

國立臺灣大學醫學院基因體暨蛋白質體研究所

碩士論文

Graduate Institute of Medical Genomics and Proteomics

College of Medicine

National Taiwan University

Master Thesis

臺灣族群個人之人類白血球抗原與殺手細胞類免疫球
蛋白受體基因體組成分析

Analysis of Personal Human Leukocyte Antigen (HLA)
and Killer-cell Immunoglobulin-like Receptor (KIR)
Genomic Profiles in the Taiwanese Population

莊惠文

Hui-Wen Chuang

指導教授：許書睿 博士

Advisor: Shu-Jui Hsu Ph.D.

中華民國 113 年 1 月

January, 2024



Acknowledgements

回首這條充滿崎嶇的研究之路，我深感生命中有太多的感激之情，想藉此致以最深切的謝意。在這特別的時刻，我想要表達我最深切的感激之情，感謝你們一直以來的支持和陪伴。沒有你們，我無法想像這段旅程會是怎麼樣的，感激實驗室的同學，包括林俐玲、江鈺婷、李振宇、周庭暄、陳品瑄、曾宜萱、簡培苗等，你們為我帶來歡笑與激勵，在學術道路上互相扶持。實驗室裡的討論總是火花四射，彼此間的幫助更是令人感激不盡，我們除了學術的部份，也會互相分享生活中的喜怒哀樂，讓我能夠在笑聲中忘卻煩憂，在淚水中找到共感。

我也要衷心感謝我的導師許書睿老師，您的引導讓我能夠在研究中找到自我探索的方向。在我迷失的時候，您給予我溫暖的指引，使我能夠提升程式分析和資料解讀的能力。同樣感謝基蛋所的陳沛隆老師，您分享的基因體學知識豐富而深刻，成為我學術路上的重要支柱。感謝三位論文口試委員：陳倩瑜老師、楊雅倩老師以及許家郎博士，您的指導使我能夠有力地梳理論文內容。

對我研究生涯中有一個重要的會議跟一個重要的機會，在醫圖免疫會議上，雙週會的方式，有著4大項免疫的主題，在會議上每位老師的指導與指引讓我避免了走彎路，同學與學長姐們也都積極的提出問題跟建議，也在演示能力上磨練自我。擔任次世代定序、生物資訊學與基因體醫學的助教機會真的很可貴，每次上課都是學識的盛宴，讓我感受到自己不斷進步的可能性。特別感謝國家高速網路中心和台灣人體生物資料庫，提供我解決技術問題和進行實驗的支持，使我的研究得以順利進行。

最後，我要感謝支持我讀研究所的家人。你們的支持是我學業途中最大的動力。在未來，我期許自己能夠持續學習，不斷提升，以更接近理想的自己。感謝各位的陪伴，讓我在這段充實的時光中茁壯成長。





摘要

人類基因組中的殺手細胞類免疫球蛋白受體 (KIR) 基因很重要，但研究起來因為基因的複雜性所以具有挑戰性。這些基因透過與人類白血球抗原 (HLA) 相互作用來影響自然殺手 (NK) 細胞的活性。這種相互作用調節免疫反應並影響各種健康狀況，例如自體免疫疾病和器官移植結果。然而，KIR 基因的遺傳結構複雜，且受遺傳因素影響個體間拷貝數 (CN) 差異較大，使得 KIR 分型變得複雜。台灣族群特有的個體層級 KIR-HLA 組合資料缺乏歷史資料，影響了我們的研究方向。

我們的研究利用了各種工具，即 HLA-VBseq、HISAT-genotype、T1K、DRAGEN、Qztype、graphKIR 和 capKIR，來解釋來自各種來源的 KIR 和 HLA 基因資料，例如、合成資料、panel、全基因體定序 (WGS)、人類泛基因組參考聯盟 (HPRC) 資料庫、台灣人體生物資料庫 (TWB)，以及結合爬蟲收集的資料進行分析。由於每種工具的性能因資料集而異，因此對其功能進行了仔細評估。在 HLA 方面，T1K 和 HISAT-genotype 基因分型因其卓越的準確性和全面的分型功能而成為 HLA 分型性能較好的工具。在 KIR 方面，graphKIR 最適合 WGS 資料類型，capKIR 適合 panel 資料類型。

使用 graphKIR 工具，我們詳細了解了台灣族群中 KIR 基因變異和不同單倍型的分佈。對 TWB 的 1,492 個樣本的分析揭示了固有的遺傳異質性，其特徵是 A 單倍型的頻率異常高，其中 AA 單倍型是最常見的 (49.7%)，BB(4.62%)，BX(45.7%)。

隨後，我們利用等位基因頻率網路資料庫 (AFND) 對台灣幾個亞人群的 KIR 基因頻率進行了全面評估。我們觀察到特定基因的差異，是有關於人口差異和研究方法差異引起的變異的問題，展示了樣本人群內 KIR 基因的複雜動態。例如，我們看 AFND 數據，會發現一些顯著的頻率差異。以 K96 資料庫中的 KIR2DL1 為例。這個基因在台灣 KIR (K96) 亞族群中只出現在約 65.6% 的情況下，這一比例顯著低於其他資料庫。這種差異可能源於每個研究中樣本收案或 PCR-SSP primer 設計的差異。這些因素突顯了基因頻率差異如何為我們提供有關 KIR 基因動態性質的重要性。



為了加深我們對 KIR-HLA 交互作用對多種疾病易感性的潛在影響的了解，我們對 TWB 的參與者資料進行了一項關聯性研究。我們的研究揭示了 KIR-HLA(T1K) 組合與多種表型的關聯，並透過邏輯迴歸分析進一步驗證。儘管存在年齡和性別等潛在限制因素，但我們的觀察有助於了解 KIR 基因和 KIR-HLA 交互作用對台灣免疫相關疾病易感性的影響。

關鍵字：殺手細胞類免疫球蛋白受體、KIR 基因型鑑定、全基因體定序、生物資訊、KIR-HLA 組合、邏輯氏回歸



Abstract

The human genome's Killer-cell Immunoglobulin-like Receptor (KIR) genes are significant yet challenging to study due to their gene complexity. These genes modulate the activity of natural killer (NK) cells through interactions with Human Leukocyte Antigens (HLA), impacting various health conditions like autoimmune diseases and organ transplantation outcomes. KIR gene complexity arises from their varied copy number (CN) amongst individuals due to genetic factors, complicating KIR typing. Due to a lack of KIR-HLA-related data for the Taiwanese population, we see an opportunity to focus our research in this direction.

Our research utilized various tools, namely HLA-VBseq, HISAT-genotype, T1K, DRAGEN, Qztype, graphKIR, and capKIR, to interpret KIR and HLA gene data from diverse sources. These sources included synthetic data, panel, whole-genome sequencing (WGS), the Human Pangenome Reference Consortium (HPRC), Taiwan Biobank (TWB) data sets, and data collected through web crawling for analysis. Since every tool's performance shifts according to the data set employed, we meticulously evaluated their capabilities. T1K and HISAT-genotyping emerged as high-performing tools for HLA due to their superior precision and comprehensive genotyping features. Regarding KIR, graphKIR is most suited for WGS data types, while capKIR is befitting panel data types.

Using graphKIR, we have gained a clear and detailed picture of the variation in KIR genes and the distribution of different haplotypes within the Taiwanese population. Our analysis of 1,492 samples from TWB revealed a significant diversity in the genetic

makeup. Specifically, the AA haplotype was most common, seen in 49.7% of the population, followed by BB and BX haplotypes at 4.6% and 45.7%, respectively.

Subsequently, we conducted an exhaustive assessment of KIR gene frequencies within several Taiwanese subpopulations using the Allele Frequency Net Database (AFND). We observed the differential occurrence of specific genes that raised questions about variants due to population differences and methodological disparities, illustrating complex dynamics of KIR genes within sampled populations. For instance, if we look at the AFND data, we find some notable frequency discrepancies. Take the KIR gene 2DL1 in the K96 database as an example. This gene shows up in only about 65.6% of the Taiwanese KIR(K96) subpopulation, a percentage significantly lower than in other databases. This discrepancy could stem from differences in the enrollment of samples or variations in the design of PCR-SSP primers used in each study. These factors accentuate how gene frequency divergence can provide important insights into the dynamic nature of KIR genes.

To deepen our understanding of the potential influence of KIR-HLA interactions on susceptibility to multiple diseases, we conducted an association study on participant data from TWB. Our research uncovered associations of KIR-HLA(T1K) combinations with numerous phenotypes, further confirmed by logistic regression analysis. Despite prevailing potential limiting factors like age and gender, our observations contribute significantly to comprehending the influence of KIR genes and KIR-HLA(T1K) interactions on susceptibility to immune-related diseases in Taiwan.

Keywords: killer-cell immunoglobulin-like receptors, KIR genotype typing, whole genome sequencing, bioinformatics, KIR-HLA combinations, logistic regression

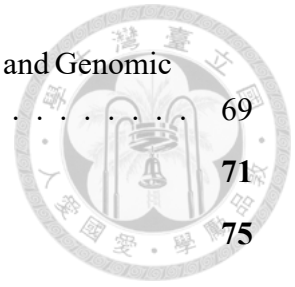


Contents

	Page
Acknowledgements	i
摘要	iii
Abstract	v
Contents	vii
List of Figures	xi
List of Tables	xiii
Denotation	xv
Chapter 1 Introduction	1
1.1 KIR	1
1.1.1 Nomenclature	1
1.1.2 Molecular structure and functions	3
1.1.3 Haplotypes	5
1.2 HLA	7
1.2.1 Nomenclature	7
1.2.2 Molecular structure and functions	9
1.2.3 Clinical role of HLA in health and disease	10
1.3 The combination between KIR and HLA	11
Chapter 2 Research Aims and Design	15
Chapter 3 Materials and Methods	17
3.1 Available datasets	17
3.2 Databases	17
3.2.1 AFND	17

3.2.2	IPD-HLA and IPD-KIR	18
3.2.3	TWB	20
3.3	KIR Software	21
3.3.1	PING2	21
3.3.2	capKIR	22
3.3.3	graphKIR	23
3.4	HLA Software	24
3.4.1	HISAT-genotype	24
3.4.2	VBSeq	25
3.4.3	T1K	27
3.4.4	DRAGEN	28
3.5	Web Crawler	29
3.6	Linux Server Hardware Configuration	30
Chapter 4 Results		33
4.1	Benchmarking HLA genotype by HPRC short read WGS dataset . . .	33
4.2	Designing WGS-based KIR genotyping protocol	40
4.3	KIR copy number typing in TWB	42
4.4	Taiwanese KIR gene frequency information in Allele Frequency Net Database (AFND)	51
4.5	Analyzing KIR-HLA pair frequencies	57
4.6	Studying the association of KIR-HLA(T1K) pairs with multiple phe- notypes using Association Studies	61
Chapter 5 Discussion		67
5.1	Discussion	67
5.1.1	Challenges with AFND Web Crawler Data Collection	67
5.1.2	Limitations and Strategies in Genotyping Tools	67
5.1.3	Influence of Age and Gender on Genotype-Phenotype Associations in Taiwanese Population	68

5.1.4	Considerations and Challenges in Phenotypic Analysis and Genomic Interpretation	69
Chapter 6 Conclusion		71
References		75



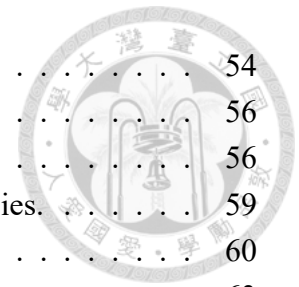




List of Figures

1.1	Leukocyte receptor complex (19q13.4) and KIR haplotypes.	2
1.2	The rule of KIR nomenclature.	2
1.3	KIR gene organization.	3
1.4	Structural characterization of KIR proteins.	4
1.5	KIR Polymorphism.	6
1.6	The rule of HLA nomenclature.	8
1.7	The HLA Complex.	10
1.8	The structure of HLA and the G group.	11
1.9	KIR-HLA Interactions.	13
2.1	Study design.	16
3.1	Trend of HLA Allele Discovery Over Time.	19
3.2	Trend of KIR Allele Discovery Over Time.	20
3.3	Sample Copy Number Figure Generated by PING2.	22
3.4	Depiction of the HISAT-genotype Algorithmic Process.	25
3.5	Schematic Diagram of the HLA-VBSeq Workflow.	27
3.6	The schematic diagram of web crawler.	30
4.1	Comparison of HPRC Tools in Two HLA Fields.	36
4.2	Comparison of 1A,2B,and 2A counts.	38
4.3	Distribution of Risk Allele Probabilities in the TWB.	39
4.4	Comparative Frequency of HLA Risk Alleles across Different Levels of Evidence Linked to Specific Chemicals.	39
4.5	Combining capKIR and graphKIR Schematic Diagram.	42
4.6	Distribution of Copy Numbers for Each KIR Gene.	46
4.7	Sorted Total Counts of KIR Genes Across Individuals.	47
4.8	Sorted Composition of AA, AB, BX Haplotypes among TWB Samples.	48
4.9	Proportional Representation of AA, AB, BX Haplotypes among TWB Samples.	48
4.10	Genotyping Results for the AA Haplotype.	49
4.11	Genotyping Results for the BB Haplotype.	49
4.12	Genotyping Results for the BX Haplotype.	50

4.13 High Gene Frequency of Selected KIR Genes.	54
4.14 Lower Gene Frequency of Specific KIR Genes.	56
4.15 Unique Frequency Representations for 3DS1 and 2DS4.	56
4.16 HLA (HISAT-genotype) Combinations With KIR Frequencies.	59
4.17 HLA (T1K) Combinations With KIR Frequencies.	60
4.18 Phenotype and their LLR Test Outcomes.	63
4.19 Phenotypes passing LLR P-Value Test	64
4.20 Phenotypes Not Passing LLR P-Value Test	65
4.21 Comparative Association Study Results with and without LLR P-value Filter	66





List of Tables

1.1	HLA Suffix.	8
3.1	HLA available datasets.	17
3.2	KIR available datasets.	17
3.3	The total of KIR alleles at version 2.12 (2022-12).	18
4.1	The Principal HLA Typing Tools Employed.	34
4.2	Distribution of Risk Alleles Among Taiwanese Individuals.	38
4.3	Distribution of 1A Risk Alleles in the Taiwanese Population	38
4.4	Comprehensive Comparative Overview of KIR Typing Tools	41
4.5	KIR Typing Results of graphKIR, capKIR, combine version in WGS and panel type on HPRC Datasets.	41
4.6	Distribution of KIR gene copy numbers in a cohort of 1492 individuals from TWB.	47
4.7	Recorded KIR Gene Information Sourced from the AFND and Relevant.	54





Denotation

AFND	Allele Frequency Net Database
BAM	Binary Alignment Map
bp	Base Pair
CN	Copy Number
capKIR	Capture-based Killer-cell Immunoglobulin-like Receptor Pipeline
EBM	Evidence-Based Medicine
EM	Expectation-Maximization
FW gene	Framework gene
GATK	Genome Analysis Toolkit
HLA	Human Leukocyte Antigen
HIFI-read	High Fidelity Read
HPRC	Human Pangenome Reference Consortium
IGV	Integrative Genome Viewer
IMGT	The International ImMunoGeneTics Information System
KIR	Killer-cell Immunoglobulin-like Receptor
MHC	Major Histocompatibility Complex
MSA	Multiple Sequence Alignment

NCHC	National Center for High-Performance Computing
NGS	Next Generation Sequencing
NK	Natural killer
PCR	Polymerase Chain Reaction
PCR-SSOP	Sequence-specific Oligonucleotide Probes
PCR-SSP	Sequence-specific Primer PCR
PING	Pushing Immunogenetics to the Next Generation
SKIRT	Structural KIR AnnoTator
TWB	Taiwan Biobank
WES	Whole-Exome Sequencing
WGS	Whole Genome Sequencing
β 2m	β 2-Microglobulin



Chapter 1 Introduction



1.1 KIR

Killer cell immunoglobulin-like receptors (KIRs) are a family of cell surface proteins found on natural killer (NK) cells, a type of lymphocyte involved in the immune response (Zinkernagel and Doherty, 1974). KIRs regulate NK cell activity by interacting with primary histocompatibility complex (MHC) molecules on target cells. This interaction can activate or inhibit NK cell cytotoxicity depending on the specific KIR-MHC combination (Asl et al., 2021).

The nomenclature of KIR genes is based on their protein structure, describing their extracellular and intracellular domains (Marsh et al., 2003). The KIR gene family includes 15 functional genes: *KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL4*, *KIR2DL5A*, *KIR2DL5B*, *KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5*, *KIR3DL1*, *KIR3DL2*, *KIR3DL3* and *KIR3DS1*. In addition, the KIR gene complex has two pseudogenes, *KIR2DP1* and *KIR3DP1* (Figure1.1).

1.1.1 Nomenclature

KIR receptors are designated based on the combination of their extracellular Ig-like domains (2D or 3D) and the length of their cytoplasmic tail (long (L), short (S), or pseudogene (P)). Differentiating factors for KIR receptors with the same extracellular domains and cytoplasmic tail length is indicated by a numerical identifier. Allelic variants are denoted by an asterisk (Figure1.2).

Genetic variations, such as substitutions, insertions, or deletions, can alter the termination site of the KIR gene, resulting in either a long or short cytoplasmic tail. These nucleotide changes have a profound impact on the function of KIR receptors. Except for *KIR2DL4*, which has dual inhibitory and activating capabilities, KIR receptors with long cytoplasmic tails are inhibitory, while those with short tails are activating.

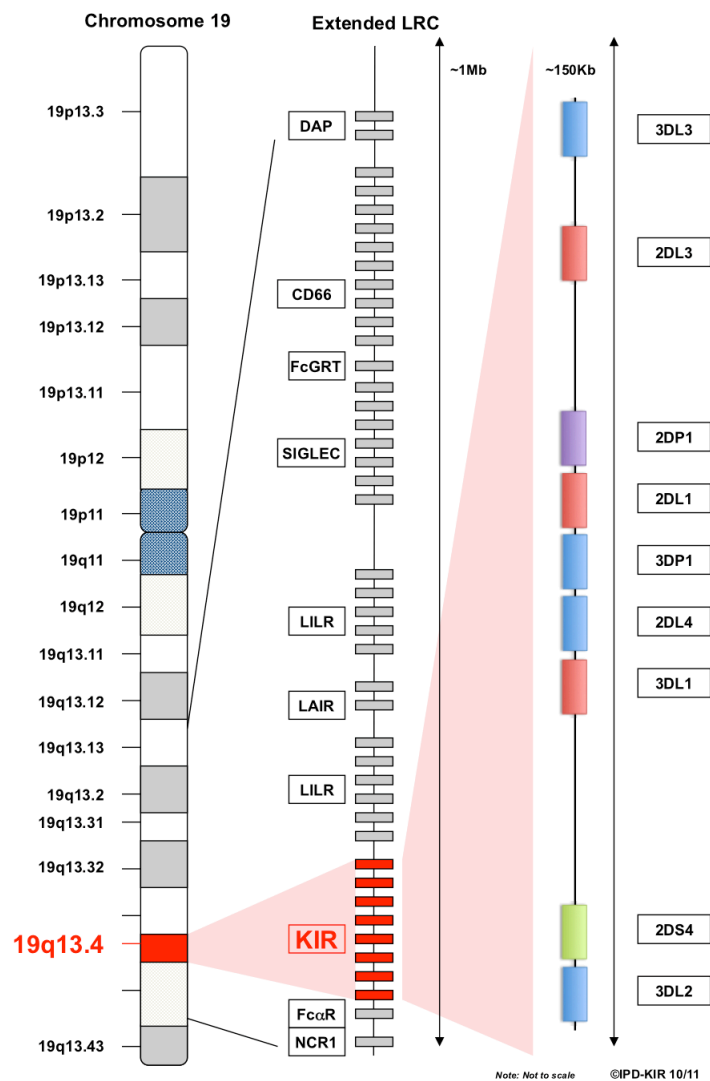


Figure 1.1: Leukocyte receptor complex (19q13.4) and KIR haplotypes. The Leukocyte Receptor Complex (LRC) is a genomic region spanning 100-200 Kb on chromosome 19 (19q13.4). This figure is from <https://www.ebi.ac.uk/ipd/kir/about/>

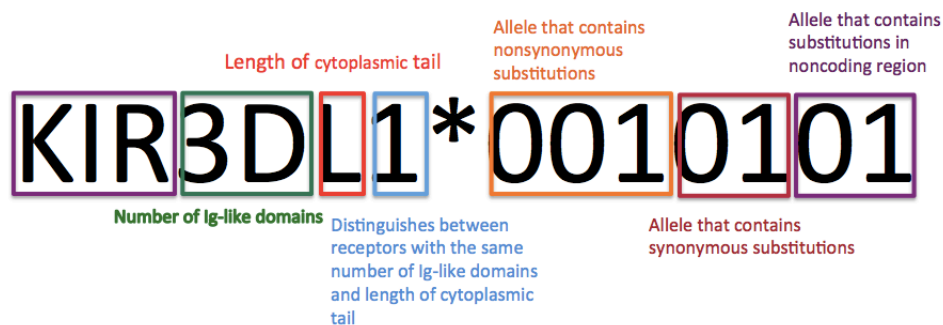


Figure 1.2: The rule of KIR nomenclature. It included Ig-like domains, cytoplasmic rail, and seven digits for expressing DNS substitution. This figure is from https://en.wikipedia.org/wiki/Killer-cell_immunoglobulin-like_receptor

1.1.2 Molecular structure and functions

The KIR gene is located on chromosome 19 and is highly polymorphic, indicating significant genetic diversity. The KIR gene complex spans a gene region approximately 4 to 16 kb in length and consists of 4-9 exons (Figure 1.3).

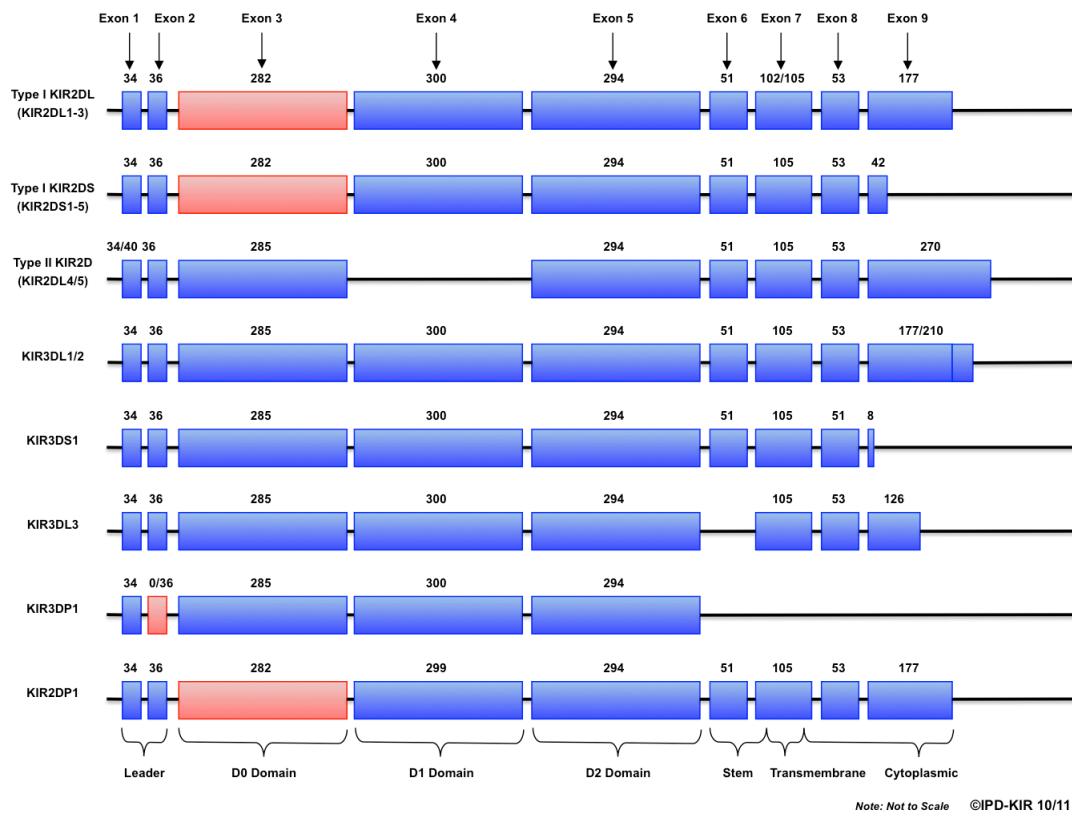


Figure 1.3: KIR gene organization.

Pseudo-exons, marked in red, exist within the genome but are excised post-transcription into RNA. As a result, these pseudo-exons are absent from the coding sequence (CDS). The diagram was sourced from <https://www.ebi.ac.uk/ipd/kir/about/>

Extracellular immunoglobulin-like domains characterize the molecular structure of KIR proteins. Based on their structure, KIR genes can be divided into type 1 KIR2D, type 2 KIR2D genes, and KIR3D genes. Type 1 KIR2D genes comprise two extracellular immunoglobulin-like domains, D1 and D2. Type 2 KIR2D genes have D0 and D2 domains. The KIR3D gene contains all three domains: D0, D1, and D2. The presence or absence of specific domains affects the binding specificity of KIR proteins to MHC-I molecules.

Another essential feature of KIR proteins is the length of their cytoplasmic tails. KIR receptors can be classified as long-tailed (KIR2DL or KIR3DL) or short-tailed (KIR2DS

or KIR3DS) by the length of their cytoplasmic domain. Long-tailed receptors contain an inhibitory signaling motif called the immunoreceptor tyrosine-based inhibitory motif (ITIM), which transmits an inhibitory signal when binding to MHC-I molecules. In contrast, short receptors lack ITIM and can transmit activating signals leading to NK cell activation (Figure 1.4) (Farag et al., 2002; Vilches and Parham, 2002).

