Data-driven analysis of carrier frequencies of autosomal recessive and X-linked diseases in the Asian population: abridged secondary publication

S Gu *, FM Lo, KWS Tsui, N Mullapudi

KEY MESSAGES

1. A robust pipeline was established to estimate and rank carrier frequencies of all known recessive genes based on genome-wide sequencing data in healthy individuals, with a focus on Asian populations.
2. Comprehensive criteria were applied to identify multiple deleterious variants including known pathogenic variants, presumed loss-of-function variants, predicted deleterious missense variants, and potentially harmful in-frame insertion and deletion mutations.
3. A high degree of correlation among different Asian population cohorts confirmed the validity of the variant selection criteria and overall analysis pipeline.

Hong Kong Med J 2025;31(Suppl 1):S14-8
HMRF project number: 08191216

[1] **S Gu**, [2] **FM Lo**, [1] **KWS Tsui**, [3] **N Mullapudi**

[1] *School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China*
[2] *Clinical Genetic Service, Department of Health, Hong Kong SAR, China*
[3] *Bionano Genomics, California, United States*

* Principal applicant and corresponding author: shengu@cuhk.edu.hk

Introduction

Single-gene disorders are common causes of neonatal and paediatric morbidity and mortality. Parents of individuals with autosomal recessive diseases and mothers of individuals with X-linked recessive diseases are carriers. Expanded newborn screening (NBS) and prenatal or pre-pregnancy expanded carrier screening (ECS) for recessive diseases are widely implemented in developed countries. Because carrier frequencies of recessive diseases vary among ethnic groups, recessive gene panels for NBS and ECS should be based on population-specific carrier frequencies in countries and regions with a single majority ethnic group.

Carrier frequencies for relatively large gene panels (eg, >100 genes) are primarily estimated using genome-wide sequencing data from unaffected individuals.[1-3] However, studies of carrier frequencies for large gene panels among individuals receiving genetic testing have been limited.[4,5] Both estimated and actual population-specific carrier frequencies are mostly based on individuals residing in the United States with self-reported ethnicities. In regions without accessible ECS, carrier frequencies for most genes remain unknown. This study aimed to develop an unbiased and robust pipeline for ranking carrier frequencies of all known recessive genes using genome-wide sequencing data. It also aimed to establish selection criteria for deleterious variants.

Methods

A combined list of type 1 to 4 variants (known pathogenic variants, presumed loss-of-function variants, predicted deleterious missense variants, and potentially harmful in-frame insertion and deletion mutations, respectively) was generated for each of the 2699 genes (Fig 1). Ethnicity-specific allele counts and total allele numbers were analysed for each variant, along with the number of individuals homozygous for each variant. The ethnicity-specific variant carrier rate (VCR) was also calculated. Each of the 2699 lists comprised the ethnicity-specific VCRs for each variant. The ethnicity-specific gene carrier rate (GCR) for each of the 2699 genes was then calculated, along with predicted genetic prevalence at the gene level (pGPg).

Spearman rank correlation coefficients and Pearson correlation coefficients were calculated. For correlation analysis between two datasets, only genes with a GCR >0 in at least one dataset were included. Statistical analyses were performed using R 3.6.0 and ggpubr 0.4.0.

Results

Selection of deleterious variants in recessive genes

Sequencing results from the publicly available Genome Aggregation Database (gnomAD) were extracted as the discovery cohort. In total, 2699 known disease-causing recessive genes were considered, including 2525 autosomal recessive genes and 174 X-linked genes. High-quality gnomAD variants aligned to the GRCh38 human genome assembly reference for each gene were processed.

