UNDERSTANDING THE FUNCTIONAL IMPACT OF GENOMIC VARIANTS

by

ANBO ZHOU

A dissertation submitted to the

School of Graduate Studies

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Quantitative Biomedicine

Written under the direction of

Jinchuan Xing

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2020

ABSTRACT OF THE DISSERTATION

Understanding the functional impact of genomic variants

By ANBO ZHOU

Dissertation Director:

Dr. Jinchuan Xing

Structural variations (SV) can lead to DNA rearrangements and frequently cause diseases such as neurological disorders. SVs account for more total nucleotide changes and occur more frequently than single nucleotide polymorphisms (SNPs) (Stankiewicz and Lupski, 2010). As we continue to gain knowledge, SV has surpassed SNPs in its effects on human evolution, population diversity, and genetic diseases (Stankiewicz and Lupski, 2010). Compared to SNP, SV is more challenging to study due to its complex configuration, large size, and repetitive arrangement. Meanwhile, sequencing technologies including Illumina and Oxford Nanopore sequencing platform are being actively developed to generate sequencing data of human whole genomes, which can then be analyzed to study genetic variations. This series of studies aims to employ contemporary sequencing technologies and computational workflows to unravel the functional impact of SVs.

Good tools are prerequisite to the successful execution of a job. My study starts from developing a pipeline construction tool called PipelineDog that can

be used throughout the work. PipelineDog is a web-based integrated development environment (IDE) that represents a novel way to arrange and define workflows while promoting code scalability and reusability. I then apply established tools and workflows to analyze a 192-invidual cohort, surveying the large structural genetic etiology of autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) co-occurrence. Lastly, the newly commercialized Nanopore sequencing technique was tested and evaluated on both existing and simulated data. The Nanopore sequencing is anticipated to improve the SV identification, as it generates longer reads and will enrich the SV determining evidence. I improved the overall SV identification accuracy by employing a random forest machine learning model to classify the combined dataset from different workflows. This analysis shed light on how to determine which SV identification workflow to use based on specific use cases for future projects.

Acknowledgements

I would like to thank my advisor, Dr. Jinchuan Xing for his guidance and support. I'd like to thank Drs. Mike Verzi, Linda Brzustowicz, Judy Flax, Karen Schindler, Tara Matise, and Kelvin Kwan their continued mentorship and collaboration. I'd like to thank Timothy Lin for the wet lab work, Drs. Yeting Zhang, Yazhou Sun, Katarzyna Tyc, Xiaolong Cao, Shuoguo Wang, Nan Wang, HongSoek Ha, JiXia Liu and all the lab members for their assistance and friendship. I want to thank my parents for their financial and emotional support during my eight years in the U.S. And I want to thank my girlfriend Jiaqi Tian for being supportive and cheerful, and making me a lot of delicious dishes.

It was not an easy journey as I was one of the first (and last) 2+2 program students to come to Rutgers to study in Biotechnology major from China. The program was not designed perfectly, and there was a risk that we cannot finish our coursework in four year as planned out. My undergraduate advisor, Dr. Zilinskas took the impossible task and tirelessly communicated between the SCUT and Rutgers, and eventually helped me graduated on time.

Getting admitted to the Ph.D. program was not easy either. I did not do well in my coursework in my first semesters at Rutgers (the program putting Microbiology, Biochemistry and Molecular Genetics all in the first semester definitely didn't help). Dr. Jinchuan Xing trusted me against all odds and allowed me to do research in his lab. I got a few publications and that gave me an advantage in applying for the Molecular Bioscience Ph.D. program at Rutgers.

I enjoyed some happy lab rotations in Dr. Tara Matise, and Dr. Linda Brzustowicz's labs. But bad things just don't stop coming on to me. As a student

with more dry lab experiences rather than wet lab, I had problems passing the qualifying exams. After failing it twice, Dr. Xing suggests me to transfer Quantitative Biomedicine program and try me luck there. Dr. Gail Arnold, my advisor at Quantitative Biomedicine, was kind enough to forgive me for not knowing what "sickle cell disease" was and accepted me to the new program.

Except for the trouble I created for myself, everyone was so nice to me. My committee chair, Dr. Mikael Verzi, provided me a big computer to play with and invited me to do research in intestine development with his talented lab members, Namit, Lei, Oscar, Kevin, Pooja, Ansu and Shannon. Dr. Karen Schindler allowed me to do research in the very unfamiliar, but very important female aneuploidy field, which I greatly appreciate. Dr. Linda Brzustowicz and Dr. Judy Flax gave me their huge hard-collected dataset to dig out gene responsible for Autism. Not to mention Dr. Xing was super open-minded and encouraged me to test out my own ideas, and at the same time immensely supportive and did not let me fall out of the track.

If things don't go wrong again, I'm planning to start working in the pharmaceutical industry after my defense. I want to thank Dr. Richard Copin for not forgetting me after 3 years since we worked together in a Hackathon and found a way to get me over to his place. And how can I forget Dr. Xing, again, as he was kind enough to even helped edit my job interview slides and arranged an entire lab meeting for me to practice it.

Looking back, I understand that my success is by no means my own, but the collective work by many. Therefore, I'll continue to live my life with gratefulness and the appreciation of the kindness of others.

Table of Contents

List of Tables

List of Illustrations

1. Introduction

1.1. Genomic variants

Deoxynucleic acid (DNA) is the fundamental genetic material that stores biology information unique to an individual. It is observed that in any two humans, 99.9% of the DNA sequences are identical. It is the 0.1% of the genomic variation that underlies why individuals are heritably different from each other, including susceptibility to diseases. A 1000 Genomes Project study in 2015 characterized over 88 million variants, including 84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants in 2,504 human genomes from 26 populations (Auton, et al., 2015).
In early days, only structural or quantity variation in chromosome, such as aneuploidies, rearrangements, heteromorphisms and fragile sites, could be observed using a microscope, hence the name microscopic structural variants. The emergence of genome-scanning array technologies and comparative DNA-sequence analyses revealed submicroscopic CNVs and other structural variations that also contribute to genomic variation, influence gene dosage and have functional impact (Feuk, et al., 2006). With the development of molecular biology and sequencing technology, more abundant small-scale variants were discovered, which include SNPs, repetitive elements such as micro- and minisatellites, and small indels. Among them, SNPs are thought to be the most abundant in numbers and well-studied (Kruglyak and Nickerson, 2001).

Together, these genomics variations define a unique individual, including their phenotypic characteristics and predisposition to diseases. Genomics variation also underpins resilience and persistence of our population in that it

create the gene pool which serves as a reservoir of genetics materials for human evolution and maintain the genetic diversity of human species (Furlan, et al., 2012).

1.2. Functional impact of genomic variants

It is important to study genomic variants because they, together with environmental factors, underlie various human genetic disorders. Human genetic disorders include single gene disorders, chromosomal imbalances, epigenetics, cancer, and complex disorders (Jackson, et al., 2018).
Many conditions and diseases depend on the variant at a single locus with inheritance pattern according to Mendel's law of segregation, independent assortment, and dominance. Therefore, these diseases are often called 'Mendelian' despite the few exceptions that some inherited disorders don't follow Mendel's law.

Because three codons determine an amino acid in the protein sequence, deletion of three or a multiple of three nucleotides from a coding sequence leads to the deletion of one or more codons. These deletions lead to one or more amino acids deletion within the protein sequence but keep the remaining amino acid sequence intact and allows the protein to still be functional. However, if a number of nucleotides which are not divisible by three are deleted from a coding region, all codons following the nucleotides will be altered, this causes a 'frameshift' mutation which will likely create an malfunctional protein.
Diseases more common in human populations are often complex multifactorial disorders like autism spectrum disorder (ASD), attention deficit/hyperactivity disorder (ADHD), schizophrenia, diabetes, and heart disease, where the disorder is caused by complex interactions of multiple genetic and environmental factors.

The impact of an individual variant in one gene may be very small, but when present together with multiple variants in other genes, in the context of a particular environment, may lead to an increased risk of disease. Complex diseases are hard to study because many genes are involved, environmental factors add complication, both common and rare variants are contributing, and not all individuals with the genotype will have the disease. In recent years, research focusing on challenging topics such as genetic contributions to complex diseases has drawn increasing attention. This is why I put my focus on studying complex disease like ASD and ADHD in chapter 3.

1.3. Methods for genomic variants discovery

Methods of detecting genomic variants have evolved over the years. SNP arrays are the early high-throughput approach which can detect >1M different human SNPs each run and are still popular today. As the technology matures, call rates (the fraction of SNPs on the array that can be reliable called) and reproducibility of SNP calls can be as good as 99.5%. In addition, the same arrays can also be used to detect copy number variants. The most commonly used methods to detect SNPs are allele discrimination by hybridization as used by Affymetrix. However, the downsides of microarrays are their design requires a priori knowledge of the genome or genomic features, they struggle in cross-hybridization between similar sequences, they rely on amplified genetic materials which generates bias, and they suffer from high signal-to-noise ratio (Bumgarner, 2013).

The first-generation sequencing, also known as Sanger Sequencing, was originally developed by Frederick Sanger and colleagues in 1975. It had advantages of being less complex and easy to scale up compared to other

competing methods therefore ultimately prevailed (Sanger and Coulson, 1975). In the Sanger sequencing method practiced today, DNAs are fragmentized into different sizes and labeled with fluorescent dye corresponding to the ending base. Then electrophoresis is used to sort the fragments based on their length. In combination of all the last base signals, it is able to determine the original sequence. The method results in reads ~800-1000 base pairs in length but has the limitation of low throughput and high cost.

To address the problems posed by Sanger Sequencing, the second-generation sequencing achieves a higher throughput by sequencing a large number of DNA molecules in parallel. The earliest of the next-generation sequencing (NGS) technology includes pyrosequencing 454 by Roche, which allows the DNA fragments to affixed to micro-beads, which are pyro-sequenced in parallel. However, the technology suffers from an inaccuracy in homopolymers. Another sequencing technology, SOLiD, has a similar library preparation to pyrosequencing 454, and determines the clonal sequence represented on each bead by sequential rounds of ligation to a collection of dinucleotide-encoded adapters. The SOLiD platform achieves a higher sequence accuracy because it interrogates each base twice in sequential rounds of ligation to dinucleotide-encoded adaptors (Hurd and Nelson, 2009).

However, the Illumina sequencing platform has been the platform of choice in NGS due to its cost-effective massive parallel, high-throughput capabilities. In Illumina platform, DNA clusters are generated through bridge amplification on a glass surface rather than agarose beads to increase density. The DNA strands go through “wash-and-scan” operation (Metzker, 2010), flooding in reagents, incorporating nucleotides into the DNA strands, stopping the incorporation

reaction, washing out the excess reagent, scanning to identify the incorporated bases and finally treating the newly incorporated bases to prepare the DNA templates for the next "wash-and-scan" cycle. The cycle is repeated until the reaction is not viable. This reversible terminator chemistry can sequence up to 100 million clusters in parallel and overcomes the homopolymer difficulties. However, because yields of the addition of each base is increasingly lower, a population of DNA molecules might lag behind in synthesis each cycle. This determined that Illumina sequencing can only generate short reads, making it most suitable for well-annotated genomes and discovering small variants.

Since NGS platforms require fragmentation and have limitations in read length, people have attempt ways to sequence the entire DNA molecule directly in the "third generation sequencing" approaches, one of which is the Oxford nanopore sequencing. The concept of nanopore sequencing is that if bases could induce different ionic current bursts during DNA traversing through a tiny channel, the ionic current can be then be captured and analyzed to infer sequencing of the molecule. In 1993, Deamer, Branton, and Kasiannowicz employed α-hemolysin (α-HL), a toxic pore-forming protein secreted by Staphylococcus aureus to attack a lipid bilayer, to detect DNA translocation through α-HL nanopore (Song, et al., 1996). The same α-HL nanopore is used in Oxford nanopore sequencing platform. The pores are inserted into a lipid bilayer which has separates small chambers connected to a cathode and an anode of a path clamp amplifier. The 1.4 nm diameter of the α-HL nanopore allows only a single strand DNA or RNA to traverse through. Different bases along the negatively charged DNA or RNA strand will cause electric current fluctuations and the signal is captured and converted to sequencing information using

computer algorithms (Wang, et al., 2014). Despite the high error rate at the current stage, the major advantage of the Nanopore sequencing is the ability to obtain ultra-long reads, the inexpensive starter bundle, and the small size sequencer that can be taken to places did not have sequencer access before. Currently, long-read sequencing approaches are mainly used to investigate genetic disorders with previously known or strongly suspected disease loci (Mantere, et al., 2019). If we could adapt previous short-read optimized algorithms or develop new algorithms for the long-read sequencing technologies, we can soon enable true whole genome sequencing routinely and allow the de novo assembly of individual whole genomes used as a generic test for genetic disorders (Mantere, et al., 2019).

1.4. Approach

To understand the impact of genomic variations in human, in this dissertation I focused on three areas:

In chapter 2, I developed a pipeline building tool for fast and extensible bioinformatics workflow development. Analysis pipelines are an essential part of bioinformatics research. In my research, for example, I frequently create *ad hoc* pipelines for prototyping and exploratory analysis purposes. However, most existing pipeline management systems or work-flow engines are too complex for rapid prototyping or learning the pipeline concept. A lightweight, user-friendly, and flexible solution is thus desirable. In this chapter, I developed a new pipeline construction and maintenance tool, PipelineDog, which is a web-based integrated development environment with a modern web graphical user interface. It offers cross-platform compatibility, project management capabilities, code formatting and error checking functions, and an online repository. It uses

an easy-to-read/write script system that encourages code reuse. With the online repository, it also encourages sharing of pipelines, which enhances analysis reproducibility and accountability. For most users, PipelineDog requires no software installation. Overall, this web application provides a way to rapidly create and easily manage pipelines.

I'm interested in autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) because they are two major neurodevelopmental disorders that frequently co-occur. However, the genetic mechanism of the co-occurrence remains unclear. As an effort to understand the genetic etiology of ASD, the New Jersey Language and Autism Genetics Study (NJLAGS) collected more than 150 families with family members affected by ASD where we found a large overlap among individuals and families affected by both ASD and ADHD. This provided a suitable dataset for my study. In chapter 3, I made use of the NJLAGS dataset and analyzed whole genome sequencing data on 272 samples from 73 families with either ASD or ADHD and identified candidate genes using linkage analysis and association analysis for SNPs and indels and burden analysis for SVs. I then further scaled down the candidate gene set using various functional annotations and reported the potential causal genes for ASD and ADHD discovering the common genetic risk factors underlying ASD and co-occurring ADHD.

Even though we learn abundant knowledge from SNPs, it is also important to study structural variations (SVs) as they account for about 1% of the differences between human genomes and play a significant role in phenotypic variation and disease susceptibility. Widely adopted next generation sequencing technologies have short read-length which limits their ability to identify SVs and

the emerging nanopore sequencing technology applies real-time single-molecule sequencing and can generate long sequence reads. Despite nanopore sequencing's potential to facilitate better SV identification, the available tools for aligning long-read data and detecting SVs have not been thoroughly evaluated. In chapter 4, I used four human nanopore datasets, including both empirical data and simulated reads, to evaluate four alignment tools and three SV detection tools. I then established a recommended workflow for analysis nanopore sequencing data for SV discovery, as well as proposing a random forest machine learning approach to combine and improve the datasets generated by different workflows. The nanopore technology keeps improving and the nanopore sequencing community is likely to grow accordingly. In turn, better benchmark call sets will be available to more accurately assess performance of available tools and facilitate further tool development.

2. PipelineDog: a simple and flexible graphic pipeline construction and maintenance tool

2.1. Introduction

Analysis pipelines (also called workflows) are routinely used in bioinformatics studies. A pipeline can contain a few related steps to be run as a serial process, or dozens of steps and need to be run as concurrent processes. Although highly curated, optimized, and quasi-standardized workflows exist for specific goals (e.g., (Van der Auwera, et al., 2013), (Wang, et al., 2011)), researchers frequently create ad hoc pipelines for prototyping and proof-of-concept purposes outside of the production environment. Considering the diversity of biomedical studies, and the overwhelmingly large number of available analysis tools, it has become a challenge to efficiently create and manage these ad hoc pipelines.

Dedicated workflow engines have been developed to address these issues (Reviewed in (Leipzig, 2016)). Popular workflow engines, such as Galaxy (Goecks, et al., 2010), Taverna (Oinn, et al., 2004), and Pegasus (Deelman, et al., 2015), offer graphical user interface (GUI) for creating pipelines, and execution environments for running pipelines. However, installing, running, and maintaining such large systems require substantial resources and expertise and they are more suitable for large research groups to implement well-tested workflows. At the same time, several relatively lightweight workflow management tools have also been developed, for example, Ruffus (Goodstadt, 2010), Snakemake (Koster and Rahmann, 2012), and BuddySuite (Bond, et al., 2017). These lightweight systems are less feature-rich, but they significantly reduced the time between pipeline conception and implementation. However,

these lightweight tools are all command-line tools, requiring working knowledge of programming languages and the installation and maintenance of the running environment. For users without extensive programming experience, a flexible lightweight tool with minimum installation and maintenance requirement is desired.

As a result, we developed PipelineDog, a lightweight tool in a modern web GUI. PipelineDog is specifically designed for users with little programming experience but has the need to perform computational analyses or learn the principles of workflow/pipeline design. For example, clinicians with little programming experience but want to explore the sequencing data analyses, or students learning the principles of workflow design. For these users, the web GUI provides sufficient functionality for constructing and debugging pipelines, and no additional software installation is needed. Only a minimal amount of reading is needed to use an easy-to-read/write PipelineDog Script and LEASH Expression system that we developed. On the other hand, for experienced users, PipelineDog can serve as a quick workflow experimental and prototyping tool. To encourage code reuse and enhance analysis reproducibility, we also provide an online repository for steps and pipelines.

2.2. System Design and implementation

The overall design goal of PipelineDog system is to increase efficiency for rapid prototyping. To achieve this goal, we aim to: 1st, improve usability by offering a modern web-based GUI and an integrated development environment (IDE); 2nd, encourage code reuse by providing a modular and human friendly scripting language (PipelineDog Script and LEASH Expression); and 3rd, enhance

community efforts and code/pipeline sharing by providing a public pipeline repository.

PipelineDog views a pipeline as a series of operations on a list of files (Figure 2.1A). Each analysis step can have a list of input files, and a list of output files. Downstream steps can use any or all of the upstream steps' output lists. Users use a PipelineDog Script to specify the pipe-line's individual steps. If needed, LEASH Expression (LEASH stands for Line Entry Automated Shuffling) can be used to dynamically select input files, alter file output directory or file names, and define specific patterns to iterate through files. Analysis steps in PipelineDog Script can then be chained or nested to create an entire workflow (Figure 2.1A).

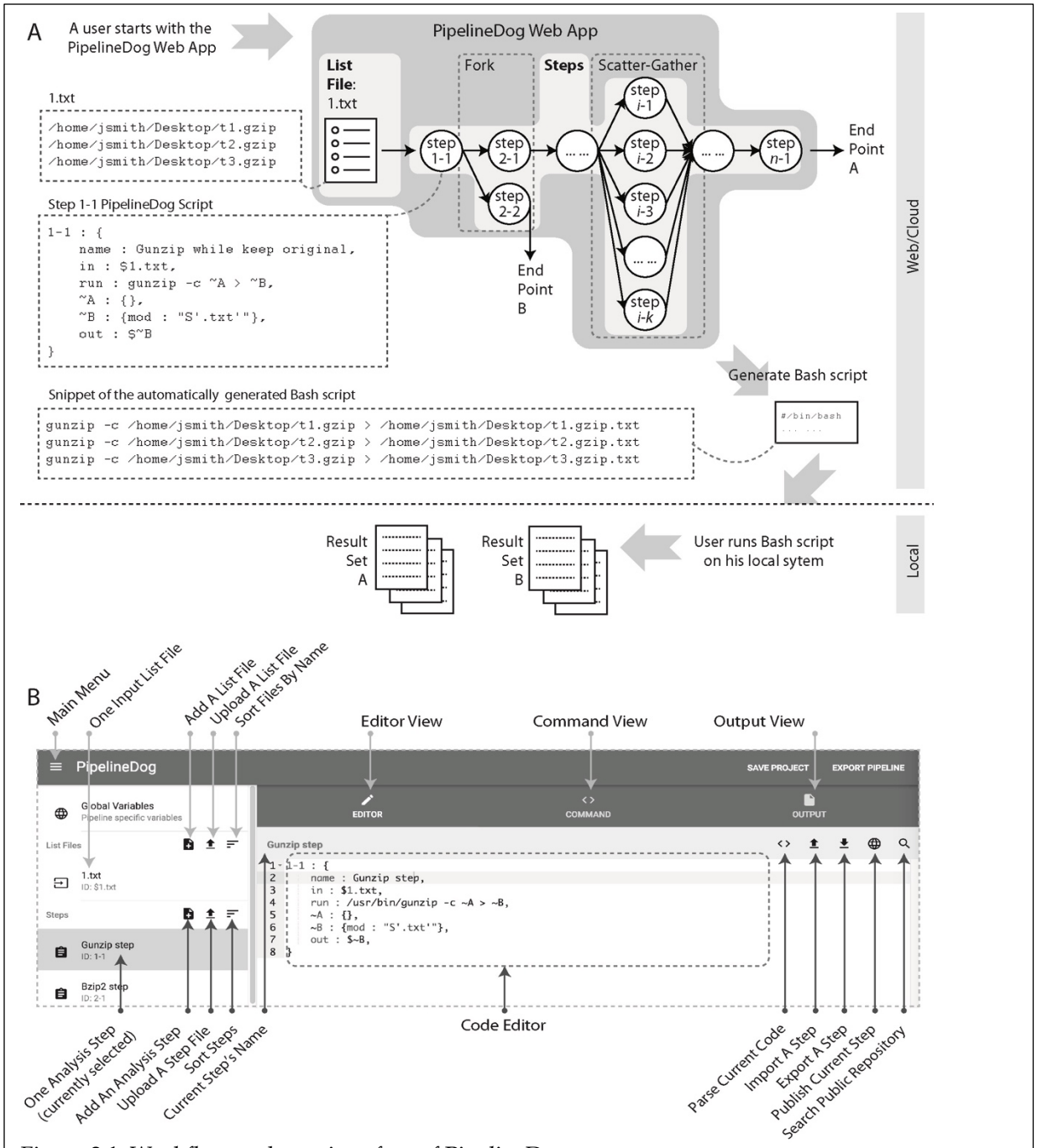


Figure 2.1. Workflow and user interface of PipelineDog.
A) Workflow of the PipelineDog web app. The List File, Step definition script, and generated Bash code are shown for a simple example to unzip files while keeping the original compressed version. B) The main interface of the PipelineDog web app. On the left side are tabs allowing a user to manage List Files and Steps. The main Code Editor panel on the right side contains format validation, parsing, and import/export functionalities of a typical IDE. The Editor section has three views: Editor, Command, and Output views, allowing a user to interactively edit a step, view the parsed commands, and view the output file paths, respectively. The Export menu on the top right corner allows a user to export a BASH script for the complete pipeline. Additional functionalities include saving a PipelineDog project as a local file, and searching for example steps or pipeline scripts in our online community-driven repository.

The PipelineDog Script and LEASH Expression system are designed for rapid prototyping needs. It requires less coding than the commonly used Common Workflow Language (CWL) or Workflow Description Language (WDL) (see Figure 2.2 for an example). PipelineDog Script allows simple implementation of both serial operations and more com-plex workflow patterns (e.g., forks and scatter-gather). PipelineDog Script and LEASH Expression are based on YAML and are designed for maximal convenience to read/write for users without extensive programming experience. The detailed technical specification for PipelineDog Script and LEASH Expression can be found in the supple-mental file and online (https://github.com/ysunlab/PipelineDog). A simple example pipeline is illustrated in Figure 2.1A with a more detail example in the "Application" section.

Unlike almost all previous systems, PipelineDog uses the Bash shell as the execution engine, which is available on most Unix-like systems, Mac OS, and is easily installed on Windows systems. By leveraging the ubiquitous Bash shell, for most users using PipelineDog requires no installation of new software.

BASH: bwa mem –t 8 –k 19 ref.fa reads1.fq reads2.fq > output.sam

PipelineDog

```
$bwa.step:
  name: bwa_mem
  in: $input_reads

  run: >-
    bwa mem
    -t $bwa.threads
    -k $bwa.min_seed_length
    $bwa.reference
    ~reads > output.sam

   ~reads:
    line: -:0
    sep: ' '

  out: output.sam
```

Workflow description language

```
task bwa_mem_tool {
  Int threads
  Int min_seed_length
  Int min_std_max_min
  File reference
  File reads

  command {
    bwa mem -t ${threads} \
            -k ${min seed length} \
            ${reference} \
            ${sep=' ' reads+} > output.sam
  }
  output {
    File sam = "output.sam"
  }
  runtime {
    docker: "broadinstitute/baseimg"
  }
}
```

Common workflow language

```
#!/usr/bin/env cwl-runner

cwlVersion: cwl:draft-3

class: CommandLineTool

hints:
  - class: ResourceRequirement
    coresMin: 4

inputs:
  - id: reference
    type: File
    inputBinding: { position: 2 }

  - id: reads
    type:
      type: array
      items: File
    inputBinding: { position: 3 }

  - id: minimum_seed_length
    type: int
    inputBinding: { position: 1, prefix: -m }

  - id: args.py
    type: File
    default:
      class: File
      path: args.py
    inputBinding:
      position: -1

outputs:
  - id: sam
    type: ["null", File]
    outputBinding: { glob: output.sam }
  - id: args
    type:
      type: array
      items: string

baseCommand: python

arguments:
  - bwa
  - mem
  - valueFrom: $(runtime.cores)
    position: 1
    prefix: -t

stdout: output.sam
```

Figure 2.2. Comparison of three workflow languages defining the same command.
The BASH command is shown at the top of the figure and the three implementations (PipelineDog, Workflow Description Language, and Common Workflow Language) are shown at the bottom.

PipelineDog is written in JavaScript and the entire PipelineDog system is packaged as a single web application (http://web.pipeline.dog). This web app itself is an IDE and provides a modern GUI with cross-system compatibility, project management features, code formatting, error checking, and an online repository (Figure 2.1B). For more advanced users, a CLI version is also

available. Additional implementation details are described in the specification section.

2.3. Discussion

The PipelineDog system provides a convenient way to create and man-age ad hoc pipelines, with the capability of producing production quality pipelines. PipelineDog Scripts are designed for users without extensive programming experience. By using the popular and ubiquitous Bash shell as the execution engine, most users can use PipelineDog with no software installation need. Additionally, we provide a repository for sharing and reusing PipelineDog scripts. With this design philosophy, we aim to encourage code reuse and the sharing of pipelines, which further enhances analysis reproducibility and accountability.

2.4. Specification

2.4.1. Project

A PipelineDog project consists of three types of definitions: global variables, list files, and step definitions. Global variables can only be defined once for the entire project, while there can be multiple list files and steps. Global variables definition stores the variable that can be accessed by all of the steps. It's recommended to put environment specific variables of each step in global variables (e.g., the Step-ID, name and in). The list file is the pipeline input which usually consists of a list of file names that are different for each pipeline run. Other parameters of the run can also be specified in the list file. The step definitions are each steps of the pipeline defined by the PipelineDog script step specifications. In a PipelineDog project, the first step reads the list files as the input, then the outputs from one step can be used by the next step in addition to the list files, thus forming a

pipeline. For detailed specification, please visit the PipelineDog Github page (https://github.com/ysunlab/PipelineDog).

2.4.2. Step

A pipeline typically contains multiple steps. A PipelineDog step definition file (as well as the entire Project definition file) is an YAML file that defines a single object with PipelineDog-specific keys. The key of the entire object is the Step-ID key. This key uniquely identifies each step and defined as two numbers separated by a dash (-). The first number is the order that steps to be executed. If the first number is the same for two steps, these two steps will be executed in parallel. In that case, the second number is used to distinguish these steps. Within the object, several additional key-value pairs are available:

1. name: a string to describe the nature/function of this analysis step.

2. in: a string or an array to provide the name of a list file, or the name of the output from a previous step.

3. run: a string to provide a template for PipelineDog to construct one or more BASH commands. The template can include LEASH targets (starts with "~") to be replaced by the following LEASH expression keys.

