



Published in final edited form as:

HLA. 2024 January ; 103(1): e15273. doi:10.1111/tan.15273.

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

Wesley M. Marin<sup>1</sup>, Danillo G. Augusto<sup>1,2,3</sup>, Kristen J. Wade<sup>1</sup>, Jill A. Hollenbach<sup>1,4,\*</sup>

<sup>1</sup>Weill Institute for Neurosciences, Department of Neurology, University of California San Francisco, San Francisco, CA, United States

<sup>2</sup>Department of Biological Sciences, University of North Carolina Charlotte, Charlotte, NC, United States

<sup>3</sup>Programa de Pós-Graduação em Genética, Universidade Federal do Paraná, Curitiba, Brazil.

<sup>4</sup>Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, United States

### Abstract

The complement component 4 gene loci, composed of the *C4A* and *C4B* genes and located on chromosome 6, encodes for complement component 4 (C4) proteins, 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. *C4A* and *C4B* gene loci exhibit copy number variation, with each composite gene varying between 0–5 copies per haplotype. *C4A* and *C4B* genes also vary in size depending on the presence of the human endogenous retrovirus (HERV) in intron 9, denoted by *C4(L)* for long-form and *C4(S)* for short-form, which affects expression and is found in both *C4A* and *C4B*. Additionally, human blood group antigens Rodgers and Chido are located on the C4 protein, with the Rodger epitope generally found on C4A protein, and the Chido epitope generally found on C4B protein. *C4A* and *C4B* copy number variation has been implicated in numerous autoimmune and pathogenic diseases. Despite the central role of C4 in immune function and regulation,

---

\* **Correspondence:** Corresponding Author jill.hollenbach@ucsf.edu.

Author Contributions

**Conceptualization:** WMM, DGA, JAH

**Data Curation:** WMM, DGA

**Formal analysis:** WMM, KJW

**Funding acquisition:** JAH

**Investigation:** WMM

**Methodology:** WMM

**Project Administration:** JAH

**Resources:** DGA, JAH

**Software:** WMM

**Supervision:** JAH

**Validation:** WMM, DGA

**Visualization:** WMM

**Writing – Original Draft Preparation:** WMM

**Writing – Review & Editing:** WMM, DGA, JAH, KJW

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

high-throughput genomic sequence analysis of *C4A* and *C4B* 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 *C4A* and *C4B* sequences from short-read sequencing data, named C4Investigator. Using paired-end targeted or whole genome sequence data as input, C4Investigator determines the overall gene copy numbers, as well as *C4A*, *C4B*, *C4(Rodger)*, *C4(Ch)*, *C4(L)*, and *C4(S)*. Additionally, C4Investigator reports the full overall *C4A* and *C4B* aligned sequence, enabling nucleotide level analysis. To demonstrate the utility of this workflow we have analyzed *C4A* and *C4B* variation in the 1000 Genomes Project Dataset, showing that these genes are highly poly-allelic with many variants that have the potential to impact C4 protein function.

## Keywords

complement component; C4; genotyping; immunogenetics; copy number; bioinformatics pipeline

---

## Introduction

The *C4A* and *C4B* genes, located in human chromosomal region 6p21.33, encodes for complement component 4 (C4) proteins, 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. *C4A* and *C4B* loci exhibit copy number variation (CNV), with each composite gene varying between 0–5 copies per haplotype, and importantly, the gene copy number of *C4A* and *C4B* correlate to C4 protein levels(5). *C4A* and *C4B* loci also vary in size depending on the presence of a complete endogenous retrovirus in the intron 9 of *C4A* and *C4B*, named HERV-K(C4) (Figure 1A), denoted by *C4(L)* for long-form and *C4(S)* for short-form, which correlates with 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 1B), which are **PCPVLD** for C4A and **LSPVIH** 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 1A.

