UNDERSTANDING WHAT'S "UNDER THE HOOD": INCREASING ACCESSIBILITY IN OMICS RESULTS

By

Janice Patterson

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

September 2015

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Janice Patterson

"Understanding What's "Under the Hood": Increasing Accessibility in Omics Results"

Has been approved

____

Thesis Advisor – Shannon McWeeney PhD

____

Committee Member – Lucia Carbone, PhD

____

Committee Member – Michael Heinrich, MD

____

Committee Member – Kemal Sonmez, PhD

Table of Contents

FIGURES

TABLES

ABBREVIATIONS

bp	*base pair*
ChIP-seq	*chromatin immunoprecipitation sequencing*
CNV	*copy number variant*
FRiP	*fraction of reads in peaks*
Gbp	*giga bases (1e9 bases)*
GIST	*gastrointestinal stromal tumor*
KIT	*c-kit proto-oncogene*
PDGFrA	*platelet-derived growth factor receptor, alpha polypeptide*

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Shannon McWeeney for her support, motivation, and immense knowledge. I have learned so much from her and I am very grateful.

A very special thanks goes out to Dr. Michael Heinrich, without his motivation and encouragement I would not have considered a graduate career. His support and guidance has been an invaluable part of my life.

I would like to thank Dr. Lucia Carbone, who provided me an opportunity to join her team for a brief time. Her insightful comments, vast knowledge, and patience were an invaluable part of my project. I am incredibly grateful to her lab for all the support and advice they provided.

Thanks also go to Dr. Kemal Sonmez, for the support, encouragement and for the hard question which drove my research and helped me to realize its full potential.

I thank my fellow lab mates for the stimulating discussions, for the moral support, and for all the fun we have had and will continue to have. They are my sisters.

I am also grateful to the Department of Medical Informatics and Clinical Epidemiology. I am extremely thankful for their sincere and valuable guidance and encouragement.

ABSTRACT

With the recognition of the need for reproducibility in the face of increasing data complexity, we implemented a series of metrics and an interactive visualization prototype to facilitate discovery and interpretation for next generation sequencing use cases: copy number variant detection with exome sequencing and chromatin immunoprecipitation followed by sequencing (ChIP-seq). This interactive framework is an important first step to provide measurable and consistent check of data validity for each step of the analysis. The goal is to allow scientists to assess the validity of their methods, ensure the accuracy of the data and guide prioritization by obtaining high confidence findings. It is important to note that the interactive component allows identification of issues or characteristics of the data that would have been missed in traditional static plots. We note this is an iterative process and with further evaluation and testing, additional metrics and contextual views can be added.

Chapter 1 Introduction & Background

The advent of “big data” has changed how and what we analyze, the inferences that can be made as well as the power of the data itself (1). The very nature of big data – with respect to its characteristics of velocity, volume, variety and veracity - has also created new challenges, particularly with interpretation and “accessibility” of the data with respect to knowledge. Given the complexity of the data structures and algorithms needed to manage and process the data, the need for rigor and reproducibility at scale is key. There is also a need for transparency with regard to how the data has been analyzed and what impact the processing steps and algorithms have on the results, as well as a need to quantify the irreproducibility. With respect to the consumers of the data (analysts, end users who are being delivered results etc.), there is the potential for errors in the multi-step analyses or in the interpretation or assessment of the results.

In a study by Prinz *et.al* from Bayer, there was an attempt to reproduce 67 projects from oncology, women’s health, and cardiovascular specialties(2). They found that only about 32% of the findings were either reproducible or in line with the reported scientific findings. Even more troubling was the finding that 65% of the repeated scientific work had conflicting and inconsistent results compared to the published findings. The inconsistencies among the studies occurred independent of domain and the prestige of the initial journal publication. Prinz and co-authors attribute this severe lack of reproducibility to a multitude of factors, such as differences in statistical testing, potentially poor scientific practices, and underpowered studies. Studies that lack power are ultimately measuring statistical noise rather true effect size (2).

Another examination of reproducibility was conducted on 53 scientific studies that had published pivotal and novel approaches in the past decade. The authors could only confirm 6 of the 53 studies (3). This lack of reproducibility was observed regardless of the journal's impact score and number of citations referencing them. The authors reported issues with misinterpretation and inappropriate statistical testing and analysis, underpowered experiments and overall failure of the checks and balances system of the scientific process(4).

In biomedicine, one source of the vast amount of data being produced is from high throughput sequencing, which has enabled new insights into the genome, transcriptome and epigenome of diverse species (5). Genome sequencing can identify variations in DNA. Methylation and histone modification can also be identified by using DNA sequencing coupled with (5) immunoprecipitation (ChIP-seq) thus measuring the interaction of DNA with protein (11). Genetic information about the transcriptome and splice variants can be extracted by sequencing the translated RNA. The variety and volume in particular has challenges that traditional scientific methods had been inadequate to handle(6).

Sources of Irreproducibility

One of the most troubling aspects of the analysis of next generation sequencing data is the large number of processing steps and the variation in the methods and tools used for analysis which can lead to differences in results between studies and impact replication of findings. The volume of the data obtained from sequencing can give researchers a false sense of security regarding the robustness of the data. However, changes in a single step in the analysis workflow leads to drastically different results. The inability to differentiate the statistical noise from the detectable signal is a major source of the irreproducibility we observe in these analyses.

High throughput sequencing has further necessitated the development of new algorithms and computational tools using a variety of techniques to manage, process and analyze the massive amounts of data that are produced. These computational methods have their own assumptions, models and parameters that can potentially lead to varying results. Consequently, drastically varying results can be obtained based solely on evaluation methods.

Quality statistics are often performed prior to processing high throughput sequencing reads, which assess potential artifacts from the experimental method or sequencing platform. FASTQC, a set of static graphs that generate statistical diagnostics, has revolutionized the interpretation of variation observed across next generation sequencing platforms. FASTQC generates quality control metrics and visualizations on the raw sequencing data generated by the sequencer(7). This simple analysis can assess and flag problems within raw data output, prior to performing hours of analysis. However, validation of subsequent analysis of the computational methods is rarely ever performed. The assessment of downstream signal that evaluates the robustness of computational methods and the reproducibility of results, is equally important to prioritize and determine the confidence of results. However, quality assessment at this point is rarely performed, and no metrics of the robustness of the data are provided. This provides no guide for assessment of stability of the findings and sensitivity to algorithmic parameters or processing steps. Computational and statistical models and algorithms should be treated as a tool or machine that is capable of finite degrees of error rather than an exact method of calculation.

Moreover, different computational methods with different underlying assumptions make it difficult to compare them. The variability of these methods, such as different algorithms, models, assumptions, and the sheer number of techniques developed has complicated the analysis and introduced deviations, and inaccurate results, stemming from the computational method alone (12, 13). The variety of computational methods, the variability in how they are implemented and used and their ongoing development, as well as the dynamic annotations and genome builds, makes development of standard operating procedures (SOP) difficult. The ability to compare the quality of computational methods and produce equivalent and reproducible results is crucial to determine the biological confidence of high throughput sequencing data.

Potential Solutions

The lack of reproducibility and transparency highlights the urgent need for frameworks for communicating and accurately depicting the data to scientists. The growing amount of these enormous and dynamic datasets calls for changes in how we handle and access data. The datasets are often too large to open as a simple spreadsheet, too complicated and multifaceted for simple visualizations. This renders the data inaccessible and essentially locked away from interpretation from a user who lacks computational training. Methods must adapt to help scientists interpret the data and facilitate discovery.

