**Title**
Development of Bioinformatics Methods to Interrogate Complex Immune Related Genomic Regions from Next Generation Sequencing Data



**Author**
Marin, Wesley Michael

**Publication Date**
2022

# Development of Bioinformatics Methods to Interrogate Complex Immune Related Genomic Regions from Next Generation Sequencing Data

by
Wesley Marin

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Jill Hollenbach
33376559E3B9479...

Jill Hollenbach
Chair

DocuSigned by:
Michael Wilson
[illegible]493...

Michael Wilson

DocuSigned by:
Seielstad, Mark
54A1F0A578614C2...

Seielstad, Mark

Committee Members

*Dedicated to Laurie Wade, Jill Hollenbach, and Andrea Zavala*

*for their dedication to learning and understanding.*

**Acknowledgements**

There have been many people that have made this dissertation possible, and it is my pleasure to highlight a select few who have made a large impact on this project professionally and in my personal life.

First and foremost, my PI, Jill Hollenbach, who gave me the space I needed to work through problems on my own; but was also always available for project guidance, mentorship and career guidance. I was the first research student in the Hollenbach lab, and it has been amazing to be a part of the development of the lab over the past 8 years. My friend and coworker, Danillo Augusto, who is a molecular biology miracle worker and has been an absolute pleasure to work with. Being able to work directly with Danillo, drawing from both his scientific experience and molecular biology expertise, has improved many elements of my projects. Best of luck Danillo with your new lab and faculty appointment, this is a well-deserved opportunity! My friend and former coworker, Ravi Dandekar, who is the only other person that has contributed to the PING codebase. It still amazes me that someone was able to decipher my code and provide improvements. My former mentor, Paul Norman, a KIR research legend who entrusted me with the development of the prototype PING workflow way back when I had no idea what I was doing. My thesis committee member, Mark Seielstad, who has offered helpful and insightful commentary on the technical aspects of my projects. My thesis committee member, Michael Wilson, who has offered helpful, forward-thinking commentary on my projects, and has a wonderful positive energy.

In my personal life, my life partner, Andrea Zavala, has been an inspiration, a source of positive energy, and my anchor through this whole process. Andrea has been a huge source of support through this journey and has been by my side through all the highs and lows. You are very special to me Andrea, and I am excited for the next chapter of our lives. My parents, Juan Marin and Laurie Wade, who instilled in

me a resourceful and determined mindset that made an undertaking like this possible. My brother, Daniel Marin, and his friends Mark and Caleb, with whom I have enjoyed many DOTA2 games when I needed a break from my work.

## Contributions

**Chapter 2:** Marin WM, Dandekar R, Augusto DG, Yusufali T, Heyn B, Hofmann J, et al. High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING. PLOS Comput Biol. 2021 Aug 2;17(8):e1008904.

**Chapter 3:** Unpublished work that has not yet been submitted for publication. Marin WM, Hollenbach JA. Software Update: Interpreting Killer-cell Immunoglobulin-like Receptor from Whole Genome Sequencing Data with PING. [submission pending]

**Chapter 4:** Unpublished work that has not yet been submitted for publication. Marin WM, Augusto DG, Hollenbach JA. High-throughput complement component 4 genomic sequence analysis with C4Investigator. [submission pending]

**Development of Bioinformatics Methods to Interrogate Complex Immune Related Genomic Regions from Next Generation Sequencing Data**

**Wesley M. Marin**

## Abstract

The killer-cell immunoglobulin-like receptor (*KIR*) gene complex, located in human chromosomal region 19q13.42, and the complement component 4 (*C4*) gene complex, located in human chromosomal region 6p21.33, encode for proteins that have vital roles in immune system function. Component genes of these complexes exhibit copy number variation (CNV), extensive nucleotide polymorphisms, and high sequence similarity with other genes of their complex. Next generation sequencing (NGS) has transformed the world of genomics, offering a high-throughput, high-fidelity and cost-effective sequencing method, however, NGS analysis of the *KIR* and *C4* regions has been thwarted due to the bioinformatics challenges imposed by their complex variation. In this work, the researcher presents the bioinformatics pipelines, PING, developed for *KIR* sequence analysis, and C4Investigator, developed for *C4* sequence analysis. These bioinformatics pipelines provide comprehensive, high-throughput characterization of human *KIR* and *C4* sequence variation from NGS data. These pipelines take in paired-end short-read sequencing data and output gene copy number for both genomic regions, high-resolution genotypes for the *KIR* complex, and high-resolution mapping of single nucleotide variants (SNVs) for the *C4* region. The performance of PING was evaluated by real-world and synthetic datasets, while the performance of C4Investigator was evaluated by real-world datasets and comparison to existing methods. Both PING and C4Investigator showed high performance for copy number determination and SNV characterization. To demonstrate the utility of the C4Investigator pipeline, the researcher applied C4Investigator to whole genome sequencing (WGS) data from the 1000 Genomes Project (1KGP) cohort (N=3199), characterizing *C4* copy number and sequence variation for the first time

in this dataset. To demonstrate the utility of the PING pipeline, the researched applied PING to targeted sequencing datasets from divergent populations (European N=363, Khoesan N=104), in addition to WGS data from the 1KGP cohort (N=215). To the best of our knowledge, PING and C4Investigator are the only bioinformatics workflows currently available for assessment of *KIR* and *C4* full genomic sequence variation from NGS data.

**Table of Contents**

**List of Figures**

**List of Tables**

**List of Abbreviations**

1KGP -- 1000 Genomes Project

ACB -- African Caribbean in Barbados

AFR -- African superpopulation

AMR -- Admixed American superpopulation

ASW -- American's of African Ancestry in SW USA

BAM -- Binary alignment map

BEB -- Bengali from Bangladesh

C1 -- Complement component 1

C2 -- Complement component 2

C3 -- Complement component 3

C4 -- Complement component 4

C5 -- Complement component 5

CDX -- Chinese Dai in Xishuanagbanna, China

CEU -- Utah Residents (CEPH) with Northern and Western European ancestry

Ch -- Chido

CHB -- Han Chinese in Beijing, China

CHS -- Southern Han Chinese

CLM -- Colombian from Medellian, Colombia

CNV -- copy number variation

ddPCR -- Digital droplet polymerase chain reaction

EAS -- East Asian superpopulation

ESN -- Esan in Nigera

EUR -- European superpopulation

FIN -- Finnish in Finland

GBR -- British in England and Scotland

GCN -- Gene copy number

GIH -- Gujarati Indian from Houston, Texas

GWD -- Mandinka in The Gambia

HERV -- Human endogenous retrovirus

HLA – Human leukocyte antigen

IBS -- Iberian population in Spain

IPD-KIR -- Immuno Polymorphism Database - killer-cell immunoglobulin-like receptor

ITU -- Indian Telugu from the UK

JPT -- Japanese in Tokyo, Japan

KHV -- Kinh in Ho Chi Minh City, Vietnam

KIR -- Killer-cell immunoglobulin-like receptor

LWK -- Luhya in Webuye, Kenya

MAC -- Membrane attack complex

MASP-2 -- Mannan-binding lectin serine protease 2

MSL -- Mende in Sierra Leone

MXL -- Mexican Ancestry from Los Angeles USA

NGS -- Next generation sequencing

NK -- Natural Killer

PEL -- Peruvian from Lima, Peru

PING -- Pushing Immunogenetics to the Next Generation

PJL -- Punjabi from Lahore, Pakistan

PUR -- Puerto Rican from Puerto Rica

Rg -- Rodger

SAS -- South Asian superpopulation

SNP -- Single nucleotide polymorphism

SNV -- Single nucleotide variant

STU -- Sri Lankan Tamil from the UK

TSI -- Toscani in Italia

UTR -- Untranslated region

WGS -- Whole genome sequencing

YRI -- Yoruba in Ibadan, Nigera

# Chapter 1: Introduction

The killer-cell immunoglobulin-like receptor (*KIR*) gene complex, located in human chromosomal region 19q13.42, and the complement component 4 (*C4*) gene complex, located in human chromosomal region 6p21.33, encode for proteins that have vital roles in immune system function. Natural killer (NK) cells express KIR proteins which interact with human leukocyte antigen (HLA) ligands, in addition to non-HLA molecules, to modulate NK cell activity(1–3). C4 proteins, C4A and C4B, are central components of the complement system, which has many roles in immune system function(4).

Next generation sequencing (NGS) has transformed the world of genomics(5), offering a high-throughput, high-fidelity and cost-effective sequencing method. However, NGS analysis of the *KIR* and *C4* regions has been thwarted due to the bioinformatics challenges imposed by their complex variation. Component genes of these complexes exhibit copy number variation (CNV)(6,7), extensive nucleotide polymorphisms, and high sequence similarity with other genes of their complex(8). These characteristics complicate and confound analysis of NGS data from these regions.

In a typical NGS alignment and interpretation workflow, sequence data are aligned to a reference and the resultant read mappings are transformed into variant calls for each mapped position. Standard tools used during this process do not account for CNVs during single nucleotide variant (SNV) calling(9). However, CNVs can have a major impact on the determination of heterozygous SNVs in mapped read processing workflows, because these workflows rely on examination of the ratios of mapped reads containing each variant. For example, if you have a standard copy two gene where one of the genes has an alternate allele, the ratio of reads mapped to that position containing the alternate allele will be around 50%; however, if you have a copy four gene where one of the genes has an alternate allele, the ratio of reads containing the alternate allele will now be around 25%. This alternate allele could then be missed by the variant calling workflow due to being below the read ratio threshold, which is assuming a

50% ratio. Another characteristic of the *KIR* and *C4* regions that confounds standard alignment processing workflows are the combination of extensive nucleotide polymorphisms with high sequence similarity between genes. These characteristics increase the potential for read misalignments, which is when a read originating from a specific location in the genome is mapped to a different location on the reference(10). Read misalignments can occur quite readily between regions with highly similar sequences, especially if there are nucleotide polymorphisms. Misaligned reads can lead to spurious variant calls which can be quite difficult to detect and correct.

In this dissertation, I outline the development of a bioinformatics pipeline for the interpretation of *KIR* sequence from NGS data in chapter 2. This work entailed the development of dynamic alignment strategies and custom mapped read processing workflows to overcome the challenges imposed by the region. Then, in chapter 3, I expand the *KIR* bioinformatics pipeline to extend functionality to whole genome sequence (WGS) data, in addition to evaluating performance on various synthetic sequence datasets to characterize the strengths and flaws of the workflow. Finally, in chapter 4, I outline the development of a bioinformatics pipeline for the interpretation of *C4* sequence from NGS data, which was built from the same custom mapped read processing workflows utilized for *KIR* sequence.

## References


1. Pende D, Falco M, Vitale M, Cantoni C, Vitale C, Munari E, et al. Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. Frontiers in Immunology [Internet]. 2019 [cited 2022 Apr 26];10. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2019.01179

2. Moretta L, Moretta A. Killer immunoglobulin-like receptors. Current Opinion in Immunology. 2004 Oct 1;16(5):626–33.

3. Downing J, D'Orsogna L. High-resolution human KIR genotyping. Immunogenetics [Internet]. 2022 Jan 20 [cited 2022 May 1]; Available from: https://doi.org/10.1007/s00251-021-01247-0

4. Wang H, Liu M. Complement C4, Infections, and Autoimmune Diseases. Frontiers in Immunology [Internet]. 2021 [cited 2022 Apr 28];12. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2021.694928

5. Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. In: Kulski JK, editor. Next Generation Sequencing - Advances, Applications and Challenges [Internet]. InTech; 2016 [cited 2022 May 1]. Available from: http://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-an-overview-of-the-history-tools-and-omic-applications

6. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. Human Molecular Genetics. 2010 Mar 1;19(5):737–51.

7. Chung EK, Yang Y, Rupert KL, Jones KN, Rennebohm RM, Blanchong CA, et al. Determining the One, Two, Three, or Four Long and Short Loci of Human Complement C4 in a Major Histocompatibility Complex Haplotype Encoding C4A or C4B Proteins. The American Journal of Human Genetics. 2002 Oct 1;71(4):810–22.

8. Marin WM, Dandekar R, Augusto DG, Yusufali T, Heyn B, Hofmann J, et al. High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING. PLOS Computational Biology. 2021 Aug 2;17(8):e1008904.

9. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008.

10. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Frontiers in Genetics [Internet]. 2015 [cited 2022 May 1];6. Available from: https://www.frontiersin.org/article/10.3389/fgene.2015.00235

# Chapter 2: High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING

## Abstract

The killer-cell immunoglobulin-like receptor (KIR) complex on chromosome 19 encodes receptors that modulate the activity of natural killer cells, and variation in these genes has been linked to infectious and autoimmune disease, as well as having bearing on pregnancy and transplant outcomes. The medical relevance and high variability of KIR genes makes short-read sequencing an attractive technology for interrogating the region, providing a high-throughput, high-fidelity sequencing method that is cost-effective. However, because this gene complex is characterized by extensive nucleotide polymorphism, structural variation including gene fusions and deletions, and a high level of homology between genes, its interrogation at high resolution has been thwarted by bioinformatic challenges, with most studies limited to examining presence or absence of specific genes. Here, we present the PING (Pushing Immunogenetics to the Next Generation) pipeline, which incorporates empirical data, novel alignment strategies and a custom alignment processing workflow to enable high-throughput KIR sequence analysis from short-read data. PING provides KIR gene copy number classification functionality for all KIR genes through use of a comprehensive alignment reference. The gene copy number determined per individual enables an innovative genotype determination workflow using genotype-matched references. Together, these methods address the challenges imposed by the structural complexity and overall homology of the KIR complex. To determine copy number and genotype determination accuracy, we applied PING to European and African validation cohorts and a synthetic dataset. PING demonstrated exceptional copy number determination performance across all datasets and robust genotype determination performance. Finally, an investigation into discordant genotypes for the synthetic dataset provides insight into misaligned reads, advancing our understanding in interpretation of short-read

sequencing data in complex genomic regions. PING promises to support a new era of studies of KIR polymorphism, delivering high-resolution KIR genotypes that are highly accurate, enabling high-quality, high-throughput KIR genotyping for disease and population studies.

## Introduction

The *killer cell immunoglobulin-like receptor* (*KIR*) complex, located in human chromosomal region 19q13.42, encodes receptors expressed on the surface of natural killer (NK) cells (1) and a subtype of T-cells (2). KIRs interact with their cognate HLA class I ligands to educate NK cells and modulate their cytotoxicity (3–5). *KIR* genes exhibit presence and absence polymorphism and gene content variation that has been implicated in numerous immune-mediated and infectious diseases (6–11). In addition, careful consideration of *KIR* gene content haplotypes for allogeneic transplantation has been shown to improve outcomes for acute myelogenous leukemia patients (12–17). Whereas evidence for the relevance of *KIR* variation in health and disease is mounting, analysis of the *KIR* family at allelic resolution has been thwarted by the complexity of the region.

The *KIR* complex evolved rapidly through recombination and gene duplication events, and in humans this has resulted in a gene-content variable cluster of 13 genes and 2 pseudogenes (18–20). Variation in *KIR* genes is characterized by extensive nucleotide polymorphisms, with 1110 alleles described to date (21). The *KIR* complex is also characterized by large-scale structural variation, including gene fusions, duplications and deletions (22,23). *KIR* haplotypes exhibit gene content variation at extraordinary levels, generating hundreds of observed haplotype structures (20,24–26).

The high variability of *KIR* makes short-read sequencing an attractive technology for interrogating the region, providing a high-throughput, high-fidelity and cost-effective sequencing method (27). Whereas the *KIR* region is relatively small, between 70-270Kbp (28), the overall sequence similarity among genes, structural variability of the region, and sequence polymorphism present major obstacles to

bioinformatics workflows. The high potential for read misalignments significantly confounds interpretation of the region in modern large-scale sequencing studies.

Previously, we introduced a laboratory method for targeted sequencing of the *KIR* gene complex (27), but the associated prototype bioinformatic pipeline for sequence interpretation presented significant workload barriers for high-throughput studies. For example, the copy number determination workflow was unable to differentiate *KIR2DL2* from *KIR2DL3*, which are sets of highly similar allelic groups of the *KIR2DL23* gene (29), and the resolution of *KIR2DS1* and *KIR2DL1* was less precise than desired due to read misalignments caused by the similarity of these two genes. Additionally, a high frequency of unresolved genotypes (not matching any described allele sequence) necessitated subsequent interpretation by a user with domain expertise. In spite of these challenges, the prototype pipeline has provided insight into *KIR* genotyping methods development (30,31), the role of *KIR* sequence variants in immune dysfunction (17,32), and *KIR* evolutionary analyses (33). To the best of our knowledge there are two other existing tools for interrogating *KIR* short-read sequencing data, KIR*IMP (34) and KPI (35,36). KIR*IMP imputes *KIR* copy number from carefully selected SNPs while KPI interprets *KIR* gene content and predicts haplotype-pairs using *in silico* probes. Neither of these methods support allele level genotyping or direct copy number assessment.

Here, we present a comprehensive *KIR* sequence interpretation workflow, termed PING (Pushing Immunogenetics to the Next Generation), which builds on our early work by incorporating empirical optimizations derived from sequencing thousands of samples, in addition to novel alignment strategies to address issues with read misalignments, to provide a comprehensive *KIR* sequence analysis tool, offering allele-level genotypes, copy number, and novel sequence analysis. This work improves on the pipeline described in Norman et al. (27) in the following ways: the copy number determination workflow was adjusted from a single-sequence per gene alignment to a comprehensive multiple-sequence per gene alignment; virtual probes used for gene content determination were refined and expanded; the

genotype determination workflow was adjusted from static single-gene filtration alignments to dynamic holistic alignments, which incorporate the so-established gene content and preliminary genotype determinations; a custom alignment processing workflow was developed to handle multiple-sequence per gene alignments; and finally, a *KIR* sequence imputation workflow was developed to enable alignment to any described *KIR* allele sequence. These innovations enable for the first time highly-automated, high-throughput *KIR* sequence analysis from short-read sequencing data, and importantly, largely obviate the need for user expertise in the *KIR* system.

## Materials and methods

The major innovations in PING, detailed below, include the use of multiple-sequence per gene alignment references that incorporate the allelic diversity of *KIR*, and genotype-matched alignment references (**Figure 2.1**). The use of a diverse reference set in the copy number module substantially improves copy number determination for *KIR2DL2*, *KIR2DL3*, *KIR2DS1* and *KIR2DL1* compared to our prototype approach. The improved performance enables an innovative alignment workflow that dynamically constructs genotype-matched alignment references based on the so-established gene content and a preliminary genotype determination. The genotypes determined by this novel genotype-aware alignment workflow are highly accurate, with few unresolved calls.

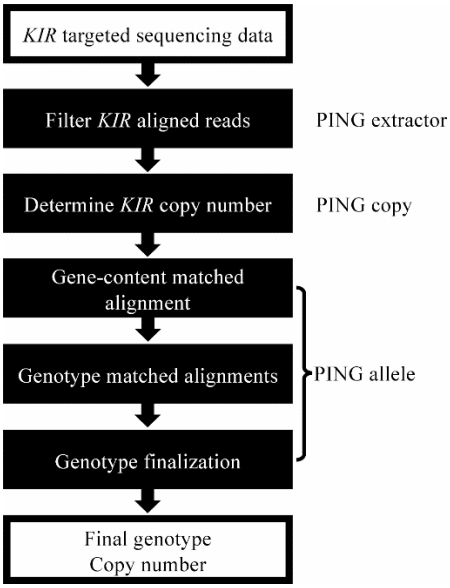


**Figure 2.1. Overview of the PING pipeline.**
The PING pipeline processes *KIR* targeted sequencing data to determine *KIR* gene copy number and allele genotypes through a series of modules. First, *KIR* aligned reads are filtered through an alignment to a set of *KIR* haplotypes in PING extractor. Second, copy number of *KIR* genes are determined through an exhaustive alignment to a diverse set of *KIR* sequences in PING copy. Finally, PING allele performs a series of alignments to determine the most congruent *KIR* genotype, which informs a final round of alignment and genotype determination. Additionally, PING reports any identified novel SNPs and new alleles (SNP combinations not found in any described *KIR* allele sequence).

## Preprocessing the database

### Imputation of uncharacterized regions and extension of untranslated regions to generate comprehensive alignment reference sequence

*KIR* allele sequences used throughout this workflow are provided by the Immuno Polymorphism Database - KIR (IPD-KIR), release 2.7.1 (37). However, many *KIR* allele sequences in IPD-KIR have only been characterized for exons, indeed, 65% of named alleles have less than 20% of their full-length sequence characterized (**Figure 2.2A**). Additionally, IPD-KIR allele sequences only include ~250bp of 5' untranslated region (UTR) sequence and ~500bp of 3' UTR sequence, reducing alignment depths across the first exon and potential regulatory regions. Thus, to maximize the utility of reference *KIR* sequences, we designed and implemented a protocol to impute the sequence of the intronic regions, followed by a protocol to extend UTRs to 1000bp each.

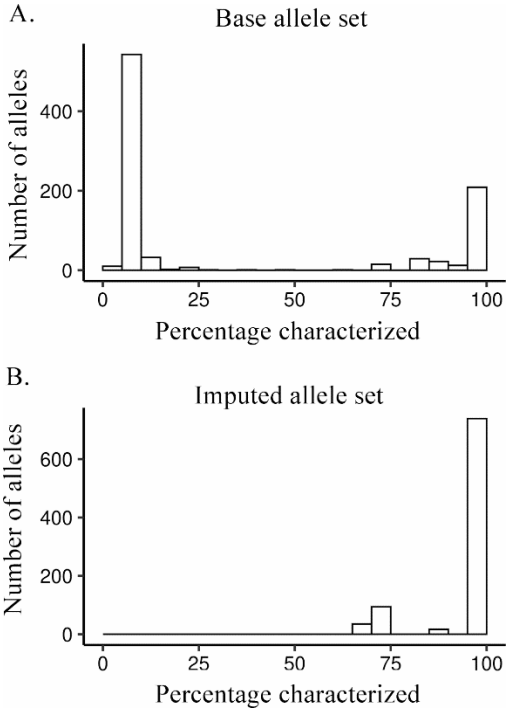


**Figure 2.2.** *KIR***sequence characterization before and after imputation.**
**(A)** Histogram of IPD-KIR allele sequence lengths, shown as percentage of longest sequence for each gene and major allele group. **(B)** Histogram of IPD-KIR allele sequence lengths after imputation, shown as percentage of longest sequence for each gene and major allelic group.

