

DECIPHER: Supporting the interpretation and sharing of rare disease phenotype-linked variant data to advance diagnosis and research

Julia Foreman¹, Simon Brent¹, Daniel Perrett¹, Andrew Bevan¹, Sarah Hunt², Fiona Cunningham³, Matthew Hurles¹, and Helen Firth⁴

¹Wellcome Sanger Institute

²EMBL-EBI

³European Bioinformatics Institute

⁴Cambridge University Hospitals NHS Foundation Trust

February 22, 2024

Abstract

DECIPHER (<https://www.deciphergenomics.org>) is a free web platform for sharing anonymised phenotype-linked variant data from rare disease patients. Its dynamic interpretation interfaces contextualise genomic and phenotypic data to enable more informed variant interpretation, incorporating international standards for variant classification. DECIPHER supports almost all types of germline and mosaic variation in the nuclear and mitochondrial genome: sequence variants, short tandem repeats, copy-number variants and large structural variants. Patient phenotypes are deposited using Human Phenotype Ontology (HPO) terms, supplemented by quantitative data, which is aggregated to derive gene-specific phenotypic summaries. It hosts data from >250 projects from ~40 countries, openly sharing ~40,000 patient records containing >51,000 variants and >172,000 phenotype terms. The rich phenotype-linked variant data in DECIPHER drives rare disease research and diagnosis by enabling patient matching within DECIPHER and with other resources, and has been cited in >2,600 publications. In this paper, we describe the types of data deposited to DECIPHER, the variant interpretation tools, and patient matching interfaces which make DECIPHER an invaluable rare disease resource.

Title

DECIPHER: Supporting the interpretation and sharing of rare disease phenotype-linked variant data to advance diagnosis and research

Authors

Julia Foreman, Simon Brent, Daniel Perrett, Andrew P. Bevan, Sarah E. Hunt, Fiona Cunningham, Matthew E. Hurles and Helen V. Firth

Author's institutional affiliations

Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom.

European Molecular Biology Laboratory, European Bioinformatics

Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ United Kingdom.

Grant Numbers

This research was funded in whole, or in part, by the Wellcome Trust [Grant numbers WT206194, WT200990/Z/16/Z, WT108749/Z/15/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Abstract and Keywords

DECIPHER (<https://www.deciphergenomics.org>) is a free web platform for sharing anonymised phenotype-linked variant data from rare disease patients. Its dynamic interpretation interfaces contextualise genomic and phenotypic data to enable more informed variant interpretation, incorporating international standards for variant classification.

DECIPHER supports almost all types of germline and mosaic variation in the nuclear and mitochondrial genome: sequence variants, short tandem repeats, copy-number variants and large structural variants.

Patient phenotypes are deposited using Human Phenotype Ontology (HPO) terms, supplemented by quantitative data, which is aggregated to derive gene-specific phenotypic summaries. It hosts data from >250 projects from ~40 countries, openly sharing ~40,000 patient records containing >51,000 variants and >172,000 phenotype terms.

The rich phenotype-linked variant data in DECIPHER drives rare disease research and diagnosis by enabling patient matching within DECIPHER and with other resources, and has been cited in >2,600 publications. In this paper, we describe the types of data deposited to DECIPHER, the variant interpretation tools, and patient matching interfaces which make DECIPHER an invaluable rare disease resource.

Rare Diseases; Genetic Disorders; Genotype Phenotype Correlation; Variant Interpretation; MatchMaker Exchange

Main text

Introduction

The population prevalence of rare disease has recently been estimated to be 3.5–5.9%, which equates to 263–446 million people affected globally. A large proportion of these rare diseases, approximately 72%, are known to have a genetic basis (Wakap *et al.* , 2020 PMID: 31527858). Advances in genomic technologies to determine causal variants, such as whole-exome sequencing, are identifying the genetic basis of disease for only 25–40% of patients (Stranneheim *et al.* , 2021, Quaio *et al.* , 2020, Sawyer *et al.* , 2016). As a result, many patients undergoing diagnostic genetic testing do not receive a genetic diagnosis, and often experience long delays which have a substantial emotional impact on the family (Miller, 2021) and significant healthcare costs (Monroe *et al.* , 2016). A genetic diagnosis has multiple benefits for the patient and their family, including better understanding of the prognosis, personalised treatment, tailored management and surveillance, improved access to health and social care, and increased reproductive choice (Wright *et al.* , 2018).

The number of rare Mendelian diseases with known molecular aetiology is estimated to be 5,000–6,000 (Hartley *et al.* , 2018), however, for the majority of disease-associated genes it is not known which variants are disease-causing, and which are benign. Different pathogenic variants in the same gene can cause different diseases, for example variants in *FGFR3* can cause multiple diseases including Muenke Syndrome, Hypochondroplasia, and Achondroplasia. Different diseases caused by variants in the same gene must be considered distinct due to their disparate clinical presentation and differing treatment options. The sharing of patient level variants and phenotypes is therefore essential to accelerate our understanding of the molecular basis of genetic disease.

DECIPHER (Firth *et al.* , 2009; Swaminathan *et al.* , 2012; Bragin *et al.* , 2014; Chatzimichali *et al.* , 2015) is a global web-based platform which shares phenotype-linked variant data from rare disease patients (Fig. 1A). It is freely available via a web interface at <https://www.deciphergenomics.org>. Approximately 40,000

of the patient records held by DECIPHER have explicit patient consent for open sharing on the website (Fig. 1B). These openly shared records contain more than 51,000 variants and more than 172,000 phenotype terms. The integration of this phenotype and variant data enables the discovery of new gene-disease and variant-disease relationships, driving diagnosis and our understanding of human biology. Since DECIPHER was established in 2004, the platform has been used and cited in more than 2,600 published manuscripts.

Patient records in DECIPHER are deposited by academic clinical centres, which are affiliated both to a hospital which oversees the treatment of patients with genetic disorders, and to a local university department of human/clinical genetics. Eligible centers (<https://www.deciphergenomics.org/join/overview>) can apply to join DECIPHER using an online application form. Data from a centre is stored within a DECIPHER project, and a senior clinician at that centre (clinical coordinator), sometimes in conjunction with a senior clinical scientist (lab coordinator), has the responsibility for approving/rejecting applications from individuals working at that centre who wish to access the data in the project.

The platform supports the deposition of almost all types of genetic variation, including sequence variants, short tandem repeats, copy-number variants (CNVs) and large structural variants. Variant interpretation interfaces are provided, including genome and protein browsers, which contextualise genetic and phenotype information to enable accurate interpretation. These interfaces integrate external datasets such as the Genome Aggregation Database (gnomAD, Karczewski *et al.*., 2020), which can be used to exclude variants seen at appreciable frequency in the general population, in addition to disease relevant datasets such as ClinVar (Landrum *et al.*., 2018 PMID: 29165669) and DECIPHER records themselves. DECIPHER also encourages the use of global standards to promote good practice, including the American College of Medical Genetics (ACMG) guidelines for sequence variant interpretation (Richards *et al.*., 2015) and ACMG/ClinGen technical standards for interpreting CNVs (Riggs *et al.*., 2020).

In the following sections we present examples of the genotype/phenotype data deposited and shared with the rare disease community. In addition we present the tools provided by DECIPHER to assess the pathogenicity of variants according to international standards, and the utility of DECIPHER to map the clinically relevant parts of the genome.

