

## ARTICLE OPEN



# Populational pan-ethnic screening panel enabled by deep whole genome sequencing

Linfeng Yang<sup>1,2,7</sup>, Zhe Lin<sup>1,2,7</sup>, Yong Gao<sup>1,2</sup>, Jianguo Zhang<sup>1,2</sup>, Huanhuan Peng<sup>2</sup>, Yaqing Li<sup>2</sup>, Jingang Che<sup>2</sup>, Lijian Zhao<sup>2,3</sup>✉ and Jilin Zhang<sup>4,5,6</sup>✉

Birth defect is a global threat to the public health systems. Mitigating neonatal anomalies is hampered by elusive molecular mechanisms of pathogenic mutations and poor subsequent translation into preventative measures. Applying appropriate strategies in China to promote reproductive health is particularly challenging, as the Chinese population comprises complex genomic diversity due to the inclusion of many ethnic groups with distinct genetic backgrounds. To investigate and evaluate the feasibility of implementing a pan-ethnic screening strategy, and guide future reproductive counselling, high-quality variants associated with autosome recessive (AR) diseases derived from the largest publicly available cohort of the Chinese population were re-analysed using a bottom-up approach. The analyses of gene carrier rates (GCRs) across distinct ethnic groups revealed that substantial heterogeneity existed potentially due to diverse evolutionary selection. The sampling population, sequencing coverage and underlying population structure contributed to the differential variants observed between ChinaMAP and the East Asian group in gnomAD. Beyond characteristics of GCR, potential druggable targets were additionally explored according to genomic features and functional roles of investigated genes, demonstrating that phase separation could be a therapeutic target for autosomal recessive diseases. A further examination of estimated GCR across ethnic groups indicated that most genes shared by at least two populations could be utilised to direct the design of a pan-ethnic screening application once sequencing and interpreting costs become negligible. To this end, a list of autosomal recessive disease genes is proposed based on the prioritised rank of GCR to formulate a tiered screening strategy.

*npj Genomic Medicine* (2023)8:38; <https://doi.org/10.1038/s41525-023-00383-8>

## INTRODUCTION

The survival rate of children remains a global concern. Although the under-five mortality rate decreased to 3.8% in 2021 from 10% in 1990, a total of five million children under five years old yet have died in 2021, mainly due to congenital anomalies (<https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-under-5-mortality-in-2020>). The congenital anomaly, covering a diverse group of disorders that single gene mutations can cause, large structural variations on chromosomes, and other environmental factors, is one of the leading causes of lethality apart from prematurity and infectious diseases<sup>1–3</sup>. Except for a few genetic abnormalities, such as Down's Syndrome and cystic fibrosis, the causative mechanisms of many birth defects remain poorly understood, including congenital heart defects, cleft palate and club foot. Many causative variants have been characterised and catalogued, further reinforced by large sequencing projects, including the 1000 genome project<sup>4</sup>, the UK Biobank cohort<sup>5</sup> and the Taiwan Biobank project<sup>6</sup>. Despite the enormous translational potential of identified variants, converting disease-associated variants with intrinsic populational diversity into preventative applications remains challenging, and it still undergoes a primitive exploring phase, especially in preventing congenital anomalies caused by genetic mutations.

Regardless of poorly elucidated underlying pathogenic mechanisms, numerous severe genetic anomalies can be prevented by imposing global or regional surveillance systems and installing

preventative carrier screening programs<sup>7–9</sup>. As an effective preventive strategy, the capacity of widely adopted extended carrier screening (ECS) panels has been drastically advanced by the fast-evolving, high-throughput next-generation sequencing techniques, enabling increased access to genetic risk assessment. Indeed, most available screening panels can now focus on 100–200 genetic diseases<sup>10,11</sup>, except for a few larger panels claiming to examine disease-causing variants in 500–600 genes. In addition, because genomic compositions across distinct populations exhibit substantial diversity, carrier frequencies of Mendelian disorders display population-specific characteristics<sup>12</sup>. Thus, many countries or regions have adopted tailored screening panels based on disease prevalence to promote reproductive health, including Mackenzie's Mission<sup>8</sup> and Victorian program<sup>13</sup> in Australia and hemoglobinopathies project in China<sup>9</sup>.

However, one major shortcoming of these ECS schemes is that their screening capacity is restricted because of including a limited number of severe genetic disorders caused by protein-coding variants at a high national-wide or ethnic-specific prevalence<sup>10,11</sup>. The remaining variants and structural variations associated with diseases at a lower prevalence and uncommon/rare diseases are barely considered during the reproductive consultation. Unavoidably, the capacity and scope of these screening panels are yet largely limited and primarily due to the economic consideration regardless of the consideration of population heterogeneity, except for panels to target diseases with high prevalence in the

<sup>1</sup>Hebei Industrial Technology Research Institute of Genomics in Maternal and Child Health, BGI-Shijiazhuang Medical Laboratory, Shijiazhuang, China. <sup>2</sup>BGI Genomics, BGI-Shenzhen, Shenzhen, China. <sup>3</sup>Medical Technology College of Hebei Medical University, Shijiazhuang, China. <sup>4</sup>Tung Biomedical Sciences Centre, City University of Hong Kong, Hong Kong S.A.R, China. <sup>5</sup>Department of Precision Diagnostic and Therapeutic Technology, The City University of Hong Kong Shenzhen Futian Research Institute, Shenzhen, China. <sup>6</sup>Department of Biomedical Sciences, City University of Hong Kong, Hong Kong S.A.R, China. <sup>7</sup>These authors contributed equally: Linfeng Yang, Zhe Lin. ✉email: zhaolijian@bgi.com; jilin.zhang@cityu.edu.hk

Ashkenazi Jewish population<sup>13</sup>. Moreover, multiple lines of evidence have demonstrated that the gene carrier rate (GCR) of many diseases causing genetic variants shall be considered when implementing the screening strategy. The polyethnic nature of Chinese population calls for a strategical utilisation<sup>14,15</sup> toward mutation spectrum and ethnic-biased prevalence to deliver integrative reproductive consultation with precision and accuracy.