Adjustment of the way data is visualized is a crucial component of exposing the data. Static two-dimensional visualizations have become inadequate in communicating and depicting big data. These visualizations don’t allow for the active interpretation and hypothesis generating interpretation that would allow us to expose large data sets. Additionally, overcomplicated plots like three-dimensional plots have been shown to be difficult to interpret and prone to misinterpretation (8). The revolution in analytics lies in interactive data visualization that allow the visualizer to filter, manipulate and interact with the data to immediately ask questions and generate hypotheses from the data (9). The development of these interactive frameworks also provides cross-platform, efficient methods for communicating the transformations and analysis of the data. The simple act of interacting with the data can communicate the filtering and characterize the multiple facets to new users. This level of transparency will improve the overall quality of scientific data and increase the reproducibility with clearly defined logical computational methods that will be become standards during analysis.

The long-term goal is to increase data transparency and develop methods for interactive visualization that will guide intuitive interpretation. Data interpretation performed with tangible statistical metrics and interactive visualizations will enable differentiation between the relevant or truly biologically significant to the irreproducible statistically irrelevant signal. The following framework will provide an evaluation of comparative computational methods. This tool will compare results from computational methods providing a statistical assessment of accuracy and reliability, while emphasizing the impact of algorithms and their parameters on data dynamics. Comparison based on shared measurable characteristics, like the sequencing depth at measured signal, can measure the robustness of the results obtained from these computational tools.

Evaluating these analysis methods and algorithms requires extensive comparison and inspection of the results using comparative metrics and visualizations. For every new algorithm or version developed the comparison has to be reevaluated, recalibrated and normalized to equivocate results and allow a fair comparison of the performance. The development of an interactive visualization framework for dynamic data abstraction and to compare computational methods is necessary.

The dynamics of interactive data visualization and exploration can expand the possibilities of static visualizations. Revolutionizing the interchange of information from simply data filtering, sorting and maneuvering to acting as a reproducible record of the transformations of analyses and providing a portable intuitive framework to communicate the transformation performed in analysis (9).

Study Aims

In this study, the focus is on developing and implementing metrics, summaries and interactive visualizations to assess performance and reproducibility of different algorithms for two high throughput workflows use cases: ChIP-seq and Copy Number. The aims are:

Aim 1: Develop and extend metrics, statistical summaries and visualizations to assess the performance and reproducibility of different algorithms and computational methods in high throughput workflows

Aim 2: Implement metrics in an interactive framework for two use cases. Use cases will be enrichment regions from chromatin immunoprecipitation (ChIP-seq) and copy number variants (CNV) from exome sequencing

Comparative metrics for computational methods is particularly difficult because of the vastly different approaches these methods use for detections. The computational workflows are often complicated multi-step, even multi-algorithmic processes involving multiple filtering steps, realignment, statistical thresholding steps, different underlying assumptions and parameter optimization. Variation and the abundance of methods make it difficult to assess and compare methods or define them for a particular data set and have resulted in the lack of consistency and have made reproducibility extremely difficult. One method I will use to compare the results is by calculating the irreproducible discovery rate (IDR). Briefly, this statistical metric uses biological replicates to measure the reproducibility of signal among replicates (10).

Methods that can assess and compare the performance of a computation, requires practical and universal measurements of comparison as well as methods of deconstruction to accurately and fairly compare and assess performance. However, the computational methods and tools used vary drastically depending on the type of next generation sequencing analysis being performed. Therefore, an analysis of comparative computational methods and metrics for assessment will be explained in context of two use cases of copy number variant (CNV) detection from whole exome sequencing and chromatin immunoprecipitation sequencing (ChIP-seq).

Copy number variation is a major source of genetic differentiation that has been associated with various phenotypes and diseases. The inherent limitations of exome sequencing make detection of the copy number variation a uniquely complicated computational problem. Exome CNV was detected using two computational methods; both using read depth as a method of detection. ExomeDepth, assumes a beta-binomial distribution of the read depths and subsequently uses a hidden Markov chain to combine copy change regions (11).

The second copy number detection method, ExomeCNV, assumes read counts have a Poisson distribution then uses the circular binary segmentation algorithm to merge copy number variant regions(12). Chromatin immunoprecipitation with massively parallel sequencing (ChIP-seq) has been used in identifying protein-DNA binding, histone modifications and nucleosome locations across the genome (13). Characterization of these binding sites allows the identification of locations essential in genetic regulation (13, 14).   Both of these methods are widely used and utilize multi-step workflows with a large number of potential algorithms and parameters and make excellent use cases for an this prototype implementation.

The interactive framework will be developed in R shiny developed by RStudio (15). Shiny is a web application framework that allows users to rapidly turn their analyses into interactive web applications. These applications can then by hosted on the web and accessible to anyone with access to the internet, without the installation of cumbersome programs and packages. Using a single program (Rstudio) to develop, analyze, and deliver data analyses allows scientist to easily and rapidly deliver their data and to prototype an interactive framework. The development of Shiny applications is also extremely versatile, with the potential to integrate HTML, Javascript and D3 as developed toolset and easily integrate them into the R shiny framework. Although, conflicts may arise in the package dependent scripts developed in R, the package dependency is among the many reasons for the ease of implementation and analysis in R. Additionally, latency issues can also arise based on the hosted server use in which the web application is ultimately launched. Despite these minor issues this application provides the ideal system for scientist to develop and launch interactive data frameworks.

Chapter 2 Exome Copy Number Detection

Introduction

Copy number variation (CNV) is a major source of genetic variation that has been associated with various phenotypes responsible for the diversity in human as well increasingly attributed to human disease. CNVs can range in sizes, from small focal length regions, as small as 50 nucleotide bases, to regions that are a kilo base or larger. Variations can be categorized as regions of deletions or duplications. Deletions are observed regions that are missing when compared to a reference genome, resulting in an assumed loss of function (LOF) for that particular location of the genome. Duplications are variations in which a particular region is duplicated resulting in supposed gain of function (GOF) of that genome location (16). The use of massively paralleled sequencing to characterize CNVs has become increasingly popular. The random distribution of the short reads along the genome with higher coverage and resolution allows for more accurate estimation of the copy number than traditional methods using fluorescent in situ hybridization, array comparative genomic hybridization and SNP arrays (17, 18).

Whole exome sequencing (WES) has become an inexpensive alternative to whole genome sequencing. The targeted sequencing method can only capture the protein-coding regions, or exons, consisting of less than 1% of the genome (17). However, WES can be effective method in detecting genes harboring copy number gains and losses that affecting the presentation of disease phenotypes. Although, the inherent limitations of WES renders CNV detection from exome sequencing incapable of capturing large cross chromosomal variations and variations

within exonic junctions, breakpoints and non-coding regions, it provides a low cost alternative to identify single nucleotide variants as well as copy number changes in coding regions (18).
A diverse set of tools has been developed to identify CNVs from WES sequencing and only a handful have been developed for somatic CNV detection (16, 18-21). The inclusion of a paired normal sample can be extremely advantageous in detecting accurate somatic copy number variants and correcting for background. The purity of the tumor and an accurate measurement of the cross contamination between tumor and normal pairs cannot often be estimated. This measurement, however, can significantly impact the sensitivity and specificity of CNVs that are detected. Additional considerations are caused by the limited capture of coding sequences, which introduces significant GC biases (16, 17). The correction of GC biases is necessary for analyzing exome sequencing, however, the following study used paired normal tissue sample from the same patient sequenced in conjunction with the tumor tissue. This method nullifies any significant biases background and biases caused by GC rich regions (18)
The computational tools developed, for somatic CNV detection from WES data, use a diverse set of methods to detect CNVs with different assumptions, algorithms and models that impact the accuracy of CNV detection (17). The majority of applications of CNV calling from WES data use the read depth of sequencing at targeted region to determine areas of gain and loss (17). Alternative methods, such as paired end or split read mapping, cannot account for the sparseness of the data obtained from WES.