As reference sequences, we used all *KIR* alleles described in the IPD - KIR (21), release 2.7.1. A subset of these sequences is not completely characterized through all exons and introns. We therefore used gene-specific alignments of known sequences, provided by IPD-KIR as multiple sequence format (MSF) files, and completed each allele sequence to comprise the invariant nucleotides together with each variable position represented by an 'N'. Using this imputation method, we generated a new set of reference alleles in which ~90% of the 905 alleles were >98% complete (**Figure 2.2B**).

To extend UTR sequence, donor sequences, sourced from the full *KIR* haplotype sequences used in PING extractor, were appended to the ends of each reference sequence to generate 1000bp long UTRs. A single 3'UTR and 5'UTR donor sequence was used for each gene and major allelic group.

**Designing a minimized reference allele set**

While performing a comprehensive alignment to the full *KIR* allele set reduces misalignments caused by reference sequence bias, it demands substantial resource utilization, as well as a large alignment and processing time cost which can prove untenable for processing large datasets. For example, copy determination processing for 10 paired-end sequences using 36 threads took 4.95 hours with a maximum Binary Alignment Map (BAM) (38) file size of 338.2MB.

To address this issue, we constructed a minimized set of reference alleles to improve resource utilization and processing time while still reducing misalignments caused by reference sequence bias. The minimized reference set consists of five alleles for each *KIR* gene and major allelic group (**Table in S2.1 table**). The use of five alleles per gene was empirically determined to be sufficient for reducing reference sequence bias, while still considerably reducing the computational burden of multiple-sequence per gene alignments. Designing this reference set was guided by selecting alleles which had fully-characterized or nearly fully-characterized sequence, the secondary criteria was maximizing SNP diversity between the reference alleles of each gene, and third was selecting reference alleles to sequester reads susceptible to off-gene mapping. For example, reference sequences to represent *KIR2DS1*002* as well as *KIR2DL1*004* were selected to sequester reads that perfectly align to both. Notable characteristics of the reference set are the separation of *KIR2DL2* from *KIR2DL3*, the separation of *KIR3DL1* from *KIR3DS1*, and the merging of *KIR2DL5A* and *KIR2DL5B*.

## PING workflow methods

**Copy number determination – PING copy**

The high sequence similarity between *KIR* genes coupled with extensive structural variation and nucleotide diversity makes copy number determination a non-trivial task. Our copy determination method is largely identical to that described in Norman et al. (27), in which copy number is determined

by comparing the number of reads that align uniquely to each *KIR* gene across a batch of samples using *KIR3DL3* as a normalizer. The improvement made by our method is the use of a comprehensive *KIR* reference composed of 905 distinct sequences from the imputed and extended allele set, instead of a single-sequence per gene reference. The use of a comprehensive reference provides a more accurate comparison of the number of reads that align uniquely to any *KIR* gene.

**_KIR_virtual probes – PING allele**

To determine the presence of target alleles or allelic groups that are prone to misidentification due to read misalignments, we have developed a set of virtual, or text-based, probes. The probe set includes those described in Norman et al. (27), as well as additional, custom probes (**Table in S2.2 Table**). Probes are designed to match sequence that is unique to the target allele or allelic group, and sequence uniqueness is determined by a grep search over the imputed and extended IPD-KIR sequence set. Application of the probe set is performed using grep over the sequencing data, counting the number of unique reads that contain sequence perfectly matching the probe. A probe hit is determined using a threshold of 10 matching reads.

**Genotype matched alignment workflow – PING allele**

The overall alignment strategy of PING is to reduce reference sequence bias through using multiple-sequence per gene references, and the use of references that reflect the gene content makeup or genotype makeup of a sample (**Figure 2.3**). Additionally, PING utilizes multiple rounds of alignment and genotype determination with varied processing parameters to reduce bias introduced by assumptions made during the processing workflow. The intermediate and final alignment and genotyping rounds are referred to as ‘initial’ and ‘final’, respectively, when differentiating processing parameters.

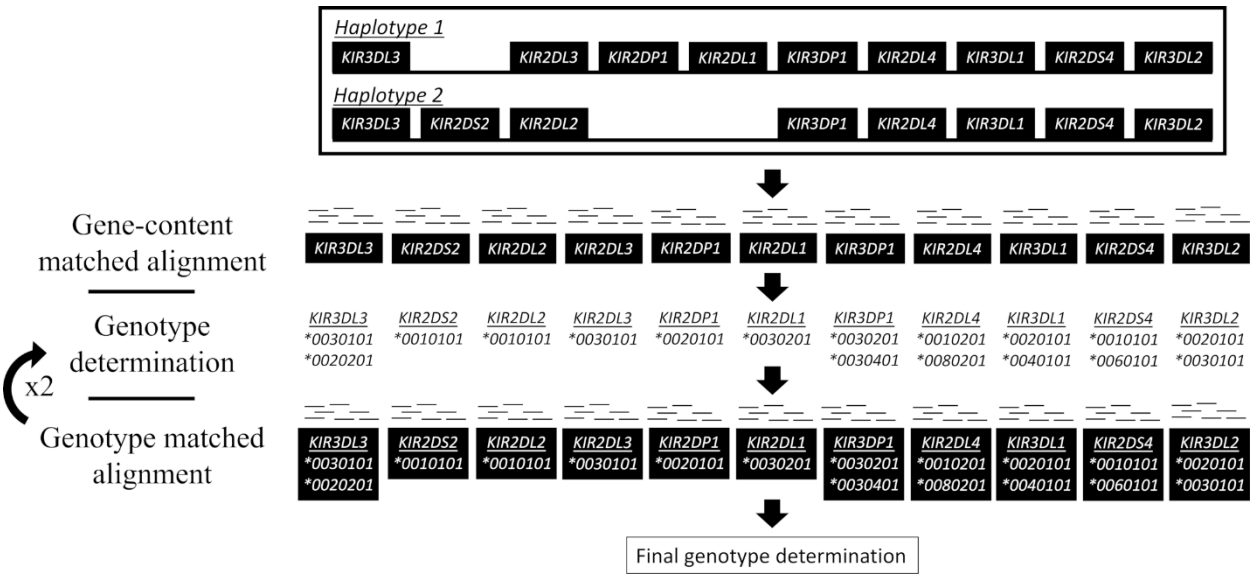


**Figure 2.3. Overview of the genotype aware alignment workflow.**
Sequence gene content, determined by PING copy, informs the selection of reference sequence from a predefined set of diverse allele sequences. An exhaustive alignment is performed to the selected allele set, from which an initial genotype determination is made. The determined genotype informs selection of reference alleles for a genotype aware alignment, followed by another round of genotype determination. The genotype aware alignment and subsequent genotype determination is repeated, and the most congruent genotypings across all alignment rounds inform reference selection for a final round of alignment. A non-exhaustive alignment is performed to the selected allele set, from which all aligned reads are processed and used for the final genotype determination.

The first step is an alignment to a multiple-sequence per gene, gene content matched reference. This gene-content aware alignment workflow constructs individualized alignment references based on the presence of certain *KIR* genes: *KIR3DP1*, *KIR2DS2*, *KIR2DL23*, *KIR2DL5A*, *KIR2DL5B*, *KIR2DS3*, *KIR2DS5*, *KIR2DP1*, *KIR2DL1*, *KIR2DL5*, *KIR3DL1S1*, *KIR2DS4*, *KIR3DL2*, *KIR2DS1* and *KIR3DL3* (assumed always present (39)). Reference sequences are selected from the diverse, minimized reference sequence set described above. We have included an option to align to the full comprehensive allele set, but this is not default behavior as these alignments are time and resource intensive.

An exhaustive alignment, an alignment in which all qualified read mappings are recorded, is performed and aligned reads are processed and formatted according to the alignment processing workflow, detailed in **S2.1 text**, selecting for reads that uniquely map to a gene or major allelic group. The

formatted uniquely-mapped read set is processed according to the genotype determination workflow, detailed below, to obtain an initial full-resolution genotype.

Second is a series of two alignments to genotype-matched references with varied processing parameters to identify the most congruent *KIR* genotype. Genotype congruence is determined by the least number of SNP mismatches between the determined allele typing(s) of a gene, and the aligned SNPs. For each genotype-matched alignment in this series, the determined allele typing(s), including any ambiguity, are used as reference sequence for the following alignment. For each alignment, genotypes are first determined at seven-digit (non-coding mutation level), then five-digit resolution (synonymous mutation level). This approach reduces the impact of uncharacterized regions of IPD-KIR allele sequences on genotype determination, as most sequences are fully characterized across exons. Genotype determination can be biased towards or against IPD-KIR alleles with uncharacterized regions depending on whether uncharacterized SNPs count as mismatches or not. To reduce time spent on genotype determination, any unambiguous typing that is perfectly matched to the aligned SNPs is locked in across all subsequent intermediate rounds of genotype determination.

The reference for the final alignment is built from the locked genotypings and the closest matched genotypings for genes without a locked genotyping. A non-exhaustive alignment is performed to the built reference, from which all aligned reads are processed and formatted according to the alignment processing workflow. The formatted read alignments are passed to the genotype determination workflow to obtain a final exonic (five-digit) resolution genotype.

**Additional methods utilized by the genotype matched alignment workflow – PING allele**

Genotype matched alignments and subsequent genotype determinations can get stuck on a mistyped allele due to persistent reference sequence bias. In other words, a false SNP call originating from misaligned reads can perpetuate itself in the genotype matched alignments due to the same allele

determination being made and the same alignment reference being used. To address this issue, we have included a method in the genotype matched alignments that will add the five allele sequences from the diverse, minimized reference set to the genotype-matched reference for any gene with an allele typing that does not perfectly match the aligned SNPs. The rationale behind this method is that mismatched allele typings are likely due to misaligned reads, and the use of the mismatched allele sequence as a reference will cause the read misalignments to be repeated in subsequent alignment and genotyping rounds. The addition of a diverse set of alignment sequences gives an avenue to break from this cycle by increasing the likelihood that a different allele typing will be made.

In building genotype-matched references PING allows any allele to be used as reference sequence, however, some allele sequences are only partially characterized even after imputation. The use of allele sequences containing uncharacterized sequence as alignment references can introduce reference sequence bias and drive read misalignments even if the reference alleles perfectly match the true genotype of the sample. To address this issue, we have included a method to add fully-characterized sequence to the alignment reference for any gene represented by only partially-characterized sequence(s). Fully-characterized alleles are pulled from the diverse, minimized reference set.

In the genotype-aware alignment workflow we found issues with false negative identifications of *KIR2DL1*004/*007/*010* due to reads cross-mapping to other gene sequences. This issue was rectified using virtual sequence probes specific to each of these *KIR2DL1* allele groups to identify **004/*007/*010* allele presence. If *KIR2DL1*010* is present, then the *KIR2DL1*010* allele sequence is added to the alignment reference. If *KIR2DL1*004* is present, then the *KIR2DL1*0040101* allele sequence is added to the alignment reference. If *KIR2DL1*007* is present, then the *KIR2DL1*007* allele sequence is added to the alignment reference. If multiple of these allele groups are present, then the *KIR2DL1*0040101* allele sequence is added to the alignment reference.

We implemented additional probes to identify alleles and structural variants prone to misidentification across *KIR2DL1*, *KIR2DL2*, *KIR2DL4*, *KIR2DS1*, *KIR3DP1* and *KIR3DS1*. For example, we implemented a probe to identify the *KIR2DL4* poly-A stretch at the end of exon 7, as well as a probe to identify *KIR3DP1* exon 2 deletion variants. The full list of probes used for reference refinement can be found in **S2.2 Table**.

**Genotype determination workflow – PING allele**

Indexed reads, detailed in **S2.1 text**, are processed to generate a depth table spanning -1000bp 5'UTR to 1000bp 3'UTR for each *KIR* gene and major allelic group. Depths are marked independently for A, T, C, G, deletions and insertions. Depth tables are processed to generate SNP tables for positions passing a minimum depth threshold (default 8 for initial genotyping and 20 for final genotyping). To identify heterozygous positions the depth of each aligned variant is divided by the highest depth variant for that position, and up to three variants (A, T, C, G, deletions and insertions) passing the ratio threshold (default 0.25 for initial and final genotyping) are recorded.

Genotypes for each gene and major allelic group are determined from the aligned SNPs using a mismatch scoring approach. First, aligned homozygous SNPs are compared to each IPD-KIR allele, with SNP mismatches counting as a score of 1 and matches as 0. The lowest scoring alleles and alleles within a set scoring buffer of the lowest score (default of 4 for the initial genotyping workflow and 1 for the final genotyping workflow), are carried over into heterozygous position scoring.

For aligned heterozygous position scoring, all possible allele combinations are enumerated according to the determined copy of the gene under consideration, up to copy 3. For each aligned position, the variant(s) for each allele combination are compared to the aligned variants, with full matches counted as a score of 0 and mismatches scored according to the number of mismatched variants. For each allele combination, the homozygous score of each component allele is added to the heterozygous score, and

the lowest scoring combinations are returned as the determined genotype. For the final genotyping workflow, only perfectly scoring combinations are accepted, with any mismatches resulting in an unresolved genotype.

The same workflow is applied to both initial and final genotyping with some important distinctions. In the initial genotyping workflow, the imputed and extended IPD-KIR allele sequences are used for SNP comparisons, uncharacterized variants within the comparison sequences are marked as full mismatches, and all aligned allele-differentiating SNP positions passing the depth threshold are compared and used for scoring. In the final genotyping workflow, the unimputed IPD-KIR allele sequences are used for SNP comparisons, uncharacterized variants within the comparison sequences are marked as matches, and only aligned exonic SNP positions passing the depth threshold are compared and used for scoring.

The final exonic resolution genotypes are processed to add null alleles to the genotype string for genes with copy 0 or 1 and combine component allele typings for the major allelic groups *KIR2DL2* and *KIR2DL3*, *KIR3DL1* and *KIR3DS1*, and *KIR2DS3* and *KIR2DS5*.

## Workflow validation

### *KIR*synthetic sequence dataset

A *KIR* synthetic dataset consisting of 50 sequences was generated using the ART next-generation sequencing read simulator (40). ART parameters were set to simulate 150-bp paired-end reads at 50x coverage, with a median DNA fragment length of 200 using quality score profiles from the HiSeq 2500 system. Eleven of the *KIR* haplotypes described in Jiang et al. (41) were used to simulate structural variation of the *KIR* region. Two of the eleven haplotypes were randomly selected with replacement to establish the copy number for each sample. Allele sequences were selected randomly without replacement from the imputed and extended set according to the copy number of each gene. Any uncharacterized regions in the selected allele sequences were replaced with sequence from a random

fully-characterized sequence from the same gene. Reads were named according to the source allele, enabling tracing of misaligned reads to their source allele and gene. The full synthetic dataset is available at: https://github.com/wesleymarin/KIR_synthetic_data.

Discordant genotype results for the synthetic dataset were investigated by identifying the source gene for each read aligned to the incorrectly genotyped gene. The results were summarized to show the total number of reads from each source gene to each aligned gene, **Table A in S2.3 Table**, and a read sharing diagram was generated using the circlize (42) package in R (43).

**Characterization of *KIR*reference cohorts for PING development**

A significant barrier to the development of bioinformatic methods for high-resolution *KIR* sequence interpretation is the lack of a well-characterized reference cohort. Without such a resource it is extremely difficult to recognize and resolve issues with read misalignments, which can result in SNP calls that appear reasonable in many cases. To resolve this issue, we have characterized a *KIR* reference cohort of 379 healthy individuals of European ancestry that had been previously sequenced using our *KIR* target capture method (44), with the results meticulously curated by manual alignment and inspection of all sequences to provide a ground truth dataset to aid pipeline development (**Table 2.4A in S2.4 Table**). Furthermore, the European samples were independently sequenced and genotyped for *KIR* by our collaborators at the DKMS registry for volunteer bone marrow donors (31). Any discordant typing or gene content results were resolved through direct examination of sequence alignments, and where necessary, confirmatory sequencing.

In order to validate our method on a second, divergent population, we also examined a previously characterized cohort of African Khoesan individuals (45), for which *KIR* alignments and genotypes were manually inspected (**Table 2.4B in S2.4 Table**).

**Copy number and genotype concordance calculations**

For the European cohort, any genotype containing an unresolved *KIR3DL3* genotyping, a genotype for which the aligned SNPs do not perfectly match currently described alleles, in the truth dataset were excluded from copy number and genotype concordance comparisons. There were 16 full genotypes excluded by this criterion. Additionally, any individual gene with an unresolved genotype in the truth dataset were excluded from copy number and genotype concordance comparisons. For the synthetic dataset there were no simulated novel alleles, so the full dataset was used for copy number and genotype concordance comparisons. For the Khoesan dataset any individual gene with an unresolved genotype in the truth dataset were excluded from genotype concordance comparisons but were included for copy number comparisons.

Copy concordance was calculated by directly comparing the determined copy values to the validation copy values (**Tables in S2.5 Table and S2.6 Table**). Genotype concordance was calculated on a per-gene basis by comparing each component allele of the determined typing to the truth genotype.

## Code availability

The PING pipeline is available at https://github.com/wesleymarin/PING (46) with the following open source license: https://github.com/wesleymarin/PING/blob/master/LICENSE. Scripts and datasets used for data analysis are available at https://github.com/wesleymarin/ping_paper_scripts. PING was developed in the R programming language and tested on a Linux system. Additional requirements are: Samtools v1.7 or higher (38), Bcftools v1.7 or higher, and Bowtie2 v2.3.4.1 or higher (47). The synthetic KIR sequence dataset is available at https://github.com/wesleymarin/KIR_synthetic_data.

## Results

### Extensive sequence identity among *KIR* genes is a major barrier for interpreting short-read sequencing data

Read misalignments due to sequence identity across the *KIR* region are a persistent challenge in *KIR* bioinformatics and often lead to spurious genotyping results. To quantify the extent of sequence identity and inform our investigation of SNPs suspected to be originating from misaligned reads, we performed a shared *k*-mer analysis using all 905 described *KIR* allele sequences in the Immuno Polymorphism Database (IPD) - KIR (21), release 2.7.1. Here, we transformed allele sequences into all distinct subsequences of sizes 50, 150, and 250 to compare sequence identity between genes. Shared k-mer proportions were calculated by dividing the number of shared k-mers by the total number of k-mers of that gene. K-mer sharing diagrams were generated using the circlize (42) package in R (43).

Our analysis showed that many genes share significant sequence identity at sequencing lengths commonly used in next generation sequencing (NGS) technology (**Figure 2.4A**). For example, *KIR2DL5A* shares 12,591 of its 15,359 distinct 150-mers (82%) with *KIR2DL5B* (**Figure 2.4B**, **Tables in S2.8 Table**), making it extremely difficult to distinguish short reads originating from these genes. Likewise, over 90% of the distinct 50-mers, and over 50% of distinct 250-mers of *KIR2DS1* are shared with other genes, the vast majority with *KIR2DL1*. This analysis allowed us to identify specific "hotspots" for read misalignments, informing post-alignment modifications (detailed previously) to minimize their impact.

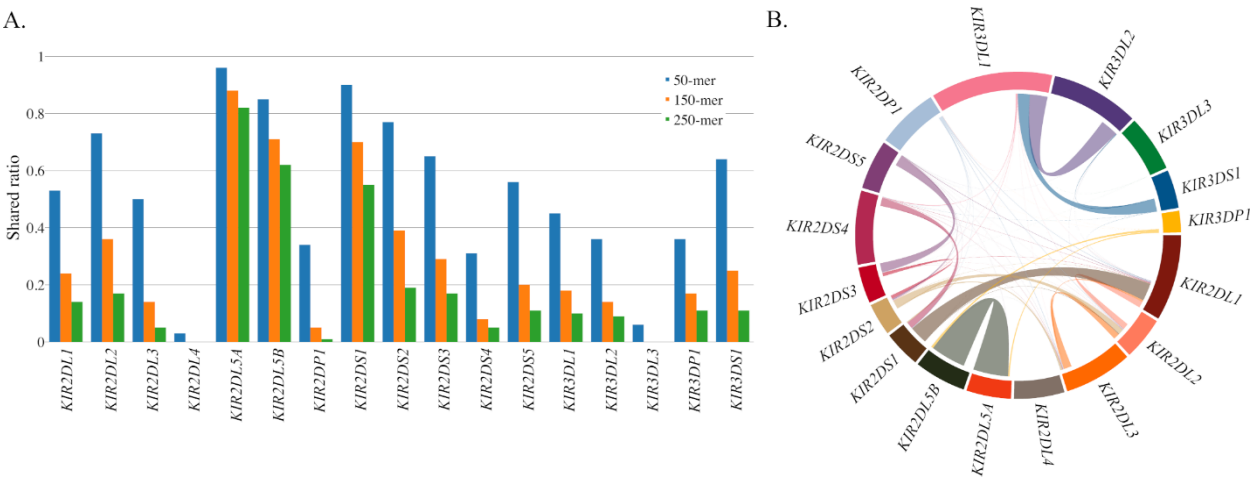


**Figure 2.4. K-mer analysis of *KIR* gene sequence similarity.**
**(A)** Ratio of distinct k-mers of size 50, 150 and 250 that are shared between the indicated *KIR* gene and others. The inverse of these bars (not shown) would indicate the proportion of k-mers that are distinct to that gene and not found in the alleles of other genes. **(B)** 150-mer connections between *KIR* genes, the size of the connecting line roughly indicates the total number of shared 150-mers.

**Development of a comprehensive *KIR* alignment reference enables accurate copy number determination of *KIR2DL1*, *KIR2DS1*, *KIR2DL2 and KIR2DL3***

We compared single-sequence per gene vs. multiple-sequence per gene reference alignments using the synthetic dataset. The reads from this dataset are labeled according to their source gene, providing a straightforward approach to quantify off-target alignments. This comparison showed substantial reductions in the frequency of read misalignments (reads mapping to an off-target gene) across *KIR2DL1*, *KIR2DL23*, *KIR2DL5A/B*, *KIR2DS1*, *KIR2DS35*, and *KIR3DL1S1*, and small reductions for *KIR3DL2*, *KIR3DL3*, and *KIR3DP1* for the multiple-sequence per gene reference (**Figure 2.5A**, **Tables in S2.9 Table**). Applying the comprehensive reference allele set to gene content and copy number determination, we achieved significant improvement over a single-sequence per gene reference for *KIR2DL1*, *KIR2DS1* and the allelic groups *KIR2DL2* and *KIR2DL3* (**S2.1 Figure**, **S2.2 Figure** and **S2.3 Figure**). The copy number for *KIR2DS1*, per example, which is highly prone to read misalignments due to similarity to *KIR2DL1* and *KIR2DS4* (**Figure 2.4B**), was clearly determined (**Figure 2.5B**).

