A systematic approach for the construction of benign missense variant "truthsets" for validation of functional assays

Charlie F. Rowlands[1], Sophie Allen[1], Alice Garrett[1,2], Miranda Durkie[3], George J. Burghel[4,5], Rachel Robinson[6], Alison Callaway[7], Joanne Field[8], Bethan Frugtniet[2], Sheila Palmer-Smith[9], Jonathan Grant[10], Judith Pagan[11], Trudi McDevitt[12], Katie Snape[2], Helen Hanson[13,14], Terri McVeigh[15], Clare Turnbull[1,15] and CanVIG-UK

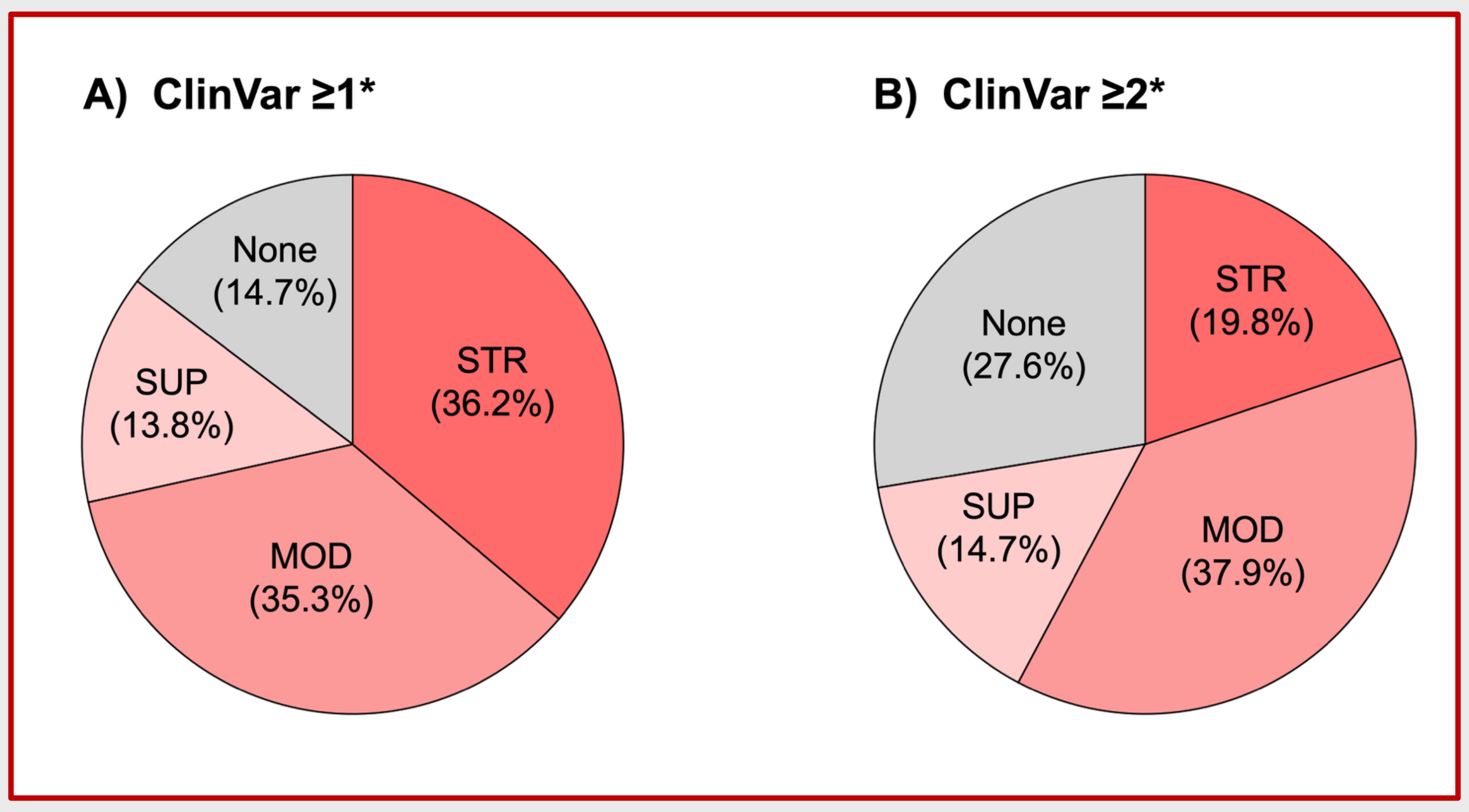
Introduction and aims

- Following ClinGen SVI guidance by Brnich et al., experimental functional assays may provide evidence towards pathogenicity (PS3) or benignity (BS3) based on their ability to **discriminate between validation variant sets** of clinical **pathogenic** and **benign** variants
- **PS3** evidence is primarily driven by the number of **benign truthset variants**
- For loss-of-function genes, most clinically classified **pathogenic variants are truncating**, and most **benign variants are synonymous**
  - These are also the most common variant types used for assay threshold calibration
- However, most clinical **VUSs** are **missense variants**
  - Missense variant impact is more likely to lie away from the extreme ends of assay readout
  - Accurate quantification of assay performance should be conducted using **validation variant sets** comprising **missense variants**
  - Clinically classified **missense variants are scarce** for most genes of clinical interest
- We thus sought to:
  - Quantify the **availability of clinically classified missense variants** in key genes of clinical interest
  - Develop a **novel approach** for construction of **benign missense truthsets** in the absence of existing classifications

