Pós-Graduação em Ciência da Computação

FRANCISCO DO NASCIMENTO JÚNIOR

SCREENVAR - A BICLUSTERING-BASED METHODOLOGY FOR EVALUATING STRUCTURAL VARIANTS

Federal University of Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE

2017

Francisco do Nascimento Júnior

SCREENVAR - A BICLUSTERING-BASED METHODOLOGY FOR EVALUATING STRUCTURAL VARIANTS

A Ph.D. Thesis presented to the Center for Informatics of Federal University of Pernambuco in partial fulfillment of the requirements for the degree of Philosophy Doctor in Computer Science.

Advisor: *Prof. Katia Silva Guimarães*

RECIFE
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

N244s Nascimento Júnior, Francisco do
ScreenVar: a biclustering-based methodology for evaluating structural variants / Francisco do Nascimento Júnior. – 2017.
99 f.: il., fig., tab.

Orientadora: Katia Silva Guimarães.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2017.
Inclui referências e apêndices.

1. Ciência da computação. 2. Biologia computacional. I. Guimarães, Katia Silva (orientadora). II. Título.

004 CDD (23. ed.) UFPE- MEI 2017-130

Francisco do Nascimento Junior

ScreenVar - A biclustering-based methodology for evaluating structural variants

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação

Aprovado em: 17/02/2017.

_____

Orientador: Profa. Dra. Katia Silva Guimarães

BANCA EXAMINADORA

_____

Prof. Dr. George Darmiton da Cunha Cavalcanti
Centro de Informática / UFPE

_____

Prof. Dr. Tsang Ing Ren
Centro de Informática / UFPE

_____

Prof. Dr. Guilherme Pimentel Telles
Instituto de Computação / UNICAMP

_____

Prof. Dr. José Fernando Garcia
Departamento de Apoio a Produção e Saúde Animal/UNESP

_____

Prof. Dr. Sergio Lifschitz
Departamento de Informática / PUC/RJ

I dedicate this thesis to all my family, friends and professors
who gave me the necessary support to get here.

Acknowledgements

"... I will guide you along the best pathway for your life. I will advise you and watch over you.".
- Psalm 32:8

First and foremost, praises and thanks to God for His blessings throughout my research work to complete this thesis successfully.

I would like to express my deep and sincere gratitude to my advisor Prof. Katia Guimaraes for the continuous support of my Ph.D study and research, for her patience, motivation, enthusiasm, pressures, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. In particular, I am grateful to her for enlightening me the first glance of Bioinformatics research since my undergraduate course.

Besides my advisor, I am extending my thanks to the members of my proposal thesis' board: Prof. Paulo Fonseca, Prof. George Cavalcanti, and Prof. José Fernando Garcia, for their time, encouragement, insightful comments, and hard questions presented on that time. And I also gratefully acknowledge the funding sources that made my Ph.D work possible. The Brazilian Educational Ministry fellowship supported my work during my first four years on the program.

I also thank my fellow doctoral students: Alixandre Santana, Glaucia Campos, and Eunice Palmeira, for the stimulating discussions, the sharing anxieties, all feedback, cooperation and of course friendship, besides all fun that we have had in the last five years, this was true glasses of water and fresh air. In addition, I would very much like to appreciate my precious polemic buddies for all time spent away from papers and experiments, but filled with movies, songs, homemade food, and many, many enjoyable quarrels.

Last but not the least, I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my brother and my sisters, aunts and uncles for their love, encouragement and continuing support to complete this research work.

Finally, my thanks to all the people who have supported me spiritually throughout writing this thesis and my life in general.

Francisco do Nascimento Júnior

Two roads diverged in a wood and I, I took the one less traveled by, and that has made all the difference.

—ROBERT FROST CITED IN DEAD POETRY SOCIETY BY PROFESSOR KEATING

Abstract

The importance of structural variants as a source of phenotypic variation has grown in recent years. At the same time, the number of tools that detect structural variations using Next-Generation Sequencing (NGS) has increased considerably with the dramatic drop in the cost of sequencing in last ten years. Then evaluating properly the detected structural variants has been featured prominently due to the uncertainty of such alterations, bringing important implications for researchers and clinicians on scrutinizing thoroughly the human genome. These trends have raised interest about careful procedures for assessing the outcomes from variant calling tools. Here, we characterize the relevant technical details of the detection of structural variants, which can affect the accuracy of detection methods and also we discuss the most important caveats related to the tool evaluation process. This study emphasizes common assumptions, a variety of possible limitations, and valuable insights extracted from the state-of-the-art in CNV (Copy Number Variation) detection tools. Among such points, a frequently mentioned and extremely important is the lack of a gold standard of structural variants, and its impact on the evaluation of existing detection tools. Next, this document describes a biclustering-based methodology to screen a collection of structural variants and provide a set of reliable events, based on a defined equivalence criterion, that is supported by different studies. Finally, we carry out experiments with the proposed methodology using as input data the Database of Genomic Variants (DGV). We found relevant groups of equivalent variants across different studies. In summary, this thesis shows that there is an alternative approach to solving the open problem of the lack of gold standard for evaluating structural variants.

Keywords: DNA Copy Number Variations. Variant detection methods. Next-generation sequencing. Biases analysis. Evaluation of variants

Resumo

A importância das variantes estruturais como fonte de variação fenotípica tem se proliferado nos últimos anos. Ao mesmo tempo, o número de ferramentas que detectam variações estruturais usando *Next-Generation Sequencing* (NGS) aumentou consideravelmente com a dramática queda no custo de seqüenciamento nos últimos dez anos. Neste cenário, avaliar corretamente as variantes estruturais detectadas tem recebido destaque proeminente devido à incerteza de tais alterações, trazendo implicações importantes para os pesquisadores e clínicos no exame minucioso do genoma humano. Essas tendências têm impulsionado o interesse em procedimentos criteriosos para avaliar os variantes identificados. Inicialmente, caracterizamos os detalhes técnicos relevantes em torno da detecção de variantes estruturais, os quais podem afetar a precisão. Além disso, apresentamos advertências fundamentais relacionadas ao processo de avaliação de uma ferramenta. Desta forma, este estudo enfatiza questões como suposições comuns à maioria das ferramentas, juntamente com limitações e vantagens extraídas do estado-da-arte em ferramentas de detecção de variantes estruturais. Entre esses pontos, há uma muito questão bastante citada que é a falta de um *gold standard* de variantes estruturais, e como sua ausência impacta na avaliação das ferramentas de detecção existentes. Em seguida, este documento descreve uma metodologia baseada em biclustering para pesquisar uma coleção de variantes estruturais e fornecer um conjunto de eventos confiáveis, com base em um critério de equivalência definido e apoiado por diferentes estudos. Finalmente, realizamos experimentos com essa metodologia usando o *Database of Genomic Variants (DGV)* como dados de entrada e encontramos grupos relevantes de variantes equivalentes em diferentes estudos. Desta forma, esta tese mostra que existe uma abordagem alternativa para o problema em aberto da falta de *gold standard* para avaliar variantes estruturais.

Palavras-chave: Variações no número de cópias. Métodos de detecção de variações estruturais. Sequenciamento de nova-geração. Avaliação de variações estruturais

List of Figures

List of Tables

List of Acronyms

aCGH	array Comparative Genome Hybridization
AS	Assembly approach
AUC	Area under ROC
BAF	B Allele Frequency
BAM	Binary Alignment/Map
CC	Cheng and Church algorithm
CNA	Copy Number Alteration
CV	Compound Variation
DGV	Database of Genomic Variants
DGV-GS	DGV Gold Standard
DNA	Deoxyribonucleic acid
GC	Genome Consortium
GWAS	Genome-wide Association Studies
MSR	Mean Squared Residue
PCR	Polymerase Chain Reaction
PEM	Paired-End Mapping approach
RD	Read-Depth approach
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SAM	Sequence Alignment/Map
SNP	Single Nucleotide Polymorphism
SR	Split-Read approach
SV	Structural Variants
SSV	Supporting Structural Variants
SVR	Structural Variant Region
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Contents

1

Introduction

This chapter starts with a general overview of the thesis and its context. The relevance of the thesis problem and the chosen solution approach are further motivated in Section 1.2. Based on this motivation, the thesis objectives are defined in Section 1.3 followed by a more detailed description of the research contributions and solution approach in Section 1.4. The chapter concludes with an outline of the structure of the thesis in Section 1.5.

1.1 General Overview

Bioinformatics is an exciting and rapidly expanding interdisciplinary research field that is attracting relevant attention from both academia and industry. A number of scientific publications, theses, and books in this area have increasingly demonstrated the capability of applying different computational techniques to solve challenging problems in bioinformatics as well as in ordering and summarizing a large body of knowledge. Sequence analysis, microarrays, gene expression, genome-wide analysis studies, gene regulation, phylogeny, and so on are subareas that have been resourced to relevant research questions, besides additional topics such as population genetics and personalized genomics.

The need for manipulating, analyzing, and visualizing the amount of data extracted from various biological systems brought computer science to this picture, in order to store and process such data, to build efficient algorithms that draw from areas such as Artificial Intelligence for recognizing patterns, and to use models to guide the research toward meaning results. The underlying idea has been to develop emerging computational methods to fulfill the myriad of demands from biology.

Over the years, biologists have learned how to analyze Deoxyribonucleic Acid (DNA) and convert millions of short DNA sequences into valuable genetic information. The DNA molecule is represented by the sequence of nucleotides (Adenine, Cytosine, Guanine and Thymine, labeled as A, C, G, and T, respectively), whose sizes can vary from a few thousand (viruses) to $6.7x10^{11}$ characters (ameba). The human genome, for instance, contains nearly 3 billion base pairs of genomic information organized into 23 chromosomes.

When the first draft of the human genome sequence was publicized in 2001, it was

openly claimed that all the differences among individuals should be attributed only to 0.1% of the genome. However, with the continuous improvement over the last 15 years in sequencing technologies and other bioinformatics skills, we now know that human genomes are highly variable. Since no two individuals share the same DNA code, investigating the similarity and respective variability of genomes among all human beings has evolved substantially over the past decade.

As a foundation for the study of genetic variation, a reference genome is considered to be some sort of representative genome of all possible genomes that an individual of that species could have. Thus, comparisons of any sequence to a reference genome have led to the identification of tens of millions of genetic variations, such as the collection of human genetic variations provided by the 1000 Genome Project (DURBIN et al., 2010).

With the completion of the initial reference human genome sequence some 17 years ago (INTERNATIONAL, 2004), attention has turned to discovering and cataloging variations among different individuals (case and control samples) and different populations. Any given individual carries 4-5 million sequence variants that are known to exist in multiple forms in our species. In addition, there are countless very rare variants, many of which probably exist in only a single or a few individuals. In fact, given the number of individuals in our species, essentially every base pair in the human genome is expected to vary in someone somewhere around the globe.

Currently, the efforts of many studies have aimed at the examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with phenotypes for a particular trait or disease. Such studies, known as Genome-Wide Association Studies (GWASs) have uncovered thousands of variants influencing major diseases and complex traits, including diabetes, dementia, cardiovascular disease, schizophrenia, breast cancer, height, and body mass index (BMI).

1.2 Motivation

The motivation of this thesis stems from the utmost importance of accuracy in analyses of biologic data closely related to personalized health care, including cancer prevention. The responsibility with these analyses lies in the existence of a continuous spectrum of the phenotypic effects of genetic variants, from adaptive traits to embryonic lethality, including morbid consequences such as developmental disorders and cancer (VALSESIA et al., 2013).

While a link between a given variant and a disease may have often been established, the relative contribution of such a variation to disease progression and the impact on drug response has yet been the object of a detailed assessment in several genome-wide analysis studies. Therefore, high levels of accuracy are the aim due to the employment of such analysis results for selecting appropriate therapies based specifically on the genetic context of a particular patient.

If, on the one hand, it is extremely important to have accurate genetic variants, on

the other hand, there is now a present uncertainty around the detection tools responsible for identifying such variations. The complete variant analysis process is complex, with multiple analysis steps, and is dependent on a variety of programs, databases, and input/output formats. Moreover, it also involves the need for handling large amounts of heterogeneous data. In the middle of such a complex context, a flood of tools for the identification of variants have been developed using different strategies and approaches.

Although this diversity of tools appears to be positive, it also brings large incongruities in the results due to the different detection tools, highlighting the need to separate the true positive events more accurately. To do this, it is necessary to include some validation criteria in order to assess such tools. However, this is the crucial problem: how to define a set of variants to work as a benchmark.

The lack of a gold standard of variants, the complex nature of biologic data, and the heterogeneity across platforms and methods make the acquisition of considerable levels of accuracy in detecting genetic variants a great challenge. In this light, many authors have investigated the concordance among structural variant detection tools and have revealed that there exist significant discrepancies between the overall variant sets called by each available variant-calling pipeline (O'RAWE et al., 2013).

Given that the importance of obtaining accurate and consistent variant calls for personalized medicine, there is a strong demand for highly accurate tools of Structural Variant (SV) identification. In general, the process of evaluating such tools intends to assess the distinction between genuine variants and random effects originating from sampling or sequencing errors. Thus, many authors have firmly stated that the lack of established benchmark data and tools is one of the largest challenging barriers for adequately evaluating detection tools. This draws attention to the need for a proper methodology for evaluating the accuracy of SVs returned by a given tool. Moreover, the relevance of this thesis is strongly corroborated by the proven effort to establish sufficient quality in available sets of genetic variants.

In addition, it is worthy to note that, from the perspective of engineering bioinformatics software, most tools, including SV detection tools, are results of development processes that posses gaps in fundamental actions (LAWLOR; WALSH, 2015), such as clearly defining the problem, the inability to reproduce the findings, the unreliability of findings, and the limitations of the data sample size, among other issues. Thus, the positive influence of thoroughly evaluated tools and data on the progress of data analysis research has been well justified. In general, the creation and widespread use of a benchmark within a research area are frequently accompanied by rapid technical progress and community building:

> *"Creating a benchmark requires a community to examine their understanding of the field, come to an agreement on what are the key problems, and encapsulate this knowledge in an evaluation. Using the benchmark results in a more rigorous examination of research contributions, and an overall improvement in the tools and techniques being developed. Throughout the benchmarking process, there is greater*

> *communication and collaboration among different researchers leading to a stronger consensus on the community's research goals."* (SIM; EASTERBROOK; HOLT, 2003)

This thesis concurs with these ideas. It is motivated by the belief that an established evaluation methodology and standard benchmarks for assessing different tools are needed for the relevant advancement of genetic variants research.

1.3 Thesis Objectives

The main objective of this thesis is the investigation of a new methodology for evaluating structural variants. Considering the lack of a real gold standard, as described previously, an important component of our study is to present a manner for alleviate this absence, as there are so many existing detection tools and other news that need assessment their outcomes.

Towards this aim, in order to answer the main research question, ***"How does one evaluate the accuracy of structural variants without a benchmark data?"***, this work visits the following key issues:

- Commonalities and differences in the methods designed for detecting structural variants.
- Low agreement among detection tools establishing the uncertainty of the variant calls.
- Most comparative studies pointing out the lack of a gold standard as a great barrier for evaluating SV detection tools.
- Data analysis using the most frequently used database of structural variants (DGV - Database of Genomic Variants).

While still considering other important questions to be answered, such as *what to evaluate, which criteria to use, how to measure those criteria, how to compare two variants*, and *how to achieve accuracy*, the thesis objectives focus on the two following general topics:

Objective 1: Develop a methodology for evaluating structural variants without using benchmark data

Objective 2: Release a set of structural variants provided by the proposed methodology to work as a benchmark.

1.4 Research Contributions

The main contributions of this thesis include the following:

- An overview of the problem of detecting structural variants using Next-Generation Sequencing data, with an emphasis on the important caveats in each stage.
- A list of relevant aspects used in comparative and analytic studies that can affect the specificity of variant detection tools; this was published in the *Transactions on Computational Biology and Bioinformatics Journal - June 2016.*
- A methodology for evaluating structural variants without using benchmark data.
- Results of DGV data analysis, in order to assess the usage of this database in a meta-analysis procedure.
- The release of a set of structural variants to be used as a benchmark for new variant calling tools.
- A list of studies in which accurate variants were published.

1.5 Thesis Organization

The thesis is organized as follow:

In chapter 2, we provide a brief overview of the primordial biological concepts inherent to the problem of detecting structural variants. Then we explain terms such as sequencing, reference genome, and genetic variation, and we also introduce popular databases and projects related to variant discovery. Moreover, this chapter includes information on the computer science part within the context of this thesis, especially the biclustering technique, which makes ScreenVar part of the introduced methodology.

In chapter 3, we enlarge the comprehension of structural variant detection through a state-of-the-art snapshot with respect to this research problem, and then through a brief survey of current methods for genetic variation detection using the most popular sequencing technology called Next-generation sequencing (NGS). Afterwards, with the investigation of several tools, an analytic perspective has revealed relevant aspects or caveats that are described in this chapter. Finally, we present a resulting list of features commonly observed during the assessment of variant detection tools.

In Chapter 4, we introduce the core contribution of this thesis, the ScreenVar methodology, which aims to provide a method for evaluating structural variants using biclustering algorithms. Then we describe each step of this methodology as well as the outcomes of several experiments performed with a consolidated database of genetic variants.

In chapter 5, we provide the most important points of discussion and contributions, along with the future works designed for this work. It is important to highlight the contents of the

appendix of this thesis, which comprises details of the ScreenVar implementation and many supplementary data obtained from the experiments.

2

Background

This chapter briefly introduces the fundamental concepts for this thesis, regarding underlying subjects present in a biclustering-based (Section 2.2.2) methodology (Section 2.3.2) for evaluating structural variants (Section 2.1.3). This thesis pursues a functional approach to evaluate such variations with respect to the accuracy. An in-depth understanding of the internal details of each part could thus be helpful, but not essential and neither applicable for this thesis. Therefore, the coverage of technical background of both areas will be kept rather short. Some points presented in this background will be complemented by each introduction section in the next two chapters (Sections 3.1 and 4.1).

2.1 Foundation in Biology

This section provides background information on the biological part of bioinformatics, highlighting important terms and concepts relevant to better understand the problem of structural variant detection.

2.1.1 Sequencing and mapping

By sequencing, we mean the process of determining the nucleotide order of a physical DNA fragment using a sequencing machine. Regardless of the approach to the genome as a whole, the process of DNA sequencing is the same. Sequencing employs a technique to separate pieces of DNA that differ from the others in length by only one base.

Three types of sequences play complementary roles in the cell: DNA sequences, RNA sequences, and protein sequences. DNA sequences are the basis of genetic material and act as the hereditary mechanism, providing the recipe for life. RNA sequences are derived from DNA sequences and play many roles in protein synthesis. Protein sequences carry out most essential processes such as tissue building, catalysis, oxygen transport, signaling, antibody defense, and transcription regulation.

An automatic sequencing machine outputs DNA sequences, called *raw sequences* by genome scientists. In *raw sequences*, the *reads* or short DNA sequences are all jumbled together, like pieces of a jigsaw puzzle in a just-opened box. Then, the *reads* are organized into larger

contigs during *assembly*. Each cell in an organism contains the same set of *chromosomes*, which are long DNA sequences. The set of chromosomes in an organism constitutes its *genome*.

Inevitably, raw sequences also contain a few gaps, mistakes, and ambiguities, making the task of assembly and mapping much more difficult. The *reference* sequence is a consensus sequence generated from a sample of donors, since such sequence does not accurately represent the set of genes of any single person. An example is the GRCh38, a human reference genome released on 24 December 2013, with roughly 3×10^9 base pairs.

Given that a reference has already been made available, the next step of a process of DNA sequencing, assembly and analysis relies heavily on mapping sequencing reads to a reference genome in order to find candidate positions of the reads. Even though it is a non-trivial task due to mutation and sequencing errors, it is very fast to look up all matching positions for those reads that are quite identical to the reference.

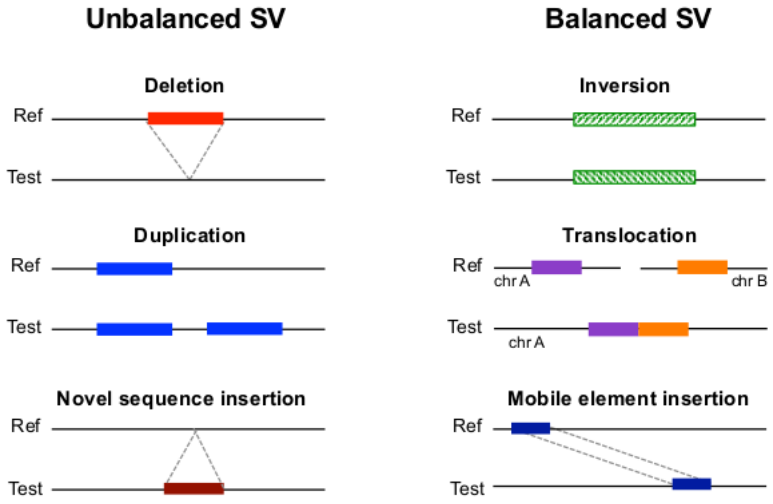
2.1.2 Human Genetic Variation

Genetic variation describes differences between the DNA sequences of individual genomes. Such differences (known as *variants*) occur in one or more individuals compared to the reference genome, or within individual cells — in the case of the study of cancer cells (case-control study), for example. Finding variants is the main goal in assembly with reference, and it is called *variant discovery* or *variant calling*.

Individuals differ from each other mostly because their DNA sequences differ, but genetic variation is not the only explanation for differences in phenotype (an observable characteristic). As a result of changes in DNA which are not corrected by repair systems, mutations are inevitable — i.e., mutation describes both a process that produces altered DNA sequences (either a change in the base sequence or in the number of copies of a specific DNA segment) and the outcome of that change (the altered DNA sequence).

The most common type of genetic variation is the single-nucleotide polymorphism (*SNP*, pronounced as "*snip*"), which is a change at a given position compared to the reference genome. Short insertions or deletions are referred to as *indels*, whereas the larger insertions and deletion are often referred to as *Copy Number Variations* (CNV), due to the alterations in the number of copies of chromosomal segments. Strictly speaking, indels should be considered to be copy number variants (a person with a gain or loss of copies instead of the normal two copies), but there is a convention to reserve the term *indels* to describe alteration up to an arbitrary 50 nucleotides. In addition, the term *Structural Variants* (SV) was assigned for large genomic rearrangements, including those involving alterations on the number of copies of genomic fragments, such as insertions and deletions (unbalanced structural variants), and those which do not, such as inversions and translocations (balanced structural variants) (Figure 2.1). Thus, unbalanced SV also includes CNV.

Figure 2.1: General classes of SV



The schematic illustrates deletions, duplications, and novel sequence insertions (unbalanced SV), and inversions, translocations, and mobile element sequence insertions (balanced SV) in a test genome (lower line) when compared with the reference genome (upper line).

2.1.3 Variants Calling

Variant calling or variant discovery is one of the key challenges in many areas of genomic research and diagnostics. Having aligned the reads of one or more individuals to a reference genome, SNP/indels/SV/CNV calling identifies variable sites, whereas genotype calling determines the genotype for each individual at each site. This is a stage of a complete data analysis process where the information sampled from both the population and the personal genomes is collected and evaluated in order to produce raw variants.

The detection of SV began more than 50 years ago with the development of the first cytogenetic techniques (karyotyping and fluorescence in situ hybridization (FISH)), with which observations of large chromosomal aberrations were observed microscopically in human cells. Then, in the middle of the 2000s, variant calling strategies achieved higher accuracy using array-based comparative genomic hybridization (arrayCGH) and SNP-array approaches. However, these approaches have suffered several inherent drawbacks, including low resolution and difficulty in detecting novel and rare mutations. With the posterior advent of microarray and sequencing technologies, the evolution of genome-wide methods for identifying all spectrum of SV resulted in significant enhancement of number, resolution and sensitivity of uncovered genetic variations.

Over the last years, Next-Generation Sequencing (NGS) has evolved into a popular strategy for genotyping and has included comprehensive characterization of variants by generating hundreds of millions short reads in a single run (METZKER, 2009). Besides keeping an inexpensive production of large volumes of sequence data, NGS also provides higher coverage and resolution, more accurate estimation of copy numbers, and more precision in detecting breakpoints. Taking these advantages into account, several variant calling tools have developed

based on different strategies for extracting features from NGS data. Along with such technological advances, the most prominent development is due to the Third Generation Sequencing (TGS) technologies, which have revolutionized genomics by enabling the sequencing of long, individual molecules of DNA and RNA.

Chapter 3 describes further details about the four abroad approaches for detecting structural variants following the emergence of NGS technologies, namely the Read Depth (RD), Paired-end Mapping (PEM), Split-read (SR) and Assembly-based (AS) methods. Moreover, such chapter includes relevant issues about overall procedure of variant calling analysis.

2.1.4 Genetic Mapping in Human Disease

The challenge of identifying genes and biological processes underlying any complex trait and diseases has motivated diverse linkage and association analyses over the years. While linkage studies use genetic mapping to identify loci associated with diseases by tracing transmission in families, genome-wide association studies (known as GWAS) use comparisons of frequencies of genetic variants among affected and unaffected individuals.

The last decade has seen rapid developments and breakthroughs in interpreting the phenotypic consequences of structural variation, especially due to the combination of the availability of full genome sequences and the growing number of GWAS. Such studies have uncovered thousands of variants influencing major diseases and complex human traits, including cardiovascular diseases, diabetes, dementia, schizophrenia, height, and body mass index (BMI).

2.1.5 Projects and catalogs

Before 2004, only a few dozen reasonably well-defined, non-disease associated, submicroscopic SV had been documented in the human genome. Since 2004, though, efforts have emerged to investigate, map, characterize and catalog SNP, indels and SV across the human genome. Through many projects around the world, a number of variants were discovered and cataloged in useful data sets, and many GWAS have used such collections in research for disease association. The following list shows some of the most relevant and commonly used projects and databases related to genetic variations.

- **1000 Genomes Project:** Started in 2007, the 1000 Genomes Project is one of the largest distributed data collection and analysis project, which aimed to sequence hundreds of human genotypes, at low coverage (4-6χ). The project stimulated the creation of a deep catalog of human genetic variation along with extensive methods to accurately discover and characterize the human variability using new sequencing technologies at an unprecedented scale and dramatically reduced cost. One of the most relevant results of this project was an integrated map of genetic variation from 1,092 individuals from 14 populations. This map consisted in 38 million SNP, 1.4 million *indels* and more than 14,000 larger deletions (ABECASIS et al.,

2012). Indeed, the 1000 Genomes Project has established as a research standard for population genetics and genomics, providing access to genotypes, sequences and genome mapping.

- **Database of Genomic Variants (DGV):** The explosion of data from diverse SV studies provoked the need of developing a public data archive. Thus, DGV was launched in 2004, comprising SV data from a few hundred individuals representing roughly 1,000 CNV and some inversions. Over more than a decade, the last version of DGV (at May 2016) has expanded to encompass information from 72 published studies with over than 6 million entries.

- **The Cancer Genome Atlas (TCGA):** This dataset was designed to catalog and discover major cancer-causing genome alterations in large cohorts of human tumors. Comprising more than two petabytes of publicly available data, TCGA has provided multi-dimensional maps of the key genomic changes in 33 types of cancer. Since June 2016, TCGA has been integrated to the Genomic Data Commons (GDC), which provides a unified data repository to enable data sharing across cancer genomic studies in support of precision medicine.

2.2 Concepts in Computer Science

This section provides background information on the computer science field of bioinformatics, showing important terms and concepts of data analysis and biclustering in order to better understand the underlying steps of the proposed methodology.

2.2.1 Data Analysis

The term *Data Analysis* has been used for quite a while, even before the advent of the computer era, as an extension of mathematical statistics, starting from development of cluster analysis and other multivariate techniques. The so-called *data avalanche* is created by the fact that there is no concise set of parameters that can fully describe a state of real-world complex systems studied nowadays by biologists, ecologists, sociologists, economists etc. On the other hand, powerful computers are able to produce, store, analyze, and visualize unlimited data sets through a wide spectrum of computational methods, including cluster analysis, pattern recognition, data mining, neural networks and so on.

A possible definition of data analysis is the process of computing various summaries and derived values from the given collection of data. Moreover, the process may become more intelligent if attempts are made to automate some of the reasoning of skilled data analysts and/or to utilize approaches developed in the *Artificial Intelligence* areas (BERTHOLD; HAND, 2003). Overall, the term *Data Analysis* is usually applied as an umbrella to cover all the various activities mentioned above, with an emphasis on mathematical statistics and its extensions.

Analysis of the data includes simple query and reporting functions, statistical analysis, more complex multidimensional analysis, and data mining (also known as knowledge discovery in databases, or KDD). Thus, among the diverse goals of data analysis, a significant interest regarding bioinformatics problems has been developed in cluster analysis, an important technique in exploratory data analysis, especially when there is no prior knowledge of the distribution of the observed data.

2.2.2 Biclustering

Biclustering consists in simultaneous partitioning of the set of objects and the set of their attributes into subsets. Assuming a given typical rectangular data matrix, it is a two-dimensional clustering in which rows correspond to objects and columns to attributes. For instance, in a practical example, if one wants to sell a laptop, one must take into consideration that one group of customers will be mainly interested in price, processor speed and screen size, while another group will be interested in dimensions, height, and design. This technique leads to finding homogeneous groups of objects, such as a subset of laptops suitable for one of two groups of customers.

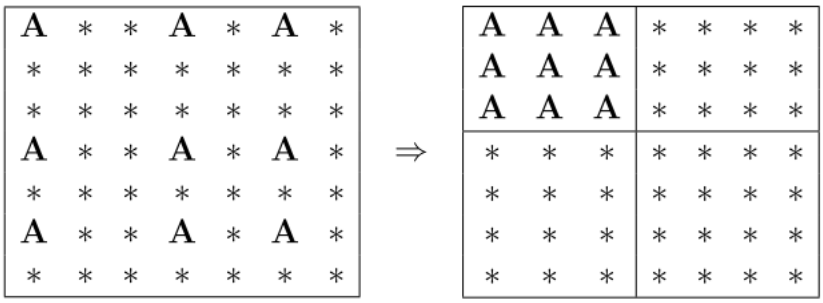


Figure 2.2: Biclustering finds objects and attributes with a similar value A and reports them as a bicluster (submatrix)

2.2.2.1 Biclustering Program Definition

Given a $m \times n$ data matrix A:

$$A_{m,n} = \begin{array}{c|cccc} & y_1 & y_2 & \cdots & y_n \\ \hline x_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ x_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_m & a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{array}$$

with objects X, attributes Y and entries a_{ij}. The goal of bicluster analysis is to find subgroups A_{IJ} of objects $I = i_1, ..., i_k, k \leq n, I \subset X$ which are as similar as possible to each other

on a subset of variables $J = j_1, \cdots, j_l, j \leq m, J \subset Y$ and as different as possible to the remaining objects and attributes. Bicluster z is then defined as $BC_z = (I_z, J_z) = A_{I_z,J_z}$.

A typical case to calculate bicluster is a high dimensional dataset with many variables, so that normal cluster algorithms lead to diffuse results due to many uncorrelated variables (KAISER, 2011). Also, biclustering is useful if there is an assumed connection of objects and some variables in the data set, that is, some objects have a certain similarity for a given set of variables.