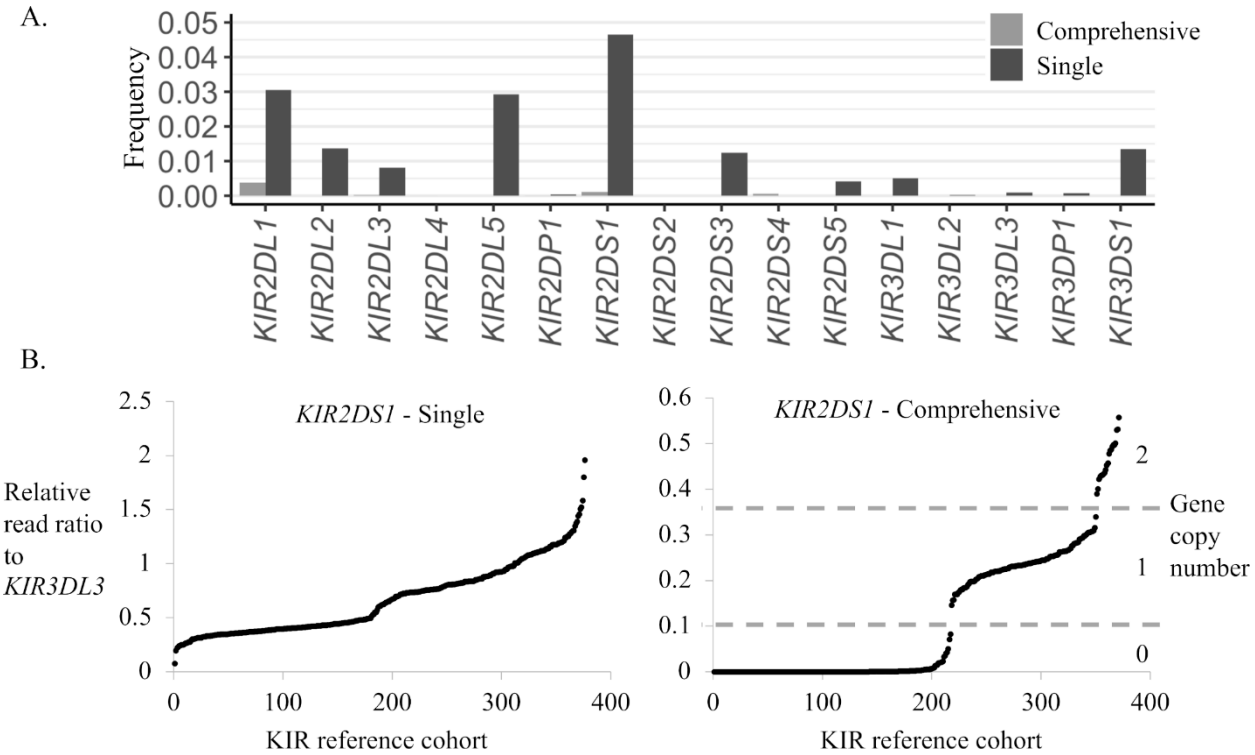


**Figure 2.5. Use of comprehensive reference improves copy determinations.**
**(A)** Frequencies of off-target read mappings using a comprehensive reference vs. a single-sequence per gene reference for the synthetic dataset. **(B)** Single-sequence reference vs. comprehensive reference copy number plot of *KIR2DS1* for the European cohort. The copy plot of the single-sequence reference alignment shows no differentiation between copy groupings while the comprehensive reference alignment shows a clear distinction between the copy 0, 1 and 2 groups.

**PING delivers accurate copy number and high-resolution allele calls**

The overall performance of PING was assessed using our European *KIR* reference cohort, a synthetic *KIR* dataset, and a Khoesan *KIR* reference cohort. Results for PING copy number determination are summarized in **Table 2.1**, showing at least 97% concordance for the European cohort for all compared genes, with most genes exhibiting more than 99% concordance. Performance for the synthetic dataset showed 100% copy concordance for all compared genes except for *KIR2DL1*, at 98%, and *KIR2DS3*, at 92%. Finally, performance for the Khoesan cohort showed at least 95% concordance for all compared genes except for *KIR2DL2*, at 61%, *KIR2DL5A/B*, at 88%, *KIR2DS5*, at 89%, and *KIR2DL1,* at 94%. Across all

datasets *KIR3DL3* was not compared due to its use as a reference gene, and for the European and Khoesan cohorts the pseudogene *KIR3DP1* was not compared due to an absence of validation data.

**Table 2.1. Copy number determination performance.**
Concordance table comparing copy numbers determined by PING for the European reference cohort, a synthetic *KIR* dataset, and a Khoesan reference cohort.

| Gene | European | N | Synthetic | N | Khoesan | N |
|---|---|---|---|---|---|---|
| *KIR3DL3* | - | - | - | - | - | - |
| *KIR2DS2* | 0.988 | 343 | 1.00 | 50 | 0.97 | 100 |
| *KIR2DL2* | 0.994 | 331 | 1.00 | 50 | 0.61 | 100 |
| *KIR2DL3* | 0.994 | 331 | 1.00 | 50 | 1.00 | 100 |
| *KIR2DL5A/B* | 0.997 | 343 | 1.00 | 50 | 0.88 | 100 |
| *KIR2DS3* | 0.988 | 343 | 0.92 | 50 | 0.97 | 100 |
| *KIR2DS5* | 0.985 | 342 | 1.00 | 50 | 0.89 | 100 |
| *KIR2DP1* | 0.982 | 338 | 1.00 | 50 | 0.95 | 100 |
| *KIR2DL1* | 0.970 | 334 | 0.98 | 50 | 0.94 | 100 |
| *KIR3DP1* | - | - | 1.00 | 50 | - | - |
| *KIR2DL4* | 0.994 | 341 | 1.00 | 50 | 1.00 | 100 |
| *KIR3DL1* | 0.997 | 340 | 1.00 | 50 | 1.00 | 100 |
| *KIR3DS1* | 0.997 | 339 | 1.00 | 50 | 0.99 | 100 |
| *KIR2DS1* | 0.988 | 342 | 1.00 | 50 | 1.00 | 100 |
| *KIR2DS4* | 0.991 | 343 | 1.00 | 50 | 0.99 | 100 |
| *KIR3DL2* | 0.988 | 326 | 1.00 | 50 | 1.00 | 100 |

Performance of genotype determination was assessed at three-digit resolution (protein level) for the European and Khoesan cohorts, and at five-digit resolution (synonymous mutation level) for the synthetic dataset (**Table 2.2**). The results were categorized as genotype matches, mismatches or unresolved genotypes, which were cases where PING could not make a genotype determination.

PING genotype determination for the European cohort showed low percentages of unresolved genotypes with few mismatches for all compared genes except for *KIR2DP1*, with 10.3% unresolved. Notable results for the European cohort were the low frequencies of unresolved genotypes across most genes, except *KIR2DP1*, and the extremely low frequencies of mismatched genotypes, below 1%, for 9 out of the 12 genes compared.

**Table 2.2. Genotype determination performance.**
Genotype determination performance table comparing the genotypes determined by PING to the validation genotypes for each dataset. Possible outcomes are 'Match', where the determined component allele matches the validation allele, 'Mismatch', where the determined component allele does not match the validation allele, or 'Unresolved', where PING was unable to determine a genotype, but the validation allele was not marked as unresolved. The coloring signifies concordance level, where green is 0-10% discordant, yellow is 10-15% discordant, and red is over 15% discordant.

| Gene | Dataset | Match | Mismatch | Unresolved | N |
|---|---|---|---|---|---|
| *KIR3DL3* | European | 0.959 | 0.009 | 0.032 | 686 |
| | Synthetic | 0.960 | 0.000 | 0.040 | 100 |
| | Khoesan | 0.887 | 0.062 | 0.050 | 80 |
| *KIR2DS2* | European | 0.975 | 0.012 | 0.013 | 686 |
| | Synthetic | 0.940 | 0.040 | 0.020 | 100 |
| | Khoesan | 0.835 | 0.005 | 0.160 | 188 |
| *KIR2DL23* | European | 0.965 | 0.003 | 0.032 | 656 |
| | Synthetic | 0.850 | 0.020 | 0.130 | 100 |
| | Khoesan | 0.810 | 0.042 | 0.149 | 168 |
| *KIR2DL5A/B* | European | 0.927 | 0.044 | 0.029 | 687 |
| | Synthetic | 0.856 | 0.106 | 0.038 | 104 |
| | Khoesan | 0.857 | 0.071 | 0.071 | 126 |
| *KIR2DS35* | European | 0.982 | 0.006 | 0.012 | 683 |
| | Synthetic | 0.923 | 0.019 | 0.058 | 104 |
| | Khoesan | 0.871 | 0.052 | 0.078 | 116 |
| *KIR2DP1* | European | 0.890 | 0.007 | 0.103 | 672 |
| | Synthetic | 0.971 | 0.000 | 0.029 | 103 |
| | Khoesan | 0.721 | 0.012 | 0.267 | 86 |

| Gene | Dataset | Match | Mismatch | Unresolved | N |
|---|---|---|---|---|---|
| *KIR2DL1* | European | 0.961 | 0.009 | 0.030 | 666 |
| | Synthetic | 0.883 | 0.019 | 0.097 | 103 |
| | Khoesan | 0.803 | 0.045 | 0.152 | 132 |
| *KIR3DP1* | European | - | - | - | - |
| | Synthetic | 0.773 | 0.055 | 0.173 | 110 |
| | Khoesan | - | - | - | - |
| *KIR2DL4* | European | 0.980 | 0.007 | 0.013 | 685 |
| | Synthetic | 1.000 | 0.000 | 0.000 | 110 |
| | Khoesan | 0.961 | 0.006 | 0.032 | 154 |
| *KIR3DL1S1* | European | 0.962 | 0.006 | 0.032 | 686 |
| | Synthetic | 0.955 | 0.000 | 0.045 | 110 |
| | Khoesan | 0.873 | 0.028 | 0.099 | 142 |
| *KIR2DS1* | European | 0.985 | 0.003 | 0.012 | 682 |
| | Synthetic | 0.900 | 0.000 | 0.100 | 100 |
| | Khoesan | 0.995 | 0.000 | 0.005 | 198 |
| *KIR2DS4* | European | 0.943 | 0.044 | 0.013 | 685 |
| | Synthetic | 0.940 | 0.020 | 0.040 | 100 |
| | Khoesan | 0.793 | 0.051 | 0.157 | 198 |
| *KIR3DL2* | European | 0.965 | 0.008 | 0.028 | 648 |
| | Synthetic | 0.990 | 0.010 | 0.000 | 100 |
| | Khoesan | 0.819 | 0.011 | 0.170 | 188 |

Determined genotypes for the synthetic dataset showed over 95% concordance for *KIR3DL3*, *KIR2DP1*, *KIR2DL4*, *KIR3DL1S1*, and *KIR3DL2*. However, the synthetic dataset showed high percentages of unresolved genotypes for *KIR2DL23*, *KIR3DP1* and *KIR2DS1,* each over 10% unresolved, and *KIR2DS35* and *KIR2DL1* showed 5.8% and 9.7% unresolved, respectively. *KIR2DL5A/B* and *KIR3DP1* showed the highest mismatched genotype percentages, at 10.6% and 5.5%, respectively, while *KIRDL3*, *KIR2DP1,* *KIR2DL4* and *KIR2DS1* each showed 0.0% mismatched genotypes.

Determined genotypes for the Khoesan cohort showed highly concordant genotypes for *KIR2DL4*, at 96.1%, and *KIR2DS1*, at 99.5%. Additionally, results for this dataset showed low mismatch frequencies for *KIR2DS2*, *KIR2DL23*, *KIR2DP1*, *KIR2DL1*, *KIR3DL1S1* and *KIR3DL2*, each below 5.0% mismatched. However, the Khoesan cohort showed moderate mismatch frequencies for *KIR3DL3*, at 6.2%, KIR2DL5A/B, at 7.1%, *KIR2DS35*, at 5.2%, and *KIR2DS4*, at 5.1%, and higher unresolved rates for *KIR2DS2*, *KIR2DL23*, *KIR2DP1*, *KIR2DL1*, *KIR2DS4* and *KIR3DL2*, each over 10.0% unresolved.

For the European and Khoesan cohorts the pseudogene *KIR3DP1* was not compared due to an absence of validation data.

Looking specifically at the concordance of resolved genotypes, the European cohort showed greater than 98.0% concordance across all compared genes except for *KIR2DL5A/B*, at 95.5%, and *KIR2DS4,* at 95.6% (**Table 2.3**). The synthetic dataset showed 100% concordance for *KIR3DL3*, *KIR2DP1*, *KIR2DL4*, *KIR3DL1S1* and *KIR2DS1*, over 95% concordance for *KIR2DS2, KIR2DL23, KIR2DS35, KIR2DL1*, *KIR2DS4* and *KIR3DL2*. The lowest performing genes in the synthetic dataset were *KIR2DL5A/B*, at 89%, and KIR3DP1, at 93%. The Khoesan cohort showed 100% concordance for *KIR2DS1*, over 95% concordance for *KIR2DS2*, *KIR2DL23*, *KIR2DP1*, *KIR2DL4*, *KIR3DL1S1* and *KIR3DL2*, and over 90% concordance for *KIR3DL3*, *KIR2DL5A/B*, *KIR2DS35*, *KIR2DL1* and *KIR2DS4*.

**Table 2.3. Resolved genotype concordance.**
PING genotype determination performance for the European reference cohort, a synthetic *KIR* dataset, and the Khoesan reference cohort for each considered *KIR* gene.

| Gene | European | N | Synthetic | N | Khoesan | N |
|---|---|---|---|---|---|---|
| *KIR3DL3* | 0.991 | 664 | 1.00 | 96 | 0.934 | 76 |
| *KIR2DS2* | 0.988 | 677 | 0.96 | 98 | 0.994 | 158 |
| *KIR2DL23* | 0.997 | 635 | 0.98 | 87 | 0.951 | 143 |
| *KIR2DL5A/B* | 0.955 | 667 | 0.89 | 100 | 0.923 | 117 |
| *KIR2DS35* | 0.994 | 675 | 0.98 | 98 | 0.944 | 107 |
| *KIR2DP1* | 0.992 | 603 | 1.00 | 100 | 0.984 | 63 |

| Gene | European | N | Synthetic | N | Khoesan | N |
|---|---|---|---|---|---|---|
| *KIR2DL1* | 0.991 | 646 | 0.98 | 93 | 0.946 | 112 |
| *KIR3DP1* | - | - | 0.93 | 91 | - | - |
| *KIR2DL4* | 0.993 | 676 | 1.00 | 110 | 0.993 | 149 |
| *KIR3DL1S1* | 0.994 | 664 | 1.00 | 105 | 0.969 | 128 |
| *KIR2DS1* | 0.997 | 674 | 1.00 | 90 | 1.000 | 197 |
| *KIR2DS4* | 0.956 | 676 | 0.98 | 96 | 0.940 | 167 |
| *KIR3DL2* | 0.992 | 630 | 0.99 | 100 | 0.987 | 156 |

Together, these results demonstrate that PING accurately provides *KIR* genotyping across distinct populations.

**Analysis of discordant determined copy number and genotype results**

The discordant copy results for *KIR2DS3* in the synthetic dataset were the result of poor differentiation between copy groups (**S2.3 Figure**). The highly discordant *KIR2DL2* copy number result for the Khoesan cohort was due to non-differentiable copy number groupings (**S2.2 Figure**). Since the *KIR2DL3* copy differentiation for this cohort was well defined, these results were used to set the *KIR2DL2* copy number prior to genotype determination using the formula KIR2DL2_copy = 2 – KIR2DL3_copy.

An investigation into the discordant genotypes for the synthetic dataset showed discordant genotype determination results for *KIR2DS35* were largely due to source reads from *KIR2DS3* aligning to *KIR2DS5* reference sequence, with a smaller number of reads from *KIR2DS5* aligning to *KIR2DS3* reference sequence (**Figure 2.6**, **Table A in S2.3 Table**). This differential read flow between the two allelic groups is reflected in the component allele typings, with six discordant *KIR2DS5* genotypings and two discordant *KIR2DS3* genotypings (**Table C in S2.7 Table**). Intragenic misalignments are a product of how the PING workflow is structured, as major allelic groups, such as *KIR2DS3* and *KIR2DS5,* are treated as independent genes during alignment and genotyping. Intragenic misalignments were also a large

contributor to *KIR3DL1S1* discordance, with reads supplied by *KIR3DL1* mapping to *KIR3DS1* reference sequence.

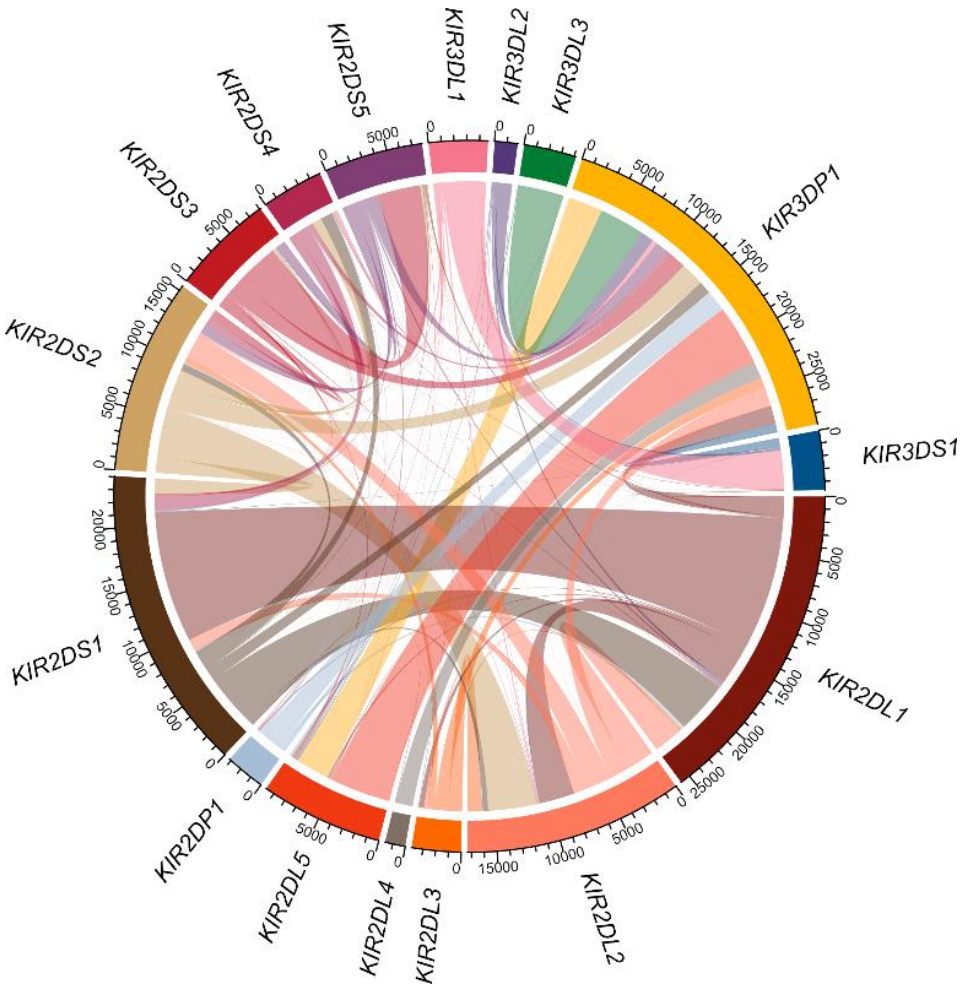


**Figure 2.6. Misaligned read sources in the synthetic dataset.**
Analysis of mismatched or unresolved genotype determination results for the synthetic sequence dataset where all misaligned reads are traced back to their source gene. The connections between genes represent the number of misaligned reads, and the color of the connection represents the source gene.

The analysis showed *KIR3DP1* as a major hub for receiving misaligned reads, with reads being contributed by each other *KIR* gene. In fact, *KIR3DP1* was largely the only receiver for misaligned reads originating from *KIR3DL3*, *KIR3DL2, KIR2DP1* and *KIR2DL4*. While the only genes receiving reads sourced from *KIR3DP1* were *KIR2DL5A* and *KIR2DL5B*.

The analysis also showed several gene pairings, where two genes largely sent and received reads from one another. Once such pairing was between *KIR2DL1* and *KIR2DS1*, where each gene were the largest contributor and receiver of reads for each other. Another pairing was between *KIR2DL2* and *KIR2DS2*, although both genes sent and received reads from several other genes.

This analysis illustrates the complex and highly interconnected nature of *KIR* and highlights the difficulty behind accurate interpretation of *KIR* short-read sequencing data.

**Performance**

The run time and resource utilization of the PING pipeline was measured on an Intel Xeon 2.20 GHz CPU using 36 threads. For ten sequences from the synthetic dataset, it took 1.92 mins for *KIR* read extraction, 34.7 mins for copy determination aligning to the minimized reference set, and 2.10 hours for genotype determination aligning to the minimized reference set. The output directory size was 1.4GB.

## Discussion

Our shared k-mer analysis of all documented *KIR* variation shows the high degree of sequence identity between *KIR* genes and illustrates the challenges imposed by the homology of *KIR* on short-read interpretation workflows. It demonstrates that some genes are more likely to exhibit read misalignment problems than others. *KIR2DP1*, *KIR3DL3* and *KIR2DL4* have relatively unique sequence, while *KIR2DS1*, *KIR2DL5A* and *KIR2DL5B* have considerable shared sequence. This type of analysis provides an informative tool for investigating irregularities in the processing of *KIR* sequence data, revealing which genes are likely to be erroneously interpreted due to read misalignments for common sequencing read lengths. While paired-end sequencing with longer reads can improve read alignment fidelity, in our own experience 290bp paired-end reads with a median insert length of approximately 600bp still exhibited considerable read misalignment problems. It is important to note that this analysis does not account for

unknown variation or intergenic sequence, two other sources of sequence variation that could potentially result in misaligned reads.

An initial determination of *KIR* gene content and copy number provides an informative scaffold for minimizing misalignments through the exclusion of reference sequence representing absent genes, as well as a system for identifying misalignments by searching for erroneously-called heterozygous SNP alleles in hemizygous genes. Thus, accurate copy number determination is a vital first step in interpreting *KIR* sequencing data. To achieve this goal, we developed a copy number determination method in PING that uses all described *KIR* alleles as an alignment resource, increasing the total number of reference sequences from 15 to 905 compared to single-sequence per gene alignments. While many of these alleles were only defined across exonic regions, rendering them ineffective for short-read alignment, we developed and implemented a protocol for intronic region imputation. The imputation method cannot resolve all uncharacterized nucleotide sequence, yet it accounts for the majority of missing sequence, greatly increasing the number of useful reference alleles. The exhaustive alignment provides a comprehensive map of the alleles to which a read may align, facilitating copy number resolution of important *KIR* allelic groups and genes that share extensive sequence similarity, such as *KIR2DL2*, *KIR2DL3*, *KIR2DL1* and *KIR2DS1*, which were inaccessible to previous bioinformatic methods (27,48). Additionally, the limited range of described UTR sequence, ~250bp 5'UTR and ~500bp 3'UTR, can reduce alignments over the first exon and potential regulatory regions (49,50).