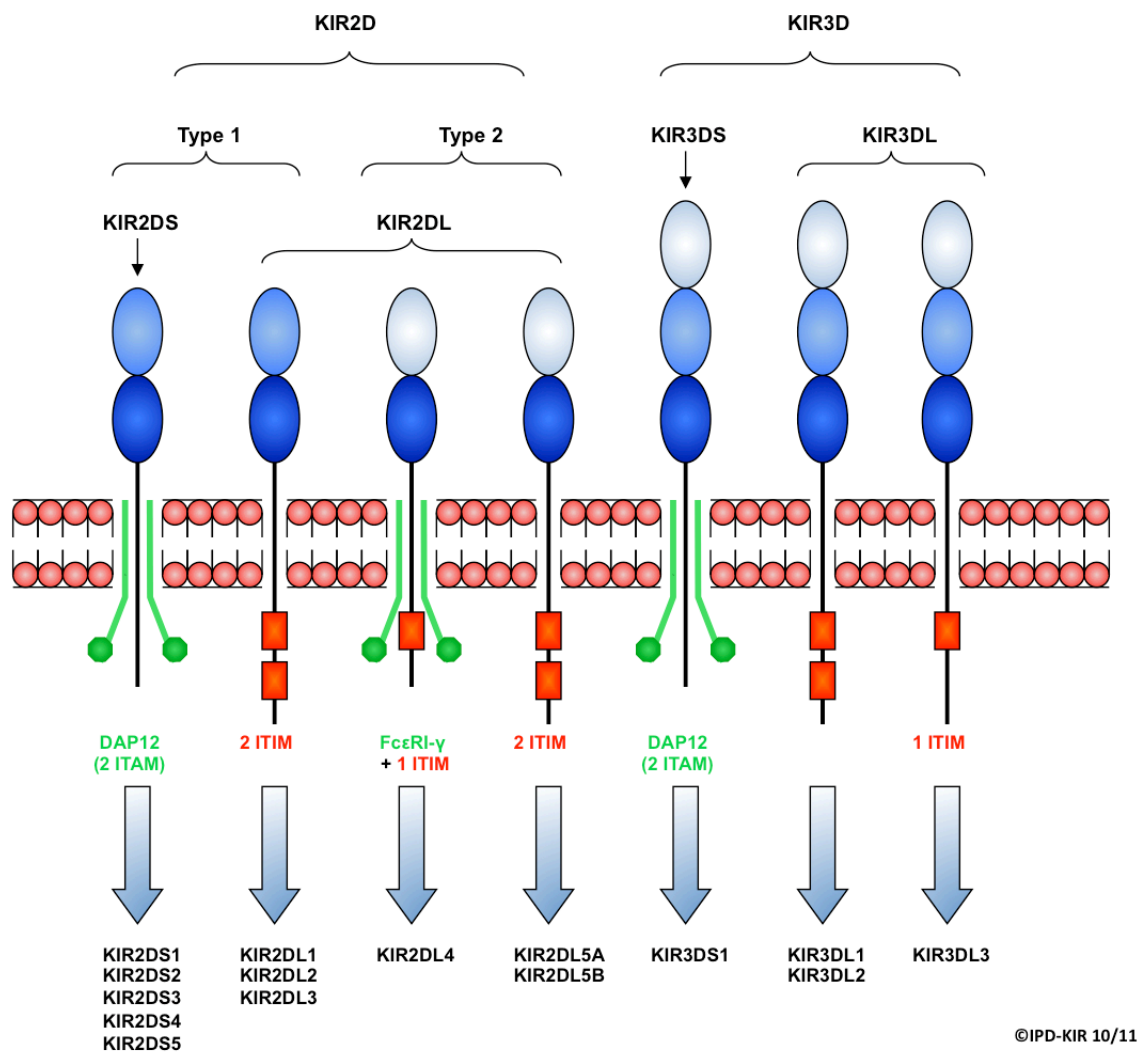


Figure 1.4: Structural characterization of KIR proteins.

KIR proteins consist of two to three immunoglobulin-like domains. The binding of activating KIRs with adapter molecules is indicated in green, and the ITIMs of inhibitory KIRs are highlighted in red boxes. The source of this illustration is <https://www.ebi.ac.uk/ipd/kir/about/>

1.1.3 Haplotypes



The KIR gene is highly polymorphic. Individuals show marked variation in the type and number of copies of the KIR gene they possess. Based on previous literature studies, the researchers defined two haplotypes, A and B, based on the distribution of common genotypes. Genes shared by both haplotypes are called framework genes, including *KIR3DL3*, *KIR3DP1*, *KIR2DL4*, and *KIR3DL2*. Framework genes are usually present in two copy numbers and are less prone to deletion and duplication events, which lead to copy number variation. However, between *KIR3DP1* and *KIR2DL4* has a recombination hotspots. Therefore, framework genes may not always have two copy numbers (Martin et al., 2000; Hsu et al., 2002).

Haplotype A is relatively fixed and mainly consists of inhibitory KIRs. It comprises the above four framework genes and *2DL3*, *2DP1*, *2DL1*, *3DL1*, and *2DS4*. Haplotype B showed more variation in the number and combination of KIR genes. It is characterized by one or more of the following genes: *2DL2*, *2DL5A/B*, *2DS1*, *2DS2*, *2DS3*, *2DS5*, and *3DS1*. Possession of any of these genes defines an individual as having haplotype B.

The KIR gene content is highly polymorphic, and a recombination hotspot exists between *3DP1* and *2DL4*. Therefore, based on this boundary, KIR haplotypes are often divided into centromeric and telomeric regions. The centromeric region is located between the framework genes *3DL3* and *3DP1*, near the center of the chromosome. KIR genes may appear in this region, including *2DL1*, *2DL2*, *2DL3*, *2DL5B*, *2DS2*, and *2DP1*. The telomeric region is located between the framework genes *2DL4* and *3DL2*, near the end of the chromosome. Possible KIR genes in this region include *2DL5A*, *2DS1*, *2DS4*, *3DL1*, and *3DS1*. Additionally, *2DS3* and *2DS5* are more particular and can exist in centromere and telomeric regions (Roe et al., 2017).

Compared with haplotype A, haplotype B contained more genes and showed more variation in the number and combination of genes (Kulkarni et al., 2008). As a result, the centromeric region of haplotype B can be further divided into B1 and B2. These two KIR haplotypes can give rise to haplogroup A/A and haplogroup B/X, respectively(Figure1.5).

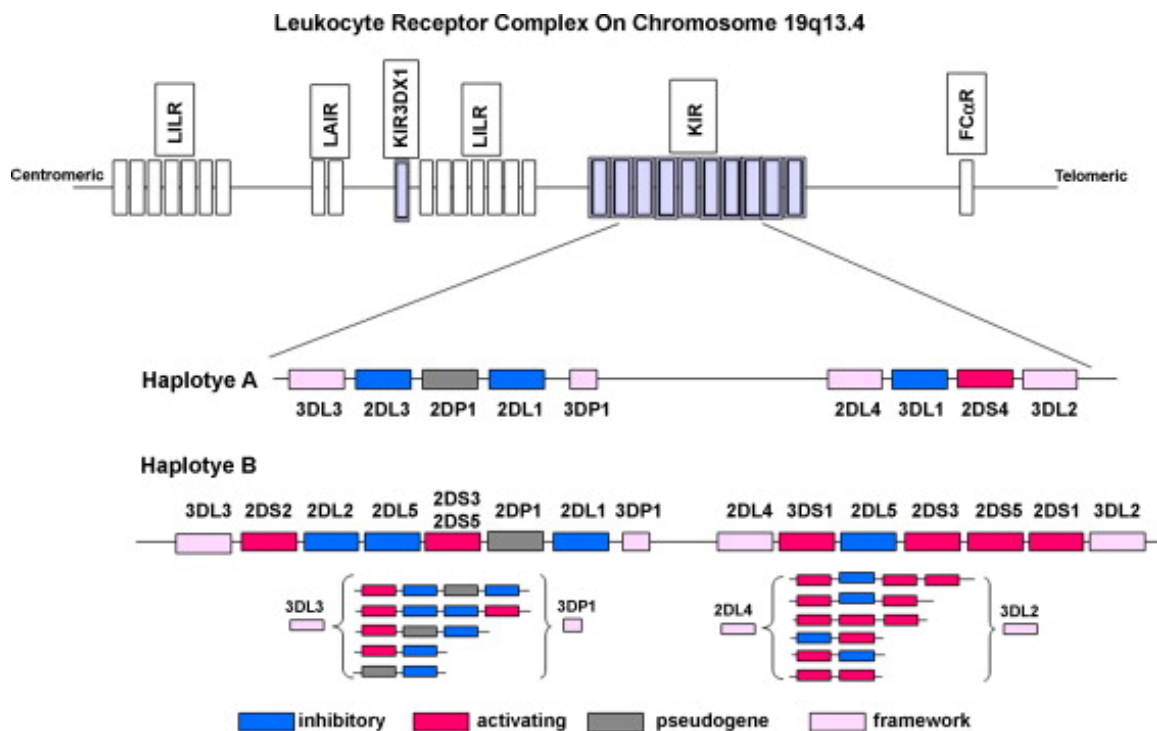


Figure 1.5: KIR Polymorphism.

The KIR gene is composed of two segments, the centromere and the telomere, which vary in their genetic composition. Two primary KIR haplotypes exist, with the A haplotype possessing a constant genetic content. Conversely, the B haplotype contains a fluctuating amount of genes, which is indicated in parentheses. Each haplotype incorporates framework genes, displayed in pink boxes.

1.2 HLA

1.2.1 Nomenclature



The human leukocyte antigen (HLA) nomenclature follows a format to identify and differentiate alleles within the human MHC genes. Here is an explanation of the different components(Figure1.6) ([Marsh et al., 2010](#)).

- HLA: This fixed prefix indicates that the allele is part of the human MHC genes.
Gene Name: The gene name follows the prefix, such as HLA-A or HLA-B, and identifies which gene the allele modifies.
- Antigen Type: The first two numbers after the gene name (e.g., HLA-A*02) represent the antigen type associated with that allele. This designation typically corresponds to the serological antigen present. Alleles with the same antigen type will not react with each other in serological tests.
- Protein Coding: The next set of digits (e.g., HLA-A*02:01) indicates the protein the allele codes. These numbers are assigned sequentially based on the order of their discovery. Any change in this number represents a different protein resulting from a nucleotide change that replaces an amino acid with another.
- Allele Variant: The third set of numbers (e.g., HLA-A*02:01:01) indicates an allele variant with a different DNA sequence that produces the same protein as the normal gene. These variants represent synonymous mutations in the coding region.
- Non-Coding Region Polymorphism: The fourth set of numbers (e.g., HLA-A*02:01:01:02) designates single or multiple nucleotide polymorphisms in a non-coding gene region.
- Suffix: In addition to the specific allele number, there are additional optional suffixes that may be added to an allele to indicate its expression status(Table 1.1).

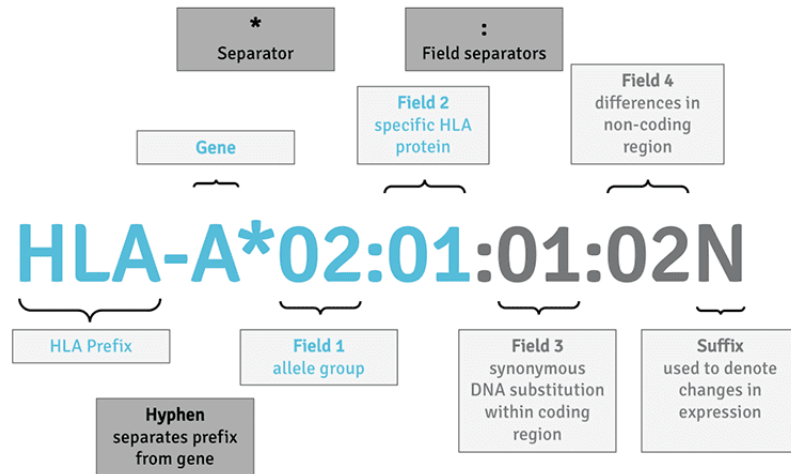


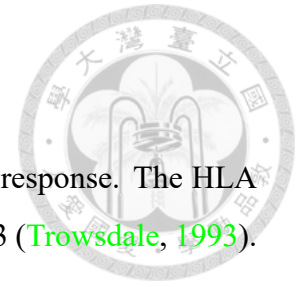
Figure 1.6: The rule of HLA nomenclature.

The HLA nomenclature system allows for organized recognition and distinction between alleles based on these elements. It provides a uniform and standard method for naming the various HLA alleles found in human MHC genes. This illustration comes from <https://www.hlaprotein.com/hla-technology/>.

Table 1.1: HLA Suffix.

Letter	Significance
N	Null allele (produces a non-functional protein)
L	Lower than regular cell surface expression
S	Soluble protein not found on cell surface
Q	Questionable (allele may affect regular expression)
C	Protein that is present in cytoplasm but not cell surface
A	Aberrant expression (uncertain if a protein is expressed)

1.2.2 Molecular structure and functions



The HLA region contains many genes involved in the immune response. The HLA region is located on the short arm of chromosome 6 at position 6p21.3 ([Trowsdale, 1993](#)). Class I, Class II and Class III comprise three areas (Figure 1.7).

- **HLA Class I**

Class I regions include the HLA-A, HLA-B, and HLA-C genes. These molecules are present on the surface of almost all nucleated cells in the body. Class I molecules play an important role in the presentation of antigens to CD8 T cells and are recognized in allograft rejection ([Choo, 2007](#)).

- **HLA Class II**

The class II region contains the HLA-DP, HLA-DQ, and HLA-DR genes. Class II molecules are constitutively expressed on antigen-presenting cells (APCs) and can be induced on other cell types during inflammation. Class II molecules consist of alpha (heavy) and beta (light) chains encoded by genes within the MHC. HLA-DR molecules have an invariant alpha chain, while the beta chain exhibits extreme polymorphism. The number of HLA-DR genes may vary among individuals, and additional genes such as HLA-DRB3, HLA-DRB4, and HLA-DRB5 may be expressed on a single haplotype ([Touraine et al., 1989](#)).

- **HLA Class III**

The class III region is located between the class I and II regions. It contains genes related to the immune response, such as complement components (C2, C4, and B factors), tumor necrosis factor (TNF), heat shock proteins, and other genes with immune functions. Genetic variants within class III regions have been associated with certain diseases, including systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) ([Gruen and Weissman, 2001](#)).

- **HLA G group**

The HLA G group refers to a subset of HLA genes called HLA-G. The HLA-G gene is located in the human genome's major histocompatibility complex (MHC) region and plays a unique role in immune regulation (Figure 1.8) ([Grubic, 2020](#)).

Group G distinguishes HLA alleles that have a or multiple synonymous variants. Group G distinguishes an HLA allele from one or more synonymous variants, mainly about the reason for antigen binding.

The HLA-G group refers to the HLA genes named HLA-G, which are part of the HLA system genes. These genes have unique properties and play specific roles in immune regulation, especially in pregnancy and certain diseases. It is particularly clinically significant (Apps et al., 2008).

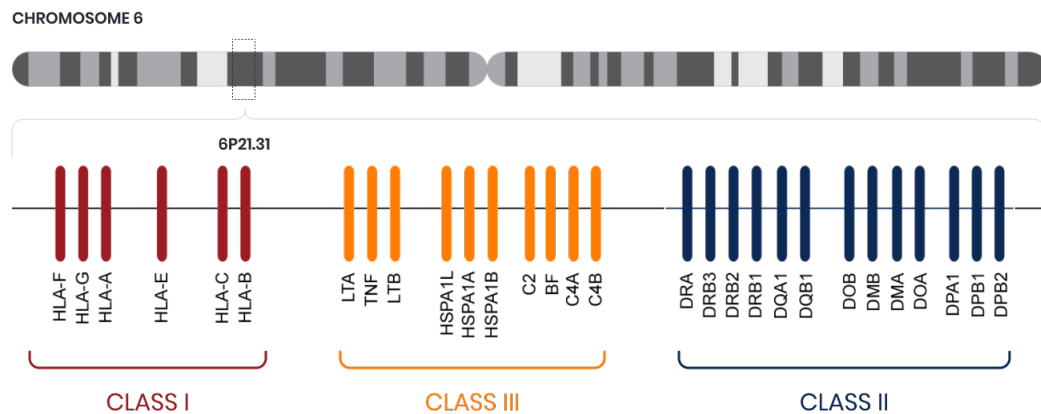


Figure 1.7: The HLA Complex.

Many key genes on chromosome 6 are involved in immune function, disease defense, and transplant compatibility. The figure is from <http://sciscogenetics.com/technology/human-leukocyte-antigen-complex/>.

1.2.3 Clinical role of HLA in health and disease

HLA complexes belong to a family of highly polymorphic genes involved in immunity and are present in disease-associated regions of the human genome. HLA regions are associated with hundreds of diseases, including autoimmune disorders, cancer, and infectious diseases. Recent advances in genotyping arrays and sequencing technologies allow better characterization of extreme polymorphisms in HLA (Gao et al., 2019).

Regarding health, HLA molecules are critical for maintaining immune homeostasis and preventing infection. They enable the immune system to distinguish self-antigens from non-self antigens, allowing a targeted immune response against pathogens while avoiding harmful reactions to the body's cells. HLA typing is commonly performed in transplant medicine to match donors and recipients, as HLA compatibility is critical for successful organ and tissue transplantation (Mahdi, 2013).

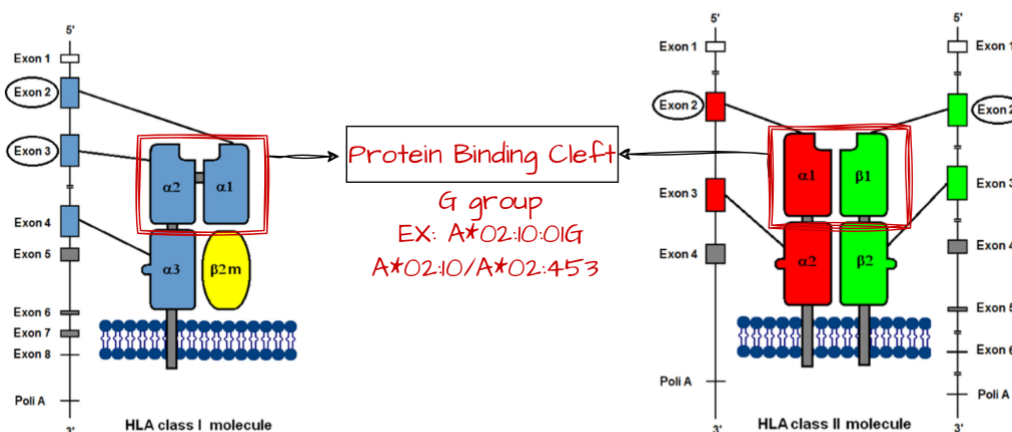


Figure 1.8: The structure of HLA and the G group.

The peptide-binding region of the HLA class I molecules, formed by the $\alpha 1$ and $\alpha 2$ domains, is encoded by exons 2 and 3 of the HLA class I genes. In contrast, within HLA class II molecules, the peptide-binding region consists of the $\alpha 1$ and $\beta 1$ domains, encoded by exon 2 of the HLA class II genes. The diagram notes the exons responsible for encoding these peptide-binding domains.

In infectious diseases, the frequency and presence of HLA alleles may vary in different populations. Certain alleles that confer resistance to specific pathogens are more prevalent in endemic areas. Genomic analysis also helps map and identify loci associated with various diseases.

HLA alleles play a crucial role in the immune response to infectious diseases. The clinical roles of HLA in health and illness are manifold. Extensive polymorphisms in HLA genes lead to individual differences in immune responses and disease susceptibility. Understanding the complex interplay between HLA and the immune response is critical to advancing personalized medicine and improving patient outcomes in various clinical settings (Mosaad, 2015).

1.3 The combination between KIR and HLA

KIR and HLA genes are highly polymorphic in humans. Many studies have found that KIR and HLA will combine to regulate the function of NK cells and the immune system (Figure 1.9). Different combinations of KIR receptors and HLA molecules lead to different susceptibility and immune responses to certain diseases in different ethnic

groups. Numerous studies have demonstrated the relevance of KIR genotypes and KIR-HLA combinations with disease interactions in various human diseases such as rheumatoid arthritis, ankylosing spondylitis, systemic lupus erythematosus, psoriatic arthritis, viral infections such as HIV ([Rajalingam, 2016](#)), hepatitis B, and C), preeclampsia ([Yang et al., 2020](#)) and cancer ([Leone et al., 2017](#)).

In addition to disease associations, knowledge of KIR-HLA combinations has also been applied to clinical treatment. For example, the specific monoclonal antibody IPH2102 has been used to treat acute myeloid leukemia (AML), multiple myeloma, and solid tumors ([Nijhof et al., 2015](#)). Furthermore, KIR-HLA matching has potential implications in solid organ transplantation and hematopoietic stem cell transplantation (HSCT) ([Ludajic et al., 2009](#)). Some studies have even shown that the influence of KIR-HLA interactions can predict post-transplant outcomes. For example, when AML patients receive hematopoietic stem cells from unrelated donors with specific KIR-HLA combinations, it can reduce the risk of relapse and improve overall clinical outcomes.

Briefly, combining KIR and HLA genes exhibits significant diversity and has implications for clinical transplantation and disease mechanisms. Examining the frequency of KIR-HLA pairings in different populations can provide insight into the role of KIRs in clinical transplantation and disease mechanisms. These associations parallel the initial observations made with HLA polymorphisms and disease, highlighting the importance of delving deeper into the underlying mechanisms. Finding out the specific combinations of KIR and HLA alleles is crucial for unraveling the mechanisms governing immune responses, disease susceptibility, and the development of therapeutic interventions. This knowledge provides valuable insights into the intricate interplay between the immune system and diseases, paving the way for personalized medicine and improved clinical outcomes.

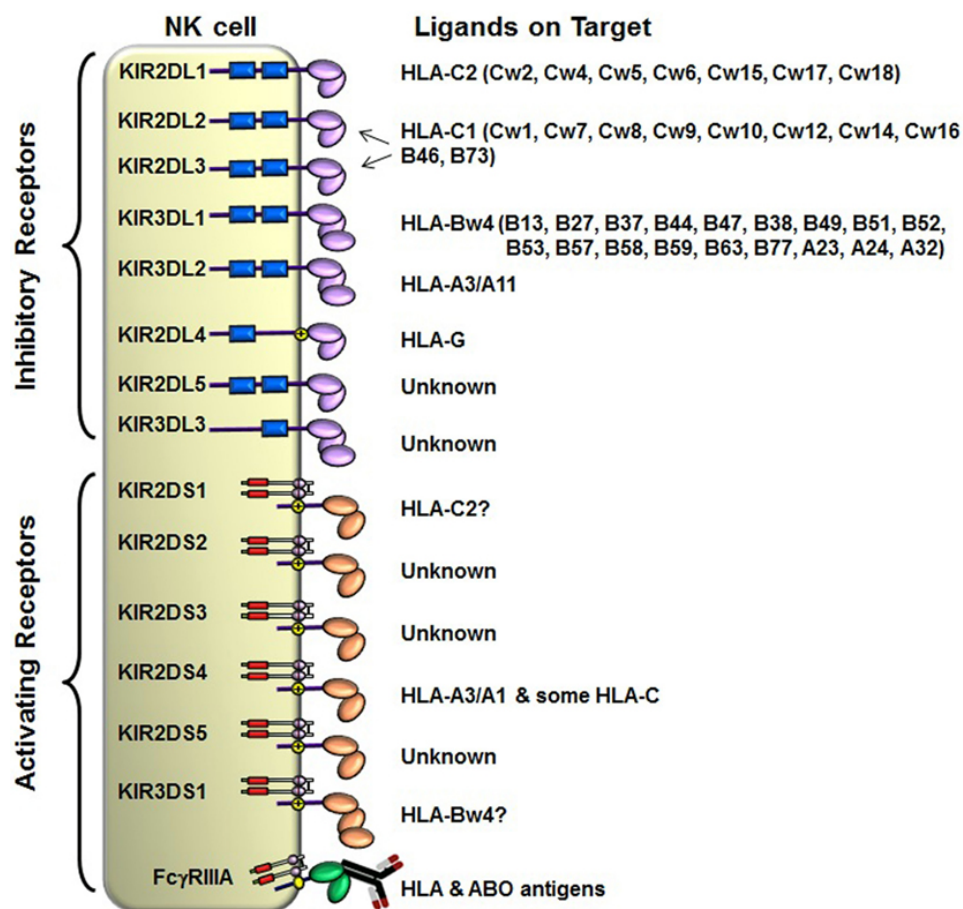


Figure 1.9: KIR-HLA Interactions.

They are unveiling the complexities of immune genes in transplantation and disease. Insights into pairings provide opportunities for personalized medicine and improved clinical outcomes ([Rajalingam, 2016](#)).





Chapter 2 Research Aims and Design

Objective:

The primary objective of this study is to determine the genetic landscape of the HLA allele and KIR copy number distribution in the Taiwanese population. Additionally, the study aims to identify potential clinical implications of HLA and KIR genes specific to the Taiwanese population (Figure 2.1).

1. **Data and Tool Collection:** The project must obtain a diverse range of data types, including panel, WGS, assembly, and simulated data. Additionally, selecting appropriate typing tools from various options is crucial. The HPRC dataset, containing assembly data, is pivotal in determining the most suitable tools for KIR and HLA typing of TWB NGS data.
2. **HLA Typing:** The accurate typing of HLA alleles involves utilizing tools such as HISAT-genotype, T1K, Dragon, hla_annotate, Qztype, and VBSeq. The analysis of the HPRC dataset enhances the precision of HLA typing. Furthermore, allele-level HLA typing is carried out on the TWB dataset using HISAT-genotypes and T1K.
3. **KIR Copy Number Determination:** To measure KIR gene copy numbers with precision, a dedicated capKIR pipeline is developed. CapKIR and graphKIR are evaluated using the HPRC dataset, incorporating the correct KIR answers obtained from the SKIRT tool. Finally, graphKIR, a suitable tool for KIR typing within the TWB dataset, is applied to determine KIR copy numbers for all 1,492 individuals.
4. **Combination of HLA and KIR Data:** I am integrating allele-level HLA typing results from HISAT-genotype and T1K with KIR copy number data derived from graphKIR.
5. **Phenotype Association Analysis:** Phenotypic information is extracted from the TWB dataset, and association analyses between combined HLA-KIR genotypes and phenotypic features are conducted using Fisher's exact tests.

6. **Interpretation and Significance:** The project concludes with the analysis and interpretation of the association results, shedding light on the potential clinical significance of HLA and KIR genes within the context of various phenotypic features. Additionally, valuable insights are provided into personalized HLA/KIR genomic profiles among individuals in the Taiwanese population.

