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

Sihao Xiao,[1,2†*], Zhentian Kai,[3] Daniel Murphy,[2,4] Dongyang Li,[1,2] Dilip Patel,[1,2] Adrianna Bielowka,[1,2] Maria E. Bernabeu-Herrero,[1,2] Awatif Abdulmogith,[1,2] Andrew D Mumford,[5] Sarah 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*]

Affiliations:

[1]National Heart and Lung Institute, Imperial College London, London W12 ONN, UK.

[2] National Institute for Health Research (NIHR) Imperial Biomedical Research Centre, London, W2 1NY UK.

[3]Topgen Biopharm Technology Co. Ltd; Shanghai, 201203, China.

[4] Women's, Children's & Clinical Support (Pharmacy), Imperial College Healthcare NHS Trust, London, W2 1NY, UK

[5]School of Molecular and Cellular Medicine, University of Bristol, Bristol BS8 1QU, 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, Aberdeen AB24 3FX, UK.

[8] William Harvey Research Institute, Queen Mary University of London, London E1 4NS, UK.

[9] Genomics England, London EC1M 6BQ, UK,

[10] Specialist Medicine, Imperial College Healthcare NHS Trust; London, W12 OHS, UK

*Correspondence c.shovlin@imperial.ac.uk (*email to include both before and after publication – this author will deal with manuscript submission).* (*include after publication* sihao.xiao@bnc.ox.ac.uk)

† Current address: Big Data Institute, University of Oxford, Oxford, UK.

‡ A full list of these authors is provided at the end of the main manuscript

Abstract: Despite whole genome sequencing (WGS), many single gene disorder cases 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 which 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 per 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 proto-oncogene *SMAD4*, where germline loss-of-function alleles cause combined HHT and colonic polyposis (MIM: 175050). Sited >5.4kb 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 in 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 ~~patients~~ 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 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 **g**enomic **r**egions **o**f **f**unctionality **f**ilter **f**or priorit**y** (*GROFFFY)* based on published experimental data particularly from ENCODE[4-6] and performed validation and discovery analyses in WGS data from ~~patients~~ individuals recruited to the 100,000 Genomes Project.[7] To accelerate clinical impact, we focused discovery analyses on noncoding regions of a proto-oncogene 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.[8,9] As indicated by its function and earlier gene names (*DPC4*-deleted in pancreatic cancer 4; *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.[8,9] 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].[~~10~~] Where *SMAD4* is identified (juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome, JPHT [MIM: 175050]),[10,11,~~12~~] this allows ~~patients~~ affected-individuals to benefit from life-long polyposis and aortopathy screening programs.[9] 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 ~~new~~ WGS analytic approach identifies a ~~novel~~ type of functional DNA variant uncaptured by usual clinical sequencing methodologies.

MATERIALS 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 Study Design are outlined in Figure 1. Following earlier recruitment of individuals with hereditary hemorrhagic telangiectasia (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.

Patient recruitment and 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) patients. The study received ethical approval from the Health Research Authority (HRA) Committee East England-Cambridge South (REC Ref 14/EE/1112), and all participants provided written consent. Anonymized raw sequencing data are available in the Genomics England Research Environment.[7,10] Separately, in a clinical diagnostic pipeline, Genomics England performed data alignments and variant classifications fed back to recruiting clinicians.

The cohort recruited with hereditary hemorrhagic telangiectasia (HHT)[10,11,12] 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 *ACVRL1*, *ENG* or *SMAD4*,[1,2,3,17,8] and discovery dataset of 98 WGS

sequences where some DNAs were expected to have heterozygous loss-of-function variants in these genes.

