# SHIELD: Snv Heuristic Identification, Exploration, and Location Detector

by

Valerie Flint

(Under the Direction of Juan B. Gutierrez)

## Abstract

RNAseq has become a preferred method of data generation, particularly when dealing with non-model organisms or organisms for which no reference genome has yet been completed. Similarly, single nucleotide polymorphisms (SNPs) and their phenotypic effects have become a favorite subject of study throughout the genetic community. While several tools exist and are being created to call and analyze SNPs from RNAseq data, there are still some gaps in the situations that the current technologies can address. The Snv Heuristic Identification, Exploration, and Location Detector (SHIELD) is a fully automated pipeline tailored to take mapped RNA-seq reads from studies with small sample sizes, find unique SNVs between user-defined groups, and generate Circos plots for visualization of the data and results. SHIELD's use is demonstrated in an analysis of *M. mulatta* subjects infected with malaria. SHIELD identified several high density SNV regions that confirm the importance of genetic variations in innate immune functions.

Index words: Malaria, *Plasmodium cynomolgi*, Bioinformatics, RNA-Seq

# SHIELD: Snv Heuristic Identification, Exploration, and Location Detector

by

Valerie Flint

B.A., Harriet L. Wilkes Honors College, Florida Atlantic University, 2010

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2018

SHIELD: Snv Heuristic Identification, Exploration, and Location Detector

by

Valerie Flint

Approved:

Major Professor: Juan B. Gutierrez

Committee: Shaying Zhao
Liang Liu
Jonathan Arnold

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

# Acknowledgments

I would like to thank Dr. Gutierrez for his guidance and patience throughout this doctoral process, particularly in the last year. I would like to thank my committee members and labmates for their advice and assistance with this dissertation. In particular I would like to thank Elizabeth Trippe, Karen Gonzalez, and Dr. Diego Moncada for their invaluable assistance with this project and the use of some of their own data. I would like to thank the MaHPIC Consortium for providing the data to analyze. Finally, I would like to thank my family and friends for their patience and support for all these years.

# Table of Contents

## List of Figures

# List of Tables

# Chapter 1

# Background and Introduction

## 1.1 SNPs, SNVs, and GWAS

Single nucleotide polymorphisms (SNPs) and variations (SNVs) occur roughly once every 300 base pairs in the human genome [1]. Both occur when a single base pair varies from a reference genome, but are only considered SNPs when at least 1% of a given population carry the deviation [2]. A point mutation may be considered a SNV even if it is actually a SNP if there is no means of confirming whether or not the variation is sufficiently fixed in a population, such as when analyzing mutations in a small cohort of subjects. All point mutations in this study will be considered to be SNVs, since it is not possible to conclude whether the mutations are sufficiently frequent or are simply post-translational modifications. Regardless, SNPS and SNVs alike are worth study as many have been found to be responsible for a wide variety of genetic diseases, largely due to contributions made by genome-wide association studies (GWAS) [3].

A GWAS compares whole genomes of individuals who have been split into two groups, those with and without the phenotype in question [4]. Pairwise comparison tests are then done at each SNP in order to determine frequencies of SNPs that are significantly different between the groups yet similar within them. These studies have been successful in identifying many variations responsible for most simple genetic diseases (that is, relatively binary diseases caused by a mutation at a single locus, such as cystic fibrosis and phenylketonuria), and some SNPs with particularly large contributions towards more complex diseases and traits (those controlled by many loci, such as height and Crohn's disease) [3, 5, 6]. However, the technique is not without its drawbacks.

Due to the nature of the study, a GWAS requires a particularly large sample size and extremely conservative P-value threshold. Many of the more commonly used SNP testing arrays contain up to 2 million SNPs, which results in an average GWAS consisting of up to 2 million comparison tests [7]. At the standard P-value threshold of 0.05, one is accepting a false positive rate of up to 5%. That error rate would accumulate after 2 million tests, however, resulting in 100,000 falsely associated SNPs. Thus the Bonferroni correction is generally applied, wherein the P-value is divided by the number of tests being performed, which results in the highly conservative threshold of $2 \times 10^{-8}$ [8]. Unfortunately, this process also requires particularly large sample sizes in order to confidently predict association, ranging anywhere from a minimum of hundreds of subjects to thousands and more depending on the SNP's allele frequency and effect size [7].

## 1.2 RNA-Seq

RNA-Seq has become a preferred method of data generation, particularly when dealing with non-model organisms or organisms for which no reference genome has yet been completed [9]. The process is generally similar to whole genome sequencing (WGS); the molecule in question is broken down and then reassembled, either against an existing reference genome or using *de novo* assembly methods. However, RNA-Seq first converts RNA into a cDNA library which is then sequenced and assembled, either *de novo* or against a reference, which can either be a genome or transcriptome [9]. RNA-Seq is a cheaper and faster method for gathering sequence data than whole genome sequencing at this time, while also being more exact and detailed than microarrays. For example, since eukaryotic organisms undergo a splicing step before translation, the resulting RNA-Seq data can reveal how various exons are connected in a given transcript.

While many tools exist and are being created to call and analyze SNPs from whole genome sequence and RNA-Seq data respectively, there are still some gaps in the situations that the current technologies can address. Admittedly, the largest obstacle to RNA-Seq data usage appears to be the ability to accurately map the reads. That is, most of the emerging technologies and programs are presented as ways to optimize the read mapping process, whether that is with a reference genome, *de novo*, or simply for an underrepresented non-model organism [10, 11, 12, 13, 14]. However, RNA-Seq data usage does not stop at mapping. There is a clear gap in the technology for an automated analysis pipeline that goes beyond the initial read mapping and can provide some results with even a small sample size. Some attempts at pipelines for further RNA-seq analysis have already been made, such as the one reported by De Wit et. al [15]. However, even the authors admit that the scripts they provide are now outdated and difficult to use, as they are a collection of individual programs that must be found and run in a particular order rather than a single cohesive pipeline. Mercury [16] is an example of an elegantly automated analysis pipeline, although it focuses on genomic data.

The SNV Heuristic Identification, Exploration, and Location Detector (SHIELD) is a fully automated pipeline tailored to take mapped RNA-seq reads and return comparisons between user-defined groups of SNVs and Circos plots for data and result visualization. Single nucleotide variants are specified in the name rather than polymorphisms because the process used by the program is appropriate for small samples; sizes at which one could confidently claim SNV discovery, but not always necessarily SNP discovery. SHIELD takes RNA-Seq reads that have already been mapped and runs them through a series of standard steps to generate SNV files tailor-made for easier downstream analysis, and provides some visual representation of the results. SHIELD can also make various group comparisons of the resulting SNV files based on the user's desired analysis. The tools used are standard, straightforward, dependable, and easily updated as versions become available. SHIELD's use is demonstrated through the analyses using data described in section 1.4.

## 1.3 Malaria

Malaria is a blood disease caused by a parasite, primarily *Plasmodium falciparum* and *Plasmodium vivax* in humans [17]. The disease is known to cause a range of symptoms in its hosts, such as fever, aches, fatigue, and severe anemia, the combination of which is frequently fatal; there were approximately 445,000 deaths due to malaria out of the estimated 196 - 263 million cases that occurred in 2016, most of which were children under the age of five [18]. The disease is most prominent in tropical areas with high mosquito counts and limited resources, particularly Africa and southeast Asia (Figure 1.1) [19, 20].

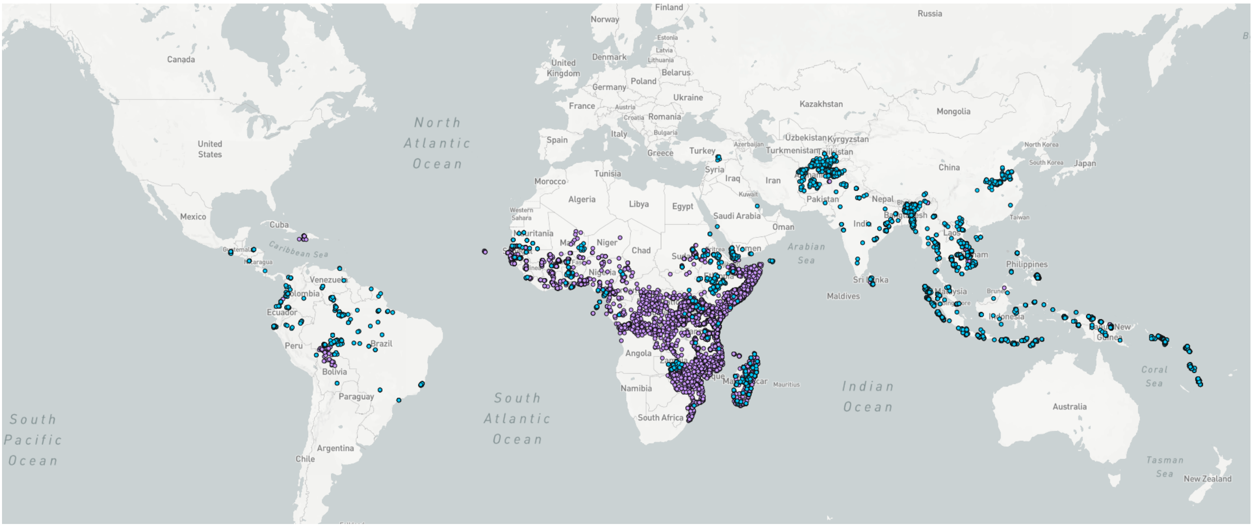


Figure 1.1: **Map of Malaria Prevalence**

All of the spots represent village level clustering surveys of malaria prevalence, with the blue dots specifying *P. vivax* and the purple dots specifying *P. falciparum*. The map was generated by the Malaria Atlas Project on April 3, 2018 [21]

.

The parasite matures from the gametocyte to the sporozoite stage from within mosquitoes of the genus *Anopheles* over the course of a week [22]. Within the digestive tract of the mosquito the male and female gemetocytes merge into zygotes, which then elongate into ookinetes. These are larger motile forms which cross through the epithelium of the mosquito's tract forming an oocyst, which is a pocket of epithelial tissue full of ookinetes. The parasites within the oocyst goes through several cycles of replication until they emerge as sporozoites and move into the mosquito's salivary glands, at which point the mosquito can infect new hosts. When the mosquito takes a blood meal, it first spits into the host a collection of enzymes to thin the blood (which is the source of the itchy bump such bites produce). Thus the sporozoites travel from the mosquito's salivary glands into the host's bloodstream and begin the asexual stage of their life cycle [22].

The injected sporozoites immediately move to the host's liver, where they remain for a few days [22]. During the liver stage, the sporozoites replicate asexually into merozoites from within the liver cells. Eventually the merozoites will rupture the liver cells and travel into the blood stream where they invade red blood cells. The merozoites continue replicating asexually in red blood cells until the cells lyse, spilling even more of the parasite into the bloodstream. The newly introduced merozoites may either continue the cycle in new blood cells or may differentiate back into male and female gametocytes. These gametocytes are then taken up by a mosquito when it lands for a blood meal and the disease is carried to a new host. The host does not display any symptoms of disease until the infection hits the blood stage [22]. The cycle has been summarized in Figure 1.2.

While the parasite is in the liver stage, the host is asymptomatic. Disease symptoms do not tend to appear until the parasite has reached the blood stage due to the body's response to the rupturing cells. Symptoms can range from what is clinically considered to be uncomplicated to complicated, depending on overall severity and organ damage. Uncomplicated malaria can result in fevers, anemia, chills, and other flu-like symptoms, while complicated malaria involves anemia severe enough to result in fatal organ failures [23]. For the purposes of this analysis, all malaria cases will hereafter be referred to as either mild or severe, depending on which set of symptoms the case falls under as outlined by the CDC (Table 1.1). Generally, however, mild malaria tends to mimic influenza or other viral symptoms, while severe malaria includes any case involving organ failure or abnormalities and can easily prove fatal. Figure 1.3 summarizes some statistics collected on recent localized malaria cases and demonstrates the correlation between severity and mortality [23, 24].

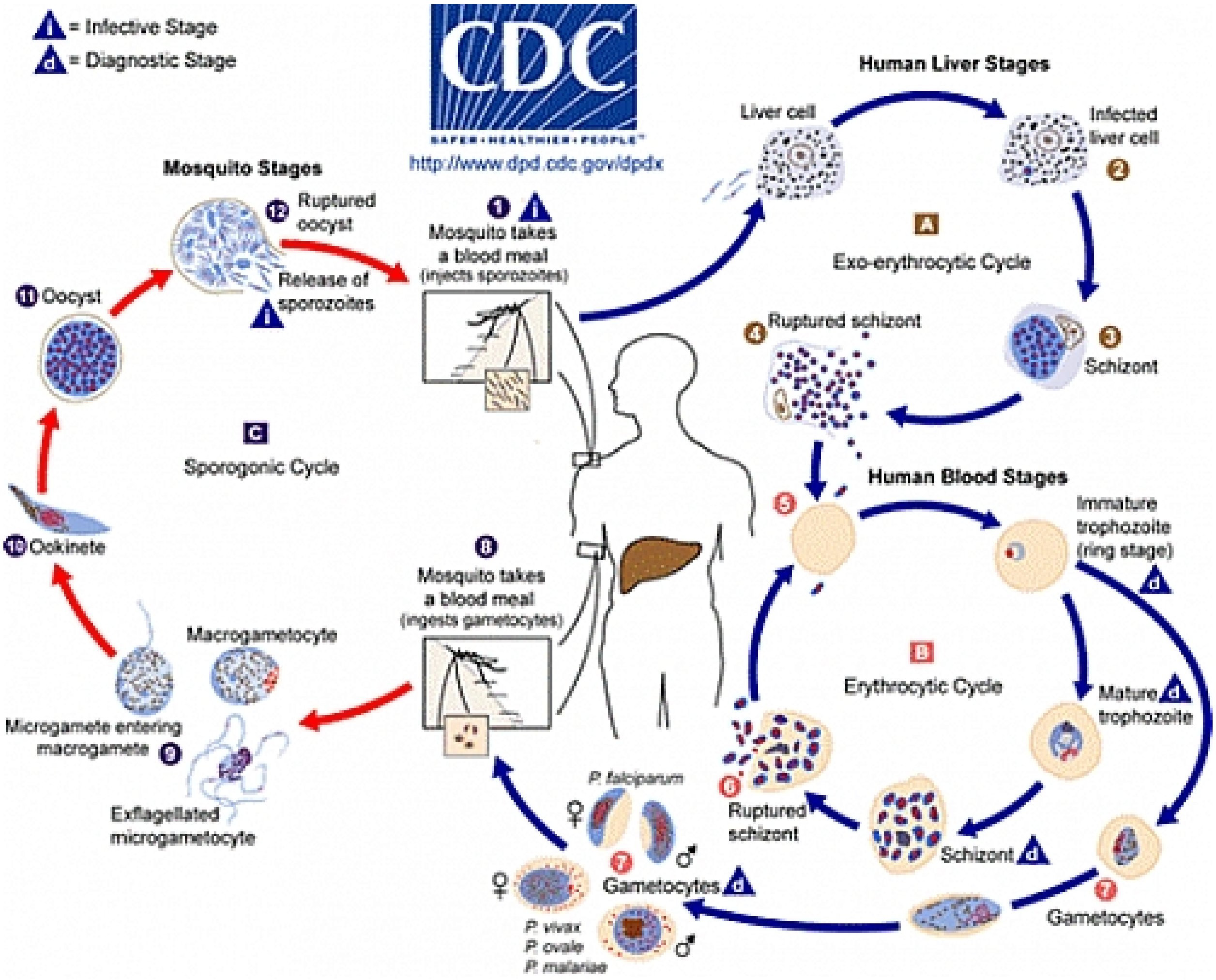


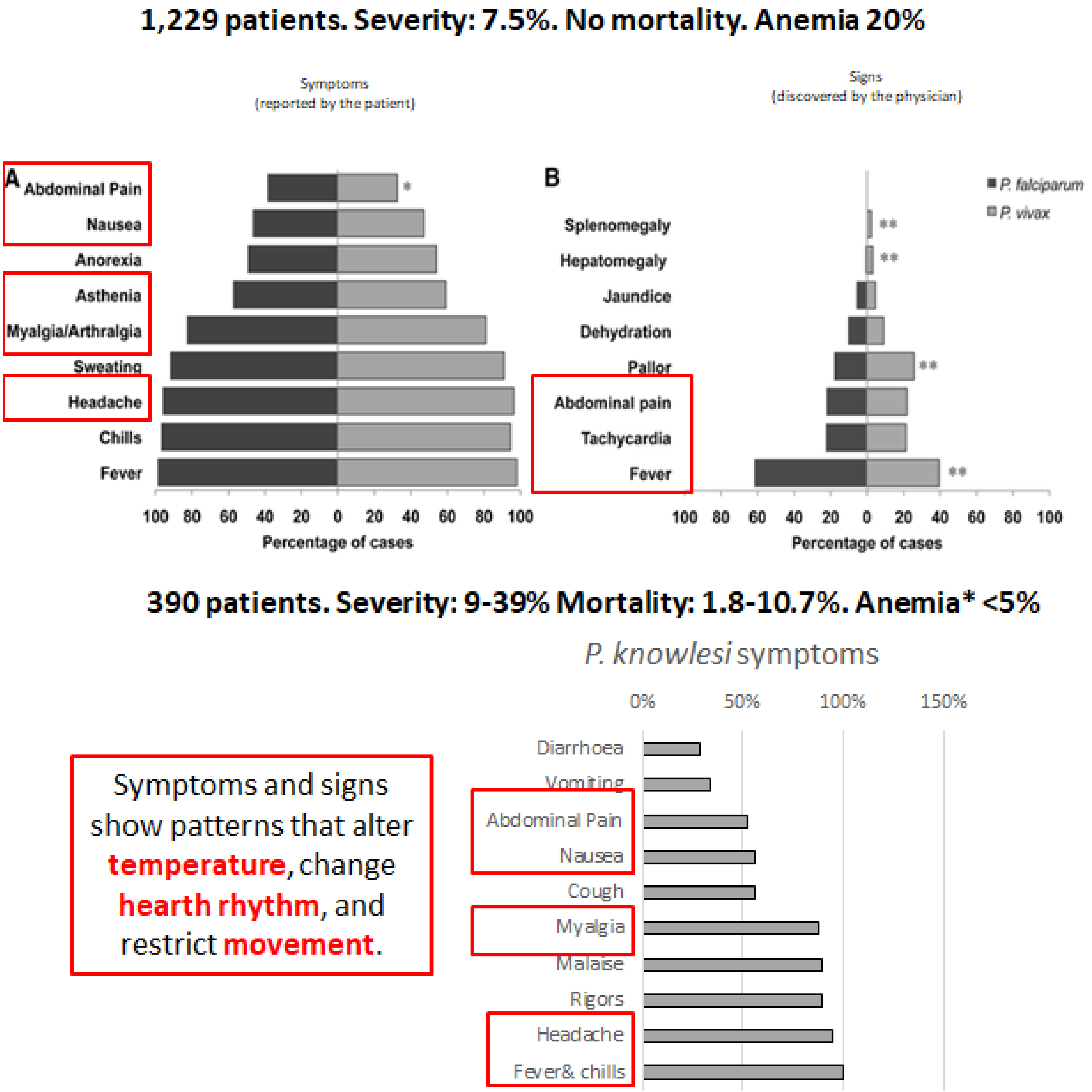
Figure 1.2: **Malaria Transmission Cycle**

This summarizes the standard transmission cycle for the malaria parasite. It enters a host via mosquito bite, replicates in the host's liver and then moves to the bloodstream to invade red blood cells. A new mosquito can then ingest some of the host's infected blood, becoming able to spread the parasite to its next host. Image credit: CDC, `https://www.cdc.gov/dpdx/malaria/index.html`, retrieved April 6, 2018.

Table 1.1: **Mild vs. Severe Malaria Symptoms**

Summary of the symptoms associated with mild (or uncomplicated) and severe malaria according to the CDC. Information retrieved from the following url on July 5, 2018: `https://www.cdc.gov/malaria/about/disease.html`

| **Mild** | **Severe** |
|---|---|
| Fever | Cerebral malaria (neurological issues, i.e. seizures and coma) |
| Chills | Severe anemia |
| Sweats | Hemoglobinuria (hemoglobin in the urine) |
| Headaches | Acute respiratory distress syndrome (ARDS) |
| Nausea and vomiting | Abnormalities in blood coagulation |
| Body aches | Low blood pressure |
| General malaise | Acute kidney failure |
| Weakness | Hyperparasitemia (parasites in more than 5% red blood cells) |
| Enlarged spleen and/or liver | Metabolic acidosis |
| Mild jaundice | Hypoglycemia |


1,229 patients. Severity: 7.5%. No mortality. Anemia 20%

390 patients. Severity: 9-39% Mortality: 1.8-10.7%. Anemia* <5%

*P. knowlesi* symptoms

Symptoms and signs show patterns that alter **temperature**, change **hearth rhythm**, and restrict **movement**.

Figure 1.3: **Basis for Malaria Interventions**

The above is a summary of some statistics from reported malaria cases, representing symptoms, severity, and mortality from all three of the *Plasmodium* species known to infect humans [23, 24].

## 1.4 MaHPIC and HAMMER

The *Malaria Host-Pathogen Interaction Center* (MaHPIC), founded in 2012, is a collaborative research effort involving investigators from Emory University, the Georgia Institute of Technology, the Centers for Disease Control and Prevention, and the University of Georgia, all funded by a combination of grants from the National Institute of Allergy and Infectious Diseases (NIAID) and the Defense Advanced Research Project Agency (DARPA). The research done in partnership with DARPA is part of DARPA's *Technologies for Host Resilience* program, specifically the *Host Acute Models of Malaria to study Experimental Resilience* project (THoR's HAMMER). The interdisciplinary group focuses on a systems biology approach to understanding the interactions between *Plasmodium* parasites and their hosts (primarily non-human primates) through a variety of carefully planned and executed experiments, as well as the collection and analysis of the resulting 'omic' datasets. The collaborative effort follows a central unifying hypothesis that "Non-Human Primate host interactions with *Plasmodium* pathogens as model systems will provide insights into mechanisms as well as indicators for human malarial disease conditions" [25]. This project utilizes transcriptomic data generated from a selection of these experiments for analysis, details of which will be given in the next chapter. Particularly, the experiments involved infecting *Macaca mulatta* hosts with either *Plasmodium cynomolgi*, *Plasmodium coatneyii*, or *Plasmodium knowlesi*. The last pathogen, *P. knowlesi*, was chosen because it is known to infect both humans and non-human primates [24]. Because *P. vivax* and *P. falciparum* require human hosts, *P. cynomolgi* and *P. coatneyi*, respectively, were chosen as the most similar species that also targets other primates [17].

# Chapter 2

# Methods and Materials

## 2.1 The SHIELD Pipeline

While many SNP analysis protocols begin by mapping raw reads back to a genome, SHIELD assumes that the user has access to the reference genome and reads that have already been mapped. This direction was chosen because it is becoming more popular in collaborative research to divide the workload, and a group responsible for analysis may not necessarily be the same group responsible for initial sequencing. While this may not be the case for all users, it is still frequently easier to add steps to the beginning of a protocol than it is begin in the middle of one. Therefore, SHIELD begins by asking the user where to find the reference genome and annotation, the mapped .bam files, the identifying tag for the species in question (particularly necessary if the reads are a mix of host and pathogen RNA), and where to set up output folders. The results of each step are stored in labeled folders so that the user has the freedom to analyze their data from any stage in the process. The folders also make it easier for the user to determine how far the program was able to run and continue running without starting over if it is interrupted for any reason. Figure 2.1 summarizes the flow of the steps taken and tools used by SHIELD.

The first few steps involve data processing in order to get the .bam files in a format best read by the SNP caller. First, samtools [26] indexes the reference genome as is required by future steps. SHIELD then sorts the contents of the files by chromosomal coordinate via Picard SortSam [27], because different mapping algorithms may leave the files sorted by other methods by default. Picard is used rather than samtools for this step because, in some cases, a user may have a mixed set of reads, some sorted by chromosomal coordinate

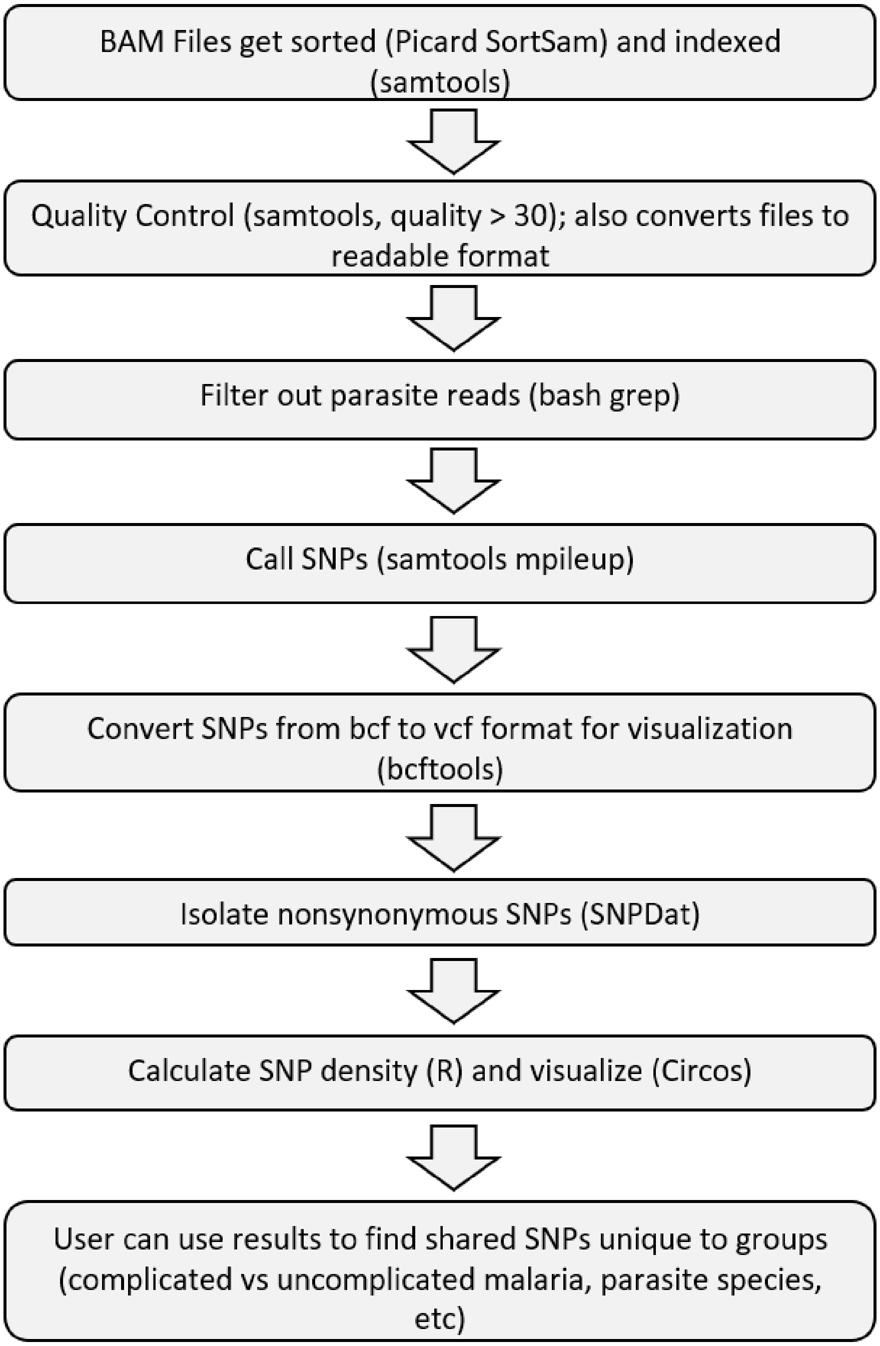

BAM Files get sorted (Picard SortSam) and indexed (samtools)
Quality Control (samtools, quality > 30); also converts files to readable format
Filter out parasite reads (bash grep)
Call SNPs (samtools mpileup)
Convert SNPs from bcf to vcf format for visualization (bcftools)
Isolate nonsynonymous SNPs (SNPDat)
Calculate SNP density (R) and visualize (Circos)
User can use results to find shared SNPs unique to groups (complicated vs uncomplicated malaria, parasite species, etc)

Figure 2.1: **SHIELD Pipeline Flowchart**

This chart displays the steps that SHIELD takes and the tools used to carry out each step.

and some by name. After a series of trials, Picard would reliably sort the given files by the parameter specified, even if they have already technically been sorted, while others would tend to return an error message and pause the program if asked to re-sort reads by the same parameter. Once sorted, the contents of the files are then indexed via the samtools index command. The sorted files have “ByLoc” added to the filenames and are stored in the first output folder (“1_sort”) alongside the indexed reads which are given the file extension “.bai”.

