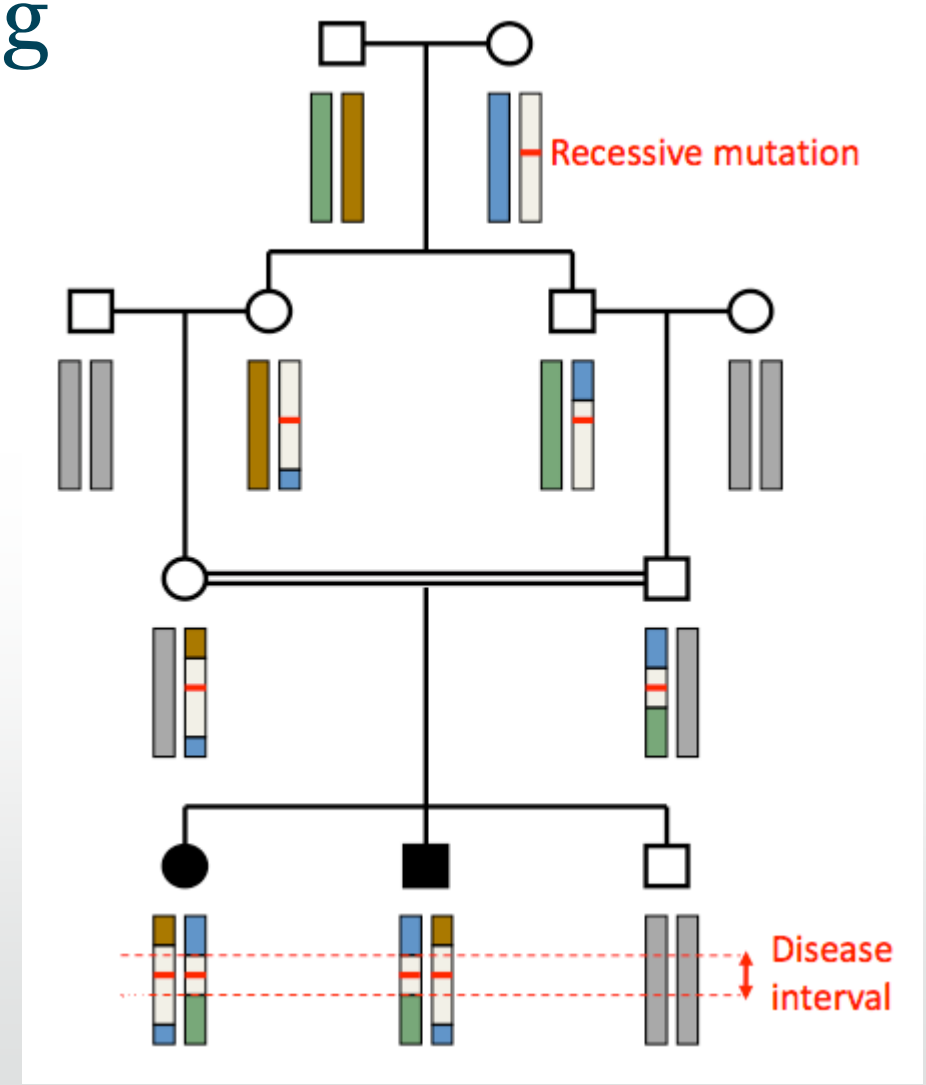
Gene	Diagnosis	Rank					
		PhenIX	Exomiser	Exomiser with CADD	OVA	eXtasy (order statistics)	eXtasy (combined max)
ARID1B	COFFIN-SIRIS SYNDROME; CSS;;FIFTH DIGIT SYNDROME	2	95	132	1037	6013	6184
KCNQ2	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 7; EIEE7	1	85	104	Not listed	1458	8508
SGCE	MYOCLONIC DYSTONIA	7	Not listed	Not listed	Not listed	239	9304
MED13L	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 15; MRT15	106	14	10	1004	2230	4511
RYR1	CONGENITAL FIBER-TYPE DISPROPORTION MYOPATHY	1	68	85	74	422	8624
SACS	SPASTIC ATAXIA, CHARLEVOIX-SAGUENAY TYPE	3	89	77	308	3264	5032
UBE3A	ANGELMAN SYNDROME	12	74	77	Not listed	178	8728
PTEN	PTEN HAMARTOMA TUMOR SYNDROME	1	1	1	Not listed	126	8822
DYNC1H1	SPINAL MUSCULAR ATROPHY, LOWER EXTREMITY, AUTOSOMAL DOMINANT; SMALED	10	85	86	20	1759	4687
SCN1A	DRAVET SYNDROME	2	27	53	72	250	8188
TCOF1	TREACHER COLLINS SYNDROME 3; TCS3;;MANDIBULOFACIAL DYSOSTOSIS, TREACHER COLLINS TYPE, AUTOSOMAL RECESSIVE	9	99	92	45	259	8858
OTX2	MICROPHTHALMIA, ISOLATED 1	5	60	70	73	Not listed	Not listed
EHMT1	KLEEFSTRA SYNDROME	10	88	95	Not listed	Not listed	Not listed
EFNB1	CRANIOFRONTONASAL SYNDROME; CFNS;;CRANIOFRONTONASAL DYSPLASIA; CFND;;CRANIOFRONTONASAL DYSOSTOSIS	1	1	1	Not listed	254	8997
HRAS	COSTELLO SYNDROME	7	1	1	52	1	9328
PTPN11	NOONAN SYNDROME 6; NS6	1	82	83	Not listed	1	9328
EIF2B1	LEUKOENCEPHALOPATHY WITH VANISHING WHITE MATTER; VWM	11	Not listed	144	Not listed	30	9216
FGFR3	MUENKE SYNDROME; MNKES	1	1	1	50	7	9281
POLG	ALPERS SYNDROME	1	89	98	402	14	8876
COMP	PSEUDOACHONDROPLASIA	1	78	90	53	10	9310



