CAGI 6 challenge: Predicting molecular events underlying disease using variant annotation, aberrant gene expression events, and human phenotype ontology.

Julien Gagneur[1,2,3], Christian Mertes[1,2], Ines Scheller[1,3], Nicholas H. Smith[1], Vicente A. Yépez[1]

1. Department of Informatics, Technical University of Munich, Garching, Germany
2. Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany
3. Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

Authors listed alphabetically.

Methods

Data acquisition and preprocessing

As provided by the CAGI 6 challenge, we downloaded the VCF files (DNA), the CRAM files (RNA), and the clinical reports containing the patients' phenotypes. The genomic data (VCF files) were obtained by sequencing DNA purified from blood by Complete Genomics (Stavropoulos et al., 2016) or by SickKids (Lionel et al., 2018). The data were aligned against the genome build GRCh37 and variant calling was completed with GATK (Van der Auwera and O'Connor, 2020). Only single nucleotide variants (SNVs) and small insertion/deletions (INDELs) were considered.
The challenge provided polyA enriched and stranded RNA sequencing data from whole blood aligned with STAR v2.6.1c (Dobin et al., 2013) against a modified version of hg19. We downloaded the CRAM files and converted them to BAM format using Samtools v1.12 (Danecek et al., 2021) with the provided FASTA file:

samtools view -T {input.FASTA} -b -o {output} {input.CRAM}

All DNA-RNA pairs were verified to match using the sampleQC module of the Detection of RNA Outliers Pipeline DROP v1.1.1 (Yépez et al., 2021).

The summary files of the patients' phenotypes were manually converted from text into the numeric identifiers of the Human Phenotype Ontology (HPO) terms (Köhler et al., 2020).

Variant annotation

First, the VCF files were normalized using BCFtools v1.12 (Danecek et al., 2021) and the provided FASTA file:

bcftools norm -f {input.FASTA} --check-ref ws -m -both {input.VCF}

Variants that created a faulty entry through the normalization step were discarded. The normalized variants were then annotated using the variant effect predictor VEP (McLaren et al., 2016) with the *everything* flag that includes gnomAD allele frequencies (Karczewski et al., 2020), protein domains, and HGVs annotation based on the ENSEMBL release 99 (Zerbino et al., 2018). In addition, we used the following VEP Plugins: SpliceAI (Jaganathan et al., 2019), CADD (Rentzsch et al., 2021), and UTRannotator (Whiffin et al., 2020), with their default configurations and required input files. Additionally, scores from a recent evolutionary model to predict the pathogenicity from variants, EVE, were added to the variants (Frazer et al., 2021).

Variant calling in RNA-seq data

No genotype file and no WGS sequencing file was provided for sample 20-10362. Therefore, variants were called on RNA-seq data of sample 20-10362 using GATK best practices for RNAseq short variant discovery as described in Zhao et al., (2019) and Yepez et al., (2021). In short, variants with a ratio of quality to depth of coverage < 2, that were strand biased (Phred-scaled fisher exact score >30), or belonging to an SNP cluster (3 or more SNPs within a 35 bp window) were filtered out, as suggested by GATK. Furthermore, variants not contained in a repeat masked region as defined by RepeatMasker v4.1.0 (Smit et al.) and with 3 or more reads supporting the alternative allele were prioritized.

Filtering and sorting variants

Based on the VEP annotation, we applied multiple filters to extract the two most impactful variants per gene and patient. We first discarded variants lying outside protein-coding genes or with a minor allele frequency (MAF) > 0.1 in any population within gnomAD (Karczewski et al., 2020). We then categorized the variants into three pathogenicity categories similar to the ACMG Guidelines (Richards et al., 2015): *high impact, medium impact,* and *rare*. The *high impact* is similar to the ACMG categories *strong*, *very strong*, and partially *moderate.* The *medium* category resembles the ACMG categories *supporting* and *moderate*. The detailed filtering criteria and cutoffs are given below with logical OR except for the MAF cutoff:

	High impact	Medium impact	Rare
MAF	0.01	0.01	0.01
ClinVar annotation	Likely pathogenic or pathogenic	Likely pathogenic or pathogenic	---
VEP Impact	HIGH	HIGH or MODERATE	---
VEP Consequence	---	Any splicing relevant variant	---
CADD Phred score	>20	>10	---
Any SpliceAI score	>0.5	>0.2	---

EVE score	>0.64	>0.5	---
UTRannotator	Any annotation	Any annotation	---

Finally, we extracted the two most impactful variants per patient-gene pair after ranking the variants by 1) is canonical, 2) our impact categories, 3) the VEP impact, and 4) the EVE score.

Aberrant expression

Aberrant expression was obtained following the corresponding module of DROP with the default parameters. Read counts were obtained by counting strand-unspecific with the DROP module on the exon-level and by aggregating on the gene-level based on the provided gene annotation file and then they were modeled using OUTRIDER (Brechtmann et al., 2018). In order to increase the statistical power for the fit, the 50 whole blood samples from the Genotype-Tissue Expression project (GTEx v6) with the highest sequencing depth were combined with the provided dataset. A median of 1 expression outlier (false-discovery rate FDR < 0.1) per sample was obtained.