The samtools view command is then invoked to convert the sorted reads to a non-binary format while simultaneously removing any reads with a quality score less than 30. This value was chosen based on the quality of the mapped reads, but more advanced users can change the quality threshold based on their own data. The resulting files are stored in the folder “2_quality” and have “Q30” added to the filenames. Afterwards, a simple linux grep command is used to filter out any extraneous reads. This step is particularly necessary if fillers are used to aid the sequencing process, or if the samples in question are a host-pathogen combination. The user has the option to specify which read tag (that is, reads specific to the host or the pathogen) they wish to keep. The output from this step has “Filter” added to the end of the filename and is stored in the folder “3_filter.”

At this point, the reads are run through the samtools mpileup tool in order to generate the format necessary for the actual SNP calling. The resulting .bcf files (given the “Raw.bcf” filename addition) are stored in the folder “4_rawSNP.” The .bcf files are compressed and encoded with information derived from the original reads and the reference genome computing SNP likelihoods. Therefore, they are further run through the bcftools call command which does the actual SNP calling and converts the file to a readable format. The output file name is not changed, the extension is merely switched to “.vcf” and the files are stored in the folder “5_calledSNP.” These .vcf files list all detected SNPs, their genomic coordinates, what the normal and mutated alleles are, as well as some population genetics details for each locus (note that these details may be biased depending on the data being used. For the purpose of our test study, we disregarded that data due to an insufficient sample size). A program written in perl called SNPDat [28] is then invoked to compare the SNP files to the reference genome and annotation files. The output, tagged “SNPdat.txt” and stored in the folder “6_SNPdat,” lists whether each SNP is synonymous or nonsynonymous among other details. More linux text editing commands are invoked to trim the files down to only the nonsynonymous SNPs and the columns listing the chromosome name, genomic coordinate, genomic feature (mRNA, exon, CDS, etc), gene name, gene transcript number, reported SNP (as shown in the codon), and amino acid change, respectively. These text files simply have “N” added to their names, and are also stored in folder 6. An extra step further reduces the files to simply the chromosome names and genomic coordinates, adds “R” to the name and stores the resulting files in the folder “7_Rinputs.” An R script then takes those files and calculates the density of SNPs per 100,000 bp window of the genome. The resulting R outputs have “Density” attached to the beginning of their names and are moved to the folder “8_Routputs.” The density files are formated to be compatible with Circos plots for graphic visualization [29]. A package of the necessary documents are included with SHIELD along with simple instructions for generating the Circos plots for a given file. SHIELD can then continue to run some analyses on the resulting files based on the user’s needs. Particularly, the pipeline can make qualitative group comparisons identifying which SNVs are common across all subjects, which SNVs are unique to two user defined groups, and how common within the groups those unique SNVs may be. Table 2.1 summarizes the outputs and the folders in which they can be found, and Figure 2.1 summarizes the steps and tools that SHIELD uses.

Table 2.1: **SHIELD's Output Folders and Filenames**

Summary of the expected output after running SHIELD

| **Folder Name** | **New Filename** | **Description** |
|---|---|---|
| $inputFiles | test.bam | The original mapped data that a user starts with. $inputFiles is the variable under which SHIELD stores the user input for .bam file location. |
| 1_sort | testByLoc.bam | Mapped data sorted by chromosomal coordinate. |
|  | testByLoc.bam.bai | Sorted data is indexed. |
| 2_quality | testByLocQ30.bam | Reads with quality scores less than 30 are removed. |
| 3_filter | testByLocQ30Filter.bam | Only reads for the target organism are kept. |
| 4_rawSNP | testByLocQ30FilterRaw.bcf | Binary output from samtools mpileup. |
| 5_calledSNP | testByLocQ30FilterRaw.vcf | Readable called SNPs. |
| 6_SNPdat | testByLocQ30FilterRawSNPdat.txt | Raw output from SNPdat listing synonymous SNPs and amino acid changes. |
|  | testByLocQ30FilterRawSNPdatN.txt | Trimmed SNPdat output, listing only the columns of interest. |
| 7_Rinputs | testByLocQ30FilterRawSNPdatNR.txt | Only lists chromosome name and SNP positions for density calculations. |
| 8_Routputs | DensitytestByLocQ30FilterRawSNPdatNR.txt | Lists SNP density per 100,000 bp window in a Circos-friendly format. |

## 2.2 The Experimental Data

*Macaca mulatta* (rhesus macaque) RNA-Seq reads that had previously been mapped back to a reference genome, as provided by the MaHPIC Consortium, were the starting point for this analysis. These reads represented transcripts from five different monkeys at peak parasitemia from MaHPIC Experiment 04 (E04). The monkeys in question had been exposed to identically cloned strains of either *Plasmodium cynomolgi*, after which various factors such as temperature, red blood cell count, parasite count, and activity were all monitored over the course of the infection (Figure 2.3) [30]. Two of the subjects were classified as expressing mild malaria symptoms, while the other three were severe. Figure 2.2 shows the summary of results published by the experimenters. Of particular note is the fact that the parasitemia levels during the initial peak remained similar for both mild and severe subjects. The only environmental difference between the two groups was the timing of the anti-malarial treatment (the severe subjects required an earlier administration), and that one was experiencing renal failure despite the treatments and required euthanasia.

Several more experiments were performed by MaHPIC and HAMMER using macaque monkeys infected with either *P. cynomolgi*, *P. knowelsi*, or *P. coatneyi*. Table 2.2 shows a summary of the various experiments and their designation, as either primary or secondary infections, the precise strain of *Plasmodium* used, and the NCBI accession numbers for the transcript data used (some transcripts were not yet assigned accession numbers at this time). Experiments 23, 24, and 25 were studying infection and reinfection with the same monkeys, and for the purposes of our analyses were counted as three different sets of monkeys. Experiment 23 lists one extra subject than experiments 24 or 25 because one subject, RJn13, was removed from the program after experiment 23 due to behavioral issues. Figure 2.3 summarizes the general experimental procedures used in each of the experiments referenced here. All experimental designs were approved by the Emory University Institutional Animal Care and Use Committee (IACUC). Experiments 6 and 30 were further approved by the MRMC

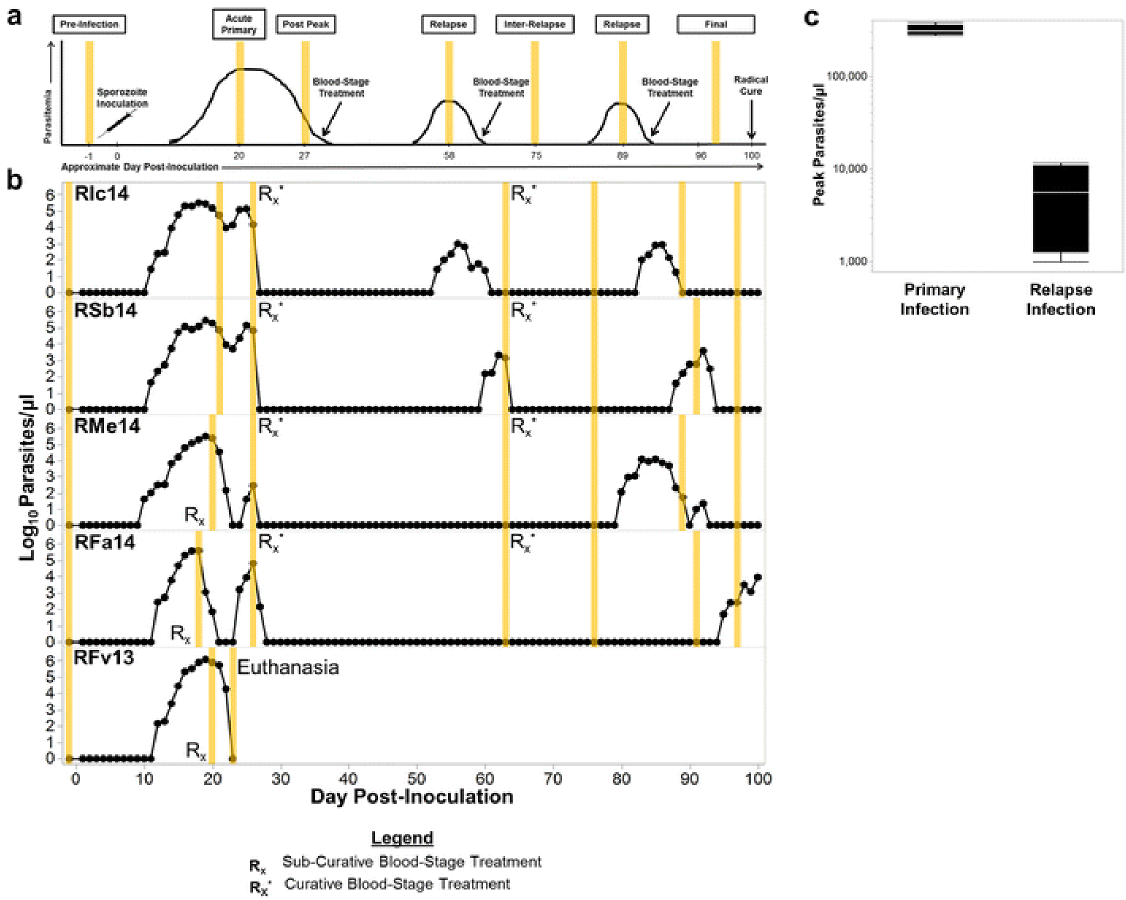


Figure 2.2: **MaHPIC E04 Summary**

Experimental summary of the five *M. mulatta* subjects used in Experiment 04. The black line shows the log parasitemia count over time (determined via daily or twice daily finger pricks), while the vertical yellow bars show the time points at which blood was drawn for sequencing. The Rx symbols show when curative treatments were administered to the subjects. From this, subjects RIc14 and RSb14 were considered to have mild malaria, while the other three were considered severe cases due to the necessity for extra treatments or euthanasia.

Office of Research Protection Animal Care and Use Review Office (ACURO). Information about these experiments was summarized from their PlasmoDB entries.

The different treatment procedures require further clarification. Subcurative treatments were always administered via artemether, and in just enough of a dosage to prevent fatal complications without completely clearing the parasite from the bloodstream. Curative treatments in the form of higher artemether dosages do clear the parasite from the blood, but not the liver, allowing relapses. Curative treatments via primaquine and chloroquine completely clear the parasite from both the blood and liver. Only one subject (REd16) in E30 received a subcurative treatment with artemether to avoid clinical complications, but no other treatments were administered. The subjects in E03 received artemether in subcurative doses during peak parasitemia and additionally as needed, then in curative doses at the end of the experiment. Subjects in E04 and E23 had the same treatment protocol; both were given subcurative treatments during the initial parasitemia peak, followed by curative artemether treatments to clear the blood after the initial infection in order to detect relapses, then curative treatments with primaquine and chloroquine at the end of the experiment. E25 followed the same protocol, but without the curative artemether treatments. Because E24 was designed to study potential relapses in reinfected individuals, only the final curative treatment of primaquine and chloroquine was necessary.

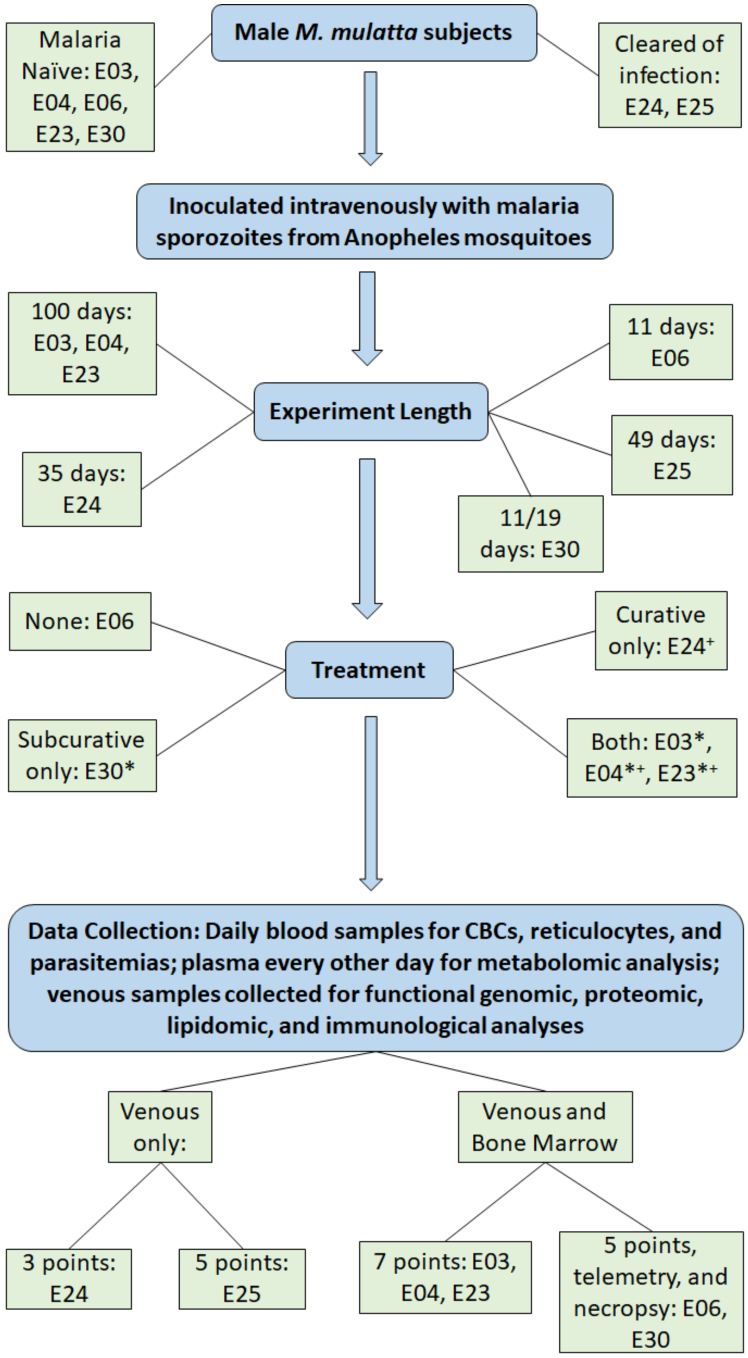

Male *M. mulatta* subjects
Malaria Naïve: E03, E04, E06, E23, E30
Cleared of infection: E24, E25
Inoculated intravenously with malaria sporozoites from Anopheles mosquitoes
100 days: E03, E04, E23
35 days: E24
Experiment Length
11 days: E06
49 days: E25
11/19 days: E30
None: E06
Subcurative only: E30*
Treatment
Curative only: E24$^{+}$
Both: E03*, E04*$^{+}$, E23*$^{+}$
Data Collection: Daily blood samples for CBCs, reticulocytes, and parasitemias; plasma every other day for metabolomic analysis; venous samples collected for functional genomic, proteomic, lipidomic, and immunological analyses
Venous only:
Venous and Bone Marrow
3 points: E24
5 points: E25
7 points: E03, E04, E23
5 points, telemetry, and necropsy: E06, E30

Figure 2.3: **Summarization of the MaHPIC Experimental Procedure**

The above flowchart summarizes the key similarities and differences in the procedures for the experiments used in this analysis. The central flow represents the core steps taken in each experiment: selection and inoculation of male *M. mulatta* subjects, a number of days with infection, some form of treatment, and data collection. (*) treatment with artemether, either curative or subcurative. (+) final curative treatment with primaquine and chloroquine.

Table 2.2: **Total Experimental Summary**

Summarization of the MaHPIC experiments and subjects used in this study. The first row for each experiment lists the infection type (either primary or secondary) and the specific *Plasmodium* species and strain used. The designation used for each subject is provided alongside the NCBI accession number for the transcript data, if it is currently publicly available. This study used data from acute parasitemia time points.

| | | |
|---|---|---|
| E03 | Primary | *P. coatneyi*, Hackeri strain |
| | RCs13 | GSM2759334 |
| | RTi13 | GSM2759335 |
| | RUn13 | GSM2759336 |
| | RWr13 | GSM2759337 |
| | RZe13 | GSM2759338 |
| E04 | Primary | *P. cynomolgi*, B/M strain |
| | RIc14 | GSM2772567 |
| | RSb14 | GSM2772573 |
| | RMe14 | GSM2772577 |
| | RFa14 | GSM2772584 |
| | RFv13 | GSM2772587 |
| E06 | Primary | *P. knowlesi*, Malayan strain |
| | RUf16 | not yet available |
| | RIh16 | not yet available |
| | RTe16 | not yet available |
| | RCl15 | not yet available |
| E23 | Primary | *P. cynomolgi*, B/M strain |
| | RAd14 | GSM2792852 |
| | RBg14 | GSM2792858 |
| | RIb13 | GSM2792863 |
| | RJn13 | GSM2792868 |
| | ROc14 | GSM2792875 |
| | ROh14 | GSM2792882 |
| E24 | Secondary | *P. cynomolgi*, B/M strain |
| | RAd14 | GSM2789835 |
| | RIb13 | GSM2789836 |
| | ROh14 | GSM2789837 |
| | ROc14 | GSM2789838 |
| | RBg14 | GSM2789839 |
| E25 | Secondary | *P. cynomolgi*, Ceylonensis strain |
| | ROh14 | GSM2795511 |
| | RAd14 | GSM2795512 |
| | RIb13 | GSM2795513 |
| | ROc14 | GSM2795514 |
| | RBg14 | GSM2795515 |
| E30 | Primary | *P. knowlesi*, Malayan strain |
| | RKy15 | not yet available |
| | REd16 | not yet available |

# Chapter 3

# Results

This chapter details the results found from the analyses performed, which are split into three major sections. Section 3.1 involves SNVs found to be common across all subjects. Variants that were found to be unique to either the mild or severe subjects are shown in section 3.2. Finally, section 3.3 provides a summary of SNV results from each of the subjects for comparison to the other analyses results.

## 3.1 Common SNVs

The nonsynonymous SNVs were collected and compared by experiment in order to identify genes for which all of the monkeys had some reported SNV. Thus, genes that had at least one SNV for all monkeys within each experiment were identified (3.1, 3.2, 3.3, 3.4, 3.5). Those intersections were then compared to find the genes that were common across all monkeys (3.6).

For Figure 3.6, it should be noted that analyses were run on 32 monkeys across seven experiments. However, venn diagrams lose legibility with more than five subjects. Therefore, for the sake of visualizing the process, a diagram is provided below showing the intersection of genes from E03, E04, E06, E23, and E30, omitting E24 and E25. The image shows 1,776 common genes, while the actual intersection contains 1,657 genes. Figure 3.7 shows the density of SNVs from within these common genes, calculated per 100,000 bp, to give a quick look at their distribution across the genome. Already it is clear that the end of chromosome 19, which corresponds with many immune response-related genes in particular, is the region with the highest SNV density.

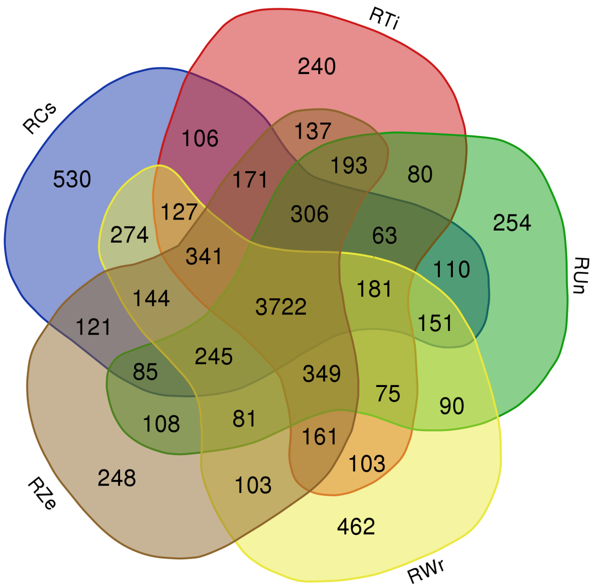


Figure 3.1: **E03 Gene Intersection**

The numbers indicate the number of genes with at least one nonsynonymous SNV reported within it. The diagram shows the various intersections of said genes for each monkey, with the central section showing that all five subjects reported at least one SNV in the same 3,722 genes. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

Gene set enrichment analysis (GSEA) of the 1,657 common genes against the Broad Institute's Molecular Signatures Database (MSigDB) curated gene sets revealed that they were significantly associated with immune system sets [31, 32, 33]. A similar analysis of the common genes against the MSigDB immune system sets revealed that many of the significantly enriched genes were involved in anti-viral responses. The top ten results from both enrichments are listed in Table 3.1, with the blue italicized entries highlighting the relevant gene sets. Other projects in this lab performed a similar analysis on genes that were significantly up-regulated for all primary infection (that is, excluding E24 and E25) *M. mulatta* subjects and found comparable results (Table 3.2). A comparison between the two gene lists revealed that all *M. mulatta* subjects were both significantly up-regulated and had

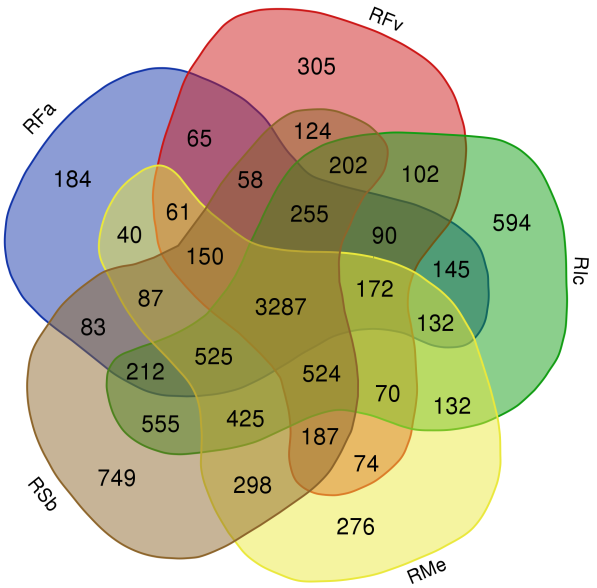


Figure 3.2: **E04 Gene Intersection**

The numbers indicate the number of genes with at least one nonsynonymous SNV reported within it. The diagram shows the various intersections of said genes for each monkey, with the central section showing that all five subjects reported at least one SNV in the same 3,287 genes. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

SNVs in the same 28 genes (Figure 3.8). Figure 3.9 shows the 28 identified genes and their expression levels in each of the primary infection subjects. Contrary to expectations, not all of the identified genes are located in regions of high SNV density (Table 3.3).

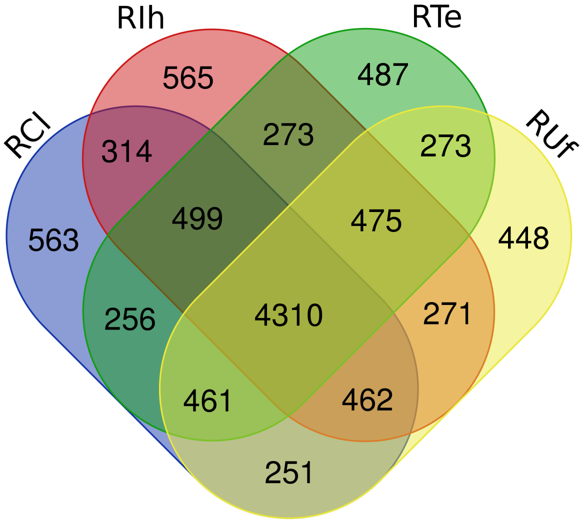


Figure 3.3: **E06 Gene Intersection**

The numbers indicate the number of genes with at least one nonsynonymous SNV reported within it. The diagram shows the various intersections of said genes for each monkey, with the central section showing that all five subjects reported at least one SNV in the same 4,310 genes. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

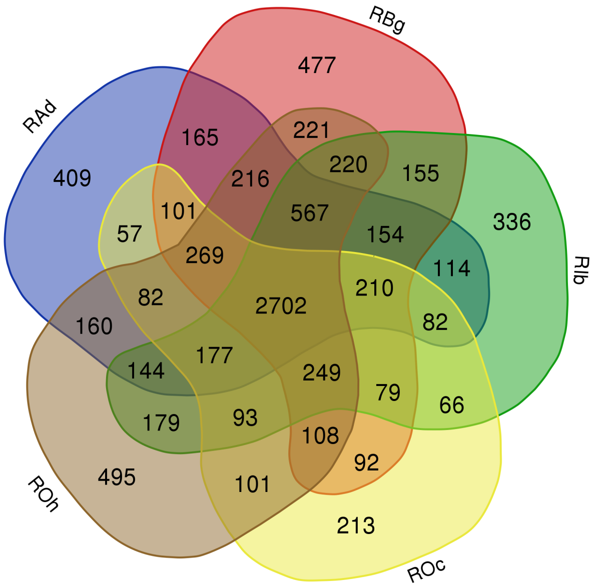


Figure 3.4: **E23 Gene Intersection**

The numbers indicate the number of genes with at least one nonsynonymous SNV reported within it. The diagram shows the various intersections of said genes for each monkey, with the central section showing that all five subjects reported at least one SNV in the same 2,702 genes. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

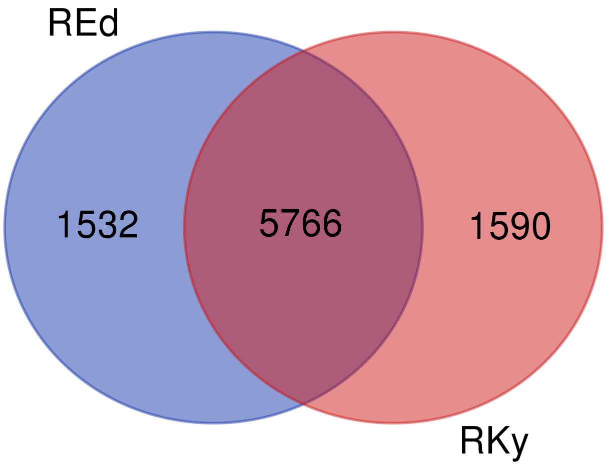


Figure 3.5: **E30 Gene Intersection**

The numbers indicate the number of genes with at least one nonsynonymous SNV reported within it. The diagram shows the various intersections of said genes for each monkey, with the central section showing that all five subjects reported at least one SNV in the same 5,766 genes. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

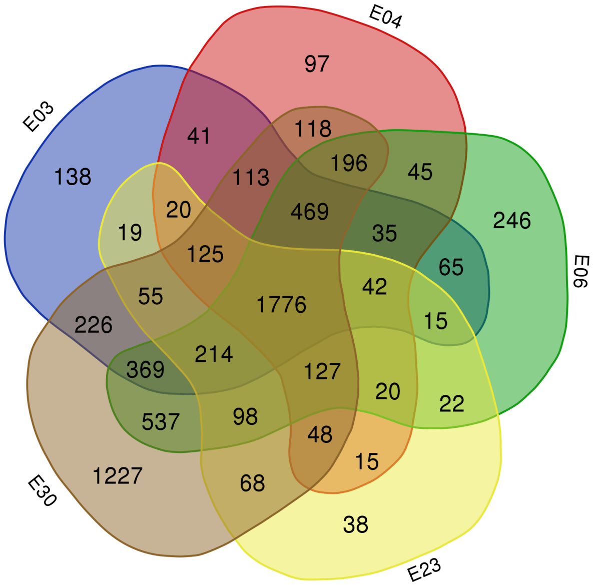


Figure 3.6: **Gene Intersection for Five Experiments**

The intersections mentioned above can be seen compared here, showing 1,776 genes common across all monkeys in these five experiments. This diagram omits results from E24 and E25, even though they were included in all analyses, for ease of visualization. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

