Functional filter for whole-genome sequencing data identifies HHT and stress-associated non-coding *SMAD4* polyadenylation site variants >5 kb from coding DNA

Authors

Sihao Xiao, Zhentian Kai, Daniel Murphy, ..., Mark J. Caulfield, Genomics England Research Consortium, Claire L. Shovlin

Correspondence

sihao.xiao@bnc.ox.ac.uk (S.X.),
c.shovlin@imperial.ac.uk (C.L.S.)

Xiao and colleagues generated a filter to prioritize the ~5 million gDNA variants/DNA from whole-genome sequencing. They identified distal 3′ UTR high-impact, non-coding, rare variants that were adjacent to RNA cleavage and polyadenylation sites, explained previously unsolved clinical phenotypes, and were functionally validated in the participants' peripheral blood mononuclear cells.

Xiao et al., 2023, The American Journal of Human Genetics *110*, 1903–1918
November 2, 2023

https://doi.org/10.1016/j.ajhg.2023.09.005

Functional filter for whole-genome sequencing data identifies HHT and stress-associated non-coding *SMAD4* polyadenylation site variants >5 kb from coding DNA

Sihao Xiao,[1,2,11,*] Zhentian Kai,[3] Daniel Murphy,[2,4] Dongyang Li,[1,2] Dilip Patel,[1,2] Adrianna M. Bielowka,[1,2] Maria E. Bernabeu-Herrero,[1,2] Awatif Abdulmogith,[1,2] Andrew D. Mumford,[5] Sarah K. Westbury,[5] Micheala A. Aldred,[6] Neil Vargesson,[7] Mark J. Caulfield,[8] Genomics England Research Consortium,[9] and Claire L. Shovlin[1,2,10,*]

Summary

Despite whole-genome sequencing (WGS), many cases of single-gene disorders remain unsolved, impeding diagnosis and preventative care for people whose disease-causing variants escape detection. Since early WGS data analytic steps prioritize protein-coding sequences, to simultaneously prioritize variants in non-coding regions rich in transcribed and critical regulatory sequences, we developed GROFFFY, an analytic tool that integrates coordinates for regions with experimental evidence of functionality. Applied to WGS data from solved and unsolved hereditary hemorrhagic telangiectasia (HHT) recruits to the 100,000 Genomes Project, GROFFFY-based filtration reduced the mean number of variants/DNA from 4,867,167 to 21,486, without deleting disease-causal variants. In three unsolved cases (two related), GROFFFY identified ultra-rare deletions within the 3′ untranslated region (UTR) of the tumor suppressor *SMAD4,* where germline loss-of-function alleles cause combined HHT and colonic polyposis (MIM: 175050). Sited >5.4 kb distal to coding DNA, the deletions did not modify or generate microRNA binding sites, but instead disrupted the sequence context of the final cleavage and polyadenylation site necessary for protein production: By iFoldRNA, an AAUAAA-adjacent 16-nucleotide deletion brought the cleavage site into inaccessible neighboring secondary structures, while a 4-nucleotide deletion unfolded the downstream RNA polymerase II roadblock. *SMAD4* RNA expression differed to control-derived RNA from resting and cycloheximide-stressed peripheral blood mononuclear cells. Patterns predicted the mutational site for an unrelated HHT/polyposis-affected individual, where a complex insertion was subsequently identified. In conclusion, we describe a functional rare variant type that impacts regulatory systems based on RNA polyadenylation. Extension of coding sequence-focused gene panels is required to capture these variants.

Introduction

Whole-genome sequencing (WGS) is an established component of medical genetic and research repertoires, but currently, the majority of its potential is unrealized. In any one individual, WGS identifies millions of DNA variants compared to reference sequences. These are present in ~20,000 protein-coding genes, and also in much less understood regions of the genome that have diverse functions including transcription into noncoding RNAs, participation in DNA chemical changes that modify transcription, and binding to other nucleic acids or proteins.[1,2]

Current WGS clinical foci are almost exclusively on a subgroup of protein-coding genes where biological function is already known. In research spheres, in order to reduce the number of variables per sample, interrogation of WGS data also commences with prioritization methods, usually based on selection of specific genomic regions. Variants in the non-coding genome, while not pre-depleted by the sequencing methodology, are effectively deleted in the early analytic stages of variant prioritization. Importantly, application of these WGS methods leave large proportions of individuals with hereditary conditions unsolved, without a genetic diagnosis.[3]

There is no accurate map of all functional genomic regions in human genomes, and it is difficult to predict *a priori,* where all regulatory elements for a specific gene locus would be located. We hypothesized, however, that it would be possible to design a more efficient variant prioritization method for WGS because markers of epigenetics and DNA-protein interactions have been applied genome-wide by molecular laboratories, and an enormous body of biological experimental data has been made publicly available. As a result, there now exist repositories of information indicating which sections of DNA are more or less likely to have a functional role in at least one examined tissue.

We designed a genomic regions of functionality filter for priority (GROFFFY) based on published experimental data

[1]National Heart and Lung Institute, Imperial College London, W12 ONN London, UK; [2]National Institute for Health Research (NIHR) Imperial Biomedical Research Centre, W2 1NY London, UK; [3]Topgen Biopharm Technology Co. Ltd., Shanghai 201203, China; [4]Women's, Children's & Clinical Support (Pharmacy), Imperial College Healthcare NHS Trust, W2 1NY London, UK; [5]School of Cellular and Molecular Medicine, University of Bristol, BS8 1QU Bristol, UK; [6]Division of Pulmonary, Critical Care, Sleep & Occupational Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [7]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, AB25 2ZD Aberdeen, UK; [8]William Harvey Research Institute, Queen Mary University of London, E1 4NS London, UK; [9]Genomics England, EC1M 6BQ London, UK; [10]Specialist Medicine, Imperial College Healthcare NHS Trust, W12 OHS London, UK
[11]Present address: Big Data Institute, University of Oxford, Oxford, UK
*Correspondence: sihao.xiao@bnc.ox.ac.uk (S.X.), c.shovlin@imperial.ac.uk (C.L.S.)
https://doi.org/10.1016/j.ajhg.2023.09.005.

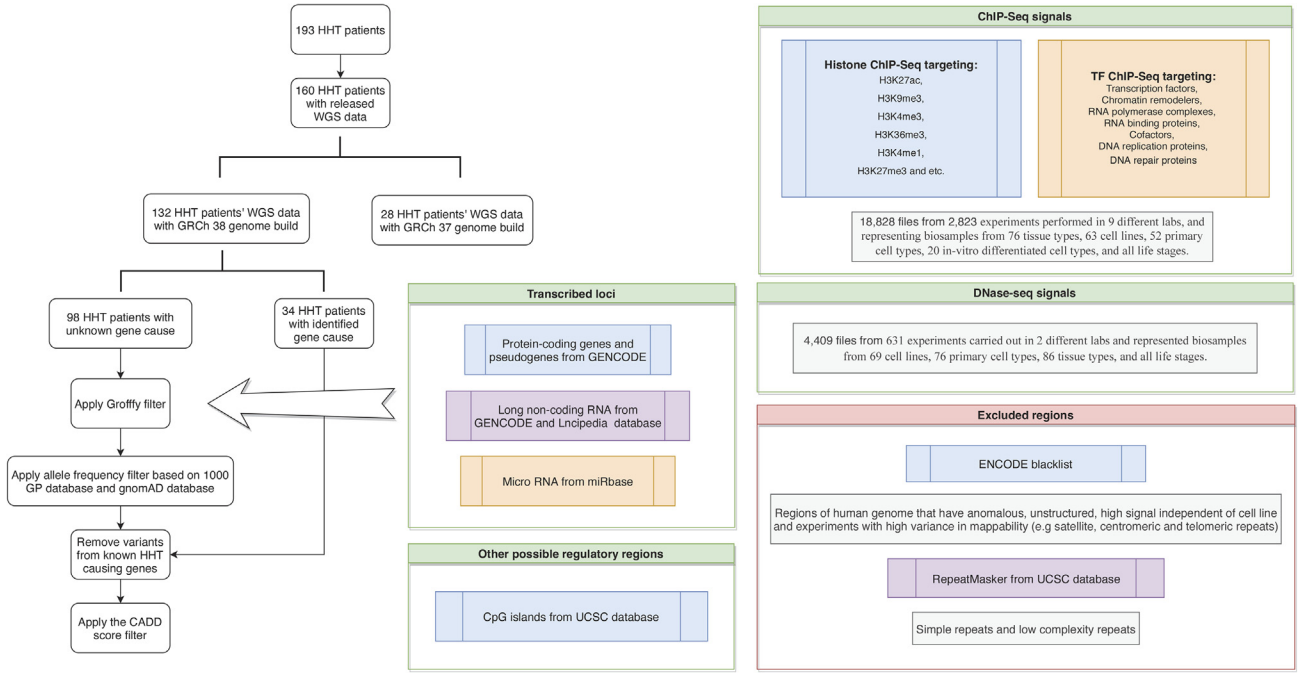

Figure 1. GROFFFY Study Protocol
Flow chart illustrating sequence of stages described in the text and Tables S1, S2, S3, and S4.

particularly from ENCODE[4–6] and performed validation and discovery analyses in WGS data from individuals recruited to the 100,000 Genomes Project.[7] To accelerate clinical impact, we focused discovery analyses on noncoding regions of a tumor-suppressor gene examined in diagnostic and screening gene panels. *SMAD4* is ubiquitously expressed and encodes the common partner SMAD which regulates signaling by transforming growth factor (TGF)-β, bone morphogenetic protein (BMP), and activin ligands.[8] As indicated by its function and earlier gene names (*DPC4* [deleted in pancreatic cancer]; *MADH4* [mothers against decapentaplegic]), the SMAD4 protein has major pathological and developmental roles.[8,9] *SMAD4* is a target of cancer genetic diagnostics because it is a driver gene for major cancers due to somatic loss[8,9] and because germline heterozygous loss causes gastrointestinal polyposis (juvenile polyposis syndrome/JPS [MIM: 174900]) where untreated hamartomatous polyps can undergo malignant transformation leading to colon, gastric, and other cancers.[9,10] Heterozygous loss also causes TGF-β/BMP-related vasculopathies including hereditary hemorrhagic telangiectasia (HHT) which usually results from a loss-of-function variant in *ACVRL1* (MIM: 600376) or *ENG* (MIM: 187300). Where *SMAD4* is identified (juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome, JPHT [MIM: 175050]),[11,12] this allows affected individuals to benefit from life-long polyposis and aortopathy screening programs.[10] Scientifically, *SMAD4* is of great interest because despite its wide-ranging roles in development and disease, little is known of its regulation.[8,9]

Here we report that this WGS analytic approach identifies a type of functional DNA variant uncaptured by usual clinical sequencing methodologies.

Material and methods

The procedures followed were in accordance with the ethical standards of the responsible committees on human experimentation (institutional and national) and proper informed consent was obtained.