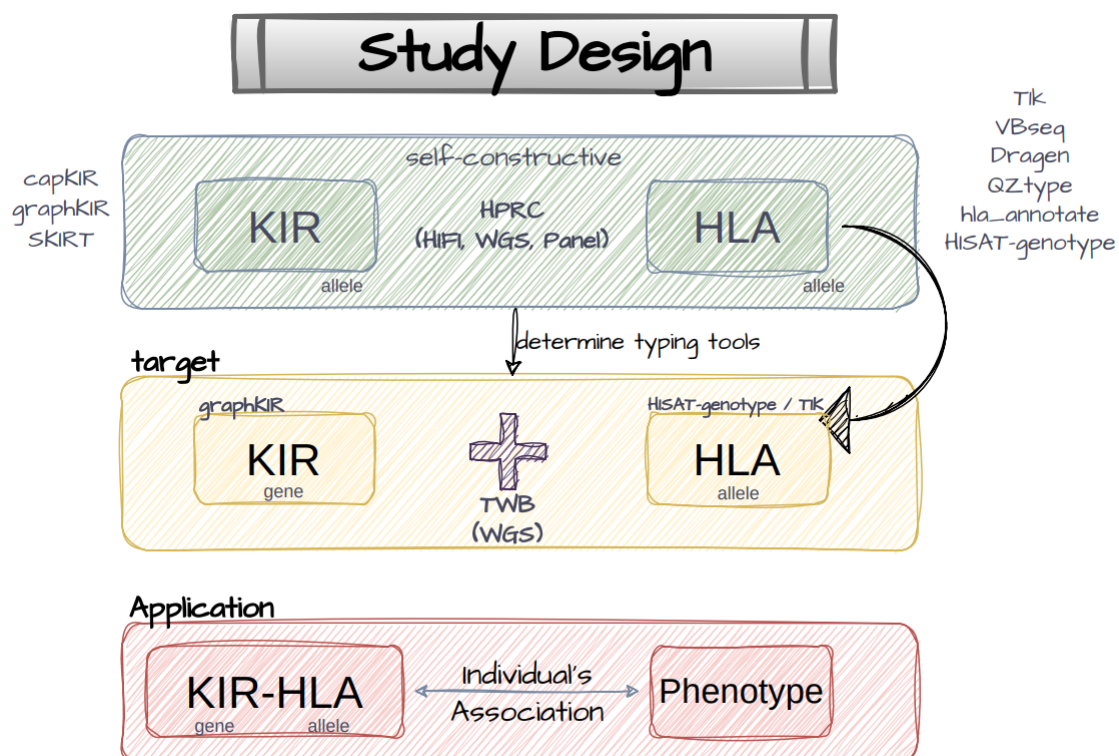


Figure 2.1: Study design.

The genetic landscape of HLA and KIR is revealing the genomic diversity of the Taiwanese population. Analysis of different data types and tools, such as HISAT-genotype and graphKIR, revealed the clinical significance of HLA and KIR genes. Integration of HLA-KIR data identifies personalized genomic profiles and reveals phenotypic associations.



Chapter 3 Materials and Methods

3.1 Available datasets

• HLA

The table records the currently available HLA datasets, as well as the number and type of classified samples, whether there is a standard answer, what software is used for analysis, and who performs the analysis, etc(Table 3.1).

• KIR

The table records the currently available KIR datasets, as well as the number and type of classified samples, whether there is a standard answer, what software is used for analysis, and who performs the analysis, etc((Table 3.2).

Table 3.1: HLA available datasets.

Sample set	Number	Data type	Answer	Software	Data source	DNA	Operator
HPRC	44	HiFi read and Contig (TGS)	allele	Annotate	HPRC	total	Sheng-Kai
	44	capture-based (new version)	allele	HISAT-genotype, VBSeq, Qztype, T1K	PL Lab	total	Sheng-Kai, Winnie
	44	WGS(NGS)	allele	HISAT-genotype, VBSeq, DRAGEN, T1K	NCBI	total	Winnie

Table 3.2: KIR available datasets.

Sample set	Number	Data type	Answer	Software	Data source	DNA	Operator
HPRC	47	HiFi read and Contig(TGS)	allele	SKIRT	HPRC		Calvin
	44	WGS (NGS)	allele	graphKIR	NCBI	total	Winnie
	44	capture-based (new version)	gene	gKIR	PL Lab		Nilson, Winnie
IHWG	56	capture-based (old version)	gene(2)	gKIR	PL Lab	total	Winnie
TWB	80 (82-2)	WGS (NGS)	gene(SSP)	graphKIR	TWB	some	Jeff, Winnie
Synthetic	100	synthetic	allele	graphKIR	Linnil1		Winnie
Synthetic	50	synthetic	allele	graphKIR	paper		Winnie

3.2 Databases

3.2.1 AFND

The Allele Frequency Net Database (AFND) is a publicly available database that serves as a comprehensive collection of frequency data for various immune-related genes

Table 3.3: The total of KIR alleles at version 2.12 (2022-12).

KIR Alleles									
Gene	2DL1	2DL2	2DL3	2DL4	2DL5A	2DL5B	2DS1	2DS2	2DS3
Alleles	185	35	68	114	45	47	33	65	71
Gene	2DS4	2DS5	3DL1	3DL2	3DL3	3DS1	2DP1	3DP1	Total
Alleles	41	88	189	168	229	91	40	108	1617

in the human genome. It was originally designed to incorporate polymorphisms in HLA, KIR, and cytokine polymorphisms information (Gonzalez-Galarza et al., 2018).

The goal of the database is to provide a central resource where researchers can contribute their frequency data while allowing users to search and access existing information. Collected data are presented in allelic, haplotype, and genotype formats, enabling comprehensive analysis of genetic variation at these important immune-related loci. The allele frequency net database (<http://www.allelefrequencies.net/hla6006a.asp>) is 2020 update. support@allelefrequencies.net (Gonzalez-Galarza et al., 2020)

3.2.2 IPD-HLA and IPD-KIR

IPD-HLA (Barker et al., 2023) and IPD-KIR (Robinson et al., 2012) are official databases that store information on genetic polymorphisms of the human immune system. These databases play a crucial role in collecting and organizing data related to allelic and sequence diversity of HLA and KIR genes.

The IPD-KIR database focuses exclusively on KIR genes, providing a comprehensive collection of star allele names and their corresponding sequences (Figure 3.2). The database is constantly expanding, and the latest version (2022-12, version 2.12.0) contains 1,617 alleles, including full-length and partially sequenced variants.

In the Table 3.3, we can see the total number of alleles for each KIR gene in the IPD-KIR library to date. The information has been obtained from <https://www.ebi.ac.uk/ipd/kir>.

As depicted in the figure, it is evident that an increasing number of researchers are delving into the field of KIR and HLA immunogenetics. This has led to the rapid finding

of new alleles, and databases have expanded astonishingly.

The count of HLA alleles, designated by WHO Nomenclature Committee for Factors of the HLA System and recorded in the IPD-IMGT/HLA Database, reveals a consistent uptrend in newly identified class I and class II alleles over the past three decades (Figure 3.1).

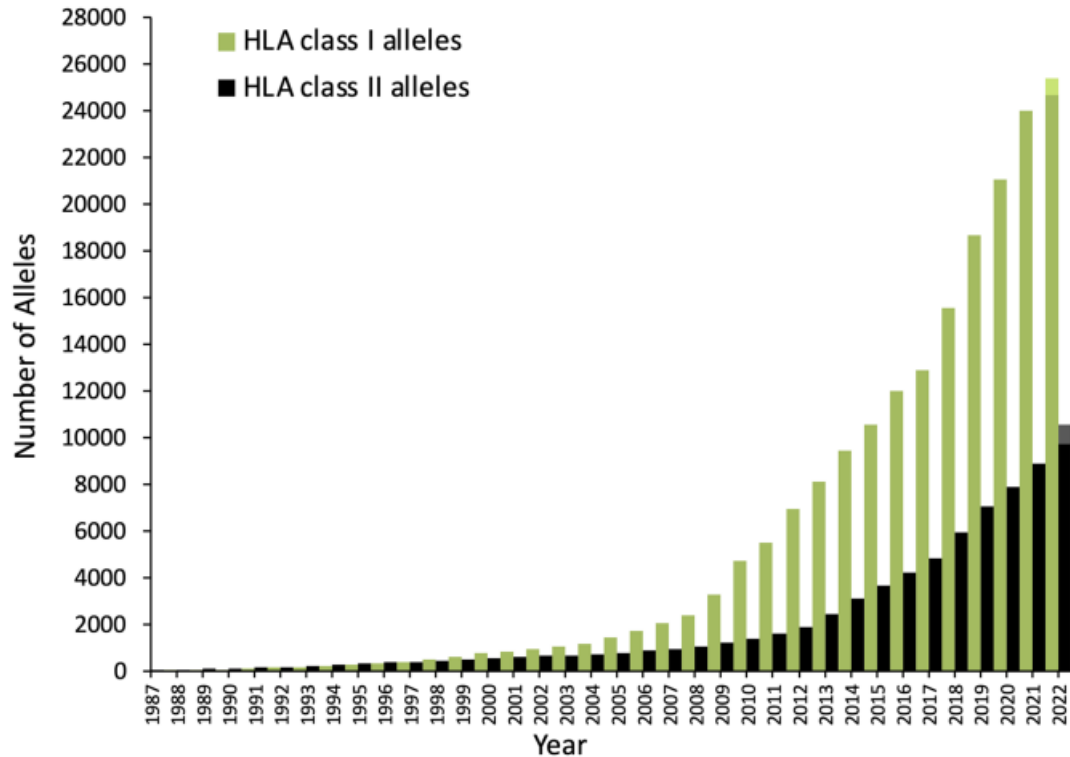


Figure 3.1: Trend of HLA Allele Discovery Over Time.

The momentum of novel allele discovery is amplified with the integration of NGS technologies, leading to a higher number of new alleles being christened in just the initial quarter of 2019 as compared to the entirety of the previous century. This figure is from <https://academic.oup.com/nar/article/48/D1/D948/5610347>.

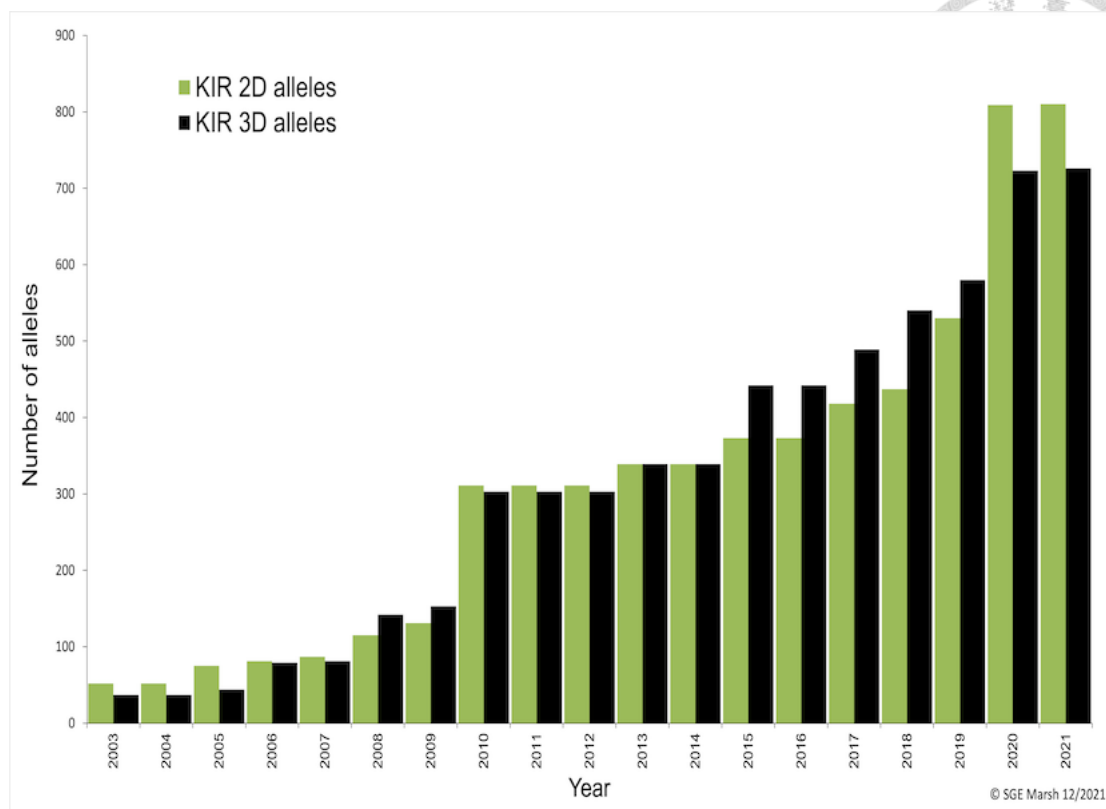


Figure 3.2: Trend of KIR Allele Discovery Over Time.

The following chart represents the count of alleles that have been identified and housed within the IPD-KIR Database, spanning from 2003 through December 2021. This figure is from <https://www.ebi.ac.uk/ipd/kir/about/statistics/growth/>.

3.2.3 TWB

Taiwan Biobank (TWB), the largest human biobank in Taiwan, was established in 2012 as a research project of Academia Sinica (Fan et al., 2008). It aims to collect and analyze biological samples from the Taiwanese population to facilitate various studies and research projects. TWB has been actively recruiting participants since December 2008 and as of December 2022, has collected data on more than 180,000 people. It is an invaluable resource for studying all aspects of human health and genetics. TWB generated WGS data by paired-end sequencing using Ion Proton, Illumina HiSeq2500 / HiSeq4000 / NovaSeq6000 platforms.

In this study, we analyzed KIR and HLA genotyping of 1492 individuals using Illumina-based WGS from TWB. Ethical approval for this study was obtained from the Institutional Review Board of the Affiliated Hospital of National Taiwan University College of Medicine (IRB approval number: TWBR11106-05).

3.3 KIR Software



3.3.1 PING2

PING (Norman et al., 2016) and PING2 (Marin et al., 2021), short for "Pushing Immunogenetics to the Next Generation," is a sophisticated tool designed to determine KIR gene copy numbers and perform allele-level genotyping with a precision of 5 digits. This tool has revolutionized the field of immunogenetics by offering a multi-step approach that involves three key modules: PING_extractor, PING_copy, and PING_allele.

The PING_extractor module utilizes a read mapping process, selectively retaining all reads mapped to a pre-defined KIR haplotype. Following this, the PING_allele module meticulously counts the number of retained reads that align perfectly with specific probes. PING2 calculates the gene's depth by determining the average probe count within the gene and normalizing it against KIR3DL3. Through the collective read counts across all samples within a cohort, PING2 expertly determines copy numbers via manual thresholding, a process guided by operators who base their decisions on observations of the copy number figures (Figure 3.3).

However, PING2 does come with certain limitations. First, it relies on manual thresholding, demanding operator intervention to fine-tune it after processing all samples in Ping_copy. Secondly, PING2 exhibits relatively slower processing speeds, particularly during KIR allele typing, due to the iterative nature of the read mapping pipeline. The tool operates using a single thread, contributing to its slower performance. Moreover, the output alleles generated by PING2 provide only partial resolution, with a precision of 5 digits instead of the desired seven digits. This limitation can impact the accuracy of downstream analyses.

Furthermore, PING2 may encounter issues and potentially crash if it fails to detect the KIR3DL3 read count, necessitating manual removal of such samples from the analysis. Lastly, PING2's virtual probes, used for copy number estimation, may require periodic updates in alignment with the latest database releases. The probes may need re-evaluation and selection to ensure accurate alignment and avoid misalignment issues.

In summary, PING2 is a valuable tool for KIR gene analysis. However, it has certain limitations, including manual thresholding, slower processing speeds, incomplete allele resolution, potential issues in specific scenarios, and the need for periodic updates and probe evaluation. Despite these drawbacks, PING2 remains an essential asset in advancing immunogenetics research.

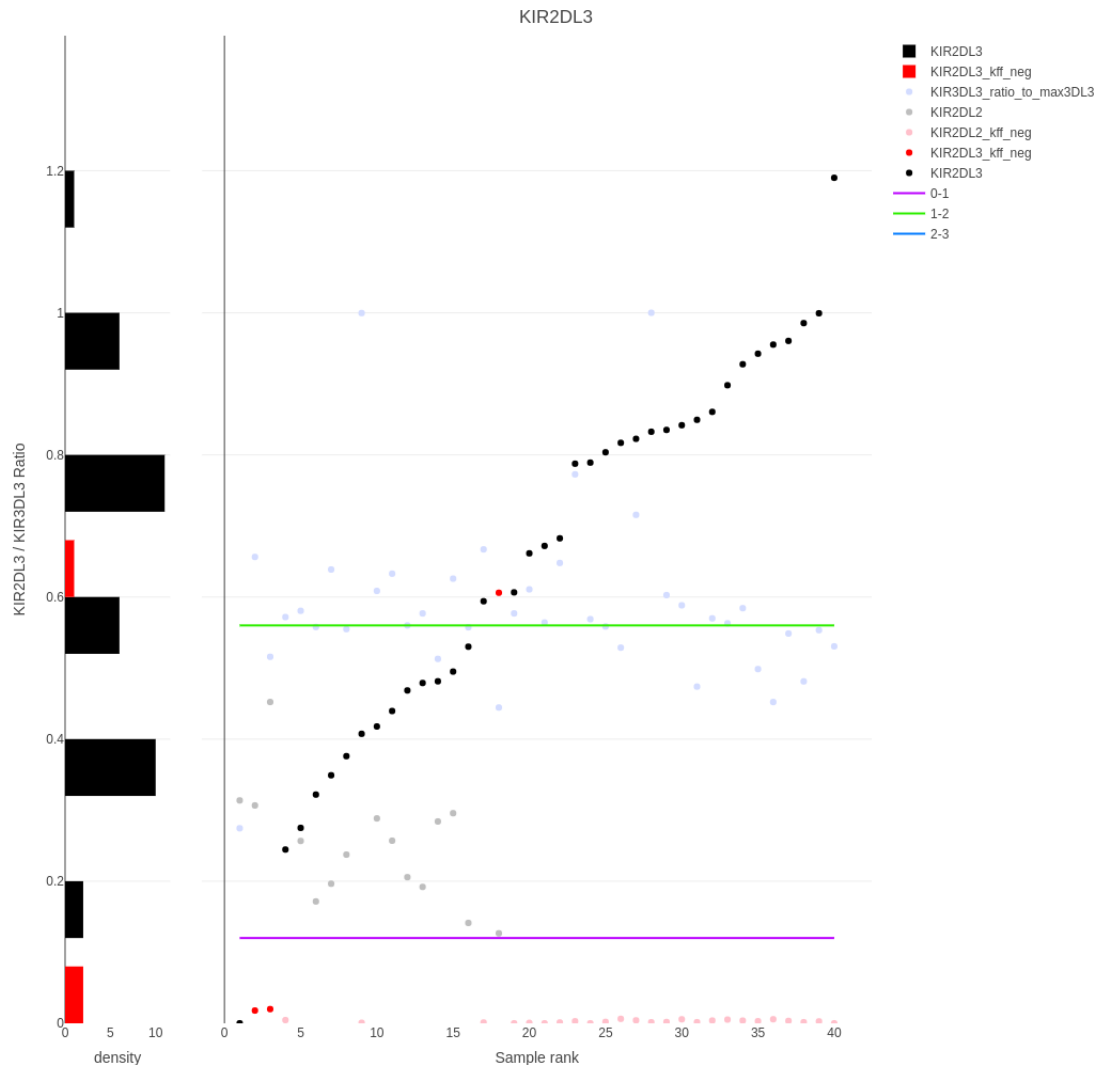


Figure 3.3: Sample Copy Number Figure Generated by PING2.

The y-axis represents the KIR3DL3-normalized KIR2DL1 gene depth for samples in the cohort. The x-axis is the HPRC samples, arranged in ascending order of scale. The dashed line is the manually assigned threshold.

3.3.2 capKIR

capKIR is a specialized data analysis pipeline developed by Wan-Chi Liu within the research facility of Dr. Ya-Chien Yang in the Department of Clinical Laboratory Sciences

and Medical Biotechnology. This pipeline is designed for the comprehensive analysis of data generated through capture-based Next-Generation Sequencing (NGS) techniques. To overcome the challenges posed by the high genetic similarity between KIR2DL1 and KIR2DS1 genes, capKIR combines reference sequences and provides tailored parameter settings specifically designed for capture-based NGS experiments conducted in Dr. Pei-Lung Chen's Lab.

These customizable settings include (1) the utilization of exon selection v.2.3, (2) the establishment of thresholds for assessing copy number variations in KIR genes, and (3) the application of an estimation coefficient known as "KIR3DL3*0.87" to standardize gene depth. These parameters have been proven to significantly improve the accuracy of identifying KIR gene copy numbers. In forthcoming capture-based NGS sequencing endeavors in this laboratory, capKIR can be readily deployed for data analysis, yielding valuable insights into KIR gene types and copy numbers.

Moreover, capKIR's capture probes are specifically designed to target HLA, AIRR, and KIR immune-related genes. The sequencing results also provide information about HLA genes in the specimens. Given the ongoing discussions regarding the interaction and compatibility of KIR and HLA genes, capKIR holds promise for further applications in KIR gene typing research.

3.3.3 graphKIR

graphKIR is a cutting-edge software developed by Hong-Ye Lin within the research facility of Dr. Chien-Yu Chen in the Department of Bio-Mechatronics Engineering. This software is designed to address the complexities associated with KIR genes within the human genome.

KIR genes exhibit high levels of polymorphism, characterized not only by striking genetic similarity between star alleles but also among the genes themselves. Additionally, KIR genes can vary in copy number and copy count between the two haplotypes in an individual, complicating accurate typing. To tackle these challenges, graphKIR has been developed to estimate KIR gene copy numbers and precisely type full-resolution (7 digits) KIR alleles from WGS data.

Key Findings and Features of graphKIR:

1. Copy Number Estimation: graphKIR uses innovative clustering methods to accurately estimate copy numbers sample-wise. This method effectively addresses the challenges posed by KIR genes' high similarity and variable copy numbers.
2. Allele Typing Accuracy: graphKIR achieves high accuracy in allele typing, outperforming competitive tools.
3. Comparative Analysis: The software was compared to other tools like PING2, GATKIR, and KPI, demonstrating superior performance. In particular, GATKIR had lower accuracy in copy number estimation, attributed to the size of the sample.
4. Graph-Based Indexing: graphKIR utilizes graph-based indexes derived from multiple sequence alignments (MSA). It offers better results compared to linear indexing, especially for similar genes.
5. Post-Mapping Filtering: The software effectively removes multiple-mapped reads, improving mapping precision and reducing false variant calls.
6. Copy Number Estimation Parameters: graphKIR evaluates various parameters affecting copy number estimation, such as read depth calculation methods, clustering algorithms, and gene depth calculation from exon-only regions.
7. Algorithms and Tools: The software uses novel clustering algorithms to estimate copy numbers and implements an enhanced version of HISAT2 EM abundance for allele typing.

3.4 HLA Software

3.4.1 HISAT-genotype

HISAT-genotype is a robust and invaluable companion tool that seamlessly complements HISAT2, the cutting-edge graph-genome read mapper (Kim et al., 2019). in 2019. While HISAT2 specializes in rapid graph mapping, HISAT-genotype furthers genomics

analysis by enabling precise alleles calling for highly polymorphic genes, including crucial ones like HLA.

The primary role of HISAT-genotype is to efficiently extract genetic variants and their corresponding star alleles from alignment files in the widely adopted .bam format. Once these variants are carefully extracted, HISAT-genotype utilizes a sophisticated Expectation-Maximization (EM) algorithm to rank them and make highly accurate allele calls. This process is fundamental in understanding complex genetic variations within these highly polymorphic genome regions.

It is essential to highlight that while HISAT-genotype excels in determining allele abundances through the EM algorithm, it does not provide an exact count of the number of alleles present in a given sample. Nevertheless, users can easily retrieve information about the two most abundant alleles from the result file, making it an invaluable resource for HLA typing(Figure 3.4).

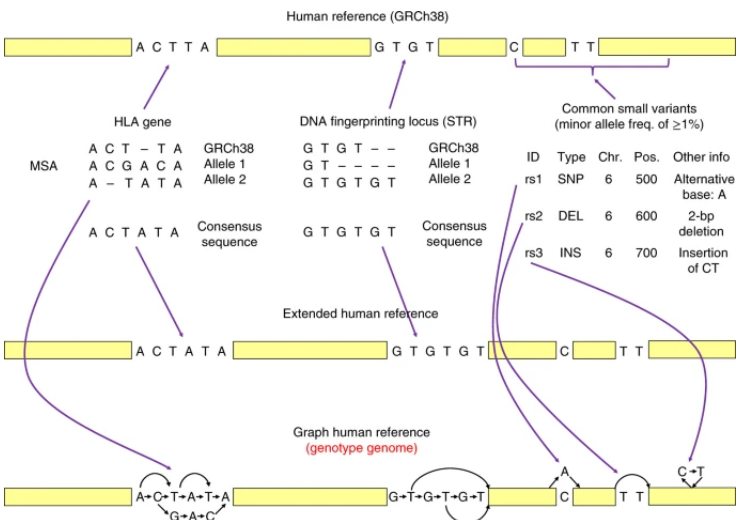


Figure 3.4: Depiction of the HISAT-genotype Algorithmic Process. This is achieved by parsing external databases, such as IMGT/HLA, which contain comprehensive information about human genes and genomic regions. These databases are meticulously converted into an intermediate format that seamlessly aligns with the sophisticated algorithms of HISAT-genotype. Genotype genomes include these variants, thereby supporting accurate genotyping.

3.4.2 VBSeq

HLA-VBSeq is designed for accurate and efficient HLA typing using whole-genome sequencing data (Nariai et al., 2015). It leverages variational Bayesian inference to opti-

mize the alignment of sequencing reads to HLA allele sequences and accurately quantify reads on HLA alleles, ultimately delivering high-resolution HLA typing results at the complete (8-digit) resolution.