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,[19] and excluded data originating in cancer cells. Regions being selected for are described in *Tables S1-S4* which provide full details of coordinate derivation from transcribed loci and candidate regulatory element (cRE) regions.[6,18,19,20,21] Following merging of 18,828 bed files from 3,454 experiments,[5,6] sequences in the ENCODE blacklist[20] and RepeatMasker[21] 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[4,5,24] and the NIH Roadmap Epigenomics Consortia.[25,22] We downloaded the call sets itemized in *Table S3* and *Table S4* from the ENCODE portal.[4,5] By merging DNA binding call sets[24] and representative DNase hypersensitivity site (rDHS) call sets,[26,23] 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' genome are usually heavily modified and rearranged.[27,24]

a) 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 representing 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.

b) 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 (V2.4.26)[~~28~~ 25] to obtain all accessible DNA regions.

c) Region coordinates for CpG islands were downloaded from the UCSC database[~~29,30~~ 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[~~31~~] system. The converted bed file was then sorted by location using BEDOPS.[~~28~~25]

Genomic coordinates for transcribed loci were extracted as follows:

d) GENCODE human genome annotation version 31[~~30~~] for GRCh38[~~19~~] was downloaded from the UCSC database[~~31,32~~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.[28,25]

e) Long non-coding RNA annotations (lncRNA) were downloaded from both GENCODE (release 31)[30] and lncipedia database (version 5.2).[20,18] LncRNA coordinates were merged together using BEDOPS[28] BEDOPS[25] to obtain all possible lncRNA gene regions.

f) Micro RNA (miRNA) annotations were downloaded from miRbase (release 22.1).[21,19] The downloaded file was in gff3 format and was converted to bed files using BEDOPS.[28,25]

Genomic coordinates of regions excluded were identified from the ENCODE blacklist which was downloaded directly from the ENCODE[20] project website in bed format, and RepeatMasker[34] which was downloaded from the UCSC database[31,32,26,27] 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 1,423,480,943bp approximating to 44.48% of GRCh38.

To assimilate for GROFFFY, separately, the Ubuntu shell (version 16.04.2 LTS based on Linux 4.4.0-64-generic x86_64 system) was launched for the Genomics England Research Environment,[10,7] 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[29] and Starling for sequence alignment, and to identify variants. The output files were in vcf files with .gz compression and decompressed using Gzip (Version 1.6).[37] Pandas module version 0.22.0 [38,30] in Python (version 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 version 2.26.0 [39-31] 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) version 96.3,[40-32] based on Perl version 5.24,[41] SAMtools version 1.[5-42-5][33] specifically SAMtools HTSlib version 1.5,[42-33] 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 version 3.5.1,[43] within R studio version 13.4.0 [44] downloaded from the Comprehensive R Archive Network[45] was used to perform statistical tests. Paired datasets were analysed 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,[22-20] and RepeatMasker[23] were subtracted from regions selected for using the ‘difference' option in BEDOPS.[28-25] 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 Figure S1 and Figure S2, stepwise filters excluded variants where general population allele frequency exceeded 0.0002 in the 1000 Genome Project[34] or gnomAD[35] 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.[~~48~~ 36] 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 prioritise 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~~12 ~~16~~15,~~49~~ 37] 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 shown to be identical. The variant IDs were integrated in a white-list. Exclusion of these white-listed variants in other patients was performed using the isin function of Pandas[30] 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[36] in the enclosed Research Environment: databases for SNV annotation (version 1.5) and small indel annotation (version 1.5) which were pre-installed in the Research Environment were indexed. As the annotation of CADD score was quite slow, the

process was put towards 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 AirLock (*Table S6*), variant genomic coordinates were visualized in the UCSC Genome Browser.[31,32]26,27 Endothelial expression of *SMAD4* was examined in whole transcriptome data from primary human BOECs.[50]–[38] Binary sequence alignment map (bam) files aligned to GRCh38[19] were analyzed in Galaxy Version 2.4.1[51]-1[39] and the Integrated Genome Browser (IGB) 9.1.8.[52] [40]

3'UTR alternative polyadenylation quantitative trait loci (3'aQTLs) from 46 tissues isolated from 467 individuals in GTEx[41] were sourced through the 3'aQTL Atlas.[54] [42] Genetic variants likely affecting gene expression in GTEx V8[41] data release were captured from UCSC[31,32] [26,27] CAVIAR tracks, which define high confidence gene expression QTLs within 1MB of gene transcription start sites (cis-eQTLs).[31,32]

All variants were independently verified by Genomics England. Impact on microRNA binding sites was examined through TargetScan Human Release 8.0[55]-0[43] and miRDB.[56]-[44] RNA structure predictions were performed using iFoldRNA v2.0[57,58]-[45,46] without restraints, and final models were visualized using Mol* Viewer[47] via the Research Collaboratory for Structural Bioinformatics Protein Data Bank server.[60]—[48]

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 ~~patient~~ individual recruited to the 100,000 Genomes Project with a JPHT ~~*SMAD4*$^{+/-}$~~ [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,[61-63]-[49-51] peripheral blood mononuclear cells (PBMCs, 'monocytes') were prepared using BD Vacutainer® CPT™ tubes (Bunzl Healthcare,

Coalville, UK) according to manufacturer's instructions with minor modifications. As detailed further in the Appendix, 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.