Study design

The main elements of the study design are outlined in Figure 1. Following earlier recruitment of individuals with HHT to the 100,000 Genomes Project (black arrows), the GROFFFY filter was designed as indicated in colored boxes and applied to the WGS data files within the 100,000 Genomes Project.

Participant recruitment and whole-genome sequencing

The 100,000 Genomes Project was set up by the UK Department of Health and Social Security in 2013 to sequence whole genomes from National Health Service (NHS) individuals and their families. The study received ethical approval from the Health Research Authority (HRA) East England-Cambridge South Research Ethics Committee (REC ref. 14/EE/1112), and all participants provided written consent. Anonymized raw sequencing data were available in the Genomics England Research Environment (see web resources).[7] Separately, in a clinical diagnostic pipeline, Genomics England performed data alignments and variant classifications fed back to NHS Genomic Medicine Cents and recruiting clinicians.[13,14]

The cohort recruited with hereditary hemorrhagic telangiectasia (HHT)[11,12] were particularly suited for *GROFFFY* methodological validation processes because a subset had not undergone prior genetic testing and because clinical pipelines were incomplete at the time of *GROFFFY* analyses. This resulted in a validation dataset of 34 WGS sequences where clinical pipelines had identified a causal variant in an HHT gene[13–18] and a discovery dataset of 98 WGS sequences where some DNAs were expected

to have heterozygous loss-of-function variants in *ACVRL1*, *ENG*, or *SMAD4*.

Generating GROFFFY

Genomic coordinates of regions included in GROFFFY were generated from publicly available databases using the Imperial College High Performing Computing service. Experimentally derived biological data were used in preference to computational predicted files with potential for false negatives. Genomic coordinates were extracted from data aligned to GRCh38, and excluded data originating in cancer cells. Regions being selected for are described in Tables S1, S2, S3, and S4 which provide full details of coordinate derivation from transcribed loci and candidate regulatory element (cRE) regions.[6,19,20] Following merging of 18,828 bed files from 3,454 experiments,[5,6] sequences in the ENCODE blacklist[21] and RepeatMasker (see web resources) were excluded. In detail:

Genomic coordinates for candidate regulatory element (cRE) regions were generated using data from the ENCODE Encyclopedia registry which includes data from both ENCODE[4,5] and the NIH Roadmap Epigenomics Consortia.[22] We downloaded the call sets itemized in Tables S3 and S4 from the ENCODE portal.[4,5] By merging DNA binding call sets and representative DNase hypersensitivity site (rDHS) call sets,[23] a rough prediction of all cREs was made. Only data aligned to GRCh38 were retained. Data generated from cancer cells were excluded as cancer cells' genomes are usually heavily modified and rearranged.[24]

(1) For DNA binding data, histone ChIP-seq and transcription factor ChIP-seq (which target histones, transcription factors, chromatin remodelers, RNA polymerase complexes, RNA binding proteins, cofactors, DNA replication proteins, and DNA repair proteins) were searched. We downloaded the call sets from the ENCODE portal[4] as indicated in Table S3. The 18,828 downloaded files (Table S3) were from 2,823 experiments performed in 9 different labs and represented biosamples from 76 tissue types, 63 cell lines, 52 primary cell types, 20 *in vitro* differentiated cell types, and all life stages. Downloaded bed files were divided into 10 subgroups with the first 9 subgroups containing 2,000 bed files and the last subgroup containing 828 bed files. Bed files from each subgroup were merged together using BEDOPS[25] and then merged bed files from each subgroup were joined together last to obtain all DNA binding regions. The merged bed file for DNA-binding regions was 49,941,695 kb (i.e., greater than 15 genomes) before sorting, reflecting many overlapping regions.

(2) For rDHSs, DNA accessibility experiments were searched in the ENCODE Encyclopedia database.[5] We downloaded 4,409 files as indicated in Table S4. These were from 631 experiments carried out in 2 different labs and represented biosamples from 69 cell lines, 76 primary cell types, 86 tissue types, and all life stages. Bed files were merged together directly using BEDOPS (v.2.4.26)[25] to obtain all accessible DNA regions.

(3) Region coordinates for CpG islands were downloaded from the UCSC database[26,27] using Rsync command tool. The downloaded file was in txt format and was converted to bed files using awk function in the Linux system. The converted bed file was then sorted by location using BEDOPS.[25]

Genomic coordinates for transcribed loci were extracted as follows.

(4) GENCODE human genome annotation v.31 for GRCh38 was downloaded from the UCSC database[26,27] using Rsync in the command line. All gene coordinates were extracted by using awk function (including protein-coding gene and pseudogenes). The downloaded file was in gff3 format and was converted to bed files using BEDOPS.[25]

(5) Long non-coding RNA annotations (lncRNA) were downloaded from both GENCODE (release 31) and LNCipedia databases (v.5.2).[19] Coordinates for lncRNAs were merged together using BEDOPS[25] to obtain all possible lncRNA gene regions.

(6) MicroRNA (miRNA) annotations were downloaded from miRbase (release 22.1).[20] The downloaded file was in gff3 format and was converted to bed files using BEDOPS.[25]

Genomic coordinates of regions excluded were identified from the ENCODE blacklist which was downloaded directly from the ENCODE project website in bed format, and RepeatMasker which was downloaded from the UCSC database[26,27] in txt format. The Linux command awk was used to grep out region coordinates, and BEDOPS[25] was used to sort the file.

The final size of the GROFFFY filter was 1,423,480,943 bp approximating to 44.48% of GRCh38.

To assimilate for GROFFFY, separately, the Ubuntu shell (v.16.04.2 LTS based on Linux 4.4.0–64-generic x86_64 system) was launched for the Genomics England Research Environment (see web resources),where the final genomics coordinates for GROFFFY were transferred. WGS variant data were examined after analysis by the Illumina WGS Service Informatics pipeline. This used Illumina Issac and Starling[28] for sequence alignment and to identify variants. The output files were in vcf files with.gz compression and decompressed using Gzip (v.1.6). Pandas module v.0.22.0 in Python (v.3.6.5) was used to process vcf files. Pegasus, the High-Performance computer cluster of Genomics England, was used to run computationally intensive jobs, submitted to the Load Sharing Facility (LSF).

The Intersect function of Bedtools v.2.26.0[29] was then used to identify WGS variants from vcf files that were in the GROFFFY bed file regions. Option -header was used to remove any headers, and option -wa was used to ensure the output file format was the same as the input vcf file. For the Intersect function, any intersection with GROFFFY was outputed to result files, even if some part of the variation was outside of the filter region. Annotations of the WGS vcf files were carried out using the Ensembl Variant Effect Predictor (VEP) v.96.3, based on Perl v.5.24, SAMtools v.1.5[30] (specifically SAMtools HTSlib v.1.5[30]), and a list of options to optimize the process (Table S2). A Python script was written to produce 66 shell scripts where each shell script contained 2 annotation jobs. R v.3.5.1, within R studio v.13.4.0, was downloaded from the Comprehensive R Archive Network and used to perform statistical tests. Paired datasets were analyzed using the non-parametric Mann Whitney (Wilcoxon rank-sum) test and multiple datasets by the Kruskal-Wallis rank-sum test with post-test Dunn's multiple comparisons.

Regions from the ENCODE Blacklist,[21] and RepeatMasker (see web resources) were subtracted from regions selected for using the "difference" option in BEDOPS.[25] All bed files were merged together to obtain the selected genomic regions. Numeric data (Table S5) and Python scripts were approved for export through the Research Environment AirLock under subproject RR42 (HHT-Gene-Stop, Table S6).

GROFFFY analysis of whole-genome sequencing data

As detailed in Figures S1 and S2, stepwise filters excluded variants where general population allele frequency exceeded 0.0002 in the 1000 Genome Project[31] or gnomAD[32] databases; synonymous variants not in splice regions; all non HHT-causal variants in the Validation Set HHT DNAs; and variants with a Combined Annotation-Dependent Depletion (CADD) score <10.[33] There was no *a priori* reason to follow any specific filter (for example, a CADD < 10 does not preclude such a variant being important), but our goal was to prioritize in the context of the current question. In detail:

An autosomal-dominant-specific disease application step was included as a high-stringency "white list" filter. For this, the annotated WGS files were retrieved for the 34 validation set DNAs where a causative variant had already been identified in known HHT genes through clinical pipelines.[13–16,34,35] Variant information was collected through unique variant IDs consisting of chromosome number, variant starting position, reference sequence, and altered sequence (e.g., chr1:111_C/TTT). To confirm that no two variants were represented by the same variant ID, the full list was compared to a set where only unique values were stored, and the two lists were identical. The variant IDs were integrated in a white list. Exclusion of these white-listed variants in other affected individuals was performed using the isin function of Pandas module: any variant in the white list was deleted from the vcf files of the target set DNAs, and the number of variants after exclusion was recorded and outputed to txt files.

For CADD score filtration and prioritization, the plugin option of VEP was used to annotate variants with CADD scores[33] in the enclosed research environment: databases for SNV annotation (v.1.5) and small indel annotation (v.1.5), which were pre-installed in the research environment, were indexed. As the annotation of CADD scores was quite slow, the process was put toward the end of the analysis pipeline, so that there were fewer variants that needed to be annotated. Prioritization by CADD score was performed by generating further customized Python scripts. The CADD PHRED-scale score for all 9 billion SNVs and millions of small indels was extracted from the information column. Variants absent from the CADD score database were represented by an empty string by default and were replaced by number 999 instead. Variants with a PHRED score less than 10 were removed so that both variants with top 10 percentiles deleteriousness and variants absent from the database were prioritized. The processed files were stored in vcf format.

Export of variant coordinates and bioinformatic analyses

Following approval for export through Genomics England AirLock (Table S6), variant genomic coordinates were visualized in the UCSC Genome Browser.[26,27] Endothelial expression of *SMAD4* was examined in whole transcriptome data from primary human blood outgrowth endothelial cells (BOECs).[36] Binary sequence alignment map (bam) files aligned to GRCh38 were analyzed in Galaxy v.2.4.1[37] and the Integrated Genome Browser (IGB) 9.1.8.[38]

3′ UTR alternative polyadenylation quantitative trait loci (3′aQTLs) from 46 tissues isolated from 467 individuals in the Genotype Tissue Expression (GTEx) project[39] were sourced through the 3′aQTL Atlas.[40] Genetic variants likely affecting gene expression in GTEx V8[39] data release were captured from UCSC[26,27] CAVIAR tracks, which define high-confidence gene expression QTLs within 1 MB of gene transcription start sites (*cis*-eQTLs).

All variants were independently verified by Genomics England. Impact on microRNA binding sites was examined through TargetScan Human Release 8.0[41] and miRDB.[42] RNA structure predictions were performed using iFoldRNA v.2.0[43,44] without restraints, and final models were visualized using Mol* Viewer[45] via the Research Collaboratory for Structural Bioinformatics Protein DataBank server.[46]

Clinical re-contact, correlations, and re-sampling

Genomics England "Contact the Clinician" forms were submitted through the Research Environment (Table S6) and clinicians who had recruited the participants were contacted and joined the research team. Clinical correlations were performed through North Thames and South West NHS Genomic Medicine Service Alliances. The affected participants were contacted by their clinicians and provided written consent for publication after reviewing the relevant sections of the manuscript.