The HLA-VBSeq process encompasses several critical phases. The process starts by aligning sequencing reads obtained from whole-genome sequencing to the reference genome (GRCh37/hg19) using the BWA-MEM alignment program, which includes dummy sequences. Subsequently, reads aligned to HLA loci and unmapped reads are filtered out from the alignment file. For paired-end sequencing data, where one mate is associated with an HLA locus while the other is not, both mates are extracted. The extracted reads are then realigned to capture all genomic HLA allele sequences in the IMGT/HLA database, enabling multiple alignments for each read. The IMGT/HLA database houses genomic DNA sequences for various HLA loci, including pseudogenes. Finally, the expected read counts on HLA alleles are estimated using variational Bayesian inference within a statistical framework designed explicitly for HLA typing.

In practical terms, users can easily employ HLA-VBSeq by downloading the HLA-VBSeq.jar file and a few additional essential files, including bamNameIndex.jar, SamToFastq.jar, parse_result.pl, hla_all.fasta, and Allelelist.txt. These supplementary files are required for processing fast data and accessing the necessary HLA allele sequences from the IMGT/HLA database. The software, developed in Java, undergoes multiple steps, beginning with the alignment of fastq sequences to the reference genome to generate a bam file. Subsequently, operations are performed on the resulting bam file to extract precise HLA typing information(Figure 3.5).

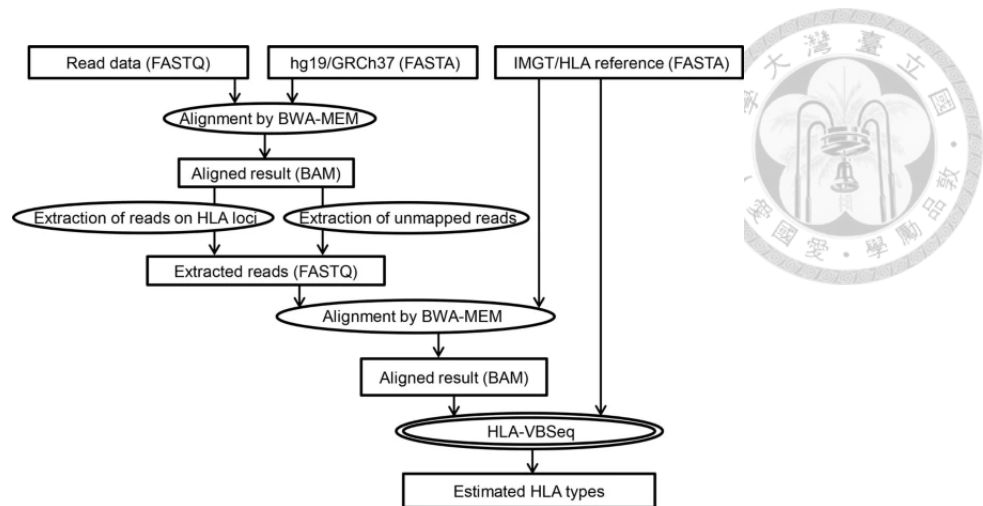


Figure 3.5: Schematic Diagram of the HLA-VBSeq Workflow. It employs provided data, the sequence of the human reference genome, and HLA allele reference sequences to predict read counts on HLA alleles for precise HLA typing.

3.4.3 T1K

T1K (Song et al., 2022), short for T1-KIR, is a versatile genotyping tool designed for the precise determination of both KIR and HLA genotypes using NGS data. At its core, T1K provides genotyping results at the allele series level, offering three-digit resolution for KIR genotyping and six-digit resolution for HLA genotyping by default. This level of resolution allows for the identification of major genetic variants within these important gene families, aiding researchers and clinicians in understanding immune system diversity and functionality.

It intelligently selects the pair of alleles that maximizes the total number of compatible reads, ensuring the most accurate genotyping possible. Furthermore, T1K employs a Poisson model to calculate quality scores for each called allele, adding an additional layer of confidence to the results and helping filter out false alleles.

T1K goes beyond simple genotyping by offering post-processing methods that enhance the utility of its results. These include the detection of novel single nucleotide polymorphisms (SNPs) on representative alleles, which can be invaluable for uncovering genetic variations specific to the studied population. Additionally, T1K provides information about single-cell level allele abundances, allowing for a deeper understanding of the genetic makeup of individual cells within a sample.

However, it's important to be aware of certain limitations associated with T1K. Firstly, T1K cannot directly compare its results with those generated by PING, limiting its ability to evaluate and benchmark its performance against other genotyping tools. Additionally, the tool reports alleles at a relatively moderate resolution level, which may not capture the full spectrum of genetic diversity in specific research or clinical contexts. T1K may also face challenges in resolving ambiguous SNPs and hybrid genes, potentially leading to reduced sensitivity and less accurate genotyping results. Finally, T1K operates under the assumption of a maximum of two alleles per gene, which may result in the exclusion of relevant alleles and may not fully represent the genetic diversity within a given population.

In summary, T1K is a valuable genotyping tool that provides essential insights into KIR and HLA genotypes, enabling researchers and clinicians to explore the genetic intricacies of the immune system.

3.4.4 DRAGEN

The DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Platform is a highly efficient solution designed to streamline the process of genome data analysis. Notably, it excels in both reducing analysis time and enhancing accuracy, thanks to its meticulously optimized algorithms. This robust platform encompasses a suite of bioinformatics pipelines specifically crafted for tasks such as mapping, alignment, classification, repeat marking, and variant calling.

One of the key strengths of DRAGEN lies in its versatility. It is adept at analyzing various types of genomic data, including Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), and RNA Sequencing (RNA-seq). This adaptability makes DRAGEN a comprehensive solution for diverse genomics applications.

It's worth noting that the capabilities and performance of DRAGEN have been rigorously tested and compared, with the results based on version v4.2.4. For more detailed information and documentation, you can refer to the official DRAGEN documentation at <https://illumina.github.io/dragen-azure-quickstart/4.2.4/>.

DRAGEN takes efficiency to an entirely new level. The program is developed in C/C++, ensuring exceptional efficiency. It employs FPGA (Schmidt and Hildebrandt,

2017) hardware acceleration for rapid and efficient parallel processing. As the field of next-generation sequencing continues to expand, the volume of generated data grows exponentially. It requires specific hardware, including 128GB of memory, 56 threads, a 2TB SSD, and an FPGA chip. It employs a specialized alignment algorithm, which differs from BWT transformation and necessitates substantial memory capacity. The index is loaded into the FPGA before alignment, allowing it to efficiently process one task at a time.

Speed and accuracy in data analysis become paramount. DRAGEN's core strength lies in its ability to perform rapid sequence alignment and variant calling. Leveraging FPGA technology and G3 data compression, it achieves remarkably fast and highly parallelized sequencing data analysis. DRAGEN can quickly complete the alignment and variant calling analysis for a whole human genome while maintaining high accuracy.

3.5 Web Crawler

Web Crawler is the process of automatically browsing a website and extracting information from its web pages (Kausar et al., 2013). It involves submitting relevant keywords or queries to a search form on a website and efficiently retrieving the desired web content. Web crawling is commonly used for various purposes, such as data mining, data analysis, content aggregation, and research.

In order to perform web crawling, several important libraries and tools are frequently used. These include urllib. Request, for making HTTP requests to fetch web pages; a RE (regular expression) library that helps with pattern matching and extracting specific information from web pages; BeautifulSoup, a library for parsing HTML and XML documents to extract Structural, a Python library for transforming data; and Pandas, a powerful data processing and analysis library (Hejing et al., 2020).

Analyzing a website's URLs helps identify patterns or rules that allow us to generate a list of the correct URLs to crawl. By understanding the structure of website URLs, we can programmatically generate URLs for different pages or categories of content, effectively scraping large amounts of data.

When crawling a website, it is imperative to mimic standard user behavior to avoid

being blocked or flagged as a robot. This is achieved by setting the User-Agent header in the HTTP request, which identifies the web crawler as a legitimate user.

Once the web pages are retrieved in HTML format, the data extraction process begins. Extract relevant information from HTML content, such as database names, sources, allele frequencies, gene frequencies, and other useful data. This extraction can be done using techniques like regular expressions, parsing HTML tags, or leveraging functionality provided by libraries like BeautifulSoup. The extracted information can then be stored in a database such as MySQL for further analysis, processing, or storage.

I use Web Crawler technology to obtain the KIR frequency data of the Taiwanese population on the AFND website and then continue to save it in MYSQL for further analysis and operation(Figure 3.6).

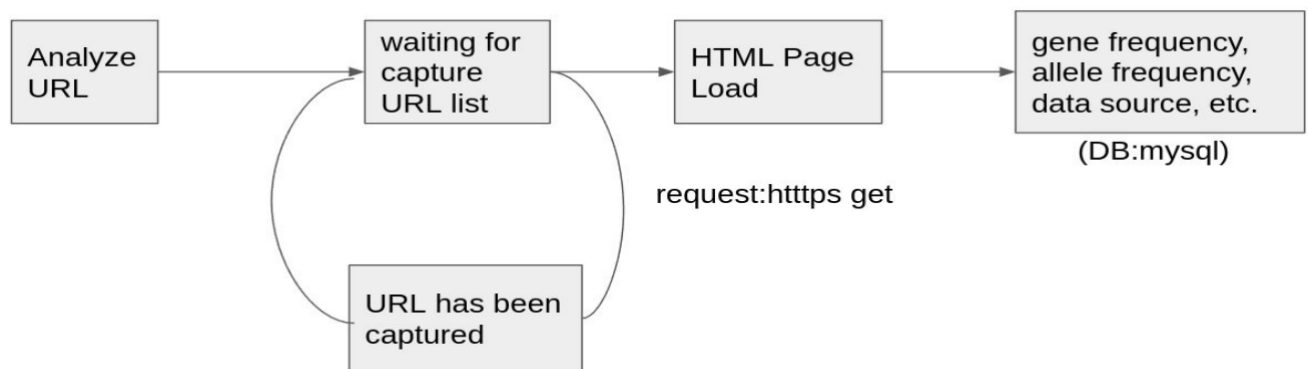


Figure 3.6: The schematic diagram of web crawler.

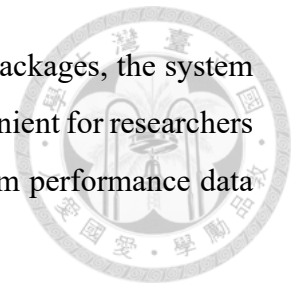
This operation deciphers and scrutinizes web pages. It simulates standard user behavior, systematically generates URLs, and collects pertinent data from HTML content for storage in a database such as MySQL. This technique was implemented to procure KIR Frequency data specific to the Taiwanese population from the AFND website.

3.6 Linux Server Hardware Configuration

The National High-Performance Computing Center (NCHC) is the organization that owns the Taiwan 3 supercomputer, one of Taiwan's self-developed supercomputers. Taiwan 3 consists of 50,400 cores distributed across 900 nodes, using Intel Xeon Platinum 8280 2.4 GHz CPUs with 28 cores each. It runs on CentOS 7.8 operating system.

Taiwan 3, as a multi-functional computing host, provides computing services in multiple fields, such as physics, chemistry, mathematics, atmospheric science, engineering

applications, and life sciences. In addition to commercial software packages, the system also provides a variety of compilers and MPI libraries, which is convenient for researchers to develop their calculation programs. We run all scripts and perform performance data analysis on Taiwania3.





Chapter 4 Results



4.1 Benchmarking HLA genotype by HPRC short read WGS dataset

In our study, the HPRC dataset served as an invaluable resource. This extensive dataset covers various data types, such as HIFI-read, WGS, and panel data. As one of the most accessible datasets, it played a crucial role in our research on the correlation between KIR-HLA pairs and different phenotypes. I want to express our heartfelt thanks to Dr. Pei-Lung Chen's Lab, with a special shout-out to Sheng-Kai Lai for their help with the HPRC panel data, HLA typing, and generating panel consensus data.

We utilized various tools for the panel and WGS datasets, such as VB-seq, HISAT-genotype, T1K, DRAGEN, and QZtype(TBG Biotechnology Corp., Taiwan http://www.tbgbio.com/en/product/product_analysis_software). Each tool has its unique features and capabilities, depending on the versions and alignment methods used and the IPD-HLA database versions they reference. To comprehensively understand these tools, we have compiled all the relevant information in Table 4.1. This table is a comprehensive summary detailing each tool's distinctive features and literature.

Table 4.1 is for a brief overview of these tools. For instance, the HLA-VBseq and HISAT-genotype tools present different versions and publication years, and they refer to other reference genomes for their function. They also diverge in the alignment method - with the former using linear alignment and the latter leveraging graph alignment. However, they refer to different IPD-HLA database versions. Notably, the HISAT-genotype includes HLA alternative contigs in its reference genome, enhancing the accuracy of its results. The other tools in the table also have unique characteristics, such as different alignment methods, reference genomes, and IMGT/HLA database versions. This diversity in the tools and their features underpins the careful selection and application of these tools fitting to the requirements of the data sets.

Table 4.1: The Principal HLA Typing Tools Employed.

	HLA-VBseq	HISAT-genotype	T1K	DRAGEN	hla_annot	Qztype
Tool version	v1(2014)	v1.3.2 (2020)	v1	v4.2	v1	v1
Publication	BMC Genomics(2015) Pharmacogenomics(2018)	Nature Biotechnology(2019)	Genome Res. (2023)	new version (2023-08)		Sponsored by TBG
Reference genome	hs37d5	hs38DH (Include HLA alternative contigs)	hg38	hg38	hla_gen.fasta hla_nuc.fasta	
Alignant	Linear	Graph	Linear	Graph	BWA minimap2	
IMGT/HLA database version	3.31.0 (2018-01)	3.26.0 (2016-10)	3.52.0 (2023-04)	3.45.0 (2021-07)	3.51.0 (2023-01)	

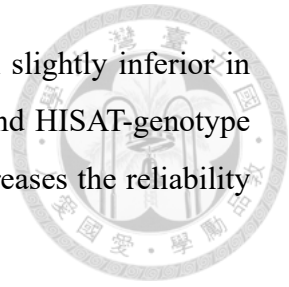
Our research adopted a comprehensive approach to analyzing HLA typing tools, culminating in selecting T1K and HISAT-genotype tools for HLA typing on Taiwanese individuals (Figure 4.1). This selection process employed a systematic comparison of HLA typing tools' performance, using data categorized into two types: WGS and panel data.

A consensus from the panel was established as the answer key to evaluate individual tools' accuracy. We depicted the specificity and limitations of these tools in Figure 4.1a and Figure 4.1b. Figure 4.1a represents a comparison not excluding tools that could not type certain genes or samples. An evident example is the DRAGEN tool, which could not type out HLA-DPA1 and HLA-DPB1.

Contrastingly, the figure of fig:17b includes these limitations. Due to this inclusive approach, the figure exhibits higher accuracy rates for some tools, marked by having NaN values, which are any un-typable or non-callable alleles excluded from the dataset. This analysis leads to different denominators for each tool's accuracy rate. Under standard conditions, as shown in Figure 4.1a, each tool should have an expected denominator of 704 (calculated from 2 alleles per gene x 8 genes x 44 samples). However, in Figure 4.1b, the denominators differ among tools due to their limitations. For instance, the denominator is 690 for 'wgs_consensus,' 504 for 'panel_DRAGEN,' 516 for 'wgs_DRAGEN,' and 692 for 'panel_Qztype.' These values reflect each tool's maximum typing ability under given limitations.

The selection of T1K was based on a comparative analysis considering its capabilities

and limitations, as well as the 'panel_consensus' results. Although slightly inferior in terms of outcomes, HISAT-genotype was also chosen. Both T1K and HISAT-genotype are accurate and have comprehensive typing capabilities, which increases the reliability of our study.



- For Figure 4.1a, the accuracy percentage was calculated differently, considering the tool's limitations in calling certain genes:

$$\text{Accuracy} = \frac{\text{Number of correct alleles determined by the tool}}{\text{Total number of correct alleles according to the panel consensus}}$$

This calculation shows the percentage of alleles the tool correctly identified compared to the panel consensus.

- For Figure 4.1b, the accuracy percentage was calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correct alleles determined by the tool}}{\text{Total alleles after removing non-callable alleles by the tool and correct alleles according to the panel consensus}}$$

This calculation provides a more realistic accuracy percentage, considering the typing limitations of each tool.

These accuracy measures allowed us to evaluate the typing tools in different scenarios, providing comprehensive insights into their performance.

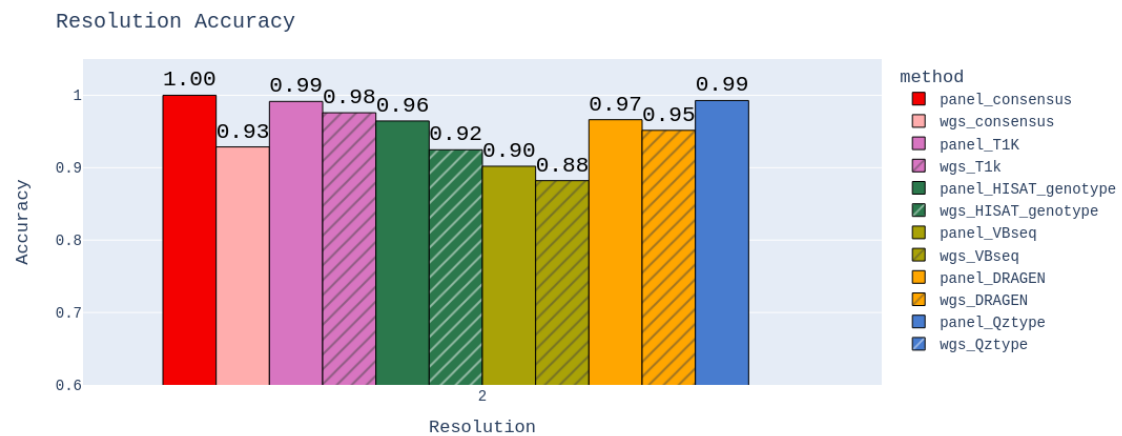
Our comprehensive analysis showed that certain tools had higher accuracy rates when non-typable or non-callable alleles were excluded from consideration. This understanding led to a more comprehensive evaluation of each tool under different scenarios and challenges they might encounter in actual research applications.

T1K and HISAT-genotype proved to be the most optimal tools in our study due to their superior accuracy and comprehensive gene typing capabilities. With their remarkable performance, even considering the limitations in typing, these tools have shown great potential for performing reliable and complete analyses for future studies.

In conclusion, our comparative analysis provides valuable insights into applying HLA gene typing tools in different contexts. Given their consistent and commendable performance, T1K and HISAT-genotype were selected as the tools of choice for analyzing the TWB WGS data.



(a) HPRC 44 Samples After Dropping NaN Values.



(b) HPRC 44 Samples with NaN Values.

Figure 4.1: Comparison of HPRC Tools in Two HLA Fields.

It reflects the findings from our meticulous summary of the HPRC typing results, which ultimately facilitated the selection of T1K and HISAT-genotype as the preferred tools for HLA typing in Taiwanese individuals. The choice was made considering the accuracy, alignment with consensus results, and overall performance demonstrated by these tools.

Evidence-based medicine (EBM) forms the basis of today's medical practice (Lebedev et al., 2020). EBM optimizes patient care decision-making processes by advocating a systematic approach to identifying the most reliable evidence from scientific research. While all academic work is supported to some degree by empirical evidence and exhibits varying degrees of scientific rigor, EBM places this work into a hierarchical categorization based on the strength of its underlying evidence.

This study focuses on exploring the concept of EBM and, in particular, examining the highest levels of evidence, namely 1A, 2A, and 2B. The investigation includes the analysis of HLA risk alleles using data from clinical variants provided and curated by PharmGKB (as of June 5, 2023). The original data source comprises information from 1492 individuals in the TWB of T1k.

Examination of this extensive data revealed different patterns of risk allele prevalence (Figure 4.2). Of the Taiwanese individuals, 1,377 had at least one risk allele. This result corresponds to a high prevalence rate of 0.923, a considerable percentage of the population (Table 4.2). When examining level 1A evidence among individuals, we found that 423 individuals had at least one 1A risk allele, corresponding to a prevalence of 0.284 (Table 4.3). Among the 1A risk alleles, HLA-B*58:01 had the highest prevalence, accounting for 12.567% of 1A risk alleles in the population (Figure 4.3).

We also wanted to categorize these risk alleles according to their level of evidence and associate them with different substances and medications with which they are associated (Figure 4.4). The result is a comprehensive classification of the influence of different risk alleles on drug sensitivity, which could be invaluable for personalized medicine. For example, the most common risk allele, HLA-B*40:01, showed evidence of stage 2A and was associated with carbamazepine. It accounts for an impressive 18.901% of the Taiwanese sample. Similarly, HLA-A*33:03 and HLA-C*03:02 (each with level 2B evidence) showed significant associations with allopurinol, with frequencies of 13.237% and 12.567%, respectively. The frequently occurring HLA-B*58:01 (evidence level 1A) was also associated with allopurinol, with a frequency of 12.567%.

The risk alleles show strong and frequent associations with different substances and medications. Our study provides a comprehensive investigation of Taiwanese individuals who have a minimum of the described risk alleles. The study is based on the principles of

EBM and uses data from 1492 individuals from the TWB, reflecting a remarkably high prevalence rate.



Table 4.2: Distribution of Risk Alleles Among Taiwanese Individuals.

At least \$num risk allele	People	Frequency
1	1377	0.923
2	955	0.640
3	422	0.283
4	233	0.156
5	66	0.044
6	34	0.023
7	1	0.001

Table 4.3: Distribution of 1A Risk Alleles in the Taiwanese Population

At least \$num 1A risk allele	People	Frequency
1	423	0.284
2	63	0.042
3	0	0.000

Comparison of 1A, 2B, and 2A Counts

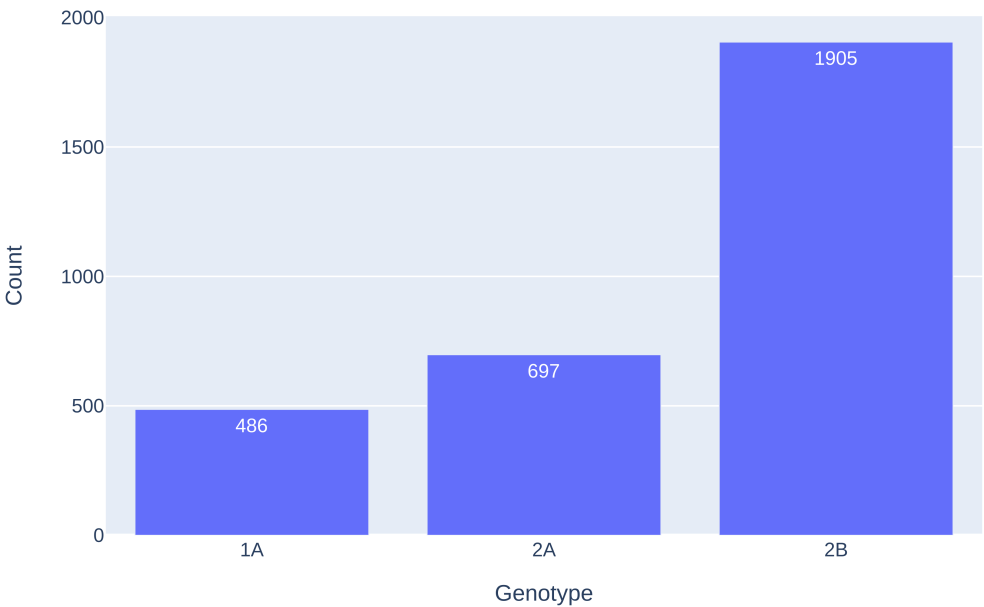


Figure 4.2: Comparison of 1A,2B,and 2A counts.
It graphically represents the variation in the likelihood of 1A, 2B, and 2A risk allele counts within the TWB. This distribution offers pivotal insights into the prevalence and impact of various risk alleles on the Taiwanese population’s health.

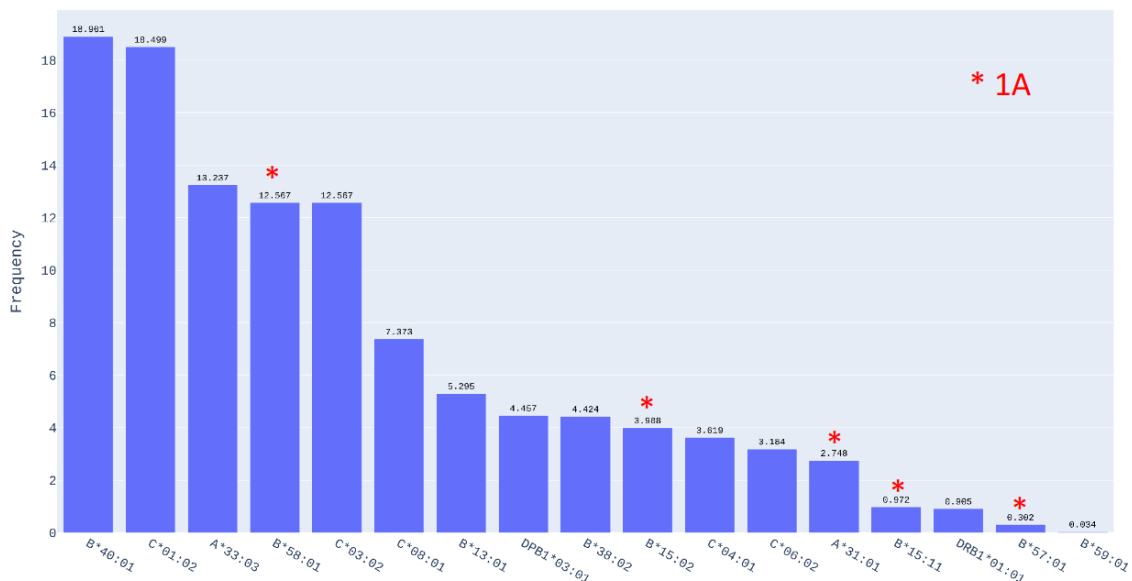


Figure 4.3: Distribution of Risk Allele Probabilities in the TWB.
The distribution of risk allele probabilities among individuals in the TWB. The x-axis represents different alleles, while the y-axis illustrates their frequencies. Alleles falling under category 1A evidence are distinctively marked with an asterisk.

