Clustering of predicted loss-of-function variants in genes linked with monogenic disease can explain incomplete penetrance

Robin N. Beaumont^{1*}, Gareth Hawkes¹, Adam C. Gunning^{1,2}, Caroline F. Wright^{1*}

¹ Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, EX1 2LU

² Exeter Genomics Laboratory, Royal Devon University Healthcare NHS Foundation Trust, Exeter, EX2 5DW

Abstract
Predicted loss-of-function variants (pLoFs) are often associated with disease. Fo
monogenic diseases, we hypothesised that pLoFs present in apparently unaffec
cluster in LoF-tolerant regions. We compared the distr Abstract
Predicted
monogen
cluster in
individua
genes in less likely
ARID1B a
translatio monogenic diseases, we hypothesised that pLoFs present in apparently unaffected individuals may
cluster in LoF-tolerant regions. We compared the distribution of pLoFs in ClinVar versus 454,773
individuals in UK Biobank and cluster in LoF-tolerant regions. We compared the distribution of pLoFs in ClinVar versus 454,773
individuals in UK Biobank and clustered the variants using Gaussian mixture models. We found that
genes in which haploinsuffi individuals in UK Biobank and clustered the variants using Gaussian mixture models. We found then genes in which haploinsufficiency causes developmental disorders with incomplete penetrance v
less likely to have a uniform genes in which haploinsufficiency causes developmental disorders with incomplete penetrance were
less likely to have a uniform pLoF distribution than other genes (P<2.2x10-6). In some cases (e.g.,
ARID1B and *GATA6*), p genes likely to have a uniform pLoF distribution than other genes (P<2.2x10-6). In some cases (e.g., ARID1B and GATA6), pLoF variants in the first quarter of the gene could be rescued by an alternative translation start s ARID1B and GATA6), pLoF variants in the first quarter of the gene could be rescued by an alternat
translation start site and should not be reported as pathogenic. In other cases (e.g., *ODC1*), patho
pLoFs were clustered o ARIDID and GATA6), pLoT variants in the first quarter of the gene could be rescued by an alternative
translation start site and should not be reported as pathogenic. In other cases (e.g., *ODC1*), pathogen
pLoFs were clust translation start site and should not be reported as pathogenic. In other cases (e.g., ODC1), pathogenic
pLoFs were clustered only at the end of the gene, consistent with a gain-of-function disease mechanism
Our results su

Introduction

place that support the use of localised constraint metrics when interpreting variants.
Our results support the use of localised constraint metrics when interpreting variants.
Contrary to expectation, many individuals in th Our results support the use of localised constraint metric interpreting variants.

Contrary to expectation, many individuals in the population harbour predicted loss of f

variants in genes where haploinsufficiency is know | (\ F t c i variants in genes where haploinsufficiency is known to cause highly penetrant monogenic conditions¹-
For example, pLoF variants in genes that cause severe developmental disorders (DD) in childhood wou
not be expected to variants in genes where haploinsufficiency is known to cause highly penetrant monogenic conditions⁺.
For example, pLoF variants in genes that cause severe developmental disorders (DD) in childhood would
not be expected t not be expected to be present at appreciable levels in the general adult population. Nonetheless, we
and others have previously shown that thousands of individuals in UK Biobank (UKB) carry pLoF variants
in DD genes and h and others have previously shown that thousands of individuals in UK Biobank (UKB) carry pLoF variat
in DD genes and have phenotypes consistent with incomplete penetrance or reduced expressivity,
though very few individua in DD genes and have phenotypes consistent with incomplete penetrance or reduced expressivity,
though very few individuals have DD diagnoses⁴⁻⁷. There are several possible explanations for this
observation. One possibil though very few individuals have DD diagnoses⁴⁻⁷. There are several possible explanations for this
observation. One possibility is that genetic or environmental modifiers alter the impact of individua
variants⁸, such t though very few individuals have DD diagnoses³⁻. There are several possible explanations for this observation. One possibility is that genetic or environmental modifiers alter the impact of individu
variants⁸, such t variants⁸, such that the penetrance of pathogenic variants identified in affected families or disease
cohorts may be over-estimated. An alternative explanation is that some pLoF variants in these genes
not cause loss of variants"
cohorts r
not cause
because
exon skip
importar
protein, a
penetrar may be over-estimated. An alternative explanation is that some pLoF variants in these genetral falses of function, either because they are technical false positives⁹ or mosaic variants, or they can be rescued through a Find the set of function, either because they are technical false positives⁹ or mosaic variants, or because they can be rescued through a variety of mechanisms, including alternative transcription¹⁰, exon skipping¹¹ not cause loss of function, either because they are technical false positives³
because they can be rescued through a variety of mechanisms, including alt
exon skipping¹¹, escape from nonsense-mediated decay (NMD)¹², Exerce intersection¹

Station re-initiation¹³. It

It ormal levels of function

product, both for estin

product, both for estin because they can be rescued through a variety of mechanisms, including alternative transcription¹³, exam exon skipping¹¹, escape from nonsense-mediated decay (NMD)¹², and translation re-initiation¹³. It is importa exon skipping**, escape from nonsense-mediated decay (NMD)**, and translation re-initiation**. It is
important to distinguish between benign pLoF variants that produce near-normal levels of functiona
protein, and pathogeni important to distinguish between benefited by the substantially reduced protein product, both for estimate
penetrance and interpreting diagnostic results.
NOTE: This preprint reports new research that has not been certifie protein and interpreting diagnostic results.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice. that are intolerant to pLoF variation^{14,15}, and regions of genes that are intolerant to missense
variation^{16,17}. However, the location of pLoF variants in genes has not been systematically investigated a
large scale du that are intolerant to pLoF variation^{24,25}, and regions of genes that are intolerant to missense
variation^{16,17}. However, the location of pLoF variants in genes has not been systematically inver
large scale due to lack variation²⁰¹². However, the location of pLoF variants in genes has not been systematically investigated at large scale due to lack of sequence data on large numbers of individuals. We used cluster analysis of exome seque exome sequence (ES) data from UKB to identify genes showing distinct patterns in the location of pL

variants. We then investigated genes showing these distinct profiles of pLoF location to determine

whether they could ex

pLoF variants are more likely to be non-uniformly distributed than missense or synonymous variants

variants. We then investigated genes showing these distinct profiles of pLoF location to determine
whether they could explain the presence of such putatively pathogenic variants in a population cohort.
Results
pLoF vari whether they could explain the presence of such putatively pathogenic variants in a population coherently and the metallity of the non-uniformly distributed than missense or synonymous variation.
We calculated the relative Results

ploF variants are more likely to be non-uniformly distributed than missense or synonymous variants