Nonetheless, the consequences of most observed non-coding variants remain poorly explained. Failing to reveal the underlying molecular mechanisms of their regulatory roles in diseases hinders the search for drug targets, thus the subsequent design of therapeutic strategies. Increasingly reported evidence has proven that non-coding variants play essential roles in pathogenesis due to the dysregulated transcriptional and post-transcriptional processes<sup>16,17</sup>. Since molecular interactions are often compartmentalised, one of the fundamental processes gaining heavy focus, particularly in pathogenesis, is the condensate formation that drives the liquid-liquid phase separation (LLPS)<sup>16,17</sup>. Whether autosomal recessive (AR) disease genes associated with LLPS could be druggable targets is still poorly investigated.

Due to the increased throughput and the drastically decreased cost of sequencing, the turning point of utilising deep whole genome sequencing (WGS) data for personalised and precision reproductive counselling is right on edge. It is now feasible and desirable to use the whole genome data of individuals to expand our capacity to explore the variants from large cohorts to provide a better risk assessment of genetic conditions during reproductive counselling. Indeed, the ECS for prenatal intervention or newborn screening to prevent neonatal disorders have been introduced in many countries<sup>18</sup>, and new tools or approaches to estimate gene carrier rate based on large datasets are emerging<sup>19</sup>. In this study, we leverage the power of ChinaMAP generated through a natural cohort with deep WGS data to demonstrate the benefit of WGS. By comparing GCR estimated from distinct populations, we report the heterogeneous spectra of variants associated with various diseases at a low prevalence in China. More interestingly, analyses of the characteristics of AR disease-causing variants further reveal feasible implementations of tiered pan-ethnic panels considering regional prevalence or rare disease subsets for precision reproduction counselling.

## RESULTS

### Causative variants of autosomal recessive diseases

To investigate the characteristics of GCR in the Chinese population, we re-analysed variants included in ChinaMAP covering the deep WGS of 10,588 individuals<sup>14</sup>. After quality control, we extracted a total of 140,109,159 variants. To uncover the signature of GCR, 2904 phenotypes corresponding to 2464 AR genes exhibiting Mendelian inheritance were extracted from Online Mendelian Inheritance in Man (OMIM) as candidate genes, where genes associated with syndromes caused by large segmental duplications or chromosomal variations were discarded. The 12.7 million candidate gene overlapping variants were extracted, including 11,826,063 SNPs and 898,110 indels. These variants were subjected to a carefully designed workflow to classify variants as deleterious (Fig. 1a and Supplementary Fig. 1, see “Methods”). We then estimated GCR based on these deleterious variants. Meanwhile, the disease-associating variants were categorised into groups depending on whether treatments were available for the associated conditions/diseases.

We identified 19,484 deleterious variants associated with AR diseases. Among these, 3409 protein-coding variants were annotated by ClinVar database. Except for 19 manually curated variants, 14,775 and 1281 were predicted by our workflow as deleterious missense and deleterious nonsense, respectively (Fig. 1a). Compared to the previously estimated GCR<sup>19</sup>, the GCR

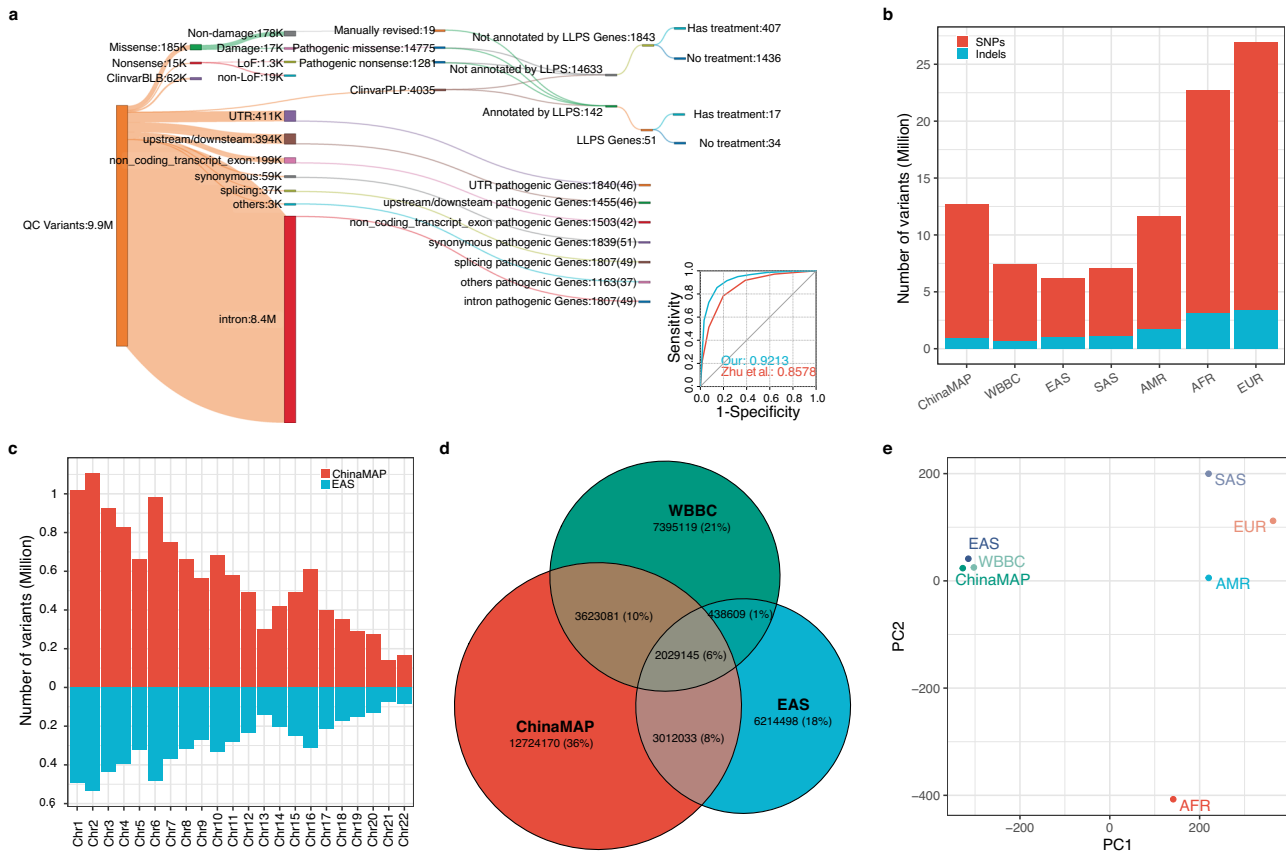
reported in our study exhibited a moderate correlation, potentially caused by the intrinsic difference of variants selection using different strategies, cutoffs and prediction tools (Supplementary Figs. 2 and 3). After a systematic comparison, we confirmed that our strategy performed better in predicting P/LP variant as indicated by AUC-ROC (Fig. 1a), and that predicting tools and corresponding thresholds used in this study were strongly evidence-based and widely used, leading to a theoretically lower false discovery rate and a less biased variant set (Supplementary Table 1 and Supplementary Fig. 4). Indeed, GCRs reported here were more consistent with previous designs (Supplementary Fig. 5)<sup>20</sup>. Interestingly, 142 coding variants within 51 AR disease genes were annotated to be involved in LLPS. Among these variants, 65 (53.72%) were catalogued in ClinVar, 55 (45.45%) were deleterious missense variants, and only one was LoF. In contrast, 8.4 million non-coding variants, accounting for 92% of the non-coding variants, were enriched in the intronic regions of 1807 AR disease-causing genes. Unexpectedly, 49 genes were predicted to participate in LLPS. This suggested a promising yet not well-studied direction to disclose the affected regulatory processes beyond misfolded proteins.

