

國立臺灣大學醫學院分子醫學研究所

碩士論文

Graduate Institute of Molecular Medicine

College of Medicine

National Taiwan University

Master Thesis

台灣單基因隱性遺傳疾病帶因者擴大篩檢分析

Expanded carrier screening status in Taiwan

劉人鳳

Jen-Feng Liu

指導教授：楊偉勛 博士

陳沛隆 博士

Advisor: Wei-Shiung Yang, Ph.D.

Pei-Lung Chen, Ph.D.

中華民國111年1月

Jan., 2022



國立臺灣大學碩士學位論文
口試委員會審定書

台灣單基因隱性遺傳疾病帶因者擴大篩檢分析
Expanded carrier screening status in Taiwan

本論文係劉人鳳君（學號 P08448003）在國立臺灣大學醫學院分子醫學研究所完成之碩士學位論文，於民國 111 年 1 月 19 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

楊偉勳

(指導教授)

陳沛隆 (簽名)

許書睿

陳清瑜

系主任、所長

潘俊良

(簽名)

誌謝

這篇論文能完成，首先要感謝楊偉勛教授，陳沛隆副教授，兩位老師營造的求學環境，不鬆不緊地引導學生，按部就班地砌上求學路上的每塊磚頭；特別是陳沛隆老師，和許書睿老師，每個星期三中午不厭其煩地領導大家，由淺入深地從基礎到臨床，從wet lab到dry lab的親身教導，學生獲益良多。

第二要感謝吳君泰老師，吳老師引領我入門，一窺基因醫學堂奧，解除我的中年危機。陳倩瑜老師和東祈學長，以及可愛的研究生群，讓我覺得又年輕了起來，重燃求學的熱情。

最後要感謝我的家人，支持我繼續求學。特別是我的太太，照顧家裏，還要忍受我半夜不睡覺，電腦的的答答作響，感謝她。還有我的父母親，他們的身教言教，讓我對這個世界持續懷抱探索的熱情。

做為一個內科醫師，基因醫學開啟了一扇窗，讓我有能力以前所未有的視野審視以前的個案，未解的難題。同時對歷史，地理，人文，各種周圍的一切，有更貼近的認識。

論文完成只是實踐的開始，我會繼續努力。

摘要

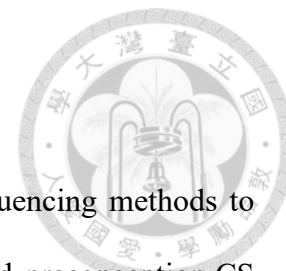


本論文的目的是將次世代定序方法應用於台灣族群的廣泛性帶因者篩檢。傳統的產前和孕前帶因者篩檢只針對特定族群中，盛行率高的特定疾病。這種方式基於種族背景和特定疾病的帶因者篩檢，在現代世界，因為族群融合迅速，往往不夠全面。其次，傳統方式需要多種評估方法來配合，才能處理多種狀況，這增加了孕前帶因者篩檢的複雜性，嚴重限制了篩檢覆蓋範圍。

使用次世代定序，上述的局限性得到大幅度的改善。它可以擴大所涵蓋的疾病基因數量並將檢測應用於跨族群的範圍。在這項研究中，我們建立一個涵蓋270基因的虛擬套組，應用臺灣生物資料庫裏的1496個全基因定序樣本予以分析各個單基因遺傳疾病的帶因率。我們發現，在台灣族群，23對父母中有1對可以得利於此項檢查，避免新生兒單基因遺傳疾病的風險。我們建議，利用次世代定序施行廣泛性帶因者篩檢，與傳統方式和陣列基因分型相比，可以提供更準確全面的資訊。瞭解個人單基因遺傳疾病帶因者者狀態不僅在生殖策略中很重要，而且在健康管理中也很重要。

關鍵詞：帶因者篩檢；單基因遺傳疾病；致病變異點；帶因者頻率；廣泛性跨族群篩檢

Abstract



The purpose of this thesis is to apply next-generation sequencing methods to expanded carrier screening(CS) in Taiwan. Traditional prenatal and preconception CS targeted specific diseases with high prevalence in defined subpopulations. CS based on ethnicity and race is often incomprehensive in the modern world due to complex population admixture. Second, multiple evaluation methods are required to process all these conditions, which contribute to the complexity of preconception screening. The technical limitation severely restricts the disease coverage in CS.

Modern genetic screening technology using NGS overcomes the limitations mentioned above by expanding the number of diseases covered and applies the testing to whole populations. In this study, a 270- gene virtual panel is applied to 1496 individual WGS samples from Taiwan biobank. 1 in 23 couples in Taiwan might benefit from this method to avoid diseased offspring. We demonstrated next-generation sequencing, offered in a pan-ethnic approach, is a better way to carrier screening. It provides more comprehensive information in a broader range on reproduction choice compared to the traditional way and array genotyping. Knowing individual carrier status is important not only in reproduction, but also in health management.