We calculated the relative location of every coding variant detected in ES data from 454,773 individuals

in UKB nesans
pLoF va
We calc
in UKB i
consequ
We ther
into sev
uniform in UKB in the coding sequence (CDS) of each gene, and the relative proportion of variants grouped by consequence class in each transcript¹⁸ (synonymous, missense, pLoF) within each quintile of the CDS. We then used Gaus consequence class in each transcript¹⁸ (synonymous, missense, pLoF) within each quintile of the CDS.
We then used Gaussian mixtures to cluster the profile of the variants of each class within the transcription of
seven c consequence class in each transcript⁴¹ (synonymous, missense, pLoF) within each quintile of the CDS.
We then used Gaussian mixtures to cluster the profile of the variants of each class within the transcription
into seven into seven clusters (Fig. 1). Of these, three clusters represented variants being distributed more-or-less
uniformly throughout the CDS (clusters 1-3; Fig. 2), and one identified genes with no variants of a
particular clas uniformly throughout the CDS (clusters 1-3; Fig. 2), and one identified genes with no variants of a
particular class (not shown). The remaining three clusters showed distinct patterns in the location of
variants, with at l

particular class (not shown). The remaining three clusters showed distinct patterns in the location variants, with at least one quintile of the CDS containing zero (or very few) variants of a given varial class, and most v partiants, with at least one quintile of the CDS containing zero (or very few) variants of a given variant
class, and most variants either being towards the first or second half of the gene (clusters 4-6; Fig 2).
We limite class, and most variants either being towards the first or second half of the gene (clusters 4-6; Fig 2).
We limited our analyses to transcripts with at least five variants of each consequence class, and only
considered M Class, and only considered MANE Select transcripts with at least five variants of each consequence class, and only considered MANE Select transcripts for each gene in our primary analysis (16,473 genes). We found that, for considered MANE Select transcripts for each gene in our primary analysis (16,473 genes). We found
that, for most genes, synonymous and missense variants fell within the uniform clusters 1-3 (99.3% a
99.4% respectively; Fi that, for most genes, synonymous and missense variants fell within the uniform clusters 1-3 (99.3% a
99.4% respectively; Fig. 3) and we excluded 114 genes with synonymous or missense variants in non
uniform clusters 4-6 f 99.4% respectively; Fig. 3) and we excluded 114 genes with synonymous or missense variants in non-
99.4% respectively; Fig. 3) and we excluded 114 genes with synonymous or missense variants in non-
uniform clusters 4-6 fr regions of the gene. In contrast, we found considerably more genes with pLoFs in the non-uniform
clusters 4-6 (n=1460, 8.9%) compared to synonymous and missense (P < 2.2e-16). These distinct pr
for pLoF location were not clusters 4-6 (n=1460, 8.9%) compared to synonymous and missense (P < 2.2e-16). These distinct pr
for pLoF location were not driven by the possible locations of pLoF variants, based on the underlyin
sequence, where only 63 for pLoF location were not driven by the possible locations of pLoF variants, based on the underlying
sequence, where only 63 genes had non-uniform distributions. Simulations showed that, while
constrained genes were more For periods, where only 63 genes had non-uniform distributions. Simulations showed that, while
constrained genes were more likely to fall into the non-uniform distribution clusters, there was an
enrichment of genes within

pl of variants are more likely to be non-uniformly distributed in genes linked with autosomal pLoF variants are more likely to be non-uniformly distributed in genes linked with autosomal dominant conditions

constrained genes were more likely to fall into the non-uniform distribution clusters, there was
enrichment of genes within these clusters compared to the expected distribution (p<0.0001; Fi_l
ploF variants are more likel **Example 12** and the set of the expected distribution (p<0.0001; Fig. 4)
 pLoF variants are more likely to be non-uniformly distributed in genes linked with autosomal
 dominant conditions

We observed that pLoF varian pLoF variants are more likely to be non-uniformly distributed in genes linked with autosomal
dominant conditions
We observed that pLoF variants in UKB were more likely to be non-uniformly distributed in autosom
dominant DD dominant DD genes (AD-DD)¹⁹ versus other genes, possibly indicating regions where pLoFs are tolerate
and do not cause severe disease. Of AD-DD genes, where pLoFs cause DD through haploinsufficiency
(421 genes)¹⁹, we ob dominant DD genes (AD-DD)²⁵ versus other genes, possibly indicating regions where pLoFs are tolerated
and do not cause severe disease. Of AD-DD genes, where pLoFs cause DD through haploinsufficiency
(421 genes)¹⁹, we o (421 genes)¹⁹, we observed 29 with no pLoFs in UKB and 153 with <5 pLoFs (Table S1). Of the remaini
239 AD-DD genes, pLoFs in 41.4% were non-uniformly distributed (Fig. 5), which contrasted with 3.3%
for autosomal reces (421 genes)²⁹, we observed 29 with no pLoFs in UKB and 153 with <5 pLoFs (Table S1). Of the remaining 239 AD-DD genes, pLoFs in 41.4% were non-uniformly distributed (Fig. 5), which contrasted with 3.3% for autosomal rece for autosomal recessive DD (AR-DD) genes. Genes linked with a range of adult-onset autosomal
dominant diseases (including cardiac conditions, heritable cancer syndromes and eye disorders), were
also more likely to have non dominant diseases (including cardiac conditions, heritable cancer syndromes and eye disorders)
also more likely to have non-uniformly distributed pLoFs (25.3%) than other genes. Simulations
that while these genes generally also more likely to have non-uniformly distributed pLoFs (25.3%) than other genes. Simulations showed
that while these genes generally had fewer pLoFs than other genes, this did not explain the non-uniform
distributions of also more likely to have non-uniformly alternative planet (25.2%) manipulate generalizations of ploFs (p<0.0001).

distributions of ploFs (p<0.0001). distributions of pLoFs (p<0.0001).
 $\frac{1}{\sqrt{2}}$ distributions of pLoFs (p<0.0001).

 μ _F μ ₇.000 are found that among 1438 genes with at least five such variants, only 51 (3.5%) were uniformly distributed. Across 156 AD-DD genes with at least five pLoFs in both UKB and ClinVar datasets we found 83 in ClinVar²⁰ we found that among 1438 genes with at least five such variants, only 51 (3.5%) were
uniformly distributed. Across 156 AD-DD genes with at least five pLoFs in both UKB and ClinVar da
we found 83 (53.2%) gen we found 83 (53.2%) genes where ploFs fell into one of the uniform clusters in UKB, compared with 150 (96.2%) in ClinVar (2-sided binomial $P < 2.2x10^{-16}$). The majority of genes clustered similarly in both datasets; for (96.2%) in ClinVar (2-sided binomial $P < 2.2x10^{-16}$). The majority of genes clustered similarly in both datasets; for example, pLoF variants in *COL4A3* (associated with Alport syndrome, MIM #104200) are uniformly distri (96.2%) in ClinVar (2-sided binomial P < 2.2x10^{-x}). The majority of genes clustered similarly in both datasets; for example, pLoF variants in *COL4A3* (associated with Alport syndrome, MIM #104200) an uniformly distribut datasets, for example, plear variants in Colembo (associated with Alport syndrome, time in 50-200) are
uniformly distributed throughout the gene in both UKB and ClinVar (Fig. 6a). In such cases, where plo
are uniformly dis