Two of these participants also consented to further blood samples together with 3 healthy volunteers and a further unsolved individual recruited to the 100,000 Genomes Project with a JPHT (MIM: 175050) clinical phenotype. This study was approved by the East of Scotland Research Ethics Service (EoSRES: 16/ES/0095), and the 6 participants provided written informed consent. Using methods we have developed to perform experimental treatments on human cells while resuspended in endogenous plasma,[47–49] peripheral blood mononuclear cells (PBMCs) were prepared using BD Vacutainer CPT tubes (Bunzl Healthcare) according to manufacturer's instructions with minor modifications. As detailed further in the supplemental methods, these were to provide comparative resources of cells in stressed and unstressed states, where alternate transcripts/exon region use might be impacted by modified efficiency of final AAUAAA cleavage and polyadenylation due to the 3′ UTR variants.

In brief, immediately after venesection, the blood was gently remixed by inverting 8–10 times and then centrifuged within 2 h of collection for 30 min at 1,600 relative centrifugal force (RCF) at room temperature. The PBMC-containing buffy coat and plasma were collected by pipetting from above the gel layer, transferred to a single 50 mL tube for each donor, and gently inverted to resuspend. After PBMC resuspension in plasma, for each donor, equal volumes were distributed to separate experimental treatment tubes, prewarmed at 37°C for 10 min, then subjected to 4 different treatment conditions for 1 h including control at 37°C and low temperature in a 32°C waterbath for 1 h to mimic the stress incurred at the threshold between mild and moderate hypothermia.[50,51] Additional stresses previously optimized in our laboratory[47–49,52,53] were inhibition of translation by cycloheximide 100 μg/mL (cycloheximide inhibits eukaryotic translation elongation by mechanisms including binding to the 60S ribosomal subunit E-site[54,55]) and a clinically relevant mild reactive oxygen species (ROS) stress using ferric citrate 10 μmol/L.[53,56] After 1 h, all tubes were centrifuged at 520 RCF at room temperature for 15 min. Cell pellets were lysed in Tri reagent (Cambridge Bioscience Ltd) before distribution to replicate tubes for paired rRNA-depleted and polyA-selected RNA sequencing library generation.

RNA sequencing and differential expression analyses

RNA extraction and quality control for 96 samples was performed by Genewiz. For RNA-seq library preparations, 48 samples were polyA selected for polyadenylated RNA enrichment, and 48 paired samples underwent ribosomal (r)RNA depletion. RNA was

fragmented and random primed for first and second strand cDNA synthesis, end repair, 5′ phosphorylation, dA-tailing, adaptor ligation, PCR enrichment, and Illumina HiSeq sequencing using paired-end 150 bp reads (Genewiz). Sequenced reads were trimmed using Trimmomatic v.0.36,[57] aligned to *Homo sapiens* GRCh38 using STAR aligner v.2.5.2b, and unique gene reads that fell within exon regions were counted using Subread package v.1.5.2 (Genewiz).

Blinded to the types of donors and treatments, Genewiz performed differential gene expression analyses using DESeq2[58] and differential exon expression using DEXSeq[59] to identify differentially spliced genes by testing for significant differences in read counts on exon regions (and junctions) of the genes. In DEXSeq,[59] read counts are normalized by size factors: contributions to the average are weighted by the reciprocal of an estimate of their sampling variance, and the expected variance used to derive weights for the "balanced" coefficients reported as estimates for the strengths of differential exon usage and DEXSeq plotting, that are of similar magnitude to the original read counts.[59] The output indicates alternative transcript isoform regulation, noting individual exon region assignment is reliable as long as only a small fraction of counting regions (bins) in the gene is called significant.[59]

Noting control variability in initial DESeq2 analyses (Figure S3), the least variable of human transcripts (the 25 genes with GINI Coefficients [GCs] < 0.15 in diverse cells[60,61]) were used to evaluate individual library quality (Figure S4) and subsequently employed for DESeq2 normalization. For these normalizations, the intra-assay coefficient of variation (CV [100*standard deviation (SD)/mean])[62] was calculated for replicate pairs using alignment per gene adjusted for total read counts per library, and analyses were restricted to libraries where >50% of GINI genes had a CV $< 10\%$ ("met CV10"). Three rRNA depletion datasets failed this quality control. The remaining datasets were DeSeq2[58] normalized using the GINI[60,61] genes as housekeepers. For this, the ratio of alignment counts for each selected housekeeper gene in each dataset to the geometric mean of that gene was calculated across the remaining 45 datasets from rRNA-depleted libraries. The median value of these ratios in each library was used to generate the "size factor" to scale that library's alignments (Table S7).

Statistics

Descriptive statistical analyses were performed using Python and STATA v.17.0 (Statacorp). Comparative statistics of the number of variants before and after filtration was performed using Mann Whitney two-group comparisons. RNA-seq expression was analyzed in STATA v.17.0 (Statacorp) and GraphPad Prism 9 (GraphPad Software), compared using Kruskal Wallis and Dunn's post test applied for selected pairwise comparisons.

Results

GROFFFY defines biologically validated regions of functionality

By including only regions where biological experiments have generated evidence in favor of functional roles, GROFFFY essentially excludes biologically less important regions of the genome. Nevertheless, the GROFFFY filter region based on positive selection of transcribed loci and candidate regulatory element (cREs), and masking of repetitive regions, included 44.4% of the human genome. A heatmap at 500 kb resolution is provided in Figure 2A. A more detailed view of GROFFFY is provided in Figure 2B.

GROFFFY substantially reduces the number of DNA variants per DNA

The scale of the bioinformatics challenge was emphasized by the pre-filtration number of DNA variants per individual which ranged from 4,786,039 to 5,070,340 (mean 4,867,167). Applying GROFFFY as a first filter reduced the mean number of variants by 2,812,015 (Figures 3A and 3B). Restricting to rare variants with population allele frequencies $< 2 \times 10^{-4}$ [63] removed means of 2,476,589[31] and 2,483,377[32] variants/DNA according to database (Figures 3A and 3B). After removing variants with a CADD[33] score <10, the mean number of unique, rare, and impactful DNA variants per genome was 21,486 (Figure 3C).

GROFFFY did not delete key variants, as shown by the validation dataset: all already-known pathogenic variants in the unfiltered dataset were retained post filtration (Figure 4A). Further, in the discovery set of 98 whole genomes, for *ACVRL1* and *ENG,* the majority of identified variants clustered to the exons and flanking regions sequenced in clinical diagnostics (Figure 4B).

Hot spot of rare deletion variants in the distal *SMAD4* 3′ untranslated region

No coding *SMAD4* variants were identified in the discovery dataset (Figure 4B). We focused on a hot spot of 3 deletion variants in the 3′ untranslated region (UTR) of *SMAD4* (Figure 4B). There were two unique variants, one of which was identified in both affected members of a single family. The variants deleted nucleotides 5,519 and 5,649 bp distal to the *SMAD4* stop codon and did not affect any microRNA binding sites.[41,42] The wild-type sequences were consistently expressed in human primary blood outgrowth endothelial cells (BOECs) derived from donors with normal SMAD4[36] (Figure 5A). General population common variant data also supported the importance of the region: while the 3′ UTR did not contain any expression quantitative trait loci (eQTLs)[39] (Figure 5B), the variants were within the only kilobase of the 3′ UTR to contain 3′ alternate polyadenylation QTLs (3′ aQTLs,[40] Figure 5C).

Variants delete nucleotides near the final *SMAD4* alternate polyadenylation site

The *SMAD4* UTR used by all coding transcripts contains 7 alternate polyadenylation site (PAS)[65] AAUAAA hexamers. These cluster in two proximal groups of 3, before a single final AAUAAA hexamer at chr18:51,083,977 (Figure 5A). This final hexamer lay immediately proximal to the two deletion variants, and as expected,[65] was flanked by an upstream AU-rich element suited to binding of proteins in the cleavage and polyadenylation (CPA) complex, and downstream repeat elements predicting intermolecular interactions in single-stranded RNA that would generate

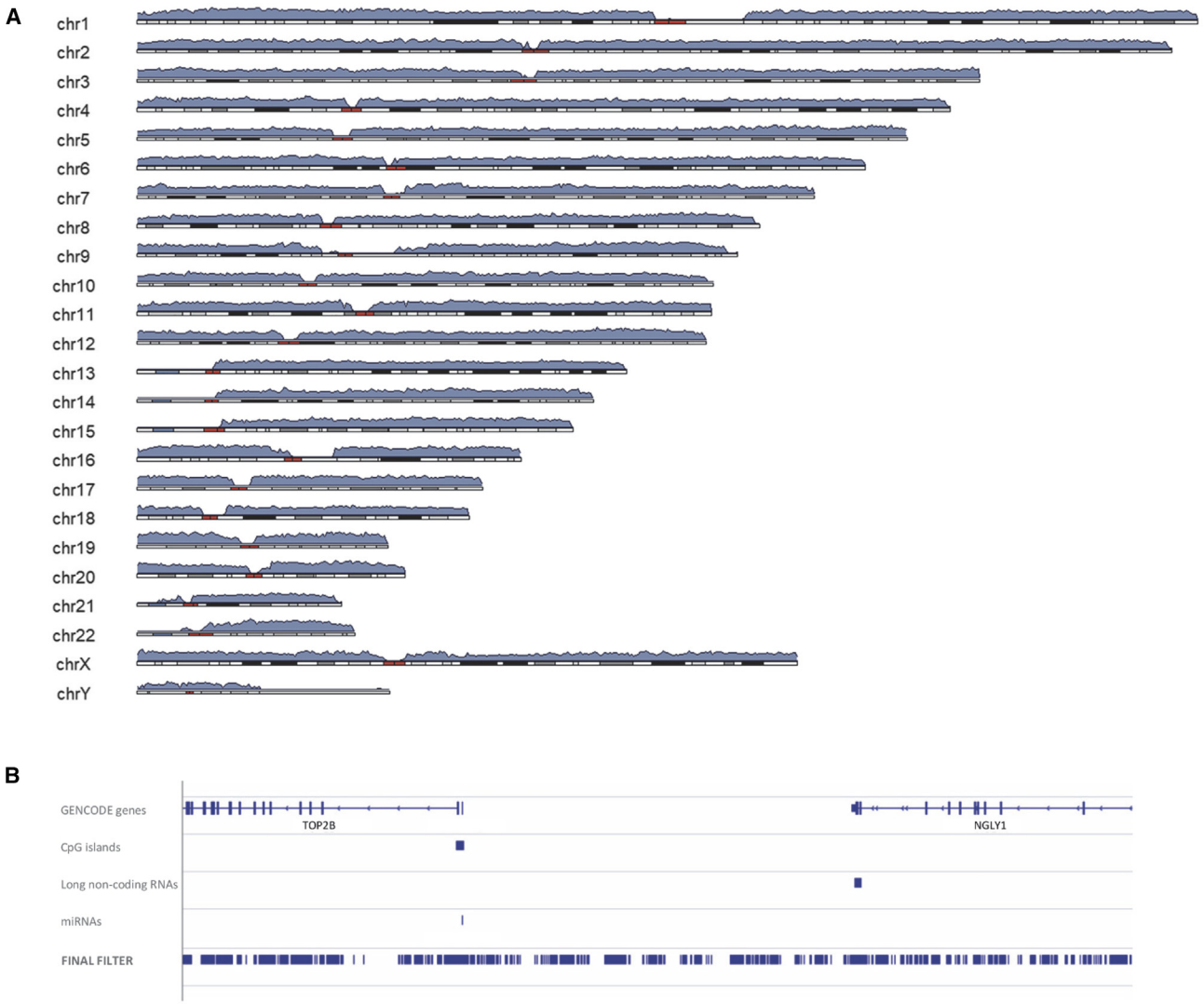

Figure 2. GROFFFY and the human genome

(A) Heatmap displaying GROFFFY categorization of Genome Reference Consortium Human (GRCh) Build 38. The heatmap maps 61,799 data points at 500 kb resolution, and heights represent the percentage of each 500 kb region included in GROFFFY.