The improved copy determination performance of PING, in addition to the expanded useful reference sequence repertoire, enables a smart, genotype-aware alignment workflow, designed to minimize read misalignments by closely matching reference sequences to the gene sequences present in the sequencing data. This alignment strategy addresses a major weakness of the filtration alignments utilized in the prototype workflow, which apply filters to retain gene-specific reads and eliminate cross-mapping reads regardless of the gene-content or sequence makeup, and thus often suffer from either

inadequate or patchy aligned read depths after filtration. There is a valid concern about carrying forward alignment biases in the genotype-aware alignment workflow, and the *KIR* system in particular is sensitive to reference bias because the combination of highly polymorphic genes and high sequence similarity between genes means that small changes in the reference sequence can have large impacts on read alignments. We have implemented several methods to counteract potential alignment biases that could be carried forward by the genotype-aware alignments. The first is the use of virtual probes to identify alleles and structural variants prone to misidentification. The second is the addition of a curated sequence set to the alignment reference for any gene with a determined genotype that does not perfectly match the aligned SNPs, these sequences were selected to cover a large amount of the allelic diversity of the corresponding gene. Even with these countermeasures we still encounter some improper novel genotype determinations, likely due to reference sequence bias. Since no novel genotypes were simulated for the synthetic dataset the Synthetic data in the Unresolved column of Table 2 represent improper novel genotype determinations. Despite these limitations, the genotype-aware workflow achieves highly accurate genotype determinations for the European dataset (**Table 2.2**), and highly accurate resolved genotype determinations across all tested datasets (**Table 2.3**).

Both the synthetic dataset and Khoesan cohort showed higher levels of unresolved genotypes compared to the European cohort (**Table 2.2**). These datasets represent challenging data to correctly interpret, with the Khoesan being an extremely divergent population with many unresolved genotypes in the validation data, and the synthetic dataset consisting of random alleles, some of which used imputed sequence. An analysis into the discordant results for the synthetic dataset (**Figure 2.6**, **S2.3 Table**) showed a complex web of cross-mapped reads. These cross-mapped reads can be extremely difficult to resolve because the high-degree of sequence shared among *KIR* genes (**Figure 2.4**) makes it almost impossible to determine correct mappings. Additionally, measures meant to prevent read misalignments, such as the use of virtual probes to refine reference sequence selection, can serve as a

double-edged sword, where the issue at hand is addressed but the changes create new sources for read misalignments.

An analysis into the discordant copy results highlights a major outstanding problem with the PING workflow since accurate copy determination is a central component of effective genotype-aware alignments, and the need for manual thresholding between copy groups introduces the component of user error. Continued development of the pipeline will address methods for automating copy determination for targeted sequencing data that matches or surpasses the accuracy achieved by manual thresholding. To compare PING against an existing method, we benchmarked against KPI (35) for determining *KIR* gene content (**Table S2.10**), achieving 100% concordance for *KIR3DP1*, *KIR2DL3*, *KIR2DL4*, *KIR3DL3* and *KIR3DL2*, over 97% concordance for *KIR2DS5*, *KIR2DP1*, *KIR2DS3*, *KIR2DS2*, *KIR3DL1*, *KIR2DL2*, *KIR2DS4*, *KIR2DL1*, and *KIR2DL5A/B*, and over 95% concordance for *KIR3DS1*, and *KIR2DS1*.

We believe improved interpretation of *KIR* sequencing data will ultimately be achieved through longer-range sequencing technologies that can extend past the range of the shared sequence motifs, and through better imputation approaches that can more fully characterize currently described *KIR* alleles to provide a more robust alignment reference. While long-read data from a platform with cost-effective methods might be difficult to interpret due to the high error rates (51), the combination of long-reads and short-reads would cover the weaknesses of the respective technologies and should provide a highly accurate *KIR* interrogation method, indeed, long-read methods have provided valuable insights into *KIR* haplotypes (28). We have not had the opportunity to test long-read technologies, but we anticipate a need for careful consideration of potential read misalignments when aligning the short and long reads together. Higher fidelity methods are currently under development, but currently the cost is prohibitive for the kind of high-throughput studies that PING was designed to address. Meanwhile, for samples for

which genotypes are not easily resolvable, we recommend direct visualization of sequence alignments potentially coupled with alternative laboratory methods to more precisely determine genotypes.

While the PING workflow is specific to interpreting sequence originating from the *KIR* complex, the underlying strategies can be extended to other problematic genomic regions. For example, multiple-sequence per gene alignment strategies provide information for discriminating between reads derived from genes with high sequence identity and extensive nucleotide polymorphisms. Additionally, genotype-aware alignment strategies reduce bias introduced by the reference sequence for reads derived from genomic regions with high structural variation.

In conclusion, PING incorporates these innovations to provide accurate, high-throughput interpretation of the *KIR* region from short-read sequencing data. Together, these modifications provide a consistent *KIR* genotyping pipeline, creating a highly automated, robust workflow for interpreting *KIR* sequencing data. To the best of our knowledge, this is the only bioinformatic workflow currently available for high-resolution *KIR* genotyping from short-read data. Given the importance of KIR variation in human health and disease, availability of a highly accurate method to assess *KIR* genotypic variation should promote important discoveries related to this complex genomic region.

## Supporting figures, tables, and text

**S2.1 Figure. European cohort copy determinations.** (JPEG)
Copy number determinations made for the European cohort. Sequences were aligned to the minimized reference set and copy thresholds were manually assigned.

**S2.2 Figure. Khoesan cohort copy determinations.** (JPEG)
Copy number determinations made for the Khoesan cohort. Sequences were aligned to the minimized reference set and copy thresholds were manually assigned.

**S2.3 Figure. Synthetic dataset copy determinations.** (JPEG)
Copy number determinations made for the Khoesan cohort. Sequences were aligned to the minimized reference set and copy thresholds were manually assigned.

**S2.1 Table. Diverse and minimized reference allele set.**

**S2.2 Table. Virtual probe table for reference modifications.**

**S2.3 Table. Analysis of read mapping errors for the synthetic dataset.**
(A) Summary of source gene read counts for reads aligning to discordantly genotyped genes. (B) Summary of aligned gene counts for reads sourced from discordantly genotyped genes.

**S2.4 Table. Validation genotype table.** (XLSX)

**S2.5 Table. PING determined copy number table.** (XLSX)

**S2.6 Table. Validation copy number table.** (XLSX)

**S2.7 Table. PING determined genotype table.** (XLSX)

**S2.8 Table. K-mer gene match table.** (XLSX)

**S2.9 Table. Synthetic dataset off-target read mappings.** (XLSX)

**S2.10 Table. Benchmarking PING and KPI gene content performance.** (XLSX)

**S2.1 Text. Genotype determination supporting methods.** (DOCX)

## References


1. Colonna M, Moretta A, Vély F, Vivier E. A high-resolution view of NK-cell receptors: Structure and function. In: Immunology Today. Elsevier Ltd; 2000. p. 428–31.

2. Björkström NK, Béziat V, Cichocki F, Liu LL, Levine J, Larsson S, et al. CD8 T cells express randomly selected KIRs with distinct specificities compared with NK cells. Blood. 2012 Oct 25;120(17):3455–65.

3. Fauriat C, Ivarsson MA, Ljunggren HG, Malmberg KJ, Michaëlsson J. Education of human natural killer cells by activating killer cell immunoglobulin-like receptors. Blood. 2010 Feb 11;115(6):1166–74.

4. Pende D, Falco M, Vitale M, Cantoni C, Vitale C, Munari E, et al. Killer Ig-like receptors (KIRs): Their role in NK cell modulation and developments leading to their clinical exploitation. Vol. 10, Frontiers in Immunology. Frontiers Media S.A.; 2019. p. 1179.

5. Kumar S. Natural killer cell cytotoxicity and its regulation by inhibitory receptors [Internet]. Vol. 154, Immunology. Blackwell Publishing Ltd; 2018 [cited 2020 Mar 5]. p. 383–93. Available from: http://doi.wiley.com/10.1111/imm.12921

6. Parham P. MHC class I molecules and KIRs in human history, health and survival. Nat Rev Immunol [Internet]. 2005 Mar [cited 2015 Feb 10];5(3):201–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15719024

7. Nelson GW, Martin MP, Gladman D, Wade J, Trowsdale J, Carrington M. Cutting Edge: Heterozygote Advantage in Autoimmune Disease: Hierarchy of Protection/Susceptibility Conferred by HLA and Killer Ig-Like Receptor Combinations in Psoriatic Arthritis. J Immunol. 2004 Oct 1;173(7):4273–6.

8. Salie M, Daya M, Möller M, Hoal EG. Activating KIRs alter susceptibility to pulmonary tuberculosis in a South African population. Tuberculosis. 2015 Dec 1;95(6):817–21.

9. Hirayasu K, Ohashi J, Kashiwase K, Hananantachai H, Naka I, Ogawa A, et al. Significant association of KIR2DL3-HLA-C1 combination with cerebral malaria and implications for co-evolution of KIR and HLA. PLoS Pathog. 2012 Mar;8(3).

10. Khakoo SI, Thio CL, Martin MP, Brooks CR, Gao X, Astemborski J, et al. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. Science (80- ). 2004 Aug 6;305(5685):872–4.

11. Martin MP, Gao X, Lee JH, Nelson GW, Detels R, Goedert JJ, et al. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. Nat Genet. 2002;31(4):429–34.

12. Giebel S, Locatelli F, Lamparelli T, Velardi A, Davies S, Frumento G, et al. Survival advantage with KIR ligand incompatibility in hematopoietic stem cell transplantation from unrelated donors. Blood [Internet]. 2003 Apr 3 [cited 2019 Oct 11];102(3):814–9. Available from: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2003-01-0091

13. Cooley S, Trachtenberg E, Bergemann TL, Saeteurn K, Klein J, Le CT, et al. Donors with group B KIR haplotypes improve relapse-free survival after unrelated hematopoietic cell transplantation for acute myelogenous leukemia. Blood [Internet]. 2009 Jan 15 [cited 2019 Oct 11];113(3):726–32. Available from: https://ashpublications.org/blood/article/113/3/726/25172/Donors-with-group-B-KIR-haplotypes-improve

14. Venstrom JM, Pittari G, Gooley TA, Chewning JH, Spellman S, Haagenson M, et al. HLA-C-dependent prevention of leukemia relapse by donor activating KIR2DS1. N Engl J Med [Internet].

2012 Aug 30 [cited 2019 Oct 11];367(9):805–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22931314

15. Nakamura R, Gendzekhadze K, Palmer J, Tsai NC, Mokhtari S, Forman SJ, et al. Influence of donor KIR genotypes on reduced relapse risk in acute myelogenous leukemia after hematopoietic stem cell transplantation in patients with CMV reactivation. Leuk Res. 2019 Dec 1;87:106230.

16. Cooley S, Weisdorf DJ, Guethlein LA, Klein JP, Wang T, Marsh SGE, et al. Donor Killer Cell Ig-like Receptor B Haplotypes, Recipient HLA-C1, and HLA-C Mismatch Enhance the Clinical Benefit of Unrelated Transplantation for Acute Myelogenous Leukemia. J Immunol. 2014 May 15;192(10):4592–600.

17. Guethlein LA, Beyzaie N, Nemat-Gorgani N, Wang T, Ramesh V, Marin WM, et al. Following transplantation for AML, donor KIR Cen B02 better protects against relapse than KIR Cen B01. J Immunol.

18. Martin AM, Freitas EM, Witt CS, Christiansen FT. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. Immunogenetics [Internet]. 2000 Apr 4 [cited 2019 Oct 11];51(4–5):268–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10803839

19. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, et al. Human diversity in killer cell inhibitory receptor genes. Immunity. 1997 Dec 1;7(6):753–63.

20. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. Proc Natl Acad Sci [Internet]. 2000 Apr 25 [cited 2019 Oct 11];97(9):4778–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10781084

21. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SGE. IPD--the Immuno Polymorphism Database. Nucleic Acids Res [Internet]. 2013 Jan [cited 2015 Feb 13];41(Database issue):D1234-40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531162&tool=pmcentrez&rendertype=abstract

22. Norman PJ, Abi-Rached L, Gendzekhadze K, Hammond JA, Moesta AK, Sharma D, et al. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. Genome Res. 2009 May 1;19(5):757–69.

23. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. [cited 2020 May 6]; Available from: http://www.ebi.ac.uk/ipd/kir/

24. Hollenbach JA, Nocedal I, Ladner MB, Single RM, Trachtenberg EA. Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations. Immunogenetics [Internet]. 2012 Oct 1 [cited 2019 Oct 11];64(10):719–37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22752190

25. Hollenbach JA, Augusto DG, Alaez C, Bubnova L, Fae I, Fischer G, et al. 16(th) IHIW: population global distribution of killer immunoglobulin-like receptor (KIR) and ligands. Int J Immunogenet [Internet]. 2013 Feb [cited 2019 Oct 11];40(1):39–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23280119

26. Hsu KC, Chida S, Geraghty DE, Dupont B. The killer cell immunoglobulin-like receptor (KIR) genomic region: Gene-order, haplotypes and allelic polymorphism. Vol. 190, Immunological Reviews. 2002. p. 40–52.

27. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. Am J Hum Genet [Internet]. 2016 Aug 4 [cited 2017 Oct 10];99(2):375–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27486779

28. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. Genes Immun [Internet]. 2017 [cited 2019 Oct 11];18(3):127–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28569259

29. Uhrberg M, Parham P, Wernet P. Definition of gene content for nine common group B haplotypes of the Caucasoid population: KIR haplotypes contain between seven and eleven KIR genes. Immunogenetics. 2002;54(4):221–9.

30. Amorim LM, Santos THS, Hollenbach JA, Norman PJ, Marin WM, Dandekar R, et al. Cost-effective and fast KIR gene-content genotyping by multiplex melting curve analysis. HLA [Internet]. 2018 Dec 1 [cited 2021 Mar 11];92(6):384–91. Available from: https://pubmed.ncbi.nlm.nih.gov/30468002/

31. Wagner I, Schefzyk D, Pruschke J, Schöfl G, Schöne B, Gruber N, et al. Allele-Level KIR Genotyping of More Than a Million Samples: Workflow, Algorithm, and Observations. Front Immunol [Internet]. 2018 Dec 4 [cited 2019 Oct 11];9:2843. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30564239

32. Anderson KM, Augusto DG, Dandekar R, Shams H, Zhao C, Yusufali T, et al. Killer Cell Immunoglobulin-like Receptor Variants Are Associated with Protection from Symptoms Associated with More Severe Course in Parkinson Disease. J Immunol [Internet]. 2020 Sep 1 [cited 2021 Mar 11];205(5):1323–30. Available from: https://www.jimmunol.org/content/205/5/1323

33. Vargas L de B, Dourado RM, Amorim LM, Ho B, Calonga-Solís V, Issler HC, et al. Single Nucleotide Polymorphism in KIR2DL1 Is Associated With HLA-C Expression in Global Populations. Front Immunol [Internet]. 2020 Aug 21 [cited 2021 Mar 11];11. Available from: /pmc/articles/PMC7478174/

34. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, et al. Imputation of KIR Types from SNP Variation Data. Am J Hum Genet [Internet]. 2015 [cited 2021 Jun 22];97(4):593–607. Available from: /pmc/articles/PMC4596914/

35. Roe D, Kuang R. Accurate and Efficient KIR Gene and Haplotype Inference From Genome Sequencing Reads With Novel K-mer Signatures. Front Immunol [Internet]. 2020 Nov 26 [cited 2021 Jun 22];11:1. Available from: /pmc/articles/PMC7727328/

36. Chen J, Madireddi S, Nagarkar D, Migdal M, Vander Heiden J, Chang D, et al. In silico tools for accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on individual-patient and population scales . Brief Bioinform [Internet]. 2021 May 20 [cited 2021 Jun 22];22(3):1–11. Available from: https://academic.oup.com/bib/article/22/3/bbaa223/5906908

37. Robinson J, Waller MJ, Stoehr P, Marsh SGE. IPD--the Immuno Polymorphism Database. Nucleic Acids Res [Internet]. 2004 Dec 17 [cited 2019 Oct 11];33(Database issue):D523–6. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki032

38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug [cited 2021 Mar 11];25(16):2078–9. Available from: /pmc/articles/PMC2723002/

39. Leaton LA, Shortt J, Kichula KM, Tao S, Nemat-Gorgani N, Mentzer AJ, et al. Conservation, extensive heterozygosity, and convergence of signaling potential all indicate a critical role for KIR3DL3 in higher primates. Front Immunol. 2019;10(JAN).

40. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. Bioinformatics [Internet]. 2012 Feb [cited 2021 Mar 11];28(4):593–4. Available from: /pmc/articles/PMC3278762/

41. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. Genome Res. 2012 Oct;22(10):1845–54.

42. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics [Internet]. 2014 Oct 1 [cited 2019 Oct 11];30(19):2811–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu393

43. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2018. Available from: https://www.r-project.org/

44. Hollenbach JA, Norman PJ, Creary LE, Damotte V, Montero-Martin G, Caillier S, et al. A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. Proc Natl Acad Sci U S A [Internet]. 2019 Apr 9 [cited 2021 Mar 11];116(15):7419–24. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1821778116

45. Nemat-Gorgani N, Guethlein LA, Henn BM, Norberg SJ, Chiaroni J, Sikora M, et al. Diversity of KIR, HLA Class I, and Their Interactions in Seven Populations of Sub-Saharan Africans. J Immunol. 2019 May 1;202(9):2636–47.

46. Marin WM, Dandekar R, Augusto DG, Yusufali T, Norman PJ, Hollenbach JA. PING [Internet]. Github. 2020 [cited 2020 Sep 29]. Available from: https://github.com/wesleymarin/PING

47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods [Internet]. 2012 Apr 4 [cited 2019 Oct 11];9(4):357–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22388286

48. Pyke RM, Genolet R, Harari A, Coukos G, Gfeller D, Carter H. Computational KIR copy number discovery reveals interaction between inhibitory receptor burden and survival. Pac Symp Biocomput [Internet]. 2019 [cited 2019 Oct 11];24:148. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417817/

49. Li H, Wright PW, McCullen M, Anderson SK. Characterization of KIR intermediate promoters reveals four promoter types associated with distinct expression patterns of KIR subtypes. Genes Immun [Internet]. 2016 Jan 1 [cited 2021 Mar 11];17(1):66–74. Available from: /pmc/articles/PMC4724278/

50. Nutalai R, Gaudieri S, Jumnainsong A, Leelayuwat C. Regulation of KIR3DL3 expression via mirna. Genes (Basel) [Internet]. 2019 Aug 1 [cited 2021 Mar 11];10(8). Available from: /pmc/articles/PMC6723774/

51. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis [Internet]. Vol. 21, Genome Biology. BioMed Central Ltd.; 2020 [cited 2021 Jun 25]. p. 1–16. Available from: https://doi.org/10.1186/s13059-020-1935-5

# Chapter 3: Software Update – Interpreting Killer-cell Immunoglobulin-like Receptor from Whole Genome Sequence Data with PING


## Abstract

Here, we demonstrate improvements to our bioinformatic pipeline, PING, which provides high-resolution genotyping of killer-cell immunoglobulin-like receptor (KIR) sequencing data, that expand the method to provide *KIR* interpretation from whole genome sequencing (WGS) data. We evaluated performance using synthetic sequence datasets and real-world data from the 1000 Genomes Project (1KGP). PING demonstrated high exonic genotyping performance on the synthetic sequence dataset meant to approximate real-world data at 95% accuracy (N=1366). This result was mirrored in the analysis of 1KGP European data (N=215) with most genes showing near or below 5% frequency of unresolved exonic genotypes, which is an important indicator for genotyping errors in real-world data. An analysis into the distributions of genotyping errors for the synthetic sequence datasets gave insights into how to further improve genotype accuracy. Similarly, an analysis into ambiguous exonic genotype frequencies for the 1KGP European data, which showed high rates of unresolved genotypes, highlighted that an effective phasing method will be an impactful future additional to the PING workflow. Together, these results demonstrate that PING can effectively provide high-resolution *KIR* genotyping on WGS data.


## Introduction

Previously, we introduced a bioinformatic pipeline, PING(1), for the high-throughput interpretation of targeted short-read sequencing data of the killer-cell immunoglobulin-like receptor (KIR) complex, located in human chromosomal region 19q13.42(2). Here, we expand that method to provide KIR interpretation from whole genome sequence (WGS) data. Our motivation for this work is to increase the

utility of WGS datasets, which has become a standard sequencing approach, and to open an avenue for advancing our understanding of *KIR* variation across diverse populations.

To accomplish this, we have made alterations to the workflow to account for differences between targeted and WGS data, and we have constructed three distinct synthetic sequence datasets that approximate WGS data, each designed to test different aspects of the workflow performance. The synthetic sequence datasets incorporate copy number variation, based on commonly observed haplotypes(3), and allelic variation, sourced from the IPD-KIR allele database(4). One dataset, termed 'syn-known', only incorporates alleles that represent those described in the allele database. The second data set, termed 'syn-novel', incorporates novel SNPs and recombinants to assess the workflow performance on novel sequence. The third data set, termed 'syn-matched', is built similarly to the syn-known dataset, however, when these samples were run through PING their component alleles informed reference sequence selection in the genotype aware alignment workflow, providing a theoretical maximum performance value for the genotype aware alignments. Finally, as a proof-of-concept for real world WGS data, we processed 215 sequences from the 1000 Genome Project (1KGP) European (EUR) superpopulation(5,6).

## Materials and Methods

### PING workflow

The PING workflow is described in detail in Marin et al.(1). Briefly, PING takes in paired-end sequencing data and undergoes a series of dynamic alignments to output gene copy number, high-resolution genotypes, and information about potential novel alleles. First, a filtration alignment isolates *KIR* specific reads, which are used as input sequence data for the rest of the workflow. Second, PING determines *KIR* gene content and copy number. The ascertained gene content informs a gene content matched alignment, which is an alignment to a reference that excludes sequences from genes determined to be

absent. An initial genotype determination informs a genotype matched alignment, which is an alignment to a reference that includes sequences that represent the determined genotype. Genotype determination and subsequent genotype matched alignments are repeated multiple times with varying parameters to identify the best fit genotype, which is used to inform a final genotype matched alignment which is processed to provide the final output. PING utilizes bowtie2(7) for alignments, and samtools(8) for alignment processing in addition to custom alignment processing methods.