Once read depth regions have been established the regions are combined using a process called segmentation. Segmentation is performed using a variety of algorithmic techniques to merge or differentiate the exons with copy number gains and losses contained within a single CNV event. The differentiation of these merging and stopping points also involves assumptions about the statistical distribution of reads in order to distinguish junctions of changing copy numbers. The regions are then interpreted statistically with various thresholds and measurements to determine the confidence of the variant, such as the log of the reads ratio of tumor to normal or Bayes Factor (19).

IDR

The perpetual development of new computational methodologies calls for comparative techniques that can measure the accuracy, reproducibility and performance of these new methods. Li *et al*. developed a potential solution to the statistical arbitrary thresholding, called the irreproducible discovery rate (IDR), which was originally applied to peaks in chromatin immunoprecipitation sequencing (ChIP-seq) experiments (10). This method assumes that real signal is reproducible and that noise is generally irreproducible among replicates.

Thus, by providing measurements of the signal that is reproducible across biological replicates we can assess the correspondence between these replicates and precisely determined the threshold at which agreement between signal dissociates. Measuring the correspondence of biological replicates and IDR, therefore, uses biological replicates to compute an empirical statistical threshold, at which variability occurs. The IDR can then be compared to across computational methods. An evaluation of the reproducible signals from different methods can provide a method to assess the robustness of the computational methods (10).

ExomeCNV

The following study uses two methods to detect exome CNV regions, both using read depth as a method of detection. ExomeCNV, uses the read coverage output from the GATK depth of coverage scripts. The ease of integration to a suite of tools capable of discovering somatic variants can save a significant amount of time when building pipelines for analysis. The suite of scripts has established a best practices guideline for mutation detection and could be integrated easily to include copy number detections(22). The distribution of read counts was assumed to follow a Poisson distribution. Next, segmentation is performed using circular binary segmentation algorithm to combine coverage and estimate the mean reads ratios (12).

ExomeCNV, additionally, provides the input of a fixed normal contamination of the tumor rate, referred to as the admixture. The level of tumor purity has a major effect on the sensitivity and specificity of the called variant regions. The defaults assume a higher level of specificity and less false positives (12). An admixture rate of zero, assuming an entirely pure tumor, was set to establish a high level of sensitivity to capture a proportion of false positives (10). The resulting CNVs are reported as the log of the reads ratios and at a given threshold for specificity and sensitivity.

ExomeDepth

The second CNV detector ExomeDepth, assumes a beta-binomial distribution of the read depths. Many CNV methods assume that reads follow a Poisson distribution with equal mean and variance. Plagnol *et al*. found that the fit significantly improved using a beta-binomial model to account for the over dispersion of reads (11).

The likelihood value for each exon is then combined using a segmentation method that implements a hidden Markov model. The model merges CNV calls across exons that use the prior probability to transition from different copy number states (11).

The tumor and normal purity can also have a significant impact on the sensitivity of the CNVs detected. This parameter as the proportion of tumor in the tumor sample was set at 0, to capture the most amount of potential CNVs that could be detected. The resulting CNVs are reported with the corresponding prior probability.

Data for Use Case: *Gastrointestinal Stromal Tumors (GIST)*

This study will analyze cancer lesions, specifically gastrointestinal stromal tumors (GIST). These tumors occur primarily in older patients and are most common non-epithelial tumor of the gastrointestinal tract(24). Its incidence has been observed worldwide at a rate of 11 and 19.6 per million people at a rate of approximately 3300 to 6000 new diagnoses per year in the United States(25). In 1998, Hirota et al found that GIST harbored mutations in the c-kit proto-oncogene, with over 75% of total GISTs harboring mutation in KIT tyrosine kinase (25, 26). Genotypic profiling found further mutations in the same kinase family and homologs of KIT, such as platelet derived growth factor PDGFrA, responsible for approximate 7-12% of GISTs. These mutations leads to the constitutive activation of KIT by arresting the conformation for ligand binding with stem cell factor (SCF), followed by homodimerization(25).

These mutations in GISTs are specific alternations responsible for the activation of the kinase that leads to uncontrolled oncogenesis. This specificity has led to the successful application and treatment with KIT kinase inhibitor imatinib mesylate (STI571, Gleevec®) as an inhibitor of the transmembrane receptor tyrosine kinase (27). However, 15% of GIST do not have a detectable KIT or PDGFrA mutation, this subset of ‘wild-type’ GIST are morphologically identical to identified mutant GIST, occurring in identical regions and expressing high levels of phosphorylated KIT (28), (Figure 1) The exact mechanism of this subset’s KIT activation is less clear but has been shown to consist of genetically heterogeneous groups associated with various mutations attributed to oncogenesis in other cancers (25). By identifying the various genetic variants associated with wild type GIST can potentially define the various mechanisms of disease and provide the potential for new biomarkers for targeted therapy.

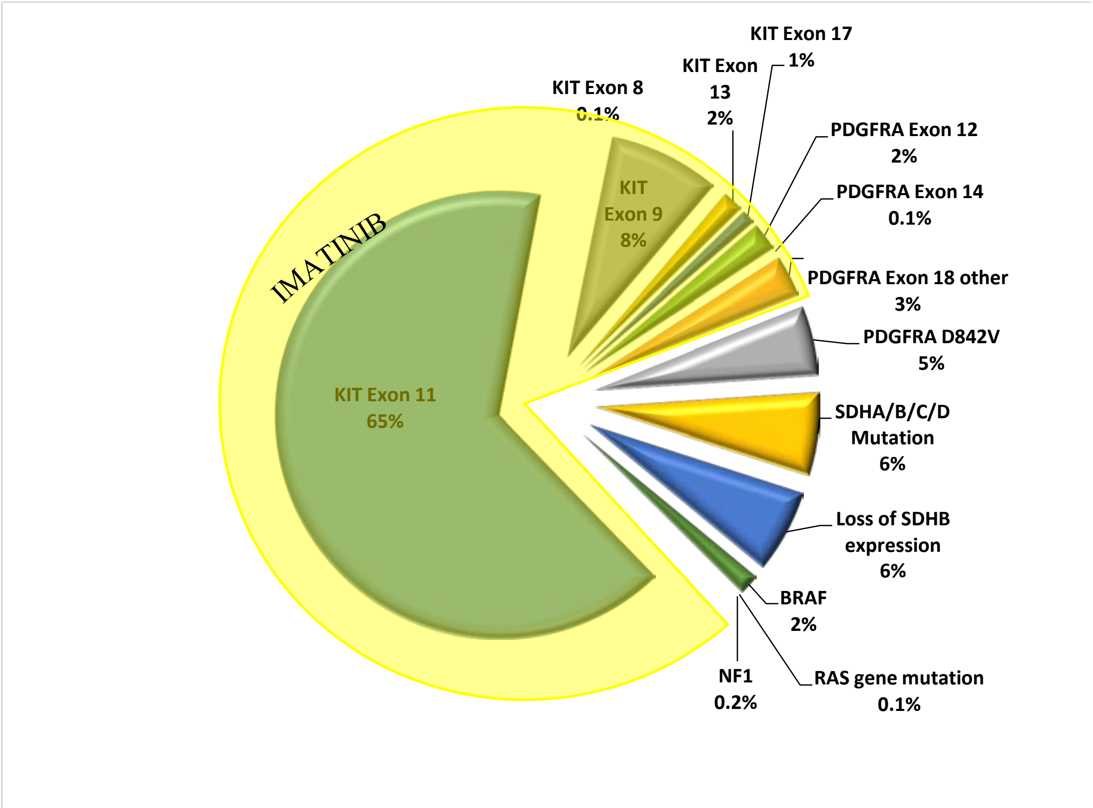


***Figure 1**: Genetic heterogeneity of gastrointestinal stromal tumors (GIST) and efficacy of imatinib mesylate (STI571, , Gleevec®).*