Boycott et al. 2013. Nat rev genet. **14**:681-691. doi:10.1038/nrg3555

Autozygosity mapping

- Useful in consanguineous scenarios
- Looking for regions of genome from the same recent ancestor (i.e. autozygous)
- Does not require trios
  - Identify tracts of homozygosity in proband



Medicine

wANNOVAR | Home | Tutorial | Example | Related projects

Basic Information

Email: r.j.pengelly@soton.ac.uk

Sample Identifier: test

Input File: + Input File | HYF3NBGXX_Filtered_Annotated.vcf

or Paste Variant Calls: paste your variant call here

OR

Submit | Reset | Monitor Progress

I agree to the Terms of Use

Disease/Phenotype

Enter Disease or Phenotype Terms: please enter your focused disease/phenotype terms

Please use semicolon or enter as separators. Like "alzheimer;brain".
Try to use multiple terms instead of a super long term
OMIM IDs are also accepted, like 114480 for 'Breast cancer'
Better Combined with wANNOVAR's disease model.

Parameter Settings

Result duration: 1 day

Reference Genome: hg19

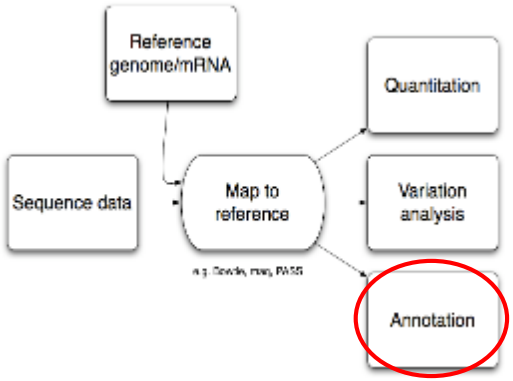
Input Fomat: VCF