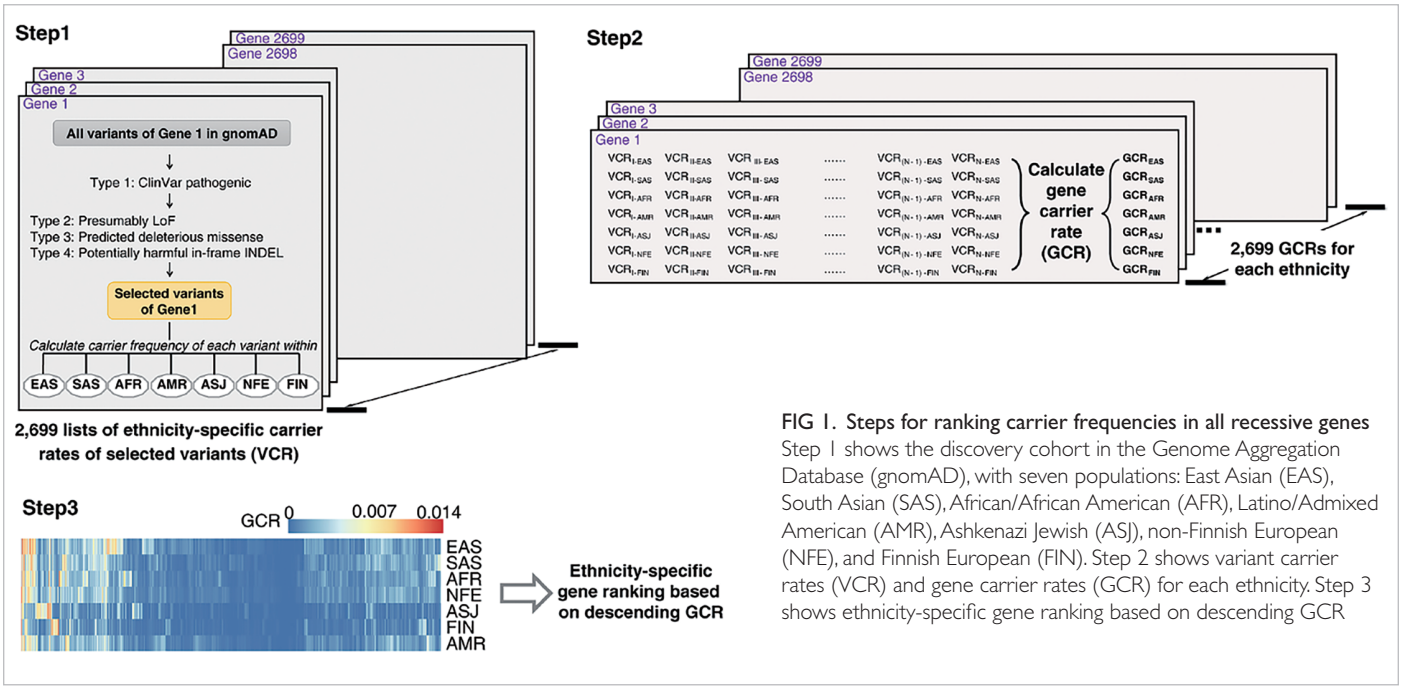

FIG 1. Steps for ranking carrier frequencies in all recessive genes
Step 1 shows the discovery cohort in the Genome Aggregation Database (gnomAD), with seven populations: East Asian (EAS), South Asian (SAS), African/African American (AFR), Latino/Admixed American (AMR), Ashkenazi Jewish (ASJ), non-Finnish European (NFE), and Finnish European (FIN). Step 2 shows variant carrier rates (VCR) and gene carrier rates (GCR) for each ethnicity. Step 3 shows ethnicity-specific gene ranking based on descending GCR

Overall, 48 198 273 gnomAD variants were identified in the 2699 genes. Among these variants, we selected those that were either reported in affected patients or potentially able to induce deleterious effects on gene function. Four types of variants were retained: known pathogenic variants (type 1), presumed loss-of-function variants (type 2), predicted deleterious missense variants (type 3), and potentially harmful in-frame insertion and deletion mutations (type 4).

Ethnicity-specific ranking of carrier frequencies in the discovery cohort

Seven sets of VCRs were identified for each ethnicity in each gene, resulting in seven sets of ethnicity-specific GCRs. Genes were then ranked by descending GCR values for each ethnicity. Consequently, genes with the highest GCRs (highest probabilities of causing recessive diseases in offspring) appeared at the top of the list for each population (Fig 1).

Ranking of carrier frequencies in validation cohorts

To verify the pipeline for ranking carrier frequencies, we analysed variants in three independent genome databases using whole-genome sequencing data from East Asian (Chinese) and South Asian (Malay and Indian) populations. These databases included the Singapore 10K Genome Project, China Metabolic Analytics Project (ChinaMAP), and Westlake BioBank for Chinese (WBBC) pilot project (Fig 2). Comparisons of results from different cohorts with similar ethnic backgrounds showed a high degree of correlation, confirming the validity of the variant selection criteria and overall analysis pipeline (Fig 3).

Carrier frequencies for genes in Hong Kong's newborn screening

In Hong Kong's NBS programme, among the 44 genes associated with 24 types of inborn errors of metabolism, six showed positive cases in the local population; their GCRs were generally high as expected. For some genes without positive cases, their GCRs were particularly low. This suggests a need to reconsider the selection of diseases and genes for a customised panel for the local population.

Discussion

We established a robust pipeline to estimate and rank carrier frequencies for all known recessive genes using genome-wide sequencing data. We developed comprehensive selection criteria for four types of potentially disease-causing variants and then confirmed the reliability of our filtering criteria.

