Improved diagnosis of rare disease patients through application of constrained coding region annotation and de novo status.

Hywel J Williams), Chris Odhams , and Genomics England Research
Chambia $Consortium²$

|
|
|
|
| 1 Tenetics and Tenetics and Generation, Tenetics and Generatic, Tenetics and Division, Dawson H
University, Cardiff CF14 4AY. 2 Genomics England, Queen Mary University of London, Dawson H
London, EC1M 6BQ. * Corresponding

University, Cardinal Crussition, Cardinal Crussity, Cardinal, Youth Crussity, Occurring,
London, EC1M 6BQ. * Corresponding author.
Identifying the pathogenic variant in a rare disease (RD) patient is the first step in endi Mostract:
Abstract:
Identifying the pathogenic variant in a rare distinguished pathogenic variant in a rare distinguished
of Mendelian disorders in RD patients. Constr
coding DNA which are devoid of functional va ノー にっこく Identifyin
diagnostic
of Mende
coding DN
constraine
pathogen Identifying the pathogenic variants in a rate in pyramic coding DNA are a well-established
diagnostic odyssey. De novo (Dn) variants affecting protein-coding DNA are a well-established
of Mendelian disorders in RD patients of Mendelian disorders in RD patients. Constrained coding regions (CCRs) are specific segments of
coding DNA which are devoid of functional variants in healthy individuals. Furthermore, the most
constrained regions, those

or Mendelian disorders in the disorders in healthy individuals. Furthermore, the most
constrained regions, those in percentile bin >95 (CCR95), are significantly enriched for functional
pathogenic variants and could theref constrained regions, those in percentile bin >95 (CCR95), are significantly enriched for functional
pathogenic variants and could therefore be useful for clinical variant prioritisation.
We aimed to evaluate the diagnostic pathogenic variants and could therefore be useful for clinical variant prioritisation.
We aimed to evaluate the diagnostic utility of incorporating Dn, CCR95 and Dn_CCR95 status into
the variant prioritisation cascade for We aimed to evaluate the diagnostic utility of incorporating Dn, CCR95 and Dn_CCF
the variant prioritisation cascade for RD patients that have undergone genomic seq
data from the Genomics England 100,000 Genomes Project v1 the variant prioritisation cascade for RD patients that have undergone genomic sequencing. Using
data from the Genomics England 100,000 Genomes Project v12, we selected 3,090 trios that have
undergone diagnostic evaluation

data from the Genomics England 100,000 Genomes Project v12, we selected 3,090 trios that have
undergone diagnostic evaluation and been analysed with an advanced Dn identification pipeline. F
this analysis we have excluded and the Genomics England 100,000 Central transfer transfer transfer in the material properties. F
this analysis we have excluded all non-autosomal variants.
Our analysis shows the diagnostic rate increased from 71% in the undergone diagnostic exactled all non-autosomal variants.

Our analysis shows the diagnostic rate increased from 71% in the full cohort to 81% when evaluating

just the CCR95 variants, 84% for Dn variants and 87% for Dn_CC Our analysis shows the diagnostic rate increased from 71%
just the CCR95 variants, 84% for Dn variants and 87% for Di
evaluation of the Dn_CCR95 variants from the undiagnosed
64% of patients (25 of 39), suggesting applicat Just the CCR95 variants, 84% for Dn variants and 87% for Dn_CCR variants. Of note, manual
evaluation of the Dn_CCR95 variants from the undiagnosed patients revealed a putative diagnosis in
64% of patients (25 of 39), sugge

just the Co_LCCR95 variants from the undiagnosed patients revealed a putative dia
64% of patients (25 of 39), suggesting application of this metric can prioritise likely pathoger
variants in undiagnosed patients.
We also 64% of patients (25 of 39), suggesting application of this metric can prioritise likely pathogenic
variants in undiagnosed patients.
We also identify a striking enrichment of signal in patients with a phenotype of neurolog variants in undiagnosed patients.

We also identify a striking enrichment of signal in patients with a phenotype of neurology and

neurodevelopmental disorders, whereby their diagnostic rate increases from 60% in the whole variants in undigitively procedure.

We also identify a striking enrichn

neurodevelopmental disorders, w

cohort to 71%, 73% and 74% in th

to the next largest phenotypic gro

are 10%, 3% 2% and 1%.

In summary, we demons neurodevelopmental disorders, whereby their diagnostic rate increases from 60% in the whole
cohort to 71%, 73% and 74% in the Dn, CCR95 and Dn_CCR95 categories respectively. This com
to the next largest phenotypic group, O

cohort to 71%, 73% and 74% in the Dn, CCR95 and Dn_CCR95 categories respectively. This compto the next largest phenotypic group, Ophthalmological disorders, where the corresponding val
are 10%, 3% 2% and 1%.
In summary, we compared to the next largest phenotypic group, Ophthalmological disorders, where the corresponding values
are 10%, 3% 2% and 1%.
In summary, we demonstrate the potential clinical utility of performing bespoke Dn analyses o are 10%, 3% 2% and 1%.
In summary, we demonstrate the potential clinical utility of performing bespoke Dn analyses of RD
patients and for incorporating CCR information into the filtering cascade to prioritise pathogenic
va are 10%, 3% 10% 10% 10% 10%.
In summary, we demonst
patients and for incorpor
variants. We believe such
the time taken to make a
samples to be analysed of patients and for incorporating CCR information into the filtering cascade to prioritise pathogenic
variants. We believe such a strategy will aid the identification of pathogenic variants and decrease
the time taken to make patients. We believe such a strategy will aid the identification of pathogenic variants and decreas
the time taken to make a diagnosis, thus increasing the overall diagnostic rate by allowing more
samples to be analysed ov the time taken to make a diagnosis, thus increasing the overall diagnostic rate by allowing more
samples to be analysed over the same time period.
TE: This preprint reports new research that has not been certified by peer the time taken to make a diagnosis, who interesting the overall diagnosis rate over a_p allowing more samples to be analysed over the same time period.
TE: This preprint reports new research that has not been certified by

Introduction:

The use of nexelone
genome seque
practice in gen
time taken to
continuing to
reached a pos
and thus there The use of next generation of the diagnosis of patients with rare diseases (RDs) is becoming routine
practice in genetic diagnostic laboratories world-wide. This has led to improvements in both the
time taken to reach a di genome in genetic diagnostic laboratories world-wide. This has led to improvements in both the taken to reach a diagnosis and the overall diagnostic rate(1-3). With the cost of sequencic continuing to fall and automation i ime taken to reach a diagnosis and the overall diagnostic rate(1-3). With the cost of sequencing
continuing to fall and automation increasing the number of samples to be sequenced, we have
reached a position where the inte

the final and automation increasing the number of samples to be sequenced, we have
reached a position where the interpretation of the data being generated is becoming a bottlened
and thus there is the potential the diagnos reached a position where the interpretation of the data being generated is becoming a bottlene
and thus there is the potential the diagnostic rate will plateau(4, 5).
The difficulty in interpreting NGS data is down to the and thus there is the potential the diagnostic rate will plateau(4, 5).
The difficulty in interpreting NGS data is down to the sheer volume of variants identified when we
compare a person's DNA to the reference genome and The difficulty in interpreting NGS data is down to the sheer volume compare a person's DNA to the reference genome and the observation
number of these variants have characteristics overlapping with path
will be benign(6, 7 The difference is an above and the observation that even though a great
compare a person's DNA to the reference genome and the observation that even though a great
number of these variants have characteristics overlapping compare a person in standard and the characteristics overlapping with pathogenic variants, most, if not a
number of these variants have characteristics overlapping with pathogenic variants, most, if not a
will be benign(6, will be benign(6, 7). In order to simplify the interpretation of NGS data, scientists rely on a suite of
annotation tools to help filter out the benign variants and leave a small list of potentially pathogenic
variants tha annotation tools to help filter out the benign variants and leave a small list of potentially pathogen
variants that can be studied in greater detail. There are now, a range of prioritisation tools availabl
to perform this variants that can be studied in greater detail. There are now, a range of prioritisation tools available
to perform this task (reviewed in (6)) and a number of commercial software programs that can semi-
automate this proc

to perform this task (reviewed in (6)) and a number of commercial software programs that can semi-
automate this process, for example: Qiagen Clinical Insights (https://digitalinsights.qiagen.com/) and
Congenica (https://w functional genetic variants. In brief, to determine these regions, the authors utilised the Genome Congenica (https://www.congenica.com/).
Even with these prioritisation tools, the task of distinguishing pathogenic from benign variants is
time consuming and novel annotation methods are required to improve this efficien Even with these prioritisation tools, the task
time consuming and novel annotation meth
novel resource was recently published whic
regions (CCRs) in the human genome(8). Th
functional genetic variants. In brief, to dete
Ag time consuming and novel annotation methods are required to improve this efficiency. To this er
novel resource was recently published which is based on creating a map of constrained coding
regions (CCRs) in the human genom the consumed and novel resource was recently published which is based on creating a map of constrained coding
regions (CCRs) in the human genome(8). These are regions of the coding genome devoid of
functional genetic vari regions (CCRs) in the human genome(8). These are regions of the coding genome devoid of
functional genetic variants. In brief, to determine these regions, the authors utilised the Genon
Aggregation Database (gmomAD(9)) re functional genetic variants. In brief, to determine these regions, the authors utilised the Ger
Aggregation Database (gmomAD(9)) resource and invoked the principles of survival bias wh
this context, involved identifying r Aggregation Database (gmomAD(9)) resource and invoked the principles of survival bias which, in
this context, involved identifying regions of coding sequence that were devoid of functional varian
(potentially protein alte Aggregation Database (gmomtal respiration Database that were devoid of functional variant (potentially protein altering) above the average 7bp in size. They next modelled these regions usind metrics such as the expected m (potentially protein altering) above the average 7bp in size. They next modelled these regions using
metrics such as the expected mutation rate based on DNA context and then ranked all the regions
into percentile bins. Th (potentially above the sypected mutation rate based on DNA context and then ranked all the regions
into percentile bins. Their analysis of these bins showed that those above the 95th percentile were
significantly enrich metric percentile bins. Their analysis of these bins showed that those above the 95th percentile were significantly enriched for known pathogenic variants in ClinVar(10) and mutations underlying developmental disorders. into percentile bins. Their analysis of these bins showed that those above the 95 - percentile were
significantly enriched for known pathogenic variants in ClinVar(10) and mutations underlying
developmental disorders. The

developmental disorders. The nature of CCRs is such that because they rely on the appearanc
just a single allele, they are ideally suited for augmenting current variant prioritisation method
when evaluating De novo (Dn) va in is a single allele, they are ideally suited for augmenting current variant prioritisation methods
when evaluating De novo (Dn) variants in studies of autosomal dominant diseases.
Dn variants are a rich source of pathoge guarding, the many suited for any suited for automorphism of their experimental for alleles.