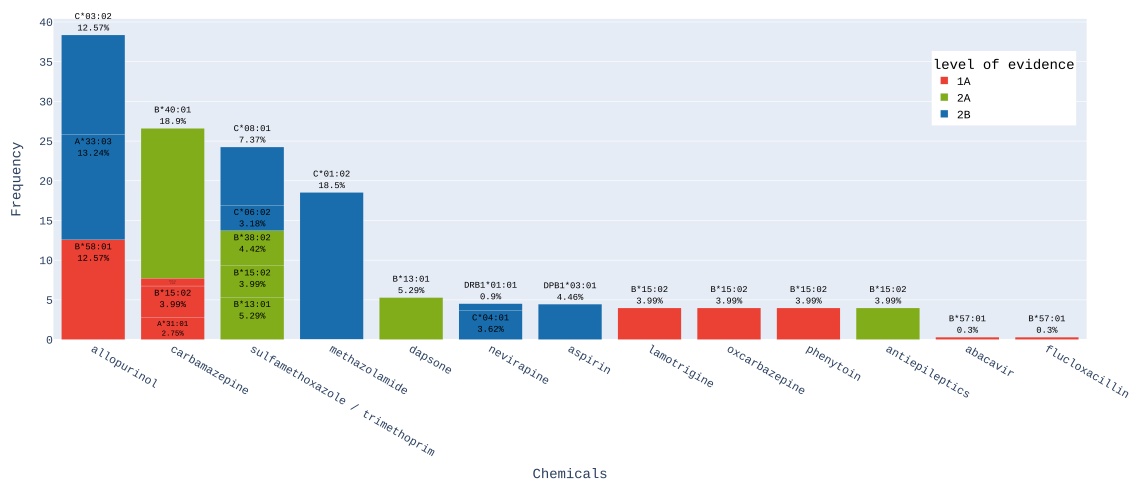


Figure 4.4: Comparative Frequency of HLA Risk Alleles across Different Levels of Evidence Linked to Specific Chemicals.

It provides a comparative visualization of the frequency of HLA risk alleles across different evidence levels (1A, 2A, 2B) associated with particular chemicals. The x-axis represents different chemicals, while the y-axis highlights their frequency. Each chemical displays stacked categories of 1A, 2A, and 2B evidence levels, offering an inclusive view of their distributions.

4.2 Designing WGS-based KIR genotyping protocol

Processing the complexity and diversity of KIR genes is crucial for many immunogenomic studies. Nevertheless, the complexity and diversity of KIR genes can make this particularly challenging, requiring the most optimal KIR typing tool.

Our research aims to identify the most appropriate tool for KIR typing in TWB for WGS. To compare KIR typing tools, we used the table 4.4, which provides a comprehensive overview of different tools, including PING, Sakaue KIR ([Sakaue et al., 2022](#)), T1K([Song et al., 2022](#)), capKIR and graphKIR. The table contains information about the alignment tools, the sample level (per sample or cohort), the reference genome, the IPD-KIR version, the detectability of new alleles, the typing level (CN, allele digit), and the source information.

Each tool has its characteristics. For example, PING uses a manual threshold method where the user has to set a threshold and then determine the CN based on the read depth distribution. As for identifying novel alleles, our results show that each tool currently has limitations and cannot recognize such alleles. It was also found that these limitations make it difficult to identify missing or fused alleles.

The results presented from KIR typing tools suggest that the findings at the copy number level are considered the most reliable. As a result, in our research, we chose to delve deeper with capKIR, a tool developed by Dr. Ya-Chien Yang's lab, and graphKIR, contributed by Hong-Ye Lin from Dr. Chien-Yu Chen's lab. Additionally, we have a benchmark from the HPRC dataset that we applied and the KIR answer provided by SKIRT, a tool originating from Tsung-Kai Hung of Dr. Pei-Lung Chen's lab ([Hung et al., 2023](#)).

A key feature of capKIR is its exon selection capability. Although capKIR requires manual threshold adjustment, it provides a standardized value after correction by Wan-Chi Liu from Dr. Ya-Chien Yang's lab so manual threshold adjustment may be unnecessary. On the other hand, the strength of graphKIR lies in its algorithm, which bases the data analysis on the 75th percentile of the read depth distribution, allowing a more comprehensive analysis of the data.

We used capKIR and graphKIR individually and in combination (capKIR+graphKIR)

for KIR typing on two data types: WGS and panel data, as shown in Figure 4.5. The table contains the CN errors detected by each tool on both data types. The table's blank indicates that the corresponding tool is not used to analyze this data type. The results showed that graphKIR had 4 CN errors in WGS data and 97 CN errors in panel data (Table 4.5). In the WGS data, capKIR +graphKIR had 165 CN errors. capKIR had 6 CN errors in the panel data.

The results of these extensive tests suggest that the performance of each tool varies depending on the data type used. GraphKIR performed significantly better on WGS data, while capKIR was the better choice for panel data and had fewer CN errors. These observations suggest that proper tool selection tailored to the data type is crucial for reliable KIR typing and achieving better results in genomics research. We therefore conclude that graphKIR is suitable for WGS data, while capKIR is more suitable for panel data for KIR typing and provides better results.

Table 4.4: Comprehensive Comparative Overview of KIR Typing Tools

Name	PING(old) , PING, PING-wgs	Sakaue KIR	T1K	capKIR	graphKIR
Alignment	5*BWA + Ch19 extract	BWA	BWA	BWA	Graph + BWA
Notice	manually thresholding	KDE	only two copy	manually thresholding exon selection	distribution p75
Sample Level	per cohort	per cohort	per cohort	per cohort	per cohort per sample
Ref Genome	hg19 + probe(allele set)	hg19	read mapping all alleles from DB	consensus	hs37d5
IPD-KIR Version	2.11.0(2.7.1 claimed in paper)	2.7.0	any by command line (2.10.0)	NA	any by command line (2.10.0)
Novel alleles	unable	unable	unable	unable	unable
Typing Level	copy number + allele(5 digits)	copy number + allele(7 digits)	copy number + allele(3 digits)	copy number	copy number + allele(7 digits)
Source	Marin, W. M. et al. (2021). Marin, W. M. et al. (2022).	Sakaue, S. et al. (2022).	Song, L. et al. (2022).	YC Lab	Linnil1

Table 4.5: KIR Typing Results of graphKIR, capKIR, combine version in WGS and panel type on HPRC Datasets.

	graphKIR	capKIR+graphKIR	capKIR
WGS	4	165	
Panel	97		6

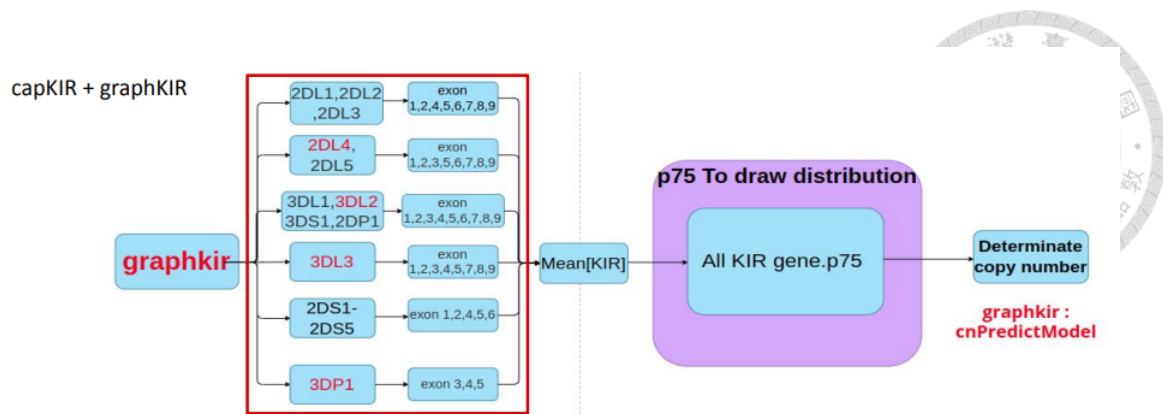


Figure 4.5: Combining capKIR and graphKIR Schematic Diagram.

It provides a detailed schematic illustration of the integrated process combining the functionalities of both capKIR and graphKIR. It outlines how these two genotyping tools work together to provide an enhanced and comprehensive method for KIR typing. The diagram represents how graphKIR's algorithm design, which analyzes data based on the 75th percentile of read depth distribution, blends with capKIR's unique feature of utilizing exon selection. This combination results in a more accurate and reliable KIR typing process.

4.3 KIR copy number typing in TWB

We employed graphKIR to analyze 1,492 samples from the TWB to evaluate the CN of KIR genes. Through careful data evaluation, we aimed to provide a full picture of CN distribution and its link to different KIR genes.

As shown in Figure 4.6, our findings highlight CN values' wide range and variability for each KIR gene. Each CN's different densities and prevalence across the TWB samples underscore the Taiwanese population's genetic diversity and unique makeup.

Further analysis of the four commonly occurring KIR genes, known as framework genes, helps to reveal haplotype prevalence (Figure 4.7). Notably, the genes from the A haplotype appeared more frequently, especially when looking at total CN counts. In contrast, genes from the B haplotype were less common. Therefore, the genetic variety of the Taiwanese population involves a significant prevalence of A haplotype genes.

The collective genotyping results from using graphKIR on the 1492 TWB samples allowed us to conduct a more detailed analysis of AA, BB, and BX haplotypes (Figures 4.8 and 4.9). We classified '3DL1' and '2DS4' as representative of the A haplotype and '2DS2', '3DS1', '2DL5', '2DS3', and '2DS5' as representatives of the B haplotype. The distribution of these haplotypes shows that nearly half (49.7%) of the total results belonged

to the AA haplotype. The BX haplotype was also quite prevalent, making up 45.7% of the total, while the BB haplotype formed a smaller part of the results at 4.6%.

By focusing on the CN values and corresponding gene counts of these haplotypes, we gained a deeper understanding of the KIR gene landscape (Figures 4.10 4.11 4.12). For instance, Figure 4.10 provides an in-depth look at the AA haplotype's gene counts and compositions. Similarly, Figures 4.11 and 4.12 display the BB and BX haplotypes, respectively.

The culmination of these analyses, visualized, offers key insights into the genetic diversity and complex interplay of the KIR genes within the Taiwanese population, laying the groundwork for more refined and targeted genetic research in the future (Table 4.6). To assess the copy number (CN) of KIR genes, we used graphKIR to analyze 1,492 samples from the TWB, a large and diverse source of genomic data. Through careful data analysis, we aimed to create a complete picture of CN distribution and its association with different KIR genes through detailed data analysis.

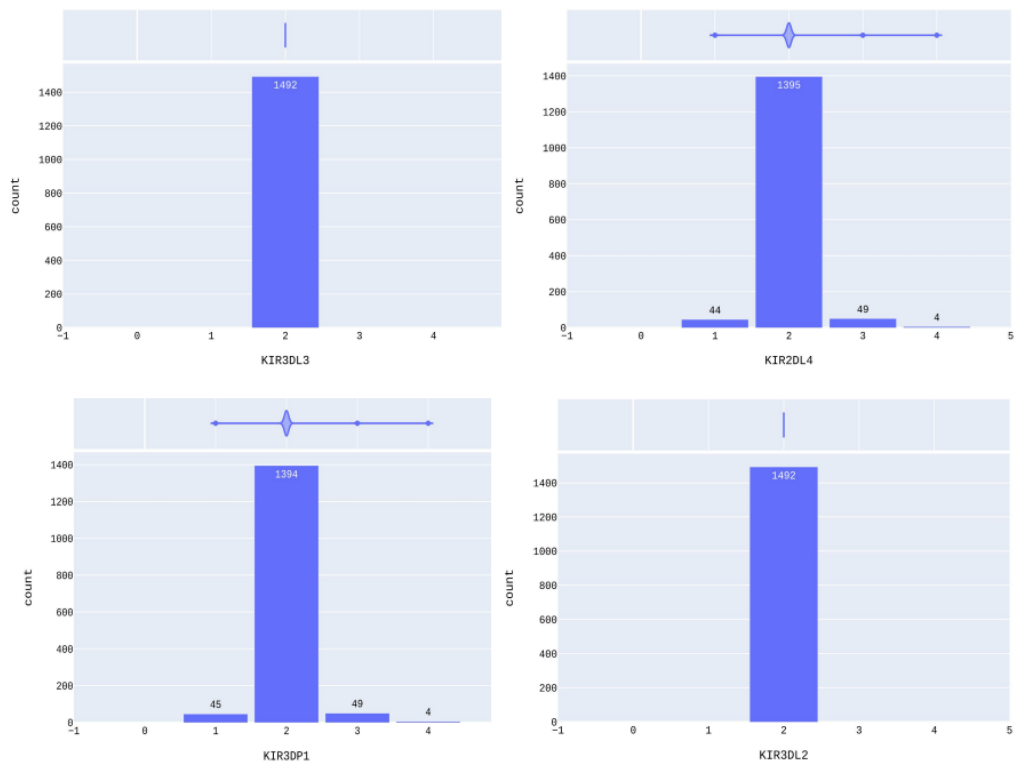
As shown in Figure 4.6, our results demonstrate the wide range and variability of CN values for each KIR gene. The different densities and prevalences of each CN in the TWB samples emphasize the Taiwanese population's genetic diversity and unique composition.

Further analysis of the four common KIR genes, known as frame genes, helps to reveal haplotype prevalence (Figure 4.7). It is noticeable that the A haplotype genes are more common, especially when considering the total number of CN. In contrast, genes of the B haplotype were less frequent. Thus, the genetic diversity of the Taiwanese population includes a significant prevalence of haplotype A genes.

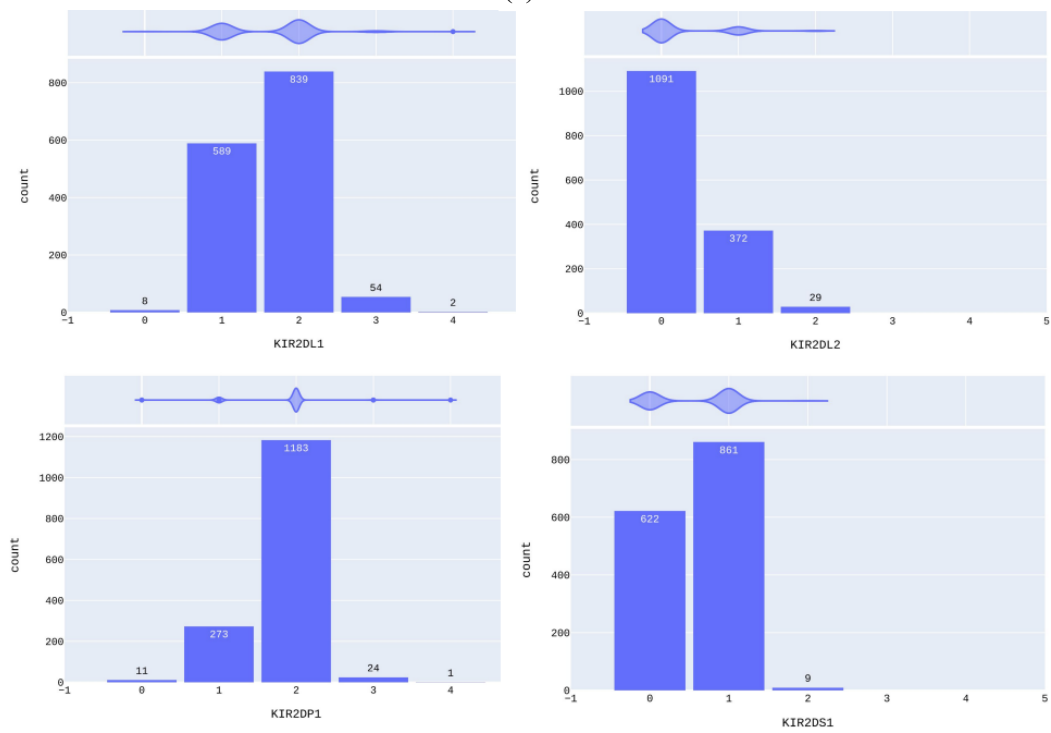
The collective genotyping results from using graphKIR on the 1492 TWB samples allowed us to analyze the AA, BB, and BX haplotypes in more detail (Figures 4.8 and 4.9). We classified '3DL1' and '2DS4' as representatives of the A haplotype and '2DS2', '3DS1', '2DL5', '2DS3' and '2DS5' as representatives of the B haplotype. The distribution of these haplotypes shows that almost half (49.7%) of the total results belonged to the AA haplotype. The BX haplotype was also quite widespread, with a share of 45.7%, while the BB haplotype accounted for a smaller proportion of the results with 4.6%.

By focusing on the CN values and corresponding gene counts of these haplotypes, we gained a deeper understanding of the KIR genetic landscape (Figures 4.10 4.11 4.12). For example, Figure 4.10 provides a detailed insight into the number and composition of AA haplotype genes. Similarly, figures 4.11 and 4.12 show the BB and BX haplotypes, respectively.

These analyses provide important insights into the genetic diversity and complex interplay of KIR genes within the Taiwanese population. It forms the basis for future refined and targeted genetic research (table 4.6).Copy



(a)



(b)

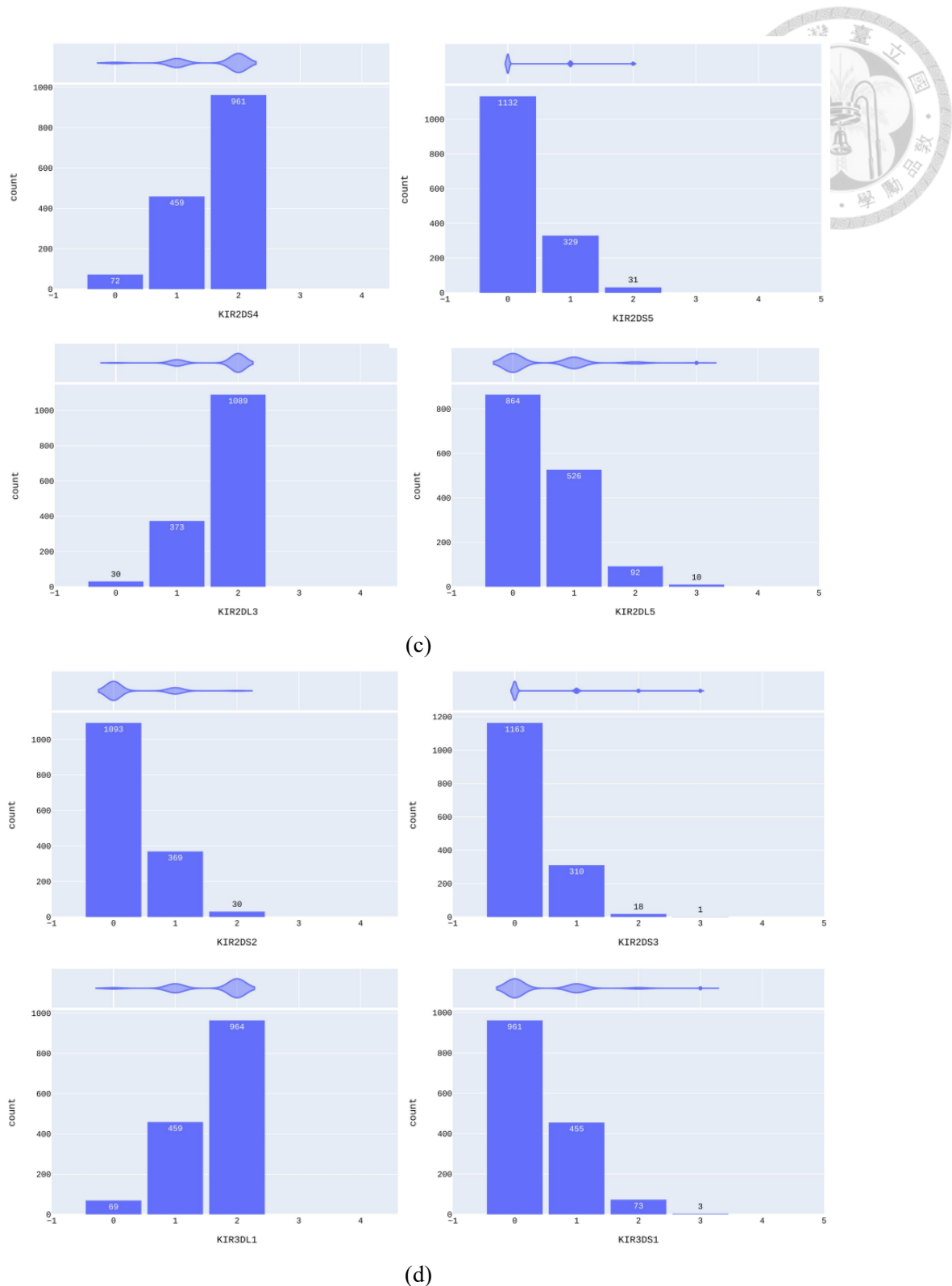


Figure 4.6: Distribution of Copy Numbers for Each KIR Gene.

Figures (a), (b), (c), and (d) present graphical representations illustrating the distribution and density of Copy Numbers for each KIR Receptor gene. These visualizations offer a clear and detailed overview of the variation and prevalence of different KIR genes within the analyzed sample set.



Table 4.6: Distribution of KIR gene copy numbers in a cohort of 1492 individuals from TWB.

	0	1	2	3	4
3DL3	0	0	1492	0	0
3DL2	0	0	1492	0	0
2DL4	0	44	1395	49	4
3DP1	0	45	1394	49	4
2DL1	8	589	839	54	2
2DL2	1091	372	29	0	0
2DP1	11	273	1183	24	1
2DS1	622	861	9	0	0
2DS4	72	459	961	0	0
2DS5	1132	329	31	0	0
2DL3	30	373	1089	0	0
2DL5	864	526	92	10	0
2DS2	1093	369	30	0	0
2DS3	1163	310	18	1	0
3DL1	69	459	964	0	0
3DS1	961	455	73	3	0

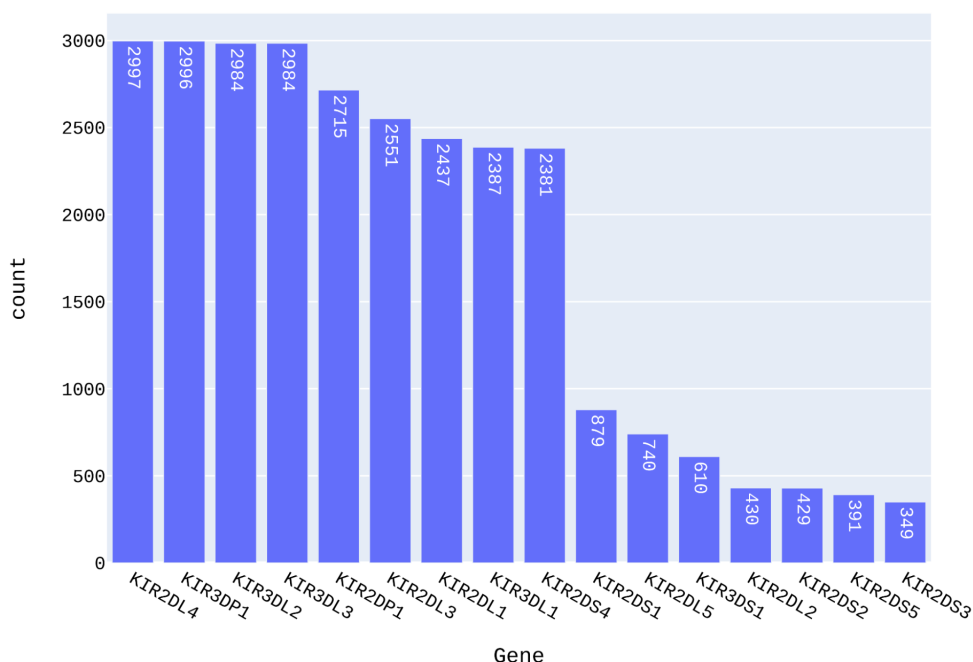


Figure 4.7: Sorted Total Counts of KIR Genes Across Individuals.

It offers a detailed visualization of the total count for each KIR gene across the 1492 individuals. The X-axis represents each KIR gene, while the Y-axis denotes the corresponding total count across the individuals. The figure is constructed with the genes sorted in an ordered manner based on their prevalence, providing a clear picture of the distribution and abundance of different KIR genes in the analyzed population.

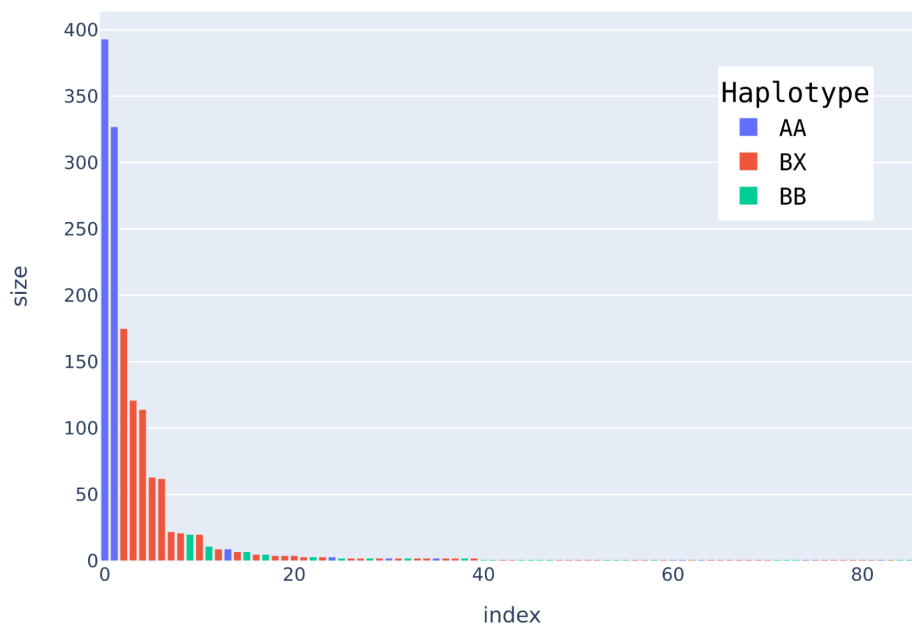


Figure 4.8: Sorted Composition of AA, AB, and BX Haplotypes among TWB Samples. It presents a histogram that depicts the count of different KIR haplotypes - AA, AB, and BX, among the 1492 TWB samples. The Y-axis represents the size, indicating the number of each specified KIR haplotype. The haplotypes are visualized in sorted order, showing their relative abundance within the genotyped population.