As protein-coding sequences only comprised a small portion of the entire genome and many proteins are not targetable by small molecules, we retrieved non-coding variants associated with clear disease-causative genes to explore their potential functional characteristics. Surprisingly, 5.01% and 4.80% of these non-coding variants were enriched within the untranslated regions and up-/downstream of 1891 well-characterised disease-causative genes, respectively. Approximately 36,655 non-coding variants, accounting for 0.45% of the genes associated non-coding variants, were particularly enriched at splicing junctions. Around 45% of these variants were canonical splicing sites, and hundreds of sites were validated by the expression profiles in GTEx (102) (Supplementary Fig. 6) and supported by ICGC (802). However, only 1908, accounting for 5.2% variants, were predicted as splicing-altering sites by SpliceAI (438 acceptor gain, 315 acceptor loss, 269 donor gain, 886 donor loss)<sup>21</sup>. More interestingly, non-coding variants were more enriched in disease-causative genes than non-OMIM genes (Wilcoxon test,  $p < 2.2e-16$ ). These observations implied that post-transcriptional regulation was deeply involved in the pathogenesis apart from that caused by altered protein products.

WBBC and gnomAD were used to assess the power of ChinaMAP data. Indeed, ChinaMAP detected more than 12 million single nucleotide polymorphisms (SNPs), which was over two-fold on average of that for each chromosome in the East Asian ethnic group (EAS) reported by gnomAD (Fig. 1b, c), thus enabling the generation of a much broader and more comprehensive view of the Chinese population. A further comparison of the mutation spectra between ChinaMAP, WBBC and EAS demonstrated that only 43% of gnomAD EAS and 42% of WBBC variants were shared with ChinaMAP, respectively, indicating potentially distinct genetic compositions among datasets (Fig. 1d). Further analysis of allele frequency confirmed that such minor discrepancy did not affect the separation of distant populations (Fig. 1e).

### Estimation of gene carrier rate enhanced by deep WGS in East Asian population

ChinaMAP covered many diverse minority groups in China, presenting the most comprehensive variations of the East Asian population to date. A substantial number of variants were not detected in the gnomAD EAS population due to technical limits or smaller sampling sizes (Figs. 1b and 2a). Thus, exploiting the dataset to obtain a systematic comparison of GCR was carried out to demonstrate the necessity of WGS to gain improved insight into reproduction counselling. Despite the slightly distinct GCR across populations, many genes of which GCR  $\geq 1/500$  in different



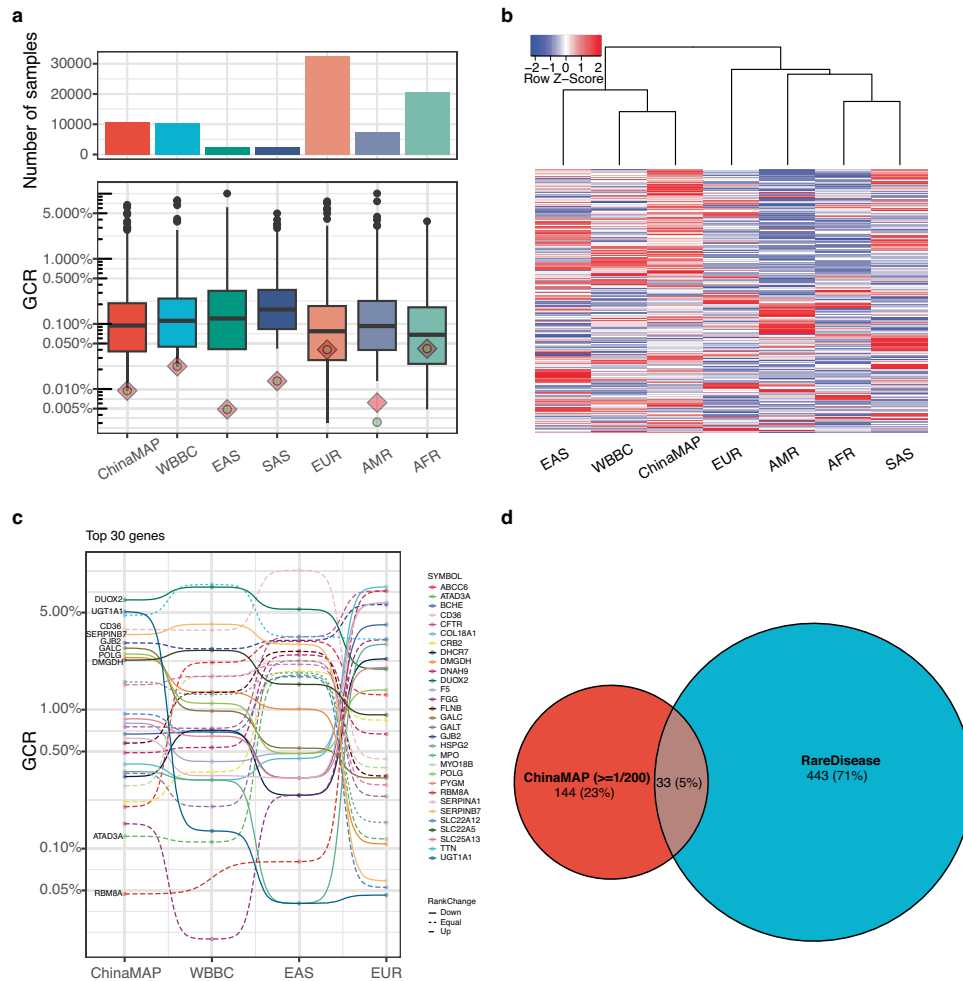
**Fig. 1** Distribution of variants and identification of AR disease-causing variants. **a** A schematic flow of classifying variants extracted from ChinaMAP. Variants that passed the initial QC are subjected to downstream annotation depending on the nature of genomic loci (protein-coding, intron, splicing junctions, UTR, etc.). The annotated pathogenic variants in protein-coding regions are further divided into the following sub-categories: treatable, untreatable, either associated with LLPS or not, where the number of corresponding genes is shown. To compare the prediction accuracy of missense variants with Zhu et al.'s method, all P/LP and B/LB missense variants extracted from ClinVar (v20221113) are taken as true sets to generate AUC-ROC plot by the R package pROC taking the number of predict tools as threshold. **b** Comparisons of the count of passed-QC variants in distinct ethnic groups included in ChinaMAP and gnomAD databases. **c** The detailed comparison of variant numbers across chromosomes between gnomAD and ChinaMAP. **d** Shared SNPs of East Asian populations between ChinaMAP, gnomAD and WBBC. **e** Principal component analysis based on allele frequency of all studied populations.