Briefly, immediately after venesection, the blood was gently remixed by inverting 8-10 times, and centrifuged within 2 hours of collection for 30 minutes at 1600 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 50ml 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 37ºC for 10 minutes, then subjected to 4 different treatment conditions for 1 hour including control at 37ºC, and low temperature in a 32ºC waterbath for 1hr to mimic the stress incurred at the threshold between mild and moderate hypothermia.[~~64~~ 52] Additional stresses previously optimized in our laboratory[~~61–63~~49–51,~~65,66~~53,54] were inhibition of translation by cycloheximide 100µg/ml (cycloheximide inhibits eukaryotic translation elongation by mechanisms include binding to the 60S ribosomal subunit E-site[~~67,68~~55,56]) and a clinically-relevant mild reactive oxygen species (ROS) stress using ferric citrate 10µmol/L.[~~66,69~~54,57] After 1hr, all tubes were centrifuged at 520 RCF at room temperature for 15 minutes. Cell pellets were lysed in Tri reagent (Cambridge Bioscience Ltd, Cambridge, UK) 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 (Leipzig, Germany). For RNASeq 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 150bp reads (Genewiz, Leipzig, Germany). Sequenced reads were trimmed using Trimmomatic v.0.36,~~[70]~~[58] aligned to Homo sapiens GRCh38~~[19]~~ using STAR aligner v2.5.2b, and unique gene reads that fell within exon regions counted using Subread package v1.5.2 (Genewiz, Leipzig, Germany).

Blinded to the types of donors and treatments, Genewiz performed differential gene expression analyses using DESeq2,~~[71]~~[59] and differential exon expression using ~~DEXSeq~~~~[72]~~DEXSeq[60] to identify differentially spliced genes by testing for significant differences in read counts on exon regions (and junctions) of the genes. In DEXSeq,~~[72]~~[60] read counts are normalised 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.~~[72]~~[60] 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.~~[72]~~[60]

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[61,62]) were used to

evaluate individual library quality (Figure S4), and subsequently employed for DESeq2 normalisation. For these normalisations, the intra-assay coefficient of variation (CV, *100*standard deviation (SD)/mean*)[~~75~~ ~~63~~] was calculated for replicate pairs using alignment per gene adjusted for total read counts per library, and analyses 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~~[~~71~~] ~~–~~DeSeq2[59] normalised using the GINI[~~73,74~~ ~~,~~ 61,62] 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 v17.0 (Statacorp, College Station, Texas). Comparative statistics of the number of variants before and after filtration was performed using Mann Whitney two-group comparisons. RNASeq expression was analysed in STATA v17.0 (Statacorp, College Station, Texas) and GraphPad Prism 9 (GraphPad Software, San Diego, CA), compared using Kruskal Wallis and Dunn's post test applied for selected pairwise comparisons.