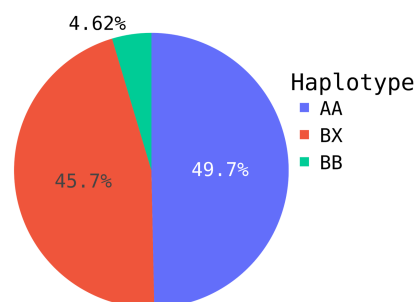
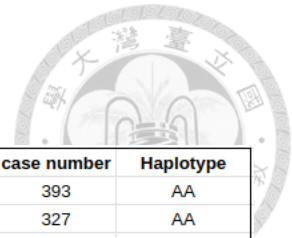


Figure 4.9: Proportional Representation of AA, AB, BX Haplotypes among TWB Samples. It offers a pie chart displaying the proportional representation of the AA, AB, and BX haplotypes observed within the TWB samples. This graphical representation offers an immediate and understandable visualization of the distribution of these haplotypes in the given population.



3DL3	2DS2	2DL2	2DL3	2DL5	2DP1	2DL1	3DP1	2DL4	3DL1	3DS1	2DS3	2DS5	2DS1	2DS4	3DL2	case number	Haplotype
2	0	0	2	0	2	2	2	2	2	0	0	0	0	2	2	393	AA
2	0	0	2	0	2	1	2	2	2	0	0	0	1	2	2	327	AA
2	0	0	2	0	3	3	2	2	2	0	0	0	0	2	2	9	AA
2	0	0	2	0	2	2	2	2	2	0	0	0	1	2	2	3	AA
2	0	0	2	0	3	2	2	2	2	0	0	0	1	2	2	2	AA
2	0	0	1	0	1	1	1	1	1	0	0	0	0	2	2	2	AA
2	0	0	2	0	2	2	2	2	2	0	0	0	0	1	2	1	AA
2	0	1	1	0	1	1	2	2	2	0	0	0	0	2	2	1	AA
2	0	0	2	0	2	1	2	2	2	0	0	0	1	1	2	1	AA
2	0	0	2	0	2	1	1	2	2	0	0	0	0	2	2	1	AA
2	0	0	2	0	1	2	2	2	2	0	0	0	0	2	2	1	AA

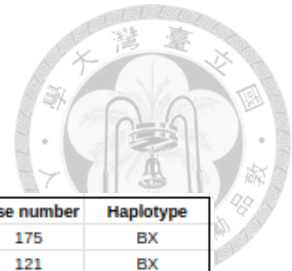
Figure 4.10: Genotyping Results for the AA Haplotype.

It provides a comprehensive analysis of the AA haplotype, focusing on the corresponding KIR gene counts and haplotypes. It offers a clear visualization of the variations within the AA haplotype, providing pivotal information on the diversity and prevalence of specific KIR genes.

3DL3	2DS2	2DL2	2DL3	2DL5	2DP1	2DL1	3DP1	2DL4	3DL1	3DS1	2DS3	2DS5	2DS1	2DS4	3DL2	case number	Haplotype
2	0	0	2	2	2	3	2	2	0	2	1	1	1	0	2	20	BB
2	0	0	2	2	2	3	2	2	0	2	0	2	1	0	2	11	BB
2	1	1	1	2	1	2	2	2	0	2	0	2	1	0	2	7	BB
2	1	1	1	2	1	2	2	2	0	2	1	1	1	0	2	5	BB
2	1	1	1	2	1	2	1	1	0	1	0	2	1	0	2	3	BB
2	0	0	2	2	2	3	2	2	0	2	2	0	1	0	2	2	BB
2	0	0	2	2	2	2	2	2	0	2	1	1	2	0	2	2	BB
2	1	1	1	3	2	3	2	2	0	2	2	1	1	0	2	2	BB
2	1	1	1	2	1	1	2	2	0	2	0	2	2	0	2	2	BB
2	2	2	0	2	0	1	2	2	0	2	0	2	1	0	2	1	BB
2	2	2	0	2	0	1	1	1	0	1	0	2	1	0	2	1	BB
2	2	2	0	2	0	0	2	2	0	2	1	1	2	0	2	1	BB
2	2	2	0	3	1	2	1	1	0	1	1	2	1	0	2	1	BB
2	1	1	1	2	1	2	1	1	0	1	2	0	1	0	2	1	BB
2	1	1	1	3	2	3	2	2	0	2	1	2	1	0	2	1	BB
2	1	1	1	2	1	1	2	2	0	2	1	1	2	0	2	1	BB
2	1	1	1	3	2	2	3	3	0	3	3	0	2	0	2	1	BB
2	1	1	1	2	2	3	1	1	0	1	1	1	1	0	2	1	BB
2	2	2	0	3	1	2	2	2	0	2	1	2	1	0	2	1	BB
2	2	2	0	2	0	1	2	2	0	2	1	1	1	0	2	1	BB
2	1	1	1	2	1	1	1	1	0	1	0	2	2	0	2	1	BB
2	0	0	2	2	3	4	2	2	0	2	0	2	1	0	2	1	BB
2	0	0	2	3	3	4	3	3	0	3	1	2	1	0	2	1	BB
2	1	1	1	2	1	2	1	1	0	1	1	1	1	0	2	1	BB

Figure 4.11: Genotyping Results for the BB Haplotype.

It mirrors the analysis done for the AA haplotype but focuses on the BB haplotype. This figure delivers a detailed examination of the BB haplotype, illustrating the distinct KIR gene counts and haplotypes within this category.

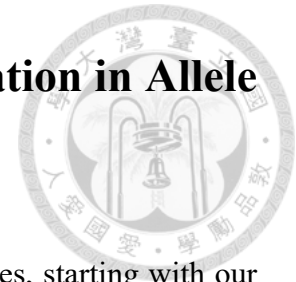


3DL3	2DS2	2DL2	2DL3	2DL5	2DP1	2DL1	3DP1	2DL4	3DL1	3DS1	2DS3	2DS5	2DS1	2DS4	3DL2	case number	Haplotype
2	0	0	2	1	2	2	2	2	1	1	0	1	1	1	2	175	BX
2	0	0	2	1	2	2	2	2	1	1	0	1	1	1	2	121	BX
2	1	1	1	0	1	1	2	2	2	0	0	0	0	2	2	114	BX
2	1	1	1	1	2	2	2	2	2	0	1	0	0	2	2	63	BX
2	1	1	1	1	1	1	2	2	1	1	0	1	1	1	2	62	BX
2	1	1	1	1	1	1	1	1	1	0	0	1	1	1	2	22	BX
2	1	1	1	1	2	2	3	3	2	1	1	0	0	2	2	21	BX
2	1	1	1	1	1	1	2	2	1	1	1	0	1	1	2	20	BX
2	1	1	1	2	2	2	2	2	1	1	1	1	1	1	2	9	BX
2	1	1	1	1	1	1	1	1	1	0	1	0	1	1	2	7	BX
2	0	0	2	0	2	2	3	3	2	1	0	0	0	2	2	5	BX
2	2	2	0	1	0	0	2	2	1	1	0	1	1	1	2	4	BX
2	1	1	1	2	2	2	3	3	1	2	1	1	1	1	2	4	BX
2	0	0	2	1	3	3	2	2	1	1	1	0	1	1	2	4	BX
2	1	1	1	1	2	2	2	2	2	0	1	0	0	1	2	3	BX
2	2	2	0	1	1	1	3	3	2	1	1	0	0	2	2	3	BX
2	2	2	0	1	0	0	1	1	1	0	0	1	1	1	2	2	BX
2	1	1	1	2	2	2	3	3	1	2	2	0	1	1	2	2	BX
2	1	1	1	2	3	2	3	3	2	1	2	0	1	2	2	2	BX
2	1	1	1	1	2	2	2	2	1	1	1	0	1	1	2	2	BX
2	2	2	0	2	1	1	3	3	1	2	1	1	1	1	2	2	BX
2	2	2	0	2	1	1	2	2	1	1	2	0	1	1	2	2	BX
2	2	2	0	2	1	1	2	2	1	1	1	1	1	1	2	2	BX
2	1	1	1	0	1	1	2	2	2	0	0	0	1	2	2	2	BX
2	2	2	0	1	1	1	2	2	2	0	1	0	0	2	2	2	BX
2	1	1	1	3	3	3	4	4	1	3	2	1	1	1	2	1	BX
2	1	1	1	1	2	2	2	2	1	1	0	1	1	1	2	1	BX
2	1	1	1	1	3	3	2	2	2	0	1	0	0	2	2	1	BX
2	1	1	1	1	2	1	2	2	2	0	1	0	1	2	2	1	BX
2	1	1	1	1	1	2	2	2	1	1	0	1	1	1	2	1	BX
2	1	1	1	1	1	1	4	4	2	2	0	1	1	2	2	1	BX
2	1	1	1	1	1	1	3	3	1	2	0	1	1	1	2	1	BX
2	1	1	1	0	1	1	3	3	2	1	0	0	0	2	2	1	BX
2	1	1	1	1	1	1	1	1	1	0	1	0	1	0	2	1	BX
2	0	0	2	1	3	2	3	3	2	1	1	0	1	2	2	1	BX
2	0	1	1	1	2	2	2	2	2	0	1	0	0	2	2	1	BX
2	0	0	2	0	2	1	3	3	2	1	0	0	1	2	2	1	BX
2	0	0	2	1	1	1	1	1	1	0	0	1	1	1	2	1	BX
2	0	0	2	1	2	1	2	2	1	1	0	1	2	1	2	1	BX
2	0	0	2	1	2	2	2	2	1	1	0	1	1	0	2	1	BX
2	0	0	2	2	3	3	3	3	1	2	1	1	1	1	2	1	BX
2	0	0	2	1	2	2	3	3	1	2	1	0	1	1	2	1	BX
2	0	0	2	2	4	3	4	4	2	2	2	0	1	2	2	1	BX
2	0	0	2	1	3	2	2	2	1	1	1	0	1	1	2	1	BX
2	1	1	1	2	2	2	2	2	1	1	2	0	1	1	2	1	BX
2	2	2	0	2	1	1	2	2	1	1	1	1	1	0	2	1	BX
2	2	2	0	3	2	2	2	2	1	1	2	1	1	1	2	1	BX
2	2	2	0	2	2	1	3	3	2	1	2	0	1	2	2	1	BX
2	2	1	0	2	1	1	2	2	1	1	1	1	1	1	2	1	BX
2	2	2	0	2	2	1	4	4	2	2	2	0	1	2	2	1	BX
2	2	2	0	1	0	0	2	2	1	1	1	0	1	1	2	1	BX
2	2	2	0	3	2	2	3	3	1	2	2	1	1	1	2	1	BX

Figure 4.12: Genotyping Results for the BX Haplotype.

It depicts the genotyping results for the BX haplotype, similar to the previous figures for AA and BB haplotypes. It underscores the variations within the BX haplotype, demonstrating the KIR gene counts and different haplotypes within this group.

4.4 Taiwanese KIR gene frequency information in Allele Frequency Net Database (AFND)



We have diligently collected data from various reputable sources, starting with our web crawl of the AFND. This database provides access to seven datasets rich in gene frequency information. Furthermore, we have acquired an additional dataset focusing exclusively on *3DSI* allele frequencies by sequencing in AFND databases. Each dataset relates to a specific subpopulation, contributing to our cohort's overall diversity and completeness.

Table 4.7 documents information relevant to the web crawler utilized in the study. It includes the database used, the sources of literature reviewed, the KIR gene detection methodology, and the labeling of the x-axis in the derived graph. Table 4.7 offers an organized overview of the behind-the-scenes data collection and processing elements that contribute to the subsequent analysis and visual representation of the findings in the study.

Furthermore, we have integrated data from two studies conducted by Taipei Medical University [Chou et al., 2020](#) and Chang Gung University [Wang et al., 2017](#), encompassing four different sub-populations, as well as being documented in the table.

The gene frequency data primarily originated from assays that utilized sequence-specific oligonucleotide probes (PCR-SSOP) or sequence-specific primers (PCR-SSP). Notably, the TWB100 dataset provides valuable insights into KIR gene frequency and is based on the analysis of 100 samples from the TWB, employing PCR-SSP. This dataset was generated in Dr. Ya-Chien Yang's laboratory, which is affiliated with the Graduate Institute of Clinical Laboratory Sciences and Medical Biotechnology at National Taiwan University.

In addition, the TWB1492 dataset represents KIR gene frequency data from 1492 samples within the TWB, and this information was derived using graphKIR software.

Figure 4.13 shows that the KIR genes *2DL1*, *2DL3*, *3DL1*, *2DL4*, *2DP1*, *3DL2*, *3DL3*, and *3DP1* each have a gene frequency of 80-100% in the analysed libraries. An exception to this pattern was observed for the KIR gene *2DL1* in the K96 database, which occurs in 65.6% of the Taiwanese KIR(K96) subpopulation, including normal people who

do not select race. Its frequency is significantly lower than that of other databases. This discrepancy may be due to differences in the population or may be due to the timing of the study and the use of PCR-SSP technology when the study was written.

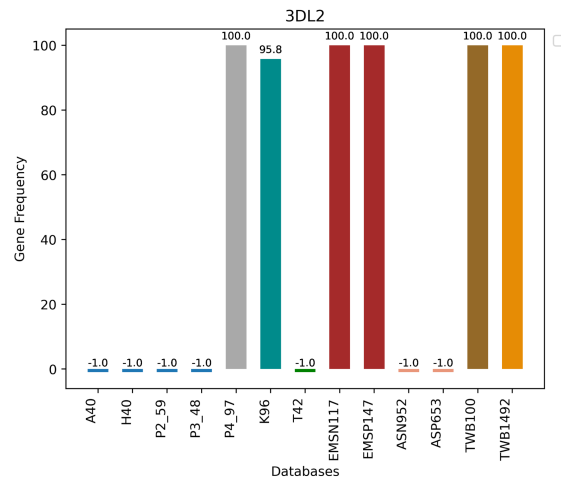
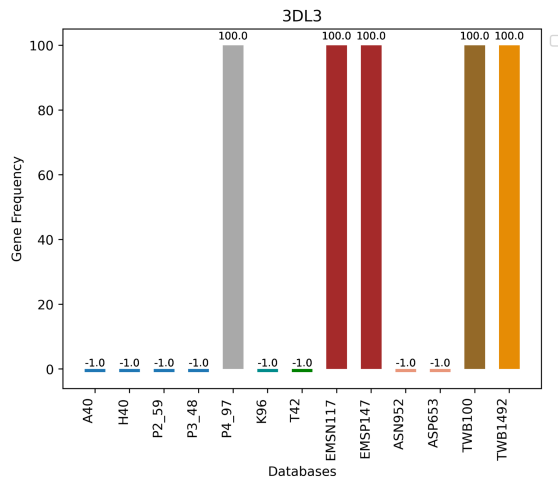
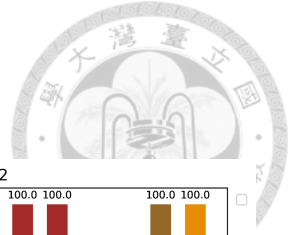
Figure 4.14 shows a visual representation of the gene frequency of certain KIR genes - *2DL5*, *2DS3*, *2DS5*, *2DS1*, *2DS2*, *2DL2*, and *3DS1* across multiple databases. Each KIR gene has a gene frequency of less than 80

In Figure 4.15, we see unique cases with the KIR genes *3DS1* and *2DS4*. The *3DS1* KIR allele was originally identified using sequencing methods. The results were converted from allele frequency to gene frequency to allow consistent comparison at the same level with gene frequency. For the Taiwanese population, we corrected the allele frequency to gene frequency for *3DS1* and added a corresponding bar (pink color) to the figure for clarity.

The gene frequency for *2DS4* is shown differently due to its partial and completely deleted forms. The frequency representation for *2DS4* in the figure follows the original representation method from the parent paper for consistency and to emphasize its unique genetic changes. This visual representation captures the nuanced dynamics of these KIR genes within the population under study.

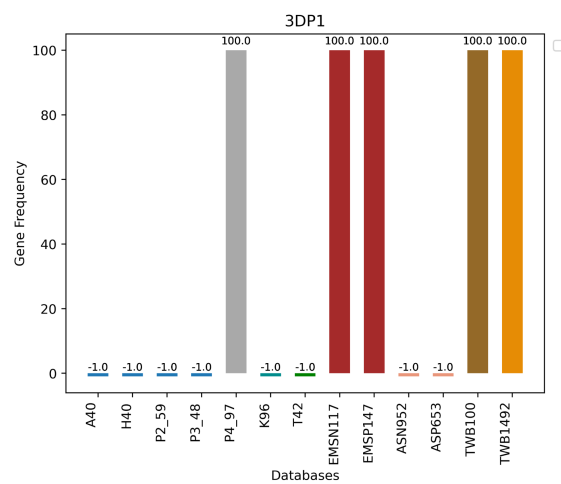
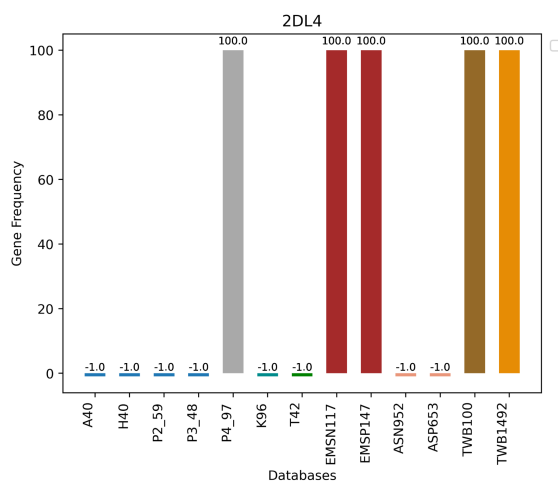
Our data collection includes a variety of sources that allow a comprehensive study of KIR gene frequencies in different subpopulations. This will enable us to analyze the differences in gene frequencies for each KIR gene in different subpopulations.

Overall, our results correct the current KIR gene frequencies and provide all KIR-associated comparisons in Taiwan, revealing a distinct range of KIR gene frequencies within this population. In addition, we extended our study by analyzing 100 TWB samples using PCR-SSP and evaluating 1,492 TWB samples using graphKIR. Comparisons of gene frequencies in different subpopulations were also performed to validate our results indirectly.



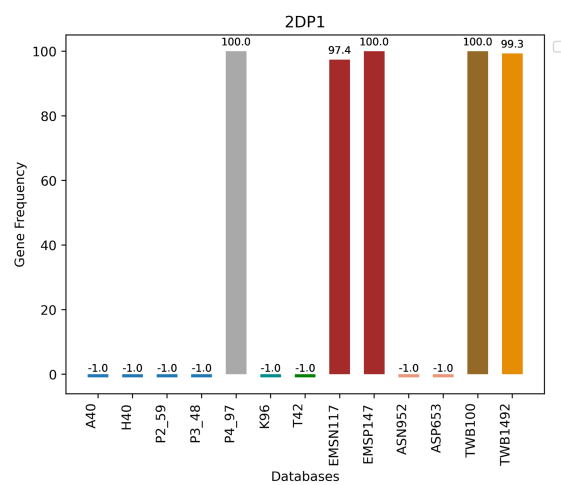
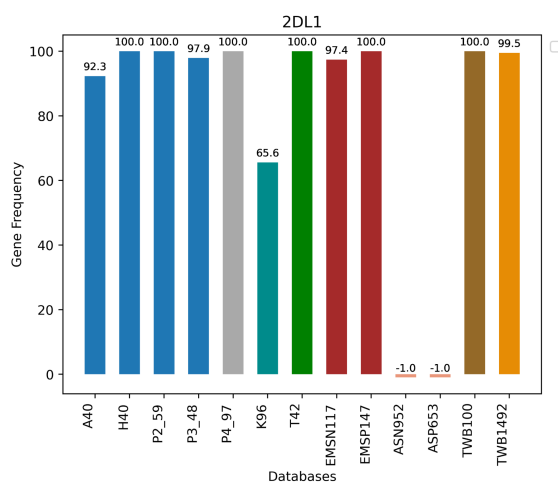
(a) 3DL3

(b) 3DL2



(c) 2DL4

(d) 3DP1



(e) 2DL1

(f) 2DP1

Table 4.7: Recorded KIR Gene Information Sourced from the AFND and Relevant.

DB name	Sample Size	Alternative name	Paper Source	Method	Population Source
Taiwan Ami KIR	40	A40	AFND[A]	SSOP	Amis
Taiwan Hakka KIR	40	H40	AFND[A]	SSOP	Hakka
Taiwan Han KIR pop 2	59	P2_59	AFND[A]	SSOP	Normal person
Taiwan Han KIR pop 3	48	P3_48	AFND[A]	SSOP	Normal person
Taiwan Han KIR pop 4	97	P4_97	AFND[B]	SSP	Normal person
Taiwan KIR	96	K96	AFND[C]	SSP	Normal person
Taiwan Taroko Atayal KIR	42	T42	AFND[A]	SSOP	Taroko Atayal
EMSN	117	EMSN117	[A]	SSOP	Normal person
EMSP	147	EMSP147	[A]	SSOP	Patient Endometriosis
ASN	952	ASN952	[B]	SSP	Normal person
ASP	653	ASP653	[B]	SSP	Patient Ankylosing Spondylitis
TWB	100	TWB100	YQ Lab	SSP	Taiwan Biobank
Taiwan Han KIR	157	Han157	AFND[D]	Sequencing	Normal person

¹ SSOP: Sequence-Specific Oligonucleotide Probe

² SSP: Sequence-Specific Primer

³ AFND[A]: [Single et al., 2007](#)

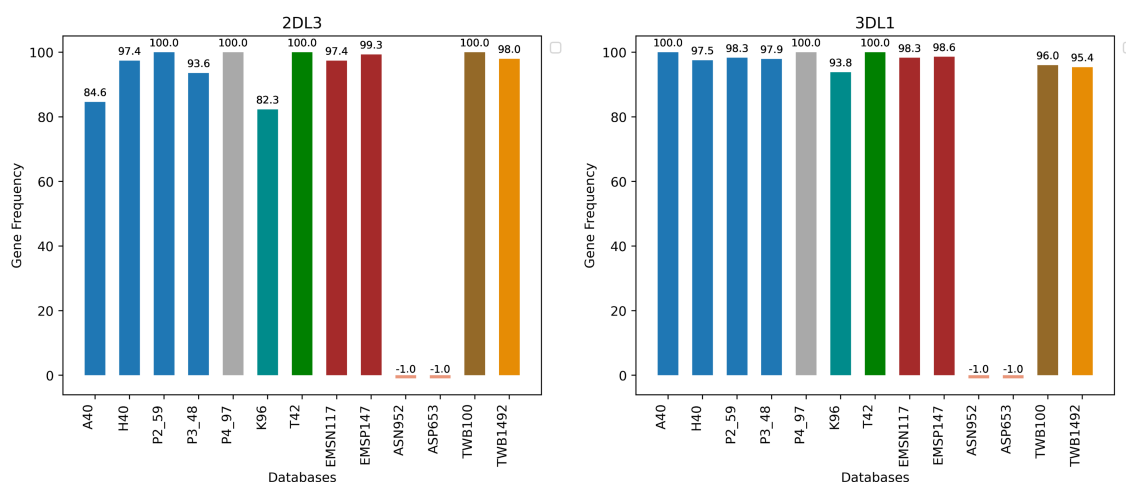
⁴ AFND[B]: [Wu et al., 2009](#)

⁵ AFND[C]: [Yen et al., 2006](#)

⁶ AFND[D]: [Norman et al., 2007](#)

^A [A]: [Wang et al., 2017](#)

^B [B]: [Chou et al., 2020](#)

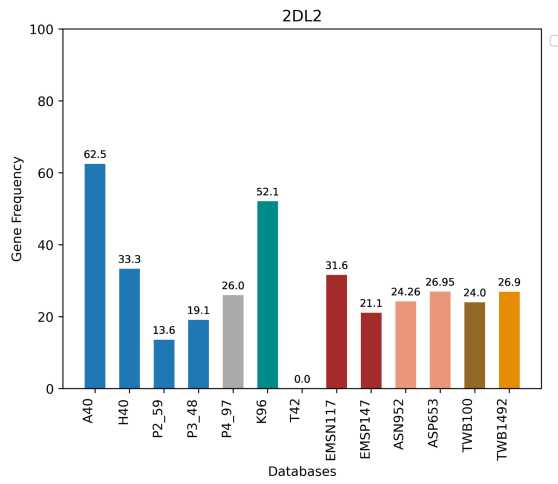


(g) 2DL3

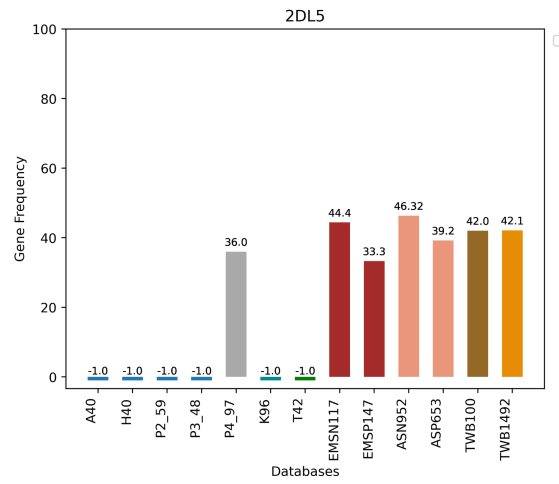
(h) 3DL1

Figure 4.13: High Gene Frequency of Selected KIR Genes.

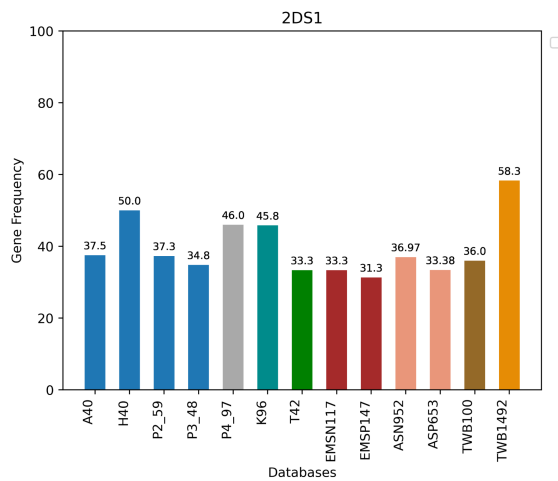
It provides a graphical representation depicting the gene frequency for specified KIR genes - *2DL4*, *3DL2*, *3DL3*, *3DP1*, *2DL1*, *2DL3*, *2DP1*, and *3DL1* within the sample set. Each gene exhibits a high frequency of 80-100%, indicating their widespread presence and possible significance within the analyzed population. The visualization offers a comprehensive overview, highlighting the genes with high prevalence rates, and serves as a pivotal insight into the genetic makeup of the studied population.