In Bioinformatics, biclustering has many significant benefits:

- It can lead to a better understanding of the biological processes. Sets of genes regulated by the same transcription factor, namely module, can be detected using biclustering.
- Multi-functionality of the genes leads us to expect subsets of genes to be co-expressed only under certain conditions and to be uncorrelated under the rest of the conditions.
- Biclustering has a great potential in detecting marker genes that are associated with certain tissues or diseases. Thus, it may lead to the discovery of new therapeutic targets.

2.2.2.2 Overview of Existing Biclustering Algorithms

The earliest biclustering algorithm that can be found in the literature is the so-called *Direct Biclustering* by Hartigan (HARTIGAN, 1972) also known as *Block Clustering*, i.e., simultaneously clustering rows and columns of a matrix. This approach relies on statistical analysis of submatrices to form the biclusters. A bicluster is considered perfect if it has zero variance, so biclusters with lower variance are considered to be better than biclusters with higher variance. This, however, leads to an undesirable effect: single-row, single-column submatrices become ideal biclusters as their variance is zero. This issue is normally resolved by finding biclusters with other desirable properties, such as minimize variance in rows, variance in columns, or biclusters following certain patterns.

There exists a diverse set of biclustering tools that follow different strategies and algorithmic concepts. The most widely used and successful techniques and their related applications can be found in some relevant surveys on biclustering (PONTES; GIRÁLDEZ; AGUILAR-RUIZ, 2015) and (BUSYGIN; PROKOPYEV; PARDALOS, 2008). Based on the examination of the biclustering methods in such studies, we selected four of them to highlight details in this section, which are included in the proposed methodology.

- **Cheng and Church (CHENG; CHURCH, 1999):** The algorithm introduced by Cheng and Church aims to find biclusters with a minimum *Mean Squared Residue (MSR) score*. This value is equal to zero if all columns of the biclusters are equal to

each other (that would imply that all rows are equal too). Cheng and Church proved that the problem of finding the largest square bicluster with MSR score lower than a given limit is a NP-hard. Thus, they used a greedy procedure starting from the entire data matrix and successively removing columns or rows contributing most to MSR score. The brute-force deletion algorithm testing the deletion of each row and column would be still quite expensive in the sense of time complexity as it would require $O((m+n)mn)$ operations. However, the authors employed a simplified search for columns and rows to delete choosing a column with maximal.

- **Plaid Models (LAZZERONI; OWEN, 2002):** Introduced by Lazzaroni and Owen, Plaid Models is a statistical approach based on exploratory analysis of multivariate data, which assumes that the level of matrix entries is sum of the uniform backgrounds and k biclusters (a superposition of layers). Motivated for analysis of gene expression, several versions of the model are described in their work, being the most general the one in which allows a gene to be in more than one biclusters or in none at all. In short, Plaid Model is a form of overlapping two-sided clustering, with an embedded ANOVA in each layer.

- **Conserved gene expression motifs (or xMOTIFs) (MURALI; KASIF, 2003):** A *xMOTIF* is a subset of genes that is simultaneously conserved across a set of samples if it is in the same state in each of the samples in the subset. xMOTIFs is a probabilistic algorithm that exploits the mathematical structure of a xMOTIF to compute the largest xMOTIF. In order to identify several xMOTIFs in the data, an iterative strategy has been adopted, where samples satisfying each xMOTIF are removed from the data, and the new largest xMOTIF is searched. This process continues until all samples satisfy some xMOTIF. This search strategy allows gene overlap and also sample overlap, whenever any sample does not take part in more than one xMOTIF with the same gene.

- **Bimax algorithm (PRELIC et al., 2006):** Binary inclusion-maximal biclustering (Bimax) algorithm is a recursive divide-and-conquer approach proposed by Prelic. Its objective is extremely simple: it finds subgroups of 1 values in a binary matrix. In special, Bimax enumerates all inclusion-maximal biclusters, which are biclusters of all ones to which no row or column can be added without introducing zeros. By definition, this strategy only works with binary matrices, but clearly non-binary data can be converted to binary data in a number of ways.

2.3 Concepts of Accuracy and Precision

The terms precision and accuracy are frequently used inconsistently. Furthermore, the misconception that high precision implies high accuracy is almost universal. Accuracy refers to

the closeness of a measured value to a standard or known value, whereas precision refers to the closeness of two or more measurements to each other. For example, a wrist clock may measure time with a precision of one second. A stop watch may time your race with a precision of one hundredth of a second. However, if the clocks change and you forget to reset the wrist watch, then you have a very precise time but is not very accurate – you will be an hour early or late for all of your meetings.

The meaning and relationships between accuracy and precision can be clarified through the common metaphor of the target (Figure 2.3):

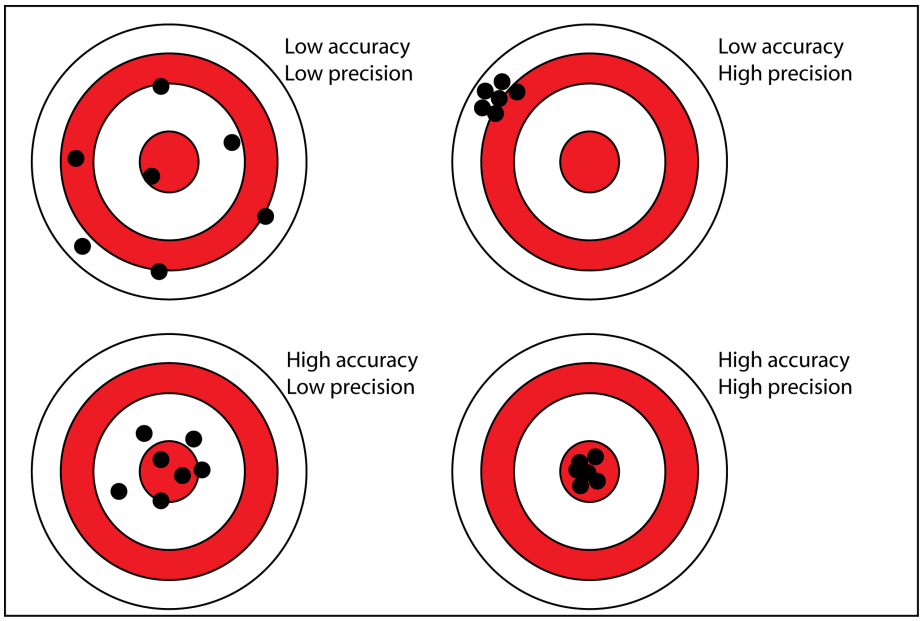


Figure 2.3: Accuracy vs Precision

- **Low accuracy & low precision:** Hits are spread across the target and consistently missing the centre.
- **High accuracy & low precision:** Hits are randomly spread across the target. On average, you get a valid group estimate, but you are inconsistent.
- **Low accuracy & high precision:** The target is hit consistently and systematically measuring the wrong value for all cases: it is consistent but wrong.
- **High accuracy & high precision:** You consistently hit the centre of the target.

2.3.1 Evaluation Methodologies

With respect to the term evaluation, we may present a definition adopted by the German Evaluation Society (DeGEval), as follows:

> *"Evaluation is the systematic investigation of an evaluand's worth or merit. Evaluands include programs, studies, products, schemes, services, organizations, policies, technologies, and research projects. The results, conclusions and recommendations shall derive from comprehensible, empirical qualitative and/or quantitative data."*

This extensive definition resulted from professional dialogues about the standardization of the use of evaluations in a diverse set of approaches, purposes and fields. From the viewpoint of scientific evaluations, some questions can direct the purpose of the evaluation. For instance, to determine the best technique to solve particular problems, to decide whether or how much new technologies improve over the state-of-the-art and finally to detect weaknesses and their causes to determine the problems to devote further research.

According to Kitchenham (1996), evaluation methods can be classified into quantitative, qualitative, and hybrid methods. *Benchmarking* belongs to hybrid methods, whose primary distinguishing feature is to carry out direct comparisons of alternatives. Formally, the same author shows the following definition about *Benchmarking*:

> *"Benchmarking is a process of running a number of standard tests/trials using a number of alternative tools/methods (usually tools) and assessing the relative performance of the tools in those tests".* (KITCHENHAM, 1996)

2.3.2 Evaluating bioinformatics tools

First, does scientific software development differ from other types of software? Creators of software widely used in computational biology discussed the factors that have contributed to their success:

> *"Scientific software often requires quite a strong insight – that is, algorithmic development. The algorithm implements novel ideas, is based on deep scientific understanding of data and the problem, and takes a step beyond what has been done previously. In contrast, a lot of commercial software is doing specific cases of fairly straightforward things – book-keeping and moving things around and so on."* (ALTSCHUL et al., 2013)

Previous work on scientific software evaluation has shown that numerical disagreement between programs of scientific computation grows at around the rate of 1% in average absolute difference per 4000 lines of implemented code and that the nature of this disagreement is non-random (HATTON; ROBERTS, 1994). Most recent scientific studies, especially in the area of bioinformatics and computational biology, deal with large and complex data sets and complicated algorithms. This complexity has made the replication of published findings difficult to pursue. In addition, not all users fully understand the intended usage and limitations of a scientific program. Errors or limitations of the computer code used could go undetected with possible negative

effects on future researches. Most importantly, there have been numerous published papers that attempt to train scientists to adopt best practices for scientific computing (WILSON et al., 2014).

Many scientists rely on the fact that the software has appeared in a peer-reviewed article, recommendations, and technical opinions, as their reason for adopting it. Nonetheless, one must not forget that there is a diversity of computational expertise within a development process, which can lead to misconceptions around all necessary steps to assure the quality of the produced tools. In other words, there is no guarantee that some tools put to use on the Internet are finished products.

Another relevant point is about biologists, bioinformaticians, statisticians, and researchers from other areas, who are not formally prepared to program software and might produce unreliable products due to the absence of specific skills to provide an adequate solution with satisfactory quality. Moreover, a complicating factor is that the reality of many of these published methods derive from thesis or were developed for research projects with time constraints, and, as a consequence of such limitation, are interrupted before the conclusion of a proper evaluation.

Due to this complex landscape, the need of applying a procedure for validating bioinformatics software is compulsory. Even though it is essential, evaluating such programs is not a trivial task as one would have imagined. It is often difficult, if not impossible, to define a gold standard mechanism to decide if the output of the target program is correct, given any possible input.

3

Unraveling variant detection problem in tangible aspects

In the previous chapter, important conceptual elements were presented to contextualize the variant detection problem. In this chapter, the focus shifts to introduce the current tools dealing the variant detection problem. In this direction ,fifty tools were identified in the literature and analyzed. Finally, issues regarding their analysis approaches, application and limitations are presented.

3.1 Introduction

The completion of the Human Genome Project has brought evidence that the DNA in the genomes of any two individuals is 99.9% identical, leaving the remaining 0.1% to be exploited in search of the source of all observed differences. With the whole sequence in hand, several computational tasks have emerged as challenges for understanding protein functions, mechanisms, and interactions. Eventually, the research focus turned to the quest for such variations and their roles, such as the increase of the risk for diseases, individual responses to medications and environmental factors, as well as phenotypic differences among individuals (height, eye color, hair color, and so on).

The initial effort in investigating human genetic variants only allowed the detection of changes on a microscopic level (no less than 3Mbp in size). The development of both experimental and computational strategies, which led to the availability of DNA sequencing technologies, allowed for the identification of genetic alterations at a nucleotide level. These alterations can be of different types and have been classified into three categories, based on their lengths: (1) SNPs(Single Nucleotide Polymorphisms), which are point mutations in the DNA, (2) Indels, which include insertions and deletions up to 50 bp in size, and (3) SV (Structural Variants), which include balanced and unbalanced events, such as long insertions and deletions, translocations, invertions, etc. (ALKAN; COE; EICHLER, 2011).

An important subgroup of the unbalanced structural variants is the copy number variations (CNV), alterations in the number of DNA segments, which are usually two, due to the underlying human evolution, diseases, or developmental disorders, leading the number of copies to become zero, one, three, or more. At least two mechanisms are responsible for these changes, such as the

process called non-allelic homologous recombination (NAHR) and microhomology-mediated events. (A discussion of these mechanisms is beyond the scope of this thesis; for further reading we suggest, for instance, the work of Hastings and Lupski (PJ Hastings, James R Lupski; IRA, 2010). CNV have been associated with neurological and neurocognitive disorders (GIRIRAJAN et al., 2013) (SEBAT et al., 2007) and with disease susceptibility (e.g. cancer, asthma, obesity) (FURUYA et al., 2015), but they can also be found in healthy individuals (ZHAO et al., 2013).

In 1968, Pepler and Smith reported, "It is now generally accepted that Down's Syndrome is due to the presence of extra genetic material of a chromosome in the G group" (PEPLER; SMITH; NIEKERK, 1968). This is only one of many evidences confirming that, for a long time, scientists have been interested in the pursuit of disease-causing genetic variations. Numerous studies have been carried out to investigate traits and complex diseases and, almost five decades later, we are still unravelling the associations between the human genome and diseases or phenotypes.

Before the advent of high-throughput sequencing (HTS) technologies, most methods for CNV detection were based on whole-genome array Comparative Genome Hybridization (aCGH), which utilized the relative frequencies of probe DNA segments between two genomes. Even with intense computational effort, hybridization-based approaches still have limited resolution (about 5-10 Mbp, for FISH (fluorescence in situ hybridization), and 10-25Kbp with 1 million probes, for aCGH) (YOON et al., 2009), being limited to short CNV detection.

Over the last few years, the newest sequencing technologies have brought revolutionary breakthroughs in areas such as the analysis of genomes through sequencing of unprecedented scale. The evolution of NGS has warranted a comprehensive characterization of CNV by generating hundreds of millions, even billions, of short reads in a single run. Hence, these advances have been responsible for numerous databases of short reads, and the resulting development of diverse tools for detecting these variants, especially for smaller SVs.

Knowledge of structural variations in the human genome has improved rapidly since many more public complete genome sequences have became available, as a result of the dramatic growth of sequencing capacity (next-generation and third-generation sequencers). Meanwhile, a number of sophisticated tools and associated pipelines for variant calling, annotation, and visualization have allowed the compilation of catalogs of human DNA variations shared in diverse databases (e.g. Database of Genomic Variants (DGV) (MACDONALD et al., 2014), dbSNP (SHERRY et al., 2001), dbVar (LAPPALAINEN et al., 2013)). A CNV map was published in February, 2015 (ZARREI et al., 2015), providing a human genome catalog of benign CNV among healthy individuals of various populations. This map was developed with data from DGV, which has collected and curated over 2 million CNV that were discovered from 55 studies.

Springing from large projects, such as HapMap (INTERNATIONAL; CONSORTIUM, 2005), 1000 Genomes Project (DURBIN et al., 2010)(TONEVA et al., 2012), and UK10K Project (SANGER, 2017), a myriad of genome-wide association studies (GWAS) have extended the list of somatic alterations in key genes uncovered in cancer studies. New findings on associations of

genetic variations were also published in regard to, for instance, susceptibility or progression of diabetes (ZANDA et al., 2014), Crohn's disease (PRESCOTT et al., 2010), and Parkinson's disease (PANKRATZ et al., 2011).

Regarding the identification of indels and CNV contained in DNA sequences, there is only a moderate agreement of findings among available softwares. Many strategies have been applied, using high-throughput sequencing data and their different data types, but detecting copy number variations is still a challenging problem. Given the importance of obtaining an accurate solution, comparative studies have been published pinpointing advantages and disadvantages in CNV calling tools. Table 3.1 lists some recently published comparison papers and surveys, with a brief word on each of them. The aim of these papers is to assist researchers in choosing the most suitable tool for their research needs. Most of the authors have stated firmly the lack of gold standard as a major challenging barrier for adequately evaluating the tools, along with the heterogeneity across sequencing platforms, computational techniques, and data formats available.

In this chapter, instead of discussing tools and their advantages or drawbacks, we focus strongly on the core features present in most CNV detection methods, addressing the strategies applied to overcome their weaknesses. This aims at providing a consistent bird's eye view with an analytical perspective covering the most popular tools. As a result, a list of eight highly relevant aspects or caveats of the CNV detection process emerges, singling out the most pressing questions in the realm of CNV analysis.

3.2 CNV Detection Tools using NGS data

An ideal CNV detection method from NGS data should accurately quantify the copy numbers of all genomic segments and define their boundaries across the whole or partial genome. Generally, such methods are incorporated into either current available pipelines or workflows (PABINGER et al., 2014), which involves integrated computational steps to execute data manipulation or analysis procedures, from raw sequences to biological meaningful results through annotated variants. In short, the main goals of these tools include to identify copy number states (gain, loss, normal) and copy number change points (a genomic location where there is a change of copy number state - breakpoints), taking a reference or target genome as the baseline for recognizing such variations.

Theoretically, detecting CNV from NGS data should be straightforward, once millions upon millions of sequence reads have been produced. However, in most cases, those reads measure only a few hundred bases, hence they rarely span a complete variant region of the genome. Moreover, sequencing is biased with respect to DNA content, which means that some regions amplify more efficiently than others. Depending on the amplification method, this could provoke unreliable quantitative results and significant differences in GC-content.

Table 3.1: List of recent comparative studies and reviews of CNV detection tools. The first three articles report results of comparative experiments.

SUMMARY	RELEVANT POINTS	REF.
Investigates statistical challenges in analyzing NGS data through a list of commonly used software.	*Dataset*: paired reads of individual NA12981 ($> 20\chi$) - from The 1000 Genome Project *Reference CNV*: from Conrad et al. CONRAD et al. (2012)	TEO et al. (2012)
Examines the abilities of algorithms designed for analysing read count data, comparing the performances of the six most widely used sequencing technologies, particularly for Read-Depth methods.	*Dataset*: sequencing data of five samples (NA19240, NA12878, NA11830, NA11840, NA12043) sequenced by three platforms at low coverage - from The 1000 Genome Project *Reference CNV*: from McCarroll et al. MCCARROLL et al. (2008)	MAGI et al. (2012)
Reports experiments performed among six CNV detection tools and their results, using synthetic and real data. Some comparative results are related to computation time/memory and levels of estimation of breakpoints or copy number.	*Dataset*: Simulated data (using reference genome hg18) and real data (chromosome 21 of the sample NA19240, at medium coverage). *Reference CNV*: DGV and outcomes from other tools	DUAN et al. (2013)
Shows an analytical study with 12 structural variants detection tools, discussing data type, required control, and SV types that are detectable by them.	NGS platforms produce elevated sequencing error rate. The number of ambiguous reads are increased by the short read sizes. Three factors can provoke an unbalanced number of mapped reads: GC-content, amplification errors, and non-uniform fragment distribution along the genome. Low coverage limits the sensitivity and specificity of the inference. For detecting insertion larger than the read length and insert size, one must opt for *de novo* assembly.	XI; KIM; PARK (2010)
A throughout review describing in details each one of the four approaches, also including a list of tools using each one.	Quality of sample preparation, library construction, sequencing instruments, and the CNV detection algorithm and its parameters can strongly influence the sensitivity and specificity of the outcomes. The high coverage variance in whole exome sequencing makes CNV estimation more challenging than whole-genome sequencing.	XI; LEE; PARK (2012)
Accesses the merits of CNV tools evaluated using both array-based data and NGS data. Includes important details about tools based on each platform, such as the size range of CNVs best detected and data type handled by all methods.	Improvements in read length will significantly impact mapping quality and *de novo* assembling. To sequence a genome at an adequate depth of coverage for reliable CNV calling (20X or greater) is considerably more expensive than using array-based platforms. Large-scale throughput is still not quite feasible, due to considerable hardware and software infrastructure required for processing all computational pipeline.	LI; OLIVIER (2013)
Shows a different study about CNV detection, providing an overview of tools used for cancer analysis, as well as a general workflow for somatic CNV detection tools. Many aspects and challenges are described, fostering the development of analytical tools for this kind of study.	Achieving a suitable accuracy for detecting CNV faced the extraordinary complexity of tumor genomes has challenged researchers to take full use of NGS data (BAF, read counts, discordant reads pairs) and to observe whether the process of NGS analysis needs specific adaptations for cancer studies.	LIU et al. (2013)
Reports results of a comparison study involving 48 tools. Challenges, strengths, and weaknesses are identified with the goal of assisting researchers in selecting the proper NGS tool. Among other insights, there are aspects about the usage of WES and WGS data.	The combination of different tools has been effective in improving CNV calling accuracy. Studies involving PEM tools combined with other PEM-based tools or with RD and AS approaches have demonstrated significant improvement in the statistical power of evidence of both CNV size and breakpoints.	ZHAO et al. (2013)
Surveys 50 software tools capable of detecting SVs in short-read WGS in order to evaluate the performance of these methods in deletion structural variant detection.	The results of the WGS simulation implied that PEM was the single best method for deteletion structural variant detection, albeit a combination of PEM, DOC and SRM methods was optimal.	NOLL et al. (2016)

3.2.1 NGS Data Analysis Workflow

NGS data analysis workflows for CNV detection includes multiple steps performed in general by a combination of vendor software, third-party tools, and custom scripts. Next, the overall workflow for NGS projects is grouped in four main tasks briefly described (Figure 3.1).

3.2.1.1 NGS Data Pre-Processing:

NGS data processing consists first in wet lab actions like image analysis, base calling, and sequence analysis. Through it all, it is known that key early decisions in library preparation

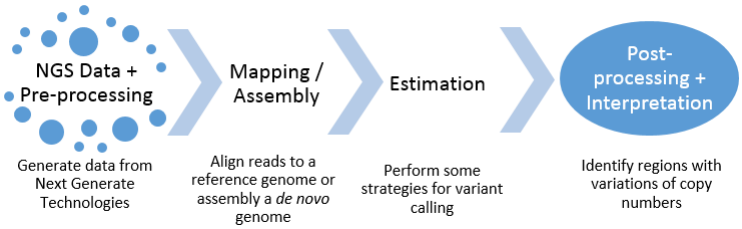


Figure 3.1: A general workflow of Copy Number Variation Detection Methods

Table 3.2: Tasks of a general workflow of Copy Number Detection Tools

TASK	COMMON ACTIONS	TYPICAL PRODUCT	MAIN CONCERNS
NGS Data Pre-Processing	Prepare library Call bases	FASTQ file	Quality of sequencing data Data type (WGS x WES) Cost-effectiveness (coverage, sampling size)
Mapping / Assembly	Align reads Apply quality procedures Recalibrate reads	BAM file	Ambiguity in mapping reads
Estimation	Identify breakpoints Identify copy number	VCF file	Low mapping quality Suitable detection tool Determination of cutoffs
Interpretation	Validate CNV callings Find associations with traits or diseases	Annotation	Lack of a gold standard Existing callings and annotations with inaccurate strategies

and sequencing technology can be strongly related to false positive rate in variant calling phase, which is discussed in detail in (QUAIL et al., 2012), a study of the three main NGS sequencing platforms (Ion Torrent, Illumina, and Roche).

Another concern involves the sequence data production, whether considering the quality and richness of data provided by Whole Genome Sequencing (WGS), despite the high cost for this production, instead of choosing Whole Exome Sequencing (WES), for its reduced cost and its increased popularity in clinical genetic studies. Along with the data type, important sequencing concepts have contributed to reach better accuracy, such as read length, single or paired end, sequencing coverage, and quality scores. Then filters for quality assessment evaluate the quality of the reads, in order to remove, trim, or correct, considering the base quality scores, when a large data set of reads are addressable in FASTQ file for the next step.

3.2.1.2 Mapping/Assembly:

Once the raw data is obtained from whole or exome sequencing, the computational intensive step of read mapping is performed. Considered as the fundamental and most costly step of this workflow, the mapping of the set of reads against a target genome is done by means of short read aligners, such as BWA (LI; RUAN; DURBIN, 2008) and Bowtie2 (LANGMEAD; SALZBERG, 2012). Alternatively, when the reference genome is unknown a priori, a *de novo* assembly is attempted through overlapping sequence reads, creating consensus sequences and eventually the entire genome. The coverage quality of a *de novo* genome assembly depends on the size and continuity of the contigs (basically, the number of gaps). This procedure requires an

efficient overlapping process, which demands higher computational power than mapping to a known target sequence. During the alignment, quality check procedures are normally applied, such as soft clipping of low-quality bases, retaining only uniquely mapped reads, and removal of potential PCR duplicates. The alignment is also refined by locally realigning any suspicious reads, including known indels, and base quality scores of realigned BAM files are recalibrated. Only then, the realigned reads are written in the typical product of this phase, a Sequence Alignment/Map (SAM) or its binary version (BAM file).

3.2.1.3 Estimation:

Once the mapped reads are available, a chosen algorithm is used to identify regions with any structural variation. Variant callers analyze BAM files to discover all sites with statistical evidence of occurrence of alternate allele. In order to achieve a higher accuracy in variant calling, some methods perform trimming and correcting tasks, like the removal of duplicated sequences and the removal/flagging of sequences with low mapping quality (MQ). In this stage, an accurate segmentation is essential to estimate the copy number of the segment. Given many aspects that cover the detection of CNV, the choice of suitable tools to be used in this step have to take into account the main requirements of the analysis. Abilities, such as dealing with germline and somatic samples, or prediction of the exact number of DNA copy, for instance, could determine the choice of one specific tool for this step. Finally, the set of output results is provided in a variant call format (VCF) file.

3.2.1.4 Post-processing/Interpretation:

In this step, genuine CNV must be distinguished from random effects originated from sampling or sequencing errors. Since it would be necessary some kind of gold standard to confirm the reliability of the CNV, assessing estimated CNV is currently an extremely challenging task. One common strategy is to verify the concordance with other results. Meanwhile, according to a reciprocal overlap criterion, some procedures perform a merging of the adjacent regions with identical copy number into one single segment, and divide regions with different copy numbers into different segments. Then, the final adjusted file is produced containing all remaining copy number variations.

3.2.2 Methods for CNV detection using NGS data

The potential for applications provided by NGS technologies has been demonstrated through several issues that can be more closely addressed in genome sequencing and functional genome research. Diverse relevant properties extracted from NGS data, such as quantity and length of reads produced, read counts, B Allele Frequency (BAF), soft-clipped reads, and discordant read pairs, became a rich source of information which is exploited for detecting CNV. Different approaches were developed focusing on the exploitation of one or a few of those

properties. The most widely used sequence-based approaches are: (1) Paired-End Mapping (PEM), (2) Read-Depth (RD) (or Depth of Coverage), (3) Split-Read (SR), (4) Assembly-based (AS), and (5) Combined-based (CB).

3.2.2.1 Paired-End Mapping (PEM) Approach

This method analyzes anomalies in the separation lengths or orientation of aligned read pairs. The overall strategy of Paired-end Mapping methods for CNV detection is to align the ends of the fragments to a reference genome and fully utilize the respective mate-pair information. Paired-ends that disagree in length or orientation indicate possible insertions, deletions, or inversions. Since 2005, when the first study to implement a paired-end sequencing approach was demonstrated by Tuzun and colleagues (TUZUN et al., 2005), many computational tools based on this approach have been released, including PEMer (KORBEL et al., 2009), VariationHunter (HORMOZDIARI et al., 2010), BreakDancer (CHEN et al., 2013), and so on, which were adopted for SV mapping in personal genomics endeavors, such as the 1000 Genomes Project. A known drawback shared by all methods from this category is that only an approximate resolution of the structural variants is possible to find, once this approach depends on the insert size to provide precise boundaries for identified variants.

3.2.2.2 Read-depth (RD) Approach

The principle of this approach consists in aligning the sampled reads to the reference genome, then piling up the aligned reads, and using the density of these alignments to calculate the read counts across sliding windows (or bins), resulting in the so-called RD-signal. The basic idea is that the read density of a given genomic region should be correlated to the copy number of that region. Genomic regions with disproportionate read counts indicate potential CNV. An assumption of data distribution is needed to model this signal and, therefore, to investigate the presence of variations. Typically, it is assumed that the reads are generated randomly, following a Poisson or modified Poisson distribution.

In 2009, a study by Chiang and collaborators was the first to adopt this approach, leading to the development of a tool called SegSeq (CHIANG et al., 2009) to define rearrangements in cancer. After that, several methods have been released, not only using the RD approach, but also combining RD and PEM on a single solution. RDXplorer/EWT (YOON et al., 2009), mrFast (ALKAN et al., 2010), CNVnator (ABYZOV et al., 2011), ReadDepth (MILLER et al., 2011), and Control-FREEC (BOEVA et al., 2011) are among the most commonly cited methods in the literature and in scientific forums.

Although the initial application for this strategy was related to tumor studies, read-depth is the only sequencing-based method to accurately identify absolute copy-number in genomes (as opposite to only infer gain/loss), even though with poor breakpoint resolution in general (ALKAN; COE; EICHLER, 2011). In addition, read-depth tools have outnumbered other

approaches (see Table A.1), due to the use of a number of different strategies commonly applied to model signal behavior. Table 3.3 points to important particularities which characterize how this subset of tools deals with data preprocessing and segmentation.

3.2.2.3 Split-read (SR) Approach

The strategy behind this approach is to detect paired reads is to apply longer sequencing reads to define the breakpoints of structural variants, based on occurrences of gapped read alignments. Gaps observed are considered as potential breakpoints of a SV, thus multiple parts of the same read characterized by different anchor points could be mapped to different SVs of the reference genome.

The focus of this strategy is to detect read pairs in which exactly one read is uniquely mapped to the reference sequence, while the other read failed to be aligned. The assumption is that the second paired read could not be mapped, even with few mismatches allowed, because it corresponds to a deletion or insertion breakpoint. The mapped read is used as an anchor and knowing both a maximum event length and the direction to search for the unmapped read; alignment of the unmapped read can be performed either by splitting it into two or three fragments.

SR methods are more sensitive when used with NGS technologies that produce short reads, since they are harder to be uniquely aligned. These methods can reach base-pair resolution in detecting small insertions and deletions. Nonetheless, the length of the variation event must be smaller than the length of a read. Third generation sequencing, with prominent longer reads, will potentially improve the method performance. Some computational tools have been developed using the SR approach, including Pindel (YE et al., 2009), SLOPE (ABEL et al., 2010), AGE (ABYZOV; GERSTEIN, 2011), and SRiC (ZHANG et al., 2011).

3.2.2.4 Sequence Assembly Approach

This approach is perhaps the most natural, since it builds on the well-known *de novo* assemblers. Theoretically, a complete genome sequencing followed by *de novo* assembly and comparison to a high-quality reference genome, could detect all structural variants, including CNV. In practice, accurate assembling is still a challenging problem, due to the genome complexity, the length and error rate of the produced reads, etc. Although *de novo* assembly of an entire human genome based on NGS data is a problem far from exhausted and demands significant computational resources, some authors have proved that it is possible to use *de novo* assembly to identify SVs. Some examples of tools based on this approach are Velvet (ZERBINO; BIRNEY, 2008), Cortex Assembler (IQBAL et al., 2012), Magnolya (NIJKAMP et al., 2012), and TIGRA (CHEN et al., 2014).

3.2.2.5 Combined Approach

Considering the specific functionality of each approach, some tools have used a combination of methods in the attempt to improve efficiency. For instance, in 2010, a tool provided by Medvedev and collaborators, CNVer, offered a combination of two known approaches, depth of coverage and paired-end mapping (MEDVEDEV et al., 2010). The common problem of non-uniqueness of the reads on RD-based methods could be mitigated by using the information of discordant mappings from paired-end mapping. Another example of the combined approach is GenomeStrip (HANDSAKER et al., 2011), a complex tool based on all three possible sources of CNV information handled by previous approaches: read-pairs, split-reads, and read-depth. It has attracted much attention recently in forums and comparative studies, which is also due to multiple operations in discovery, refinement, and genotyping. This tool was originally conceived to support the 1000 Genomes Project, and it is currently maintained by the Broad Institute (Broad Institute, 2017).