RESULTS

GROFFFY defines biologically validated regions of functionality

By only including 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 500kb 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 (Figure 3A, Figure 3B). Restricting to rare variants with population allele frequencies $<2x10^{-4}$ ~~[76]~~ [64] removed means of 2,476,589 ~~[46]~~ [34] and 2,483,377 ~~[47]~~ [35] variants/DNA according to database (Figure 3A, Figure 3B). After removing variants with a ~~CADD[48]~~ CADD[36] 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 ~~novel~~ 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 5,649bp distal to the *SMAD4* stop codon, and did not affect any microRNA binding sites.[~~55,56~~43,44] The wild-type sequences were consistently expressed in human primary blood outgrowth endothelial cells (BOECs) derived from donors with normal ~~SMAD4~~[~~50~~] SMAD4[38] (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)[~~53~~41] (Figure 5B), the variants were within the only kilobase of the 3'UTR to contain 3' alternate polyadenylation QTLs (3' aQTLs,[~~54~~ 42] 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)[~~77~~ 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,[~~77~~ 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 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[57,58 45,46] visualized using Mol* ~~Viewer~~[59] Viewer[47] via the Research Collaboratory for Structural Bioinformatics Protein Data Bank server,[60 –48] indicated that both deletion variants disturbed secondary structures that substantially altered the sequence context for CPA activity. In wildtype sequence, the AAUAAA hexamer was in a near-linear conformation with stacked pyrimidine and purine rings evident on magnified views (Figure 7Ai, Movie 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, Movie S2). In contrast, the second variant which deleted 4 nucleotides 134bp downstream of the AAUAAA hexamer, disrupted and unfolded the downstream structured region expected to be the major RNA polymerase II ~~roadblock~~[77] roadblock[65] (Figure 7B).

Clinical correlation

All three ~~patients~~ individuals with Variants 1 and 2 had clinically-confirmed HHT.[11 10-18 17,79 67] After identification of the *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, 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 ~~patient~~ 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 Data Supplement, peripheral blood mononuclear cells (PBMCs) were isolated from affected individuals with the 3 *SMAD4* variants and controls, and cultured in conditions predicted to modify 3'UTR use, before RNA sequencing. DESeq2[59] analyses of PolyA-selected RNAs indicated that *SMAD4* polyadenylated transcripts increased after a 1hr 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 high (*Figure S3*). This reduced after normalising with low GINI[73,74] [61,62] coefficient genes (*Figure S4*).

Whether normalised to read counts per library, or GINI[73,74] [61,62] 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[60] (Figure 8Bi). In controls, *SMAD4* transcript expression was modified following 1hr cycloheximide 100µg/mL, with lower use of exon region (ER)60 containing the AAUAAA site and variants, consistent with shorter 3'UTRs after stress (*Figure S5, Figure S6*). Despite this, ER60 use was further reduced in the Variant 1 donor after CHX (Figure 8Ci, Figure 8Di) with increased use of penultimate exon regions ER52-55 (Figure 8Di, Figure 78Ei), supporting different 1hr changes in RNA splicing on stress.

Total *SMAD4* RNA was higher in the "roadblock unfolding" Variant 2 donor than 3 normal controls across all conditions (Figure 8Aii**)**. Although exon region use was similar to controls in untreated PBMCs (Figure 8Bii, Figure 8Cii**)**, after 1hr cycloheximide, compared to controls there was higher use of regions corresponding to two of the 3' aQTLs (Figure 8Dii, Figure 8Eii). We concluded that the contrasting overall expression patterns were consistent with the opposing predictions following RNA modelling of Variants 1 and 2; that Variant 2 data also supported different 1hr 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 *SMAD4* positive control due to HHT-JP syndrome (colonic and gastric polyposis; HHT nosebleeds; HHT mucocutaneous telangiectasia, pulmonary AVMs treated by embolization, and antecedent HHT-JP 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 controls (Figure 8Aiii), and similar to Variant 1 (Figure 8Ai/iii) with additional similarities to Variants 1 and 2 post cycloheximide (Figure 8E). A new team member blinded to the findings and project, was invited to examine the raw *SMAD4* WGS data in the binary alignment map (bam) file, and identified a single 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 RNASeq 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 RNASeq 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 containing 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,[9] in addition to standard HHT screening and pre-symptomatic interventions.[10,11,12] 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 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 patients 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 patients 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[65,81,69] 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.[77,65] Our data suggest this will be important for regulation of *SMAD4*, a ubiquitous and essential protein with diverse functions,[8,9] where ~7kb of 3' UTR is transcribed at high levels in coding and non-coding transcripts (Figure 5, Figure 8, *Fig. S5*, *Fig. 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.[82,70] As recently reviewed,[77,65] 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 1hr stress responses when the cell has to rely predominantly on reuse of existing RNA transcripts, highlight further mechanisms to explore. 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 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 patients. 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 **s**tress **i**mpact, **f**unctional **a**lternate **p**olyadenylation **s**ite (SIFAPS) variants contribute to diseases will only be exposed 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.