*C4A* and *C4B* 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 *C4A* and *C4B* 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 *C4A* and *C4B* CNV is accomplished through both digital droplet polymerase chain reaction (ddPCR)(11,20) and real time PCR quantification (15,21,22) which are capable of quantifying gene copy number for overall *C4A* and *C4B*, *C4A(L)*, *C4A(S)*, *C4B(L)* and *C4B(S)*. While these methods produce accurate results for *C4A* and *C4B* gene copy number and phasing with long and short form, they are intractable for identifying additional sequence variation at scale, including loss of functional variations(23,24) and recombinations (25,26), and are completely blind to novel sequence variation. High-throughput genomic sequence analysis of *C4A* and *C4B* variants has been impeded by the complex genetic variation exhibited by these genes. One recent tool for assessing *C4A* and *C4B* sequence variation is the analysis workflow hosted on Terra ([https://app.terra.bio/#workspaces/mccarroll-genomestrip-terra/C4AB\\_Analysis](https://app.terra.bio/#workspaces/mccarroll-genomestrip-terra/C4AB_Analysis)) (27), which was developed using the Genome STRiP software (28) to analyze *C4A* and *C4B* from whole genome sequencing (WGS) data. However, this tool is currently unpublished and is restricted to analysis of copy number variation of *C4A* and *C4B* specific SNPs and the HERV-K(C4).

Most C4 analysis workflows are targeted at characterizing the region of *C4A* and *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 amino acid sequence that, when mutated, have drastic functional consequences (Figure 1B). First are amino acid positions 477 and 478; mutations at these positions can disrupt C5 convertase activity (29,30), an important step in the classical and lectin complement cascade pathways that form the membrane attack complex (MAC). Positions 756 and 757 are the site of C1/MASP-2 cleavage(31) to produce C4a and C4b, which is the initial modification made to C4 proteins to initiate the complement cascade. Positions 1405–1427 and 1716–1732 are binding sites for C1/MASP-2 (32,33). Positions 763–770 make up a binding site for C2a (34), 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 (35), 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 1A) (24).

Due to the importance of C4 in complement cascade activity, coupled with the high degree of allotypic variation (36,37), we believe that full genomic sequence characterization of *C4* is of vital importance to advancing our understanding of its in human health. To investigate *C4A* and *C4B* variation using genomic sequencing data, we have developed a bioinformatic pipeline for comprehensive, high-throughput characterization of human *C4A* and *C4B* 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 *C4A* and *C4B*, *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 1000 Genomes Project (1KGP) high depth 30x WGS data(38,39), a dataset consisting of 3,202 samples, characterizing *C4A* and *C4B* 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

### 1.1 C4Investigator 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 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.(40), was developed to integrate the *C4A* and *C4B* alignments into the overall alignment. From the overall alignment, *C4A* and *C4B* copy numbers are determined by comparing the median alignment depth across *C4A* and *C4B* to the average depth of the Tenascin XB (*TNXB*), a nearby copy-stable gene. Gene copy numbers of *C4A*, *C4B*, *C4(Ch)*, *C4(Rg)*, *C4(L)* and *C4(S)* are determined by multiplying the ratios of *C4A/C4B* specific SNPs, *Rg/Ch* specific SNPs and the HERV-K(C4) 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* and *C4B* specific SNPs to the HERV-K(C4) region, this method is unable to phase *C4A* and *C4B* with long and short-form.

In addition to gene copy number analysis, C4Investigator outputs the full overall *C4A-C4B* aligned sequences as an SNP table.

The pipeline is available at: <https://github.com/hollenbach-lab/C4Investigator>.

**1.1.1 Alignment workflow**—The structural variation of *C4A* and *C4B* loci and the high-degree of sequence similarity between them necessitates a custom alignment and processing workflow. The first step of the workflow is a Bowtie2(41) 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 *C4A-C4B* alignment used for downstream analysis. The output of this workflow is a *C4A* and *C4B* 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.

**1.1.2 C4A and C4B copy number determination**—The median depth of the overall *C4A-C4B* alignment is normalized by the median depth of *TNXB*, which determines the overall copy number of *C4A* and *C4B*. 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 gene copy number to determine individually the copy numbers of *C4A* and *C4B*. 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)* copy number. Finally, the depth ratio of the HERV-K(C4) insertion, across positions 19.276-19.6642, is multiplied by the overall 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(23,42) 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.

