SUPPLEMENTARY MATERIAL

Detailed Methods

Data access

Data can be accessed at the European Genome-phenome Archive (https://www.ebi.ac.uk/ega) using accession number EGAS00001003302.

Patient cohorts

867 adult and child probands of Northern European ancestry presenting with TOF, together with their parents (when available), were recruited from nine centres: seven in the UK, Leuven (Belgium), and Sydney (Australia), as previously described (1). Ethical approval was obtained from the local institutional review boards at each participating centre prior to blood or saliva sample collection and informed consent was obtained from all subjects or their parents/legal guardians. Patients who exhibited clinical features of recognised syndromes, extra-cardiac abnormalities or learning difficulties were excluded from the study. Almost all cases came from families with a single affected member; where more than one family member was affected, only the proband's exome data contributes to the results reported here. All samples were screened for the 22q11.2 deletion associated with TOF prior to exome sequencing (2) using multiplex ligation-dependent probe amplification (MRC-Holland), and were excluded if the deletion was present.

Whole exome sequencing (WES)

WES was carried out at the McGill University/Genome Quebec Innovation Centre (MUGQIC) using the Illumina HiSeq2000 (3). Exome capture was performed using Agilent SureSelectXT Human All Exon 50 Mb version 4 kit (Agilent). Exome data were analyzed using the GenPipes DNAseq pipeline (https://bitbucket.org/mugqic/genpipes). Paired-end sequencing reads (100 bp) were trimmed using Trimmomatic to obtain a high quality set of reads for sequence alignment (SAM/BAM) file generation (4). The trimmed reads were aligned to a reference genome (build 37/hg19) using BWA-mem (v0.6.2) (5). This resulted in 90% of the target region being covered a minimum of 30 times. Samples with evidence of unsatisfactory sequencing data quality or erroneous mapping were excluded and the remaining 829 cases were taken forward for further annotation.

Annotation and classification of genetic variants

The genome analysis tool kit (GATK v3.7) was used for base quality score recalibration and indel realignment prior to variant calling using the HaplotypeCaller (6-8). Only variants with a quality score (QS) >200 contributed to the analyses and additional filters removed multiple indels within a 50bp window in the same sample as well as limiting the number of variants with the same start site to a maximum of three. Functional annotation of the variants was performed using SnpEff (9). Additional annotation was provided from OMIM and Genomic Evolutionary Rate Profiling (GERP) (35 species alignment) as well as population frequencies from the 1000 genomes project (10) and the Genome Aggregation Database (gnomAD) (11) using the Gemini framework (12). Variants that caused premature truncation of the protein coding sequence (nonsense and frameshift mutations) were classified as pathogenic. The likely pathogenicity of non-synonymous variants that were not truncating was assessed using CADD (13).

Only variants absent in the gnomAD database were included in the analyses. Nonsense variants and missense variants with a scaled CADD score ≥ 20 were included, which are predicted to represent the top 1% of deleterious variants in the human genome (13). Variants meeting our criteria were henceforth designated 'deleterious'. The relatedness statistic between samples was calculated using the methodology of Yang *et al* (2010) and implemented using VCFtools (14,15).

Reference exomes

In addition to gnomAD, we used reference exome data, analysed using the same variant calling pipeline as our case data, to avoid, as far as possible, systematic biases due to technical factors or analytic methodology. Reference data included 1000 exomes from UK population samples (the ICR1000 Exome series) participating in the 1958 UK birth cohort collection (16) obtained from MUGQIC, as well as 35 in-house non-TOF exomes. In addition, 217 exomes that had been sequenced using the same capture kit and methodology as the cases, at the Manchester Centre for Genomic Medicine (MCGM), were analysed. In the absence of a cohort sequenced using the same capture kit and methodology, contemporaneously, in the same centre, we did not carry out a case-control comparison. Rather, we used the reference exomes to exclude any variants that might have been found in our cases through artefact specific to our data generation and analysis approach. Accordingly, any variants that were observed in the 1252 reference exome samples were eliminated.