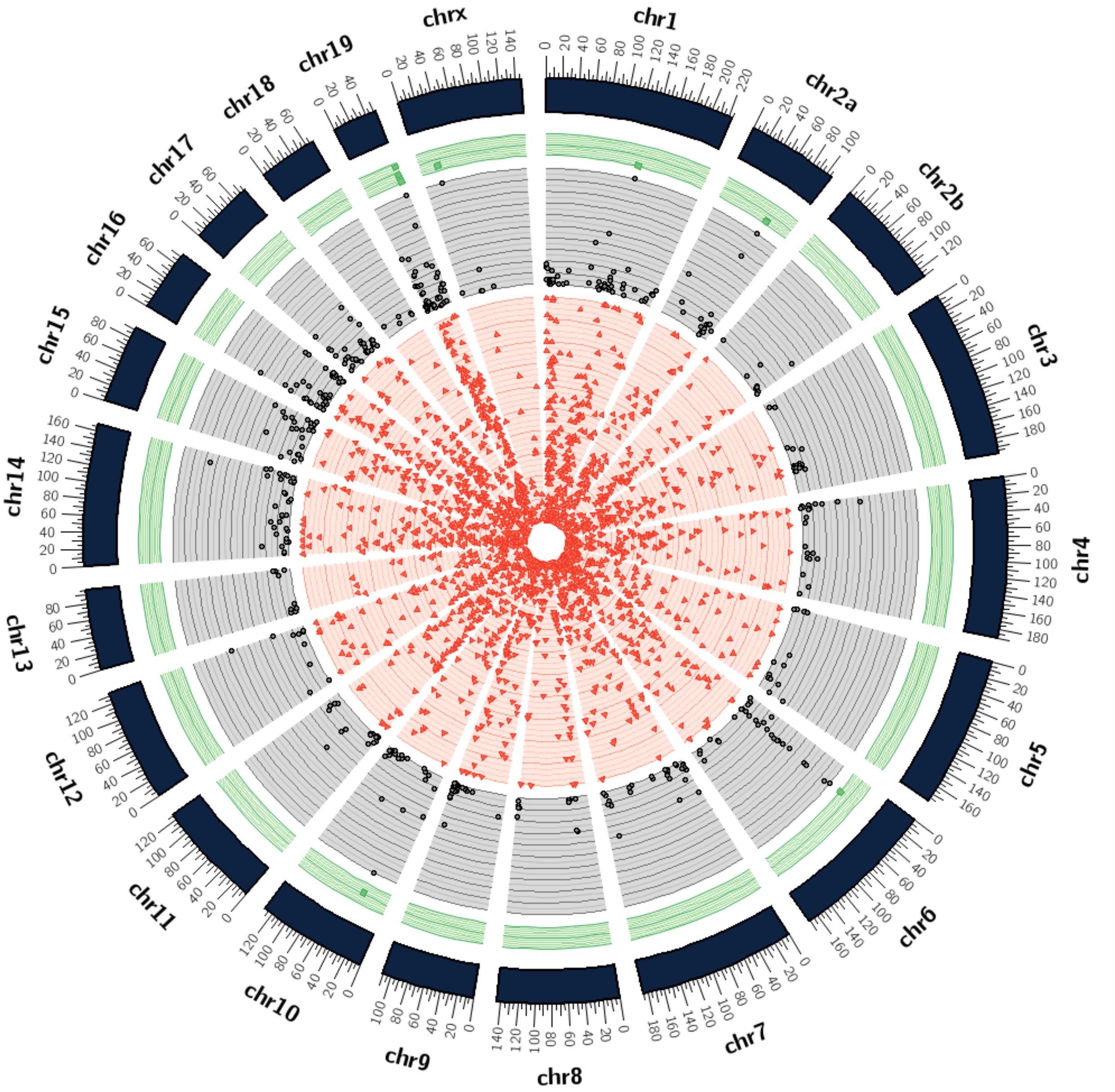


Figure 3.7: **Genome Distribution of SNV Density from Genes Common Across All Experiments**

The density of SNVs per 100,000 bp window when only considering SNVs in the 1,657 previously reported to be common across all monkeys. The red innermost ring represents densities between 0 and 0.01, the middle gray ring shows densities between 0.011 and 0.099, and the green outermost ring displays densities of 0.1 and higher.

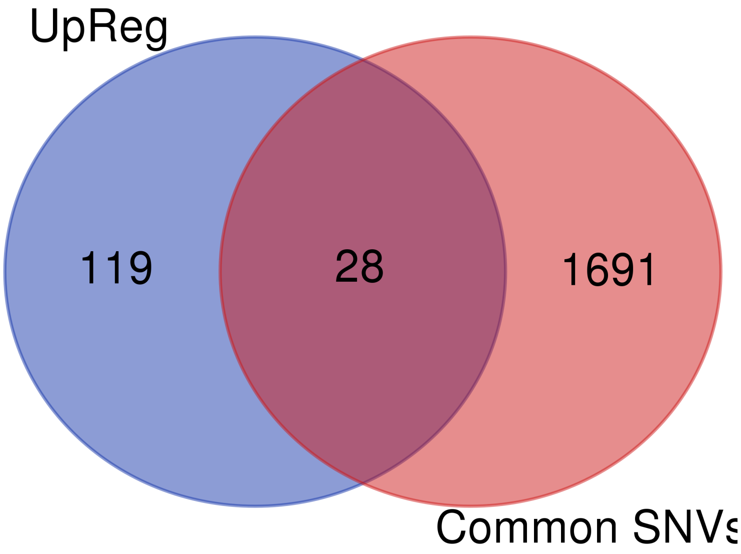


Figure 3.8: **Intersection of Up-Regulated Genes and Genes With Common SNVs**

This displays the intersection between the genes for which every subject reported at least one SNV and the genes that were reported to be significantly up-regulated. Generated on May 28, 2018 via the online tool developed by Ghent University, available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

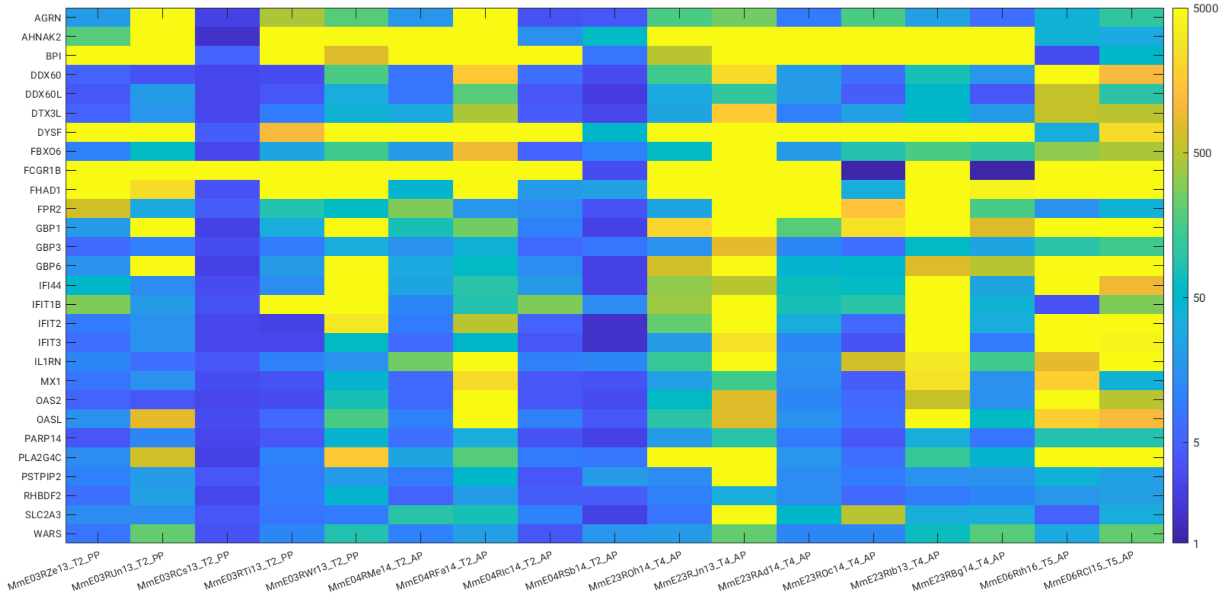


Figure 3.9: **Absolute Expression Fold-Change in 28 Genes**

A separate analysis found a list of 151 genes with significantly up-regulated expression from all primary infection *M. mulatta* subjects. This heatmap shows the absolute expression levels for the 28 genes for which all subjects also reported at least one SNV.

Table 3.1: **Gene Set Enrichment Analysis for Genes with Common SNVs**

The top ten GSEA results for genes with common SNVs across experiments when compared to both the curated gene sets and the immune system sets. Sets of interest, immune system in the first section and anti-viral in the second, are shown in blue and italicized.

| **Enrichment of Genes with Common SNVs** |
|---|
| **Curated Gene Sets** |
| *REACTOME IMMUNE SYSTEM* |
| *REACTOME CYTOKINE SIGNALING IN IMMUNE SYSTEM* |
| *REACTOME ADAPTIVE IMMUNE SYSTEM* |
| *REACTOME INTERFERON SIGNALING* |
| REACTOME HEMOSTASIS |
| REACTOME METABOLISM OF LIPIDS AND LIPOPROTEINS |
| KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY |
| KEGG AMINOACYLE TRNA BIOSYNTHESIS |
| KEGG ENDOCYTOSIS |
| REACTOME TRNA AMINOACYLATION |
| **Immunologic Signatures** |
| GSE369 PRE VS POST IL6 INJECTION SOCS3 KO LIVER |
| GSE22886 DAY0 VS DAY7 MONOCYTE IN CULTURE |
| GSE22886 CD8 TCELL VS BCELL NAIVE |
| *GSE24671 CTRL VS SENDAI VIRUS INFECTED MOUSE SPLENOCYTES* |
| *GSE29618 PDC VS MDC DAY7 FLU VACCINE* |
| *GSE10325 LUPUS BCELL VS LUPUS MYELOID* |
| GSE22886 NAIVE BCELL VS MONOCYTE |
| GSE3982 NEUTROPHIL VS CENT MEMORY CD4 TCELL |
| *GSE7548 NAIVE VS DAY7 PCC IMMUNIZATION CD4 TCELL* |
| *GSE10325 LUPUS CD4 TCELL VS LUPUS MYELOID* |

Table 3.2: **Gene Set Enrichment Analysis for Commonly Up-Regulated Genes**

The top GSEA results for genes that were commonly significantly up-regulated when compared to both the curated gene sets (top) and the immune system sets (bottom). Sets of interest, immune system in the first section and anti-viral in the second, are shown in blue and italicized.

| **Enrichment of Genes Significantly Up-Regulated** |
|---|
| **Curated Gene Sets** |
| *REACTOME INTERFERON SIGNALING* |
| *REACTOME CYTOKINE SIGNALING IN IMMUNE SYSTEM* |
| *REACTOME IMMUNE SYSTEM* |
| *REACTOME INTERFERON ALPHA BETA SIGNALING* |
| *REACTOME INTERFERON GAMMA SIGNALING* |
| PID IL4 2PATHWAY |
| *REACTOME ANTIVIRAL MECHANISM BY IFN STIMULATED GENES* |
| NABA MATRISOME ASSOCIATED |
| NABA MATRISOME |
| *REACTOME INNATE IMMUNE SYSTEM* |
| *REACTOME RIG I MDA5 MEDIATED INDUCTION OF IFN ALPHA BETA PATHWAY* |
| **Immunologic Signatures** |
| *GSE13485 DAY3 VS DAY7 YF17D VACCINE PBMC DN* |
| *GSE13485 CTRL VS DAY7 YF17D VACCINE PBMC DN* |
| *GSE13485 DAY1 VS DAY7 YF17D VACCINE PBMC DN* |
| *GSE13485 CTRL VS DAY3 YF17D VACCINE PBMC DN* |
| GSE42724 NAIVE BCELL VS PLASMABLAST UP |
| *GSE14000 UNSTIM VS 4H LPS DC DN* |
| *GSE14000 UNSTIM VS 4H LPS DC TRANSLATED RNA DN* |
| *GSE18791 UNSTIM VS NEWCASTLE VIRUS DC 10H DN* |
| *GSE18791 UNSTIM VS NEWCASTLE VIRUS DC 6H DN* |
| *GSE13485 PRE VS POST YF17D VACCINATION PBMC DN* |

Table 3.3: **Locations for the Genes of Interest**

Below is a list of the 28 genes found in the intersection of the two analyses (Common SNVs and significant up-regulation of expression), and the chromosomes where they are found. When compared with Figure 3.7, some genes appear in regions of high SNV density, as would be expected, but several do not.

| **Gene** | **Chromosome** | **Gene** | **Chromosome** |
|---|---|---|---|
| AGRN | Chr01 | AHNAK2 | Chr14 |
| BPI | Chr15 | DDX60 | Chr04 |
| DDX60L | Chr04 | DTX3L | Chr03 |
| DYSF | Chr02a | FBXO6 | Chr01 |
| FCGR1B | Chr01 | FHAD1 | Chr01 |
| FPR2 | Chr19 | GBP1 | Chr01 |
| GBP3 | Chr01 | GBP6 | Chr01 |
| IFI44 | Chr01 | IFIT1B | Chr10 |
| IFIT2 | Chr10 | IFIT3 | Chr10 |
| IL1RN | Chr02a | MX1 | Chr07 |
| OAS2 | Chr12 | OASL | Chr12 |
| PARP14 | Chr03 | PLA2G4C | Chr19 |
| PSTPIP2 | Chr18 | RHBDF2 | Chr17 |
| SLC2A3 | Chr12 | WARS | Chr14 |

## 3.2 Mild VS Severe SNVs

The subjects were divided into two groups based on whether they exhibited mild or severe symptoms in the course of their experiments. Results from E24 and E25 were omitted from this particular analysis as they were performed with the same subjects as E23. Coincidentally, there were then 11 mild and 11 severe subjects. Table 3.4 lists how each subject was categorized. A list was compiled of all SNVs expressed by the mild group and compared to a similar list from the severe group, and the SNVs unique to each group were identified. The genes containing the unique SNVs were identified, and GSEA was performed on both sets against the MSigDB curated gene sets. Unlike the previous analysis, the results from this initial enrichment did not reveal an obvious subset to focus on. Also, contrary to the initial expectation, there was no clear separation between the mild and severe enrichment results; both displayed enrichment in the same types of gene sets, namely those involved in the matrisome and transmembrane proteins. Table 3.5 provides a summary of the GSEA results for genes uniquely mutated between the mild and severe subjects.

Table 3.4: **Mild and Severe Subjects**

The subjects below come from E03, E04, E06, E23, and E30. Their categorization was determined by their clinical symptoms.

| **Mild** | **Severe** |
|---|---|
| RAd14 | RCl15 |
| RBg14 | REd16 |
| RCs13 | RFa14 |
| RIb13 | RFv13 |
| RIc14 | RIh16 |
| RJn13 | RKy15 |
| ROc14 | RMe14 |
| ROh14 | RTe16 |
| RSb14 | RTi13 |
| RWr13 | RUf16 |
| RZe13 | RUn13 |

Although GSEA revealed little difference between the mild and severe groups, SNV density proved to be quite different. The density of the SNVs unique to each group was calculated

Table 3.5: **GSEA on Genes with SNVs Unique to Mild or Severe Subjects**

GSEA against MSigDB's curated gene sets revealed little difference between the mild (top) and severe (bottom) groups' uniquely mutated genes. Both display enrichment for the same gene sets, namely those involving the matrisome and other transmembrane proteins.

| **MILD** |
|---|
| NABA MATRISOME |
| REACTOME GPCR LIGAND BINDING |
| NABA MATRISOME ASSOCIATED |
| KEGG SYSTEMIC LUPUS ERYTHEMATOSUS |
| REACTOME SIGNALING BY GPCR |
| REACTOME GPCR DOWNSTREAM SIGNALING |
| REACTOME RNA POL I PROMOTER OPENING |
| PID RB 1PATHWAY |
| NABA SECRETED FACTORS |
| KEGG PRION DISEASES |

| **SEVERE** |
|---|
| NABA MATRISOME |
| NABA MATRISOME ASSOCIATED |
| NABA SECRETED FACTORS |
| REACTOME GPCR LIGAND BINDING |
| KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION |
| REACTOME SIGNALING BY GPCR |
| REACTOME CLASS A1 RHODOPSIN LIKE RECEPTORS |
| KEGG RIBOSOME |
| REACTOME METABOLISM OF PROTEINS |
| REACTOME PEPTIDE LIGAND BINDING RECEPTORS |

and displayed on Circos plots to show their distributions across the genome (Figures 3.10 and 3.11). While both groups appear to contain SNVs in genes involved in the same processes, the severe group contains far more SNVs within more genes than the mild group.

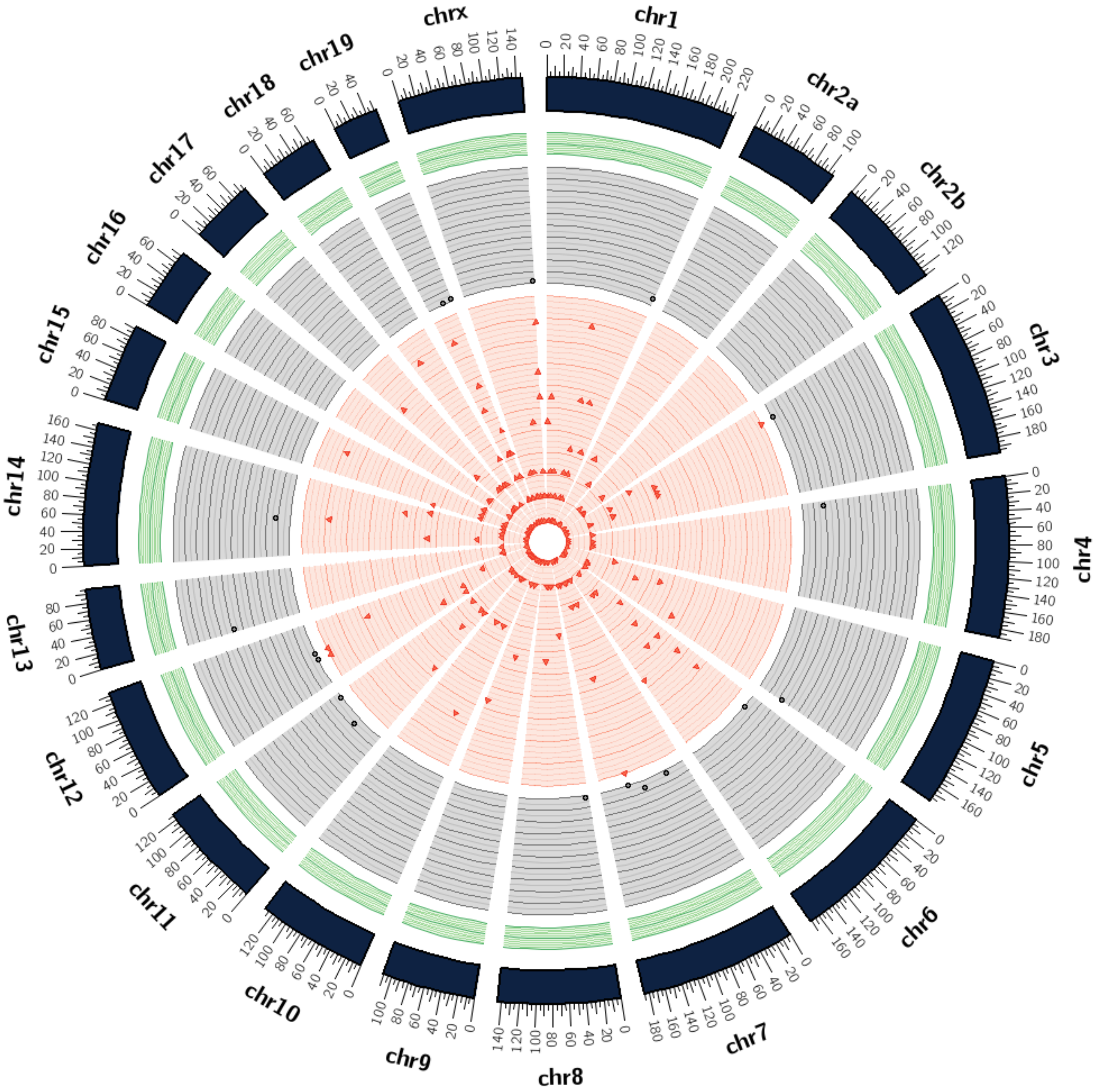


Figure 3.10: **Density of SNVs Unique to Mild Subjects**

This Circos plot shows the distribution of SNV density across the genome for the SNVs unique to the 11 mild subjects. Density was calculated per 100,000 bp. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

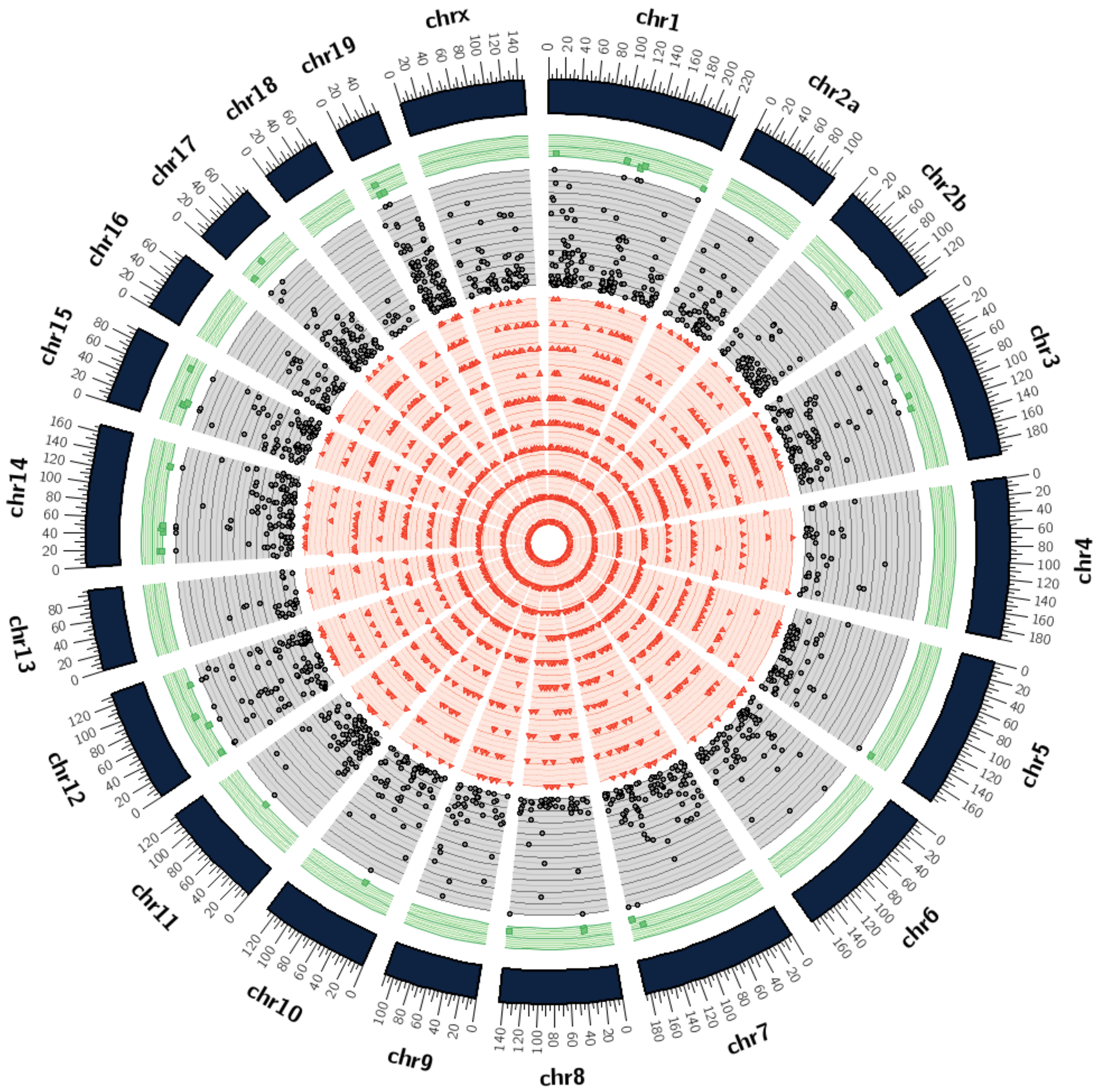


Figure 3.11: **Density of SNVs Unique to Severe Subjects**

This Circos plot shows the distribution of SNV density across the genome for the SNVs unique to the 11 severe subjects. Density was calculated per 100,000 bp. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

## 3.3 The Individual Subjects

The following series of Circos plots show the general SNV densities for each of the individuals used in this project. These densities were calculated with all of the nonsynonymous SNVs reported for a given individual without filtering for shared or unique results. These plots are provided for completeness and for comparison with the results provided above. While each individual has a slightly different SNV density distribution, they all share a peak at chromosome 19, which validates the peak seen when comparing the common SNVs (Figure 3.7).

### 3.3.1 Experiment 03

The following plots (Figures 3.12, 3.13, 3.14, 3.15, and 3.16) represent whole genome SNV densities per 100,000 base pairs in each of the five *M. mulatta* subjects infected with *P. coatneyi* malaria.