Dn variants are a rich source of pathogenicity in RD patients. To date, the majority of large coho

studies have focused on the Dn variants are a rich source of pathogenicity in RD patients. To date, the majority is
studies have focused on the impact of Dn variants in RD patients with a neurodevel
associated phenotype and have shown unequivocally t Entrin variants are a rich source of pathogenic pathogenic in extending studies have focused on the impact of Dn variants in RD patients with a neurodevelopmental-associated phenotype and have shown unequivocally that they patients(11-17). Similar studies in non-neurodevelopmental-associated RD patients have shown Dr
variants have a substantial impact but at a lower level, for example ~8% in patients with congenita
heart disease(18). To unde

power of the Genomics England (GeL), 100,000 Genomes Project (100KGP)(19) to perform such an pariants have a substantial impact but at a lower level, for example ~8% in patients with congenital
heart disease(18). To understand the impact of pathogenic Dn variants in RD patients from across
the full phenotypic spec heart disease(18). To understand the impact of pathogenic Dn variants in RD patients from across
the full phenotypic spectrum requires a large, systematically ascertained and analysed cohort.
To investigate the utility of the full phenotypic spectrum requires a large, systematically ascertained and analysed cohort.
To investigate the utility of applying CCR information in a real-life RD study, we have leveraged the
power of the Genomics Eng To investigate the utility of applying CCR information in a real-life RD study, we have leveraged
power of the Genomics England (GeL), 100,000 Genomes Project (100KGP)(19) to perform such
evaluation. The 100KGP is a landma power of the Genomics England (GeL), 100,000 Genomes Project (100KGP)(19) to perform such an evaluation. The 100KGP is a landmark genomics project based in the UK that is aligned with the National Health Service (NHS). The power of the 100KGP is a landmark genomics project based in the UK that is aligned with the
evaluation. The 100KGP is a landmark genomics project based in the UK that is aligned with the
National Health Service (NHS). The National Health Service (NHS). The RD component covers the full spectrum of RD phenotypes are
therefore well suited for assessing variant prioritisation tools that will have a general applicabilit
Currently, the diagnostic therefore well suited for assessing variant prioritisation tools that will have a general applicability.
Currently, the diagnostic rate for large-scale genomic RD studies is around 25% but this measure
differs substantiall Currently, the diagnostic rate for large-scale genomic RD studies is around 25% but this measure
differs substantially depending on the patient phenotype, ranging from as low as 3% in patients will
differs substantially de Currently, the angles for the for large-scale genomic RD states is around 25% but the included
differs substantially depending on the patient phenotype, ranging from as low as 3% in patients v
and the patients of the patie differs substantially depending on the patient phenotype, ranging from as low as 3% in patients with

the take home message is clearly that the majority of patients are left in a diagnostic odyssey and
more work is needed to improve the diagnostic rate.
The purpose of this study is two-fold; firstly we want to test if the more work is needed to improve the diagnostic rate.
The purpose of this study is two-fold; firstly we want to test if the addition of CCR data and Dn
status into current variant prioritisation pathways can improve the iden The purpose of this study is two-fold; firstly we want
status into current variant prioritisation pathways car
variants (thereby reducing time-taken-to-diagnosis) a
extra information helps to identify novel pathogenic v
Me status into current variant prioritisation pathways can improve the identification of pathogenic
variants (thereby reducing time-taken-to-diagnosis) and secondly, we want to test whether this
extra information helps to ide status into current current current variants processes) and secondly, we want to test whether this extra information helps to identify novel pathogenic variants.
Extra information helps to identify novel pathogenic variant

extra information helps to identify novel pathogenic variants.
Verthods:
Data access: The Genomics England research environment (RE) was used for all data analyses. W
used the genomic data corresponding to data release v12 Methods:

Data access: The Genomics England research environment (RE

used the genomic data corresponding to data release v12, acc

using RStudio(v1.4.1103)(21) library RLabkey. We used two so

First, we used the *gmc exit* |
| | |
| | | *Data acces*
used the g
using RStu
First, we u
the clinica
to what ex
GMC from Butd access: The Genomics England research environment (RE) was asea for all data analyses. We
used the genomic data corresponding to data release v12, accessed through LabKey and extracted
using RStudio(v1.4.1103)(21) lib using RStudio(v1.4.1103)(21) library RLabkey. We used two sources of patient data for this project.
First, we used the *gmc_exit_questionnaire* table which contains the diagnostic results provided by
the clinical scientist First, we used the *gmc_exit_questionnaire* table which contains the diagnostic results provided by
the clinical scientists from the Genomic Medicine Centres (GMCs) for each patient. This data inform
to what extent a famil First, we used the *gmc_exit_questionnaire* table which contains the diagnosite results provided by
the clinical scientists from the Genomic Medicine Centres (GMCs) for each patient. This data inforr
to what extent a famil to what extent a family's presented case can be explained by the combined variants reported to the GMC from Genomics England and the Clinical Interpretation Providers (CIPs). It also includes information on any segregation GMC from Genomics England and the Clinical Interpretation Providers (CIPs). It also includes
information on any segregation testing performed, the confidence in the identification and
pathogenicity of each variant and, the

information on any segregation testing performed, the confidence in the identification and
pathogenicity of each variant and, the clinical validity of each variant or variant pair in genera
clinical utility in a specific p pathogenicity of each variant and, the clinical validity of each variant or variant pair in gene
clinical utility in a specific patient. One Exit Questionnaire is completed per case.
Secondly, for the Dn variant analysis w pathology in a specific patient. One Exit Questionnaire is completed per case.
Secondly, for the Dn variant analysis we used tables denovo_cohort_information which provided the condition of each participant run through the Secondly, for the Dn variant analysis we used tables denovo_cohort_information
ID information for each participant run through the Dn pipeline and denovo_flag,
gives a list of all the Dn variants called and their confidenc Secondly, for the Data participant run through the Dn pipeline and denovo_flagged_variants which
gives a list of all the Dn variants called and their confidence level. Full information on the GeL Dn
variant research datase

gives a list of all the Dn variants called and their confidence level. Full information on the GeL Dn
variant research dataset is available at
https://cnfl.extge.co.uk/display/GERE/De+novo+variant+research+dataset.
The phe gives a list of all the Danian Burstall the Demand and the Danian Burstall on the DeL D
variant research dataset is available at
https://cnfl.extge.co.uk/display/GERE/De+novo+variant+research+dataset.
The phenotype data as Variant research dataset is arranged that
https://cnfl.extge.co.uk/display/GERE/
The phenotype data associated with th
Ontology (HPO) (25) terms provided by
disease_phenotype table at three succe
description of the general The phenotype data associated with the 100KGP participants is based on the Ontology (HPO) (25) terms provided by the submitting GMC and is made avidisease_phenotype table at three successive levels of detail; *Disease Grou* The phenotype (HPO) (25) terms provided by the submitting GMC and is made available in the
disease_phenotype table at three successive levels of detail; *Disease Group* gives a higher order
description of the general pheno Christopy (HPD) (25) terms presented by the submining electric materials of disease_phenotype table at three successive levels of detail; *Disease Group* gives a highe description of the general phenotype class e.g. Skelet description of the general phenotype class e.g. Skeletal disorders or Cardiovascular disorders,
Disease Sub Group gives a finer scale description of the phenotype e.g. Skeletal dysplasias or
Cardiomyopathy and, Specific Di Disease Sub Group gives a finer scale description of the phenotype e.g. Skeletal dysplasias or
Cardiomyopathy and, *Specific Disease* provides a detailed description of the specific disease e.
Osteogenesis imperfecta or Di

Disease Sub Group gives a finer scale description of the phenotype e.g. Skeletal dysplasias or
Cardiomyopathy and, *Specific Disease* provides a detailed description of the specific disease ϵ
Osteogenesis imperfecta or Cardiomyopathy and, Specific Disease provides a detailed description of the specific disease e.g.

Osteogenesis imperfecta or Dilated Cardiomyopathy.