Clustering analysis

To assess the significance of variant clustering we used the W_d statistic and test described by Lange (1997). Clustering analysis, which is a case-only analysis, follows the assumption that each position in the coding sequence of the exome is equally likely to carry a variant. It assesses whether there is a statistically significant excess of variants with respect to the number expected in the coding sequence of a particular gene, given the number of variants observed in all analysed coding sequences (17). We assessed the clustering of potentially deleterious variants in all protein coding genes covered by the exome capture kit, excluding genes involved in segmental duplications. Gene lengths were obtained from Ensembl v75 using the coding length of the longest transcript. P-values are corrected for multiple comparisons.

Permutation Test

For the top 100 genes identified in the clustering analysis, we additionally carried out a permutation test. The permutation test estimates the probability of finding at least as many variants as observed, given the relative length of the coding region in the individual gene and the total number of variants found exome-wide by randomly distributing the observed number of variants 100,000 times (or, in the case of the top three genes, 1 million times). Results, which confirmed those from the clustering analyses, are presented in supplementary table 6.

Power calculation

To estimate the feasibility of our approach we assumed that 20,000 genes were going to be assessed and that *a priori* the probability of a unique, deleterious variant to be found in each of these genes is equal (i.e. $5x10^{-5}$ for each gene). We further assumed that for a gene involved in disease predisposition, this probability is increased. We explore the ability of our test to detect a clustering

depending on the increase in probability (described as fold increase) depending on sample size and on the average number of unique, deleterious variants per genome. The probability of detecting a cluster was done simulating and analysing the distribution of rare mutations 10,000 times for each combination of factors (probability increase, average number of variants of interest per genome) and a sample size of 800. We explore the range between 4 and 32 unique, deleterious variants per genome and a 6 to 32-fold increase in the probability of a gene to contain a unique variant (Supplementary figure 1).

Confirmation of exome sequencing and identification of *de novo* variants

DNA fragments to be sequenced were obtained by PCR amplification and their specificity was verified by agarose gel electrophoresis. Sanger sequencing was performed to validate variants. All variants sequenced with a QS >200 were confirmed (data not shown). DNA samples from the parents of thirteen probands with *NOTCH1* variants were available and these samples were sequenced as described above to determine if the variants were inherited or *de novo*.

Tissue culture

HeLa cells (Sigma, 93021013) were cultured in DMEM (Thermo Fisher Scientific, 41965-039) with 10% fetal bovine serum (Sigma, F9665) and penicillin/streptomycin (Sigma, P0781) 1:1000. They were passaged by washing in PBS and detached using TrpLE Express (Thermo Fisher Scientific, 12605-010) before quenching in prewarmed media and seeding.

NOTCH1 constructs and luciferase assays

The pcDNA3_NOTCH1 plasmid (a gift from Iannis Aifantis, New York University, US) is a full-length NOTCH1 expression construct containing codons 1 to 2555 of human *NOTCH1*, followed by a FLAG-tag sequence (18). *NOTCH1* variants G200R, C607Y and N1875S were introduced by site-directed mutagenesis using the QuikChange Lightning kit (Agilent). Recombinant Human Jagged 1 Fc Chimera Protein (R&D systems) reconstituted in PBS was immobilised on tissue culture plates at a concentration of 2.5µg/ml. PBS only was used for a no-ligand control. HeLa cells were seeded onto the immobilised ligand and transiently transfected with pcDNA3_NOTCH1 or mutant NOTCH1 constructs together with the RBPJ–luciferase reporter construct (Qiagen) using Lipofectamine 3000 transfection reagent (Thermo Fisher). Luciferase activity was measured using the Dual-Glo Luciferase Reporter Assay System (Promega) on a CLARIOstar microplate reader (BMG LABTECH). Firefly luciferase activity was normalised to co-transfected internal control, Renilla luciferase (Qiagen). Results are presented as mean ± SEM from four biological replicates. Statistical significance was assessed by two-tailed paired *t*-tests where $P<0.05$ was considered statistically significant.