DECIPHER patient records

DECIPHER associates variants and phenotypes through individual patient records, each of which are connected to a particular depositing centre. DECIPHER itself cannot re-identify individuals, and technical and organisational measures are in place to safeguard data. These measures are reviewed and updated in line with evolving best practice.

On deposition, each patient record is given a DECIPHER Patient ID as a reference, which is shown on the website and forms part of the URL for the patient record (e.g. <https://www.deciphergenomics.org/patient/283351> - note that URLs of the form <https://decipher.sanger.ac.uk/patient/283351> continue to be supported). Each patient record also has an internal ID (e.g. a lab number), which is only displayed to users of the depositing centre. The internal ID allows the depositing centre (only) to link the record to an individual patient.

Through the DECIPHER platform it is possible to send a patient's clinician a message to request further information about the patient, for example in the case where there is a potential patient match, or if a researcher is carrying out a functional study on the gene in which that patient's variant is situated. Below we will describe in more detail the clinical and research utility of this notification system.

Deposition and breadth of sharing

DECIPHER has been carefully designed to ensure that the depth and breadth of sharing is proportionate to the scientific/clinical needs and level of consent. For example, a user who does not belong to a DECIPHER project can only access the openly shared patient data, while data which is visible to registered users who are logged in reflects their project and consortium memberships.

Patient genotype and phenotype data can be deposited to DECIPHER in three ways:

1. Via the web interface for an individual patient's data.
2. By uploading Excel or csv files via the web interface (bulk upload) for data from multiple patients.
3. Using the deposition API to allow programmatic uploading of data and synchronisation of data across systems (e.g. synchronisation between a centre's electronic health records and the patient records in that centre's DECIPHER project).

DECIPHER users at the depositing centre determine the sharing level of each patient record and variant. Patient records, and individual variants within these records, can be kept private to the depositing centre. This allows DECIPHER's tools to be used for assessing variant pathogenicity to inform the conversation with the patient before seeking consent for wider sharing. With explicit patient consent, patient records are shared openly, with the data available to anyone who visits the website. DECIPHER also supports consortium sharing. This allows sharing of patient records between a defined group of centres, where there is an expectation of collaboration for patient care, again before explicit patient consent for open sharing has been obtained. DECIPHER currently hosts six consortia, which share more than 63,000 patient records. Consortia include the United Kingdom National Health Service consortium, the Deciphering Developmental Disorders (DDD) consortium which shares research data from the DDD study (Wright *et al.* , 2014), and a data-sharing consortium covering New South Wales and Western Australia.

DECIPHER is a live interface and data deposited is available to view, interpret, and share in real time. Patient records can be added and edited iteratively as more information becomes available, e.g. additional phenotype terms, the inheritance status of a variant, or new functional data. Information can be added to a record by a clinician and clinical scientist working asynchronously and in different locations.

Genetic Data

As our knowledge of rare disease genetics develops and the interaction between genetic loci are more fully understood, there is a pressing need for the visualization of all types of genetic variation within a single interface. DECIPHER fulfills this need, supporting many types of genetic variation including sequence variants, CNVs, aneuploidy, uniparental disomy (UPD), inversions, insertions and short tandem repeats (STRs) (Fig. 2).

Variant deposition: Variants are deposited using genomic coordinates. Sequence variants can also be deposited using a relevant subset of HGVS nomenclature (den Dunnen *et al.* , 2006), and will be normalised (left aligned, parsimonious) during the deposition process (Tan *et al.* , 2015). For known STRs, the disease-relevant STR can be selected from a dropdown in the web interface. Additional information about the variant such as inheritance, genotype, pathogenicity, and contribution to phenotype can also be recorded.

Mosaicism: For *de novo* mosaic variants, it is possible to record the mosaicism observed in each tissue, as a percentage. This information is clinically important as it can help explain the variability of clinical symptoms, for example the difference between nevus sebaceous or Schimmelpenning syndrome (where extracutaneous abnormalities are present), caused by *HRAS* and *KRAS* variants (Groesser *et al.* , 2012).

Mitochondrial variants: DECIPHER supports the deposition and interpretation of variants in the nuclear and mitochondrial genomes. Mitochondrial diseases are the most common form of inherited neuro-metabolic disorders, and are caused by mutations in the nuclear or mitochondrial genomes. In addition, nuclear genetic factors have been shown to influence clinical outcomes for mitochondrial DNA mutations (Boggan *et al.* , 2019). Thus the display of both genomes in a single interface is clinically important. In DECIPHER it is possible to record homoplasmy or the percentage of heteroplasmy per tissue, which is clinically essential as it has been shown to contribute to disease progression (Grady *et al.* , 2018).

Variant haplotypes: Variants may work *in cis* to create or modify a disease allele or *in trans* to cause a biallelic disorder. For this reason DECIPHER users can assign variants to a haplotype, e.g. for compound heterozygous variants, the variants will be shown as *in trans* . As our understanding of rare disease genetics improves, the representation of its complexity is becoming even more essential. It is known that genetic

modifiers alleviate or exacerbate the severity of the disease (Rahit and Tarailo-Graovac 2020) and there are recent examples where rare pathogenic haplotypes have been shown to cause disease, such as an albinism-causing TYR haplotype (Campbell *et al.* , 2019).

Pathogenicity predictors: For all sequence variants deposited to DECIPHER, predictions from the Ensembl Variant Effect Predictor (VEP; McLaren *et al.* , 2016) are displayed across all Ensembl/GENCODE transcripts. Predictions include the consequence (e.g. missense, frameshift), the protein change, and several pathogenicity scores: SIFT (Sim *et al.* , 2012), PolyPhen-2 (Adzhubei *et al.* , 2013), CADD (Kircher *et al.* , 2014), REVEL (Ioannidis *et al.* , 2016), and SpliceAI (Jaganathan *et al.* , 2019). DECIPHER seeks advice from experts in the field and refers to benchmarking studies for pathogenicity predictors (e.g. Gunning *et al.* , 2021) prior to the inclusion of additional scores, assisting in the application of good practice.

Reference genome: All genomic information is displayed in the GRCh38 assembly version of the human genome, allowing the most up-to-date genome and transcript information to be used to enable accurate variant interpretation. The display of genomic data in GRCh38 permits DECIPHER to promote the use of Matched Annotation from NCBI and EMBL-EBI (MANE) transcripts, where the RefSeq and Ensembl/GENCODE transcripts from a protein-coding gene pair are identical (5' UTR, coding region, and 3' UTR). DECIPHER currently promotes and highlights MANE Select transcripts, one high-quality representative transcript per protein-coding gene that is well-supported by experimental data and represents the biology of the gene (https://tark.ensembl.org/web/mane_project). Describing variants relative to a single, recommended transcript, along with sequence variant normalisation, assists in the standardisation of variant reporting.

Reference conversion tools: Deposition with GRCh37/hg19 coordinates is still supported: prior to normalisation, DECIPHER remaps GRCh37 coordinates onto the GRCh38 assembly, using an algorithm based on the UCSC LiftOver tool (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>, Kuhn *et al.* , 2013). A range of tools are also provided to allow users to visualise the differences between assemblies. These include GRCh37 and GRCh38 comparative genome browsers, gene lists for variants lifted over by DECIPHER which display genes that no longer overlap the variant, and a liftover mapping genome browser track (Fig. 3).

