doi: 10.1093/nar/gkz457

MTR-Viewer: identifying regions within genes under purifying selection

Michael Silk[1,2,3], Slavé Petrovski[4,5] and David B. Ascher[1,2,3,6,*]

[1]Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, VIC 3052, Australia, [2]ACRF Facility for Innovative Cancer Drug Discovery, Bio21 Institute, University of Melbourne, Melbourne, VIC 3052, Australia, [3]Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, VIC 3004, Australia, [4]Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK, [5]Department of Medicine, The University of Melbourne, Austin Health and Royal Melbourne Hospital, Melbourne, VIC 3050, Australia and [6]Department of Biochemistry, University of Cambridge; Cambridge CB2 1GA, UK

Received March 10, 2019; Revised May 09, 2019; Editorial Decision May 10, 2019; Accepted June 03, 2019

ABSTRACT

Advances in genomic sequencing have enormous potential to revolutionize personalized medicine, however distinguishing disease-causing from benign variants remains a challenge. The increasing number of human genome and exome sequences available has revealed areas where unfavourable variation is removed through purifying selection. Here, we present the MTR-Viewer, a web-server enabling easy visualization at the gene or variant level of the Missense Tolerance Ratio (MTR), a measure of regional intolerance to missense variation calculated using variation from 240 000 exome and genome sequences. The MTR-Viewer enables exploration of MTR calculations, using different sliding windows, for over 18 000 human protein-coding genes and 85 000 alternative transcripts. Users can also view MTR scores calculated for specific ethnicities, to enable easy exploration of regions that may be under different selective pressure. The spatial distribution of population and known disease variants is also displayed on the protein's domain structure. Intolerant regions were found to be highly enriched for ClinVar pathogenic and COSMIC somatic missense variants (Mann–Whitney U test $P < 2.2 \times 10^{-16}$). As the MTR is not biased by known domains and protein features, it can highlight functionally important regions within genes overlooked or inaccessible by traditional methods. MTR-Viewer is freely available via a user friendly web-server at http://biosig.unimelb.edu.au/mtr-viewer/.

INTRODUCTION

Exome sequencing is becoming a routine tool to guide personalized medicine of genetic diseases (1,2), including in the diagnosis of many Mendelian genetic diseases and to guide cancer treatment decisions (3). While this has lead to a growing library of variants with evidence of pathogenicity (4–6), many variants in a patient's exome remain of uncertain significance. *In silico* predictors of deleteriousness are used to prioritize likely candidate variants, but it remains a major challenge to discriminate pathogenic from benign variants (7).

Large exome (8) and genome (9) sequencing projects have yielded references of variation across the human genome providing the means to measure patterns of variability within genes (10,11). It has been demonstrated previously that measuring depletion of standing variation within genes can be used to identify novel disease-associated genes (10,11). With the current sample sizes of sequenced individuals, we can begin to measure depletion of variation at a regional level within these genes.

We have shown that the Missense Tolerance Ratio (MTR), a measure of regional intolerance to missense variation, can capture this regional level information (12). The MTR is a direct measure of purifying selection of missense variation within a gene, calculated as a ratio between the observed proportion of missense variants compared to an expected proportion, estimated under the assumption of no selection occurring on that sequence context. A sliding window summation is used to provide accurate regional measurements. We have previously shown that regions measured as intolerant to missense variation are significantly enriched for pathogenic missense variants in epilepsy genes (12).

We introduce the MTR-Viewer, a web-server for evaluating missense variant deleteriousness by examining its surrounding regional intolerance. Missense variants that exist within regions that are measured as being intolerant regions are more likely to be pathogenic. The MTR-Viewer

[*]To whom correspondence should be addressed. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au

provides an easy-to-use interface for viewing a selected gene/transcript MTR estimates, also supporting ethnicity-based differences in purifying selection as well as the ability to query individual variants, including via an API, and view disease and background variants on the protein domain structure (http://biosig.unimelb.edu.au/mtr-viewer).

MATERIALS AND METHODS

Data sets

Population variation was sourced from gnomAD (8), the DiscovEHR dataset (13) and the UK Biobank (14), collaborative efforts to aggregate human exome and genome sequences. The amalgamated datasets from a total of 240 000 exome and genome sequences were filtered for only single-point variation with a quality control 'PASS' flag, as previously described (12).