There is currently many open-source/freely-available tools for accurately estimating DNA copy number variations in NGS data, using different strategies, statistic models, programming languages, input/output format, and so on. This heterogeneity is further enhanced by new challenges as, for instance, the advent of increasingly longer reads. Table A.1 shows a panel to illustrate this variety, featuring some of the most popular tools in use, and Table 3.3 shows major features of the read-depth based methods.

Table 3.3: Major features of most popular tools based on Read-Depth

Tool	Data preprocessing	Segmentation
CBSBR	Multiple sequencing data Negative Binomial Distribution Penalized least square regression	Continuation block-wise single best replacement (extended CBS)
cn.MOPS	Quality control Differently sized windows Sample normalization GC-Correction	Mixture of Poissons (to separate signal from noise) Circular Binary Segmentation (CBS) or DNACopy
CNAnorm	Mapping quality Window size tuned for data available GC-Correction Genome-wide normalization Contamination correction	Smooth segmentation DNACopy
CNASeg	Matched control GC-correction Smoothed RD signal (Wavelet Transform) Mapping quality filtering	HMM Segment merging based on statistical testing
CNV-Seq	Random sampling assumption Matched control	Statistical testing (multiple) Adaptable size window Log-Ratio based
CNVeM	Clustering of discordant reads GC-Correction Mappability: all possible mapping positions	Mean-shift approach Expectation-Maximization (EM)
CNVer	Random sampling assumption GC-Correction Reduce sequencing bias with discordant read pairs	Expectation Maximum Minimum cost flow
CNVnator	Do not require matched control Equally sized windows GC-Correction	Mean-shift approach
Control-FREEC	Matched control (optional) GC-correction Mappability correction	LASSO algorithm Empirical cutoff
JointSLM	Random sampling assumption Multiple sampling Matched control GC-correction	Shifting Level Model (SLM) extended / HMM Fixed window
RDXplorer	Random sampling assumption Matched control (optional) GC-correction Fixed window Mapping quality filtering	Significance testing (Event-wise testing)
ReadDepth	Negative-binomial distribution assumption GC-correction / LOESS regression Mappability correction Discordant read pairs	Circular Binary Segmentation (CBS) Adaptable size window Optimized cutoffs
SegSeq	Matched control	Circular Binary Segmentation (CBS) Optimized cutoffs

3.3 Tangible issues/aspects of CNV Detection Tools

This section discusses issues commonly concerned in any CNV detection tool, related to the effects of using NGS data and the importance of technical decisions taken on the design of these tools, which may strongly affect their experiments and outcomes. Since many frequent new versions have been released, it is extremely important to recognize which questions and answers have motivated or supported diverse improvements in the most recent methods. The aspects chosen for discussion are the following.

1. Handling NGS data: error rate and coverage
2. Choosing a data type for sampling
3. Using a suitable fragment distribution
4. Correcting GC-Content bias
5. Correcting mappability bias
6. Performing a segmentation algorithm
7. Adapting to the advent of longer reads
8. Evaluating in the absence of a gold standard

3.3.1 Handling NGS data: error rate and coverage

Before discussing merits and demerits from methods properly, it is worth examining the concerns related to the usage of NGS data for discovering structural variants, since this technology has displaced consolidated experimental approaches, including aCGH and SNP array, and the traditional Sanger sequencing. Regarding the latter, although NGS platforms generate considerably shorter sequence reads than Sanger technology, it is much faster, easier to operate, and less expensive. These massively parallel platforms normally produce huge volumes of reads (millions or even billions) at greater coverage depths than Sanger, because of the short size of their reads (average read length from 50bp to 350bp, depending on the platform). Sanger technologies can generate reads of 1Kbp or more, sequencing only a few thousands nucleotides in a week, while NGS technologies allow the sequencing of a whole genome of an individual within a couple of hours. An immediate consequence is the impact in handling the magnitude of data, then requiring substantial extra investment in computational resources. Apart from these challenges, the nature of the human genome, with its complex structure, brings other implications, due to large quantities (over 50% of DNA sequence) of regions rich in repeats and segmental duplications.

3.3.1.1 Error rate

Another important factor is the reliability of the sequencing methods, highly related to systematic and stochastic sequencing errors. Those errors are dangerous when considered alone, for they become indistinguishable from a real variant. This question can normally be managed by increasing the number of sequencing reads, since the error rate is known and quantifiable through extensive calibration of the machines. Furthermore, a high sequencing coverage can also mitigate the damage caused by high error rates, i.e. raw reads can be recalibrated and obtain better base quality scores as a result.

3.3.1.2 Coverage

Sequence coverage, defined as the average number of times each nucleotide is represented in an aligned read, is correlated to other factors relevant to accurate mapping reads, such as the error rate of the sequencing method, the alignment algorithm used, the repeat complexity, and the read length. Many studies have proved the correlation between the depth of coverage and the level of specificity and sensitivity in CNV detection, so that either low-coverage (e.g. 4X to 8X) or high-coverage (e.g. 40X to 100X) certainly could positively or negatively affect the capability of a read-depth based method to detect certain types of variations, for instance. In the AS approach, because of the *de novo* genome sequencing and assembly needed, a higher coverage is mandatory.

The cost of a project is undeniably an important factor to consider, so much so that some projects opted to sequence samples at low coverage for cost reducing. The 1000 Genomes Project, for instance, used two- to six-fold coverage resulting in an expected reduction of the power to discover structural variants. However, even dealing with low coverage, it is possible to obtain high sensitivity and specificity for calling structural variants through the split-read approach. The genomic coverage can be irregular along the genome, yielding low local coverage, regardless of overall coverage, due to regions of the genome that are not easily fragmented for sequencing. Thus, read-depth methods are vulnerable to false positives even after bias corrections (GC-content and mappability) (SIMS et al., 2014). However, in general, the most serious effect related to coverage in NGS data and CNV detection is that low-depth can introduce sequence errors, which can be propagated through analyses of genetic variation, leading to wrong conclusions in CNV detection.

3.3.2 Considering choices for data sample

A crucial issue for planning NGS projects consists in deciding between whole versus partial sequencing, which is strongly related to the cost and the time required to complete the analysis.

Although high-throughput sequencing of the entire genome became possible about a decade ago, researchers and clinicians were historically mostly interested in specific genomic

regions, such as particular genes or even parts of them. Besides, as whole-genome sequencing has high computational cost and analytical complexity, WES technologies have been widely used for molecular diagnostics of pathogenic variants. Even in face of some intrinsic challenges associated to WES data, like the sparse nature of the target data and the non-uniform depth of coverage among targeted regions, important contributions to Mendelian and complex diseases emerged regarding human genetic mutations in the coding regions.

However, recent progress in genomic studies has identified many structural variants in non-coding regions of the human genome, leading to conclude that the majority of genetic variants associated with complex traits lie outside genic regions (ZHANG; LUPSKI, 2015). This research trend has lead the focus back to exploring WGS data in order to achieve a more robust identification of CNV.

Another critical consideration is the choice among using single sample of one individual, multiple genomes for population analysis, or different samples (case/control) of one individual. Depending on the study design, one can be much better suited than the others. This issue is discussed in more detail in Section 3.4.

3.3.3 Using a suitable fragment distribution

Algorithms based on the RD approach for detecting CNV heavily rely on the assumption that the sequencing process is uniform, i.e., the number of reads mapped to a region is assumed to follow a Poisson distribution and to be proportional to the number of copies. However, certain biases as GC-content and mappability make this assumption unrealistic. In practice, neither sampling nor mapping of the reads is uniform because of these experimental biases. Miller and colleagues (MILLER et al., 2011) proved that the observed distribution violates the Poisson distribution assumption of equal mean and variance, conjecturing that the negative-binomial distribution is a better approximation for the over-dispersed Poisson distribution.

Assuming that shotgun sampling of DNA fragments is random implies that the CNV calls made by the methods are not due to different sequencing bias between two sets of data compared. This assumption could only be held valid when both data are generated using the same sequencing method. Nevertheless, considering different sequencing techniques, the randomness of sampling may not hold. In this case, to verify the validity of the initial assumption, some methods apply statistical tests to compare the distribution of GC-frequencies (for instance, the Kolmogorov-Smirnov test), in order to make sure that there is no significant difference between the two distributions. In spite of the probable violation, virtually all methods have used the Poisson or normal distribution assumption without subsequent evaluation of the suitability of the choice.

3.3.4 Correcting GC-Content bias

There is a well-documented dependency between GC-content bias and the number of reads mapped to different genomic regions in a sequence, causing these regions to be under- or over-sampled. NGS technologies are potentially affected by this bias in sample preparation, sequencing, alignment, and assembly. Indeed, DNA amplification, as part of the library preparation procedure, can severely bias the GC-content of sequences. This implies that higher levels of GC-content can distort the coverage of the genome, i.e., it can dominate the signal of interest for analyses that focus on measuring fragment abundance within a genome, the copy number estimation.

Of the four approaches studied, the most affected by the GC-content bias is RD, due to an unimodal relationship between this bias and the read depth signal, in which both high and low GC levels decrease the sequence depth of a region. This correlation has been well discussed since early studies (DOHM et al., 2008)(HILLIER et al., 2008). A common way to reduce this effect is to increase the overall sequence depth.

The GC-content effect can be hard to isolate from the true signal, once it is not consistent among repeated experiments. The control of this effect has become a widely-present procedure in workflows for NGS analyses. Algorithms based on Poisson models, data smoothing, binning, and other techniques (BENJAMINI; SPEED, 2012) have been used for normalizing the original signal and correcting the GC-content bias. For instance, mrFast (ALKAN et al., 2010) and ReadDepth apply a statistical correction technique to normalize the read-depth signal of each window. Alternatively, methods that require control sample rely on the assumption that the GC bias is implicitly corrected, since the variations from bias affect both tumor and normal samples similarly. CNV-Seq (XIE; TAMMI, 2009), SegSeq, CNASeg (IVAKHNO et al., 2010), JointSLM (MAGI et al., 2011), among others, share this assumption.

3.3.5 Correcting mappability bias

Other sequencing-related bias in NGS data is the mappability (also known as uniqueness), which indicates that, for a given region of the genome, the chances that a read originating from this region is unambiguously mapped back to it. Finding the original position cannot always be done uniquely due to repetitive regions, mutations, structural rearrangements (insertions/deletions), or sequencing errors, even using the best existing short read mapping algorithms available. Even one or two mutations or sequencing errors in one short read is enough to lead to a wrong location in mapping (LI; RUAN; DURBIN, 2008). Thus, regions with higher mappability have more unique sequences and produce less ambiguity, and vice versa.

After mapping the reads, the CNV detection tools utilize the mapping quality score, assigned by the aligner to each mapped read, to deal with multi-reads, i.e., reads mapped to multiple locations. There are three main different strategies: discarding the read, choosing a random position out of all of equally good match position, and just reporting all possible

positions. Some CNV detection tools included procedures to avoid discarding multi-reads, and consequently leading to false positive deletion calls.

As the mappability only represents the confidence of individual reads and does not point to regions of the genome where the reads can be confidently mapped, other attempts for defining this metric have been introduced using fixed length k-mers. Mappability scores have been shown as crucial information to distinguish regions of the genome that can be reliably mapped from those that cannot. One great advantage in this a priori processing is to investigate tradeoffs between many settings of experiments (read length, error rate, paired/single-end, etc.). The most common programs for computing the mappability score are GEM mappability (from the GEM (GEnome Multitool)) (DERRIEN et al., 2012) and Genome Mappability Analyzer (GMA) (LEE; SCHATZ, 2012).

The simplest method for correction using mappability scores is to skip the regions with low mappability, filtering by some threshold, so that only reads within high mappability regions are used to call CNV. This strategy is used by Control-FREEC. Another correction strategy is to recalculate the read counts of a given bin through dividing the raw read counts by regional mappability. This procedure results in both ambiguous reads discarded and unambiguous reads in low mappability regions overweighted for CNV detection. ReadDepth uses both procedures in order to prevent overcorrection in regions with very low mappability.

Nevertheless, no matter what strategy is used, the ambiguous reads will likely create some biases in the read count signal and may cause mistakes in CNV detection (TEO et al., 2012). Hence, existing mappers have interpreted multiple mapping reads in different ways, particularly when they are designed to discard all the reads involved, with implications on the outcomes obtained through quantitative analyses.

3.3.6 Performing a segmentation algorithm

Right assumption of data distribution for modelling the data variation is essential for most CNV detection programs to be able to distinguish genuine CNV from random effects. Segmentation techniques allow splitting RD signal into segments, determining boundaries on each change of DNA copy numbers. The probability of a segment having an altered number of copies in general depends mainly on the choice of a suitable data distribution for the signal and a simple threshold method.

Before segmenting, a partition step splits the entire genome into windows that have mapped reads enough to estimate the read depth signal. This procedure has been developed using various mechanisms, with different window-sizes, by different tools. For instance, CNV-Seq adopts a statistical test to determine an optimal size, while RDXplorer belongs to the class of methods that use fixed, user-defined window size. As far as base-pair resolution is desirable for breakpoint estimation, one must remember that the window size limits this resolution. Too large a window would sacrifice the resolution; on the other hand, too small a window would not give

enough power for detecting regions rich in segmental duplication. Therefore, it is worth tuning parameters related to partition, in order to reach the desired resolution in the results.

Most common segmentation algorithms used in aCGH and SNP array have been adapted for NGS data, including Circular Binary Segmentation and Hidden Markov Models, while others have been developed in recent years, as, for example, Bayesian Information Criterion (BIC) and regression tree-based algorithms (XI et al., 2011).

3.3.7 Adapting to the advent of longer reads

As the sequencing technology advances to make longer reads possible, some CNV detection methods have been adjusted to detect larger and more complex variants according to the new lengths of generated reads, especially with respect to the proper memory size for processing. Indeed, sequencing longer reads has increased the assessment of variations embedded in long repeat structures, such as balanced inversions.

It is widely agreed that CNV detection accuracy is improved with longer reads. Zhang and collaborators introduced SRiC (ZHANG et al., 2011), a split-read method, which, according to them, would be more useful with the production of longer reads in third-generation sequencing technologies. This was demonstrated through simulations performed with unbiased proportion of all types of SVs across different length-scales. However, only insertions are affected positively by longer reads; deletions showed comparable results, keeping marginal improvements at varying lengths. Indeed, the choice of the ideal read length is still a rather open-end question.

Higher mappability also depends mostly on the length of the sequence reads, besides the number of mismatches allowed. For instance, in the 1000 Genomes Project, about 20% of the reference genome was considered inaccessible, due to many ambiguously placed reads (TEO et al., 2012). This difficulty is due to the different families of motifs present in the human genome, which inflicts this complexity in certain regions, causing poor mappability.

According to Alkan and collaborators in a review published in 2011, it was estimated that roughly 1.5% of the human genome still could not be covered uniquely even with read lengths of 1Kb (ALKAN; COE; EICHLER, 2011). Recently, the powerful long reads have been highlighted by currently available technologies from Illumina, Oxford Nanopore, and Pacific Biosciences (PacBio). It has been reported that the most established one is Single Molecule Real Time (SMRT), from PacBio, which can generate reads as long as 54Kb, with an average read length over 10Kb, though with high error rate by reads (11-15%) (LEE; GURTOWSKI; YOO, 2014). In a recent article (MYERS, 2014), Gene Myers released DALIGNER, a new aligning tool designed for very noisy long reads. This solution is based on two essential properties: (1) The distribution along the genome of the reads produced being close to a Poisson sampling, and (2) An almost perfect randomness in the location of errors within reads.

It is important to note that the favorable effects brought by longer reads are not always observed. A few studies have shown different views about the impact of these input data for CNV

detection, particularly related to coverage. For instance, according to Krishnan and colleagues, to detect most copy number alterations in cancer samples, one does not need longer reads but acceptable coverage (KRISHNAN et al., 2012); also, according to Abyzov and colleagues, at constant coverage, longer reads lower sensitivity to smaller CNV in RD-based methods (ABYZOV et al., 2011).

3.3.8 Evaluating in the absence of a gold standard

Although simulation studies cannot provide accurate evaluation of CNV detection methods, since the true landscape of variations shows complex structures, which are very hard to simulate, synthetic data can give some relevant insights. It is quite common for developers to use DGV as a benchmarking for assessing CNV detection methods, because this database is distinguished as one of the most reliable and useful current storages of CNV. Many authors have produced extensive simulation schemes for recreating different settings of copy number changes by altering real datasets, in order to supply a properly tailored and repeatable variety of scenarios under control. Considering this goal, some specific tools were developed to generate artificial benchmarking datasets. For instance, Hong and colleagues presented a computational tool (HONG et al., 2014), which generates datasets of test sequences with inserted CNV originated from DGV spanning a large range of sizes (75bp to 10Mbp), types (losses and gains), and random locations, as well as short indels and SNPs.

Such automatic procedures allow saving the set of included CNV, to be used in retesting. Additionally, they may ensure that the quantities of each type of CNV included is proportional to the ratios of the variations present on real datasets. One common resource used as quantifying reference is the HapMap collection, in which nearly 80% of the all annotated variations are of type loss region, 15% of type gain region, and 5% of mixed region (MAGI et al., 2012).

On the other side, to demonstrate performance using real data, most methods have adopted as resource the mapped reads derived from the collections of genomes available through projects like the 1000 Genomes Project and HapMap, or from individual complete human genome sequences, such as Craig Venter's (LEVY et al., 2007) and James Watson's (WHEELER et al., 2008), which were generated using Sanger and 454 platforms, respectively. Furthermore, in order to validate their results, most methods compare the CNV detected with variants available on these mentioned resources, treating them as ground truth. One special advantage of the 1000 Genomes Project is its variety of available data, generated by several platforms (Illumina, 454 Roche, ABI Solid System), different read lengths (25bp - 400), and low (up to 4X) or high (5X, 10X, 25X, 50X, 70X) coverage. Despite the fact that higher coverage leads to better performance in terms of specificity and sensitivity, analyzing CNV in low-coverage data will continue to be relevant in the future due to lower financial and computational costs.

3.4 CNV detection tools for cancer studies

The conventional data choice in CNV detection consists in general in single individual samples, using a reference genome for identifying variations, without borrowing information from other samples or matched controls. CNVnator, ReadDepth, and CNVeM (WANG et al., 2013) are examples of tools that work this way. Population-based sampling, on the other hand, pursuits particularities among individuals from a specific race, locality, or a group sharing a particular phenotype. Focusing on clinical applications, the analysis of common CNV population scale includes the identification of recurrent genomic variants in patient subgroups that are hardly observed in healthy individuals.

To deal with this type of sampling, detection tools based on multiple samples, such as RDXplorer, cn.MOPS (KLAMBAUER et al., 2012), JointSLM, and commonLAW (HORMOZDIARI et al., 2011), integrate complementary information in order to improve the detection power. For example, CBSBR (DUAN; DENG; WANG, 2014) solves the concurrency of CNV across multiple samples through a regression model to fit multiple RD signals. Other tools are JointSLM, which relies on modelling the signals as a sum of independent stochastic processes, supplemented by a HMM (MAGI et al., 2011), and cn.MOPS, that models the signal with a mixture of Poison models that generates a separate model for each DNA locus.

A type of sampling that is growing drastically in usage is the control sample, used as a reference for the mapping. Such samples are particularly useful in disease studies, when sequence reads from both disease and normal samples of the same individual are compared. Examples of methods that use this type of data sampling are: CNASeg, BIC-Seq (XI et al., 2011), CNAnorm (GUSNANTO et al., 2012), and rSW-Seq (KIM et al., 2010). Some methods accept either single samples or case/control samples, as, for instance, Control-FREEC. Usually this type of sampling is applied in cancer studies to locate copy number alterations (CNA), which differs from CNV because CNA are mutations that occur in tumor tissues (as opposed to normal ones) (OSTROVNAYA; NANJANGUD; OLSHEN, 2010). Detecting CNV usually refers to finding the number of copies of a particular gene that differs from one individual to others, normally named germline variants, while detecting CNA means pursuing somatic changes to chromosomal structure that result in gain or loss in copies of sections of DNA. The procedure for finding these alterations consists in comparing normal sample with tumor sample from the same individual.

There are relevant differences between germline and somatic variations, especially due to the complexities of tumor samples. For instance, their lengths and diversities in genome are very distinguishable, besides the specific concern in dealing with the existence of an inevitable normal cell contamination in tumor cell. Because of that and other challenges that cause some misperceptions in the signal variation, there are excellent germline CNV detection tools that are not suitable for CNA detection (LIU et al., 2013). Tools exclusively focused on CNA detection have been developed, such as CNASeg, CNAnorm, and ReadDepth.

In cancer studies a tool for detecting CNAs usually relies on matched normal samples (i.e.

case/control approach), which is required in most methods in order to help identify heterozygous SNP loci, and to filter out benign CNV in patients. There are undeniable advantages in using matched control to identify patient-specific CNV and to focus exactly on somatic alterations. Nonetheless, there are two negative issues about this. One is that sequencing appropriate control samples is not always possible, thus becoming a challenge for users when it is required. Another is the disadvantage of handling a double amount of data (LIU et al., 2013). Considering the two issues just mentioned, there are at least two tools that avoid this requirement of control sample data. One is ReadDepth, which uses only tumor data, and the other is Control-FREEC that can call CNAs with or without control samples. When the matched sample is absent, Control-FREEC uses profiles based on GC-content previously calculated in the normalization step.

One last, but not less important, point to consider in CNA detection tools is about normal cell contamination in tumor sample, which is inevitable, especially in clinical situations, when the material is obtained from tissues of tumor samples (GUSNANTO et al., 2012). Therefore, without the chance of obtaining pure tumor samples, it is necessary to resort to estimation in order to correct it. Some tools have included procedures to deal with that, such as CNAnorm and Control-FREEC. On simulations to compare these procedures, the results showed that, unlike CNAnorm, the performance of Control-FREEC progressively got worse as the contamination level increased.

3.5 Relevant aspects for comparisons

In this section, some relevant aspects are discussed which should be used to compare features commonly observed during the assessment of CNV detection tools.

3.5.1 Performance: Execution time and memory usage

NGS and its ability to perform massive parallel sequencing in a single run brought an unprecedented opportunity to sequence many genomes at a relatively inexpensive cost. With this huge production of data, the technology needs improvement to better handle, store, and analyze such data. This also implies in technical issues designed to achieve the best computational speed using less memory, which means to consider, for instance, whether the programming language supports parallel computation or how memory should be managed. For instance, considering an input data of 247Mbp at 34X coverage, the computational time required by six different methods ranged from aproximatly 6.7 minutes to 2 hours and 15 minutes, while peak memory reached roughly from 4 to 24.5 Mbytes, in a desktop computer with dual-core 2.8 GHz x86 64-bit processor and 6 GB memory (DUAN et al., 2013).

At least three factors are extremely related to efficiency in speed and in memory usage: coverage, WES/WGS, and CNV calling length. A higher coverage implies in a larger number of generated reads. For instance, in the comparative experiment performed by Magi and colleagues for analyzing the performance of three CNV detection tools, the number of reads in the database

used varied from 13.44 (at 1.1X coverage) to 2738.03 million (at 39.1X) (MAGI et al., 2012). A popular alternative for reducing the huge number of reads is to focus on a target region for analysis (WES). In another aspect, CNV calling length has a strong impact, since, usually, the longer it is, the more time-consuming the methods become. Therefore, most methods narrow the size of CNV calling, in order to avoid peaks of execution time, due to the decrease in the accuracy of the aligning process with larger gaps.

3.5.2 Bias control

In general, the low sensitivity in CNV detection is mainly caused by short reads mapped wrongly to the reference genome. This absence of uniqueness in mapping could be associated to regions with diverse particularities, such as high GC-Content, repeated/segmental duplicated regions, low read coverage, low base quality scores, mutations, and sequencing errors. Regarding these challenges, it is essential that each method include efficient strategies to deal with these factors during the alignment step of data processing.

The local GC-content and the genomic mappability are the two main sources of biases that affect substantially the variant calling. There is no consensus as to the best technique to mitigate or to remove the GC effect in a sample. A thorough and well summarized study appears in a very interesting paper by Benjamini and Speed (BENJAMINI; SPEED, 2012). There is also no agreement on the major source of the GC bias, but empirical evidence supports the hypothesis that amplifications are the most important cause for this bias. The key point is to find a more suitable model for the GC curve and apply it in the development of a method for correcting the skew effect.

About the mappability, it is certain that it depends mostly on the length of sequence reads, sequencing approach (single reads vs paired-end sequencing), and the number of mismatches allowed, as well as the gap parameters of the alignment algorithm used. To avoid discarding low-mappability regions, it is important that the methods use some strategy for increasing coverage in these regions. The use of longer reads, along with paired-end libraries, has often been used to increase the chance of reads to be mapped uniquely. However, even adjusting the parameters, the ambiguity will probably be present, due to the nature of the data, increasing false positives in CNV calling.

3.5.3 Variety of applications

Since there is yet no single solution able to cover the full range of analysis involving CNV, a variety of available tools are used for solving the CNV detection problem partially. Analytical studies under various scenarios have been assayed, for instance, with different types of samples (individual, case/control, and population), both NGS data types (WGS/WES), and at different levels of resolution. Hence, for achieving a higher accuracy in all CNV classes along with rather diverse scenarios, it is crucial to take into account the stronger abilities of each tool

and, therefore, consider the combination of strategies for improving the accuracy.

It is worth noting that any comparative procedure is incomplete without taking into account the inherent difficulty of ensuring that two different CNV events could represent the same variant. In spite of the systematic inspection of quality assurance performed during the inclusion of an event in DGV database, some imprecision still exists relative to size and location of the available variants. Thus, to consider two variants as being the same, most studies of CNV analysis assess the equality as long as there is at least 50% reciprocal overlap (MILLS et al., 2011).

3.5.3.1 Detection of insertions/duplications

It is quite consensual that insertions are harder than deletions for all approaches (TEO et al., 2012). PEM-based methods can detect insertions only when the distance between mapped read pairs is shorter than the fragment length, i.e., these methods have a length upper bound of an insertion detected as the average fragment length minus the length of the reads. Alternatively, SR-based methods are not completely capable of detecting insertions, once the read length of the current technologies limit the size of the detected insertion. Hence, in general, read-depth tools are considered the best option, since they can detect very large insertions and duplications.

3.5.3.2 Detection of deletions

RD-based methods also show good sensitivity of detecting deletions, although regions with multi-reads could be falsely detected as a deletion, in the case that these reads are removed. With PEM methods, false deletions can also be identified instead of true large insertions, caused by the discordant mappings. The effectiveness of the SR approach in detecting deletions shows the same limitation related to insertions, the limitation of the read length. Again, tools based on read-depth are recommended for this application.

3.5.3.3 Copy number estimation

Another source of diversity among approaches is about their ability to estimate the absolute number of copies for each segment. Read-depth is the only sequencing-based approach to accurately predict exact DNA copy numbers, thanks to the analysis of proportionality of mapped read. Paired-end methods cannot estimate well, because they perform poorly on repeat and segmental duplication-rich regions (ZHAO et al., 2013). Read-depth tools are good options for estimating the number of copies, due to the hypothesized correlation between depth of coverage of a genomic region and the copy number of the region (TEO et al., 2012).

3.5.3.4 Detection of breakpoints

A crucial distinction among approaches is the breakpoint resolution achievable, since a precise characterization of breakpoints strongly contributes to the accuracy of CNV callings in general. Split-read methods are highlighted for their capability of detecting breakpoints at base pair resolution, though limited by the read length. RD-based methods, on the other hand, have a considerable limitation on the detection of precise breakpoints, even increasing coverage.

3.5.4 Aligner-dependence

It is possible that some methods show some aligner-dependence on their sensitivity and specificity, since the process of read alignment plays an important role in all strategies for detecting CNV. The mappability bias, for instance, and the ambiguity caused by it depend strongly on the aligner and the parameters used when aligning the reads to an existing reference sequence. Read alignments also strongly impact the detection of exact breakpoints, which depends on the upstream aligner to map short sequencing reads to the reference genome.

Many alignment tools for short reads have been developed over the years and, with the constant push to improve the read length by advances in sequencing technologies, some improved algorithms have emerged. An in-depth review of some aligner tools was done by Pabinger and collaborators, showing relevant issues for the selection of an alignment program for variant analysis (PABINGER et al., 2014).

3.6 Conclusion and Perspectives

The current bottleneck of genomic projects is not the sequencing of the DNA itself anymore, but lies in addressing issues related to data management and the highly sophisticated computational analysis of the experimental genomic data. Examining the way to achieve reliable detection of genetic variants, since the early solutions based on aCGH, it is possible to highlight diverse challenges (addressed or not) that can better explain the state-of-the-art methods.

To stress the importance of all issues discussed in this study, we can conclude it with an outlook on the near future of the CNV detection problem, showing two challenging issues that arise from the fast evolution of the key technologies involved. On the one hand, there is the intense computational infrastructure needed to satisfactorily support large-scale experiments with adequate population samples, long reads, and high coverage in whole-genome sequencing. On the other hand, tool developers must note the urgency of investing in fast computation procedures for improving the performance of each individual step of the pipeline of CNV analysis.

Most available tools deal with different scopes, including different data types (WGS or WES), applications (individual, case/control, or population studies), and resolution of CNV size and breakpoint, besides differences at design level (implementation, operating system, and I/O format). All this variety brings caveats in elaborating a fair evaluation procedure for these tools.

Since there is no universally suitable gold standard tool and true benchmark data, an unbiased evaluation of CNV tools remains a largely open issue.

Moreover, the lack of a generally agreed upon benchmark data also weakens the estimation of the sensitivity of the methods. Large sequencing projects like the 1000 Genomes, the UK10K, and HapMap, have provided better understanding about the links between genetic changes and traits or diseases. Even though, the imprecision of the breakpoints identified makes the variants provided by such projects not perfectly adequate for working as a true benchmark.

With the rapid development of different techniques and analytical methods, there are improvement gaps to be filled, ranging from data generation to computational analyses, to achieve the main goal, which is an adequate and accurate clinical interpretation. Despite a number of software packages and associated pipelines enabled for detecting CNV, this problem is still a long way from being properly solved.

Finally, all findings presented here strongly point towards what one must be concerned with in order to comprehend the current context and possible future directions in CNV detection. As we are dealing with applications for human health, the next step after variant analyses consists on carrying out genetic association tests at every variation existent, thus achieving important clinical related results for diagnosis and therapy.

4

ScreenVar: a methodology for evaluating structural variants

The absence of a benchmark data of structural variants has been firmly stated in the literature as a barrier for designing experimental validation of new SV detection tools. In this thesis, a putative set of consolidated variants is defined as an important result of the proposed methodology ScreenVar. This chapter introduces the steps of ScreenVar and discusses the results of the experiment using data from DGV.

4.1 Introduction

Initiatives such as HapMap (INTERNATIONAL; CONSORTIUM, 2005), 1000 Genomes Project (TONEVA et al., 2012)(DURBIN et al., 2010), and UK10K (SANGER, 2017) have aimed to extend the knowledge about human genetic variation, with respect to uncovering the effects of the variability caused by copy number variations (CNV) and balanced rearrangements on human diseases, complex traits, and evolution. Indeed, these projects have relied on the growing development of the strategies designed to the process of discovering SV.