Keywords : carrier screening; single-gene disorders; pathogenic variants; carrier frequencies; pan-ethnic screening

Contents



口試委員審定書.....	i
誌謝	ii
摘要	iii
Abstract	iv
Contents	v
List of tables	vi
List of figures	vii
Introduction.....	1
Material and methods.....	3
Results.....	5
Discussion.....	8
Conclusion.....	10
Figures.....	11
Tables	12
References.....	19



List of Tables

Table 1 Estimated carrier frequency comparison	12
Table 2 <i>GJB2</i> variants in TWBv20	13
Table 3 270 gene panel.....	17
Table 4 Reproductive risk detection rate with different panels.....	18
Table 5 Analysis of 494-people subset of HBA1/2 carrier status....	18
Table 6 The SMN1/SMN2 copy number (CN) distribution from 494 Taiwanese.....	18

List of figures

Figure 1. Workflow scheme	11
Figure 2. At-risk couple rate calculation	11



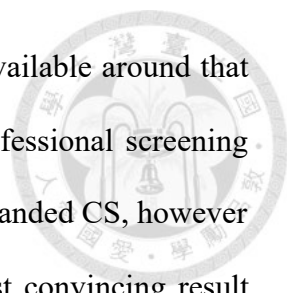


[Introduction]

The landscape of preconception carrier screening (CS) is rapidly changing in the era of next-generation sequencing (NGS). Traditional prenatal and preconception CS targeted specific diseases with high prevalence in defined subpopulations. For example, in United states, CS for eight disorders (familial dysautonomia, Tay-Sachs disease, Canavan disease, Fanconi anemia, Niemann-Pick type A, etc.) are recommended by the American College of Medical Genetics and Genomics (ACMG) and the American College of Obstetricians and Gynecologists (ACOG) for Ashkenazi Jewish (AJ) heritage¹. ACMG, ACOG also recommend pan-ethnic screening for cystic fibrosis, spinal muscular atrophy (SMA) and hemoglobinopathies^{2 3 4}. CS based on ethnicity and race is often incomprehensive in the modern world due to complex population admixture. For instance, in a U.S. study⁵, as we targeted Ashkenazi Jewish (AJ) diseases (eight disorders mentioned above), 81.6% (4434/5435) of carriers identified did not report AJ ancestry. Therefore we need a pan-ethnic approach to deal with it.

Traditionally multiple evaluation methods are applied to process all these conditions. Mean corpuscular volume (MCV) and hemoglobin electrophoresis are used to detect hemoglobinopathy carriers. Clotting factor tests are used for hemophilia carriers. Cystic fibrosis carrier detection has utilized array genotyping in the last decade. MLPA (Multiple-ligation dependent probe amplification) is used for SMA carrier detection. These contribute to the complexity of preconception screening.

Modern genetic screening technology using NGS overcomes the limitations mentioned above by expanding the number of diseases covered and applies the testing to whole populations⁶. ACOG and ACMG have endorsed offering expanded CS to couples who are considering pregnancy or are already pregnant, regardless of ethnicity



since 2017⁷ ⁸. Expanded carrier screening became commercially available around that time and rapidly demonstrated the superiority over traditional professional screening guidelines ⁹. Initially array genotyping was the mainstream for expanded CS, however it has been replaced with next generation sequencing ¹⁰. The most convincing result came from Westemyer M et al. ⁵. They evaluated carrier status for 381,014 U.S. individuals with NGS, and with a clinically-significant 274-gene list. Compared with standard screening, NGS-based CS provides additional information that may impact reproductive choices. Pan-ethnic CS leads to substantially increased identification of at-risk couples. And these data support offering NGS-based CS to all reproductive-aged women.

Taiwan is a country composed of people from multiple geographical and racial origins. We think that pan-ethnic approach is the best way for preconception CS. Here we present the result of a virtual panel, composed of the 274 genes mentioned above, ⁵ utilizing 1496 WGS data from Taiwan Biobank. It is the first time we use NGS data for estimation of carrier frequency in Taiwan population. For academic interest, we compared the frequency of pathogenic variants with the U.S. cohort ⁵ and the recently published Taiwan domestic carrier frequency result, using 103,106 array genotyping data¹¹. We also calculate the at-risk couple rate identified with the virtual panel. We think the data is valuable for reproduction strategy. And beyond that, we noticed some indigenous disease causing genes that are not included in the 274 gene panel. For their high prevalence and clinical actionability, they should be included in the local CS panel.

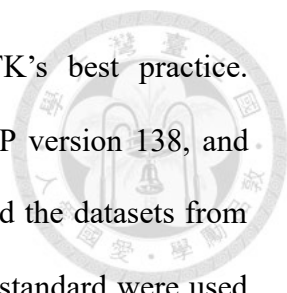
[Materials and methods]

Data source

All 1,496 Taiwanese WGS data were collected and de-identified at Taiwan Biobank (TWB). The samples were sequenced by Illumina HiSeq 2500, 4000, and Novaseq system. Sequencing of DNA extracted from each blood sample generated about 90 GB of data with an average coverage of 30X.

Genotype calling, validation, and variant annotation

Variant detection and joint genotype calling analyses were done in collaboration with Mr. Dong-Chi Wu in Chien Yu CHen's lab. Variant detection and joint genotype calling analyses were done based on the Sentieon DNAscope pipeline (Sentieon Inc., version 201808¹²), which is a commercial implementation of GATK's best practice¹³. In details, the sequence reads in FASTQ format of each sample were first aligned against the human reference genome (GRCh37/ucsc.hg19.fasta) using BWA-MEM (Burrows-Wheeler Aligner with Maximal Exact Match algorithm, version 0.7.15-r1140¹⁴). The output alignment file was then sorted by Samtools¹⁵ the Sentieon Dedup. Once the duplicated reads had been removed, we use the Sentieon Realigner to reinforce local realignment around each indel region, and the Sentieon QualCal to recalibrate the base quality scores. Single nucleotide variants (SNV) and short insertions/deletions (INDEL) were then called in genomic variant call format (GVCF) by Haplotype. To compile all variants in the 1,496 subjects, the Sentieon GVCFTyper was used to jointly call 1,496 subjects as a cohort. Next, these variants underwent variant refinement processes using Sentieon VarCal plus ApplyVarCal. Sentieon VarCal plus ApplyVarCal is a machine-learning-based variant quality sequence recalibration (VQSR) algorithm, in



