

Supplementary Data

Variant Annotation

There are a number of stages in which variants are annotated. VEP (McLaren *et al.*, 2016) and GEMINI (Paila *et al.*, 2013) annotate variants with gene, transcript and impact severity, as well as allele frequencies from dbSNP, ExAC (Lek *et al.*, 2016), 1000 Genomes (1000 Genomes Project Consortium *et al.*, 2015), and EVS (Tennessen *et al.*, 2012), as well as *in silico* predictions from PolyPhen-2 (Adzhubei *et al.*, 2010), SIFT (Kumar *et al.*, 2009) and CADD (Kircher *et al.*, 2014). Additional SNV and Indel annotations are managed by Seave; variant allele frequencies in healthy controls: MGRB [<https://sgc.garvan.org.au/initiatives/mgrb>] (Lacaze *et al.*, 2018; McNeil *et al.*, 2017; 45 and Up Study Collaborators *et al.*, 2008); diseases: ClinVar (Landrum *et al.*, 2013), MITOMAP (Ruiz-Pesini *et al.*, 2007), COSMIC (Forbes *et al.*, 2010, 2015); links to phenotypes and disorders: OMIM (Amberger *et al.*, 2015), COSMIC Cancer Gene Census (CGC) (Futreal *et al.*, 2004), Orphanet [<http://www.orpha.net>] (Orphanet, 2017), and Genomics England PanelApp [<https://panelapp.genomicsengland.co.uk>]; and pre-computed *in silico* annotations from RVIS (Petrovski *et al.*, 2013), and dbNSFP (Liu *et al.*, 2013), which provides PROVEAN (Choi *et al.*, 2012), FATHMM (Shihab *et al.*, 2014), MetaSVM/MetaLR (Dong *et al.*, 2015), and GERP++ (Davydov *et al.*, 2010). To keep these annotations up-to-date, we provide tools for downloading and updating many of these resources.

Supplementary Figures



Pull up a chair and **grab some data**.

First, you need to select some data in a database. Click the database row you would like to query. Databases with pedigree information can utilise advanced queries and it is recommended you add this information to your databases.



Available databases

Show 25 entries

Database	Group	Sample Names	Samples	Variants	Size	Date	GEMINI	Actions
AshkenaziTrio.Oslo.hc.joint.vqsr.vcp	Public	HG002;HG003;HG004	3	98303	685.28 MB	04/12/2015	v0.11.0	
NA12878trio.hc.vqsr.decomposed.nc	Public	NA12878;NA12891;NA12893	3	6474526	19.35 GB	14/07/2017	v0.18.3	

Showing 1 to 2 of 2 entries Previous 1 Next

Seave is running GEMINI version 0.19.1.

To see private databases, you need to log in.

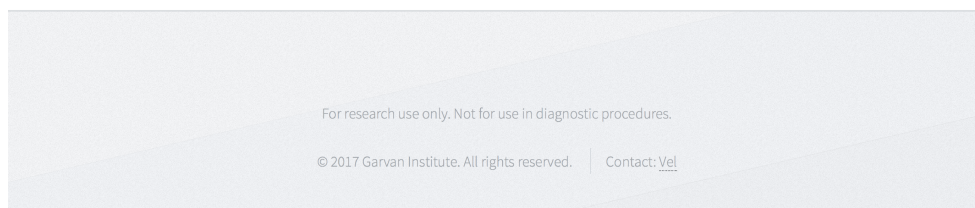



Figure S1: Seave screenshot: database selection page. This page shows the available databases for query and information about them, including the number of samples and variants. Upon logging in, this page displays databases from all groups the user has access to.

Databases
Familial Filters
SEAVE
Data Sources
Log In

Your database contains **families**.

You can choose to use familial information to conduct variant filtration on members of a single family using predefined analysis methods. Alternatively, you can choose to analyse the entire dataset.



Database selected
NA128789.hc.vqsr.decomposed.normalised.vep.db

Select a family to analyse

Family information
NA12878 (Female) - Affected
NA12891 (Male) - Unaffected
NA12892 (Female) - Unaffected
Please ensure this information is correct before proceeding.


Select an analysis type

None Homozygous/Recessive Heterozygous Dominant Compound Heterozygous

Familial filtering


Click each of the headings below if you would like more information regarding the filtration mechanism and for different example familial scenarios.