Considering the variability within a molecular assay workflow (extraction, quantification, molecular testing, and data analysis) as well as the diversity of available tools involved in a NGS analysis and also the complex nature of human genome, this resultant scenario is undeniably a source of difficulties in establishing unbiased comparisons among identified variants. Without going into specifics, this uncertainty stem from the inherent differences in sequencing technologies, data collection, read-align methods, variant-calling algorithms, and so on.

In addition, given the existence of multiple data processing pipelines for CNV detection, some comparative studies have been developed in order to examine the degree of concordance among the currently existing pipelines. We can place particular emphasis on two studies related to discordance between variant calling pipelines using *indels* calls. One of them is the description and experimentation of a method called ReliableGenome (RG) for partitioning genomes into high and low concordance regions with respect of surveyed variant analysis pipelines (POPITSCH; SCHUH; TAYLOR, 2016). Such study provides strong evidence that the degree of concordance depends predominantly on genomic context, including genomic region, variant type, read depth, and varies by analytic pipelines.

Another comparative study investigated the question *"how closely do the results from multiple pipeline agree with each other?"*, showing that, when examining the pipelines GATK, SOAPindels and SAMTools, the agreement rate for *indels* calls was very low at 3.0% (O'RAWE et al., 2013). It can be explained by the imprecision in breakpoint resolution, since single variants can be obtained with a very high prediction confidence and, in turn, indels and SV callings have higher uncertainty levels. Moreover, most available reads generated by such platforms have been short-sized, measuring a few hundred bases in size, thus collaborating to a probable ambiguity in mapping tasks.

Furthermore, there are diverse studies that draw attention to different caveats that have been present throughout the stages of NGS data analysis, which may strongly affect the outcomes (NASCIMENTO; GUIMARAES, 2016) (TEO et al., 2012). In short, no single discovery approach can cover the entire spectrum of SV in the genome, and the process of evaluating a SV detection tool depends largely on the platforms and tools used. Thus, many authors have firmly stated the lack of established benchmark data and tools as a major challenging barrier for adequately evaluating such tools (ALKAN; COE; EICHLER, 2011) (VALSESIA et al., 2013) (LIU et al., 2013) (GIANNOULATOU et al., 2014).

The accuracy of a SV detection tool lies basically on its ability to assess the distinction between genuine variants and random effects originated from sampling or sequencing errors, to an extent that any significant results of a reliable tool are inherently repeatable. Thus, to address the issue of evaluating the accuracy of a given detection tool, one way is to focus on verifying the concordance of the outcomes with other providers. In other words, one important concern is not whether a given tool give correct variants, but rather how closely it agrees with the others.

In this chapter, we introduce a biclustering-based methodology called ScreenVar, which aims to screen all variants obtained from a set of different resources, and then to determine which variants are accurate in face of a subset of these sources, taking into account a specific equivalence criterion. The overall idea of this methodology involves a rearrangement of the initial data through a cross tabulation by found equivalent variants and its respective references. For this purpose, ScreenVar incorporates different ways to check the equivalence of two variants coupled with the use of biclustering algorithms.

In our experiments, ScreenVar was applied to discover if there are accurate variants within chromosome 1, under a variety of parameters. The experiments resulted in outcomes that could be ranked according to their number of equivalent variants, their number of sources, and their accuracy computed by a determined quality measure. The list of the best ranked results contains, for instance, a case comprised by a set of 588 variants supported by 8 studies (*Vogler 2010*, *Suktitipat 2014*, *Redon 2006*, *de Smith 2007*, *Lou 2015*, *Cooper 2011*, *Conrad 2009*, *Coe 2014*), which was obtained with the least strict analyzed equivalence criterion. In addiction, it was selected the variants associated to the outcomes with high score and belonged to a putative benchmarking dataset provided by DGV.

Indeed, an important contribution of this study is to provide some different lists of

putative accurate variants in order to enhance evaluation strategies of new tools. Moreover, another relevant contribution consists in bringing a scalable solution that allows researchers to ascertain the accuracy of a chosen detection tool under own conditions, since input data of different formats can work together, whenever applicable.

4.2 Materials and Methods

This section describes the proposed methodology for evaluating the agreement of structural variants derived by different calling studies. Initially, we outline the steps of ScreenVar and, then, we show a brief description of the biclustering technique, the relevant aspects of the acquisition of input data, a concept for equivalent variants, and the validation strategies aimed to produce a collection of candidate gold standard elements.

4.2.1 Proposing an evaluation methodology

To lay a foundation for thorough tool evaluations, specific questions need to be answered, such as *what to evaluate*, *which criteria to use*, *how to measure those criteria* and *how to verify the accuracy*. The answers to such questions contributed to building the foundational ideas of this thesis:

1. **What is being evaluated?** Structural variants.
2. **Which criteria will be used?** The localization and the type of each variant.
3. **How to verify the accuracy of analyzed structural variants?** Biclustering techniques.

(1) Briefly, **structural variants** are DNA alteration events occurred in a given location. Different formats have been used to represent such variants in data discovery and storage. Adaptable formats, the complex nature of variants and large spectrum of variant types comprise a landscape with an intrinsic problem: *how to verify whether two events are equivalent?*

(2) In order to proceed with the evaluation of SVs, the next step is to define which criteria can be used to compare them. Each variant record consists of many fields, including chromosome, position, identifications, reference and alternate allele etc. Due to this data complexity, this thesis only considers a short subset of fields that can provide the adequate information for the type of comparison to be applied in the evaluation process. In this case, the analysis will deal with **the localization and the type of each variant**.

(3) This work concentrates its efforts on investigating the accuracy of SV using **biclustering algorithms** in order to scrutinize the agreement among structural variants identified over different resources.

We then describe a new methodology called ScreenVar for evaluating structural variants based on biclustering techniques. ScreenVar comprises the four following steps:

1. Preprocessing input data;
2. finding equivalent variants;
3. applying a biclustering algorithm;
4. validating biclusters.

4.2.2 Preprocessing the input data

The input for ScreenVar consists of a collection of variants with the following fields: identification, chromosome, start position, end position, variation type (gain, loss, inversion, insertion etc.), and variant calling source (a given study or tool responsible for calling the variant). Different variant calling sources could feed into the phase of acquisition of data, such as the published studies 1000 Genomes Consortium Phase 1 (ABECASIS et al., 2012), Wong 2012 (WONG et al., 2013) etc.; a variant calling tool — CNVnator (ABYZOV et al., 2011); ReadDepth (MILLER et al., 2011); Pindel (YE et al., 2009); a well-structured collection of variants — DGV (MACDONALD et al., 2014), TCGA (ATLAS, 2008)), or even a putative gold standard, whenever applicable.

Thus, ScreenVar handles raw data by mapping each variant to its respective source, forming a matrix where the lines represent variants and the columns represent their references. Formally, we define this input as a $m \times n$ binary matrix, A = $[a_{ij}]$, with rows corresponding to variants $v_1, v_2, v_3, ..., v_m$ and columns corresponding to references $r_1, r_2, r_3, ..., r_n$ such that $a_{ij} = 1$ if and only if v_i is identified by source r_j.

4.2.3 Finding equivalent variants

One important issue in the context of variant detection is the notion of equivalent variants. It is quite common to find variants identified by different experiments which overlap, sometimes partially, in the same stretch or neighborhood of the genome. There is a common uncertainty in measurements supporting the same variants detected by different methods, especially if they result from older (and lower-resolution) studies. Hence, there must be a careful overlapping control to find possible associations among these events due to their imprecise boundaries (outer/inner start and stop points).

To decide whether two variant regions, r_1 and r_2, correspond to the same event, we use the concept of *Minimum Reciprocal Overlap (MRO)* defined as:

$$MRO(r_1, r_2) = min\left(\frac{length(overlap(r_1, r_2))}{length(r_1)}, \frac{length(overlap(r_1, r_2))}{length(r_2)}\right)$$

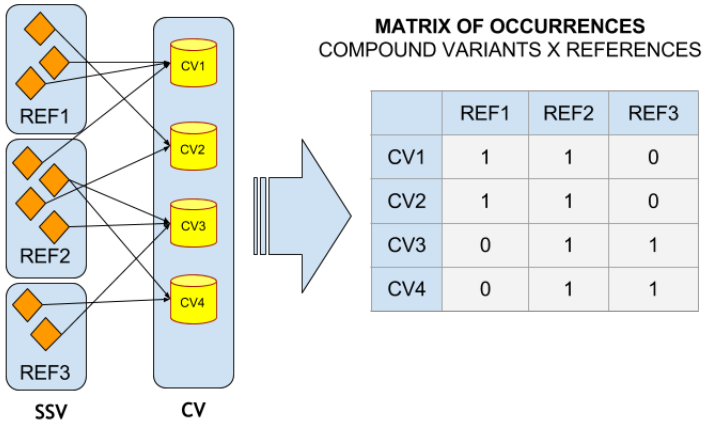
This measure provides a standard way of determining the similarity in the chromosomal location of two variant regions. As an example, when using the *MRO* measure with a threshold of 0.5, at least half of r_1 must be overlapping with r_2, and vice-versa.

In order to properly deal with similar genetic variants, taking into account fields such as location, classification and sampling data, we defined two criteria to consider two variants as equivalent: *Global* and *Local* criteria, both supported by a parameter for lower limiting the *MRO* value. To be accepted as equivalent under the *Global criterion*, two variants must fulfill two conditions: (1) to overcome the determined threshold for *MRO*, and (2) to be of the same type. While the *Local criterion* only requires the first condition, limiting the reciprocal overlap rate, regardless of their types.

To proceed with the step of finding the equivalent variants, ScreenVar scans all variants present in the matrix A, comparing with each other according to previously defined criteria. In addition, to lead a reduction of such huge number of pairwise combinations among all variants, two user-defined thresholds for controlling the *length of overlap* and *MRO* were created. In general, the greater these values, larger and more similar variants are delivered for further tasks in ScreenVar. Thus, choosing suitable values for these parameters can determine the acceptance level of equivalence between each pair of involved variants.

ScreenVar splits such equivalent variants into groups, each of which is referred to as a *Compound Variant* (CV) (Fig 4.1). As the methodology compares all variants against each other, possibly originated from different sources, all members in a CV are equivalents among themselves. Thus, a variant in a given CV can be seen as an event supported by different references/studies. In order to better represent such degree of relationship between variants and references, we established that the number of different references supporting a single CV represents the *support level* of this CV. As the proposed methodology aims to examine the concordance degree across the data sources, high assurance levels should lead, to a certain extent, to finding accurate variants.

Figure 4.1: ScreenVar clusters equivalent variants to elaborate a matrix of occurrence by references.



	REF1	REF2	REF3
CV1	1	1	0
CV2	1	1	0
CV3	0	1	1
CV4	0	1	1

The next action is to handle the matrix A produced in the first step in order to remodel it, considering now the compound variants found. The new matrix is a rather smaller, concise sample space. We represented the consolidated matrix as a $m \times n$ binary matrix, B = $[b_{ij}]$, with rows corresponding to compound variants $cv_1, cv_2, cv_3, ..., cv_m$ and columns corresponding to

references $r_1, r_2, r_3, ..., r_n$ such that $b_{ij} = 1$ if s_j supports cv_i (that is, there is a member of the compound variants cv_i referred to the source s_j) and $b_{ij} = 0$ otherwise.

4.2.4 Applying a biclustering algorithm

The use of clustering techniques is broadly applied to reveal a genuine similarity in data profiles, analyzing patterns existing in such data in order to partition them according to some investigated properties. In our case, applying a standard clustering approach to matrix B would produce clusters of compound variants regarding all studies, or, on the other hand, clusters of studies involving all compound variants. It should be observed that while the clustering problem always creates disjoint clusters that cover all the input set, biclusters may cover only a part of the matrix. Actually, however, the need of this ScreenVar step is to analyze not only the columns representing the properties of items, but also the rows as the items themselves. Thus, with such demand for forming an associated cluster of variants and bind it to the cluster of references, it is suitable to deal with it as a biclustering problem.

Biclustering is an unsupervised technique that can be applied to simultaneously group rows and columns of a matrix, in order to find certain coherent patterns (BUSYGIN; PROKOPYEV; PARDALOS, 2008). Formally, given a rectangular matrix $X_{M\times N} = x_{ij}$, with $M \times N$ numeric values, a bicluster Y = (R, C) is a sub-matrix of X, where R and C are the subsets of row and column indexes, respectively. A biclustering algorithm aims to discover a set of biclusters so that each result satisfies some specific criterion of homogeneity.

Although being a proved NP-hard problem (CHENG; CHURCH, 1999), there are various methods designed to undertake this challenging task, as it is the case with heuristic-search based solution. Several algorithms have been proposed to identify biclusters and some state-of-the-art methods were considered in this thesis (see Table 4.1).

Table 4.1: Biclustering methods

ALGORITHM	ACRONYM	REFERENCE
Biclustering of expression data by Cheng and Church	CC	Cheng and Church (CHENG; CHURCH, 1999)
Plaid Models	PM	Lazzeroni and Owen (LAZZERONI; OWEN, 2002)
Conserved gene expression Motifs	Xmotifs	Murali and Kasif (MURALI; KASIF, 2003)
Bimax algorithm	Bimax	Prelic et. al. (PRELIC et al., 2006)

After the generation of matrix B, ScreenVar performs a given biclustering algorithm in order to find homogeneous subsets of rows and columns simultaneously. We selected four biclustering algorithms to be used in ScreenVar, namely Cheng and Church (CC), Plaid Models, Conserved gene expression Motifs, Bimax. Each one should be executed with matrix B as input and the resulting biclusters be ordered by a quality score. Such list of resulting biclusters should indicate exact homogeneous groups involving variants strongly connected by a subset of references.

4.2.5 Validating biclusters

Some validation measures can be applied to measure the quality of a resulting bicluster. Typically, these metrics are organized into two main types, namely internal and external measures. Internal measures consider only information intrinsic to the dataset and external measures use prior knowledge about groups of objects as extrinsic information (HANDL; KNOWLES; KELL, 2005).

In recent years, different measures have been proposed to score the accuracy of biclusters. In this work, we chose the statistical measure called ChiaKaruturi measure (CHIA; KARUTURI, 2010), contained in the *Biclust* toolbox[1] that also releases the biclustering methods used in this work. The key feature of the ChiaKaruturi measure lies in the estimation of three different types of co-expression: strong row effects, strong column effects or row and column joint effects. Along with such effective measurement, this score quantifies the differential between the rows and columns involved and the remaining ones.

The idea underlying the addition of this quality function in such a methodology is to rank the outcomes with respect to their joint effects. Biclusters with undesirable effects caused by single-row or single-column should be avoided, since it is against the expected relevance among variants and references jointly. Thus, beyond the need of a low global variance in homogeneous groups, ScreenVar seeks to focus on biclusters with a suitable number of lines and columns.

Finally, ScreenVar checks whether the variants contained in the well-ranked biclusters are present in a validation database provided by DGV (DGV-GS). Such crosschecking procedure of ScreenVar intends to product two special sets: (i) the variants contained in both collections, which implies in validated events, and (ii) the variants found in our method but not in DGV-GS, which can indicate new findings for a putative gold standard.

One of the objectives of this thesis is the development of a comprehensive methodology for evaluating structural variants. Then, the delivery of a tool based on such methodology is surely the main contribution of this work (see details of the ScreenVar tool in Appendix B).

4.3 Results

We performed a thorough analysis of the performance and behavior of the method ScreenVar regarding the reliability of variants through some specific issues such as:

1. Insights on the suitability of the input data;
2. distribution of equivalent variants across distinct studies;
3. analysis of resulting biclusters;
4. validation of the resulting biclusters with a validated subset of DGV.

[1] Package for bicluster analysis in R (https://cran.r-project.org/package=biclust)

4.3.1 Insights on the suitability of the input data

The primordial requirement for meta-analysis procedures is the acquisition of a sufficiently large and adequate number of entries in order to reduce the play of chance. Thus, ScreenVar needs input data that guarantees a reasonable coverage of location, types, and different discovery methods. In that respect, DGV has easily accomplished its initial goals, as discussed below.

DGV is a curated database with the goal of providing a catalog of the variants discovered in analytic studies. It is organized in two large groups, supporting structural variants (SSV) and structural variant regions (SVR). SSV represent variants identified in a single sample/individual and SVR are regions formed by the combination of multiple SSV sharing the same start and end positions (MACDONALD et al., 2014). The archive of DGV used in these experiments was released in May 2016, corresponding to NCBI Genome Reference hg19 and containing roughly 7 million variants distributed in 72 studies.

It was verified that performing all 7 million variants would be impracticable, since ScreenVar includes pairwise analyses and a compilation of such results in a matrix, taking excessive memory usage and execution time. Because of that, we selected the chromosome with greater number of variants for our experiments, chromosome 1, which represents a total of 565,300 SSV associated to 66 studies.

There is a remarkable heterogeneous scenario around the studies of DGV, comprised by different genome reference, samples, methods, and platforms. Hence, the initial concern is to verify possible biases in such context that may affect the pursuit of reliable variants. Then, some analyses are shown as follows, with the examination of the chosen subset of DGV under some perspectives, including the source, the classification (the variant types), size and localization (chromosome, start, end). The aim is surely to collaborate for building a more comprehensive scenario and fostering respective further conclusions.

4.3.1.1 Source: analytic study and samples

DGV holds information about the calling process of each analytic study, including the discovery methods, samples/individuals, platform and other settings. From 72 studies in DGV, 66 have endorsed some variants of chromosome 1, with an average number of 8,565 SSV ($\sim 1.5\%$) per study. As shown in Table 4.2, there is a high concentration of variants on the first ten studies with more SSV (the full list of 66 studies is presented in Appendix B.2), gathering a total of 521,315 SSV (92%). This concentration occurs especially due to the *1000 Genome Consortium Phase 3* study, which carries 266,299 variants ($\sim 47\%$ of total).

Such large quantities of variants in a single study can be justified by the adoption of recent sequencing technologies and high sample sizes. Regarding the experimental design, a large sample size is required to achieve sufficient statistical power for rare variant studies. While sample size increases, the number of novel variants per sequenced individual will decrease

Table 4.2: Analytic studies cataloged in DGV/Chromosome 1

REFERENCE	METHOD	TOTAL	%
1000 GC Phase 3	Sequencing	266299	47.1%
1000 GC Phase 1	Merging, Oligo aCGH, PCR, Sequencing	93733	16.5%
Coe 2014	Oligo aCGH, SNP array	41312	7.3%
Cooper 2011	Oligo aCGH, SNP array	36707	6.5%
Wong 2012b	Sequencing	18775	3.3%
Campbell 2011	Oligo aCGH	17390	3.0%
Sudmant 2013	Oligo aCGH, Sequencing	16870	2.9%
Altshuler 2010	SNP array	15296	2.7%
1000 GC Pilot Project	Digital array, Oligo aCGH, PCR, Sequencing	9228	1.6%
Uddin 2014	SNP array	5705	1.0%

(DURBIN et al., 2010). In the input dataset in question, the average sample size is approximately 1,054, with a maximum of 29,084 samples in the *Coe 2014* study (COE et al., 2014).

Another important consideration is the resolution achieved in these studies, which can strongly affect the performance of *ScreenVar* as a whole. Most studies currently available do not have base level precision and thus they provide their boundaries in terms of breakpoint range. This uncertainty directly affects the step of finding equivalent variants of *ScreenVar*, very likely leading to false concordance among studies. Therefore, as long as the variants can be base pair resolution, ScreenVar could integrate such data more suitably and seek valuable insights.

The diversity of methods in such studies is intrinsically linked to the varying resolution achieved in the respective discovery process. One example of the relevance of high resolution was demonstrated in the construction of a CNV map produced for documenting the variability of the human genome (ZARREI et al., 2015). One of the primary factors for study selection was the accurate breakpoint resolution, discarding data sets from studies based on lower-resolution arrays. Thus, it is worthy highlighting that the selection in such case was performed looking for studies based on sequencing.

To show this aspect in our analyzed input data, Table 4.3 summarizes the quantities of variants by methods, in which more than 51% of SSV occurred in studies using sequencing, as expected, due to ongoing advances in NGS data. Moreover, the concentration of variants in sequencing methods and, therefore into studies, can unevenly guide meta-analysis results (see the full list of methods used in DGV in Appendix B.3).

At this stage, we do not apply any restrictions for selecting specific sources, but we can retake some relevant features of the studies in the interpretation phase of the results of ScreenVar, as appropriate.

4.3.1.2 Location, type and size distributions

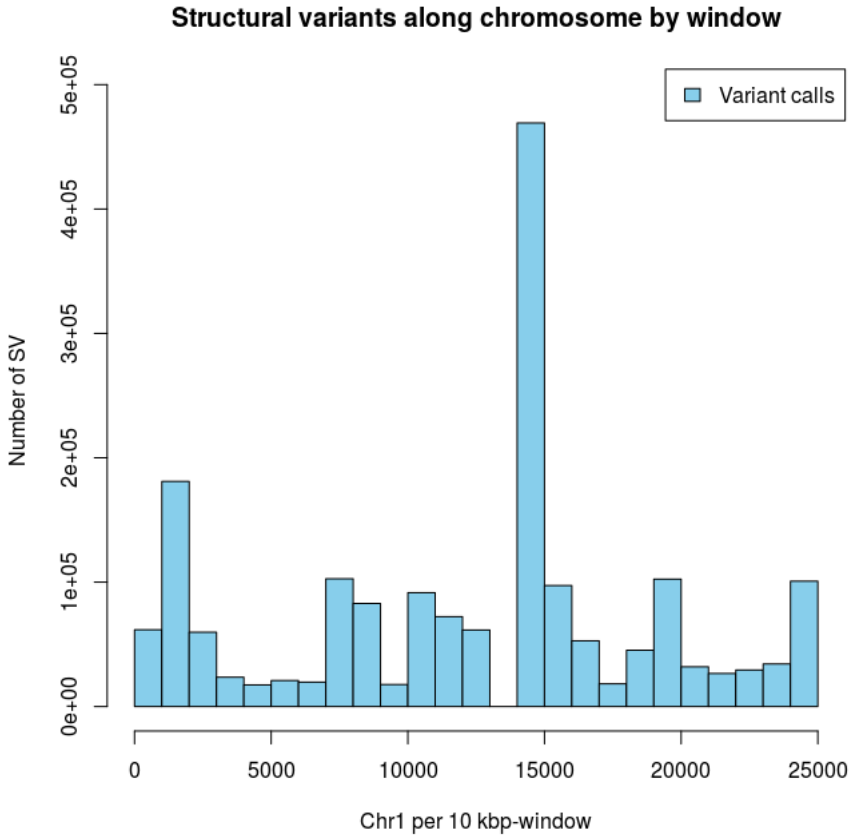
In terms of the location distribution of SSV along chromosome 1, it is important to keep in mind the effect of known biases like mappability and GC-content in NGS data analyses. An odd behavior presented in Figure 4.2 should result from such effects. The number of variant calls were displayed unevenly spread, specifically across the windows of $130 - 150 \times 10^6$ bp. These

Table 4.3: Methods used in studies cataloged in DGV/Chromosome 1

METHOD	TOTAL
Sequencing	293715 (51.9%)
Merging, Oligo aCGH, PCR, Sequencing	93733 (16.5%)
Oligo aCGH, SNP array	79483 (14.0%)
SNP array	30147 (5.3%)
Oligo aCGH	27195 (4.8%)
Oligo aCGH, Sequencing	17864 (3.1%)
Digital array, Oligo aCGH, PCR, Sequencing	9228 (1.6%)
Merging, SNP array	4638 (0.8%)
BAC aCGH, SNP array	2705 (0.4%)
Sequencing, SNP array	1626 (0.2%)

windows are comprised by an abrupt absence of variant calls in first part and quite high quantities in the adjacent window, probably because this gap region corresponds to the centromere of chromosome 1.

Figure 4.2: Histogram of locations of the structural variants along chromosome 1



In addition, substantial differences in type and size distributions are shown in Table 4.4. Deletions are the vast majority in chromosome 1, achieving more than 76% of the total, while duplications only represent roughly 20%, and the others remain close to 1% of the variants. Indeed, many authors have already confirmed that it is quite common to have more deletions than others, due to the upper limitation for detecting insertions through the length of insert fragment

(TEO et al., 2012). Regarding the distribution size, the interquartile ranges[2] for inversion, insertion and complex events showed very low size spread. However, the gain, loss and gain+loss types together enlarged the range to thousands, revealing a relevant size dispersion around the variants of such input data.

Table 4.4: Variant type/size in studies cataloged in DGV/Chromosome 1

TYPE	NUMBER OF SSV	INTERQUARTILE RANGE OF SIZE
Loss / Deletion	431948 (76,41)	6895 (7595 - 700)
Gain / Duplication	116252 (20,56)	59325 (69010 - 9685)
Insertion	9851 (1,74)	13 (14 - 1)
Inversion	6136 (1,09)	51 (660 - 609)
Gain + Loss	825 (0,15)	108120 (124700 - 16580)
Complex	288 (0,05)	0 (6000 - 6000)

4.3.1.3 Identifying replicates

As we have mentioned earlier, there are some studies with high sample sizes, generally focused on complex traits, for which sequencing a large number of cases/controls or subjects is required. Therefore, such research works tend to report large numbers of replicate variants. The existing redundancy was detected in our experiments due to a poor performance in overlapping all variants with each other, in which several compound variants were identified gathering over a thousand identical variants, resulting in a drastic drop in performance.

Aiming at eliminating this redundancy, a new restriction was employed in order to filter only distinct variants, regarding location (chromosome, start/end positions), classification (variant type and sub-type), and source (reference). The number of variants decreased from 565,300 to almost 60,000 SSV, representing a little more than 10%. Interestingly, one of the studies responsible for these redundant items was the 1000 Genome Project Phase 3, with a striking mark of 3,419 replicated items for the same variant. In addition, with regards to the chosen benchmark data, of the initial 431,000 variants of DGV-GS, only 27,865 variants could be identified as unique, thus representing roughly 6.4% of initial DGV-GS.

4.3.2 Analysis of the equivalent variants across/among independent studies

For experimenting the ScreenVar stage responsible for finding equivalent variants, we selected many values for the parameters *length of overlap* (1, 100, 500, 1000, 2000, 5000 and 10000) and *MRO* (0.5, 0.6, 0.7, 0.8, 0.9, 0.97, 0.99), and also with the application of an equivalence criterion defined above. One of the key features of this stage is that it allows to characterize the input data set by a smaller number of representative variants through the distribution of the resulting compound variants. Such experiments have shown a decrease of the entire initial set of 565,300 SSV to 7-10 thousand new compound elements, varying according to

[2] Also called *midspread*, which is a measure of statistical dispersion: the difference between upper and lower quartiles

the equivalence conditions used. Not all values of each parameter represented relevant changing in results between themselves. So, we opted to present the further analyses from the most distinct results, in this case, those using values 1 and 100 for the length of overlap, and the values 0,7, 0,9, and 0,99 for the *MRO*, as summarized in Table 4.5. It is important to note that the growing values assigned to the *MRO* (from 70% to 99%) directly impacted the decrease of the number of found CVs, as expected.

Table 4.5: Parameters used for finding equivalent variants

	PARAMETERS			RESULTS	
#	EQUIVALENCE CRITERION	OVERLAP LENGTH	RECIPROCAL OVERLAP RATE	NUMBER CV	NUMBER SSV
EQ1	*Full*	1	0.99	7214	18791
EQ2	*Local*	1	0.99	6289	15599
EQ3	*Full*	100	0.90	10767	36097
EQ4	*Local*	100	0.90	10440	33506
EQ5	*Full*	100	0.70	10666	42779
EQ6	*Local*	100	0.70	10509	40358

Considering the list of ten studies with more input SSV, it is possible to visualize in Figure 4.3 how the resulting CV were distributed across each of them. It is possible to highlight the high number of SSV associated to the *Wong2012b* study, which achieved almost 17,000 for the parameter settings ***EQ5*** and ***EQ6***. With this excellent rate of more than 90% of its SSV, which are considered equivalent to some variant in other studies, *Wong's study* leads as one of most reliable studies in the DGV catalog. The list of all studies with their respective numbers of equivalent SSV and ratio of input SSV as equivalent are available in Appendix B.4.

Other results not as exceptional but yet remarkable have also been found by ScreenVar, such as *Coe2014*, *Cooper2011*, and *Conrad2009*. In contrast, there have been cases with a very low proportion of equivalence, which is still an interesting outcome, since such situations would not only indicate reliability regarding agreement but can also highlight rarely studied variants. For example, the *Kidd-2010* and *Itsara-2009* studies showed low ratios of 0/881 and 9/1291, respectively, using the most relaxed equivalence criterion (*EQ6*).

From another point of view, by definition, each CV is a result of comprising many SSV and therefore it can be associated to different studies. Such measurement has termed the *support level* of a CV and the higher this value, the more studies give support to all SSV contained in such CV levels ranging from 0 to 32 for the input data set found (the entire list of support levels is listed in Appendix B.5). Level 0 (zero) represents a cluster of SSV whose equivalents were found in the study itself. However, the main interest is on the high values of *support level*, which can indicate sets of putative reliable variants. Under equivalence condition EQ5, 274 variants present in 32 different studies were simultaneously uncovered. The list of these 32 references is in Appendix B.7.

Figure 4.3: Number of SSV found in each study

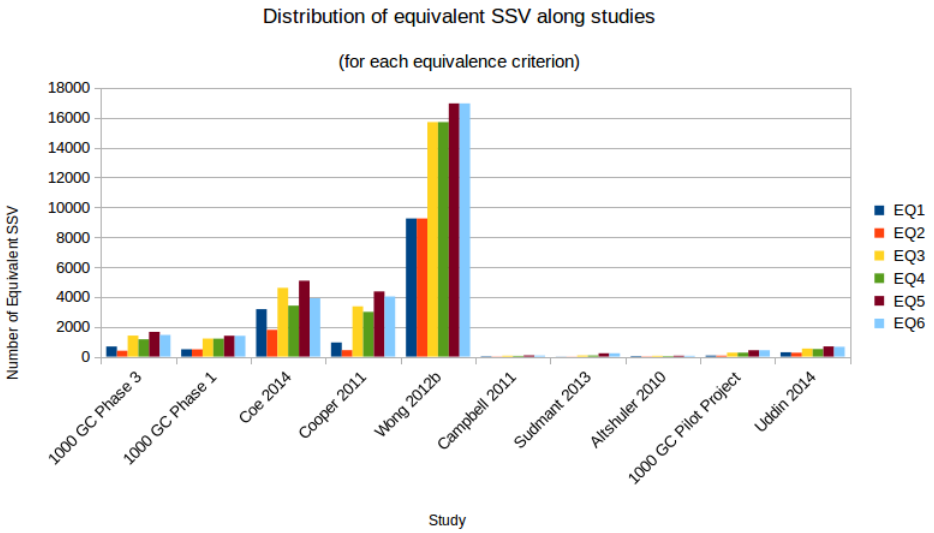


Table 4.6: Five-number summaries of variables extracted from the resulting biclusters

VARIABLE	MINIMUM	1ST QUARTILE	MEDIAN	3RD QUARTILE	MAXIMUM
Number of CV	2	45	220	445	9108
Number of SSV	4	394	1702	4143	32263
Number of Studies	2	2	2	3	28
Score	-1.03	2.58	3.24	3.85	4.62

4.3.3 Analysis of the resulting biclusters

A key characteristic of ScreenVar is not only to address the issue of clustering equivalent variants but also to perform a two-way analysis with such variants and their respective supporting studies. Then, the six sets of generated CVs related to each equivalence criterion were applied to four prominent biclustering algorithms with 10 iterations.