(B) Higher-resolution image from a randomly chosen region of the genome (chr3:25,597,986–25,783,443). The top 4 tracks illustrate sources, from top: GENCODE[6] gene annotations, CpG islands,[26,27] long non-coding RNAs,[19] and miRNAs.[20] The lowest track illustrates the final filter. Note that this filter contains both intra- and intergenic regions for the region and that the raw data were not subjected to any processed annotation tracks.

secondary structures to block the progress of RNA polymerase II (Figure 6A). One GROFFFY-filtered variant deleted the 16 nucleotides sited +3 to +18 from the PAS hexamer with 5 further single nucleotide substitutions, and the second deleted 4 nucleotides in the downstream repetitive element region (Figure 6B; Table S8).

The deletion variants disrupt RNA secondary structures required for cleavage and polyadenylation

iFoldRNA secondary structures[43,44] visualized using Mol* Viewer[45] via the Research Collaboratory for Structural Bioinformatics Protein DataBank server[46] indicated that both deletion variants disturbed secondary structures that substantially altered the sequence context for CPA activity. In wild-type sequence, the AAUAAA hexamer was in a near-linear conformation with stacked pyrimidine and purine rings evident on magnified views (Figure 7Ai, Video S1). Strikingly, with the neighboring complex deletion variant, the AAUAAA nucleotides acquired new inter-molecular interactions, lost the stacked alignment of bases, and were incorporated into inaccessible secondary structures (Figure 7Aii, Video S2). In contrast, the second variant, which deleted 4 nucleotides 134 bp downstream of the AAUAAA hexamer, disrupted and unfolded the downstream structured region expected to be the major RNA polymerase II roadblock[65] (Figure 7B).

Clinical correlation

All three individuals with variants 1 and 2 had clinically confirmed HHT.[11–18,35,66] After identification of the

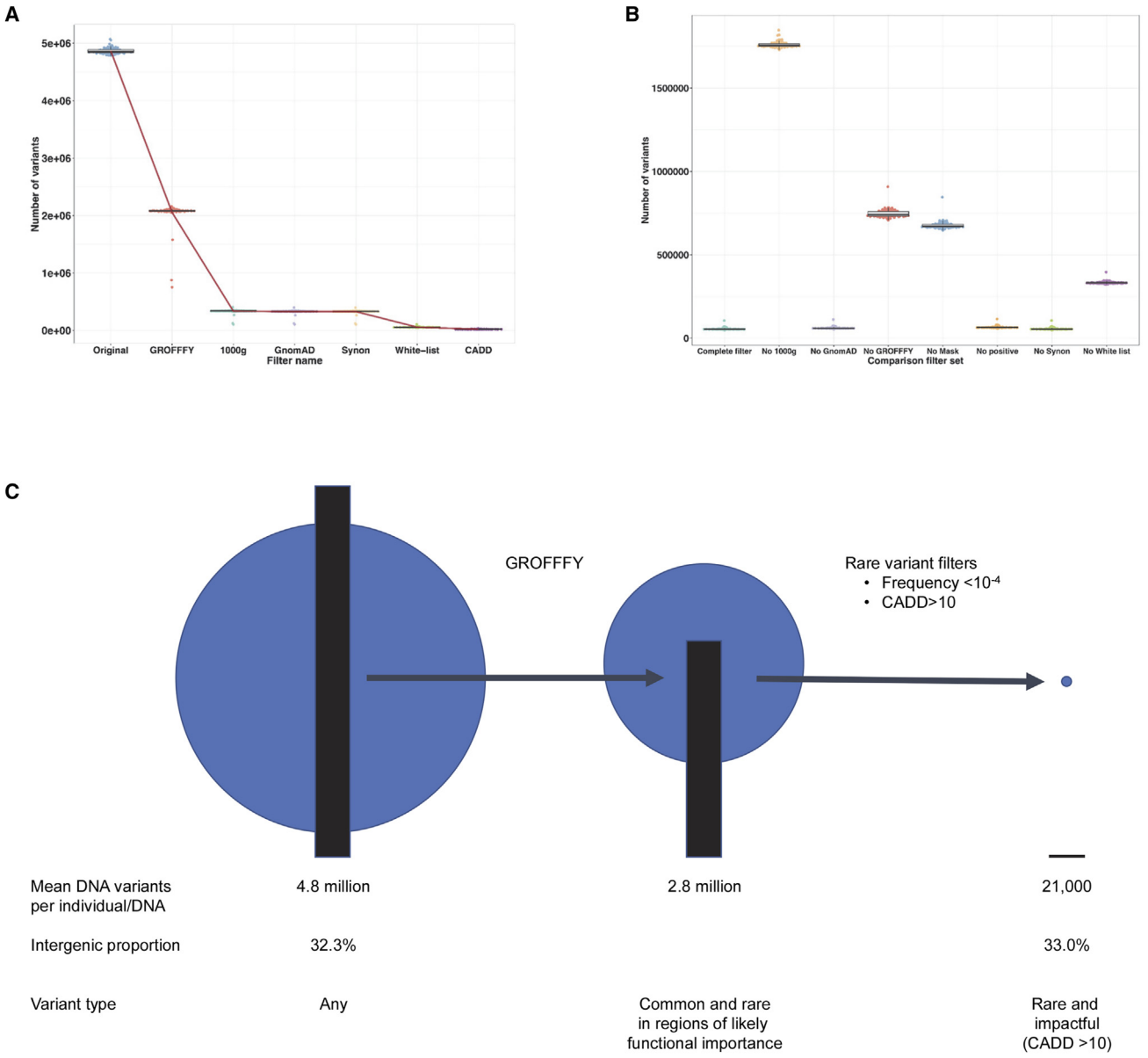

Figure 3. Application of GROFFFY to whole-genome sequences

(A) Serial application of GROFFFY; allele frequency filters based on frequencies in the 1000 Genomes (1000g)[31] or gnomAD[32]; synonymous (Synon) filter, and white-listed filter (see material and methods and Tables S1, S2, S3, and S4 for further details). Where error bars are not visible at the illustrated scale, exact numeric data are provided in Table S5.

(B) Number of variants remaining per DNA after applying each comparator filter set.

(C) Number, site, and type of DNA variants present in 98 human whole genomes before and after application of GROFFFY and other filters, scaled in one dimension (black bars) and two dimensions (blue circles). CADD, combined annotation-dependent depletion score where >10 represents a variant in the top 10% of deleteriousness.[33] Irrespective of other filters applied, GROFFFY and its individual components significantly reduced the number of variants compared to the other tested filter sets (Figures S1 and S2).

SMAD4 variants, recruiting clinicians also reported *SMAD4*-compatible clinical phenotypes: The first-degree relatives with variant 2 had no other identified cause to HHT. They each experienced daily nosebleeds and had classical HHT telangiectasia, and one had pulmonary arteriovenous malformations requiring treatment and hemihypertrophy (left-right axis defect). Gastrointestinal and aortopathy screening had not been considered. The individual with variant 1 did have a missense variant in *ACVRL1,* though in addition to severe nosebleeds needing blood transfusion and intravenous iron, classical HHT telangiectasia, and pulmonary arteriovenous malformations, they had multiple colonic and rectal polyps requiring excision over a 6 year period of observation.

Peripheral blood mononuclear cell *SMAD4* RNA expression

As described in the supplemental methods, peripheral blood mononuclear cells (PBMCs) were isolated from affected individuals with the 3 *SMAD4* variants and

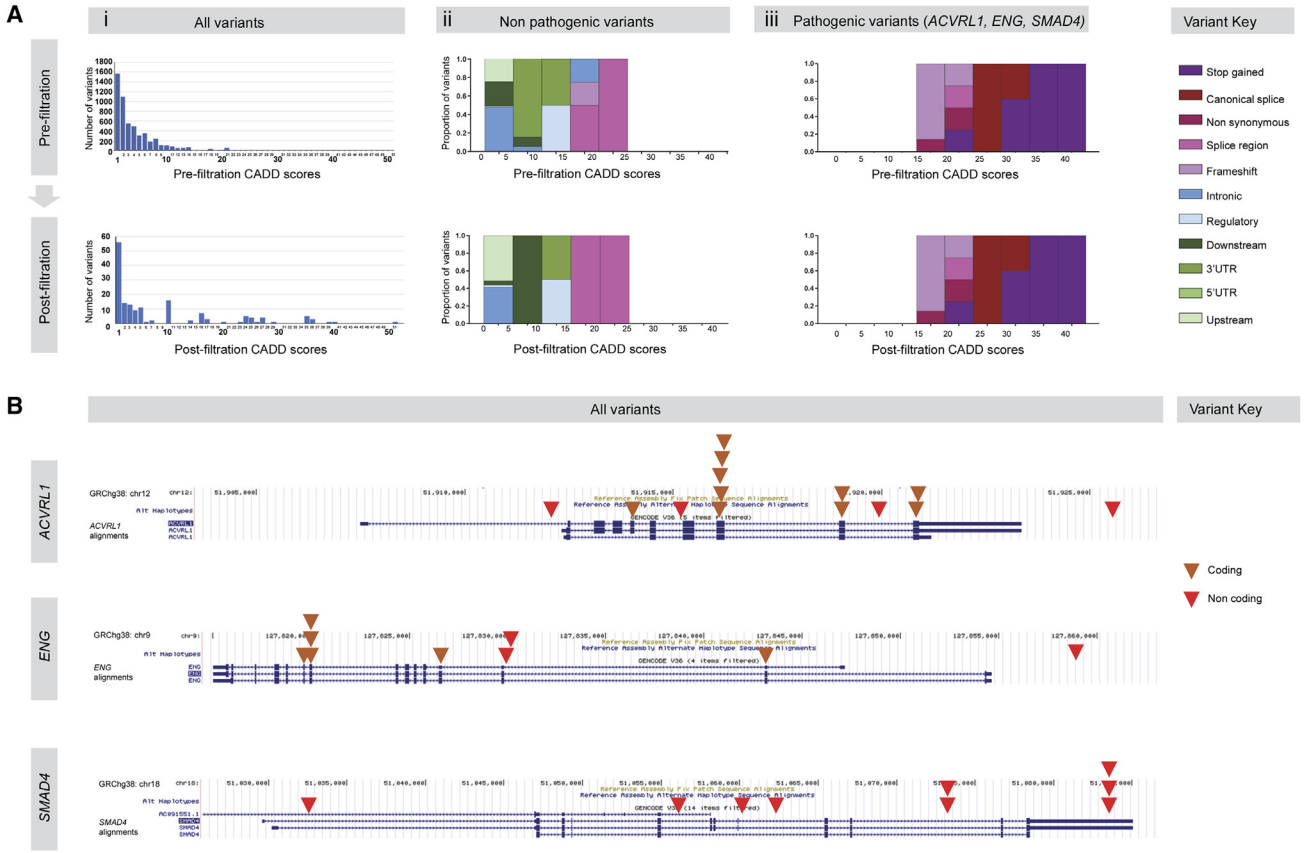

Figure 4. GROFFFY variant-level validation

Comparison of HHT gene variants in the validation and discovery datasets before and after application of GROFFFY.