Carrier frequencies of recessive diseases vary markedly among populations. The widespread use of next-generation sequencing has enabled the acquisition of carrier frequencies for large number of genes through both estimates based on large-scale genome-wide sequencing data and observations from ECS results. The former approach primarily focuses on individual genes or specific disease spectrums, whereas the latter approach focuses on genes included in the ECS panel. We performed

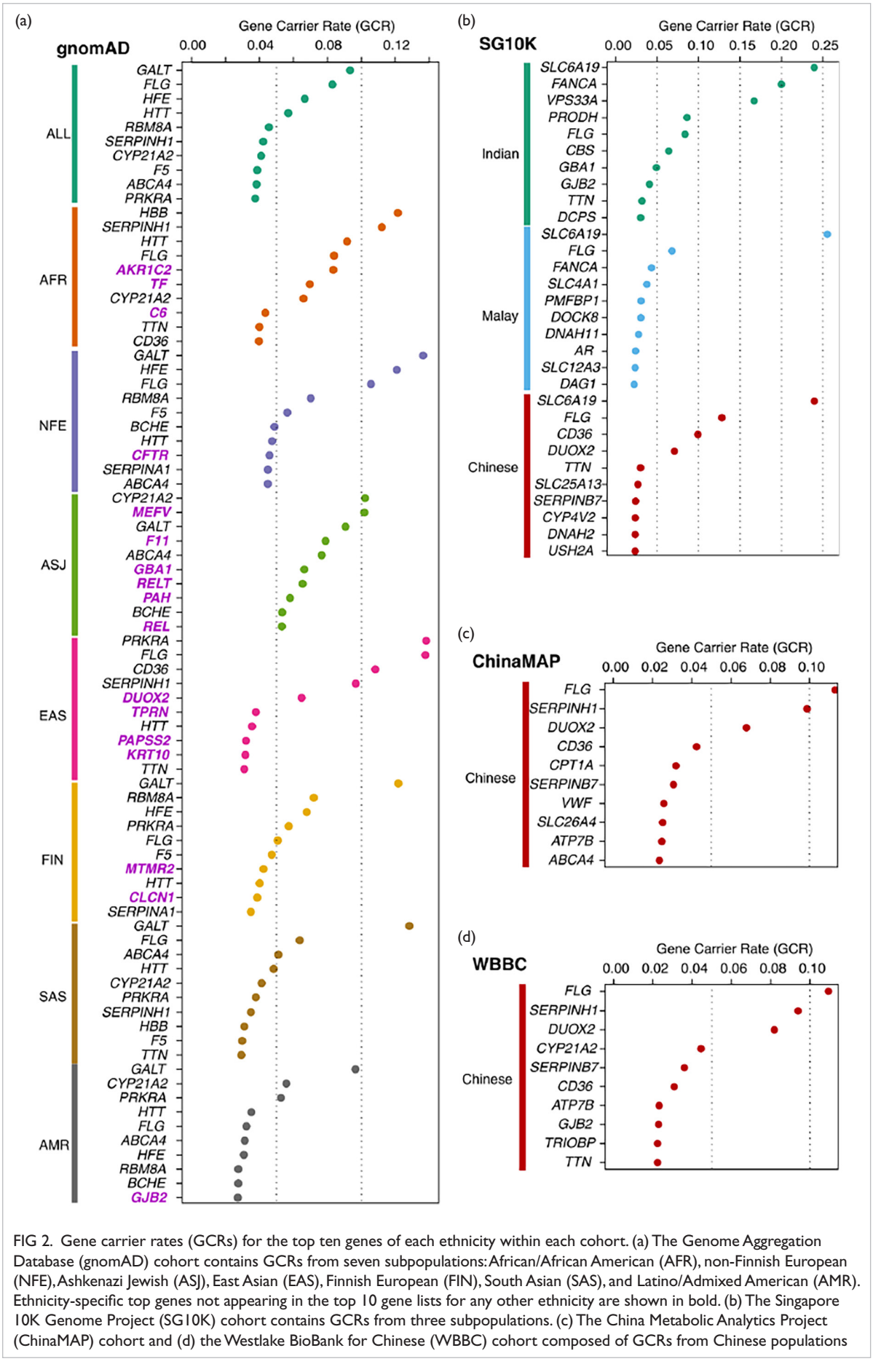

FIG 2. Gene carrier rates (GCRs) for the top ten genes of each ethnicity within each cohort. (a) The Genome Aggregation Database (gnomAD) cohort contains GCRs from seven subpopulations: African/African American (AFR), non-Finnish European (NFE), Ashkenazi Jewish (ASJ), East Asian (EAS), Finnish European (FIN), South Asian (SAS), and Latino/Admixed American (AMR). Ethnicity-specific top genes not appearing in the top 10 gene lists for any other ethnicity are shown in bold. (b) The Singapore 10K Genome Project (SG10K) cohort contains GCRs from three subpopulations. (c) The China Metabolic Analytics Project (ChinaMAP) cohort and (d) the Westlake BioBank for Chinese (WBBC) cohort composed of GCRs from Chinese populations