**1.1.3 C4A and C4B sequence analysis**—The overall depth table is processed to generate a SNP table for positions passing a minimum depth threshold (6 for whole genome sequence data and 20 for targeted sequence data). Heterozygous positions are identified using a depth ratio of 0.5 normalized by the determined *C4A* and *C4B* gene copy number. The output of this step is an overall SNP table with the combined sequence for *C4A* and *C4B*.

## 1.2 Application and validation of C4Investigator

### 1.2.1 Targeted sequencing dataset generation and analysis with

**C4Investigator**—Targeted-capture next-generation sequencing (NGS) was applied 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(43).

A total of 100 ng of high-quality DNA is fragmented using the Twist EF Kit 2.0 1 (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 the targeted sequencing datasets using a minimum depth of 20X for variant calling and a ratio of 0.50, normalized by the total copy of *C4A* and *C4B*, for heterozygous position identification.

**1.2.2 ddPCR genotyping for validation of C4Investigator**—Copy numbers for *C4A*, *C4B*, *C4(L)* and *C4(S)* were determined by ddPCR as described previously (11)

for the INDIGO cohort subset described above. 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.

**1.2.3 1000 Genomes Project analysis and C4Investigator validation via comparison to existing annotation workflow.**—As an additional means of validation, C4Investigator copy number results for the 1KGP dataset were compared to results from the analysis workflow utilizing Genome STRiP(38) implemented in Terra (27). For each 1KGP individual, short reads aligned to *C4A* and *C4B* and the nearby region were extracted from GRCh38 aligned CRAM files using the coordinates outlined in Table S1 using Samtools(44). The extracted reads were converted to paired-end FASTQ files using Bazam(45). 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 *C4A* and *C4B*, for heterozygous position identification. Copy number results were stratified by superpopulation. Population totals and abbreviations are outlined in Table 1.

Then, C4Investigator annotations were compared to the Genome STRiP Terra workflow annotations for this set of 1KGP individuals. For overall *C4A* and *C4B*, all results across both datasets were compared. For *C4A* and *C4B* comparison, samples marked as *C4A1*, *C4A2*, *C4B1*, or *C4R1*, which represented rare variants, by the Genome STRiP Terra workflow were excluded, representing a total of 55 samples excluded from the 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.

Chi-squared testing was carried out to compare observed *C4A* and *C4B* copy number frequencies between superpopulations. For each gene, copy numbers with low observed frequencies were summed together so that each observation had a value >0. Binned frequencies were then analyzed using the *chisq.test* function in R.

## Results

### 1.3 Performance evaluation – ddPCR copy number comparison

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 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.

### 1.4 Performance evaluation – C4A/B Terra copy number comparison

To benchmark C4Investigator performance against another bioinformatic workflow, we compared results for the 1000 Genomes Project dataset (N=3199) against results from the unpublished Terra workflow(27), a bioinformatic pipeline that utilizes Genome STRiP(38) to quantify *C4A* and *C4B* copy numbers.



Overall copy determination performance was highly concordant with the 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 *C4A* and *C4B* variants quantified by 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 S1A). Further examination into the *C4A* and *C4B* 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 S1B). 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 Terra workflow ratios were clustered near the center of the copy intervals (Figure S2). These results suggest that poor kmer quality and differences in the functionalities of C4Investigator and Terra may be driving the observed, albeit limited, discordancies.

### 1.5 1000 Genomes Project – *C4A* and *C4B* copy number analysis

Analysis of *C4A* and *C4B* copy number variation across superpopulations showed most individuals across all superpopulations had 4 copies overall, 2 copies of *C4A*, and 2 copies of *C4B*, and there were very few individuals with 0 copies (Figure 2). The African (AFR) and European (EUR) superpopulations had higher occurrences of 3 overall copies of *C4*, almost double that observed in the other superpopulations, and lower occurrences of 5 and 6 overall copies of *C4A* and *C4B*. In contrast, the South Asian (SAS) superpopulation had the lowest occurrence of 3 overall copies of *C4A* and *C4B*, 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%. Chi-square testing of copy number frequencies across superpopulations confirmed a significant association between copy number and superpopulation for each of the *C4* genes (Table S2).

### 1.6 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 3, and SNP variation that is not represented in the main assembly of the GRCh38 reference in Figure 3.

Analysis of allele frequencies for *C4A* and *C4B* non-synonymous sequence variation showed large variations in frequencies across populations (Table 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.