Gene and protein sequences were acquired from the Ensembl database (v95) (15) using the R Bioconductor biomaRt package (16). Transcripts were only used where they contained at least one single-point variant in gnomAD and had non-ambiguous sequences. Ensembl transcript ID's were queried for their matching HGNC gene symbols (17) and Refseq transcript ID's (18) using the biomaRt package.

The observed proportion of missense variation was compared to an expected proportion of missense variation calculated under the assumption of neutrality where no positive/negative selection is occurring. All possible single-point mutations within all gene transcripts were labelled by the Variant Effect Predictor (Release 95) (15) as either missense or synonymous.

For validation purposes, the MTR scores were also calculated in the absence of the DiscovEHR dataset. DiscovEHR missense variants not reported in gnomAD or the UK Biobank, and thus independent of the formulation of the MTR, were used as a control set of neutral variants.

For validation, ClinVar (19) missense variants were retrieved from the NCBI FTP database at ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/ and subset to pathogenic / likely pathogenic and benign / likely benign variants with no conflicting evidence.

For validation, COSMIC (20) missense variants were retrieved from their website at https://cancer.sanger.ac.uk/cosmic/download and filtered for confirmed somatic missense variants.

For further validation, the MTR scores were examined using the FATHMM inherited disease variant dataset and FATHMM cancer-associated missense variants dataset (21). These were compared to the results from the MPC (V2), a prediction of missense variant deleteriousness combining functional and regional missense intolerance information, downloaded from ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/regional_missense_constraint/.

Calculation of the missense tolerance ratio

The proportion of missense variants to synonymous variants was calculated for both the observed variation in gnomAD and the expected variation under neutrality using the annotations from all possible variants in a given transcript, as previously described (12). This was calculated over each Ensembl transcript using a sliding window of 21-, 31- and 41-codons. While using smaller window sizes can provide finer resolution, they can suffer from jitter caused by limited information per window. For this reason we recommend to use 31-codon as the default (12).

For a given window $W_i^{H,J}$ and with selected window size w,

$$\begin{aligned} &\text{where } i = \text{amino acid position} \\ &H = \max\left(1,\ i - (w-1)/2\right) \\ &J = \min\left(\text{transcript length},\ i + (w-1)/2\right), \end{aligned} \quad (1)$$

Within each window (Equation 1), the missense and synonymous variants are each summed at each amino acid position y_i for both the observed and expected datasets (Equation 2).

$$y_i = \sum_{x_m \in W_i^{H,J}} x_m \quad (2)$$

$$\forall x \in \{\text{missense\_obs},\ \text{synonymous\_obs},$$
$$\text{missense\_exp},\ \text{synonymous\_exp}\}$$

Thus for each amino acid position, the MTR is calculated as follows.

$$\mathrm{MTR}_i = \frac{\text{missense\_obs}_i \ / \ (\text{missense\_obs}_i \ + \ \text{synonymous\_obs}_i)}{\text{missense\_exp}_i \ / \ (\text{missense\_exp}_i \ + \ \text{synonymous\_exp}_i)} \quad (3)$$

FDR-adjusted binomial exact test

To identify significantly intolerant regions, an exact binomial test was performed at each residue position to test whether the regional observed proportion of missense variants significantly deviates from the expected proportion.

$$P(X) = \frac{n!}{(n-x)!x!}(p)^x(q)^{n-x} \quad (4)$$

$$\begin{aligned} &\text{where } n = \text{missense\_obs} + \text{synonymous\_obs} \\ &x = \text{missense\_obs} \\ &p = \text{missense\_exp}/(\text{missense\_exp} + \text{synonymous\_exp}) \\ &q = 1 - p \end{aligned}$$

The exome-wide binomial exact test was then adjusted for False Discovery Rate (FDR) using the Benjamini–Hochberg method (22,23). FDR <0.1 was selected through empirical observation as accurately identifying intolerant regions.

WEBSERVER

We have implemented the MTR-Viewer as a user-friendly and freely available web-server (http://biosig.unimelb.edu.au/mtr-viewer/). The webserver was developed using Python Flask (v1.0.2), formatted using Bootstrap (v4.1.3) with data stored using PostgreSQL 10.5. The Pfam API (24) is used to provide graphical domain representations for the accompanying Lollipop plots (25), obtained by translating the HGNC gene symbols (17) to UniProt accession numbers using the UniProt REST API (26). The web application is hosted on an Apache2 web-server running Ubuntu 16.04. MTR calculations were performed in R and plotting within the web application is performed using Python Bokeh (v1.0.1) (27).