GIST has been shown to develop of resistance to treatment and is defined by primary and secondary resistance methods. Primary resistance develops within the first six months of treatment and is often a result of the varying sensitivities of the different KIT genotypes to imatinib(25). Additionally wild-type GIST displaying resistance has been shown to contain mutations downstream of the KIT that can be targeted directly with other specific inhibitors(29).

Secondary resistance, progression of disease after six months of treatment can occur through a multitude of mechanisms. The acquisition of secondary resistance mutations in KIT or PDGFrA often occurs in the same gene or allele as the driving primary KIT or PDGFrA mutation(30). Further evidence has shown intra and inter lesion genetic heterogeneity in GIST tumor. Leigl et al found 82% of GIST lesion contained secondary mutations after first-line treatment with tyrosine kinase inhibitors (TKI)(31). This level of oncogenic variability and the polyclonal nature of these tumors significantly impact the efficacy of the TKI inhibitors. The importance of defining the various oncogenic mechanisms is crucial to provide effective inhibitors to the disease.

The following study analyzed 41 GIST tumor lesions from 23 patients for cohort level summaries. Two lesions from single patient both extracted from the omentum were used to represent biological replicates for IDR analysis. The primary mutations showed two identical KIT mutations in both lesions, although inter tumor differentiation may occur between these lesions, the identical genetic morphology and localization provides a model with an expected minimal variation. Further metrics and visualizations will be used to assess these samples and to determine the underlying morphology that is consistent between them. Additionally, these metrics and visualization will present a comparative analysis of two copy number detection methods to assess their performance.

Methods

Whole Exome Sequencing Alignment

Alignment was performed against the UCSC Hg19 human reference using BWA MEM v 0.7.9a-r786 with default parameters. Picard Tools v1.119 was used to mark duplicated reads followed by realignment around potential insertion-deletion events using and GATK v3.2-2-gec30cee INDEL realigner and subsequent recalibration of base quality scores. Coverage of the exome was summarized also using GATK v3.2-2 Depth of Coverage, counting fragments only within the target regions specified by the Roche system capture kit, Nimblegen SeqCap Exome version 2 (**Table 1**).

Copy Number Detectors

ExomeCNV version 1.4 was implemented using R version 3.2.1 (12). This CNV detection method began by initially calculating the log coverage from the depth of coverage files generated by GATK. The variants were then detected using ExomeCNV with an input parameter expecting zero normal contamination in the tumor sample, assuming a pure tumor sample.

The tumor purity was not determined for the sequenced samples, although a heterogeneous tumor and normal mixtures is likely. Increasing the fraction of the potential normal contamination of tumor would in fact lead to less noise, increasing the specificity of the CNVs detected. However, the methods used in this analysis are dependent on a measurable noise component to distinguish the reproducible component signals from reproducible signal among biological replicates. Therefore, to avoid arbitrary assignment of tumor purity and to obtain the entire spectrum of potential signal components, essential for calculating the IDR, CNV detection was performed assuming no normal contamination in the tumor sample. Segmentation was carried about by ExomeCNV optimized for the area under curve, maximizing the sensitivity and specificity (12).

ExomeDepth version 1.1.5 was implemented using R version 3.2.1 (11). Count data was generated using the bam corresponding bam file and Nimblegen SeqCap Exome version 2 captured exome intervals. The CNV were called and segmented using the normal sample from the patient as the reference. Additional parameters included a tumor proportion of 100% to maximize sensitivity and a HMM state transition probability of 1e-03.

Results and Discussion

The 41 GIST lesions and corresponding paired normal form 23 patients had an average sequencing depth of 78x. **Table 1** summarizes alignment metrics for the two lesions that will be used as biological replicates in this analysis.

	Tumor1	Tumor2	Normal
Total Reads	194281725	186158109	1.94E+08
Unmapped Reads	2883984	2556289	2842003
Duplicated Reads	105080570	96771053	99903730
Total Coverage (Fragments)	5015975655	5.014E+09	5.35E+09
Average Coverage (Fragments/base)	106.6	106.55	113.79

Table 1. *Whole Exome Sequencing Results. Parameters from Whole Exome Sequencing (WES) of 2 tumor lesions from a single patient and the matched normal tissue.*

Context-guided Visualization

We note that during the iterative evaluation of the prototype for CNV, we were prompted by the interactive aspects of the system to develop and expand the initial queries. Importantly, it became clear that there were actually 3 context-specific comparisons that needed to be assessed: consistency in CNV calls across methods for single patient sample, consistency across tumor samples within the same patient, and consistency in CNV calls across paired samples and cohort-level summaries.

Six of the seven metrics could be applied to the three perspectives whereas the IDR required the use of biological replicates and was ideally applied to only the paired sample context. The metrics were combined into an interactive framework toolset consisting of three applications to address the contextual use cases, with interactive assessment of the reactive metrics from **Table 2**. A fourth framework was developed for the IDR, to initially compare the results obtained from both CNV detectors at both copy number losses and gains. However, because of lack of confidence score reported by one of the CNV detectors, the IDR could only be applied to one of the CNV detection methods, ExomeDepth.

Seven metrics and visualizations were assessed and implemented for WES for copy number detection (**Table 2**). Assessment of these metrics provides information on robustness of copy number detectors and would assist identifying high confidence CNVs. This assessment of the CNV methods will examine each of these metrics in our evaluation of ExomeCNV and Exome Depth the context-specific comparisons to visualize and assess these variant detectors.

Visualizations and Metrics for Comparison Exome Copy Number Variant (CNV)	
Distribution and sizes of CNVs	Comparative assessment of the region sizes can define an algorithms intended target region and can reveal biases from over fitting.
Number of LOF and GOF CNVs identified by each CNV detection method	Comparison of the number of regions can indicate information about the sensitivity and specificity of a particular detection method
Overlap of CNV LOF and GOF for different analysis methods.	Compare overlap of CNVs found by multiple detections methods.
Fraction of Reads in variant region	Determine whether significant regions are biased by coverage.
Circos plot LOF/GOF comparing analysis methods along genome	Visual comparison of size and location along the genome of CNVs detected by each method.
GC Enrichment Distributions	Diagnostic for copy number variant detection to ensure GC correction are being made and no severe biases
IDR Correspondence Curve and threshold	Irreproducible discovery rate to determine significance threshold of peaks based on reproducibility of biological replicates (10).