For the Constrained Coding Region (ref) analysis we downloaded the data For the Constrained Coding Region (ref) analysis we d
Quinlan (https://s3.us-east-2.amazonaws.com/ccrs/co
Data cleaning: This was performed primarily using RSi
included substantial manual manipulation to harmoni
gmc_exit_q Quinlan (https://s3.us-east-2.amazonaws.com/ccrs/ccrs/ccrs.autosomes.v2.20180420.bed.gz).
Data cleaning: This was performed primarily using RStudio with the library tidyverse but also
included substantial manual manipulati Data cleaning: This was performed primarily using RStudio with the library tidyverse but also
included substantial manual manipulation to harmonise the data to a consistent format. For the
gmc_exit_questionnaire, patients Butd cleaning: This was performed primarity using NStudio with the library trayverse but also
included substantial manual manipulation to harmonise the data to a consistent format. For t
gmc_exit_questionnaire, patients we gmc_exit_questionnaire, patients were only retained if there was information for a genomic variand the reference genome build. Patients were classified in the 'case_solved_family' column as:
Solved (referred to as yes), Pa gmc_exit_questionnaire, patients were only retained if there was information for a genomic variant
and the reference genome build. Patients were classified in the 'case_solved_family' column as:
Solved (referred to as yes) Solved (referred to as yes), Partial (referred to as partially), Unknown and Unsolved. In subseque
analyses we combined the Solved and Partial cases into a group termed Diagnosed as each varia
has been determined to be pat Solved (referred to a product of a particle to a proup term of Diagnosed as each variant
has been determined to be pathogenic by a clinical scientist, with the Unknown and Unsolved cases
combined to form the Undiagnosed gr analyses we combined to be pathogenic by a clinical scientist, with the Unknown and Unsolved case
combined to form the Undiagnosed group.
To build our cohort for analysis we split the data from the *gmc_exit_questionnaire*

has been determined to form the Undiagnosed group.
To build our cohort for analysis we split the data from the *gmc_exit_questionnaire* table into
genomic builds GRCh37 and GRCh38 and for both, we removed any duplicates an To build our cohort for analysis we split the
genomic builds GRCh37 and GRCh38 and fo
Discreen the Units SRCh37 and GRCh38 and fo To build our cohort for analysis we split the data from the gmc_exit_questionnaire table into
genomic builds GRCh37 and GRCh38 and for both, we removed any duplicates and those vari
example table wari $g_{\rm{eff}}$ and ϵ builds and for both, we remove $\sigma_{\rm{eff}}$ and those variants are $\sigma_{\rm{eff}}$

It is made available under a [CC-BY-NC-ND 4.0 International license](http://creativecommons.org/licenses/by-nc-nd/4.0/) . **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. medRxiv preprint doi: [https://doi.org/10.1101/2022.08.19.22278944;](https://doi.org/10.1101/2022.08.19.22278944) this version posted August 19, 2022. The copyright holder for this preprint

or Partial, we manually curated the data using the information available to leave only those variants
that were pathogenic. For the Unknown and Unsolved patients, we kept all the variants that had
been returned by the GMC. that were pathogenic. For the Unknown and Unsolved patients, we kept all the variants that had
been returned by the GMC. To unite the two genome builds into one cohort (gmc_ALL38), we used
the UCSC Genome Browser LiftOver

been returned by the GMC. To unite the two genome builds into one cohort (gmc_ALL38), we use
the UCSC Genome Browser LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) and conver
the coordinates for all GRCh37 varia be UCSC Genome Browser LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) and converte
the coordinates for all GRCh37 variants into GRCh38.
To build our Dn cohort we first extracted the denovo_cohort_information tab the coordinates for all GRCh37 variants into GRCh38.
To build our Dn cohort we first extracted the denovo_cohort_information table and only kept
samples labelled as, member = Offspring and affection_status = AFFECTED. For To build our Dn cohort we first extracted the denovo
samples labelled as, member = Offspring and affectio
denovo_flagged_variants table we only kept variants
amended the denovo_flagged_variants table with par
denovo_cohort samples labelled as, member = Offspring and affection_status = AFFECTED. For the
denovo_flagged_variants table we only kept variants with a Stringent Filter score =1. We then
amended the denovo_flagged_variants table with samples labelled as a straight stable we only kept variants with a Stringent Filter score =
amended the denovo_flagged_variants table with participant_id information from t
denovo_cohort table using the Trio Id as a scaffo amended the denovo_flagged_variants table with participant_id information from the
denovo_cohort table using the Trio Id as a scaffold and then split these samples by genome bu
GRCh37 and GRCh38. Using the UCSC Genome Brow denovo_cohort table using the Trio Id as a scaffold and then split these samples by gen
GRCh37 and GRCh38. Using the UCSC Genome Browser LiftOver tool (http://genome.u
bin/hgLiftOver) we converted the GRCh37 variants into denotes the novel table using the UCSC Genome Browser LiftOver tool (http://genome.ucsc.edu/cg
bin/hgLiftOver) we converted the GRCh37 variants into GRCh38 and combined the variants
(Dn_ALL38).
As only a subset of the gmc_

(Dn_ALL38).
As only a subset of the gmc_ALL38 were included in the Dn analysis, we used the participan
information from DN_ALL38 to derive a final cohort that contained diagnostic information f
those patients present in bo (DD₂ (DD₂ (DD₃).
As only a subtinformation
those patien
To identify w
and extracte
For the CCR i

To identify which of the variants from the gmc_Dn_ALL cohort were Dn we used the Dn_ALL38 data
and extracted those variants that matched for participant id and genomic position (Dn gmc).

information from DN_ALL38 to derive a final cohort that contained diagnostic information from
those patients present in both cohorts, this is the cohort we used for our analysis (gmc_Dn_ALL)
To identify which of the varia those patients present in both cohorts, this is the cohort we used for our analysis (gmc_Dn_ALL).
To identify which of the variants from the gmc_Dn_ALL cohort were Dn we used the Dn_ALL38 data
and extracted those variants To identify which of the variants from the gmc_Dn_ALL cohort were Dn we used the Dn_ALL38 dand extracted those variants that matched for participant_id and genomic position (Dn_gmc).
For the CCR intersection analysis we fi and extracted those variants that matched for participant_id and genomic position (Dn_gmc).
For the CCR intersection analysis we first removed all the VARTRUE variants from the bed file as
these correspond to known variant For the CCR intersection analysis we first removed all the VARTRUE variants from the bed file a
these correspond to known variants in the gnomAD database(8). Because the CCR region
coordinates are in genome build GRCh37 we For these correspond to known variants in the gnomAD database(8). Because the CCR region
coordinates are in genome build GRCh37 we used the UCSC Genome Browser LiftOver tool
(http://genome.ucsc.edu/cgi-bin/hgLiftOver) to c coordinates are in genome build GRCh37 we used the UCSC Genome Browser LiftOver too (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the coordinates to GRCh38. We the bedtools-intersect (22) to extract those variant (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the coordinates to GRCh38. We the
bedtools-intersect (22) to extract those variants that resided within a CCR and included the
percentile score. Using the percentile s bedtools-intersect (22) to extract those variants that resided within a CCR and included their
percentile score. Using the percentile score we were able to filter our variants to only those that
intersected a CCR with a p

percentile score. Using the percentile score we were able to filter our variants to only those t
intersected a CCR with a percentile score of \geq 95 (CCR95).
Combining the results of the Dn_gmc and CCR95 intersection we percentile score. The percentile score of 295 (CCR95).

Combining the results of the Dn_gmc and CCR95 intersection we were able to derive a list of Dn

variants from the gmc_ALL38 cohort that intersected a CCR95 (Dn_CCR95) interfected a Combining the results of the Dn_gmc and CCR95 intersected variants from the gmc_ALL38 cohort that intersected a CC
Disease terms: We used the patient_phenotype table in Linformation for the gmc_ALL38 cohort. variants from the gmc_ALL38 cohort that intersected a CCR95 (Dn_CCR95).
Disease terms: We used the patient_phenotype table in LabKey to extract the phenotype
information for the gmc_ALL38 cohort. The disease terms are pro Disease terms: We used the patient _phenotype table in LabKey to extract that information for the gmc_ALL38 cohort. The disease terms are provided at the and becoming more focused (Disease Group, Disease Sub Group and Spec *Disease terms:* We used the patient-phenotype table in LabKey to extract the phenotype information for the gmc_ALL38 cohort. The disease terms are provided at three levels sta and becoming more focused (*Disease Group, Di* information for the gmc_nation for the gmc_nation and becoming more focused (*Disease Group, Disease Sub Group and Specific Disease*). As the phenotype assigned to each patient becomes more specific the overall number of t and becoming more focused (Disease Group, Disease 3ab Group and Specific Disease). As the
phenotype assigned to each patient becomes more specific the overall number of terms incre
greatly and consequently the number of pa

phenotype assigned to the number of patients with each term decreases. We therefore set a cut
greatly and consequently the number of patients with each term decreases. We therefore set a cut
off \geq 5 patients per term an they could be used as an additional filter to identify pathogenic variants more specifically and thus off ≥5 patients per term and chose to focus of the highest descriptive level, *bisease Group* to ensure
we had sufficient numbers in each group to make meaningful comparisons.
Analysis strategy: We focused on assessing the Analysis strategy: We focused on assessing the proportion of variants prese
annotation categories; 1) Dn variant, 2) intersecting a CCR95 and, 3) a Dn va
CCR95. The aims of this were twofold: Firstly we wanted to explore w Analysis strategy: We rocused on assessing the proportion of variants present in the following three
annotation categories; 1) Dn variant, 2) intersecting a CCR95 and, 3) a Dn variant intersecting a
CCR95. The aims of this CCR95. The aims of this were twofold: Firstly we wanted to explore whether there was an enrichment of pathogenic variants in one of the three annotation categories to examine wheth
they could be used as an additional filte enrichment of pathogenic variants in one of the three annotation categories to examine w
they could be used as an additional filter to identify pathogenic variants more specifically a
potentially allow a diagnosis to be re enrichment of pathogenic variants in they could be used as an additional filter to identify pathogenic variants more specifically and thus
potentially allow a diagnosis to be reached quicker. Secondly, we wanted to explore potentially allow a diagnosis to be reached quicker. Secondly, we wanted to explore if, in the
undiagnosed samples, the application of both the Dn and CCR95 annotations would filter the
number of variants down to a small e potentially allow a diagnosed samples, the application of both the Dn and CCR95 annotations would filter the number of variants down to a small enough number to highlight novel disease variants. number of variants down to a small enough number to highlight novel disease variants.