which the training sets for VQSR were identical to the GATK's best practice. Specifically, the datasets from the 1000G phase1, omni2.5, dbSNP version 138, and HapMap version 3.3 were included as the training sets for SNV and the datasets from the 1000G phase1, dbSNP version 138, and Mills-and-1000G gold standard were used as the training sets for INDEL. VQSR included a list of sequence level annotations such as QD, MQ, MQRankSum, ReadPosRankSum, and FS. In the processes of VQSR, multi-allelic variants were normalized into multiple bi-allelic variants at the same position through decomposition and left-aligned using bcftools (v1.9)^{15,16}. To benchmark the variant calling pipeline, we repeated the same pipeline with an additional seven Genome in A Bottle (GIAB) samples (HG001~HG007) that were jointly called with 1,496 TWB subjects. We stratified the variant call sets according to the VQSR tranche 100.0, 99.9, 99.8, 99.7, 99.6, 99.5, and 99.0. We further adopted the call rate, allele number (AN), and depth of coverage (DP) criteria for classifying all reference alleles into three quality classifications. To calculate the recall and precision of each VQSR tranche, we randomly validated hundreds of variants in four samples (NGS2_20150510B, NGS2_20150510C, NGS2_20150510D, NGS2_20150510F) by Sanger sequencing. We are particularly interested in the variants absent after joint calling or only present after joint calling. Due to limited DNA, we separated samples into two groups (BC and DF) for cross-validation. The detailed estimated false discovery rates (FDR) for each VQSR tranche were described in the submitted paper (Wu *et. al.* 2022). Based on the estimated FDR, and precision rates, VQSR tranche 99.7 was chosen as PASS because its recall rate was very good without compromising precision rate too much.

Variant annotation was done by ANNOVAR (version 20180416)¹⁷ with ClinVar (v20210501), variants annotated as pathogenic or likely pathogenic in the ClinVar

database were included for analysis. Then we used the Mendelian gene list in the Clinical Genomic Database(CGD)(v20211021) (<http://research.nhgri.nih.gov/CGD/>) as a filter for the final variant analysis, which includes 4,325 genes with medically significant genetic data and available interventions. The workflow is presented as Fig. 1.

To further elucidate the possible role of structural variant calling tools, as a trial, we used Manta (version 1.6.0) ¹⁸ and AnnotSV (version 3.0.4) ¹⁹ to identify and annotate the structural variants (SV) in *HBA1/2* genes (n = 494, a subset of 1,496). Furthermore, the *SMN1/SMN2* copy number was analyzed by SMNCopyNumberCaller (version 1.1.1) ²⁰.

[Results]

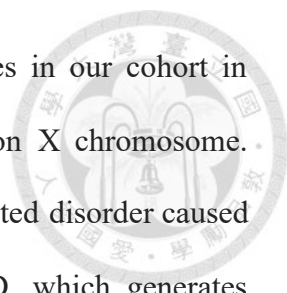
Carrier status of variants on monogenic disease-causing genes

We compared our carrier data with the United States cohort ⁵ and another domestic research with the array genotyping method (TWBv2)¹¹. The main difference is listed in Table 1. The most striking difference resides in the *GJB2* gene, a famous hearing impairment contributor. The estimated carrier frequency in the Taiwan biobank NGS database reached 16.7%. *GJB2:p.V37I* variant contributes to 92% of them. *GJB2* pathogenic/likely pathogenic (P/LP) carrier frequency in the US cohort is only 6.25% ⁵, less than one-half of the Taiwan counterpart. In contrast, the *GJB2* carrier rate falls to 1.59% with TWBv2 array genotyping ¹¹. The information of *GJB2* variants in TWBv2 is in Table 2.

Another example is the *SLC25A13* gene, related to citrullinemia. Pathogenic/likely pathogenic carrier frequency is 2.57% in our series, whereas it is only 0.40% in the U.S. cohort ⁵, and 1.9% in the TWBv2 series ¹¹. Yet another one is the *PTS*

gene, defective *PTS* variants would lead to phenylketonuria. The pathogenic/likely pathogenic carrier frequency of *PTS* is 0.66% in our cohort, which is only 0.12% in the U.S. counterpart⁵. The frequency is compatible with clinical observation of PKU patients in Taiwan, where BH4-deficiency(defective *PTS*) PKU patients account for up to 1/3 of total PKU patients²¹²². In contrast, defective *PTS* PKU patients only account for 1-2% Caucasian cohorts. *PTS* variant is not shown in TWBv2 array data.