(A) Validation dataset. (i) Total number of variants with indicated CADD[33] scores (note differences in scale before and after filtration). (ii) Non-pathogenic variants by CADD[33] score categories: molecular subtypes are indicated in the key. (iii) Pathogenic variants by CADD[33] score categories, and molecular subtype as in key. Note identical plots in (iii) pre- and post-filtration because all pathogenic variants were still present after filtration.

(B) Discovery dataset. Location of GROFFFY-captured variants in the major HHT genes[18] (i) *ACVRL1*, (ii) *ENG*, (iii) *SMAD4*. The cartoons include screenshots of GRCh38 from the University of California Santa Cruz (UCSC) Genome Browser[26,27] and major transcripts. Red inverted triangles indicate location of variants after application of all filters (brown for coding/splice regions, bright red for non-coding regions).

controls, and cultured in conditions predicted to modify 3′ UTR use, before RNA sequencing. DESeq2[58] analyses of PolyA-selected RNAs indicated that *SMAD4* polyadenylated transcripts increased after a 1 h hypothermic stress, and this was also seen in 2 individuals with the *SMAD4* variants (Figure S3). However, for the rRNA-depleted libraries representing "total" RNA, variability between control samples assessed by initial DESeq2 analyses was high (Figure S3). This reduced after normalizing with low GINI[60,61] coefficient genes (Figure S4).

Whether normalized to read counts per library[67] or GINI[60,61] genes, total *SMAD4* RNA expression was lower in the "inaccessible AAUAAA" variant 1 donor than 3 controls (Figure 8Ai). Decrements were also apparent in untreated PBMC exon regions by DEXSEQ[59] (Figure 8Bi). In controls, *SMAD4* transcript expression was modified following 1 h cycloheximide 100 μg/mL, with lower use of exon region (ER)60 containing the AAUAAA site and variants, consistent with shorter 3′ UTRs after stress (Figures S5 and S6). Despite this, ER60 use was further reduced in the variant 1 donor after CHX (Figures 8Ci and 8Di) with increased use of penultimate exon regions ER52-55 (Figures 8Di and 8Ei), supporting different 1 h changes in RNA splicing on stress.

Total *SMAD4* RNA was higher in the "roadblock unfolding" variant 2 donor than 3 unaffected control subjects across all conditions (Figure 8Aii). Although exon region use was similar to control subjects in untreated PBMCs (Figures 8Bii and 8Cii), after 1 h cycloheximide, compared to controls there was higher use of regions corresponding to two of the 3′ aQTLs (Figures 8Dii and 8Eii). We concluded that the contrasting overall expression patterns were consistent with the opposing predictions following RNA modeling of variants 1 and 2; that variant 2 data also supported different 1 h changes in RNA splicing after CHX stress, but that precise transcript changes would need to be the subject of future RNA studies.

Validation of positive control variant

The third donor had been recruited as a positive control due to JPHT syndrome (MIM: 175050) with colonic and

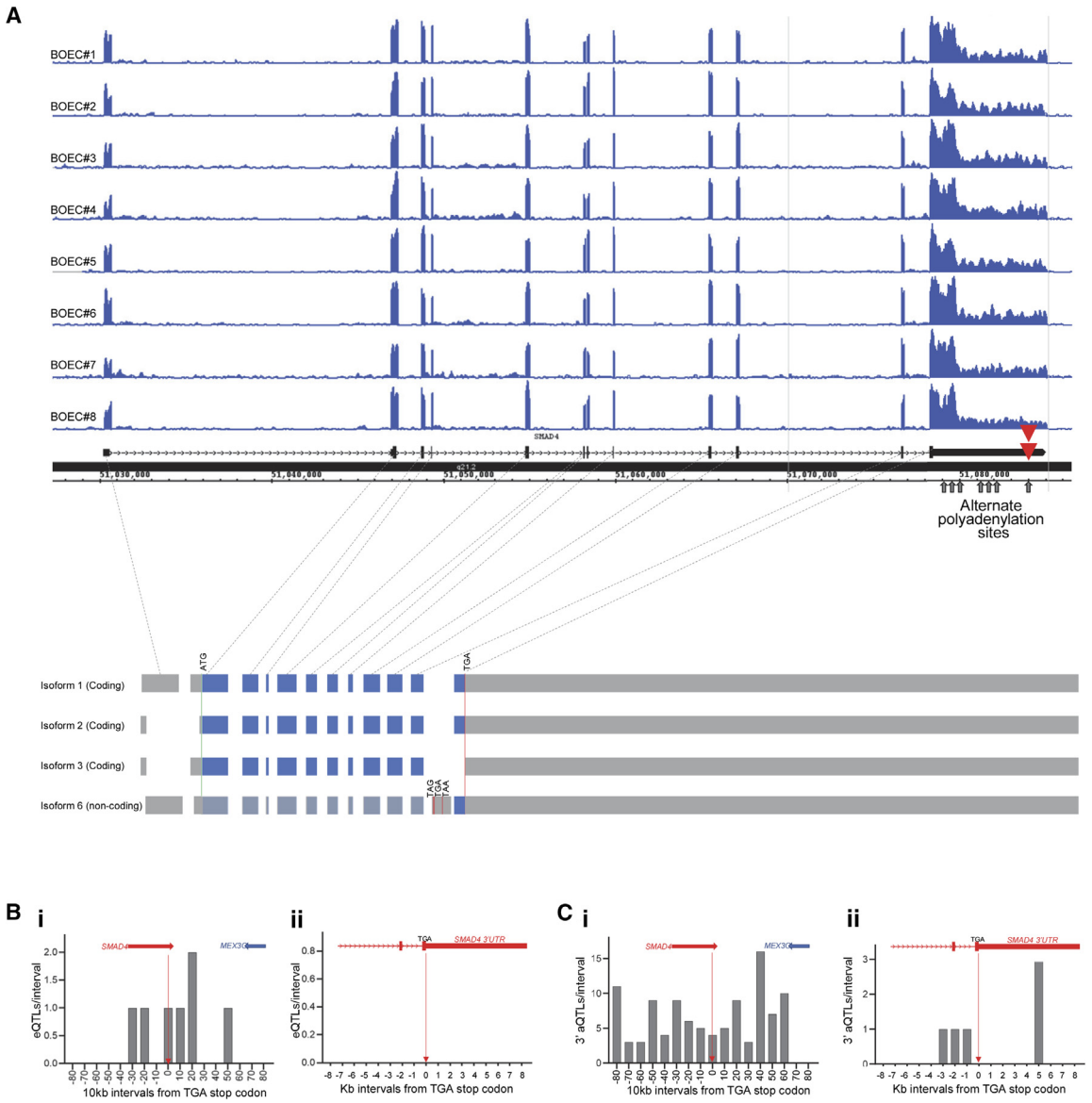

Figure 5. Expression of *SMAD4* in primary human cells

(A) Endothelial *SMAD4* total RNA expression. RNA-seq data from 8 different cultures of blood outgrowth endothelial cells (BOECs) with normal *SMAD4* sequence.[36] The consistent peaks sharply define exon boundaries in GRCh38. The cartoon below links the RNA-seq expression by gray dotted lines to the main (upper) and alternate *SMAD4* RefSeq[64] splice isoforms that share the final UTR-containing exon, with exons to scale. Blue, coding; gray, non-coding regions. Sites of start and stop codons, unique filtered variants (red triangles), and the seven alternate cleavage and polyadenylation sites at c.3121, c.3487, c.3791, c.5186, c.5452, c.5615, and c.7709 are also highlighted. Note although isoform 6 shares the majority of nucleotides with isoforms 1, 2, and 3, it does not share the same ribosomal reading frame and contains a unique penultimate exon with stop codons in all 3 reading frames that enhance fidelity as a non-coding transcript.

(B) Number of general population *SMAD4* expression QTLs (eQTLs) per interval of DNA flanking the TGA stop codon, as listed by UCSC CAVIAR tracks[26,27] for data from the Genotype Tissue Expression(GTEx) project.[39] The graphs are centered on the *SMAD4* natural stop codon site (vertical red arrow), with relevant gene loci indicated to scale horizontally above graphs. (i) Overview of *SMAD4* locus and flanking regions at 10 kb intervals. (ii) Magnified view of penultimate and final exons/adjacent introns at 1 kb intervals.

(C) Number of general population 3′ UTR alternative polyadenylation QTLs (3′aQTLs[40]) per kilobase of DNA flanking the TGA stop codon, as determined in GTEx.[39] (i) Overview of *SMAD4* locus and flanking regions at 10 kb intervals. (ii) Magnified view of penultimate and final exons/adjacent introns at 1 kb intervals.

gastric polyposis, HHT nosebleeds, HHT mucocutaneous telangiectasia, pulmonary AVMs treated by embolization, and antecedent JPHT family history. However, no *SMAD4* variant had been identified by clinical service panel testing, the 100,000 Genomes Project clinical pipelines, or by GROFFFY. Total PBMC *SMAD4* expression levels were lower than control subjects (Figure 8Aiii), and similar to variant 1 (Figures 8Ai and 8Aiii) with additional similarities to variants 1 and 2 post cycloheximide (Figure 8E). A new team member blinded to the findings and project

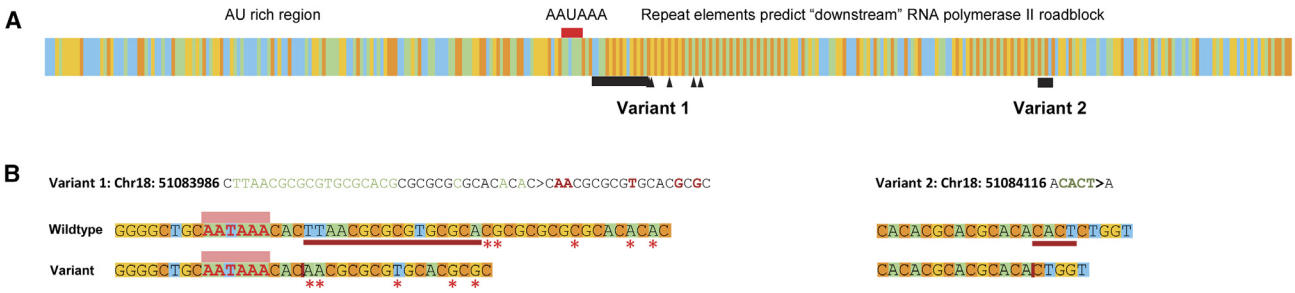

Figure 6. Schematic of *SMAD4* 3′ UTR variants in the context of RNA function

(A) Color-coded nucleotides 7561–7920 of the *SMAD4* main coding transcript GenBank: NM_005359.[64] These span the final AAUAAA hexamer (red bar) and include the upstream AU-rich (blue/green) region, downstream repetitive elements, and sites of the two GROFFFY-identified variants. Deleted residues are indicated by black bars, missense substitutions as black triangles. For FASTA format sequences, see Table S8.