Immunoblotting

Transfected HeLa cells were lysed and prepared for immunoblotting in Laemmli buffer. Immunoblotting was performed using Biorad Mini-PROTEAN gels and transfer kits according to manufacturers' instructions. Nitrocellulose membranes were incubated with antibodies in 5% BSA blocking buffer. The antibody used for the detection of NOTCH1 was mouse anti-flag M2 (1:1000; Cat No - F3165, Sigma) and rabbit β-actin (1:2000; D6A8, Cat No - 8457 Cell Signaling) was used as a loading control. Goat secondary antibodies conjugated to horseradish peroxidase (HRP), anti-

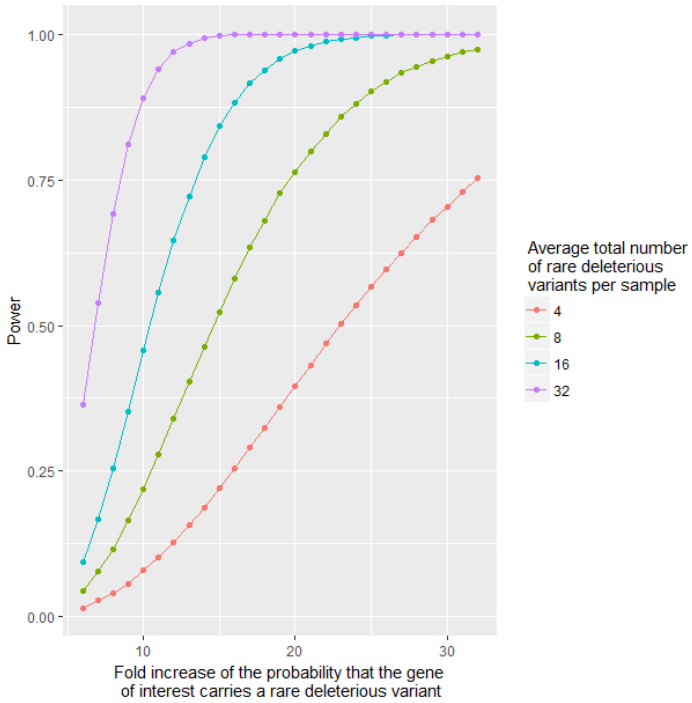
Mouse-HRP (1:5000; Cat No - P0447, Dako) and anti-Rabbit-HRP (1:20,000; Cat No - P0448, Dako), were used to detect primary antibodies. HRP-conjugated antibodies were detected by ECL (Thermo Fisher Scientific). ImageJ was used for the quantification of cleaved versus uncleaved protein and statistical significance was assessed by two-tailed paired *t*-tests where $P<0.05$ was considered statistically significant.