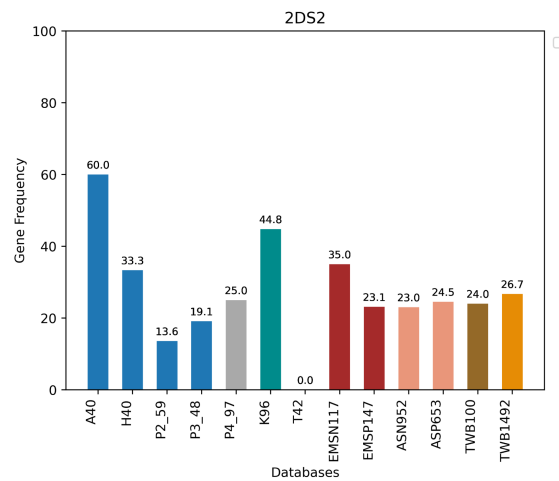
(a) 2DL2



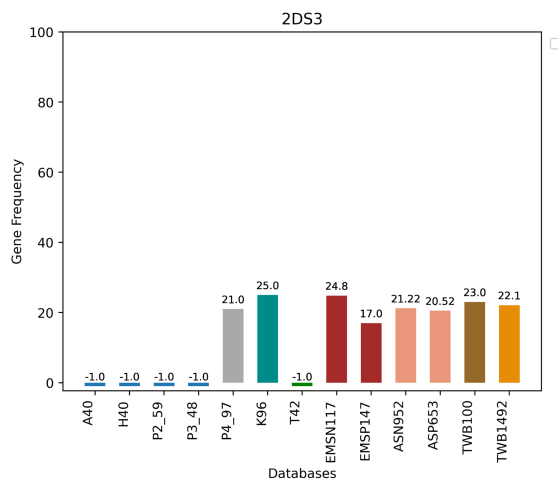
(b) 2DL5



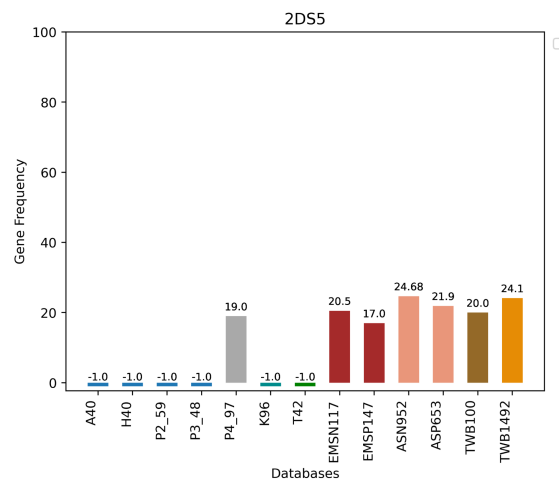
(c) 2DS1



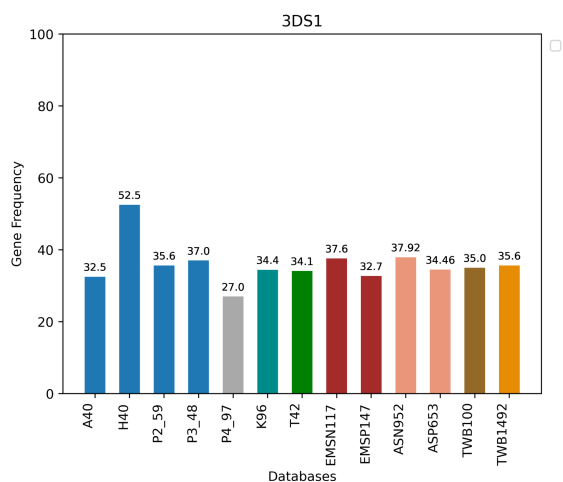
(d) 2DS2



(e) 2DS3



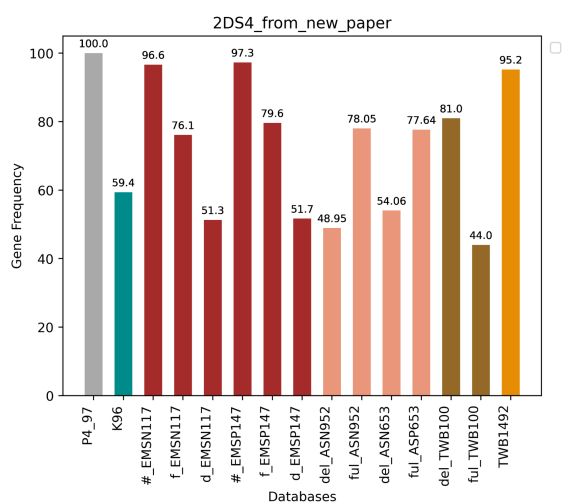
(f) 2DS5



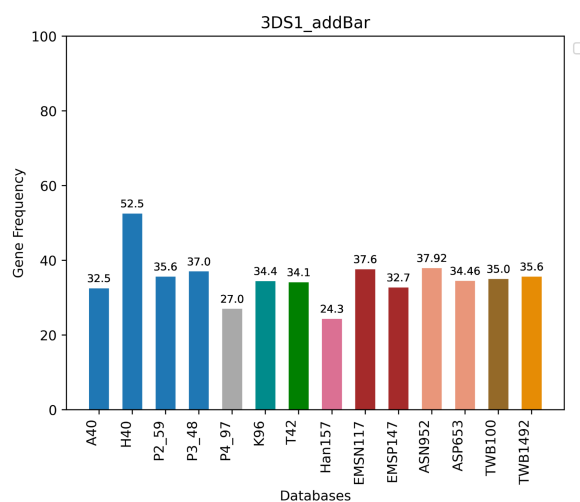
(g) 3DS1

Figure 4.14: Lower Gene Frequency of Specific KIR Genes.

It presents a detailed visualization of the gene frequency for selected KIR genes - *2DL5*, *2DS3*, *2DS5*, *2DS1*, *2DS2*, *2DL2*, and *3DS1* within the analyzed sample set. Each of these genes presents a gene frequency of less than 80%, indicating their varying presence in the studied population. The figure provides valuable insights, emphasizing the genes with lower prevalence rates, and contributes to a comprehensive understanding of the diversity and distribution of KIR genes in the observed population.



(a) 2DS4



(b) 3DS1

Figure 4.15: Unique Frequency Representations for 3DS1 and 2DS4.

It presents a distinctive gene frequency representation for KIR genes *3DS1* and *2DS4*. This figure offers a comprehensive view of the specific gene frequency conversions and representations utilized for these two genes, providing insights into their variations. Unique gene frequency presentations for *3DS1* and *2DS4* give the study a nuanced perspective on the diversity and distribution of these KIR genes within the analyzed population.

4.5 Analyzing KIR-HLA pair frequencies



The relationship between HLA and KIR genes is crucial in immune function and responses. In this study, we examined the frequencies of HLA-KIR combinations in the TWB dataset to gain insights into their prevalence and potential implications for human health. We focused on HLA alleles, excluding combinations involving KIR pseudogenes and framework genes. This approach ensured the accuracy of the data by considering only functional KIR genes.

By analyzing the frequencies of HLA (HISAT-genotype) and KIR pairs, we investigated 1,474 pair combinations, as illustrated in Figure ???. This visualization provides a clear understanding of the distribution of pair frequencies. Notably, the most common HLA Class I combination is KIR2DL1_A*11:01, accounting for 44.37% (654/1474) of cases, closely followed by KIR2DL1_C*07:02 at 35.41% (520/1474). Among Class II combinations, KIR2DL1_DPA1*02:02 is the most prevalent, with a frequency of 80.66% (1189/1474), followed by KIR2DL1_DPB1*05:01 at 62.07% (915/1474).

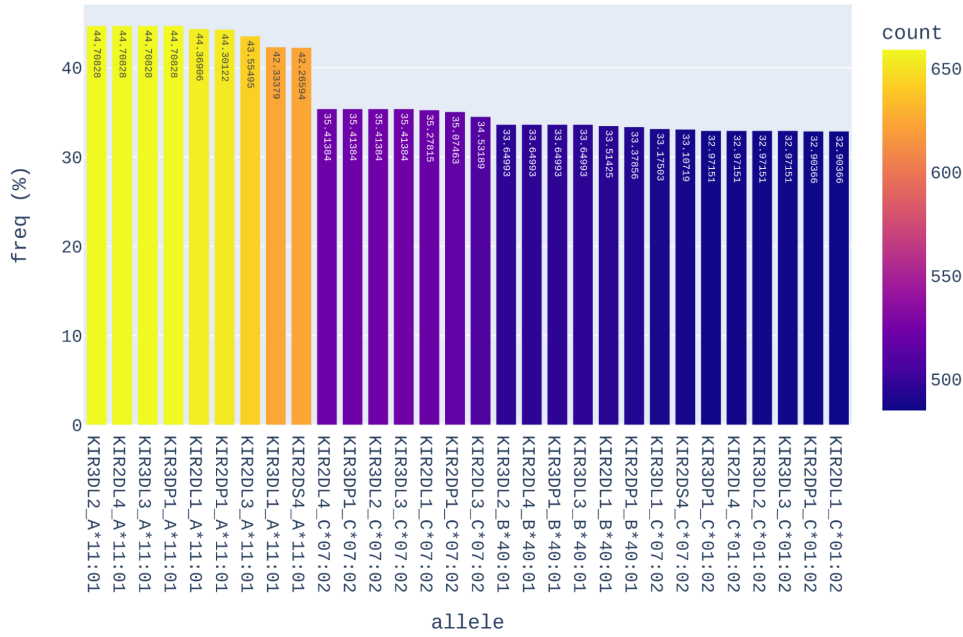
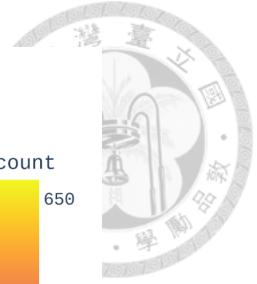
Looking at HLA (T1K)-KIR combinations, as depicted in Figure 4.17, the most frequent HLA Class I combination is KIR2DL1_A*11:01, comprising 44.50% (664/1492) of total combinations. Notably, KIR2DL1_C*07:02 closely follows with an occurrence of 35.72% (533/1492). Among Class II combinations, KIR2DL1_DPA1*02:02 is the most frequently observed at 80.56% (1202/1474), and KIR2DL1_DPB1*05:01 is the second most prevalent at 66.69% (995/1492).

The detailed visualization of HLA-KIR pairs and their frequency distributions (Figures 4.16 and 4.17) contributes to our understanding of population-based HLA-KIR combinations. The analysis of HLA-KIR pair frequencies using both T1k and HISAT-genotype produced similar results, proving the robustness of this comparison. While HISAT-genotype could type out 1,474 samples, T1k could handle all 1,492 samples, highlighting the T1k's broader reach in this analysis. Despite this difference in sample handling, the gene rankings for T1k and HISAT-genotypes reflected significant consistency.

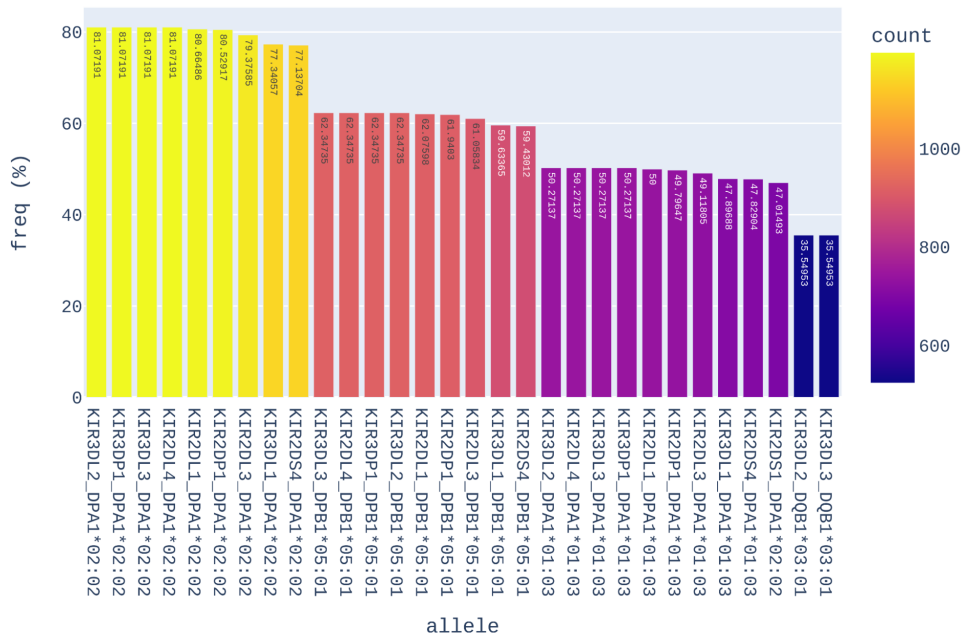
Our analysis using T1k and HISAT-genotype methods documents a largely consistent HLA-KIR gene pairing pattern across the Taiwanese population. This correlation makes

our data more reliable and provides strong evidence for understanding the prevalent HLA-KIR combinations in the Taiwanese cohort.



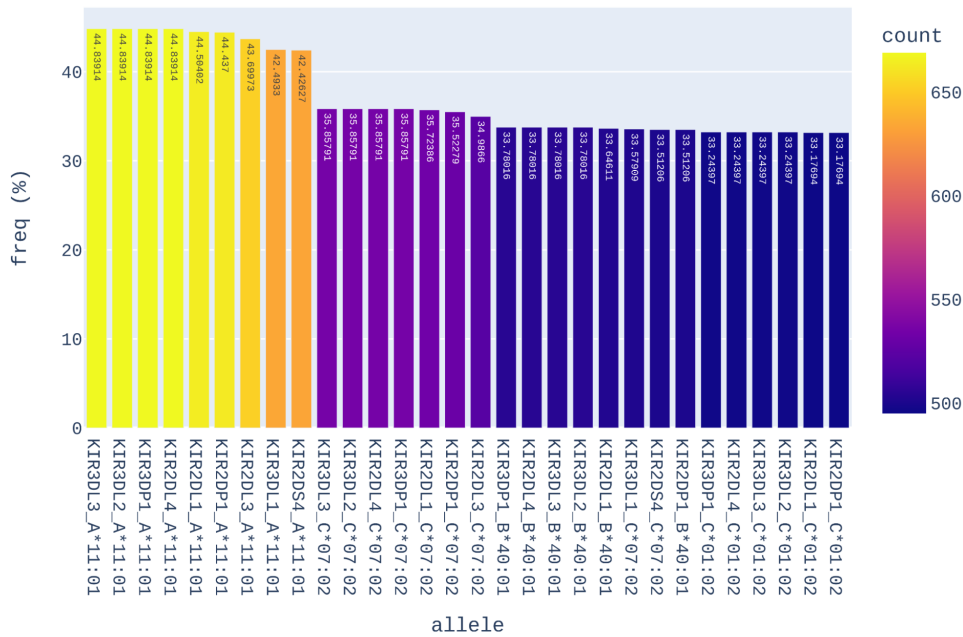
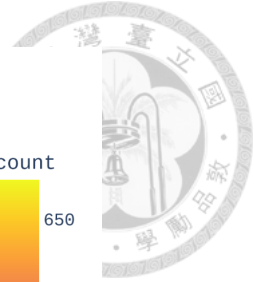


(a) HISAT-genotype classI

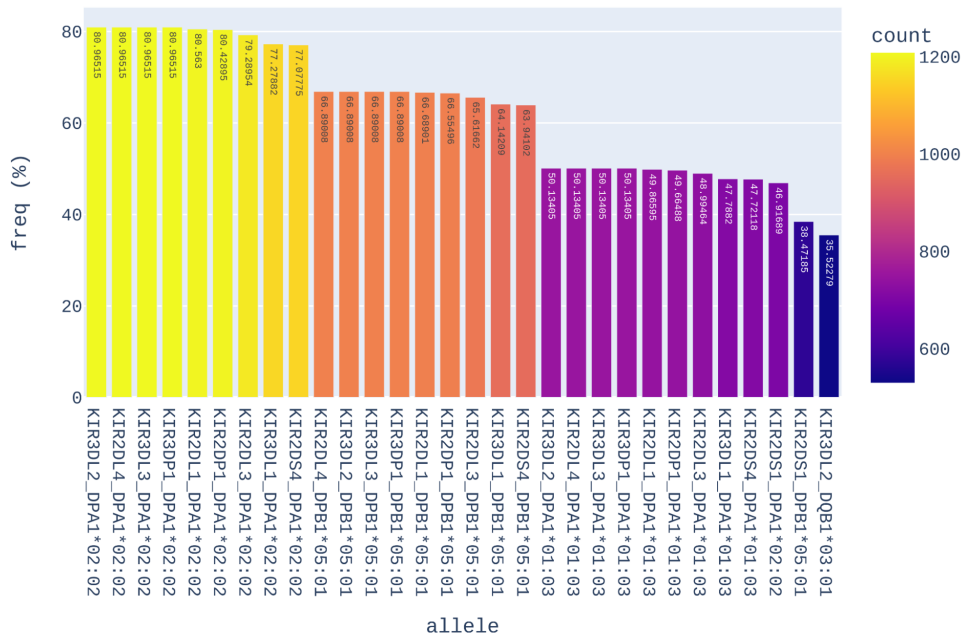


(b) HISAT-genotype classII

Figure 4.16: HLA (HISAT-genotype) Combinations With KIR Frequencies. It offers a comprehensive visual representation of the combinations of HLA (HISAT-genotype) with KIR genes and their corresponding frequencies. The X-axis represents the unique combinations of KIR with HLA, while the Y-axis denotes the respective frequencies as a percentage for these pairings. With 1474 instances considered, the visualization employs a color-coding scheme to signify the counts of each specific KIR-HLA combination, facilitating an easier understanding of their prevalence.



(a) T1k classI

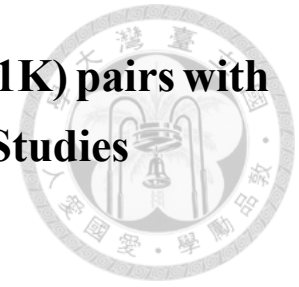


(b) T1k classII

Figure 4.17: HLA (T1K) Combinations With KIR Frequencies.

It provides a graphical overview of HLA (T1K) combinations with KIR genes and their associated frequencies. The X-axis showcases the combinations of KIR with HLA, while the Y-axis illustrates their frequencies as a percentage. A color gradient represents the counts of each unique combination out of 1492 instances. This pictorial presentation efficiently conveys the prevalence of specific KIR-HLA combinations within the analyzed data.

4.6 Studying the association of KIR-HLA(T1K) pairs with multiple phenotypes using Association Studies



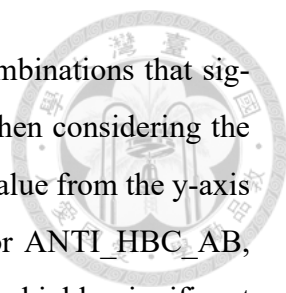
Association studies provide a thorough approach that can probe the possible influence of genetic variation on various diseases simultaneously and on a large scale. These studies often focus on the involvement of genetic factors, like KIRs and their corresponding HLAs, in immune responses and the associated risks for various diseases. These KIR-HLA(T1K) pairs play critical roles in controlling our immune responses, and changes in their interactions are tied to susceptibility to multiple diseases. Therefore, our research utilized association studies to investigate the relationship between KIR-HLA(T1K) combinations and specific traits to illuminate the genetic basis of different conditions.

To enhance the accuracy of our study, we analyzed the TWB dataset using a logistic regression model. We included KIR-HLA(T1K) combinations and their associated phenotypes in the analysis. Gender and age were also considered variables to address potential confounding factors and better understand the relationship between KIR-HLA(T1K) combinations and individual phenotypes.

The study included minor positive counts above 15 to ensure accurate analytical findings. This approach provides the representation of at least 1% of the population and excludes extremely rare alleles. By doing so, the analysis avoids any potential bias and maintains the reliability and relevance of the results.

The phenotypes analyzed in this study spanned a vast spectrum, including but not limited to anti-HCV AB, HBsAg, HBeAg, anti-HBs AB, anti-HBc AB, HbA1C, diabetes, gout, allergies, and asthma 4.18. This wide-ranging phenotype inclusion enables a comprehensive overview of the implications of KIR-HLA(T1K) combinations on various aspects of an individual's health.

Findings from this study include the use of logistic regression models for 31 phenotypes. These models were validated for accuracy using log-likelihood ratio (LLR) p-value testing, a widely accepted statistical tool to confirm the model's fit to the presented data. The identified top 20 KIR-HLA(T1K) combinations with the most significant associations were visualized to make understanding the major associations easier.

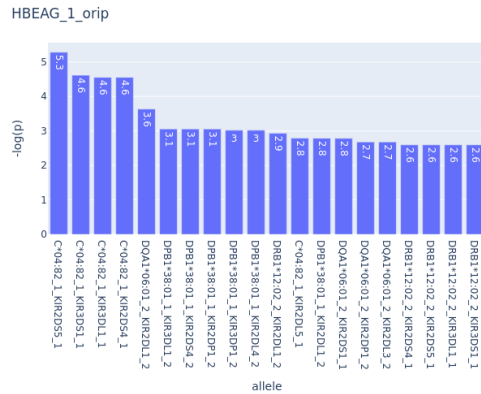


These graphical presentations aid in identifying KIR-HLA(T1K) combinations that significantly impact various phenotypes (Figure 4.19). For instance, when considering the phenotype ANTI_HBC_AB, we can observe the log-transformed p-value from the y-axis while the x-axis displays alleles that have passed the LLR test. For ANTI_HBC_AB, the leading allele is DPB1*107:01_1_KIR2DL4_2, demonstrating a highly significant association with a p-value of 0.00041. These observations indicate that the KIR-HLA combination DPB1*107:01_1_KIR2DL4_2 plays a substantial role in the expression of ANTI_HBC_AB, reinforcing the importance of both KIR-HLA combinations and genetic factors in health and disease.

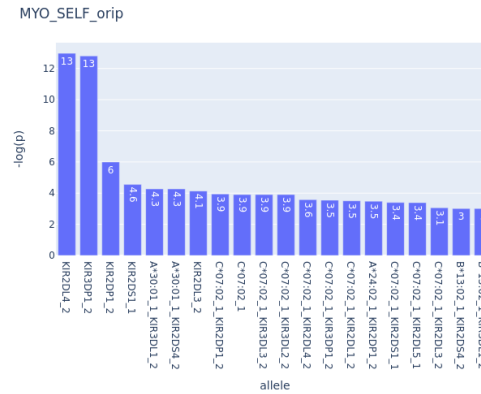
Our methodology has successfully uncovered the intricate connection between KIR-HLA(T1K) pairs and various individual traits. These discoveries enhance our understanding of how genetic factors impact disease risk and can help shape future research endeavors.

I want to thank everyone who has helped with this research and made it possible. I am grateful for the hard work of those who provided the datasets used in this study. For more information on our study's methodology, please visit our GitHub repository at https://github.com/Ritabear/all_kir_hla.

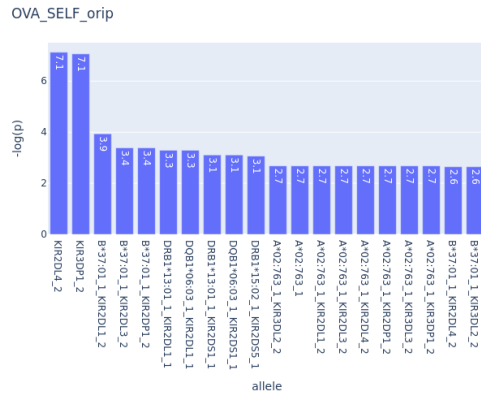
Refinement of this thesis through ChatGPT (OpenAI, 2023).



(e) HBEAG

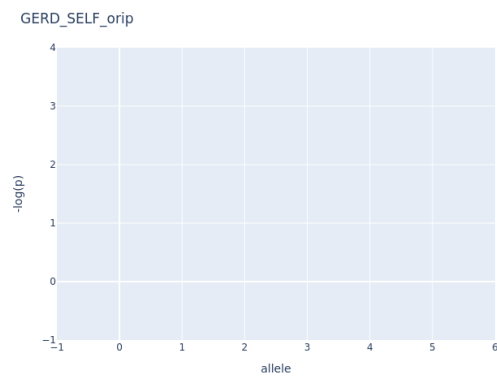


(f) MYO

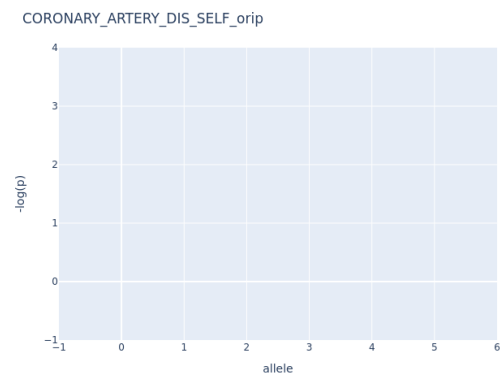


(g) OVA

Figure 4.19: Phenotypes are passing the LLR P-Value Test. This figure displays the significant associations between KIR-HLA(T1K) combinations (shown on the X-axis) and various phenotypes that pass the LLR p-value test for model accuracy. The Y-axis represents the negative logarithm of the p-value from the logistic regression for each combination, which allows for an easier visual interpretation of the significance level. A higher value on the Y-axis denotes a smaller p-value, indicating a stronger association. Of all the combinations under investigation, phenotypes ANTI_HBC_AB, CERVICAL, DYS, ENDO, HBEAG, MYO, and OVA successfully passed the verification for model accuracy. This advancement enhances the reliability and credibility of the logistic regression's p-values for these phenotypes.