Alterations made for processing WGS data included decreasing the minimum alignment depth to 6 for both initial and final genotyping and the addition of more virtual probes for correcting commonly misidentified genotypes (**Table S3.1**).

The PING WGS workflow is available at: https://github.com/wesleymarin/PING/tree/wgs_snakemake

### Synthetic sequence datasets

Synthetic sequence datasets were generated using ART(9). The workflow for generating these datasets follows the outline described in Marin et al.(1), including the simulation of structural variation and missing sequence imputation. Adjustments made for simulating WGS data included lowering the coverage depth from 50x to 30x, and lowering read length from 150-bp to 140-bp.

The syn-known synthetic sequence dataset (N=100) was generated with structural variation and allelic variation but no novel variation (**Table 3.1**). The goal of this dataset was to assess the performance of PING on known and described alleles, which is generally the most common use case.

The syn-novel synthetic sequence dataset (N=100) was generated with structural variation and allelic variation with novel sequence variants. All generated allele sequences have novel variation, split evenly between novel SNPs and recombinations. Allele sequences were randomly assigned to have either a novel SNP or a recombination sequence between another allele from the same gene. The novel SNP or recombination point was introduced at a random position between 0.25-0.75 of the full allele sequence.

The syn-matched synthetic sequence dataset (N=100) was generated with structural variation and allelic variation but no novel variation. The goal of this dataset was to assess the maximum performance of PING by aligning the synthetic sequence data to their true genotypes. The PING workflow was run over this dataset with modifications to use the true genotypes as input to the genotype-aware alignment workflow.

**Table 3.1. Descriptions of the synthetic sequence datasets.**

| Dataset | Characteristics | Application |
|---|---|---|
| syn-known | Structural variation and allelic variation | Assess performance on characterized alleles |
| syn-novel | Structural variation, allelic variation, novel SNPs and novel recombinants | Assess performance on novel sequence |
| syn-matched | Structural variation and allelic variation | Genotypes inform genotype matched alignment as performance benchmark |

Each of these synthetic datasets are available at: https://github.com/wesleymarin/KIR_synthetic_data.

## Thousand Genome Project analysis

1KGP project analysis was limited to 215 individuals from the European superpopulation (**Table S3.2**). Sequence files were 30x high-coverage WGS data(5,6). *KIR* aligned reads were extracted from CRAM files aligned to GRCh38 and converted to paired-end FASTQ format via samtools(8) and bazam(10) using coordinates described in **Table S3.3**.

## Performance assessment

The performance of PING across the synthetic datasets was assessed using alignment coverage and number of genotype errors. Alignment coverage represents the number of bases in the final PING alignment that are above the minimum depth threshold (default of 6) compared to the total number of

bases in the synthetic FASTA sequence. This comparison is represented as a ratio between 0-1, where 0 represents no alignment coverage and 1 represents complete alignment coverage. The genotype error score represents SNP mismatches between the final alignment determined by PING and the synthetic FASTA sequence. A genotype error score of 0 means there was no difference between the aligned SNPs and the original sequence. These metrics were measured independently for exons, introns and UTRs for each *KIR* gene and major allelic group for each sample in each of the datasets.

Performance on real-world data was assessed by the frequency of unresolved exonic genotypes, which are genotypes that do not match any described allele sequence, the frequency of ambiguous exonic genotypes, which are genotypes with multiple possible allele typings that match the aligned SNPs, and average alignment coverage, represented as a ratio between 0-1. While we do expect unresolved genotypes that represent true novel sequence in real-world data, they are also a common outcome of read misalignments. Genotype ambiguity is both an outcome of incomplete alignment coverage, since positions with inadequate coverage are not considered during genotype determination, and a lack of an effective phasing method. For comparison, we also applied the unresolved genotype frequency and ambiguous genotype frequency metrics to the syn-known and syn-matched datasets.

## Results

**Table 3.2. Summary table of gene alignment coverage and genotype errors by gene feature for each synthetic sequence dataset.**
Coverage is calculated on a scale of 0-1, with 0 being no coverage and 1 being perfect coverage. Genotype error is calculated by summing the total number of SNP mismatches between the genotype determined through alignment and the original synthetic FASTA sequence, a genotype error score of 0 is perfect. The perfect column summarizes the ratio of genotypes with full alignment coverage or no genotype errors.

| | feature | dataset | mean | median | sd | min | max | perfect | N |
|---|---|---|---|---|---|---|---|---|---|
| coverage | exon | syn-novel | 0.999 | 1.000 | 0.009 | 0.798 | 1.000 | 0.99 | 1340 |
| | exon | syn-known | 1.000 | 1.000 | 0.004 | 0.866 | 1.000 | 1.00 | 1366 |
| | exon | syn-matched | 0.998 | 1.000 | 0.014 | 0.781 | 1.000 | 0.96 | 1366 |
| | intron | syn-novel | 0.997 | 1.000 | 0.022 | 0.663 | 1.000 | 0.95 | 1340 |
| | intron | syn-known | 0.998 | 1.000 | 0.018 | 0.597 | 1.000 | 0.97 | 1366 |
| | intron | syn-matched | 0.982 | 1.000 | 0.076 | 0.539 | 1.000 | 0.88 | 1366 |

| | feature | dataset | mean | median | sd | min | max | perfect | N |
|---|---|---|---|---|---|---|---|---|---|
| | UTR | syn-novel | 0.952 | 0.959 | 0.026 | 0.751 | 0.987 | 0.00 | 1340 |
| | UTR | syn-known | 0.953 | 0.959 | 0.023 | 0.825 | 0.988 | 0.00 | 1366 |
| | UTR | syn-matched | 0.951 | 0.958 | 0.033 | 0.483 | 0.988 | 0.00 | 1366 |
| genotype errors | exon | syn-novel | 0.5 | 0 | 3.8 | 0 | 37 | 0.92 | 1340 |
| | exon | syn-known | 0.5 | 0 | 3.7 | 0 | 37 | 0.95 | 1366 |
| | exon | syn-matched | 0.3 | 0 | 3.3 | 0 | 37 | 0.98 | 1366 |
| | intron | syn-novel | 27.0 | 2 | 197.3 | 0 | 3301 | 0.43 | 1340 |
| | intron | syn-known | 28.9 | 0 | 258.8 | 0 | 4679 | 0.56 | 1366 |
| | intron | syn-matched | 20.7 | 0 | 231.2 | 0 | 4616 | 0.61 | 1366 |
| | UTR | syn-novel | 0.7 | 0 | 5.3 | 0 | 68 | 0.89 | 1340 |
| | UTR | syn-known | 0.7 | 0 | 6.1 | 0 | 67 | 0.93 | 1366 |
| | UTR | syn-matched | 0.5 | 0 | 5.1 | 0 | 61 | 0.95 | 1366 |

PING displayed high alignment coverage performance across exons for the synthetic datasets, with 100% of the syn-known dataset and 99% of the syn-novel dataset showing perfect coverage (**Table 2.2**), compared to 96% perfect coverage for the syn-matched dataset. Intron coverage was also high, with 97% of the syn-known dataset and 95% of the syn-novel dataset showing perfect coverage, compared to 88% perfect coverage for the syn-matched dataset. The UTR region coverage was very similar for all datasets at around 95% mean coverage.

For genotype determination performance PING displayed high exon performance for the synthetic datasets, with 95% of the syn-known dataset and 92% of the syn-novel dataset showing perfect genotypings, compared to 98% for the syn-matched dataset. Intronic genotype determination performance was poor across the synthetic datasets, with only 56% of the syn-known dataset and 43% of the syn-novel dataset showing perfect genotypings, compared to 61% for the syn-matched dataset. Analysis of UTR genotype determination showed 93% of the syn-known dataset and 89% of the syn-novel dataset were perfectly genotyped, compared to 95% for the syn-matched dataset.

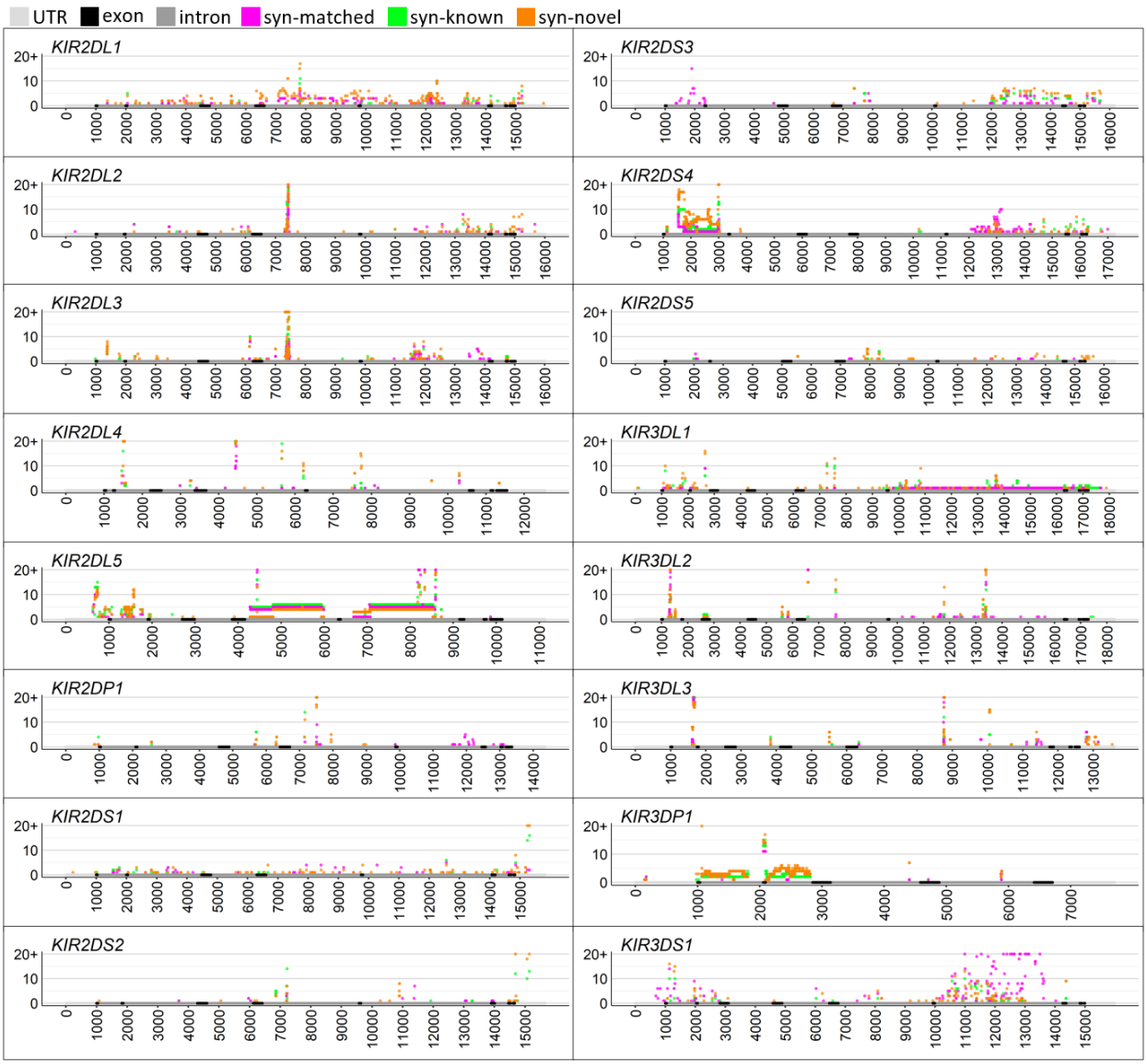


**Figure 3.1. Distributions of genotype errors by gene position for each KIR gene and major allelic group across the synthetic datasets.**
Positions of UTR, exon, and intron sequence were marked. The x-axis represents gene position and the y-axis represents total genotype errors.

Examination of genotype errors by gene and base position for the synthetic datasets showed common error types that we grouped into three categories. The first was scattered errors, where the errors seem randomly distributed, for example *KIR2DL1*. The second was hotspot errors, which were singular locations that showed a high number of errors, for example *KIR2DL2* and *KIR2DL3*. Finally, there were

structural variant errors, where large scale insertion or deletion sequence was misinterpreted, for example *KIR3DP1* and *KIR2DL5*.

*KIR2DL1* alignments mainly displayed scattered errors across the entire sequence with a substantial hotspot error in intron 5 (**Figure 3.1**). *KIR2DL2* and *KIR2DL3* showed a substantial hotspot error in intron 5 and scattered errors across introns 6 and 7. *KIR2DL4* displayed a large hotspot error in intron 2, and smaller hotspot errors in introns 5 and 6. *KIR2DL5* showed concentrated hotspot and scattered errors across the 5'UTR, exon 1, and intron 1, as well as structural variant errors and more hotspot errors across introns 5 and 6. *KIR2DP1* showed some hotspot errors in intron 5 but was relatively error free. *KIR2DS1* and *KIR2DS2* showed scattered errors and some hotspot errors yet were largely error free. *KIR2DS3* showed scattered errors across introns 1, 6 and 7, and the 3'UTR. *KIR2DS4* showed structural variant errors and hotspot errors across intron 1 and scattered errors across introns 6 and 7. *KIR2DS5* was largely error free. *KIR3DL1* showed scattered errors and hotspot errors across many introns. *KIR3DL2* showed a handful of error prone hotspots, one in intron 1 and the second in intron 6; both hotspots had over 40 genotyping errors in each dataset. *KIR3DL3* showed error prone hotspots in intron 1 and intron 5/6. *KIR3DS1* showed scattered genotyping errors predominately along intron 6. *KIR3DP1* showed genotyping errors over the long structural variant regions in intron 1 and 2 for the syn-known and syn-novel datasets, while the syn-matched dataset performed much better across this region.

Further examination of *KIR2DL5* structural variant errors across introns 5 and 6, which exhibited problems in all synthetic sequence datasets, were found to be due to low alignment depth of the structural variant (variant depth of 2 compared to non-variant depth ~80). This issue was traced back to the *KIR* read filtration step where reads originating from the structural variant were mistakenly filtered out.

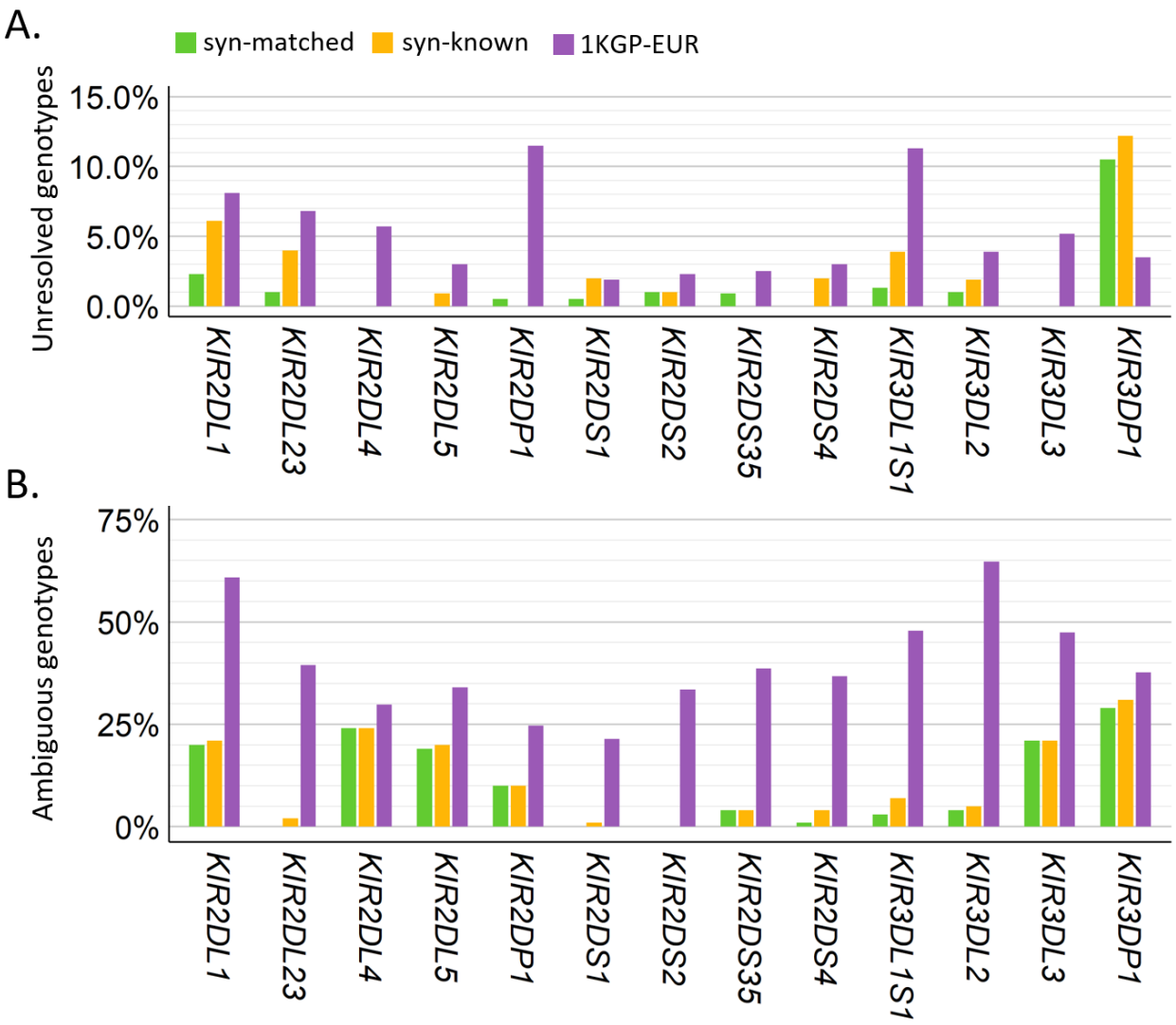


**Figure 3.2. Performance evaluation for 1KGP and synthetic datasets.**
**(A)** Summary of unresolved exonic genotype frequencies, a performance metric where lower values were generally better for real-world data and 0% was best for synthetic data, for the syn-matched and syn-known datasets as well as the 1KGP European dataset. **(B)** Summary of ambiguous exonic genotype frequencies, an outcome-based performance metric for alignment coverage where lower values were best, for the syn-matched and syn-known datasets as well as the 1KGP European dataset. EUR = European.

Examination of unresolved exonic genotype frequencies, a performance metric where lower frequencies were generally better for real-world data and 0% was best for synthetic data, showed strong performance for *KIR2DL5*, *KIR2DS1*, *KIR2DS2*, *KIR2DS35*, *KIR2DS4*, *KIR3DL2* and *KIR3DL3* across all examined datasets, with unresolved genotype frequencies near or below 5% (**Figure 3.2A**). Generally, PING exhibited the best unresolved genotype performance on the syn-matched dataset, followed by the

syn-known dataset and finally the 1KGP European dataset. An exception to this trend was *KIR3DP1*, where the 1KGP European data showed an unresolved genotype frequency of 3.5% yet both synthetic datasets showed frequencies above 10%. PING exhibited the worst real-world data performance on *KIR2DP1* and *KIR3DL1S1*, with unresolved genotype frequencies around 11% for each.

Examination of ambiguous exonic genotype frequencies, an outcome-based performance metric where lower frequencies were best across all datasets, showed large disparities between the synthetic datasets and the 1KGP European dataset (**Figure 3.2B**). PING performed very well on the synthetic datasets by this metric, with many genes having close to 0% ambiguous genotypes. However, performance on real-world data was very different with almost all genes showing over 25% ambiguous genotypes except for *KIR2DP1* and *KIR2DS1*, which were both over 20%. The worst real-world data performance was for *KIR3DL2* at 65% ambiguous genotype frequency.

**Table 3.3. Summary of alignment coverage by gene feature for each *KIR*gene for the 1KGP European dataset.**
Coverage is represented by a ratio of 0-1.

| | exon | | | intron | | | UTR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | median | sd | mean | median | sd | mean | median | sd | n |
| *KIR2DL1* | 0.90 | 0.97 | 0.15 | 0.90 | 0.96 | 0.13 | 0.86 | 0.92 | 0.14 | 210 |
| *KIR2DL2* | 0.82 | 0.84 | 0.14 | 0.74 | 0.73 | 0.12 | 0.72 | 0.74 | 0.19 | 104 |
| *KIR2DL3* | 0.87 | 0.92 | 0.16 | 0.82 | 0.89 | 0.15 | 0.80 | 0.88 | 0.19 | 200 |
| *KIR2DL4* | 0.95 | 1.00 | 0.10 | 0.95 | 0.98 | 0.07 | 0.89 | 0.92 | 0.11 | 215 |
| *KIR2DL5* | 0.86 | 0.91 | 0.15 | 0.74 | 0.70 | 0.15 | 0.68 | 0.70 | 0.17 | 106 |
| *KIR2DP1* | 0.91 | 0.97 | 0.13 | 0.90 | 0.96 | 0.12 | 0.77 | 0.83 | 0.17 | 210 |
| *KIR2DS1* | 0.80 | 0.83 | 0.16 | 0.73 | 0.74 | 0.14 | 0.54 | 0.54 | 0.17 | 85 |
| *KIR2DS2* | 0.83 | 0.86 | 0.14 | 0.77 | 0.77 | 0.12 | 0.76 | 0.82 | 0.20 | 106 |
| *KIR2DS3* | 0.86 | 0.90 | 0.15 | 0.76 | 0.75 | 0.13 | 0.66 | 0.68 | 0.15 | 60 |
| *KIR2DS4* | 0.86 | 0.94 | 0.17 | 0.79 | 0.87 | 0.15 | 0.79 | 0.85 | 0.17 | 202 |
| *KIR2DS5* | 0.81 | 0.85 | 0.14 | 0.72 | 0.71 | 0.09 | 0.62 | 0.63 | 0.14 | 71 |
| *KIR3DL1* | 0.89 | 0.95 | 0.13 | 0.87 | 0.96 | 0.14 | 0.81 | 0.89 | 0.18 | 202 |
| *KIR3DL2* | 0.95 | 0.98 | 0.08 | 0.94 | 0.96 | 0.05 | 0.90 | 0.93 | 0.07 | 215 |
| *KIR3DL3* | 0.98 | 1.00 | 0.03 | 0.97 | 0.98 | 0.03 | 0.86 | 0.88 | 0.07 | 215 |
| *KIR3DP1* | 0.96 | 1.00 | 0.08 | 0.96 | 0.99 | 0.09 | 0.92 | 0.94 | 0.07 | 215 |
| *KIR3DS1* | 0.80 | 0.82 | 0.14 | 0.77 | 0.76 | 0.12 | 0.72 | 0.74 | 0.14 | 82 |

Examination of alignment coverage for the 1KGP European dataset showed exons generally have the best coverage, followed by introns and finally UTRs (**Table 3.3**). *KIR2DL4*, *KIR3DL2*, *KIR3DL3* and *KIR3DP1* showed the best mean exonic coverage, each above 95%, while *KIR3DS1*, *KIR2DS1*, *KIR2DS5* and *KIR2DL2* showed the worst exonic coverage, each around 80%. Mean intronic alignment coverage was best for *KIR2DL4*, *KIR3DL2*, *KIR3DL3* and *KIR3DP1*, each at 94% or above, while most other genes had between 70-80% coverage. Mean UTR coverage was best for *KIR3DL2*, *KIR3DP1* and KIR2DL4, each near 90%, and was the worst for *KIR2DS1* at 54%.