Non-uniform distributions of pLoF variants may explain incomplete penetrance through a variety of molecular mechanisms including alternative splicing and translation re-initiation

are uniformly distributed throughout a gene in both population and clinical datasets, this approach is
not able to determine why some pLoF variants are likely to be benign whilst others are pathogenic.
Non-uniform distribu not able to determine why some pLoF variants are likely to be benign whilst others are pathogenic.
Non-uniform distributions of pLoF variants may explain incomplete penetrance through a variety of molecular mechanisms inc Non-uniform distributions of pLoF variants may explain incomplete penetrance through a variety
molecular mechanisms including alternative splicing and translation re-initiation
In 43.6% of genes with at least five pLoF var differed substantively between UKB and ClinVar (i.e., one was uniform whilst the other was non-
uniform). We hypothesised that these might represent examples where variant location could expli
incomplete penetrance. We exa

Select transcript in UKB, of which five are in early exons not present in the MANE Plus Clinical transcript incomplete penetrance. We examined this list of genes for examples where the difference in
distributions was robust, based on visual inspection of the underlying pLoF variant distributions and
sequence data, and investigat distributions was robust, based on visual inspection of the underlying pLoF variant distributio
sequence data, and investigated potential mechanistic explanations.
One mechanism which might explain incomplete penetrance is distributions which might explain incomplete penetrance is the existence of alternative transcript
one mechanism which might explain incomplete penetrance is the existence of alternative transcript
where benign pLoF varian Sequence in the penetrance is the explain of the where benign plof variants are clustered in exons that are excluded for example, $TP63$ (associated with Hay-Wells syndrome, MIM #1062
Select transcript in UKB, of which five where benign pLoF variants are clustered in exons that are excluded from other functional transcripts.
For example, *TP63* (associated with Hay-Wells syndrome, MIM #106260) has seven pLoFs in the MANE
Select transcript in For example, *TP63* (associated with Hay-Wells syndrome, MIM #106260) has seven pLoFs in the MANE
Select transcript in UKB, of which five are in early exons not present in the MANE Plus Clinical transcrip
(Fig. 6b). This t For example, The Gassociated with Hay-Wells syndrome, MIM #1002007 has seven plot s in the MANE
Select transcript in UKB, of which five are in early exons not present in the MANE Plus Clinical transcript
(Fig. 6b). This tr (Fig. 6b). This transcript (ENST00000354600) has an alternative later start codon but contains all the ClinVar pathogenic pLoF variants. In addition to MANE Plus Clinical transcripts, which may be the most obvious candida ClinVar pathogenic pLoF variants. In addition to MANE Plus Clinical transcripts, which may be the mo
obvious candidates for alternative transcripts to explain the presence of pLoFs in apparently healthy
individuals, other obvious candidates for alternative transcripts to explain the presence of pLoFs in apparently healthy individuals, other transcripts may show higher expression levels and explain non-penetrance of some disease genes. Examp individuals, other transcripts may show higher expression levels and explain non-penetrance of some
disease genes. Examples highlighted by our analysis include two large genes, $ARID1B$ (associated with
Coffin-Siris syndrom disease genes. Examples highlighted by our analysis include two large genes, *ARID1B* (associated with
Coffin-Siris syndrome, MIM #135900) and *NSD1* (associated with Sotos syndrome, MIM #117550), wh
have 13 and 31 pLoFs i disease genes. Examples highlighted by our analysis include two large genes, AMD1D (associated with
Coffin-Siris syndrome, MIM #135900) and *NSD1* (associated with Sotos syndrome, MIM #117550), whi
have 13 and 31 pLoFs in Coffin-Siris syndrome, MIM #133500) and NSD1 (associated with Sotos syndrome, MIM #117350, Which
have 13 and 31 pLoFs in UKB respectively. For *ARID1B*, 10/13 pLoFs fall before Met584 of the MANE
Select transcript (ENST000 have 13 and 31 pLoFs in OKB respectively. For AMD1D, 10/13 pLoFs fan before wiet 504 of the WANE
Select transcript (ENST00000346085), which also corresponds to the start of an alternative transcript
(ENST00000414678) that (ENST00000414678) that shows higher expression in GTeX v7 than the MANE Select transcript²¹ (Fig. 6
For *NSD1*, all the pLoFs in UKB occur either in the large last exon or the first exon of the MANE Select
transcript (E (ENST00000414678) that shows higher expression in GTeX v7 than the MANE Select transcript⁺⁺ (Fig. 6c).
For *NSD1*, all the pLoFs in UKB occur either in the large last exon or the first exon of the MANE Select
transcript For NSD1, all the peors in OKB occur either in the large last exord or the mist exord or the MANE Select
transcript (ENST00000439151) (Fig. 6d). The first exon is excluded from an alternative transcript,
ENST00000354179, w ENST00000354179, which has much higher expression levels in GTeX v7 than the MANE Select
transcript. While the final exon is included in both transcripts, it lies downstream of the functiona
domains of the protein, and sin transcript. While the final exon is included in both transcripts, it lies downstream of the functio
domains of the protein, and since pLoFs in the final exon usually escape NMD, a functional C-te
truncated protein could be

transcript. While the protein, and since ploFs in the final exon usually escape NMD, a functional C-termi
truncated protein could be produced. In both cases, the ploF variants in ClinVar are fairly uniform
distributed thro truncated protein could be produced. In both cases, the pLoF variants in ClinVar are fairly uniformly
distributed throughout the rest of the gene, but lie outside of these exons.
It is also important to consider not only a Instributed throughout the rest of the gene, but lie outside of these exons.
It is also important to consider not only alternative transcription but also translation re-initiation in
explaining incomplete penetrance, as al distributed throughout the rest of the gene, substituted the rest of the same transplaining incomplete penetrance, as alternative start sites on the same trap
pLoF variants (though these may not be annotated as such). For It is also important to complete penetrance, as alternative start sites on the same transcript could rescue son
pLoF variants (though these may not be annotated as such). For example, *GATA6* (associated with
pancreatic ag pLoF variants (though these may not be annotated as such). For example, *GATA6* (associated with
pancreatic agenesis and congenital heart defects, MIM #600001) has 10 pLoFs in UKB, of which eight a
located before Met147 (F pancreatic agenesis and congenital heart defects, MIM #600001) has 10 pLoFs in UKB, of which eighorated before Met147 (Fig. 6e). This contrasts with ClinVar variants, which all lie after Met147. pancreatic agencies and congenitative action, MIM #6000, Just 20 plot and they be minded agencies and located before Met147 (Fig. 6e). This contrasts with ClinVar variants, which all lie after Met147. located before Met147 (Fig. 6e). This contrasts with ClinVar variants, which all lie after Met147.