Based on TWB 1496 NGS cohort data, we applied the same method used in Westemeyer et al.⁵ to calculate the combined at-risk couple rate in Taiwan. At-risk couple rate is calculated as the summation of the square of AR gene carrier rate and summation of XL gene carrier rate(as demonstrated in Fig 2). The 270 gene panel (274 genes in Westemeyer et al.⁵, omit 4 genes *HBA1/2*, *DMD*, *SMN1*, *FMRI*) (Table 2) would identify the risk for a genetic disorder in the offspring of 1 in 28 couples (3.55%). If *DUOX2*, *G6PD* were added to the local carrier screening list, the risk identification would be 1 in 13 couples (7.16%). (Table 3) There are 3 weak pathogenic variants *CFTR*(rs551227135), *GALT*(rs2070074), *GJB2*(rs72474224). *CFTR*(rs551227135) is one of the classic *CFTR* IVS8-5T variants. It will cause abnormal splicing with skip of exon 9 of *CFTR*, and full length mRNA expression decreased to 11%(heterozygotic) and 6%(homozygote) compared with 7T/7T individuals²³. *GALT*(rs2070074)has been referred to as the Duarte variant and carries about 5-20% of wild type enzyme activity²⁴. *GJB2*(rs72474224)(also referred to as *GJB2-V37I*)is famous for its moderate effect with incomplete penetrance in hearing impairment. Little phenotypic presentations were noted with homozygotic combination of these variants. If we eliminated the weak-pathogenic variant combination condition, at-risk couple rate would be 1 in 23 couples(4.39%)(Table 3).



We identified 2 clinically significant disease causing genes in our cohort in addition to the 270 gene panel: *G6PD* and *DUOX2*. *G6PD* is on X chromosome. Glucose-6-phosphate dehydrogenase (*G6PD*) deficiency is an inherited disorder caused by a genetic defect in the red blood cell (RBC) enzyme *G6PD*, which generates NADPH and protects RBCs from oxidative injury. Identification of *G6PD* deficiency and patient education regarding medications and environmental exposure is critical to preventing life threatening hemolysis in neonatal period. Female carrier rate is estimated at 4.5 % in our cohort. The *DUOX2* gene provides instructions for making an enzyme called dual oxidase 2. Dual oxidase 2 helps generate a chemical called hydrogen peroxide, which is required in the thyroid gland for the production of thyroid hormones. Its dysfunction contributes to congenital hypothyroidism. Carrier frequency is estimated to be 2.34% in our cohort.

For the highest SV related hereditary disease in Taiwan: thalassemia, we used Manta¹⁸ and AnnotSV¹⁹ to identify *HBA1/2* pathogenic variants and in a subset of the TWB cohort (N=494). The alpha thalassemia carrier frequency was 6.88%(5.06% α^0 , 1.81% α^+)(Table 4). Alpha thalassemia α^0 means lacking 2 of 4 functional gene segments, --SEA (allele frequency 2.2%) is the most prevalent genotype of α^0 in our cohort. Alpha thalassemia α^+ means lacking 1 of 4 functional gene segments and α^{WS} (rs41479347)(allele frequency 0.51%) is the most prevalent one. For the other high prevalence hereditary disease spinal muscular atrophy (SMA), the *SMN1/SMN2* copy number was analyzed by SMNCopyNumberCaller²⁰. 10 carriers with only one copy of *SMN1* were identified out of 494 members, which meant 2.02% carrier frequency of SMA(Table 5).

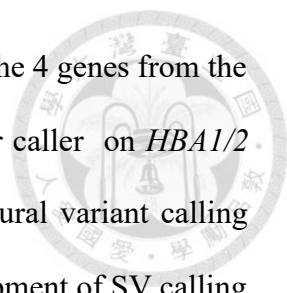
[Discussion]

In this study, we utilized WGS data to estimate the carrier risk in Taiwan. We analyzed a 270 gene panel to estimate the accumulated monogenic disease risk around 3.55%, implying approximately 1 of 28 couples might benefit from the carrier screening. This provides our young couples invaluable information of reproduction strategy information. And we also noticed carrier rate differences between Taiwan and the U.S. cohort. Several factors including ancestry, environment, founder effect contribute to these differences.

Like our U.S. cohort, sequencing methods surpassed array genotyping methods in the use of expanded carrier screening. In comparison with Wei et al.¹¹ in Table 1, the sequencing method explores more carriers with a more comprehensive result. With the lowering price of WES or targeted capture sequencing, it is possible to develop a local panel of expanded carrier screening.

However, one caveat of using sequencing methods for carrier screening is the misleading information of pseudogenes. *OTOA:p.Glu787Ter(rs200988634)* is a frequently met false-positive variant, confounded by *OTOAP1* pseudogene. *CFTR:p.Lys464Asn(rs397508198)*, *TUBB2A:p.Ala248Val(rs2808001)* and *TUBB2B:Ala248Val (rs777598117)* are other examples. We have to use IGV tool to detect pseudogene specific haplotypes to confirm pseudogene confounding. For example, we used *CFTR:c.1392+6insC* and *CFTR:c.1392+12G>A* as indicators of pseudogene presence²⁵.

Besides, another caveat for carrier screening is the inability to detect long fragment variants from short-reads data, e.g. CNV, large deletion, trinucleotide repeats, and gene conversion. With the short-reads data, we often underestimate the carrier frequency of 4 important genes: *HBA1/2*, *DMD*, *FMRI*, *SMN1*, of which many known



pathogenic variants are relatively large. For this reason, we deleted the 4 genes from the 274 gene panel. And we tried Manta, AnnotSV, SMNCopyNumber caller on *HBA1/2* and *SMN1* variant calling in a 494-individual subset. These structural variant calling tools are quite reliable in these 2 genes. Further trials on the development of SV calling algorithms on *DMD* and *FMRI* should be done soon.