## Discussion

PING displayed consistently high exon interpretation performance across the synthetic sequence datasets based on the coverage and genotype error metrics (**Table 3.2**). This performance did not track across introns, where examination of genotype errors by position showed a high concentration of errors across intronic regions for many *KIR* genes (**Figure 3.1**). One cause of this discrepancy is the difference in completeness of exonic and intronic sequence in the IPD-KIR allele database(4), where many *KIR* sequences are only described across exons(1). While imputation has improved intronic alignments, this analysis shows there are still improvements to be made and highlights how missing data in the allele database can impact alignment fidelity.

A peculiar outcome of the alignment coverage analysis of the synthetic datasets was the generally lower alignment coverage of the syn-matched dataset compared to the syn-known and syn-novel datasets (**Table 3.2**). This occurred because the PING workflow was modified for the syn-matched dataset to only align to the component alleles of each samples true genotype, which resulted in alignment to sequences that were not fully characterized and ultimately leading to regions of missing alignment coverage. This was different for the syn-known and syn-novel datasets, which were run normally through PING and had the possibility of aligning to more completely characterized sequences.

We were pleased to observe that many results of the syn-known dataset tracked closely with the results of the syn-matched dataset. The purpose of the syn-matched dataset was to benchmark perfect performance of the genotype aware alignment strategy, which was accomplished by using the true genotypes of the samples to inform the genotype aware alignments. It was promising to see similar performance for the syn-known dataset for exonic coverage and genotype errors (**Table 3.2**), as well as for unresolved genotype frequencies and ambiguous genotype frequencies (**Figure 3.2**). This result suggests that the rounds of genotype determination and subsequent genotype aware alignments PING utilizes are effective at approximating the true genotype of a sample.

PING displayed high performance on real-world data from the 1KGP European dataset across most *KIR* genes based on the low frequencies of unresolved genotypes. While there are still improvements to be made for *KIR2DP1* and *KIR3DL1S1*, both showing above 10% unresolved genotype frequencies, and to a lesser extent *KIR2DL1* and *KIR2DL23*, it was very promising that most genes had frequencies near or below 5%. While we expected to observe some measure of unresolved genotypes in real-world data, which can represent novel sequence, we did not expect to find high frequencies in European data since this has been a highly studied superpopulation. A helpful comparator in this analysis were the syn-known and syn-matched datasets, which did not have novel variation and should have had no unresolved genotypes. Results from these datasets gave an approximation of the frequency of unresolved genotypes that were caused by read misalignments and processing errors. However, since the allelic makeup of the synthetic datasets and the 1KGP European dataset were obviously different this comparison is just an estimate from which no concrete values can be derived.

A major shortcoming of PINGs performance on real-world data was the frequency of ambiguous genotypes, which are genotypes that have multiple possible alleles that match the aligned SNPs. This was an area where there was a stark difference in PING's performance between the synthetic and the real-world data. There are several possible explanations for this result. One is that the real-world data

does not have as robust of alignment coverage as the synthetic data, which would result in fewer SNPs being utilized for genotype determination. Indeed, an analysis of alignment coverage showed that the 1KGP European dataset had worse alignment coverage than the synthetic datasets (**Tables 3.1, 3.2**). However, *KIR3DL2*, *KIR3DL3* and *KIR3DP1* had the highest exonic alignment coverage for the 1KGP data (**Table 3.3**) yet still displayed high frequencies of ambiguous genotypes (**Figure 3.2B**). Another explanation for the increase in ambiguous genotypes is that the random allele composition of the synthetic datasets led to fewer phasing problems then the real-world dataset that is composed of evolutionarily related alleles.

While ambiguous genotypes can be addressed bioinformatically through haplotype estimation(11), these methods rely on observational data and can be difficult to properly apply to unobserved populations. Implementation of a read-backed phasing approach to help address ambiguity issues will be a vital step to improving the utility of PING.

In conclusion, we have demonstrated PING can effectively interpret WGS data to provide high-resolution *KIR* genotypes using a mix of synthetic data and real-world data from the 1000 Genomes Project. To the best of our knowledge, PING is still the only published and proven bioinformatic pipeline for accurate high-resolution *KIR* genotyping from next generation sequencing data. We are confident that this work will greatly increase the utility of PING, which will continue to serve as a platform to advance our knowledge of *KIR* variation.

## Supplemental figures and tables

**S3.1 Table. Virtual probe data for reference modifications.** (XLSX)

**S3.2 Table. Thousand Genomes sample identifications and locations.** (TSV)

**S3.3 Table. KIR genomic coordinates.** (BED)

## References


1. Marin WM, Dandekar R, Augusto DG, Yusufali T, Heyn B, Hofmann J, et al. High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING. PLOS Comput Biol. 2021 Aug 2;17(8):e1008904.

2. Pende D, Falco M, Vitale M, Cantoni C, Vitale C, Munari E, et al. Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. Front Immunol [Internet]. 2019 [cited 2022 Apr 26];10. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2019.01179

3. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. Genome Res. 2012 Oct 1;22(10):1845–54.

4. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SGE. IPD—the Immuno Polymorphism Database. Nucleic Acids Res. 2013 Jan;41(Database issue):D1234–40.

5. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios [Internet]. bioRxiv; 2021 [cited 2022 Apr 23]. p. 2021.02.06.430068. Available from: https://www.biorxiv.org/content/10.1101/2021.02.06.430068v1

6. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.

8. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 16;10(2):giab008.

9. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012 Feb 15;28(4):593–4.

10. Sadedin SP, Oshlack A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. Genome Biol. 2019 Apr 18;20(1):78.

11. Amorim LM, Augusto DG, Nemat-Gorgani N, Montero-Martin G, Marin WM, Shams H, et al. High-Resolution Characterization of KIR Genes in a Large North American Cohort Reveals Novel Details of Structural and Sequence Diversity. Front Immunol. 2021 May 7;12:674778.

# Chapter 4: High-throughput complement component 4 genomic sequence analysis with C4Investigator


## Abstract

The complement component 4 (*C4*) gene locus, composed of the *C4A* and *C4B* genes and located on chromosome 6, encodes for C4 protein, a key intermediate in the classical and lectin pathways of the complement system. The complement system is an important modulator of immune system activity and is also involved in the clearance of immune complexes and cellular debris. The *C4* gene locus exhibits copy number variation (CNV), with each composite gene varying between 0-4 copies per haplotype, *C4* genes also vary in size depending on the presence of the HERV retrovirus in intron 9, denoted by *C4(L)* for long-form and *C4(S)* for short-form, which modulates expression and is found in both *C4A* and *C4B*. Additionally, human blood group antigens Rodgers (Rg) and Chido (Ch) are located on the C4 protein, with the Rg epitope generally found on C4A protein, and the Ch epitope generally found on C4B protein. C4 CNV has been implicated in numerous autoimmune and pathogenic diseases. Despite the central role of C4 in immune function and regulation, high-throughput genomic sequence analysis of *C4* variants has been impeded by the high degree of sequence similarity and complex genetic variation exhibited by these genes. To investigate *C4* variation using genomic sequencing data, we have developed a novel bioinformatic pipeline for comprehensive, high-throughput characterization of human *C4* sequence from short-read sequencing data, named C4Investigator. Using paired-end targeted or whole genome sequence data as input, C4Investigator determines gene copy number for overall *C4*, *C4A*, *C4B*, *C4(Rg)*, *C4(Ch)*, *C4(L)*, and *C4(S)*, additionally, C4Ivestigator reports the full overall *C4* aligned sequence, enabling nucleotide level analysis of *C4*. To demonstrate the utility of this workflow we have analyzed *C4* variation in the 1000 Genomes Project Dataset (1KGP), showing that the *C4* genes are highly poly-allelic with many variants that have the potential to impact C4 protein function.

## Introduction

The *C4* gene locus, composed of the *C4A* and *C4B* genes and located in human chromosomal region 6p21.33, encodes for complement component 4 (C4) protein, a key intermediate in the classical and lectin pathways of the complement system(1). The complement system is an important modulator of immune system activity, can activate the innate and adaptive immune response systems(2–4) and is also involved in the clearance of immune complexes and cellular debris. The *C4* gene locus exhibits copy number variation (CNV), with each composite gene varying between 0-4 copies per haplotype, and importantly, the gene copy number of *C4A* and *C4B* correlate to C4 protein levels(5). *C4* genes also vary in size depending on the presence of the HERV-K(C4) retrovirus in intron 9 (**Figure 3.1A**), denoted by *C4(L)* for long-form and *C4(S)* for short-form, which modulates expression and is found in both *C4A* and *C4B* resulting in four distinct genomic forms of *C4* (*C4A(L)*, *C4B(L)*, *C4A(S)*, and *C4B(S)*)(5).

C4 is mainly expressed by liver cells, white blood cells, and intestinal epithelial cells(6), but also by central nervous system cells(7). C4 is expressed as two isotypes, C4A and C4B, encoded by the *C4A* and *C4B* genes, respectively. The isotypes have nearly identical sequence but are differentiated by a short peptide sequence motif at positions 1120-1125 (**Figure 4.1B**), which are **PC**PV**LD** for C4A and **LS**PV**IH** for C4B. Additionally, human blood group antigens Rodgers (Rg) and Chido (Ch) are located on the C4 protein at positions 1207-1210(8–10). The Rg epitope is generally found on C4A protein, and the Ch epitope is generally found on C4B protein. The relative locations of the C4A/B specific single nucleotide polymorphisms (SNPs) and the Rg/Ch major epitope encoding SNPs are shown in **Figure 4.1A**.

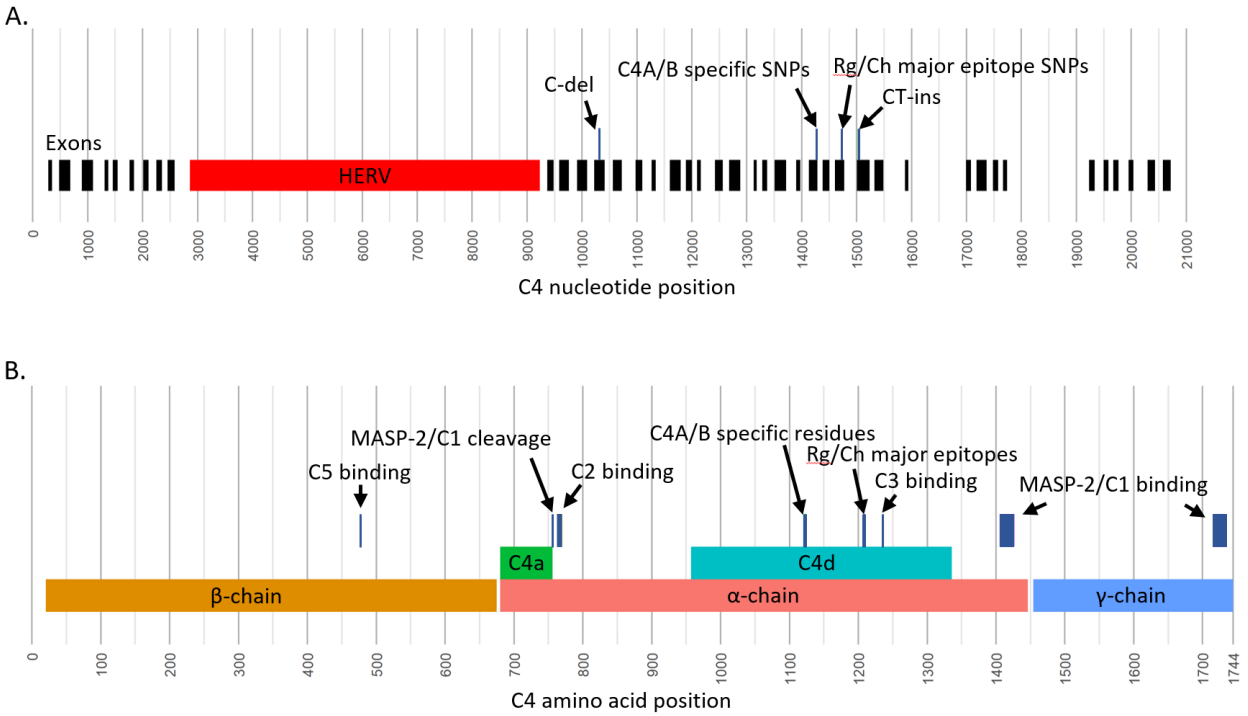


**Figure 4.1. Sequence features of C4 genes and C4 proteins.**
**(A)** Positions of *C4A* and *C4B* genomic sequence features shown for a long-form of *C4*. Exon positions are marked in black, the HERV retroviral sequence is marked in red, and select sequence variants are shown above the exons. Positions are based on the *C4* alignment reference, which includes 5'UTR and 3'UTR sequence. The C-del variant and the CT-ins variant are frame-shift mutations that result in premature terminations. **(B)** Positions of C4A/C4B protein sequence features. The major chains, α, β, and γ, are shown in the bottom row, the cleavage products, C4a and C4d, are shown on the middle row, and important binding locations and sequence variants are shown in the top row. The amino acid positions include the leading 19 amino acid signal peptide.

*C4* CNV has been implicated in the neurological diseases schizophrenia(11,12) and Alzheimer's(13), and there is a large body of evidence connecting *C4A* deficiency and the development of systemic lupus erythematosus (SLE)(14–16), an autoimmune disease. Additionally, while the role of *C4* CNV has yet to be studied in the context of COVID-19 pathology, recent studies have implicated complement hyperactivation with severe SARS-CoV-2 complications(17–19).

Currently, interrogation of *C4* CNV is accomplished through digital droplet polymerase chain reaction (ddPCR)(11,20), which is capable of quantifying gene copy number for overall *C4*, *C4A(L)*, *C4A(S)*, *C4B(L)* and *C4B(S)*. While this method produces accurate results for *C4A* and *C4B* gene copy number and

phasing with long and short form, it is intractable for identifying additional sequence variation at scale, including loss of function mutations(21,22) and recombinations(23,24), and is completely blind to novel sequence variation. High-throughput genomic sequence analysis of *C4* variants has been impeded by the complex genetic variation exhibited by these genes. One recent tool for assessing *C4* sequence variation is the *C4A*/*B* analysis workflow hosted on Terra (25), which was developed using the Genome STRiP software (26) to analyze *C4* from WGS data. However, this tool is currently unpublished and is restricted to analysis of copy number variation of *C4A*/*C4B* specific SNPs and the HERV retrovirus.

Most C4 analysis workflows are targeted at characterizing the region of *C4A*/*C4B* specific SNPs, which encode for an important active site that causes C4A and C4B to have unique biochemistries. However, there are many other vital locations along C4 sequence that when mutated have drastic functional consequences (**Figure 4.1B**). First are amino acid positions 477 and 478; mutations at these positions can disrupt C5 convertase activity (27,28), an important step in the classical and lectin complement cascade pathways that results in the formation of the membrane attack complex (MAC). Positions 756 and 757 are the site of C1/MASP-2 cleavage(29) to produce C4a and C4b, which is the initial modification made to C4 to initiate the complement cascade. Positions 1405-1427 and 1716-1732 are binding sites for C1/MASP-2 (30,31). Positions 763-770 make up a binding site for C2a (32), an intermediary of the classical and lectin cascade pathways that binds with C4b to make a C3 convertase. Positions 1236 and 1238 are known binding positions for C3b (33), an intermediary that binds with the C4b·C2a complex to make a C5 convertase. Finally, there are known frame-shift mutations on exon 13 and 29 that both result in premature terminations (**Figure 4.1A**) (22).

Due to the importance of C4 in complement cascade activity, coupled with the high degree of allotypic variation (34,35), we believe that full genomic sequence characterization of *C4* is of vital importance to advancing our understanding of its in human health. To investigate *C4* variation using genomic sequencing data, we have developed a bioinformatic pipeline for comprehensive, high-throughput

characterization of human *C4* copy number and sequence variation from short-read sequencing data, named C4Investigator. Using whole genome sequence data as input, C4Investigator determines gene copy number for overall *C4*, *C4A*, *C4B*, *C4(Rg)*, *C4(Ch)*, *C4(L)*, and *C4(S)*; additionally, C4Investigator reports full genomic sequence and highlights frame-shift mutations and potential recombinations.

To demonstrate the utility of C4Investigator, we have applied the workflow to the Thousand Genomes Project high depth 30x WGS data(36,37), a dataset consisting of 3,199 samples, characterizing *C4* copy number and sequence variation for the first time in this dataset to provide a snapshot of population-level differentiation at this important genomic region.

## Materials and methods

### C4Investigator workflow overview

Due to the high degree of sequence similarity between *C4A* and *C4B*, the C4Investigator workflow combines alignments of these two genes into an overall *C4* alignment. A long-form *C4A* sequence and a short-form *C4B* sequence are used as a reference for this alignment. A custom alignment processing workflow, similar to that outlined in Marin et al.(38), was developed to integrate the *C4A* and *C4B* alignments into the overall *C4* alignment. From the overall alignment, *C4* copy number is determined by comparing the median alignment depth across *C4* to the average depth of the Tenascin XB (*TNXB)* gene, a nearby copy-stable gene. Gene copy number of *C4A*, *C4B*, *C4(Ch)*, *C4(Rg)*, *C4(L)* and *C4(S)* are determined by multiplying the ratios of *C4A*/*B* specific SNPs, *Rg*/*Ch* specific SNPs and the HERV insertion region, to the overall *C4* copy. *C4A*-*Ch* and *C4B*-*Rg* recombinants are identified using read-based phasing. A limitation of this approach is that because of the genomic distance between the *C4A*/*B* specific SNPs to the HERV region, this method is unable to phase *C4A*/*B* with long and short-form.

In addition to gene copy number analysis, C4Investigator outputs the full overall *C4* aligned sequence as a SNP table.

The pipeline is available at: https://github.com/wesleymarin/C4Investigator

## C4 alignment workflow

The structural variation of the *C4* gene locus and high-degree of sequence similarity between *C4A* and *C4B* necessitates a custom alignment and processing workflow. The first step of the workflow is a Bowtie2(39) alignment to a reference consisting of a short-form of *C4B*, the long-form of *C4A*, and *TNXB*, which is used as a close proximity normalizer gene. Subsequently, the reads aligned to both *C4A* and *C4B* are combined, formatted, and indexed according to the aligned read formatting procedure outlined in Marin et al. (2021) to generate an overall *C4* alignment used for downstream analysis. The output of this workflow is a *C4* depth table spanning from position -285 5'UTR to position 341 3'UTR with depths marked independently for A, T, C, G, deletions, and insertions.

## Copy number determination

The median depth of the overall *C4* alignment is normalized by the median depth of *TNXB* to determine the overall *C4* gene copy number. The relative depth ratios of the *C4A* and *C4B* specific SNPs, at positions E26.129, E26.132, E26.140, E26.143, and E26.145, are multiplied by the overall *C4* gene copy number to determine the *C4A* and *C4B* gene copy number. Similarly, the *Rg* and *Ch* major epitope specific SNPs, at positions E28.111, E28.116, E28.125, and E28.126, are processed to determine the *C4(Rg)* and *C4(Ch)* gene copy number. Finally, the depth ratio of the *HERV* insertion, across positions I9.276-I9.6642, is multiplied by the overall *C4* gene copy number to determine the long-form and short-form copy number.

Exon 29 TC insertion sequence depth ratio is multiplied by the overall *C4* copy to determine the copy of loss of function alleles, this value is subtracted from *C4A* gene copy number to give the functional *C4A* copy number. While it is possible for the TC insertion to exist in a *C4B* sequence, this variant is very

rare(40) and there is no solid evidence of it in the datasets we analyzed. A similar approach is utilized for the exon 13 C deletion in *C4B* to give the functional *C4B* copy number.

## Sequence analysis

The overall *C4* depth table is processed to generate a SNP table for positions passing a minimum depth threshold (6 for whole genome sequence data). Heterozygous positions are identified using a depth ratio of 0.5 normalized by the determined *C4* gene copy number. The output of this step is an overall *C4* SNP table with combined sequence for *C4A* and *C4B*.

## Variant phasing

One of the major challenges of interpreting *C4* genomic sequencing data is phasing variants. The high sequence similarity between *C4A* and *C4B*, coupled with the high variability of copy number of C4 overall, make phasing variants with *C4A* and *C4B* extremely challenging. In C4Investigator we have implemented a phasing algorithm utilizing paired-end reads to phase variants near the *C4A* and *C4B* specific SNPs. Importantly, this approach covers the *Rg* and *Ch* major epitope specific SNPs, which are 440bp apart from the *C4A*/*B* specific SNPs, facilitating phasing between *C4A*/*C4B* and *Rg*/*Ch* and identification of the *C4A*-*Ch* and *C4B*-*Rg* recombinants.

## Targeted sequencing dataset generation

To validate the C4Investigator workflow, we applied targeted-capture next-generation sequencing (NGS) in a cohort of 38 African Americans and 37 European Americans from the United States. These healthy individuals were unrelated and part of the INDIGO (The Immunogenetics for Neurological DIseases working GrOup) cohort.

A total of 100 ng of high-quality DNA is fragmented using the Twist EF Kit 2.0 l (Twist Bioscience), incubating for 5 minutes at 37 °C. Subsequently, the fragmented DNA have their ends repaired, poly-A tail added, and are ligated through PCR to Illumina compatible dual index adapters uniquely barcoded.

After ligation, fragments are purified with 0.8X ratio Ampure XP magnetic beads (Beckman Coulter) followed by double size selection (0.42X and 0.15X ratios) to select libraries of approximately 800 bp. Finally, libraries are amplified and purified with magnetic beads. After quantification by quantitative PCR, 60 ng of each sample are precisely pooled using ultrasonic acoustic energy, and the enrichment targeted capture is performed with hybridization kits from Twist Bioscience. Briefly, the libraries are bound to 33,620 biotinylated 120 bp probes target the entire MHC (chr6:28525013-33457522, hg38). By using streptavidin magnetic beads, the targeted fragments are captured and then amplified and purified. Enriched libraries are analyzed in BioAnalyzer (Agilent) and quantified by digital-droplet PCR. Finally, enriched libraries are sequenced using NovaSeq6000 (Illumina) with paired-end 150bp sequencing protocol.