This experiment resulted in 742 biclusters, which were scored using the accuracy function of Chia-Karuturi for measuring the co-expression with regards to both the compound variants and the studies simultaneously. The overall behavior of the resulting biclusters is demonstrated through five-number summaries for numbers of CV, numbers of SSV, number of studies, and scores (see Table 4.6). Such statistical measures provide information about how spread are the dimension and the accuracy of the outcomes regardless of the biclustering algorithm and equivalence criteria used. Although large intervals between the 1st and the 3rd quartile for the number of SSV are not sufficient on their own in order to indicate excellent results, this characterizes a possible scenario for multiple size, type and location of the underlying variants.

In addition, concerning the relevance of each equivalence criterion on the evaluation of the biclusters, the experiments provided results with higher average scores for the stricter criteria (*EQ1* and *EQ2*), and, similarly, lower average score for the least rigorous one (*EQ6*), as summarized in Table 4.7. To a certain extent, this was expected because rigid conditions tend to produce more cohesive and smaller sets after clustering, as verified by the low average number

of CV and references for the former cases.

Table 4.7: Summary of generated biclusters

CRITERION	#BICLUSTERS	AVG #CV	MAX #CV	AVG #REFS	MAX #REFS	AVG SCORE	MAX SCORE
EQ1	100	59	405	2.2	3	4.2	4.62
EQ2	101	109.88	5278	2.2	3	4.32	4.62
EQ3	113	836.01	9108	2.22	3	3.15	3.93
EQ4	116	1024.78	8883	2.22	3	2.91	3.84
EQ5	164	798.8	7999	4.07	28	2.49	4.15
EQ6	148	933.7	7959	3.51	28	2.3	3.59

Results grouped by each six defined equivalence criterion: Number of biclusters, average/maximum number of compound variants, average/maximum number of studies, average/maximum score.

Focusing on the list of the most well-evaluated biclusters (last quartile, i.e, score greater than 3.85), which are listed in Table 4.8, all of them are comprised by only two or three studies. Half of these biclusters gather a number of SSV higher than the median value (1,702 SSV), even though associated to very low number of studies (two and three). Particularly, the outcomes with the highest scored-value are very small subsets: ID=1 {4 SSV x 3 studies} and ID=2 {10 SSV x 2 studies}.

At a first, it is possible to highlight potential meaningful biclusters by combining different aspects such as dimension, score, spreading of the variants, unrelated studies, equivalence criterion, and so on. Large dimensions could possibly represent good results, but it is still necessary to verify, for instance, the existence of some bias in the agreement among the involved studies. Moreover, the ratio of numbers of compound variants and SSV in a bicluster can point to a high concentration of equivalent variants.

On the other hand, there are a few dozen biclusters with 4-10 studies, which were left out of the former list by the limit score. This time, two conditions were used to select another set of biclusters: the score higher than the median (a lower limit) and the number of studies higher than 3 (the highest value in previous selection). Table 4.9 lists the five biclusters found. As it can be observed, this list involves larger biclusters, achieving a relevant quantity of 8 studies, even though the numbers of SSV have decreased. Indeed, as a significant support level is extremely important to ensure the reliability of these grouped variants, biclusters #29 and #30 deserve to be taken into full account. Thus, to an enhanced data analysis, the lists of all variants associated to these two biclusters are shown in Appendix B.8 and B.12.

4.3.4 Validation of the resulting biclusters with a validated subset of DGV

A lack of existing benchmarking data sets means that it is not clear what variants will be suitable to compare with the identified results. However, it is quite important to consider some validated sets of structural variants, such as DGV Gold Standard (DGV-GS). This collection is comprised by a filter whose role is to select a subset of the highest resolution variants according to the combination of three conditions: (i) share at least 50% reciprocal overlap, (ii) supported by

Table 4.8: List of the most well-evaluated biclusters

ID	# STUDIES	# CV	# SSV	CRITERION	SCORE	STUDIES
1	3	2	4	EQ1,EQ2	4.62	Alkan 2009, Conrad 2009, Perry 2008
2	2	5	10	EQ2	4.62	Locke 2006, Sharp 2005
3	2	6	16	EQ1	4.57	Locke 2006, Sharp 2005
4	2	26	83	EQ2	4.48	Cooper 2011, Conrad 2009
5	2	405	1502	EQ1	4.48	1000 GC Phase 3, 1000 GC Phase 1
6	2	396	1464	EQ2	4.47	1000 GC Phase 1, 1000 GC Phase 3
7	2	55	269	EQ1	4.43	Conrad 2009, Ju 2010
8	2	10	37	EQ2	4.43	Cooper 2011, Wong 2012b
9	2	44	132	EQ1, EQ2	4.41	Conrad 2009, Perry 2008
10	2	85	576	EQ2	4.34	Wong 2012b, Conrad 2009
11	3	3	20	EQ2	4.27	Conrad 2009, Cooper 2011, Wong 2012b
12	3	11	48	EQ1	4.25	Conrad 2009, Park 2010, Perry 2008
13	2	41	274	EQ1	4.21	Conrad 2009, Park 2010
14	2	1055	6894	EQ5	4.15	1000 GC Phase 3, 1000 GC Phase 1
15	2	8	24	EQ2	4.05	Cooper 2011, Perry 2008
16	2	743	3549	EQ3	3.93	Coe 2014, Cooper 2011
17	2	832	6462	EQ5	3.91	Conrad 2009, Cooper 2011
18	2	3	8	EQ2, EQ1	3.88	Alkan 2009, Conrad 2009
19	2	300	1680	EQ4, EQ3	3.85	Redon 2006, Coe 2014
20	3	394	2010	EQ4	3.75	Coe 2014, Vogler 2010, Cooper 2011
21	2	1968	6193	EQ4	3.75	Vogler 2010, Cooper 2011
22	2	550	3336	EQ3	3.74	Conrad 2009, Cooper 2011
23	3	396	2471	EQ3	3.73	Coe 2014, Vogler 2010, Cooper 2011
24	2	657	6406	EQ4	3.68	Wong 2012b, Alsmadi 2014
25	2	13	58	EQ1	3.65	Park 2010, Perry 2008
26	2	1532	10174	EQ5	3.59	1000 GC Phase 1, 1000 GC Phase 3
27	2	1524	9497	EQ6	3.59	1000 GC Phase 1, 1000 GC Phase 3
28	2	1534	10174	EQ5	3.58	1000 GC Phase 3, 1000 GC Phase 1

The data is ordered by the score and it only contains biclusters which achieved score greater than 3.58 (3rd quartile).

Table 4.9: List of the biclusters with the highest number of studies

ID	# STUDIES	# CV	# SSV	CRITERION	SCORE	STUDIES
29	8	25	588	EQ6	3.25	Vogler 2010, Suktitipat 2014, Redon 2006, de Smith 2007, Lou 2015, Cooper 2011, Conrad 2009, Coe 2014
30	8	17	495	EQ6	3.32	Vogler 2010, Redon 2006, Suktitipat 2014, de Smith 2007, Lou 2015, Coe 2014, Conrad 2009, Cooper 2011
31	5	111	3244	EQ5	3.44	Wong 2012b, Thareja 2015, Boomsma 2014, Alsmadi 2014, Dogan 2014
32	5	81	2374	EQ5	3.52	Wong 2012b, Thareja 2015, Dogan 2014, Alsmadi 2014, Boomsma 2014
33	4	58	823	EQ5	3.52	Perry 2008, Cooper 2011, Park 2010, Conrad 2009

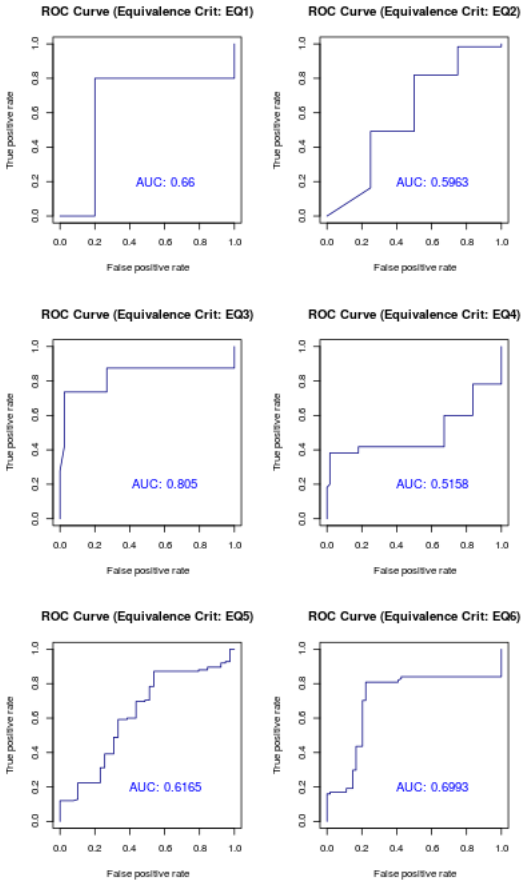
The list contains the biclusters with score greater than 3.24 (median value) and the number of studies exceeded 3 (the maximum value found in the list of most-evaluated biclusters).

at least two high resolution studies and (iii) found in at least two samples. The data was obtained from the release of May-2016/hg19, which contains 428,747 SSV, equivalent to approximately 75% of the number of input SSV.

First, the overall True Positive Rate (TPR) of the results of this work reached almost 63%. Yet, when filtering results by the accuracy score, in which only biclusters with score greater than the median (3.24) were included, the TPR rose to 77%. In order to observe the effect of calibrating this score over specificity and sensitivity of such results, it is possible to visualize the plotted ROC curves. There is a ROC curve for each equivalence criterion, as shown in Figure 4.4. To generate these graphics, the Chia-Karuturi score worked as a threshold in order to determine

whether a given instance should be classified as valid or not. It is also worth noting that the equivalence criterion EQ3 had the best performance according to the highest value of AUC (area under the ROC curve).

Figure 4.4: ROC curves with different equivalence criteria



Another list of biclusters is presented in Table 4.8, which is composed by the most well-evaluated biclusters (3rd quartile considering the accuracy measure), and added the number of SSV validated using as reference DGV-GS. Thus, according to Table 4.10, most of these biclusters achieved more than 70% in TPR.

Additionally, there is supplementary material with the lists of SSV obtained from DGV-GS related to the used input data set (Appendix B.6). This list was ordered in descending percentage in order to highlight 35 studies with no occurrence of SSV in such gold standard. Possibly, the reason for these studies being out of the gold standard may indicate a significantly low quality of these studies regarding the selective criteria of DGV.

4.4 Discussion and future directions

During the experiments, the use of ScreenVar accomplished the initial purpose of this work, which was to provide sets of reliable variants along with the observation of diverse aspects with respect to the behavior of each stage of the methodology. A first challenge addressed in

Table 4.10: List of the biclusters with higher validation ratios

ID	# IN-SSV	# GS-SSV	SCORE	STUDIES
5	4	4 (100%)	4.62	Alkan 2009, Conrad 2009, Perry 2008
18	576	563 (97.74%)	4.34	Conrad 2009, Wong 2012b
24	37	36 (97.3%)	4.43	Cooper 2011, Wong 2012b
22	83	80 (96.39%)	4.48	Conrad 2009, Cooper 2011
4	20	19 (95%)	4.27	Conrad 2009, Cooper 2011, Wong 2012b
25	24	22 (91.67%)	4.05	Cooper 2011, Perry 2008
21	274	251 (91.61%)	4.21	Conrad 2009, Park 2010
13	6406	5834 (91.07%)	3.68	Alsmadi 2014, Wong 2012b
16	1464	1316 (89.89%)	4.47	1000 GC Phase 1, 1000 GC Phase 3
23	58	52 (89.66%)	3.65	Perry 2008, Park 2010
3	48	43 (89.58%)	4.25	Conrad 2009, Park 2010, Perry 2008
15	1502	1345 (89.55%)	4.48	1000 GC Phase 1, 1000 GC Phase 3
20	132	118 (89.39%)	4.42	Conrad 2009, Perry 2008
9	9497	8276 (87.14%)	3.59	1000 GC Phase 1, 1000 GC Phase 3
14	3336	2897 (86.84%)	3.74	Conrad 2009, Cooper 2011
28	7	6 (85.71%)	3.92	Alkan 2009, Conrad 2009
10	6894	5864 (85.06%)	4.15	1000 GC Phase 1, 1000 GC Phase 3
7	10174	8640 (84.92%)	3.58	1000 GC Phase 1, 1000 GC Phase 3
8	10174	8640 (84.92%)	3.59	1000 GC Phase 3, 1000 GC Phase 1
2	2010	1699 (84.53%)	3.75	Coe 2014, Vogler 2010, Cooper 2011
11	6462	5446 (84.28%)	3.91	Cooper 2011, Conrad 2009
19	269	218 (81.04%)	4.43	Conrad 2009, Ju 2010
1	2471	1937 (78.39%)	3.73	Coe 2014, Vogler 2010, Cooper 2011
12	3549	2719 (76.61%)	3.93	Coe 2014, Cooper 2011
6	6193	4520 (72.99%)	3.75	Cooper 2011, Vogler 2010
17	1293	897 (69.37%)	3.84	Coe 2014, Redon 2006
26	16	0 (0%)	4.57	Locke 2006, Sharp 2005
27	10	0 (0%)	4.62	Locke 2006, Sharp 2005

List ordered by the ratio between Input-SSV and GS-SSV

ScreenVar was to establish proper flexibility for the input data. This is related to the fact that ScreenVar facilitates the integration of different sources of detecting variants, aiming to discover how they agree. For instance, it is possible to crosscheck information related to outcomes of mature tools such as CNVnator or CNV-Seq along with findings from a newly introduced tool. A simple preprocessing phase can easily convert different data formats to a given format according to the required fields of ScreenVar. Still about data input, it is essential to emphasize the relevance of having high resolution data along with a significant quantity of them, since the presented methodology can be seen as a strategy based on meta-analysis.

The task of clustering similar variants showed expected impacts according to the chosen equivalence criterion. A relevant implication about this step is that stricter conditions leads to better assessed biclusters, whereas softer criteria may entail a higher number of studies in agreement about a given set of variants. Thus, the user can set up parameters for overlapping variants, in order to opt by capturing either extremely cohesive, smaller groups, or bigger groups but not completely homogeneous.

This methodology shows some differences compared to the idea behind the CNV map built by Zarrei and colleagues (ZARREI et al., 2015), which have also analyzed variants from different studies. While they gathered sets of variants sharing a milder condition (at least 50% reciprocal overlap), ScreenVar is configured to work with any threshold for this rate (as experimented in this work, which were carried out with 70%, 90% and 99%), without mentioning the adoption of specific rules for recognizing the equivalence. Regarding the number of studies,

they restricted their results to exactly two distinct studies as a stringency condition, whereas the proposed methodology delegates this task to a data mining technique in order to discover how many and what studies are more appropriate to form a good cluster. All in all, the differences between these two works are focused on the ability of combining any number of variants and studies simultaneously.

The release of a list of reliable variants, as introduced in this thesis, is particularly valuable, since there are no precise and currently available validated SV. Thus, an ongoing benefit of using ScreenVar lies on assessing the quality of new SV detection tools and, therefore, providing further enhancements in the quality of available sets of genetic variants. As we pointed out in the Results section, ScreenVar found interesting results encompassing two cohesive groups with eight studies supporting numerous variants (around 500 events).

5

Discussions and contributions

5.1 Summary

Despite improvements to NGS technologies and SV detecting tools, the accurate identification of genetic variants still remains a challenge. This is due to a large variety of obstacles, including different technologies and protocols addressing each step of the thorough variant calling process, different resolutions in identified variants, and the lack of an established gold standard for such data. These obstacles directly affect the validation of new detection tools, especially on designing assays for experimental validation. Besides, the most relevant impact can be encountered when such currently available genetic variations are present in clinical analyses.

The exploration of the possibilities for identifying a set of reliable variants tackled three issues: unification of output format of different tools and databases, the definition of an equivalence criterion, and concordance analysis among variants and respective sources.

5.1.1 Unification of output format used in tools and databases

Through the integration of structural variants identified from different resources, the initial need would be to deal with variants represented in two distinct output formats commonly adopted by detection tools, which are GFF (generic feature format) and VCF (variant calling format). Since variants are intrinsically relative elements, each different assembly (*hg18* and *hg19*) can lead to uncertainty with respect to the equivalence among variants. Regarding DGV releases, there is a proper data format, including graphical, tabular and text-based formats.

To apply ScreenVar, it was necessary to define a unified collection with fields related to minimum required information about variants. This minimum field list was formed according to some degree of flexibility in input data so that a tool using VCF could be integrated with records from DGV (tabular format), for instance. Consequently, a proper data model was built and a new database released called *ScreenVarDB* (see respective entity-relationship model in Appendix - Figure B.2). This database is a collection involving the relationship among variants split into three levels of refinement: input, equivalent and reliable variants.

5.1.2 Definition of an equivalence criterion

In order to design the process of evaluating structural variants, defining a foundation for deciding when two variants can be considered equivalent was needed. This was an intriguing issue due to diverse factors, among which we highlighted the imprecision of a variant localization, the complex nature of human genome, the own relative nature of a variant (gain for a sample/loss for another) because of the dependency with the concerned reference genome and sample, and so on. Thus, defining an equivalence criteria provided an important decision for the other tasks of ScreenVar.

As our experiments have shown, the degree of replicated variants in DGV is very high, as shown by the example of a single variant with almost 3,5 thousand identical copies. Thus, an equivalence criterion should play the role of grouping approximated variants, considering different degrees of stringency, as defined by the *Local* and *Global Criteria*. Both criteria relied on reciprocal overlapping among the variants and thresholds to control the acceptability.

5.1.3 Concordance analysis among variants and respective sources

Concordance analysis is a strategy to measure the reliability among information and respective sources. In this thesis, the evaluation methodology consisted in rearranging the data related to equivalent variants found and their studies in order to build a matrix and then apply a biclustering algorithm. This aimed at carrying out a sort of meta-analysis seeking for agreement among studies with respect to identified variants. Therefore, the large number of resulting biclusters brought to light the possibility of having putative reliable variants depending on the considered studies and parameterization. As introduced in the previous chapter, there are well-evaluated biclusters comprised by approximately five hundred variants and supported by three to eight different studies.

The relevance of this contribution is reinforced by the fact that the chosen criteria were very stringent and yet a suitable number of variants was selected. Moreover, it is clear that adjustments or constraints can be included to suit a given researcher need. The methodology allows employing other equivalence criteria, parameters, biclustering algorithm and also to determine specific studies and/or variants to deal with.

5.1.4 Limitations

Notwithstanding, there are limitations of the proposed approach. First, it is important to note that a procedure based on meta-analysis can lead to misleading conclusions due to the highly dependence of the input data. Particular attention to data distribution is required, especially across the underlying studies, along with the method used for discovering the analyzed variants. Moreover, pursuing a gold standard for SV cannot handle with exclusively computational procedures without dealing with dry lab validations. The main application of this tool consists in pointing out putative benchmark data for filling the literature gap regarding the absence of robust

truth sets of SVs. Technically, there is an implication of performing ScreenVar with large input data sets. The experiments with variants associated to chromosome 1 were processed in roughly 12 hours using $\leq$ 4GB of memory. The main concern in terms of processing time focuses on the step of finding equivalent variants, which executes pairwise comparisons among all variants. Unfortunately, it was not possible to perform an analysis involving more than one chromosome in order to investigate translocation variants.

5.2 Future works

Important improvements can be obtained from adding selective filters in input data through an engine that allows choosing only studies involving a given method employed (e.g., sequencing) or a certain resolution achieved. Including a stage for doing this can reduce the number of false equivalent variants found.

Another point is to adapt this methodology to work as an ensemble of tools through possible new experiments designed to receive outputs from the execution of multiple SV detection tools, all targeting the same region of DNA. This trial should lead to ScreenVar pinpointing which tools agree with each other, forming thus an certain ensemble of reliable tools. Additionally, the resulting set of variants can be adequately qualified as supported by a chosen set of tools, and so, an ongoing database of reliable variants can gradually be produced. The expected results in the future are that updated and useful releases of diverse putative gold standard variants for the researchers are ade available. Thus, in an overview, ScreenVar can directly receive the outcomes of many detection tools and then it can produce set of agreed results, according to all inherent parameters and criteria.

5.3 Concluding remarks

Detecting structural variants is a complex problem which researchers have tried to solve with different strategies until the present without accurate results and with common low concordance among the tools. Hence, a large amount of effort has been employed to produce an adequate integration of the data sets in order to acquire meaningful insights. The inaccuracy of the results obtained derives from the fact that different pipelines have been applied to the variant analysis, in general, being affected by rapid advances in sequencing technologies. As an immediate consequence, there is an unprecedented incoming of data with higher levels of coverage and resolution and also increased availability.

After data mining of hundreds of thousands of variants in DGV using the proposed methodology, relevant collections of hundreds of variants could be considered reliable by supporting consolidated published studies. Unlike simulated variants, the findings indicated by our experiments using ScreenVar can provide a true landscape of variants since they are from real outcomes. Thus, this work has reached the expected objective of introducing a methodology

capable of combining information from massive data, relying richly upon diversity size, variant type, and location, and then being a provider of equivalent variants highly supported by different sources. Finally, this has produced a comprehensive approach to deal with the need of arranging benchmark data for use in the validation process of SV detection tools.

5.4 Important lessons learned

In this section, we would like to provide insights about the work performed that may help other researchers in this area. During each moment of this research, studying, reading, coding and writing, many difficulties were turned into learned lessons, and, besides the scientific contributions, we can also share other collaborations.

The challenge of a multidisciplinary area was recurrent along all the process. Undeniably, Bioinformatics has challenging problems to investigate, which have drawn attention from biologists, computer scientists, statisticians, mathematicians, and so on. However, most researchers have an academic degree in only one area, leaving the training for the other required fields of knowledge to be accomplished individually. This may produce diverse obstacles in the grasp of the overall context, such as fuzzy scope, fragile requirements, the dependency of a nonexistent user etc. Particularly, as computer science researchers, we can point out the absence of specificity in biological definitions. In general, there are many considerations to better comprehend the scenario, demanding a long contextualization to apply a certain concept or rule. Thus, we should reinforce the crucial need of training in each of the areas involved, at least Computer Science, Biology, and Statistics, earlier and more effectively.

Regarding development skills, an obvious, but very relevant point is the enormity of bioinformatics' databases. It is extremely important to be precise and careful in coding programs to handle large, complex and inconsistent databases, such as DGV and other biological databases. Furthermore, in this work, the combination of R-script with operational system Linux was a successful decision due to the availability of several libraries of Bioinformatic and having safety at the completion of jobs.

References

ABECASIS, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. **Nature**, [S.l.], v.491, n.7422, p.56–65, nov 2012.

ABEL, H. J. et al. SLOPE: a quick and accurate method for locating non-snp structural variation from targeted next-generation sequence data. **Bioinformatics**, [S.l.], v.26, n.21, p.2684–2688, nov 2010.

ABYZOV, A. et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. **Genome Research**, [S.l.], v.21, n.6, p.974–984, jun 2011.

ABYZOV, A.; GERSTEIN, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. **Bioinformatics**, [S.l.], v.27, n.5, p.595–603, mar 2011.

ALKAN, C.; COE, B. P.; EICHLER, E. E. Genome structural variation discovery and genotyping. **Nature reviews. Genetics**, [S.l.], v.12, n.5, p.363–376, may 2011.

ALKAN, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. **Nature Genetics**, [S.l.], v.41, n.10, p.1061–1067, 2010.

ALTSCHUL, S. et al. The anatomy of successful computational biology software. **Nature Biotechnology**, [S.l.], v.31, n.10, p.894–897, 2013.

ATLAS, T. C. G. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. **Nature**, [S.l.], v.455, n.7216, p.1061–1068, 2008.

BENJAMINI, Y.; SPEED, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. **Nucleic Acids Research**, [S.l.], v.40, n.10, p.1–14, 2012.

BERTHOLD, M.; HAND, D. **Intelligent Data Analysis**: an introduction. [S.l.]: Springer, 2003.

BOEVA, V. et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. **Bioinformatics**, [S.l.], v.27, n.2, p.268–269, jan 2011.

Broad Institute. **Broad Institute of MIT and Harvard [homepage] - https://www.broadinstitute.org/**. 2017.

BUSYGIN, S.; PROKOPYEV, O.; PARDALOS, P. M. Biclustering in data mining. **Computers and Operations Research**, [S.l.], v.35, n.9, p.2964–2987, 2008.

CHEN, K. et al. BreakDancer: an algorithm for high resolution mapping of genomic structural variation ken. **Nature methods**, [S.l.], v.6, n.9, p.677–681, 2013.

CHEN, K. et al. **TIGRA: a targeted iterative graph routing assembler for breakpoint assembly**. [S.l.: s.n.], 2014. 310–317p. v.24, n.2.

CHENG, Y.; CHURCH, G. Biclustering of expression data. **International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology, Department of Genetics, Harvard Medical School, Boston, MA 02115, USA**, [S.l.], v.8, p.93–103, 1999.

CHIA, B. K. H.; KARUTURI, R. K. M. Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. **Algorithms for molecular biology : AMB**, [S.l.], v.5, p.23, 2010.

CHIANG, D. Y. et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. **Nature methods**, [S.l.], v.6, n.1, p.99–103, 2009.

COE, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. **Nature genetics**, [S.l.], v.46, n.10, p.1063–1071, 2014.

CONRAD, D. F. et al. Origins and functional impact of copy number variation in the human genome. **Nature**, [S.l.], v.464, n.7289, p.704–712, 2012.

DERRIEN, T. et al. Fast computation and applications of genome mappability. **PloS one**, [S.l.], v.7, n.1, p.e30377, jan 2012.

DOHM, J. C. et al. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. **Nucleic Acids Research**, [S.l.], v.36, n.16, 2008.

DUAN, J.; DENG, H. W.; WANG, Y. P. Common copy number variation detection from multiple sequenced samples. **IEEE Transactions on Biomedical Engineering**, [S.l.], v.61, n.3, p.928–937, 2014.

DUAN, J. et al. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. **PloS one**, [S.l.], v.8, n.3, p.e59128, jan 2013.

DURBIN, R. M. et al. A map of human genome variation from population-scale sequencing. **Nature**, [S.l.], v.467, p.1061–1073, 2010.

FURUYA, T. et al. CNVs Associated with Susceptibility to Cancers : a mini-review. **Journal of Cancer Therapy**, [S.l.], n.May, p.413–422, 2015.

GIANNOULATOU, E. et al. Verification and validation of bioinformatics software without a gold standard: a case study of bwa and bowtie. **BMC Bioinformatics**, [S.l.], v.15, n.Suppl 16, p.S15, 2014.

GIRIRAJAN, S. et al. Refinement and Discovery of New Hotspots of Copy-Number Variation Associated with Autism Spectrum Disorder. **The American Journal of Human Genetics**, [S.l.], v.92, n.2, p.221–237, 2013.

GUSNANTO, A. et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. **Bioinformatics**, [S.l.], v.28, n.1, p.40–47, jan 2012.

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. **Bioinformatics**, [S.l.], v.21, n.15, p.3201–3212, 2005.

HANDSAKER, R. E. et al. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. **Nature genetics**, [S.l.], v.43, n.3, p.269–276, 2011.

HARTIGAN, J. A. Direct Clustering of a Data Matrix. **Journal of the American Statistical Association**, [S.l.], v.67, n.337, p.123–129, 1972.

HATTON, L.; ROBERTS, A. How accurate is scientific software? **IEEE Transactions on Software Engineering**, [S.l.], v.20, n.10, p.785–797, 1994.

HILLIER, L. W. et al. Whole-genome sequencing and variant discovery in C. elegans. **Nature methods**, [S.l.], v.5, n.2, p.183–188, 2008.

HONG, S. et al. Shape–based retrieval of CNV regions in read coverage data. **... Journal of Data Mining ...**, [S.l.], v.9, n.3, p.254–276, 2014.

HORMOZDIARI, F. et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. **Bioinformatics**, [S.l.], v.26, n.12, p.350–357, 2010.

HORMOZDIARI, F. et al. Simultaneous structural variation discovery among multiple paired-end sequenced genomes. **Genome Research**, [S.l.], v.21, n.12, p.2203–2212, 2011.

INTERNATIONAL, T.; CONSORTIUM, H. A haplotype map of the human genome. **Nature**, [S.l.], v.437, n.7063, p.1299–1320, 2005.

INTERNATIONAL, T. H. G. S. C. Finishing the euchromatic sequence of the human genome. **Nature**, [S.l.], v.431, n.7011, p.931–945, 2004.

IQBAL, Z. et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. **Nature Genetics**, [S.l.], v.44, n.2, p.226–232, feb 2012.

IVAKHNO, S. et al. CNAseg-a novel framework for identification of copy number changes in cancer from second-generation sequencing data. **Bioinformatics**, [S.l.], v.26, n.24, p.3051–3058, dec 2010.

KAISER, S. Biclustering: methods, software and application. **Ph.D Thesis**, [S.l.], p.178, 2011.

KIM, T.-M. et al. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. **BMC bioinformatics**, [S.l.], v.11, p.432, jan 2010.

KITCHENHAM, B. Evaluating software engineering methods and tool part 1: the evaluation context and evaluation methods. **ACM SIGSOFT Software Engineering Notes**, [S.l.], v.21, n.1, p.11–15, 1996.

KLAMBAUER, G. et al. Cn.MOPS: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. **Nucleic Acids Research**, [S.l.], v.40, n.9, p.e69, may 2012.

KORBEL, J. O. et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. **Genome biology**, [S.l.], v.10, n.2, p.R23, jan 2009.

KRISHNAN, N. M. et al. COPS: a sensitive and accurate tool for detecting somatic copy number alterations using short-read sequence data from paired samples. **PLoS ONE**, [S.l.], v.7, n.10, p.e47812, jan 2012.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, [S.l.], v.9, n.4, p.357–359, 2012.

LAPPALAINEN, I. et al. DbVar and DGVa: public archives for genomic structural variation. **Nucleic Acids Research**, [S.l.], v.41, n.D1, p.936–941, 2013.

LAWLOR, B.; WALSH, P. Engineering bioinformatics: building reliability, performance and productivity into bioinformatics software. **Bioengineered**, [S.l.], v.6, n.4, p.193–203, 2015.

LAZZERONI, L.; OWEN, A. Plaid Models for Gene Expression Data. **Statistica Sinica**, [S.l.], v.12, p.61–86, 2002.

LEE, H.; GURTOWSKI, J.; YOO, S. Error correction and assembly complexity of single molecule sequencing reads. **bioRxiv**, [S.l.], p.1–17, 2014.

LEE, H.; SCHATZ, M. C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. **Bioinformatics (Oxford, England)**, [S.l.], v.28, n.16, p.2097–105, aug 2012.

LEVY, S. et al. The diploid genome sequence of an individual human. **PLoS Biology**, [S.l.], v.5, n.10, p.2113–2144, 2007.

LI, H.; RUAN, J.; DURBIN, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. **Genome Research**, [S.l.], v.18, n.11, p.1851–1858, nov 2008.