Aberrant splicing

Aberrant splicing was also obtained using DROP with the default parameters. Split reads and non-split reads spanning exon-intron boundaries were counted and converted into the splicing ratio-based metrics percent-spliced in ψ and splicing efficiency θ (Pervouchine et al., 2013). These metrics were then modeled independently using FRASER (Mertes et al., 2021). Junctions with an FDR < 0.1 and an absolute differential splicing effect greater than 0.3 were considered significant. We further subsetted the results to junctions with strong effects on both the donor and acceptor sites using a metric similar to the Jaccard index and discarded junctions lying on “blacklist” regions (Amemiya et al., 2019). A median of 25 genes per sample showed aberrant splicing.

Mono-allelic expression

Mono-allelic expression (MAE) was also computed following the module of DROP. For each heterozygous single-nucleotide variant, reads aligning to each allele were counted. These reads were modeled using a negative binomial test. Variants with alternative allele ratios > 0.8 or < 0.2 and with an FDR < 0.05 are considered to be mono-allelically expressed.

Semantic similarity

First, the semantic similarity was computed between all available HPO terms (https://hpo.jax.org, (Köhler et al., 2020)) and the HPO terms from each sample using the *compareHPSets* function from the R package Phenotype Consensus Analysis PCAN (Godard and Page, 2016). Then, these scores were grouped by gene. Finally, a single aggregating semantic similarity score was computed per gene-sample combination using the *hpSetCompSummary* function from the R package PCAN.

Learning a disease impacting score for each gene per patient

To score the impact on disease for every gene per patient, we considered a classification problem as follows. Non-coding genes and genes without a rare variant per individual were discarded upfront. We defined a training set combining all remaining gene-individual combinations from the CAGI dataset. The dataset was prepared as follows:

- an element is a pair (gene, individual)
- all genes with at least one rare variant (MAF<0.01) were considered, whether they are expressed or not in the RNA-seq sample
- We use the following gene-level annotations:
    - is the gene expressed in the tissue
    - is the gene aberrantly down-regulated
    - is the gene aberrantly up-regulated
    - is the gene aberrantly spliced
    - semantic similarity score
    - is the gene reported to cause Mendelian disorders (OMIM, Amberger et al., 2019)
- And the following variant-level annotations for the top 2 variants from the filtering step:
    - Alternative allele ratio
    - MAF
    - CADD score
    - SpliceAI score
    - EVE score
- In case the top variant was homozygous, these scores were repeated. In case there was no second variant, we added dummy values to avoid missing ones (Alternative allele ratio 0.5, MAF 0.2, CADD 9, SpliceAI 0.1, EVE 0.4).

We furthermore generated the same features for 111 individuals with rare mitochondrial disorders for which the causal gene, variants, DROP results, and HPO terms were available (Kopajtich et al., 2021; Stenton et al., 2021; Yepez et al., 2021). This added 111 positive gene-individual pairs. Because i) the features are based on semantic similarity rather than HPO terms themselves, and ii) the number of positive pairs is approximately equal to the number of patients, we reasoned that modeling the probability of a gene-individual pair to be positive in this dataset would function as a rough proxy for the probability of a gene to be disease-causing in the given individual.

Next, we trained a gradient boosted tree model to discriminate between the 111 positive gene-individual pairs and the rest of the dataset. To this end, we used XGBoost (Chen and Guestrin, 2016), an algorithm that is adequate for strong class imbalanced classification problems. We selected the hyperparameters that yielded the highest area under the ROC curve using a 5-fold cross-validation scheme. The hyperparameters are: *eta = 0.1, max_depth = 2, gamma = 4, subsample = 1, colsample_bytree = 1, eval_metric = 'auc'*. This model was then trained on the full dataset with those hyperparameters. The trained model estimates the probability for each gene-individual pair to be of the positive class. That is the score we report in

our first submission. Uncertainties of the predictions were estimated as the standard error across prediction scores generated by training the model on 10 bootstraps with replacement of the full dataset.

Manual Curation

In addition to building the automated gene prioritization method outlined above, we also prioritized genes based on manual inspection of the variants. Using the ACMG standard and guidelines to focus on literature support in population, disease-specific, and sequence databases in addition to the predictive qualities of in-silico tools. We manually altered our model's scores to reflect the ACMG labels. We call 'potentially biallelic' variants in cases of heterozygous variants, where we cannot conclude whether they lie in the same allele. We settled on a discrete set of scores that override the scores predicted by the model:

- 1: biallelic variants, all classified as *very strong or strong evidence of pathogenicity.*
- 0.8: biallelic variants, all classified as *moderate evidence of pathogenicity* or higher.
- 0.5: biallelic or potentially biallelic variants, all classified as *supporting evidence of pathogenicity* or higher.
- 0.2: monoallelic variant classified as *very strong or strong evidence of pathogenicity*, or inconclusive variants in a gene with aberrant gene expression or that matches the phenotype.