C4Investigator was run over both targeted sequencing datasets using a minimum depth of 20 for variant calling and a ratio of 0.50, normalized by the total copy of *C4*, for heterozygous position identification. C4Investigator results were compared to ddPCR results to provide validation for C4 interpretation from targeted sequence data.

### ddPCR genotyping

Gene copy number for *C4A*, *C4B*, *C4(L)* and *C4(S)* were determined by ddPCR as described previously(11) for 38 samples of African ancestry and 37 samples of European ancestry to provide a copy determination comparison.

Gene copy number results determined by C4Investigator were compared to ddPCR determined results to quantify the copies of *C4A*, *C4B*, *C4(L)* and *C4(S)* that were identified by both methods.

### Thousand Genome Project analysis

Reads aligned to *C4* and the nearby region were extracted from GRCh38 aligned CRAM files using the coordinates outlined in **Table S4.1** using Samtools(41). The extracted reads were converted to paired-

end FASTQ files using Bazam(42). C4Investigator was run over the paired-end fastq files using a minimum depth of 6 for variant calling and a ratio of 0.50, normalized by the total copy of *C4*, for heterozygous position identification. *C4* copy number results were stratified by superpopulation. Population totals and abbreviations are outlined in **Table 4.1**.

**Table 4.1. 1KGP population abbreviations and size.**
Superpopulations are written in bold and the total samples for superpopulations are the sums of the component populations.

| Population | N |
|---|---|
| **European (EUR)** | **633** |
| British in England and Scotland (GBR) | 91 |
| Finnish in Finland (FIN) | 99 |
| Iberian population in Spain (IBS) | 157 |
| Utah Residents with Northern and Western European ancestry (CEU) | 179 |
| Toscani in Italia (TSI) | 107 |
| | |
| **East Asian (EAS)** | **582** |
| Southern Han Chinese (CHS) | 161 |
| Chinese Dai in Xishuanagbanna, China (CDX) | 92 |
| Kinh in Ho Chi Minh City, Vietnam (KHV) | 122 |
| Han Chinese in Bejing, China (CHB) | 103 |
| Japanese in Tokyo, Japan (JPT) | 104 |
| | |
| **Admixed American (AMR)** | **490** |
| Puerto Rican from Puerto Rica (PUR) | 139 |
| Colombian from Medellian, Colombia (CLM) | 132 |
| Peruvian from Lima, Peru (PEL) | 122 |
| Mexican Ancestry from Los Angeles USA (MXL) | 97 |
| | |
| **South Asian (SAS)** | **601** |
| Punjabi from Lahore, Pakistan (PJL) | 146 |
| Bengali from Bangladesh (BEB) | 131 |
| Sri Lankan Tamil from the UK (STU) | 114 |
| Indian Telugu from the UK (ITU) | 107 |
| Gujarati Indian from Houston, Texas (GIH) | 103 |
| | |
| **African (AFR)** | **893** |
| African Carribean in Barbados (ACB) | 116 |
| Mandinka in The Gambia (GWD) | 178 |
| Esan in Nigera (ESN) | 149 |
| Mende in Sierra Leone (MSL) | 99 |

| Population | N |
|---|---|
| Yoruba in Ibadan, Nigera (YRI) | 178 |
| Luhya in Webuye, Kenya (LWK) | 99 |
| American's of African Ancestry in SW USA (ASW) | 74 |

## Validation

C4Investigator performance was validated against ddPCR results for divergent populations. The first dataset consisted of 37 samples and was of European ancestry, the second dataset consisted of 38 samples and was of African ancestry. Both datasets were generated using targeted sequencing. Copy number results were compared for *C4A*, *C4B*, *C4(S)*, and *C4(L)* for each dataset.

C4Investigator copy number results for the 1KGP dataset were compared to results from the *C4A*/*B* analysis workflow utilizing Genome STRiP(36) implemented in Terra (25). Results were compared across overall *C4*, *C4A*, *C4B*, *C4(L)* and *C4(S)* results. For overall *C4* all results across both datasets were compared. For *C4A* and *C4B* comparison, samples marked as *C4A1*, *C4A2*, *C4B1*, or *C4R1,* which represented rare *C4* sequence variants, by the Genome STRiP Terra workflow were excluded, this excluded a total of 55 samples from comparison. For *C4(L)* and *C4(S)* all results were compared. *C4A1*, *C4A2*, *C4B1*, and *C4R1* results for C4Investigator were generated by confirming correct phase across positions E26.128 – E26.145, based on the k-mers provided for these variants by the Terra workflow, then determining the copy number of these variants based on the relative SNP depth.

# Results

## Performance evaluation – comparison to ddPCR

**Table 4.2. Evaluation of C4Investigator copy number determination performance compared to ddPCR for European and African datasets.**
*C4(S)* = *C4* short-form, *C4(L)* = *C4* long-form.

| Ancestry | *C4A* | *C4B* | *C4(S)* | *C4(L)* |
|---|---|---|---|---|
| African | 1.00 N=76 | 1.00 N=66 | 0.89 N=61 | 0.91 N=81 |
| European | 1.00 N=82 | 1.00 N=70 | 0.94 N=34 | 0.98 N=118 |

Evaluation of C4Investigator copy number determination performance compared to ddPCR results for European and African datasets show perfect concordance between the two methods for *C4A* and *C4B* copy number determination (**Table 4.2**), 94% for *C4(S)* and 98% for *C4(L)* for the European dataset, and 89% for *C4(S)* and 91% for *C4(L)* for the African dataset.

### Performance evaluation – comparison to *C4A*/*B* Terra

To benchmark C4Investigator performance against another bioinformatic workflow, we compared results for the 1000 Genomes Project dataset (N=3199) against results from the unpublished *C4A/B* Terra workflow(25), a bioinformatic pipeline that utilizes Genome STRiP(36) to quantify *C4* copy number.

Overall *C4* copy determination performance was highly concordant with the *C4A*/*B* Terra workflow, at 99.95% (N=12977). *C4A* and *C4B* copy identification concordance was 99.12% (N=6942) for *C4A* and 98.96% (N=5976) for *C4B*. *C4(L)* and *C4(S)* copy identification concordance was 99.60% (N=8700). Comparing the additional *C4* variants quantified by *C4A*/*B* Terra workflow showed an overall concordance of 96.6% (N=59).

Investigation into the discordant *C4A* and *C4B* samples showed the ratios of *C4A* were near the copy thresholds for both methods (**Figure S4.1A**), further examination into the *C4A*/*B* Terra k-mer quality scores showed the discordant samples had a median quality of 9, while concordant samples had a median quality of 62.7 (**Figure S4.1B**). A similar analysis was performed for the *C4(L)* and *C4(S)* discordant samples, which showed the C4Investigator ratios were near the copy thresholds, while the *C4A*/*B* Terra workflow ratios were clustered near the center of the copy intervals (**Figure S4.2**).

## 1000 Genomes Project – copy number analysis

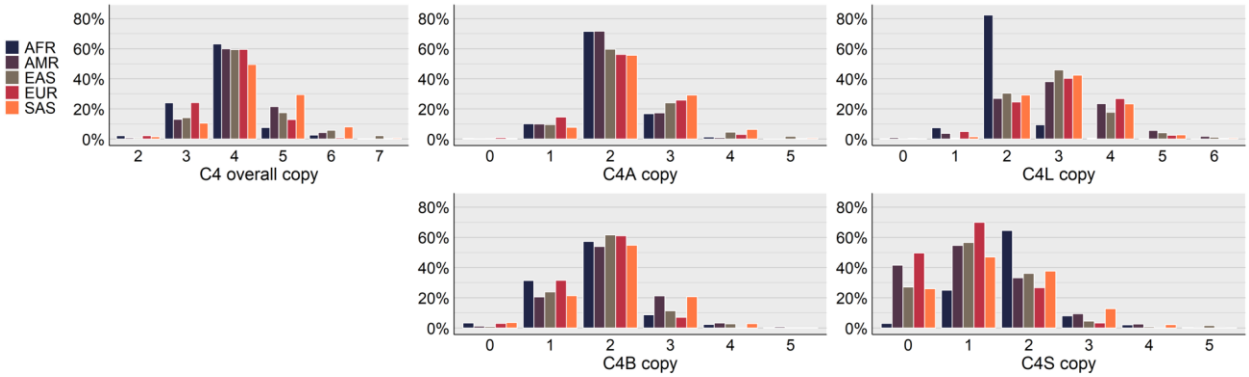


**Figure 4.2. Superpopulation distributions of *C4* copy number results for the 1KGP dataset.**
*C4* overall copy represents the total copy number of *C4A* and *C4B*, *C4S* represents the total copy number for the short-forms of *C4A* and *C4B*, and *C4L* represents the total copy number for the long-forms of *C4A* and *C4B*. AFR = African, AMR = Admixed American, EAS = East Asian, EUR = European, SAS = South Asian.

Analysis of *C4* copy number variation across superpopulations showed most individuals across all superpopulations had 4 copies of *C4* overall, 2 copies of *C4A,* and 2 copies of *C4B*, and there were very few individuals with 0 copies of *C4A* or *C4B* (**Figure 4.2**). Outside of these similarities there were stark differences observed between the superpopulations. The African (AFR) and European (EUR) superpopulations had much higher occurrences of 3 overall copies of *C4*, almost double that observed in the other superpopulations, and much lower occurrences of 5 and 6 overall copies of *C4*. In contract, the South Asian (SAS) superpopulation had the lowest occurrence of 3 overall copies of *C4*, but the highest of 5 and 6. One of the largest differences observed was with *C4L* copy 2 for the AFR superpopulation, which was observed at over double the rate of the other superpopulations; this superpopulation also had substantially lower *C4L* copy 3 occurrence and virtually no occurrence of 4 copies. The *C4S* copy 0 occurrence for the AFR superpopulation was negligible, while other superpopulations were over 20%.

## 1000 Genomes Project – SNP analysis

The SNP tables output by C4Investigator, which represent combined *C4A* and *C4B* sequence, were parsed to identify sequence variation, and any identified exonic nucleotide variants are evaluated for amino acid coding change. From these results we have summarized non-synonymous mutations in **Table**

**4.3**, and SNP variation that is not represented in the main assembly of the GRCh38 reference in **Figure 4.3**.

**Table 4.3. Population specific minor allele frequencies for *C4A* and *C4B* unphased, non-synonymous exonic sequence variants.**
For this analysis we did not distinguish between *C4A* and *C4B*. This table shows amino acid frequencies, the amino acid position and nucleotide position, the nucleotide frequencies, and population allele frequencies for the minor allele. Major amino acids and nucleotides represent the most frequent variant in most populations while minor amino acids and nucleotides represent the second most frequent variant. This data was filtered to only show variants with allele frequencies >= 2% for any population. Blank values represent absence of the variant. See **Table 4.1** for population abbreviations.

| major | minor | aa | nuc | major | minor | EUR | | | | | EAS | | | | | AMR | | | | SAS | | | | | AFR | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aa | aa | pos | pos | nuc | nuc | GBR | FIN | IBS | CEU | TSI | CHS | CDX | KHV | CHB | JPT | PUR | CLM | PEL | MXL | PJL | BEB | STU | ITU | GIH | ACB | GWD | ESN | MSL | YRI | LWK | ASW |
| L | V | 141 | E3_157 | C | G | 6.1 | 8.4 | 3.2 | 6.1 | 4.1 | 35.8 | 53.3 | 33.7 | 25.8 | 13.3 | 8.7 | 7.7 | 16 | 15.8 | 4 | 16.1 | 8 | 8.8 | 6.3 | 7 | 5.1 | 10.6 | 3.5 | 10.5 | 10.7 | 10.7 |
| T | I | 229 | E6_60 | C | T | | | | | | | | 2.1 | | | | | | | | | | | | | | | | | | |
| K | M | 325 | E9_62 | A | T | | | | | | | | 4.4 | 2.9 | 4.4 | | | | | | | | | | | | | | | | |
| M | I | 328 | E9_72 | G | A | | | | | | | | 4 | 2.9 | 4.2 | | | | | | | | | | | | | | | | |
| P | L | 478 | E12_92 | C | T | | | | | | | | | | | | | | | | | | | | 3.5 | 3.8 | 4.7 | | 4.1 | | |
| H | P | 549 | E13_122 | A | C | 2.6 | 5.9 | 2.1 | 5.5 | 4.1 | 14.3 | 18.7 | 10.2 | 12.8 | 8 | 3 | 3.9 | 6.8 | 7 | 4 | 3.8 | 4 | 5.5 | 3.1 | | | | | | | 2.5 |
| P | L | 726 | E17_106 | C | T | | | | | | | | | | | | | | | | | | | | | | | | | 3.1 | |
| R | H | 791 | E18_103 | G | A | | | | | | | | | | | | | | | | | | | | 2.2 | | | | | | |
| R | Q | 916 | E21_155 | G | A | | | | | | | | | | | | | | | | | | | | | 3.1 | | 4.3 | | 2.1 | |
| E | D | 959 | E23_23 | A | C | | 2.2 | | | | 8.6 | 17.7 | 13 | 4.6 | | | | 2.7 | | | 7.4 | | 2.9 | | | 2.3 | 2.3 | | 2.4 | | |
| A | S | 1286 | E29_180 | G | T | 7.8 | 4.5 | 4.6 | 8.2 | 3.8 | 3.6 | | | 5.6 | 13.1 | 3.1 | 2.4 | | 2.2 | 5.7 | 4.8 | 4.6 | 6.4 | 5.9 | | | | | | | |
| A | P | 1413 | E33_6 | G | C | | | | | | | | | | | | | | | | | | | | 2.2 | 4.5 | | 10.2 | 4 | 2.9 | 2.1 |
| P | S | 1530 | E36_4 | C | T | | | | | | | | | | | | | | | | | | | | 2.2 | | | 10.2 | 4 | 3.7 | 2.8 |

Analysis of allele frequencies for *C4A* and *C4B* non-synonymous exonic sequence variation showed large variations in frequencies across populations (**Table 4.3**). The variant p.H549P was very common in the EAS superpopulation, and was found in most populations, but very rare in the AFR superpopulation. The variant p.L141V was the major allele in the CDX population, was highly frequent across the EAS superpopulation, and was found at appreciable frequencies across all populations. The variants p.T229I, p.K325M, and p.M328I were only found in the EAS superpopulation. And the variants p.P478L, p.P726L, p.R791H, p.R916Q, p.A1413P, and p.P1530S were only found in the AFR superpopulation.

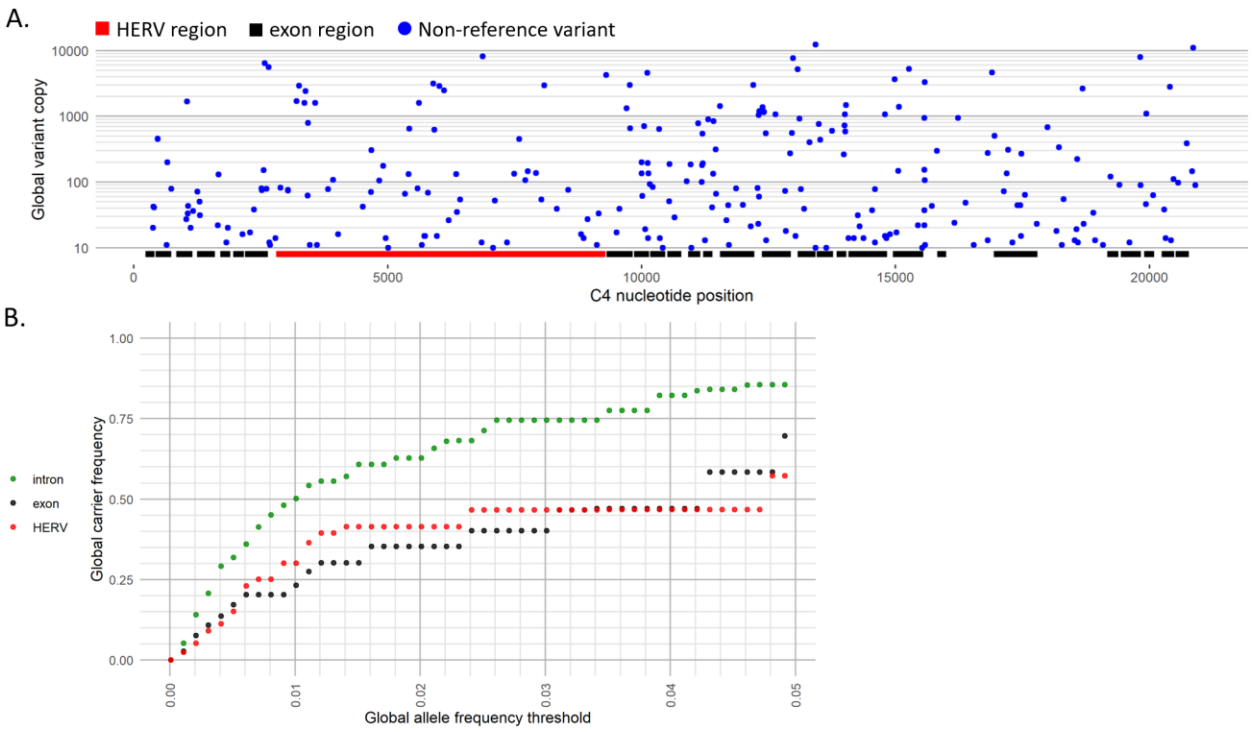


**Figure 4.3. SNV variation across the 1KGP dataset.**
**(A)** Total copy of combined *C4A* and *C4B* non-reference variants, which are variants not represented in the main assembly of GRCh38, by *C4* position for the 1KGP dataset. The copy number of all non-reference variants for a position across the 1KGP dataset are summed to get the non-reference variant copy, which was then filtered to only show variant positions with total copy of at least 10. Positions of *C4* exon and HERV regions are marked. **(B)** Global carrier frequencies for non-reference variants in the 1KGP dataset for increasing global allele frequency thresholds from 0.00-0.05 for introns, exons, and the HERV region. The y-axis represents the total proportion of carriers that carry a non-reference allele that is at or below the global allele frequency threshold on the x-axis. For example, nearly 25% of the 1KGP dataset carried exonic variants with a global allele frequency of 1% or lower.

An analysis into non-reference SNVs, which are variants not represented in the main assembly of GRCh38, for the 1KGP dataset across *C4A* and *C4B* showed 251 variant positions with total non-reference variant copy of at least 10 (**Figure 4.3A, Table S4.2**). Examination of the positional distribution of these variants across *C4A* and *C4B* showed 50 exonic variant positions accounting for 0.955% of all exonic positions (N=5235), 138 intronic variant positions accounting for 1.56% of all intronic positions (N=8831, exclusive of HERV), and 59 HERV variant positions accounting for 0.927% of all HERV positions (N=6367).

An examination of the proportion of the 1KGP dataset that carry rare variants showed that almost 25% of the samples carried exonic variants with global allele frequencies at or below 1% (**Figure 4.3B**, **Table S4.3**), and about 50% carried intronic variants. Looking at the carrier distribution of more common variants showed that about 70% of the samples carried exonic variants with global allele frequencies below 5%, and about 85% carried intronic variants.

## 1000 Genomes Project – recombinant analysis

**Table 4.4.** *C4ACh* **and** *C4BRg* **carrier frequencies by population.**
Carrier frequencies were calculated by the total *C4A* and *C4B* carrier count per population. *C4A-Ch* = *C4A-Chido*, *C4B-Rg* = *C4B-Rodger*. See **Table 4.1** for population abbreviations.

| | EUR | | | | | EAS | | | | | AMR | | | | SAS | | | | | AFR | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GBR | FIN | IBS | CEU | TSI | CHS | CDX | KHV | CHB | JPT | PUR | CLM | PEL | MXL | PJL | BEB | STU | ITU | GIH | ACB | GWD | ESN | MSL | YRI | LWK | ASW |
| *C4A-Ch* | 1.1 | 1.0 | 1.9 | 1.1 | 0 | 0.6 | 2.2 | 4.9 | 2.9 | 4.8 | 5.8 | 4.5 | 7.4 | 5.2 | 2.1 | 0 | 0.9 | 0.9 | 1.0 | 11.2 | 20.2 | 8.1 | 37.4 | 20.2 | 14.1 | 13.5 |
| *C4B-Rg* | 0 | 0 | 6.4 | 1.7 | 3.7 | 0 | 1.1 | 0.8 | 1.9 | 0 | 5.0 | 6.8 | 3.3 | 5.2 | 4.1 | 1.5 | 7.0 | 4.7 | 4.9 | 0 | 4.5 | 0 | 0 | 0 | 0 | 2.7 |
| N | 91 | 99 | 157 | 179 | 107 | 161 | 92 | 122 | 103 | 104 | 139 | 132 | 122 | 97 | 146 | 131 | 114 | 107 | 103 | 116 | 178 | 149 | 99 | 178 | 99 | 74 |

Analysis of carrier frequencies for *C4A*/*C4B* and Rodger/Chido recombinants, *C4A-Ch* and *C4B-Rg*, showed higher overall frequencies of the *C4A-Ch* recombinant compared to *C4B-Rg* (**Table 4.4**). The *C4A-Ch* recombinant was highly prominent in the AFR superpopulation, with a 37.4% carrier frequency in the MSL population, 20% in GWD and YRI, 14.1% in LWK, 13.5% in ASW, 11.2% in ACB, and 8.1% in ESN. The AMR superpopulation also showed appreciable *C4A-Ch* carrier frequencies, the highest being the PEL population at 7.4%, followed by PUR at 5.8%, MXL at 5.2% and CLM at 4.5%. While carrier frequencies of the *C4B-Rg* recombinant were generally lower overall, with many populations showing no carriers, the frequencies of this recombinant were not negligible, with 8 of the populations displaying at least 4.5% carrier frequency. The AMR and SAS superpopulations showed the highest frequencies of the *C4B-Rg* recombinant, the highest being the STU population at 7.0%, followed by CLM at 6.8%.

## Performance evaluation – *C4A*/*C4B* and Rodger/Chido phasing

Phasing completeness between the *C4A*/*C4B* specific SNP group and the *Rg*/*Ch* specific SNP group was estimated by comparing the number of samples with read-backed phasing for the non-recombinant

variants, *C4A-Rg* and *C4B-Ch,* to the total number of samples carrying *C4A-Rg* and *C4B-Ch*, respectively. Phasing completeness for *C4A-Rg* was 97.69% (N=3167) and *C4B-Ch* was 96.60% (N=3113).