ethnic groups were shared by at least two populations (Fig. 2b and Supplementary Fig. 7), suggesting an ancestral origin that potentially underwent selections<sup>22</sup>. However, the GCR of *DMGDH*, *CD36* and *GJB2*, the variants of which were linked to diseases Dimethylglycine dehydrogenase deficiency, Platelet glycoprotein IV deficiency, and Deafness (autosomal recessive 1A), respectively, exhibited substantial variations across ethnic groups. For instance, variants on gene *CD36* that could cause Platelet glycoprotein IV deficiency were ranked at the top in East Asian groups. In contrast, its observed GCR was drastically lower in several other groups, confirming that diversity caused by potential population-specific selections existed between some groups (Fig. 2c). However, the rank of many genes remained unchanged in other ethnic groups compared to ChinaMAP. In addition, numerous genes with highly ranked GCR were also observed at the leading ends within several other populations, confirming that AR disease genes could commonly affect many ethnic groups other than the non-negligible diversity (Fig. 2b and Supplementary Fig. 7b).

By comparing GCRs estimated from these datasets, we found that the sequencing depth (30X for ChinaMAP, 18X for gnomAD v3) and sample size affected the estimation of GCR in several manners (Fig. 2a). The sample size dictated the detection power, while the under-detected variants caused by insufficient sequencing depth could be compensated by increasing the sampling size<sup>23</sup>. We speculated that the partial congruence with previously estimated GCR of EAS in gnomAD was due to the insufficient

detection of variants and potentially false claim of ethnicity. To prove this, we performed PCA analysis on the ChinaMAP, EAS of gnomAD and WBBC at allele frequency level (Fig. 1e) and GCR correlation test between distinct ethnic groups (Supplementary Fig. 8 and Supplementary Table 11) to estimate the impact of low-sequencing coverage and insufficient recovery of rare variants. Indeed, approximately 3.6 million SNPs detected in both ChinaMAP and WBBC were missing in gnomAD, a much larger number compared to that between gnomAD and these two datasets, implying that a lower detection rate of rare variants in the EAS group of gnomAD cohort (Fig. 1d and Supplementary Fig. 9).

After intersecting the selected genes with a list of rare diseases compiled by the Chinese Center for Disease Control and Prevention, we found very few overlapping genes, suggesting that many of the rare diseases lacked appropriate genetic dissections. Alternatively, this potentially indicated the inflation of GCR due to either the carrier effect or heterogeneous genetic composition that could impact the estimation of GCR (Fig. 2c). ChinaMAP uniquely enabled the recovery of genes with higher GCR (Supplementary Fig. 10), confirming that EAS in gnomAD had a distinct genetic background compared to ChinaMAP. We also observed that the estimated GCR of treatable and untreatable diseases across several ethnic groups surprisingly differed significantly (Supplementary Fig. 11). This reinforced the necessity of introducing reproductive counselling for preventative purposes,



**Fig. 2 Unique characteristics of GCR estimated from ChinaMAP.** **a** Distribution of GCR estimated from all investigated public datasets. Bottom: The distribution of GCR estimated from clear AR disease-causing variants; The green dot and red square represent the GCR that covers 99% and 95% of studied genes, respectively. Top: The corresponding sample size of each ethnic group in the studied databases. **b** Hierarchical clustering of genes with  $GCR \geq 1/500$  in ChinaMAP are selected to compare with corresponding GCRs in other studied populations. **c** Top 30 genes with drastic GCR changes, the order of which is determined based on the maximum pairwise distance of GCR between populations for a given gene, are selected to demonstrate ethnic-specific GCR features (solid: higher GCR in the East Asian group; dashed-line: lower GCR in the East Asian group). Additionally, the relative rank change of selected genes compared to ChinaMAP is indicated by line type. **d** The overlap of gene lists between ChinaMAP-derived AR disease-causing genes and genes associated with rare diseases officially listed in China.

allocating resources to find novel drug targets, and carrying out innovative research for rare diseases.

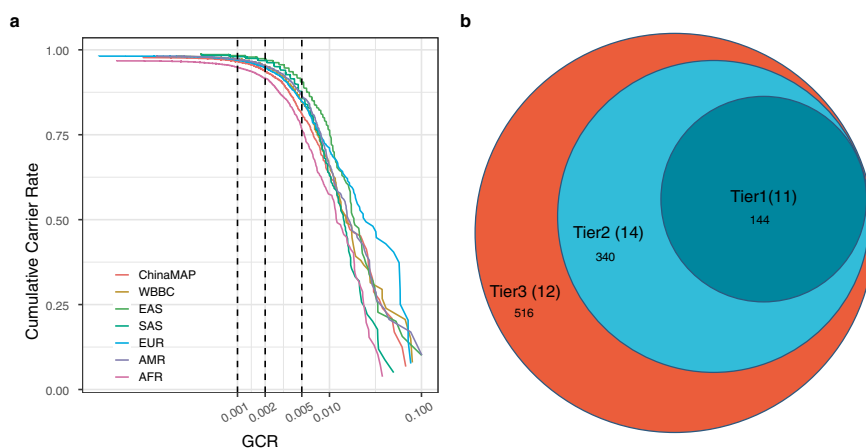
### Discovery of potential druggable targets for rare diseases

Although responsible genetic variants of a few neonatal conditions were characterised, many conditions lacked approved treatments, including effective drugs and clinical therapies (Fig. 1a). Most rare diseases are not treatable, calling for unavoidable attention to explore potential druggable targets for these untreatable diseases besides implementing preventative measures.