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 3A, Table S3). 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-K(C4)), and 59 HERV-K(C4) variant positions accounting for 0.927% of all HERV-K(C4) 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 3B, Table S3), 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.

### 1.7 1000 Genomes Project – recombinant analysis

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). 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%.

### 1.8 Performance evaluation – phasing of C4A and C4B with Rodger and Chido I

Phasing completeness between the *C4A* and *C4B* specific SNP group with the SNPs that define Rg and Ch 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 *C4A* and *C4B* copy number determination to ddPCR results showed high concordance between the two methods across divergent populations (Table 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* and *C4B* Terra workflow, another bioinformatic pipeline, on the 1KGP WGS dataset showed high concordance between the two workflows, especially for the overall copy numbers. An investigation into discordant copy number results showed that the discordant samples had lower base quality scores on average (Figure S1B), with neither method showing clear copy number results for the discordant samples (Figure S1A). In contrast, the investigation into discordant HERV-K(C4) results showed a marked difference between the two methods, with the Terra workflow showing clear copy numbers for these samples while C4Investigator had unclear determinations (Figure S2). This is likely due to the additional structural variant processing of the Terra workflow, which incorporates Genome STRiP (38), a workflow specifically developed for identifying copy number variation in WGS data. The Terra workflow for *C4A* and *C4B* strictly focuses on identifying copy number variation, which 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 *C4A* and *C4B* genes.

An analysis of *C4A* and *C4B* copy number variation between superpopulations (Figure 2) demonstrated some specific patterns, such as a median overall *C4A* and *C4B* 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 individuals carrying 2 copies of *C4L* in the AFR superpopulation, and the general imbalance between overall *C4A* and *C4B* copy number 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 haplotype 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 *C4A* and *C4B* 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). 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 (33) (Figure 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(30). 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 these genes are highly poly-allelic, with extensive variation across introns, exons, and the HERV-K(C4) region (Figure 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 3B). This analysis demonstrates the value of nucleotide level analysis of *C4A* and *C4B*, 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. This issue of variant phasing may be solved in the future, with the development of long-read sequencing protocols designed for *C4A* and *C4B* analysis. Though the C4Investigator algorithm is designed specifically for short read data, the copy number ratio conceptual framework could be re-implemented with long read specific tools and statistical models in future work.

Until such solutions are developed, 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 the distance between the SNPs that define *C4A* and *C4B* SNPs with those that define the *Rg* and *Ch* is only 440bp, we can determine the 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), and appreciable carrier frequencies for the *C4B-Rg* recombinant and the AMR and SAS superpopulations. Since *C4A-Rg* and *C4B-Ch* is the predominant linkage pattern observed in human populations (8), the observation of the *C4A-Ch* recombinant is unexpected. However, to-date, few studies have thoroughly characterized *C4A* and *C4B* variation with the resolution we achieved in populations of African ancestry, thus, this is a novel finding. Interestingly, complement system activation has been shown to be altered in individuals with sickle cell disease (SCD) (46,47), which also has a high prevalence in African populations, due to its protective effect against malaria (48,49). It is possible that the unique combination of *C4A*-specific SNPs with the *Chido* group SNPs has a role in mediating the molecular response to sickle cell disease. Future studies may be able to better interpret this novel observation.

In conclusion, C4Investigator fills a critical role in the investigation of *C4A* and *C4B* variation, processing WGS data to provide 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 *C4A* and *C4B* copy number varies between superpopulations, that alleles with low global allele frequencies can have high population specific frequencies, the presence and distribution of recombinant variants, and

population specific carrier frequencies for rare alleles. Additionally, we have demonstrated that C4Investigator can identify 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 *C4A* and *C4B* from WGS data, and as such, promises to contribute to our understanding of the role of this genomic region in human health and disease.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments and funding

We would like to thank Michael Wilson and Mark Seielstad for constructive comments. We would like to acknowledge that this work exists as a chapter of Wesley Marin's doctoral dissertation (44). This work was supported by the National Institutes of Health (NIH-R01AI128775). The funders had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Data Availability Statement

The datasets analyzed for this study can be found in the International Genome Sample Resource data portal at <https://www.internationalgenome.org/data>. The C4Investigator workflow is available at <https://github.com/Hollenbach-lab/C4Investigator>. And the scripts used to analyze the data are available at [https://github.com/wesleymarin/C4investigator\\_scripts](https://github.com/wesleymarin/C4investigator_scripts).