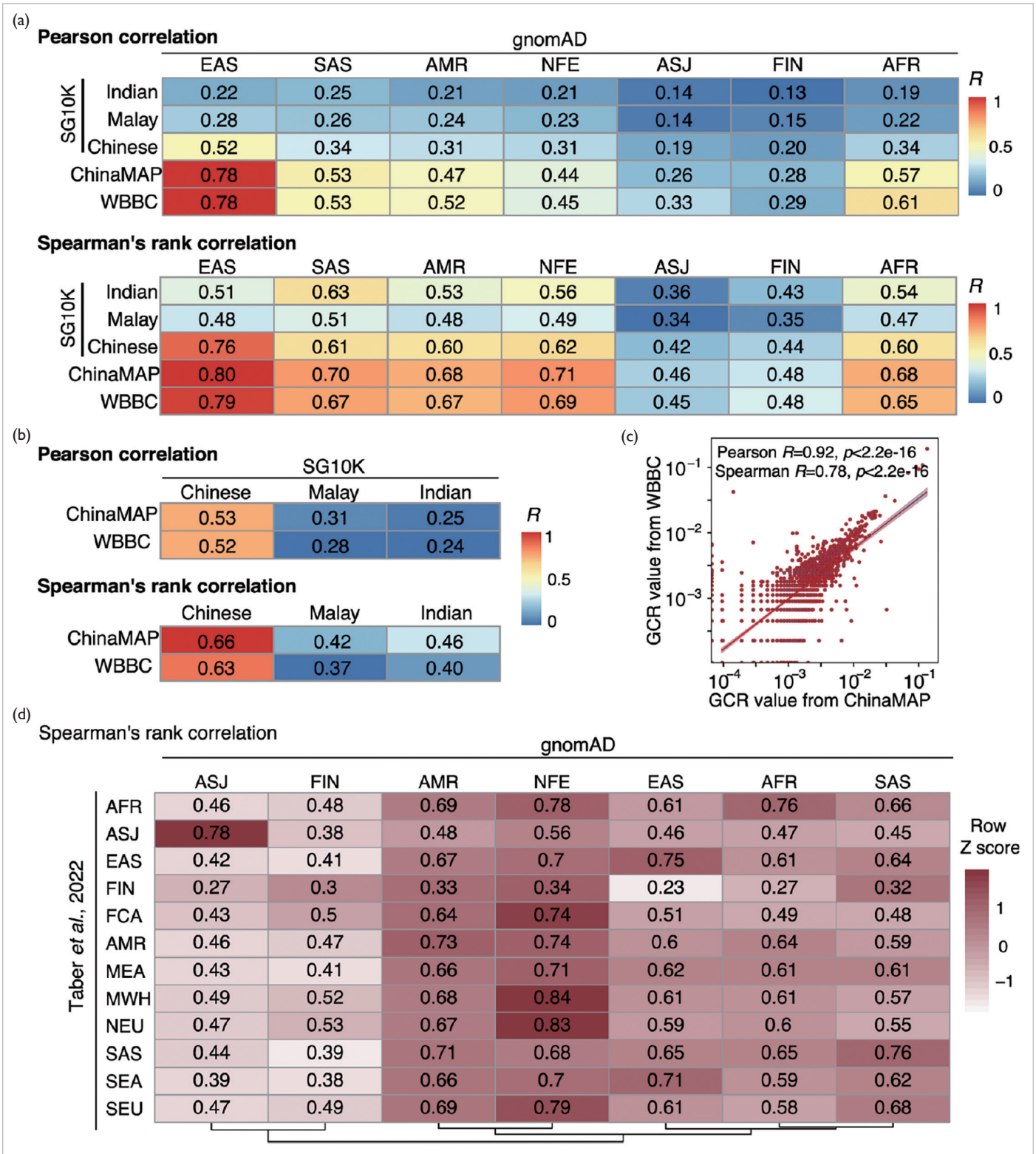

FIG 3. Comparison of carrier frequencies among cohorts. (a) Comparison between Genome Aggregation Database (gnomAD) populations and Singapore 10K Genome Project (SG10K) subpopulations, China Metabolic Analytics Project (ChinaMAP) Chinese, or Westlake BioBank for Chinese (WBBC) Chinese. (b) Comparison between SG10K subpopulations and ChinaMAP Chinese or WBBC Chinese. (c) Comparison between ChinaMAP Chinese and WBBC Chinese. (d) Comparison between calculated carrier frequencies in gnomAD populations and actual ethnicity-specific carrier frequencies based on expanded carrier screening