LI, W.; OLIVIER, M. Current analysis platforms and methods for detecting copy number variation. **Physiological genomics**, [S.l.], v.45, n.1, p.1–16, jan 2013.

LIU, B. et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. **Oncotarget**, [S.l.], v.4, n.11, p.1868–81, 2013.

MACDONALD, J. R. et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. **Nucleic Acids Research**, [S.l.], v.42, n.D1, p.986–992, 2014.

MAGI, A. et al. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. **Nucleic Acids Research**, [S.l.], v.39, n.10, p.e65, may 2011.

MAGI, A. et al. Read count approach for DNA copy number variants detection. **Bioinformatics**, [S.l.], v.28, n.4, p.470–478, feb 2012.

MCCARROLL, S. A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. **Nature genetics**, [S.l.], v.40, n.10, p.1166–1174, 2008.

MEDVEDEV, P. et al. Detecting copy number variation with mated short reads. **Genome Research**, [S.l.], v.20, n.11, p.1613–1622, nov 2010.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics**, [S.l.], v.11, n.1, p.31–46, 2009.

MILLER, C. A. et al. ReadDepth: a parallel r package for detecting copy number alterations from short sequencing reads. **PLoS ONE**, [S.l.], v.6, n.1, p.e16327, jan 2011.

MILLS, R. et al. Mapping copy number variation by population-scale genome sequencing. **Nature**, [S.l.], v.470, n.7332, p.59–65, 2011.

MURALI, T. M.; KASIF, S. Extracting Conserved Gene Expression Motifs From Gene Expression Data. In: PAC. SYMP. BIOCOMPUT. **Anais...** [S.l.: s.n.], 2003. p.77–88.

MYERS, G. Efficient Local Alignment Discovery amongst Noisy Long Reads. In: **Algorithms in Bioinformatics**. [S.l.: s.n.], 2014. p.52–67.

NASCIMENTO, F.; GUIMARAES, K. S. Copy Number Variations Detection: unravelling the problem in tangible aspects. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, [S.l.], v.5963, n.c, p.1–1, 2016.

NIJKAMP, J. F. et al. De novo detection of copy number variation by co-assembly. **Bioinformatics**, [S.l.], v.28, n.24, p.3195–3202, dec 2012.

NOLL, A. C. et al. Clinical detection of deletion structural variants in whole-genome sequences. **npj Genomic Medicine**, [S.l.], v.1, n.January, p.16026, 2016.

O'RAWE, J. et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. **Genome medicine**, [S.l.], v.5, n.3, p.28, 2013.

OSTROVNAYA, I.; NANJANGUD, G.; OLSHEN, A. B. A classification model for distinguishing copy number variants from cancer-related alterations. **BMC bioinformatics**, [S.l.], v.11, p.297, 2010.

PABINGER, S. et al. A survey of tools for variant analysis of next-generation genome sequencing data. **Briefings in Bioinformatics**, [S.l.], v.15, n.2, p.256–278, mar 2014.

PANKRATZ, N. et al. Copy number variation in familial parkinson disease. **PLoS ONE**, [S.l.], v.6, n.8, 2011.

PEPLER, W. J.; SMITH, M.; NIEKERK, W. a. van. An unusual karyotype in a patient with signs suggestive of Down's syndrome. **Journal of medical genetics**, [S.l.], v.5, n.1, p.68–71, 1968.

PJ Hastings, James R Lupski, S. M. R.; IRA, G. Mechanisms of change in gene copy number. **Nat Rev Genet**, [S.l.], v.10, n.8, p.551–564, 2010.

PONTES, B.; GIRÁLDEZ, R.; AGUILAR-RUIZ, J. S. Biclustering on expression data: a review. **Journal of Biomedical Informatics**, [S.l.], v.57, p.163–180, 2015.

POPITSCH, N.; SCHUH, A.; TAYLOR, J. C. ReliableGenome: annotation of genomic regions with high/low variant calling concordance. **Bioinformatics**, [S.l.], v.33, n.September 2016, p.btw587, 2016.

PRELIC, A. et al. A systematic comparison and evaluation of biclustering methods for gene expression data. **Bioinformatics**, [S.l.], v.22, n.9, p.1122–1129, 2006.

PRESCOTT, N. J. et al. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. **Human Molecular Genetics**, [S.l.], v.19, n.9, p.1828–1839, 2010.

QUAIL, M. et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. **BMC Genomics**, [S.l.], v.13, n.1, p.1, 2012.

SANGER, W. T. **UK10K - Home page - http**://www.uk10k.org/. Online; accessed January 04, 2017.

SEBAT, J. et al. Strong association of de novo copy number mutations with autism. **Science (New York, N.Y.)**, [S.l.], v.316, n.5823, p.445–9, 2007.

SHERRY, S. T. et al. dbSNP: the ncbi database of genetic variation. **Nucleic acids research**, [S.l.], v.29, n.1, p.308–311, 2001.

SIM, S. E.; EASTERBROOK, S.; HOLT, R. C. Using Benchmarking to Advance Research: a challenge to software engineering. In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING, 25., Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2003. p.74–83. (ICSE 03).

SIMS, D. et al. Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, [S.l.], v.15, n.2, p.121–32, feb 2014.

TEO, S. M. et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing. **Bioinformatics**, [S.l.], v.28, n.21, p.2711–2718, nov 2012.

TONEVA, I. et al. The 1000 Genomes Project: data management and community access. **Nature ...**, [S.l.], v.9, n.March, p.459–462, 2012.

TUZUN, E. et al. Fine-scale structural variation of the human genome. **Nature genetics**, [S.l.], v.37, n.7, p.727–732, 2005.

VALSESIA, A. et al. The growing importance of CNVs: new insights for detection and clinical interpretation. **Frontiers in Genetics**, [S.l.], v.4, n.MAY, p.92, jan 2013.

WANG, Z. et al. CNVeM: copy number variation detection using uncertainty of read mapping. **Journal of Computational Biology**, [S.l.], v.20, n.3, p.224–236, mar 2013.

WHEELER, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. **Nature**, [S.l.], v.452, n.7189, p.872–6, 2008.

WILSON, G. et al. Best practices for scientific computing. **PLoS biology**, [S.l.], v.340, n.May, p.1–18, 2014.

WONG, L. P. et al. Deep whole-genome sequencing of 100 southeast Asian malays. **American Journal of Human Genetics**, [S.l.], v.92, n.1, p.52–66, 2013.

XI, R. et al. PNAS Plus: copy number variation detection in whole-genome sequencing data using the bayesian information criterion. **Proceedings of the National Academy of Sciences**, [S.l.], v.108, n.46, p.E1128–E1136, nov 2011.

XI, R.; KIM, T.-M.; PARK, P. J. Detecting structural variations in the human genome using next generation sequencing. **Briefings in functional genomics**, [S.l.], v.9, n.5-6, p.405–415, dec 2010.

XI, R.; LEE, S.; PARK, P. J. A survey of copy-number variation detection tools based on high-throughput sequencing data. **Current Protocols in Human Genetics**, [S.l.], v.Chapter 7, n.SUPPL.75, p.Unit7.19, oct 2012.

XIE, C.; TAMMI, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. **BMC bioinformatics**, [S.l.], v.10, p.80, jan 2009.

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. **Bioinformatics**, [S.l.], v.25, n.21, p.2865–2871, nov 2009.

YOON, S. et al. Sensitive and accurate detection of copy number variants using read depth of coverage. **Genome Research**, [S.l.], v.19, n.9, p.1586–1592, sep 2009.

ZANDA, M. et al. A Genome-Wide Assessment of the Role of Untagged Copy Number Variants in Type 1 Diabetes. **PLoS Genetics**, [S.l.], v.10, n.5, 2014.

ZARREI, M. et al. A copy number variation map of the human genome. **Nature Publishing Group**, [S.l.], v.16, n.February, p.172–183, 2015.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. **Genome Research**, [S.l.], v.18, n.5, p.821–829, may 2008.

ZHANG, F.; LUPSKI, J. R. Non-coding genetic variants in human disease. **Human Molecular Genetics**, [S.l.], v.24, n.July, p.ddv259, 2015.

ZHANG, Z. D. et al. Identification of genomic indels and structural variations using split reads. **BMC genomics**, [S.l.], v.12, n.1, p.375, jan 2011.

ZHAO, M. et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives - springer. **BMC bioinformatics**, [S.l.], v.14 Suppl 1, n.Suppl 11, p.S1, jan 2013.

Appendix

A

Supplementary Information: Unraveling the CNV Detection problem in tangible aspects

Table A.1: Relevant information about most popular Copy Number Variation detection tools ordered by approach and number of citations

Method	Language	Last update	Reference	Citations[1]
ASSEMBLY-BASED TOOLS				
Velvet (2008) - https://www.ebi.ac.uk/~zerbino/velvet/	C	Aug, 2014	ZERBINO; BIRNEY (2008)	2052/3790
Cortex Assembler (2008) - http://cortexassembler.sourceforge.net/	C	April, 2011	IQBAL et al. (2012)	64/151
Tigra (2014) - http://gmt.genome.wustl.edu/packages/tigra-sv/	C++	Sep, 2012	CHEN et al. (2014)	10/19
Magnolya (2012) - http://sourceforge.net/projects/magnolya/	Python	Nov, 2014	NIJKAMP et al. (2012)	6/14
PAIRED-END MAPPING TOOLS				
BreakDancer (2009) - http://gmt.genome.wustl.edu/packages/breakdancer/	Perl, C++	Aug, 2014	CHEN et al. (2013)	256/513
VariationHunter / CommonLAW (2010) - http://variationhunter.sourceforge.net/Home	C++	Jul, 2012	HORMOZDIARI et al. (2010)	67/135
SPLIT-READS TOOLS				
Pindel (2009) - http://gmt.genome.wustl.edu/packages/pindel/	C++	Sep, 2014	YE et al. (2009)	243/477
AGE (2011) - http://sv.gersteinlab.org/age/	C++	Sep, 2011	ABYZOV; GERSTEIN (2011)	24/44
SLOPE (2010) - http://www-genepi.med.utah.edu/suppl/SLOPE/index.html	C++	-	ABEL et al. (2010)	16/29
SRiC (2011) - N/A	N/A	N/A	ZHANG et al. (2011)	15/28
READ-DEPTH TOOLS				
MrFast (2009) - http://mrfast.sourceforge.net/	C++	Oct, 2014	ALKAN et al. (2010)	192/425
RDXplorer / EWT (2009) - http://rdxplorer.sourceforge.net/	Perl, Java	May, 2011	YOON et al. (2009)	148/292
SegSeq (2009) - http://www.broad.mit.edu/cancer/pub/solexa_copy_numbers/	Matlab	Jan, 2009	CHIANG et al. (2009)	130/283
CNVnator (2011) - http://sv.gersteinlab.org/cnvnator/	C++	Feb, 2014	ABYZOV et al. (2011)	130/244
CNV-Seq (2009) - http://tiger.dbs.nus.edu.sg/cnv-seq	Perl, R	Aug, 2014	XIE; TAMMI (2009)	107/244
Control-FREEC (2011) - http://bioinfo-out.curie.fr/projects/freec/	C++	Jun, 2014	BOEVA et al. (2011)	38/82
ReadDepth (2011) - https://code.google.com/p/readdepth/	R	Apr, 2011	MILLER et al. (2011)	37/62
cn.MOPS (2012) - http://www.bioinf.jku.at/software/cnmops/	R	Aug, 2014	KLAMBAUER et al. (2012)	35/70
BIC-Seq (2010) - http://compbio.med.harvard.edu/Supplements/PNAS11.html	R	Jan, 2013	XI et al. (2011)	35/67
CNASeg (2009) - http://www.compbio.group.cam.ac.uk/software/cnaseg	R	Sep, 2010	IVAKHNO et al. (2010)	27/57
CNAnorm (2012) - http://www.precancer.leeds.ac.uk/software-and-datasets/cnanorm/	R	May, 2014	GUSNANTO et al. (2012)	22/52
JointSLM (2011) - http://nar.oxfordjournals.org/content/suppl/2011/02/16/gkr068.DC1/JointSLM_R_Package.zip	R	Feb, 2011	MAGI et al. (2011)	18/32
rSW-Seq (2010) - http://compbio.med.harvard.edu/Supplements/BMCBioinfo10-2.html	C	Dec, 2010	KIM et al. (2010)	14/28
CNVeM (2013) - http://www.sph.umich.edu/csg/szoellner/software/	C	Sep, 2008	WANG et al. (2013)	3/7
CBSBR (2014) - http://www.mathworks.com/matlabcentral/fileexchange/36518-continuation-block-wise-sparse-approx	Matlab	May, 2012	DUAN; DENG; WANG (2014)	0/0
COMBINED TOOLS				
Genome Strip - RD+PEM+SR (2011) - http://www.broadinstitute.org/software/genomestrip/	N/A	2015	HANDSAKER et al. (2011)	82/151
CNVer - RD+PEM (2010) - http://compbio.cs.toronto.edu/CNVer/	C++	Jul, 2011	MEDVEDEV et al. (2010)	53/105

B

Supplementary Information: ScreenVar - A methodology for evaluating structural variants

The tool ScreenVar was implemented using R-scripts in compliance with the requirements described in the proposed methodology.

The diagram in Figure B.1 represents the flow of the data along all steps of ScreenVar. To understand the implementation of this tool, it is essential to observe the required input and parameters, and its output:

- Input: a text file containing genetic variants
- Parameters:
  - MRO: threshold for reciprocal overlap in order to find equivalent variants
  - Overlap size: threshold for the overlap length between each pair of variants
  - Minimum rows: Restriction for the dimension of resulting biclusters
  - Minimum cols: Restriction for the dimension of resulting biclusters
  - Maximum number of groups: Stopping criterion used in biclustering executions
- Output:
  - Graphical files with histograms summarizing all generated biclusters regarding number of studies, number of variants, score values, and so on.
  - Text files with summary data of all execution.

An important item of such process is the data model used for storing the data produced in each phase in order to allow further intermediate filtering and underlying analyses. The entity-relationship modeled for this tool is shown in Figure B.2. Moreover, a data summary of the data resulting after all experiments performed on this research is described in Table B.1. The data discussed in this thesis can be downloaded at http://bit.ly/screenvardb.

Figure B.1: Workflow of ScreenVar

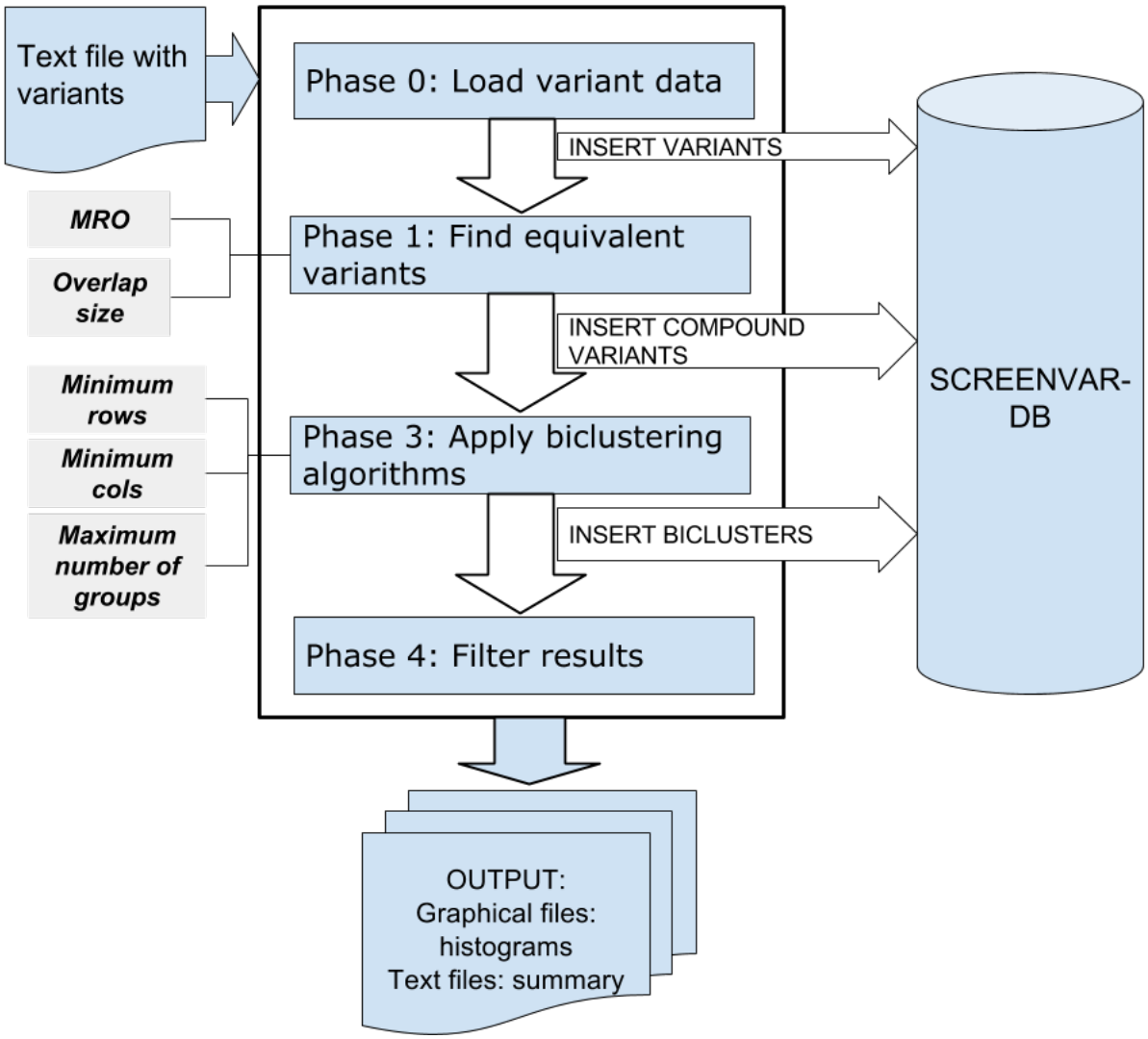


Figure B.2: Entity-relationship model of ScreenVar-DB

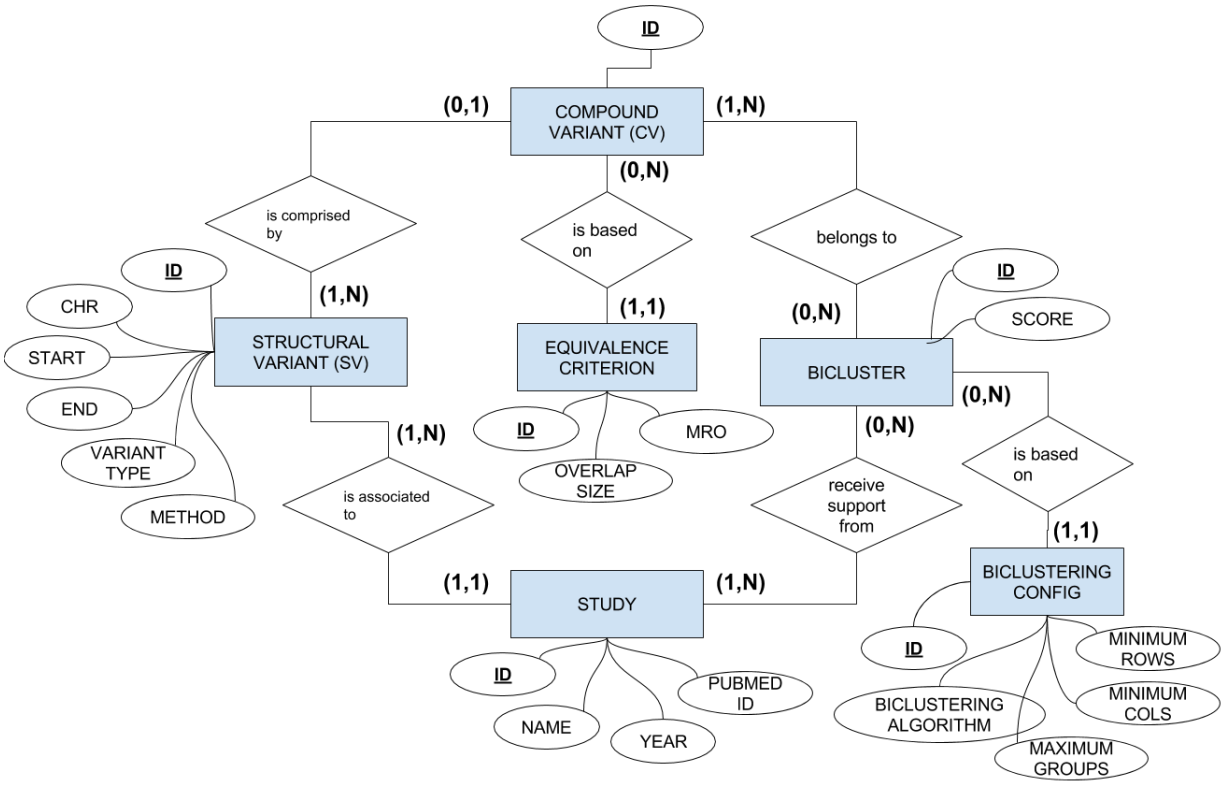


Table B.1: Data summary of ScreenVar-DB

ENTITY	DESCRIPTION	DATA
Structural variants	Structural variants received to be analyzed	565,300 records obtained from DGV
Equivalence Criterion	Criteria defined to cluster structural variants by equivalence condition	6 pre-defined configurations (listed in Table 4.5)
Study	Reference to which each structural variants is associated	72 records from DGV
Compound Variants	A set of equivalent structural variants	55,885 CV generated by ScreenVar
Bicluster	Two-dimensional relation involving a set of structural variants and a set of studies generated according to a given criterion	2,743 biclusters generated by ScreenVar
Biclustering Config	List of values defined for parameters related to biclustering execution	240 pre-defined records (4 biclustering methods x 6 equivalence criterion configurations x 10 iterations)

Table B.2: Complete list of studies cataloged by DGV: all studies present in DGV with respective numbers of variants for each type (deletion, duplication, insertion, inversion, and others)

Study	Method	Sample	Del	Dup	Ins	Inv	Others	Total
1000 GC Phase 1	Merging,Oligo aCGH,PCR,Sequenc	1151	93733	0	0	0	0	93733
1000 GC Phase 3	Sequencing	2504	205923	54605	2	5769	0	266299
1000 GC Pilot Project	Digital array,Oligo aCGH,PCR,S	185	0	1862	7345	21	0	9228
Ahn 2009	Sequencing	1	206	87	0	26	0	319
Alkan 2009	Oligo aCGH,Sequencing	3	18	61	0	0	0	79
Alsmadi 2014	Sequencing	2	2718	775	90	122	0	3705
Altshuler 2010	SNP array	1184	13980	1316	0	0	0	15296
Arlt 2011	Sequencing,SNP array	1	123	3	55	12	0	193
Banerjee 2011	SNP array	1250	61	25	0	0	0	86
Bentley 2008	Sequencing	1	405	0	0	0	0	405
Boomsma 2014	Sequencing	767	865	0	0	0	0	865
Campbell 2011	Oligo aCGH	2366	7050	10340	0	0	0	17390
Coe 2014	Oligo aCGH,SNP array	29084	25629	15118	0	0	565	41312
Conrad 2006	Oligo aCGH,SNP array	60	173	0	0	0	0	173
Conrad 2009	Oligo aCGH	40	2412	2681	0	0	76	5169
Cooper 2008	SNP array	9	31	14	0	0	0	45
Cooper 2011	Oligo aCGH,SNP array	17421	33173	3399	0	0	135	36707
De Smith 2007	Oligo aCGH	51	436	443	0	0	0	879
Dogan 2014	Sequencing	1	552	170	15	0	0	737
Forsberg 2012	SNP array	6	0	1	0	0	0	1
Gusev 2009	SNP array	270	11	0	0	0	0	11
Hinds 2006	Oligo aCGH,PCR	95	5	0	0	0	0	5
Iafrate 2004	BAC aCGH,FISH	39	15	14	0	0	0	29
Itsara 2009	Oligo aCGH,SNP array	1557	1021	270	0	0	0	1291
Jakobsson 2008	SNP array	443	146	51	0	0	0	197
John 2014	Sequencing	1	167	198	56	2	0	423
Ju 2010	Sequencing	1	70	26	0	0	0	96
Kidd 2008	FISH,Multiple complete digesti	9	322	0	465	99	0	886
Kidd 2010	Oligo aCGH,Sequencing	9	0	0	881	0	0	881
Kidd 2010b	Sequencing	9	65	7	0	14	0	86
Kim 2009	Oligo aCGH,Sequencing,SNP arra	2	55	43	1	0	0	99
Korbel 2007	FISH,Oligo aCGH,PCR,Sequencing	2	102	0	69	15	0	186
Levy 2007	Merging,Oligo aCGH,Sequencing,	2	330	2	376	13	0	721
Locke 2006	BAC aCGH	265	72	300	0	0	0	372
Lou 2015	Sequencing,SNP array	369	1181	252	0	0	0	1433
McCarroll 2006	SNP array	269	309	0	0	0	0	309
McCarroll 2008	SNP array	270	3171	754	0	0	0	3925
McKernan 2009	Sequencing	1	394	15	136	10	0	555
Mills 2006	Sequencing	24	271	0	59	0	0	330
Mokhtar 2014	SNP array	34	66	48	0	0	0	114
Pang 2010	Oligo aCGH,Sequencing,SNP arra	3	316	55	108	7	0	486
Pang 2013b	Sequencing	1	126	18	2	0	0	146
Park 2010	Oligo aCGH	31	1476	662	0	0	0	2138
Perry 2008	Oligo aCGH	31	578	1041	0	0	0	1619
Perry 2008b	BAC aCGH,FISH,PCR	62	74	135	0	0	0	209
Pinto 2007	SNP array	771	139	186	0	0	0	325
Redon 2006	BAC aCGH,SNP array	270	1073	1597	0	0	35	2705
Schrider 2013	PCR,Sequencing	946	2	0	3	0	0	5
Schuster 2010	Oligo aCGH,Sequencing	1	2	27	0	0	0	29
Sebat 2004	ROMA	31	8	7	0	0	0	15
Shaikh 2009	SNP array	2026	2642	518	0	0	0	3160
Sharp 2005	BAC aCGH,FISH	48	121	10	0	0	0	131
Simon-Sanchez 2007	qPCR,SNP array	181	2	6	0	0	0	8
Sudmant 2013	Oligo aCGH,Sequencing	97	51	16819	0	0	0	16870
Suktitipat 2014	SNP array	3017	623	123	0	0	0	746
Teague 2010	BAC aCGH,Oligo aCGH,Optical ma	4	113	0	167	2	284	566
Thareja 2015	Sequencing	1	578	172	12	19	0	781
Tuzun 2005	BAC aCGH,PCR,Sequencing	1	11	0	9	2	0	22
Uddin 2014	SNP array	873	4902	803	0	0	0	5705
Vogler 2010	Merging,SNP array	1109	3964	662	0	0	12	4638
Wang 2007	SNP array	112	162	65	0	0	0	227
Wang 2008	Sequencing	1	186	0	0	3	4	193
Wheeler 2008	Oligo aCGH,Sequencing	3	3	2	0	0	0	5
Wong 2007	BAC aCGH	95	752	463	0	0	2	1217
Wong 2012b	Sequencing	96	18775	0	0	0	0	18775
Young 2008	MLPA,PCR,Sequencing	52	8	1	0	0	0	9

Table B.3: Complete methods using in studies cataloged by DGV: all methods used in studies present in DGV with respective numbers of variants for each type (deletion, duplication, insertion, inversion, and others)

Method	Del	Dup	Ins	Inv	Others	Total	
Sequencing	231301	56073	372	4	5965	4	293715
Merging,Oligo aCGH,PCR,Sequenc	93733	0	0	0	0	0	93733
Oligo aCGH,SNP array	59996	18787	0	700	0	700	79483
SNP array	26243	3904	0	0	0	0	30147
Oligo aCGH	11952	15167	0	76	0	76	27195
Oligo aCGH,Sequencing	74	16909	881	0	0	0	17864
Digital array,Oligo aCGH,PCR,S	0	1862	7345	0	21	0	9228
Merging,SNP array	3964	662	0	12	0	12	4638
BAC aCGH,SNP array	1073	1597	0	35	0	35	2705
Sequencing,SNP array	1304	255	55	0	12	0	1626
BAC aCGH	824	763	0	2	0	2	1589
FISH,Multiple complete digesti	322	0	465	0	99	0	886
Merging,Oligo aCGH,Sequencing,	330	2	376	0	13	0	721
Oligo aCGH,Sequencing,SNP arra	371	98	109	0	7	0	585
BAC aCGH,Oligo aCGH,Optical ma	113	0	167	284	2	284	566
BAC aCGH,FISH,PCR	74	135	0	0	0	0	209
FISH,Oligo aCGH,PCR,Sequencing	102	0	69	0	15	0	186
BAC aCGH,FISH	136	24	0	0	0	0	160
BAC aCGH,PCR,Sequencing	11	0	9	0	2	0	22
ROMA	8	7	0	0	0	0	15
MLPA,PCR,Sequencing	8	1	0	0	0	0	9
qPCR,SNP array	2	6	0	0	0	0	8
Oligo aCGH,PCR	5	0	0	0	0	0	5
PCR,Sequencing	2	0	3	0	0	0	5

Table B.4: Number of equivalent variants by DGV studies: Totals of compound variants found after executing ScreenVar for each defined equivalence criterion (EQ1 to EQ6)