## References

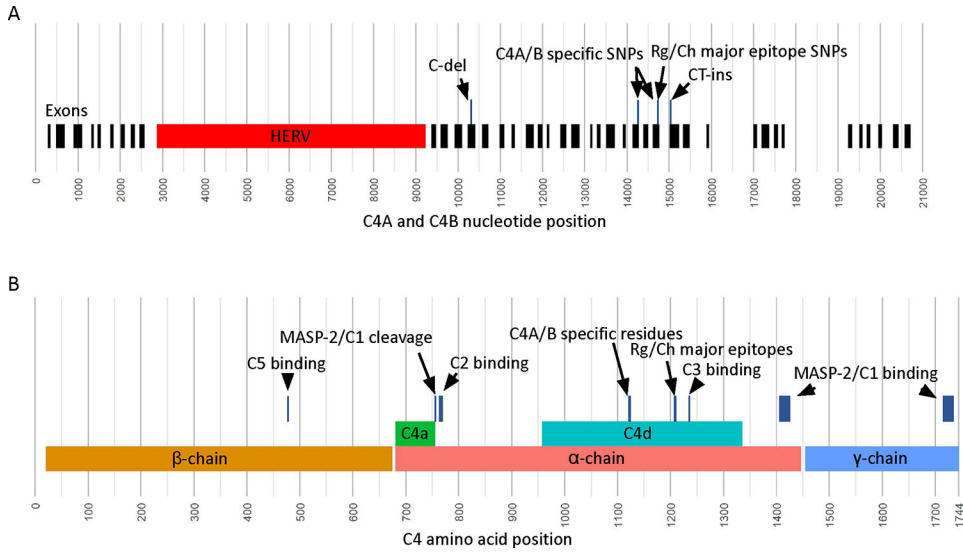
1. Wang H, Liu M. Complement C4, Infections, and Autoimmune Diseases. *Frontiers in Immunology* [Internet]. 2021 [cited 2022 Apr 28];12. Available from: <https://www.frontiersin.org/article/10.3389/fimmu.2021.694928>
2. Toapanta FR, Ross TM. Complement-mediated activation of the adaptive immune responses: role of C3d in linking the innate and adaptive immunity. *Immunol Res.* 2006;36(1–3):197–210. [PubMed: 17337780]
3. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The complement system and innate immunity. *Immunobiology: The Immune System in Health and Disease* 5th edition [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. *Frontiers in Immunology.* 2015;6:257. [PubMed: 26074922]
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. *The Journal of Immunology.* 2003 Sep 1;171(5):2734–45. [PubMed: 12928427]
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 Research.* 1998 Oct 26;809(1):31–8. [PubMed: 9795119]
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. [PubMed: 20795316]
10. Mougey R An update on the Chido/Rodgers blood group system. *Immunohematology*. 2019 Dec;35(4):135–8. [PubMed: 31935328]
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. [PubMed: 26814963]
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. [PubMed: 31439935]
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. [PubMed: 27758680]
14. Macedo ACL, Isaac L. Systemic Lupus Erythematosus and Deficiencies of Early Components of the Complement Classical Pathway. *Frontiers in Immunology*. 2016;7:55. [PubMed: 26941740]
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. *Advances in Rheumatology*. 2019 Aug 6;59(1):36. [PubMed: 31387635]
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. *The American Journal of Human Genetics*. 2007 Jun 1;80(6):1037–54. [PubMed: 17503323]
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. *Frontiers in Immunology*. 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. [PubMed: 34804054]
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. [PubMed: 32878183]
21. Szilagyi A, Blasko B, Szilassy D, Fust G, Sasvari-Szekely M, Ronai Z. Real-time PCR quantification of human complement C4A and C4B genes. *BMC Genetics*. 2006 Jan 10;7(1):1.
22. Paakkanen R, Vauhkonen H, Eronen KT, Järvinen A, Seppänen M, Lokki ML. Copy Number Analysis of Complement C4A, C4B and C4A Silencing Mutation by Real-Time Quantitative Polymerase Chain Reaction. *PLoS One*. 2012 Jun 21;7(6):e38813. [PubMed: 22737222]
23. 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. *The Journal of Immunology*. 1999 Mar 15;162(6):3687–93. [PubMed: 10092831]
24. 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. [PubMed: 19279649]
25. 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. *The Journal of Immunology*. 1998 Oct 1;161(7):3438–43. [PubMed: 9759862]
26. Jaatinen T, Eholuoto M, Laitinen T, Lokki ML. Characterization of a De Novo Conversion in Human Complement C4 Gene Producing a C4B5-Like Protein. *The Journal of Immunology*. 2002 Jun 1;168(11):5652–8. [PubMed: 12023363]