4. LEASH expressions: a series of objects to provide specific instructions to PipelineDog on how to modify and replace the LEASH targets inside the "run" template with user provided parameters. They also start with "~" to match the LEASH targets.

5. out: a string or an array to specify the list of new files that will be generated after this step has been successfully executed (so that later steps can get access to this list of new Line Entries)

6. comment: an optional string that provide additional information about this step for the users, typically when it is too long to put inside the name field. This will appear as commented code in the exported BASH script.

Besides Step-ID, two key-value pairs are required in Step Definition: name and run. A LEASH target and its corresponding LEASH expression are the key mechanism that allows PipelineDog to automatically format a command line argument list.

2.4.3. LEASH expression

LEASH stands for Line Entry Automated SHuffling, indicating that the line entries will be reformatted and rearranged automatically according to a set of LEASH expressions. LEASH expression is the core driver of PipelineDog, because it processes the commands based on the user-specified parameters for a specific pipeline and replace the LEASH targets marked in the run key with the dynamic parts defined by the expression. LEASH expression is defined after the run key, also using a YAML object with keys namely:

1. file: select one or more list files that will be included in the pipeline, or an output from a previous step.

2. line: select one or more line entries from the selected list file(s), and also specify how these entries should be arranged or repeated.

3. mod: modify each line entry with the optional prefix and/or suffix string(s), and also select which basic parts of an entry to retain and modify.

4. mods: use either mods or mod. Mods (stands for mod-simplified) uses a format similar to what you would want to see in the actual constructed command. It fulfills most but not all of the functions of mod, but with a much easier-to-remember syntax.

5. sep: define the separator character of the line entries if they are in the same command.

Together, the LEASH expression provides an automated way to reformat and rearrange line entries.

2.4.4. Implementation Details

Both GUI version and command line interface (CLI) version of PipelineDog are written in JavaScript, with Browserify (http://browserify.org) employed to unify the package system between these two versions. It utilizes the React framework (https://facebook.github.io/react) with JSX syntax for the front-end user interface and ES2015 (http://www.ecma-international.org/ecma-262/6.0) JavaScript syntax for the backend logic and algorithms. Material UI component set (http://www.material-ui.com) for React is used as the foundation of the user interface design. The general UI design also follows the material design (https://material.google.com) principles defined by Google. The web user interface is hosted on Github (https://github.com) as a static web page application. The CLI version of the app is hosted on npm (https://www.npmjs.com) as a downloadable Node.js package. The "save code to local" functionality, which receives the code data and generates a file for user to download, is hosted on Heroku (https://www.heroku.com) as a Node.js application. The online code repository backend is implemented using the Firebase (https://firebase.google.com) platform, which provides 4 a real-time database and a simple administrative system. YAML format is parsed using the library js-yaml (https://github.com/nodeca/js-yaml).

3. Analysis of Common Genetic Risk Factors in ASD and ADHD Co-occurring Families

3.1. Introduction

Autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) are two neurodevelopmental disorders of high prevalence and severity. ASD is characterized by deficits in social interaction and social communication, and by restricted, repetitive and stereotyped patterns of behavior, interests, and activities (Polderman, et al., 2014). ADHD is characterized by inattention and hyperactive/impulsive symptoms (Polderman, et al., 2014). The two disorders have shown a high frequency where 20-50% of children with ADHD meet the criteria for ASD and 30-80% of ASD children meet the criteria for ADHD (Rommelse, et al., 2010). Both conditions can cause a severe negative impact on the patients' life quality, and only worsen when co-occurring. Thus, a deep understanding of the genetic etiology of familial co-occurrence of ASD and ADHD is important to enhance treatment planning.

Extensive research has been conducted on both disorders and demonstrates overlapping genetic factors between the two conditions (Doherty and Owen, 2014; Ghirardi, et al., 2018; Johnson, et al., 2015; Rommelse, et al., 2010; Ronald and Hoekstra, 2011). For example, ADHD candidate causal genes *DRD3* and *MAOA* are cautiously positively associated with ASD (Rommelse, et al., 2010). In genome-wide association studies (GWAS), 16 single nucleotide variants (SNVs) related to ADHD were found to be possibly involved in ASD and 25 SNVs related to ASD are possibly involved in ADHD (Rommelse, et al., 2010). In copy number variation (CNV) studies, CNVs segregated with ADHD were also found

to be enriched for ASD candidate genes. Correspondingly, ASD patients' family members also carry the ADHD diagnosed CNVs (Elia, et al., 2010; Martin, et al., 2014). While the studies have suggested shared genetic risk factors as one possible reason for the co-occurrence of the two disorders, they either used microarray or whole exome sequencing. In addition those studies are based on case-control study and did not leverage the power of family studies. Therefore, the evidence of shared genetic risk factors between ASD and ADHD from these studies are mostly weak and inconclusive.

To further explore the etiology of ASD/ADHD concomitance, we aimed to use whole-genome sequencing (WGS) analysis to identify SNVs and structural variants (SVs) responsible for the ASD/ADHD phenotypes segregated in the New Jersey Language & Autism Genetics Study (NJLAGS) (Bartlett, et al., 2014). NJLAGS is a project studying the genetic influences of autism development and has conducted WGS on 272 samples from 73 families, each with at least one ASD proband. All family members have been characterized by extensive behavioral assessments of language and social functioning, restrictive and repetitive behaviors, and other co-occurring behaviors. Despite the original study focus on ASD, these families exhibit elevated rates of ADHD, both in ASD probands and non-ASD family members. Importantly, more than half of the ASD probands have also been diagnosed with ADHD. The high rate of co-occurrence of ASD and ADHD in the NJLAGS families provides a unique opportunity to identify genetic factors underlying the co-occurring ASD and ADHD.

3.2. Methods

3.2.1. Sample collection and phenotype assessment

The sample in this study includes 79 families from a previous NJLAGS study (Bartlett, et al., 2014) (Wave 1) plus an additional 36 families collected after the 2014 publication (Wave 2). The original aim of the NJLAGS study was to find genetic variation of potential relevance to ASD and language impairment as well as disorders related to autism and language. Families were required to have at least one person with autism and one additional person with a language-learning impairment. During Wave 1, autism probands were required to have a strict diagnosis of Autistic Disorder based on the Autism Diagnostic Interview (ADI-R), Autism Diagnostic Observation Schedule (ADOS), and the Diagnostics and Statistical Manual-IV (DSM-IV). A second proband had to meet criteria for Specific Language Impairment, a disorder where language development is delayed or deviant and cannot be explained by any other neurodevelopmental diagnosis. For Wave 2, NJLAGS sought to understand if the original strict autistic disorder diagnostic criteria were necessary, and therefore, the autism proband requirements were intentionally relaxed to become more in line with the less restrictive, newer DSM-5 criteria for Autism Spectrum Disorder (DSM-5, 2012). All subjects gave informed consent or assent conforming to the guidelines for treatment of human subjects from the Intuitional Review Board at Rutgers, The State University of New Jersey (IRB number: 13-112Mc).

For the present study, we focused on co-occurrence of ASD and ADHD. The ADHD affection status of ASD probands was determined by responses to questions specific to ADHD on three different NJLAGS questionnaires: 1) Medical History Questionnaire, 2) the Family History Questionnaire, and 3) the Language Correlates Questionnaire. Affection status of all other family members

was determined from the Language Correlates Questionnaire and the Family History Questionnaire.

3.2.2. Genotyping

Wave 1 Affymetrix Axiom array genotypes have been described previously (Bartlett et al. 2014). For Wave 2, SNP data was generated with the Illumina Infinium PsychArray-24 v1 array (Illumina, San Diego, CA), which includes 593,260 SNPs. In this study we focused on SNP with population MAF (minor allele frequency) >1%. Quality control on SNP genotypes was conducted by array batch and by array type, as described previously (Bartlett et al. 2014). The quality control criteria include individual/SNP genotype completion, relationship checking, Mendelian errors checking, and ancestry inference. The linkage analysis only included samples that clustered with the CEU samples from the HapMap reference data, as determined by EIGENSTRAT using the recommend parameters in the documentation (Patterson, et al., 2006).

Initially, a subset of 10,899 SNPs in common across the Wave 1 and 2 arrays was chosen for linkage analysis, which minimize marker-to-marker linkage disequilibrium while retaining a high MAF (>30%) to provide suitable genomic coverage of recombination events in the pedigrees. Overlap in some genomic regions was too low to retain acceptable information content as measured by MERLIN (Abecasis, et al., 2002). In those regions, array-specific SNPs were included. This procedure did not pose an issue with missing data within families since every family was genotyped using only a single array type.

3.2.3. Statistical Analysis

Linkage analyses were conducted with KELVIN 2.3.3 (kelvin.mathmed.org). KELVIN implements the posterior probability of linkage (PPL) metric to estimate

the probability that a genetic location is linked with a tested trait (Vieland, et al., 2011). Primary linkage analysis of the phenotypes was conducted on each wave separately, and the linkage evidence was "sequentially updated" across the waves using Bayes rule to provide a single metric for linkage evidence. A secondary linkage analysis was conducted using all families jointly in a single "pooled" analysis of each trait. By comparing the sequentially updated result to the pooled result, we could qualitatively infer the role of heterogeneity in the dataset. Since stratifying on an irrelevant trait will on average produce the same result as a pooled analysis (Govil and Vieland, 2008), if we observe appreciably higher sequentially updated PPL over pooled results, we may infer that heterogeneity demarcated by wave is present in the data.

Based on previous simulations of the null distribution in the NJLAGS sample when correcting for three phenotypes (Bartlett, et al., 2014), a PPL of 0.32 or greater is consistent with a genome-wide error rate of $p<0.001$, a PPL of 0.26 corresponds to $p<0.01$, and a PPL of 0.11 corresponds to $p<0.05$. These threshold values are similar to previous studies of the false positive rate of the PPL after of correction for testing multiple phenotypes (Bartlett, et al., 2002; Logue, et al., 2003).

3.2.4. DNA sequencing

DNA extraction was performed by RUCDR either from blood DNA (WB) or Lymphoblastoid cell lines (LCL). The sequencing was done in four batches by three vendors (Table 3.1). All samples were sequenced using Illumina paired-end reads with a spec of 30x read depth.

Batch	Sequenced	SNV/Indel	SV
Knome	25	13	16

Hudson Alpha	20	19	20
Hudson Alpha	150	139	147
Genewiz	102	101	97
Total	297	272	280

Table 3.1. Number of samples from each vendor.

Some of these samples were excluded for different analyses because of quality issues. A few were dropped because the subjects withdrew from the study. For samples that were sequenced in more than one batch, the best quality run was used for the analysis.

3.2.5. Small variant (SNV/indel) and structural variant calling

Alignment of paired-end fastq files was performed using the BWA-MEM algorithm (v0.7.12) to the Human Genome Reference Consortium Build 37 (hg19) using default parameters. The output was converted to BAM format using SAMtools view (v0.1.19). BAM files from read alignment were then processed using the GATK v3.5.0 variant calling pipelines followed the best practice recommendation for alignment processing and variant calling (DePristo, et al., 2011; Wang and Xing, 2013). Starting from sorted and indexed individual BAM files, a series of GATK alignment processing procedures were conducted, including PCR duplicate removal and base quality score recalibration. Variants were called per individual using HaplotypeCaller before joint called by GenotypeGVCF. All samples from different sequencing batches were joint called along with the 1000 Genomes project European ancestry samples (CEU, GBR, FIN) from the Utah Genome Project as controls to reduce the batch-effect for downstream analysis. After variant call, we employed variant quality score recalibration using VariantRecalibrator and ApplyRecalibration as outlined in

the GATK protocol using gold standard variant data from the HapMap and the 1000 Genomes projects.

The realigned reads were split into individual chromosomes to facilitate parallel computing on a slurm HPC platform. MetaSV (Mohiyuddin, et al., 2015), along with its components (Breakseq2 (Abyzov, et al., 2015), breakdancer (Chen, et al., 2009), CNVnator (Abyzov, et al., 2011), and Pindel (Ye, et al., 2009)) were run on the samples for SV discovery and local realignment by spades and AGE were carried out to further improve breakpoint resolution. MetaSV then combined all evidence produced and merged them into a single call set. The output of each chromosome from the same individual was then merged back to one file using VCFtools (Danecek, et al., 2011). Basic statistics for the SV calls were calculated by SURVIVOR (Sedlazeck, et al., 2017) for quality control.

3.2.6. Pedigree trimming

Pedigree information was organized in the PED file format. Only families affected by ADHD or ASD were included in this analysis. For running pVAAST, the pedigrees were then processed and trimmed using custom scripts to include only individuals from one pair of ancestral parents per pedigree for running the dominant mode, or a two-generation subset of the pedigree for running the recessive mode. The individuals retained were selected to maximize the number of sequenced and affected samples.

3.2.7. Variant annotation and selection

SNVs and indels were first annotated by VAT in the VAAST package (2.0.2) and then condensed into cdr files to represent one family per file by VST in the VAAST package. The variants were filtered to only include those that have an MAF <5% in the ExAC dataset excluding psychiatric cohorts

(http://exac.broadinstitute.org/). For the control samples, 635 GTEx whole-genome sequenced samples were obtained and condensed into a single group. The SV calls were annotated by AnnotSV (Geoffroy, et al., 2018) which gives a severity score for each SV. A custom script was written to discard all the SVs that had an MAF >5% in gnomAD, the 1000 Genomes Project, or preliminary results from the Center for Common Disease Genomics (CCDG) study (Abel, et al., 2018).

3.2.8. Gene prioritization

For SNV and indels, the gene prioritization tool pVAAST (v0.02) was used to find candidate genes based on the aggregative score for each variant within the coding region (Hu, et al., 2013; Hu, et al., 2014). A pVAAST score was calculated for each gene from its variants' linkage pattern, association strength, allele frequency, and functional prediction. Six pVAAST analyses were performed in total for different traits, regions, and inheritance modes (Table 3.2), including two analyses for ADHD linage regions and four whole-genome analyses. The whole-genome analyses were performed at 10^5 permutations per gene, and the linkage region analyses (chr12: 38200001- 71500000 and chr17: 3300001- 10700000) were performed at 10^6 permutations per candidate gene.

For SVs, a custom script was written to convert results from an SV-based report into a gene-based report. In both case and control individuals, the annotated severity scores of all SVs overlapping with genes were aggregated for each gene. The affected genes were ranked based on the ratio of aggregated scores of the SVs overlapping the genes in case samples versus control samples.

Trait	Region	Inheritance Mode
ADHD	Linkage region	Dominant

ADHD	Linkage region	Recessive
ADHD	Whole genome	Dominant
ADHD	Whole genome	Recessive
ASD or ADHD	Whole genome	Dominant
ASD or ADHD	Whole genome	Recessive

Table 3.2. Differences in pVAAST runs that generated.

3.2.9. Gene annotation, pathway and enrichment analysis

A custom gene-based annotation program was written to collect information from online databases, annotate, and filter the candidate gene sets. The online databases and resources used are listed in Table 3.3.

Database	Information	Link	Reference
SFARI	Consensus genes	https://www.sfari.org/resource/sfari-gene/	(Abrahams, et al., 2013)
ADHDgene	Consensus genes	http://adhd.psych.ac.cn/	(Zhang, et al., 2012)
iPSYCH	Consensus genes	https://ipsych.dk/en/	(Schork, et al., 2019)
Autism Sequencing Consortium & iPSYCH	Consensus genes	https://doi.org/10.1016/j.cell.2019.12.036	(Satterstrom, et al., 2020)
DISEASE	Disease association by literature text-mining	https://diseases.jensenlab.org/Search	(Pletscher-Frankild, et al., 2015)
gnomAD	tolerance to LoF (pLI score)	https://gnomad.broadinstitute.org/	(Karczewski, et al., 2020)
IMPC	Known phenotype changes in mouse gene knockouts	https://www.mousephenotype.org/	(Koscielny, et al., 2014)
GTEx	Differential gene expression human tissues	https://gtexportal.org/home/	(GTEx-Consortium, 2017)

brainSpan	Differential gene expression in developing brain	https://developingmouse.brain-map.org/	(Miller, et al., 2017)
GIANTS	Tissue-specific protein interaction	https://hb.flatironinstitute.org/	(Greene, et al., 2015)
Panther	Gene enrichment analysis	http://pantherdb.org/	(Thomas, et al., 2003)
KEGG	Pathway analysis	https://www.kegg.jp/kegg/mapper.html	(Kanehisa, et al., 2020)
STRING	Pathway analysis	https://string-db.org/	(Szklarczyk, et al., 2020)

Table 3.3. Online databases integrated in the annotation, pathway and enrichment analysis.

The final merged and interset candidate gene sets were then uploaded to web-based tools including Panther (Thomas, et al., 2003), KEGG (Kanehisa, et al., 2020), and STRING (Szklarczyk, et al., 2020) for gene overrepresentation and pathway analyses.

3.3. Results

3.3.1. ASD and ADHD have high co-occurrence rate in NJLAGS families

In our samples collection, 111 families were recruited requiring at least one ASD proband within the family. The study was conducted with microarray analysis and WGS analysis in parallel (Figure 3.1). The WGS analysis prioritized genes base on SNVs/indels and SVs called from sequencing data, and took advantage of the discovered linkage region from the microarray analysis to scale-down the linkage-region specific candidate genes sets. The candidate gene sets were then subjected to enrichment and pathway analyses. Among the 73 families involved in WGS, 47 (64.3%) families have individuals who are also affected by ADHD. Out of the 98 individuals affected by ASD, 41 (41.8%) are also affected by ADHD

(Table 3.4). Males are more likely to be affected by both ASD and ADHD compared to females (T-test P=0.03) (Table 3.4). Of the 493 total individuals within the 73 families, 125 are affected by either ASD or ADHD (Table 3.4).

	All	Male	Female	Families
ASD	98	77	21	73
ADHD	68	46	22	47
ASD and ADHD	41	30	11	36
ASD or ADHD	125	93	32	73
All samples	493	281	212	73

Table 3.4. Summary of families involved in WGS and their phenotypes.

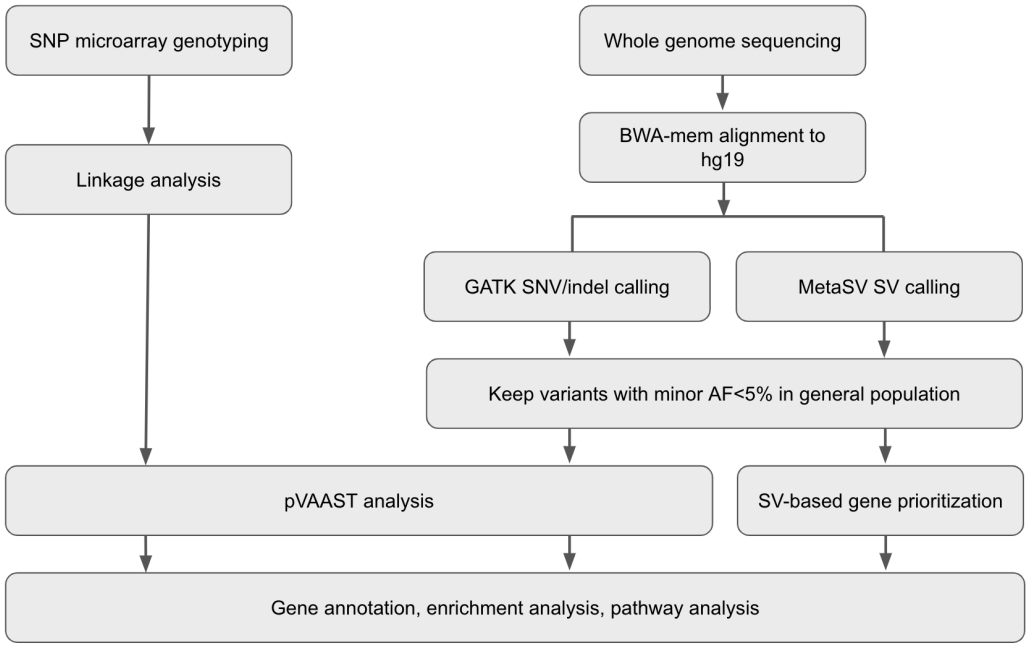


Figure 3.1. Study overall workflow.

3.3.2. Linkage analysis highlight peaks for ADHD

A total of 524 persons from 111 families were genotyped for linkage analysis. The fully constructed the pedigrees contains 707 persons, due to the need of adding missing persons, such as grandparents, to link sibships of cousins. Families have

an average of 4.2 people genotyped and an average of 5.7 people (including ungenotyped people).

Multipoint linkage analysis of the data was performed with results summarized in Figures 3.2-3.4 with peaks summarized in Table 3.5. The ADHD phenotype was linked to 17p, meeting the conventional standard for declaring linkage ($p<0.001$) as determined by a simulation study to estimate the empirical null distribution (Figure 3.2). As the pooled PPL was larger than the sequentially updated PPL, we may infer that the locus is largely homogenous, in the sense that these results did not offer evidence that data from either of the two waves are inconsistent with linkage. As such, the locus did not depend on the strictness of the ASD criteria used in ascertainment. The locus on 12q met the criteria for suggestive linkage to ADHD (Figure 3.2). Similar to the locus on 17p, the pooled PPL for 12q was larger than the sequentially updated PPL, indicating that the linkage with ADHD did not differ across the two waves of ASD criteria used for recruitment in our study. Several additional loci met the criteria nominal linkage. The ADHD phenotype is nominally linked to 3p (pooled PPL=0.11, updated PPL=0.23) with evidence for heterogeneity across ASD criteria and 19q (pooled PPL=0.20, updated PPL=0.14) with no evidence for heterogeneity across ASD criteria. The phenotype "ADHD or ASD" had two nominal linkage peaks, both suggesting heterogeneity, on 19p (pooled PPL=0.11, updated PPL=0.18) and 20q (pooled PPL=0.02, updated PPL=0.19) (Figure 3.3). The "ADHD and ASD" phenotype did not show evidence for linkage, likely due to the small sample size of the phenotype (Figure 3.4).

Trait	Chromosome	cM	Cytoband	PPL Pooled	PPL Updated
ADHD	3	95-97	3p13	0.11	0.23

ADHD	12	59-86	12q12-15	0.27	0.15
ADHD	17	12-32	17p13.1-2	**0.38**	0.17
ADHD	19	54-62	19q12-13.1	0.20	0.14
ADHD or ASD	19	5-17	19p13.3	0.11	0.18
ADHD or ASD	20	74-104	20q13.13-13.33	0.02	0.19

Table 3.5. Summary of regions of interest.

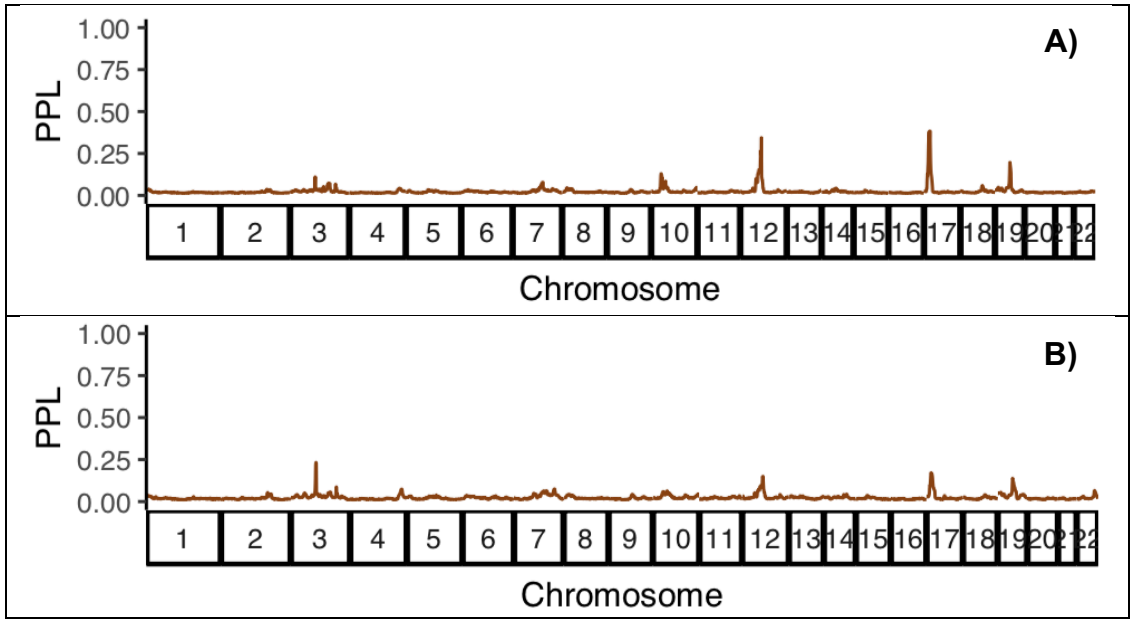


Figure 3.2. Linkage peaks for ADHD in across the genome.
(A) pooled samples (B) sequential update.

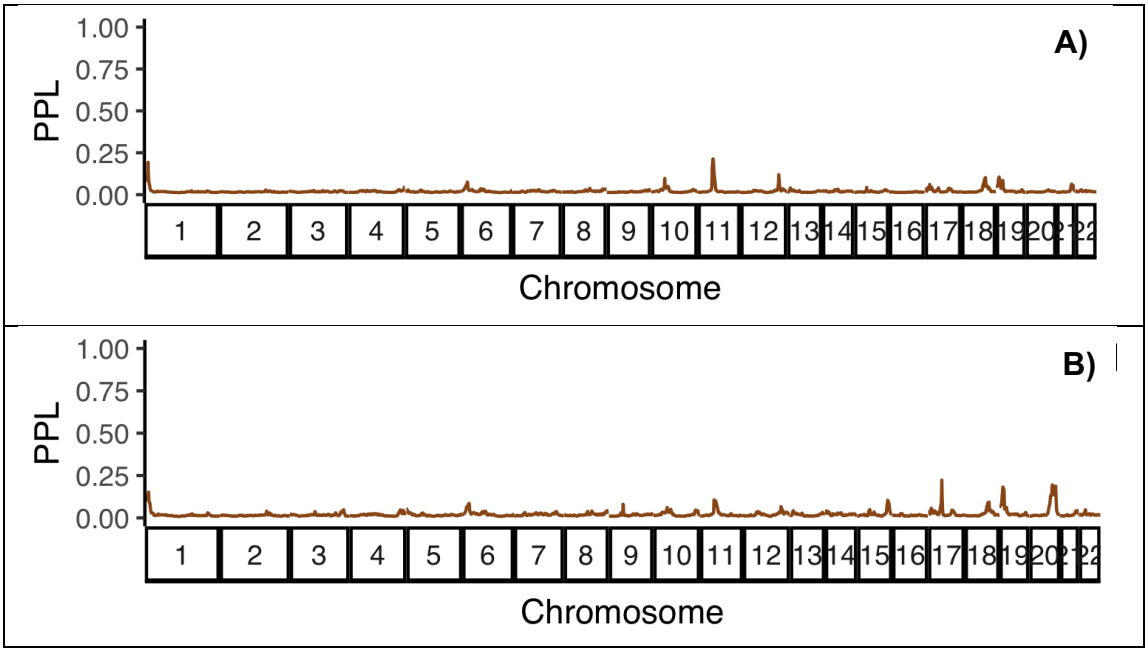


Figure 3.3. Linkage peaks for ASD or ADHD in across the genome.
(A) pooled samples (B) sequential update.