Web Resources section

Genome Reference Consortium Human Build 38:
https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/

GENCODE Human Genome Release 31: https://www.gencodegenes.org/human/

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/

RStudio: https://www.rstudio.com/

RepeatMasker: http://www.repeatmasker.org

The Comprehensive R Archive Network: https://cran.r-project.org/

The National Genomics Research and Healthcare Knowledgebase v5 (2019) Genomics England. doi:10.6084/m9.figshare.4530893.v5.

Appendix

Supplementary Data include a single file of Supplementary Methods, 7 Supplementary Figures, and 6 Supplementary Tables; 2 Supplementary Tables provided as separate resource files; and 2 Movies.

Declaration of interests

The authors declare no competing interests

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. AA was supported by Prince Sultan Military Medical City, Saudi Arabia. MAA 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: SX, CLS

Methodology: SX, ZK, DP, AB, MEB, AA, MAA, NV, MJC, GERC, CLS

Investigation: SX, ZK, ADM, SW, CLS

Visualization: SX, CLS

Funding acquisition: SX, AA, ADM, SW, MAA, MJC, CLS

Project administration: GERC, CLS

Supervision: DP, MEB, MAA, CLS

Writing – original draft: CLS

Writing – review & editing: SX, ZK, DP, AB, MEB, AA, ADM, SW, MAA, NV, MJC, GERC, CLS

SX devised and generated the GROFFFY approach, devised all scripts to generate GROFFFY, and generated all GROFFFY numeric data, Figures 1, 2 and 3, Figures S1 and S2, and Tables S1, S2, S3, S4 S5, and S6. ZK advised on Linux and script generation. DM interrogated Donor 3 bam files. DL assisted in PBMC cultures. DP, AB, MBH, and MAA performed BOEC cultures and RNA sequencing. AA designed primers for validations. AM contributed to patient recruitment. SW contributed to clinical correlations. NV advised on *SMAD4* regulation. GERC performed all sequencing. MJC contributed to specific project set up at Genomics England. CLS 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 RNASeq and variant level data analyses; generated Figures 4, 5, 6, 7 and

8, Figures S3, S4, S5, S6 and S7, Tables S6, S7 and S8, and wrote the manuscript. All authors have reviewed and approved the final manuscript.

Data and code availability:

The publicly available file accession numbers used to generate the code are provided in full within the Data Supplement and have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA596860, referencing the WGS data source under accession number SAMN13640532. Primary 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.

References and Notes

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(12), 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(D1), D726—32

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(D1), D794—D801, accessed for data 25.07.2019

6 Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D. et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol.7 Suppl 1 S4,1—9.

7 The National Genomics Research and Healthcare Knowledgebase v5, Genomics England. doi:10.6084/m9.figshare.4530893.v5. 2019

8 Hayashi, A., Hong J. and Iacobuzio-Donahue C.A. (2021). The pancreatic cancer genome revisited. Nat Rev Gastroenterol Hepatol. 18(7), 469—481.

9 Larsen Haidle, MacFarland, J.S.P. and Howe, J.R. (2022) Juvenile Polyposis Syndrome. 2003 May 13 [updated 2022 Feb 3]. In: Adam M.P., Everman D.B., Mirzaa, G.M. editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993—2022.

~~10 The National Genomics Research and Healthcare Knowledgebase v5, Genomics England. doi:10.6084/m9.figshare.4530893.v5. 2019~~

~~11~~10 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(1):104370.

~~12~~11 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, et al. (2020). Second International Guidelines for the Diagnosis and Management of Hereditary Hemorrhagic Telangiectasia. Ann Intern Med. 173(12):989—1001.

~~13~~12 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(12):859—862.

~~14~~13 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 A. 188(3):959—964.

~~15~~14 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(13):3956—3969.

~~16~~15 Shovlin, C.L., Almaghlouth, F., Alsafi, A., Coote, N.C., Rennie, C.R., Wallace, G.M.F., Govani, F.S., and Genomics England Research Consortium. (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 In Press

~~17~~16 Sharma L., Almaghlouth F., Mckernan, H., Springett, J., Tighe, H.C., Genomics England Research Consortium and Shovlin C.L. (2023) Iron deficiency responses and integrated

compensations in patients according to hereditary haemorrhagic telangiectasia ACVRL1, ENG and SMAD4 genotypes. Haematologica. In Press

~~18~~17 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(17):1907—1918.