Quantification of paucity of classified missense variants

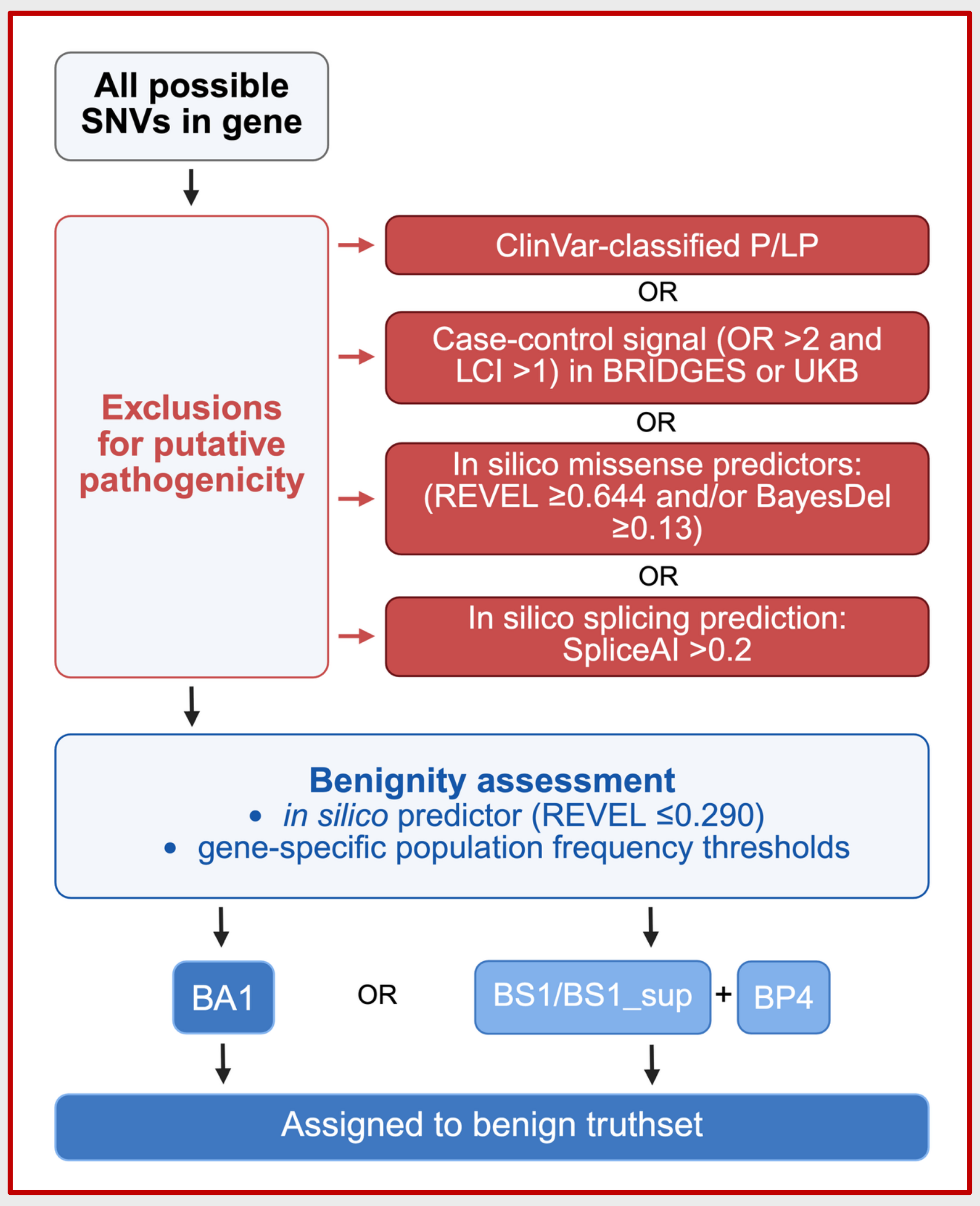
- We collated available ClinVar **benign missense classifications** (≥1* and ≥2*) in 116 inherited cancer susceptibility genes (CSGs)
- We quantified the **maximum PS3 strength attainable** if using these variants as a truthset
  - In this analysis we assumed **perfect concordance** of truthset classifications with assay output and that **all identified variants were present** in assay output.

- We observed that these truthsets were sufficiently large to allow application of **PS3 at a maximum strength of "strong"** for:
  - **42/116 (36.2%)** of genes when based on ClinVar **≥1* classifications**
  - Only **23/116 (19.8%)** when restricting to the higher-confidence **≥2* classifications**.