Gene Definition: RefSeq Gene

Individual analysis: Individual analysis

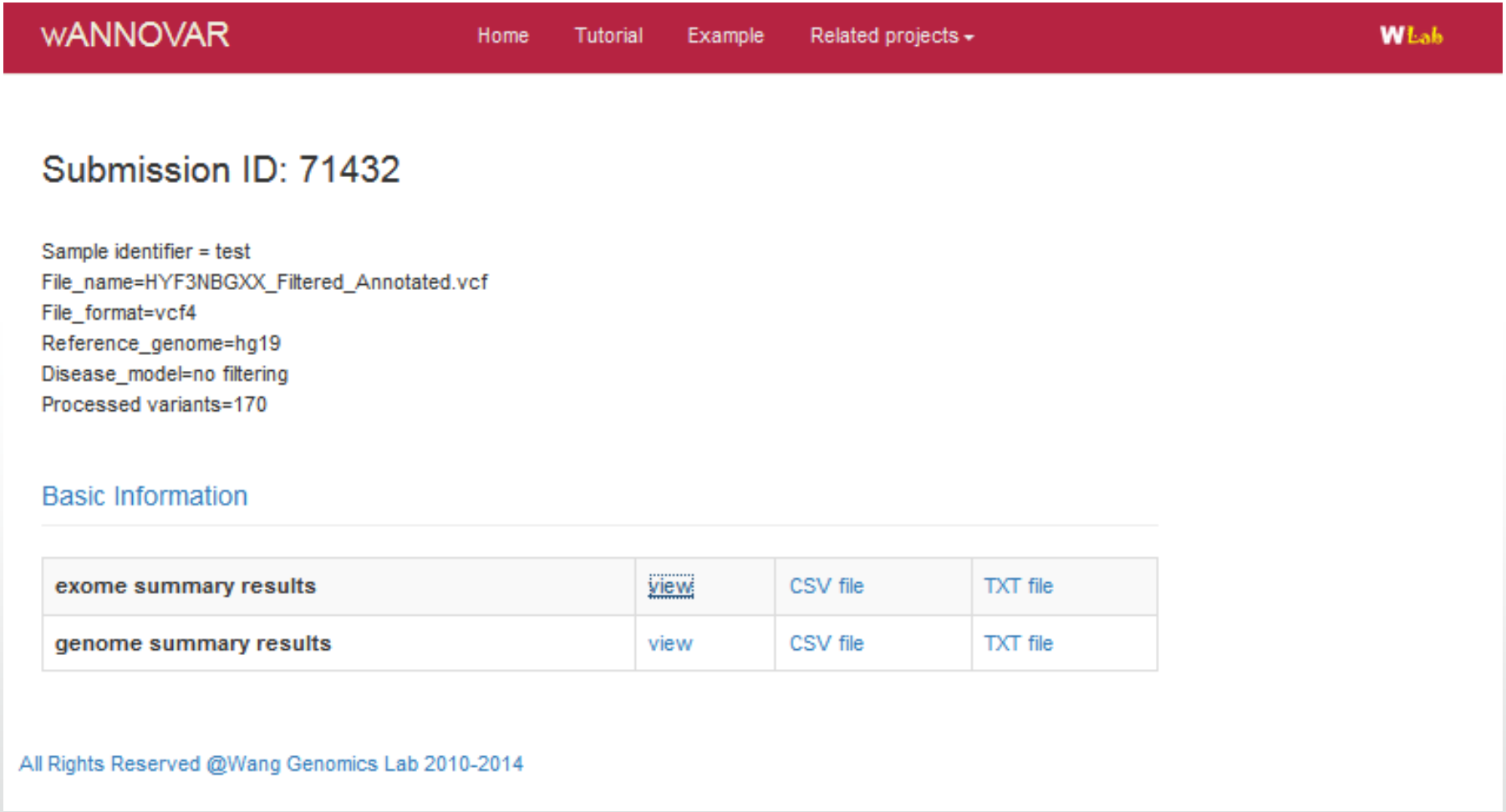
Disease Model: none

UNIVERSITY OF Southampton

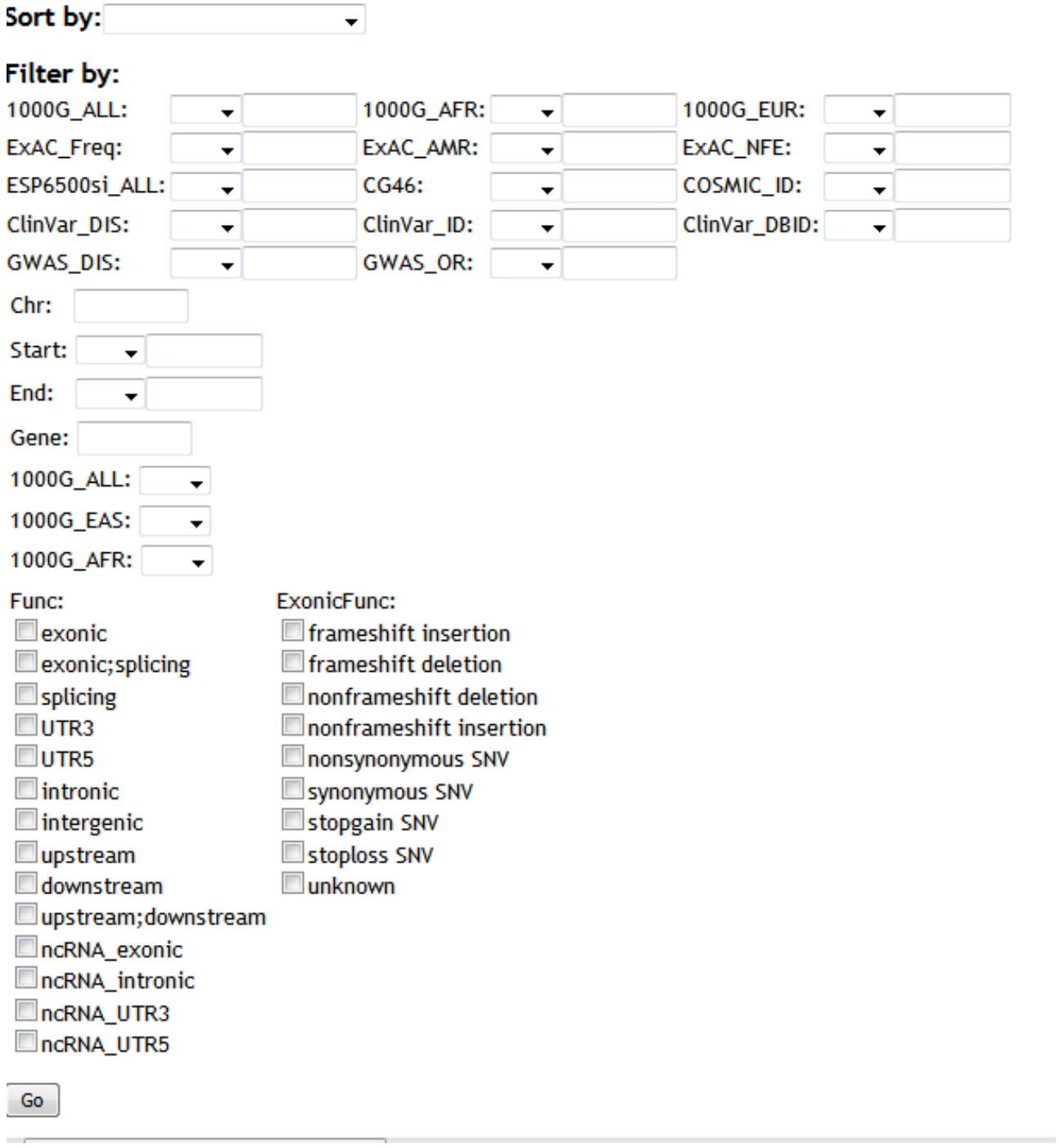


http://wannovar.usc.edu/

UNIVERSITY OF Southampton

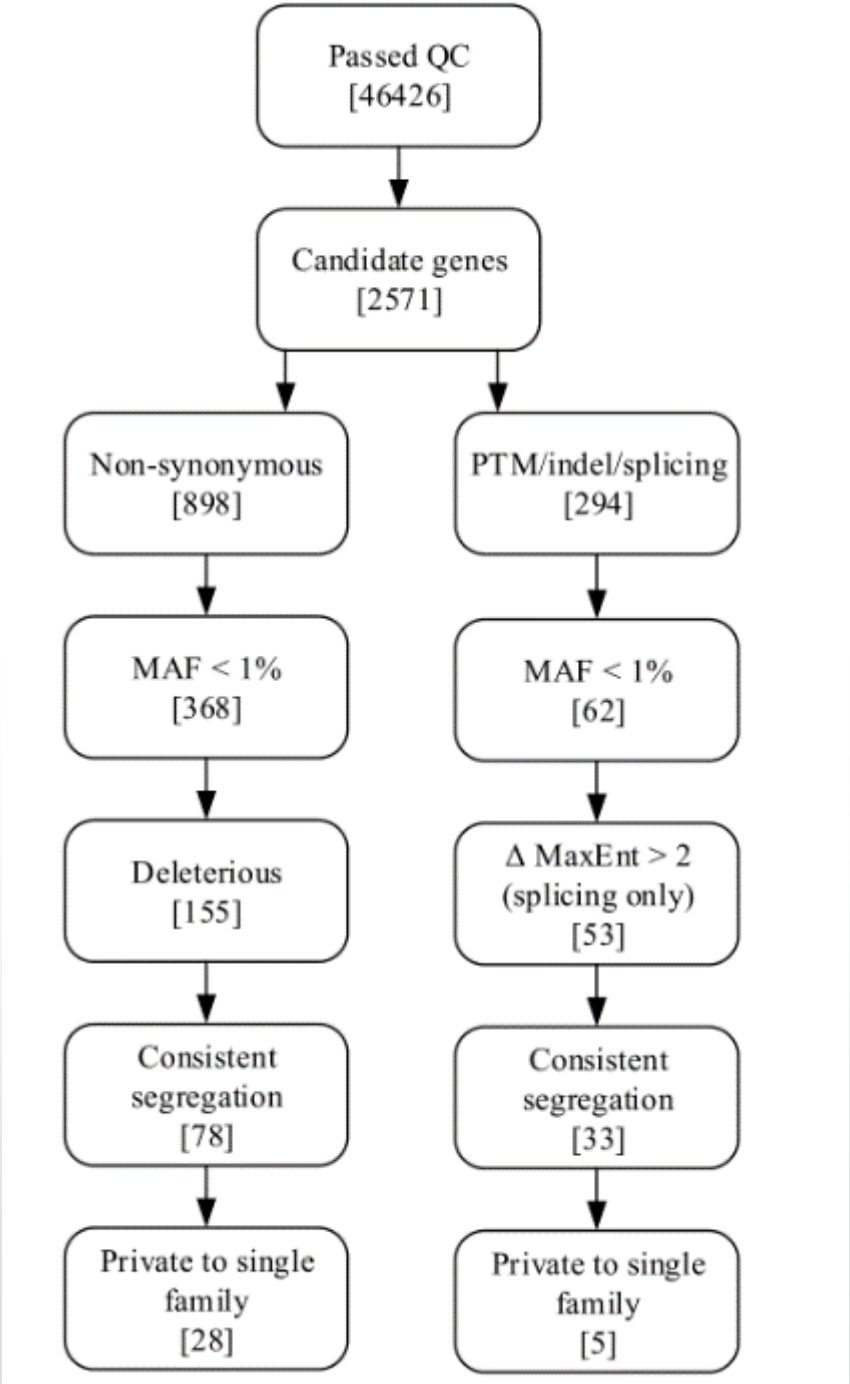


Func	Gene	ExonicFunc	AAChange	Conserved	SegDup	ESP5400 ALL	1000g2012feb ALL	dbSNP135	AVSIFT	LJB PhyloP	LJB PhyloP Pred	LJB SIFT	LJ SI Pr
exonic	SLC16A1	nonsynonymous SNV	NM_001166496:c.T1470A:p.D490E	13		0.666	0.66	rs1049434	1	0.862	N	0.22	T
exonic	LMNA	synonymous SNV	NM_005572:c.C51T:p.S17S	1075		0.00939	0.01	rs11549668					
exonic	LMNA	synonymous SNV	NM_001257374:c.G276A:p.L92L	369		0.00595	0.0037	rs12117552					
exonic	LMNA	synonymous SNV	NM_001257374:c.T525C:p.A175A			0.203	0.16	rs538089					
exonic	LMNA	synonymous SNV	NM_001257374:c.C631T:p.L211L	124									
exonic	LMNA	synonymous SNV	NM_001257374:c.T1002C:p.D334D	344		0.263	0.21	rs505058					
exonic	LMNA	synonymous SNV	NM_001257374:c.C1230T:p.C410C	142		0.00149	0.0014	rs149339264					