~~19 Genome Reference Consortium Human Build 38. Available at https://www.ncbi.nlm.nih.gov/assembly?term=GRCh38&cmd=DetailsSearch~~

~~20~~18 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(D1):D135-D139.

~~21~~19 Kozomara, A. Birgaoanu M., Griffiths-Jones S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. 47(D1):D155—62

~~22~~20 Amemiya, H.M., Kundaje, A. and Boyle A.P. (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 9(1):9354

~~23~~21 Smit, A.F.A., Hubley, R., Green, P. (2013-2015) RepeatMasker Open-4.0.~~Institute for Systems Biology RepeatMasker available at: http://www.repeatmasker.org, accessed for data 25.09.2019~~

~~24 Encode Experiment search. Available at~~

~~25~~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. Nature Biotechnol 28;1045—8

~~26~~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(18):2847—52.

~~27~~24 Koch, L. (2017) Cancer genetics: A 3D view of genome rearrangements. Nat Rev Genet 18(8):456.

~~28~~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(14):1919—20

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(6):996—1006.

~~32~~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(D1):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(16):2041—3

~~36~~29 Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H. Y., Källberg, M., Kumar, S. A., Liao, A., Little, K. M., Strömberg, M. P., and Tanner, S. W. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics (Oxford, England), 29(16), 2041–2043

30 McKinney W. (2010). Data Structures for Statistical Computing in Python: 51—6. Available from: http://conference.scipy.org/proceedings/scipy2010/mckinney.html

~~39~~31 Quinlan, A. R., and Hall, I. M (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics; 26(6):841–2.

~~40~~32 Ensembl Variant Effect Predictor (VEP), available at http://grch37.ensembl.org/info/docs/tools/vep/index.html

33 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(16):2078—9.

34 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(7):1—8

~~47~~35 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(7809):434—443.

~~48~~36 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(D1):D886—D894.

~~49~~37 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(5):405—24

~~50~~38 Bernabeu-Herrero, M.E., Patel, D., Bielowka, A., Chaves Guerrero, P., Marciniak S.J., Noseda, M., Aldred, M.A., Shovlin. C.L. 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. BioRxiv 471269 v2. 27th June 2023. doi: https://doi.org/10.1101/2021.12.05.471269

~~51~~39 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;W1: W537—W544

~~52~~40 Freese, N.H., Norris, D.C. and Loraine, A.E. (2016) Integrated genome browser: visual analytics platform for genomics. Bioinformatics. 32(14):2089—95

~~53~~41 GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45(6):580—5.

~~54~~42 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(7):994-1005.

~~55~~43 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(6472):eaav1741

~~56~~44 Chen, Y. Wang X. (2020) miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res; 48(D1):D127—D131.

~~57~~45 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.

~~58~~46 Krokhotin, A., Houlihan K., and Dokholyan N.V. (2015). iFoldRNA v2: folding RNA with constraints. Bioinformatics 31 2891—2893.

~~59~~47 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(W1): W431—W437.

~~60~~48 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, accessed at https://www.rcsb.org/3d-view.

~~61~~49 Shi, C. (2012) Developing a new mutational analysis system using hereditary haemorrhagic telangiectasia as a genetic model. Imperial College London MSc Thesis (supervisor CL Shovlin).

~~62~~50 Li, Y. (2013) Gene mutations and transcripts in hereditary haemorrhagic telangiectasia (HHT). Imperial College London MSc Thesis (supervisor CL Shovlin)..

~~63~~51 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

~~64~~52 Duong, H. and Patel, G. (2022). Hypothermia. In: StatPearls Treasure Island (FL): StatPearls.

~~65~~53 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(4):184—96.

~~66~~54 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(2):e0147990.

~~67~~55 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(3):209—217.

~~68~~56 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 U S A. 118(48):e2111862118.

~~69~~57 Kartikasari, A.E., Georgiou, N.A., Visseren, F.L., van Kats-Renaud, H., van Asbeck, B.S., and Marx, J.J. (2006) Endothelial activation and induction of monocyte adhesion by nontransferrin-bound iron present in human sera. FASEB J. 20(2):353—5.