The Den and COR95 and CCR95 and CCR number of variants down to a small enough number to a small enough number of variants.
The following number of variants of variants of variants of variants. The following state is a small enough of
The following state is a

across the clinical spectrum or whether certain disease phenotypes were more applicable to this
approach by splitting our cohort into separate disease classes and looking for enrichment of signal.
Results:
Sample cohort: F

approach by splitting our cohort into separate disease classes and looking for enrichment of signa
Results:
Sample cohort: For our analyses we used the 100KGP v12 dataset which is comprised of 73,880 R
genomes from 71,597 approach by splitting our conditions of partic and to the constraint of the signal.
Results:
Sample cohort: For our analyses we used the 100KGP v12 dataset which is comprised of 73,880 RD
genomes from 71,597 participants. Sample
genome
been pro
segregat
classed a
families. Sample conort: For our analyses we used the 100KGP v12 dataset which is comprised of 73,000 KD
genomes from 71,597 participants. Following mapping and variant calling(20), 33,315 families have
been processed through an aut been processed through an automated analytic pipeline to filter down the variants to rare,
segregating and predicted damaging candidate variants in coding regions. These variants have been
classed as tier 1, 2 or 3 (see (2 been processed and predicted damaging candidate variants in coding regions. These variants has
classed as tier 1, 2 or 3 (see (20) for full description) and used to make a clinical diagnosis in
families.
Although Dn status

segregating and predicted through generation and used to make a clinical diagnosis in 30,419 families.

Although Dn status can be predicted through simple analysis of the trio data it does not provide a

robust output and families.
Although Dn status can be predicted through simple analysis of the trio data it does not provide a
robust output and therefore, for our Dn analyses we analysed 13,353 trios using a bespoke Dn
pipeline that utili

Although
robust ou
pipeline t
The coho
a clinical
because Although Data diagnosis and herefore, for our Dn analyses we analysed 13,353 trios using a bespoke Dn
pipeline that utilised the raw sequence data for each member of the trio (see methods).
The cohort we used for this anal pipeline that utilised the raw sequence data for each member of the trio (see methods).
The cohort we used for this analysis was derived from the overlapping patients that had under
a clinical diagnosis and had been run th pipeline that used for this analysis was derived from the overlapping patients that had
a clinical diagnosis and had been run through the bespoke Dn analysis pipeline (n = 3,631
because the CCR data has been generated for a clinical diagnosis and had been run through the bespoke Dn analysis pipeline (n = 3,631). However
because the CCR data has been generated for the autosomes only, we removed all samples with a
variant assigned to the X-Ch

because the CCR data has been generated for the autosomes only, we removed all samples with a
variant assigned to the X-Chromosome or mitochondrial genome which resulted in a final cohort of
3,090 families.
Diagnostic rate because the Chromosome or mitochondrial genome which resulted in a final cohort of
3,090 families.
Diagnostic rate calculation: Following data cleaning we calculated the diagnostic rate for the
gm_Dn_ALL cohort as a whole 3,090 families.

Diagnostic rate calculation: Following data cleaning we calculated the diagnostic rate for the

gm_Dn_ALL cohort as a whole was 71% (Table 1 and Supplementary Table 1). When we stratify the

data for CCR95 *S*
Siagnostic rate
gm_Dn_ALL coldata for CCR95
with a CCR95 ir
If we instead lo
see that the dia

Brughostic rate calculation: Following data cleaning we calculated the diagnostic rate for the
gm_Dn_ALL cohort as a whole was 71% (Table 1 and Supplementary Table 1). When we strat
data for CCR95/Dn annotation, we show th gman and a for CCR95/Dn annotation, we show the causative variant is Dn in 40% of patients, intersects
with a CCR95 in 18% of the patients and is a Dn intersecting a CCR95 in 12% of patients (Table 1).
If we instead look Material CR95 in 18% of the patients and is a Dn intersecting a CCR95 in 12% of patients (Table 1).
If we instead look at each annotation class and calculate the diagnostic rate within each group, w
see that the diagnostic If we instead look at each annotation class and calculate the diagnostic rate within each group, w
see that the diagnostic rate increases to 84% for Dn variants, 81% for CCR95 variants and 87% for
CCR95 intersecting varia see that the diagnostic rate increases to 84% for Dn variants, 81% for CCR95 variants and 87% for C
CCR95 intersecting variants (Table 1). That is, for example, if we extract all the Dn CCR95 intersecting
variants (n=293)

- F / ()
- F / ()
- F / within the Neurology and neurodevelopmental disorders domain showed a highly significant
enrichment of pathogenic variants in all annotation classes (Table 2). This ranged from 53% in t
whole cohort to 69% in the Dn cohor which ment of pathogenic variants in all annotation classes (Table 2). This ranged from 53% in
whole cohort to 69% in the Dn cohort (Enrichment p_adj = 2.24x10⁻¹⁰), 76% in CCR95 cohort
(Enrichment p_adj = 7.47x10⁻¹⁰)

enrichment of pathogenic channel of anticomment paral $\frac{1}{2}$ (table 2). 76% in CCR95 cohort
(Enrichment p_aadj = 7.47x10⁻¹⁰) and 80% in the Dn_nCCR95 cohort (Enrichment p_aadj = 9.48x10⁻⁰⁹).
In comparison, those whole cohort to 69% in the Dn cohort (Enrichment p_adj = 2.24x10⁻⁻⁻), 76% in CCR95 cohort (Enrichment p_adj = 9.48x:
(Enrichment p_adj = 7.47x10⁻¹⁰) and 80% in the Dn_CCR95 cohort (Enrichment p_adj = 9.48x:
In compari (Enrichment p_adj = 7.47x10⁻⁻) and 80% in the Dn_CCR95 cohort (Enrichment p_adj = 9.48x10⁻⁻).
In comparison, those patients with a phenotype in the Ophthalmological disorders domain (the ne:
largest *Disease Group*), Iargest Disease Group), showed a highly significant negative enrichment in all annotation classes
(Table 2). This group comprised 10% of the whole cohort, 3% of the Dn cohort (Enrichment p_adj =
Table 2). This group compri (Table 2). This group comprised 10% of the whole cohort, 3% of the Dn cohort (Enrichment p_adj)
(Table 2). This group comprised 10% of the whole cohort, 3% of the Dn cohort (Enrichment p_adj) Γ , this group comprised 10% of the whole cohort, 3% of the whole cohort, Γ adj Γ

It is made available under a [CC-BY-NC-ND 4.0 International license](http://creativecommons.org/licenses/by-nc-nd/4.0/) . medRxiv preprint doi: [https://doi.org/10.1101/2022.08.19.22278944;](https://doi.org/10.1101/2022.08.19.22278944) this version posted August 19, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

【 ()
()
()
() Our study
295 (CCR95
enriches fo
when takin
metrics, thi
that incorp
the time ta \geq 95 (CCR95) in the variant prioritisation cascade used to clinically diagnose rare disease patients
enriches for pathogenic variants (from 71% to 81%). This effect is similar to the enrichment seen
when taking Dn stat 2022 (CCR) in the variants (from 71% to 81%). This effect is similar to the enrichment seen

when taking Dn status into account (from 71% to 84%) and is further increased when combing bo

metrics, that is, Dn variants resi enriches for pathogens that in account (from 71% to 84%) and is further increased when combing bo
metrics, that is, Dn variants residing within a CCR95 (from 71% to 87%) (Table 1). We hypothesise
that incorporating this me metrics, that is, Dn variants residing within a CCR95 (from 71% to 87%) (Table 1). We hypothesise
that incorporating this metric into the genomic clinical diagnostic filtering cascade would decrease
the time taken to ident

that incorporating this metric into the genomic clinical diagnostic filtering cascade would decrease
the time taken to identify pathogenic variants and therefore allow more patients to be screened
over a set period of tim the time taken to identify pathogenic variants and therefore allow more patients to be screened
over a set period of time, thus resulting in an increase in the diagnostic rate.
A major advantage of the CCR method is that over a set period of time, thus resulting in an increase in the diagnostic rate.
A major advantage of the CCR method is that it offers a focused annotation metric for discrete
regions of a gene instead of methods such as A major advantage of the CCR method is that it offers a focused annotation n
regions of a gene instead of methods such as pLI (probability of being loss-of-
or Z scores(24) that are provided in gnomAD(9), which annotate th A major advantage of methods such as pLI (probability of being loss-of-function intolerant or Z scores(24) that are provided in gnomAD(9), which annotate the whole gene. The reason the important is because, even in highly or Z scores(24) that are provided in gnomAD(9), which annotate the whole gene. The reason this is
important is because, even in highly constrained genes (pLI \geq 0.9, positive Z score), there are often
discrete regions or Dimportant is because, even in highly constrained genes (pLI \geq 0.9, positive Z score), there are often discrete regions that show low constraint with many functional variants seen in gnomAD(9).
Therefore, if a poten important is because, even in the many functional variants seen in gnomAD(9).