Phenotypic Data

DECIPHER supports detailed phenotype data capture (Fig. 4A) which enables the in-depth comparison of patient phenotypes, as well as the delineation of new syndromes. Much of the phenotype is represented using Human Phenotype Ontology (HPO) terms - a standardised, controlled vocabulary which supports deep phenotyping (Köhler *et al.* , 2019). This allows phenotypic information to be described unambiguously, and for phenotypic similarity between patients to be established computationally by comparing related terms in the ontology. This is essential for finding potential patient matches. The DECIPHER phenotype deposition interface provides a search tool, allowing HPO terms to be added to a patient record quickly and easily. DECIPHER also supports the recording of the absence of clinically relevant phenotypes, and of manifestations of HPO terms (clinical modifiers), such as severity, age of onset, and pace of progression. This information can be helpful to users trying to determine the accuracy of a patient match, especially when the number of patient phenotypes is small.

In collaboration with ophthalmologists, DECIPHER has developed forms for groups of HPO phenotypes for the eye community, to assist phenotyping in the clinic. These forms contain a predetermined list of HPO terms which can be marked absent or present, and include common retinal and non-retinal disease, and symptoms and signs (extraocular features, ocular features, and electrodiagnostic testing and imaging).

Family history: In the case of inherited disorders, it is important to capture family phenotype history. In DECIPHER, users can record whether or not relevant family members are affected with similar or related phenotypes. Presence of absence of HPO terms can also be indicated for each family member if known.

Quantitative data: In addition to HPO terms, DECIPHER supports quantitative phenotype data capture (Fig. 4B). Developmental milestones (age of social smile, sat independently, walked independently, first

words) and anthropometric measurements (growth, visual function, fundus imaging) can be deposited. Aggregated observations from open-access patient records are shared openly (See Quantitative phenotypic data to confirm fit with diagnosis section and Fig. 6A). DECIPHER also provides an interface to record birth and pregnancy information, such as age of the mother/father at birth of the patient, consanguinity, maternal illness, and gestation (which is also used to adjust growth charts); this information is not currently shared openly, but is shared within a consortium.

Genotypic summaries to assist variant interpretation

DECIPHER provides a suite of tools to assist in assessing the pathogenicity of variants, including genome and protein browsers.

Protein browser: A protein browser is available for protein-coding genes, showing a genotypic summary which helps users to determine if a variant is located in a mutational hot spot or established functional domain (Fig. 5A). The protein browser is fully interactive and is customisable via a settings menu. In the centre of the protein browser, Pfam domains (Mistry *et al.*, 2021) are displayed allowing users to identify distinct functional/structural elements of the protein. Clinically relevant variants from DECIPHER and ClinVar are plotted above and below the Pfam domains, with annotated pathogenicity and predicted molecular consequence (e.g. missense, likely loss-of-function (LOF)) indicated through colouring. In addition to the location of the variants being shown, for likely LOF variants, the location of the protein truncating codon is indicated, since this information is essential in determining if a transcript will escape nonsense mediated decay (NMD). A predicted (NMD) track is also displayed. The location of variation in the general population is shown through display of gnomAD missense and LOF tracks. Regional missense constraint data is also available (regional missense constraint improves variant deleteriousness prediction, Samocha *et al.*, <https://www.biorxiv.org/content/10.1101/148353v1>), in addition to exon structure. Protein secondary structures (e.g. locations of helices and turns) and the locations of 3D structures (where available from the Protein Data Bank in Europe (PDBe)) are displayed at the bottom of the protein browser. Clicking on these 3D structures will display an interactive 3D protein viewer (Marco Biasini, 2015, pv v1.8.1. Zenodo. <https://doi.org/10.5281/zenodo.20980>) which provides zooming, panning and rotation, and hovering over an amino acid with a pointing device identifies the visualised amino acid and position (similar behaviour exists for ligands). DECIPHER variants are shown in this 3D view, allowing users to determine, for example if the variants are all within a DNA binding pocket or enzyme active site.

When looking at the protein browser from a patient record with a sequence variant, the location of the patient's variant is displayed by a vertical line, allowing easy orientation. In the case of a patient with a CNV, the protein browser is accessible from the CNV's genes tab, which displays a table of genes that overlap the CNV, along with other relevant information such as gene/disease association information and predictive scores. Clicking on a row displays further information about that gene, including the protein browser. An additional track is shown on the protein browser, indicating which part of the protein overlaps the CNV.

Genome browser: The Genoverse genome browser (<http://genoverse.org>), developed by the DECIPHER team, is a portable, interactive, customizable genome browser which allows the user to explore data. It displays a number of tracks containing information relevant to variant pathogenicity assessment such as genes associated with disease phenotypes (as curated and maintained by Online Mendelian Inheritance in Man (OMIM), <https://omim.org>, Amberger *et al.*, 2019), protein ortholog sequences from Ensembl indicating conservation, transcripts (as maintained by Ensembl), and regional missense constraint. Information from population resources such as gnomAD and Database of Genomic Variants (DGV) Gold (Church *et al.*, 2010) are displayed to enable users to determine if their patient's variant has been observed in healthy individuals. Disease relevant variant tracks are also available, which include DECIPHER sequence variants and CNVs, ClinVar sequence and structural variants, and variants from Human Gene Mutation Database (HGMD) public (Stenson *et al.*, 2020). The tracks which are displayed by default are tailored according to the type of variant being assessed.

Tools supporting diagnostic assessment

Assessing pathogenicity according to international standards

DECIPHER supports the annotation and sharing of variant pathogenicity using ACMG guidelines for sequence variants and ACMG/ClinGen technical standards for CNVs, which helps to standardise the classification of variants across centres. When interpreting a CNV it is possible for users to choose to assess the variant using sequence variant guidelines, which may be more applicable for small CNVs since the distinction between a sequence variant and a CNV is blurred (Brandt *et al.*, 2019).

Criteria selection: In both pathogenicity interfaces (Fig. 5A, 5B), types of evidence (such as population data and functional data) are displayed, along with the relevant evidence criteria used to determine if data supports the variant being pathogenic or benign. Relevant criteria can be selected with a single click. Some of the criteria have additional information links. These either provide information about how the criteria can be used according to the original paper (e.g. *de novo* CNV evidence), or in the case of sequence variants they provide information about ClinGen Sequence Variant Interpretation (SVI) Working Group guidelines (e.g. recommendation for functional assays (PS3/BS3) Brnich *et al.*, 2019). As new guidelines become available these pathogenicity interfaces are updated to provide the latest relevant recommendations. Criteria strengths can be modified as required in the interface.

Relevant evidence: Within the interfaces there is a customised section displaying ‘evidence to consider’ which provides information relating to the specific evidence type being assessed. For example, for computational and predictive data evidence, predictive pathogenicity scores (SIFT, PolyPhen-2, CADD, REVEL, SpliceAI) are displayed. Links are also provided to relevant DECIPHER interpretation interfaces, for example to the in-built tolerated population variation calculator, which can be used to determine if a variant observed in the reference sample is too common to cause a given Mendelian disorder of interest (Whiffin *et al.*, 2017). External links (e.g. PubMed literature search) are also provided.