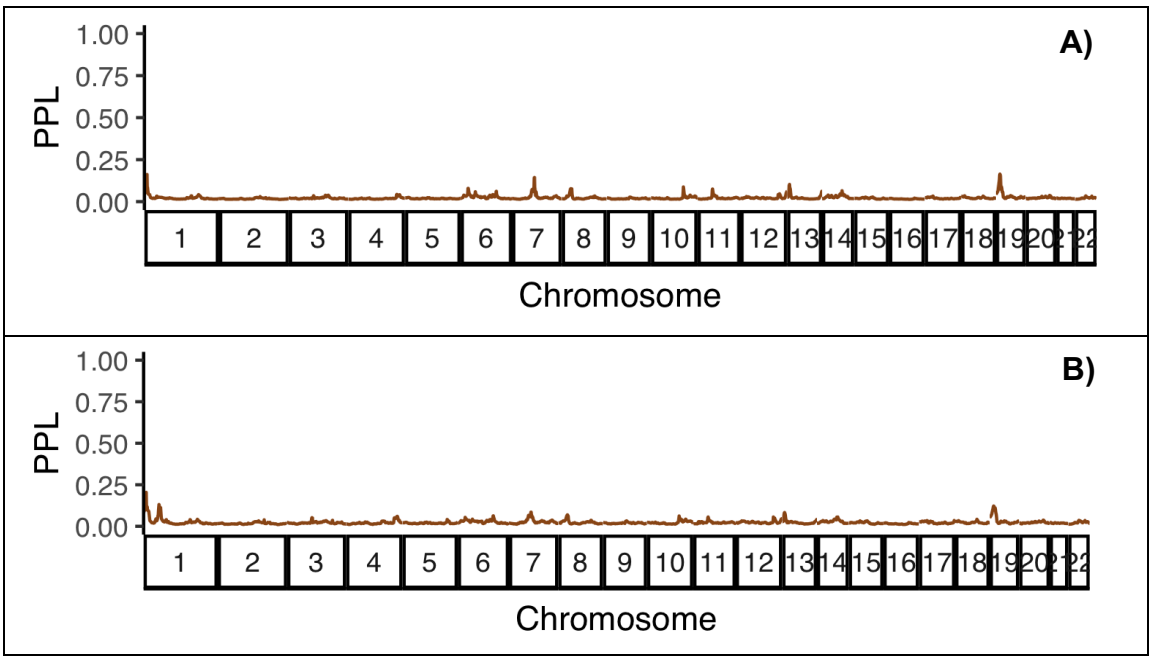


Figure 3.4. Linkage peaks for ASD and ADHD in across the genome.
(A) pooled samples (B) sequential update.

3.3.3. SNP and SV discovery

We selected 272 individuals from the recruited families to undergo whole genome sequencing to ~30x coverage. The sequencing reads that passed quality filter were then mapped to the human genome reference hg19 and genomic variants are called jointly using GATK for SNP and indel. SV calling has been historically difficult (Ho, et al., 2019), therefore we employed an ensemble algorithm, MetaSV, to call SV per-sample.

The number of variants called in small variants and SVs were consistent sample-wise. For small variants, around 4 million SNVs were called for most of the samples, and the indels ranged from 700,000 to 1.1 million. For SVs, an average of 1,103,764 SVs are discovered for each individual, ranging from 400,000 to 1,950,000 across the samples. The majority of the SVs were within 50-1000 bp range (Figure 3.5) and the largest categories were insertions, deletions and inversions (Table 3.8).

Sample	# Reads	Yield (Mbases)	Source
04C32333	328,664,498	99257.00	LCL
04C32335	446,837,505	134946.00	LCL
04C34428	315,196,704	95192.00	LCL
04C35408	356,957,465	107803.00	LCL
04C37033	373,706,310	112861.00	LCL
04C38287	355,184,708	107263.00	LCL
05C40411	297,112,370	89731.00	LCL
05C40413	341,072,614	103006.00	LCL
05C43286	290,261,577	87657.00	LCL
05C43287	324,033,739	97857.00	LCL
05C44020	358,181,538	108167.00	LCL
05C45100	323,833,491	97800.00	LCL
05C45345	337,065,159	101795.00	LCL

05C46007	393,919,714	118961.00	LCL
05C46009	323,188,608	97603.00	LCL
05C46337	374,070,641	112966.00	LCL
05C46426	317,105,568	95763.00	LCL
05C47471	400,689,785	121011.00	LCL
05C47975	316,762,005	95663.00	LCL
05C47977	449,953,232	135883.00	LCL
05C50238	339,997,951	102678.00	LCL
05C50316	422,678,222	127654.00	LCL
06C52243	376,703,496	113767.00	LCL
06C58552	348,772,656	105329.00	LCL
06C58874	351,078,366	106027.00	LCL
06C61307	335,693,680	101376.00	LCL
07C64741	424,597,242	128226.00	LCL
07C65307	316,804,294	95674.00	LCL
07C65371	376,474,743	113693.00	LCL
07C66696	396,828,216	119842.00	LCL
07C67896	374,602,996	113128.00	LCL
07C67897	313,764,671	94756.00	LCL
07C69064	315,723,832	95345.00	LCL
07C69065	332,692,849	100474.00	LCL
07C69182	316,371,330	95546.00	LCL
07C70334	339,863,346	102637.00	LCL
08C71374	360,805,840	108960.00	LCL
08C71765	502,915,302	151879.00	LCL
08C73255	345,239,441	104264.00	LCL
08C75152	532,460,411	160806.00	LCL
08C75153	377,001,593	113853.00	LCL
08C75494	323,735,410	97769.00	LCL
08C76448	385,947,691	116556.00	LCL
08C76449	384,295,223	116058.00	LCL
09C95711	354,926,092	107188.00	WB
10C117871	321,561,392	97114.00	WB
10C117872	521,755,604	157563.00	WB
10C117873	344,230,601	103957.00	WB
10C117874	341,077,777	103006.00	LCL

10C117899	356,111,591	107546.00	WB
2743-SL-0021	337,633,333	101290.00	LCL
2743-SL-0022	344,200,000	103260.00	LCL
2743-SL-0023	345,733,333	103720.00	LCL
2743-SL-0024	334,766,667	100430.00	LCL
2743-SL-0026	331,233,333	99370.00	LCL
2743-SL-0027	344,100,000	103230.00	LCL
2743-SL-0028	331,800,000	99540.00	LCL
2743-SL-0029	334,766,667	100430.00	LCL
2743-SL-0030	336,366,667	100910.00	LCL
2743-SL-0031	333,133,333	99940.00	LCL
2743-SL-0032	333,566,667	100070.00	LCL
2743-SL-0033	352,933,333	105880.00	LCL
2743-SL-0034	353,066,667	105920.00	LCL
2743-SL-0035	354,133,333	106240.00	LCL
2743-SL-0036	382,866,667	114860.00	LCL
2743-SL-0037	337,466,667	101240.00	LCL
2743-SL-0038	314,800,000	94440.00	LCL
2743-SL-0039	339,033,333	101710.00	LCL
2743-SL-0040	334,166,667	100250.00	LCL
LP6005170-DNA_A04	698,500,000	139700.00	WB
LP6005170-DNA_B01	631,000,000	126200.00	WB
LP6005170-DNA_B03	673,000,000	134600.00	WB
LP6005170-DNA_C01	589,000,000	117800.00	WB
LP6005170-DNA_C03	658,000,000	131600.00	WB
LP6005170-DNA_D01	665,000,000	133000.00	WB
LP6005170-DNA_D02	761,500,000	152300.00	WB
LP6005170-DNA_D03	736,000,000	147200.00	WB
LP6005170-DNA_F02	786,000,000	157200.00	WB
LP6005170-DNA_G01	678,500,000	135700.00	WB
LP6005170-DNA_G02	575,500,000	115100.00	WB
LP6005170-DNA_G03	714,500,000	142900.00	WB
LP6005170-DNA_H03	763,500,000	152700.00	WB
MH0151131	368,167,362	111188.00	WB
MH0152513	302,233,936	91274.00	WB
MH0152515	368,718,640	111352.00	LCL

MH0152516	322,782,456	97481.00	LCL
MH0152518	282,544,44	85326.00	LCL
MH0160130	325,919,921	98430.00	LCL
MH0160135	360,204,895	108784.00	LCL
MH0160147	356,700,711	107722.00	WB
MH0160342	455,592,486	137589.00	LCL
MH0161789	328,519,983	99215.00	WB
MH0161791	357,817,154	108062.00	LCL
MH0161794	319,319,390	96440.00	LCL
MH0161798	333,749,002	100791.00	WB
MH0165172	376,946,231	113841.00	WB
MH0165173	283,233,594	85537.00	LCL
MH0165177	423,243,939	127821.00	LCL
MH0166945	357,653,711	108014.00	WB
MH0168145	372,506,472	112500.00	WB
MH0168147	321,254,387	97019.00	WB
MH0168149	452,492,299	136654.00	WB
MH0168151	394,533,316	119149.00	WB
MH0170044	342,434,914	103413.00	WB
MH0170045	395,575,630	119466.00	WB
MH0170047	284,341,243	85871.00	WB
MH0170049	351,954,750	106293.00	WB
MH0181201	384,572,569	116140.00	WB
MH0181202	363,512,308	109780.00	WB
MH0181203	514,652,831	155425.00	WB
MH0181214	331,254,878	100041.00	WB
MH0181699	353,759,815	106831.00	WB
MH0181700	372,710,572	112561.00	WB
MH0181701	354,053,405	106924.00	WB
MH0181702	368,805,960	111380.00	WB
MH0181703	324,356,475	97957.00	WB
MH0181704	378,531,552	114311.00	WB
MH0181705	409,229,861	123586.00	WB
MH0181706	279,393,55	84374.00	WB
MH0186884	365,783,208	110466.00	WB
MH0186896	373,811,036	112891.00	WB

MH0186904	346,790,153	104731.00	WB
MH0186906	356,894,636	107781.00	WB
MH0186908	333,011,717	100566.00	WB
MH0188328	418,949,626	126523.00	WB
MH0196857	363,725,927	109846.00	WB
MH0197641	454,412,999	137231.00	WB
MH0197642	363,067,952	109646.00	WB
MH0197643	331,414,673	100087.00	LCL
MH0197645	360,820,477	108969.00	LCL
MH0197646	352,090,074	106331.00	WB
MH0197647	334,708,730	101082.00	WB
MH0197648	338,809,566	102320.00	WB
SL127696	931,101,286	279330.38	WB
SL127697	938,948,176	281684.45	WB
SL127698	1,004,738,046	301421.41	WB
SL127699	961,542,892	288462.86	LCL
SL127700	971,036,636	291310.99	LCL
SL127701	N/A	N/A	LCL
SL127703	837,494,428	251248.32	WB
SL127704	909,652,332	272895.70	WB
SL127705	946,242,780	283872.83	WB
SL127706	950,235,436	285070.63	LCL
SL127707	872,547,776	261764.33	LCL
SL127708	946,009,476	283802.84	LCL
SL127709	890,052,436	267015.73	WB
SL127710	917,169,078	275150.72	LCL
SL127711	909,540,138	272862.04	WB
SL127712	950,414,932	285124.48	WB
SL127713	923,038,402	276911.52	LCL
SL127714	949,431,850	284829.55	LCL
SL127715	909,505,558	272851.66	LCL
SL127716	929,503,370	278851.01	LCL
SL127717	911,168,678	273350.60	WB
SL127718	907,091,774	272127.53	LCL
SL127719	926,856,194	278056.85	LCL
SL127720	892,910,764	267873.22	LCL

SL127721	945,177,282	283553.18	LCL
SL127722	913,474,670	274042.40	LCL
SL127723	764,364,208	229309.26	WB
SL127724	789,895,372	236968.61	LCL
SL127725	912,223,848	273667.15	WB
SL127726	887,682,762	266304.82	LCL
SL127727	889,293,154	266787.94	LCL
SL127728	1,082,201,226	324660.36	LCL
SL127729	931,522,432	279456.73	LCL
SL127730	843,208,808	252962.64	LCL
SL127731	970,589,090	291176.72	LCL
SL127732	931,285,818	279385.74	LCL
SL127733	949,706,594	284911.97	LCL
SL127734	930,389,832	279116.95	LCL
SL127735	878,048,870	263414.66	LCL
SL127736	N/A	N/A	LCL
SL127737	N/A	N/A	WB
SL127738	N/A	N/A	LCL
SL127739	N/A	N/A	WB
SL127740	N/A	N/A	LCL
SL127741	N/A	N/A	LCL
SL127742	N/A	N/A	LCL
SL127743	N/A	N/A	LCL
SL127744	910,690,302	273207.09	WB
SL127745	898,596,104	269578.83	WB
SL127746	888,964,398	266689.31	LCL
SL127747	909,152,394	272745.71	LCL
SL127748	924,679,232	277403.77	LCL
SL127749	903,399,968	271019.99	LCL
SL128036	923,797,126	277139.13	LCL
SL128038	N/A	N/A	LCL
SL128039	953,413,070	286023.92	WB
SL128040	942,069,524	282620.85	LCL
SL128041	948,562,266	284568.68	LCL
SL128042	945,878,062	283763.41	LCL
SL128043	936,969,380	281090.81	LCL

SL128044	887,455,604	266236.68	LCL
SL128045	907,020,864	272106.25	LCL
SL128046	917,336,710	275201.01	LCL
SL128047	895,291,620	268587.48	LCL
SL128048	887,574,868	266272.46	LCL
SL128049	921,928,650	276578.59	LCL
SL128051	921,818,548	276545.56	LCL
SL128054	906,921,822	272076.54	LCL
SL128055	903,651,286	271095.38	LCL
SL128056	907,201,336	272160.40	LCL
SL128057	922,786,450	276835.93	LCL
SL128058	899,786,148	269935.84	LCL
SL128059	905,001,352	271500.40	LCL
SL128060	914,678,850	274403.65	LCL
SL128061	934,512,058	280353.61	LCL
SL128062	920,129,618	276038.88	LCL
SL128063	919,965,014	275989.50	WB
SL128065	922,118,692	276635.60	LCL
SL128066	922,081,254	276624.37	LCL
SL128067	909,678,974	272903.69	LCL
SL128068	895,407,288	268622.18	LCL
SL128069	912,547,088	273764.12	LCL
SL128070	904,588,182	271376.45	LCL
SL128071	793,877,762	238163.32	LCL
SL128072	882,292,778	264687.83	LCL
SL128073	929,051,706	278715.51	LCL
SL128074	936,200,276	280860.08	LCL
SL128075	739,771,128	221931.33	WB
SL128076	931,373,298	279411.98	LCL
SL128077	911,743,406	273523.02	LCL
SL128078	913,101,256	273930.37	LCL
SL128079	888,097,114	266429.13	LCL
SL128080	779,471,954	233841.58	LCL
SL128081	651,156,586	195346.97	LCL
SL128082	931,139,148	279341.74	LCL
SL128083	856,215,060	256864.51	LCL

SL128084	906,140,794	271842.23	LCL
SL128085	894,279,458	268283.83	LCL
SL128086	898,018,560	269405.56	LCL
SL128087	809,099,442	242729.83	WB
SL128089	945,573,700	283672.11	LCL
SL128090	947,480,050	284244.01	LCL
SL128091	902,498,276	270749.48	LCL
SL128092	948,038,380	284411.51	LCL
SL128093	942,790,568	282837.17	LCL
SL128095	931,483,494	279445.04	LCL
SL128096	945,074,882	283522.46	LCL
SL128097	903,049,170	270914.75	LCL
SL128098	835,220,760	250566.22	LCL
SL128099	858,708,572	257612.57	LCL
SL128100	772,837,876	231851.36	LCL
SL128101	879,154,874	263746.46	LCL
SL128102	914,381,866	274314.56	LCL
SL128103	937,965,982	281389.79	LCL
SL128104	916,630,458	274989.13	LCL
SL128107	858,830,604	257649.18	WB
SL128108	972,216,730	291665.01	LCL
SL128109	952,430,006	285729.00	LCL
SL128110	976,013,430	292804.02	LCL
SL128111	983,507,804	295052.34	LCL
SL128112	943,874,894	283162.46	LCL
SL128113	N/A	N/A	LCL
SL128114	N/A	N/A	LCL
SL128115	975,525,764	292657.72	LCL
SL128116	817,510,130	245253.03	LCL
SL128117	921,564,384	276469.31	LCL
SL128118	922,902,170	276870.65	LCL
SL128119	954,377,186	286313.15	WB
SL128120	935,200,224	280560.06	LCL
SL128121	891,957,594	267587.27	LCL
SL128122	928,833,192	278649.95	LCL
SL128124	N/A	N/A	LCL

SL128125	N/A	N/A	LCL
SL128126	N/A	N/A	LCL
SL128127	N/A	N/A	LCL
SL128128	950,430,242	285129.073	LCL
SL128129	991,130,180	297339.054	LCL
SL128130	988,269,882	296480.965	WB
SL128131	976,100,982	292830.295	LCL

Table 3.6. Sequencing report for each sample.

Sample	SNV Count	Indel Count	Other Count
04C32333	4054928	890268	53176
04C32335	4107110	934570	57210
04C34428	3987367	866460	51133
04C35408	4041561	895996	53838
04C37033	4093388	919889	54970
04C38287	4038243	898708	53091
05C40411	4060040	901796	55547
05C40413	3989129	895022	53811
05C43286	4048717	877268	52472
05C43287	4018590	879923	51242
05C44020	4066768	886541	53341
05C45100	4081513	887723	53445
05C45345	4052362	890762	52666
05C46007	4058680	905995	54694
05C46009	4027111	884519	52533
05C46337	4060592	896727	54817
05C46426	4045509	880031	51255
05C47471	4250145	903566	53749
05C47975	4242628	915713	52952
05C47977	3966461	955099	57529
05C50238	4016053	873948	52523
05C50316	4170299	901381	54825
06C52243	4081633	918746	55158
06C58552	4042626	902327	53456

06C58874	4005434	886110	52714
06C61307	4129355	873509	52031
07C64741	4066562	930575	56774
07C65307	4013187	879154	51201
07C65371	4018657	899524	54265
07C66696	4100972	895589	53195
07C67896	4070716	905978	54278
07C67897	4043433	882300	52216
07C69064	4040566	878945	51732
07C69065	4065896	877239	50802
07C69182	4140972	876608	50680
07C70334	4072680	896976	52845
08C71374	4109790	903204	54379
08C71765	4093070	949985	58823
08C73255	4096021	896524	52750
08C75152	4063296	949751	58769
08C75153	4142515	904066	54524
08C75494	4031252	898315	52181
08C76448	4021549	894801	53548
08C76449	4030520	895963	54053
09C95711	3977658	891656	54447
10C117871	3980624	880607	51073
10C117872	4082153	963958	58752
10C117873	4070008	895017	53466
10C117874	4054848	882667	52323
10C117899	4035222	899225	52852
2743-SL-0021	4125585	742699	47196
2743-SL-0022	4162215	763956	49810
2743-SL-0023	4208489	744148	48048
2743-SL-0024	4118828	738789	47594
2743-SL-0026	4164432	739160	48815
2743-SL-0027	4134490	739959	48030
2743-SL-0028	4156355	734208	47213
2743-SL-0029	4180991	786168	50578
2743-SL-0030	4163782	748881	48718
2743-SL-0031	4158619	738618	47635

2743-SL-0032	4177400	747311	48247
2743-SL-0033	4145608	759753	50334
2743-SL-0034	4173532	753620	48555
2743-SL-0035	4352587	755293	49063
2743-SL-0036	4150038	749612	48625
2743-SL-0037	4184604	748154	48866
2743-SL-0038	4118849	747024	49041
2743-SL-0039	4140904	732761	47127
2743-SL-0040	4094646	742193	48033
LP6005170-DNA_A04	4169478	731361	48541
LP6005170-DNA_B01	4137256	701289	46778
LP6005170-DNA_B03	4135729	713635	46983
LP6005170-DNA_C01	4187582	811343	49920
LP6005170-DNA_C03	4124678	745579	48729
LP6005170-DNA_D01	4153478	697956	46442
LP6005170-DNA_D02	4114500	729266	48958
LP6005170-DNA_D03	3786254	760397	49902
LP6005170-DNA_F02	4126507	798971	53255
LP6005170-DNA_G01	4114176	911492	55462
LP6005170-DNA_G02	4140030	738068	46897
LP6005170-DNA_G03	4199118	718314	48323
LP6005170-DNA_H03	4109264	816020	55225
MH0151131	4106897	901343	54425
MH0152513	4139379	904429	53802
MH0152515	4142707	918514	54896
MH0152516	4136444	910955	53111
MH0152518	4247495	914764	54927
MH0160130	4111428	893054	52871
MH0160135	4106793	893880	53555
MH0160147	4145909	894554	54514
MH0160342	4160256	936849	57342
MH0161789	4300817	872422	52496
MH0161791	4101113	893180	53431
MH0161794	4107197	863535	50658
MH0161798	4134154	864961	51033
MH0165172	4123693	896477	53624

MH0165173	4163623	886933	53357
MH0165177	4090208	923642	55844
MH0166945	4121492	886299	52995
MH0168145	4094339	886493	51239
MH0168147	4160405	859925	49298
MH0168149	4155450	903926	52877
MH0168151	4415214	910383	54924
MH0170044	4113521	933848	51554
MH0170045	4228066	1017588	57364
MH0170047	4162736	914220	50556
MH0170049	4125358	902151	53395
MH0181201	4254271	889656	52324
MH0181202	4109519	890344	51989
MH0181203	4112943	934848	56587
MH0181214	4175253	876121	50604
MH0181699	4180012	882062	52276
MH0181700	4145491	884611	50275
MH0181701	4113273	881445	51512
MH0181702	4131210	898910	53197
MH0181703	4126704	873092	51291
MH0181704	4051474	886700	51862
MH0181705	4142900	900191	52622
MH0181706	4174273	880277	51684
MH0186884	4153184	896389	52994
MH0186896	4120779	883783	52605
MH0186904	4111921	883233	51746
MH0186906	4170333	887031	52376
MH0186908	4132595	874767	51697
MH0188328	4109172	908290	54145
MH0196857	4053468	867430	50203
MH0197641	4052711	956338	55525
MH0197642	4130594	922090	55302
MH0197643	4008009	915925	54830
MH0197645	4187023	928835	55512
MH0197646	4035752	922685	56459
MH0197647	4163585	903779	54505

MH0197648	4045330	908816	54697
SL127696	4016324	1106476	60035
SL127697	4060717	1115926	60167
SL127698	4127130	1129446	61314
SL127699	4090799	1116864	60233
SL127700	4081225	1121471	60870
SL127701	4008247	1092978	60188
SL127703	4044301	1079642	59841
SL127704	4100766	1123370	61498
SL127705	4079368	1120542	61045
SL127706	4116794	1120973	61229
SL127707	4152466	1109406	61471
SL127708	4129328	1124980	61642
SL127709	4228099	1093556	60356
SL127710	4114873	1109479	60573
SL127711	4172233	1140752	62356
SL127712	4387263	1115682	60643
SL127713	4319179	1125974	60684
SL127714	4112978	1117431	60911
SL127715	4050328	1127113	62366
SL127716	4095273	1111177	60874
SL127717	4142492	1076463	60429
SL127718	4262062	1106248	61397
SL127719	4142377	1093964	60799
SL127720	4149302	1101387	60038
SL127721	3820043	1123907	60699
SL127722	4141131	1102111	60260
SL127723	4229083	1093572	60403
SL127724	4230588	1063401	59656
SL127725	4181892	978920	52260
SL127726	4136415	1093485	59399
SL127727	4368007	1087342	60033
SL127728	4156492	1077499	60057
SL127729	4129504	1133261	61887
SL127730	4112064	1101595	59698
SL127731	4175182	1112539	61126

SL127732	4104082	1117117	61081
SL127733	4115006	1111948	61838
SL127734	4123327	1116449	61558
SL127735	4073743	1080982	60094
SL127736	4196935	1140913	61649
SL127737	4170542	1111765	60125
SL127738	4011137	1086011	59717
SL127739	4101511	1104891	60707
SL127740	4272188	1109382	61014
SL127741	4060702	1123828	61536
SL127742	4145224	1082937	59653
SL127743	4203434	1077830	60262
SL127744	4277476	1077961	59366
SL127745	4176506	1077458	60059
SL127746	4163623	1068927	59099
SL127747	4090208	1072995	59086
SL127748	4121492	1076993	59191
SL127749	4094339	1081047	59801
SL128036	4160405	1114557	60789
SL128038	4155450	1114540	61101
SL128039	4415214	1159906	62630
SL128040	4113521	1094357	60606
SL128041	4228066	1109366	60403
SL128042	4162736	1097524	60731
SL128043	4125358	1093376	60694
SL128044	4254271	1113282	61309
SL128045	4109519	1092197	60726
SL128046	4162096	1099831	60799
SL128047	4112943	1090878	60153
SL128048	4175253	1106436	60647
SL128049	4180012	1053462	61885
SL128051	4145491	1093198	60361
SL128054	4113273	1042625	57334
SL128055	4131210	1073156	59295
SL128056	4126704	1083196	60144
SL128057	4051474	1041092	58641

SL128058	4142900	1079179	60350
SL128059	4174273	1089099	59809
SL128060	4153184	1072892	59039
SL128061	4120779	1065794	59328
SL128062	4111921	1019438	57317
SL128063	4290292	1089531	60311
SL128065	4170333	1055752	57592
SL128066	4132595	1052553	58622
SL128067	4109172	1040262	57942
SL128068	4053468	954470	52731
SL128069	4052711	950517	53423
SL128070	4130594	966705	53887
SL128071	4008009	919636	50291
SL128072	4187023	981174	54497
SL128073	4035752	955552	53595
SL128074	4163585	978832	53565
SL128075	4045330	972688	54035
SL128076	4016324	930630	49915
SL128077	3973358	894949	47614
SL128078	4060717	943127	51788
SL128079	4127130	992093	53351
SL128080	4090799	994091	53933
SL128081	4081225	959383	51567
SL128082	4008247	959957	52282
SL128083	4044301	950615	52074
SL128084	4100766	1065703	59085
SL128085	4079368	1064009	59396
SL128086	4116794	1046595	59494
SL128087	4152466	1087522	60336
SL128089	4129328	1068966	59486
SL128090	4228099	1104113	60441
SL128091	4114873	1075555	59036
SL128092	4172233	1029202	56646
SL128093	4387263	1070212	59267
SL128095	4319179	1079879	59550
SL128096	4112978	1027551	57453

SL128097	4050328	1007775	56642
SL128098	4095273	1045759	57654
SL128099	4278840	1041288	57468
SL128100	4142492	1124862	61436
SL128101	4262062	1145641	62063
SL128102	4142377	1126129	60826
SL128103	4149302	1124647	62466
SL128104	3820043	996501	53281
SL128107	4141131	1103025	61541
SL128108	4229083	1156191	62491
SL128109	4230588	1133788	61984
SL128110	4181892	1131858	61187
SL128111	4136415	1121239	61307
SL128112	4368007	1163165	62326
SL128113	4156492	1116918	61623
SL128114	4129504	1069096	58634
SL128115	4112064	1114051	60700
SL128116	4175182	1096075	61238
SL128117	4104082	1101910	60405
SL128118	4115006	1098001	60318
SL128119	4123327	1094298	60861
SL128120	4073743	1088383	59422
SL128121	4196935	1125944	61408
SL128122	4170542	1142557	61742
SL128124	4011137	966533	53525
SL128125	4101511	1078262	59318
SL128126	4272188	1112506	60072
SL128127	4060702	949756	52559
SL128128	4145224	1124065	60788
SL128129	4203434	1133413	61422
SL128130	4277476	1140558	62653
SL128131	4176506	1122667	62191

Table 3.7. Summary of variants called from the NJLAGS WGS data.