Therefore, if a potentially pathogenic variant is located in one of these low constrained regions the

variant could be misinterpreted as path Therefore, if a potentially pathogenic variant is located in one of these low constrained regio
variant could be misinterpreted as pathogenic based on the gene-wide annotation. Converse
there are instances where genes sho variant could be misinterpreted as pathogenic based on the gene-wide annotation. Conversely,
there are instances where genes showing low constraint (pLI \leq 0.5, negative Z score) contain small,
highly constrained regio variant could be missing the missing low constraint (pLI \leq 0.5, negative Z score) contain sm
highly constrained regions that are devoid of functional variation in gnomAD(9) which can lead t
pathogenic variant, from wi

highly constrained regions that are devoid of functional variation in gnomAD(9) which can lead to a
pathogenic variant, from within this region, being missed if the gene-wide annotation is used (see(8
Figure 1).
Furthermor highly constant, from within this region, being missed if the gene-wide annotation is used (see (8
Figure 1).
Furthermore, because the number of variants fulfilling the Dn_CCR95 criteria is low, it is feasible to
use this pathogenic variant, from within the region, being missed in the generation and the level (1944)
Figure 1).
Furthermore, because the number of variants fulfilling the Dn_CCR95 criteria is low, it is feasible to
use this met Furthermouse

meater is m

patients c

variants in

(Supplemourrepath

meuropath

meuropath Furthermore, we dentify the number of an example, we identified 39 variants in
patients classed as undiagnosed that were in the Dn_CCR95 category (Table 1). Inspection of these
variants in a research setting led to a likel patients classed as undiagnosed that were in the Dn_CCR95 category (Table 1). Inspection of tl
variants in a research setting led to a likely diagnosis for 25 of the undiagnosed patients (64%)
(Supplementary Table 4). For parameters in a research setting led to a likely diagnosis for 25 of the undiagnosed patients (64%)
(Supplementary Table 4). For example, in an undiagnosed patient with HPO(25) terms including
microcephaly, intellectual di (Supplementary Table 4). For example, in an undiagnosed patient with HPO(25) terms including
microcephaly, intellectual disability, small for gestational age, short stature, motor axonal
neuropathy and congenital microcep (Supprementary) in the Herbertary Manuscule present with the corresponding microcephaly, intellectual disability, small for gestational age, short stature, motor axonal neuropathy and congenital microcephaly; we identified mortectual disability, meaning of the disability, we identified a missense variant in the gene MC
(p.Y448C) that was predicted to be deleterious by SIFT(26), probably damaging by PolyPhe
and had a CADD(28) Phred score of 3 neuropathy and congenital microcephaly, we identified a missense variant in the gene MORC2
(p.Y448C) that was predicted to be deleterious by SIFT(26), probably damaging by PolyPhen-2(2)
and had a CADD(28) Phred score of 30 (probably finite to probably here some of 30. Patients with *MORC2* pathogenic variants present with a phenotype consisting of developmental delay, intellectual disability, growth retardation, microcephaly, variable cranio and had a CADD(20) Fired score of 30. Patients with Monez pathogenic variants present with a
phenotype consisting of developmental delay, intellectual disability, growth retardation,
microcephaly, variable craniofacial dys phicrocephaly, variable craniofacial dysmorphism, and in some individuals electrophysiology
abnormalities suggestive of neuropathy (29). The phenotypic overlap along with the dama
of this variant make this a strong candida microcephaly, variable cranication by interpretism, and in some individuals electrophysics gradion abnormalities suggestive of neuropathy (29). The phenotypic overlap along with the damaging of this variant make this a str abilities suggestive of neuropathy(29). The phenotypic overlap along with the damaging means
of this variant make this a strong candidate to be the pathogenic variant for this patient. of this variant make this a strong candidate to be the pathogenic variant for this patient.

urinary tract disorders Disease Group that had a missense variant in the gene *SETD5* (p.R348L) v
was predicted to be deleterious by SIFT(26), probably damaging by PolyPhen-2(27) and had a
CADD(28) Phred score of 32. The H urinary tract disorders Disease Group that had a missense variant in the gene SETD5 (p.R348E) which
was predicted to be deleterious by SIFT(26), probably damaging by PolyPhen-2(27) and had a
CADD(28) Phred score of 32. The CADD(28) Phred score of 32. The HPO(25) terms for this patient were abnormality of the eye,
hypodontia, pectus excavatum, global developmental delay, abnormal facial shape and mild s
stature, none of which fit with the Di CADD Mypodontia, pectus excavatum, global developmental delay, abnormal facial shape and mild s
stature, none of which fit with the Disease Group they were assigned to. Patients with *SETD5*
pathogenic variants display var my personally perfect and the Unit of Scanner Properties (Stature, none of which fit with the Disease Group they were assigned to. Patients with *SETD5* pathogenic variants display variable features including intellectual

stature, none of which it with the Biscase Group they were assigned to. Patients with SETD5
pathogenic variants display variable features including intellectual disability, facial dysmorphis
cardiac and skeletal abnormalit pathogenic variable featy variable features including intellectual distable, yielding yielding cardiac and skeletal abnormalities, behavioural problems and short stature(30). We therefore believe the SETD5 variant in this believe the SETD5 variant in this patient is likely to be pathogenic.
Although we cannot be sure why this patient was included in the Renal and urinary tract disore
Disease Group, this example highlights the power of combi Although we cannot be sure why this patient was included in the R
Disease Group, this example highlights the power of combining Dn
disease/gene agnostic approach to identify highly likely pathogenic
therefore help in overc Disease Group, this example highlights the power of combining Dn and CCR95 data as it allows a
disease/gene agnostic approach to identify highly likely pathogenic variants. This approach will
therefore help in overcoming p Disease/gene agnostic approach to identify highly likely pathogenic variants. This approach will
therefore help in overcoming problems of misdiagnoses or when patients have multi-system
abnormalities and do not fit a singl therefore help in overcoming problems of misdiagnoses or when patients have multi-system
abnormalities and do not fit a single phenotype grouping. Also, because applying this filter resu
very few variants qualifying for in abnormalities and do not fit a single phenotype grouping. Also, because applying this filter re
very few variants qualifying for inspection, it's use as a first-pass stringent filter in clinical diag
laboratories has the p

above the variants qualifying for inspection, it's use as a first-pass stringent filter in clinical diagnostic
laboratories has the potential to rapidly identify pathogenic variants from patient cohorts, freeing up
time to very filter than one patients overall.
To explore our data further we next decided to look at recurrent Dn sites for instances where a Dn
had been called pathogenic in one patient but not in another or where a recurrent Dn time to screen more patients overall.
To explore our data further we next decided to look at recurrent Dn sites for instances where a Dn
had been called pathogenic in one patient but not in another or where a recurrent Dn To explore our data further we next d
had been called pathogenic in one pat
greater than one undiagnosed patient
could help diagnose patients where, f
variant pathogenic. We identified 44 I
were identified in only diagnose To explore a recurrent Dn was seen in
had been called pathogenic in one patient but not in another or where a recurrent Dn was seen in
greater than one undiagnosed patient (Supplementary Table 5). We reasoned that this app greater than one undiagnosed patient (Supplementary Table 5). We reasoned that this approach
could help diagnose patients where, for whatever reason, there was not enough evidence to call a
variant pathogenic. We identifie greater than one was not enough evidence to call
variant pathogenic. We identified 44 Dn variants that were recurrent in 100 individuals. Of these,
were identified in only diagnosed patients, six were a mixture where at le

variant pathogenic. We identified 44 Dn variants that were recurrent in 100 individuals. Of these, 3
were identified in only diagnosed patients, six were a mixture where at least one patient was
diagnosed and the other was were identified in only diagnosed patients, six were a mixture where at least one patient was
diagnosed and the other was undiagnosed, and in two instances, both patients were undiagnosed.
For the latter group, our analysi Were in this and the other was undiagnosed, and in two instances, both patients were undiagnosed and the other was undiagnosed, and in two instances, both patients were undiagnosed For the latter group, our analysis highli For the latter group, our analysis highlighted a Dn canonical splice site variant (c.2493+1 G>A) in the
Floating-Harbour syndrome gene(31) *SRCAP* (NM_006662.3) in two undiagnosed patients with
intellectual disability. The Floating-Harbour syndrome gene(31) *SRCAP* (NM_006662.3) in two undiagnosed patients with
intellectual disability. The evidence available (CADD Phred score 36) makes this a highly likely
pathogenic variant in these patien Floating-Harbour syndrome gene(31) SRCAP (NML_000002.3) in two ditalaghosed patients with
intellectual disability. The evidence available (CADD Phred score 36) makes this a highly likely
pathogenic variant in these patient interactual disability. These patients. At the second loci, we identified two undiagnosed patients a Dn missense variant in the gene *FBXO11* in which Dn variants are known to cause an intellectual disorder with dysmorphic pathogenic variant in the gene FBXO11 in which Dn variants are known to cause an intellectual
developmental disorder with dysmorphic facies and behavioural abnormalities (IDDFBA)(32). The Dn
variant we identified at codon a Dn missense variant in the gene FBXO11 in which Dn variants are known to cause an intellectual
developmental disorder with dysmorphic facies and behavioural abnormalities (IDDFBA)(32). The I
variant we identified at codo and the detection of the Dn variant weigher in a disorganised protein domain. However, the predicted detection and the CB percentile and the CB percent in group of the Dn variant (ClinVar ID:559601) at this codon which cau acid which is not present in gnomAD, is predicted by SIFT(26) to be tolerated, benign by PolyPhen-
2(27) and has a CADD(28) Phred score of 20.1. In ClinVar there is a known Pathogenic/Likely
pathogenic missense variant (C 68, meaning we would not be confident in calling this variant Pathogenic/Likely pathogenic without
additional evidence. 2(27) and has a capacity of the DISS9601) at this codon which causes an amino acid clom Arginine to Serine which is also not present in gnomAD. Arginine is a basic amino acid atherefore a change to Glycine (small) or Serin pathom Arginine to Serine which is also not present in gnomAD. Arginine is a basic amino acid and therefore a change to Glycine (small) or Serine (nucleophilic) could both potentially have functional effects, especially as therefore a change to Glycine (small) or Serine (nucleophilic) could both potentially have function
effects, especially as they occur within a disorganised protein domain. However, the predicted
deleteriousness of the Dn v therefiects, especially as they occur within a disorganised protein domain. However, the predicted
deleteriousness of the Dn variant we identified is weak and the CCR percentile score for this region i
68, meaning we would