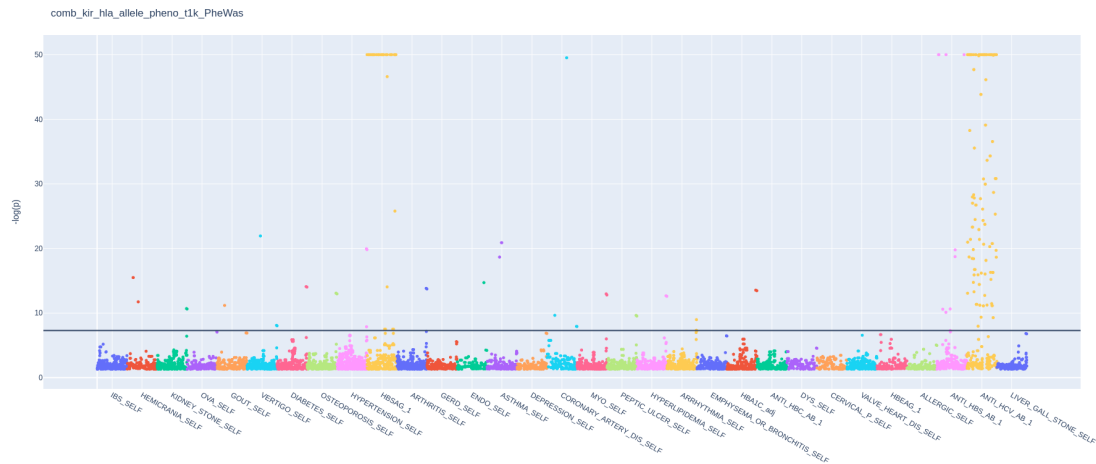
(a) GERD



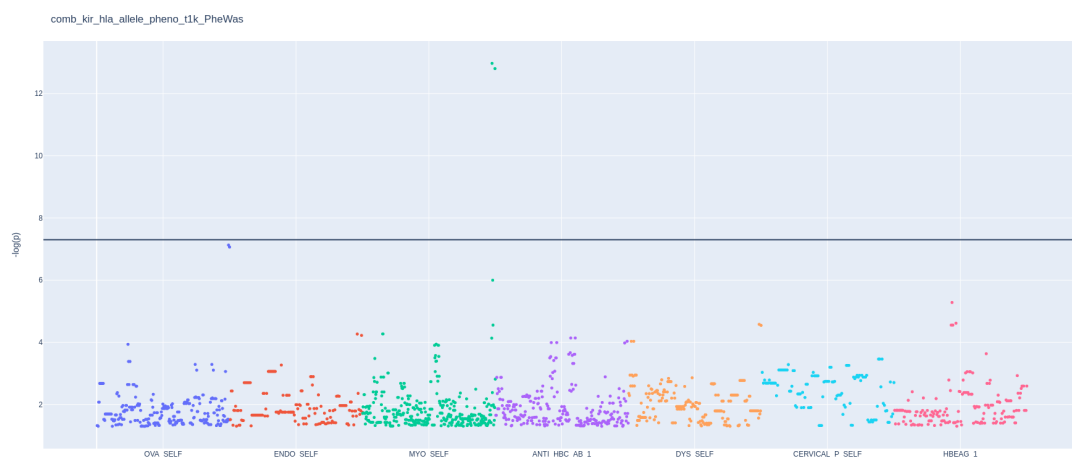
(b) CORONARY_ARTERY

Figure 4.20: Phenotypes Not Passing LLR P-Value Test.

This graph illustrates the KIR-HLA(T1K) combinations linked with phenotypes that did not meet the significance level in the LLR p-value test for the accuracy of the logistic regression model.



(a) Association Study Results Without LLR P-value Filter



(b) Association Study Results With LLR P-value Filter

Figure 4.21: Comparative Association Study Results with and without LLR P-value Filter.

Impact of LLR P-value Filtering on Genotype-Phenotype Associations: This figure displays association study results, evaluating associations between various genotypes (KIR, HLA, KIR-HLA combinations) and a wide range of phenotypes. Figure 4.21a demonstrates an unfiltered snapshot of these associations, depicting the preliminary landscape of genotype-phenotype interactions. Conversely, Figure 4.21b presents a more stringent view by applying an LLR p-value filter, zeroing in on the seven phenotypes with statistical significance.

Chapter 5 Discussion



5.1 Discussion

This study explored Taiwanese KIR gene frequencies, inspected HLA-KIR pairings, and analyzed their potential associations with a broad range of individual phenotypes. While this study offers substantial insights, various areas could be elaborated on, improved, or clarified.

5.1.1 Challenges with AFND Web Crawler Data Collection

The primary data collected from the AFND, largely based on SSOP and SSP, is a critical cornerstone of our research. However, the performance of these methodologies is inherently dependent on specific primers used, which can introduce variations arising from different laboratory protocols. Additionally, certain KIR genes appear underrepresented or absent in some datasets. This challenges making comprehensive conclusions and comparative studies across diverse populations presents a challenge.

Another point worth discussing is the evident discrepancy in KIR gene frequency, particularly for KIR gene *2DL1*. As illustrated in Figure 4.13 and 4.14, this gene demonstrates a significantly lower frequency in the K96 database than others. This divergence could be attributed to genuine genetic variations, differences in sample collection, or methodology issues related to this gene's specific detection. Understanding this discrepancy could necessitate additional research or a more detailed overview of the methodologies originally employed in data collection.

5.1.2 Limitations and Strategies in Genotyping Tools

In examining KIR-HLA interactions, our study relies on computational tools such as graphKIR and HLA typing tools such as HISAT-genotype and T1K. It is important to note

that these tools do not guarantee 100% accuracy, especially when dealing with the output of similar genes such as *2DL1* and *2DS1* or HLA allele typing results.

We strategically focus on CN results rather than allele-level profiles in the graph's KIR output to obtain more reliable KIR-HLA combinations. The reason for this decision is that CN values generally provide higher reliability. The graphKIR pipeline relies on alignments produced by the HISAT2 graph mapper, which means that limitations inherent to HISAT2 (such as its inability to handle repetitive elements typically found within the 300-1000bp region of KIR genes) may affect our analysis, resulting in occasional mapping errors. The graphKIR does not consider phasing information between different read pairs, further complicating accuracy. Therefore, graphKIR may be unable to determine the appropriate allelic variants when the distance between variants exceeds the short read length. This limitation may lead to mistyping of certain alleles.

Similar uncertainties exist in HLA typing tools (e.g., HISAT-genotype, T1K), where allelic typing results are not always accurate. We address these nuances despite these limitations, allowing us to adopt a more informed and precise approach.

5.1.3 Influence of Age and Gender on Genotype-Phenotype Associations in Taiwanese Population

In our investigation, seven phenotypes, including ANTI_HBC_AB, CERVICAL, DYS, ENDO, HBEAG, MYO, and OVA, demonstrated significant associations with specific KIR-HLA(T1K) combinations derived from logistic regression models, as presented in Figure 4.19.

However, it is interesting to note that the majority of p-values for each phenotype did not surpass the GWAS stringent p-value threshold of 5×10^{-8} , as depicted in the association study graph (Figure 4.21b). The outcome suggests that while these associations have statistical significance, the physiological implications may be less profound than anticipated.

A particular point of consideration is the nationwide implementation of the Hepatitis B vaccination program in Taiwan since July 1992. The program's widespread coverage and effectiveness may have significantly influenced the association of KIR-HLA(T1K)

combinations with Hepatitis B-related phenotypes, rendering age a significant factor. This aspect could be attributed to less pronounced p-values for genotype-phenotype associations for Hepatitis B.

Furthermore, five out of the seven significant phenotypes are female-related health conditions. This observation suggests that gender is important in shaping these associations, leading to heightened p-values.

Thus, while genotype-phenotype associations were observable and statistically significant for certain phenotypes, their physiological impact might be overshadowed by influential factors such as age and gender. This finding is crucial, as it underscores the need for a multifactorial approach toward understanding genetic influence on various health conditions and disease susceptibilities.

Our study demonstrates significant associations between KIR-HLA(T1K) combinations and multiple phenotypes in the Taiwanese population. However, the physiological implications of these associations may be modulated by confounding factors such as age and gender, underscoring the complex landscape of genetic influence on health and disease. This complexity necessitates an integrated examination of various influencing factors in genetic research to decipher the multi-dimensional nature of genotype-phenotype correlations.

5.1.4 Considerations and Challenges in Phenotypic Analysis and Genomic Interpretation

Our research utilized phenotype data from the TWB, a rich resource that encapsulates a wide range of data obtained through direct measurements and questionnaires. The selection of this data source, while extensively enriching, necessitates the consideration of potential uncertainties. With questionnaire-based data, the accuracy predominantly relies on the preciseness and reliability of respondents' inputs, presenting the potential for pre-embedded error. Subjectivity and inherent bias can contribute to inaccuracies. It thus becomes critical to navigate the interpretation of such data with due diligence and caution.

The study employed logistic regression to analyze the association between phenotypes and KIR-HLA combinations. However, it is crucial to acknowledge the statistical

considerations that inherently accompany this approach. We chose a minor positive count cut-off at 15 to eliminate excessively infrequent alleles that could skew associations. Still, this decision simultaneously risks excluding potential genetic variants with true but rare impacts on certain phenotypes.

The sample size constraint also posed an inherent limitation. Although we referred to a substantial dataset of Taiwanese individuals, increasing the sample size is almost always preferable. A larger sample size ensures heightened statistical power, and better population representation, enhances the detection of rare alleles, and procures more precise estimates of allele frequencies.

Overall, these considerations affirm the necessity for meticulous attention to the multifaceted elements both in the practical aspects of the study and in the interpretational components. By doing so, we ensure the findings drawn are reliable, valid, and, most importantly, reflective of the actualities of the complex genomic landscapes.

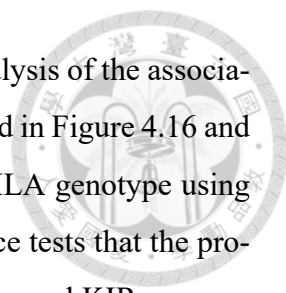
Chapter 6 Conclusion



The primary goal of this research is to explore and elucidate the complex characteristics of Taiwanese KIR gene frequencies, their KIR-HLA pair combinations, and the associations between these pairs and multiple phenotypes in individuals. The study design leverages a comprehensive dataset and statistical techniques to explore these interactions and their impacts on individual phenotypes.

We employed a three-pronged approach to fulfill this objective based on potent computational methods and vast empirical data. These facets of the research encompass the analysis of Taiwanese KIR gene frequencies and KIR-HLA pair frequencies and use logistic regression to study the relation of these pairs with multiple phenotypes. We employ a three-pronged approach to fulfill this objective based on potent computational methods and extensive empirical data. These facets of the research encompass the analysis of Taiwanese KIR gene frequencies and KIR-HLA pair frequencies and using logistic regression to study the relation of these pairs with multiple phenotypes. The analysis of KIR gene frequencies reveals valuable insights about the KIR genotypic makeup of the Taiwanese population. Careful analysis of multiple datasets provides a comprehensive understanding of KIR gene frequencies in Taiwanese populations. Primary data comes from the AFND, a reputable source of gene frequency information. We further supplement this with additional data from two Taiwanese studies at Taipei Medical University and Chang Gung Memorial University. By including data from Dr. Ya-Chien Yang's laboratory, we add different dimensions to our analysis. We also have the TWB1492 dataset, involving a sample of 1492 individuals within the TWB, analyzed using graphKIR software.

Through exhaustive data collection and precision-enhanced visualization (as shown in Figures 4.13, 4.14, and 4.15), we successfully plot the frequency distributions of multiple KIR genes across various subpopulations. Our findings indicate a nuanced diversity concerning KIR gene distributions, thereby highlighting the genetically unique attributes associated with different groups within the Taiwanese population.



We expand our research's horizon by conducting an in-depth analysis of the associations between KIR genes and their respective HLA ligands, as depicted in Figure 4.16 and Figure 4.17. This exploration is undergirded by benchmarking the HLA genotype using the HPRC as referenced in Figure 4.1. It is clear from the performance tests that the proficiency of each tool varies depending on the data type used. Notably, graphKIR emerges as a better performer with WGS data, while capKIR displays fewer CN errors with panel data. Therefore, depending on the data type, the appropriate tool selection remains crucial for reliable KIR typing, enhancing the study outcomes.

Moreover, given the tools' differential performance, we recommend using graphKIR for WGS data and capKIR for panel data for optimal KIR typing. This customized selection of tools is pivotal in yielding superior outcomes in genomic research.

Lastly, we conduct an exhaustive association study to study the impact of KIR-HLA interactions on multiple phenotypes. Methodologically, we leverage logistic regression tools with inclusion variables such as age and gender, and the logistic regression model is utilized to enhance the accuracy and interpretability of the obtained results. To improve the accuracy and validity of our findings, minor allele counts greater than 15 are factored into the study to ensure the observed results represent at least 1% of the population. In terms of phenotyping, a comprehensive range of well-distinguished phenotypic features such as anti-HCV AB, HBsAg, HBeAg, diabetes, gout, arthritis, hypertension, and blood lipids, among others, are incorporated in our analysis, as depicted in the results table (Table 4.18).

Our study yields extensive data on KIR gene frequencies and their interactions with HLA alleles in Taiwanese populations. These findings deepen our understanding of these gene combinations in the context of individual phenotypes and reveal the compelling relevance of these genetic factors in human health and disease. Furthermore, the data collected in this study provide a solid foundation for subsequent studies to delve into the world of KIR gene interactions. Comprehensive analysis of KIR gene frequencies, KIR-HLA pair association, and association study applications highlights the impact of these genes on the human immune system and disease susceptibility.

Our combined analyses provide a unique piece of this complex puzzle, promoting a more comprehensive understanding of the vast diversity of the genomic landscape that

controls human health and disease. Hopefully, this study can provide a useful reference for future research.







References

- Richard Apps, Lucy Gardner, and Ashley Moffett. A critical look at hla-g. Trends in immunology, 29(7):313–321, 2008.
- Khadijeh Dizaji Asl, Kobra Velaei, Ali Rafat, Hamid Tayefi Nasrabadi, Ali Akbar Movas-saghpour, Majid Mahdavi, and Hojjatollah Nozad Charoudeh. The role of kir positive nk cells in diseases and its importance in clinical intervention. International Immunopharmacology, 92:107361, 2021.
- Dominic J Barker, Giuseppe Maccari, Xenia Georgiou, Michael A Cooper, Paul Flicek, James Robinson, and Steven GE Marsh. The ipd-imgt/hla database. Nucleic Acids Research, 51(D1):D1053–D1060, 2023.
- Sung Yoon Choo. The hla system: genetics, immunology, clinical testing, and clinical implications. Yonsei medical journal, 48(1):11–23, 2007.
- Ya-Ching Chou, Chi-Huang Chen, Ming-Jer Chen, Ching-Wen Chang, Pi-Hua Chen, Mu-Hsien Yu, Yi-Jen Chen, Eing-Mei Tsai, Peng-Sheng Yang, Shyr-Yeu Lin, et al. Killer cell immunoglobulin-like receptors (kir) and human leukocyte antigen-c (hla-c) allorecognition patterns in women with endometriosis. Scientific reports, 10(1):4897, 2020.
- Chien-Te Fan, Jui-Chu Lin, and Chung-His Lee. Taiwan biobank: a project aiming to aid taiwan’ s transition into a biomedical island. 2008.
- Sherif S Farag, Todd A Fehniger, Loredana Ruggeri, Andrea Velardi, and Michael A Caligiuri. Natural killer cell receptors: new biology and insights into the graft-versus-leukemia effect. Blood, The Journal of the American Society of Hematology, 100(6): 1935–1947, 2002.
- Jinping Gao, Caihong Zhu, Zhengwei Zhu, Lili Tang, Lu Liu, Leilei Wen, and Liangdan Sun. The human leukocyte antigen and genetic susceptibility in human diseases. journal of Bio-X Research, 2(03):112–120, 2019.

Faviel F Gonzalez-Galarza, Antony McCabe, Eduardo J Melo dos Santos, Louise Takeshita, Gurpreet Ghattaoraya, Andrew R Jones, and Derek Middleton. Allele frequency net database. HLA typing: methods and protocols, pages 49–62, 2018.

Faviel F Gonzalez-Galarza, Antony McCabe, Eduardo J Melo dos Santos, James Jones, Louise Takeshita, Nestor D Ortega-Rivera, Glenda M Del Cid-Pavon, Kerry Ramsbottom, Gurpreet Ghattaoraya, Ana Alfirovic, et al. Allele frequency net database (afnd) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic acids research, 48(D1):D783–D788, 2020.

Zorana Grubic. Human leucocyte antigens in blood transfusion. ISBT Science Series, 15 (1):164–173, 2020.

Jeffrey R Gruen and Sherman M Weissman. Human mhc class iii and iv genes and disease associations. Front Biosci, 6:D960–972, 2001.

Wu Hejing, Liu Fang, Zhao Long, Shao Yabin, and Cui Ran. Application research of crawler and data analysis based on python. International Journal of Advanced Network, Monitoring and Controls, 5(2):64–70, 2020.

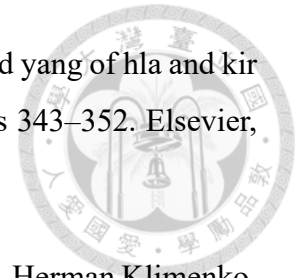
Katharine C Hsu, Shohei Chida, Daniel E Geraghty, and Bo Dupont. The killer cell immunoglobulin-like receptor (kir) genomic region: gene-order, haplotypes and allelic polymorphism. Immunological reviews, 190(1):40–52, 2002.

Tsung-Kai Hung, Wan-Chi Liu, Sheng-Kai Lai, Hui-Wen Chuang, Yi-Che Lee, Hong-Ye Lin, Chia-Lang Hsu, Chien-Yu Chen, Ya-Chien Yang, Jacob Shujui Hsu, et al. Genetic diversity and structural complexity of the killer-cell immunoglobulin-like receptor gene complex: A comprehensive analysis using human pangenome assemblies. bioRxiv, pages 2023–11, 2023.

Md Abu Kausar, VS Dhaka, and Sanjeev Kumar Singh. Web crawler: a review. International Journal of Computer Applications, 63(2):31–36, 2013.

Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nature biotechnology, 37(8):907–915, 2019.

Smita Kulkarni, Maureen P Martin, and Mary Carrington. The yin and yang of hla and kir in human disease. In Seminars in immunology, volume 20, pages 343–352. Elsevier, 2008.



Georgy Lebedev, Eduard Fartushnyi, Igor Fartushnyi, Igor Shaderkin, Herman Klimenko, Pavel Kozhin, Konstantin Koshechkin, Ilya Ryabkov, Vadim Tarasov, Evgeniy Morozov, et al. Technology of supporting medical decision-making using evidence-based medicine and artificial intelligence. Procedia Computer Science, 176:1703–1712, 2020.

Patrizia Leone, Valli De Re, Angelo Vacca, Franco Dammacco, and Vito Racanelli. Cancer treatment and the kir–hla system: an overview. Clinical and experimental medicine, 17: 419–429, 2017.

Katarina Ludajic, Yesilda Balavarca, Heike Bickeboeller, Agathe Rosenmayr, I Fae, Gottfried F Fischer, Michal Kouba, David Pohlreich, Peter Kalhs, and Hildegard T Greinix. Kir genes and kir ligands affect occurrence of acute gvhd after unrelated, 12/12 hla matched, hematopoietic stem cell transplantation. Bone marrow transplantation, 44(2): 97–103, 2009.

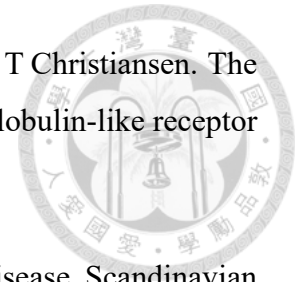
Batool Mutar Mahdi. A glow of hla typing in organ transplantation. Clinical and translational medicine, 2:1–5, 2013.

Wesley M Marin, Ravi Dandekar, Danillo G Augusto, Tasneem Yusufali, Bianca Heyn, Jan Hofmann, Vinzenz Lange, Jürgen Sauter, Paul J Norman, and Jill A Hollenbach. High-throughput interpretation of killer-cell immunoglobulin-like receptor short-read sequencing data with ping. PLoS computational biology, 17(8):e1008904, 2021.

Steven GE Marsh, Peter Parham, Bo Dupont, Daniel E Geraghty, John Trowsdale, Derek Middleton, Carlos Vilches, Mary Carrington, Campbell Witt, Lisbeth A Guethlein, et al. Killer-cell immunoglobulin-like receptor (kir) nomenclature report, 2002. Immunogenetics, 55:220–226, 2003.

Steven GE Marsh, ED Albert, WF Bodmer, RE Bontrop, B Dupont, HA Erlich, M Fernández-Viña, DE Geraghty, R Holdsworth, CK Hurley, et al. Nomenclature for factors of the hla system, 2010. Tissue antigens, 75(4):291, 2010.

Annalise M Martin, Elizabeth M Freitas, Campbell S Witt, and Frank T Christiansen. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (kir) gene cluster. Immunogenetics, 51:268–280, 2000.



YM Mosaad. Clinical role of human leukocyte antigen in health and disease. Scandinavian journal of immunology, 82(4):283–306, 2015.

Naoki Nariai, Kaname Kojima, Sakae Saito, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, Jun Yasuda, and Masao Nagasaki. Hla-vbseq: accurate hla typing at full resolution from whole-genome sequencing data. In BMC genomics, volume 16, pages 1–6. Springer, 2015.

Inger S Nijhof, Jeroen J Lammerts van Bueren, Berris van Kessel, Pascale Andre, Yannis Morel, Henk M Lokhorst, Niels WCJ van de Donk, Paul WHI Parren, and Tuna Mutis. Daratumumab-mediated lysis of primary multiple myeloma cells is enhanced in combination with the human anti-kir antibody iph2102 and lenalidomide. Haematologica, 100(2):263, 2015.

Paul J Norman, Laurent Abi-Rached, Ketevan Gendzekhadze, Daniel Korbel, Michael Gleimer, Don Rowley, Dan Bruno, Christine VF Carrington, Dasdayanee Chandanayingyong, Yih-Hsin Chang, et al. Unusual selection on the kir3dl1/s1 natural killer cell receptor in africans. Nature genetics, 39(9):1092–1099, 2007.

Paul J Norman, Jill A Hollenbach, Neda Nemat-Gorgani, Wesley M Marin, Steven J Norberg, Elham Ashouri, Jyothi Jayaraman, Emily E Wroblewski, John Trowsdale, Raja Rajalingam, et al. Defining kir and hla class i genotypes at highest resolution via high-throughput sequencing. The American Journal of Human Genetics, 99(2):375–391, 2016.

OpenAI. ChatGPT (Mar 14 version). Online, 2023. URL <https://chat.openai.com/chat>. Large language model.

Raja Rajalingam. The impact of hla class i-specific killer cell immunoglobulin-like receptors on antibody-dependent natural killer cell-mediated cytotoxicity and organ allograft rejection. Frontiers in immunology, 7:585, 2016.

James Robinson, Jason A Halliwell, Hamish McWilliam, Rodrigo Lopez, and Steven GE Marsh. Ipd—the immuno polymorphism database. Nucleic acids research, 41(D1): D1234–D1240, 2012.

David Roe, Cynthia Viera-Green, Chul-Woo Pyo, Kevin Eng, Richard Hall, Rui Kuang, Stephen Spellman, Swati Ranade, DE Geraghty, and Martin Maiers. Revealing complete complex kir haplotypes phased by long-read sequencing technology. Genes & Immunity, 18(3):127–134, 2017.

Saori Sakaue, Kazuyoshi Hosomichi, Jun Hirata, Hirofumi Nakaoka, Keiko Yamazaki, Makoto Yawata, Nobuyo Yawata, Tatsuhiko Naito, Junji Umeno, Takaaki Kawaguchi, et al. Decoding the diversity of killer immunoglobulin-like receptors by deep sequencing and a high-resolution imputation method. Cell Genomics, 2(3), 2022.

Bertil Schmidt and Andreas Hildebrandt. Next-generation sequencing: big data meets high performance computing. Drug discovery today, 22(4):712–717, 2017.

Richard M Single, Maureen P Martin, Xiaojiang Gao, Diogo Meyer, Meredith Yeager, Judith R Kidd, Kenneth K Kidd, and Mary Carrington. Global diversity and evidence for coevolution of kir and hla. Nature genetics, 39(9):1114–1119, 2007.

Li Song, Gali Bai, X Shirley Liu, Bo Li, and Heng Li. T1k: efficient and accurate kir and hla genotyping with next-generation sequencing data. bioRxiv, pages 2022–10, 2022.

JL Touraine, H Betuel, C Pouteil-Noble, and C Royo. Hla class ii antigens: structure, function, and expression in immunodeficiencies, autoimmune diseases, and allograft rejection. Advances in nephrology from the Necker Hospital, 18:325–334, 1989.

John Trowsdale. Genomic structure and function in the mhc. Trends in Genetics, 9(4): 117–122, 1993.

Carlos Vilches and Peter Parham. Kir: diverse, rapidly evolving receptors of innate and adaptive immunity. Annual review of immunology, 20(1):217–251, 2002.

Chin-Man Wang, Sheng-Hung Wang, Yeong-Jian Jan Wu, Jing-Chi Lin, Jianming Wu, and Ji-Yih Chen. Human leukocyte antigen c* 12: 02: 02 and killer immunoglobulin-like receptor 2dl5 are distinctly associated with ankylosing spondylitis in the taiwanese. International journal of molecular sciences, 18(8):1775, 2017.

G-Q Wu, Y-M Zhao, X-Y Lai, K-L Yang, F-M Zhu, W Zhang, W Wang, Y Luo, Y-M Tan, Y-J Wang, et al. Distribution of killer-cell immunoglobulin-like receptor genes in eastern mainland chinese han and taiwanese han populations. Tissue antigens, 74(6): 499–507, 2009.

Xiuhua Yang, Yahui Yang, Yiru Yuan, Lin Liu, Tao Meng, et al. The roles of uterine natural killer (nk) cells and kir/hla-c combination in the development of preeclampsia: a systematic review. BioMed research international, 2020, 2020.

J-H Yen, C-H Lin, W-C Tsai, C-C Wu, T-T Ou, C-J Hu, and H-W Liu. Killer cell immunoglobulin-like receptor gene's repertoire in rheumatoid arthritis. Scandinavian journal of rheumatology, 35(2):124–127, 2006.

Rolf M Zinkernagel and Peter C Doherty. Restriction of in vitro t cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature, 248(5450):701–702, 1974.