Input

The MTR-Viewer can be used in two different ways: either as a gene transcript viewer for MTR estimates across the entire protein-coding sequence or to query specific missense variants for the MTR scores at their position.

The gene viewer query page (Supplementary Figure S1) allows a user to input a specific HGNC gene symbol, which will default to our canonical-selected transcript, or to directly enter an Ensembl transcript ID or Refseq transcript ID. Names are not case sensitive.

The variant query page features a text box for users to input one or multiple missense variants on separate lines in formats chromosome-position-reference-alternative allele, chromosome-position or transcript-protein position. Positions are based on the GRCh37 reference genome. A query can also be performed using an API at http://biosig.unimelb.edu.au/mtr-viewer/api?q=⟨query⟩ or as a CSV file upload.

Users may also search for a gene or transcript using an input box in the navigation bar on the results page, which will assume a gene is being queried unless formatted as a variant using delimiters.

Output

The gene viewer results page (Figure 1) displays the MTR scores across the gene transcript as a line-graph. Line sections are coloured red where the FDR-adjusted binomial exact test <0.1, quantifying MTR deviation from neutrality (MTR = 1). Window sizes of 21 codons, 31 codons (default) and 41 codons can be displayed. Ethnicity-specific MTR estimates, calculated by filtering observed variation by ethnicity, can be overlaid. These are available only for ethnicities with over 15 000 exomes and for window sizes of 31 and 41 codons to account for the smaller sample size. Currently, this includes European Non-Finnish, Latino and South Asian populations. Hovering over the MTR line displays the amino acid position and corresponding MTR estimate. Buttons are available to drag-to-zoom, pan and download the table of raw data for the current transcript in flat file form (MTR estimates for individual genomic variants) or as an MTR table (MTR scores for amino acid positions) (see Figure 1).

Lollipop plots are also shown for the canonical transcript of the selected gene if a matching Pfam graphical representation is available (Figure 1). A lollipop plot is displayed to show the underlying distribution of gnomAD missense (yellow) and synonymous (green) variation and, if the gene is a ClinVar pathogenic gene, a second lollipop plot showing ClinVar annotated pathogenic (red) and benign (blue) variants.

The variant query results page (Supplementary Figure S2) displays a table of the input variants with their corresponding MTR estimates for all Ensembl gene transcripts (v95) that the variant is contained in. Variants with no match are reported in the results table. The view button will redirect the user to the gene viewer for that transcript and label the variant on the MTR line graph.