In addition, the prevalence of certain diseases was primarily dictated by population-specific genetic compositions. Most affected genes varied across the population. For example, genes *UGT1A1*, *SERPINB7* and *ABCG5* were responsible for Crigler-Najjar syndrome, palmoplantar keratosis, phytosterolaemia, exhibited significantly lower GCR in non-Asian ethics compared to that of the Asian ethnic group, suggesting heterogeneous GCR at the population level needed to consider a population-centric priority for the variable genes when implementing preventative strategies or diagnostic strategies<sup>6,12,24</sup>. For investigation, both Asian groups

seemed to be more susceptible to diseases Platelet glycoprotein IV deficiency, Lissencephaly 5, Deafness, autosomal recessive 111, Dyssegmental dysplasia, Silverman-Handmaker type/Schwartz-Jampel syndrome, type 1 and Citrullinemia, adult-onset type II / Citrullinemia, type II, neonatal-onset which were caused by mutations in *CD36*, *LAMB1*, *MPZL2*, *HSPG2* and *SLC25A13* supported by a slightly higher GCR compared to other ethnic groups<sup>12,15</sup>, restating the importance of strategic and peculiar drug development to target affected population (Supplementary Fig. 12).

Growing evidence indicated that LLPS participated in various regulatory processes by forming membrane-less organelles in cells, including transcriptional and translational dysregulation in pathogenesis<sup>16,17,25</sup>. As diseases caused by mutations within the protein-coding genes only account for a minor portion of the polymorphism, and many roles of non-coding variants were not examined thoroughly due to their poorly investigated functional roles, we expanded the search scope to identify non-coding variants associated with well-annotated pathogenic genes and potentially dysregulated condensate alterations during LLPS. The rationale was that intrinsically disordered regions pervasively



**Fig. 3** Design of pan-ethnic screening panel. **a** The cumulative carrier rate (CCR) of the top 2000 AR disease-causing genes is visualised. The thresholds of GCR used to select genes are indicated by grey dashed lines. **b** The proposed screening panel with three tiers, the number of included genes (LLPS associated) are chosen based on suggested cutoff 1/200 and 1/500 GCR, and the top 1000.

existed in protein sequences and played essential roles in various biological processes through regulating non-membrane organelles. In total, 265,009 variants, accounting for 3.23% of the functionally uncharacterised variants, were involved in LLPS by querying against the list of predicted LLPS associating variants<sup>16</sup>, implying the direction of future drug design and exploration. Indeed, mutations in myosin VIIA and filaggrin could impair their ability to form condensates, leading to subsequent dysregulation of forming motor protein clusters in stereocilia<sup>25</sup> and assembling keratohyalin granule in keratinocytes, respectively<sup>26</sup>. However, the GCR distribution was not significantly different across populations (Supplementary Fig. 13). A closer examination of the genes that could be impacted by dysregulated LLPS further revealed that at least 60% of diseases without available treatments were associated with the non-membrane organelle forming biomolecules predicted by previously reported approaches<sup>16</sup> (Supplementary Fig. 14a). The functional enrichment of genes involved in LLPS indicated that genes were associated with extracellular matrix related functions. By performing GO enrichment of 14,487 non-coding variants adjacent to protein-coding genes (UTRs and splicing regions, see “Methods”), we also found that non-coding variants most likely affected genes (83 genes in the top 10 GO) that were involved in amino acid metabolic processes in addition to sensory perception, indicating a potential direction, such as secretome, to search for drug targets. Additionally, we observed that 265 genes fell into the non-treatable category (Supplementary Fig. 14b).

### Rescale pan-ethnic preventative intervention

Finally, we leveraged the power of the WGS-based mutation profile to attempt to exploit its potential in guiding the implementation of preventative measures. The cumulative GCR across distinct populations was surveyed, and it was clear that the number of highly ranked AR-diseases-causing genes were restrained within a certain range regardless of the population-specific genomic composition (Fig. 3a), suggesting that these genes were essential and were potentially selected though they have undergone diverse selection. Based on such observation, genes with a GCR threshold over 1/200 or 1/500 were examined to impute the panel design strategy considering the theoretical prevalence of a disease.

We found that only 20 AR disease-causing genes were exclusively detected by ChinaMAP with a threshold GCR > 1/200 (Supplementary Fig. 7), and all genes prioritised by ChinaMAP were unexpectedly shared with other ethnic groups when the threshold was lowered to GCR > 1/500, potentially because of the

aggregated allele frequency of multiple ethnic groups. The comparison of GCRs derived from differential genes and their rankings (Supplementary Fig. 12 and Supplementary Fig. 10) revealed that such difference was likely due to the sampling or intrinsically distinct population structures in the studied populations. These genes showed significantly higher GCR than those in European (EUR) and African (AFR) populations.

Next, a strategy with a three-tier design was proposed to optimise the practical feasibility and rationale (Fig. 3b). Tier 1 and Tier 2 included 144 and 340 genes showing 1/200 and 1/500 GCR in the Chinese population, respectively. However, when the top 1000 genes of each population were extracted for comparison, over 83.26% of AR disease genes were shared by at least two populations (Supplementary Fig. 7c), restating the population heterogeneity could be circumvented once sequencing cost became a minor factor to be considered. The genes in Tier 2 have included genes to be screened during the carrier screening. Indeed, 55% to 80% of genes formed by previously reported panel designs were covered by the proposed panel at Tier 2 (Supplementary Fig. 15 and Supplementary Table 8). However, AR diseases caused by copy number variations, such as in *HBA1/HBA2*, *DMD*, *SMN1*, *F8*, *F9*, *IDS*, *MTM1*, *GLA*, *IL2RG* and *OTC* etc., were excluded in our current analysis. Such genes could be selectively supplemented to augment proposed panels based on prevalence in the population.