## Discussion

Comparison of C4Investigator *C4* copy number determination to ddPCR results showed high concordance between the two methods for *C4A* and *C4B* copy number determination across divergent populations (**Table 4.2**). *C4(L)* and *C4(S)* copy determination performance was acceptable for the European dataset, but poor for the African dataset.

Comparison of C4Investigator to the *C4A/B* Terra workflow, another bioinformatic pipeline, on the 1KGP WGS dataset showed high concordance between the two workflows, especially for overall *C4* copy. An investigation into discordant *C4A/B* results showed that the discordant samples had lower base quality scores on average (**Figure S4.1B**), with neither method showing clear copy number results for the discordant samples (**Figure S4.1A**). In contrast, the investigation into discordant HERV results showed a marked difference between the two methods, with the *C4A/B* Terra workflow showing clear copy numbers for these samples while C4Investigator had unclear determinations (**Figure S4.2**). This is likely due to the additional structural variant processing of the *C4A/B* Terra workflow, which incorporates Genome STRiP (36), a workflow specifically developed for identifying copy number variation in WGS data. The *C4A/B* Terra is strictly focused on identifying copy number variation, a task that it appears to perform very well. In contrast, C4Investigator takes a different approach, focusing on identifying nucleotide variants in a copy variable system through the utilization of custom alignment processing algorithms, which has enabled the identification and quantification of SNP variation across the *C4* genes.

An analysis into C4 copy number variation between superpopulations (**Figure 4.2**) demonstrated some specific patterns, such as a median overall *C4* copy number of 4, and a median copy number for *C4A* and *C4B* of 2 each, but also important distinctions between populations, such as the strikingly high number

of *C4L* copy 2 genotypes in the AFR superpopulation, and the general imbalance between overall *C4* copy of 3 and 5, which was unique for each superpopulation. Differences of this nature might suggest evolutionary pressure or unique genomic makeups that are specific to the different superpopulations and modulate the fitness of different *C4* gene structures.

An essential innovation of C4Investigator is demonstrated by its capacity to reveal important differences in sequence variation between populations, with likely important functional implications. An analysis of non-synonymous exonic sequence variants demonstrated that *C4* sequence makeup can differ greatly between populations, with some variants with seemingly rare global allele frequencies showing high allele frequencies in specific populations. For example, the p.A1413P and the p.P1530S mutations were absent in most populations, but both had 10.2% allele frequency in the MSL population (**Table 3.3**). The fact that both mutations have the same allele frequency raises the question of if these mutations are in-phase, unfortunately, there is a 2046bp gap between these variants which was outside the scope of our phasing approach. However, an examination of the individuals that carried each mutation showed a high overlap, where 28 individuals carried both mutations compared to total 33 individuals carrying the p.A1413P mutation and 31 individuals carrying the p.P1530S mutation. A structural interrogation of C4·MASP-2 binding shows the p.A1413P mutation occurs in the middle of a MASP-2 exosite(31) (**Figure 4.1**), while the change from alanine to proline would not likely change the electrostatic interactions between C4 and MASP-2, it could potentially alter the structure of the binding site. Another sequence variant with potential to impact function is the p.P478L mutation, which causes severe reduction of hemolytic activity by disruption of C5 binding(28). Similar analyses in the context of disease association studies are likely to reveal important insights into immune-mediated pathogenesis.

An analysis into *C4A* and *C4B* non-reference variants demonstrated that the *C4* genes are highly poly-allelic across introns, exons and the HERV region (**Figure 4.3A**). Further examination into rare variant carrier frequencies demonstrated that exonic variants under 5% global allele frequency are carried by

around 70% of the 1KGP samples (**Figure 4.3B**). This analysis demonstrates the value of nucleotide level analysis of *C4*, which reveals important features of genomic variation not otherwise evident with existing methods.

One important aspect of SNP variation identification is the ability to phase variants. However, phasing high-copy variants (gene copy number > 2) is very complex and it is difficult to be certain of phasing completeness due to the high potential for missing information. Due to the high sequence similarity between *C4A* and *C4B*, the alignments must be treated as a single gene, exacerbating the high-copy phasing problem. We have implemented read-backed phasing that enables us to determine whether two variants in proximity are in-phase, but the potential for missing information means in many cases we cannot make the determination that two variants are *not* in-phase; essentially, we can make more confident true positive phasing calls than true negative. Because of the distance between the *C4A*/*C4B* SNPs and the *Rg*/*Ch* SNPs, 440bp, we can determine presence of recombinants between the two SNP groups. An estimate of phasing completeness between *C4A*-*Rg* and *C4B*-*Ch* showed this phasing approach only missed a small percentage of samples. Utilization of this phasing approach to identify *C4A*-*Ch* and *C4B*-*Rg* recombinants showed high *C4A*-*Ch* carrier frequencies across the AFR superpopulation (**Table 4.4**), and appreciable carrier frequencies for the *C4B*-*Rg* recombinant and the AMR and SAS superpopulations.

In conclusion, C4Investigator fills a critical role in the investigation of *C4* variation, processing WGS data to provide *C4* copy number variation and full genomic sequence information. Here, we have demonstrated the utility of this workflow on the Thousand Genomes Project dataset, revealing that *C4* copy number varies between superpopulations, that alleles with low global allele frequencies can have high population specific frequencies, the presence and distribution of *C4* recombinant variants, and population specific carrier frequencies for rare alleles. Additionally, we have demonstrated that C4Investigator can identify *C4* variation that is known to alter C4 function. To the best of our knowledge,

C4Investigator is the only bioinformatic workflow currently available for nucleotide level characterization of *C4* from WGS data, and as such, promises to contribute to our understanding of the role of this genomic region in human health and disease.

## Supplemental figures and tables

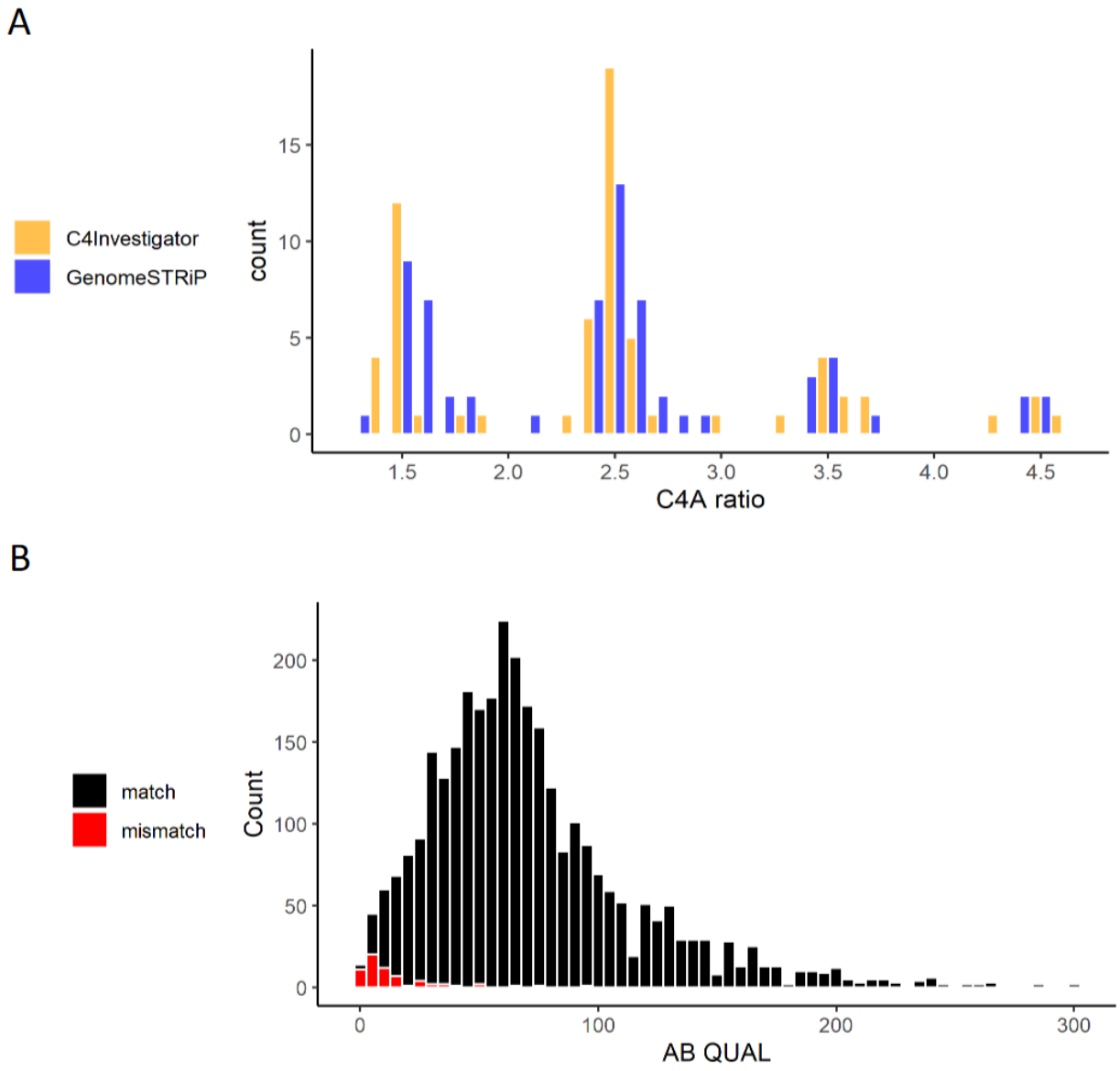


**Figure S4.1. C4Investigator discordance analysis for *C4A/B***
(A) Histogram of normalized ratios for *C4A* read/k-mer count for the C4investigator and Genome STRiP workflows for discordant samples from the 1000 Genomes Project dataset. (B) Histogram of quality scores for the concordant (match) and discordant (mismatch) samples when comparing results for the C4Investigator and Genome STRiP workflows for the 1000 Genomes Project dataset.

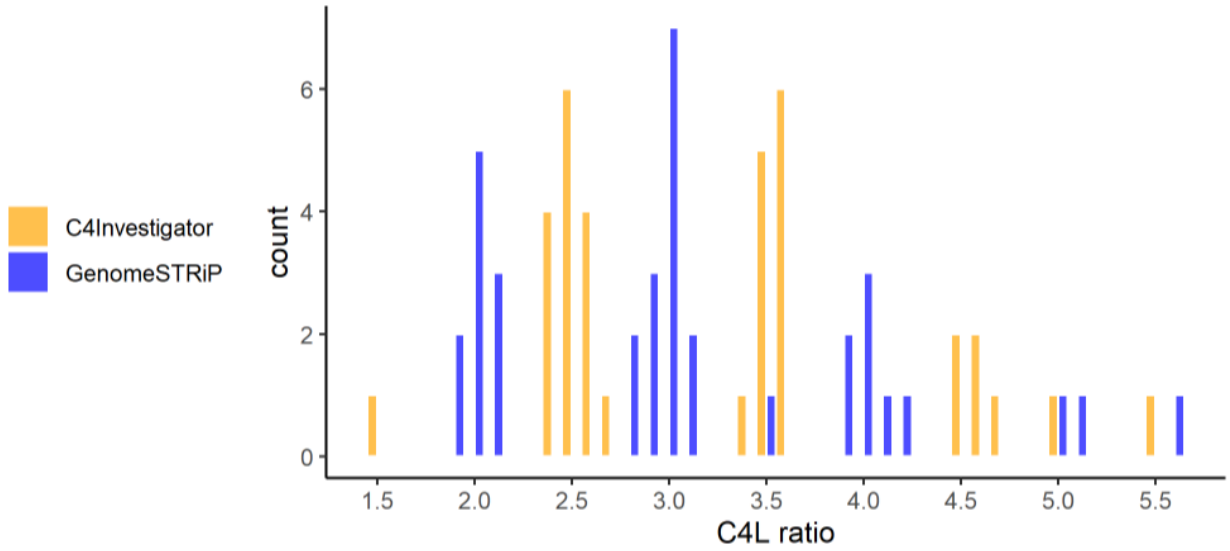


**Figure S4.2. C4Investigator discordance analysis for** *C4(L)/(S).*
Histogram of normalized ratios for *C4(L)* read/k-mer counts for the C4Investigator and GenomeSTRiP workflows for discordant samples from the 1000 Genomes Project dataset.

**S4.1 Table. GRCh38 C4 coordinates.** (CSV)

**S4.2 Table. Global allele count of non-reference** ***C4*****SNVs in the 1KGP dataset.** (CSV)

**S4.3 Table. Global carrier frequencies for uncommon non-reference variants in the 1KGP dataset.**

(CSV)

## References


1. Wang H, Liu M. Complement C4, Infections, and Autoimmune Diseases. Front Immunol [Internet]. 2021 [cited 2022 Apr 28];12. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2021.694928

2. Toapanta FR, Ross TM. Complement-mediated activation of the adaptive immune responses: role of C3d in linking the innate and adaptive immunity. Immunol Res. 2006;36(1–3):197–210.

3. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The complement system and innate immunity. Immunobiol Immune Syst Health Dis 5th Ed [Internet]. 2001 [cited 2022 Jan 4]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK27100/

4. Merle NS, Noe R, Halbwachs-Mecarelli L, Fremeaux-Bacchi V, Roumenina LT. Complement System Part II: Role in Immunity. Front Immunol. 2015;6:257.

5. Yang Y, Chung EK, Zhou B, Blanchong CA, Yu CY, Füst G, et al. Diversity in Intrinsic Strengths of the Human Complement System: Serum C4 Protein Concentrations Correlate with C4 Gene Size and Polygenic Variations, Hemolytic Activities, and Body Mass Index. J Immunol. 2003 Sep 1;171(5):2734–45.

6. Isenman DE. Chapter 17 - C4. In: Barnum S, Schein T, editors. The Complement FactsBook (Second Edition) [Internet]. Academic Press; 2018 [cited 2022 Jan 4]. p. 171–86. (Factsbook). Available from: https://www.sciencedirect.com/science/article/pii/B9780128104200000171

7. Walker DG, Kim SU, McGeer PL. Expression of complement C4 and C9 genes by human astrocytes. Brain Res. 1998 Oct 26;809(1):31–8.

8. Chido/Rodgers Blood Group System. In: Human Blood Groups [Internet]. John Wiley & Sons, Ltd; 2013 [cited 2022 Jan 4]. p. 400–9. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118493595.ch17

9. Mougey R. A review of the Chido/Rodgers blood group. Immunohematology. 2010;26(1):30–8.

10. Mougey R. An update on the Chido/Rodgers blood group system. Immunohematology. 2019 Dec;35(4):135–8.

11. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016 Feb;530(7589):177–83.

12. Woo JJ, Pouget JG, Zai CC, Kennedy JL. The complement system in schizophrenia: where are we now and what's next? Mol Psychiatry. 2020 Jan;25(1):114–30.

13. Zorzetto M, Datturi F, Divizia L, Pistono C, Campo I, De Silvestri A, et al. Complement C4A and C4B Gene Copy Number Study in Alzheimer's Disease Patients. Curr Alzheimer Res. 2017;14(3):303–8.

14. Macedo ACL, Isaac L. Systemic Lupus Erythematosus and Deficiencies of Early Components of the Complement Classical Pathway. Front Immunol. 2016;7:55.

15. Pereira KMC, Perazzio S, Faria AGA, Moreira ES, Santos VC, Grecco M, et al. Impact of C4, C4A and C4B gene copy number variation in the susceptibility, phenotype and progression of systemic lupus erythematosus. Adv Rheumatol. 2019 Aug 6;59(1):36.

16. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, et al. Gene Copy-Number Variation and Associated Polymorphisms of Complement Component C4 in Human Systemic Lupus Erythematosus (SLE): Low Copy Number Is a Risk Factor for and High Copy Number Is a

Protective Factor against SLE Susceptibility in European Americans. Am J Hum Genet. 2007 Jun 1;80(6):1037–54.

17. Afzali B, Noris M, Lambrecht BN, Kemper C. The state of complement in COVID-19. Nat Rev Immunol. 2021 Dec 15;

18. Zinellu A, Mangoni AA. Serum Complement C3 and C4 and COVID-19 Severity and Mortality: A Systematic Review and Meta-Analysis With Meta-Regression. Front Immunol. 2021;12:2184.

19. Savitt AG, Manimala S, White T, Fandaros M, Yin W, Duan H, et al. SARS-CoV-2 Exacerbates COVID-19 Pathology Through Activation of the Complement and Kinin Systems. Front Immunol. 2021 Nov 5;12:767347.

20. Jaimes-Bernal CP, Trujillo M, Márquez FJ, Caruz A. Complement C4 Gene Copy Number Variation Genotyping by High Resolution Melting PCR. Int J Mol Sci. 2020 Aug 31;21(17):6309.

21. Lokki ML, Circolo A, Ahokas P, Rupert KL, Yu CY, Colten HR. Deficiency of Human Complement Protein C4 Due to Identical Frameshift Mutations in the C4A and C4B Genes. J Immunol. 1999 Mar 15;162(6):3687–93.

22. Wu YL, Hauptmann G, Viguier M, Yu CY. Molecular Basis of Complete Complement C4 Deficiency in Two North-African Families with Systemic Lupus Erythematosus (SLE). Genes Immun. 2009 Jul;10(5):433–45.

23. Martínez-Quiles N, Paz-Artal E, Moreno-Pelayo MA, Longás J, Ferre-López S, Rosal M, et al. C4d DNA Sequences of Two Infrequent Human Allotypes (C4A13 AND C4B12) and the Presence of Signal Sequences Enhancing Recombination. J Immunol. 1998 Oct 1;161(7):3438–43.

24. Jaatinen T, Eholuoto M, Laitinen T, Lokki ML. Characterization of a De Novo Conversion in Human Complement C4 Gene Producing a C4B5-Like Protein. J Immunol. 2002 Jun 1;168(11):5652–8.

25. Handsaker RE, Kashin S, Wysoker A, McCarroll SA. Showcase workspace for GenomeSTRiP C4 A/B analysis on the 1000 Genomes WGS data set [Internet]. [cited 2022 Mar 30]. Available from: https://app.terra.bio/#workspaces/mccarroll-genomestrip-terra/C4AB_Analysis

26. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015 Mar;47(3):296–303.

27. Ebanks RO, Jaikaran AS, Carroll MC, Anderson MJ, Campbell RD, Isenman DE. A single arginine to tryptophan interchange at beta-chain residue 458 of human complement component C4 accounts for the defect in classical pathway C5 convertase activity of allotype C4A6. Implications for the location of a C5 binding site in C4. J Immunol. 1992 May 1;148(9):2803–11.

28. McLean RH, Niblack G, Julian B, Wang T, Wyatt R, Phillips JA, et al. Hemolytically inactive C4B complement allotype caused by a proline to leucine mutation in the C5-binding site. J Biol Chem. 1994 Nov;269(44):27727–31.

29. Rossi V, Teillet F, Thielens NM, Bally I, Arlaud GJ. Functional Characterization of Complement Proteases C1s/Mannan-binding Lectin-associated Serine Protease-2 (MASP-2) Chimeras Reveals the Higher C4 Recognition Efficacy of the MASP-2 Complement Control Protein Modules *. J Biol Chem. 2005 Dec 23;280(51):41811–8.

30. Perry AJ, Wijeyewickrema LC, Wilmann PG, Gunzburg MJ, D'Andrea L, Irving JA, et al. A Molecular Switch Governs the Interaction between the Human Complement Protease C1s and Its Substrate, Complement C4. J Biol Chem. 2013 May 31;288(22):15821–9.

31. Kidmose RT, Laursen NS, Dobó J, Kjaer TR, Sirotkina S, Yatime L, et al. Structural basis for activation of the complement system by component C4 cleavage. Proc Natl Acad Sci. 2012 Sep 18;109(38):15425–30.

32. Pan Q, Ebanks RO, Isenman DE. Two Clusters of Acidic Amino Acids Near the NH2 Terminus of Complement Component C4 α'-Chain Are Important for C2 Binding. J Immunol. 2000 Sep 1;165(5):2518–27.

33. Kim YU, Carroll MC, Isenman DE, Nonaka M, Pramoonjago P, Takeda J, et al. Covalent binding of C3b to C4b within the classical complement pathway C5 convertase. Determination of amino acid residues involved in ester linkage formation. J Biol Chem. 1992 Feb;267(6):4171–6.

34. WHO-IUIS nomenclature sub-committee. Revised nomenclature for human complement component C4. J Immunol Methods. 1993 Jul 6;163(1):3–7.

35. Zhou D, Rudnicki M, Chua GT, Lawrance SK, Zhou B, Drew JL, et al. Human Complement C4B Allotypes and Deficiencies in Selected Cases With Autoimmune Diseases. Front Immunol [Internet]. 2021 [cited 2022 Mar 30];12. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2021.739430

36. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

37. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios [Internet]. bioRxiv; 2021 [cited 2022 Apr 23]. p. 2021.02.06.430068. Available from: https://www.biorxiv.org/content/10.1101/2021.02.06.430068v1

38. Marin WM, Dandekar R, Augusto DG, Yusufali T, Heyn B, Hofmann J, et al. High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING. PLOS Comput Biol. 2021 Aug 2;17(8):e1008904.

39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.

40. Ittiprasert W, Kantachuvesiri S, Pavasuthipaisit K, Verasertniyom O, Chaomthum L, Totemchokchyakarn K, et al. Complete deficiencies of complement C4A and C4B including 2-bp insertion in codon 1213 are genetic risk factors of systemic lupus erythematosus in Thai populations. J Autoimmun. 2005 Aug 1;25(1):77–84.

41. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 16;10(2):giab008.

42. Sadedin SP, Oshlack A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. Genome Biol. 2019 Apr 18;20(1):78.

# Chapter 5: Conclusions

Through this work I have shown that NGS data representing complex genomic regions, such as *KIR* and *C4*, necessitate custom bioinformatic processing. And that, through informed processing of alignments, analysis of NGS data from these gene complexes can be done accurately and in high-throughput. The CNV exhibited by these complexes complicated variant identification by breaking common assumptions of variant depth ratios, which I have addressed through the development of custom variant identification methods that are informed by the determined gene copy number. Additionally, the component genes of these complexes exhibit high sequence similarity and nucleotide polymorphisms, characteristics that increase read mapping errors and erroneous genotypes. I have decreased these errors and improved genotyping accuracy through comprehensive reference alignments combined with custom alignment processing workflows. Together, these methods unlock full genomic sequence analysis of the *KIR* and *C4* regions for the first time, promoting research into these diverse and vital immune related complexes in addition to outlining strategies for improving alignments to other, similarly complex gene systems.

**Publishing Agreement**

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Wesley Marin
5A7F31DE5F6543A...
Author Signature

5/23/2022
Date