Supplementary References

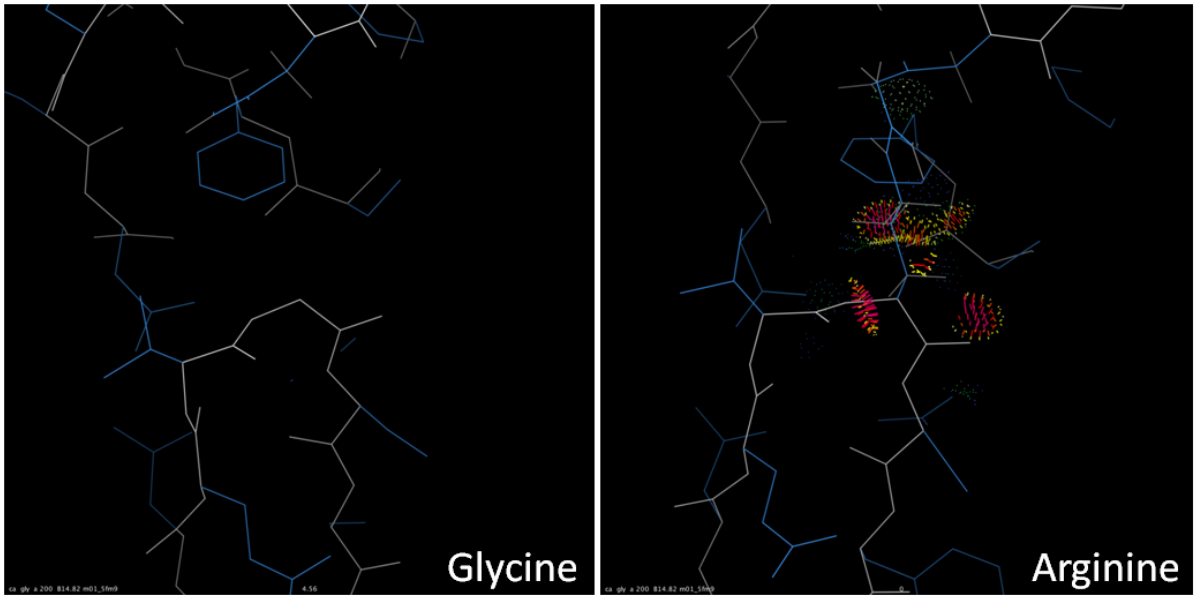
1. Cordell HJ, Mamasoula C, Postma AV, et al. Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Human Molecular Genetics*. 2013;**22**:1473–1481.
2. Mercer-Rosa L, Rychik J, Zhao H, Zhang X, Yang W, Shults J, Goldmuntz E. 22q11.2 Deletion Status and Disease Burden in Children and Adolescents With Tetralogy of Fallot. Clinical Perspective. *Circulation: Cardiovascular Genetics*. 2015;**8**:74.
3. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;**456**:53–59.
4. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;**30**:2114–2120.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;**25**:1754–1760.
6. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2014;**43**:11.10.1–33.
7. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 2011;**43**:491–498.
8. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010;**20**:1297–1303.
9. Cingolani P, Platts A, Le Lily Wang, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;**6**:80–92.
10. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*. 2015;**526**:68–74.
11. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;**536**:285–291.
12. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput Biol*. 2013;**9**:e1003153.
13. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*. 2014;**46**:310–315.
14. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*. 2010;**42**:565.
15. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Group 1GPA. The variant call format and VCFtools. *Bioinformatics*. 2011;**27**:2156.
16. Ruark E, Münz M, Renwick A, Clarke M, Ramsay E, Hanks S, Mahamdallie S, Elliott A, Seal S, Strydom A, Gerton L, Rahman N. The ICR1000 UK exome series: a resource of gene variation in an outbred population. *F1000Research*. 2015;**4**:883.
17. Lange K. Mathematical and Statistical Methods for Genetic Analysis. *Statistics for Biology and Health*. 1997.

18. Sulis ML, Williams O, Palomero T, Tosello V, Pallikuppam S, Real PJ, Barnes K, Zuurbier L, Meijerink JP, Ferrando AA. NOTCH1 extracellular juxtamembrane expansion mutations in T-ALL. *Blood*. 2008;**112**:733.