Although there is only a single known transcript for this gene, *GATA6* can be produced through
translation re-initiation from Met147, creating a second recognised protein isoform²² that is shorter but
still retains the translation re-initiation from Met147, creating a second recognised protein isoform²² that is shorter but still retains the functional domains. Unlike the many other AD-DD genes, the phenotypes linked with GATA6 haploins GATA6 haploinsufficiency are both specific and severe enough that we considered it implausible they
would not be recorded in the linked electronic health records of UKB participants; importantly, we not
hat none of the 30

Pathogenic pLoF variants at the end of genes may point towards a gain-of-function disease mechanism

GATA6 haplomsufficiency are both specific and severe enough that we considered it implausible they
would not be recorded in the linked electronic health records of UKB participants; importantly, we not
that none of the 30 that none of the 30 carriers have any indication of either pancreatic agenesis or cardiac malformations.
 Pathogenic pLoF variants at the end of genes may point towards a gain-of-function disease
 Pathogenic ploF varian Pathogenic ploF variants at the end of genes may point towards a gain-of-function disease
 Pathogenic ploF variants at the end of genes may point towards a gain-of-function disease

Finally, we also found a small numbe UKB but non-uniformly distributed in ClinVar. For example, in *ODC1* (associated with Bachmann-Bupp syndrome, MIM #165640), all 11 pLoFs in UKB occur before the penultimate exon, whilst ClinVar pathogenic variants all occ oko but non-uniformly distributed in chrivar. For example, in ODC1 (associated with bachmann-Bupp
syndrome, MIM #165640), all 11 ploFs in UKB occur before the penultimate exon, whilst ClinVar
pathogenic variants all occur pathogenic variants all occur in the last or penultimate exons (Fig. 6f). Here, despite being annotation plof, there is no evidence that haploinsufficiency causes disease, and pathogenic variants at the the gene are likely ploF, there is no evidence that haploinsufficiency causes disease, and pathogenic variants at the end of
the gene are likely to result in a gain-of-function (GoF), for example, by causing resistance to normal
degradation²

plane are likely to result in a gain-of-function (GoF), for example, by causing resistance to normal
degradation²³.
Discussion
Using cluster analysis, we have identified 1460 genes which show distinct patterns of pLoF lo the gradation²³.
 Discussion

Using cluster analysis, we have identified 1460 genes which show distinct patterns of pLoF location

within UKB, of which 16.4% are in genes where haploinsufficiency causes monogenic disea degradation²³.
Discussion
Using cluster an
within UKB, of
generally assur
characterised g
mechanisms th
examples show **Discussion**
Using cluste
within UKB,
generally as
characterise
mechanism
examples sl
understand within UKB, of which 16.4% are in genes where haploinsufficiency causes monogenic diseases that a
generally assumed to be fully penetrant. We have also highlighted specific examples of well clinicall
characterised genes, i generally assumed to be fully penetrant. We have also highlighted specific examples of well clinically
characterised genes, including *GATA6* and *ARID1B*, where we were able to suggest potential molecular
mechanisms that generally assumed to be fully a *GATA6* and *ARID1B*, where we were able to suggest potential molecula
mechanisms that may explain the presence of pLoF variants in apparently healthy individuals. These
examples show the im

characterised genes, including on 7 and ARID1B, where we were able to suggest potential molecular
mechanisms that may explain the presence of pLoF variants in apparently healthy individuals. These
examples show the importa mechanisms that may example show the importance of examining alternative transcription and alternative translation to
understand the clinical impact of pLoFs.
Haploinsufficient genes can be divided into three groups based understand the clinical impact of pLoFs.

Haploinsufficient genes can be divided into three groups based on the distribution of population ger

variation in UKB: (1) those where we observe too few pLoF variants to be able Haploinsufficient genes can be divided in
variation in UKB: (1) those where we obs
(37.4%); (2) those where we observe dis
and (3) those where we observe a broad
these groups, the low numbers of pLoFs
being genuinely highl wariation in UKB: (1) those where we observe too few pLoF variants to be able to cluster them effectively (37.4%); (2) those where we observe distinct non-uniform patterns of pLoF variant distribution (20.9%); and (3) tho (37.4%); (2) those where we observe distinct non-uniform patterns of pLoF variant distribution (20.9%); and (3) those where we observe a broadly uniform distribution of pLoF variants (41.6%). For the first of these groups and (3) those where we observe a broadly uniform distribution of pLoF variants (41.6%). For the first of these groups, the low numbers of pLoFs in UKB may be the result of haploinsufficiency in these genes being genuinely these groups, the low numbers of pLoFs in UKB may be the result of haploinsufficiency in these genes
being genuinely highly penetrant. For the second group we have demonstrated how this distribution ca
explain incomplete p these genuinely highly penetrant. For the second group we have demonstrated how this distribution cexplain incomplete penetrance of pLoF variants in many of these genes. The final group of genes (whe we observe uniform di explain incomplete penetrance of pLoF variants in many of these genes. The final group of genes (where we observe uniform distributions of pLoFs in UKB) is perhaps most puzzling; although a subset may exhibit patterns of explain in the observe uniform distributions of pLoFs in UKB) is perhaps most puzzling; although a subset may
exhibit patterns of pLoF variant locations that are below the resolution captured by the quintiles used in
our exhibit patterns of pLoF variant locations that are below the resolution captured by the quintiles use our clustering approach, this is unlikely to be the case for all of them. Similarly, although a subset m cause unrecog our clustering approach, this is unlikely to be the case for all of them. Similarly, although a subset may cause unrecognised developmental disorders in some individuals, this is unlikely to be true for the majority given cause unrecognised developmental disorders in some individuals, this is unlikely to be true for the majority given the known ascertainment bias towards healthy individuals in UKB²⁴. However, some genes (such as $SRCAP^{25}$ majority given the known ascertainment bias towards healthy individuals in UKB²⁴. However, some
genes (such as *SRCAP*²⁵) contain pLoFs that cause clinically distinct DDs via different mechanisms bi
on their location, majority given the known ascertainment bias towards healthy individuals in UKB²⁻³. However, some
genes (such as *SRCAP*²⁵) contain pLoFs that cause clinically distinct DDs via different mechanisms ba
on their location, genes (such as *SRCAP²²*) contain pLoFs that cause clinically distinct DDs via different mechanisms based
on their location, with phenotypes ranging from mild to severe. For other genes with uniformly
distributed pLoFs, distributed pLoFs, the presence of incompletely penetrant pLoF variants may instead indicate the
presence of modifiers, potentially in other genes or nearby non-coding regions. Understanding t
mechanisms which modify the p presence of modifiers, potentially in other genes or nearby non-coding regions. Understanding the
mechanisms which modify the penetrance of these genes will require sequence data on large nun
of affected individuals to com prechanisms which modify the penetrance of these genes will require sequence data on large numk
of affected individuals to compare to healthy controls and is beyond the scope of this study, but wo
enable assessment of geno of affected individuals to compare to healthy controls and is beyond the scope of this study, but would
enable assessment of genotype-phenotype correlations and disease mechanisms at a sub-genic level.
enable assessment of enable assessment of genotype-phenotype correlations and disease mechanisms at a sub-genic level.