(B) Variant 1 (chr18:51,083,986 CTTAACGCGCGTGCGCACGCGCGCGCACACA>CAACGCGCGTGCACGCG) and variant 2 (chr18:51,084,116 ACACT>A) in detail. The AAUAAA hexamer is shown by DNA sequence (AATAAA) and highlighted by a pink box, deleted residues by red underline (wild-type) or vertical red line (variants), and missense substitutions as red stars. Table S8 provides further sequence details.

was invited to examine the raw *SMAD4* WGS data in the binary alignment map (bam) file and identified a single *SMAD4* exonic variant in the donor's DNA (Figure S7). This was sited between variants 1 and 2 in the 3′ UTR, with the complex insertion/rearrangement separated from variant 1 by only two bases (Figure S7).

Discussion

We have presented and validated a system that synthesizes biologically generated signals of function in order to filter out variants in DNA regions with no such evidence of functionality. This generically applicable method was highly effective in reducing the number of WGS variants from almost 5 million per individual to an average of ~21,000. Critically, the method retained pathogenic variants already known in a validation dataset and identified ultra-rare, disease-associated variants in the distal *SMAD4* 3′ UTR. These variants disrupted RNA secondary structures required for cleavage and polyadenylation, and subsequent RNA-seq and clinical correlations supported *SMAD4* etiology.

Study strengths include the development and application of an unbiased, genome-wide method with no prior assumptions. Of other variant filtration methods already used in WGS, most depend on union and intersection rules of existing annotation tracks. The candidate *cis*-regulatory elements file produced by ENCODE has been particularly favored with its specific predictions of each possible CRE position and size. By using the raw biological data providing broader areas for inclusion, GROFFFY may better suit the purpose of a first pass filter for definition of variants worthy of further study than computational predicted files with potential false negatives. Simultaneous evaluation of WGS data from nearly 100 individuals with a similar clinical phenotype enabled resource direction to unstudied non-coding sequences where multiple rare, high-impact variants were identified. Study strength was further augmented by replicate RNA-seq expression data from primary human endothelial cells, the cell type responsible for the *SMAD4* clinical phenotype (HHT) where causal loss-of-function variants were being sought, and Genomics England clinician contact pathways that identified *SMAD4*-specific phenotypes after the draft manuscript was approved for submission. This also enabled recontact, leading to evidence from sequenced individuals' PBMCs that support perturbations in *SMAD4* RNA expression. We do not expect the PBMC responses to be a complete model of the variant effects in all pathobiological contexts, but they are presented in order to provide functional evidence of molecular impact. Additionally, extensive open-source datasets and code enabled exploration of common human variation responsible for *SMAD4* QTLs that contained exons where expression was impacted by the identified variants, while the variants themselves highlighted an emerging field in biology that has had limited recognition in medicine.

A potential study weakness, the presented discovery elements that focus on a single gene, can be justified because of the immediate pathway to translational impact. In addition to somatic cancer genetic diagnostics, early diagnosis of a germline heterozygous *SMAD4* loss-of-function allele offers proven methods to save lives and emergency healthcare resources by institution of gastrointestinal (from adolescence) and aortic screens,[10] in addition to standard HHT screening and pre-symptomatic interventions.[11,12] It is not possible to perform further segregation analyses in these families as all known affected relatives were in the antecedent generations and deceased. Two of these ultra-rare variants have been detected previously (rs1599209874, absent in gnomAD, TOMMO MAF of 0.00006; rs1375437193, gnomAD MAF of 0.000071). Since the phenotypes are late onset, it is very likely that such variants could be identified in members of the population who did not yet have a clinical diagnosis. Thus, while detailed mechanistic dissection can be

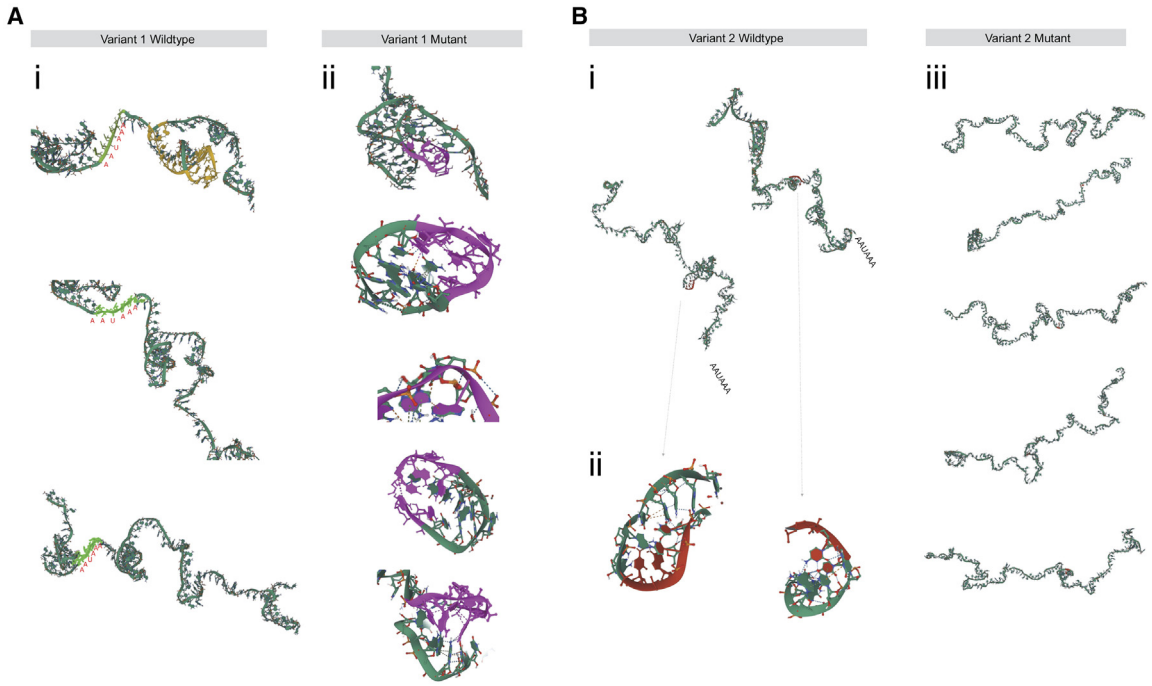

Figure 7. Replicate iFoldRNA structures

(A) Variant 1. (i) Three representative iFoldRNA[43,44] simulations of the wild-type 150 nucleotides spanning the AAUAAA hexamer (light green), selected as those best illustrating the 3-dimensional relationships in two dimensions. The site of the 32-nucleotide deletion/insertion is highlighted in yellow within the upper structure where they are best demarcated. The lower structures provide further simulations highlighting in brighter green, the accessibility of the near-linear AAUAAA sequences. A camera spin is provided in Video S1. (ii) Magnified view of five separate simulations of variant 1 sequence with AAUAAA hexamer site highlighted in purple. All 5 simulations were consistent and showed the AAUAAA hexamer now inaccessible, incorporated into a secondary structure. A camera spin is provided in Video S2.

(B) Variant 2. (i/ii) Two representative simulations of wild-type sequence. The site of the 4 nucleotides deleted in the variant are highlighted in red in panoramic (i) and magnified (ii) views. (iii) Five simulations of the variant sequence.

the subject of future work, we suggest the presented data support immediate extension of the *SMAD4* regions included in biological and virtual gene panels for patients with HHT, juvenile polyposis, and cancer to include the 3′ UTR sequences flanking the final AAUAAA hexamer. For the HHT-affected individuals harboring the identified variants, there seems sufficient evidence for them to be considered as "likely *SMAD4* HHT" for at least one round of endoscopic and echocardiographic surveillance, while further functional studies are pending. For other HHT-affected individuals where conventional screening of HHT genes has not identified a causal variant, the possibility of undetected *SMAD4* variation can be considered.

Alternate polyadenylation has not been explored to date for *SMAD4* or for other heritable diseases beyond triplet expansion neurodegenerative diseases[65,68] Long 3′ UTRs with their abundance of regulatory motifs provide greater opportunity for regulatory control than short 3′ UTRs, while switching between alternate polyadenylation sites to provide shorter or longer 3′ UTRs is increasingly recognized to modify protein translation, for example differentially transporting mRNAs to condensates which can result in translation repression or enrichment in specified cellular regions or states.[65] Our data suggest this will be important for regulation of *SMAD4*, a ubiquitous and essential protein with diverse functions,[8–10] where ~7 kb of 3′ UTR is transcribed at high levels in coding and non-coding transcripts (Figures 5, 8, S5, and S6). Recent data highlight that polyadenylation sites differ in strength: weaker proximal CPA sites are used in genes with cell type-specific transcription (requiring transcriptional enhancers to strengthen CPA activity), while distal and single PAS sites are strongest to ensure mature mRNAs are produced.[69] As recently reviewed,[65] cleavage and polyadenylation occurs while RNA polymerase II (Pol II) is transcribing a gene and is regulated by Pol II elongation dynamics. Pol II pausing immediately downstream to a final AAUAAA hexamer CPA cleavage site is necessary in order to enable CPA complex assembly and co-transcriptional addition of the "poly-A tail" that is essential for mRNA generation and subsequent protein translation (Figure 6). If at the final polyadenylation site, the full cycle of polymerase pausing, CPA complex binding, and cleavage/polyadenylation is impaired, different sites and efficiency of polyadenylation would modify function. Our current data examining 1 h stress responses (when the cell has to rely predominantly on reuse of existing RNA transcripts) highlight further mechanisms to explore.