Heterozygous dominant



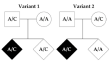
All affected individuals have a heterozygous genotype and all unaffected individuals do not have a heterozygous genotype. Equivalent to autosomal dominant in the autosome and X-linked dominant in the X chromosome.

Homozygous/hemizygous recessive




All affected individuals have a homozygous alternate genotype and all unaffected individuals do not have a homozygous alternate genotype. Equivalent to autosomal recessive in the autosome and X-linked recessive in the X chromosome.

Compound heterozygous



Affected individuals all share a heterozygous genotype and for one position in a gene one unaffected individual shares this heterozygous genotype with the affecteds and in another position another unaffected individual shares the heterozygous genotype with the converse not being true.

De novo dominant



All affected individuals share a heterozygous variant and all unaffected individuals either share a homozygous reference or homozygous alternate genotype.

None

The none analysis type returns all variants in the database where at least one of the samples in the family selected has a variant.

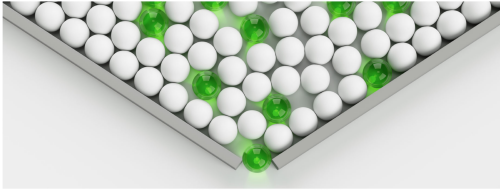
For research use only. Not for use in diagnostic procedures.
© 2017 Garvan Institute. All rights reserved. [Contact Us](#)

Figure S2: Seave screenshot: family and analysis selection page. After clicking a database to query, this page optionally allows selecting a family within it to query. If a family is selected, further options to select an analysis type (i.e. inheritance pattern) appear.

[Databases](#)
[Familial Filters](#)
SEAVE
[Data Sources](#)
[Log In](#)

OK, you have some data. Now filter it.

Select from the filtration options below.



Database selected
NA128781to_hc.vqsr.decomposed.normalised.vep.vcf

Family selected
NA128781to

Inclusion genomic location(s)
Search region(s)
e.g. chr2:15483-25583;chr1:37211-67824;chr5:MT
Separate multiple regions to search with a semicolon. To search all regions, leave this box blank. Any genes specified will be restricted to these coordinates.

Search gene list(s)
ACMG 56 genes (56)
ACMG cancer genes (AD only) (22)
ACMG cancer genes (AR + AD) (28)
ACMG cancer genes (AR only) (1)
Arrhythmic_Syndromes_Aug_2015_Fatkin (4)
Clear

Search custom gene list
e.g. BRCA1,PIK3CA,TP53
Separate multiple genes with a semicolon, comma or space. To search all genes, leave this box blank.

Impact
Restrict variants by impact
 Loss of Function
 High Impact
 High & Moderate Impact
 Coding
 All Impact

Minimum scaled CADD score

All variants without CADD scores are returned. For no minimum scaled CADD score, set this value to 0.

Quality
Minimum sequencing depth in all samples selected

For no minimum sequencing depth, set this value to 0.
 Minimum variant quality

For no minimum variant quality, set this value to 0.

Exclude failed variants

Exclusion genomic location(s)
Exclude region(s)
e.g. chr2:15483-25583;chr1:37211-67824;chr5:MT
Separate multiple regions to exclude with a semicolon. To search all regions, leave this box blank. Any genes specified will be restricted to these coordinates.

Exclude gene list(s)
ACMG 56 genes (56)
ACMG cancer genes (AD only) (22)
ACMG cancer genes (AR + AD) (28)
ACMG cancer genes (AR only) (1)
Arrhythmic_Syndromes_Aug_2015_Fatkin (4)
Clear

Exclude custom gene list
e.g. BRCA1,PIK3CA,TP53
Separate multiple genes with a semicolon, comma or space. To not include any genes, leave this box blank.

Prevalence
Frequency in control databases
 1000 Genomes
 1%
 ESP
 1%
 ExAC
 1%
Variants will be returned that are either below the allele frequency set or not present in the database. For no minimum allele frequency, set the value to 0%.

Exclude dbSNP Common Exclude dbSNP Flagged

Variant type(s)
 SNPs
 INDELS
 Both

Maximum number of variants to return

For research use only. Not for use in diagnostic procedures.
© 2017 Garvan Institute. All rights reserved. [Contact Us](#)

Figure S3: Seave screenshot: filtration options/query page. The query page allows the user to specify filtration options to restrict the number of variants returned. Restrictions can be by genomic location, impact on genes, prevalence in control populations and sequencing quality.