Enable assessment of genotype-phenotype correlations and disease mechanisms at a sub-genic level. enable assessment of genotype-phenotype correlations and disease mechanisms at a sub-genic level.

may explain incomplete penetrance, there are some notable limitations. The use of quintiles to
normalise the position of variants within genes means there will be patterns which are missed by our
clustering approach, as th mormalise the position of variants within genes means there will be patterns which are missed be
clustering approach, as their distribution is below the resolution captured by quintiles. Also, as
demonstrated by the simula clustering approach, as their distribution is below the resolution captured by quintiles. Also, as
clemonstrated by the simulation analyses, there are a number of genes which will fall into clusters wit
distinct patterns o demonstrated by the simulation analyses, there are a number of genes which will fall into clust
distinct patterns of pLoF distribution by chance, rather than being driven by underlying biologic
mechanisms. Identifying thes distinct patterns of pLoF distribution by chance, rather than being driven by underlying biological
mechanisms. Identifying these genes and separating them from those where the patterns of pLoFs are
informative requires ad mechanisms. Identifying these genes and separating them from those where the patterns of pLoF
informative requires additional data and may not always be possible with high confidence. Howev
believe that we have demonstrate informative requires additional data and may not always be possible with high confidence. However, w
believe that we have demonstrated the utility of our approach, which will improve with larger datasets
Additionally, whil believe that we have demonstrated the utility of our approach, which will improve with larger datasets.
Additionally, while none of the individuals in UKB carrying pLoFs in the genes highlighted have been
diagnosed with an Additionally, while none of the individuals in UKB carrying pLoFs in the genes highlighted have been
diagnosed with any of the conditions in question, there may be relevant phenotypes not captured in the
UKB data. Increasi

diagnosed with any of the conditions in question, there may be relevant phenotypes not captured in UKB data. Increasing the sample size would also allow us to increase the robustness of the clustering especially for highly UKB data. Increasing the sample size would also allow us to increase the robustness of the clustering,
especially for highly constrained genes with few carriers in UKB.
We have shown how genes associated with assumed fully especially for highly constrained genes with few carriers in UKB.
We have shown how genes associated with assumed fully penetrant childhood-onset conditions throu
haploinsuffiency can have regions where predicted pathogeni Expeciency of the magin, performance genes intuitive carrier in the the building time thaploins uffiency can have regions where predicted pathogenic v
disease. Excluding such variants from both diagnostic pipelines a
cruci haploinsuffiency can have regions where predicted pathogenic variants are tolerated and don't cause
disease. Excluding such variants from both diagnostic pipelines and studies of disease penetrance is
crucial. For example, disease. Excluding such variants from both diagnostic pipelines and studies of disease penetrance is
crucial. For example, within *GATA6* and *ARID1B*, we suggest that pLoFs occurring in the first quarter of
the CDS of the crucial. For example, within *GATA6* and *ARID1B*, we suggest that pLoFs occurring in the first quarter
the CDS of the MANE Select transcript (corresponding to the first 146 and 583 amino acids of the
proteins, respectivel crucial. For example, within GATA6 and ARID1B, we suggest that plots soccurring in the first quarter of
the CDS of the MANE Select transcript (corresponding to the first 146 and 583 amino acids of the
proteins, respectivel proteins, respectively) do not cause disease and should not be reported diagnostically. We have a
demonstrated the benefits of using regional rather than gene-wide constraint metrics to understa
potential impact of pLoF va produced the benefits of using regional rather than gene-wide constraint metrics to understand
potential impact of pLoF variants, and our results may be helpful in determining whether pLoF varian
in genes associated with m

Acknowledgements

demonstrated the benefits of plot variants, and our results may be helpful in determining whether plot variants
in genes associated with monogenic conditions cause disease.
Acknowledgements
We thank Andrew Wood for assista potential impact of plants in an extendion scale disease.
 Acknowledgements

We thank Andrew Wood for assistance with the UK Biobank exome sequencing data, and members of

the Exeter rare variant group for helpful feedba **Acknowledgements**
We thank Andrew Wood for assistance with the UK Biobank ex
the Exeter rare variant group for helpful feedback on this work
the UK Biobank Resource under Application Number 49847 and
acknowledge the use o the Exeter rare variant group for helpful feedback on this work. This research has been conducted usin
the UK Biobank Resource under Application Number 49847 and 9072. The authors would like to
acknowledge the use of the U the UK Biobank Resource under Application Number 49847 and 9072. The authors would like to
acknowledge the use of the University of Exeter High-Performance Computing facility in carrying out this
work. This study was suppo acknowledge the use of the University of Exeter High-Performance Computing facility in carrying
work. This study was supported by the National Institute for Health and Care Research Exeter
Biomedical Research Centre. The v work. This study was supported by the National Institute for Health and Care Research Exeter
Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of
the NIHR or the Department Biomedical Research Centre. The views expressed are those of the authors and not necessarily
the NIHR or the Department of Health and Social Care. This work was supported by the Medic
Research Council [MR/T00200X/1].
Metho Biomedia Protection and Social Care. This work was supported by the Medical
Research Council [MR/T00200X/1].
Methods
Classification of variants in UK Biobank
Variants from exome sequencing (ES) data within the UK Biobank w

Research Council [MR/T00200X/1].
Methods
Classification of variants in UK Biobank
Variants from exome sequencing (ES) data within the UK Biobank were called centrally by the UK
using graphTyper²⁶. We used the Ensembl VEP Methods
Classification of variants in UK Biok
Variants from exome sequencing (E.
using graphTyper²⁶. We used the En
with their predicted functional cons Methods
Classificat
Variants f
using grap
with their
UKB due t
(VEP class Classification of variants in OK Biobank
Variants from exome sequencing (ES) da
using graphTyper²⁶. We used the Enseml
with their predicted functional conseque
UKB due to low depth. Within each prote
(VEP classification using graphTyper²⁶. We used the Ensembl VEP v104¹⁸ with the LOFTEE plugin¹⁴ to annotate the variants
with their predicted functional consequences. We excluded variants which were flagged for removal by
UKB due to lo using graphTyper^{ev}. We used the Ensembl VEP v104²⁶ with the LOFTEE plugin²⁴ to annotate the variants
with their predicted functional consequences. We excluded variants which were flagged for removal by
UKB due to low WEB due to low depth. Within each protein-coding transcript, we grouped variants into synonymous
(VEP classification "synonymous_variant"), missense ("missense_variant"), and pLoF (variants classified
MEP classification "s UKB due to low depth. Within each protein-coding transcript, we grouped variant η , and plof (variants classification "synonymous_variant"), missense ("missense_variant"), and plof (variants classification "synonymous_va \mathcal{V}^{max} synonymous_variant \mathcal{V}^{max}), missense ("missense"), and plot (variants classified variants classified variants classified variants classified variants classified variants constants constants constan