27. 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](https://app.terra.bio/#workspaces/mccarroll-genomestrip-terra/C4AB_Analysis)
28. 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. [PubMed: 25621458]
29. 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. *The Journal of Immunology.* 1992 May 1;148(9):2803–11. [PubMed: 1573269]
30. 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. *Journal of Biological Chemistry.* 1994 Nov;269(44):27727–31. [PubMed: 7961694]
31. 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 \*. *Journal of Biological Chemistry.* 2005 Dec 23;280(51):41811–8. [PubMed: 16227207]
32. 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. [PubMed: 23592783]
33. 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. *Proceedings of the National Academy of Sciences.* 2012 Sep 18;109(38):15425–30.
34. Pan Q, Ebanks RO, Isenman DE. Two Clusters of Acidic Amino Acids Near the NH2 Terminus of Complement Component C4  $\alpha'$ -Chain Are Important for C2 Binding. *The Journal of Immunology.* 2000 Sep 1;165(5):2518–27. [PubMed: 10946278]
35. 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. *Journal of Biological Chemistry.* 1992 Feb;267(6):4171–6. [PubMed: 1740458]
36. WHO-IUIS nomenclature sub-committee. Revised nomenclature for human complement component C4. *Journal of Immunological Methods.* 1993 Jul 6;163(1):3–7. [PubMed: 8335957]
37. 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. *Frontiers in Immunology* [Internet]. 2021 [cited 2022 Mar 30];12. Available from: <https://www.frontiersin.org/article/10.3389/fimmu.2021.739430>
38. 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. [PubMed: 26432245]
39. Byrka-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>
40. 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. [PubMed: 34339413]
41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012 Apr;9(4):357–9. [PubMed: 22388286]
42. Ittiprasert W, Kantachavesiri 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. *Journal of Autoimmunity.* 2005 Aug 1;25(1):77–84. [PubMed: 15998580]
43. 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's Disease. *J Immunol.* 2020 Sep 1;205(5):1323–30. [PubMed: 32709660]
44. 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. [PubMed: 33590861]
  45. Sadedin SP, Oshlack A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. *Genome Biology.* 2019 Apr 18;20(1):78. [PubMed: 30999943]
  46. Yoo JJ, Graciaa SH, Jones JA, Zuo Z, Arthur CM, Leong T, et al. Complement Activation during Vaso-Occlusive Pain Crisis in Pediatric Sickle Cell Disease. *Blood.* 2021 Nov 23;138:858. [PubMed: 34036317]
  47. Roumenina LT, Chadebech P, Bodivit G, Vieira-Martins P, Grunenwald A, Boudhabhay I, et al. Complement activation in sickle cell disease: Dependence on cell density, hemolysis and modulation by hydroxyurea therapy. *American Journal of Hematology.* 2020;95(5):456–64. [PubMed: 31990387]
  48. Elguero E, Délicat-Loembet LM, Rougeron V, Arnathau C, Roche B, Becquart P, et al. Malaria continues to select for sickle cell trait in Central Africa. *Proceedings of the National Academy of Sciences.* 2015 Jun 2;112(22):7051–4.
  49. Adigwe OP, Onoja SO, Onavbavba G. A Critical Review of Sickle Cell Disease Burden and Challenges in Sub-Saharan Africa. *J Blood Med.* 2023 May 31;14:367–76. [PubMed: 37284610]
  44. Marin WM. Development of Bioinformatics Methods to Interrogate Complex Immune Related Genomic Regions from Next Generation Sequencing Data. [Doctoral dissertation, University of California, San Francisco]. [eScholarship.org](https://escholarship.org) and the California Digital Library. 2022.