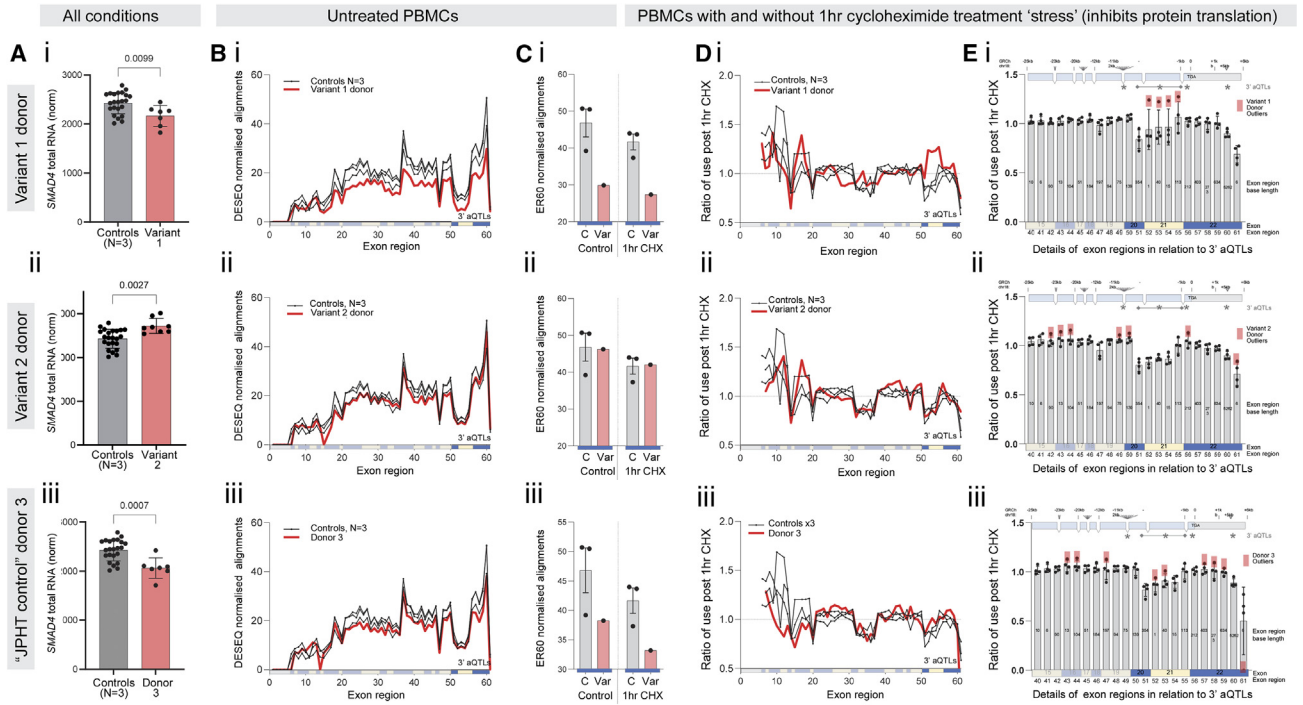

Figure 8. ***SMAD4*** **RNA expression in ribosomal (r)RNA-depleted libraries from 3 controls compared to affected donors from the 3 separate HHT-affected families**

Shown are variant 1 (i), variant 2 (ii), and an unsolved JPHT (MIM: 175050)-positive control (iii). For preceding methodological data on the rRNA depleted libraries, see Figures S3–S6. RNA from peripheral blood mononuclear cells (PBMCs) cultured with and without 3 different 1 h stresses (Figure S3): *SMAD4* splice site changes met DEXSeq2[59] significance after cycloheximide (CHX).

(A) Total *SMAD4* RNA from control (gray, n = 23) and *SMAD4* variant-affected donors (red: i) variant 1 n = 7; (ii) variant 2 n = 8; (iii) donor 3, n = 7) following DESeq2 normalization[58,67] using GINI housekeeper genes.[60,61] Note contrast between variant 1 (i) and 2 (ii) donors but similarities between variant 1 (i) and donor 3 (iii).

(B–E) DEXSEQ[59] splicing patterns across 61 exon regions in 22 *SMAD4* exons.

(B) Exon region (ER) use in untreated PBMCs by donor, plotting data from the affected individuals (red) and the same 3 control subjects (black). Exons are color coded to highlight 3′aQTL loci.[40]

(C) Use of ER60, the variant-containing 3′ UTR region in untreated and CHX-treated PBMCs: note again similarities in (i) and (iii).

(D) The ratio of exon region use between CHX-treated and untreated PBMCs, plotted as in (B).

(E) The ratios in the final 8 exons (ER40–ER61) containing all ERs differentially used by the variant 1 and 2 donors after CHX. Each graph is annotated with the genomic DNA origins and kb markers (upper bar); sites of the 3′aQTLs (*, see Figure 5C); and variant outlier values (red) that were not accompanied by increased polyadenylated transcripts (Figure S3).

These include 3′ UTR variant impacts on alternate splice site selection and maintenance of polyadenylated transcripts that may be less successfully achieved in the setting of stress conditions necessitating rapid changes (Figure S3A). The potential to facilitate future development of 3′ UTR therapeutics is augmented given the fact that repetitive regions of pol II "roadblocks" provide fertile and previously hidden substrates for impactful human DNA variation.

In conclusion, we present and validate a filter that reduces the overwhelming number of variants identified by WGS while retaining functional genome variation of importance to clinical diagnostics. Exposure of non-coding variants in the top 10 percentile of deleteriousness and clusters in unexplored genomic regions enhances the near-term value of WGS. The GROFFFY filter enabled identification of rare *SMAD4* variants that disrupt the final site for RNA cleavage and polyadenylation, necessary for protein production. However, the full extent to which rare stress impact, functional alternate polyadenylation site (SIFAPS) variants contribute to diseases will be exposed only if untranslated sequences spanning the sites are included in virtual and physical diagnostic gene panels. Wider use of WGS and inclusion of 3′aQTL UTR regions in exome-based sequencing are recommended to capture relevant disease-specific variants.

Data and code availability

The publicly available file accession numbers used to generate the code are provided in full in supplemental methodsTables S3 and S4 and are available at the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA 596860. Primary WGS data from the 100,000 Genomes Project, which are held in a secure research environment, are available to registered users. Please see https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access for further information.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.09.005.

Acknowledgments

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The work was cofounded by the National Institute for Health Research Imperial Biomedical Research Centre, the D'Almeida Charitable Trust, and Imperial College Healthcare NHS Trust. A.A. was supported by Prince Sultan Military Medical City, Saudi Arabia. M.A.A. was supported by the National Institutes of Health (grant R35HL140019). The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. We thank the National Health Service staff of the UK Genomic Medicine Centres and the participants for their willing participation; the Genomics England Clinical Research Interface team, specifically Susan Walker, for separately reviewing bam file variant sequences; Charlotte Bevan, Michael Hubank, and Santiago Vernia for helpful discussions and manuscript review; and our academic and public partners within the NIHR Imperial BRC's Social Genetic and Environmental Determinants of Health (SGE) theme. We specifically thank the presented families for confirmation of their clinical phenotypes and consent to share in this manuscript. The views expressed are those of the authors and not necessarily those of funders, the NHS, the NIHR, or the Department of Health and Social Care.

Author contributions

Conceptualization, S.X., C.L.S.; methodology, S.X., Z.K., D.M., D.P., A.M.B., M.E.B.-H., A.A., M.A.A., N.V., M.J.C., GERC, C.L.S.; investigation, S.X., Z.K., D.M., D.L., A.D.M., S.K.W., C.L.S.; visualization, S.X., C.L.S.; funding acquisition, C.L.S.; project administration, GERC, C.L.S.; supervision, D.P., M.E.B.-H., M.A.A., C.L.S.; writing – original draft, C.L.S.; writing – review & editing, S.X., Z.K., D.M., D.L., D.P., A.M.B., M.E.B.-H., A.A., A.D.M., S.K.W., M.A.A., N.V., M.J.C., GERC, C.L.S.

S.X. devised and generated the GROFFFY approach, devised all scripts to generate GROFFFY, and generated all GROFFFY numeric data, Figures 1, 2, 3, S1, and S2, and Tables S1, S2, S3, S4, S5, and S6. Z.K. advised on Linux and script generation. D.M. interrogated donor 3 bam files. D.L. assisted in PBMC cultures. D.P., A.M.B., M.E.B.-H., and M.A.A. performed BOEC cultures and RNA preparations. A.A. designed primers for validations. A.D.M. contributed to recruitment of affected individuals. S.K.W. contributed to clinical correlations. N.V. advised on *SMAD4* regulation. M.J.C. contributed to specific project set up at Genomics England. GERC performed all whole-genome sequencing and alignments. C.L.S. recruited patients and performed clinical correlations; devised concepts and advised on GROFFFY approaches; devised and performed PBMC cultures; devised and performed in-house endothelial and PBMC RNA-seq and variant level data analyses; generated Figure 3, 4, 5, 6, 7, 8, S3, S4, S5, S6, and S7, and Tables S6–S8, and wrote the manuscript. All authors have reviewed and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Received: May 28, 2023
Accepted: September 8, 2023
Published: October 9, 2023

Web resources

Data Structures for Statistical Computing in Python, http://conference.scipy.org/proceedings/scipy2010/mckinney.html
Ensembl Variant Predictor: http://grch37.ensembl.org/info/docs/tools/vep/index.html
GENCODE Human Genome Release 31, https://www.gencodegenes.org/human/
Genome Reference Consortium Human Build 38, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/
LINUX, https://opensource.com/resources/linux
Online Mendelian Inheritance in Man, https://www.omim.org/
Perl 5.24, https://docs.activestate.com/activeperl/5.24/get/relnotes/
R: The R Project for Statistical Computing, https://www.r-project.org/
RepeatMasker: http://www.repeatmasker.org
RStudio, https://www.rstudio.com/
The Comprehensive R Archive Network, https://cran.r-project.org/
The National Genomics Research and Healthcare Knowledgebase v5 (2019) Genomics England, https://doi.org/10.6084/m9.figshare.4530893.v5
The Protein DataBank, https://www.rcsb.org/3d-view