by First allele frequency. We further separated pLoF variants into SNVs, and frameshift insertions or deletions (indels). For our main analyses we considered all pLoF variants together; pLoF SNVs were used as a sensitivity deletions (indels). For our main analyses we considered all pLoF variants together; pLoF SNVs were
as a sensitivity analysis to ensure patterns were not driven by indels spanning quintiles of the gene.
Using start and end as a sensitivity analysis to ensure patterns were not driven by indels spanning quintiles of the gene.
Using start and end locations for exons within each transcript from Ensembl (downloaded on 01-04-
2022), we calculated Using start and end locations for exons within each transcript from Ensembl (downloaded on 01-04-
2022), we calculated the relative position of each variant within the coding sequence of each transc
considering only exonic 2022), we calculated the relative position of each variant within the coding sequence of each transcrienconsidering only exonic variants (i.e., excluding splice donor and splice acceptor variants). We then divided each tra considering only exonic variants (i.e., excluding splice donor and splice acceptor variants). We then
divided each transcript into quintiles, and for each class of variants (synonymous, missense, pLoF) we
calculated the nu considering only an
adivided each transcript into quintiles, and for each class of variants (synonymous, missense, pLoF) we
alculated the number of variants within each quintile as a proportion of the number of variants of

calculated the number of variants within each quintile as a proportion of the number of variants of tha
class.
Cluster analysis
We applied principal components analysis (PCA) to the proportion of variants in each quintil class.
Cluster analysis
We applied principal components analysis (PCA) to the proportion of variants in each quintile of the
gene transcript, separated by variant class. PCA loadings were calculated based on locations of
s Cluste
We ap
gene t
synon
pLoFs,
We clu Cluster analysis
We applied prince
gene transcript,
synonymous, mi
pLoFs, simulation
We clustered the
allowing seven c
of variant within

gene transcript, separated by variant class. PCA loadings were calculated based on locations of
synonymous, missense, and all pLoF variants. We then projected the PCA onto SNV pLoFs, possible
pLoFs, simulations, and ClinVa synonymous, missense, and all pLoF variants. We then projected the PCA onto SNV pLoFs, poss
pLoFs, simulations, and ClinVar variants.
We clustered the PCA profile of variants of each class within each transcript using Gaus synonymous, missense, and ClinVar variants.
ploFs, simulations, and ClinVar variants.
We clustered the PCA profile of variants of each class within each transcript using Gaussian Mixture
allowing seven clusters, trained us We clustered the PCA profile of variants.
allowing seven clusters, trained using the
of variant within each transcript we obta
inclusion in each cluster. We projected th
into the clusters to obtain their most like
for mult allowing seven clusters, trained using the profile of synonymous, missense and all pLoFs. For each class
of variant within each transcript we obtained the most likely cluster, as well as the probability for its
inclusion i of variant within each transcript we obtained the most likely cluster, as well as the probability for its inclusion in each cluster. We projected the SNV pLoFs, possible pLoFs, simulations and ClinVar variants into the clu inclusion in each cluster. We projected the SNV pLoFs, possible pLoFs, simulations and ClinVar varian
into the clusters to obtain their most likely cluster and probabilities. Seven clusters were chosen to al
for multiple c into the clusters to obtain their most likely cluster and probabilities. Seven clusters were chosen to allo
for multiple clusters with different non-uniformly distributed variants. We performed sensitivity
analyses varying for multiple clusters with different non-uniformly distributed variants. We performed sensitivity
analyses varying the number of clusters to ensure that our results were robust to the number of clusters
chosen.
Sensitivity

Possible locations of pLoF variants

For multiple clusters with different non-uniformly method is a matter of performance constrained particle in
the multiple consense.
Sensitivity analyses
Possible locations of pLoF variants
We examined the coding sequen chosen.
 Sensitivity analyses
 Possible locations of pLoF variants

We examined the coding sequence of each transcript and calculated the locations of all possible pLoF

SNVs. We clustered these in the same way as obse mood
Sensitivi
Possible
We exar
SNVs. W
patterns
pLoF SN' **Possible locations of**
Possible locations of
We examined the co
SNVs. We clustered f
patterns we identifie
pLoF SNVs. The clust
Simulations SNVs. We clustered these in the same way as observed pLoF variants (see below) to verify that any
patterns we identified were not driven by the underlying coding sequence, and the possible locations
pLoF SNVs. The clusteri

Simulations

patterns we identified were not driven by the underlying coding sequence, and the possible locatio pLoF SNVs. The clustering of these variants was compared to that of observed pLoF SNVs.
Simulations
To estimate the rate at ploF SNVs. The clustering of these variants was compared to that of observed ploF SNVs.
Simulations
To estimate the rate at which genes with a given number of variants of a given class clustered into each
cluster, we use Simulations
To estimate the rate at which genes with a given number of variants of a given class cluste
cluster, we used simulations to create synthetic sets of genes with varying numbers of var
the relative positions of a Cluster, we used simulations to create synthetic sets of genes with varying numbers of variants. We took
the relative positions of all variants in the UKB ES data, and randomly selected a number of these based
on the numbe the relative positions of all variants in the UKB ES data, and randomly selected a number of these based
on the number of pLoF variants within each gene in UKB. We repeated this 10,000 times. Simulated
genes were then clus

the number of ploF variants within each gene in UKB. We repeated this 10,000 times. Simulated genes were then clustered to estimate the number of genes falling into each cluster by chance.
ClinVar variants
We downloaded cl on the number of peak cannot the number of genes falling into each cluster by chance.
 ClinVar variants

We downloaded clinically annotated variants from ClinVar (02/10/2022) and calculated the proportiof

pathogenic/lik ClinVar variants
We downloaded clinically annotated variants from ClinVar (02/10/2022) and calculated the prop
of pathogenic/likely pathogenic pLoF variants within each quintile of each transcript and cluster
to compare to Chinvar variants
We downloaded
of pathogenic/lik
to compare to th of pathogenic/likely pathogenic pLoF variants within each quintile of each transcript and clustered then
to compare to the UK Biobank pLoF clusters.
 of pathogenic_/likely pathogenic pLOF variants within each quintie of each transcript and clusters in the UK Biobank pLoF clusters. to compare to the UK Biobank pLoF clusters.

Disease gene lists

Gene2Phenotype database to investigate whether these could elucidate the variable penetrance c
these genes. Gene lists were downloaded from https://www.ebi.ac.uk/gene2phenotype/ (accesse
06-2021) and split into those causi First discussion of these genes. Gene lists were downloaded from https://www.ebi.ac.uk/gene2phenotype/ (accessed 06-2021) and split into those causing severe developmental disorders (DD), and those causing later onset dise 06-2021) and split into those causing severe developmental disorders (DD), and those causing later
onset diseases (cancer, cardiac, eye, and skin). These were further subdivided into monoallelic
(autosomal dominant) and bi onset diseases (cancer, cardiac, eye, and skin). These were further subdivided into monoallelic
(autosomal dominant) and biallelic (autosomal recessive) genes with "absent gene product"
mechanisms; G2P genes with other inh onset and the cancer (cancer, cancer, cancer, cancer, cancer, cancer, cardiac, cardiac, cardiac, cardiac, cardiac, cardiac, cardiac, mechanisms; G2P genes with other inheritance classes or disease mechanisms were excluded. (autosomal dominant) and biallelic controller inheritance classes or disease mechanisms were exclude
mechanisms; G2P genes with other inheritance classes or disease mechanisms were exclude
the product of the product of the mechanisms; G2P genes with other inheritance classes or disease mechanisms were excluded.