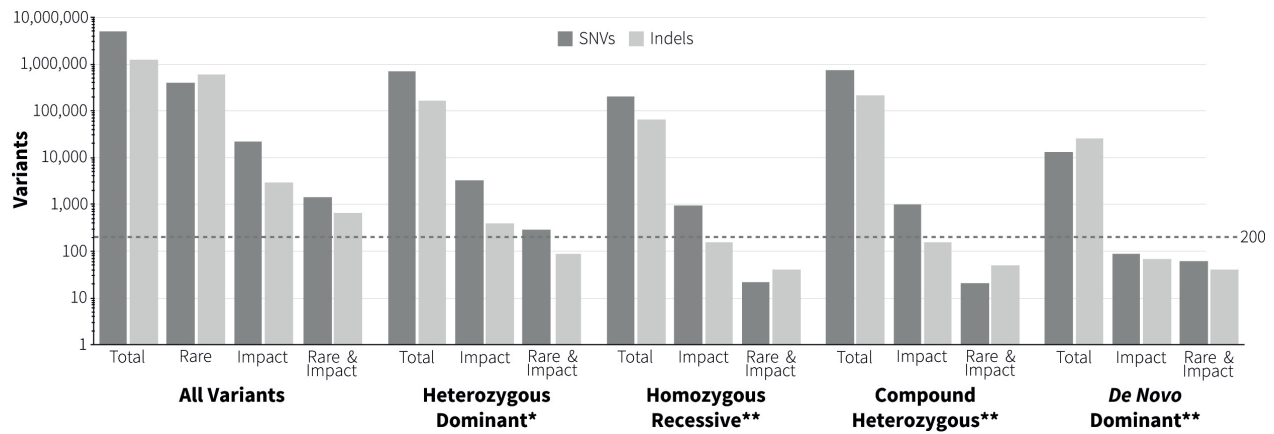


Figure S4: Variant counts from whole genome sequencing in the NA12878 trio restricted by combinations of rarity, gene impact (damaging) and inheritance patterns. Counts derived using the best practices GATK pipeline on raw data from the Illumina Platinum Genomes project (Eberle *et al.*, 2017), mapped to the b37+decoy reference genome, decomposed and normalised with vt (Tan *et al.*, 2015) and queried with GEMINI (Paila *et al.*, 2013). Rarity is defined as a maximum allele frequency of 1% in 1000 Genomes, ExAC and ESP. Impact is defined as medium or high impact, as defined by the Ensembl impact variant annotation. *NA12878 and NA12891 were marked as affected for the purposes of this analysis. **NA12878 was marked as affected for the purposes of this analysis.

Great. It's time for some **results**.

The table below displays your variants. *Click any row* to fetch all GEMINI information for that variant in a separate table.

Show: **10** entries

Search:

Variant	Quality	Gene	Type	Impact	KCCG Exomes AF	KCCG Genomes AF	ClinVar Variation ID	ClinVar Clinical Significance	ClinVar Trait	Impact Summary
chr14.g.50088557C>G	1037.13000488	MGAT2	SNP	missense_variant	0	0	313257	Uncertain significance	Congenital disorder of	
chr9.g.2096706A>T	744.130004883	SMARCA2	SNP	missense_variant	0	0	366215	Likely benign	Nicotinoides-Baraitser syn	
chr1.g.89660991C>G	548.130004883	GBP4	SNP	missense_variant	0	0	No Result	No Result	No Result	
chr1.g.23116181T>TAA	252.949996948	FAM89A	insertion	splice_region_variant	0	0	No Result	No Result	No Result	
chr3.g.183882362C>G	700.130004883	DVL3	SNP	missense_variant	0	0	No Result	No Result	No Result	
chr4.g.119259448C>T	444.130004883	PRSS12	SNP	missense_variant	0	0	No Result	No Result	No Result	
chr5.g.27028494CA>C	333.140014648	CDH9	Deletion	splice_region_variant	0	0.16 (25/160)	No Result	No Result	No Result	
chr6.g.27115124GACA>G	646.130004883	HIST1H2AH	Deletion	inframe_deletion	0	0	No Result	No Result	No Result	
chr6.g.29835838C>CCA	3246.89990234	HLA-K	insertion	splice_region_variant	0	0.15 (23/152)	No Result	No Result	No Result	
chr6.g.128505804A>C	568.130004883	PTPRK	SNP	missense_variant	0	0	No Result	No Result	No Result	

Showing 1 to 10 of 33 entries

Previous 1 2 3 4 Next

[Download query results \(.tsv format\)](#)
Increase/decrease table width
Show/hide GEMINI query

Figure S6: Seave screenshot: results table expanded with additional columns. The variants table on the results page can be expanded to show more annotation information. Any overflowing information can be read by hovering over the table cell and reading the tooltip that appears, as shown in this screenshot.