References

1. Ransohoff, J.D., Wei, Y., and Khavari, P.A. (2018). The functions and unique features of long intergenic non-coding RNA. Nat. Rev. Mol. Cell Biol. *19*, 143–157.
2. Marchal, C., Sima, J., and Gilbert, D.M. (2019). Control of DNA replication timing in the 3D genome. Nat. Rev. Mol. Cell Biol. *20*, 721–737.
3. Halley, M.C., Ashley, E.A., and Tabor, H.K. (2022). Supporting undiagnosed participants when clinical genomics studies end. Nat. Genet. *54*, 1063–1065.
4. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.
5. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.
6. Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol. *7*, S4.1–S4.9.
7. Turnbull, C., Scott, R.H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F.B., Halai, D., Baple, E., Craig, C., Hamblin, A., et al. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ *361*, k1687.
8. Dai, C.J., Cao, Y.T., Huang, F., and Wang, Y.G. (2022). Multiple roles of mothers against decapentaplegic homolog 4 in

tumorigenesis, stem cells, drug resistance, and cancer therapy. World J. Stem Cells *14*, 41–53.
9. Hayashi, A., Hong, J., and Iacobuzio-Donahue, C.A. (2021). The pancreatic cancer genome revisited. Nat. Rev. Gastroenterol. Hepatol. *18*, 469–481.
10. Haidle, L., MacFarland, J.S.P., and Howe, J.R. (2022). Juvenile Polyposis Syndrome. In GeneReviews® [Internet], M.P. Adam, D.B. Everman, and G.M. Mirzaa, eds. (University of Washington, Seattle), pp. 1993–2022.
11. Shovlin, C.L., Buscarini, E., Sabbà, C., Mager, H.J., Kjeldsen, A.D., Pagella, F., Sure, U., Ugolini, S., Torring, P.M., Suppressa, P., et al. (2022). The European Rare Disease Network for HHT Frameworks for management of hereditary haemorrhagic telangiectasia in general and speciality care. Eur. J. Med. Genet. *65*, 104370.
12. Faughnan, M.E., Mager, J.J., Hetts, S.W., Palda, V.A., Lang-Robertson, K., Buscarini, E., Deslandres, E., Kasthuri, R.S., Lausman, A., Poetker, D., et al. (2020). Second International Guidelines for the Diagnosis and Management of Hereditary Hemorrhagic Telangiectasia. Ann. Intern. Med. *173*, 989–1001.
13. Clarke, J.M., Alikian, M., Xiao, S., Kasperaviciute, D., Thomas, E., Turbin, I., Olupona, K., Cifra, E., Curetean, E., Ferguson, T., et al. (2020). Low grade mosaicism in hereditary haemorrhagic telangiectasia identified by bidirectional whole genome sequencing reads through the 100,000 Genomes Project clinical diagnostic pipeline. J. Med. Genet. *57*, 859–862.
14. Balachandar, S., Graves, T.J., Shimonty, A., Kerr, K., Kilner, J., Xiao, S., Slade, R., Sroya, M., Alikian, M., Curetean, E., et al. (2022). Identification and validation of a novel pathogenic variant in GDF2 (BMP9) responsible for hereditary hemorrhagic telangiectasia and pulmonary arteriovenous malformations. Am. J. Med. Genet. *188*, 959–964.
15. Joyce, K.E., Onabanjo, E., Brownlow, S., Nur, F., Olupona, K., Fakayode, K., Sroya, M., Thomas, G.A., Ferguson, T., Redhead, J., et al. (2022). Whole genome sequences discriminate hereditary hemorrhagic telangiectasia phenotypes by non–HHT deleterious DNA variation. Blood Adv. *6*, 3956–3969.
16. Shovlin, C.L., Almaghlouth, F.I., Alsafi, A., Coote, N., Rennie, C., Wallace, G.M., Govani, F.S., and Research Consortium, G.E. (2023). Updates on diagnostic criteria for hereditary haemorrhagic telangiectasia in the light of whole genome sequencing of "Gene Negative" individuals recruited to the 100,000 Genomes Project. J. Med. Genet. *16*. 2023-109195.
17. Sharma, L., Almaghlouth, F., Mckernan, H., Springett, J., Tighe, H.C., and Shovlin, C.L. (2023). Iron deficiency responses and integrated compensations in patients according to hereditary haemorrhagic telangiectasia *ACVRL1*, *ENG* and *SMAD4* genotypes. Haematologica. https://doi.org/10.3324/haematol.2022.282038.
18. Shovlin, C.L., Simeoni, I., Downes, K., Frazer, Z.C., Megy, K., Bernabeu-Herrero, M.E., Shurr, A., Brimley, J., Patel, D., Kell, L., et al. (2020). Mutational and phenotypic characterization of hereditary hemorrhagic telangiectasia. Blood *136*, 1907–1918.
19. Volders, P.J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. *47*, D135–D139.
20. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47*, D155–D162.
21. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354.
22. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.
23. Mundade, R., Ozer, H.G., Wei, H., Prabhu, L., and Lu, T. (2014). Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle *13*, 2847–2852.
24. Koch, L. (2017). Cancer genetics: A 3D view of genome rearrangements. Nat. Rev. Genet. *18*, 456.
25. Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.
26. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
27. Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., et al. (2023). The UCSC Genome Browser database: 2023 update. Nucl. Acid Res. *51*, D1188–D1195.
28. Raczy, C., Petrovski, R., Saunders, C.T., Chorny, I., Kruglyak, S., Margulies, E.H., Chuang, H.Y., Källberg, M., Kumar, S.A., Liao, A., et al. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics *29*, 2041–2043.
29. Quinlan, A.R., and Hall, I.M.((2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
30. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
31. Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P.; and 1000 Genomes Project Consortium (2017). 1000 Genomes Project Consortium, Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience *6*, 1–8.
32. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
33. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.
34. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., 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*, 405–424.
35. Mukhtar, G., and Shovlin, C.L. (2023). Unsupervised machine learning algorithms identify expected haemorrhage relationships but define unexplained coagulation profiles mapping to thrombotic phenotypes in hereditary haemorrhagic telangiectasia. EJHaem *4*, 602–611.

36. Bernabeu-Herrero, M.E., Patel, D., Bielowka, A., Chaves Guerrero, P., Marciniak, S.J., Noseda, M., Aldred, M.A., and Shovlin, C.L. (2023). Heterozygous transcriptional signatures unmask variable premature termination codon (PTC) burden alongside pathway-specific adaptations in blood outgrowth endothelial cells from patients with nonsense DNA variants causing hereditary hemorrhagic telangiectasia. Preprint at bioRxiv. https://doi.org/10.1101/2021.12.05.471269.
37. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. *46*, W537–W544.
38. Freese, N.H., Norris, D.C., and Loraine, A.E. (2016). Integrated genome browser: visual analytics platform for genomics. Bioinformatics *32*, 2089–2095.
39. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
40. Li, L., Huang, K.L., Gao, Y., Cui, Y., Wang, G., Elrod, N.D., Li, Y., Chen, Y.E., Ji, P., Peng, F., et al. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat. Genet. *53*, 994–1005.
41. McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., and Bartel, D.P. (2019). The biochemical basis of microRNA targeting efficacy. Science *366*, eaav1741.
42. Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. *48*, D127–D131.
43. Ding, F., Sharma, S., Chalasani, P., Demidov, V.V., Broude, N.E., and Dokholyan, N.V.((2008). Large scale simulations of 3D RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. RNA *14*, 1164–1173.
44. Krokhotin, A., Houlihan, K., and Dokholyan, N.V. (2015). iFoldRNA v2: folding RNA with constraints. Bioinformatics *31*, 2891–2893.
45. Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koča, J., and Rose, A.S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. *49*, W431–W437.
46. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
47. Shi, C. (2012). Developing a New Mutational Analysis System Using Hereditary Haemorrhagic Telangiectasia as a Genetic Model (Imperial College London). MSc Thesis.
48. Li, Y. (2013). Gene Mutations and Transcripts in Hereditary Haemorrhagic Telangiectasia (HHT) (Imperial College London). MSc Thesis.
49. Shurr, A.Y.L., Maurer, C., Turbin, I.G., Bernabeu-Herrero, M.E., Aldred, M., Patel, D., and Shovlin, C.L. (2019). Addressing the problem of variants of uncertain significance in genetic diagnosis of vascular pulmonary disease: a role for transcript expression in blood monocytes? Thorax *74*, A152.
50. Duong, H., and Patel, G. (2022). Hypothermia. In StatPearls (StatPearls Publishing). https://www.ncbi.nlm.nih.gov/books/NBK545239/.
51. Perman, S.M., Bartos, J.A., Del Rios, M., Donnino, M.W., Hirsch, K.G., Jentzer, J.C., Kudenchuk, P.J., Kurz, M.C., Maciel, C.B., Menon, V., et al. (2023). Temperature Management for Comatose Adult Survivors of Cardiac Arrest: A Science Advisory from the American Heart Association. Circulation *148*, 982–988.
52. Govani, F.S., Giess, A., Mollet, I.G., Begbie, M.E., Jones, M.D., Game, L., and Shovlin, C.L. (2013). Directional next-generation RNA sequencing and examination of premature termination codon mutations in endoglin/hereditary haemorrhagic telangiectasia. Mol. Syndromol. *4*, 184–196.
53. Mollet, I.G., Patel, D., Govani, F.S., Giess, A., Paschalaki, K., Periyasamy, M., Lidington, E.C., Mason, J.C., Jones, M.D., Game, L., et al. (2016). Low dose iron treatments induce a DNA damage response in human endothelial cells within minutes. PLoS One *11*, e0147990.
54. Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B., and Liu, J.O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat. Chem. Biol. *6*, 209–217.
55. Shen, L., Su, Z., Yang, K., Wu, C., Becker, T., Bell-Pedersen, D., Zhang, J., and Sachs, M.S. (2021). Structure of the translating Neurospora ribosome arrested by cycloheximide. Proc. Natl. Acad. Sci. USA *118*, e2111862118.
56. Kartikasari, A.E.R., Georgiou, N.A., Visseren, F.L.J., van Kats-Renaud, H., van Asbeck, B.S., and Marx, J.J.M. (2006). Endothelial activation and induction of monocyte adhesion by nontransferrin-bound iron present in human sera. FASEB J *20*, 353–355.
57. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
58. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.
59. Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.
60. Wright Muelas, M., Mughal, F., O'Hagan, S., Day, P.J., and Kell, D.B. (2019). The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. Sci. Rep. *9*, 17960.
61. O'Hagan, S., Wright Muelas, M., Day, P.J., Lundberg, E., and Kell, D.B. (2018). GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. Cell Syst. *6*, 230–244.e1.
62. Reed, G.F., Lynn, F., and Meade, B.D. (2002). Use of coefficient of variation in assessing variability of quantitative assays. Clin. Diagn. Lab. Immunol. *9*, 1235–1239.
63. 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.
64. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Farrell, C.M., Feldgarden, M., Fine, A.M., Funk, K., et al. (2023). Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res. *51*, D29–D38.
65. Mitschka, S., and Mayr, C. (2022). Context-specific regulation and function of mRNA alternative polyadenylation. Nat. Rev. Mol. Cell Biol. *7*, 1–18.
66. Shovlin, C.L. (2010). Hereditary haemorrhagic telangiectasia: pathophysiology, diagnosis and treatment. Blood Rev. *24*, 203–219.

67. Evans, C., Hardin, J., and Stoebel, D.M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief. Bioinform. *19*, 776–792.
68. Li, Y., Li, J., Wang, J., Zhang, S., Giles, K., Prakash, T.P., Rigo, F., Napierala, J.S., and Napierala, M. (2022). Premature transcription termination at the expanded GAA repeats and aberrant alternative polyadenylation contributes to the Frataxin transcriptional deficit in Friedreich's ataxia. Hum. Mol. Genet. *31*, 3539–3557.
69. Kwon, B., Fansler, M.M., Patel, N.D., Lee, J., Ma, W., and Mayr, C. (2022). Enhancers regulate 3′ end processing activity to control expression of alternative 3′UTR isoforms. Nat. Commun. *13*, 2709.