-
- 1. Estimating diagnosis in panel-based genomic analysis Science in the https://www.sciencedirect.com/science/article/pii/S1098360022008176?via
2. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 i
- Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in th
Biobank. *Nature* 586, 749–756 (2020).
Klemenzdottir, E. O. *et al.* A population-based survey of FBN1 variants in Iceland reveal.
 2. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the OK
Biobank. Nature 586, 749–756 (2020).
3. Klemenzdottir, E. O. *et al.* A population-based survey of FBN1 variants in Iceland re Biobank. Nature **586**, 749–756 (2020).
Klemenzdottir, E. O. *et al.* A population
underdiagnosis of Marfan syndrome. *E*
01455-0.
Kingdom, R. *et al.* Rare genetic variants 3. Klemenzdottir, E. O. et al. A population-based survey of FBN1 variants in Iceland reveals
underdiagnosis of Marfan syndrome. *Eur. J. Hum. Genet.* 1–8 (2023) doi:10.1038/s41431
01455-0.
4. Kingdom, R. *et al.* Rare gene
- underdiagnosis of Marian syndrome. *Eur. J. Hum.* Genet. 1–8 (2023) doi:10.1038/s41431-023-
01455-0.
Kingdom, R. *et al.* Rare genetic variants in genes and loci linked to dominant monogenic
developmental disorders cause m Kingdom,
developm
Genet. 10
Gardner, 4. Kingdom, R. et al. Rare genetic variants in genes and loci linked to dominant monogented developmental disorders cause milder related phenotypes in the general population. And Genet. 109, 1308–1316 (2022).
5. Gardner, E
- developmental disorders cause innuel Telated phenotypes in the general population. Am. J. Hum.
Genet. 109, 1308–1316 (2022).
Gardner, E. J. *et al.* Reduced reproductive success is associated with selective constraint on h
- Gardner, E. J. *et al.* Reduced rep
Gardner, E. J. *et al.* Reduced rep
genes. *Nature* 603, 858–863 (20
Pizzo, L. *et al.* Rare variants in th
phenotypes in individuals carryi 5. Gardner, E. J. et al. Reduced reproductive success is associated with selective constraint on humanizier and genes. Nature 603, 858–863 (2022).
6. Pizzo, L. et al. Rare variants in the genetic background modulate cognit pizzo, L. *et al.* Rare variants in the ge
phenotypes in individuals carrying di
Genet. **21**, 816–825 (2019).
Kurki, M. I. *et al.* Contribution of rare 6. Pizzo, L. et al. Rare variants in the genetic background modulate cognitive and developmental
phenotypes in individuals carrying disease-associated variants. *Genet. Med. Off. J. Am. Coll. M.*
Genet. **21**, 816–825 (20
- phenotypes in individuals carrying disease-associated variants. Oenet. Med. Off. J. Am. Coll. Med.
Genet. 21, 816–825 (2019).
Kurki, M. I. *et al.* Contribution of rare and common variants to intellectual disability in a s Senet. 21, 816–825 (2015).
Kurki, M. I. *et al.* Contribution
of Northern Finland. *Nat. Co*
Genetic modifiers of rare va
https://www.medrxiv.org/c
- 7. Kurki, M. I. et al. Contribution of rare and common variants to intellectual disability in a sub-isolate
of Northern Finland. Nat. Commun. **10**, 410 (2019).
8. Genetic modifiers of rare variants in monogenic development
- of Northern Finland. *Nut.* Commun. **10**, 410 (2019).
Genetic modifiers of rare variants in monogenic dev
https://www.medrxiv.org/content/10.1101/2022.1
A Systematic Survey of Loss-of-Function Variants in
https://www.scien 8. https://www.medrxiv.org/content/10.1101/2022.12.15.22283523v1.
8. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes | Scier
https://www.science.org/doi/10.1126/science.1215040.
10. de Klerk, A Systematic Survey of Loss-of-Function Variants in Human Protein-Co
https://www.science.org/doi/10.1126/science.1215040.
de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcription, process
from RNA sequencing. *Trend*
- 9. A Systematic Care y of Loss-of-Function Variants in Human Protein-Coding Gener | Care | Control
https://www.science.org/doi/10.1126/science.1215040.
10. de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcription, p de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcrip
from RNA sequencing. *Trends Genet*. **31**, 128–139 (2015) 10. de Klerk, E. A. C. Alternative maritale manufation, processing, and translations. $\frac{1}{2}$. from RNA sequencing. Trends Genet. 31, 128–139 (2015).
-
- 11. Dietz, H. C., Kolakada, D., Cortazar, M. A. & Jagannathan, S. How to get away with nonsense:

Mechanisms and consequences of escape from nonsense-mediated RNA decay. WIREs RNA 11,

e1560 (2020). during splice site selection. Nat. Genet. **6**, 183–188 (1994).
Dyle, M. C., Kolakada, D., Cortazar, M. A. & Jagannathan, S.
Mechanisms and consequences of escape from nonsense-n
e1560 (2020).
The regulatory potential of up 12. Dyle, M. C., M. C., M. C., M. C., Cortaza, M. C. Coreans, M. A. B. Cortains and consequences of escape from nonsense-mediated RNA decay. WIREs RNA 1
13. The regulatory potential of upstream open reading frames in eukar
- Mechanisms and consequences of escape from nonsense-mediated RNA decay. WMES RNA 11,
e1560 (2020).
The regulatory potential of upstream open reading frames in eukaryotic gene expression Weth
- 2014 WIREs RNA Wiley On /
The regulatory
- 2014 - WIREs
https://wires.c
Karczewski, K. 13. The regulatory potential of upstream open reading frames in europe gene expression of tensors

13. The regulatory of upstream open reading frames in europe since the set of upstream of the reading frames in the set of

- https://wires.onlinelibrary.wiley.com/doi/
Karczewski, K. J. *et al.* The mutational cons
humans. *Nature* **581**, 434–443 (2020).
Lek, M. *et al.* Analysis of protein-coding gel
- Karczewski, K. J. *et al.* The mutational constraint spectrum quar
humans. *Nature* **581**, 434–443 (2020).
Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,7
(2016). 14. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456
humans. Nature 581, 434–443 (2020).
15. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. humans. Nature 581, 434–443 (2020).
Lek, M. *et al.* Analysis of protein-codin_!
(2016).
Samocha, K. E. *et al.* Regional missense
148353 Preprint at https://doi.org/10.
- 15. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. bioMxiv 536, 030336
(2016).
16. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction.
148353 Pre (2009).
Samoch
148353
Havrilla_.
the hun
- 16. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction.
148353 Preprint at https://doi.org/10.1101/148353 (2017).
17. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A
the human genome. Nat. Genet. 51, 88–95 (2019).
McLaren, W. et al. The Ensembl Variant Effect Predictor. Ger
Thormann, A. et al. Flexible and scalable diagn
-
- 11. Havrida, J. M., Pedersen, B. S., Layer, R. M. M. S., Landing, R. M. Andre L. Lawer, C. S., Layer, R. A.
18. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122 (2016).
19. Thormann, A. *et* the human genome. Nat. Senet. 31, 88–95 (2019).
McLaren, W. *et al.* The Ensembl Variant Effect Pred
Thormann, A. *et al.* Flexible and scalable diagnostic
Ensembl VEP. Nat. Commun. **10**, 2373 (2019).
ClinVar: improving ac 19. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using (Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
20. ClinVar: improving access to variant interpretations and supporting evidenc
- 19. Thormann, A. et al. Flexible and scalable diagnostic intering of genomic variants asing G2P with
Ensembl VEP. Nat. Commun. 10, 2373 (2019).
20. ClinVar: improving access to variant interpretations and supporting eviden
- Ensembl VET. Mat. Commun. 10, 2373 (2013).
ClinVar: improving access to variant interpreta
Research | Oxford Academic. https://academi
The GTEx Consortium atlas of genetic regulato
https://www.science.org/doi/10.1126/scien 20. Clinton inpressing access to variant interpretation and supporting entantly parameter interpretations.