exonic;splicing	LMNA;LMNA	synonymous SNV	NM_001257374:c.C1362T:p.H454H	216		0.194	0.21	rs4641					
exonic	SLC19A2	synonymous SNV	NM_006996:c.G639A:p.K213K	260		0.00093	0.0023	rs137970656					
exonic	ENAH	synonymous SNV	NM_001008493:c.T1062C:p.P354P	55									
exonic	ENAH	synonymous SNV	NM_001008493:c.G759A:p.R253R	218		0.0354	0.03	rs1340868					
exonic	ENAH	nonsynonymous SNV	NM_001008493:c.G651T:p.E217D	82						0.5	N	0.89	T
exonic	ENAH	synonymous SNV	NM_001008493:c.G615A:p.E205E	82									
exonic	KLF11	nonsynonymous SNV	NM_001177716:c.A134G:p.Q45R	209		0.0941	0.06	rs35927125	0.11	0.245	N	0.64	T
exonic	KLF11	synonymous SNV	NM_001177716:c.A1134T:p.V378V	426		0.808	0.84	rs11687357					
exonic	ALMS1	synonymous SNV	NM_015120:c.G57A:p.E19E										
exonic	ALMS1	synonymous SNV	NM_015120:c.G60A:p.E20E			0.000892	0.01	rs183407241					
exonic	ALMS1	synonymous SNV	NM_015120:c.A75G:p.E25E					rs13009043					
exonic	ALMS1	synonymous SNV	NM_015120:c.G975A:p.S325S			0.000103							
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.C1174T:p.R392C			0.389	0.34	rs3813227	0.18	0.143	N	1.0	D