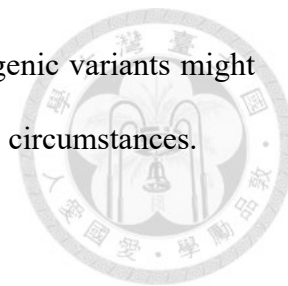
Two main pathogenic variant contributors are *GJB2:p.Val37Ile* (rs72474224, MAF:8.6%) and *CFTR:c.1210-11_1210-10insG* (rs551227135, MAF:0.9%). Both are considered “mild” pathogenic variants. Phenotype penetration of *GJB2*(rs72474224) is highly variable in Taiwan clinical experience. *CFTR*(rs551227135) is one of the classical *CFTR* IVS8-5T variants, which are responsible for non-classical cystic fibrosis presentation(CABVD, recurrent pancreatitis, late-onset CF).

To sum up, we used the 270 gene panel, together with *DUOX2* and *G6PD*. At risk couple rate would be 7.16%. That means 1 out of 13 couples might benefit from our expanded carrier screening panel. Even after adjustment to eliminate homozygotic weak-pathogenic combinations(*CFTR, GALT, GJB2*), at risk couple rate is still 1 in 23 couples.(Table 2)

In addition, these numbers give inspiration to daily medical practice. Some autosomal recessive diseases might be overlooked because of atypical or incomplete clinical presentation. With these carrier risk data, physicians would be more alert of covert Wilson disease, cystic fibrosis, and other rare diseases. And we could not fully exclude the possibility that aging, environmental factors, epigenetic factors might damage the remaining functional copy of genes in these carriers. Knowing individual carrier status is important not only in reproduction, but also in health management.

Lastly, as we screened all the pathogenic variants in monogenic genes listed in

CGD(clinical genomic database), many autosomal dominant pathogenic variants might be detected incidentally. ACMG SF v30²⁶ should be applied in these circumstances.



[Conclusion]

This study of 1496 individuals WGS data for pathogenic variants with up to 270 genes evaluated reaffirms the view that NGS can help in expanded carrier screening. Considering an increasingly diverse and multicultural population in Taiwan, NGS-based CS has the potential to impact the lives of at-risk couples in a way that current prenatal CS guidelines and traditional carrier screening have failed to achieve.

Figures



Fig 1. Workflow scheme

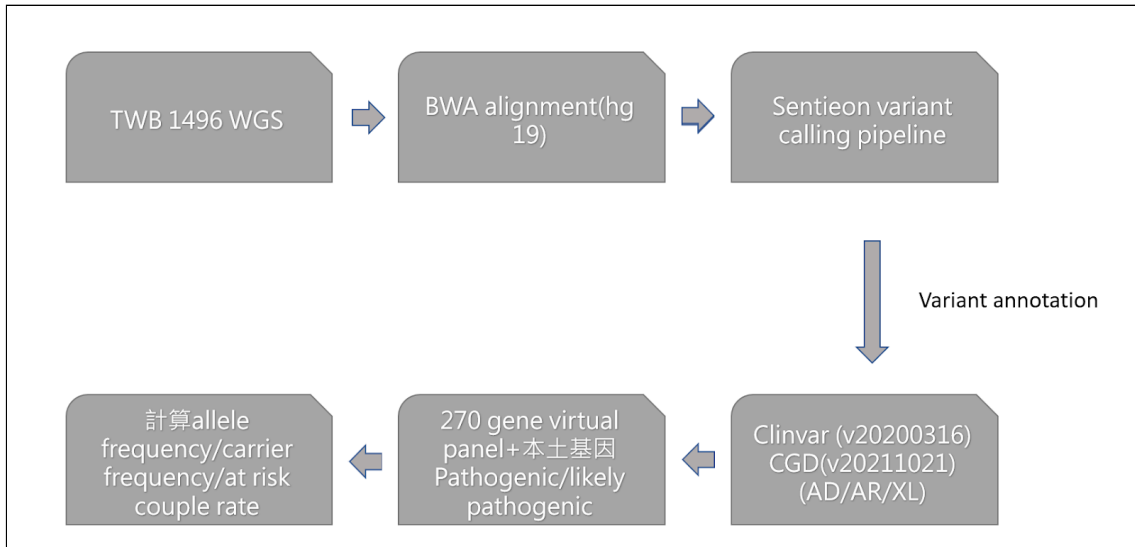


Fig 2. At-risk couple rate calculation

Example of at-risk couple rate calculation

GENE name	Inheritance pattern	Carrier frequency
A1	AR	a1
A2	AR	a2
A3	AR	a3
B1	XL	b1
B2	XL	b2

Here we take an example of a 5-gene panel
 At risk couple rate= $(a1)^2+(a2)^2+(a3)^2+b1+b2$

Tables



Table 1. Estimated carrier frequency comparison

Gene	Disease Name	Carrier frequency		
		Our cohort	Westemeyer et al. ⁵	Wei et al. ¹¹
<i>GJB2</i>	Hearing Loss(non-syndromic)	16.68%	6.25%	1.59%
<i>G6PD</i>	G6PD deficiency#	3.4%		2.49%
<i>VPS13B</i>	Cohen Syndrome	2.78%	0.33%	
<i>CFTR</i>	Cystic Fibrosis*	2.73%	3.85%	
<i>GALT</i>	Galactosemia	2.64%	0.67%	
<i>SLC25A13</i>	Citrin Deficiency	2.57%	0.40%	1.94%
<i>DUOX2</i>	congenital hypothyroidism	2.34%		
<i>GALC</i>	Krabbe Disease	2.18%	1.27%	1.67%
<i>SLC26A4</i>	Pendred Syndrome	2.04%	1.43%	1.70%
<i>ATP7B</i>	Wilson Disease	1.98%	1.52%	1.77%
<i>PAH</i>	Phenylketonuria	1.32%	2.50%	0.48%
<i>HBB</i>	Beta-Hemoglobinopathies	1.32%	3.23%	0.59%
<i>PTS</i>	PTPS Deficiency(PKU)	0.66%	0.12%	