deleteriousness of the Dn variant we identified is weak and the CCR percentile score for this re
68, meaning we would not be confident in calling this variant Pathogenic/Likely pathogenic wit
additional evidence.
Although deleteration of the DN variant in the DN variant Path was the DN protection in the DN program.
The DN variant Pathogenic/Likely pathogenic without additional evidence.
Although our study shows the benefit of incorporating For the state of the contract of the benefit of incorporating CCR percentile score information and Dn
status in the filtering cascade for identifying pathogenic rare disease variants we need to be aware
of the constraints Although our study s
status in the filtering
of the constraints rel
diagnostic rate was f:
100KGP pilot project
gmc_exit_questionna Although the filtering cascade for identifying pathogenic rare disease variants we need to be awa
of the constraints relating to the cohort we have used. Firstly, in the cohort we have studied the
diagnostic rate was far h status in the filtering status in the filtering pathogen of the constraints relating to the cohort we have used. Firstly, in the cohort we have studied the diagnostic rate was far higher (71%) than that seen in other rare diagnostic rate was far higher (71%) than that seen in other rare disease cohorts, including the
100KGP pilot project (25%)(20). This may be due to many reasons; first, the data in the 100KGP
gmc_exit_questionnaire table c diagnostic rate to many reasons; first, the data in the 100KGP of the metallical project (25%)(20). This may be due to many reasons; first, the data in the 100KGP gmc_exit_questionnaire table contains entries for 30,419 fa 100 Ferry project (25%), the may be due to many reasons, the data in the 100KGP
gmc_exit_questionnaire table contains entries for 30,419 families, however, many of these ent gmc_exit_questionnaire table contains entries for α , however, many of these entries entries entries entries

once we had filtered the data to include only patients that were also included in the Dn cohort and
filtered out any variants on the X chromosome and mitochondrial genome we had a cohort of 3,090
patients for our analysis. filtered out any variants on the X chromosome and mitochondrial genome we had a cohort of 3,090
patients for our analysis. Because the variants contained in the gmc_exit_questionnaire are returne
by the gmc centres that re patients for our analysis. Because the variants contained in the gmc_exit_questionnaire are returned
by the gmc centres that recruited the patients following diagnostic evaluation, we can assume that
no Tier 1 or Tier 2 va patients for our analysis. Because the patients following diagnostic evaluation, we can assume that
no Tier 1 or Tier 2 variants were present in the undiagnosed patients that would explain their
phenotype. The Tiering of t by the gmc centre that were present in the undiagnosed patients that would explain their
no Tier 1 or Tier 2 variants were present in the undiagnosed patients that would explain their
phenotype. The Tiering of the data is

phenotype. The Tiering of the data is performed by GeL commercial partners and so without
reanalysing the whole dataset from the raw data we are unable to estimate how many Dn or (
variants may be present in these un-annot phenotype. The the tering of the datable performed by 122 commercial partners and some presents in the ready sing the whole dataset from the raw data we are unable to estimate how many Dn or variants may be present in thes variants may be present in these un-annotated patients.
Secondly, the patients recruited to the 100KGP were partly composed of research cohorts and
patients who previously had negative genetic tests and therefore may repre Secondly, the patients recruited to the 100KGP were part
patients who previously had negative genetic tests and th
patients not fully representative of the diverse patients s
truly estimate the clinical utility of applying patients who previously had negative genetic tests and therefore may represent a distinct cohordients not fully representative of the diverse patients seen in NHS clinical diagnostic centres.

truly estimate the clinical u patients not fully representative of the diverse patients seen in NHS clinical diagnostic centres. To
truly estimate the clinical utility of applying Dn and CCR95 status we would need to incorporate
these annotations to a particular extendions to a standard NHS diagnostic laboratory patient cohort and perform a direct these annotations to a standard NHS diagnostic laboratory patient cohort and perform a direct comparison of the diagnostic r fit these annotations to a standard NHS diagnostic laboratory patient cohort and perform a direct
comparison of the diagnostic rate with these annotations and without them. This would also allo
us to estimate if the applic these annotations and without them. This would also all
us to estimate if the application of these annotations decreases the time taken to reach a diagn
and if so how much it improves the overall diagnostic rate. If this w us to estimate if the application of these annotations decreases the time taken to reach a diagnosis
and if so how much it improves the overall diagnostic rate. If this was to be done, it would be
imperative we include CCR

and if so how much it improves the overall diagnostic rate. If this was to be done, it would be
imperative we include CCR annotations for the X chromosome as we could potentially be missing ou
on a number of diagnoses that imperative we include CCR annotations for the X chromosome as we could potentially be miss
on a number of diagnoses that reside on that chromosome.
Finally, the inclusion of *Disease Group* data showed there was a large b on a number of diagnoses that reside on that chromosome.
Finally, the inclusion of *Disease Group* data showed there was a large bias towards patients with a
neurology and neurodevelopmental disorders phenotype (>50% of t Finally, the inclusion of *Disease Group* data showed there wandworldly, the inclusion of *Disease Group* data showed there wandworldlend on the position of pathogenicity for patients with s the highly significant enrichm Finally, the inclusion of Disease Group data showed there was a large bias towards pattents with a
neurology and neurodevelopmental disorders phenotype (>50% of the cohort) which may have
biased the results somewhat (Tabl biased the results somewhat (Table 2). This is especially pertinent for the Dn analysis as Dn variation are a well-known source of pathogenicity for patients with such phenotypes (11-17), as shown be highly significant en are a well-known source of pathogenicity for patients with such phenotypes (11-17), as shown by
the highly significant enrichment we saw of Dn variants in this group (p_adj = 2.24x10⁻¹⁰). It should
be noted however, tha

the highly significant enrichment we saw of Dn variants in this group $(p_adj = 2.24x10^{10})$. It should
be noted however, that we also observed a highly significant enrichment of variants from this grou
that intersected a CCR the highly significant enrichment we saw of Dn variants in this group (p_adj = 2.24x10 -). It should
be noted however, that we also observed a highly significant enrichment of variants from this grou
that intersected a CC that intersected a CCR95 region (p_adj = 7.47x10⁻¹⁰), 28% of which were not Dn variants.
This observation is in stark contrast to the pattern of enrichment we saw for patients within the
ophthalmological disorders *Dise* that intersected a CCR95 region (p_adj = 7.47x10⁻⁻), 28% of which were not Dn variants.
This observation is in stark contrast to the pattern of enrichment we saw for patients with
ophthalmological disorders *Disease Grou* ophthalmological disorders *Disease Group*. For these patients we saw a highly significant
underrepresentation of Dn, CCR95 and Dn_CCR95 variants (Table 2). For patients within the skeldisorders *Disease Group*, we observe ophthalmological disorders Disease Group. For these patients we saw a highly significant
underrepresentation of Dn, CCR95 and Dn_CCR95 variants (Table 2). For patients within t
disorders Disease Group, we observed an under disorders *Disease Group,* we observed an underrepresentation of variants intersecting CCR95
regions which was also observed for Dn_CCR95 variants but not for Dn variants alone (Table 2). For
the remaining *Disease Group*

disorders Disease Group, we observed an underrepresentation of variants intersecting echosting
regions which was also observed for Dn_CCR95 variants but not for Dn variants alone (Table 2
the remaining *Disease Group* phen regions where maining *Disease Group* phenotypes, the numbers were much smaller which reduced our
power to identify any significant enrichments.
Why we see such an enrichment for Dn and CCR95 variants in the neurology and
 the remaining *Disease Group* phenotypes, the numbers were much smaller which reduced our
power to identify any significant enrichments.
Why we see such an enrichment for Dn and CCR95 variants in the neurology and
neurodev Why we see such an enrichment for Dn and CC
neurodevelopmental disorders *Disease Group*
scope of this study. Nonetheless, it does sugge
neurodevelopmental disorders phenotype, the
sequencing approach and that this should we interpretent and is interest of the suggest that for patients is open to speculation but
scope of this study. Nonetheless, it does suggest that for patients with a neurolog
neurodevelopmental disorders phenotype, the gr neurodevelopmental disorders Disease Group patients is open to speculation but is beyond the
scope of this study. Nonetheless, it does suggest that for patients with a neurology and
neurodevelopmental disorders phenotype,

scope of the greatest diagnostic return will come from sequencing approach and that this should be the first-tier test used by clinical diagnostic laboratories.
Iaboratories.
Where a Dn analysis is not possible, due to una sequencing approach and that this should be the first-tier test used by clinical diagnostic
laboratories.
Where a Dn analysis is not possible, due to unavailability of one or both parents or because of c
constraints, it is sequencies.
September a Dn analysis is not possible, due to unavailability of one or both parents or becau
constraints, it is noteworthy that the use of CCR95 status alone will enable the identificat
highly likely pathogen More a Dn a