Sample	DEL	INS	DUP	INV	ITX	Total
04C32333	308555	617785	15917	280833	153586	1376676
04C32335	345889	821653	17435	299049	179143	1663169
04C34428	325345	647911	17748	327581	85473	1404058
04C35408	330154	643617	16038	238991	155347	1384147
04C37033	341598	738047	16423	242348	175353	1513769
04C38287	335467	722986	17670	297756	131389	1505268
05C40411	311286	633918	14780	223295	176615	1359894
05C40413	300165	646233	17008	278306	118062	1359774
05C43286	293989	627395	15036	271998	149482	1357900
05C43287	293114	617084	14846	229597	137254	1291895
05C44020	320609	699256	17641	307397	89051	1433954
05C45100	300636	617715	14202	212059	153391	1298003
05C45345	294560	660471	15756	263532	140593	1374912
05C46007	343301	759241	16252	277727	152774	1549295
05C46009	289819	659089	16454	253580	153297	1372239
05C46337	343519	725035	18302	337192	122348	1546396
05C46426	286091	670885	19593	380630	56986	1414185
05C47471	349106	774198	18578	327324	119276	1588482
05C47975	297732	629206	16123	274233	162771	1380065
05C47977	377234	866056	17968	306048	168069	1735375
05C50238	327728	666620	15924	282470	151228	1443970
05C50316	365285	835468	18643	323690	126039	1669125
06C52243	344483	762260	16717	250062	181671	1555193
06C55728	332323	684320	17246	296558	104710	1435157
06C58552	307070	650494	15859	257066	160006	1390495
06C58874	346981	703337	16258	219057	156398	1442031
06C61307	302989	647770	14922	265607	101634	1332922
07C64741	323970	768800	16588	280155	157203	1546716
07C65307	321972	647444	17769	324694	100438	1412317
07C65371	175792	318603	7076	121084	68094	690649
07C66696	346204	765710	19311	305435	135704	1572364
07C67896	317382	724968	17346	291026	145250	1495972
07C67897	322707	637746	17009	292699	106965	1377126
07C69064	291249	617688	15941	298236	129215	1352329
07C69065	292224	683788	18276	325192	117063	1436543

07C69182	302457	642095	15157	326270	77107	1363086
07C70334	306187	687496	14386	276985	144983	1430037
08C71374	321925	654469	14763	204591	132484	1328232
08C71765	341023	843259	16242	297523	165479	1663526
08C73255	290493	659276	16480	270895	93837	1330981
08C75152	375017	917272	18757	269890	104422	1685358
08C75153	335237	738574	16551	271362	121821	1483545
08C75494	302482	641254	14656	277440	93730	1329562
08C76448	325537	715994	16559	284300	122649	1465039
08C76449	326108	761138	17083	290732	118098	1513159
09C95711	331763	621911	14439	205616	109720	1283449
2743-SL-0021	314614	185868	21298	101669	4879	628328
2743-SL-0022	316580	208171	23184	107127	5625	660687
2743-SL-0023	287786	183864	25799	94750	5095	597294
2743-SL-0024	312764	216248	38065	121296	5247	693620
2743-SL-0025	311111	185720	25182	100814	4468	627295
2743-SL-0026	299355	188321	20475	123781	4516	636448
2743-SL-0027	281556	173692	21000	107415	4800	588463
2743-SL-0028	297275	172636	18792	88421	5216	582340
2743-SL-0029	322736	242513	29446	109497	5819	710011
2743-SL-0030	268199	183932	22703	100409	5076	580319
2743-SL-0031	297943	184674	21853	93741	5239	603450
2743-SL-0032	323080	197574	30484	92004	5038	648180
2743-SL-0033	312021	216001	21105	99594	5476	654197
2743-SL-0034	315585	198476	23607	112941	4946	655555
2743-SL-0035	315282	199380	26263	103562	5039	649526
2743-SL-0036	267265	189159	23482	96424	5094	581424
2743-SL-0037	287304	185268	25216	92056	5406	595250
2743-SL-0038	284185	178648	21347	105309	5077	594566
2743-SL-0039	286110	179424	20530	108139	5312	599515
2743-SL-0040	291028	179930	26126	97502	5303	599889
LP6005170-DNA_A02	166472	84759	48513	113171	4337	417252
LP6005170-DNA_A03	231964	100241	27117	149702	6916	515940
LP6005170-DNA_A04	287258	152489	27570	156638	7096	631051
LP6005170-DNA_B01	334899	188015	62334	189930	11368	786546
LP6005170-DNA_B02	264430	132033	26867	124295	8906	556531

LP6005170-DNA_B03	255062	121559	28140	147733	7427	559921
LP6005170-DNA_C01	297321	158659	47851	138685	9002	651518
LP6005170-DNA_C02	198141	133494	52346	117669	15223	516873
LP6005170-DNA_C03	240662	93861	23496	128430	5004	491453
LP6005170-DNA_D01	314857	187846	50283	166437	10743	730166
LP6005170-DNA_D02	277982	179144	42834	133100	12116	645176
LP6005170-DNA_E01	230659	104328	33940	148968	4176	522071
LP6005170-DNA_G02	211302	123730	46287	151493	8777	541589
LP6005170-DNA_H03	317342	189005	39117	196628	8439	750531
LP6005417-DNA_A01	377685	377667	100542	207820	126039	1189753
LP6005417-DNA_B01	284923	184077	49128	143265	19782	681175
MH0151131	326842	684558	17171	254865	164994	1448430
MH0152513	322536	661398	16448	296500	121062	1417944
MH0152515	321623	693250	15325	243272	123180	1396650
MH0152516	301194	591276	14529	234674	138599	1280272
MH0152518	305488	604474	14836	240933	134421	1300152
MH0160130	278116	638026	17562	325811	108458	1367973
MH0160135	339395	677788	16612	245703	114941	1394439
MH0160147	315859	639884	13928	246456	129548	1345675
MH0160342	385454	785202	14998	267383	116462	1569499
MH0161789	329342	636177	15963	278547	108690	1368719
MH0161791	316312	712522	19326	352055	80268	1480483
MH0161794	313507	639048	17855	277394	123374	1371178
MH0161798	326116	656661	16434	281713	80355	1361279
MH0165172	324812	720603	17979	297929	166602	1527925
MH0165173	320470	669085	16363	309389	152122	1467429
MH0165177	354800	835542	20210	293793	145473	1649818
MH0166945	336686	717900	18554	290392	130906	1494438
MH0168145	182119	346787	8348	116448	86021	739723
MH0168147	298703	707103	19696	412226	68353	1506081
MH0168149	361903	917545	20381	400857	72926	1773612
MH0168151	340710	763770	16175	273505	179694	1573854
MH0170044	313594	691934	16771	275323	148182	1445804
MH0170045	352711	790247	20781	311554	108109	1583402
MH0170047	295572	622048	16035	283276	160095	1377026
MH0170049	319785	666290	16253	261831	160166	1424325

MH0181201	333713	748236	16992	261516	178987	1539444
MH0181202	309908	709680	16688	285027	180586	1501889
MH0181203	375296	950257	19082	296625	155759	1797019
MH0181214	282653	638178	16062	243989	145924	1326806
MH0181699	319909	701245	16948	288357	138953	1465412
MH0181700	328850	771379	19845	375711	113949	1609734
MH0181701	317894	677262	15372	277637	164004	1452169
MH0181702	355099	744403	17828	257865	134333	1509528
MH0181703	314448	635317	14698	271900	140457	1376820
MH0181704	301096	758960	20083	320547	129959	1530645
MH0181705	383913	870258	18148	301596	135425	1709340
MH0181706	298035	646683	15652	274747	151515	1386632
MH0186884	328181	703746	17022	350834	135438	1535221
MH0186896	329450	759178	20169	314128	130327	1553252
MH0186904	305682	681859	15350	289467	193136	1485494
MH0186906	332209	719320	18374	324231	129647	1523781
MH0186908	302069	685921	17599	294406	126888	1426883
MH0188328	359375	834407	19551	330073	117688	1661094
MH0196857	304490	715683	17246	308762	153413	1499594
MH0197641	382112	915460	21992	348593	127743	1795900
MH0197642	317814	756306	18513	314533	133608	1540774
MH0197643	293910	640555	15190	249658	125130	1324443
MH0197645	329914	748128	20243	302044	129528	1529857
MH0197646	313225	613346	14939	215072	141562	1298144
MH0197647	318475	759162	21834	323625	88542	1511638
MH0197648	305226	682440	19612	325322	102404	1435004
SL127696	319499	321053	13562	110315	18412	782841
SL127697	299809	316797	13669	95483	15819	741577
SL127698	299539	302869	13158	87805	17215	720586
SL127699	324730	316724	16166	97957	19623	775200
SL127700	311894	312843	13041	136401	19558	793737
SL127701	388808	589751	15906	125792	17954	1138211
SL127702	355018	577500	15377	94448	14261	1056604
SL127703	378524	777321	15456	143683	11080	1326064
SL127704	306649	267845	13573	97674	23475	709216
SL127705	331233	317635	14264	97358	22853	783343

SL127706	321481	312885	15248	109525	30514	789653
SL127707	362685	352807	18138	111242	27873	872745
SL127708	333852	382719	16418	139272	28211	900472
SL127709	291844	298957	12853	104661	24747	733062
SL127710	307638	328105	15569	116444	21956	789712
SL127711	280422	301158	12220	117696	18651	730147
SL127712	273275	261427	14626	95531	14825	659684
SL127713	303945	309360	14896	95191	15704	739096
SL127714	329041	300275	15597	117301	20499	782713
SL127715	334102	302777	15345	98349	22105	772678
SL127716	299610	350060	17951	90252	17155	775028
SL127717	341540	340737	14130	94255	20531	811193
SL127718	276668	362706	15263	109730	16945	781312
SL127719	275193	320825	14489	95417	24399	730323
SL127720	294174	363811	15061	118362	15784	807192
SL127721	327643	376663	15803	111732	15354	847195
SL127722	327652	372182	13905	117594	18953	850286
SL127723	305801	288991	14100	96986	17333	723211
SL127724	317545	366991	12913	116116	17380	830945
SL127725	295011	396216	13429	129041	14531	848228
SL127726	303389	383102	14087	109627	15015	825220
SL127727	292306	355963	13216	124305	20946	806736
SL127728	275633	254548	16368	103820	14599	664968
SL127729	330613	266845	16037	93134	25502	732131
SL127730	422547	291823	18614	126187	32793	891964
SL127731	334735	296984	17546	125929	30312	805506
SL127732	313027	261045	16952	108512	18001	717537
SL127733	269413	276504	12946	97381	18582	674826
SL127734	256438	280124	15355	102583	18012	672512
SL127735	247479	253396	13016	81450	24403	619744
SL127736	360956	438504	18138	116812	21848	956258
SL127737	327724	477654	15433	128781	14064	963656
SL127738	328590	462438	14248	118852	21224	945352
SL127739	353376	450388	15390	114406	15493	949053
SL127740	314847	571882	15731	130089	13416	1045965
SL127741	357511	485683	14794	130478	12238	1000704

SL127742	331423	567482	15504	106597	13194	1034200
SL127743	321185	650262	15907	131023	11045	1129422
SL127744	347820	468145	14888	133535	14269	978657
SL127745	331502	449183	13003	112950	13006	919644
SL127746	332522	507350	13299	132061	13627	998859
SL127747	357635	512235	13171	126794	15181	1025016
SL127748	359290	479833	16129	112449	17950	985651
SL127749	332302	523024	16352	117839	15653	1005170
SL128036	295509	367117	13621	87005	22321	785573
SL128037	366640	574184	16240	118213	24488	1099765
SL128038	301573	414468	14467	108894	15348	854750
SL128039	310588	552911	15200	108587	12497	999783
SL128040	339245	656820	16177	112398	16073	1140713
SL128041	341082	756785	14321	114566	14519	1241273
SL128042	365663	931549	18189	129346	17643	1462390
SL128043	334698	716502	16893	98625	26521	1193239
SL128044	314832	356955	14342	117922	14405	818456
SL128045	315066	363845	15476	99203	17402	810992
SL128046	316183	373736	14989	109724	18472	833104
SL128047	323454	412353	13012	125055	15308	889182
SL128048	292128	366585	16160	110108	13613	798594
SL128049	310003	403676	14119	117344	15857	860999
SL128050	296926	369824	12407	94296	13830	787283
SL128051	323526	383163	12990	120414	13096	853189
SL128052	291522	283734	12166	84655	14971	687048
SL128053	307490	299510	14942	101780	15321	739043
SL128054	288873	299397	13930	110722	13257	726179
SL128055	277129	280972	13753	90768	16691	679313
SL128056	311145	311131	13819	123026	15585	774706
SL128057	290008	290590	11757	113245	16903	722503
SL128058	320062	325291	12791	115150	18240	791534
SL128059	284886	304273	13269	101269	17328	721025
SL128060	348847	441146	14221	115886	18661	938761
SL128061	337034	387405	12945	110360	16080	863824
SL128062	338077	403435	13985	97123	18593	871213
SL128066	360304	478477	13296	110669	22591	985337

SL128067	298731	380608	12995	92604	16206	801144
SL128068	755658	446527	15452	116426	22686	1356749
SL128069	687816	491600	17923	132762	22058	1352159
SL128070	602718	514381	15112	151866	24062	1308139
SL128071	500847	513752	13091	118665	17400	1163755
SL128072	522486	601535	16964	138770	16430	1296185
SL128073	631096	765717	16655	148433	20305	1582206
SL128074	657009	854856	20232	144916	21326	1698339
SL128075	472442	630675	16740	138915	22482	1281254
SL128076	689232	403226	14379	134175	23970	1264982
SL128077	673890	327960	12245	126093	11494	1151682
SL128078	1115148	497742	14902	138140	22130	1788062
SL128079	965901	613031	15199	139724	22356	1756211
SL128080	886243	587308	17431	148674	21905	1661561
SL128081	586988	561585	15202	189142	17825	1370742
SL128082	583676	538785	15116	145612	19477	1302666
SL128083	775225	468724	16265	140547	18423	1419184
SL128084	322591	402666	13647	125500	19117	883521
SL128085	381716	490232	15747	121604	21260	1030559
SL128086	380485	540497	14117	135826	20553	1091478
SL128087	328594	457291	13227	110869	15924	925905
SL128088	352845	488843	14927	126386	16305	999306
SL128089	350122	445019	14609	134870	20197	964817
SL128090	308172	447980	15236	99733	21291	892412
SL128091	347401	450522	14784	124679	20015	957401
SL128092	380678	597542	15729	152849	18124	1164922
SL128093	388654	675187	18438	143907	22354	1248540
SL128094	432967	812219	17324	124621	24466	1411597
SL128095	375635	780324	17090	129377	19944	1322370
SL128096	362596	671441	13297	115270	16418	1179022
SL128097	380091	802530	13601	121272	16227	1333721
SL128098	351105	547819	13270	125411	20272	1057877
SL128099	376376	528460	16917	143577	28946	1094276
SL128100	299227	286284	16396	95546	21107	718560
SL128101	352286	329810	18430	136236	24340	861102
SL128102	300380	326821	15121	100525	20723	763570

SL128103	304315	356105	16378	97064	17224	791086
SL128104	313613	307885	14630	106636	20147	762911
SL128105	333850	349016	15193	129015	20112	847186
SL128106	308832	324333	13511	111671	15229	773576
SL128107	304781	321008	14975	97437	15619	753820
SL128108	414487	700498	15606	104362	40055	1275008
SL128109	332117	392754	14770	91230	20405	851276
SL128110	333880	370809	13196	106520	18632	843037
SL128111	359045	400363	18056	97519	22718	897701
SL128112	322617	368511	16274	119762	19682	846846
SL128113	277078	368717	13998	103267	14963	778023
SL128114	370373	543617	13360	119095	12456	1058901
SL128115	332108	404811	13904	118695	17694	887212
SL128116	375815	510995	15995	119492	14508	1036805
SL128117	399967	641587	17033	138038	15564	1212189
SL128118	363888	646047	17216	127958	19254	1174363
SL128119	445312	684677	15792	139337	23897	1309015
SL128120	536792	1113158	20958	149005	37393	1857306
SL128121	514419	999057	17802	139102	37361	1707741
SL128122	521685	1239824	19511	132003	37692	1950715
SL128123	443368	1131994	19413	143988	40151	1778914
SL128124	348791	521724	15488	122844	14722	1023569
SL128125	309110	546596	14099	112004	29233	1011042
SL128126	346305	554201	16261	136735	23535	1077037
SL128127	352007	547833	13838	117182	24430	1055290
SL128128	330305	390979	16684	105326	17857	861151
SL128129	354799	427012	17557	107064	22284	928716
SL128130	330170	399517	16532	108032	16058	870309
SL128131	305958	440516	15974	116452	12573	891473

Table 3.8. Summary of SVs called from the NJLAGS WGS data.

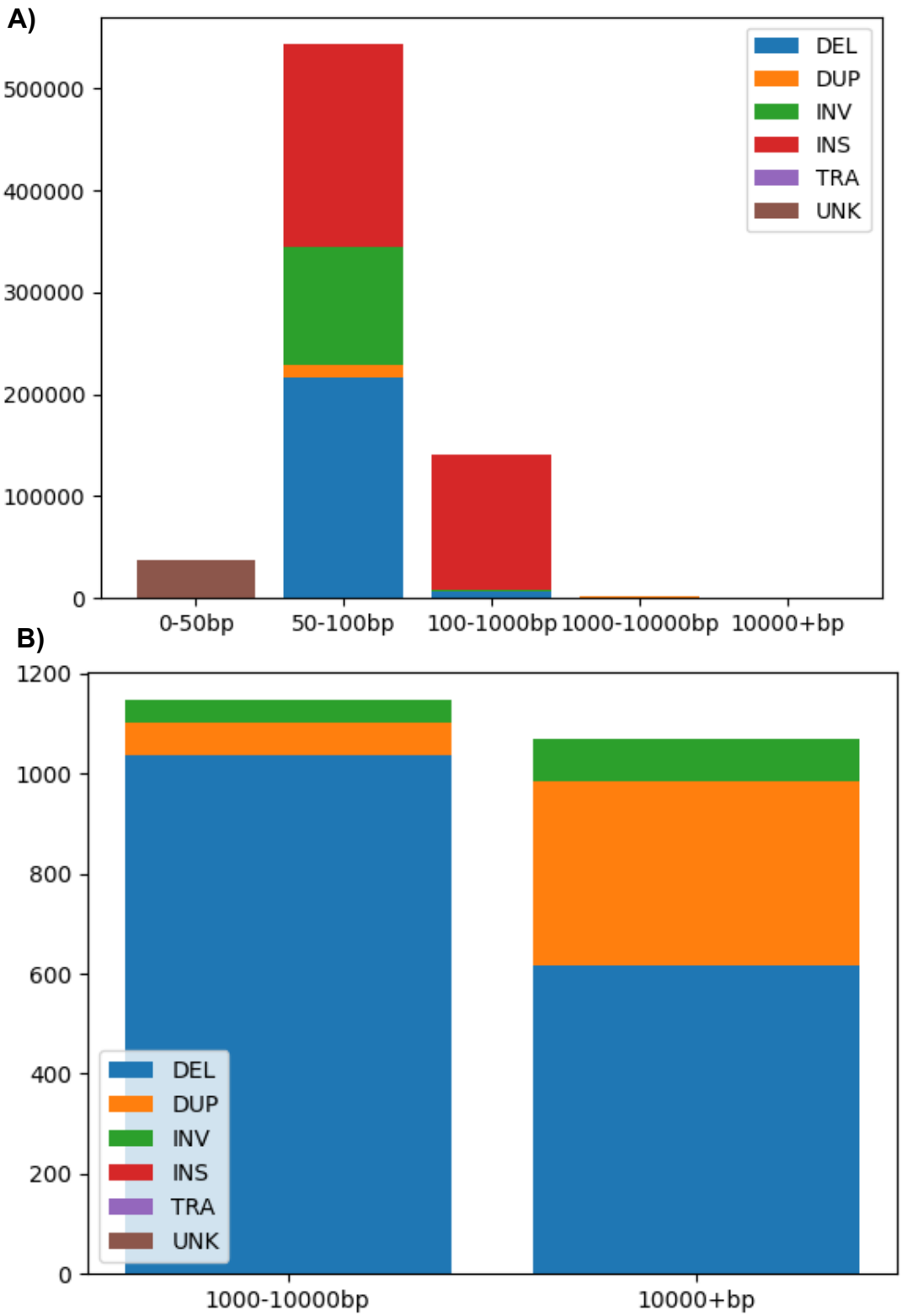


Figure 3.5. SV size and type distribution.
A) all data. B) zoomed scale of SVs of 1000-10000bp and 10000+bp in length.

3.3.4. Candidate genes associated with ADHD in linkage regions

For small variant analysis, variants annotated to have an MAF >5% in general population were excluded. To identify the risk genes for ASD and ADHD, we utilize linkage study based on the pedigree structure, and association study based on case/control group. The two analyses are integrated in a single probabilistic model named pVAAST (Hu, et al., 2014), which prioritize genes for given phenotypes. We first performed pVAAST analysis on the chromosome 12 and 17 significant linkage regions. After Bonferroni correction of the p-values, we discovered 43 genes that fall in the significance criteria (p-value<0.05) for the dominance mode and 6 genes for the recessive mode. Among them, 27 genes are located on chromosome 12, and 22 are located on chromosome 17 (Table 3.9). Among the 49-candidate gene set, 3 of the genes were already collected in one or more disease gene databases (*KDM6B*, *PTPRB*, *PER1*).

The highest-ranking gene is Lysine Demethylase 6B (*KDM6B*) in the chromosome 17 linkage region. In NJLAGS dataset, *KDM6B* has two mutations segregating in 2 different families. One of the mutations (17-7751888-C-G) is a missense mutation (p.Thr761Ser) and predicted to be possibly damaging by PolyPhen2. The other mutation (17-7749972-G-T) is also a missense mutation (p.Val209Leu). *KDM6B* was found to express in brain based on the GTEx project and intolerance to mutations (pLI = 1). The *KDM6B* protein demethylates trimethylated lysine-27 on histone H3, and pathogenic alterations in histone lysine methylation and demethylation genes have been associated with multiple neuro-developmental disorders in previous studies (Stolerman, et al., 2019). *KDM6B* is also annotated in neuro-developmental disorder databases such as

SFARI and iPSYCH, and is identified as one of the 102 risk genes in a large-scale exome ASD study (35,584 total samples with 11,986 samples affected by ASD) (Satterstrom, et al., 2020).

Another candidate gene from the chromosome 17 linkage region, Neuralized E3 Ubiquitin Protein Ligase 4 (*NEURL4*), has a missense mutation (17-7224433-C-T, p.Gly1120Arg) segregating in one of the families. *NEURL4* is also expressed in brain and intolerant to mutations (pLI ~= 1). The *NEURL4* protein is a scaffold protein, which maintains normal centriolar homeostasis and preventing formation of ectopic microtubular organizing centers (Li, et al., 2012). Although no publications directly links *NEURL4* with neurodevelopmental disorders, evidences have shown that microtubule related genes are related to neuronal migration and dendritic functioning (Chang, et al., 2018). Knockout of *NEURL4* in mice is also found to cause neurological phenotypes such as decreased prepulse inhibition (Dickinson, et al., 2016).

In chromosome 12, we found the circadian rhythm controller gene Timeless Circadian Regulator (*TIMELESS*) with a missense mutation (12-56827209-C-A, p.Ala129Ser) segregating in 2 families and predicted deleterious from SIFT (Ng and Henikoff, 2003), highlighting its potential association with ADHD. Sleep disorders are common in both ASD and ADHD patients (Ming and Walters, 2009). In a previous study, a screening of 28 ASD patients (14 with sleeping disorders) with 23 controls has found mutations in circadian-relevant genes including *TIMELESS* are more frequent in patients with ASD than in controls (Yang, et al., 2016).

We also found gene Keratin 75 (*KRT75*) had a SIFT-predicted deleterious missense mutation (12-52818419-C-T, p.Gly513Asp) segregating in one of the

families under the recessive model in chromosome 12. *KRT75* is a gene associated with cytostructural function, and was found to have variant in all affected individuals with bipolar disorder of an Amish family study and the gene was affected by one of 10 potentially pathogenic alleles that was tested in a larger Amish cohort (Strauss, et al., 2014). *KRT75* is also predicted by HumanBase disease machine learning algorithm based on tissue specific gene networks to be the most associated with autism spectrum disorder (confidence score 0.47) (Greene, et al., 2015).

In addition to SNVs and indels, we also identified SVs using the WGS data. Within the linkage regions, an average of 9,747 SVs were identified per sample, with a minimum of 2,469 and a maximum of 20,539. Gene prioritization showed 13 genes are affected more severely in case versus control among the families (Table. 3.10). The genes most enriched with pathogenic SVs in affected samples versus non-affected samples includes genes in the Akt signaling pathway (*CDK2*, *NR4A1*) suggested by KEGG pathway analysis.

Rank	Gene	Mode	Chr	PValue	Score	LOD Score	LOD PValue
1	KDM6B	Dominant	chr17	8.95E-04	103.44	0.31	2.00E-01
2	MYH13	Dominant	chr17	8.95E-04	94.93	0.23	2.00E-01
3	USP6	Dominant	chr17	8.95E-04	94.1	0.19	3.00E-01
4	MYBBP1A	Dominant	chr17	8.95E-04	92.41	0.78	1.00E-01
5	MLL2	Dominant	chr12	8.95E-04	88.5	1.3	3.00E-02
6	KRT74	Dominant	chr12	8.95E-04	82.06	0.24	2.00E-01
7	MYO1A	Dominant	chr12	8.95E-04	72.93	1.7	4.00E-04
8	COL2A1	Dominant	chr12	8.95E-04	70.84	0.26	2.00E-01
9	PTPRB	Dominant	chr12	8.95E-04	69.98	0.2	2.00E-01
10	FAM186A	Dominant	chr12	8.95E-04	68.8	0.13	4.00E-01
11	ITGA7	Dominant	chr12	8.95E-04	65.31	0.26	2.00E-01
12	STAT2	Dominant	chr12	8.95E-04	58.34	0.02	2.00E-01

13	NEURL4	Dominant	chr17	8.95E-04	55.97	0.26	2.00E-01
14	TIMELESS	Dominant	chr12	8.95E-04	54.99	1	1.00E-02
15	CCDC65	Dominant	chr12	8.95E-04	54.04	0.53	6.00E-02
16	PFAS	Dominant	chr17	8.95E-04	52.31	0.23	1.00E-01
17	NLRP1	Dominant	chr17	8.95E-04	46.14	0.9	2.00E-02
18	EIF4B	Dominant	chr12	8.95E-04	40.3	0.29	5.00E-02
19	NELL2	Dominant	chr12	8.95E-04	32.33	0.11	1.00E-01
20	WNT10B	Dominant	chr12	8.95E-04	26.5	0.01	1.00E-01
21	ALG10B	Dominant	chr12	1.79E-03	35.78	0.12	1.00E-01
22	PELP1	Dominant	chr17	2.69E-03	49.3	0.37	1.00E-01
23	CHRNE	Dominant	chr17	2.69E-03	30.17	0.15	1.00E-01
24	GRASP	Dominant	chr12	2.69E-03	30.02	0.13	1.00E-01
25	ARHGEF15	Dominant	chr17	4.48E-03	30.98	0.98	4.00E-03
26	TROAP	Dominant	chr12	4.48E-03	28.79	0.17	1.00E-01
27	TRPV3	Dominant	chr17	7.16E-03	39.49	0.31	1.00E-01
28	P2RX5	Dominant	chr17	7.16E-03	35.82	0.56	6.00E-02
29	DNAJC22	Dominant	chr12	7.16E-03	28.07	0.13	8.00E-02
30	AMAC1L3	Dominant	chr17	8.95E-03	16.37	0.83	2.00E-03
31	ACADVL	Dominant	chr17	1.79E-02	24.81	0.51	6.00E-02
32	ZZEF1	Dominant	chr17	1.79E-02	39.71	0.56	2.00E-01
33	KRT76	Dominant	chr12	1.79E-02	38.8	0.3	1.00E-01
34	TEKT1	Dominant	chr17	1.79E-02	29.6	0.08	1.00E-01
35	ANKRD33	Dominant	chr12	1.79E-02	24.89	0.21	9.00E-02
36	PER1	Dominant	chr17	1.79E-02	32.14	0.69	7.00E-02
37	LYZ	Dominant	chr12	1.79E-02	12.03	2.13	2.00E-06
38	USP43	Dominant	chr17	2.69E-02	35.14	0.23	2.00E-01
39	KRT72	Dominant	chr12	3.58E-02	28.9	0.3	9.00E-02
40	C17orf74	Dominant	chr17	3.58E-02	24.67	0.23	9.00E-02
41	OR6C70	Dominant	chr12	3.58E-02	23.51	0.25	5.00E-02
42	KRT3	Dominant	chr12	4.48E-02	26.84	0.56	6.00E-02
43	OR3A3	Dominant	chr17	4.48E-02	19.58	0.04	1.00E-01
1	PELP1	Recessive	chr17	8.95E-04	66.7	0.79	3.00E-03
2	OR6C4	Recessive	chr12	8.95E-04	64.9	1.23	3.00E-04
3	C12orf54	Recessive	chr12	8.95E-04	47.19	0.6	5.00E-03
4	SHPK	Recessive	chr17	1.79E-03	26.61	1.83	6.00E-06
5	ESPL1	Recessive	chr12	2.69E-03	45.65	0.62	8.00E-03

6	KRT75	Recessive	chr12	1.79E-02	30.74	0.6	5.00E-03

Table 3.9. pVAAST linkage region candidate genes.