*rs551227135(carrier frequency 1.9%) is so called *CFTR* IVS8-5T variant in previous literature. It is considered a “mild” *CFTR* mutation. When in trans with a known pathogenic *CFTR* mutation (e.g. $\Delta F508$), the IVS8-5T allele is responsible for non-classic CF phenotype, such as bilateral absence of vas deferens, recurrent pancreatitis, and late onset cystic fibrosis.²⁷ Individuals homozygous for the IVS8-5T allele as the sole variation of the whole *CFTR* coding sequence may present as non-classic CF with sinopulmonary disease and male infertility in Taiwan.²⁸

#Carrier frequency of G6PD pathogenic variants is estimated from TWB current female:male ratio 1.8:1.

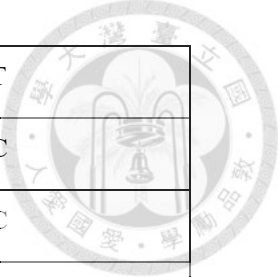
Table 2. *GJB2* variants in TWBv20



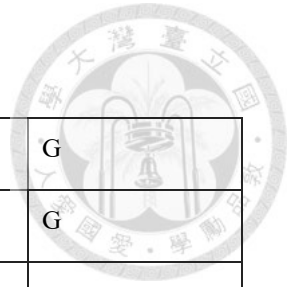
RS ID	Chr	Position	Strand	Cytoband	Ref	Alt
rs111033194	13	20188912	+	q12.11	T	G
rs587783647	13	20188932	+	q12.11	TATC	-
rs587783646	13	20188949	+	q12.11	AC	-
rs111033294	13	20188965	+	q12.11	T	C
rs76838169	13	20188974	+	q12.11	A	G
rs104894406	13	20188977	+	q12.11	C	A
rs111033335	13	20188982	+	q12.11	TCCAGAC AC	GAATGTCATG AACACTG
rs771748289	13	20188986	+	q12.11	G	A
rs1057517521	13	20189006	+	q12.11	TG	-
rs770116143	13	20189017	+	q12.11	TC	-
rs876657693	13	20189049	+	q12.11	ACAGTGT TGGG	-
rs1057517508	13	20189068	+	q12.11	A	-
rs770330002	13	20189068	+	q12.11	A	T
rs773528125	13	20189070	+	q12.11	-	CGTT
rs201004645	13	20189071	+	q12.11	C	T
rs774518779	13	20189076	+	q12.11	C	T
rs200104362	13	20189079	+	q12.11	T	C
rs111033360	13	20189083	+	q12.11	C	T
rs376898963	13	20189089	+	q12.11	G	A
rs28931592	13	20189106	+	q12.11	T	A
rs772264564	13	20189117	+	q12.11	A	T



rs111033420	13	20189126	+	q12.11	G	T
rs767178508	13	20189143	+	q12.11	C	T
rs80338948	13	20189155	+	q12.11	G	A
rs397516877	13	20189156	+	q12.11	G	T
rs76434661	13	20189166	+	q12.11	C	T
rs786204690	13	20189174	+	q12.11	G	T
rs397516874	13	20189212	+	q12.11	G	A
rs80338947	13	20189222	+	q12.11	CTC	-
rs150529554	13	20189227	+	q12.11	C	T
rs756484720	13	20189247	+	q12.11	TT	-
rs797045596	13	20189255	+	q12.11	CCCCTTG ATGAACT TC	-
rs111033253	13	20189256	+	q12.11	CCCTTGAT GAACTT	-
rs111033204	13	20189282	+	q12.11	AT	-
rs143343083	13	20189284	+	q12.11	G	A
rs111033299	13	20189299	+	q12.11	C	T
rs730880338	13	20189312	+	q12.11	-	A
rs80338945	13	20189313	+	q12.11	A	G
rs781534323	13	20189336	+	q12.11	G	C
rs727504302	13	20189343	+	q12.11	T	G
rs199883710	13	20189344	+	q12.11	G	A
rs80338943	13	20189347	+	q12.11	G	-
rs80338944	13	20189351	+	q12.11	C	T
rs104894395	13	20189352	+	q12.11	C	T
rs104894397	13	20189353	+	q12.11	A	G



rs28931593	13	20189358	+	q12.11	C	T
rs121912968	13	20189364	+	q12.11	T	C
rs200023879	13	20189374	+	q12.11	G	C
rs104894403	13	20189386	+	q12.11	C	G
rs111033203	13	20189388	+	q12.11	T	C
rs886037849	13	20189389	+	q12.11	A	G
rs750188782	13	20189391	+	q12.11	CACACGT TCTTGCA GC	-
rs104894410	13	20189407	+	q12.11	C	T
rs111033297	13	20189413	+	q12.11	G	A
rs80338942	13	20189415	+	q12.11	A	-
rs104894412	13	20189420	+	q12.11	G	T
rs587783645	13	20189424	+	q12.11	C	T
rs104894398	13	20189443	+	q12.11	C	A
rs72561723	13	20189448	+	q12.11	C	T
rs1057517491	13	20189448	+	q12.11	C	-
rs141774369	13	20189472	+	q12.11	A	G
rs587783644	13	20189475	+	q12.11	A	G
rs371024165	13	20189488	+	q12.11	G	A
rs104894396	13	20189511	+	q12.11	C	T
rs397516873	13	20189514	+	q12.11	ATCTTTCC AATGCTG GTGGAGT GTTTGTC ACACCCC C	-
rs879253741	13	20189516	+	q12.11	C	A
rs886037624	13	20189520	+	q12.11	CCAATGC TGGTG	T