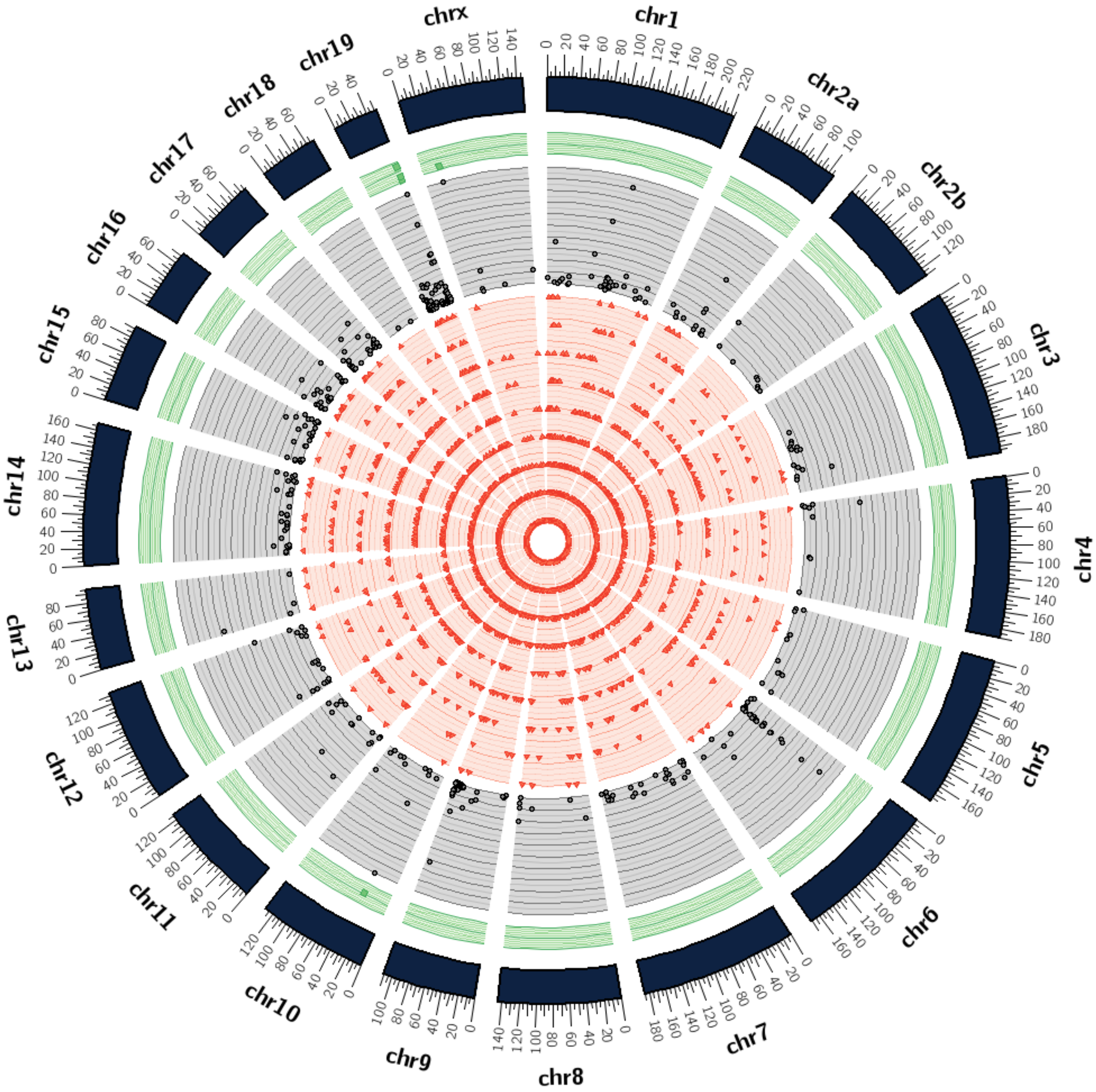


Figure 3.12: **Subject RCs13, peak parasitemia timepoint, E03**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

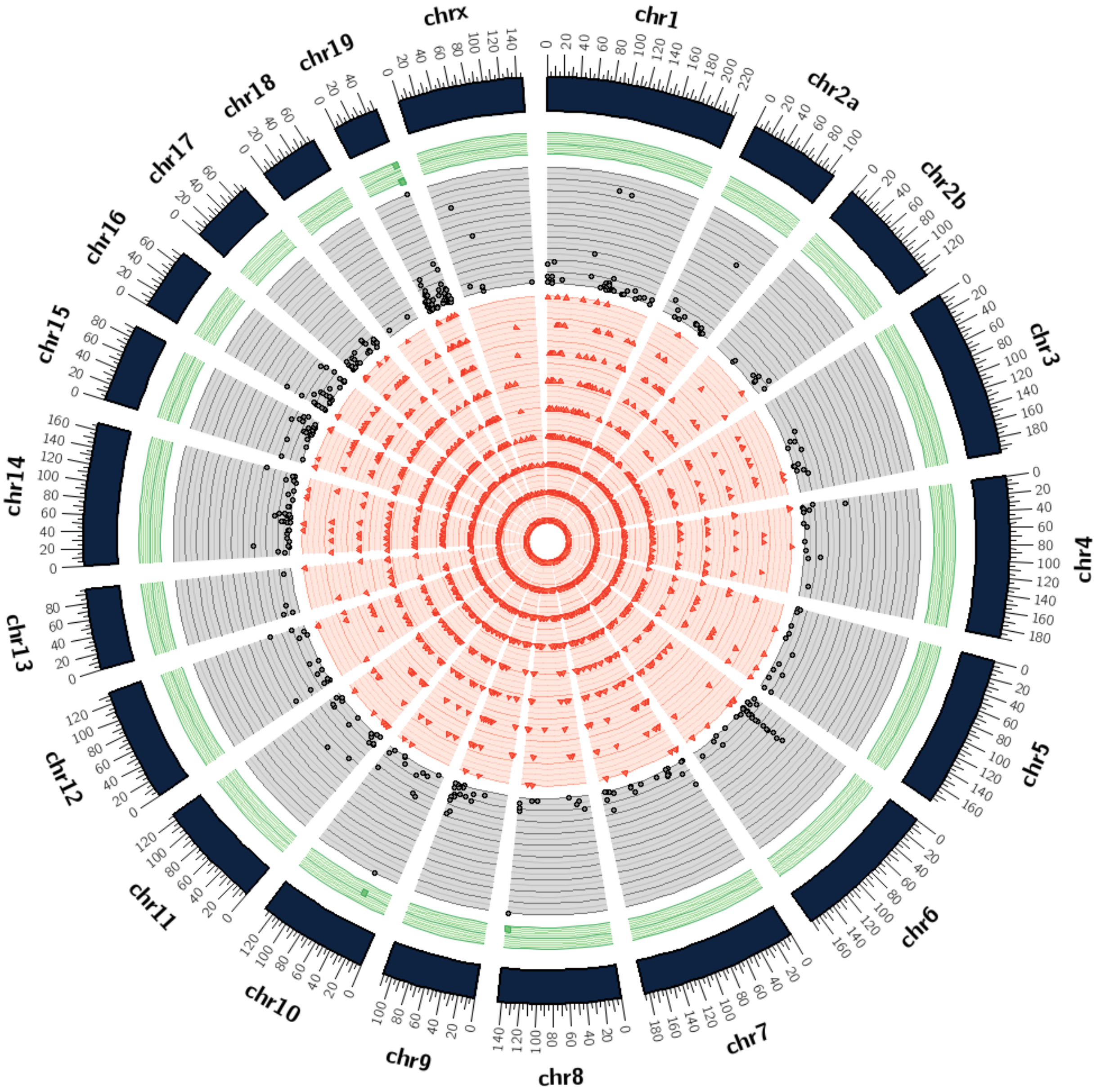


Figure 3.13: **Subject RTi13, peak parasitemia timepoint, E03**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

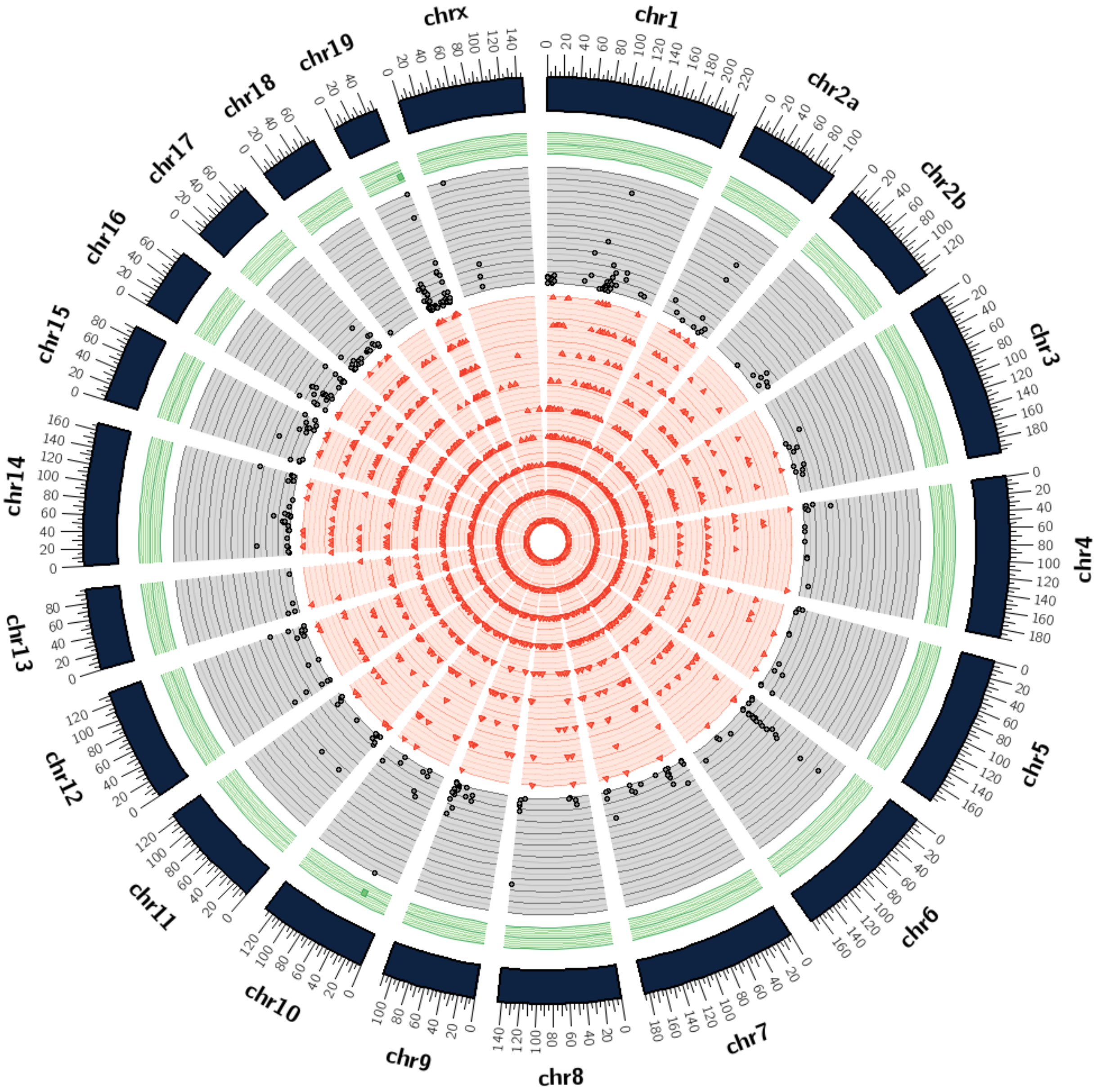


Figure 3.14: **Subject RUn13, peak parasitemia timepoint, E03**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 10 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

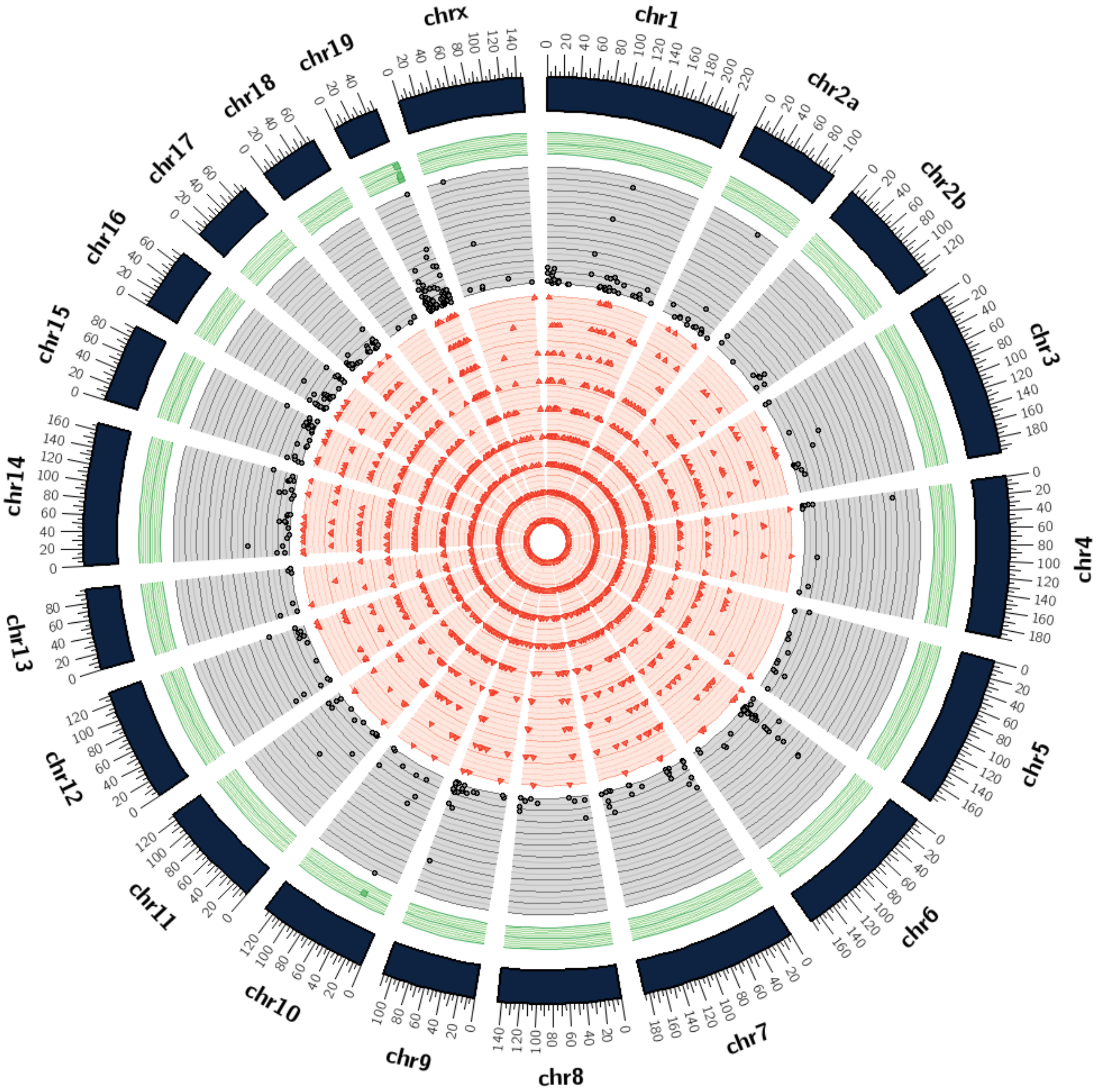


Figure 3.15: **Subject RWr13, peak parasitemia timepoint E03**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 10 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

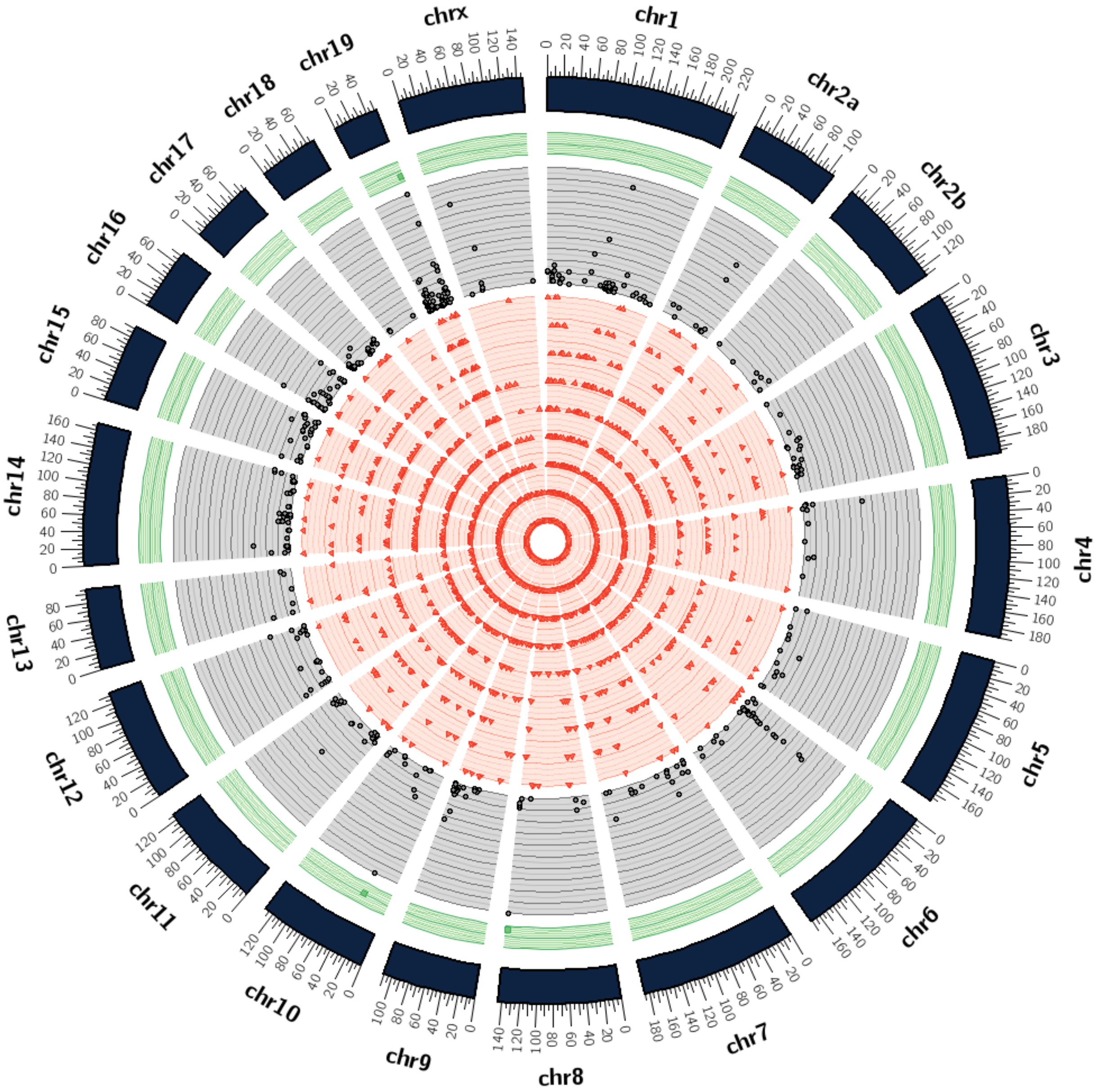


Figure 3.16: **Subject RZe13, peak parasitemia timepoint, E03**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

### 3.3.2 Experiment 04

The following plots (Figures 3.17, 3.18, 3.19, 3.20, and 3.21) represent whole genome SNV densities per 100,000 base pairs in each of the five *M. mulatta* subjects infected with *P. cynomolgi* malaria.

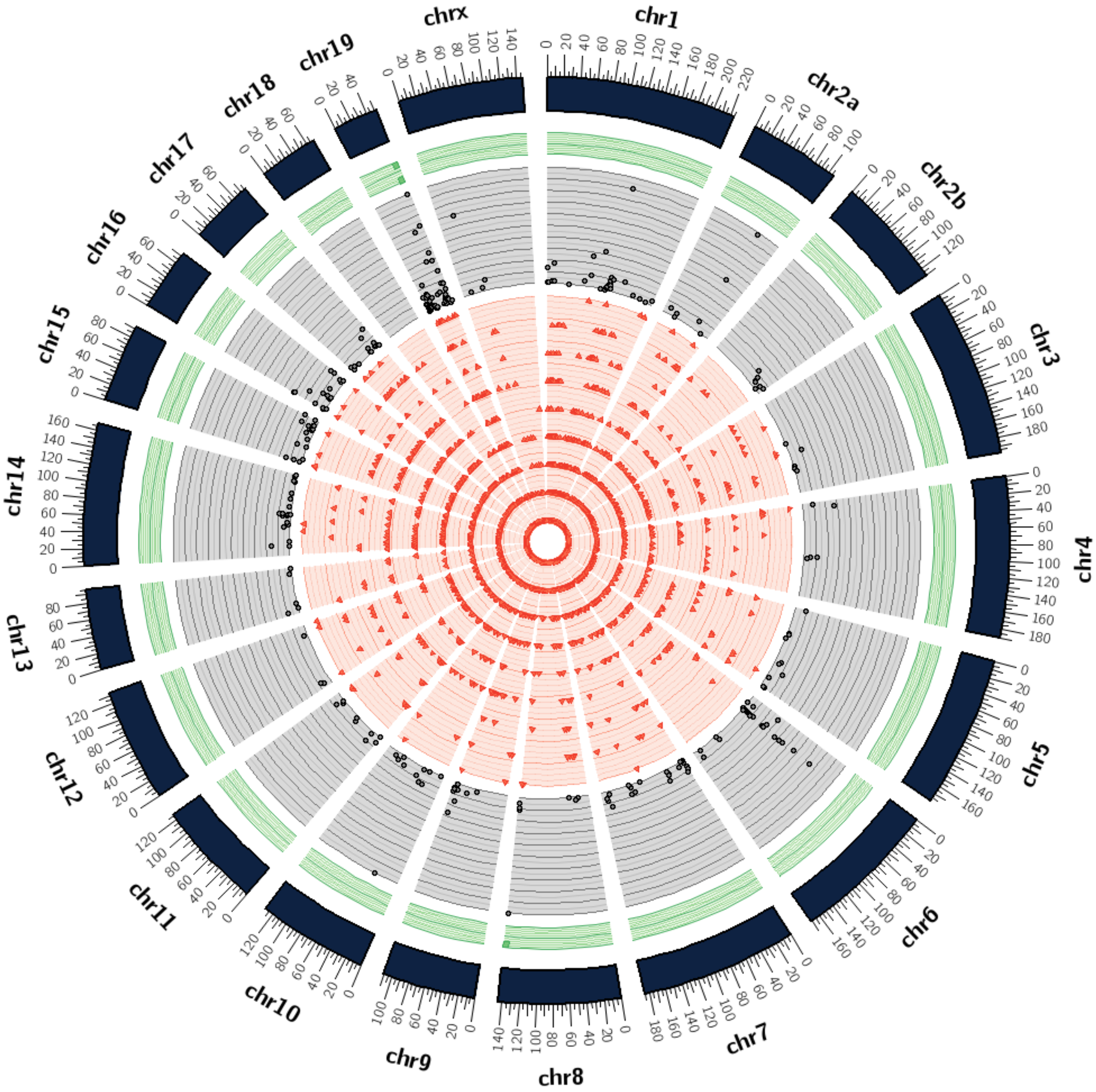


Figure 3.17: **Subject RFa14, peak parasitemia timepoint, E04**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 8 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

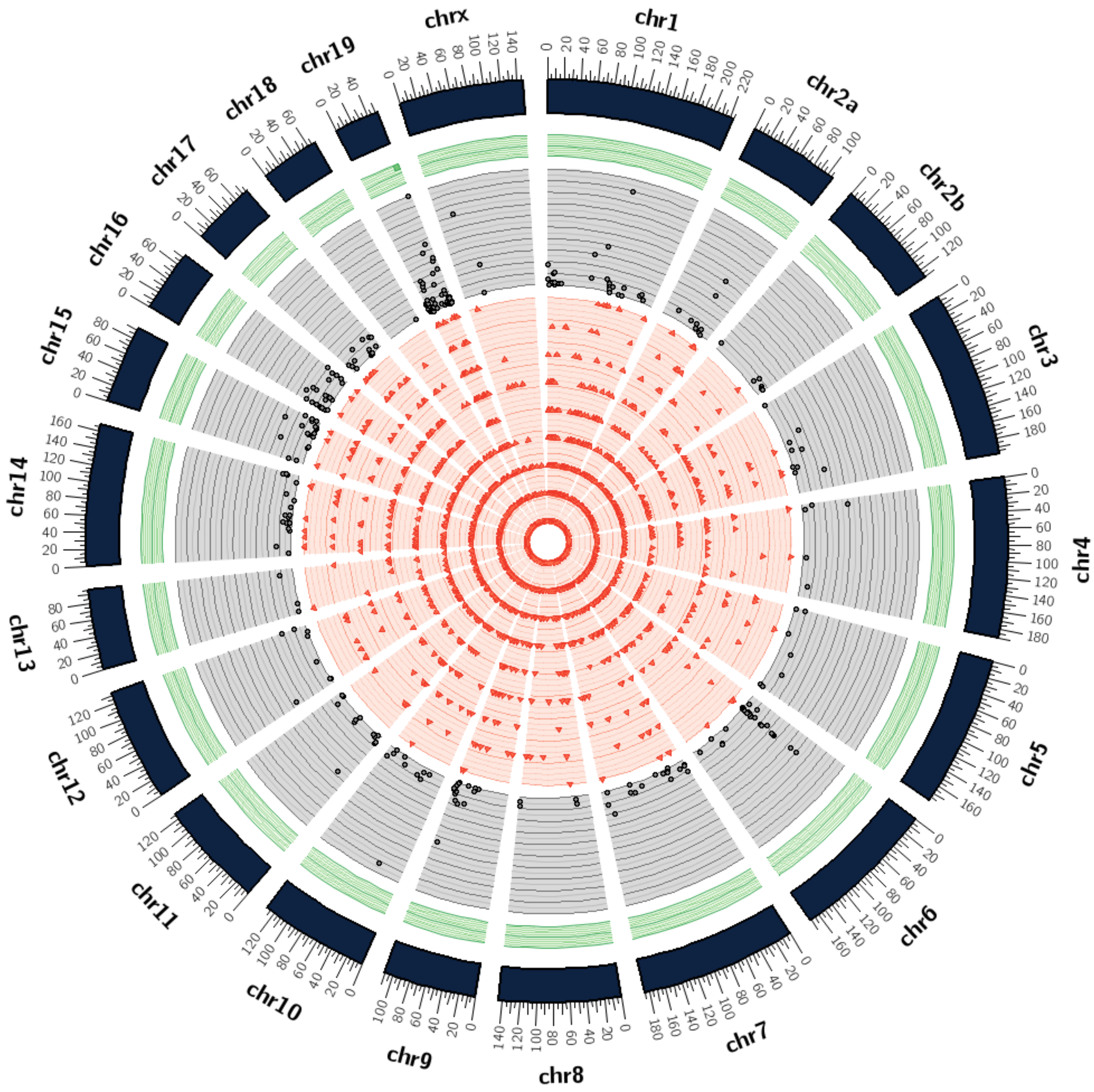


Figure 3.18: **Subject RFv13, peak parasitemia timepoint, E04**

This subject was severe. The subject displays particularly high SNV densities only at chromosome 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

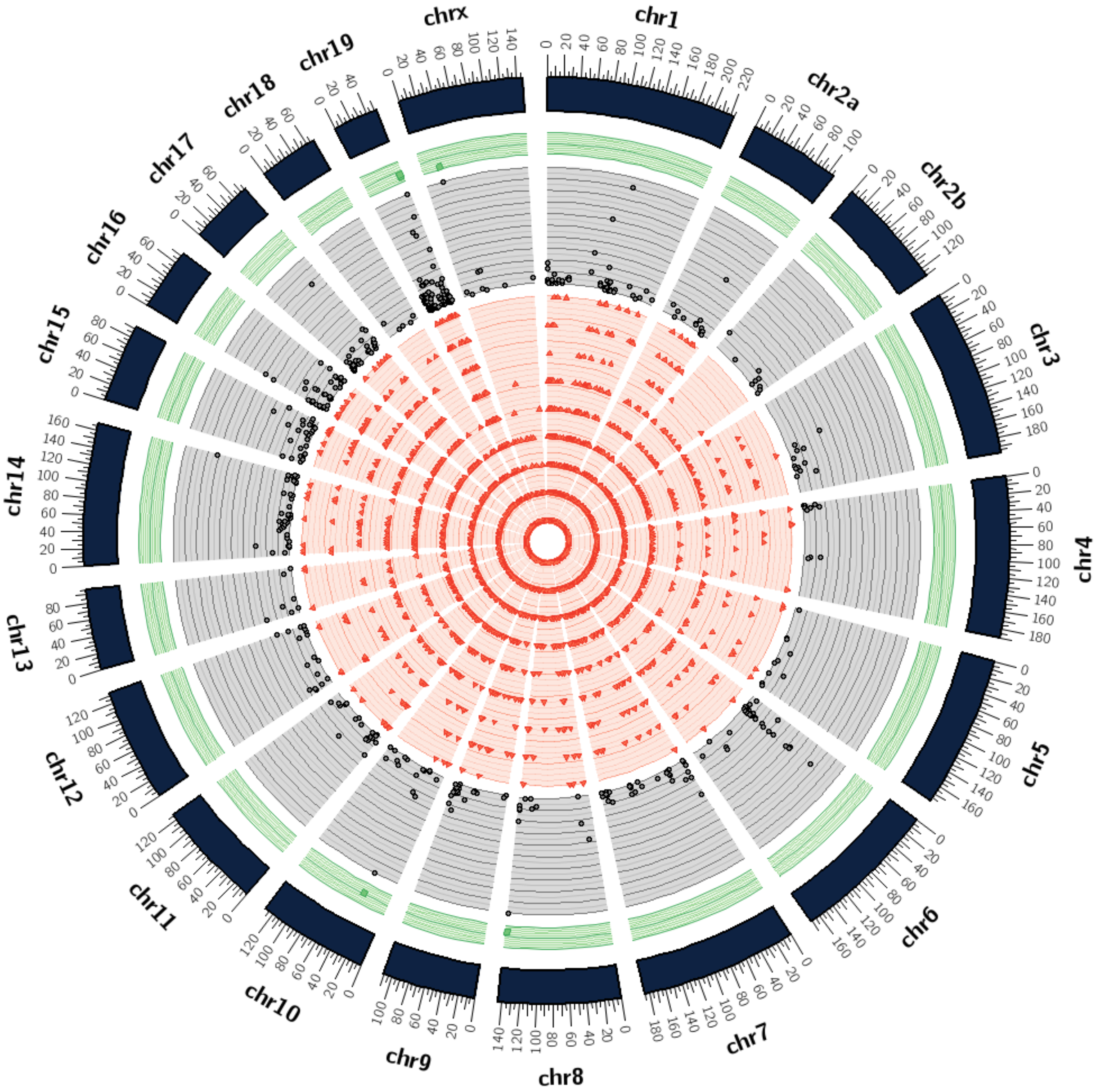


Figure 3.19: **Subject RIc14, peak parasitemia timepoint, E04**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 8, 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