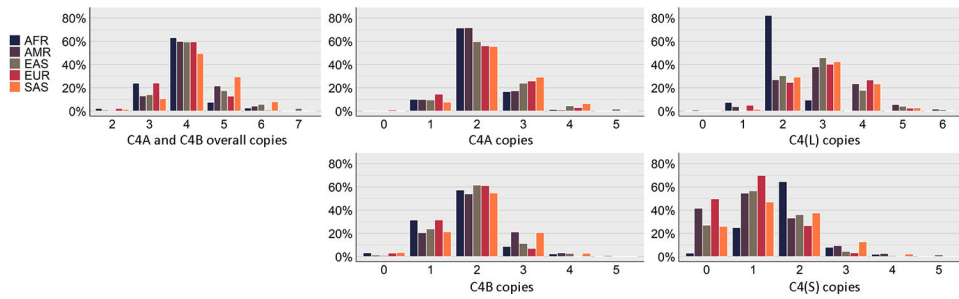
**Figure 1. Sequence features of C4A and C4B genes and C4 proteins.**  
**(A)** Positions of C4A and C4B genomic sequence features shown for a long-form of the genes. Exon positions are marked in black, the HERV-K(C4) sequence is marked in red, and select sequence variants are shown above the exons. Positions are based on the C4A and C4B combined alignment reference, which includes 5' UTR and 3' UTR sequence. The C-del variant and the TC-ins variant are frame-shift mutations that result in premature terminations. **(B)** Positions of C4A and C4B protein sequence features. The major chains,  $\alpha$ ,  $\beta$ , and  $\gamma$ , 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.

Author Manuscript

Author Manuscript

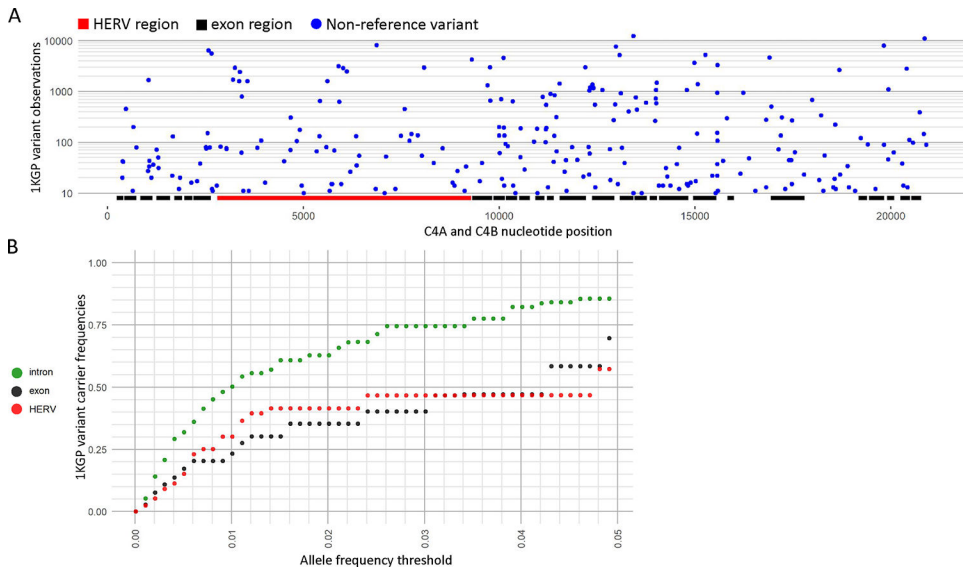
Author Manuscript

Author Manuscript



**Figure 2. Superpopulation distributions of *C4A* and *C4B* copy number results for the 1KGP dataset.**

*C4A* and *C4B* 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.



**Figure 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 *C4A* and *C4B* 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 exon and HERV-K(*C4*) 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-K(*C4*) 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.