VALIDATION

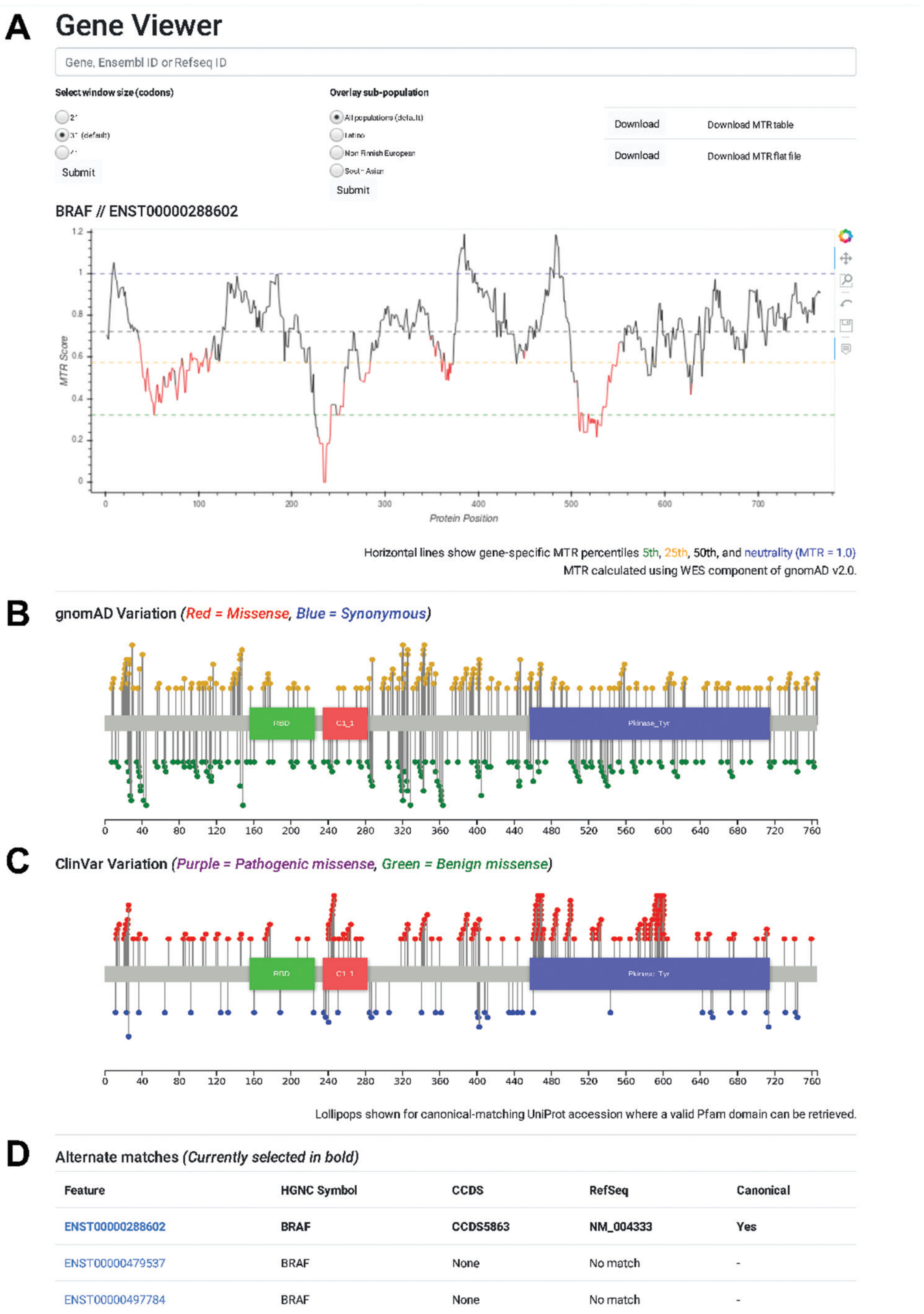
To further validate the utility of the MTR scores to differentiate pathogenic variants, we examined their distribution across the ClinVar pathogenic missense variant and Catalogue Of Somatic Mutations In Cancer (COSMIC) datasets.

The MTR scores of unique ClinVar pathogenic-assigned missense variants ($n = 29\,330$, Average MTR = 0.77, MTR Standard Deviation = 0.24) were compared to the MTR scores of unique ClinVar benign-assigned missense variants ($n = 18\,582$, Average MTR = 0.92, MTR Standard Deviation = 0.14) (Figure 2A). In addition, the pathogenic-assigned variants were also compared to the MTR scores from a novel set of missense variants not observed in gnomAD from the DiscovEHR reference cohort, and filtered to those within ClinVar genes ($n = 195\,735$, average MTR = 0.87, MTR Standard Deviation = 0.18). ClinVar pathogenic-assigned variants were significantly more likely to occur in MTR missense depleted regions than the ClinVar benign variants or the novel population-based DiscovEHR missense variants (Mann–Whitney U test P values of $< 2.2 \times 10^{-16}$).

A comparison of confirmed somatic COSMIC variants ($n = 231\,724$, average MTR = 0.74, MTR Standard Deviation = 0.28) to DiscovEHR population variation within COSMIC genes ($n = 47\,589$, Average MTR = 0.85, MTR Standard Deviation = 0.19) was also performed to identify whether there is significant enrichment of cancer-ascertained somatic mutations within intolerant regions (Figure 2B). COSMIC variants were found to be significantly more likely to occur in intolerant regions (Mann–Whitney U test, $P < 2.2 \times 10^{-16}$).

The discriminatory power of the MTR scores was also assessed using the FATHMM SwissProt/TrEMBL training dataset and the FATHMM cancer-associated training dataset. When we evaluate missense variants with MTR scores less than 0.25 or 0.5, we found that 2.0% and 8.6% respectively of disease causing missense variants, but only 0.1% and 0.9% respectively from neutral variants reside in these regions (odds ratio [OR] = 13.76; Fisher's exact test $P < 2.2 \times 10^{-16}$, odds ratio [OR] = 10.11; Fisher's exact test $P < 2.2 \times 10^{-16}$). Similarly, we found that 2.1% and 9.7% of cancer asociated missense variants, but only 0.3% and 1.7% from neutral variants, have MTR less than 0.25 or 0.5 respectively (odds ratio [OR] = 6.49; Fisher's exact test $P < 2.2 \times 10^{-16}$, odds ratio [OR] = 6.36; Fisher's exact test $P < 2.2 \times 10^{-16}$). This showed that low-MTR scored regions are highly enriched for pathogenic variation.