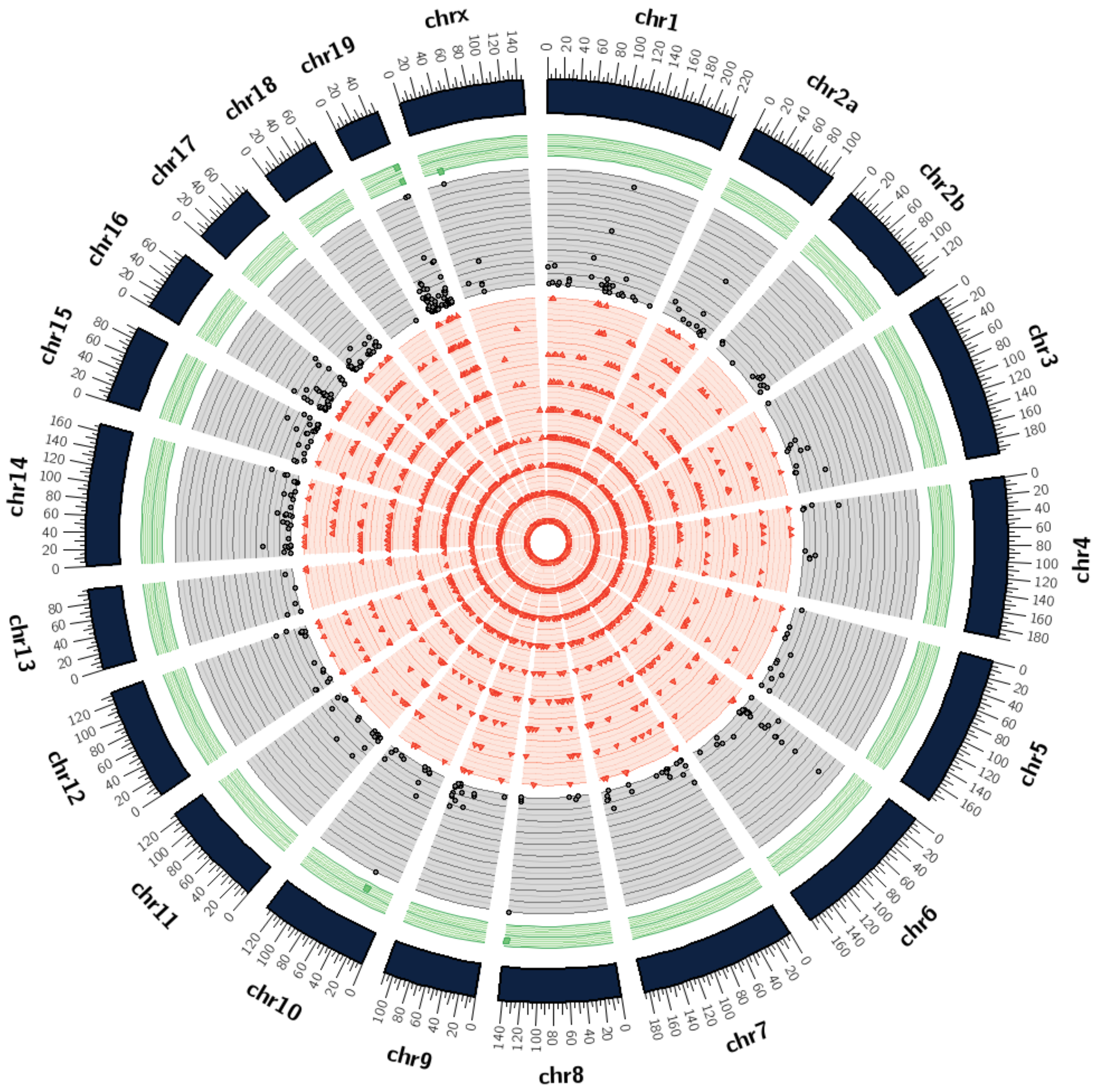


Figure 3.20: **Subject RMe14, peak parasitemia timepoint, E04**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 8, 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

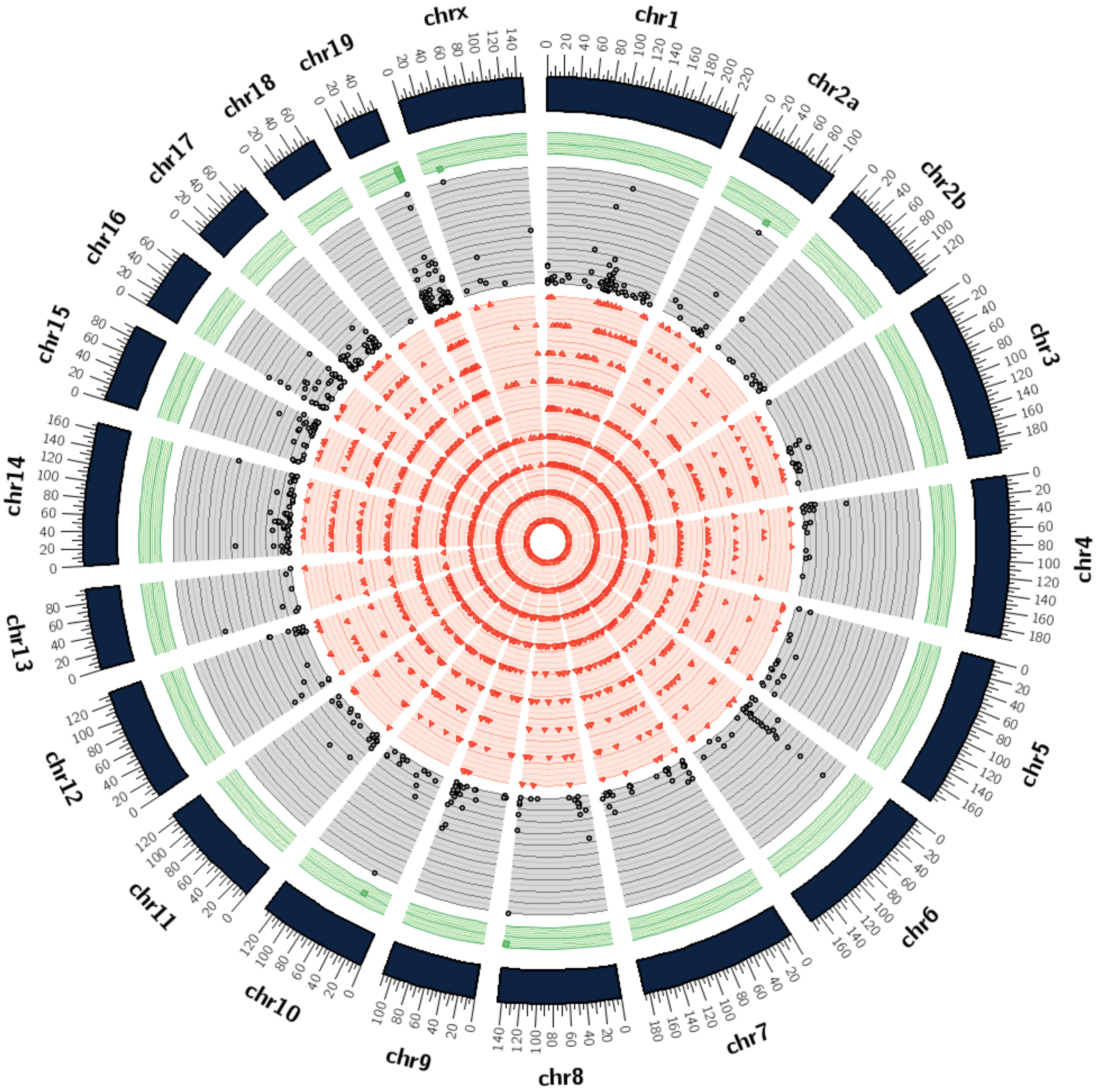


Figure 3.21: **Subject RSb14, peak parasitemia timepoint, E04**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 2a, 8, 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

### 3.3.3 Experiment 06

The following plots (Figures 3.22, 3.23, 3.24, 3.25, and 3.21) represent whole genome SNV densities per 100,000 base pairs in each of the five *M. mulatta* subjects infected with *P. knowlesi* malaria.

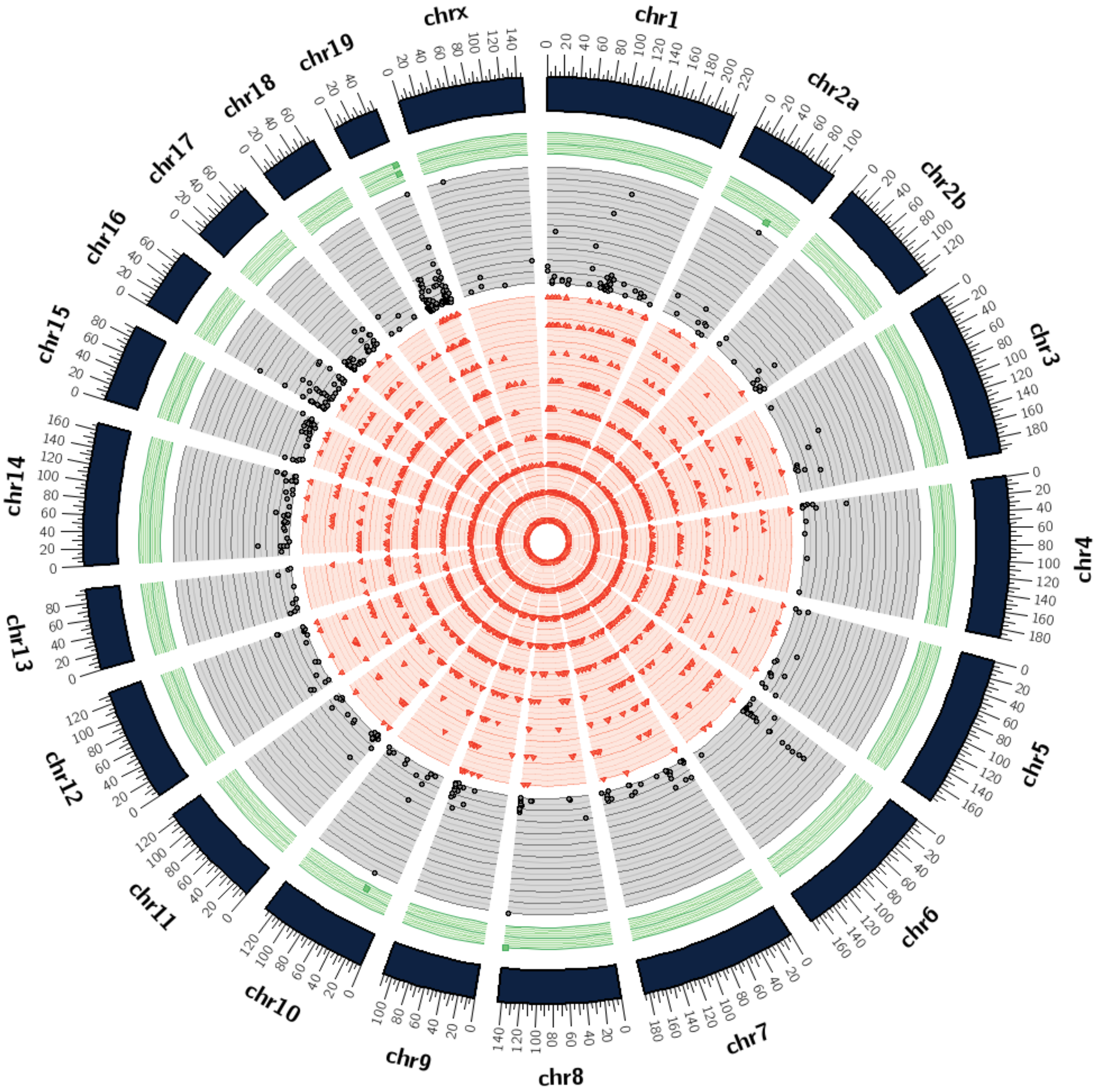


Figure 3.22: **Subject RCl15, peak parasitemia timepoint, E06**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 2a, 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

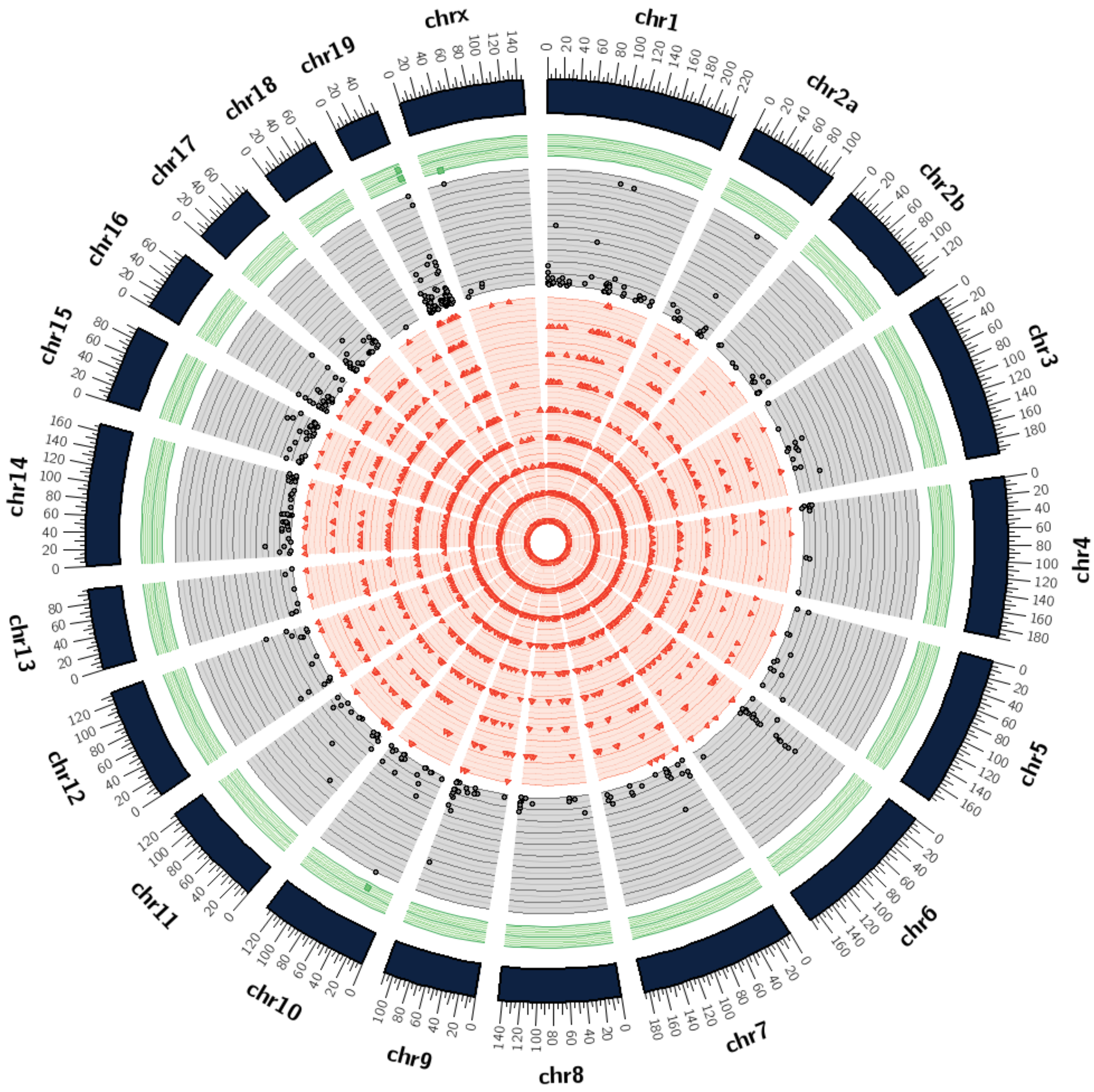


Figure 3.23: **Subject RIh16, peak parasitemia timepoint, E06**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

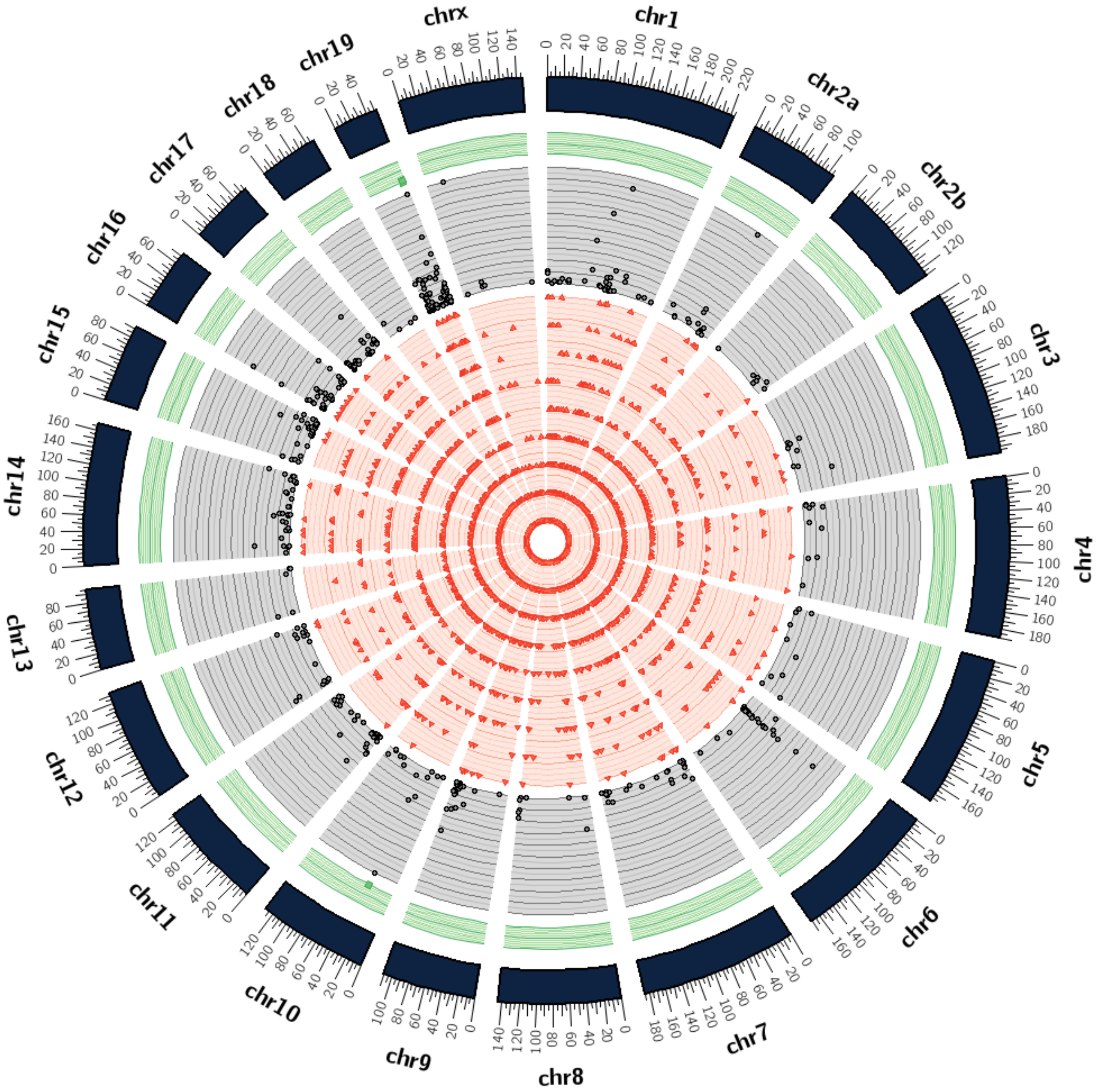


Figure 3.24: **Subject RTe16, peak parasitemia timepoint, E06**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 10 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

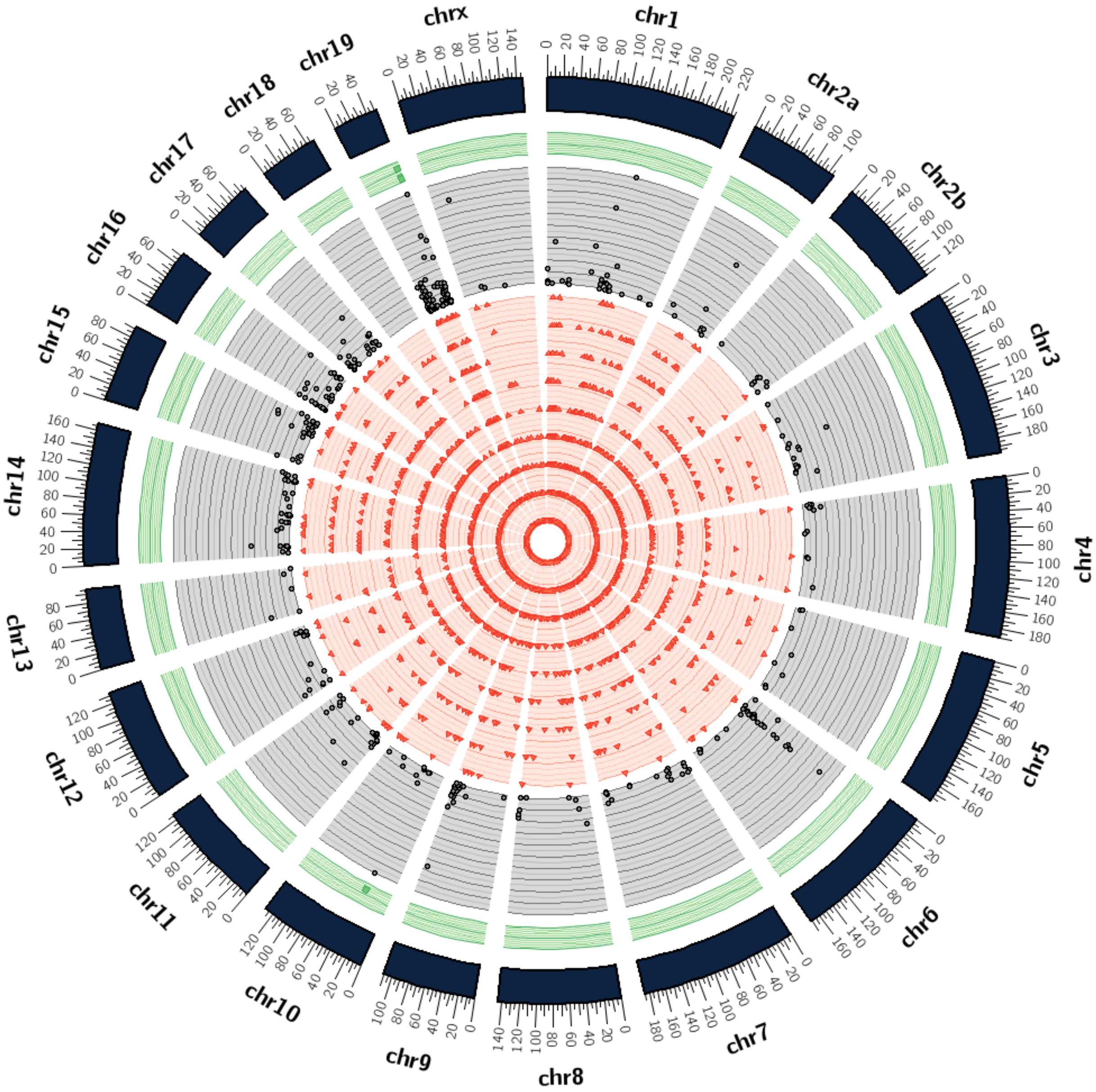


Figure 3.25: **Subject RUf16, peak parasitemia timepoint, E06**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 10 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

### 3.3.4 Experiment 23

The following plots (Figures 3.26, 3.27, 3.28, 3.29, 3.30, and 3.31) represent whole genome SNV densities per 100,000 base pairs in each of the six *M. mulatta* subjects infected with *P. cynomolgi* malaria.