Sonstraints, it

inghly likely p

(n=485), 81%

be particularl Where a Constraints, it is noteworthy that the use of CCR95 status alone will enable the identification of
highly likely pathogenic variants. Our analysis shows that, of all the variants that intersected a CCRS
(n=485), 81 highly likely pathogenic variants. Our analysis shows that, of all the variants that intersected a C
(n=485), 81% (n=395) were clinically diagnosed pathogenic variants (Table 1). This observation of
be particularly relevan (n=485), 81% (n=395) were clinically diagnosed pathogenic variants (Table 1). This observation could
be particularly relevant to clinical diagnostic laboratories in low/middle economic countries where
particularly relevan (n) be particularly relevant to clinical diagnostic laboratories in low/middle economic countries where
the particularly relevant to clinical diagnostic laboratories in low/middle economic countries where
the particularly be particularly relevant to clinical diagnostic laboratories in low/middle economic countries where

with a neurology and neurodevelopmental disorders phenotype.
In summary, we have constructed a cohort of rare disease patients (n=3,090) that have undergone
WGS as part of the GeL 100KGP, have been clinically assessed by a In summary, we have constructed a cohort of rare disease patient
WGS as part of the GeL 100KGP, have been clinically assessed by a
have undergone a bespoke Dn variant calling pipeline. We have s
account a variant's Dn stat IN SIGN as part of the GeL 100KGP, have been clinically assessed by a diagnostic laboratory and who
have undergone a bespoke Dn variant calling pipeline. We have sought to determine if taking into
account a variant's Dn st have undergone a bespoke Dn variant calling pipeline. We have sought to determine if taking into
account a variant's Dn status, whether it intersects a CCR95 or both (Dn_CCR95) could improve ou
ability to identify clinical have undergone a besponse intersects a CCR95 or both (Dn_CCR95) could improve ou
ability to identify clinically pathogenic variants. We show how the rate of diagnosis in our whole
cohort (71%) increases when we look at jus ability to identify clinically pathogenic variants. We show how the rate of diagnosis in our whole
cohort (71%) increases when we look at just those variants that are classed as intersecting a CCR95
(81%), Dn variants (84% cohort (71%) increases when we look at just those variants that are classed as intersecting a CCR
(81%), Dn variants (84%) or a combination of both (87%) (Table 1). This observation seems to be
driven primarily by patients contributed as a first (84%) or a combination of both (87%) (Table 1). This observation seems to be driven primarily by patients with a neurology and neurodevelopmental disorders phenotype. We therefore suggest that, where (81%), Dividibility (81%), Dividibility or a combined view (81%), (Table 2). Wherefore primarily by patients with a neurology and neurodevelopmental disorders phenotype. We therefore suggest that, where possible patients w dependence that, where possible patients with such a phenotype should receive WGS trio
sequencing as a first tier test but where parent availability or cost make this unfeasible, the
identification of variants intersecting therefore suggest that, where persons patients with such a phenotype such a sequencing as a first tier test but where parent availability or cost make this unfeasible, the identification of variants intersecting a CCR95 is

identification of variants intersecting a CCR95 is an alternative route to aid in the detection (
likely clinically pathogenic variants.
Further work should aim to apply this approach in systematic way in a standard clinic likely clinically pathogenic variants.
Further work should aim to apply this approach in systematic way in a standard clinical diagnostic
laboratory to assess its utility and to determine if it can decrease the time taken Further work should aim to apply the
laboratory to assess its utility and to
and therefore, allow more diagnose
Acknowledgements: Further work is the full and there are the time taken to reach a diagnos
and therefore, allow more diagnoses to be made, thus increasing the overall diagnostic rate.
Acknowledgements:
This research was made possible throug

and therefore, allow more diagnoses to be made, thus increasing the overall diagnostic rate.
Acknowledgements:
This research was made possible through access to the data and findings generated by the 100,000
Genomes Projec and interfere, allead there diagnoses to the made, plate interfere, and the diagnostic rate.
Acknowledgements:
This research was made possible through access to the data and findings generated by the 10
Genomes Project. Th ノー(くりこく This research was ma
Genomes Project. Th
owned company of the
National Institute for
and the Medical Rese
Project uses data pro
care and support. The Senomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health). The 100,000 Genomes Project is funded by the National Institute for Health Resea owned company of the Department of Health). The 100,000 Genomes Project is funded by the
National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK
and the Medical Research Council have National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research
and the Medical Research Council have also funded research infrastructure. The 100,000 Genome
Project uses data provided by patien National Institute For Mealth Research Council have also funded research infrastructure. The 100,000 Genomes
Project uses data provided by patients and collected by the National Health Service as part of their
care and sup Project uses data provided by patients and collected by the National Health Service as part of their
care and support.
The 100,000 General research infrastructure. The 100,000 General research in the 100,000 General resear Provided by patients and support.
Property and complete by the National Health Service as part of the Nat

Case Stud. 2018;4(6).
5. Boycott KM, Hartley T, Biesecker LG, Gibbs RA, Innes AM, Riess O, et al. A Diagnosis for All
Rare Genetic Diseases: The Horizon and the Next Frontiers. Cell. 2019;177(1):32-7. S. Boycott KM, F
5. Boycott KM, F
Rare Genetic Diseases Rare Genetic Diseases: The Horizon and the Next Frontiers. Cell. 2019;177(1):32-7.

Sample Range Am, Riess of Anti-Rare Genetic Diseases: The Horizon and the Next Frontiers. Cell. 2019;177(1):32-7.

care and support. $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ H ごくきごく March Merces
1. Boyd
Cooperation
genetics. 2012
2018;379(22
3. Tayle

^{2.} Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. American journal of human
genetics. 2017;100(5):695-705.
2. Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, et al. E
Geneti Expension to Enable the Diagnosis of Spinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, et al. Eff
Genetic Diagnosis on Patients with Previously Undiagnosed Disease. N Engl J Med.
2018;379(22):213 2. Splinter K, Adams DR, B.
Genetic Diagnosis on Patients w
2018;379(22):2131-9.
3. Taylor JC, Martin HC, Lis
success of clinical genome seque
2015;47(7):717-26.
4. Frésard L, Montgomery
Case Stud. 2018;4(6). Genetic Diagnosis on Patients with Previously Undiagnosed Disease. N Engl J Med.
2018;379(22):2131-9.
3. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, et al. Factors influencing
success of clinical genome 2018;379(22):2131-9.
2018;379(22):2131-9.
3. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, et al. Factor
success of clinical genome sequencing across a broad spectrum of disorders. Nat G
2015;47(7):717-26 3. Taylor JC, Mar
success of clinical geno
2015;47(7):717-26.
4. Frésard L, Mor
Case Stud. 2018;4(6).
5. Boycott KM, H
Rare Genetic Diseases Success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet.
2015;47(7):717-26.
4. Frésard L, Montgomery SB. Diagnosing rare diseases after the exome. Cold Spring Harb
Case Stud. 2018;4(6).
5. Boy

success of the Northern Controller and the spectrum of an extrement of the species of the species of the exome.

4. Frésard L, Montgomery SB. Diagnosing rare diseases after the exome. Cold Sprin

Case Stud. 2018;4(6).

5. 2012;47(7):717-26.
4. Frésard L, N
Case Stud. 2018;4(6
5. Boycott KM
Rare Genetic Diseas

It is made available under a [CC-BY-NC-ND 4.0 International license](http://creativecommons.org/licenses/by-nc-nd/4.0/) . **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. medRxiv preprint doi: [https://doi.org/10.1101/2022.08.19.22278944;](https://doi.org/10.1101/2022.08.19.22278944) this version posted August 19, 2022. The copyright holder for this preprint

6.

disease. Nat Rev Genet. 2017;18(10):599-612.

7. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome

sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55.
 7. Bamshad MJ, Ng SB, Bigham AW, Tabo
sequencing as a tool for Mendelian disease ger
8. Havrilla JM, Pedersen BS, Layer RM, Qı
human genome. Nat Genet. 2019;51(1):88-95.
9. Karczewski KJ, Francioli LC, Tiao G, Cun
constrai 9. Bamshad Markover, Mathematic American Sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):74!

8. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regio

huma

S. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in
human genome. Nat Genet. 2019;51(1):88-95.
9. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutati 8. Human genome. Nat Genet. 2019;51(1):88-95.

9. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational

constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;58 man generatorski KJ, Francioli LC, Tiao G, Cum
constraint spectrum quantified from variation i
10. Landrum MJ, Lee JM, Benson M, Browr
access to variant interpretations and supportin
11. Acuna-Hidalgo R, Veltman JA, Hoisch

9. Constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-41
10. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving
access to variant interpretati Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving
access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-d7
11. Acuna-Hidalgo R, Veltman JA, access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-d
11. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de
novo mutations in health and di Accuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de
novo mutations in health and disease. Genome biology. 2016;17(1):241.
12. Brunet T, Jech R, Brugger M, Kovacs R, Alhaddad B, Leszins novo mutations in health and disease. Genome biology. 2016;17(1):241.
12. Brunet T, Jech R, Brugger M, Kovacs R, Alhaddad B, Leszinski G, et al. De novo variants in
neurodevelopmental disorders-experiences from a tertiary novo mutations in health and disease. By Alhaddad B, Leszinski G, eneurodevelopmental disorders-experiences from a tertiary care center. C
28. Disorders DD. Prevalence and architecture of de novo mutations
Nature. 2017;542 neurodevelopmental disorders-experiences from a tertiary care center. Clin Genet. 2021;100(1):2
28.
13. Disorders DD. Prevalence and architecture of de novo mutations in developmental disorc
Nature. 2017;542(7642):433-8.
1

28.

13. Disorders DD. Prevalence and architecture of de novo mutations in developmental disorder

Nature. 2017;542(7642):433-8.

14. Heyne HO, Singh T, Stamberger H, Abou Jamra R, Caglayan H, Craiu D, et al. De novo varia --
13.
Nat
14. in n
15. disc
Pre Nature. 2017;542(7642):433-8.
14. Heyne HO, Singh T, Stamberger H, Abou Jamra R, Caglayan H, Craiu D, et al. De novo variants
in neurodevelopmental disorders with epilepsy. Nat Genet. 2018;50(7):1048-53.
15. Kaplanis J, Sa 14. Heyne HO, Singh T, Star

in neurodevelopmental disorde

15. Kaplanis J, Samocha KE,

disorders discovered by combin

Prevalence and architecture of (

2020;586(7831):757-62.

16. Pode-Shakked B, Barel (

experience wit

in neurodevelopmental disorders with epilepsy. Nat Genet. 2018;50(7):1048-53.
15. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, et al. Evidence for 28 genetic
disorders discovered by combining healthcare in neurodevelopmental anexions KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, et al. Evid
disorders discovered by combining healthcare and research data
Prevalence and architecture of de novo mutations in developmental disor 15. Kaplandian Screen Screen

16. Prevalence and architecture of de novo mutations in developmental disorders. Nature.

2020;586(7831):757 Prevalence and architecture of de novo mutations in developmer
2020;586(7831):757-62.
16. Pode-Shakked B, Barel O, Singer A, Regev M, Poran H, Eliy
experience with publicly funded clinical exome sequencing for ne
multiple 2020;586(7831):757-62.
2020;586(7831):757-62.
16. Pode-Shakked B, Barel O, Singer A, Regev M, Poran H, Eliyahu A, et al. A single c
experience with publicly funded clinical exome sequencing for neurodevelopmental disorders 16. Pode-Shakked B,
experience with publicly
multiple congenital anon
17. Samocha KE, Rok
for the interpretation of
18. Sevim Bayrak C,
congenital heart disease
19. Torjesen I. Genoi

experience with publicly funded clinical exome sequencing for neurodevelopmental disorders
multiple congenital anomalies. Sci Rep. 2021;11(1):19099.
17. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, e experience multiple congenital anomalies. Sci Rep. 2021;11(1):19099.
17. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framewor
for the interpretation of de novo mutation in human disease. Na multiple congenital and KE, Robinson EB, Sanders SJ, Stevens C, S
for the interpretation of de novo mutation in human disease
18. Sevim Bayrak C, Zhang P, Tristani-Firouzi M, Gelb B
congenital heart disease patients identi

17. Samon Maria Controllary Controllary School (17. Sevim Bayrak C, Zhang P, Tristani-Firouzi M, Gelb BD, Itan Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. Genome Med For Sevim Bayrak C, Zhang P, Tristani-Firouzi M, Gelb BD, Itan Y. De novo variants in exom
congenital heart disease patients identify risk genes and pathways. Genome Med. 2020;12(1)
19. Torjesen I. Genomes of 100,000 peopl congenital heart disease patients identify risk genes and pathways. Genome Med. 2020;12(1):9.
19. Torjesen I. Genomes of 100,000 people will be sequenced to create an open access researc
resource. Bmj. 2013;347:f66 Congenium and the patients identify risk and pather and periodic intervals and pather and pathers are resource. Bmj. 2013;347:f6690.

20. Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000

G 19. Torsource. Bmj. 2013;347:f6690.
19. Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000
19. Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med.
19. T 20. Smedley D, Smith KR, M
Genomes Pilot on Rare-Disease
2021;385(20):1868-80.
21. Team R. RStudio: Integra
22. Quinlan AR, Hall IM. BEI
Bioinformatics (Oxford, England
23. Lek M, Karczewski KJ, M
coding genetic variation Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med.
2021;385(20):1868-80.
21. Team R. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. 2020.
22. Quinlan

2021;385(20):1868-80.
21. Team R. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. 2020.
22. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic fear
Bioinformatics (Oxford, E 21. Team R. RStudion
22. Quinlan AR, Hal
Bioinformatics (Oxford,
23. Lek M, Karczew
coding genetic variation
24. Samocha KE, Ro
for the interpretation o
25. Köhler S, Garga 22. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic fe
Bioinformatics (Oxford, England). 2010;26(6):841-2.
23. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Ana

Bioinformatics (Oxford, England). 2010;26(6):841-2.
23. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein
coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91. 23. Lek M, Karczewski KJ, Minikel EV, Samocha K
coding genetic variation in 60,706 humans. Nature. 2
24. Samocha KE, Robinson EB, Sanders SJ, Steve
for the interpretation of de novo mutation in human
25. Köhler S, Gargano

Composity of the interpretation of de novo mutation in human disease. Nat Genet. 2014
For the interpretation of de novo mutation in human disease. Nat Genet. 2014
25. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-S

22. Samon Kinder Mannes Matten Controlling and Sease. Nat Genet. 2014;46(9):944-50.
25. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The
25. Köhler S, Gargano M, Matentzoglu N, Carmo

27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and
server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et a
Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207-d17.
26. Kumar P, Henikoff S, Ng PC. Predicting the effects 25. Komar Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207-d17.
26. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on
protein function using the SIFT algorithm. Nat Pro Human Phenotype Enterty, Martin Controllect Reserves of coding non-synonym protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.
27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.
27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method a
server for predicting damaging missense mutations. Nat M protein function in September 2. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, B
server for predicting damaging missense mutations. Nat Methods. 2010;7
28. Rentzsch P, Witten D, Cooper GM, Shendure J, Kirch 27. Server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
28. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
deleteriousness of variants throughout the human genome. server for predicting analoging interactions and the Machine of Account 2013, (1, 1, 2014)
28. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
deleteriousness of variants throughout the human g deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(
d94.
. deleteriousness of variants throughout throughout throughout the human generation and the human generation $(1-p)$:
d94.

It is made available under a [CC-BY-NC-ND 4.0 International license](http://creativecommons.org/licenses/by-nc-nd/4.0/) . **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. medRxiv preprint doi: [https://doi.org/10.1101/2022.08.19.22278944;](https://doi.org/10.1101/2022.08.19.22278944) this version posted August 19, 2022. The copyright holder for this preprint

29. Novo Variants in the ATPase Module of MORC2 Cause a Neurodevelopmental Disorder with Grover Netardation and Variable Craniofacial Dysmorphism. American journal of human genetics.
2020;107(2):352-63.
30. Anderson E, Lam Z,

Retardation and Variable Craniofacial Dysmorphism. American journal of human genetics.
2020;107(2):352-63.
30. Anderson E, Lam Z, Arundel P, Parker M, Balasubramanian M. Expanding the phenotype of
SETD5-related disorder an Retardation and Retardation and Retardation 2020;107(2):352-63.

Retards Craniof Cranio and Parker M, Balasubramanian M. Expanding the phen-

SETD5-related disorder and presenting a novel association with bone fragility. C 2021;107(2):2021
2021;100(3):352-4.
2021;100(3):352-4.
31. Hood RL, Line
Mutations in SRCAP,
syndrome. American
32. Lee CG, Seol
with dysmorphic faci

SETD5-related disorder and presenting a novel association with bone fragility. Clin Genet.
2021;100(3):352-4.
31. Hood RL, Lines MA, Nikkel SM, Schwartzentruber J, Beaulieu C, Nowaczyk MJ, et al.
Mutations in SRCAP, encodi 2021;100(3):352-4.

SET And RL, Lines MA, Nikkel SM, Schwartzentruber J, Beaulieu C, Nowaczyk MJ, et a

Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Hark

syndrome. American journal of 2022
2021 - Hood RL, Lin
Mutations in SRCAP
syndrome. America
32. Lee CG, Seo
with dysmorphic fac
Med Genet A. 2020, Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbo
syndrome. American journal of human genetics. 2012;90(2):308-13.
32. Lee CG, Seol CA, Ki CS. The first familial case of inherited int syndrome. American journal of human genetics. 2012;90(2):308-13.
32. Lee CG, Seol CA, Ki CS. The first familial case of inherited intellectual developmental d
with dysmorphic facies and behavioral abnormalities (IDDFBA) wi spannian Lee CG, Seol CA, Ki CS. The first familial case of inherited interprise with dysmorphic facies and behavioral abnormalities (IDDFBA) with a
Med Genet A. 2020;182(11):2788-92.
Med Genet A. 2020;182(11):2788-92. with dysmorphic facies and behavioral abnormalities (IDDFBA) with a novel FBXO11 variant. Am J
Med Genet A. 2020;182(11):2788-92.
 with dysmorphic facies and behavioral abit distributions (IDDFBA) with a normality (IDDFBA) Med Genet A. 2020;182(11):2788-92.