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.G1267A:p.V423I			0.00385	0.0005	rs45630557	0.98	0.805	N	1.0	D
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.A1868G:p.H623R			0.0194	0.01	rs41291187	0.12	0.138	N	0.96	D
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.T2012G:p.V671G			0.881	0.86	rs2037814	0.14	0.772	N	1.0	D
exonic	ALMS1	synonymous SNV	NM_015120:c.C2187T:p.F729F			0.51	0.54	rs7598901					
exonic	ALMS1	synonymous SNV	NM_015120:c.C2532T:p.D844D			0.0331	0.04	rs77517267					
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.C3304G:p.P1102A			0.000519			0.1	0.853	N	0.94	T
exonic	ALMS1	synonymous SNV	NM_015120:c.A3891G:p.Q1297Q			0.0592	0.05	rs112034360					
exonic	ALMS1	synonymous SNV	NM_015120:c.A4176G:p.Q1392Q			0.365	0.32	rs6546836					
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.G4241C:p.G1414A			0.389	0.34	rs6546837	1	0.036	N	0.92	T
exonic	ALMS1	synonymous SNV	NM_015120:c.G4956A:p.Q1652Q										
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.A5356G:p.N1786D			0.0115	0.01	rs45608038	0.35	0.000859	N	0.94	T