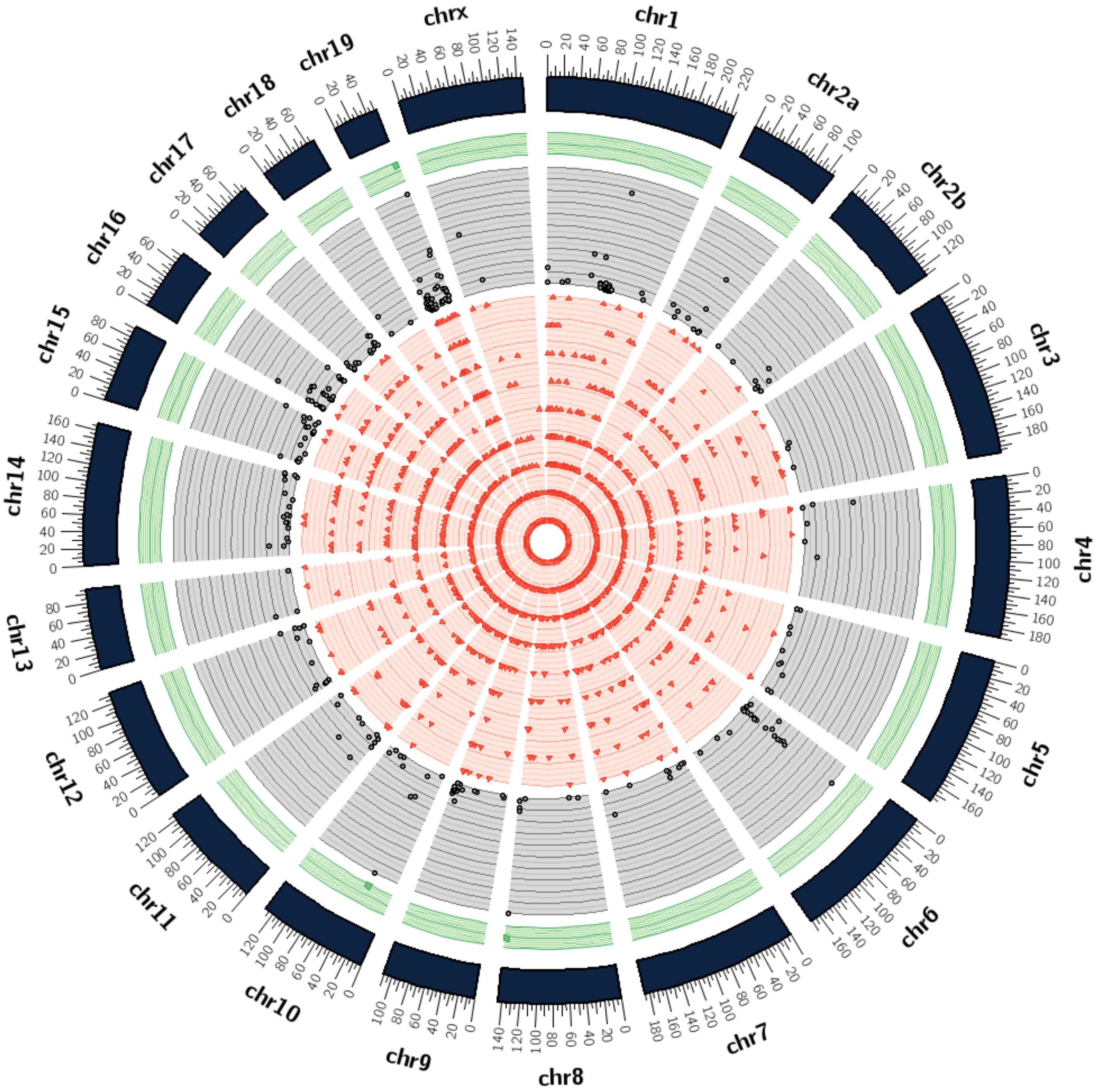


Figure 3.26: **Subject RAd14, peak parasitemia timepoint, E23**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

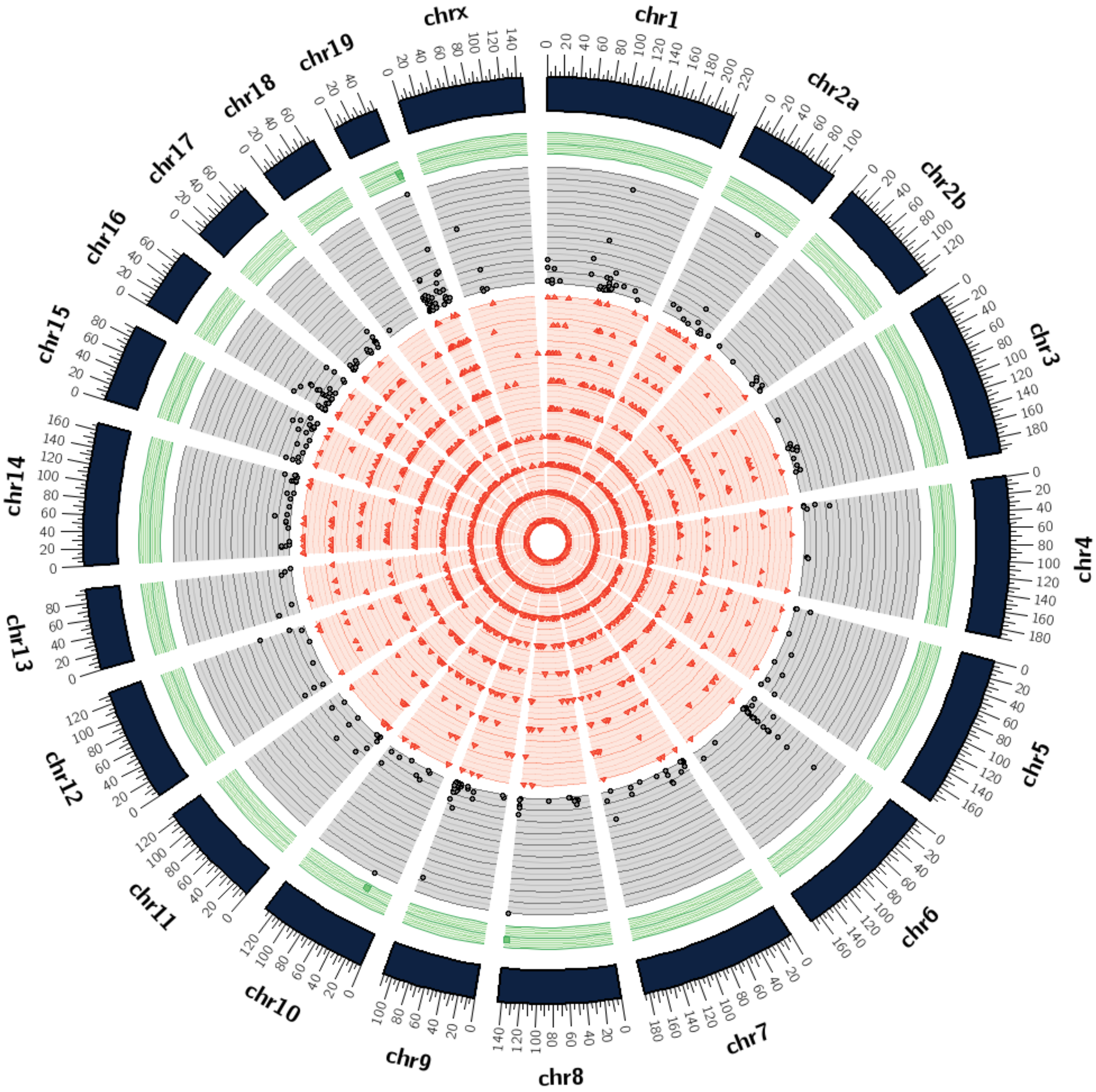


Figure 3.27: **Subject RBg14, peak parasitemia timepoint, E23**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

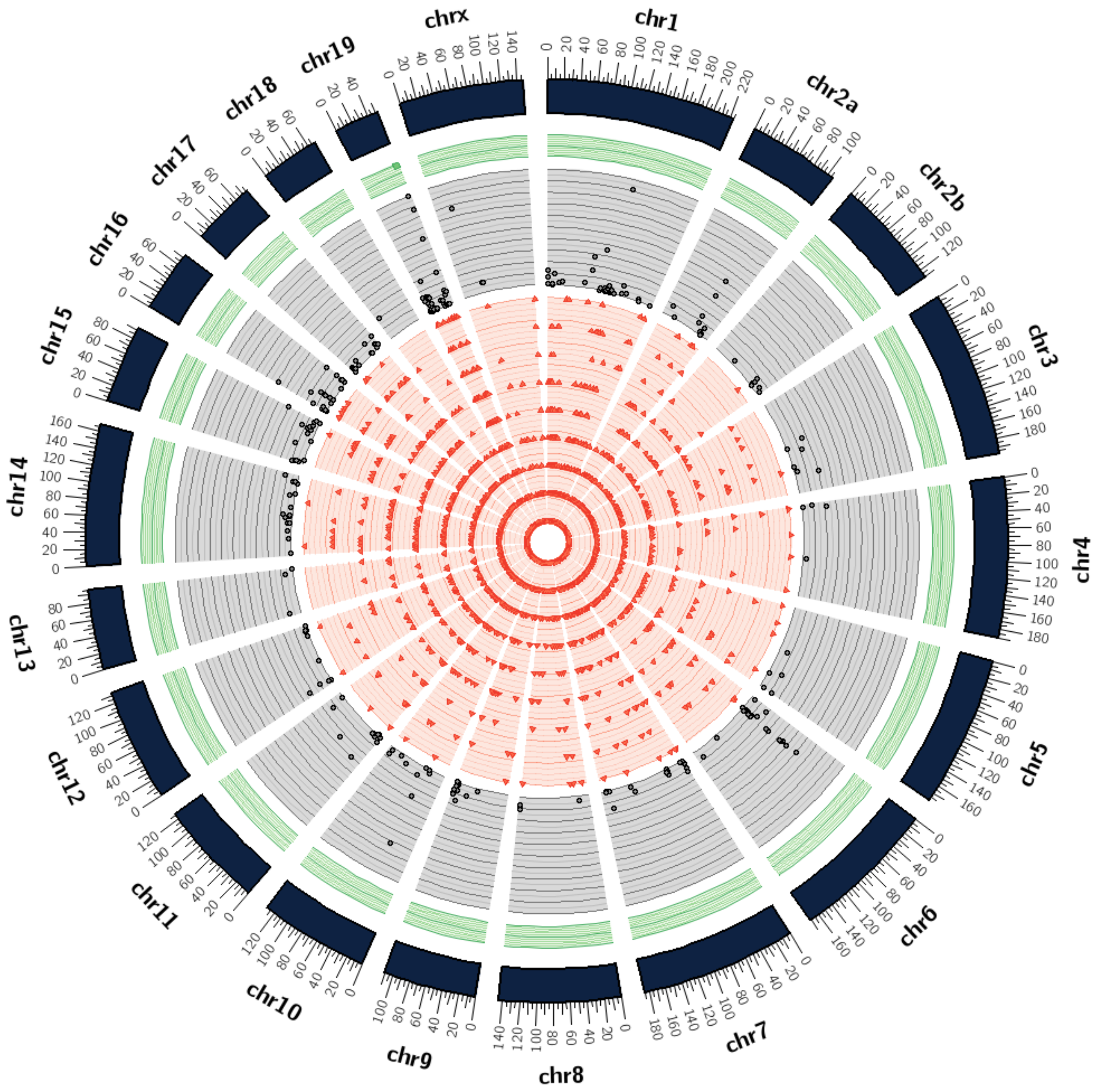


Figure 3.28: **Subject RIb13, peak parasitemia timepoint, E23**

This subject was mild. The subject displays particularly high SNV densities only at chromosome 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

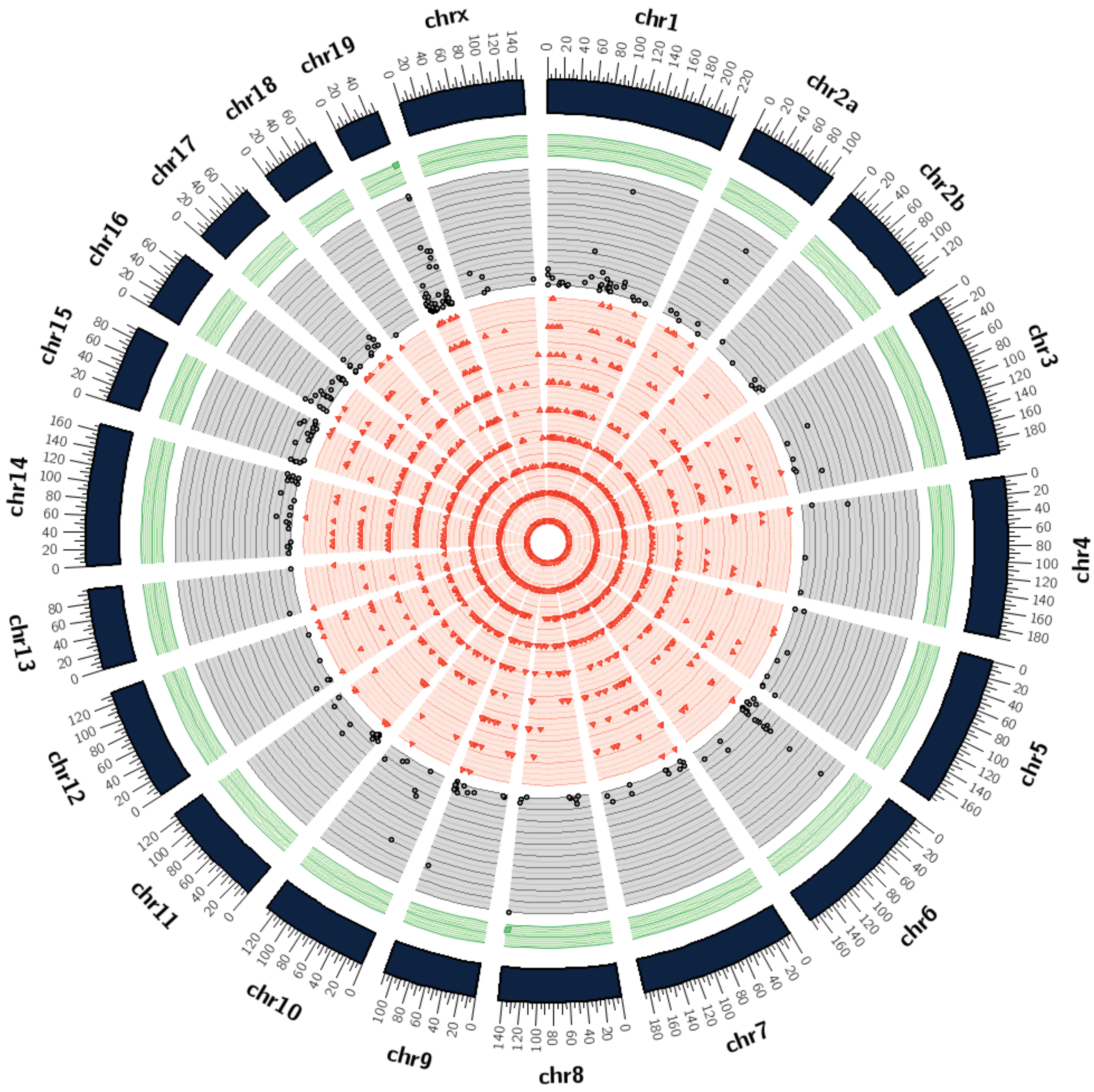


Figure 3.29: **Subject RJn13, peak parasitemia timepoint, E23**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 8 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

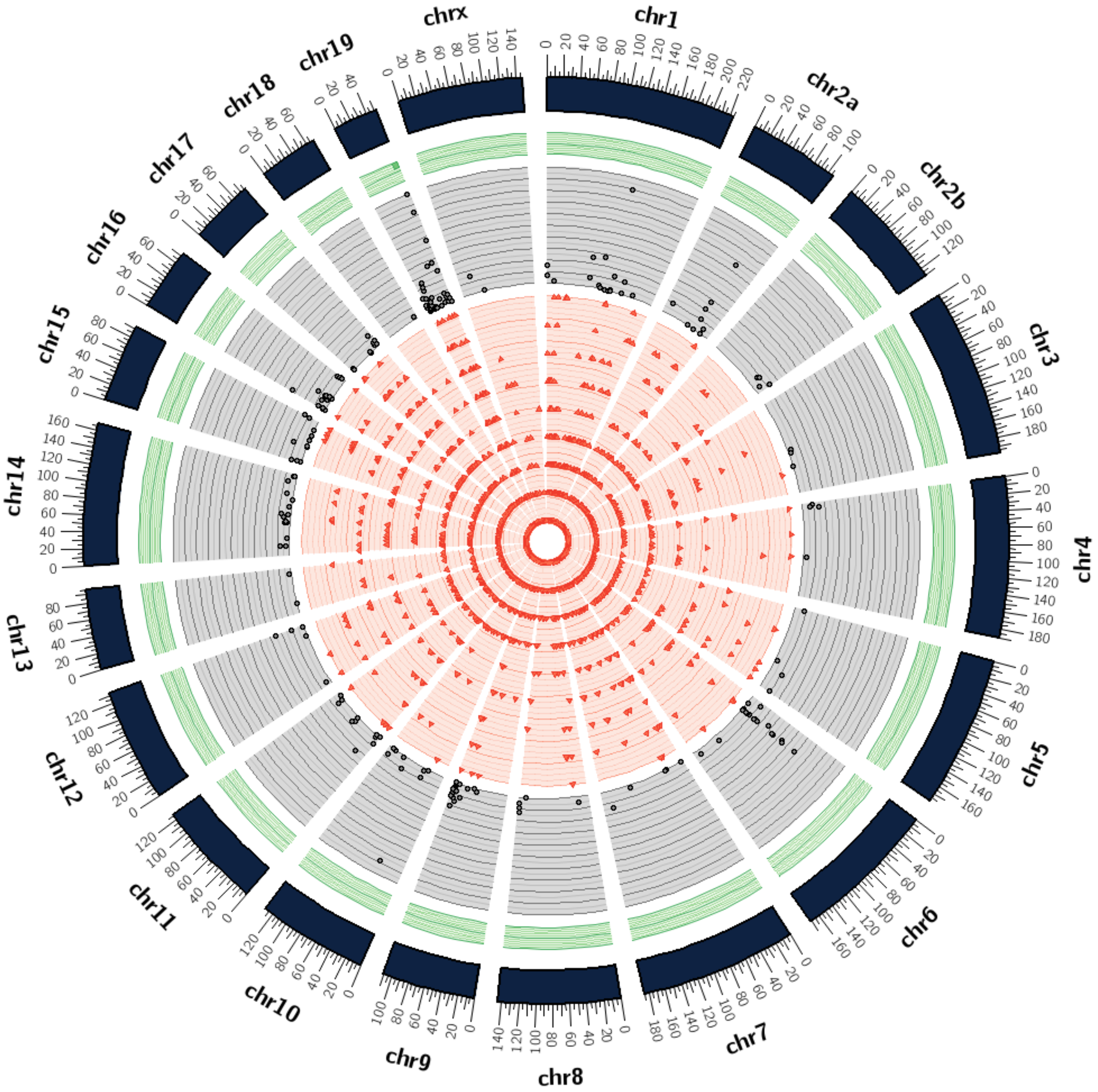


Figure 3.30: **Subject ROc14, peak parasitemia timepoint, E23**

This subject was mild. The subject displays particularly high SNV densities only at chromosome 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

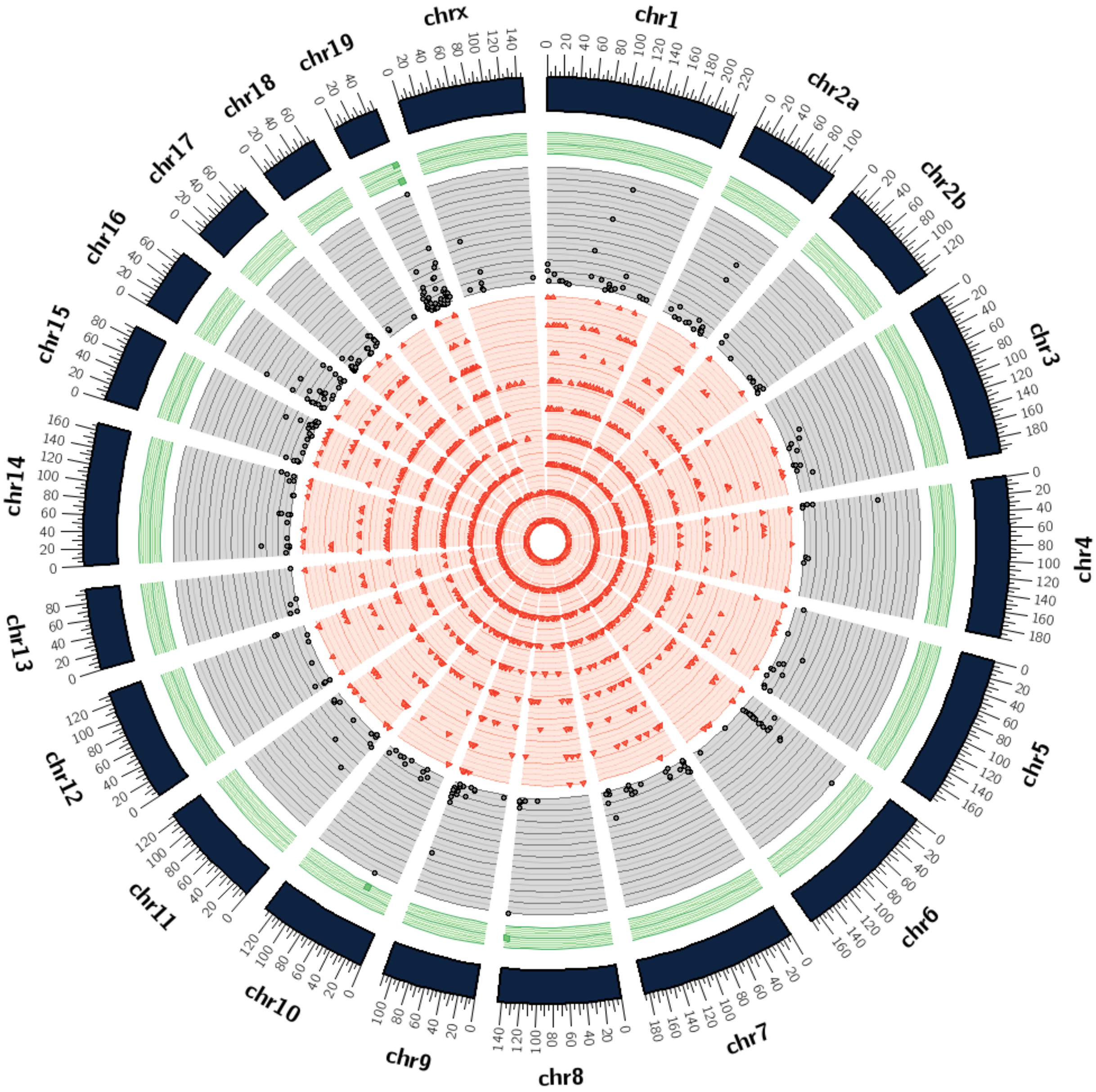


Figure 3.31: **Subject ROh14, peak parasitemia timepoint, E23**

This subject was mild. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

### 3.3.5 Experiment 24

The following plots (Figures 3.32, 3.33, 3.34, 3.35, and 3.36) represent whole genome SNV densities per 100,000 base pairs in each of the five *M. mulatta* subjects reinfected with *P. cynomolgi* malaria.

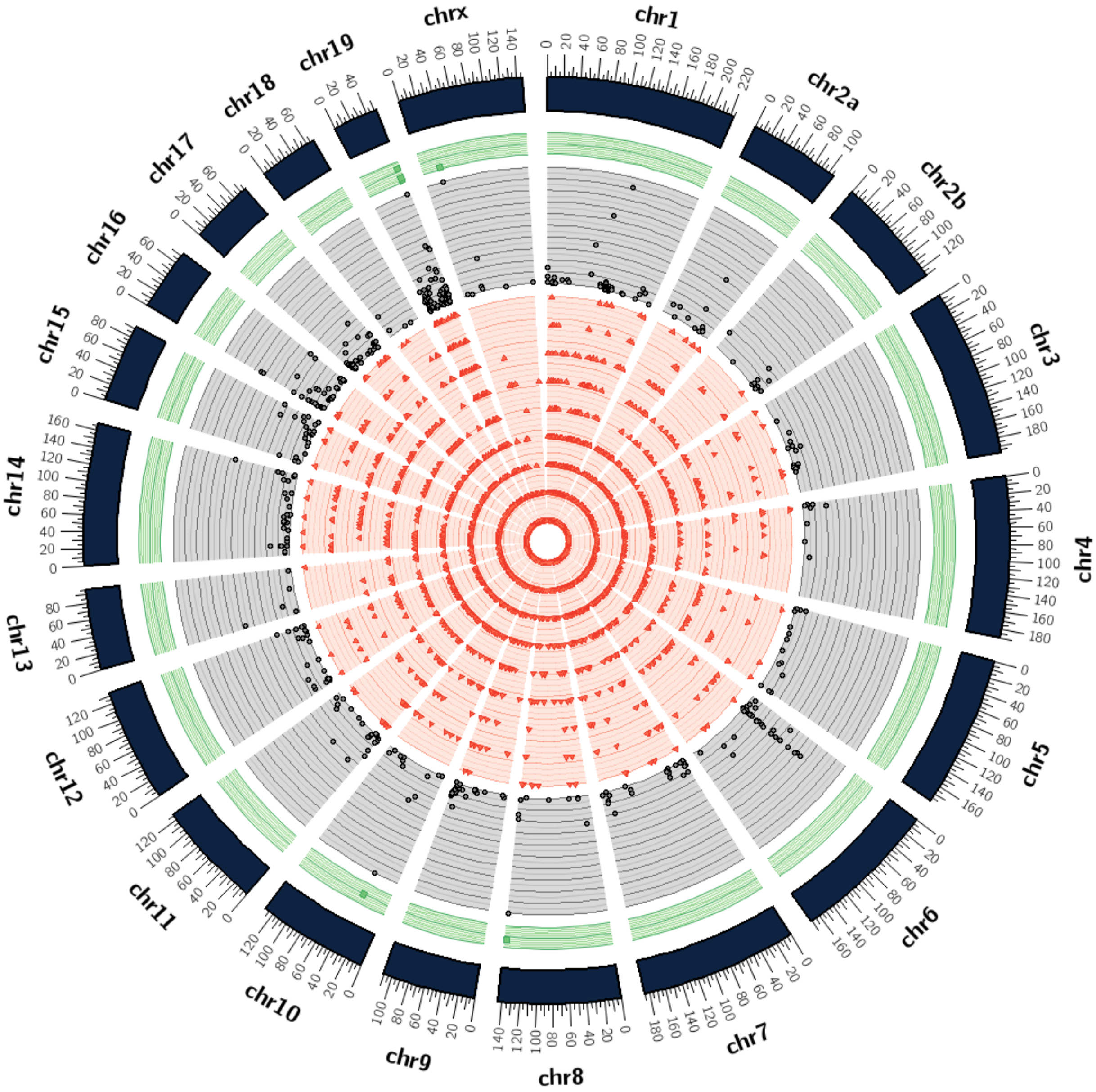


Figure 3.32: **Subject RAd14, peak parasitemia timepoint, E24**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

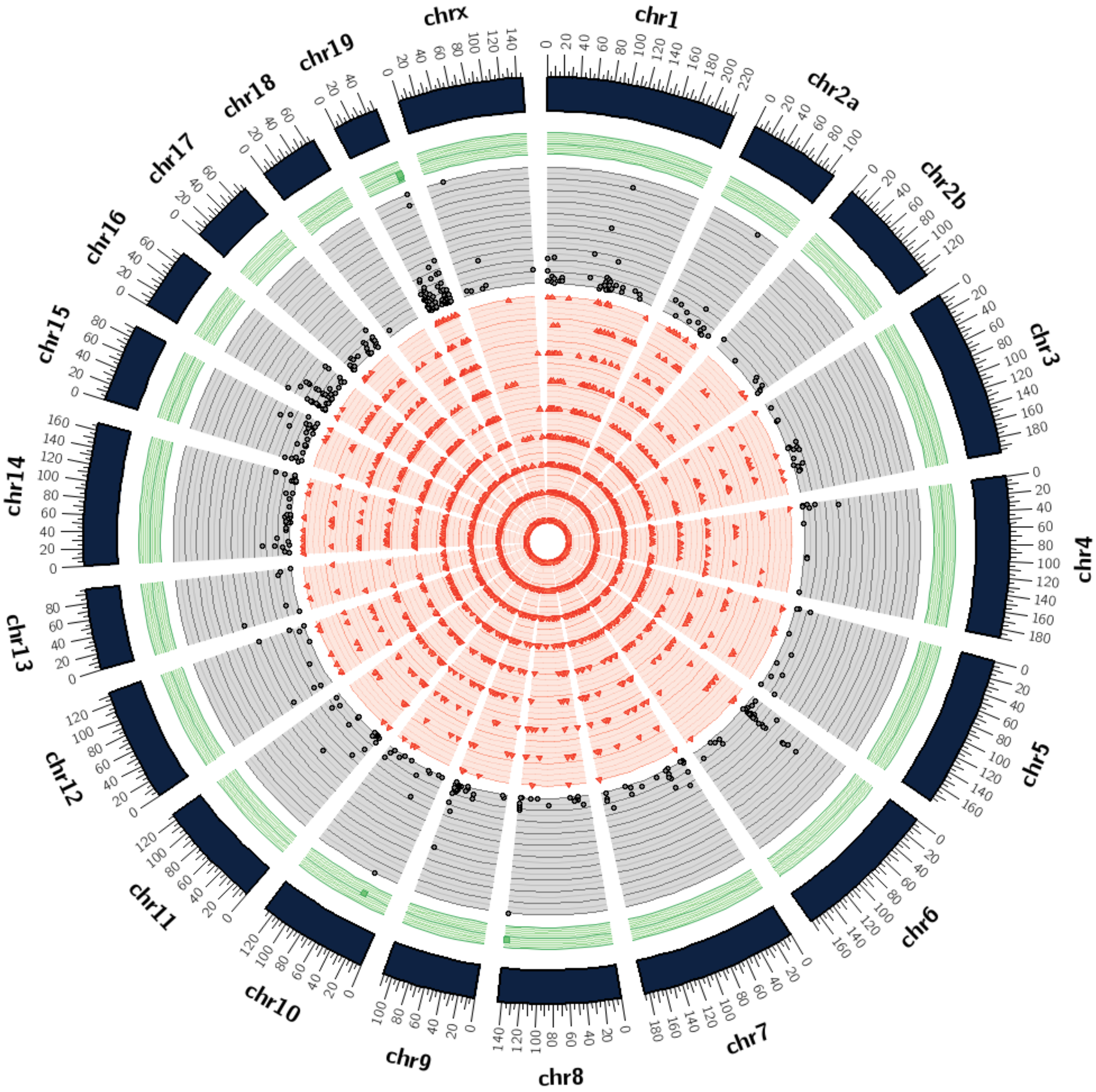


Figure 3.33: **Subject RBg14, peak parasitemia timepoint, E24**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