Table 2*. Visualizations and Metrics for Comparison of Exome Copy Number Variants (CNV).*

Distribution of CNV Sizes

The number of copy number variants detected by ExomeCNV was less than the CNVs called by ExomeDepth. ExomeCNV found 251 and 202 CNV events, while ExomeDepth called 110 and 121 CNV events for Tumor 1 and Tumor 2, respectively (**Table 3**). **Figure 2** indicates the size and distribution of the two copy number variant methods with both tumor replicates. The number of large CNV events, greater than 100kb, was larger with ExomeCNV, while moderately sized events between 10kb and 100kb were slightly lower. The increased frequency of large >100kb events may suggest a liberalness in the merging and segmentation of the two methods. ExomeCNV segmentation method using circular binary segmentation (CBS) may be more liberal in combining individual CNV segments the hidden Markov model that uses the likelihood across multiple exons. The larger number of events observed by ExomeCNV versus ExomeDepth can impacted by the statistical threshold and the model used. ExomeCNV's models the reads from sequencing using a Poisson model, which assumes the mean and variance are equal. Issues of over dispersion and deviation from model assumptions can impact the calls.

Figure 2 additionally shows a concerning lack of small or focal (≤1kb), copy number variants called by ExomeDepth for both tumor replicates. Somatic CNVs frequencies in cancer have been shown to be inversely proportional to length, except for chromosome arm length CNVs (20, 32). Chromosomal arm length CNVs are observed to occur with marked increase in frequency in somatic CNVs (32). The lack of these large mega base long variants detected by ExomeDepth is most likely caused by the initiating parameters during the segmentation step of CNV. Exome Depth uses HMM and assigns the hidden Markhov states to each segment, determining whether that segment is merged or segmented from adjacent regions, and is set at a default of 50kb regions, which can impact the results.

	ExomeCNV				ExomeDepth			
	Tumor1		Tumor2		Tumor1		Tumor2	
Total called regions	251		202		207		202	
CNV Type	Gains	Losses	Gains	Losses	Gains	Losses	Gains	Losses
Number of CNV events	102	149	103	99	17	190	36	166
Minimum Size (bases)	114	139	122	136	1970	82	498	73
Maximum Size (bases)	142.2 x10^6	87.9 x10^6	180.7 x10^6	120.6 x10^6	70600	55100	36000	45500
Average Size (bases)	18.5 x10^6	6.98 x10^6	18.9 x10^6	10 x10^6	89000	2120000	275000	1340000

Table 3. *Overview Frequency and size of CNVs detected by ExomeCNV and ExomeDepth (11, 12)*

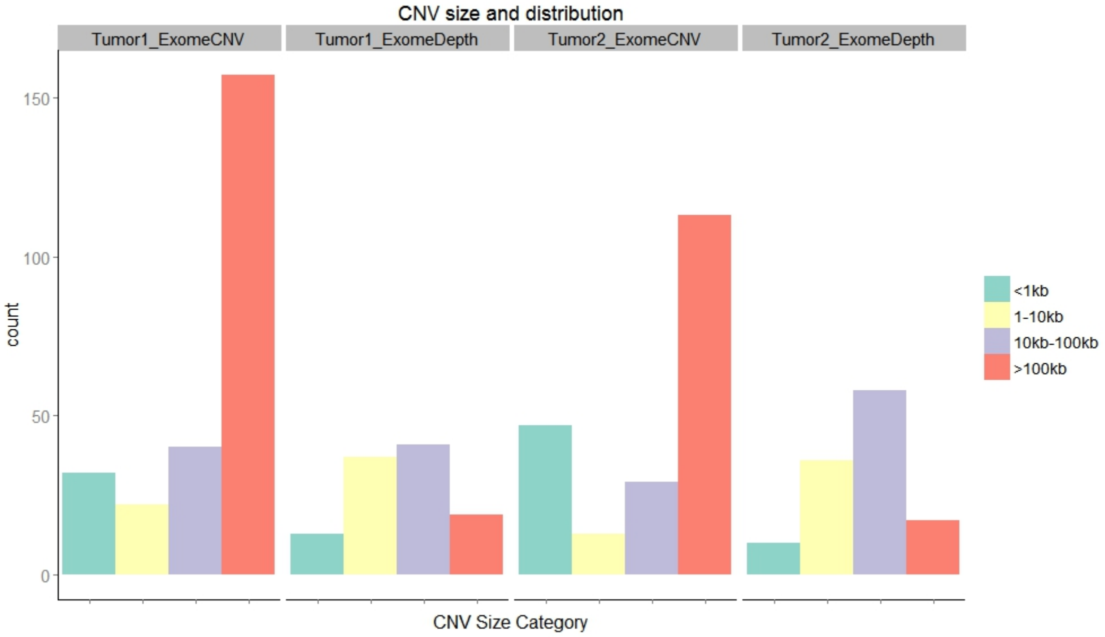


Figure 2. *Size and Distribution of CNV Size for two biological gastrointestinal stromal tumor (GIST) duplicates, Tumor 1 and Tumor 2 from a single patient evaluated using two copy number variant detectors ExomeCNV (12) and ExomeDepth (11).*

Number of LOF and GOF CNVs identified by each CNV detection method and

The frequency of the copy number gains or duplications is significantly less than copy number losses detected with the different methods (**Table 3**). ExomeDepth only detected 31 of the combined CNV events, ExomeCNV, comparatively found 205 of the 452 events consisting of duplications.

The size of the region identified by ExomeDepth is also markedly smaller, with sizes ranging from an average of 1283 bases to 1.34 mega bases (Mb), while comparatively ExomeCNV detected events from 114 bases to 180.7 Mb. The average chromosome arm in the human genome reference Hg19 is approximately 65 Mb. The broad spectrum of events captured by

ExomeCNV and the significantly smaller frequency and size of CNV events identified by ExomeDepth suggests that there is a stricter statistical threshold for identifying CNV events as well as stricter thresholding of segmentation.

Fraction of Reads in variant region

The fraction of reads in CNV regions is drastically different for the two CNV detection methods (**Table 4**). ExomeCNV regions is shown in **Figure 3** to provide a copy number state for each region in the entire genome, including regions assigned as a normal or diploid state. The results can then be filtered based on a measurable variant region, such as the reads ratio of tumor to normal. However, ExomeDepth only provides regions that observed some change in copy number. The method of segmentation may be involved in the observed discrepancy in the abundance of signal. Circular binary segmentation is used for ExomeCNV, in which regions are recursively merged or differentiated across the entire genome, regardless of their copy number state or statistical differentiation. The hidden Markov chain, method used by ExomeDepth for segmentation, uses the particular copy number state at a region to determine whether to merge or segment the region into a CNV state. If no CNV state change is observed, neither the region nor the state of the region is flagged within the indicated CNV results.