- **Exclusion of variants** from truthsets if experimental **functional evidence was used** in their classification is vital to **avoid circularity**
- Comments accompanying ClinVar benign missense submissions were examined more closely for eight hereditary breast and ovarian cancer genes of interest (*BRCA1, BRCA2, PALB2, ATM, CHEK2, RAD51C, RAD51D* and *BRIP1*)
- Of the 113 variants in the non-*BRCA* genes, for only **45 (39.8%)** was it possible to ascertain that the classification was **not reliant on functional data**

Additional copies


All possible SNVs in gene

Exclusions for putative pathogenicity

ClinVar-classified P/LP

OR

Case-control signal (OR >2 and LCI >1) in BRIDGES or UKB

OR

In silico missense predictors: (REVEL ≥0.644 and/or BayesDel ≥0.13)

OR

In silico splicing prediction: SpliceAI >0.2

Benignity assessment
- *in silico* predictor (REVEL ≤0.290)
- gene-specific population frequency thresholds

BA1 OR BS1/BS1_sup + BP4

Assigned to benign truthset

We devised a gene-specific **systematic approach** for generation of benign missense truthsets, as **benign truthset size** is the primary **driver of applicable PS3 strength**.

All possible missense variants in a gene of interest are filtered to remove any with indicators of potential pathogenicity based on existing **classifications**, **case-control** signal or ***in silico* predictions**.

Remaining variants can then be assigned to benign truthsets if their **population frequency** exceeds gene-specific thresholds and have ***in silico* scores** indicative of benignity, in accordance with ACMG/AMP combining criteria for benignity.

Systematic construction of benign missense truthsets for eight cancer susceptibility genes

- We derived gene-specific population frequency thresholds and used our systematic approach to construct benign missense validation variant sets for our eight selected hereditary breast and ovarian cancer (HBOC) genes of clinical interest.

	ClinVar ≥1*			ClinVar ≥2*			Systematic generation		
Gene	# vars.	PS3 EPs	PS3 strength	# vars.	PS3 EPs	PS3 strength	# vars.	PS3 EPs	PS3 strength
BRCA1	302	7.80	**STR**	183	7.11	**STR**	37	4.93	**STR**
BRCA2	320	7.88	**STR**	201	7.24	**STR**	190	7.16	**STR**
PALB2	25	4.40	**STR**	11	3.27	**MOD**	106	6.37	**STR**
ATM	51	5.37	**STR**	25	4.40	**STR**	159	6.92	**STR**
CHEK2	5	2.20	**MOD**	1	0	**None**	13	3.50	**MOD**
RAD51C	4	1.89	**SUP**	2	0.95	**None**	38	4.97	**STR**
RAD51D	4	1.89	**SUP**	3	1.50	**SUP**	37	4.93	**STR**
BRIP1	24	4.34	**STR**	5	2.20	**MOD**	44	5.17	**STR**

- For all eight genes, systematically generated truthsets generated hypothetical maximum PS3 evidence strengths that were greater than or equal to those generated by ClinVar-based truthsets alone.

- For all genes except *BRCA1* and *BRCA2*, the raw number of variants generated through systematic benign missense truthset construction was larger than any ClinVar-based truthset
  - This reflects the longer timecourse of *BRCA* gene testing and the existence of highly evolved *BRCA*-specific interpretation guidance

Conclusions

- Non-VUS ClinVar missense variant classifications remain scarce and poorly annotated for most genes, limiting the utility of ClinVar as a source of truthset variants for quantification of PS3 strength

- We have devised a gene-specific approach to systematic identification of benign truthset variants that generates truthsets allowing application of PS3 at equal or greater strength than ClinVar-based truthsets across all genes investigated.

Contact

Charlie Rowlands, PhD
Division of Genetics and Epidemiology
The Institute of Cancer Research
15 Cotswold Road, Sutton, London, UK
charlie.rowlands@icr.ac.uk

Author affiliations

1. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK; 2. St George's University Hospitals NHS Foundation Trust, Tooting, London, UK; 3. Sheffield Diagnostic Genetics Service, NEY Genomic Laboratory Hub, Sheffield Children's NHS Foundation Trust, Sheffield, UK; 4. Manchester Centre for Genomic Medicine and NW Laboratory Genetics Hub, Manchester University Hospitals NHS Foundation Trust, Manchester, UK; 5. Division of Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK; 6. Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Leeds, UK; 7. Central and South Genomics Laboratory Hub, Wessex Genomics Laboratory Service, University Hospital Southampton NHS Foundation Trust, Salisbury, UK; 8. Genomics and Molecular Medicine Service, Nottingham University Hospitals NHS Trust, Nottingham, UK; 9. Wales Genomic Health Centre, Cardiff and Vale University Health Board, Cardiff, UK; 10. Laboratory Genetics, Queen Elizabeth University Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK; 11. South East Scotland Clinical Genetics, Western General Hospital, Edinburgh, UK; 12. Department of Clinical Genetics, CHI at Crumlin, Dublin, Ireland; 13. Peninsula Regional Genetics Service, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK; 14. Department of Clinical and Biomedical Sciences, University of Exeter Medical School, Exeter, United Kingdom; 15. The Royal Marsden NHS Foundation Trust, Fulham Road, London, UK