References

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., Abecasis, G. R., Gibbs, R. A., Wang, J., Li, Y., Gupta, N., Lander, E. S., Eichler, E. E., Alkan, C., Banks, E., Li, H., Barnes, B., Shaw, R., Kidd, J. M., Hormozdiari, F., Hurles, M. E., Abyzov, A., Gabriel, S. B., MacArthur, D. G., Gravel, S., Homer, N., Flicek, P., Montgomery, S. B., Altshuler, D. M., and Ruiz-Linares, A. (2015). A global reference for human genetic variation. *Nature Publishing Group*, **526**(7571), 68–74.
- 45 and Up Study Collaborators, Banks, E., Redman, S., Jorm, L., Armstrong, B., Bauman, A., Beard, J., Beral, V., Byles, J., Corbett, S., Cumming, R., Harris, M., Sitas, F., Smith, W., Taylor, L., Wutzke, S., and Lujic, S. (2008). Cohort profile: the 45 and up study. *International journal of epidemiology*, **37**(5), 941–947.
- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, **7**(4), 248–249.
- Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, **43**(D1), D789–D798.
- Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, **7**(10), e46688.
- Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, **6**(12), e1001025.
- Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics*, **24**(8), 2125–2137.
- Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., and Bentley, D. R. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, **27**(1), 157–164.
- Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R., and Futreal, P. A. (2010). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, **39**(Database), D945–D950.
- Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015). COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, **43**(Database issue), D805–11.
- Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature reviews. Cancer*, **4**(3), 177–183.
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Publishing Group*, **46**(3), 310–315.
- Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**(7), 1073–1081.
- Lacaze, P., Pinese, M., Kaplan, W., Stone, A., Brion, M.-J., Woods, R. L., McNamara, M., McNeil, J. J., Dinger, M. E., and Thomas, D. M. (2018). The Medical Genome Reference Bank: a whole-genome data resource of 4,000 healthy elderly individuals. Rationale and cohort design. *bioRxiv*, pages 1–15.

- Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, **42**(D1), D980–D985.
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., and Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature Publishing Group*, **536**(7616), 285–291.
- Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation*, **34**(9), E2393–402.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, **17**(1), 122.
- McNeil, J. J., Woods, R. L., Nelson, M. R., Murray, A. M., Reid, C. M., Kirpach, B., Storey, E., Shah, R. C., Wolfe, R. S., Tonkin, A. M., Newman, A. B., Williamson, J. D., Lockery, J. E., Margolis, K. L., Ernst, M. E., Abhayaratna, W. P., Stocks, N., Fitzgerald, S. M., Trevaks, R. E., Orchard, S. G., Beilin, L. J., Donnan, G. A., Gibbs, P., Johnston, C. I., Grimm, R. H., and ASPREE Investigator Group (2017). Baseline Characteristics of Participants in the ASPREE (ASpirin in Reducing Events in the Elderly) Study. *The journals of gerontology. Series A, Biological sciences and medical sciences*, **72**(11), 1586–1593.
- Orphanet (2017). The portal for rare diseases and orphan drugs. <http://www.orpha.net>.
- Paila, U., Chapman, B. A., Kirchner, R., and Quinlan, A. R. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Computational Biology*, **9**(7), e1003153–8.
- Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, **9**(8), e1003709–13.
- Ruiz-Pesini, E., Lott, M. T., Procaccio, V., Poole, J. C., Brandon, M. C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., and Wallace, D. C. (2007). An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Research*, **35**(Database issue), D823–8.
- Shihab, H. A., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., and Gaunt, T. R. (2014). Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human Genomics*, **8**, 11.
- Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics*, **31**(13), 2202–2204.
- Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad GO, Seattle GO, and NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**(6090), 64–69.