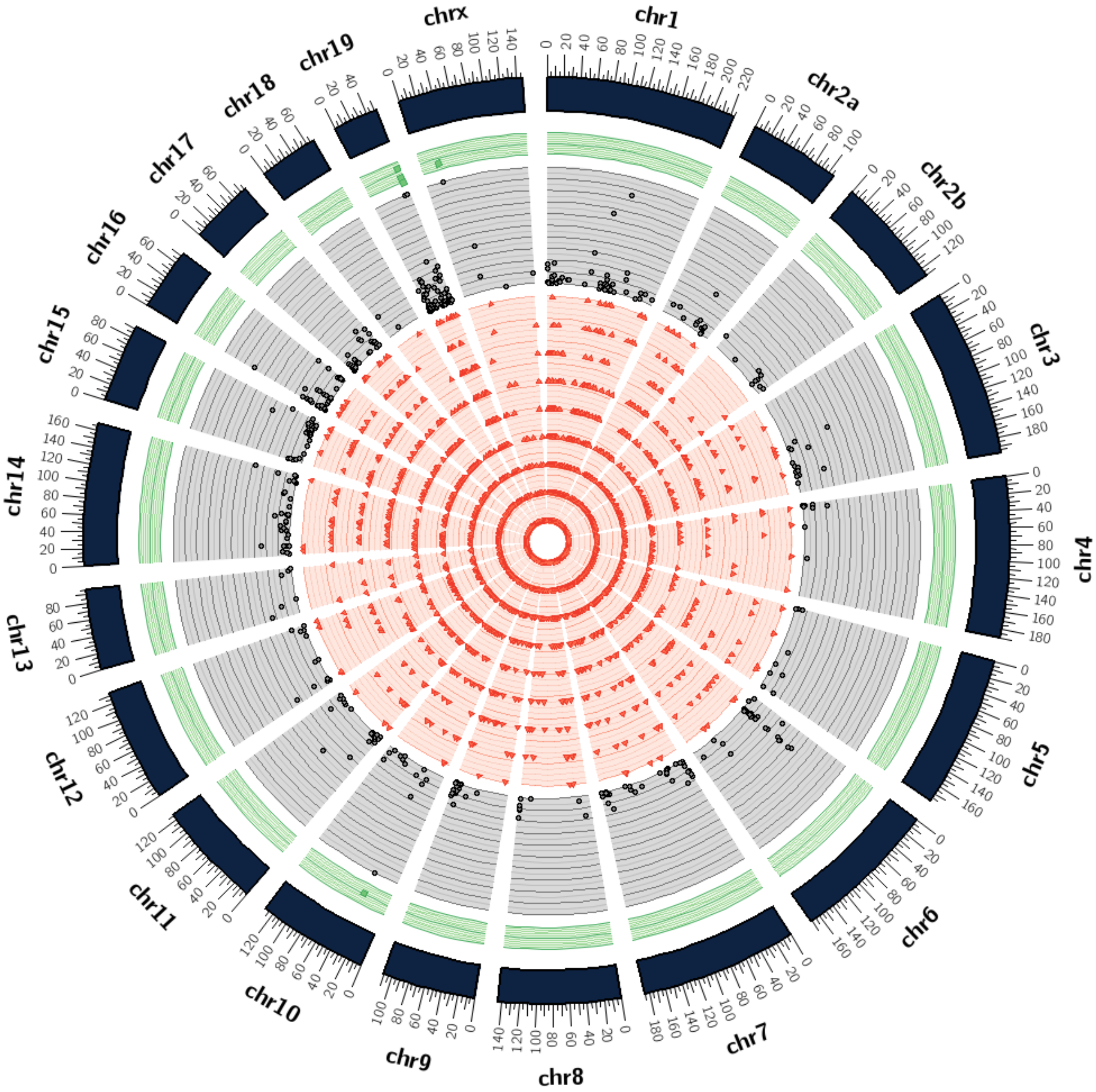


Figure 3.34: **Subject RIb13, peak parasitemia timepoint, E24**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

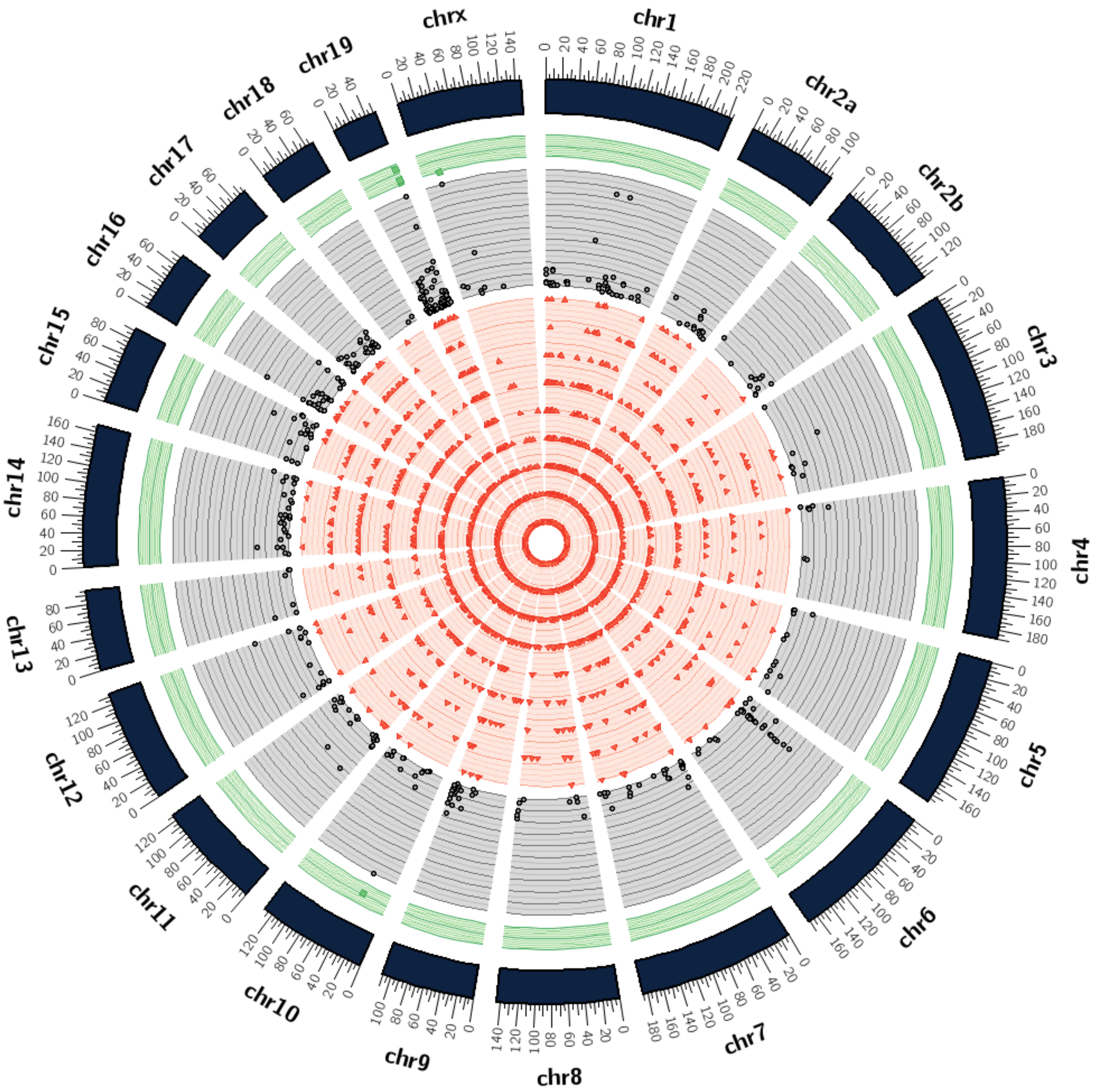


Figure 3.35: **Subject ROc14, peak parasitemia timepoint, E24**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

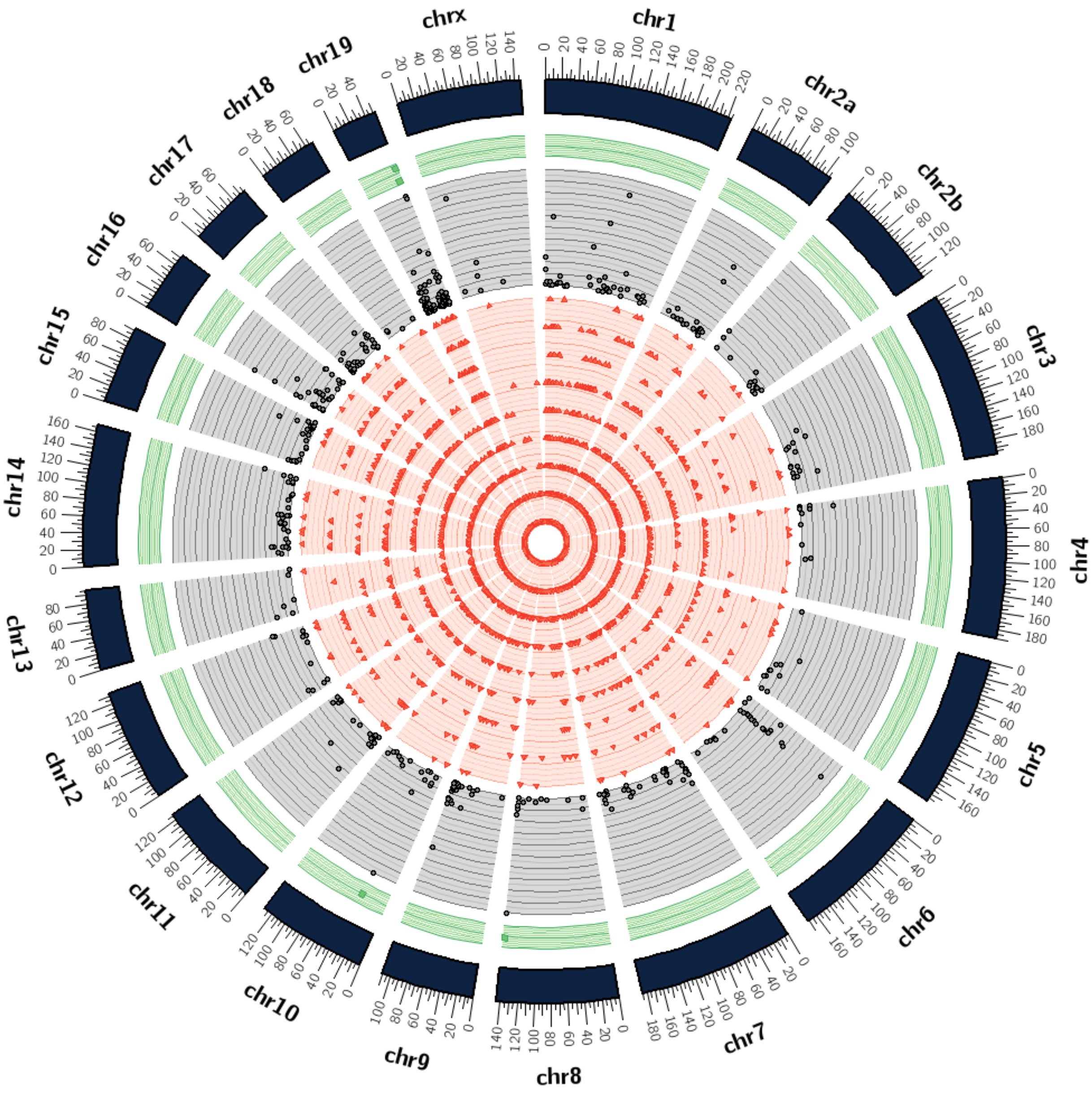


Figure 3.36: **Subject ROh14, peak parasitemia timepoint, E24**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

### 3.3.6 Experiment 25

The following plots (Figures 3.37, 3.38, 3.39, 3.40, and 3.41) represent whole genome SNV densities per 100,000 base pairs in each of the five *M. mulatta* subjects infected with a different strain of *P. cynomolgi* malaria.

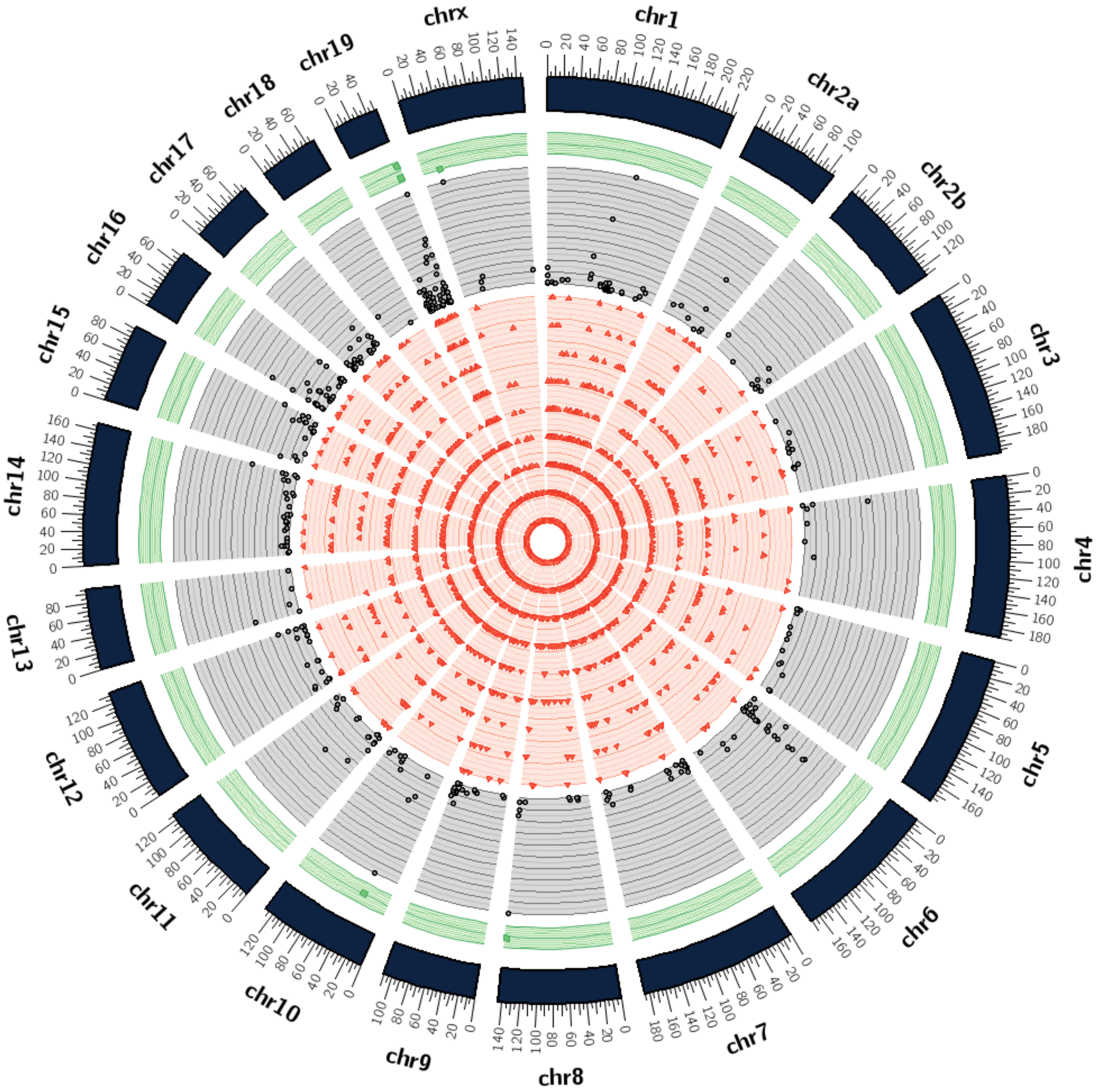


Figure 3.37: **Subject RAd14, peak parasitemia timepoint, E25**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 8, 10, 19, and X. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

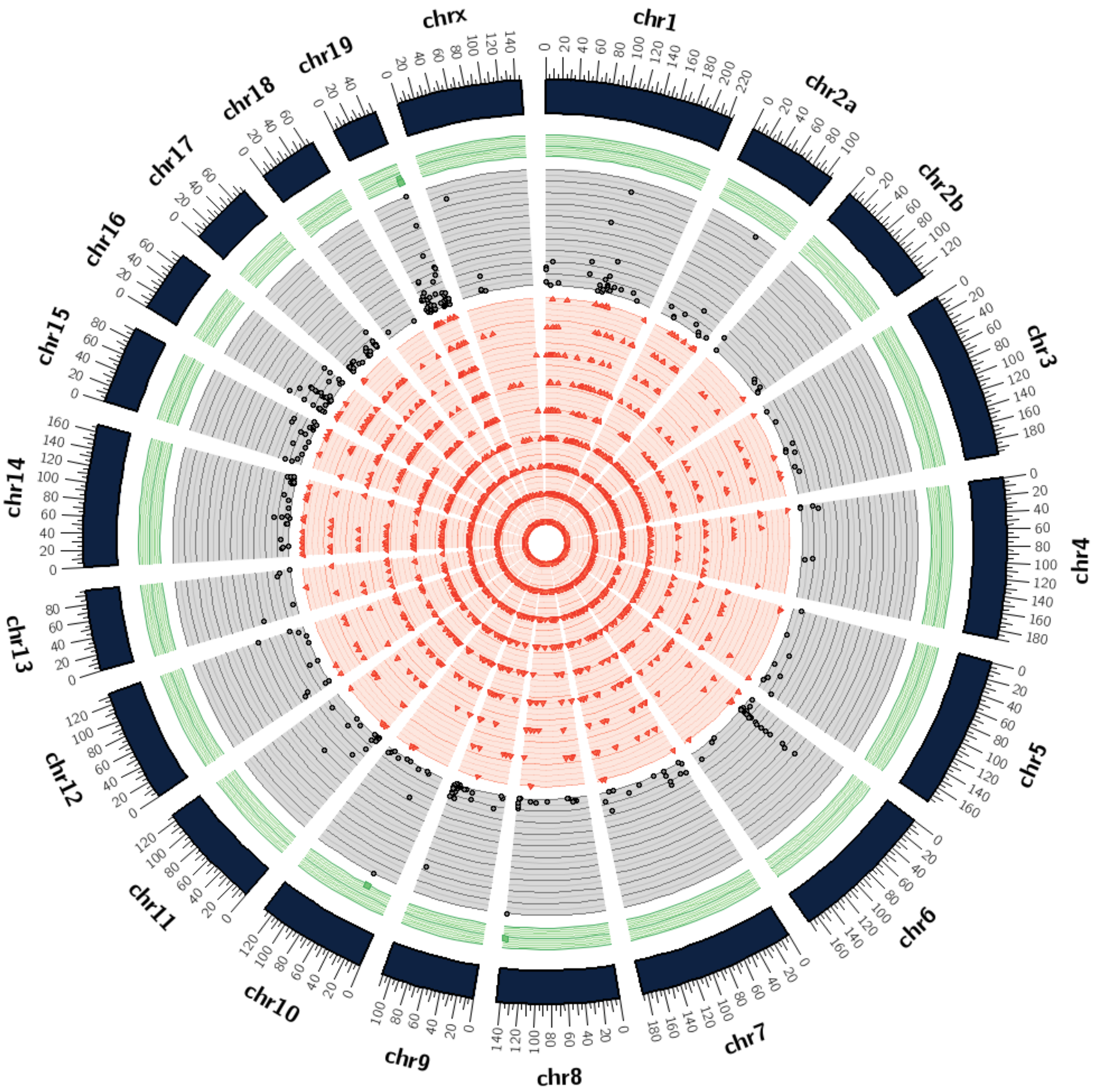


Figure 3.38: **Subject RBg14, peak parasitemia timepoint, E25**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

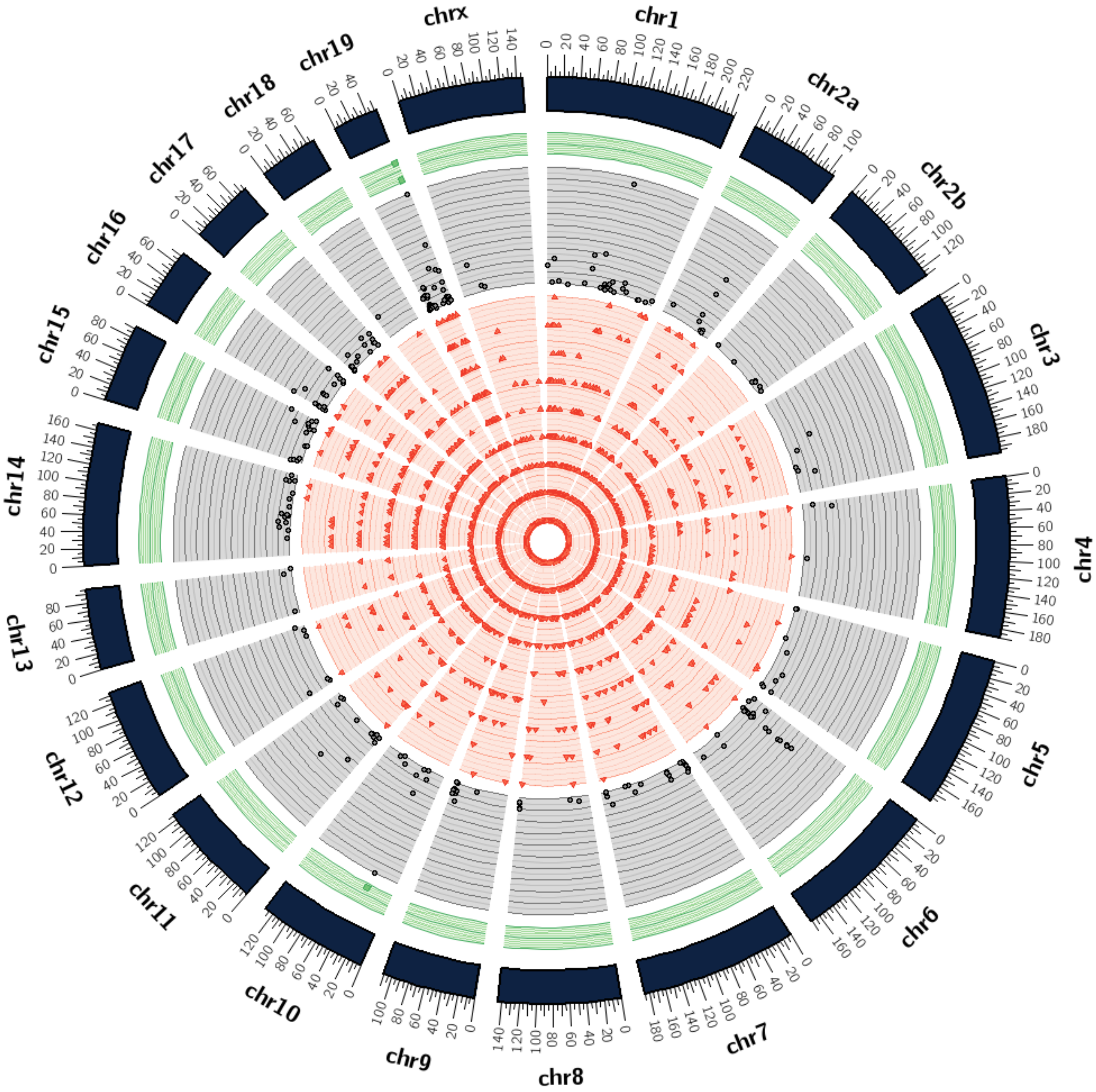


Figure 3.39: **Subject RIb13, peak parasitemia timepoint, E25**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 10 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

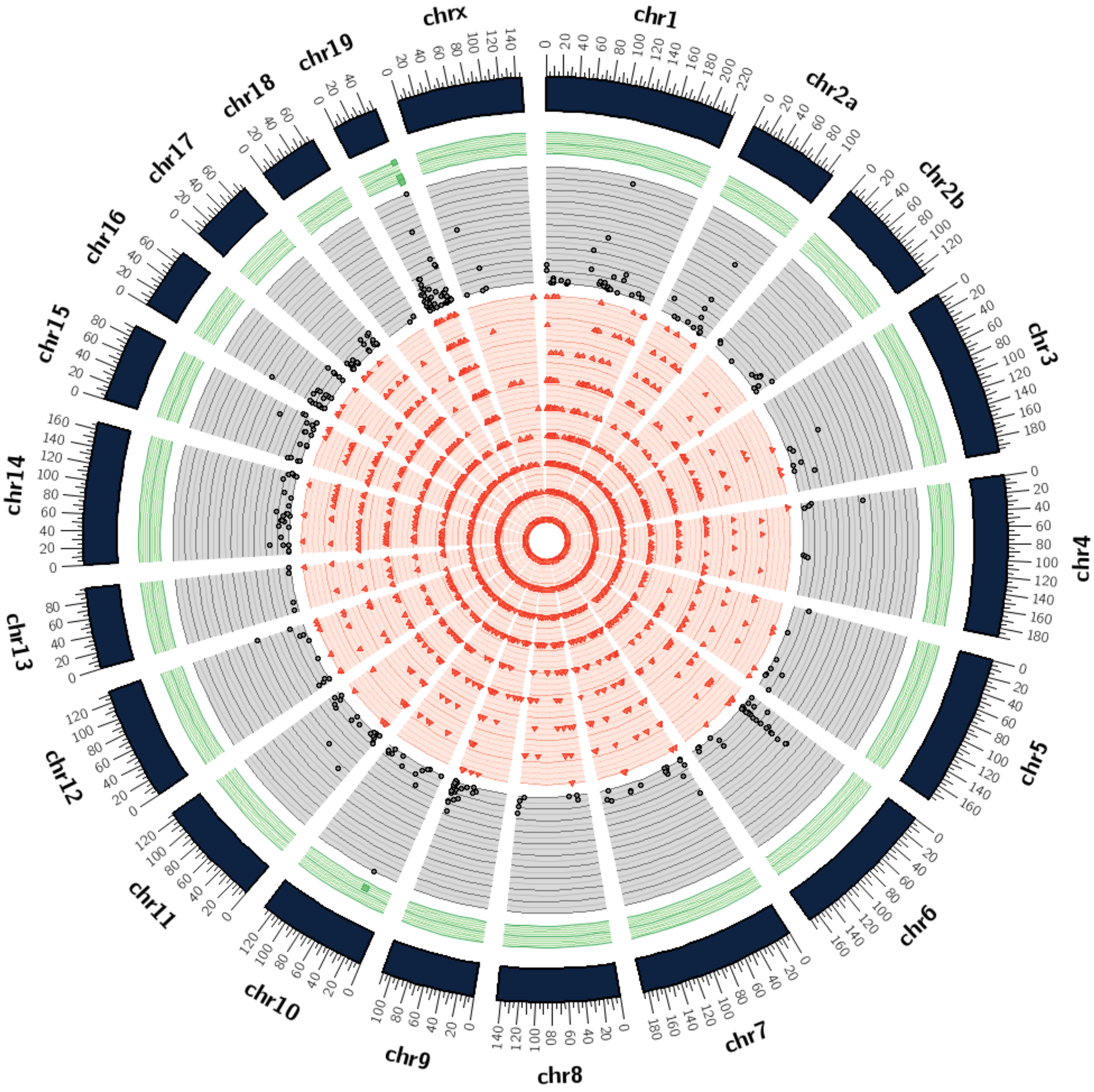


Figure 3.40: **Subject ROc14, peak parasitemia timepoint, E25**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 10 and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

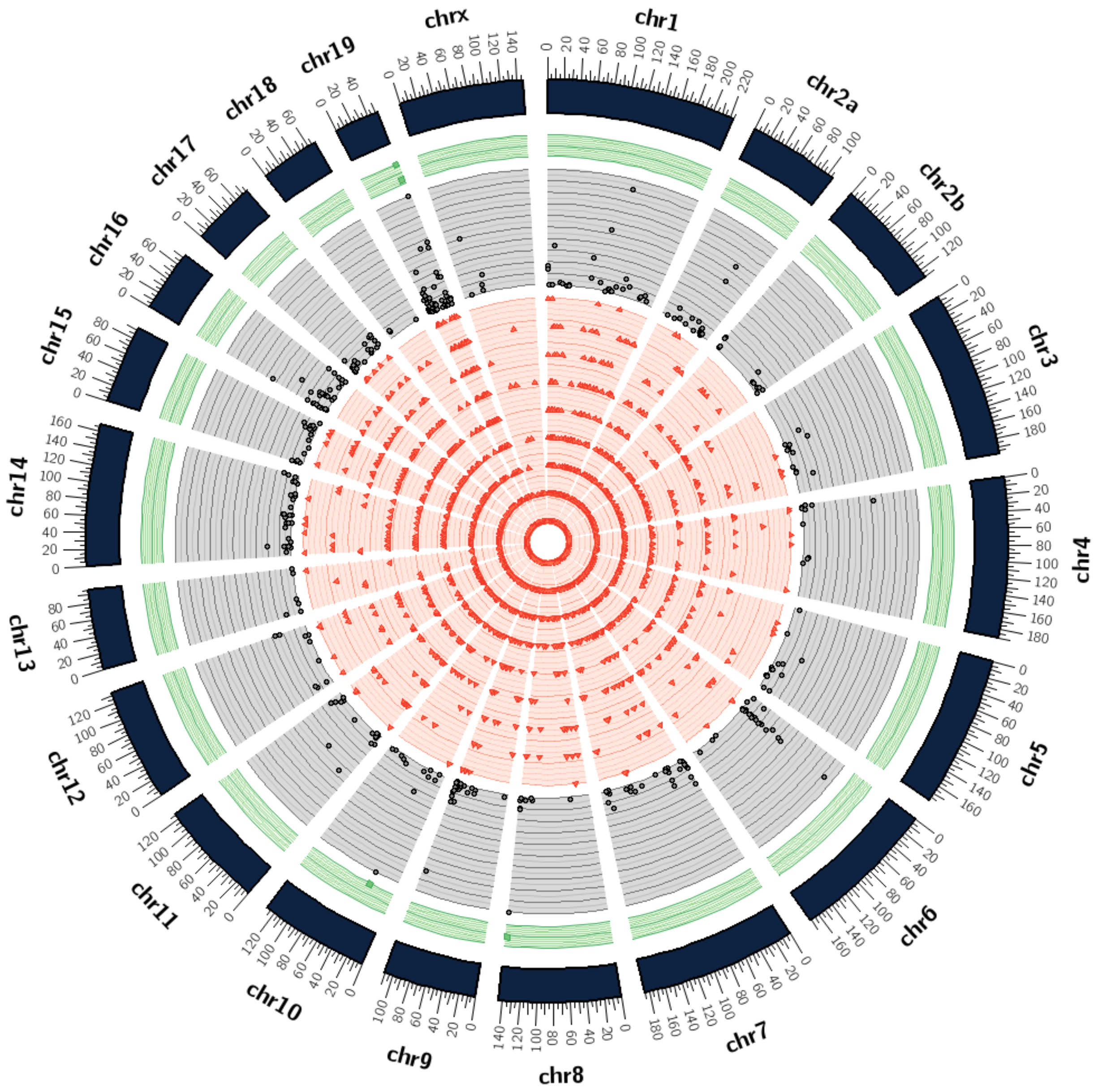


Figure 3.41: **Subject ROh14, peak parasitemia timepoint, E25**

Data from this experiment was not used in the severity analysis. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

### 3.3.7 Experiment 30

The following plots (Figures 3.42 and 3.43) represent whole genome SNV densities per 100,000 base pairs in both of the *M. mulatta* subjects infected with *P. knowlesi* malaria.