We empirically selected FDR <0.1 to define regions with a significantly different proportion of observed missense variants. 10.5% of the FATHMM disease-associated variants and 9.6% of the cancer-associated variants are found in these regions, compared with 2.4% and 3.2% neutral variants, showing a significant enrichment of disease-associated variation in both datasets (odds ratio [OR] = 4.69; Fisher's exact test $P < 2.2 \times 10^{-16}$, odds ratio [OR] = 3.23; Fisher's exact test $P < 2.2 \times 10^{-16}$).



Select window size (codons)

○ 21
● 31 (default)
○ 41

Overlay sub-population

● All populations (default)
○ Latino
○ Non Finnish European
○ South Asian

Download MTR table

Download MTR flat file

BRAF // ENST00000288602

Horizontal lines show gene-specific MTR percentiles 5th, 25th, 50th, and neutrality (MTR = 1.0)
MTR calculated using WES component of gnomAD v2.0.

gnomAD Variation *(Red = Missense, Blue = Synonymous)*

ClinVar Variation *(Purple = Pathogenic missense, Green = Benign missense)*

Lollipops shown for canonical-matching UniProt accession where a valid Pfam domain can be retrieved.

Alternate matches *(Currently selected in bold)*

Feature	HGNC Symbol	CCDS	RefSeq	Canonical
ENST00000288602	**BRAF**	**CCDS5863**	**NM_004333**	**Yes**
ENST00000479537	BRAF	None	No match	-
ENST00000497784	BRAF	None	No match	-

Figure 1. MTR-Viewer gene query results page. (**A**) A line graph displays the MTR distribution for example gene *BRAF* with regions in red indicating observed variation differs significantly from neutrality. (**B**) Lollipop plots show the underlying gnomAD missense and synonymous variation and (**C**) ClinVar known pathogenic and known benign variants for the gene. (**D**) Alternate transcripts are displayed below with matching RefSeq transcript ID's.

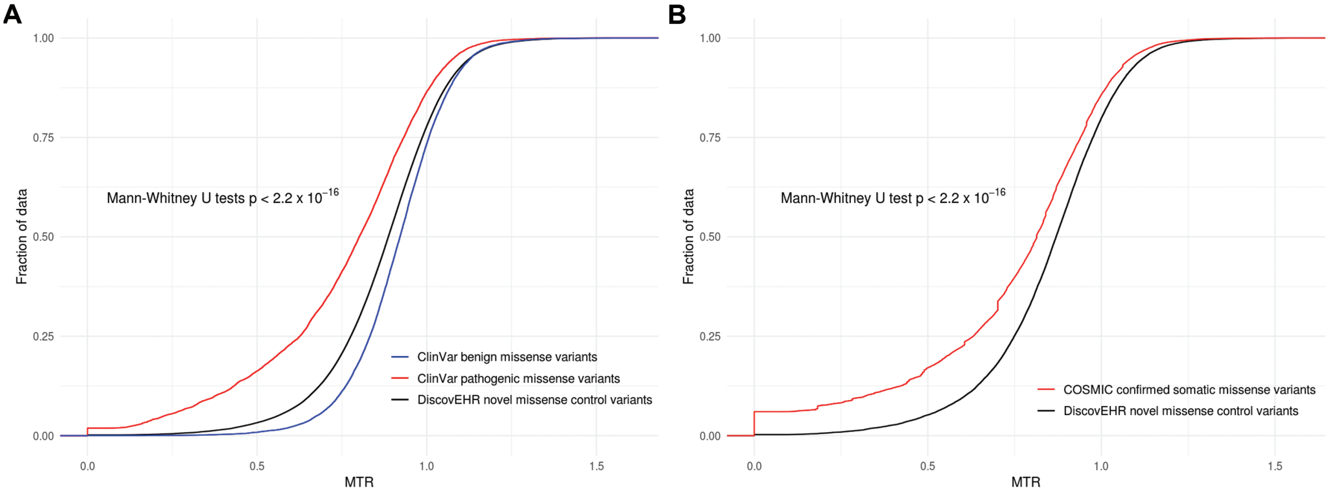


Figure 2. Distribution of MTR scores for known disease variants compared to background. (**A**) Cumulative distribution of MTR scores for ClinVar pathogenic variants (red), ClinVar benign variants (blue) and DiscovEHR novel missense control variants (black). (**B**) Cumulative distribution of MTR scores for COSMIC somatic missense variants (red) compared with DiscovEHR novel missense control variants (black).