Research | Oxford Academic. https://academic.oup.com/nar/article/46/D1/D1062/4641904.

21. The GTEx Consortium atl The GTEx Consortium atlas of genetic regulatory effects across human tissues | Science.
https://www.science.org/doi/10.1126/science.aaz1776. 21. The GTEx Construction at the GTEx Construction at the state in the state of actions https://www.science.org/doi/10.1126/science.aaz1776. https://www.science.org/doi/10.1126/science.aaz1776.
-
- 22. Chia, C. Y. et al. GATA6 Cooperates with EOMES/SMAD2/3 to Deploy the Gene Regulatory Network
Governing Human Definitive Endoderm and Pancreas Formation. *Stem Cell Rep.* 12, 57–70 (2019).
23. Biochemical features of pr Biochemical features of primary cells from a pediatric patient with a gain-of-function ODC1 genetic
mutation | Biochemical Journal | Portland Press. https://portlandpress.com/biochemj/article-
abstract/476/14/2047/219620/B 22. Biochemical features of primary cells from a gain of primary cells from a pediatric patient of primary cells from a stract/476/14/2047/219620/Biochemical features of primary cells from a stredirected From a stredirecte
-
- Schoeler, T. *et al.* Participati
analyses. *Nat. Hum. Behav.*
Rots, D. *et al.* Truncating SR
distinct neurodevelopmenta 24. Schoeler, T. et al. Participation bias in the OK Biobank distorts genetic associations and downstream
analyses. Nat. Hum. Behav. 7, 1216–1227 (2023).
25. Rots, D. et al. Truncating SRCAP variants outside the Floating-H analyses. *Nut. Hum. Behav. 1*, 1216–1227 (2023).
Rots, D. *et al.* Truncating SRCAP variants outside t
distinct neurodevelopmental disorder with a spec
Genet. **108**, 1053–1068 (2021).
Eggertsson, H. P. *et al.* GraphTyper 25. Rots, D. et al. Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a
distinct neurodevelopmental disorder with a specific DNA methylation signature. Am. J. Hum.
Genet. 108, 1053–1068 (2021).
26.
- denet. 108, 1053–1068 (2021).
Genet. 108, 1053–1068 (2021).
Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variatio
using pangenome graphs. *Nat. Commun.* 10, 5402 (2019). Genet. 108, 1053–1008 (2021).
Eggertsson, H. P. *et al.* GraphTy
using pangenome graphs. *Nat.* (26. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping or structural variation
using pangenome graphs. Nat. Commun. 10, 5402 (2019). using pangenome graphs. Nat. Commun. 10, 5402 (2019).

 \overline{a}

For each class of variant (synonymous, missense, pLoF) we present the relative properties variants of that class were included in each of the 6 clusters where variants seventh cluster identifying genes with no variants of

Figure 4: Proportions of genes in cluster 4-6 in UKB compared with simulations.

The proportion of genes where pLoFs in OK Biobank are clustered into clusters 4-6 against the number of pLoFs in that gene is shown in blue. The proportion of simulated genes where pLoFs clustered into clusters 4-6 against the number of variants in the gene is shown in red. The green line shows the number of genes with each number of pLoFs in UK Biobank.

It is made available under a CC-BY 4.0 International license. **(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/10.1101/2023.10.11.23296535;](https://doi.org/10.1101/2023.10.11.23296535) this version posted October 12, 2023. The copyright holder for this preprint

rigue 5: Proportion of genes with pLot 3 in each cluster, including subsets of disease genes, locations of
possible pLoFs, and simulation analyses.
The relative proportion of genes where pLoFs are included in each of the s The relative proportion of genes where p
pLoF variants is shown for different sets genes with at least five pLoFs); AD-DD G2
through haploinsufficiency); AR-DD G2P genessive mechanisms); other AD G2P genes cancer syndromes pLoF variants is shown for different sets of genes, locations of possible pLoFs, and simulations: All (all
genes with at least five pLoFs); AD-DD G2P genes (genes where pLoFs cause developmental delay
through haploinsuffic present with at least five pLoFs); AD-DD G2P genes (genes where pLoFs cause developmental delay
through haploinsufficiency); AR-DD G2P genes (genes where pLoFs cause developmental delay throug
recessive mechanisms); other genes where ploFs cause developmental delay the plot is cause the plot of the plot is cause developmental delay the
recessive mechanisms); other AD G2P genes (genes where ploFs cause adult onset diseases, inclu
cancer synd through the matrix of the AD G2P genes (genes where pLoFs cause adult onset diseases, including
cancer syndromes and heritable cardiac, eye or skin conditions); other AR G2P genes (genes where
pLoFs cause adult onset disea cancer syndromes and heritable cardiac, eye or skin conditions); other AR G2P genes (genes where
pLoFs cause adult onset diseases through recessive mechanisms); genes with high probability of LoF
intolerance (pLl¹⁵) sco ploFs cause adult onset diseases through recessive mechanisms); genes with high probability of Lol
intolerance (pLl¹⁵) scores >0.9; possible LoF variants based on the underlying sequence of each gene
simulations of all

It is made available under a CC-BY 4.0 International license. **(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/10.1101/2023.10.11.23296535;](https://doi.org/10.1101/2023.10.11.23296535) this version posted October 12, 2023. The copyright holder for this preprint

Figure 6: Profiles of variants of each class within selected genes

Locations of variants of each class in UK Biobank individuals, and ClinVar pathogenic/likely pathogenic
variants in COL4A3, TP63, ARID1B, NSD1, GATA6, and ODC1 are shown. The top panel of each figure variants in COL4A3, TP63, AMD 1B, NSD1, GATA6, and ODC1 are shown. The top panel of each figure shows a frequency density plot of the relative position of variants of each class in UKB, plus ClinVar pathogenic/likely pathogenic variants. The middle panels show rug plots of the relative positions of each variant of each class in separate panels. The bottom panels show the locations of start codons, and a diagram of either the relative positions of domains within the protein (for COL4A3, GATA6, ODC1), or a depiction of the exons included in the labelled transcript (TP63, ARID1B, NSD1). The location of the final exon is indicated by the dark bar above the transcript diagram.