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.A5623G:p.I1875V			0.386	0.33	rs6546838	1	0.0262	N	0.76	T
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.C6122T:p.T2041I						0	0.984	C	0.99	D
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.T6209C:p.I2070T			0.137	0.10	rs10496192	0	0.0271	N	0.9	T
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.C6299T:p.S2100L			0.0232	0.02	rs28730854	0.02	0.981	C	0.99	D
exonic	ALMS1	nonsynonymous SNV	NM_015120:c.T6333A:p.S2111R			0.39	0.34	rs6724782	1	0.0136	N	0.99	D


Sort by:
Filter by:
1000G_ALL:
1000G_AFR:
1000G_EUR:
ExAC_Freq:
ExAC_AMR:
ExAC_NFE:
ESP6500si_ALL:
CG46:
COSMIC_ID:
ClinVar_DIS:
ClinVar_ID:
ClinVar_DBID:
GWAS_DIS:
GWAS_OR:
Chr:
Start:
End:
Gene:
1000G_ALL:
1000G_EAS:
1000G_AFR:
Func:
exonic
exonic;splicing
splicing
UTR3
UTR5
intronic
intergenic
upstream
downstream
upstream;downstream
ncRNA_exonic
ncRNA_intronic
ncRNA_UTR3
ncRNA_UTR5
ExonicFunc:
frameshift insertion
frameshift deletion
nonframeshift deletion
nonframeshift insertion
nonsynonymous SNV
synonymous SNV
stopgain SNV
stoploss SNV
unknown
Go