These values were assigned after the authors looked at the corresponding evidence on a case-by-case level. Using this overriding scoring, we modified 22 scores across 19 samples. The modified scores along with all the unmodified ones are the ones we report in our second submission. As a standard error for the score was requested for the challenge, we provided a rather arbitrary one for all the manually changed values. We set it to 0.1, which is the order of magnitude of the standard errors estimated by the bootstrap for our XGBoost model.

Submitted results tables

We have submitted two tables. The first one contains the scores of the XGBoost model for the top 100 genes with the highest score per sample. The second contains the same information as the first, but with the 22 manually curated scores overriding the model predictions. We report at most two variants per row. For cases with aberrant expression, we report the type, the fold change and the false discovery rate (FDR). For cases with aberrant splicing, we report the type and the genomic coordinates of all aberrantly spliced junctions. For cases with MAE, we report the alternative allele ratio. We do not include the RefSeq transcripts as we report the full genomic coordinates of the variants and splicing events using the GRCh37 human genome build. We developed gene-level scores but no specific variant-level scores. Therefore, we report for the variants the probability and standard deviation estimates of the gene.

Code availability

The code to preprocess the raw data, find the expression outliers, and execute the XGBoost model is publicly available under www.github.com/gagneurlab/cagi6_sickkids.

References

Amberger, J. S., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res.* 47, D1038–D1043. doi:10.1093/nar/gky1151.

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* 9, 9354. doi:10.1038/s41598-019-45839-z.

Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V. A., Avsec, Ž., Herzog, M., et al. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* 103, 907–917. doi:10.1016/j.ajhg.2018.10.025.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 785–794. doi:10.1145/2939672.2939785.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008. doi:10.1093/gigascience/giab008.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., et al. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature* 599, 91–95. doi:10.1038/s41586-021-04043-8.

Godard, P., and Page, M. (2016). PCAN: phenotype consensus analysis to support disease-gene association. *BMC Bioinformatics* 17, 518. doi:10.1186/s12859-016-1401-2.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535-548.e24. doi:10.1016/j.cell.2018.12.015.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7.

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., et al. (2020). The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* 49, D1207–D1217. doi:10.1093/nar/gkaa1043.

Kopajtich, R., Smirnov, D., Stenton, S. L., Loipfinger, S., Meng, C., Scheller, I., et al. (2021). Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. Genetic and Genomic Medicine doi:10.1101/2021.03.09.21253187.

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., et al. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.* 20, 435–443. doi:10.1038/gim.2017.119.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122. doi:10.1186/s13059-016-0974-4.

Mertes, C., Scheller, I. F., Yépez, V. A., Çelik, M. H., Liang, Y., Kremer, L. S., et al. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* 12, 529. doi:10.1038/s41467-020-20573-7.

Pervouchine, D. D., Knowles, D. G., and Guigo, R. (2013). Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* 29, 273–274. doi:10.1093/bioinformatics/bts678.

Picard toolkit (2019). Broad Institute Available at: http://broadinstitute.github.io/picard/.
Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* 13, 31. doi:10.1186/s13073-021-00835-9.
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). 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. *Genet. Med.* 17, 405–423. doi:10.1038/gim.2015.30.
Smit, A., Green, P., and Hubley, R. RepeatMasker. Available at: http://www.repeatmasker.org/ [Accessed May 4, 2020].
Stavropoulos, D. J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., et al. (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *Npj Genomic Med.* 1, 1–9. doi:10.1038/npjgenmed.2015.12.
Stenton, S. L., Shimura, M., Piekutowska-Abramczuk, D., Freisinger, P., Distelmaier, F., Mayr, J. A., et al. (2021). Diagnosing pediatric mitochondrial disease: lessons from 2,000 exomes. doi:10.1101/2021.06.21.21259171.
Van der Auwera, G. A., and O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Inc Available at: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/ [Accessed December 29, 2021].
Whiffin, N., Karczewski, K. J., Zhang, X., Chothani, S., Smith, M. J., Evans, D. G., et al. (2020). Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat. Commun.* 11, 2523. doi:10.1038/s41467-019-10717-9.
Yepez, V. A., Gusic, M., Kopajtich, R., Mertes, C., Smith, N. H., Alston, C., et al. (2021). Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *medRxiv*. doi:10.1101/2021.04.01.21254633.
Yépez, V. A., Mertes, C., Müller, M. F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., et al. (2021). Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* 16, 1276–1296. doi:10.1038/s41596-020-00462-5.
Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi:10.1093/nar/gkx1098.
Zhao, Y., Wang, K., Wang, W., Yin, T., Dong, W., and Xu, C. (2019). A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics* 20, 160. doi:10.1186/s12864-019-5533-4.