While the MTR is solely a measure of missense depletion, using the FATHMM training datasets, it was compared to the trained predictors MPC and PolyPhen-2, which utilize functional information (Supplementary Tables S1 and S2). The MTR had the highest Matthew's correlation coefficient over the FATHMM cancer-associated dataset, and was comparable to the MPC scores over the FATHMM disease-associated using the authors' defined cut-off of MPC >2.

CONCLUSION

Here we present the MTR-Viewer, a web-server to explore regional intolerance to missense variation across human protein-coding genes from 240 000 exome and genome sequences. By providing a measure and visualization of purifying selection occurring within a given gene transcript, patient-ascertained variants can be better prioritized based on whether they reside in intolerant regions (12). The MTR-Viewer is freely available as a user-friendly web server at http://biosig.unimelb.edu.au/mtr-viewer/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Australian Government Research Training Program (to M.S.); NHMRC R.D. Wright Career Development Fellowship [1126877 to S.P.]; AstraZeneca (to S.P.); Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council (MRC) [MR/M026302/1 to D.B.A.]; National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; Department of Biochemistry and Molecular Biology, University of Melbourne [to M.S. and D.B.A.]. Funding for open access charge: MRC.

Conflict of interest statement. S.P. is an employee of AstraZeneca.

REFERENCES

1. Choi,M., Scholl,U.I., Ji,W., Liu,T., Tikhonova,I.R., Zumbo,P., Nayir,A., Bakkaloglu,A., Ozen,S., Sanjad,S. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*,**106**, 19096–19101.
2. Yang,Y., Muzny,D.M., Reid,J.G., Bainbridge,M.N., Willis,A., Ward,P.A., Braxton,A., Beuten,J., Xia,F., Niu,Z. *et al.* (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, **369**, 1502–1511.
3. Beltran,H., Eng,K., Mosquera,J.M., Sigaras,A., Romanel,A., Rennert,H., Kossai,M., Pauli,C., Faltas,B., Fontugne,J. *et al.* (2015) Whole-exome sequencing of metastatic cancer and biomarkers of treatment response. *JAMA Oncol.*, **1**, 466–474.
4. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
5. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
6. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
7. MacArthur,D.G., Manolio,T.A., Dimmock,D.P., Rehm,H.L., Shendure,J., Abecasis,G.R., Adams,D.R., Altman,R.B., Antonarakis,S.E., Ashley,E.A. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
8. 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. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
9. Genomes Project, C., 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. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
10. 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 Genet.*, **9**, e1003709.
11. Samocha,K.E., Kosmicki,J.A., Karczewski,K.J., O'Donnell-Luria,A.H., Pierce-Hoffman,E., MacArthur,D.G., Neale,B.M. and Daly,M.J. (2017) Regional missense constraint improves variant deleteriousness prediction. bioRxiv, 148353. http://dx.doi.org/10.1101/148353, 12 June 2017, preprint: not peer reviewed.

12. Traynelis,J., Silk,M., Wang,Q., Berkovic,S.F., Liu,L., Ascher,D.B., Balding,D.J. and Petrovski,S. (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.*, **27**, 1715–1729.
13. Dewey,F.E., Murray,M.F., Overton,J.D., Habegger,L., Leader,J.B., Fetterolf,S.N., O'Dushlaine,C., Van Hout,C.V., Staples,J., Gonzaga-Jauregui,C. *et al.* (2016) Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, **354**, 1549.
14. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P., Green,J., Landray,M. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.
15. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
16. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
17. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
18. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
19. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
20. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
21. Shihab,H.A., Gough,J., Mort,M., Cooper,D.N., Day,I.N. and Gaunt,T.R. (2014) Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics*, **8**, 11.
22. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
23. Haynes,W. (2013) Benjamini–Hochberg Method. In: Dubitzky,W, Wolkenhauer,O, Cho,K-H and Yokota,H (eds). *Encyclopedia of Systems Biology*. Springer, NY, p. 78.
24. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
25. Jay,J.J. and Brouwer,C. (2016) Lollipops in the clinic: information dense mutation plots for precision medicine. *PLoS One*, **11**, e0160519.
26. UniProt,C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
27. Bokeh Development Team (2014) Bokeh: Python library for interactive visualization. http://www.bokeh.pydata.org.