	Tumor 1	FRIR Tumor1	Tumor2	FRIR Tumor2
Total Reads	197287594		188805115	
ExomeCNV	195569545	0.991	187256272	0.992
ExomeDepth	3754029	0.019	3336425	0.018

***Table 4**. Fraction of reads in CNV region. The total amount of reads called in detected region by copy number detector.*

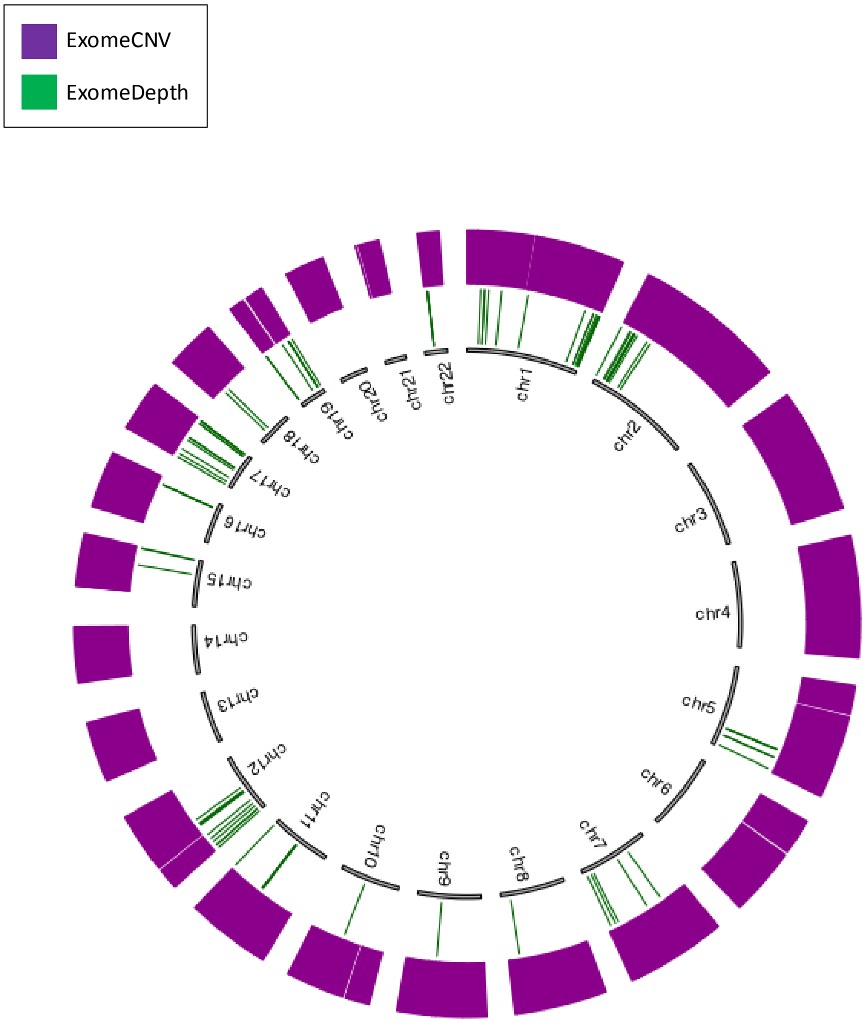


Figure 3. *Circos plot of Tumor 1 from of the copy number results by two CNV detection methods, ExomeCNV and ExomeDepth. Screenshot of R shiny application CNV_Method_Comparison, focusing on method comparisons within the same sample and subject for high confidence calls.*

The results reported from copy number detection methods differ substantially. ExomeCNV provides the copy number for each regions of coverage over the entire genome, with only the reads ratio to provide a method of filtering for regions of potential copy variant. However, the reported regions for ExomeDepth provide the reads ratio along with the Bayes factor for that region. The Bayes factor provides a measure of the confidence level for a particular variant call. No measure of confidence was observed, however, for ExomeCNV methods, with only the option of filtering by reads ratio available.

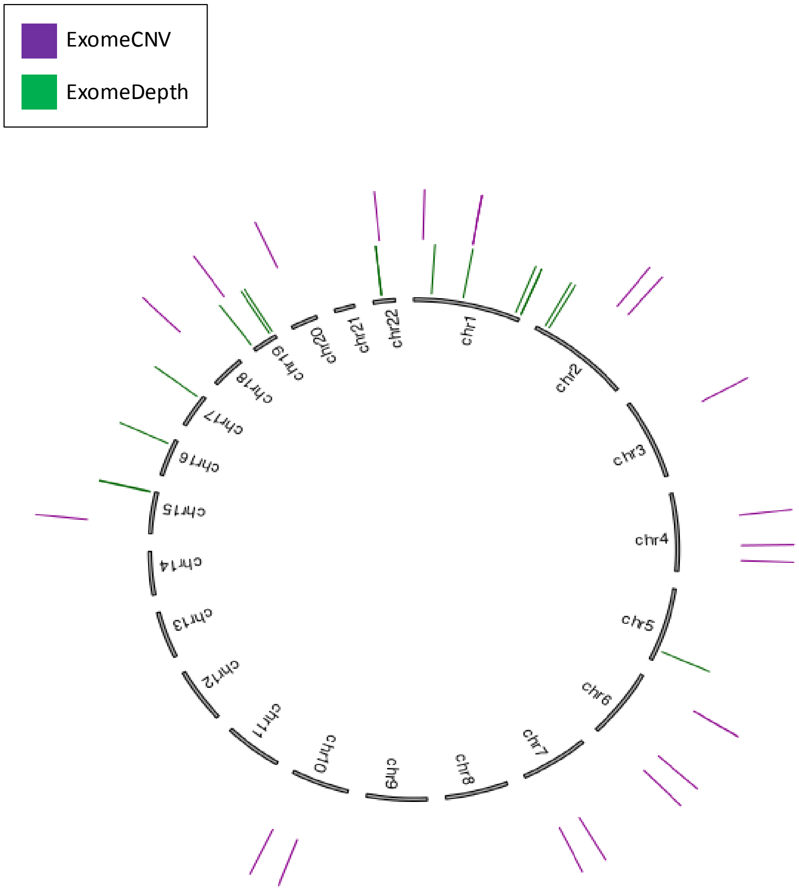


***Figure 4**. Circos of Tumor 1 of CNV regions called by two CNV detection methods, ExomeCNV and ExomeDepth. Screenshot of R shiny application CNV_Method_Comparison, focusing on method comparisons within the same sample.*

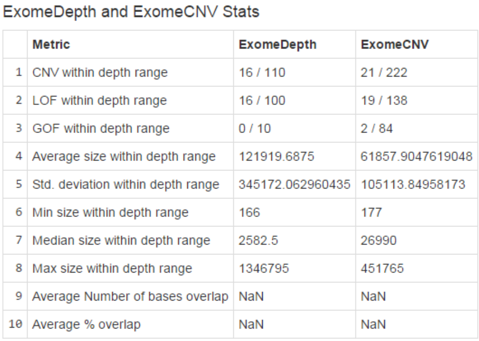
Filtered by $\log_2$ reads ratio < = -1 or $\log_2$ reads ratio >= 0.585.

The results were filtered by reads ratio (R) with $\log_2$ reads ratio < = -1 or $\log_2$ reads ratio >= 0.585, corresponding to reads ratios of 0.5 and 1.5, respectively. Filtering by the reads ratio led to a significant decrease in the amount of regions observed by ExomeCNV, the unfiltered variants depicted in **Figure 3** and the filtered variants in **Figure 4**. While ExomeDepth regions went from an average of 204 copy number regions called to an average of 44 regions when filtered by reads ratio. The ExomeCNV regions decreased from an average of 227 regions to approximately 24 variant regions after filtering for an extreme reads ratio.

Additionally, although there are some regions in which ExomeCNV and ExomeDepth CNVs are adjacent or flanking variant calls, for example in chromosome 1, 19 and 22 (**Figure 4**), there are no regions that overlap when filtered by reads ratio.