For genes included in Tier 3, although the theoretical prevalence of their associated AR diseases was much lower in the current Chinese cohort, they could be considered as a complementary list for couples to consider during reproductive counselling and expand the capacity carrier screening panel to a pan-ethnic panel once sequencing cost became neglectable and more reliable medical interpretation could be accessed through advanced or specialised artificial intelligence.

### DISCUSSION

ChinaMAP is one of the largest publicly accessible cohorts providing relatively comprehensive deep WGS-based variant spectra of the Chinese population to understand the heterogeneity, which could potentially guide the implementation of strategies to prevent neonatal anomalies and assist reproductive counselling.

The approach developed in this study aims to assess the feasibility of introducing a pan-ethnic screening strategy focusing on the Chinese population. A thorough comparison with existing approaches has yet to be carried out. However, the correlation

between gnomAD EAS and ChinaMAP is at the same level as previously reported observations<sup>19</sup>. An in-depth comparison between approaches demonstrates that GCRs of most genes estimated by ClinVar P/LP showed a high correlation between studies except for a few genes with variants that received major revision of allele frequency in ClinVar. The deviation of GCR could be introduced by several factors, including distinct input data, database revision, variant selection strategy, and heterogeneous strategies applied to predict missense and loss-of function variants. The primary source of discordant GCR estimation in ChinaMAP originates from the variant prediction procedure, where a more stringent strategy in this study has been employed by selecting 100% concordant prediction of nine tools instead, leading to the exclusion of ambiguous variants (Supplementary Fig. 3). Appropriately prioritising genes associated with rare diseases at low prevalence needs to integrate epidemiological information, which is the major limitation once sequencing cost drops to a negligible level.

The substantial difference in estimated GCR between gnomAD and ChinaMAP could be caused by the insufficient sampling bias of the East Asian group included by gnomAD or the heterogeneous compositions of individuals surveyed in ChinaMAP. The sample size of ChinaMAP dataset is four-fold higher than the EAS in gnomAD, which generates the deep WGS data of more than 10 thousand Chinese individuals, granting a more comprehensive capture of the variant spectra. Additionally, the sequencing approach is another potential source of incongruence between datasets as many variants are non-coding variants in ChinaMAP. The higher sequencing coverage used by the ChinaMAP project has a theoretically better capability to detect uncommon variants. It is well-known that China has a complex population structure that involves 55 minority groups. Thus, intrinsic sampling bias could be another source. However, the difference could not be resolved in our study as access to the allele frequency of ChinaMAP individuals was limited. To gain a holistic picture with better socioeconomic benefit, the future screening based on WGS with at 30X shall be considered once the sequencing cost is not a determinant.

The heterogeneous population in China raises concerns on data reusability and translational potential. At the 1/500 GCR threshold, the heterogeneity counterintuitively has refined the power of the screening panel, potentially due to aggregated allele frequency has eliminated population-specific bias caused by selection. Although two-thirds of the genes associated with LLPS have been detected for potential druggable targets, accounting for only a tiny proportion of untreatable diseases, these proteins could be included in the future direction of drug development. One functional study has demonstrated that mutations in *MYO7A* weakens its ability to form *MYO7A/USH1C/USH1G* complex which impairs phase separation, resulting in an abnormal tip-link densities and causing hearing and vision loss in Usher syndrome patients<sup>25</sup>. Another example is that mutations in flaggrin (*FLG*), which are associated with human skin barrier disorders, could alter the properties of proteins leading to an abolished formation of keratophyalin granules that can subsequently compromise skin defence<sup>26</sup>, although the role of mutated *FLG* in AR is yet to be characterised. Nevertheless, many exciting discoveries related to LLPS remain to be clarified with substantial research on the mutation affected LLPS and require clear dissection of corresponding mechanistic insights. More innovative approaches will be invented to target these variants to find solutions for rare diseases<sup>27</sup>. The resulting knowledge from our analysis could relieve the socioeconomic burdens and, more importantly, benefit families carrying untreatable rare diseases. The detailed molecular mechanism, however, requires more experiments to disclose the altered functional roles in cells, which is beyond the scope of this study.

Finally, interpreting the present screening results of ECS to assist reproduction counselling in China depends on prior knowledge established based on non-Chinese populations. This potentially could cause biased evaluations and lead to undesired reproductive decisions. Our analyses provide a more compressive view of the GCR in the Chinese population to guide the design of preventative measures and even drug searching direction. As the advance in sequencing techniques will continuously reduce the sequencing cost, biobanks with larger cohorts and more comprehensive information will emerge rapidly and globally. The rationale and feasibility of implementing a pan-ethnic screening strategy for preventative reproductions also need to consider the rapid iteration of artificial intelligence. With an automated, precise and yet reliable interpretation driven by state-of-art artificial intelligence algorithms, the economic cost will eventually reduce to a level where almost all rare diseases can be screened to promote human reproductive health.

## METHODS

### Candidate genes of autosomal recessive diseases

To focus on the autosomal recessive (AR) disorder caused by SNPs/Indels, a list of candidate genes was collected from OMIM database (<https://www.omim.org/downloads>) based on the following rules:

- includes AR-associated genes,
- excludes genes labelled by non-disease phenotypes or associated with multifactorial disorders, or the relationship between phenotype and gene is provisional,
- exclude genes of which phenotype is a chromosome deletion or duplication syndrome.

In total, 2464 candidate AR genes were obtained to perform subsequent analysis (Supplementary Table 1).

### Variants collection of different population

The China Metabolic Analytics Project (ChinaMAP) dataset was used to estimate the GCR of AR disorders in the Chinese population. The ChinaMAP project has released a variants frequency spectrum of 10,588 Chinese individuals determined by deep whole genome sequencing<sup>14</sup>, a site-only VCF containing allele frequencies was downloaded from mBiobank ([www.mbiobank.com/download/](http://www.mbiobank.com/download/)). To obtain high-quality variants,  $QUAL \leq 100$  was discarded, common variants with  $AF \geq 0.05$  were also filtered out. In addition, to narrow down the margin error of low allele frequency (0.001) to  $\pm 0.0005$  in 95% of interval confidence, allele with  $AN < 10,229$  was filtered out<sup>28</sup>.