UNIVERSITY OF Southampton

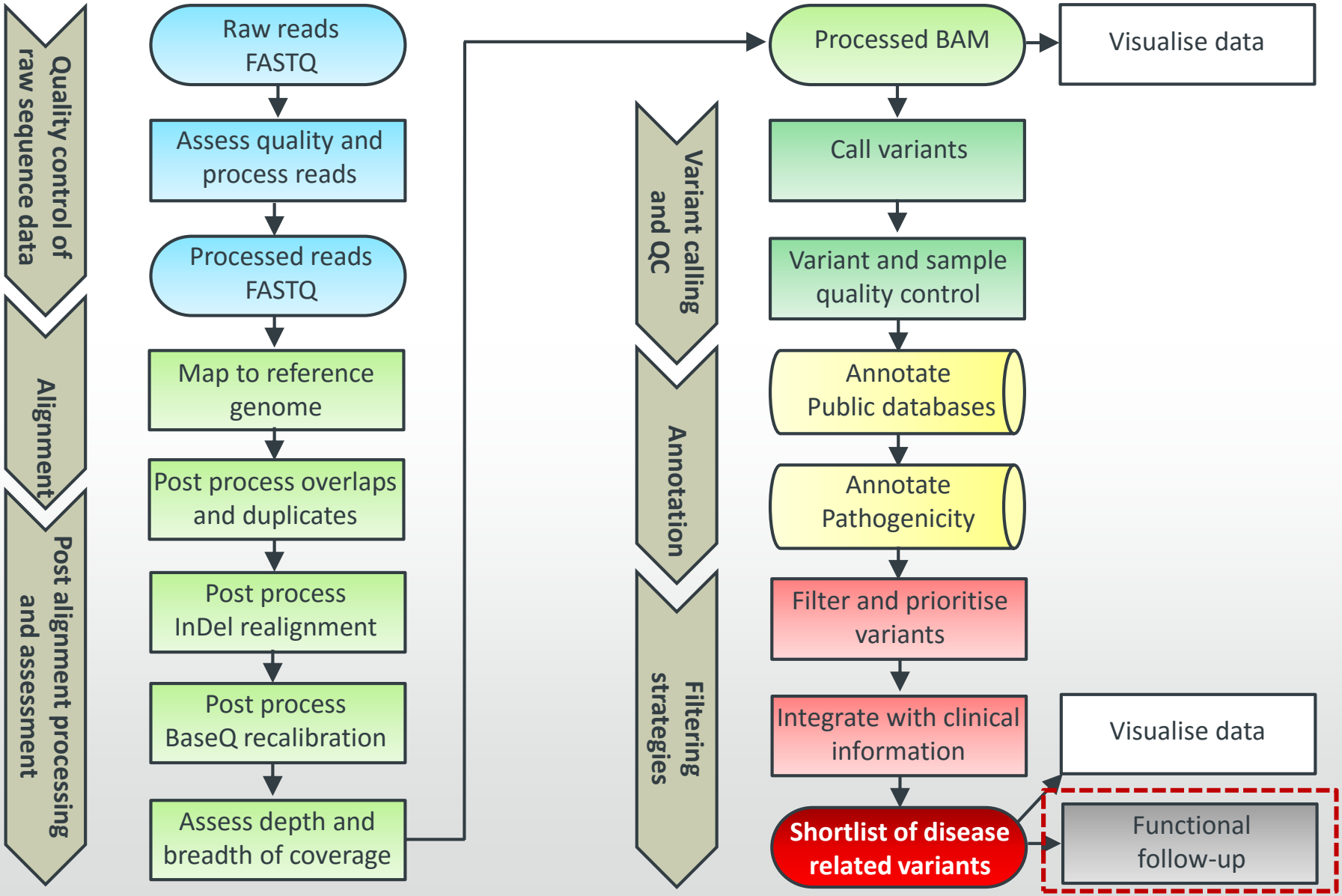
Filtering



Variant assessment

- Does variant look real?
- Is variant in gene associated with phenotype?
  - Predictive phenotyping
- Is it a known pathogenic variant?
- Does it alter the protein in a way reported to be damaging?
- Is the variant rare (not just in patient's population)?
- Is the variant predicted to be deleterious (multiple scores)?
- Is the variant being pathogenic biologically plausible?

Analysis workflow



UNIVERSITY OF Southampton

Follow up

- Literature review
- Animal models
- Biochemical assays
- Cell based assays
- Patient cohorts
- *In silico* modelling

Summary

- Filtering is required to produce usable variant shortlist
- Allele frequencies are an effective filter
- Predictive scores are informative for prioritisation
- Segregation is a powerful tool
- Phenotype led approaches can be helpful, but a work in progress