Calculation of variant pathogenicity: As criteria are added, DECIPHER uses these to calculate the variant pathogenicity. For sequence variants, this is calculated according to the combining rules detailed in the original 2015 ACMG guidelines. In addition, DECIPHER calculates the posterior probability of pathogenicity and classification according to the ClinGen SVI Working Group’s Bayesian classification framework, which provides a mathematical foundation for the combining rules (Tavtigian *et al.*, 2018). DECIPHER highlights cases where these classifications disagree, and ultimately all pathogenicity assessments are made by depositors using their professional discretion. For CNVs, the evidence can be scored according to ACMG/ClinGen technical standards instead.

ClinGen Expert Panel specifications: For some genes there are ClinGen Variant Curation Expert Panel Specifications, which recommend adaptations of the sequence variant ACMG guidelines (e.g. Rett and Angelman-like Disorders Variant Curation Expert Panel for *MECP2*, *CDKL5*, *FOXP1*, *UBE3A*, *SLC9A6*, and *TCF4*). When interpreting variants in genes for which these recommendations exist, detailed information about how to apply the criteria is provided along with a link to the relevant Clinical Domain Working Group, so that patients with variants in these genes benefit from interpretation in accordance with these recommended standards.

Confirming variant-phenotype association and diagnosis

DECIPHER provides an assessment interface (Fig. 5D) which is designed to be used in a multidisciplinary team meeting to evaluate whether one or more variants explain the clinical features seen in a patient, and record if a diagnosis has been made (or excluded). Depositors can report evidence from several evidence lines, such as the age at presentation or additional clinical investigation, to weigh evidence for or against a genotype-phenotype relationship. An OMIM gene-disease pair and assertion is recorded, for example ‘genetic diagnosis confirmed’, ‘uncertain genetic diagnosis’, or ‘non-penetrant (or pre-symptomatic) for a dominant genetic disorder’. The output of the assessment is a date-stamped report providing the patient’s variants and phenotypes, in addition to the diagnosis and evidence on which that diagnosis was made.

There are many published examples of patients having compound phenotypes due to pathogenic variants

in more than one gene, for example, in Ferrer *et al.* , 2019, the patient had three independent rare disease diagnoses due to pathogenic variants in *SIN3A* (Witteveen–Kolk syndrome), *FLG* (dermatitis), and *EDAR* (ectodermal dysplasia). A recent study has suggested that multiple molecular diagnoses occur in approximately 5% of cases (Posey *et al.* , 2017). The assessment interface allows multiple assessments to be created for a patient, allowing the genetic basis of compound phenotypes to be recorded and shared.

Quantitative phenotypic data to confirm fit with diagnosis

Quantitative phenotype data and gene-specific centile charts: Quantitative phenotype data (developmental milestones or anthropometric measurements) can be recorded in DECIPHER, and are aggregated on a gene-by-gene basis and shared openly (Fig. 6A). In order for this information to be shown for a given gene there must be at least five patients with both quantitative phenotype data and openly shared sequence variants annotated as pathogenic/likely pathogenic. Once this threshold is met, DECIPHER automatically aggregates and shares the information as a series of graphs on which expectations for the predominantly healthy population ('Normal'), the DECIPHER population as a whole, and the gene-specific data is plotted. Anthropometric measurements are plotted around the standard deviation (adjusted for sex and gestation, where possible), while developmental milestones are plotted against time. The standard deviation for each population is displayed at the bottom of the graph as a boxplot. For users logged into DECIPHER and looking at a patient record from their centre, a vertical line indicates their patient's measurement or age at attainment of the milestone, allowing them to easily judge whether it is consistent with a pathogenic/likely pathogenic variant in the gene. The display of the DECIPHER population allows users to determine if a particular measurement is particularly discriminative for a given disorder. These gene-specific centile charts can also be used in the clinic to determine how a child is developing relative to other children with the same disorder.

Composite facial images: For certain genes there also are composite faces, which highlight facial dysmorphologies specific to a gene. These anonymised composite face images have been created from individuals with *de novo* mutations in the affected genes that were collected through the DDD study (Deciphering Developmental Disorders, Study, 2017).

Identifying patient matches to supporting diagnosis

Matchmaking within DECIPHER

The phenotype-linked variant data in DECIPHER allows for effective patient matching. DECIPHER presents a powerful, flexible matching patient interface (Fig. 6B), which allows users to view DECIPHER records which overlap a deposited copy-number, sequence, or insertion variant, or a gene. The matching patient interface displays useful summary information about the potential matches, for example, for sequence variants, this is consequence, inheritance, and pathogenicity. To allow users to quickly identify the most prominent clinical features in overlapping patients, a list of phenotypes present in multiple matching patients is displayed. When viewing this interface from a patient record, additional lists showing which of the patient's phenotypes are present or not recorded in matching patients are also displayed. This allows users to easily determine if there is a good phenotypic match between their patient and other matching patients. Beneath this is a table containing information about the individual matching patient records. The table columns can be sorted and all matching phenotypes are shown in bold.

Customisable data display: A series of filters are provided in the matching patient interface so that users can drill into the most relevant patient data. This allows users to filter on, for example, functional similarity, consequence, inheritance, and/or pathogenicity. This can be particularly useful when different variant consequences are associated with different syndromes (e.g. *SCN2A* , where loss of function variants are associated with nonspecific severe intellectual disability, and missense variants with infantile epileptic encephalopathy).

Functionally identical variants: If the same variant has previously been deposited to DECIPHER, a 'Functionally Identical Variant' interface is present, displaying variant pathogenicity and evidence, in addition to phenotype information from these patient records. This ensures that users are alerted to other patients

carrying the same variant, and assists in the standardisation of variant classification across centres.

Discriminative phenotypes: The wealth of the phenotype-genotype linked data in DECIPHER also allows the aggregation of data associated with pathogenic variants in disease genes. Within DECIPHER, aggregated phenotype data is used to identify the most discriminating phenotypes associated with disease genes (Fig. 6C). Recognising distinctive clinical characteristics associated with a disorder can be key to a diagnosis. The interface presents a table displaying the percentage of phenotyped patients with sequence variants in a gene of interest with a particular phenotype, compared with the percentage of phenotyped patients in DECIPHER with the same phenotype, and the odds ratio and p -value from a Fisher's exact test, which indicate the most discriminative phenotypes associated with a gene.

Clinician contact: If a matching patient is discovered, it is possible to contact the clinician responsible for the patient's care through DECIPHER. DECIPHER depositors are able to send messages directly, and since October 2014, over 4,500 collaboration requests have been sent amongst these registered DECIPHER users. In the case where a user is not registered with DECIPHER, the DECIPHER team first moderates such contact requests, and if the request appears to be legitimate and appropriate, forwards the message to the clinician responsible for the patient, asking them to contact the requestor directly to discuss collaboration. Over 2,900 such contact requests have been sent since January 2018.