~~70~~58 Bolger, A.M. Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114—20. v0.36

~~71~~59 Anders S., and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106

~~72~~60 Anders, S., Reyes, A., and Huber, W. (2012) Detecting differential usage of exons from RNA-seq data. Genome Res. 22(10):2008—17

~~73~~61 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(1):17960.

~~74~~62 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(2):230—244.e1

~~75~~63 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(6):1235—9

~~76~~64 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(7616):285—91

~~77~~65 Mitschka, S. and Mayr, C. (2022). Context-specific regulation and function of mRNA alternative polyadenylation. Nat Rev Mol Cell Biol. 7:1—18.

~~78~~66 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(D1):D29—D38.

~~79~~67 Shovlin, C.L. (2010). Hereditary haemorrhagic telangiectasia: pathophysiology, diagnosis and treatment. Blood Rev. 24(6):203—19.

~~80~~68 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(5):776—792

~~81~~69 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(20):3539—3557.

~~82~~70 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(1):2709.

The Genomics England Research Consortium Members comprised on 8th May 2022:

Ambrose, J. C. 1 ; Arumugam, P.1 ; Bevers, R.1 ; Bleda, M. 1 ; Boardman-Pretty, F. 1,2 ; Boustred, C. R. 1 ; Brittain, H.1 ; Brown, M.A.; Caulfield, M. J.1,2 ; Chan, G. C. 1 ; Giess A. 1; Griffin, J. N. ; Hamblin, A.1; Henderson, S.1,2; Hubbard, T. J. P. 1 ; Jackson, R. 1 ; Jones, L. J. 1,2; Kasperaviciute, D. 1,2 ; Kayikci, M. 1 ; Kousathanas, A. 1; Lahnstein, L. 1 ; Lakey, A.; Leigh, S. E. A. 1 ; Leong, I. U. S. 1 ; Lopez, F. J. 1 ; Maleady-Crowe, F. 1 ; McEntagart, M.1; Minneci F. 1 ; Mitchell, J. 1 ; Moutsianas, L. 1,2 ; Mueller, M. 1,2 ; Murugaesu, N. 1; Need, A. C. 1,2 ; O'Donovan P. 1; Odhams, C. A. 1 ; Patch, C. 1,2 ; Perez-Gil, D. 1 ; Pereira, M. B.1 ; Pullinger, J. 1 ; Rahim, T. 1 ; Rendon, A. 1 ; Rogers, T. 1 ; Savage, K. 1 ; Sawant, K. 1; Scott, R. H. 1 ; Siddiq, A. 1 ; Sieghart, A. 1 ; Smith, S. C. 1; Sosinsky, A. 1,2 ; Stuckey, A. 1 ; Tanguy M. 1 ; Taylor Tavares, A. L.1; Thomas, E. R. A. 1,2 ; Thompson, S. R. 1 ; Tucci, A. 1,2 ; Welland, M. J. 1 ; Williams, E. 1 ; Witkowska, K. 1,2 ; Wood, S. M. 1,2; Zarowiecki, M. 1 .

1. Genomics England, London, UK

2. William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK.

FIGURE LEGENDS

Fig. 1 GROFFFY Study Protocol: Flow chart illustrating sequence of stages described in the text and online Data Supplement

Figure 2: GROFFFY and the human genome. A) Heatmap displaying GROFFFY categorization of Genome Reference Consortium Human [GRCh] Build 38.[19] The heatmap maps 61,799 data points at 500kb resolution, and heights represent the percentage of each 500kb region included in GROFFFY. **B)** Higher resolution image from a randomly chosen region of the genome (on chromosome 3: chr3:25,597,986-2-25,783,443). The top 4 tracks illustrate sources, from top: GENCODE[6] gene annotations, CpG islands,[31,32]26,27 long non-coding RNAs[18,30] and miRNAs.[21]19 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.

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)[46,34] or gnomAD;[47,35] synonymous (Synon.) filter, and white-listed filter (see Methods and *Tables S1-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.[48–36] Irrespective of other filters applied, GROFFFY, and its individual components, significantly reduced the number of variants compared to the other tested filter sets (*Figure S1, Figure S2*).