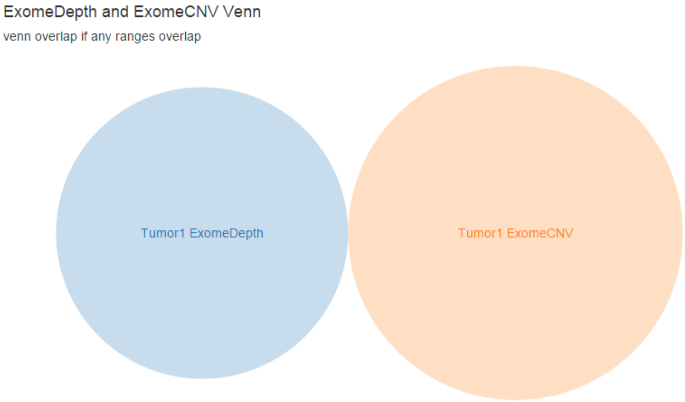
Overlap of CNV for different analysis methods

Figure 5 shows the statistical summaries and Venn diagram showing the correspondence of these log reads ratio (logR) filtered regions. None of the logR thresholded CNV ranges observed overlapped with one another across both methods. The lack of congruency emphasizes the fundamental differences in these detection methods, which led to different results. Because of the lack of overlap between log reads ratio thresholded regions the subsequent CNV summaries were not filtered by the log reads ratio threshold.



ExomeDepth and ExomeCNV Stats

	Metric	ExomeDepth	ExomeCNV
1	CNV within depth range	16 / 110	21 / 222
2	LOF within depth range	16 / 100	19 / 138
3	GOF within depth range	0 / 10	2 / 84
4	Average size within depth range	121919.6875	61857.9047619048
5	Std. deviation within depth range	345172.062960435	105113.84958173
6	Min size within depth range	166	177
7	Median size within depth range	2582.5	26990
8	Max size within depth range	1346795	451765
9	Average Number of bases overlap	NaN	NaN
10	Average % overlap	NaN	NaN



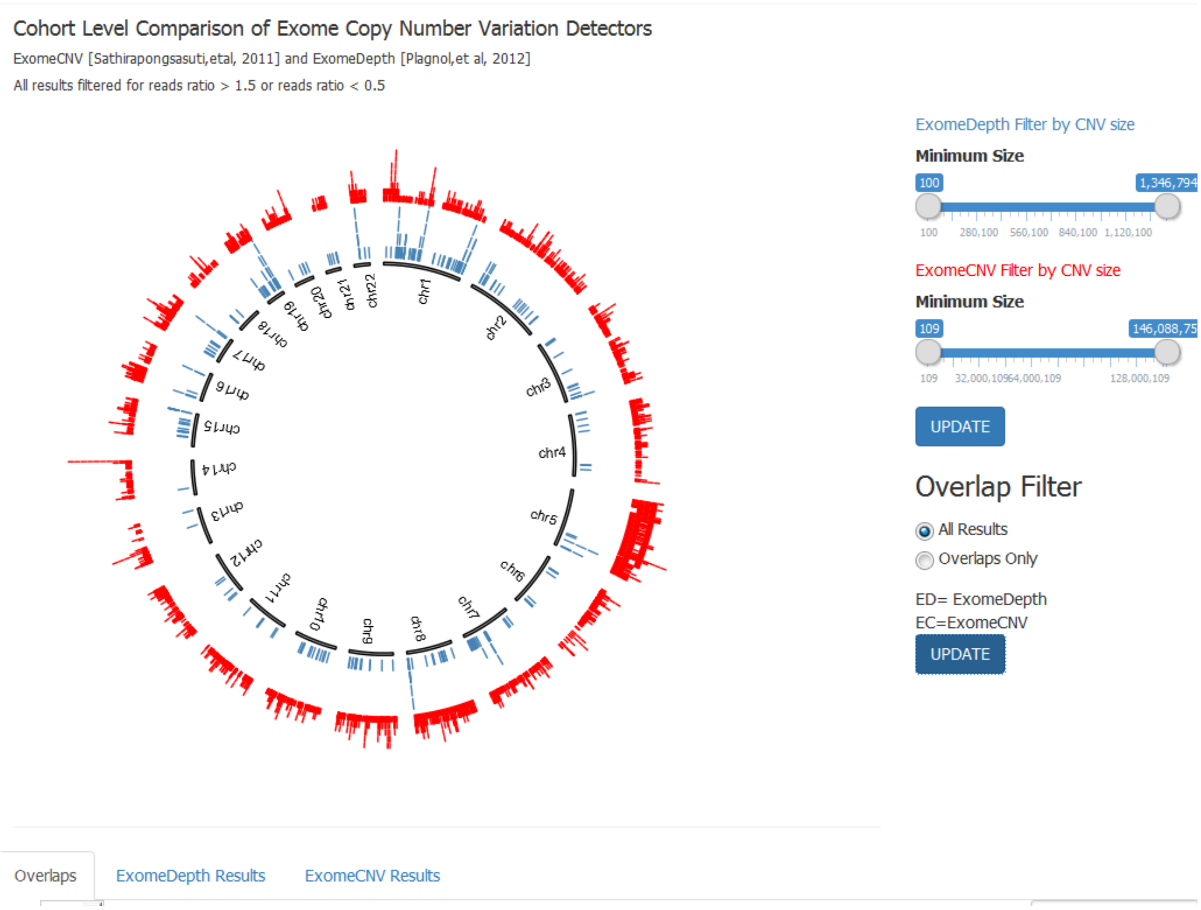
ExomeDepth and ExomeCNV Venn

venn overlap if any ranges overlap

Table 5. *Metrics of the overlap.*

Figure 5. *Venn diagram of overlap. Overlap of two copy number variant calls, ExomeDepth and ExomeCNV, for a single GIST lesion filtered by reads ratio. Screenshot of R shiny application CNV_Method_Comparison, focusing on method comparisons within the same sample.*

The cohort context interactive visualizations (**Figure 6**) summarized CNVs across the 41 lesions analyzed, potentially identifying consensus variant regions that are common to the GIST copy variant profile. For example 13 LOF regions were detected by both ExomeDepth and ExomeCNV that overlapped in chromosome 22q. The loss of 22q is a common GIST copy variation (25). The cohort level summary can determine whether individual biases within a sample prevents congruency in highly confident or obvious variant regions. For example a poor quality paired normal in a single sample comparison of methods may decrease the sensitivity of the CNVs that can be detected. However, the cohort level summary reinforces that the lack of overlap at reads ratio threshold.



Cohort Level Comparison of Exome Copy Number Variation Detectors

ExomeCNV [Sathirapongsasuti,etal, 2011] and ExomeDepth [Plagnol,et al, 2012]

All results filtered for reads ratio > 1.5 or reads ratio < 0.5

ExomeDepth Filter by CNV size

Minimum Size: 100 – 1,346,794

ExomeCNV Filter by CNV size

Minimum Size: 109 – 146,088,75

Overlap Filter

● All Results

○ Overlaps Only

ED= ExomeDepth
EC=ExomeCNV

***Figure 6**. Screenshot of R shiny application CNV_cohort, comparing the copy number variants (CNVs) called by two variant detection methods, ExomeDepth and ExomeCNV, across 41 gastrointestinal stromal tumor (GIST) lesions from 23 patients.*

The method comparison with a single sample context is able to observe and compare results CNV detection methods at a basic level. A comparison of copy number regions that overlap between the two detections methods, ExomeDepth and ExomeCNV, were detected in a single sample, and were unfiltered by reads ratio can be observed in **Figure 6**.