**Table 1.**

1000 Genomes Project population abbreviations and size.

	<b>Population</b>	<b>N</b>
<b>European (EUR)</b>		<b>633</b>
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
<b>East Asian (EAS)</b>		<b>582</b>
Southern Han Chinese (CHS)		161
Chinese Dai in Xishuanagbanna, China (CDX)		92
Kinh in Ho Chi Minh City, Vietnam (KHV)		122
Han Chinese in Beijing, China (CHB)		103
Japanese in Tokyo, Japan (JPT)		104
<b>Admixed American (AMR)</b>		<b>490</b>
Puerto Rican from Puerto Rica (PUR)		139
Colombian from Medellin, Colombia (CLM)		132
Peruvian from Lima, Peru (PEL)		122
Mexican Ancestry from Los Angeles USA (MXL)		97
<b>South Asian (SAS)</b>		<b>601</b>
Punjabi from Lahore, Pakistan (PJI)		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
<b>African (AFR)</b>		<b>893</b>
African Carribean in Barbados (ACB)		116
Mandinka in The Gambia (GWD)		178
Esan in Nigera (ESN)		149
Mende in Sierra Leone (MSL)		99
Yoruba in Ibadan, Nigera (YRI)		178
Luhya in Webuye, Kenya (LWK)		99
American's of African Ancestry in SW USA (ASW)		74

**Table 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

<u>Ancestry</u>	<u>C4A</u>	<u>C4B</u>	<u>C4(S)</u>	<u>C4(L)</u>
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

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 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 global variant while minor amino acids and nucleotides represent the second most frequent variant. This data was filtered to only show variants with allele frequencies  $\geq 2\%$  for any population. Blank values represent absence of the variant. See Table 1 for population abbreviations.

	aa pos	141	229	325	328	478	549	726	791	916	959	1286	1413	1530
	major aa	L	T	K	M	P	H	P	R	R	E	A	A	P
	minor aa	V	I	M	I	L	P	L	H	Q	D	S	P	S
	nuc pos	E3 157	E6 60	E9 62	E9 72	E12 92	E13 122	E17 106	E18 103	E21 155	E23 23	E29 180	E33 6	E36 4
	major nuc	C	C	A	G	C	A	C	G	G	A	G	G	C
	minor nuc	G	T	T	A	T	C	T	A	A	C	T	C	T
	<b>GBR</b>	6.1					2.6					7.8		
	<b>FIN</b>	8.4					5.9				2.2	4.5		
<b>EUR</b>	<b>IBS</b>	3.2					2.1					4.6		
	<b>CEU</b>	6.1					5.5					8.2		
	<b>TSI</b>	4.1					4.1					3.8		
	<b>CHS</b>	35.8					14.3				8.6	3.6		
	<b>CDX</b>	53.3					18.7				17.7			
<b>EAS</b>	<b>KHV</b>	33.7	2.1	4.4	4		10.2				13			
	<b>CHB</b>	25.8		2.9	2.9		12.8				4.6	5.6		
	<b>JPT</b>	13.3		4.4	4.2		8					13.1		
	<b>PUR</b>	8.7					3					3.1		
<b>AMR</b>	<b>CLM</b>	7.7					3.9					2.4		
	<b>PEL</b>	16					6.8				2.7			
	<b>MXL</b>	15.8					7					2.2		
	<b>PJL</b>	4					4					5.7		
	<b>BEB</b>	16.1					3.8				7.4	4.8		
<b>SAS</b>	<b>STU</b>	8					4					4.6		
	<b>ITU</b>	8.8					5.5				2.9	6.4		
	<b>GIH</b>	6.3					3.1					5.9		
	<b>ACB</b>	7				3.5		2.2					2.2	2.2
	<b>GWD</b>	5.1				3.8			3.1	2.3		4.5		
<b>AFR</b>	<b>ESN</b>	10.6				4.7				2.3				
	<b>MSL</b>	3.5							4.3				10.2	10.2
	<b>YRI</b>	10.5				4.1				2.4			4	4



<u>LWK</u>	10.7		3.1	2.1	2.9	3.7
<u>ASW</u>	10.7	2.5			2.1	2.8

---

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 4.**  
***C4A-Ch* and *C4B-Rg* 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 1 for population abbreviations.

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