Gene	Chr	Description	pLI
SNORD118	chr17	small nucleolar RNA, C/D box 118	NA
STK38L	chr12	serine/threonine kinase 38 like	0.97
GRASP	chr12	general receptor for phosphoinositides 1 associated scaffold protein	0.93
SPRYD3	chr12	SPRY domain containing 3	0.95
DNAJC14	chr12	DnaJ heat shock protein family (Hsp40) member C14	1.00
PA2G4	chr12	proliferation-associated 2G4	1.00
ATF7IP	chr12	activating transcription factor 7 interacting protein	1.00
AEBP2	chr12	AE binding protein 2	0.95
CDK2	chr12	cyclin dependent kinase 2	0.96
XPOT	chr12	exportin for tRNA	1.00
SLC38A2	chr12	solute carrier family 38 member 2	0.97
NR4A1	chr12	nuclear receptor subfamily 4 group A member 1	0.96
SENP1	chr12	SUMO specific peptidase 1	1.00

Table 3.10. SV candidate genes within linkage regions.

3.3.5. Novel genes are found to be associated with ADHD and ASD or ADHD in the whole genome

We then performed pVAAST analysis for the whole genome. In determining number of permutations, we traded precision in calculating p-values for greater time efficiency. Therefore, we did not perform Bonferroni correction for p-values of genes in the whole genome scale, instead, we took the first 100 genes of each candidate gene set predicted by pVAAST (first 10 shown in Table 3.11, Table 3.12).

Ataxin 2 (*ATXN2*) appears in top rankings of both recessive models in "ASD or ADHD" and "ADHD only" phenotype. It bears 2 small inframe insertions of 3bp and 12bp segregating in 3 NJLAGS families. Intolerant to mutation (pLI ~= 1), *ATXN2* encodes the protein Ataxin-2 which is involved in epidermal growth factor receptor trafficking. *ATXN2* is also a candidate gene in ADHDgene database, and previously identified disease association of *ATXN2* includes neurodegenerative disorders such as spinocerebellar ataxia type 2 (SCA2) (Paciorkowski, et al., 2011), Parkinson's disease (Shan, et al., 2001), amyotrophic lateral sclerosis (ALS) (Ross, et al., 2011). Interestingly, *ATXN1*, another gene that is known to cause ataxia when disrupted, and is also related to brain development (Lu, et al., 2017), is ranked 12th in the dominant model gene set of ASD or ADHD phenotype. In addition to causing ataxia, *SACS*, another ataxia gene, forms neurofilaments comprising the structural framework that establishes the size and shape of the axons (Parfitt, et al., 2009), which echoes the cellular function of *NEURL4* and *KRT75*.

The Potassium Calcium-Activated Channel Subfamily N Member 3 (*KCNN3*) gene is a strong candidate gene suggested by the "ASD or ADHD" phenotype as it occupies the top ranking on both dominant and recessive models. It segregates in 2 families with 2 small deletions of 3bp, is intolerant to mutations (pLI = 0.86), and is expressed heavily in brain. *KCNN3* belongs to the KCNN family of potassium channels and encodes an calcium-activated channel thought to regulate neuronal excitability by contributing to the slow component of synaptic afterhyperpolarization (AHP) (O'Leary, et al., 2016). Previous studies have linked *KCNN3* to schizophrenia (Grube, et al., 2011) and bipolar disorder (Ujike, et al., 2001), but not to ASD or ADHD.

In SV analysis (Table 3.13), The Prostaglandin-Endoperoxide Synthase 2 (*PTGS2*) ranked high for both phenotypes, and it is found to be associated with ASD in a Korean population cohort (Yoo, et al., 2008). Elongator Acetyltransferase Complex Subunit 4 (*ELP4*), a subunit of a histone acetyltransferase complex that associates with RNA polymerase II transcriptional elongation, occupies the top ranking in "ASD or ADHD" phenotype, and a previous case-control study of ASD also identified a significant excess of CNVs in *ELP4* in case versus control (P=2.7 E-03) (Addis, et al., 2015).

Rank	Gene	Model	Chr	Score	PValue	LOD Score	LOD PValue
1	C9orf150	Dominant	chr9	249.2	1.00E-05	0.7	4.00E-02
2	FADS6	Dominant	chr17	246.39	1.00E-05	0.11	1.00E-01
3	EP400	Dominant	chr12	242.66	1.00E-05	0.9	1.00E-01
4	CTSA	Dominant	chr20	222.98	1.00E-05	0.01	3.00E-01
5	MUC5B	Dominant	chr11	209.92	1.00E-05	0.02	1.00E+00
6	FAM75C1	Dominant	chr9	199.8	1.00E-05	0.79	2.00E-01
7	POTEG	Dominant	chr14	197.52	1.00E-05	0.48	2.00E-01
8	PDE4DIP	Dominant	chr1	189.61	1.00E-05	1.25	2.00E-01
9	NBPF10	Dominant	chr1	187.41	1.00E-05	0.85	4.00E-01
10	PRKCSH	Dominant	chr19	184.77	1.00E-05	0.11	3.00E-01
1	ATXN2	Recessive	chr12	467.52	1.00E-05	0.62	4.00E-02
2	MMP17	Recessive	chr12	331.4	1.00E-05	0.73	4.00E-02
3	CCDC144NL	Recessive	chr17	242.24	1.00E-05	1.18	6.00E-03
4	CTBP2	Recessive	chr10	211.85	1.00E-05	1.9	2.00E-03
5	CNOT1	Recessive	chr16	148.71	1.00E-05	0.73	7.00E-03
6	FLT3	Recessive	chr13	133.12	1.00E-05	0.12	8.00E-02
7	OR11H12	Recessive	chr14	127.41	1.00E-05	1.23	8.00E-04
8	FAM38A	Recessive	chr16	114.69	1.00E-05	1.29	8.00E-04
9	ACACB	Recessive	chr12	111.85	1.00E-05	1.73	2.00E-04
10	ATP12A	Recessive	chr13	108.34	1.00E-05	0.6	2.00E-02

Table 3.11. Top 10 candidate genes for ADHD under dominant and recessive model in whole genome region.

Rank	Gene	Model	Chr	Score	PValue	LOD Score	LOD PValue
1	KCNN3	Dominant	chr1	326.77	1.00E-05	0.54	2.00E-01
2	C9orf150	Dominant	chr9	322.95	1.00E-05	0.12	3.00E-01
3	EP400	Dominant	chr12	310.49	1.00E-05	3.1	7.00E-03
4	MAP3K4	Dominant	chr6	303.38	1.00E-05	0.3	3.00E-01
5	ANKRD36	Dominant	chr2	285.05	1.00E-05	0.26	9.00E-01
6	PODXL	Dominant	chr7	282.74	1.00E-05	0.24	3.00E-01
7	MUC5B	Dominant	chr11	275.46	1.00E-05	0.58	1.00E+00
8	FAM75C1	Dominant	chr9	271.63	1.00E-05	0.79	4.00E-01
9	MEOX2	Dominant	chr7	267.63	1.00E-05	0.14	3.00E-01
10	CACNA1A	Dominant	chr19	267.44	1.00E-05	0.46	6.00E-01
1	KCNN3	Recessive	chr1	685.52	1.00E-05	0.85	8.00E-03
2	CEP170	Recessive	chr1	582.6	1.00E-05	1.12	1.00E-02
3	ATXN2	Recessive	chr12	564.77	1.00E-05	0.62	1.00E-02
4	CNOT1	Recessive	chr16	189.12	1.00E-05	0.73	3.00E-02
5	LOC649330	Recessive	chr1	153.35	1.00E-05	0.54	2.00E-01
6	CACNA1H	Recessive	chr16	148.86	1.00E-05	1.19	1.00E-02
7	AKAP13	Recessive	chr15	133.47	1.00E-05	2.07	3.00E-04
8	HYDIN	Recessive	chr16	131.03	1.00E-05	1.14	6.00E-03
9	MC1R	Recessive	chr16	130.99	1.00E-05	0.07	6.00E-02
10	SACS	Recessive	chr13	121.4	1.00E-05	0.87	1.00E-02

Table 3.12. Top 10 candidate genes for ASD or ADHD under dominant and recessive model in whole genome region.

Gene	Phenotype	Chr	Description	pLI
PRPF38A	ADHD	chr1	pre-mRNA processing factor 38A	1.00
ELP1	ADHD	chr9	elongator complex protein 1	NA
CFTR-AS1	ADHD	chr7	CFTR antisense RNA 1	NA
SYVN1	ADHD	chr11	synoviolin 1	1.00
STK38L	ADHD	chr12	serine/threonine kinase 38 like	0.97
KANSL3	ADHD	chr2	KAT8 regulatory NSL complex subunit 3	0.99
SIGMAR1	ADHD	chr9	sigma non-opioid intracellular receptor 1	0.13
PTGS2	ADHD	chr1	prostaglandin-endoperoxide synthase 2	1.00

LRRFIP2	ADHD	chr3	LRR binding FLII interacting protein 2	0.00
LOC102724058	ADHD	chr2	uncharacterized LOC102724058	NA
ELP4	ASD or ADHD	chr11	elongator acetyltransferase complex subunit 4	0.01
PRPF38A	ASD or ADHD	chr1	pre-mRNA processing factor 38A	1.00
KANSL3	ASD or ADHD	chr2	KAT8 regulatory NSL complex subunit 3	0.99
ELP1	ASD or ADHD	chr9	elongator complex protein 1	NA
CFTR-AS1	ASD or ADHD	chr7	CFTR antisense RNA 1	NA
SYVN1	ASD or ADHD	chr11	synoviolin 1	1.00
STK38L	ASD or ADHD	chr12	serine/threonine kinase 38 like	0.97
SIGMAR1	ASD or ADHD	chr9	sigma non-opioid intracellular receptor 1	0.13
PTGS2	ASD or ADHD	chr1	prostaglandin-endoperoxide synthase 2	1.00
LRRFIP2	ASD or ADHD	chr3	LRR binding FLII interacting protein 2	0.00

Table 3.13. Top 10 candidate genes for ADHD only and ASD or ADHD from SV data analysis in whole genome region.

3.3.6. Candidate genes participate in pathways related to neurological disorders

To integrate evidences from various sources and create a single candidate gene set, a consensus filtering among the significant linkage-region genes and the first 100 whole-gnome genes from the pVAAST candidate gene sets, and 2 SV gene set, one for "ADHD only" and one for "ASD or ADHD", was performed to create a high confidence set. Twenty-nine genes are present in at least four gene sets and we consider these 29 genes as the high-confidence set (Table 3.14). Another 670-gene set was created by merging all candidate gene sets to create a large gene set for overrepresentation analysis.

Gene	Chr	Description	pLI
TRPV3	chr17	transient receptor potential cation channel subfamily V member 3	0.00
USP6	chr17	ubiquitin specific peptidase 6	0.00
STARD9	chr15	StAR related lipid transfer domain containing 9	NA

SHPK	chr17	sedoheptulokinase	0.00
SACS	chr13	sacsin molecular chaperone	0.00
PELP1	chr17	proline, glutamate and leucine rich protein 1	1.00
PDIA2	chr16	protein disulfide isomerase family A member 2	0.00
PDE4DIP	chr1	phosphodiesterase 4D interacting protein	NA
OR6C4	chr12	olfactory receptor family 6 subfamily C member 4	0.00
OR11H1	chr22	olfactory receptor family 11 subfamily H member 1	0.77
NLRC5	chr16	NLR family CARD domain containing 5	0.00
MYH13	chr17	myosin heavy chain 13	0.00
MYBBP1A	chr17	MYB binding protein 1a	0.00
KMT2D	chr12	lysine methyltransferase 2D	1.00
METAP1	chr4	methionyl aminopeptidase 1	0.04
C1D	chr2	C1D nuclear receptor corepressor	0.36
KRT74	chr12	keratin 74	0.00
KDM6B	chr17	lysine demethylase 6B	1.00
HYDIN	chr16	HYDIN axonemal central pair apparatus protein	NA
GRASP	chr12	general receptor for phosphoinositides 1 associated scaffold protein	0.93
FAM86B2	chr8	family with sequence similarity 86 member B2	0.67
PIEZO1	chr16	piezo type mechanosensitive ion channel component 1	0.54
FAM186A	chr12	family with sequence similarity 186 member A	NA
ESPL1	chr12	extra spindle pole bodies like 1, separase	1.00
TSPOAP1	chr17	TSPO associated protein 1	NA
ARID1B	chr6	AT-rich interaction domain 1B	1.00
AKAP13	chr15	A-kinase anchoring protein 13	0.85
ACADVL	chr17	acyl-CoA dehydrogenase very long chain	0.00
ABCC3	chr17	ATP binding cassette subfamily C member 3	0.00

Table 3.14. Final candidate gene set integrating 8 previous gene sets via consensus filtering.

The high-confidence 29-gene set together with the combined 670-gene set was analyzed using online resources, including gene association databases, GWAS catalogs, literature search databases, tolerance to loss of function mutation scores, mouse knock-out experiments, expression databases, and pathway and enrichment analysis. A total of 85 genes from the 670 gene set (12.69%) was

described by previous research or existing neurodevelopmental disorder gene databases. Four genes from the 29 gene set are described previously (13.79%), with 3 of them (*KDM6B*, *ARID1B*, *METAP1*) overlapping with the Autism Sequencing Consortium and iPSYCH study's 102 risk genes (Satterstrom, et al., 2020).

Gene overrepresentation test were performed for the 670 combined gene set. In PANTHER overrepresentation test, we observed gene enrichment in pathways such as cell projection organization (1.96E-02), intracellular signal transduction (1.32E-02), microtubule-based process (4.73E-03). In KEGG pathway mapping analysis, Williams-Beuren syndrome (5) is the most represented disease, followed by Spinocerebellar ataxia (4) and deafness (4). Most represented pathways include metabolic pathways (43), MAPK signaling pathway (22), pathways in cancer (21), Alzheimer disease (17) and PI3K-Akt signaling pathway (17). Huntington disease (13), Alzheimer disease (9), calcium signaling (7) and microtubule-based transport (6) are the most perturbed neurological diseases related molecular networks.

The 29-consensus gene set analyzed in STRING functional association network (Figure 3.6) reveals linkage between the histone lysine-specific demethylase *KDM6B*, histone-lysine N-methyltransferase *KMT2D*, and chromatin remodeler gene *ARID1B*. The endocytic receptor *LRP1*, tentatively participating in kinase-dependent intracellular signaling, neuronal calcium signaling as well as neurotransmission, interacts with the intracellular estrogen-binding protein *PDIA2*. In addition to STRING analysis, ion channel proteins such as *TRPV3* and *PIEZO1* are co-mentioned in publication abstracts.

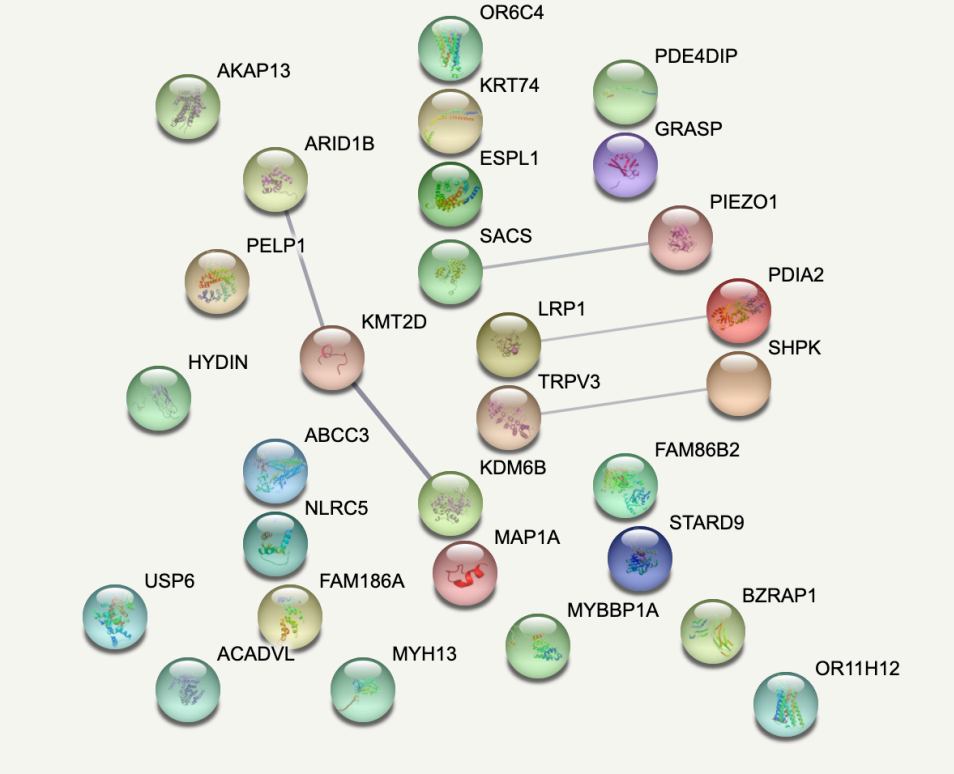


Figure 3.6. The 29-consensus gene interaction network generated in STRING database revealed connections between candidate genes.

3.4. Discussion

In this study, we have identified a novel linkage peak (17p13.1-2) for ADHD phenotype using the microarray genotyping data from 524 samples in 111 families. Using whole genome sequencing data from selected 272 samples, we discovered 49 significant genes within the linkage region and explored their gene functions. Expanding the analysis region to the whole genome revealed more risk factors for both "ADHD only" and "ASD or ADHD" phenotype. We observed multiple neurological disorder related pathways overrepresented by the merged 670-gene set and multiple disrupted gene networks by the high-confidence 29-gene set. Among the genes we found evidence in the NJLAGS dataset, 85 were also identified in previous studies, demonstrating the power of

combing linkage analysis with genotyping microarray and rare variant and SV analysis using whole genome sequencing to uncover risk genes associated with disease phenotypes.

By studying ADHD affected patients in families with ASD probands, we established genetic risk factors underlying ASD and ADHD cooccurrence. Three of the genes (*KDM6B*, *PTPRB*, *PER1*) that are significant in ADHD patients discovered within the ADHD linkage region are already described in previous ASD focused studies (Abrahams, et al., 2013; Schork, et al., 2019; Zhang, et al., 2012). Pathways we discovered that are important in ADHD etiology are also vital in ASD etiology. By providing an extensive candidate gene set, we are not only reaffirming the linkage between ASD and ADHD, but also suggesting novel potential risk pathways for further investigation. By expanding the search region to whole genome and expanding the phenotype to include ASD samples, we identified a spectrum of gene and pathways, previously known or novel, that are involved in ASD or ADHD etiology.

We noted several important pathways recurring in multiple gene sets in our functional analysis, namely microtubule related pathways, the MAPK pathway, the Akt pathway, ion channel related pathways, histone methylation, acetylation and chromatin remodeling pathways, and the circadian pathway. These pathways revealed interesting aspects of mechanisms in ASD and ADHD etiology as an architecture of interconnected network of hundreds of genes and pathways (Iakoucheva, et al., 2019), and underlined the complexity of genetic causalities in neurodevelopmental disorders. A recent article argued that several neurodevelopmental disorders share common genetic risk factors and there might not be "Autism-specific" genes (Myers, et al., 2020). Our results provide

support for this view. We raise awareness of important but underappreciated pathways such as the circadian pathway involved in both ASD and ADHD, and has phenotypic effects in cognition, mood, and reward-related behaviors. Researches documented that the prevalence of insomnia range from 50% to 80% in ASD patients, compared to 9–50% in age-matched typically developing children (Charrier, et al., 2017). However, the underlying mechanism of association and genes involved remain to be ascertained.

The limitation of sequencing capacity has led to the sequencing of only a selected number of ADHD affected individuals, as the NJLAGS dataset was collected primarily around ASD and language-impaired probands. Most recently, updated information regarding the ADHD status of NJLAGS family members was completed via follow-up questionnaires. Family members, primarily children, who were too young to receive an ADHD diagnosis during study recruitment, now have confirmed ADHD diagnoses. They will be included as affected for ADHD in any new genetic analyses. Moving forward, we anticipate extensive sequencing efforts to continue for the NJLAGS project resulting in sequencing of all NJLAGS families. As more ADHD affected individuals are sequenced, the more complete pedigree information and stronger statistical power will allow us to further discover potential common risk factors underlying the ASD and ADHD etiology.

4. Evaluating nanopore sequencing data processing pipelines for structure variation identification

4.1. Introduction

Structural variation (SV) is a major type of genomic variation. SV is usually fined as genomic alterations that are larger than 50 base pairs (bps) in size, such as insertions, deletions, duplications, inversions, and translocations. In humans, SVs account for the majority of the differences among individual genomes at the nucleotide level (Chaisson, et al., 2019; Korbel, et al., 2007; Sudmant, et al., 2015). SVs have a profound impact on the genome architecture and are associated with a variety of diseases, including neurological diseases and cancer (Carvalho and Lupski, 2016; Yang, et al., 2013). Therefore, studying SVs and their functional implications are critical to understand the genomic architecture and the underlying genetic factors for many diseases.

DNA sequencing became one of the primary methods for SV identification in recent years (Chaisson, et al., 2019; Korbel, et al., 2007; Sudmant, et al., 2015). Since 2005, a cost-effective, high-throughput generation of sequencing technology, termed next-generation sequencing (NGS), has been widely used in genomic research (Goodwin, et al., 2016; Kircher and Kelso, 2010). However, for SV identification NGS has limitations due to its short read-length (usually less than 200 bps), and most evidence supporting an SV event are indirect (e.g., read-depth, mis-match read pairs) (Treangen and Salzberg, 2011).

The arrival of a third generation of sequencing technology, characterized by real-time, single DNA/RNA molecule sequencing, allows for much longer read lengths, potentially addressing some of the limitations of NGS. These long reads

are believed to be advantageous towards the study of repetitive regions and SVs (Chaisson, et al., 2019). One sequencing technology that has generated a lot of interest is the nanopore sequencing technology commercialized by Oxford Nanopore Technologies (ONT) (Bayley, 2015; Jain, et al., 2016). Unlike many other sequencing methods, nanopore sequencing does not require the detection of fluorophore which typically indicates a product of chemical or enzymatic reaction. Instead, single-stranded DNA/RNA molecules are directly sequenced by measuring a current disruption as the molecule passes through a nanopore (Bayley, 2015). Long reads obtained from the nanopore sequencing offer possibilities to detect SVs in a single continuous read instead of being inferred through indirect evidences from short reads. In the last several years, new computational tools have been developed specifically for long-read data and several studies have identified SVs using the nanopore data (Cretu Stancu, et al., 2017; Jain, et al., 2018; Miao, et al., 2018; Wouter, et al., 2018). However, since the ONT sequencers were only recently launched, the tools available for aligning long-read data and detecting SVs have not yet been thoroughly evaluated.

In this study, we evaluated several aligners and SV callers on the nanopore data using three human nanopore datasets, including both empirical sequencing data and simulated reads. By comparing SV calls from seven aligner-SV caller combinations to high quality SV call sets, we evaluated the performance of long-read aligners, SV callers, and their overall combined performance.

4.2. Results

4.2.1. Selection of benchmarking dataset

For benchmarking, it is preferable to use several different datasets. In this study, we used three datasets: nanopore sequencing of the human sample NA12878

(referred to as NA12878 in the following text); simulated nanopore reads based of the human genome assembly CHM1(referred to as CHM1 in the following text); and simulated SV events and nanopore reads based on the chromosome 20 of the human reference genome GRCh38 (referred to as Chr20 in the following text).

The sample NA12878 was sequenced at ~30x coverage depth by the nanopore whole genome sequencing consortium (Jain, et al., 2018). For the SV true set we used the SV call set generated by the Genome in a Bottle Consortium (Zook, et al., 2019). This call set was based on the whole genome sequencing data at ~44x coverage using the Pacific Biosciences (PacBio) platform. SV calls were generated using three SV detection methods, including a local assembly pipeline (Chaisson, et al., 2015) (Table 4.1).

Pipelines
PBHoney, raw reads, blasr1.3.1
Custom pipeline, raw reads, blasr1.3.1
PBHoney, error-corrected reads, blasr1.3.1
Custom pipeline, error-corrected reads, blasr1.3.1
Local Assembly
Custom pipeline, error-corrected reads, blasr1.3.2
Custom pipeline, raw reads, blasr1.3.2

Table 4.1. SV calls of the NA12878 true set are integrated from seven call sets.

The CHM1 genome was assembled from a human haploid hydatidiform mole using reference-guided assembly (Steinberg, et al., 2014). Based on the CHM1 assembly, we simulated the nanopore sequencing reads to ~50x coverage (see Methods). Mapping the simulated nanopore reads resembles mapping empirical sequencing reads from an individual with a CHM1 genome. As a

corresponding true SV call set for this sample, we used a SV call set generated using the PacBio platform (Chaisson, et al., 2015).

The SV true sets for NA12878, CHM13, and CHM1 samples are dependent on their respective analysis pipelines and were filtered to select SVs with high accuracy. Therefore, it is likely that these true sets are incomplete which could affect the false positive rate estimates for SV calling pipelines. To address this issue, we simulated the chromosome 20 of the human reference genome GRCh38 with pre-defined SVs and generated nanopore sequencing reads at ~50x coverage for pipeline evaluation.

To assess overall properties of the true sets, we collected several statistics of the true sets (Table 4.2). All true sets have more insertions than deletions. CHM1 and CHM13 true sets have more than 2-fold higher number of calls compared to the NA12878 set. SV size distribution analysis showed that most SVs are less than 500 bps in length (Figure 4.1), and only a small number of SVs were larger than 10,000 bps (Table 4.3). For all sets, a peak could be observed at ~300 bps, an expected size for *Alu* transposable elements (Figure 4.1).

	NA12878 deletion	**NA12878 insertion**	**CHM13 deletion**	**CHM13 insertion**	**CHM1 deletion**	**CHM1 insertion**	**Chr20 deletion**	**Chr20 insertion**
SV count	4352	5783	10,671	20,497	10,784	15,158	96	181
Median size (bp)	312	300	304	318	69	103	318	296
Longest size (bp)	97,696	41,311	26,862	32,727	18,511	71,339	10,937	41,310
Shortest size (bp)	34	32	50	51	31	32	50	40

Table 4.2. Summary statistics of the SV true sets.

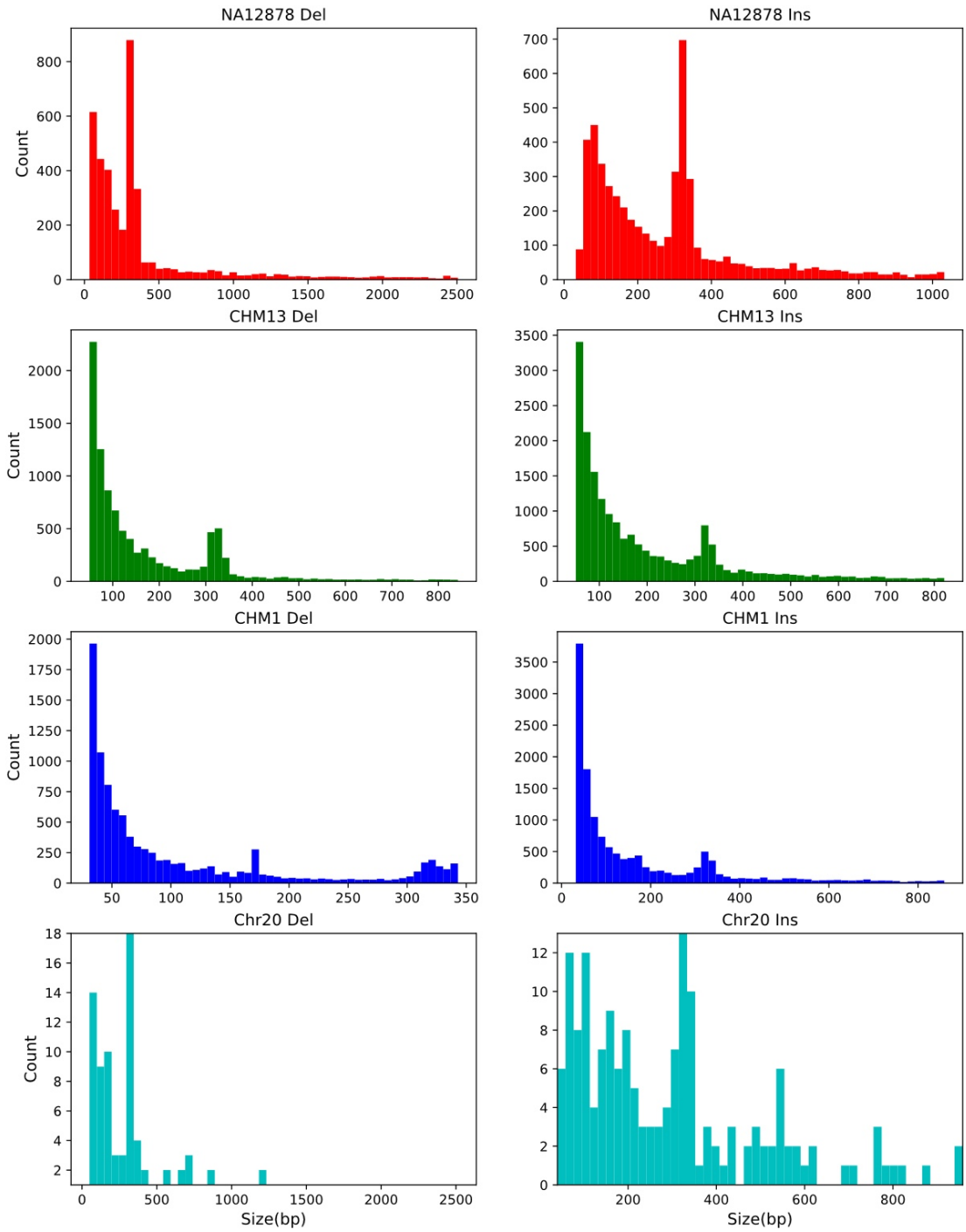


Figure 4.1. True set indel size distribution.
The size distribution of insertions and deletions in the four SV true sets are shown. The indels were filtered to remove the top 10% largest calls to improve the visibility. Counts of larger SVs are listed in Table 4.1.