rs1057517519	13	20189523	+	q12.11	A	G
rs111033217	13	20189538	+	q12.11	T	G
rs80338939	13	20189547	+	q12.11	C	-
rs111033451	13	20189563	+	q12.11	G	A
rs111033401	13	20189573	+	q12.11	C	T
rs371086981	13	20189580	+	q12.11	A	G
rs73431552	13	20191295	+	q12.11	G	A
rs80338940	13	20192782	+	q12.11	C	T



Table 3. 270 gene panel

270 gene panel	<p><i>ABCB11, ABCC8, ABCD1, ACAD9, ACADM, ACADVL, ACAT1, ACOX1, ACSF3, ADA, ADAMTS2, AGA, AGL, AGPS, AGXT, AIRE, ALDH3A2, ALDOB, ALG6, ALMS1, ALPL, AMT, AQP2, ARSA, ARSB, ASL, ASNS, ASPA, ASS1, ATM, ATP6V1B1, ATP7A, ATP7B, ATRX, BBS1, BBS10, BBS12, BBS2, BCKDHA, BCKDHB, BCS1L, BLM, BSND, BTD, CAPN3, CBS, CDH23, CEP290, CERKL, CFTR, CHM, CHRNE, CIITA, CLN3, CLN5, CLN6, CLN8, CLRN1, CNGB3, COL4A3, COL4A4, COL4A5, COL7A1, CPS1, CPT1A, CPT2, CRB1, CTNS, CTSK, CYBA, CYBB, CYP11B2, CYP17A1, CYP19A1, CYP27A1, DCLRE1C, DHCR7, DHDDS, DLD, DNAH5, DNAI1, DNAI2, DYSF, EDA, EIF2B5, EMD, ESCO2, ETFA, ETFDH, ETHE1, EVC, EYS, F11, F9, FAH, FAM161A, FANCA, FANCC, FANCG, FH, FKRP, FKTN, G6PC, GAA, GALC, GALK1, GALT, GAMT, GBA, GBE1, GCDH, GFMI, GJB1, GJB2, GLA, GLB1, GLDC, GLE1, GNE, GNPTAB, GNPTG, GNS, GPR56, GRHRP, HADHA, HAX1, HBB, HEXA, HEXB, HFE2, HGSNAT, HLCS, HMGCL, HOGA1, HPS1, HPS3, HSD17B4, HSD3B2, HYAL1, HYLS1, IDS, IDUA, IKBKAP, IL2RG, IVD, KCNJ11, LCA5, LDLR, LDLRAP1, LHX3, LIFR, LIPA, LOXHD1, LPL, LRPPRC, MAN2B1, MCCC1, MCCC2, MCOLN1, MED17, MEFV, MESP2, MFSD8, MKS1, MLC1, MMAA, MMAB, MMACHC, MMADHC, MPI, MPL, MPV17, MTHFR, MTM1, MTRR, MTPP, MUT, MYO7A, NAGLU, NAGS, NBN, NDRG1, NDUFAF5, NDUFS6, NEB, NPC1, NPC2, NPHS1, NPHS2, NR2E3, NTRK1, OAT, OPA3, OTC, PAH, PCCA, PCCB, PCDH15, PDHA1, PDHB, PEX1, PEX10, PEX2, PEX6, PEX7, PFKM, PHGDH, PKHD1, PMM2, POMGNT1, PPT1, PROPI, PSAP, PTS, PUS1, PYGM, RAB23, RAG2, RAPSN, RARS2, RDH12, RMRP, RPE65, RPGRIP1L, RS1, RTEL1, SACS, SAMHD1, SEPSECS, SGCA, SGCB, SGCG, SGSH, SLC12A3, SLC12A6, SLC17A5, SLC22A5, SLC25A13, SLC25A15, SLC26A2, SLC26A4, SLC35A3, SLC37A4, SLC39A4, SLC4A11, SLC6A8, SLC7A7, SMARCAL1, SMPD1, STAR, SUMF1, TCIRG1, TECPR2, TFR2, TGM1, TH, TMEM216, TPP1, TRMU, TSFM, TTPA, TYMP, USH1C, USH2A, VPS13A, VPS13B, VPS45, VRK1, VSX2, WNT10A</i></p>
----------------	---

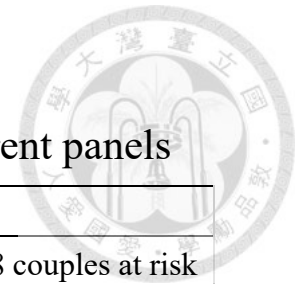


Table 4. Reproductive risk detection rate with different panels

Expanded carrier gene panel	accumulated risk	
270 gene panel	3.55%	1 in 28 couples at risk
270 gene panel+ <i>DUOX2</i> + <i>G6PD</i> #	7.16%	1 in 13 couples at risk
272 gene panel-adjustment(<i>CFTR</i> / <i>GALT</i>)*	7.06%	1 in 14 couples at risk
272 gene panel-adjustment(<i>CFTR</i> / <i>GALT</i> / <i>GJB2</i>)*	4.39%	1 in 23 couples at risk