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 scores (note logarithmic scale pre filtration versus linear scale post filtration). **ii)** Non-pathogenic variants by CADD score categories: Molecular subtype are indicated in the key. **iii)** Pathogenic variants by CADD score categories, and molecular subtype as in key. Note identical plots in **iii** pre and post filtration because all pathogenic variants were still present post filtration. **B)** Discovery dataset: Location of GROFFFY-captured variants in the major HHT genes[187] –**i)** *ACVRL1*, **ii)** *ENG*, **iii)** *SMAD4*. The cartoons include screenshots of GRCh38 from the University of California Santa Cruz (UCSC) Genome Browser, [31,32] and major transcripts. Red inverted triangles indicate location of variants after application of all filters (dark red for coding/splice regions, bright red for non-coding regions).

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

A) Endothelial *SMAD4* total RNA expression: RNASeq data from 8 different cultures of blood outgrowth endothelial cells (BOECs) with normal *SMAD4* sequence.[50 38] The consistent peaks sharply define exon boundaries in GRCh38.[19] 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. The cartoon below links the RNASeq expression by grey dotted lines.to the main (upper) and alternate *SMAD4* RefSeq[78] RefSeq[66] splice isoforms that share the final UTR-containing exon, with exons to scale. Blue: coding, grey: non-coding regions. 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[54]) per interval of DNA flanking the TGA stop codon, as listed by USCSC CAVIAR tracks[31,32 26,27] for data from the Genotype Tissue Expression project (GTEx).[53 41] 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 10kb intervals. **ii)** Magnified view of penultimate and final exons at 1kb intervals.

C) Number of general population 3'UTR alternative polyadenylation QTLs (3'aQTLs[54] 3'aQTLs[42]) per kilobase of DNA flanking the TGA stop codon, as determined in GTEx.[53 41] **i)** Overview of *SMAD4* locus and flanking regions at 10kb intervals. **ii)** Magnified view of penultimate and final exons at 1kb intervals.

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 NM_005359. 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 variants. Deleted residues are indicated by black bars, missense substitutions as black triangles. For FASTA format sequences, see *Table S8.* **B)** Variant 1 (chr18:51083986 CTTAACGCGCGTGCGCACGCGCGCGCGCACA>CAACGCGCGTGCACGCG and Variant 2 (chr18:51084116 ACACT>A) in detail. The AAUAAA hexamer is shown by DNA sequence (AATAAA) and highlighted by a pink box, deleted residues by red underline (wildtype) or vertical red line (variants), and missense substitutions as red stars. *Table S8* provides further sequence details.

Figure 7: **Replicate iFoldRNA structures**: iFoldRNA[57,58 45,46]simulations as visualized in Mol* Viewer.[59,6047,48] **A)** Variant 1: **i)** Three representative simulations of the wildtype 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 Movie 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 Movie S2. **B)** Variant 2**: i/ii)** Two representative

simulations of wildtype 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.

Figure 8: *SMAD4* RNA expression in ribosomal (r)RNA-depleted libraries from 3 controls compared to affected donors from the 3 separate HHT families: (i) Variant 1, **ii)** Variant 2, **iii)** an unsolved clinical *SMAD4* positive control. 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 1hr stresses (*Figure S3)*: *SMAD4* splice site changes met DEXSeq2[59] significance after cycloheximide (CHX). **A)** Total *SMAD4* RNA from control (grey, N=23) and ~~*patient*~~ *SMAD4* variant-affected donors (red: **i)** Variant 1 N=7; **ii)** Variant 2 N=8; **iii)** Donor 3, N=7) following DESeq2 normalisation[59,~~80~~ 68] using GINI housekeeper genes.[~~73,74~~61,62] Note contrast between **i/ii** (Variant 1 and 2 donors), but similarities between **i/iii** (Variant 1 and Donor 3). **B-E)** DEXSEQ[60] splicing patterns across 61 exon regions in 22 *SMAD4* exons. **B)** Exon region (ER) use in untreated PBMCs by donor, plotting data from the individual patients (red) and the same 3 controls (black). Exons are colour-coded to highlight 3' aQTL loci.[~~54~~ 42] **C)** Use of ER60, the variant-containing 3'UTR region in untreated and CHX-treated PBMCs: note again **i/iii** similarities. **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*).

982