Matchmaking through the Matchmaker Exchange

DECIPHER is a founding member of the Matchmaker Exchange (MME, <https://matchmakerexchange.org>), a Global Alliance for Genomics and Health (GA4GH) driver project which enables the federated discovery of similar rare disease patient data in connected databases. This worldwide collaboration allows automated matchmaking of genetic and/or phenotypic data between databases, via an application programming interface (API). Through MME, DECIPHER is currently connected to Broad-seqr (<https://seqr.broadinstitute.org/matchmaker/matchbox>, Arachchi et al., 2018), GeneMatcher (<https://genematcher.org>, Sobreira et al., 2015), MyGene2 (<https://www.mygene2.org/>, MyGene2, 2016), PhenomeCentral (<https://phenomecentral.org>, Buske et al., 2015) and RD-Connect (<https://platform.rd-connect.eu>, Lochmüller et al., 2018). Since 2020, DECIPHER depositors have made approximately 1,500 requests for matches from connected databases and received details of more than 4,100 potential patient matches. In the same time period, DECIPHER has received more than 55,000 requests for matches from connected databases, and has returned details of more than 255,000 potential patient matches.

Within DECIPHER, users with write access to a patient are able to query the MME. It is essential that the patient record in DECIPHER has explicit consent for open sharing, as some connected databases have dual notification i.e. they provide their user with details of any potential patient match, and unshared patient records will not be available to users of the other databases. Once MME is queried and the connected databases have responded, details of potential patient matches are displayed within the DECIPHER interface. Potential matches from each database are displayed in a tabular format with matching phenotypes in bold, assisting users in determining the level of phenotype similarity (Fig. 6D). DECIPHER supports the querying of MME for patients with at least one open-access sequence or a copy-number variant which overlaps one gene. Other types of variants present in the patient record will not be included in the MME request.

When a MME request is sent to DECIPHER which contains genomic information, all open-access patient sequence or copy-number variants which overlap a single gene, and all DDD consortium research variants (see Driving rare disease research section) are evaluated for similarity based on functional overlap. Many of the variant requests received from connected databases provide genomic coordinates in GRCh37, and in these cases DECIPHER performs liftover to convert the coordinates to GRCh38 prior to identifying matches. A score for each potential patient match is provided, ranging from 0 to 1, with 1 indicating a better match. DECIPHER's scoring algorithm for genomic matches takes into account the Ensembl VEP predicted consequence, assessing the severity and similarity of the consequence to those provided in the request.

If only phenotypic data is provided, all open-access patients with phenotypes are evaluated for a match. This takes into account all HPO ancestor terms for both the patient in the request, and patients within

DECIPHER. These matches are scored by generating an Intersection over Union score comparing the HPO ancestor terms of the request patient and the patient in DECIPHER.

DECIPHER returns the 20 highest scoring matches per MME request. In the case where there are many matches, the patients' chromosomal sex is taken into account in addition to the score, in order to prioritise the best possible matches. The returned matches include variant, phenotype (including absent phenotypes) and diagnosis information.

Driving rare disease research

The ~40,000 openly consented patient records in DECIPHER contain more than 51,000 variants and ~172,000 phenotypes, and represents a rich dataset to drive rare disease research. Since its inception in 2004, DECIPHER has been cited more than 2,600 times in peer reviewed publications (Fig. 7A); a testimony to its impact on rare disease research. In some cases there is a large genotypic patient series, which allows, for example, the full spectrum of phenotypes associated with a gene to be recognised. At the time of writing, the genes with the most open access sequence variants were *NF1* (162), *ANKRD11* (123), *ARID1B* (107), *KMT2A* (107), and *DDX3X* (78) (Fig. 7B).

Search: To identify the most relevant patient records and gene information DECIPHER offers a powerful search function allowing users to search using many different categories including gene, phenotype, HPO identifier, genomic position (in GRCh37 or GRCh38), chromosome band, pathogenicity, and inheritance. Advanced searches are supported, such as searching for multiple terms either from the same category (e.g. multiple phenotypes) or different categories (e.g. gene plus phenotype). Results are displayed in a tabular format, in addition to genome browser-based representations.

Driving discovery: The genotype-linked phenotypic data allows, for example, new variant-disease associations to be discovered, such as loss-of-function variants in *ARFGEF1* causing developmental delay and epilepsy (Thomas *et al.* , 2021). The dataset also enables the extension of phenotypes for new syndromes to be uncovered (e.g. Witteveen-Kolk syndrome a *SIN3A* -related disorder Balasubramanian *et al.* , 2021), in addition to well established syndromes (e.g. *ALG13* congenital disorder of glycosylation Alsharhan *et al.* , 2021). It also permits the understanding of contiguous gene effects, such as that around *ERF* which causes a novel craniosynostosis syndrome with varying degrees of intellectual disability (Calpena *et al.* , 2021).

DDD Research variants: In addition to the openly consented patient data, DECIPHER openly shares the DDD research variants, which are variants of unknown significance identified in undiagnosed probands with developmental disorders in the DDD study. These include functional *de novo* variants and rare loss-of-function homozygous, compound heterozygous, and hemizygous variants in genes that are neither developmental disorder genes, nor OMIM-morbid genes. At present this dataset comprises nearly 5,000 variants. High-level phenotype terms are provided for each variant (Fig. 7C). The number of patients with each variant in the DDD dataset is displayed, in addition to the number of patients identified in the GeneDx and Radboud University Medical Center *de novo* variant dataset as described by Kaplanis *et al.* , 2020. This dataset enables the discovery of new gene-disease associations.

Bulk data for research: The openly consented patient data is available for bulk download for research purposes, subject to a data access agreement. In bulk the data can be used, for example, for developing new analytical methods, in understanding patterns of polymorphism, and in refining critical intervals to map genes involved in specific phenotypes and diseases. The dataset has recently been used to associate phenotypes with functional systems (Jabato *et al.* , 2021), and to develop a new tool to assist clinical interpretation of CNVs (Requena *et al.* , 2021). DECIPHER also shares the data in bulk for display, subject to a Data Display Agreement. This allows third-party variant analysis companies and the academic genome browser providers such as Ensembl and UCSC to display the data, maximising the possibility of finding patient matches.

Summary

DECIPHER is a free web-based platform which enables the visualisation of genomic and phenotypic relationships to aid variant interpretation, diagnosis, and discovery. The platform supports the interpretation

and sharing of almost all types of genetic variation, providing variant interpretation interfaces which contextualise the genotypic and phenotypic data. These interfaces include a genome browser, protein browser, matching patient variant displays, and tools to assess the variant according to internationally-accepted standards. Potential matching patients in other connected databases can also be identified through the MME. The platform enables the flexible and proportionate sharing of patient-level data, so that the depth and breadth of sharing is tailored to the scientific/clinical needs and the level of patient consent attained. DECIPHER currently openly shares ~40,000 rare disease patient records, and supports the more limited sharing of >63,000. DECIPHER is under continuous development, ensuring that it keeps up to date with the fast moving field of rare genetic disease. New user-facing features are released approximately every six weeks, along with updates to reference data sources (such as the Ensembl/GENCODE gene set, HPO, ClinVar). DECIPHER enables clinical use of selected new datasets and tools developed by the research community. This makes them directly available to clinicians and clinical scientists, thereby assisting in the rapid translation of research into the diagnostic arena. Since its inception in 2004, the platform has made a huge impact on rare genetic disease research, and is cited in more than 2,600 publications. The rich phenotype-linked variant data hosted by DECIPHER, and the tools it provides, enable DECIPHER to advance its mission of mapping the clinically relevant parts of the genome.