	No. of SVs	Recall	Precision	F1 Score	0-99bp	100-499bp	500-999bp	1-5kb	5-10kb	>=10kb

NA12878-Deletion										
minimap2-nanosv	68661	72.13%	8.34%	14.96%	62144	5482	361	485	134	55
minimap2-sniffles	37666	71.55%	17.32%	27.89%	31979	4763	316	427	139	42
ngmlr-nanosv	66128	70.11%	8.08%	14.49%	58596	6387	398	534	158	55
ngmlr-sniffles	34573	66.25%	17.34%	27.49%	28439	5158	342	451	137	46
graphmap-nanosv	38683	48.28%	13.25%	20.79%	34408	3664	229	336	46	0
graphmap-sniffles	20444	47.45%	26.15%	33.72%	15966	3735	315	393	35	0
last-picky	33814	53.13%	13.00%	20.89%	24889	7599	854	342	108	22
truth	4352				771	2489	311	557	172	52
CHM13-Deletion										
minimap2-nanosv	12155	50.25%	37.47%	42.93%	7572	3589	383	450	99	62
minimap2-sniffles	11147	46.81%	44.19%	45.46%	6874	3271	376	456	109	61
ngmlr-nanosv	11060	49.95%	39.33%	44.01%	6586	3394	396	518	109	57
ngmlr-sniffles	9424	44.97%	46.52%	45.73%	5627	2765	340	501	121	70
graphmap-nanosv	8627	36.00%	44.38%	39.76%	5015	2697	293	474	105	43
graphmap-sniffles	8155	32.16%	47.74%	38.43%	4172	2895	359	556	121	52
last-picky	6119	26.48%	42.93%	32.76%	3192	2194	252	374	90	17
truth	10671				4490	4736	486	750	166	43

CHM1-Deletion										
minimap2-nanosv	18641	40.96%	25.12%	31.14%	11249	5438	1099	726	96	33
minimap2-sniffles	14741	34.83%	30.88%	32.74%	9280	4494	549	341	55	22
ngmlr-nanosv	14600	39.20%	29.77%	33.84%	8357	4977	679	489	68	30
ngmlr-sniffles	12005	34.62%	35.02%	34.82%	7103	3934	499	382	60	27
graphmap-nanosv	12998	28.83%	30.18%	29.49%	7985	4243	473	285	12	0
graphmap-sniffles	8954	22.77%	34.71%	27.50%	4806	3533	351	256	8	0
last-picky	13089	24.10%	21.06%	22.48%	8576	3752	395	302	47	17
truth	10784				6581	3457	317	329	89	11
Chr20-Deletion										
minimap2-nanosv	279	97.92%	33.69%	50.13%	42	156	42	31	7	1
minimap2-sniffles	275	96.88%	33.82%	50.13%	46	153	39	30	5	2
ngmlr-nanosv	238	96.88%	39.08%	55.69%	43	151	23	16	4	1
ngmlr-sniffles	272	95.83%	33.95%	50.14%	44	151	39	31	4	3
graphmap-nanosv	270	94.79%	33.70%	49.73%	35	159	41	32	3	0
graphmap-sniffles	313	92.71%	28.62%	43.73%	57	181	42	31	2	0
last-picky	790	92.71%	11.37%	20.25%	540	170	42	30	5	3
truth	96				14	49	13	15	4	1

NA12878-Insertion										
minimap2-nanosv	39260	65.12%	25.71%	36.86%	22024	14709	1208	1117	155	47
minimap2-sniffles	7428	47.29%	51.62%	49.36%	3403	3474	341	192	13	5
ngmlr-nanosv	21971	58.95%	39.06%	46.99%	10389	8974	1436	1047	114	11
ngmlr-sniffles	5860	39.96%	56.03%	46.65%	2025	2957	589	235	37	17
graphmap-nanosv	22046	54.00%	44.29%	48.66%	11682	8452	1050	847	15	0
graphmap-sniffles	3426	32.11%	72.73%	44.55%	885	1918	324	299	0	0
last-picky	4574	29.64%	50.77%	37.43%	703	2547	852	460	9	3
truth	5783				1089	3461	622	570	28	13
CHM13-Insertion										
minimap2-nanosv	25885	64.71%	63.80%	64.25%	12722	11320	1008	759	70	6
minimap2-sniffles	10989	55.92%	64.52%	59.91%	5526	4689	492	264	16	2
ngmlr-nanosv	20617	49.00%	60.58%	54.18%	9665	9552	937	421	40	2
ngmlr-sniffles	9915	41.79%	61.69%	49.82%	4496	4449	648	316	2	4
graphmap-nanosv	22068	57.13%	69.09%	62.54%	11746	9147	738	437	0	0
graphmap-sniffles	7167	41.93%	71.39%	52.83%	3364	3266	313	224	0	0
last-picky	7952	20.88%	55.66%	30.36%	3977	3464	370	139	1	1
truth	20497				7280	9903	1637	1437	195	45

CHM1-Insertion										
minimap2-nanosv	194	12.88%	22.65%	16.42%	34	108	32	18	2	0
minimap2-sniffles	175	11.12%	25.43%	15.47%	28	106	27	14	0	0
ngmlr-nanosv	171	11.55%	23.46%	15.48%	37	117	15	2	0	0
ngmlr-sniffles	133	10.13%	25.95%	14.58%	29	101	2	0	0	1
graphmap-nanosv	303	9.85%	23.83%	13.94%	81	168	40	14	0	0
graphmap-sniffles	179	7.26%	26.97%	11.45%	44	98	25	12	0	0
last-picky	158	6.64%	22.70%	10.27%	30	104	23	0	0	1
truth	181				29	106	29	15	1	1
Chr20-Insertion										
minimap2-nanosv	40536	90.61%	94.94%	92.73%	20918	15644	2044	1623	196	111
minimap2-sniffles	14421	88.95%	99.38%	93.88%	7298	5145	946	869	121	42
ngmlr-nanosv	28426	76.24%	99.35%	86.28%	14272	10358	1916	1585	216	79
ngmlr-sniffles	11960	66.85%	99.18%	79.87%	5244	4344	1194	1033	103	42
graphmap-nanosv	35734	90.06%	99.64%	94.61%	18142	13193	1997	2062	302	38
graphmap-sniffles	10006	88.95%	98.80%	93.61%	4639	3713	686	811	130	27
last-picky	6974	76.80%	97.20%	85.80%	1976	3134	1082	743	30	9
truth	14779				5061	6848	1354	1292	178	46

Table 4.3. SV call set evaluation.

4.2.2. Aligner and SV caller selection

Multiple aligners and SV callers were downloaded and tested on the three nanopore datasets (Table 4.4, Table 4.5). After testing, we excluded several tools from downstream analysis for a variety of reasons (see Table 4.5 for details). As a result, we examined four aligners (Minimap2, NGMLR, GraphMap, LAST) and three SV callers (Sniffles, NanoSV, Picky). We selected these tools based on their usability, compatibility, maintenance status, and popularity.

Name	Type	Version	Release year	Threads	Language	Description	Citation
GraphMap	Aligner	0.5.2	2016	16	C++	Aligns nanopore long reads with circular genome handling	(Sović, et al., 2016)
LAST	Aligner	941	2011	16	C++	Modified BLAST, outputs MAF format	(Kiełbasa, et al., 2011)
minimap2	Aligner	2.1	2017	16	C	Aligns error-prone long reads, faster and more accurate than BWA	(Li, 2018)
NGMLR	Aligner	0.2.6	2017	16	C++	Works with nanopore long reads to generate high-quality SV calls	(Sedlazeck, et al., 2018)
NanoSV	SV caller	1.2.0	2017	16	Python	Identifies and clusters split reads based on genomic positions and orientations to identify breakpoint	(Cretu Stancu, et al., 2017)

						junctions of SVs	
Picky	SV caller	0.2.a	2017	16	Perl	"Pick"-and-stitch segments from LAST alignments into representative alignments with a greedy algorithm	(Gong, et al., 2018)
Sniffles	SV caller	1.0.8	2017	16	C++	Detects all types of SVs using split-read alignments, high-mismatch regions, and depth of coverage	(Sedlazeck, et al., 2018)

Table 4.4. Evaluated aligners and SV callers.

Name	Type	Version	Release Year	Language	Description	Citation
Meta-aligner	Aligner	N/A	2017	C++	Run failed	(Nashta-ali, et al., 2017)
MashMap	Aligner	2	2017	C++	Only aligns long reads	(Jain, et al., 2017)
BLASR	Aligner	5.3.2	2012	C++	Output not compatible with other tools	(Chaisson and Tesler, 2012)
SMRT-SV	SV Caller	N/A	2017	Python	Requires BLASR	(Huddleston, et al., 2017)
HySA	SV Caller	N/A	2017	Perl	Not packaged into a tool	(Fan, et al., 2017)
PBHoney	SV Caller	15.8.24	2014	Python	Not maintained	(English, et al., 2014)

Table 4.5. Aligners and SV callers excluded from the analysis.

4.2.3. Aligner resource consumption and performance

First, we compared the computational resource consumptions of the four aligners: minimap2, NGMLR, GraphMap, and LAST (Figure 4.2A). Overall, each aligner performed similarly across datasets. Among the four aligners, minimap2 was the fastest by a large margin compared to other aligners, while GraphMap was the slowest. GraphMap also consumed the most memory. The file system operations were similar among all aligners (Figure 4.2A, FS Operations). Next, we compared the quality of the aligned reads, such as the total mapped bases, mismatch rate, and genome coverage (Table 4.6). LAST's output was not included in this analysis because its output was directly piped to the Picky for SV detection. Mapping coverage for NA12878 was ~24x for all aligners, compared to the raw sequencing coverage depth of ~30x. CHM13 had a higher coverage than NA12878, at ~42x. CHM13 also had a lower mismatch rate than NA12878, regardless of the aligner used. This difference might reflect the longer read-length and the newer base-calling program used in the CHM13 dataset. The two simulated datasets, CHM1 and Chr20, has ~40x and ~50x coverage, respectively (Table 4.6).

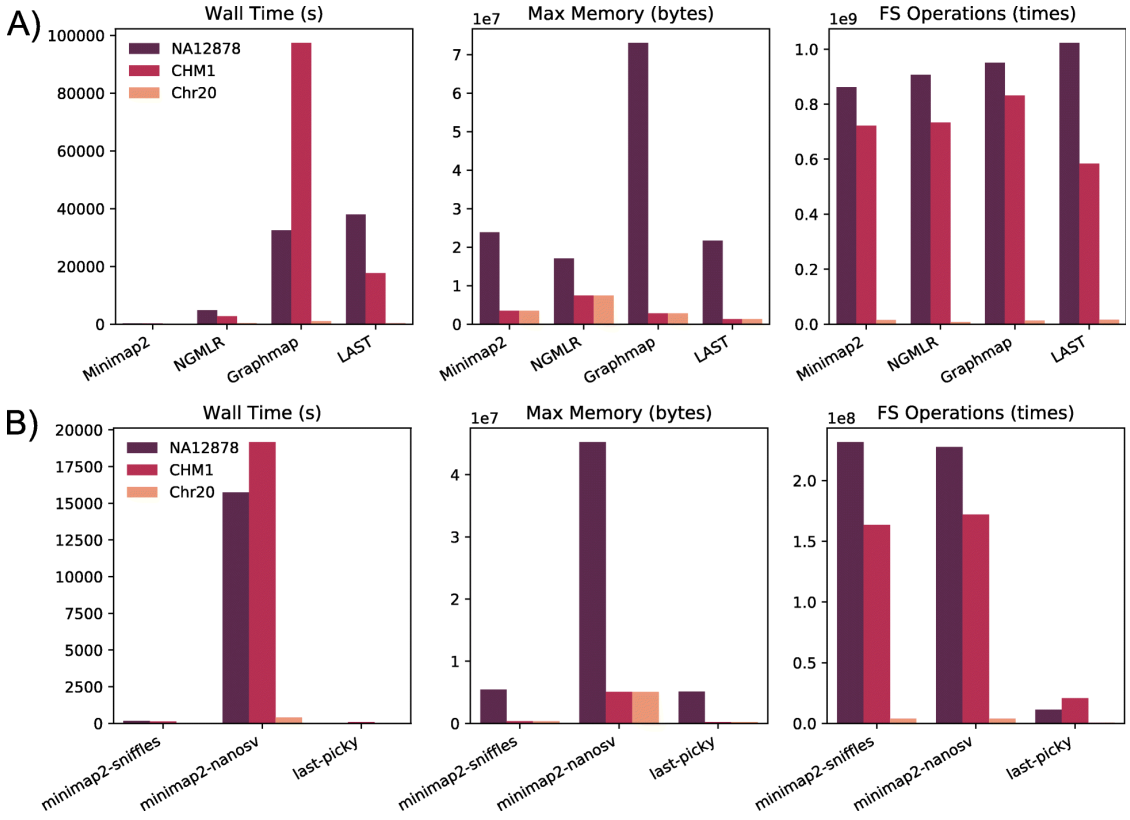


Figure 4.2. Resource consumption.
A) Aligner. B) SV caller. The computing time (Wall Time), maximum memory usage (Max Memory), and file system operations (FS Operations) are shown. LAST's Wall Time included Picky's representative read selection step because Picky requires a specific output format from LAST in place of the default output. SV callers are noted with respect to the aligner used to map the reads. The CHM13 dataset was analyzed under multiple cluster configurations and therefore was not included in the computational resource evaluation.

Aligner	Dataset	Bases mapped (Gb)	Mismatch rate	Coverage
minimap2	NA12878	77.5	1.97E−01	24.4
NGMLR	NA12878	73.6	1.92E−01	23.4
GraphMap	NA12878	80.2	2.17E−01	25.1
minimap2	CHM13	144.7	1.12E−01	43.7
NGMLR	CHM13	137.3	1.05E−01	42.0
GraphMap	CHM13	139.6	1.24E−01	42.7

minimap2	CHM1	128.6	1.35E−01	39.6
NGMLR	CHM1	127.6	1.35E−01	39.5
GraphMap	CHM1	130.4	1.52E−01	39.7
minimap2	Chr20	3.3	1.35E−01	48.5
NGMLR	Chr20	3.2	1.34E−01	47.4
GraphMap	Chr20	3.3	1.54E−01	49.1

Table 4.6. Alignment statistics.

4.2.4. SV calling pipeline resource consumption and call set evaluation

We then compared computational resource consumptions for the three SV callers: NanoSV, Sniffles, and Picky (Figure 4.2B). SV caller results were collected based on minimap2 alignments for NanoSV and Sniffles, and LAST alignment for Picky. Time and memory usage results highlighted that NanoSV consumed substantially more resources than the other two SV callers. The main time-consuming step of the NanoSV calling was calculating the depth of coverage at the potential SV breakpoints. Picky performed fewer file system operations partially because the "select representative reads" step was performed in combination with LAST before the SV calling step.

Because the percentage of mapped reads were similar among all aligners, we chose minimap2, NGMLR, and GraphMap as aligners to test in combination with Sniffles and NanoSV. The LAST alignment output format was not fully compatible with Sniffles and NanoSV, so we only evaluated LAST with Picky. LAST was chosen to run with Picky also because of its claimed synergy with Picky and it was incorporated in the default Picky workflow (Gong, et al., 2018). In total we tested seven SV calling pipelines: Minimap2-NanoSV, NGMLR-

NanoSV, GraphMap-NanoSV, Minimap2-Sniffles, NGMLR-Sniffles, GraphMap-Sniffles, and LAST-Picky.

Each SV caller called different types of SVs with different abundance as shown in Table 4.7. Deletion was the most abundant category, followed by insertions and duplications. The remaining categories, including inversions, translocations, etc, all contain small number of calls. Because only a small number of duplications were called and the true sets only contains insertions and deletions, the SV calls were grouped into two main categories: deletions and insertions. As such, duplications were merged with insertions. Other types of SVs (e.g., inversions, translocations) from the call sets were not included in the evaluation.

SV Type	minimap2-sniffles	minimap2-nanosv	ngmlr-sniffles	ngmlr-nanosv	graphmap-sniffles	graphmap-nanosv	last-picky
DEL	38,364	82,989	34,877	82,175	21,166	52,019	41,525
DUP	111	491	634	523	0	0	2,535
INS	7,645	39,698	5,377	22,096	3,629	22,289	2,102
Others	280	0	280	0	29	0	19
DEL: deletion; DUP: duplication; INS: insertion; Others: inversion, translocation, etc.							

Table 4.7. Counts of different types of NA12878 SVs called by the seven pipelines.

The size distribution of the call sets showed more small SVs than large SVs, a pattern similar to the true sets (Figure 4.3, Table. 4.3). NanoSV called more insertions and deletions than Sniffle and Picky. In the simulated Chr20 dataset, Picky called more small deletions than any other pipeline. This is likely due to Picky's goal to maximize sensitivity and the high coverage in the Chr20 set resulted in a high false-positive rate.

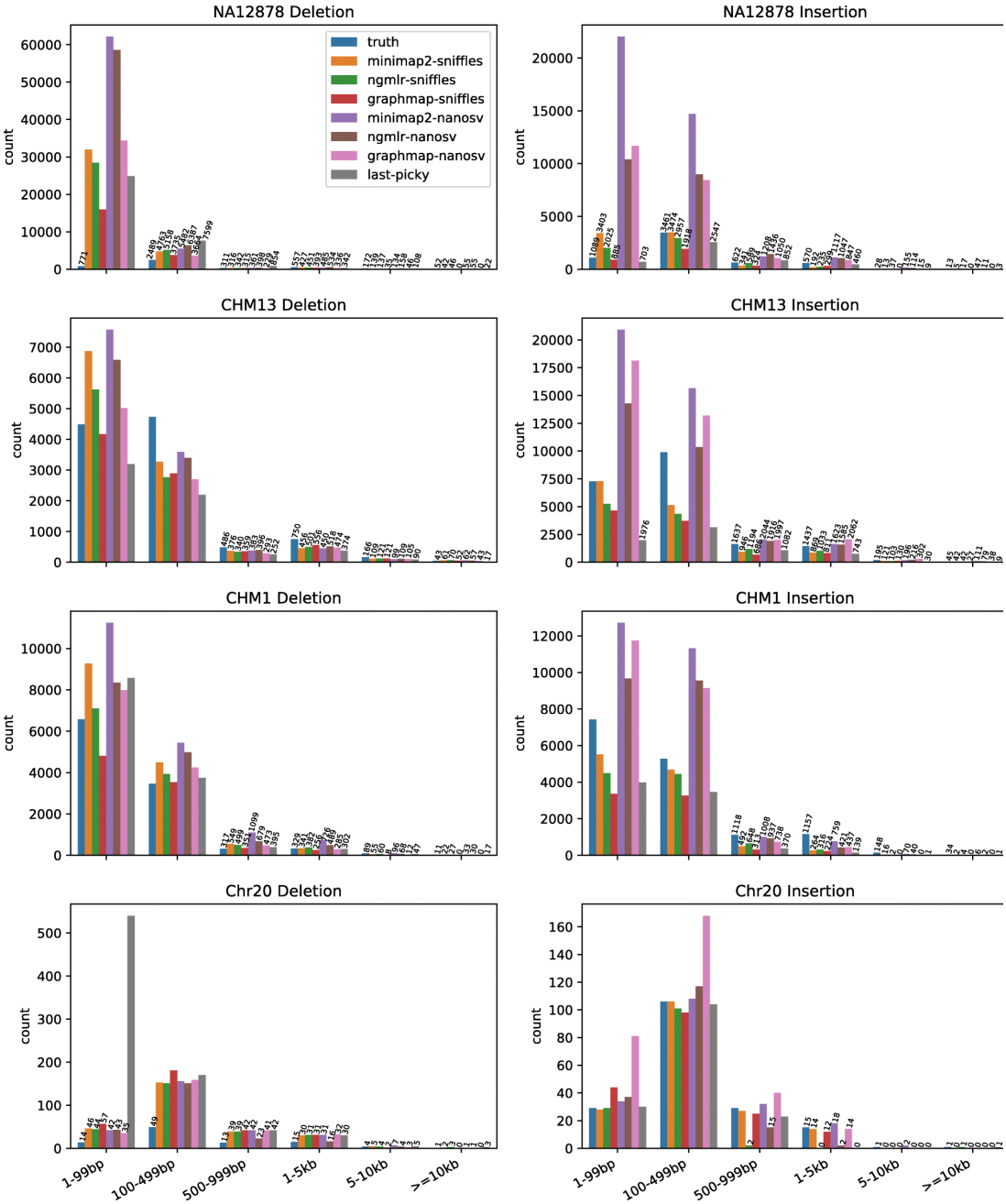


Figure 4.3. Insertion and deletion call set size distribution.
The number of insertions and deletions in six size categories is shown for the true sets and calls from seven SV calling pipelines for the four datasets.

To evaluate the quality of the SV calls, we calculated the precision, recall, and F1 score for each call set (Table 4.3). The precision-recall graph showed that the four datasets occupy distinct areas (Figure 4.4). The calls from the Chr20

dataset clustered on the right side of the plot, indicating that all call sets have high recall rates, although the precision was much higher for insertions than deletions. LAST-Picky deletion call set had the most false positive calls (precision rate 11%), while NGMLR-Sniffles insertion calls had the lowest recall (73%). The NA12878 call sets, especially insertions (Figure 4.4, cyan color), are in the central area of the graph and have the widest spread among different pipelines. The observed spread suggests that different pipelines had different precision versus recall advantages. As such, NanoSV call sets demonstrated highest recall rates (Figure 4.4, cyan colored circle, square and cross), with Minimap2-NanoSV being the highest (Figure 4.4, cyan colored circle). Sniffles and Picky, on the other hand, had better precision rates, with the highest being GraphMap-Sniffles (Figure 4.4, cyan colored diamond). The CHM13 dataset clustered in the center area (Figure 4.4, orange and yellow colors), suggesting different pipelines performed more consistent in this dataset. For CHM13 Minimap2-NanoSV had the highest recall rate and GraphMap-Sniffles had the highest precision. Finally, the CHM1 insertion call sets occupied the bottom-left area, which made it the worst call set given the true set, especially for the recall rates. CHM1 deletions were called with a small recall advantage over insertions (Figure 4.4, red and magenta colors, respectively).

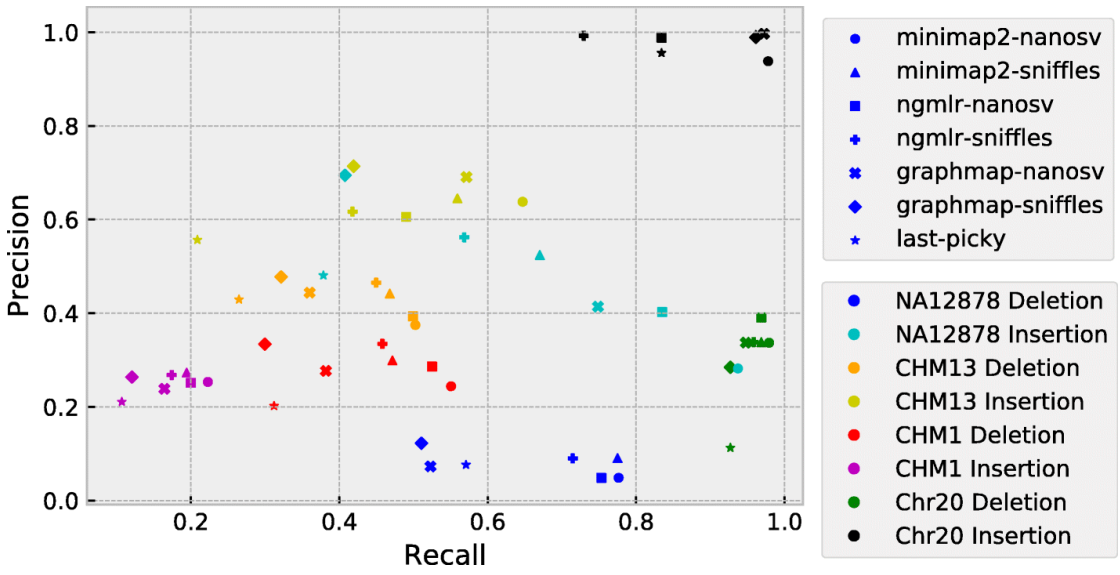
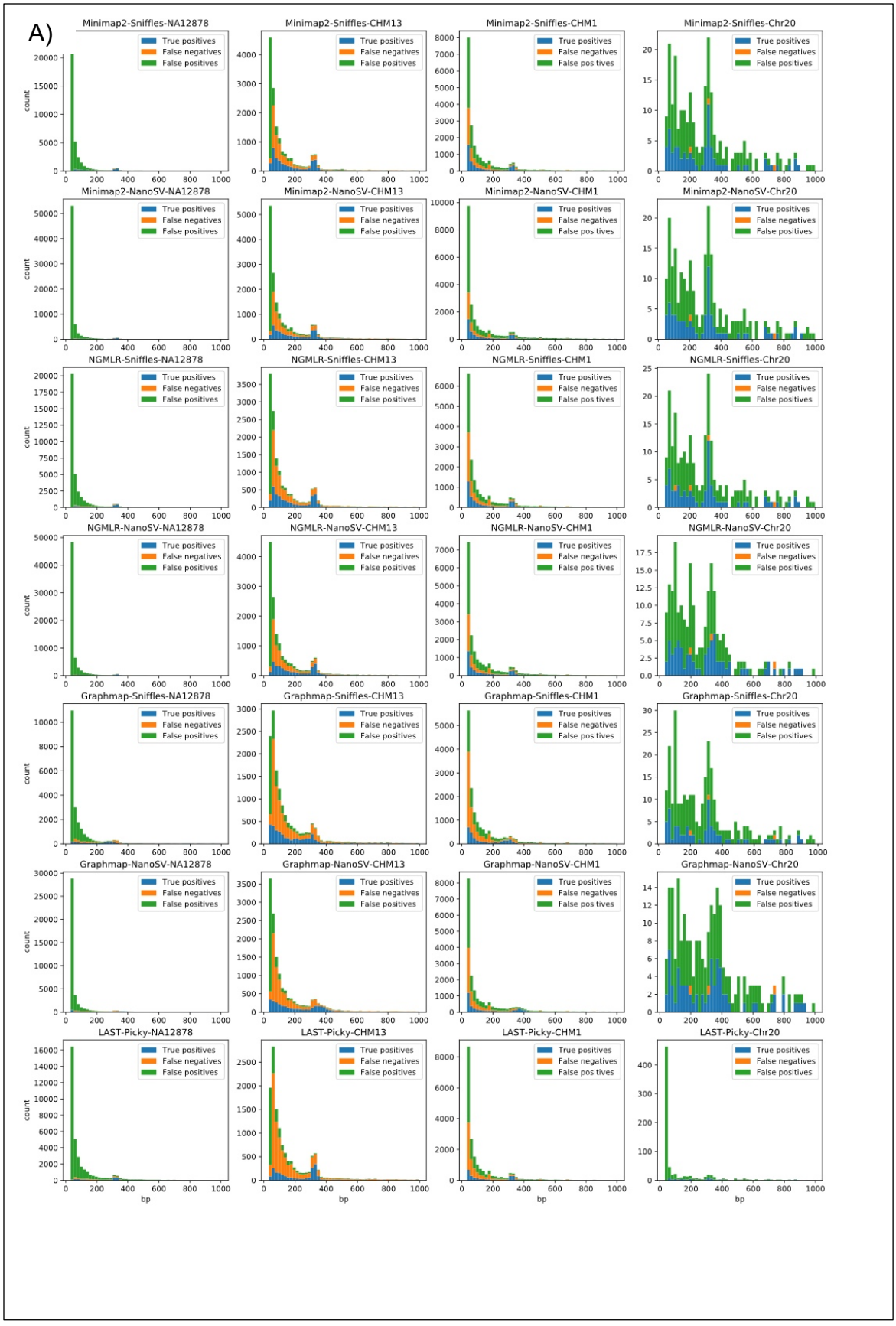


Figure 4.4. Precision-recall graph of SV calling pipelines.
Pipelines are represented by shapes, and datasets are represented by colors as specified in the legend.