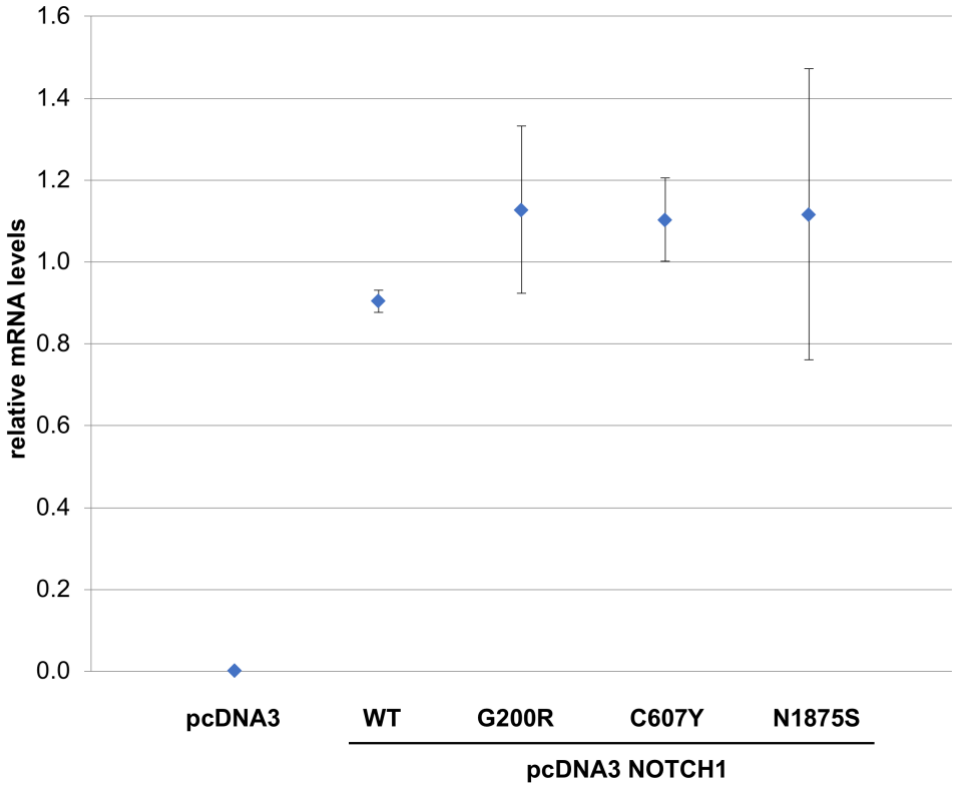
Supplementary Figures



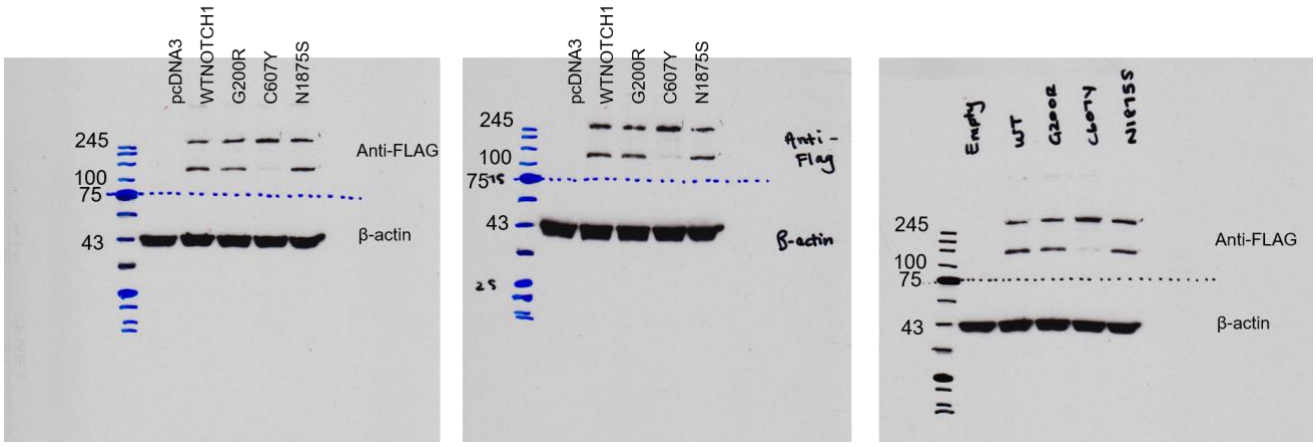
Supplementary figure I: Power for a sample with 800 cases. In our series there were 16683 variants across 829 samples (i.e. on average 20 variants per sample) and we aimed to detect increases of 10 fold or larger.



Supplementary figure II: The structural impact of NOTCH1 variant G200R. The arginine side-chain is modelled in place of the original glycine. Potential van der Waal atomic clashes between the arginine side-chain and the surrounding protein are highlighted in pink and give an indication of how well the substitution can be accommodated with respect to the existing local structure (17).



Supplementary Figure III: mRNA expression of NOTCH1 following transfection of vectors as indicated. Error bars: mean ±SEM from three biological replicates, each with three technical replicates. Statistical significance was assessed using two-tailed paired *t*-tests.



Supplementary Figure IV: Full scans of immunoblot biological replicates, including molecular weight markers.

Supplementary Tables

Supplementary Table I: All genes ordered by levels of significance following the clustering analysis of unique, deleterious variants.

Supplementary Table II: The unique, deleterious *NOTCH1* variants identified in the TOF patient cohort.

Supplementary Table III: 166 NOTCH pathway-associated genes, ordered by levels of significance following the clustering analysis of unique, deleterious variants.

Supplementary Table IV: The unique, deleterious *FLT4* variants identified in the TOF patient cohort.

Supplementary Table V: The inheritance status of *FLT4* variants following Sanger sequencing of available parent samples.

Supplementary Table VI: Results of the permutation test for the top 100 candidate genes identified by the clustering analyses.

Supplementary Table VII: A summary of *in vivo* and *in vitro* functional data currently available for the candidate genes featured in Table 1.