Study	Total in Chr	EQ 1	EQ 2	EQ 3	EQ 4	EQ 5	EQ 6
1000 GC Phase 3	266299	679 (0,25%)	392 (0,15%)	1407 (0,53%)	1160 (0,44%)	1652 (0,62%)	1450 (2%)
1000 GC Phase 1	93733	490 (0,52%)	483 (0,52%)	1209 (1,29%)	1203 (1,28%)	1396 (1,49%)	1390 (2%)
Coe 2014	41312	3175 (7,69%)	1793 (4,34%)	4600 (11,13%)	3411 (8,26%)	5080 (12,3%)	3905 (2%)
Cooper 2011	36707	945 (2,57%)	431 (1,17%)	3360 (9,15%)	2993 (8,15%)	4362 (11,88%)	4019 (2%)
Wong 2012b	18775	9248 (49,26%)	9246 (49,25%)	15701 (83,63%)	15699 (83,62%)	16949 (90,27%)	16947 (2%)
Campbell 2011	17390	28 (0,16%)	11 (0,06%)	62 (0,36%)	48 (0,28%)	80 (0,46%)	69 (2%)
Sudmant 2013	16870	9 (0,05%)	8 (0,05%)	74 (0,44%)	73 (0,43%)	224 (1,33%)	223 (2%)
Altshuler 2010	15296	40 (0,26%)	17 (0,11%)	56 (0,37%)	40 (0,26%)	63 (0,41%)	54 (2%)
1000 GC Pilot Project	9228	68 (0,74%)	66 (0,72%)	267 (2,89%)	267 (2,89%)	428 (4,64%)	428 (2%)
Uddin 2014	5705	291 (5,1%)	256 (4,49%)	536 (9,4%)	507 (8,89%)	683 (11,97%)	654 (2%)
Conrad 2009	5169	1034 (20%)	878 (16,99%)	2366 (45,77%)	2269 (43,9%)	2749 (53,18%)	2666 (2%)
Vogler 2010	4638	499 (10,76%)	348 (7,5%)	770 (16,6%)	650 (14,01%)	919 (19,81%)	813 (2%)
McCarroll 2008	3925	44 (1,12%)	22 (0,56%)	69 (1,76%)	48 (1,22%)	84 (2,14%)	66 (2%)
Alsmadi 2014	3705	261 (7,04%)	260 (7,02%)	599 (16,17%)	598 (16,14%)	846 (22,83%)	845 (2%)
Shaikh 2009	3160	76 (2,41%)	35 (1,11%)	261 (8,26%)	221 (6,99%)	483 (15,28%)	447 (2%)
Redon 2006	2705	291 (10,76%)	147 (5,43%)	497 (18,37%)	388 (14,34%)	600 (22,18%)	502 (2%)
Park 2010	2138	132 (6,17%)	88 (4,12%)	358 (16,74%)	317 (14,83%)	458 (21,42%)	419 (2%)
Perry 2008	1619	153 (9,45%)	103 (6,36%)	430 (26,56%)	389 (24,03%)	625 (38,6%)	588 (2%)
Lou 2015	1433	60 (4,19%)	44 (3,07%)	115 (8,03%)	101 (7,05%)	160 (11,17%)	148 (2%)
Itsara 2009	1291	2 (0,15%)	2 (0,15%)	4 (0,31%)	4 (0,31%)	9 (0,7%)	9 (2%)
Wong 2007	1217	81 (6,66%)	2 (0,16%)	102 (8,38%)	27 (2,22%)	173 (14,22%)	108 (2%)
Kidd 2008	886	24 (2,71%)	24 (2,71%)	164 (18,51%)	164 (18,51%)	329 (37,13%)	329 (2%)
Kidd 2010	881	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (2%)
de Smith 2007	879	91 (10,35%)	71 (8,08%)	178 (20,25%)	165 (18,77%)	221 (25,14%)	210 (2%)
Boomsma 2014	865	48 (5,55%)	47 (5,43%)	382 (44,16%)	382 (44,16%)	546 (63,12%)	546 (2%)
Thareja 2015	781	117 (14,98%)	117 (14,98%)	299 (38,28%)	299 (38,28%)	436 (55,83%)	436 (2%)
Suktitipat 2014	746	73 (9,79%)	51 (6,84%)	154 (20,64%)	147 (19,71%)	219 (29,36%)	213 (2%)
Dogan 2014	737	94 (12,75%)	94 (12,75%)	258 (35,01%)	258 (35,01%)	353 (47,9%)	353 (2%)
Levy 2007	721	45 (6,24%)	45 (6,24%)	106 (14,7%)	106 (14,7%)	132 (18,31%)	132 (2%)
Teague 2010	566	36 (6,36%)	32 (5,65%)	120 (21,2%)	116 (20,49%)	179 (31,63%)	177 (2%)
McKernan 2009	555	14 (2,52%)	14 (2,52%)	69 (12,43%)	69 (12,43%)	153 (27,57%)	153 (2%)
Pang 2010	486	74 (15,23%)	57 (11,73%)	149 (30,66%)	143 (29,42%)	190 (39,09%)	186 (2%)
John 2014	423	42 (9,93%)	42 (9,93%)	160 (37,83%)	160 (37,83%)	214 (50,59%)	214 (2%)
Bentley 2008	405	13 (3,21%)	13 (3,21%)	154 (38,02%)	154 (38,02%)	258 (63,7%)	258 (2%)
Locke 2006	372	7 (1,88%)	6 (1,61%)	10 (2,69%)	9 (2,42%)	15 (4,03%)	14 (2%)
Mills 2006	330	61 (18,48%)	61 (18,48%)	74 (22,42%)	74 (22,42%)	86 (26,06%)	86 (2%)
Pinto 2007	325	101 (31,08%)	96 (29,54%)	180 (55,38%)	175 (53,85%)	207 (63,69%)	202 (2%)
Ahn 2009	319	12 (3,76%)	12 (3,76%)	101 (31,66%)	101 (31,66%)	187 (58,62%)	187 (2%)
McCarroll 2006	309	3 (0,97%)	2 (0,65%)	10 (3,24%)	10 (3,24%)	25 (8,09%)	25 (2%)
Wang 2007	227	13 (5,73%)	5 (2,2%)	23 (10,13%)	17 (7,49%)	37 (16,3%)	32 (2%)
Perry 2008b	209	16 (7,66%)	6 (2,87%)	23 (11%)	14 (6,7%)	31 (14,83%)	22 (2%)
Jakobsson 2008	197	22 (11,17%)	16 (8,12%)	49 (24,87%)	44 (22,34%)	87 (44,16%)	83 (2%)
Arlt 2011	193	8 (4,15%)	8 (4,15%)	56 (29,02%)	56 (29,02%)	101 (52,33%)	101 (2%)
Wang 2008	193	12 (6,22%)	12 (6,22%)	71 (36,79%)	71 (36,79%)	163 (84,46%)	163 (2%)
Korbel 2007	186	6 (3,23%)	6 (3,23%)	71 (38,17%)	71 (38,17%)	99 (53,23%)	99 (2%)
Conrad 2006	173	5 (2,89%)	4 (2,31%)	22 (12,72%)	22 (12,72%)	56 (32,37%)	56 (2%)
Pang 2013b	146	34 (23,29%)	34 (23,29%)	55 (37,67%)	55 (37,67%)	67 (45,89%)	67 (2%)
Sharp 2005	131	12 (9,16%)	5 (3,82%)	12 (9,16%)	5 (3,82%)	14 (10,69%)	7 (2%)
Mokhtar 2014	114	18 (15,79%)	16 (14,04%)	25 (21,93%)	23 (20,18%)	28 (24,56%)	26 (2%)
Kim 2009	99	6 (6,06%)	6 (6,06%)	34 (34,34%)	34 (34,34%)	40 (40,4%)	40 (2%)
Ju 2010	96	68 (70,83%)	13 (13,54%)	92 (95,83%)	51 (53,13%)	96 (100%)	58 (2%)
Banerjee 2011	86	51 (59,3%)	1 (1,16%)	51 (59,3%)	2 (2,33%)	53 (61,63%)	5 (2%)
Kidd 2010b	86	41 (47,67%)	40 (46,51%)	52 (60,47%)	51 (59,3%)	56 (65,12%)	55 (2%)
Alkan 2009	79	8 (10,13%)	5 (6,33%)	15 (18,99%)	13 (16,46%)	20 (25,32%)	18 (2%)
Cooper 2008	45	4 (8,89%)	3 (6,67%)	15 (33,33%)	14 (31,11%)	23 (51,11%)	22 (2%)
Iafrate 2004	29	2 (6,9%)	0 (0%)	2 (6,9%)	1 (3,45%)	2 (6,9%)	1 (2%)
Schuster 2010	29	1 (3,45%)	1 (3,45%)	7 (24,14%)	7 (24,14%)	12 (41,38%)	12 (2%)
Tuzun 2005	22	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (9,09%)	2 (2%)
Sebat 2004	15	1 (6,67%)	1 (6,67%)	4 (26,67%)	4 (26,67%)	6 (40%)	6 (2%)
Gusev 2009	11	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (18,18%)	2 (2%)
Young 2008	9	1 (11,11%)	0 (0%)	1 (11,11%)	0 (0%)	1 (11,11%)	0 (2%)
Simon-Sanchez 2007	8	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (12,5%)	1 (2%)
Hinds 2006	5	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (60%)	3 (2%)
Schrider 2013	5	0 (0%)	0 (0%)	1 (20%)	1 (20%)	1 (20%)	1 (2%)
Wheeler 2008	5	1 (20%)	1 (20%)	5 (100%)	5 (100%)	5 (100%)	5 (2%)
Forsberg 2012	1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (2%)

Table B.5: Number of equivalent variants by support levels: All values found for each support level and presents respective totals of equivalent variants

Support level	EQ 1	EQ 2	EQ 3	EQ 4	EQ 5	EQ 6
0	11020	8765	10997	10148	7871	7503
1	4236	3585	7626	7253	8085	7967
2	1372	1217	3682	3360	3631	3512
3	662	595	2858	2650	2575	2449
4	490	448	2420	2231	2082	2038
5	359	356	2138	2000	2075	1960
6	273	270	983	912	2595	2585
7	123	116	1043	1031	1887	1820
8	149	146	554	533	1488	1393
9	1	1	898	898	1285	1144
10	0	0	439	437	987	1044
11	1	91	688	549	824	663
12	92	8	516	413	495	513
13	11	1	259	273	889	849
14	2	0	180	138	829	782
15	0	0	189	178	836	625
16	0	0	114	102	528	470
17	0	0	39	30	726	669
18	0	0	12	10	431	383
19	0	0	5	5	415	232
20	0	0	4	8	218	203
21	0	0	17	11	176	159
22	0	0	25	16	229	213
23	0	0	2	2	237	104
24	0	0	7	24	216	163
25	0	0	49	10	221	207
26	0	0	58	115	116	82
27	0	0	147	50	92	58
28	0	0	18	118	51	25
29	0	0	129	1	140	139
30	0	0	1	0	230	176
31	0	0	0	0	45	228
32	0	0	0	0	274	0

Table B.6: Number of SSV in DGV-GS by studies: All studies present in DGV with respective numbers of SSV in input data set, the number of SSV in DGV-GS along with the proportion between both these values

Study	Total SSV	SSV in GS-DGV
McCarroll 2008	3925	3837 (97,76%)
Ju 2010	96	91 (94,79%)
Vogler 2010	4638	4311 (92,95%)
1000 GC Phase 1	93733	85568 (91,29%)
Wang 2008	193	172 (89,12%)
Altshuler 2010	15296	13441 (87,87%)
Coe 2014	41312	35629 (86,24%)
1000 GC Phase 3	266299	223686 (84%)
Mokhtar 2014	114	89 (78,07%)
Alkan 2009	79	59 (74,68%)
Campbell 2011	17390	12715 (73,12%)
Boomsma 2014	865	611 (70,64%)
Park 2010	2138	1508 (70,53%)
Wong 2012b	18775	12808 (68,22%)
Cooper 2011	36707	24648 (67,15%)
Bentley 2008	405	271 (66,91%)
Cooper 2008	45	29 (64,44%)
Kidd 2010b	86	54 (62,79%)
Conrad 2009	5169	3208 (62,06%)
Ahn 2009	319	188 (58,93%)
Perry 2008	1619	904 (55,84%)
Schuster 2010	29	14 (48,28%)
Arlt 2011	193	93 (48,19%)
Dogan 2014	737	325 (44,1%)
Uddin 2014	5705	2362 (41,4%)
Kim 2009	99	40 (40,4%)
Mills 2006	330	103 (31,21%)
McKernan 2009	555	167 (30,09%)
Levy 2007	721	161 (22,33%)
Pang 2010	486	68 (13,99%)
Sudmant 2013	16870	1587 (9,41%)
1000 GC Pilot Project	9228	0 (0%)
Alsmadi 2014	3705	0 (0%)
Banerjee 2011	86	0 (0%)
Conrad 2006	173	0 (0%)
de Smith 2007	879	0 (0%)
Forsberg 2012	1	0 (0%)
Gusev 2009	11	0 (0%)
Hinds 2006	5	0 (0%)
Iafrate 2004	29	0 (0%)
Itsara 2009	1291	0 (0%)
Jakobsson 2008	197	0 (0%)
John 2014	423	0 (0%)
Kidd 2008	886	0 (0%)
Kidd 2010	881	0 (0%)
Korbel 2007	186	0 (0%)
Locke 2006	372	0 (0%)
Lou 2015	1433	0 (0%)
McCarroll 2006	309	0 (0%)
Pang 2013b	146	0 (0%)
Perry 2008b	209	0 (0%)
Pinto 2007	325	0 (0%)
Redon 2006	2705	0 (0%)
Schrider 2013	5	0 (0%)
Sebat 2004	15	0 (0%)
Shaikh 2009	3160	0 (0%)
Sharp 2005	131	0 (0%)
Simon-Sanchez 2007	8	0 (0%)
Suktitipat 2014	746	0 (0%)
Teague 2010	566	0 (0%)
Thareja 2015	781	0 (0%)
Tuzun 2005	22	0 (0%)
Wang 2007	227	0 (0%)
Wheeler 2008	5	0 (0%)
Wong 2007	1217	0 (0%)
Young 2008	9	0 (0%)

Table B.7: List of supporting structural variants of the compound variant with greatest support level (32 studies): All SSV achieved with the highest value for support level, in case of the maximum value was 32 studies

Studies	Equivalent SSV (members of a single compound variant)
1000 GC Phase 1, 1000 GC Phase 3, Ahn 2009, Alsmadi 2014, Altshuler 2010, Arlt 2011, Boomsma 2014, Campbell 2011, Coe 2014, Conrad 2006, Conrad 2009, Cooper 2011, de Smith 2007, Dogan 2014, Ju 2010, Kidd 2008, Kidd 2010b, Kim 2009, Korbel 2007, Lou 2015, McCarroll 2006, McKernan 2009, Mokhtar 2014, Pang 2010, Pang 2013b, Park 2010, Perry 2008, Sudmant 2013, Teague 2010, Uddin 2014, Vogler 2010, Wang 2008, Wong 2012b	essv5397580, essv9967976, essv31631, nssv3989690, nssv3994264, essv5002005, nssv626339, essv9762651, nssv2858524, nssv3462749, nssv3462770, nssv3462777, nssv3462784, nssv3462864, nssv3462869, nssv3462885, nssv3462979, nssv3463022, nssv3463261, nssv3463449, nssv3463751, nssv3463825, nssv3463947, nssv3464253, nssv3464717, nssv3465014, nssv3465249, nssv3465450, nssv3465830, nssv3466353, nssv3466551, nssv3466556, nssv3467025, nssv3467054, nssv3467155, nssv3467302, nssv3467366, nssv3467517, nssv3467565, nssv3467575, nssv3467624, nssv3467737, nssv3467770, nssv3467790, nssv3467970, nssv3469289, nssv3469589, nssv3469641, nssv3470453, nssv3470554, nssv3470856, nssv3473002, nssv3473024, nssv3473557, nssv3474331, nssv3474675, nssv3475184, nssv3475865, nssv3475906, nssv3475934, nssv3475962, nssv3476348, nssv3476366, nssv3476740, nssv3476802, nssv3477328, nssv3478970, nssv3479077, nssv3479987, nssv3480458, nssv3480561, nssv3481107, nssv3699284, nssv3701280, nssv3701281, nsv1001710, nsv1004356, nsv1004497, nsv1004974, nsv1005131, nsv1005581, nsv1007261, nsv1008865, nsv1010149, nsv1011218, nsv1011940, nsv1012858, nsv1014678, nsv997354, nsv997915, nsv998810, nssv467098, nssv467102, essv33962, essv35698, essv38799, essv38946, essv48139, essv49845, essv55301, essv55735, essv65461, essv70883, essv73288, esv15735, nssv716273, nssv716274, nssv716472, nssv716526, nssv716536, nssv716540, nssv716541, nssv716542, nssv716546, nssv716553, nssv716561, nssv716616, nssv716641, nssv716642, nssv716644, nssv716663, nssv716673, nssv716688, nssv716690, nssv716730, nssv716766, nssv717424, nssv717426, nssv717428, nssv717437, nssv717441, nsv546552, nsv546553, essv101386, essv101502, essv92699, essv94191, essv96805, essv99457, nssv2997160, nssv3002290, nssv1420391, nssv10026, nssv2178, nssv4300, nssv9515, nssv585384, nssv1418741, nssv466029, nssv466040, nssv4027312, nssv4027490, nssv4027501, nssv471557, essv5275117, essv5369149, essv9838628, essv3565326, essv7100023, nssv1421457, nssv1423530, nssv1430349, nssv12095, nssv12777, nssv2758365, nssv618632, nssv622766, essv9768990, essv9769012, essv9769023, essv9769056, essv9769101, essv9769156, essv9769189, essv9769267, essv9769289, essv9769367, essv9769534, essv9769711, essv9769789, essv9771256, essv9771289, essv7005437, essv7005447, essv7005463, essv7005565, essv7005568, essv7028879, essv7028890, essv7030490, essv7030845, essv7030857, essv7031812, essv7031834, essv7031956, essv26730, essv6666931, essv6669034, essv6669654, essv6675252, essv6679222, essv6682870, essv6689303, essv6692829, essv6696087, essv6697025, essv6703903, essv6707301, essv6714295, essv6718205, essv6719276, essv6722049, essv6725900, essv6733526, essv6736018, essv6738858, essv6742190, essv6744984, essv6747827, essv6749020, essv6750647, essv6753544, essv6756591, essv6761885, essv6764203, essv6766588, essv6769578, essv6773493, essv6776980, essv6780938, essv6785044, essv6789207, essv6797509, essv6801653, essv6804446, essv6804931, essv6807418, essv6810380, essv6813207, essv6816646, essv6817120, essv6821239, essv6829056, essv6832648, essv6836236, essv6843935, essv6844288, essv6853314, essv6859251, essv6864017, essv6868756, essv6871756, essv6877265, essv6877700, essv6880465, essv6886022, essv6889051, essv6892370, essv6895888, essv6898646, essv6902584, essv6914037, essv6917174, essv6920150, essv6921783, essv6925739, essv6929123, essv6937731, essv6941847, essv6950615, essv6950662, essv6954830, essv6961646, essv6967989

Table B.8: Supporting structural variants associated to a well-evaluated bicluster (ID=29, 25 CV x 8 studies) (1/4): All 588 SSV that were identified as part of a same bicluster, having as support the following studies: Vogler 2010, Suktitipat 2014, Redon 2006, de Smith 2007, Lou 2015, Cooper 2011, Conrad 2009, Coe 2014

CV ID	Min start	Max end	List equivalent SSV
CV-285	12834253	12934346	essv100377, essv101326, essv101537, essv1569, essv2359, essv25781141, essv25781694, essv25787832, essv25787908, essv25787995, essv5455860, essv6560, essv6990745, essv6991078, essv6991744, essv6991967, essv7004977, essv93105, essv93282, essv95999, essv9856931, esv16950, nssv1173678, nssv1173679, nssv1173680, nssv1173681, nssv1421443, nssv1424726, nssv1431793, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464712, nssv3465020, nssv3465164, nssv3466322, nssv3468171, nssv3471035, nssv3473439, nssv3473824, nssv3476384, nssv3478201, nssv3478916, nssv3480618, nssv3480932, nssv4029064, nssv710027, nssv710031, nssv710033, nssv710039, nssv710040
CV-289	12835868	12934346	essv100377, essv101326, essv101537, essv1080, essv1569, essv2359, essv25781141, essv25781694, essv25787832, essv25787908, essv25787995, essv5455860, essv6560, essv6990745, essv6991078, essv6991744, essv6991967, essv7004977, essv7033189, essv93105, essv93282, essv95999, essv9856931, esv16950, nssv1173680, nssv1173681, nssv1421068, nssv1421443, nssv1424726, nssv1431793, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464481, nssv3464712, nssv3465020, nssv3465164, nssv3466322, nssv3468171, nssv3471035, nssv3473439, nssv3473824, nssv3476384, nssv3478201, nssv3478916, nssv3479202, nssv3480618, nssv3698020, nssv4029064, nssv710031, nssv710033, nssv710039, nssv710040
CV-296	12838732	12934346	essv100377, essv101326, essv101537, essv1080, essv1569, essv2359, essv25779746, essv25780439, essv25781141, essv25781694, essv25787832, essv25787908, essv25787995, essv25791015, essv5455860, essv6560, essv6990745, essv6991078, essv6991744, essv6991967, essv7004977, essv7033189, essv93105, essv93282, essv95999, essv9856931, esv16950, nssv1421068, nssv1421443, nssv1424726, nssv1431793, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464481, nssv3464712, nssv3465020, nssv3465164, nssv3466322, nssv3468171, nssv3471035, nssv3473439, nssv3473824, nssv3476384, nssv3478201, nssv3478916, nssv3480618, nssv3698020, nssv4029064, nssv547699, nssv710031, nssv710033, nssv710039, nssv710040, nssv710080, nssv710082
CV-300	12839440	12934346	essv100377, essv101326, essv101537, essv1569, essv2359, essv25780439, essv25781141, essv25781694, essv25787832, essv25787908, essv25787995, essv25791015, essv5455860, essv6560, essv6990745, essv6991078, essv6991744, essv6991967, essv7004977, essv93105, essv93282, essv95999, essv9856931, esv16950, nssv1421443, nssv1424726, nssv1431793, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464481, nssv3464712, nssv3465020, nssv3465164, nssv3466322, nssv3468171, nssv3471035, nssv3473439, nssv3478201, nssv3480618, nssv3481472, nssv3698020, nssv4029064, nssv710031, nssv710033, nssv710039, nssv710040, nssv710080, nssv710082
CV-303	12839564	12934346	essv101537, essv1080, essv1569, essv2359, essv25779746, essv25780439, essv25781141, essv25781694, essv25787832, essv25787908, essv25787995, essv25789560, essv25791015, essv5455860, essv6560, essv6990745, essv6991078, essv6991744, essv6991967, essv7004962, essv7004977, essv7033189, essv93282, essv95999, essv9856931, esv16950, nssv1173688, nssv1173690, nssv1421068, nssv1421443, nssv1424726, nssv1431793, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464481, nssv3464712, nssv3465020, nssv3465164, nssv3466322, nssv3468171, nssv3471035, nssv3473439, nssv3478201, nssv3480618, nssv3698020, nssv4028323, nssv4029064, nssv547699, nssv710031, nssv710033, nssv710039, nssv710040, nssv710080, nssv710082
CV-308	12839977	12920040	essv2359, essv25781141, essv25781234, essv25787832, essv25787908, essv25787995, essv6560, essv7004962, essv7004977, essv93282, essv95999, essv9838624, esv16950, nssv1424726, nssv1426371, nssv1431793, nssv1437927, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464712, nssv3465020, nssv3465164, nssv3468171, nssv3471035, nssv3473439, nssv3474437, nssv4028323, nssv4029064, nssv710031, nssv710033, nssv710040
CV-313	12841915	12918674	essv1080, essv2359, essv25781141, essv25781234, essv25787832, essv25787908, essv25787995, essv6560, essv7004962, essv7004977, essv95999, essv9838624, esv16950, nssv1426371, nssv1437927, nssv1440799, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464712, nssv3464724, nssv3465020, nssv3465164, nssv3468171, nssv3471035, nssv3473439, nssv3474437, nssv4028323, nssv4029064, nssv710030, nssv710031, nssv710033, nssv710040
CV-1174	148916177	149459615	essv101256, essv14154, essv21629, essv25788953, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv6431, essv6996122, essv6996133, essv6996155, essv6996600, essv6996611, essv6996622, essv6997489, essv6997500, essv6997722, essv7006218, essv7006220, essv7006226, essv7006297, essv7006309, essv96191, essv9837877, nssv21402, nssv25970, nssv3482912, nssv3482964, nssv3483427, nssv3483540, nssv3483541, nssv3483633, nssv3483897, nssv3483927, nssv3483946, nssv3484265, nssv3484476, nssv3485358, nssv3486035, nssv3486170, nssv3486259, nssv3486287, nssv3486784, nssv3488039, nssv3488053, nssv3488077, nssv3489566, nssv3490081, nssv3490111, nssv3490819, nssv3493105, nssv3493789, nssv3494550, nssv3494893, nssv3495361, nssv3498902, nssv3500097, nssv3500770, nssv3500897, nssv3501870, nssv3702228, nssv3703992, nssv3703993, nssv3704088, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv451753, nssv723095, nssv723125, nssv723126, nssv723129, nssv723244, nssv723247, nssv723249, nssv723251, nssv723269, nsv1003007
CV-1175	148916177	149521828	essv101256, essv21629, essv25788953, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv6431, essv6996122, essv6996133, essv6996155, essv6996600, essv6996611, essv6996622, essv6997489, essv6997500, essv6997722, essv7006218, essv7006220, essv7006226, essv7006309, nssv21402, nssv25970, nssv3482912, nssv3482964, nssv3483427, nssv3483541, nssv3483927, nssv3484476, nssv3485358, nssv3485496, nssv3486035, nssv3486170, nssv3486259, nssv3486287, nssv3488053, nssv3488077, nssv3488713, nssv3489566, nssv3490081, nssv3490111, nssv3490819, nssv3493789, nssv3494893, nssv3495361, nssv3498902, nssv3500097, nssv3500770, nssv3500897, nssv3501870, nssv3704088, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv723095, nssv723125, nssv723126, nssv723129, nssv723244, nssv723247, nssv723249, nssv723251, nssv723269, nssv723340, nssv723345, nssv723346, nsv1003007
CV-1176	148916177	149521846	essv101256, essv21629, essv25792902, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv6431, essv6996122, essv6996155, essv6996611, essv6996622, essv6996811, essv6997489, essv6997500, essv6997722, essv7006218, essv7006220, essv7006226, essv7006309, nssv21402, nssv25970, nssv3482912, nssv3482964, nssv3483427, nssv3483541, nssv3483927, nssv3484476, nssv3485358, nssv3485496, nssv3486170, nssv3486287, nssv3488053, nssv3488077, nssv3488713, nssv3490081, nssv3490111, nssv3490819, nssv3493789, nssv3494893, nssv3495361, nssv3498902, nssv3500097, nssv3500770, nssv3500897, nssv3501870, nssv3704088, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv723095, nssv723125, nssv723126, nssv723129, nssv723244, nssv723247, nssv723249, nssv723251, nssv723269, nssv723340, nssv723345, nssv723346, nsv1003007
CV-1204	148947698	149732729	essv100936, essv1715, essv19363, essv19885, essv20523, essv25790783, essv25792902, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv5662, essv6985823, essv6996155, essv6996622, essv6996811, essv6996844, essv7006228, essv7006251, essv7006252, essv97263, nssv23694, nssv25699, nssv27030, nssv3483916, nssv3484219, nssv3484736, nssv3486792, nssv3488304, nssv3488713, nssv3490111, nssv3490114, nssv3491002, nssv3493160, nssv3493254, nssv3493487, nssv3494530, nssv3496414, nssv3497834, nssv3499738, nssv3704174, nssv4029809, nssv723095
CV-1244	149024802	149432369	essv11482, essv14239, essv18916, essv19038, essv25779316, essv25788953, essv5001910, essv6981512, essv6982967, essv6996511, essv6996600, essv6996611, essv6997044, essv6997055, essv6997066, essv6997244, essv6997366, essv6997434, essv6997445, essv6997489, essv6997645, essv6997656, essv6997678, essv6997711, essv7006268, essv7006297, essv78072, essv7932, essv942, essv94654, essv9837828, esv16514, nssv1795806, nssv21402, nssv25970, nssv2849373, nssv3482753, nssv3482792, nssv3482797, nssv3483131, nssv3483222, nssv3483260, nssv3483325, nssv3483346, nssv3483397, nssv3483423, nssv3483582, nssv3483642, nssv3483897, nssv3484091, nssv3484107, nssv3484217, nssv3484538, nssv3484544, nssv3484771, nssv3484863, nssv3485169, nssv3485358, nssv3485570, nssv3485825, nssv3485952, nssv3485997, nssv3486287, nssv3486345, nssv3486424, nssv3486621, nssv3486746, nssv3486972, nssv3487653, nssv3487780, nssv3488077, nssv3488099, nssv3488871, nssv3489845, nssv3490028, nssv3490029, nssv3490081, nssv3490123, nssv3490202, nssv3490319, nssv3490665, nssv3490750, nssv3490819, nssv3490997, nssv3491137, nssv3491620, nssv3491812, nssv3492029, nssv3492689, nssv3492993, nssv3493913, nssv3494893, nssv3494978, nssv3495023, nssv3495361, nssv3495407, nssv3495467, nssv3495835, nssv3495925, nssv3496262, nssv3498191, nssv3498902, nssv3498972, nssv3500322, nssv3501135, nssv3702635, nssv3702636, nssv3703992, nssv3703993, nssv3704088, nssv3704115, nssv4029794, nssv4029797, nssv4029798, nssv4029799, nssv4029802, nssv4029805, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv4029812, nssv723111, nssv723124, nssv723125, nssv723126, nssv723129, nssv723212, nssv723240, nssv723243, nssv723244, nssv723247, nssv723249, nssv723251, nssv723259, nssv723269, nssv723273, nssv723336, nssv723340, nssv723345, nssv723346, nssv723364, nssv723428, nsv1003007

Eight studies give support to this bicluster: Coe 2014, Conrad 2009, Cooper 2011, de Smith 2007, Lou 2015, Redon 2006, Suktitipat 2014, Vogler 2010

Table B.9: Supporting structural variants associated to a well-evaluated bicluster (ID=29, 25 CV x 8 studies) (2/4): All 588 SSV that were identified as part of a same bicluster, having as support the following studies: Vogler 2010, Suktitipat 2014, Redon 2006, de Smith 2007, Lou 2015, Cooper 2011, Conrad 2009, Coe 2014