We next determined the rates of true positive, false negative and false positive calls in each call set stratified by indel size (Figure 4.5). All pipelines performed the best for insertions in the Chr20 dataset, achieving a high true positive rate (Figure 4.5B). For deletions, all Chr20 call sets contained many false positive calls, especially the LAST-Picky call set. Individual call datasets also showed different performance in different size distributions. In the NA12878 dataset, most pipelines identified many false positive calls for SVs smaller than 200 bps, especially for deletions (Figure 4.5). One possible reason for the high false positive rates of the small SVs could be that nanopore sequencing reads have a high error rate at homopolymer and low complexity regions. To test the effect of these repetitive regions, we subsequently excluded SVs overlapping simple repeats and low complexity regions in the reference genome. The NA12878 filtered call sets indeed showed improvements for precisions, especially for deletions. However, filtering calls in the repetitive region also

reduced the recall rates of the call sets (Figure 4.6). For the CHM13 call sets, all pipelines generally had more false negative calls when calling small SVs. CHM1 dataset displays similar pattern to the CHM13 dataset, but showing slightly lower true positive rate, especially for insertions.


A)
Minimap2-Sniffles-NA12878
Minimap2-Sniffles-CHM13
Minimap2-Sniffles-CHM1
Minimap2-Sniffles-Chr20
Minimap2-NanoSV-NA12878
Minimap2-NanoSV-CHM13
Minimap2-NanoSV-CHM1
Minimap2-NanoSV-Chr20
NGMLR-Sniffles-NA12878
NGMLR-Sniffles-CHM13
NGMLR-Sniffles-CHM1
NGMLR-Sniffles-Chr20
NGMLR-NanoSV-NA12878
NGMLR-NanoSV-CHM13
NGMLR-NanoSV-CHM1
NGMLR-NanoSV-Chr20
Graphmap-Sniffles-NA12878
Graphmap-Sniffles-CHM13
Graphmap-Sniffles-CHM1
Graphmap-Sniffles-Chr20
Graphmap-NanoSV-NA12878
Graphmap-NanoSV-CHM13
Graphmap-NanoSV-CHM1
Graphmap-NanoSV-Chr20
LAST-Picky-NA12878
LAST-Picky-CHM13
LAST-Picky-CHM1
LAST-Picky-Chr20
True positives
False negatives
False positives
count
bp

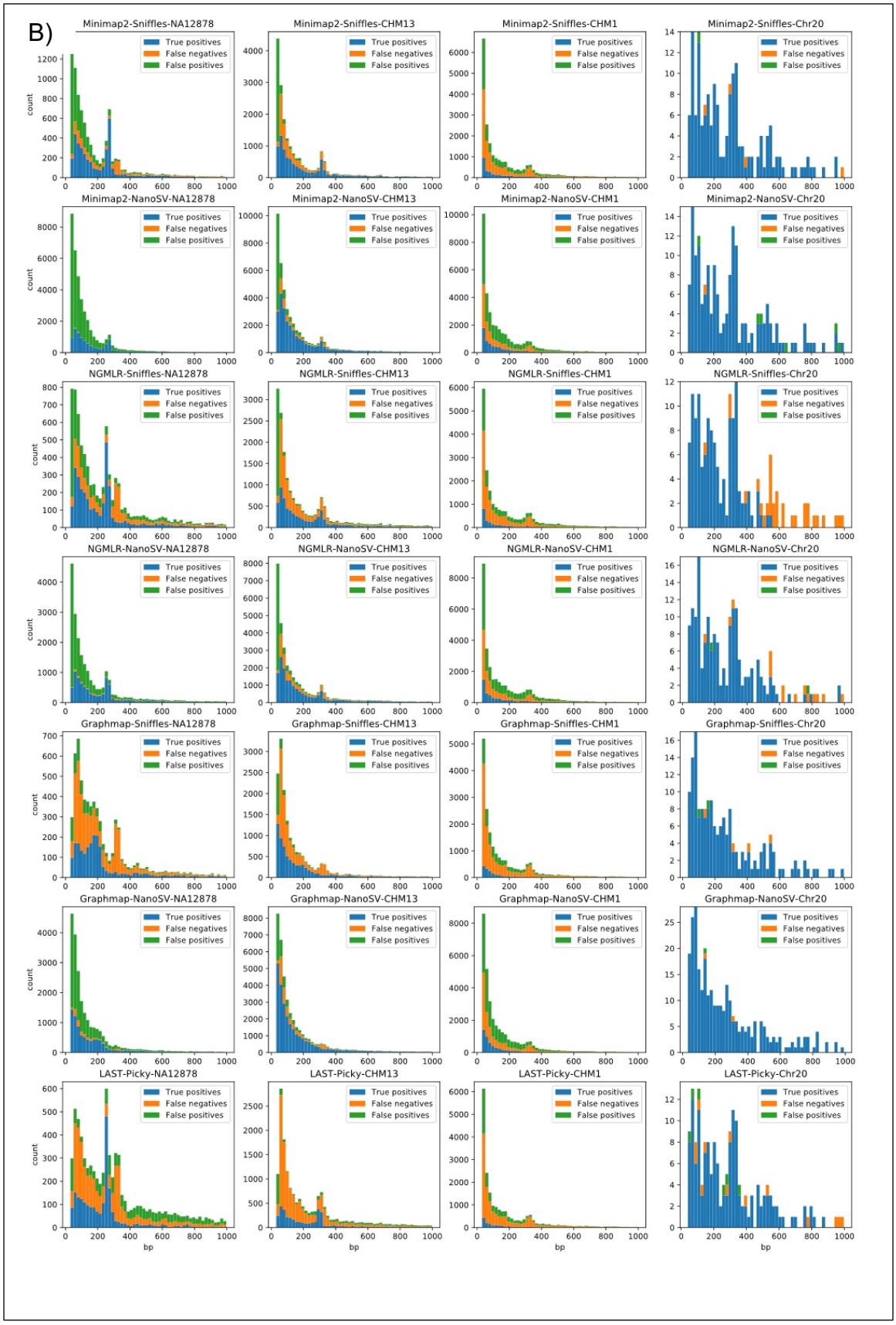

B)
Minimap2-Sniffles-NA12878
Minimap2-Sniffles-CHM13
Minimap2-Sniffles-CHM1
Minimap2-Sniffles-Chr20
Minimap2-NanoSV-NA12878
Minimap2-NanoSV-CHM13
Minimap2-NanoSV-CHM1
Minimap2-NanoSV-Chr20
NGMLR-Sniffles-NA12878
NGMLR-Sniffles-CHM13
NGMLR-Sniffles-CHM1
NGMLR-Sniffles-Chr20
NGMLR-NanoSV-NA12878
NGMLR-NanoSV-CHM13
NGMLR-NanoSV-CHM1
NGMLR-NanoSV-Chr20
Graphmap-Sniffles-NA12878
Graphmap-Sniffles-CHM13
Graphmap-Sniffles-CHM1
Graphmap-Sniffles-Chr20
Graphmap-NanoSV-NA12878
Graphmap-NanoSV-CHM13
Graphmap-NanoSV-CHM1
Graphmap-NanoSV-Chr20
LAST-Picky-NA12878
LAST-Picky-CHM13
LAST-Picky-CHM1
LAST-Picky-Chr20
True positives
False negatives
False positives
count
bp

Figure 4.5. Quality of each SV call set by size.
A) deletions; B) insertions. In each size bin, the calls are divided into True positives (blue), False negatives (orange), and False positives (green), based on the comparison with the true set. Only SVs smaller than <1,000 bps are shown to improve visibility.

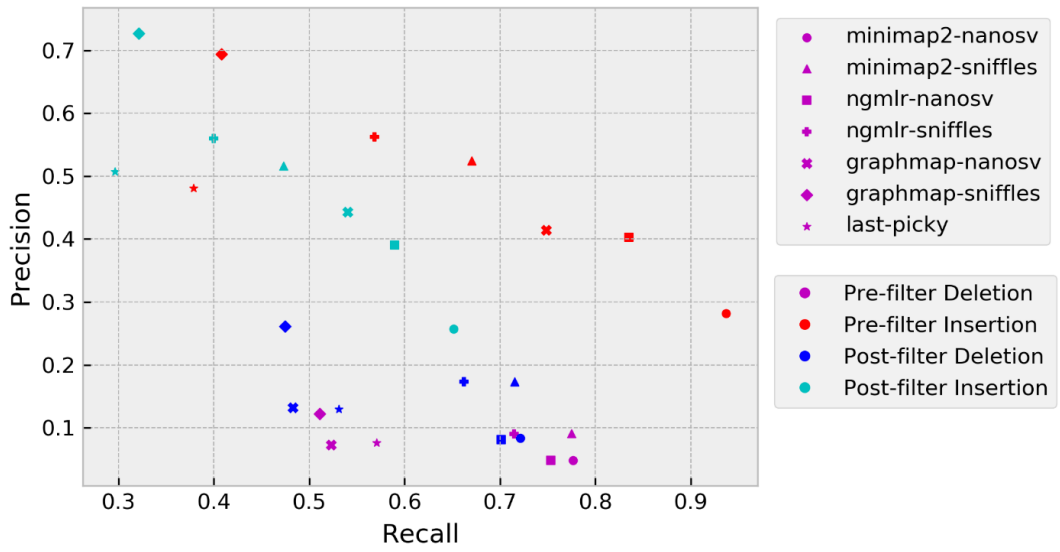


Figure 4.6. Precision-recall graph of NA12878 SV calls before and after filtering repetitive genomic regions.
SV call sets were filtered to exclude SVs overlapping "Simple repeat" or "Low complexity" regions in the human genome version hg38 based on RepeatMasker (http://www.repeatmasker.org/genomes/hg38/RepeatMasker-rm405-db20140131/hg38.fa.out.gz). Pipelines are represented with shapes and datasets are represented with colors.

To evaluate the overall performance of each pipeline and select the best pipeline, we calculated F1 score for insertions and deletions called by each pipeline in each dataset. F1 scores were comparable among all pipelines for a given dataset and SV type (i.e., insertion or deletion), but varied greatly among datasets and between insertion and deletion (Figure. 4.7, Table 4.3). The best pipeline varied depending on the dataset and the type of SVs. Out of the eight dataset-SV type combinations, NanoSVs and Sniffles each had the highest F1 score in four combinations. In contrast, LAST-Picky had the lowest F1 scores in six combinations.

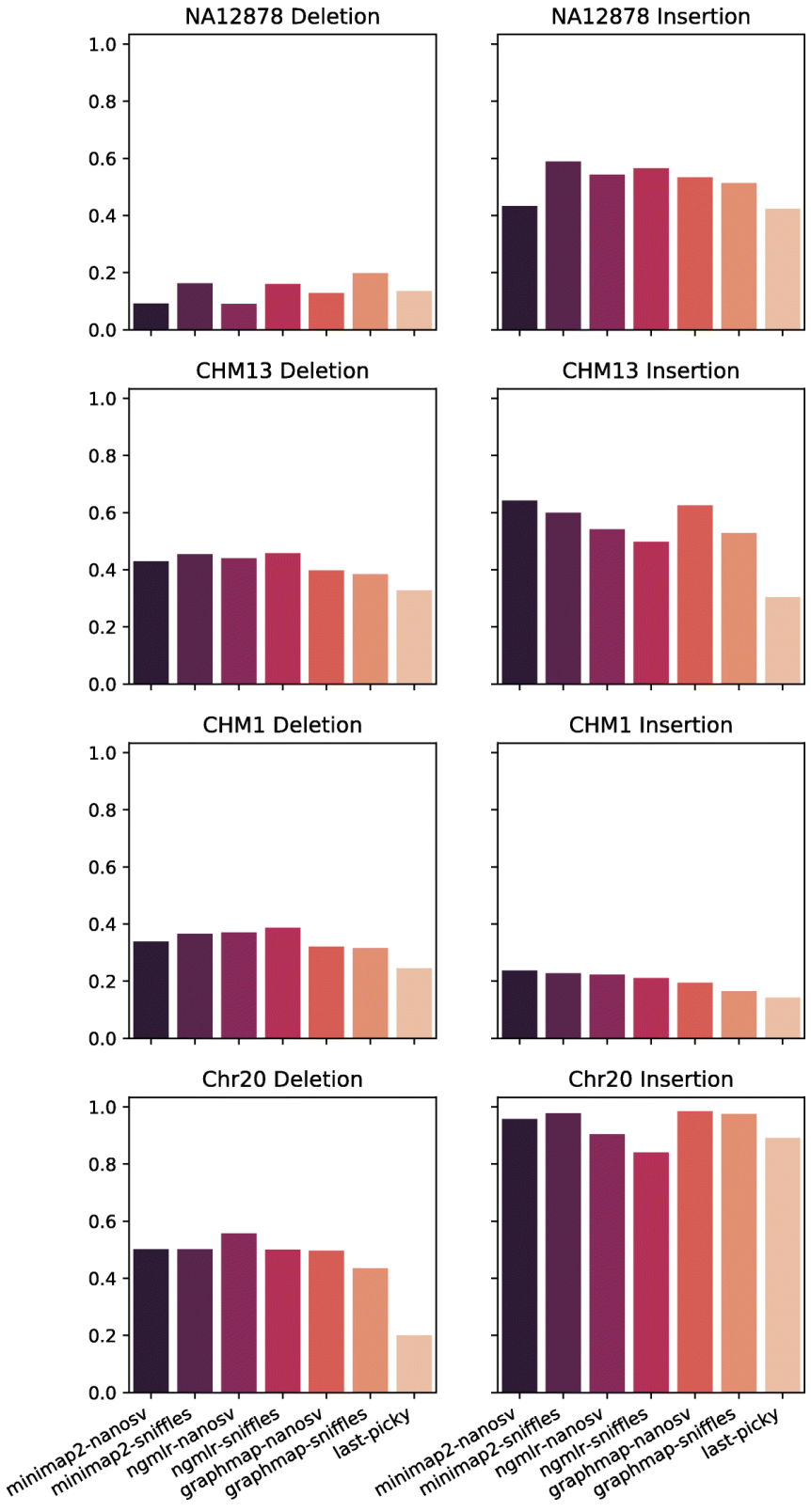


Figure 4.7. F1 scores for SV calling pipelines.
F1 scores for the seven pipelines are shown for insertion and deletion calls of each dataset.

To evaluate the impact of the sequencing depth on indel calls, we created subsets of each dataset by randomly selecting reads to achieve 50x, 40x, 30x, 20x or 10x sequencing coverages and calculated the F1 score of the Minimap2-Sniffles pipeline at different coverages (Figure. 4.8). In all datasets, F1 scores stayed relatively constant until 20x coverage and dropped dramatically at 10x coverage. One possible reason for the F1 score drop-off below 20x coverage could be that all SV callers apply a minimum number of supporting reads cut off (e.g., we used 10 for Sniffles and Picky) and other quality requirements. Therefore, the coverage close to or lower than the cut off would dramatically affect the performance of the callers.

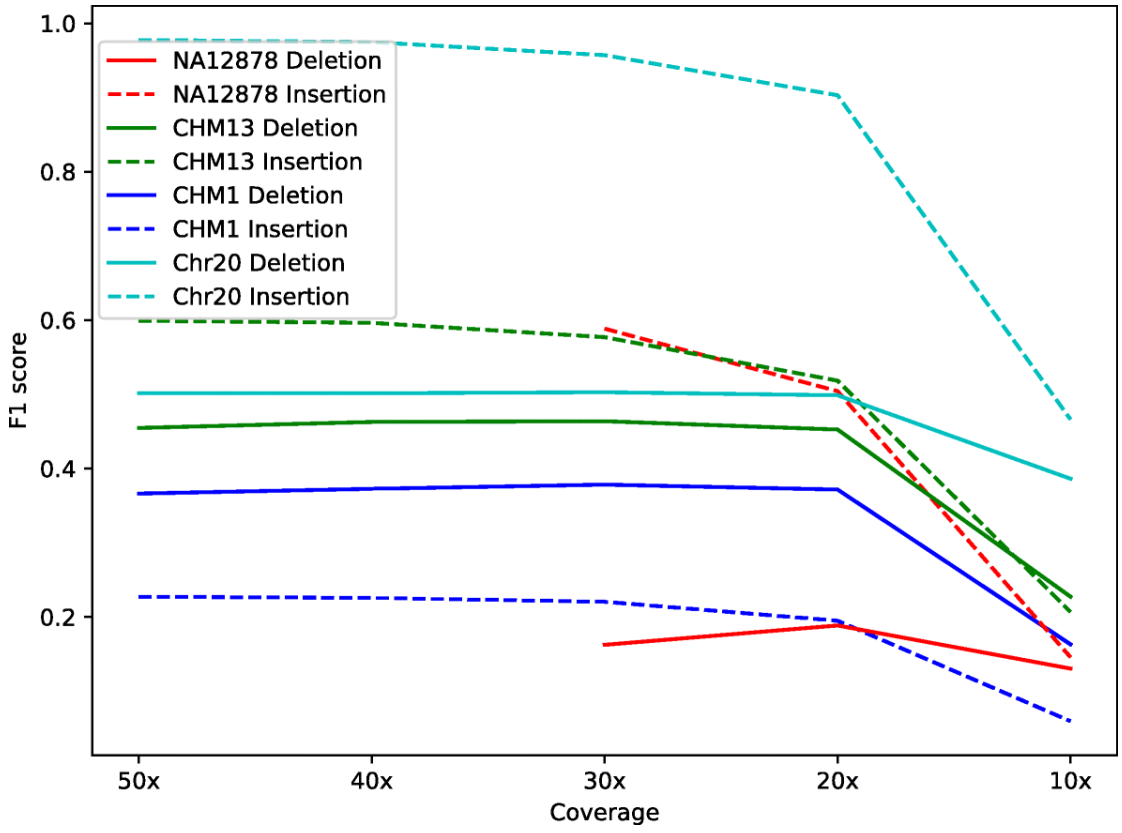


Figure 4.8. Impact of the sequencing coverage on the F1 score.
For each dataset, subsampling was performed on the original dataset to generate lower coverage datasets. Solid line, deletions; dashed line, insertions. The data is based on the Minimap2-Sniffles pipeline.

4.2.5. Consensus call set analysis and machine learning prediction

Next, we compared the SV calls among different pipelines. Overall, call sets from different pipelines each had many unique calls. As shown in the Venn diagrams of deletion calls in the NA12878 dataset, a large number of calls did not overlap between pipelines (Figure. 4.9). Even for pipelines using the same aligner or the same SV caller, the discrepancies remained large (Figure. 4.9).

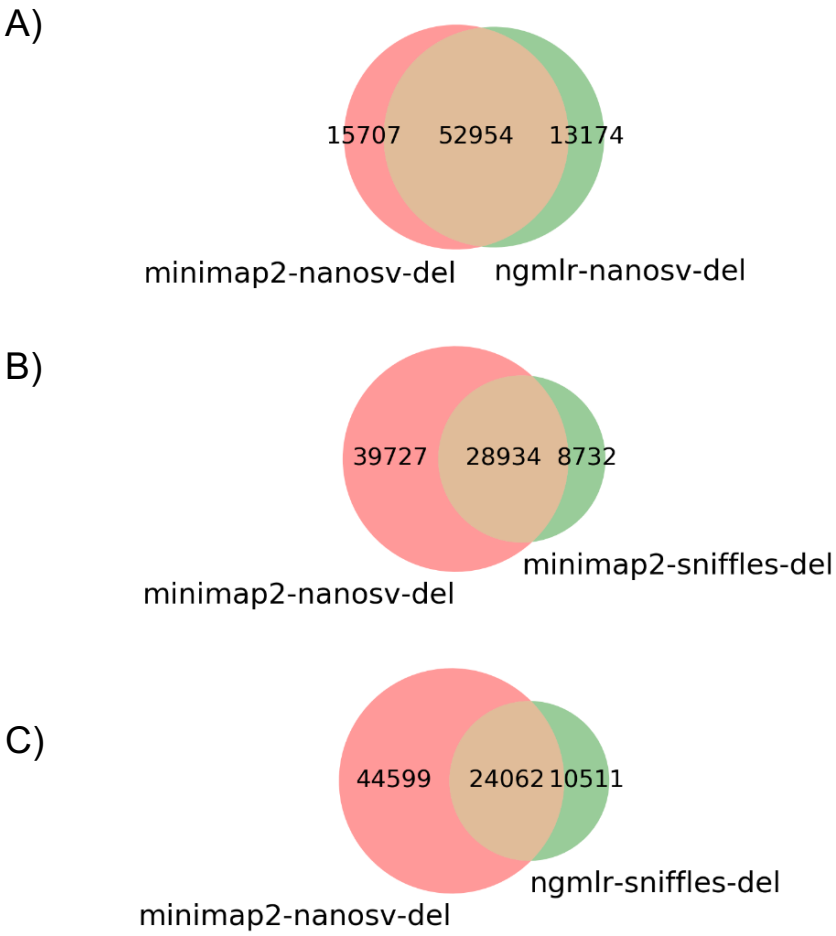


Figure 4.9. Overlapping SV calls between different pipelines.
A) same mapper different callers; B) different mappers same caller; C) different mappers different callers.

The large proportion of unique calls in each pipeline suggested there is a potential to improve SV calling quality by integrating calls from multiple pipelines. To evaluate the effect of integrating multiple call sets, we merged all call sets for each dataset, while tracking the number of call sets for each merged SV call. For deletions, requiring evidence from multiple pipelines improved the F1 scores of the call sets (Figure. 4.10A). The F1 scores for deletions in all four datasets reached a peak when requiring overlaps of six or seven pipelines. For insertions, applying the consensus pipeline filter also increased the F1 scores, and calls shared among two or three pipelines resulted in the best F1 scores (Figure. 4.10A).

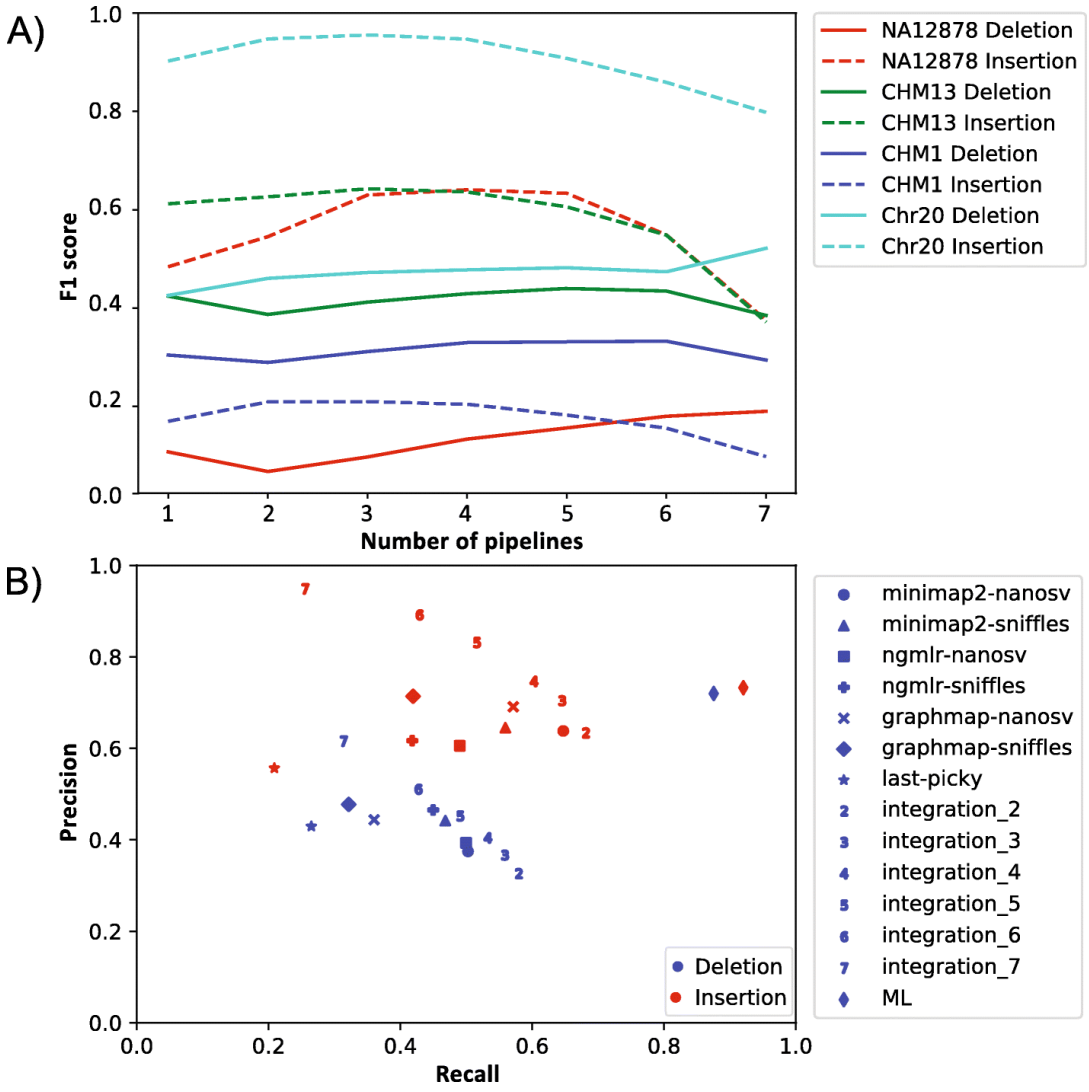


Figure 4.10. SV call set integration.
A) Consensus approach. Solid line, deletions; dashed line, insertions. F1 scores for insertions and deletions identified by a specified number of pipelines are shown. B) Precision-recall graph of call sets from SV calling pipelines and integration approaches. Blue, deletions; red, insertions. Results from individual pipelines and the machine learning approach are represented by shapes as specified in the legend. Results from the consensus approach are represented by numbers. For example, "2" represents the consensus call set from two callers.

Overall, selecting calls supported by multiple pipelines showed improvement of F1 scores but the improvement patterns were not consistent. Thus, we applied a more sophisticated call set integration approach by training a machining learning model based on the random forest algorithm. We selected seven SV features provided in the output of the SV callers, such as SV length,

number of supporting reads, mapping quality, and confidence interval of the breakpoint (Table 4.8). Using the CHM13 dataset as a test set, we achieved F1 scores of 0.79 for deletions and 0.81 for insertions, a substantial improvement over the best simple integration method (0.47 for deletion and 0.67 for insertion). Unlike the simple integration method, the machine learning approach was able to improve recall rate without sacrificing the precision (Fig. 4.10B). Among the seven features, the most important contributing feature was SV length, which accounted for ~50% of the evidence, followed by the depth P-value, read support, and mapping quality (Table 4.8). Similar to CHM13, the machine learning approach also produced improvement for most other data sets (Table 4.9). Because the depth P-value is only provided by NanoSV, while the read support was provided by Sniffles and Picky (Table 4.8), the machine learning approach allowed us to consider additional information provided by different callers to produce a high-confidence call set.