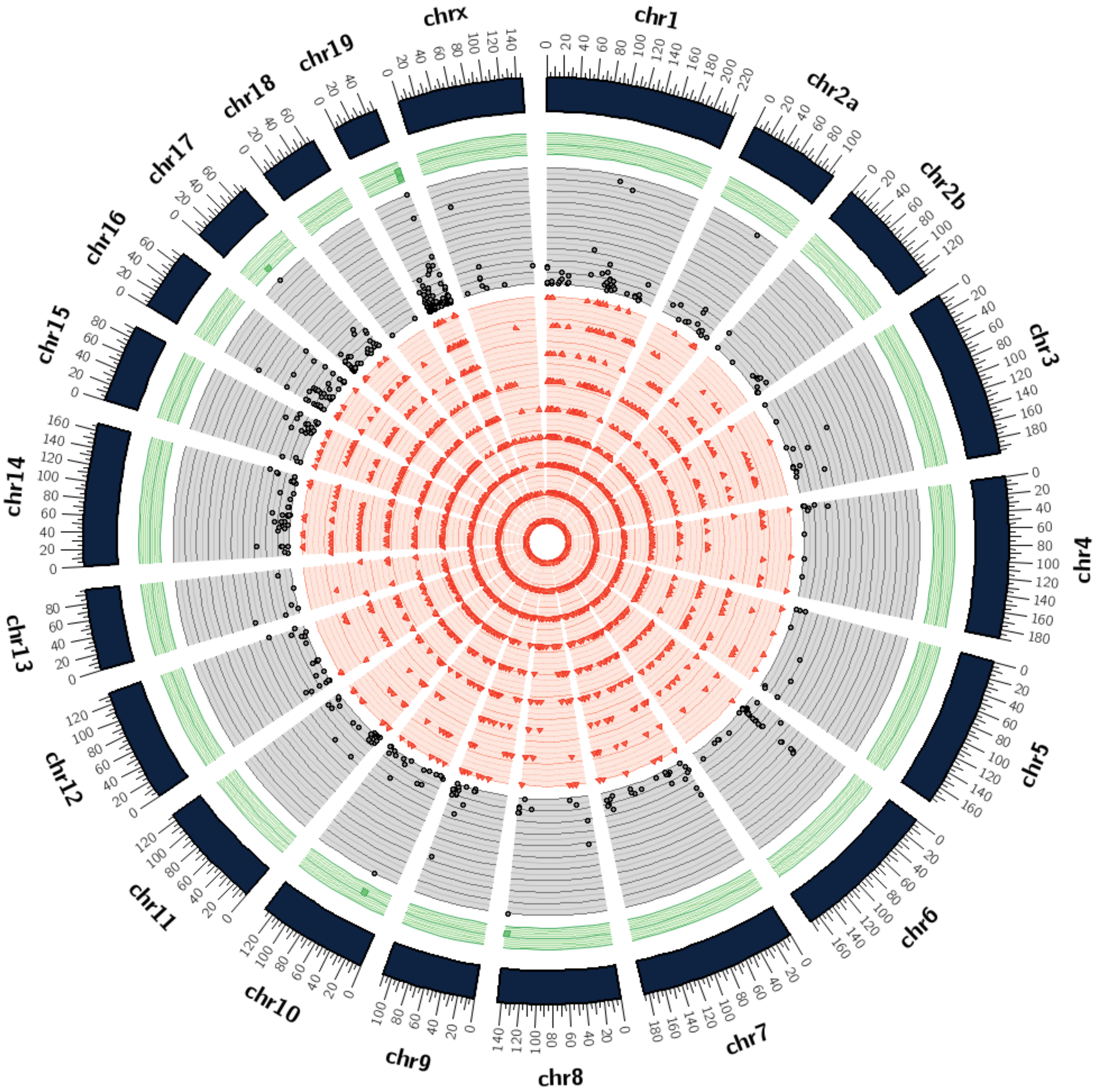


Figure 3.42: **Subject REd16, peak parasitemia timepoint, E30**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 8, 10, 17, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

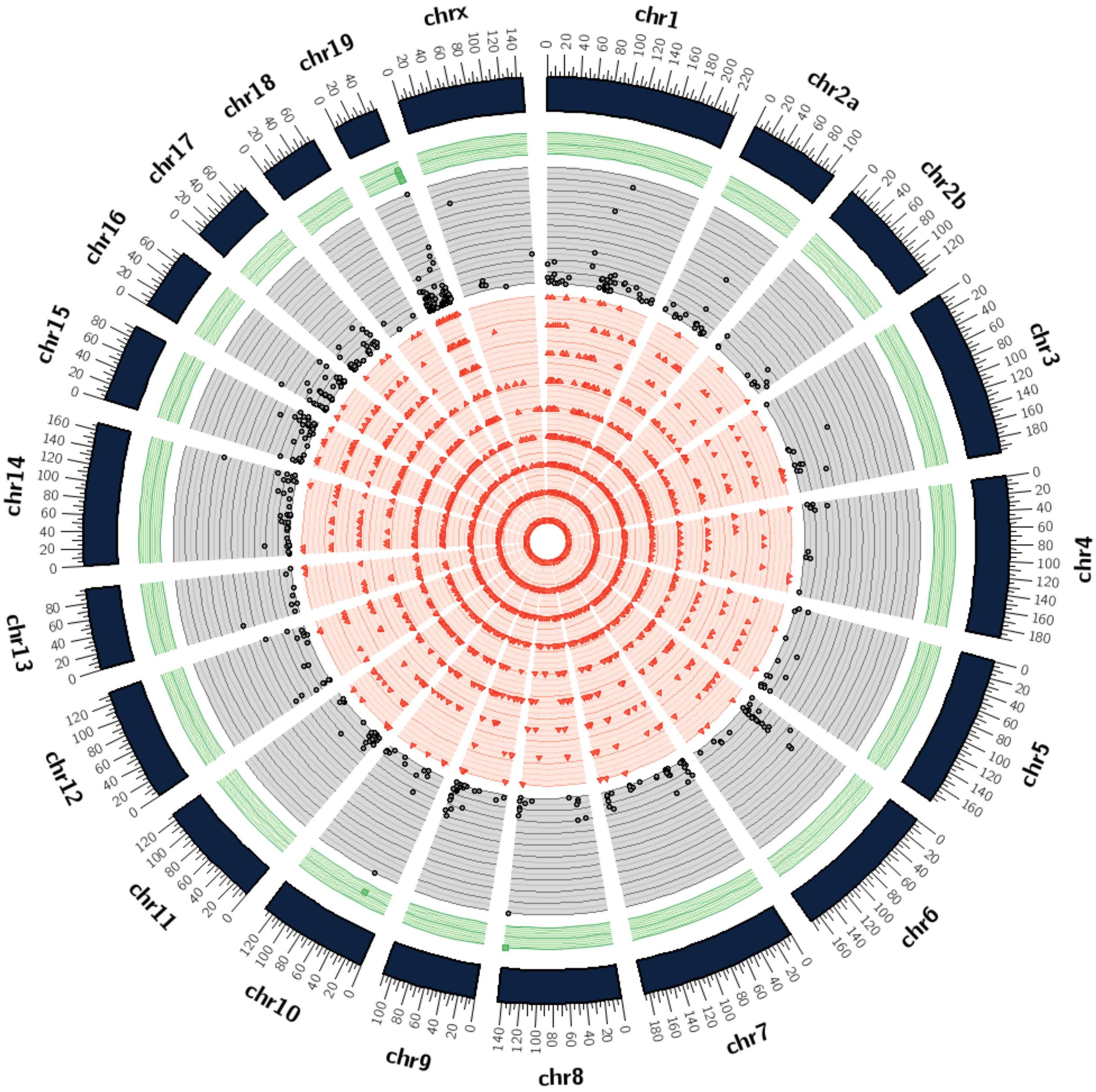


Figure 3.43: **Subject RKy15, peak parasitemia timepoint, E30**

This subject was severe. The subject displays particularly high SNV densities at chromosomes 8, 10, and 19. The red innermost ring represents densities between 0.00001 and 0.0009, the middle gray ring shows densities between 0.001 and 0.01, and the green outermost ring displays densities of 0.011 and higher.

# Chapter 4

## Discussion and Implications

### 4.1 Caveats

No system is perfect, a rule from which this project is not exempt. Perhaps the most apparent caveat is the limited sample size as compared with the statistical power needed to make claims of significance. As it currently exists, SHIELD cannot identify statistically significant SNVs or SNPs. The program is specifically designed around exploratory analysis for studies with small sample sizes. Some amount of further analysis is required of the user, though the output is designed to be as versatile and malleable as possible to accommodate many possible routes of examination.

Another caveat is the reference genome that was used in this analysis. As mentioned before, this project began with RNA-Seq reads that had already been mapped back to a reference genome. In order to minimize unnecessary variables, all analyses were done against the same reference. This reference was sequenced internally, and is not yet publicly available. Furthermore, this genome is different enough from the current NCBI *M. mulatta* reference genome to make coordinate-based comparisons with databases impossible (i.e., there is currently no way to compare the SNVs with SNP databases to see if any of them have already been reported). However, this is an issue isolated to this analysis and unrelated to SHIELD's performance.

### 4.2 Common SNVs

The 1,657 genes for which every monkey in all seven experiments (E03, E04, E06, E23, E24, E25, and E30) commonly reported at least one SNV were identified. The density distribution of SNVs from within those genes was plotted and revealed chromosomes 1, 2a, 6, 10, 19, and X as areas with particularly high densities (Figure 3.7). Out of that list, 28 genes of interest were found to also be significantly up-regulated during the peak parasitemia time point. While many of those 28 genes are in highly dense SNV regions, several are not (Table 3.3). For example, the genes BPI and PARP14 are located on chromosomes 15 and 3, respectively, both of which appear to have particularly low SNV densities by comparison. This could support the idea that such genes and the reported SNVs within them are worth closer examination.

These genes (as well as the genes determined to be significantly upregulated) were also identified via GSEA to be significantly involved with the immune system (Tables 3.1 and 3.2). However, it is known that genes involved in immunity, particularly innate immunity, are under higher selective pressure than other genomic regions, which would in turn result in higher shared SNV densities [34, 35]. Therefore, the fact that SNPs are expected and found in such genes suggests the validity of results from this analysis.

## 4.3 Mild vs Severe

As shown in Figures 3.10 and 3.11, there is a clear difference between the amount of SNVs uniquely expressed in the mild and severe subject groups. However, it is unclear why at this time. Because the two groups show different SNVs in different genes in the same genesets (Table 3.5), this suggests that the pathways are being regulated differently, which could result in the difference in clinical responses. If mutations occur in regulatory genes, particularly in receptors, the immune system could enter a feedback loop, resulting in an over reactive immune response. Such a response tends to be the source of clinically severe symptoms; the host's system begins attacking healthy red blood cells as well as infected ones with no apparent regulatory feedback to end the response. The fact that the severe group returned so many more SNVs in genes involved in such regulatory pathways could suggest that the parasite itself is causing some post-translational changes in the host. However, further targeted testing would need to be done in order to confirm or deny this.

Cytokine-cytokine receptor interactions are particularly important to response severity, and are one gene set that was uniquely enriched in the severe group. From within that set, SNVs in genes such as IL6, IL10, and IFNA were also unique to the severe group. These genes code for interleukins and interferon, respectively, and are influential in regulating inflammatory and other innate immune reactions. In fact, mutations in IL6 and other type I interferons have recently been connected to malaria severity in humans [36]. Figure 4.1 depicts the role of various type I interferons in a case of severe malaria, including IL6, IL10, and IFNA [36]. This not only validates findings from this analysis, but also serves as further justification for the use of non-human primates, particularly *M. mulatta*, as a model organism.

Furthermore, similar analyses can be done on the parasite itself to see if there is any correlation between the parasite genome and the clinical outcomes of the subjects. Perhaps there is some selection occurring as the parasite replicates within a host in response to the host's immune reaction. Already there is evidence to support the idea that the parasite responds quickly to changes in the host's environmental conditions [37]. Following any of these lines of inquiry could provide some invaluable insights to host-pathogen relationships and disease severity for diseases beyond malaria.

## 4.4 Future Directions With SHIELD

SHIELD was designed with the intent to distribute. Thus, the pipeline could be encapsulated and published as a web service in the near future for use by other research groups. Any study with RNA-Seq reads, a reference genome, and a genome annotation can use SHIELD to generate useful SNV files for further study, and the described analyses can be easily repeated. This presents an opportunity to inquire about pan-pathogenic commonalities among genes identified as more likely to be affected by variations. In the future, SHIELD could include additional methods of analysis such as the dynamic network biomarker (DNB) method, which was designed to detect early warning signals of primary tumor cell metastasis [38].

The model behind this method could theoretically be applied to host-pathogen systems as well to determine the early warning signs of a severe immune reaction, or perhaps even predict susceptibility.

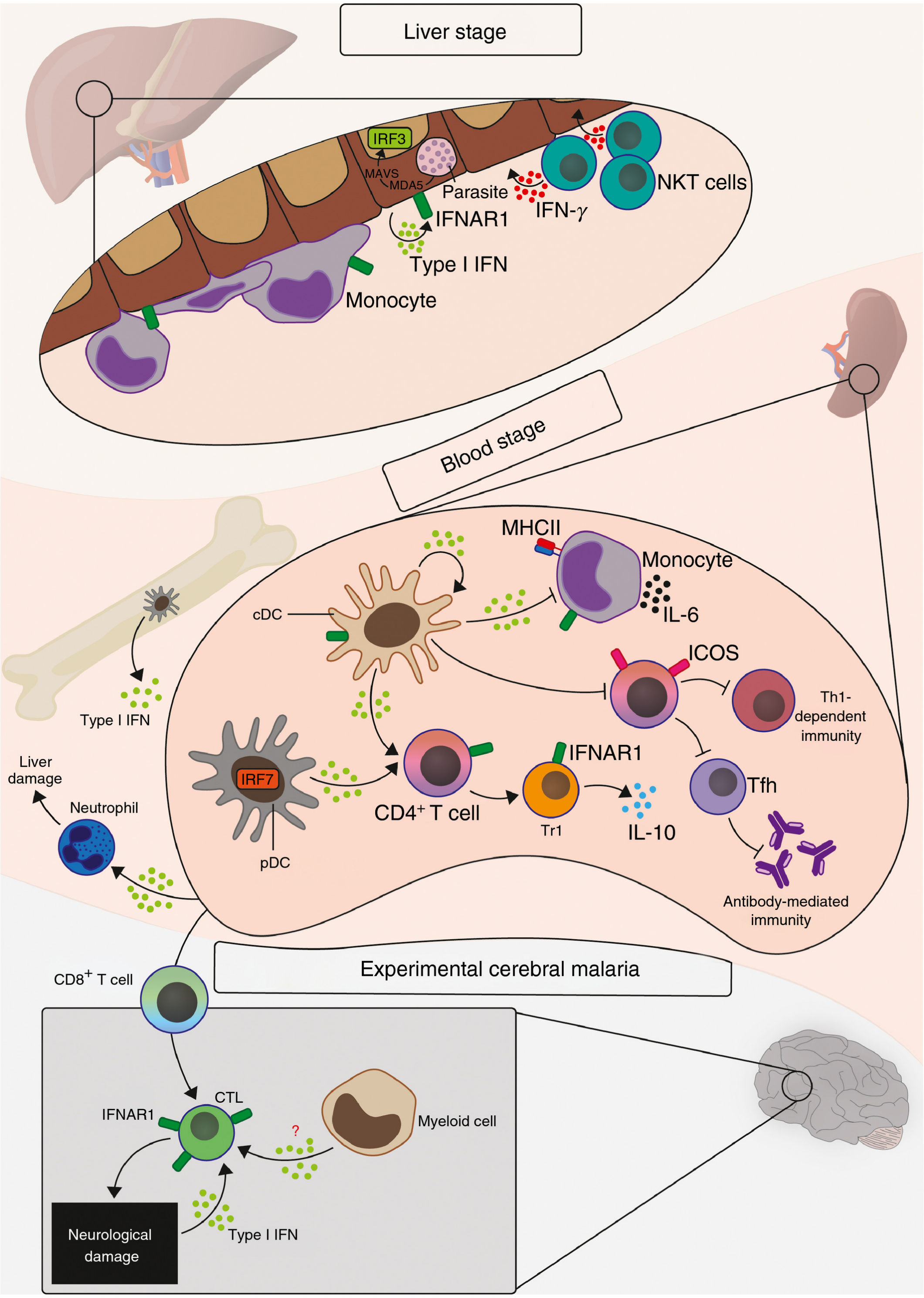


Figure 4.1: **Role of Type I Interferons in Malarial Infections**

Depiction of the various type I interferon responses during a severe malarial response. Image taken from Sebina et al [36]. Examining the named interferons reveals associations between specific interferons and response severity. Among those listed are IL6, IL10, and IFNA, which were mentioned above.

# Bibliography

[1] NIH, "What are single nucleotide polymorphisms (SNPs)?" Jul. 2018. [Online]. Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp

[2] E. Rees, M. C. ODonovan, and M. J. Owen, "Genetics of schizophrenia," *Current Opinion in Behavioral Sciences*, vol. 2, pp. 8–14, 2015.

[3] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1001–D1006, 2014.

[4] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome Research*, vol. 17, pp. 1520–1528, 2007.

[5] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common SNPs explain a large proportion of the heritability for human height," *Nature Genetics*, vol. 42, p. 565, Jun. 2010.

[6] J. Z. Liu, S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, T. Shah, S. Abedian, J. H. Cheon, J. Cho, N. E. Daryani, L. Franke, Y. Fuyuno, A. Hart, R. C. Juyal, G. Juyal, W. H. Kim, A. P. Morris, H. Poustchi, W. G. Newman, V. Midha, T. R. Orchard, H. Vahedi, A. Sood, J. J. Y. Sung, R. Malekzadeh, H.-J. Westra, K. Yamazaki, S.-K. Yang, International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, J. C. Barrett,

A. Franke, B. Z. Alizadeh, M. Parkes, T. B K, M. J. Daly, M. Kubo, C. A. Anderson, and R. K. Weersma, “Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations,” *Nature Genetics*, vol. 47, p. 979, Jul. 2015.

[7] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, “10 Years of GWAS Discovery: Biology, Function, and Translation,” *The American Journal of Human Genetics*, vol. 101, pp. 5–22, 2017.

[8] T. A. Pearson, “How to Interpret a Genome-wide Association Study,” *JAMA*, vol. 299, no. 11, p. 1335, Mar. 2008.

[9] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature reviews. Genetics*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[10] J. T. Hill, B. L. Demarest, B. W. Bisgrove, B. Gorsi, Y.-C. Su, and H. J. Yost, “MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq,” *Genome Research*, vol. 23, pp. 687–697, 2013.

[11] J. A. C. Wilson, N. A. Prow, W. A. Schroder, J. J. Ellis, H. E. Cumming, L. J. Gearing, Y. S. Poo, A. Taylor, P. J. Hertzog, F. Di Giallonardo, L. Hueston, R. Le Grand, B. Tang, T. T. Le, J. Gardner, S. Mahalingam, P. Roques, P. I. Bird, and A. Suhrbier, “RNA-Seq analysis of chikungunya virus infection and identification of granzyme A as a major promoter of arthritic inflammation,” *PLOS Pathogens*, vol. 13, no. 2, p. e1006155, Feb. 2017.

[12] F. Lu, A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney, M. D. Casler, E. S. Buckler, and D. E. Costich, “Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol,” *PLoS Genetics*, vol. 9, no. 1, p. e1003215, Jan. 2013.

[13] J. Duitama, P. K. Srivastava, and I. I. Mandoiu, "Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data," *BMC Genomics*, vol. 13, no. 56, pp. 1–10, 2012.

[14] J. C. Glaubitz, T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler, "TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline," *PLoS ONE*, vol. 9, no. 2, p. e90346, Feb. 2014.

[15] P. De Wit, M. H. Pespeni, J. T. Ladner, D. J. Barshis, F. Seneca, H. Jaris, N. O. Therkildsen, M. Morikawa, and S. R. Palumbi, "The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis," *Molecular Ecology Resources*, vol. 12, no. 6, pp. 1058–1067, Nov. 2012.

[16] J. G. Reid, A. Carroll, N. Veeraraghavan, M. Dahdouli, A. Sundquist, A. English, M. Bainbridge, S. White, W. Salerno, C. Buhay, F. Yu, D. Muzny, R. Daly, G. Duyk, R. A. Gibbs, and E. Boerwinkle, "Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline," *BMC Bioinformatics*, vol. 15, no. 1, p. 30, 2014.

[17] R. G. Coatney, W. E. Collins, M. Warren, and P. G. Contacos, *The Primate Malarias*. U.S. Department of Health, Education, and Welfare, 1971.

[18] "World Malaria Report 2017," World Health Organization, Tech. Rep., 2017.

[19] P. W. Gething, A. P. Patil, D. L. Smith, C. A. Guerra, I. R. Elyazar, G. L. Johnston, A. J. Tatem, and S. I. Hay, "A new world malaria map: Plasmodium falciparum endemicity in 2010," *Malaria Journal*, vol. 10, no. 1, p. 378, 2011.

[20] P. W. Gething, I. R. F. Elyazar, C. L. Moyes, D. L. Smith, K. E. Battle, C. A. Guerra, A. P. Patil, A. J. Tatem, R. E. Howes, M. F. Myers, D. B. George, P. Horby, H. F. L. Wertheim, R. N. Price, I. Mueller, J. K. Baird, and S. I. Hay, "A Long Neglected

World Malaria Map: Plasmodium vivax Endemicity in 2010," *PLoS Neglected Tropical Diseases*, vol. 6, no. 9, p. e1814, 2012.

[21] S. I. Hay and R. W. Snow, "The Malaria Atlas Project: Developing Global Maps of Malaria Risk," *PLoS Medicine*, vol. 3, no. 12, p. e473, Dec. 2006.

[22] T. Ponnudurai and A. H. W. Lensen, "Feeding behaviour and sporozoite ejection by infected Anopheles stephensi," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 85, pp. 175–180, 1991.

[23] M. Arvalo-Herrera, M. Lopez-Perez, L. Medina, A. Moreno, J. B. Gutierrez, and S. Herrera, "Clinical profile of Plasmodium falciparum and Plasmodium vivax infections in low and unstable malaria transmission settings of Colombia," *Malaria journal*, vol. 14, no. 1, p. 154, 2015.

[24] S. B. Millar and J. Cox-Singh, "Human infections with Plasmodium knowlesizoonotic malaria," *Clinical Microbiology and Infection*, vol. 21, no. 7, pp. 640–648, 2015.

[25] Emory, "Ongoing MaHPIC Research Projects," May 2018. [Online]. Available: http://www.systemsbiology.emory.edu/research/projects/index.html

[26] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[27] "Picard," May 2017. [Online]. Available: http://broadinstitute.github.io/picard

[28] A. G. Doran and C. J. Creevey, "Snpdat: Easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms," *BMC Bioinformatics*, vol. 14, no. 1, p. 45, 2013.

[29] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, “Circos: an Information Aesthetic for Comparative Genomics,” *Genome Research*, 2009.

[30] The MaHPIC Consortium, C. Joyner, A. Moreno, E. V. S. Meyer, M. Cabrera-Mora, J. C. Kissinger, J. W. Barnwell, and M. R. Galinski, “Plasmodium cynomolgi infections in rhesus macaques display clinical and parasitological features pertinent to modelling vivax malaria pathology and relapse infections,” *Malaria Journal*, vol. 15, no. 1, Dec. 2016.

[31] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdttir, P. Tamayo, and J. P. Mesirov, “Molecular signatures database (MSigDB) 3.0,” *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[32] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[33] A. Liberzon, C. Birger, H. Thorvaldsdttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, “The Molecular Signatures Database Hallmark Gene Set Collection,” *Cell Systems*, vol. 1, no. 6, pp. 417 – 425, 2015.

[34] M. Deschamps, G. Laval, M. Fagny, Y. Itan, L. Abel, J.-L. Casanova, E. Patin, and L. Quintana-Murci, “Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes,” *American Journal of Human Genetics*, vol. 98, no. 1, pp. 5–21, Jan. 2016.

[35] S. J. McTaggart, D. J. Obbard, C. Conlon, and T. J. Little, "Immune genes undergo more adaptive evolution than non-immune system genes in Daphnia pulex," *BMC Evolutionary Biology*, vol. 12, p. 63, May 2012.

[36] I. Sebina and A. Haque, "Effects of Type I Interferons in malaria," *Immunology*, 2018.

[37] N. Mideo, S. E. Reece, A. L. Smith, and C. J. E. Metcalf, "The Cinderella syndrome: why do malaria-infected cells burst at midnight?" *Trends in Parasitology*, vol. 29, no. 1, pp. 10–16, Jan. 2013.

[38] B. Yang, M. Li, W. Tang, W. Liu, S. Zhang, L. Chen, and J. Xia, "Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma," *Nature Communications*, vol. 9, no. 1, Dec. 2018.