CV ID	Min start	Max end	List equivalent SSV
CV-1245	149024802	149459615	essv14239, essv18916, essv25779316, essv25788953, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv5001910, essv6996600, essv6996611, essv6996622, essv6997244, essv6997366, essv6997434, essv6997445, essv6997489, essv6997500, essv6997678, essv6997711, essv6997722, essv7006297, essv7006309, essv942, essv94654, essv9837839, nssv1449962, nssv1795806, nssv21402, nssv25970, nssv3482912, nssv3483350, nssv3483582, nssv3483642, nssv3483897, nssv3484091, nssv3484217, nssv3484544, nssv3484771, nssv3485358, nssv3486287, nssv3486424, nssv3486746, nssv3486764, nssv3487780, nssv3488077, nssv3488099, nssv3488871, nssv3489845, nssv3490028, nssv3490029, nssv3490081, nssv3490202, nssv3490819, nssv3493913, nssv3494893, nssv3495361, nssv3498902, nssv3498972, nssv3500322, nssv3501135, nssv3501870, nssv3703992, nssv3703993, nssv3704088, nssv4029797, nssv4029798, nssv4029799, nssv4029802, nssv4029805, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv4029812, nssv723124, nssv723125, nssv723126, nssv723129, nssv723240, nssv723243, nssv723244, nssv723247, nssv723249, nssv723251, nssv723269, nssv723336, nssv723340, nssv723345, nssv723346, nssv723364, nssv723428, nsv1003007
CV-1248	149024802	149732729	essv100936, essv1715, essv19363, essv19885, essv20523, essv25790783, essv25792902, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv5662, essv6985823, essv6996811, essv6996844, essv6997489, essv6997500, essv6997722, essv7006251, essv7006252, essv7006309, essv97263, nssv21402, nssv23694, nssv25699, nssv25970, nssv27030, nssv3482912, nssv3483350, nssv3484219, nssv3485358, nssv3486287, nssv3486764, nssv3486792, nssv3490081, nssv3490114, nssv3490819, nssv3491002, nssv3493160, nssv3493254, nssv3493487, nssv3494530, nssv3494544, nssv3495361, nssv3496414, nssv3497834, nssv3499738, nssv3501870, nssv3704174, nssv3704205, nssv4029807, nssv4029808, nssv4029809, nssv723126, nssv723129, nssv723249, nssv723251, nssv723269, nssv723346
CV-1249	149024802	149773350	essv100936, essv1715, essv19363, essv19885, essv20523, essv25790783, essv25792902, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv5662, essv6985823, essv6996844, essv6997500, essv6997534, essv6997722, essv7006251, essv7006252, essv7006306, essv7006309, essv97263, nssv21402, nssv23694, nssv25699, nssv25970, nssv27030, nssv3482912, nssv3482965, nssv3484219, nssv3484555, nssv3484570, nssv3486792, nssv3488597, nssv3489328, nssv3490081, nssv3490114, nssv3491002, nssv3493160, nssv3493254, nssv3493487, nssv3494259, nssv3494530, nssv3496414, nssv3497834, nssv3499738, nssv3501870, nssv3704174, nssv3704182, nssv3704203, nssv3704208, nssv4029808, nssv4029809, nssv547768, nssv723129, nsv1000267
CV-1256	149024808	149459615	essv14239, essv18916, essv25779316, essv25788953, essv33321, essv33753, essv35957, essv38722, essv38955, essv40849, essv41126, essv43839, essv44471, essv45569, essv5001910, essv6997244, essv6997366, essv6997434, essv6997445, essv6997489, essv6997500, essv6997678, essv6997711, essv6997722, essv7006268, essv7006297, essv7006309, essv942, essv94654, essv9837839, nssv1449962, nssv1795806, nssv21402, nssv25970, nssv3482797, nssv3482912, nssv3483350, nssv3483582, nssv3483642, nssv3483897, nssv3484091, nssv3484217, nssv3484544, nssv3484771, nssv3485358, nssv3486287, nssv3486424, nssv3486746, nssv3486764, nssv3487780, nssv3488077, nssv3488099, nssv3488871, nssv3489845, nssv3490028, nssv3490029, nssv3490081, nssv3490202, nssv3490819, nssv3493913, nssv3494893, nssv3495361, nssv3498902, nssv3498972, nssv3500322, nssv3501135, nssv3501870, nssv3703992, nssv3703993, nssv3704088, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv4029812, nssv723124, nssv723125, nssv723126, nssv723129, nssv723240, nssv723243, nssv723244, nssv723247, nssv723249, nssv723251, nssv723269, nssv723336, nssv723340, nssv723345, nssv723346, nssv723364, nssv723428, nsv1003007
CV-1318	149036512	149334231	essv11482, essv18916, essv19038, essv25779316, essv25788983, essv25798108, essv5001910, essv55468, essv6981512, essv6982967, essv6989889, essv6996867, essv6996911, essv6996922, essv6996933, essv6996978, essv6997044, essv6997055, essv6997066, essv6997244, essv6997366, essv6997434, essv6997445, essv6997545, essv6997556, essv6997578, essv6997589, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv6997733, essv7006261, essv7006268, essv78072, essv7932, essv92827, essv94654, essv9837828, nssv1173153, nssv1434888, nssv3482753, nssv3482792, nssv3482797, nssv3482913, nssv3483041, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483325, nssv3483397, nssv3483443, nssv3483503, nssv3483582, nssv3483642, nssv3483825, nssv3484023, nssv3484060, nssv3484091, nssv3484107, nssv3484217, nssv3484247, nssv3484414, nssv3484538, nssv3484544, nssv3484718, nssv3484771, nssv3484863, nssv3484907, nssv3484933, nssv3485085, nssv3485169, nssv3485324, nssv3485432, nssv3485570, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3485998, nssv3486028, nssv3486203, nssv3486345, nssv3486621, nssv3486746, nssv3486814, nssv3486821, nssv3486940, nssv3486972, nssv3487034, nssv3487555, nssv3487569, nssv3487653, nssv3487780, nssv3487823, nssv3487908, nssv3488871, nssv3489491, nssv3489748, nssv3489778, nssv3489845, nssv3490028, nssv3490029, nssv3490046, nssv3490123, nssv3490202, nssv3490665, nssv3490996, nssv3490997, nssv3491137, nssv3491317, nssv3491428, nssv3491473, nssv3491620, nssv3491812, nssv3491915, nssv3492689, nssv3492993, nssv3493189, nssv3493913, nssv3494028, nssv3494285, nssv3494348, nssv3494815, nssv3494978, nssv3495023, nssv3495155, nssv3495296, nssv3495346, nssv3495393, nssv3495407, nssv3495423, nssv3495467, nssv3495835, nssv3495925, nssv3496262, nssv3497988, nssv3498229, nssv3498972, nssv3499009, nssv3499158, nssv3499407, nssv3499744, nssv3500322, nssv3500765, nssv3500839, nssv3501135, nssv3502111, nssv3502134, nssv3502527, nssv3702635, nssv3702636, nssv3704060, nssv3704061, nssv3704062, nssv3704071, nssv3704115, nssv3704135, nssv3704136, nssv3704144, nssv3704164, nssv3704165, nssv4029812, nssv723195, nssv723196, nssv723200, nssv723212, nssv723240, nssv723243, nssv723254, nssv723255, nssv723257, nssv723258, nssv723259, nssv723268, nssv723272, nssv723273, nssv723284, nssv723285, nssv723287, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723349, nssv723352, nssv723353, nssv723354, nssv723359, nssv723361, nssv723362, nssv723363, nssv723364, nssv723378, nssv723380, nssv723386, nssv723392, nssv723395, nssv723417
CV-1319	149036512	149339573	essv11482, essv18916, essv19038, essv25779316, essv25788983, essv25798108, essv5001910, essv55468, essv6981512, essv6982967, essv6989889, essv6996867, essv6996911, essv6996922, essv6996933, essv6996978, essv6997044, essv6997055, essv6997066, essv6997244, essv6997366, essv6997434, essv6997445, essv6997545, essv6997556, essv6997578, essv6997589, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv6997733, essv7006261, essv7006268, essv78072, essv7932, essv92827, essv94654, essv9837828, nssv1173153, nssv1434888, nssv3482753, nssv3482797, nssv3482913, nssv3483041, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483325, nssv3483397, nssv3483443, nssv3483503, nssv3483582, nssv3483642, nssv3483897, nssv3484023, nssv3484060, nssv3484091, nssv3484107, nssv3484217, nssv3484247, nssv3484414, nssv3484538, nssv3484544, nssv3484718, nssv3484771, nssv3484863, nssv3484907, nssv3484933, nssv3485085, nssv3485169, nssv3485324, nssv3485432, nssv3485570, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3485998, nssv3486028, nssv3486203, nssv3486345, nssv3486621, nssv3486746, nssv3486814, nssv3486821, nssv3486940, nssv3486972, nssv3487034, nssv3487555, nssv3487569, nssv3487653, nssv3487780, nssv3487908, nssv3488871, nssv3489491, nssv3489748, nssv3489778, nssv3489845, nssv3490028, nssv3490029, nssv3490046, nssv3490123, nssv3490202, nssv3490665, nssv3490997, nssv3491137, nssv3491317, nssv3491428, nssv3491473, nssv3491620, nssv3491915, nssv3492689, nssv3493189, nssv3493913, nssv3494028, nssv3494285, nssv3494348, nssv3494815, nssv3494978, nssv3495023, nssv3495155, nssv3495296, nssv3495346, nssv3495393, nssv3495407, nssv3495423, nssv3495467, nssv3495835, nssv3496262, nssv3497988, nssv3498229, nssv3498972, nssv3499009, nssv3499158, nssv3499407, nssv3499744, nssv3500322, nssv3500765, nssv3500839, nssv3501135, nssv3502111, nssv3502134, nssv3502527, nssv3703992, nssv3703993, nssv3704060, nssv3704061, nssv3704062, nssv3704071, nssv3704115, nssv3704135, nssv3704136, nssv3704144, nssv3704164, nssv3704165, nssv4029812, nssv723196, nssv723200, nssv723212, nssv723240, nssv723243, nssv723254, nssv723255, nssv723257, nssv723258, nssv723259, nssv723272, nssv723273, nssv723284, nssv723285, nssv723287, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723349, nssv723352, nssv723353, nssv723354, nssv723359, nssv723361, nssv723362, nssv723363, nssv723364, nssv723378, nssv723380, nssv723386, nssv723392, nssv723395, nssv723417

Eight studies give support to this bicluster: Coe 2014, Conrad 2009, Cooper 2011, de Smith 2007, Lou 2015, Redon 2006, Suktitipat 2014, Vogler 2010

Table B.10: Supporting structural variants associated to a well-evaluated bicluster (ID=29, 25 CV x 8 studies) (3/4): All 588 SSV that were identified as part of a same bicluster, having as support the following studies: Vogler 2010, Suktitipat 2014, Redon 2006, de Smith 2007, Lou 2015, Cooper 2011, Conrad 2009, Coe 2014

CV ID	Min start	Max end	List equivalent SSV
CV-1320	149036512	149340200	essv11482, essv18916, essv19038, essv25779316, essv25788983, essv25798108, essv5001910, essv55468, essv6981512, essv6982967, essv6989889, essv6996867, essv6996911, essv6996922, essv6996933, essv6996978, essv6997044, essv6997055, essv6997066, essv6997244, essv6997366, essv6997434, essv6997445, essv6997545, essv6997556, essv6997578, essv6997589, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv6997733, essv7006261, essv7006268, essv7006297, essv78072, essv7932, essv92827, essv94654, essv9837828, nssv1173153, nssv1434888, nssv3482753, nssv3482797, nssv3482913, nssv3483041, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483397, nssv3483443, nssv3483503, nssv3483582, nssv3483642, nssv3483897, nssv3484023, nssv3484060, nssv3484091, nssv3484107, nssv3484217, nssv3484247, nssv3484414, nssv3484538, nssv3484544, nssv3484718, nssv3484771, nssv3484863, nssv3484907, nssv3484933, nssv3485085, nssv3485169, nssv3485324, nssv3485432, nssv3485570, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3485998, nssv3486028, nssv3486203, nssv3486345, nssv3486621, nssv3486746, nssv3486814, nssv3486821, nssv3486940, nssv3486972, nssv3487034, nssv3487555, nssv3487653, nssv3487780, nssv3487908, nssv3488871, nssv3489491, nssv3489748, nssv3489778, nssv3489845, nssv3490028, nssv3490029, nssv3490046, nssv3490123, nssv3490202, nssv3490665, nssv3490997, nssv3491317, nssv3491428, nssv3491473, nssv3491620, nssv3491915, nssv3492689, nssv3493189, nssv3493913, nssv3494028, nssv3494348, nssv3494815, nssv3494978, nssv3495023, nssv3495155, nssv3495296, nssv3495346, nssv3495393, nssv3495407, nssv3495423, nssv3495467, nssv3495835, nssv3496262, nssv3497988, nssv3498229, nssv3498972, nssv3499009, nssv3499158, nssv3499407, nssv3499744, nssv3500322, nssv3500765, nssv3500839, nssv3501135, nssv3502111, nssv3502134, nssv3502527, nssv3703992, nssv3703993, nssv3704060, nssv3704061, nssv3704062, nssv3704071, nssv3704115, nssv3704135, nssv3704136, nssv3704144, nssv3704164, nssv3704165, nssv4029812, nssv723196, nssv723200, nssv723212, nssv723240, nssv723243, nssv723254, nssv723255, nssv723257, nssv723258, nssv723259, nssv723272, nssv723273, nssv723285, nssv723287, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723349, nssv723352, nssv723353, nssv723354, nssv723359, nssv723362, nssv723363, nssv723364, nssv723378, nssv723380, nssv723386, nssv723392, nssv723395, nssv723417
CV-1321	149036512	149373529	essv11482, essv18916, essv19038, essv25779316, essv25788953, essv25788983, essv5001910, essv55468, essv6981512, essv6982967, essv6989889, essv6996933, essv6996978, essv6997044, essv6997055, essv6997066, essv6997244, essv6997366, essv6997434, essv6997445, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv7006261, essv7006268, essv7006297, essv78072, essv7932, essv92827, essv94654, essv9837828, nssv1434888, nssv3482753, nssv3482913, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483397, nssv3483443, nssv3483582, nssv3483642, nssv3483897, nssv3484091, nssv3484107, nssv3484217, nssv3484247, nssv3484414, nssv3484538, nssv3484544, nssv3484771, nssv3484863, nssv3484907, nssv3484933, nssv3485085, nssv3485169, nssv3485570, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3486028, nssv3486203, nssv3486345, nssv3486424, nssv3486621, nssv3486746, nssv3486821, nssv3486940, nssv3486972, nssv3487653, nssv3487780, nssv3488077, nssv3488871, nssv3489491, nssv3489748, nssv3489845, nssv3490029, nssv3490046, nssv3490202, nssv3490665, nssv3490997, nssv3491620, nssv3492689, nssv3493913, nssv3494348, nssv3494893, nssv3494978, nssv3495023, nssv3495296, nssv3495346, nssv3495407, nssv3495467, nssv3495835, nssv3496262, nssv3497988, nssv3498972, nssv3499009, nssv3499158, nssv3499407, nssv3500322, nssv3500765, nssv3500839, nssv3501135, nssv3502111, nssv3703992, nssv3703993, nssv3704115, nssv3704164, nssv3704165, nssv4029812, nssv723200, nssv723212, nssv723240, nssv723243, nssv723244, nssv723258, nssv723259, nssv723273, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723340, nssv723353, nssv723354, nssv723362, nssv723363, nssv723364, nssv723378, nssv723380, nssv723386, nssv723395, nsv1003007
CV-1322	149036512	149376652	essv11482, essv18916, essv19038, essv25779316, essv25788953, essv25788983, essv5001910, essv55468, essv6981512, essv6982967, essv6989889, essv6996978, essv6997044, essv6997055, essv6997066, essv6997244, essv6997366, essv6997434, essv6997445, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv7006261, essv7006268, essv7006297, essv78072, essv7932, essv92827, essv94654, essv9837828, nssv1434888, nssv3482753, nssv3482913, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483397, nssv3483443, nssv3483582, nssv3483897, nssv3484091, nssv3484107, nssv3484217, nssv3484247, nssv3484538, nssv3484544, nssv3484771, nssv3484863, nssv3484907, nssv3484933, nssv3485085, nssv3485169, nssv3485570, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3486028, nssv3486203, nssv3486345, nssv3486424, nssv3486621, nssv3486746, nssv3486821, nssv3486940, nssv3486972, nssv3487653, nssv3487780, nssv3488077, nssv3488871, nssv3489748, nssv3489845, nssv3490029, nssv3490046, nssv3490202, nssv3490665, nssv3490997, nssv3491620, nssv3492689, nssv3493913, nssv3494348, nssv3494893, nssv3494978, nssv3495023, nssv3495296, nssv3495346, nssv3495407, nssv3495467, nssv3495835, nssv3496262, nssv3497988, nssv3498972, nssv3499158, nssv3499407, nssv3500322, nssv3500765, nssv3500839, nssv3501135, nssv3502111, nssv3703992, nssv3703993, nssv3704115, nssv3704165, nssv4029812, nssv723200, nssv723212, nssv723240, nssv723243, nssv723244, nssv723247, nssv723258, nssv723259, nssv723273, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723340, nssv723353, nssv723354, nssv723362, nssv723363, nssv723364, nssv723380, nssv723386, nssv1003007
CV-1332	149036524	149370974	essv11482, essv18916, essv19038, essv25779316, essv25788953, essv25788983, essv55468, essv6981512, essv6982967, essv6989889, essv6997434, essv6997445, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv7006297, essv78072, essv7932, essv92827, essv94654, nssv1434888, nssv3482753, nssv3482913, nssv3483041, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483397, nssv3483443, nssv3483582, nssv3484091, nssv3484107, nssv3484217, nssv3484247, nssv3484414, nssv3484538, nssv3484544, nssv3484863, nssv3484907, nssv3484933, nssv3485085, nssv3485169, nssv3485570, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3485998, nssv3486028, nssv3486203, nssv3486345, nssv3486424, nssv3486621, nssv3486746, nssv3486821, nssv3486940, nssv3486972, nssv3487555, nssv3487653, nssv3487780, nssv3488871, nssv3489491, nssv3489748, nssv3489845, nssv3490046, nssv3490665, nssv3490997, nssv3491620, nssv3492689, nssv3493913, nssv3494348, nssv3494978, nssv3495023, nssv3495296, nssv3495346, nssv3495407, nssv3495467, nssv3495835, nssv3496262, nssv3497988, nssv3498972, nssv3499009, nssv3499158, nssv3499407, nssv3500322, nssv3500765, nssv3500839, nssv3501135, nssv3502111, nssv3704115, nssv3704164, nssv3704165, nssv4029812, nssv723200, nssv723212, nssv723240, nssv723243, nssv723244, nssv723258, nssv723259, nssv723273, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723353, nssv723354, nssv723359, nssv723362, nssv723363, nssv723364, nssv723378, nssv723380, nssv723386, nssv723395
CV-2140	161476489	161681552	essv100802, essv101296, essv12956, essv13240, essv13501, essv17988, essv18558, essv19379, essv19737, essv20462, essv20874, essv22275, essv22791, essv23720, essv24352, essv25032, essv25779317, essv25801433, essv33649, essv3508, essv3577, essv3586300, essv38332, essv3903, essv47029, essv5007, essv50525, essv5641, essv6057, essv68133, essv6982257, essv6982345, essv6986310, essv7001566, essv7001599, essv7001633, essv7001644, essv7001666, essv7006608, essv70602, essv74890, essv7633, essv77909, essv79773, essv81284, essv929, essv93182, essv93746, essv93879, essv94830, essv96235, essv97445, essv97924, essv9838861, essv98751, esv12861, esv2763869, esv2764263, nssv1417711, nssv1418784, nssv1420724, nssv1421202, nssv1421483, nssv1425536, nssv1440156, nssv19108, nssv19430, nssv20409, nssv21757, nssv22769, nssv24878, nssv25954, nssv25958, nssv26873, nssv2852234, nssv3482857, nssv3482924, nssv3482999, nssv3483033, nssv3483380, nssv3483447, nssv3483551, nssv3483672, nssv3484055, nssv3484065, nssv3484399, nssv3484406, nssv3484478, nssv3484815, nssv3485712, nssv3486096, nssv3486110, nssv3486336, nssv3486819, nssv3486923, nssv3488032, nssv3488409, nssv3488764, nssv3489019, nssv3490643, nssv3491312, nssv3493103, nssv3493731, nssv3494071, nssv3494109, nssv3495534, nssv3495565, nssv3496609, nssv3496688, nssv3496987, nssv3497337, nssv3500421, nssv3500783, nssv3501236, nssv3501251, nssv3501742, nssv3501770, nssv3502321, nssv3704748, nssv3704750, nssv3704771, nssv3704775, nssv3704783, nssv3704788, nssv4029940, nssv540130, nssv617996, nssv654186, nssv671510, nssv672481, nssv677351, nssv677518, nssv677746, nssv678767, nssv679036, nssv686593, nssv693456, nssv693890, nssv697045, nssv727684
CV-2142	161478524	161681552	essv100802, essv101296, essv12956, essv13240, essv13501, essv17988, essv18558, essv19379, essv19737, essv20462, essv20874, essv22275, essv22791, essv23720, essv24352, essv25032, essv25779317, essv25801433, essv33649, essv3508, essv3577, essv3586300, essv38332, essv3903, essv47029, essv5007, essv50525, essv5641, essv6057, essv68133, essv6982257, essv6982345, essv6986310, essv7001566, essv7001599, essv7001633, essv7001644, essv7001666, essv7006608, essv70602, essv74890, essv7633, essv77909, essv79773, essv81284, essv929, essv93182, essv93879, essv94830, essv96235, essv97445, essv97924, essv9838861, essv98751, esv12861, esv2763869, esv2764263, nssv1417711, nssv1418784, nssv1420724, nssv1421202, nssv1421483, nssv1425536, nssv1440156, nssv19108, nssv19430, nssv20409, nssv21757, nssv22769, nssv25954, nssv25958, nssv26873, nssv2852234, nssv3482857, nssv3482924, nssv3482999, nssv3483033, nssv3483380, nssv3483447, nssv3483551, nssv3483672, nssv3484055, nssv3484065, nssv3484399, nssv3484406, nssv3484478, nssv3484815, nssv3485022, nssv3485712, nssv3486096, nssv3486110, nssv3486336, nssv3486819, nssv3486923, nssv3488032, nssv3488409, nssv3488764, nssv3489019, nssv3490643, nssv3491312, nssv3493103, nssv3493731, nssv3494071, nssv3494109, nssv3495534, nssv3495565, nssv3496609, nssv3496688, nssv3496987, nssv3497337, nssv3500421, nssv3500783, nssv3501236, nssv3501251, nssv3501742, nssv3501770, nssv3502321, nssv3704748, nssv3704750, nssv3704771, nssv3704775, nssv3704783, nssv3704788, nssv4029940, nssv540130, nssv617996, nssv654186, nssv671510, nssv672481, nssv677351, nssv677518, nssv677746, nssv678767, nssv679036, nssv686593, nssv693456, nssv693890, nssv697045, nssv727684

Eight studies give support to this bicluster: Coe 2014, Conrad 2009, Cooper 2011, de Smith 2007, Lou 2015, Redon 2006, Suktitipat 2014, Vogler 2010

Table B.11: Supporting structural variants associated to a well-evaluated bicluster (ID=29, 25 CV x 8 studies) (4/4): All 588 SSV that were identified as part of a same bicluster, having as support the following studies: Vogler 2010, Suktitipat 2014, Redon 2006, de Smith 2007, Lou 2015, Cooper 2011, Conrad 2009, Coe 2014

CV ID	Min start	Max end	List equivalent SSV
CV-2145	161481796	161681552	essv100802, essv101296, essv12956, essv13240, essv13501, essv17988, essv18558, essv19379, essv19737, essv20874, essv22275, essv22791, essv23720, essv24352, essv25779317, essv25801433, essv33649, essv3508, essv3577, essv3586300, essv3903, essv47029, essv5007, essv5641, essv6982257, essv6982345, essv6986310, essv7001599, essv7001633, essv7001644, essv7001666, essv7006608, essv74890, essv7633, essv79773, essv929, essv93879, essv94830, essv96235, essv97445, essv97924, essv9838861, essv98751, esv2764263, nssv1417711, nssv1418784, nssv1421202, nssv1425536, nssv1440156, nssv21757, nssv26873, nssv2852234, nssv3482857, nssv3482924, nssv3482999, nssv3483033, nssv3483380, nssv3483447, nssv3483551, nssv3483672, nssv3484055, nssv3484065, nssv3484399, nssv3484406, nssv3484478, nssv3484815, nssv3485022, nssv3485712, nssv3486096, nssv3486110, nssv3486336, nssv3486819, nssv3486923, nssv3488032, nssv3488409, nssv3488764, nssv3489019, nssv3490643, nssv3493103, nssv3494071, nssv3494109, nssv3495534, nssv3495565, nssv3496609, nssv3496688, nssv3496987, nssv3497337, nssv3500421, nssv3500783, nssv3501236, nssv3501251, nssv3501742, nssv3501770, nssv3502321, nssv3704748, nssv3704750, nssv3704771, nssv3704775, nssv3704783, nssv3704788, nssv4029940, nssv617996, nssv654186, nssv671510, nssv672481, nssv677351, nssv677518, nssv677746, nssv678767, nssv679036, nssv686593, nssv693456, nssv693890, nssv727684, nssv727685

Eight studies give support to this bicluster: Coe 2014, Conrad 2009, Cooper 2011, de Smith 2007, Lou 2015, Redon 2006, Suktitipat 2014, Vogler 2010

Table B.12: Supporting structural variants associated to a well-evaluated bicluster (ID=30, 17 CV x 8 studies): All 495 SSV that were identified as part of a same bicluster, having as support the following studies: Vogler 2010, Redon 2006, Suktitipat 2014, de Smith 2007, Lou 2015, Coe 2014, Conrad 2009, Cooper 2011.

List of SSV
essv100377, essv100802, essv100936, essv101256, essv101296, essv101326, essv101537, essv1080, essv11482, essv12956, essv13240, essv13501, essv14154, essv14239, essv1569, essv1715, essv17988, essv18558, essv18916, essv19038, essv19363, essv19379, essv19737, essv19885, essv20462, essv20523, essv20874, essv21629, essv22275, essv22791, essv2359, essv23720, essv24352, essv25032, essv25779316, essv25779317, essv25780439, essv25781141, essv25781234, essv25781694, essv25787832, essv25787908, essv25787995, essv25788953, essv25788983, essv25790783, essv25791015, essv25792902, essv25801433, essv33321, essv33649, essv33753, essv3508, essv3577, essv3586300, essv35957, essv38332, essv38722, essv38955, essv3903, essv40849, essv41126, essv43839, essv44471, essv45569, essv47029, essv5001910, essv5007, essv50525, essv5455860, essv55468, essv5641, essv5662, essv6057, essv6431, essv6560, essv68133, essv6981512, essv6982257, essv6982345, essv6982967, essv6985823, essv6986310, essv6989889, essv6990745, essv6991078, essv6991744, essv6991967, essv6996122, essv6996133, essv6996155, essv6996600, essv6996611, essv6996622, essv6996811, essv6996844, essv6997244, essv6997366, essv6997434, essv6997445, essv6997489, essv6997500, essv6997534, essv6997622, essv6997645, essv6997656, essv6997678, essv6997711, essv6997722, essv7001566, essv7001599, essv7001633, essv7001644, essv7001666, essv7004962, essv7004977, essv7006218, essv7006220, essv7006226, essv7006228, essv7006251, essv7006252, essv7006268, essv7006297, essv7006306, essv7006309, essv7006608, essv7033189, essv70602, essv74890, essv7633, essv77909, essv78072, essv7932, essv79773, essv81284, essv92827, essv929, essv93105, essv93182, essv93282, essv93746, essv93879, essv942, essv94654, essv94830, essv95999, essv96191, essv96235, essv97263, essv97445, essv97924, essv9837839, essv9837877, essv9838624, essv9838861, essv9856931, essv98751, esv12861, esv16950, esv2763869, esv2764263, nssv1173678, nssv1173679, nssv1173680, nssv1173681, nssv1417711, nssv1418784, nssv1420724, nssv1421068, nssv1421202, nssv1421443, nssv1421483, nssv1424726, nssv1425536, nssv1426371, nssv1431793, nssv1434888, nssv1437927, nssv1440156, nssv1440799, nssv1449962, nssv1795806, nssv19108, nssv19430, nssv20409, nssv21402, nssv21757, nssv22769, nssv23694, nssv24878, nssv25699, nssv25954, nssv25958, nssv25970, nssv26873, nssv27030, nssv2852234, nssv3462863, nssv3462954, nssv3463447, nssv3464085, nssv3464481, nssv3464712, nssv3464724, nssv3465020, nssv3465164, nssv3466322, nssv3468171, nssv3471035, nssv3473439, nssv3473824, nssv3474437, nssv3476384, nssv3478201, nssv3478916, nssv3479202, nssv3480618, nssv3480932, nssv3481472, nssv3482753, nssv3482797, nssv3482857, nssv3482912, nssv3482913, nssv3482924, nssv3482964, nssv3482965, nssv3482999, nssv3483033, nssv3483041, nssv3483131, nssv3483132, nssv3483222, nssv3483260, nssv3483350, nssv3483380, nssv3483397, nssv3483427, nssv3483443, nssv3483447, nssv3483540, nssv3483541, nssv3483551, nssv3483582, nssv3483633, nssv3483642, nssv3483672, nssv3483897, nssv3483916, nssv3483927, nssv3483946, nssv3484055, nssv3484065, nssv3484091, nssv3484107, nssv3484217, nssv3484219, nssv3484247, nssv3484265, nssv3484399, nssv3484406, nssv3484414, nssv3484476, nssv3484478, nssv3484538, nssv3484544, nssv3484555, nssv3484570, nssv3484736, nssv3484771, nssv3484815, nssv3484863, nssv3484907, nssv3484933, nssv3485022, nssv3485085, nssv3485169, nssv3485358, nssv3485496, nssv3485570, nssv3485712, nssv3485825, nssv3485945, nssv3485952, nssv3485997, nssv3485998, nssv3486028, nssv3486035, nssv3486096, nssv3486110, nssv3486170, nssv3486203, nssv3486259, nssv3486287, nssv3486336, nssv3486345, nssv3486424, nssv3486621, nssv3486746, nssv3486764, nssv3486784, nssv3486792, nssv3486819, nssv3486821, nssv3486923, nssv3486940, nssv3486972, nssv3487555, nssv3487653, nssv3487780, nssv3488032, nssv3488039, nssv3488053, nssv3488077, nssv3488099, nssv3488304, nssv3488409, nssv3488597, nssv3488713, nssv3488764, nssv3488871, nssv3489019, nssv3489328, nssv3489491, nssv3489566, nssv3489748, nssv3489845, nssv3490028, nssv3490029, nssv3490046, nssv3490081, nssv3490111, nssv3490114, nssv3490202, nssv3490643, nssv3490665, nssv3490819, nssv3490997, nssv3491002, nssv3491312, nssv3491620, nssv3492689, nssv3493103, nssv3493105, nssv3493160, nssv3493254, nssv3493487, nssv3493731, nssv3493789, nssv3493913, nssv3494071, nssv3494109, nssv3494259, nssv3494348, nssv3494530, nssv3494544, nssv3494550, nssv3494893, nssv3494978, nssv3495023, nssv3495296, nssv3495346, nssv3495361, nssv3495407, nssv3495467, nssv3495534, nssv3495565, nssv3495835, nssv3496262, nssv3496414, nssv3496609, nssv3496688, nssv3496987, nssv3497337, nssv3497834, nssv3497988, nssv3498902, nssv3498972, nssv3499009, nssv3499158, nssv3499407, nssv3499738, nssv3500097, nssv3500322, nssv3500421, nssv3500765, nssv3500770, nssv3500783, nssv3500839, nssv3500897, nssv3501135, nssv3501236, nssv3501251, nssv3501742, nssv3501770, nssv3501870, nssv3502111, nssv3502321, nssv3698020, nssv3702228, nssv3703992, nssv3703993, nssv3704088, nssv3704115, nssv3704164, nssv3704165, nssv3704174, nssv3704182, nssv3704203, nssv3704205, nssv3704208, nssv3704748, nssv3704750, nssv3704771, nssv3704775, nssv3704783, nssv3704788, nssv4028323, nssv4029064, nssv4029797, nssv4029798, nssv4029799, nssv4029802, nssv4029805, nssv4029806, nssv4029807, nssv4029808, nssv4029809, nssv4029812, nssv4029940, nssv451753, nssv540130, nssv547768, nssv617996, nssv654186, nssv671510, nssv672481, nssv677351, nssv677518, nssv677746, nssv678767, nssv679036, nssv686593, nssv693456, nssv693890, nssv697045, nssv710027, nssv710030, nssv710031, nssv710033, nssv710039, nssv710040, nssv710080, nssv710082, nssv723095, nssv723124, nssv723125, nssv723126, nssv723129, nssv723200, nssv723212, nssv723240, nssv723243, nssv723244, nssv723247, nssv723249, nssv723251, nssv723258, nssv723259, nssv723269, nssv723273, nssv723296, nssv723300, nssv723310, nssv723311, nssv723336, nssv723340, nssv723345, nssv723346, nssv723353, nssv723354, nssv723359, nssv723362, nssv723363, nssv723364, nssv723378, nssv723380, nssv723386, nssv723395, nssv723428, nssv727684, nssv727685, nsv1000267, nsv1003007

Eight studies give support to this bicluster: Vogler 2010, Redon 2006, Suktitipat 2014, de Smith 2007, Lou 2015, Coe 2014, Conrad 2009, Cooper 2011