Abbreviations: AFR=African or African-American, AMR=Hispanic (corresponding to Latino/Admixed American population in gnomAD), ASJ=Ashkenazi Jewish, EAS=East Asian, FCA=French Canadian or Cajun, FIN=Finnish, MEA=Middle Eastern, MWH=Mixed or Other White, NEU=Northern European, SAS=South Asian, SEA=Southeast Asian, SEU= Southern European

a comprehensive analysis of all known recessive genes. Our analysis pipeline can be readily adapted to prospective novel recessive genes. The overall number of androgen receptor genes with recognisable phenotypes is estimated to be between 9000 and 10 100, suggesting that currently known androgen receptor genes represent only approximately 20% of the total.

In an 18-month retrospective study of NBS for 24 inborn errors of metabolism in Hong Kong, where 86.5% of newborns are Chinese (East Asian) and the remaining are primarily Southeast or South Asian (Filipino, Indian, Nepalese, or Pakistani), nine patients with positivity for six inborn errors of metabolism were recorded. Carrier frequencies for these six diseases were particularly high in the East Asian and South Asian populations. Specifically, citrullinemia type II and carnitine uptake deficiency were confirmed in more than one Chinese patient, ranking 14th and 12th among carrier frequencies for all 2699 recessive genes in the WBBC cohort. Similarly, these two genes ranked 21st and 20th in the ChinaMAP cohort. Nevertheless, some of the remaining 18 diseases, for which no positive cases were identified, had low carrier frequency rates in both East and South Asian populations, indicating that more targeted selection for the region's NBS panel is warranted.

The design of NBS and ECS panels requires a careful balance between comprehensiveness and cost-effectiveness. Regarding NBS panels, only treatable diseases with relatively high prevalence should be included. Regarding ECS, as costs for next-generation sequencing decrease, increasing numbers of genes are added to various panels; there have been suggestions to include whole-exome sequencing or whole-genome sequencing in preconception carrier screening. However, larger panels reduce sequencing depth for individual genes, leading to missed variant calls. Moreover, these panels place unnecessary burdens on variant interpretation and genetic counselling, which are the most time- and cost-intensive processes. Therefore, genes and diseases included in NBS and ECS panels should be precisely selected and customised to the specific needs of each region or territory. Recent guidelines suggest a pan-ethnic, universal ECS panel for countries with mixed-race populations and a high likelihood of interracial couples. In countries and regions with a single majority ethnic group, a more focused panel would be more economically efficient and sufficient.

Conclusion

A robust pipeline was established to estimate and rank carrier frequencies; this pipeline is readily adaptable to new genome-wide sequencing data and prospective novel recessive genes. Because carrier frequencies in a given population constitute critical information for NBS and ECS design, our data-driven analysis provides a scientific basis and guidelines for such practices.

Funding

This study was supported by the Health and Medical Research Fund, Health Bureau, Hong Kong SAR Government (#08191216). The full report is available from the Health and Medical Research Fund website (https://rfs2.healthbureau.gov.hk).

Disclosure

The results of this research have been previously published in:

1. Zhu W, Wang C, Mullapudi N, et al. A robust pipeline for ranking carrier frequencies of autosomal recessive and X-linked Mendelian disorders. NPJ Genom Med 2022;7:72.

References

1. Guo MH, Gregg AR. Estimating yields of prenatal carrier screening and implications for design of expanded carrier screening panels. Genet Med 2019;21:1940-7.
2. Hanany M, Rivolta C, Sharon D. Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases. Proc Natl Acad Sci U S A 2020;117:2710-6.
3. Tan J, Wagner M, Stenton SL, et al. Lifetime risk of autosomal recessive mitochondrial disorders calculated from genetic databases. EBioMedicine 2020;54:102730.
4. Haque IS, Lazarin GA, Kang HP, Evans EA, Goldberg JD, Wapner RJ. Modeled fetal risk of genetic diseases identified by expanded carrier screening. JAMA 2016;316:734-42.
5. Johansen Taber K, Ben-Shachar R, Torres R, et al. A guidelines-consistent carrier screening panel that supports equity across diverse populations. Genet Med 2022;24:201-13.