Acknowledgements

The authors thank the patients and their families for their permission to include their information in DECIPHER. The authors would also like to thank all registered DECIPHER users for depositing and seeking consent to share patient data. The DECIPHER project was given a favourable NRES REC opinion by Cambridge South (previously Cambridgeshire 4 REC), REC reference 04/MRE05/50, in November 2004. DECIPHER submits annual progress reports to ensure this favourable opinion applies for the duration of the research. DECIPHER is supported by Wellcome funding, grant WT206194. Helen Firth is supported by The Wellcome award WT200990/Z/16/Z Designing, developing and delivering integrated foundations for genomic medicine. Fiona Cunningham and Sarah E. Hunt receive funding from the Wellcome Trust (grant number WT108749/Z/15/Z) and the European Molecular Biology Laboratory. This research was funded in whole, or in part, by the Wellcome Trust [Grant numbers WT206194, WT200990/Z/16/Z, WT108749/Z/15/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. Matthew Hurles is a co-founder, shareholder and non-executive director of Congenica Ltd., a diagnostic software company. Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Conflict of Interests

Matthew Hurles is a co-founder, shareholder and non-executive director of Congenica Ltd., a diagnostic software company.

References

- Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*, Chapter 7, Unit7 20. doi:10.1002/0471142905.hg0720s76
- Alsharhan, H., He, M., Edmondson, A. C., Daniel, E. J. P., Chen, J., Donald, T., . . . Sobering, A. K. (2021). ALG13 X-linked intellectual disability: New variants, glycosylation analysis, and expanded phenotypes. *J Inherit Metab Dis*, 44(4), 1001-1012. doi:10.1002/jimd.12378
- Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*, 47(D1), D1038-D1043. doi:10.1093/nar/gky1151
- Arachchi, H., Wojcik, M. H., Weisburd, B., Jacobsen, J. O. B., Valkanas, E., Baxter, S., . . . Rehm, H. L. (2018). matchbox: An open-source tool for patient matching via the Matchmaker Exchange. *Hum Mutat*, 39(12), 1827-1834. doi:10.1002/humu.23655
- Balasubramanian, M., Dingemans, A. J. M., Albaba, S., Richardson, R., Yates, T. M., Cox, H., . . . Kleefstra,

- T. (2021). Comprehensive study of 28 individuals with SIN3A-related disorder underscoring the associated mild cognitive and distinctive facial phenotype. *Eur J Hum Genet*, 29(4), 625-636. doi:10.1038/s41431-020-00769-7
- Boggan, R. M., Lim, A., Taylor, R. W., McFarland, R., & Pickett, S. J. (2019). Resolving complexity in mitochondrial disease: Towards precision medicine. *Mol Genet Metab*, 128(1-2), 19-29. doi:10.1016/j.ymgme.2019.09.003
- Bragin, E., Chatzimichali, E. A., Wright, C. F., Hurles, M. E., Firth, H. V., Bevan, A. P., & Swaminathan, G. J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res*, 42(Database issue), D993-D1000. doi:10.1093/nar/gkt937
- Brandt, T., Sack, L. M., Arjona, D., Tan, D., Mei, H., Cui, H., . . . Meck, J. M. (2020). Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. *Genet Med*, 22(2), 336-344. doi:10.1038/s41436-019-0655-2
- Brnich, S. E., Abou Tayoun, A. N., Couch, F. J., Cutting, G. R., Greenblatt, M. S., Heinen, C. D., . . . Clinical Genome Resource Sequence Variant Interpretation Working, G. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med*, 12(1), 3. doi:10.1186/s13073-019-0690-2
- Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., . . . Brudno, M. (2015). PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum Mutat*, 36(10), 931-940. doi:10.1002/humu.22851
- Calpena, E., McGowan, S. J., Blanco Kelly, F., Boudry-Labis, E., Dieux-Coeslier, A., Harrison, R., . . . Wilkie, A. O. M. (2021). Dissection of contiguous gene effects for deletions around ERF on chromosome 19. *Hum Mutat*, 42(7), 811-817. doi:10.1002/humu.24213
- Campbell, P., Ellingford, J. M., Parry, N. R. A., Fletcher, T., Ramsden, S. C., Gale, T., . . . Sergouniotis, P. I. (2019). Clinical and genetic variability in children with partial albinism. *Sci Rep*, 9(1), 16576. doi:10.1038/s41598-019-51768-8
- Chatzimichali, E. A., Brent, S., Hutton, B., Perrett, D., Wright, C. F., Bevan, A. P., . . . Swaminathan, G. J. (2015). Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. *Hum Mutat*, 36(10), 941-949. doi:10.1002/humu.22842
- Church, D. M., Lappalainen, I., Sneddon, T. P., Hinton, J., Maguire, M., Lopez, J., . . . Flicek, P. (2010). Public data archives for genomic structural variation. *Nat Genet*, 42(10), 813-814. doi:10.1038/ng1010-813
- Deciphering Developmental Disorders, S. (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542(7642), 433-438. doi:10.1038/nature21062
- den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., . . . Taschner, P. E. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*, 37(6), 564-569. doi:10.1002/humu.22981
- Ferrer, A., Schultz-Rogers, L., Kaiwar, C., Kemppainen, J. L., Klee, E. W., & Gavrilova, R. H. (2019). Three rare disease diagnoses in one patient through exome sequencing. *Cold Spring Harb Mol Case Stud*, 5(6). doi:10.1101/mcs.a004390
- Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., . . . Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*, 84(4), 524-533. doi:10.1016/j.ajhg.2009.03.010
- Grady, J. P., Pickett, S. J., Ng, Y. S., Alston, C. L., Blakely, E. L., Hardy, S. A., . . . McFarland, R. (2018). mtDNA heteroplasmy level and copy number indicate disease burden in m.3243A>G mitochondrial disease. *EMBO Mol Med*, 10(6). doi:10.15252/emmm.201708262

- Groesser, L., Herschberger, E., Ruetten, A., Ruivenkamp, C., Lopriore, E., Zutt, M., . . . Hafner, C. (2012). Postzygotic HRAS and KRAS mutations cause nevus sebaceous and Schimmelpenning syndrome. *Nat Genet*, 44(7), 783-787. doi:10.1038/ng.2316
- Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet*, 58(8), 547-555. doi:10.1136/jmedgenet-2020-107003
- Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., . . . Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet*, 99(4), 877-885. doi:10.1016/j.ajhg.2016.08.016
- Jabato, F. M., Seoane, P., Perkins, J. R., Rojano, E., Garcia Moreno, A., Chagoyen, M., . . . Ranea, J. A. G. (2021). Systematic identification of genetic systems associated with phenotypes in patients with rare genomic copy number variations. *Hum Genet*, 140(3), 457-475. doi:10.1007/s00439-020-02214-7
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., . . . Farh, K. K. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176(3), 535-548 e524. doi:10.1016/j.cell.2018.12.015
- Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., . . . Retterer, K. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, 586(7831), 757-762. doi:10.1038/s41586-020-2832-5
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434-443. doi:10.1038/s41586-020-2308-7
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3), 310-315. doi:10.1038/ng.2892
- Kohler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J. P., . . . Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res*, 47(D1), D1018-D1027. doi:10.1093/nar/gky1105
- Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief Bioinform*, 14(2), 144-161. doi:10.1093/bib/bbs038
- Lochmuller, H., Badowska, D. M., Thompson, R., Knoers, N. V., Aartsma-Rus, A., Gut, I., . . . consortium, E. U. (2018). RD-Connect, NeurOmics and EURenOmics: collaborative European initiative for rare diseases. *Eur J Hum Genet*, 26(6), 778-785. doi:10.1038/s41431-018-0115-5
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol*, 17(1), 122. doi:10.1186/s13059-016-0974-4
- Miller, D. (2021). The diagnostic odyssey: our family’s story. *Am J Hum Genet*, 108(2), 217-218. doi:10.1016/j.ajhg.2021.01.003
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., . . . Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res*, 49(D1), D412-D419. doi:10.1093/nar/gkaa913
- Monroe, G. R., Frederix, G. W., Savelberg, S. M., de Vries, T. I., Duran, K. J., van der Smagt, J. J., . . . van Haaften, G. (2016). Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet Med*, 18(9), 949-956. doi:10.1038/gim.2015.200
- Monroe, G. R., Frederix, G. W., Savelberg, S. M., de Vries, T. I., Duran, K. J., van der Smagt, J. J., . . . van Haaften, G. (2016). Effectiveness of whole-exome sequencing and costs of the traditional diagnostic