The gnomAD<sup>29</sup> v3.1.2 data was used to compare the difference of GCR among populations with distinct genomic compositions. Allele frequencies of candidate genes of different subgroups stored in the VCF files were collected from gnomAD (<https://gnomad.broadinstitute.org/downloads>). The 'Non-Cancer' group was used to calculate GCR. Variants without "PASS" tag in the FILTER column or  $AF > 0.05$  were discarded. A list of candidate variants of five populations, including EAS, SAS, EUR, AMR and AFR, was selected by setting at least one allele cutoff ( $AC > 0$ ) for each population.

The Westlake BioBank for Chinese (WBBC) cohort<sup>30</sup>, which recruited 14,726 participants in the pilot project, was included to survey the potential diversity within the Chinese population. The samples were primarily collected from Jiangxi, Shandong and Zhejiang provinces. Site-only VCFs of autosome chromosomes were also obtained (<https://wbcc.westlake.edu.cn/downloads.html>). All variants in the files were considered high quality.

## Variants annotation

All variant files from public databases were normalised using bcftools (bcftools norm -m -)<sup>31</sup> before being annotated by VEP<sup>32</sup> with ensemble database release version 108, the most severe transcript of each gene was considered to identify potential deleteriousness. The ClinVar<sup>33</sup> database (release version V20221113) was also applied to annotate the variants. An in-house script was used to extract variants located in the candidate genes. These subsets of annotated variants were used to identify the subsequent deleterious variants.

## Deleterious variants

Deleterious variants of each population were identified separately following a carefully designed workflow (Supplementary Fig. 1). First, variants annotated as benign or likely benign by ClinVar were discarded, while those annotated as pathogenic or likely pathogenic with at least one star of review status were kept. The remaining variants, including pathogenic/likely pathogenic with conflicting annotation in ClinVar database, were further classified into three subgroups: (i). missense variants were predicted by nine computational tools with general cutoff: CADD  $\geq 20$ <sup>34,35</sup>, Eigen $\geq 1$ <sup>36</sup>, REVEL  $\geq 0.75$ <sup>37</sup>, DANN  $> 0.5$ <sup>38,39</sup>, Polyphen2 = 'D'<sup>40</sup>, SIFT = 'D'<sup>41</sup>, MetaSVM = 'D'<sup>42</sup>, MutationAssessor = 'H' or 'M'<sup>43,44</sup>, PROVEAN = 'D'. Only variants with all tools passed the cutoff were considered as deleterious. (ii). nonsense variants (including frameshift\_variant, stop\_gained, splice\_donor\_variant, splice\_acceptor\_variant and start\_lost), which classified as highly-confident loss-of-function by LOFTEE<sup>29</sup>, were predicted by autoPVS1<sup>45</sup> in advance. Only variants with 'Strong' or 'Very Strong' adjusted strength were considered as deleterious. (iii). non-missense/nonsense variants were discarded. Tools and corresponding cutoff settings listed above were carefully chosen based on the performance reported by previous studies<sup>46,47</sup> (Supplementary Table 10). Moreover, we filtered out variants by gene based on AF. For genes with at least one pathogenic variant identified by ClinVar, the max AF of ClinVar variant was used as a cutoff, otherwise using 0.005. Finally, for the ChinaMAP dataset, we manually revised missense variants that failed by the above process but with conflicting pathogenic/likely pathogenic interpretations in the ClinVar database. All putatively deleterious variants of each population were listed in Supplementary Tables 2–8.

## Candidate autosomal recessive gene annotation

To distinguish untreatable diseases from treatable diseases, we searched for potential treatments from "Treatments for genetic disorders" ([www.rx-genes.com](http://www.rx-genes.com)) by the name of genes carrying deleterious variants. Genes with any treatment returned by the database were considered as treatable, otherwise untreatable. Additionally, we used a set of deleterious variants, including predicted (missense and LoF) variants and those catalogued in ClinVar, that affect liquid-liquid phase separation (LLPS) as described in Salman F Banani's study<sup>16</sup>, to identify genes related to non-membrane organelle formation. If a deleterious variant in this study was also reported by Banani et al., the corresponding gene was considered as LLPS gene.

## GO enrichment

To investigate the potential of drug development, a set of LLPS-related genes which contain more than 200 adjacent non-coding variants (including 5'- and 3'-UTR, up- and downstream and splicing region variants) were collected for GO enrichment using clusterProfiler<sup>48</sup>, all significantly enriched ( $p$ -value  $< 0.01$  and adjust  $p$ -value  $< 0.05$ ) GO terms were ranked according to the count of input genes for visualisation.

## Gene carrier rate

To estimate GCR, the variant carrier rate (VCR) was first calculated according to Eq. (1) introduced in<sup>49</sup>:

$$VCR = 1 - \frac{AC - 2 * Hom}{0.5 * AN} \quad (1)$$

*Hom* represents the number of homozygous individuals. To facilitate the calculation, a simplified estimation was applied to all candidate genes under the assumption that homozygous variants are rare for severe AR conditions, where *Hom* count could be ignored. To confirm this hypothesis, GCRs with and without accounting for homozygous of the WBBC dataset were used to investigate the effect of *Hom* alleles. Assuming that homozygous individuals only accounted for a relatively small proportion of a population, the results showed that the two calculation methods are almost identical (Supplementary Fig. 16). Hence, the GCR can be estimated through Eq. (2):

$$GCR_g = 1 - \prod_{i=0}^v (1 - VCR_i) = 1 - \prod_{i=0}^v \left(1 - \frac{AC}{0.5 * AN}\right) \quad (2)$$

GCR of each population was summarised in Supplementary Tables 2–8.

## Panel design and comparison

To assess the robustness of proposed screening panels selected based on distinct cutoffs, two potential panels described by Xi et al.<sup>50</sup> and Wei et al.<sup>6</sup> were used to carry out a thorough comparison (Supplementary Tables 2–8, Supplementary Fig. 15).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## DATA AVAILABILITY

The datasets used and/or analysed during the current study are included in this published article as supplementary data files or reproducible through the provided source code.

## CODE AVAILABILITY

The code for this study is available on GitHub and can be accessed via the link: [https://github.com/LinZhe-SZ/CHN\\_ECS](https://github.com/LinZhe-SZ/CHN_ECS).