Feature	Description	SV caller			Contribution	
		Sniffles	NanoSV	Picky	Deletion (%)	Insertion (%)
SVLEN	Length of the SV	Yes	Yes	Yes	52	55
DEPTHPVAL	*P* value of the significance test of the depth of coverage at possible breakpoint junctions	No	Yes	No	20	15
RE	Read support	Yes	No	Yes	7	14
MAPQ	Median mapping quality of read pairs	Yes	Yes	No	10	8
CIEND	Confidence interval around the END position	No	Yes	Yes	3	3

CIPOS	Confidence interval around the POS position	No	Yes	Yes	2	2
PRECISE	Precise structural variant	Yes	Yes	Yes	5	1

Table 4.8. SV features and their contributions in the random forest classifier for CHM13. "Yes/No" under SV callers indicates whether a feature is provided by an SV caller.

Dataset	Accuracy	Recall	Precision	F1 score	Feature contribution						
					SVLEN	MAPQ	CIPOS	CIEND	PRECISE	RE	DEPTH PVAL
CHM13-Deletion	0.78	0.83	0.75	0.79	52.13%	10.08%	1.80%	2.91%	5.37%	7.31%	20.40%
CHM13-Insertion	0.74	0.92	0.73	0.82	55.40%	8.47%	2.42%	3.03%	1.32%	14.41%	14.95%
NA12878-Deletion	0.94	0.59	0.69	0.64	62.78%	1.89%	13.54%	6.98%	1.73%	1.84%	11.24%
NA12878-Insertion	0.69	0.48	0.64	0.55	41.90%	3.78%	20.23%	25.35%	3.26%	5.48%	0.00%
CHM1-Deletion	0.76	0.19	0.74	0.31	19.58%	4.78%	15.80%	6.78%	5.19%	12.65%	35.22%
CHM1-Insertion	0.75	0	N/A	N/A	39.96%	20.36%	6.12%	15.76%	3.51%	10.78%	3.51%
Chr20-Deletion	0.77	0.36	0.66	0.47	13.86%	4.07%	9.84%	9.09%	1.15%	13.40%	48.58%
Chr20-Insertion	0.98	1.00	0.98	0.99	11.79%	60.37%	3.75%	19.57%	0.00%	4.52%	0.00%
N/A: undefined metrics due to no predicted true samples.											

Table 4.9. Statistics of random forest classifier on all datasets.

4.3. Discussion

Improvements in our ability to detect and evaluate SVs in the genome is crucial to improve our understanding of the functional impact of SVs. While next-generation sequencing technologies have revolutionized genomics, their short

read-length has hindered the ability to reliably detect SVs. Recently, ONT released its nanopore-based sequencers that are capable of generating long reads, potentially improving our ability to detect SVs. Using public high-coverage nanopore sequencing data and simulated data, we evaluated multiple aligners and SV callers to assess SV identification performance using nanopore long-read sequencing data.

We benchmarked four aligners: an older and established aligner LAST and three more recently developed long-read aligners (minimap2, NGMLR, and GraphMap). Alignment time and memory usage varied widely between the four aligners while differences with respect to the mapped reads were moderate. Minimap2 was the fastest aligner tested with the most mapped bases. Therefore, we recommend minimap2 as a default aligner for general use. Unlike the newer aligners, which output the alignments in SAM (Sequence Alignment Map) format, LAST uses MAF (multiple alignment format) format. Although we tested converting the MAF format to SAM format, the resulted alignments are not fully compatible with SV callers expecting a SAM format input (data not shown). Therefore, we only evaluated the LAST-Picky pipeline.

The SV call sets differed dramatically among the pipelines, for both deletions and insertions. Unless the user is limited by specific requirements for SV calling, we recommend using minimap2 paired with Sniffles for the initial assessment of the data. This combination of tools showed the fastest processing time and a balanced overall performance in detecting both deletions and insertions. Our results is similar to a recent study on a different human sample (Wouter, et al., 2018). On the other hand, for a specific project, the choice of the pipeline could depend on the need of the user for either high recall rate or high

precision. Sniffles call sets showed the highest precision for most of the datasets tested, while NanoSV call sets generally had a higher recall rate, largely attributed to the higher number of SVs identified by NanoSV. Therefore, Sniffles should be used when high precision is the priority, while NanoSV should be considered if high sensitivity is desired and additional false positive calls can be tolerated.

All four datasets we used in this study have their own advantages and limitations for SV caller evaluation. For the Chr20 simulation dataset, we incorporated SVs based on the SV distribution from a real call set and used empirical error profile from an ONT sequencing run to simulate reads that resemble a true human sample. The advantage of such simulated dataset is that we know the true SVs that can be used to evaluate different pipelines. Nevertheless, the simulated reads are based solely on the chromosome 20 and are unlikely to capture the true heterogeneity of the entire human genome. This could in part explain the better performance of the Chr20 call sets compared to call sets from the other three datasets. For the NA12878, the CHM13, and the CHM1 genome, we evaluated our SV calls against high-coverage datasets (40-60X coverage) generated using the PacBio sequencing technology (Chaisson, et al., 2015; Zook, et al., 2019). These three datasets are among the a few available long-read datasets that attempt to produce high confidence SV calls by employing several different SV calling pipelines and the *de novo* assembly approach. Although SV calls in the three PacBio datasets are likely to have a high accuracy, these datasets are limited in several ways. For example, some of the benchmark datasets only include deletions and insertions, whereas SV callers we employed also generated other types of SV calls. In addition, these datasets are

based on the PacBio sequencing platform, which has its own limitations in terms of both sequencing technology and analysis tools. For example, one of the SV callers used to generate the benchmark, PBHoney (English, et al., 2014), is an older SV caller and it is not actively maintained at the moment. Indeed, the vast majority of NA12878 deletions that are called by all seven pipelines were absent from the SV true set. One such deletion region is chr1:117,029,131-117,029,278, for which minimap2 alignment shows multiple nanopore sequencing reads with evidence of a deletion, while the PacBio BLASR alignment showed only low-quality alignments in the region (i.e. with large number of mismatches) (Figure 4.11). Therefore, some of these SVs are likely to be real in the nanopore data but false negative in the benchmark set. As long-read sequencing technology matures, more comprehensive true SV call sets will become available and improve the evaluation. More importantly, experimental validation of some SV calls is necessary to empirically assess the accuracy of the calls.

Table 4.10. Nanopore and PacBio sequencing alignments comparison at an SV region. IGV plot of an NA12878 SV region (chr1:117,029,131 - 117,029,278, red box) that was identified by all seven pipelines but absent in the true set. Top: SV call diagram of the seven pipelines; middle: nanopore sequence minimap2 alignments; bottom: PacBio sequence alignment.

With the different datasets, we also assessed the impact of genome coverage on the SV identification among the SV callers. We sought to determine the minimum depth of coverage required to obtain a reasonable SV calling quality, given the limitation of budget and computational resources in research projects. For all three datasets, 20x coverage appeared to be the minimum coverage required to maintain the performance of the tools as judged by the F1 score. Given both the sequencing technology and the computational tools are under active development, we expect the coverage requirement will also be reduced in the future.

The SV calling results from the pipelines tested here showed that there is a room for improvement for the tools in terms of both recall and precision. In the meantime, one potential way to improve performance of the currently available

SV callers is to use an integrative approach and combine calls from multiple pipelines. We evaluated the integration principle using two approaches: one simple consensus approach, and one machine learning approach using the random forest algorithm that uses seven features from the SV caller outputs. Our results showed that both approaches can improve the F1 scores of the call sets. However, when combining the quality features provided by multiple call sets, the machine learning approach provided a much better overall performance compared to the simple consensus approach (Fig. 4.10B). This result suggests that when a true set is available for training, a machine-learning approach can be a good way to produce high-quality call set from multiple callers. In general, these results demonstrated the value of an integrative approach and further supported the need for the systematic evaluation and development of integrative approaches. Several SV integration tools with more sophisticated integration algorithm, such as MetaSV (Mohiyuddin, et al., 2015), svclassify (Parikh, et al., 2016), and Parliament (English, et al., 2015), have been developed for integrating SV calling results from multiple sequencing technologies and SV callers, including single molecule sequencing technologies. Similar algorithm can be applied to single-molecular sequencing SV callers and generate high-quality consensus SV call set.

4.4. Conclusion

Nanopore sequencing is a rapidly developing technology in terms of both sequencing technology and data analysis. For SV analysis, several new aligners and SV callers have been developed to leverage the long-read sequencing data. In addition, assembly-based approaches can also be used for SV identification. We have established a workflow for evaluating mappers and SV callers. We

found that SV callers' performance diverges between SV types. Therefore, our recommendations are tailored to the specific applications. For an initial analysis, we recommend minimap2 and Sniffles due to their high speed and relatively balanced performance calling both insertions and deletions. For more detailed analysis, we recommend running multiple tools and integrating their results for the best performance. When a high-quality true set can be defined, a machine learning approach, such as the one we proposed here, can be used to further improve the call set. Most analysis tools for nanopore sequencing are recently developed, and both accuracy and sensitivity can be improved. We expect resources from ONT and the nanopore sequencing community to accumulate as the technology improves and its user base grows. With more data being generated, better benchmark call sets will be available to more accurately assess the tool performance and facilitate future tool development.

4.5. Methods

4.5.1. Data set generation

The nanopore sequencing data of NA12878 in FASTQ format was obtained from the release 3 of the nanopore whole genome sequencing consortium repository (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/nanopore-human-genome/rel_3_4.md) (Jain, et al., 2018). The data was sequenced on the Oxford Nanopore MinION using 1D ligation kit. The SV call set for NA12878 was downloaded from (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz) (Zook, et al., 2019). This call set was based on the whole genome sequencing data of NA12878 at about 44x coverage using the PacBio

platform. The SV call set were generated using three SV detection methods, including a local assembly pipeline (Chaisson, et al., 2015). Only SV calls with a "PASS" flag in the "FILTER" field was included in the analysis. This dataset was lifted over from human reference genome GRCh37 to GRCh38 using liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

The CHM13 genome nanopore sequencing reads was downloaded from the release 2 of the nanopore whole genome sequencing consortium (https://s3.amazonaws.com/nanopore-human-wgs/chm13/nanopore/rel2/rel2.fastq.gz). The SV calls was obtained from dbVar (ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/vcf/nstd137.GRCh38.variant_call.vcf.gz).

The CHM1 genome assembly was downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/306/695/GCA_000306695.2_CHM1_1.1/GCA_000306695.2_CHM1_1.1_genomic.fna.gz). The nanopore sequence reads were simulated from the CHM1 assembly using NanoSim (ver 2.1.0) (Yang, et al., 2017). To generate a training dataset for nanopore sequencing read profile, DNA sample of the individual HuRef (Levy, et al., 2007) was purchased from Coriell (NS12911, Camden, NJ, USA). The HuRef sample was sequenced in our lab to about 1x coverage with an ONT MinION sequencer (Additional File 1: Supplemental Text: HuRef Sequencing). The sequencing reads were then used to generate the read profile by NanoSim *read_analysis.py* command (Yang, et al., 2017). Using the read profile and the CHM1 genome as the input, NanoSim *simulator.py* command simulated *in-silico* reads to about 50x target coverage (50,000,000 sequences) from the CHM1 genome. A high-quality

SV dataset for CHM1 was generated using the PacBio technology by the local-assembly approach (Chaisson, et al., 2015). This data was downloaded from (http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/data/GRCh37/insertions.bed, http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/data/GRCh37/deletions.bed). The dataset was lifted over from GRCh37 to GRCh38 using liftOver.

The R package RSVSim (ver. 1.24.0) (Bartenhagen and Dugas, 2013) was used to simulate deletions and insertions in the chromosome 20 of the human reference genome GRCh38. The number and size of each simulated SV were set to be identical to the NA12878 true set above (181 insertions and 96 deletions on the chromosome 20). NanoSim was used to simulate reads to about 50x target coverage (1,200,000 reads) based on the same read profile trained by the HuRef reads.

4.5.2. Read mapping and SV identification

The aligners and SV callers (Table 4.4) were downloaded and compiled on a High-Performance Computing (HPC) cluster based on the Ubuntu 14.04 system. Each node has 2 AMD Opteron 6272 2.1GHz 16-core processors and 256Gb RAM. The reads were mapped by the candidate aligners and SVs were called by the SV callers using outputs from each of the aligner when possible. The computational resource consumptions were recorded using GNU command "/usr/bin/time -v". The depth of coverage of an alignment file was calculated by SAMtools *depth* command (ver. 1.6) (Li, et al., 2009). The percentage of mapped reads, number of mapped bases, and mismatch rate of an alignment file were calculated by SAMtools' *stats* command (ver. 1.6).

Evaluation of insertions and deletion call sets for each of the three datasets was performed using BEDTools (ver. 2.27.1) (Quinlan and Hall, 2010). Deletions were compared with the true sets using BEDTools *intersect* command requiring at least 50% overlap between the two regions. Insertions were compared using BEDTools *window* command allowing 100 bps overlap upstream/downstream of the insertion positions. Precision rate, recall rate, and F1 score were calculated for each SV call set against their respective true set. Plots are generated using the matplotlib and seaborn library in Python3.

4.5.3. Call set filtering

For both true sets and call sets, several filtering and processing steps were performed to generate comparable datasets. First, SV calls from unincorporated contigs and the mitochondrial genome were filtered out to generate call sets for SVs on autosomes (chromosome 1-22), chromosome X, and chromosome Y. Next, insertions, duplications, and deletions were selected from all call sets. Insertion and duplication calls were combined as one category (referred as "insertions) for comparison. SVs were then filtered for size between 30 bps and 100,000 bps. The resulted SV calls were sorted using BEDTools *sort* command, and merged using BEDTools *merge* command.

4.5.4. Coverage analysis

Random subsampling of the FASTA files in each analysis was performed using the seqtk toolset (https://github.com/lh3/seqtk) based on the minimum number of reads needed to reach an expected coverage depth ranging from 10x to each dataset's original coverage, increasing by 10x each time. Subsampled reads at each coverage depth were mapped by minimap2, and the SVs were

called by Sniffles. The call sets were evaluated with the respective true set, and F1 score was calculated for each coverage depth in each comparison category.

4.5.5. Consensus call set

To generate a consensus call set for each dataset, call sets from all pipelines for each dataset were concatenated to a single file. BEDTools *merge* function (Quinlan and Hall, 2010) was then used to merge the concatenated calls into a consensus call set. The number of pipelines identified each consensus SV was stored. The consensus SVs were then filtered based on the number of pipelines that identified them, ranging from 2 to 7, and compared to their respective true sets.

4.5.6. Random forest classifier

SV calls from all seven pipelines for each pipeline were combined and labeled "True" or "False" based on whether they overlapped with the corresponding true set. The combined call set was randomly split into a training set (20% of the calls) and a testing set (80% of the calls) using the python package scikit-learn (v0.21.3, parameter "train_size=0.2"). The labeled SVs were learned and predicted by XGBoost (v0.90) random forest classifier (Chen and Guestrin, 2016) using the features selected from the "INFO" tag in the VCF files (Table 4.8). Precision and recall rate of the predictions were calculated by scikit-learn metrics.

5. Conclusion

In this era of transformation and innovation in computational genomics, we are rethinking how to organize workflows, applying statistics algorithms to solve complex problems, while introducing new sequencing technologies. Through my works on both method development and DNA sequencing analysis, I invented novel tools and workflows for genomic variants discovery and explored the role of genomic variants in neurodevelopmental disorders. My research was built upon the works from many others, and I hope I, in turn, made a meaningful contribution to this fast-evolving field of computational genomics.

6. Bibliography

Abecasis, G.R., *et al.* Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30(1):97-101.

Abel, H.J., *et al.* Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. 2018.

Abrahams, B.S., *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). In, *Mol Autism*. 2013. p. 36.

Abyzov, A., *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications* 2015;6(1):1-12.

Abyzov, A., *et al.* CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;21(6):974-984.

Addis, L., *et al.* Microdeletions of ELP4 Are Associated with Language Impairment, Autism Spectrum Disorder, and Mental Retardation. *Hum Mutat* 2015;36(9):842-850.

Auton, A., *et al.* A global reference for human genetic variation. *Nature* 2015;526(7571):68-74.

Bartenhagen, C. and Dugas, M. RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics* 2013;29(13):1679-1681.

Bartlett, C.W., *et al.* A major susceptibility locus for specific language impairment is located on 13q21. *Am J Hum Genet* 2002;71(1):45-55.

Bartlett, C.W., *et al.* A genome scan for loci shared by autism spectrum disorder and language impairment. *Am J Psychiatry* 2014;171(1):72-81.

Bayley, H. Nanopore sequencing: from imagination to reality. *Clinical chemistry* 2015;61(1):25-31.

Bond, S.R., *et al.* BuddySuite: Command-line toolkits for manipulating sequences, alignments, and phylogenetic trees. *Mol Biol Evol* 2017.

Bumgarner, R. DNA microarrays: Types, Applications and their future. *Curr Protoc Mol Biol* 2013;0 22:Unit-22 21.

Carvalho, C.M. and Lupski, J.R. Mechanisms underlying structural variant formation in genomic disorders. *Nature reviews. Genetics* 2016;17(4):224-238.

Chaisson, M.J. and Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *Bmc Bioinformatics* 2012;13.

Chaisson, M.J.P., *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 2015;517(7536):608-611.

Chaisson, M.J.P., *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019;10(1):1784.

Chang, Q., *et al.* Role of Microtubule-Associated Protein in Autism Spectrum Disorder. *Neurosci Bull* 2018;34(6):1119-1126.

Charrier, A., *et al.* Clock Genes and Altered Sleep-Wake Rhythms: Their Role in the Development of Psychiatric Disorders. *Int J Mol Sci* 2017;18(5).

Chen, K., *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* 2009;6(9):677-681.

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: ACM; 2016. p. 785-794.

Cretu Stancu, M., *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* 2017;8(1):1326.

Danecek, P., *et al.* The variant call format and VCFtools. In, *Bioinformatics*. 2011. p. 2156-2158.

Deelman, E., *et al.* Pegasus, a workflow management system for science automation. *Future Generation Computer Systems* 2015;46:17-35.

DePristo, M.A., *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43(5):491-498.

Dickinson, M.E., *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* 2016;537(7621):508-514.

Doherty, J.L. and Owen, M.J. Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice. *Genome Medicine* 2014;6.

Elia, J., *et al.* Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes (vol 15, pg 637, 2010). *Mol Psychiatry* 2010;15(11):1122-1122.

English, A.C., *et al.* Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC genomics* 2015;16:286.

English, A.C., Salerno, W.J. and Reid, J.G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* 2014;15(1):180.

Fan, X., *et al.* HySA: a Hybrid Structural variant Assembly approach using next-generation and single-molecule sequencing technologies. *Genome research* 2017;27(5):793-800.

Feuk, L., Carson, A.R. and Scherer, S.W. Structural variation in the human genome. *Nature Reviews Genetics* 2006;7(2):85-97.

Furlan, E., *et al.* Small population size and extremely low levels of genetic diversity in island populations of the platypus, Ornithorhynchus anatinus. *Ecol Evol* 2012;2(4):844-857.

Geoffroy, V., *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* 2018;34(20):3572-3574.

Ghirardi, L., *et al.* The familial co-aggregation of ASD and ADHD: a register-based cohort study. *Mol Psychiatry* 2018;23(2):257-262.

Goecks, J., *et al.* Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;11(8):R86.

Gong, L., *et al.* Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nature methods* 2018;15(6):455-460.

Gong, L., *et al.* Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat Methods* 2018;15(6):455-460.

Goodstadt, L. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics* 2010;26(21):2778-2779.

Goodwin, S., McPherson, J.D. and McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics* 2016;17(6):333-351.

Govil, M. and Vieland, V. Practical Considerations for Dividing Data into Subsets Prior to PPL Analysis. In, *Hum Hered*. 2008. p. 223-237.

Greene, C.S., *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* 2015;47(6):569-576.

Grube, S., *et al.* A CAG repeat polymorphism of KCNN3 predicts SK3 channel function and cognitive performance in schizophrenia. *EMBO Mol Med* 2011;3(6):309-319.

GTEx-Consortium. Genetic effects on gene expression across human tissues. *Nature* 2017;550(7675):204-213.

Ho, S.S., Urban, A.E. and Mills, R.E. Structural variation in the sequencing era. *Nature Reviews Genetics* 2019;21(3):171-189.

Hu, H., *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* 2013;37(6):622-634.

Hu, H., *et al.* A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nat Biotechnol* 2014;32(7):663-669.

Huddleston, J., *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research* 2017;27(5):677-685.

Hurd, P.J. and Nelson, C.J. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics* 2009;8(3):174-183.

Iakoucheva, L.M., Muotri, A.R. and Sebat, J. Getting to the Cores of Autism. *Cell* 2019;178(6):1287-1298.

Jackson, M., *et al.* The genetic basis of disease. *Essays Biochem* 2018;62(5):643-723.

Jain, C., *et al.* A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases. In.: Springer, Cham; 2017. p. 66-81.

Jain, M., *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 2018;36(4):338-345.

Jain, M., *et al.* The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016;17(1):239.

Johnson, M.H., *et al.* Annual research review: Infant development, autism, and ADHD--early pathways to emerging disorders. *J Child Psychol Psychiatry* 2015;56(3):228-247.

Kanehisa, M., *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 2020;45(D1).

Karczewski, K.J., *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. 2020.

Kiełbasa, S.M., *et al.* Adaptive seeds tame genomic sequence comparison. *Genome research* 2011;21(3):487-493.

Kircher, M. and Kelso, J. High-throughput DNA sequencing--concepts and limitations. *Bioessays* 2010;32(6):524-536.

Korbel, J.O., *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;318(5849):420-426.

Koscielny, G., *et al.* The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. In, *Nucleic Acids Res*. 2014. p. D802-809.

Koster, J. and Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 2012;28(19):2520-2522.

Kruglyak, L. and Nickerson, D.A. Variation is the spice of life. *Nature Genetics* 2001;27(3):234-236.

Leipzig, J. A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics* 2016;18(3):bbw020-bbw020.

Levy, S., *et al.* The diploid genome sequence of an individual human. *PLoS Biol* 2007;5(10):e254.

Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34(18):3094-3100.

Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Li, J., *et al.* Neurl4, a novel daughter centriole protein, prevents formation of ectopic microtubule organizing centres. *EMBO Rep* 2012;13(6):547-553.

Logue, M.W., *et al.* Bayesian analysis of a previously published genome screen for panic disorder reveals new and compelling evidence for linkage to chromosome 7. *Am J Med Genet B Neuropsychiatr Genet* 2003;121b(1):95-99.

Lu, H.C., *et al.* Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. *Nat Genet* 2017;49(4):527-536.

Mantere, T., Kersten, S. and Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet* 2019;10:426.

Martin, J., *et al.* Biological overlap of attention-deficit/hyperactivity disorder and autism spectrum disorder: evidence from copy number variants. *J Am Acad Child Adolesc Psychiatry* 2014;53(7):761-770 e726.

Metzker, M.L. Sequencing technologies — the next generation. *Nature Reviews Genetics* 2010;11(1):31-46.

Miao, H., *et al.* Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas* 2018;155:32.

Miller, J.A., *et al.* Neuropathological and transcriptomic characteristics of the aged brain. 2017.

Ming, X. and Walters, A.S. Autism spectrum disorders, attention deficit/hyperactivity disorder, and sleep disorders. *Curr Opin Pulm Med* 2009;15(6):578-584.

Mohiyuddin, M., *et al.* MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* 2015;31(16):2741-2744.

Myers, S.M., *et al.* Insufficient Evidence for "Autism-Specific" Genes. *Am J Hum Genet* 2020;106(5):587-595.

Nashta-ali, D., *et al.* Meta-aligner: long-read alignment based on genome statistics. *BMC Bioinformatics* 2017;18(1):126-126.

Ng, P.C. and Henikoff, S. SIFT: predicting amino acid changes that affect protein function. In, *Nucleic Acids Res*. 2003. p. 3812-3814.

O'Leary, N.A., *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44(D1):D733-745.

Oinn, T., *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 2004;20(17):3045-3054.

Paciorkowski, A., *et al.* Massive expansion of SCA2 with autonomic dysfunction, retinitis pigmentosa, and infantile spasms. In, *Neurology*. 2011. p. 1055-1060.

Parfitt, D.A., *et al.* The ataxia protein sacsin is a functional co-chaperone that protects against polyglutamine-expanded ataxin-1. *Hum Mol Genet* 2009;18(9):1556-1565.

Parikh, H., *et al.* svclassify: a method to establish benchmark structural variant calls. *BMC genomics* 2016;17:64.

Patterson, N., Price, A.L. and Reich, D. Population structure and eigenanalysis. *PLoS Genet* 2006;2(12):e190.

Pletscher-Frankild, S., *et al.* DISEASES: text mining and data integration of disease-gene associations. *Methods* 2015;74:83-89.

Polderman, T.J., *et al.* The co-occurrence of autistic and ADHD dimensions in adults: an etiological study in 17,770 twins. *Transl Psychiatry* 2014;4:e435.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841-842.

Rommelse, N.N., *et al.* Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder. *Eur Child Adolesc Psychiatry* 2010;19(3):281-295.

Ronald, A. and Hoekstra, R.A. Autism Spectrum Disorders and Autistic Traits: A Decade of New Twin Studies. *Am J Med Genet B* 2011;156b(3):255-274.

Ross, O.A., *et al.* Ataxin-2 repeat-length variation and neurodegeneration. *Hum Mol Genet* 2011;20(16):3207-3212.

Sanger, F. and Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 1975;94(3):441-448.

Satterstrom, F.K., *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* 2020;180(3):568-584.e523.

Schork, A.J., *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat Neurosci* 2019;22(3):353-361.

Sedlazeck, F.J., *et al.* Tools for annotation and comparison of structural variation. *F1000Research* 2017;6.

Sedlazeck, F.J., *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;15(6):461-468.

Shan, D.E., *et al.* Spinocerebellar ataxia type 2 presenting as familial levodopa-responsive parkinsonism. *Ann Neurol* 2001;50(6):812-815.

Song, L., *et al.* Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science (New York, N.Y.)* 1996;274(5294):1859-1866.

Sović, I., *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications* 2016;7:11307-11307.

Stankiewicz, P. and Lupski, J.R. Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine* 2010;61(1):437-455.

Steinberg, K.M., *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome research* 2014;24(12):2066-2076.

Stolerman, E.S., *et al.* Genetic variants in the KDM6B gene are associated with neurodevelopmental delays and dysmorphic features. *American journal of medical genetics. Part A* 2019;179(7):1276-1286.

Strauss, K.A., *et al.* A population-based study of KCNH7 p.Arg394His and bipolar spectrum disorder. *Hum Mol Genet* 2014;23(23):6395-6406.

Sudmant, P.H., *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;526(7571):75-81.

Szklarczyk, D., *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* 2020;47(D1).

Thomas, P.D., *et al.* PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. 2003.

Treangen, T.J. and Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics* 2011;13(1):36-46.

Ujike, H., *et al.* Association study of CAG repeats in the KCNN3 gene in Japanese patients with schizophrenia, schizoaffective disorder and bipolar disorder. *Psychiatry Res* 2001;101(3):203-207.

Van der Auwera, G.A., *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013. p. 11.10.11-11.10.33.

Vieland, V.J., *et al.* KELVIN: A Software Package for Rigorous Measurement of Statistical Evidence in Human Genetics. *Human Heredity* 2011;72(4):276-288.

Wang, S. and Xing, J. A Primer for Disease Gene Prioritization Using Next-Generation Sequencing Data. *Genomics Inform* 2013;11(4):191-199.

Wang, Y., *et al.* RseqFlow: workflows for RNA-Seq data analysis. *Bioinformatics* 2011;27(18):2598-2600.

Wang, Y., Yang, Q. and Wang, Z. The evolution of nanopore sequencing. *Frontiers in genetics* 2014;5:449-449.

Wouter, D.C., *et al.* Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *bioRxiv* 2018:434118.

Yang, C., *et al.* NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience* 2017;6(4):1-6.

Yang, L., *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* 2013;153(4):919-929.

Yang, Z., *et al.* Circadian-relevant genes are highly polymorphic in autism spectrum disorder patients. *Brain Dev* 2016;38(1):91-99.

Ye, K., *et al.* Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. In, *Bioinformatics*. 2009. p. 2865-2871.

Yoo, H.J., *et al.* Association between PTGS2 polymorphism and autism spectrum disorders in Korean trios. *Neurosci Res* 2008;62(1):66-69.

Zhang, L., *et al.* ADHDgene: a genetic database for attention deficit hyperactivity disorder. *Nucleic Acids Res* 2012;40(Database issue):D1003-1009.

Zook, J.M., *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 2019;37(5):561-566.