- trajectory in children with intellectual disability. *Genet Med*, 18(9), 949-956. doi:10.1038/gim.2015.200
- MyGene2. (2016). Website aims to accelerate gene discovery, diagnosis, treatment: MyGene2.org fosters open sharing among families, researchers, and clinicians. *Am J Med Genet A*, 170(6), 1388-1389. doi:10.1002/ajmg.a.37746
- Posey, J. E., Harel, T., Liu, P., Rosenfeld, J. A., James, R. A., Coban Akdemir, Z. H., . . . Lupski, J. R. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N Engl J Med*, 376(1), 21-31. doi:10.1056/NEJMoa1516767
- Quaio, C., Moreira, C. M., Novo-Filho, G. M., Sacramento-Bobotis, P. R., Groenner Penna, M., Perazzio, S. F., . . . Baratela, W. (2020). Diagnostic power and clinical impact of exome sequencing in a cohort of 500 patients with rare diseases. *Am J Med Genet C Semin Med Genet*, 184(4), 955-964. doi:10.1002/ajmg.c.31860
- Rahit, K., & Tarailo-Graovac, M. (2020). Genetic Modifiers and Rare Mendelian Disease. *Genes (Basel)*, 11(3). doi:10.3390/genes11030239
- Requena, F., Abdallah, H. H., Garcia, A., Nitschke, P., Romana, S., Malan, V., & Rausell, A. (2021). CNVxplorer: a web tool to assist clinical interpretation of CNVs in rare disease patients. *Nucleic Acids Res*, 49(W1), W93-W103. doi:10.1093/nar/gkab347
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (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), 405-424. doi:10.1038/gim.2015.30
- Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., . . . Martin, C. L. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet Med*, 22(2), 245-257. doi:10.1038/s41436-019-0686-8
- Sawyer, S. L., Hartley, T., Dymont, D. A., Beaulieu, C. L., Schwartzentruber, J., Smith, A., . . . Boycott, K. M. (2016). Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin Genet*, 89(3), 275-284. doi:10.1111/cge.12654
- Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, 40(Web Server issue), W452-457. doi:10.1093/nar/gks539
- Sobreira, N., Schiettecatte, F., Valle, D., & Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*, 36(10), 928-930. doi:10.1002/humu.22844
- Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., . . . Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD((R))) : optimizing its use in a clinical diagnostic or research setting. *Hum Genet*, 139(10), 1197-1207. doi:10.1007/s00439-020-02199-3
- Stranneheim, H., Lagerstedt-Robinson, K., Magnusson, M., Kvarnung, M., Nilsson, D., Lesko, N., . . . Wedell, A. (2021). Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med*, 13(1), 40. doi:10.1186/s13073-021-00855-5
- Swaminathan, G. J., Bragin, E., Chatzimichali, E. A., Corpas, M., Bevan, A. P., Wright, C. F., . . . Firth, H. V. (2012). DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet*, 21(R1), R37-44. doi:10.1093/hmg/dds362
- Tan, A., Abecasis, G. R., & Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics*, 31(13), 2202-2204. doi:10.1093/bioinformatics/btv112

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., . . . ClinGen Sequence Variant Interpretation Working, G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*, 20(9), 1054-1060. doi:10.1038/gim.2017.210

Thomas, Q., Gautier, T., Marafi, D., Besnard, T., Willems, M., Moutton, S., . . . Vitobello, A. (2021). Haploinsufficiency of ARFGEF1 is associated with developmental delay, intellectual disability, and epilepsy with variable expressivity. *Genet Med*. doi:10.1038/s41436-021-01218-6

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., . . . Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med*, 19(10), 1151-1158. doi:10.1038/gim.2017.26

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., van Kogelenberg, M., . . . study, D. D. D. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, 385(9975), 1305-1314. doi:10.1016/S0140-6736(14)61705-0

Wright, C. F., FitzPatrick, D. R., & Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*, 19(5), 253-268. doi:10.1038/nrg.2017.116

Figure legends

Figure 1. A: The DECIPHER community is a global network of academic clinical centres with expertise in genetics. Depositing centres are able to send messages directly to other registered users about patient matches through DECIPHER. Since October 2014 over 4,500 such messages have been sent. Here, each line represents a collaboration request sent between depositing centres. Unregistered users' messages, sent through DECIPHER, are not included in this image. **B:** The DECIPHER database currently openly shares approx. 40,000 rare disease patient records, built up over time.

Figure 2. DECIPHER supports the deposition and sharing of almost all types of genetic variation.

Figure 3. All genomic data is visualised in GRCh38, but deposition is still supported in GRCh37/hg19. Tools are provided to visualise the differences between assemblies. These include comparative genome browsers and gene lists for variants lifted over by DECIPHER, and a liftover mapping genome browser track.

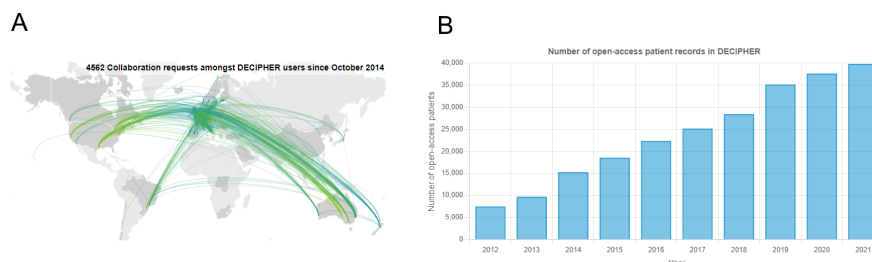
Figure 4. A: DECIPHER enables the deposition of phenotypes using HPO terms. **B:** DECIPHER supports the deposition of developmental milestones and anthropometric measurements, e.g. occipitofrontal (head) circumference.