Received: 19 June 2023; Accepted: 7 November 2023;

Published online: 20 November 2023

## REFERENCES

- Dolk, H., Loane, M. & Garne, E. The prevalence of congenital anomalies in Europe. *Adv. Exp. Med. Biol.* **686**, 349–364 (2010).
- Ostrander, B. & Bale, J. F. Congenital and perinatal infections. *Handb. Clin. Neurol.* **162**, 133–153 (2019).
- Chen, L. J., Chiou, J. Y., Huang, J. Y., Su, P. H. & Chen, J. Y. Birth defects in Taiwan: a 10-year nationwide population-based, cohort study. *J. Formos. Med. Assoc.* **119**, 553–559 (2020).
- Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Halldórsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
- Wei, C. Y. et al. Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *NPJ Genom. Med.* **6**, 10 (2021).
- Gregg, A. R. et al. Screening for autosomal recessive and X-linked conditions during pregnancy and preconception: a practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1793–1806 (2021).

8. Kirk, E. P. et al. Gene selection for the Australian Reproductive Genetic Carrier Screening Project ("Mackenzie's Mission"). *Eur. J. Hum. Genet.* **29**, 79–87 (2021).
9. Zhao, S. et al. Pilot study of expanded carrier screening for 11 recessive diseases in China: results from 10,476 ethnically diverse couples. *Eur. J. Hum. Genet.* **27**, 254–262 (2019).
10. Arjunan, A. et al. Evaluation and classification of severity for 176 genes on an expanded carrier screening panel. *Prenat. Diagn.* **40**, 1246–1257 (2020).
11. Beauchamp, K. A. et al. Systematic design and comparison of expanded carrier screening panels. *Genet. Med.* **20**, 55–63 (2018).
12. Xiao, Q. & Lauschke, V. M. The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *NPJ Genom. Med.* **6**, 41 (2021).
13. Leibowitz, R. et al. Reproductive genetic carrier screening for cystic fibrosis, fragile X syndrome and spinal muscular atrophy: patterns of community and healthcare provider participation in a Victorian screening program. *Aust. J. Prim. Health* **28**, 580–587 (2022).
14. Cao, Y. et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
15. Pan, Y. et al. Comparative genomic and transcriptomic analyses reveal the impacts of genetic admixture in Kazaks, Uyghurs, and Huis. *Mol. Biol. Evol.* **40**, msad054 (2023).
16. Banani, S. F. et al. Genetic variation associated with condensate dysregulation in disease. *Dev. Cell* **57**, 1776–1788.e1778 (2022).
17. Mensah, M. A. et al. Aberrant phase separation and nucleolar dysfunction in rare genetic diseases. *Nature* **614**, 564–571 (2023).
18. Seydel, C. Baby's first genome. *Nat. Biotechnol.* **40**, 636–640 (2022).
19. Zhu, W. et al. A robust pipeline for ranking carrier frequencies of autosomal recessive and X-linked Mendelian disorders. *NPJ Genom. Med.* **7**, 72 (2022).
20. Johansen Taber, K. et al. A guidelines-consistent carrier screening panel that supports equity across diverse populations. *Genet. Med.* **24**, 201–213 (2022).
21. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e524 (2019).
22. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15**, 379–393 (2014).
23. Fumagalli, M. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE* **8**, e79667 (2013).
24. Hanks, S. C. et al. Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. *Am. J. Hum. Genet.* **109**, 1653–1666 (2022).
25. He, Y., Li, J. & Zhang, M. Myosin VII, USH1C, and ANKS4B or USH1G Together Form Condensed Molecular Assembly via Liquid-Liquid Phase Separation. *Cell Rep.* **29**, 974–986.e974 (2019).
26. Quiroz, F. G. et al. Liquid-liquid phase separation drives skin barrier formation. *Science* **367**, eaax9554 (2020).
27. Greene, D. et al. Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nat. Med.* **29**, 679–688 (2023).
28. Cochran, W. G. *Sampling Techniques* 3rd edn.
29. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
30. Cong, P. K. et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).
31. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
32. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
33. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
34. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
35. Niroula, A. & Vihinen, M. How good are pathogenicity predictors in detecting benign variants? *PLoS Comput. Biol.* **15**, e1006481 (2019).
36. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
37. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
38. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
39. Zeng, Z., Aptekmann, A. A. & Bromberg, Y. Decoding the effects of synonymous variants. *Nucleic Acids Res.* **49**, 12673–12691 (2021).
40. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter **7**, Unit 7.20 (2013).
41. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
42. Kim, S., Jhong, J. H., Lee, J. & Koo, J. Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData Min.* **10**, 2 (2017).
43. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
44. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
45. Xiang, J., Peng, J., Baxter, S. & Peng, Z. AutoPVS1: an automatic classification tool for PVS1 interpretation of null variants. *Hum. Mutat.* **41**, 1488–1498 (2020).
46. Katsonis, P., Wilhelm, K., Williams, A. & Lichtarge, O. Genome interpretation using in silico predictors of variant impact. *Hum. Genet.* **141**, 1549–1577 (2022).
47. Li, J. et al. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.* **46**, 7793–7804 (2018).
48. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
49. Guo, M. H. & Gregg, A. R. Estimating yields of prenatal carrier screening and implications for design of expanded carrier screening panels. *Genet. Med.* **21**, 1940–1947 (2019).
50. Xi, Y. et al. Expanded carrier screening in Chinese patients seeking the help of assisted reproductive technology. *Mol. Genet. Genom. Med.* **8**, e1340 (2020).

## ACKNOWLEDGEMENTS

The start-up grant (project no. 9610580) to Dr. Jilin Zhang by City University of Hong Kong supported the study.

## AUTHOR CONTRIBUTIONS

J.Z., L.Z., L.Y. and J.-G.Z. conceptualised, designed, and steered the project; J.Z. and J.C. performed data collection. L.Y., Z.L. and Y.G. led the primary data analyses. J.-G.Z., H.P. and Y.-Q.L. collected and interpreted diseased data. J.Z., L.Y. and Z.L. interpreted the data and wrote the manuscript. L.Y. and Z.L. are equal contributors to this work and are designated as co-first authors. All authors approved the submission of the final version.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41525-023-00383-8>.

**Correspondence** and requests for materials should be addressed to Lijian Zhao or Jilin Zhang.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023