*after elimination of *CFTR*(rs551227135) / *GALT*(rs2070074) / *GJB2*(rs72474224)

homozygous weak pathogenic variant condition

#at risk couple rate for X linked disease is equal to female carrier rate

Table 5. Analysis of 494-people subset of *HBA1/2* carrier status(using Manta and AnnotSV)

Thalassemia carrier	α^+				α^0		
	$-\alpha^{3.7}$	α^{CS} (rs41464951)	α^{QS} (rs41397847)	α^{WS} (rs41479347)	--SEA	--THAI	--FIL
N= 494	2	1	1	5	22	2	1

Table 6. The *SMN1*/*SMN2* copy number (CN) distribution from 494 Taiwanese.

<i>SMN2</i> CN	<i>SMN1</i> CN					
		0 (SMA patient)	1[∞] (SMA carrier)	2	3	4
0		0	0	21	5	0
1		0	0	145	18	1
2		0	4	278	4	0
3		0	6	9	2	0
4		0	0	0	0	0

¶ SMNCopyNumberCaller fails to identify *SMN1*/*SMN2* status in 1 of 494 individuals

∞There are 10 carriers out of 494 cases. 4 of them are *SMN1*:1/*SMN2*:2; 6 of them are *SMN1*:1/*SMN2*:3.



[References]

1. Gross, S. J., Pletcher, B. A., Monaghan, K. G. & Professional Practice and Guidelines Committee. Carrier screening in individuals of Ashkenazi Jewish descent. *Genet. Med.* **10**, 54–56 (2008).
2. Committee Opinion No. 691: Carrier Screening for Genetic Conditions. *Obstet. Gynecol.* **129**, e41–e55 (2017).
3. Prior, T. W. & Professional Practice and Guidelines Committee. Carrier screening for spinal muscular atrophy. *Genet. Med.* **10**, 840–842 (2008).
4. Grody, W. W. *et al.* Laboratory standards and guidelines for population-based cystic fibrosis carrier screening. *Genetics in Medicine* vol. 3 149–154 (2001).
5. Westemeyer, M. *et al.* Clinical experience with carrier screening in a general population: support for a comprehensive pan-ethnic approach. *Genet. Med.* **22**, 1320–1328 (2020).
6. Kraft, S. A., Duenas, D., Wilfond, B. S. & Goddard, K. A. B. The evolving landscape of expanded carrier screening: challenges and opportunities. *Genet. Med.* **21**, 790–797 (2019).
7. College of Obstetricians and, A. ACOG (American College of Obstetricians and Gynecologists) educational bulletin. Adult manifestation of childhood sexual abuse, number 259, July 2000. *Clinical ... Int. J. Gynaecol. Obstet.* (2001).
8. Grody, W. W. *et al.* ACMG position statement on prenatal/preconception expanded carrier screening. *Genet. Med.* **15**, 482–483 (2013).
9. Ben-Shachar, R., Svenson, A., Goldberg, J. D. & Muzzey, D. A data-driven evaluation of the size and content of expanded carrier screening panels. *Genetics in Medicine* vol. 21 1931–1939 (2019).
10. Chokoshvili, D., Vears, D. & Borry, P. Expanded carrier screening for monogenic

- disorders: where are we now? *Prenat. Diagn.* **38**, 59–66 (2018).
11. 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).
12. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools-A fast and accurate solution to variant calling from next-generation sequence data. *BioRxiv* (2017).
13. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
14. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
15. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
16. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
17. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
18. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
19. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
20. Chen, X. *et al.* Spinal muscular atrophy diagnosis and carrier screening from





- genome sequencing data. *Genet. Med.* **22**, 945–953 (2020).
21. Liu, K.-M. *et al.* Long-term follow-up of Taiwanese Chinese patients treated early for 6-pyruvoyl-tetrahydropterin synthase deficiency. *Arch. Neurol.* **65**, 387–392 (2008).
 22. Chien, Y.-H. *et al.* Treatment and outcome of Taiwanese patients with 6-pyruvoyltetrahydropterin synthase gene mutations. *J. Inherit. Metab. Dis.* **24**, 815–823 (2001).
 23. Noone, P. G. *et al.* Lung disease associated with the IVS8 5T allele of the CFTR gene. *Am. J. Respir. Crit. Care Med.* **162**, 1919–1924 (2000).
 24. Elsas, L. J., 2nd, Langley, S., Paulk, E. M., Hjelm, L. N. & Dembure, P. P. A molecular approach to galactosemia. *Eur. J. Pediatr.* **154**, S21–7 (1995).
 25. El-Seedy, A. *et al.* Consequences of partial duplications of the human CFTR gene on cf diagnosis: mutations or ectopic variations. *J. Cyst. Fibros.* **12**, 407–410 (2013).
 26. Miller, D. T. *et al.* ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1381–1390 (2021).
 27. Cottin, V. *et al.* Late CF caused by homozygous IVS8-5T CFTR polymorphism. *Thorax* **60**, 974–975 (2005).
 28. Wu, C.-C. *et al.* Mutation spectrum of the CFTR gene in Taiwanese patients with congenital bilateral absence of the vas deferens. *Hum. Reprod.* **20**, 2470–2475 (2005).