Figure 5. A: DECIPHER has developed a protein browser which summarises genotypic data. Tracks include: Pfam domains, DECIPHER and ClinVar variants, gnomAD variants, and region of predicted nonsense mediated decay (NMD) escape. **B:** DECIPHER supports the annotation and sharing of sequence variant pathogenicity assessments using ACMG guidelines. A pathogenicity evidence interface is available for depositors. Relevant criteria are selected by clicking on the criteria displayed on the left under "Available evidence types". "Selected criteria" are displayed on the right, along with "Evidence to consider". "Further information" links provide recommendations for the use of criteria. In this example, a variant in *SLC9A6* is being annotated and ClinGen Variant Curation Expert Panel Specifications exist for this gene. Detailed information about these recommendations are displayed by clicking on the "Gene recommendation" links - expert panel recommendations for *de novo* criterion PS2 are displayed. As criteria are added, DECIPHER calculates the variant pathogenicity according to criteria-combining rules detailed in the original 2015 guidelines, and according to the ClinGen SVI Working Group's Bayesian classification framework. **C:** DECIPHER supports the annotation of copy-number variants according to ACMG/ClinGen technical standards. Similar to the sequence variant interface, "Available evidence types" are displayed on the left, with "Selected evidence" and "Evidence to consider" displayed on the right. As criteria are selected, the classification score and pathogenicity are calculated and displayed at the bottom of the interface. **D:** An assessment interface is provided which is designed to be used in a multidisciplinary team meeting to evaluate whether one or more variants explain the clinical features seen in a patient, and record if a diagnosis has been made (or exclu-

ded). Depositors can report several lines of evidence, to weigh evidence for or against a genotype-phenotype relationship. An OMIM gene-disease pair and assertion is recorded.

Figure 6. A: Quantitative phenotype data (such as developmental milestones or anthropometric measurements) is recorded in DECIPHER, and aggregated on a gene-by-gene basis. The data is shared openly in a series of graphs which displays expectations for the healthy population (‘Normal’), the DECIPHER population as a whole, and the gene-specific data. For certain genes, such as *EP300* (displayed here), there are composite faces, which highlight facial dysmorphologies. **B:** The matching patient interface allows users to view DECIPHER records which overlap a deposited copy-number, sequence, or insertion variant, or a gene. In this example, the matching patients overlap *EP300*. Summary information is shown in a series of pie charts, along with phenotypes present in multiple matching patients. The individual patient records are displayed at the bottom of the interface. Filters are available to assist in finding the most relevant patient matches. **C:** Within DECIPHER, aggregated phenotype data is used to identify the most discriminating phenotypes associated with disease genes. A table shows the percentage of phenotyped patients with sequence variants in a gene of interest, with a particular phenotype, compared with the percentage of phenotyped patients in DECIPHER with the same phenotype. The odds ratio and *p*-value from a Fisher’s exact test are displayed. In this example, data for *KMT2A* is displayed and sorted by *p*-value. **D :** Users with write access to an open-access patient record are able to query the MatchMaker Exchange to search for potential patient matches. DECIPHER is currently connected to Broad-seqr, GeneMatcher, MyGene2, PhenomeCentral and RD-Connect. Details of potential patient matches are displayed within DECIPHER (patient IDs have been removed in this example).

Figure 7. A: Since its inception in 2004, DECIPHER has been cited in more than 2,600 publications. **B:** The genes with the most open access sequence variants in DECIPHER (at the time of writing). **C:** DECIPHER openly shares variants of unknown significance identified in undiagnosed probands in the Deciphering Developmental Disorders study (research variants). For each variant a page provides details of the variant and high-level phenotype terms. The number of patients with each variant in the DDD dataset is displayed, in addition to the number of patients identified in the GeneDx and Radboud University Medical Center *de novo* variant dataset as described by Kaplanis *et al* ., 2020.

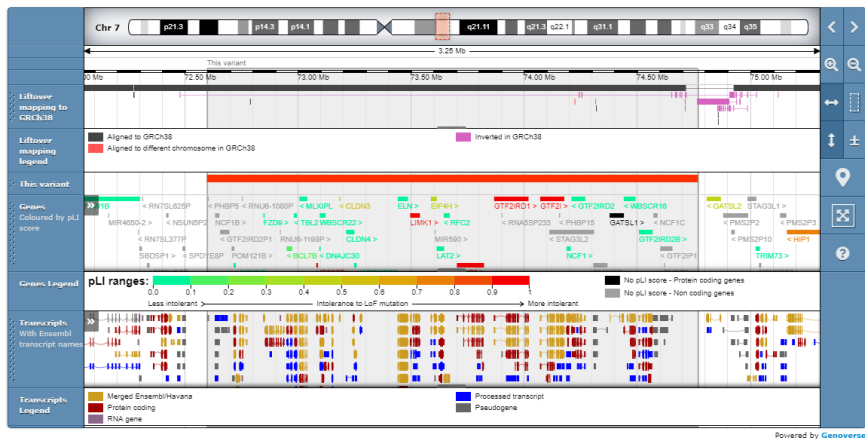


Sequence Variant	Uniparental Disomy
Sequence Variant	Isodisomy
Copy-Number Variant	Heterodisomy
Deletion	Unknown
Duplication	Inversion
Duplication/Triplication	Inversion
Triplication	Insertion
Amplification (> Triplication)	Mobile Element: Alu
Aneuploidy/Segmental aneuploidy	Mobile Element: LINE1
Nullisomy	Mobile Element: SVA
Monosomy	Retrogene
Disomy	Other Insertion
Trisomy	Short Tandem Repeat
Tetrasomy	Short Tandem Repeat

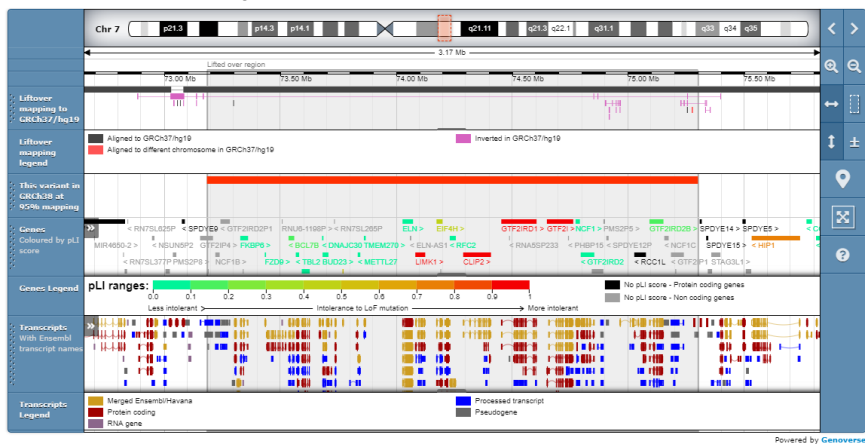
The GRCh37/hg19 location of this variant overlaps the following genes, which its GRCh38 location does not. Protein coding genes are shown in bold.

- **WBSCR16** (ENSG00000174374.9)
- **GATSL1** (ENSG00000183086.10)

GRCh37/hg19 2.17 Mb



GRCh38 2.11 Mb - 2% shorter than in GRCh37/hg19





A

Patient <input type="text" value="4"/>	Mother <input type="text" value="0"/>	Father <input type="text" value="0"/>	Add...
--	---------------------------------------	---------------------------------------	--------



Patient phenotypes

[Show simple term list](#)



Abnormality of the eye

Hypertelorism  

Abnormality of head or neck



Depressed nasal bridge  

Abnormality of the musculoskeletal system

Microcephaly  

- Progressive

Abnormality of the nervous system

Global developmental delay  

B