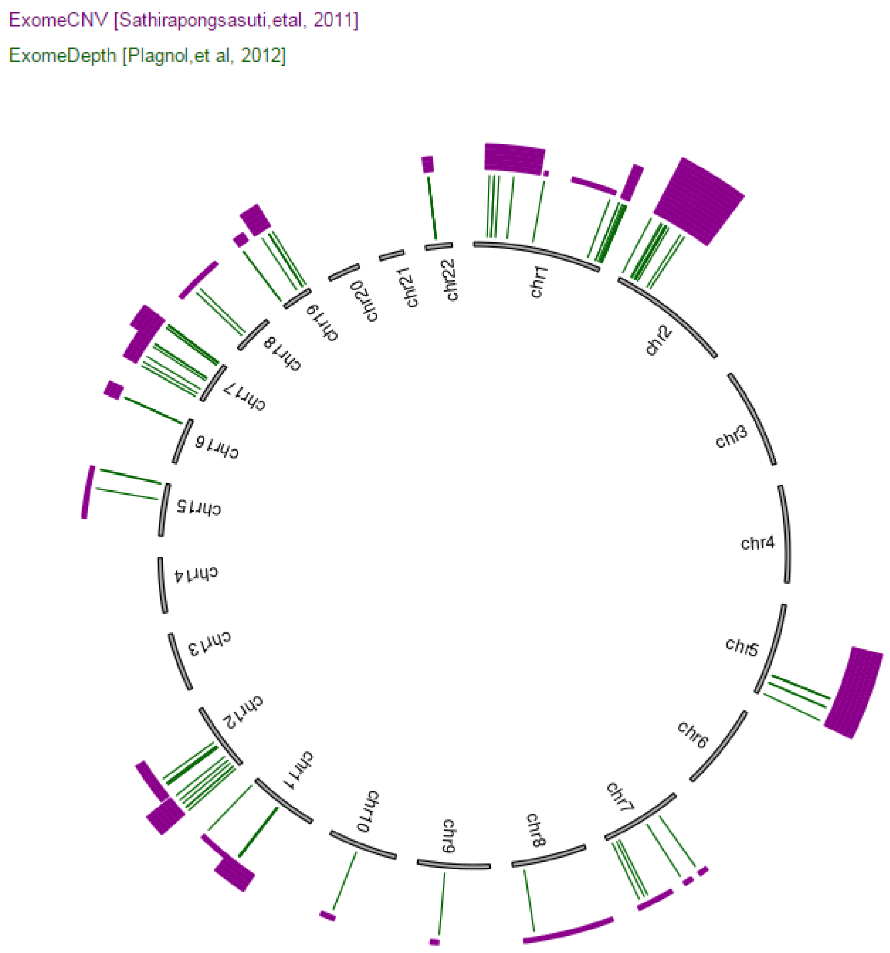
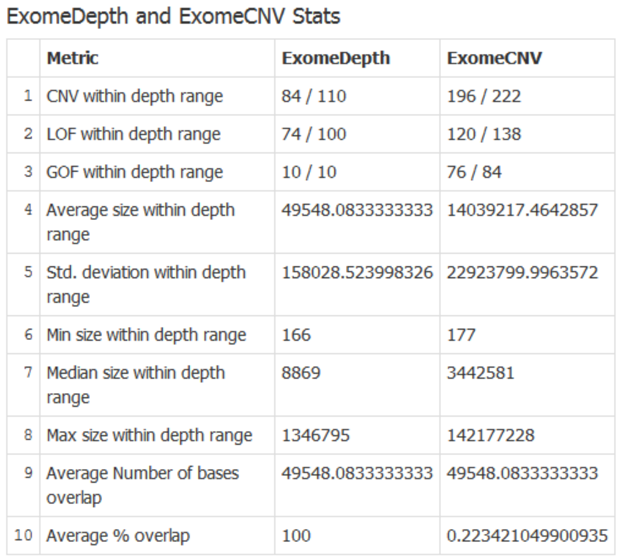


Figure 7. *Circos of Tumor 1 of CNV regions called by two CNV detection methods, ExomeCNV and ExomeDepth. Screenshot of R shiny application CNV_Method_Comparison, focusing on method comparisons within the same sample. This is not filtered by log reads ratio, filtered by overlapping regions between CNV detection methods.*



ExomeDepth and ExomeCNV Stats

	Metric	ExomeDepth	ExomeCNV
1	CNV within depth range	84 / 110	196 / 222
2	LOF within depth range	74 / 100	120 / 138
3	GOF within depth range	10 / 10	76 / 84
4	Average size within depth range	49548.0833333333	14039217.4642857
5	Std. deviation within depth range	158028.523998326	22923799.9963572
6	Min size within depth range	166	177
7	Median size within depth range	8869	3442581
8	Max size within depth range	1346795	142177228
9	Average Number of bases overlap	49548.0833333333	49548.0833333333
10	Average % overlap	100	0.223421049900935

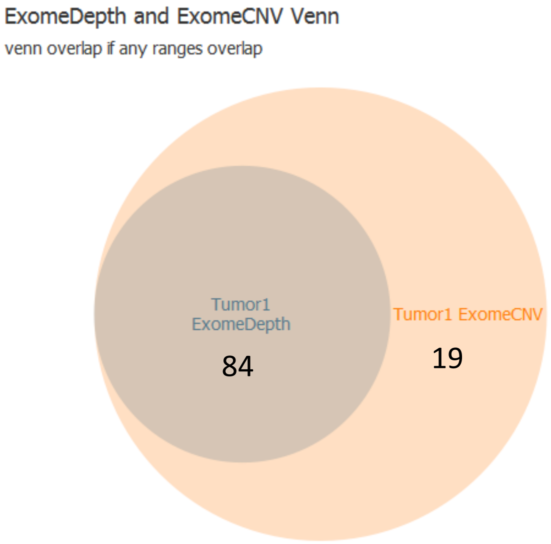


Table 6. *Metrics of the overlap for ExomeDepth and ExomeCNV*

Figure 8. *Venn diagram of overlap. Overlap of two copy number variant calls, ExomeDepth and ExomeCNV, for a single GIST lesion filtered for overlapping regions between the two methods. Screenshot of R shiny application CNV_Method_Comparison, focusing on method comparisons within the same sample..*

The sex chromosomes X and Y were removed from the analysis since highly repetitive regions can cause biases in mapping and variant detection (33).

Table 6 emphasizes the large discrepancy in the size of regions called, at 50 kilo bases for ExomeDepth and 14 mega bases for ExomeCNV. This difference in the size of regions was apparent in the observation of the distribution of reads but reinforced in the interactive visualization framework developed within this context.

Additionally there are regions of high confidence, in consensus areas for both CNV detection methods (**Figures 5,7,8, Tables 5-6).** Low confidence regions can also be flagged by identifying regions where variant detectors assign the opposite type of variant. These contradictory CNV regions, for example, when ExomeDepth calls a duplication and ExomeCNV identifies a deletion in overlapping regions can be flagged as potential areas where the algorithm or statistical model dealt with the differentiating read depth in very different ways.

These overlapping regions (**Figure 7,8**) were found to consist entirely of ExomeDepth regions which were contained entirely within ExomeCNV regions. These regions consisted of ranges of congruency, in which an ExomeDepth range is nested entirely within a larger ExomeCNV variant range with an congruent assignation of copy number deletion or duplication. However, the congruent regions only consisted of 27 of the 84 overlapping regions. The majority of the overlapping regions were discrepant in the copy number variant that were called for a particular range. The large amount of discrepancy is explained by the drastic differences in these methodologies. The statistical model for significance of read ratios and the method of segmentation differs but the fundamental calculation of the reads ratio differs as well. ExomeCNV defines the reads ratio as in Equation 1.

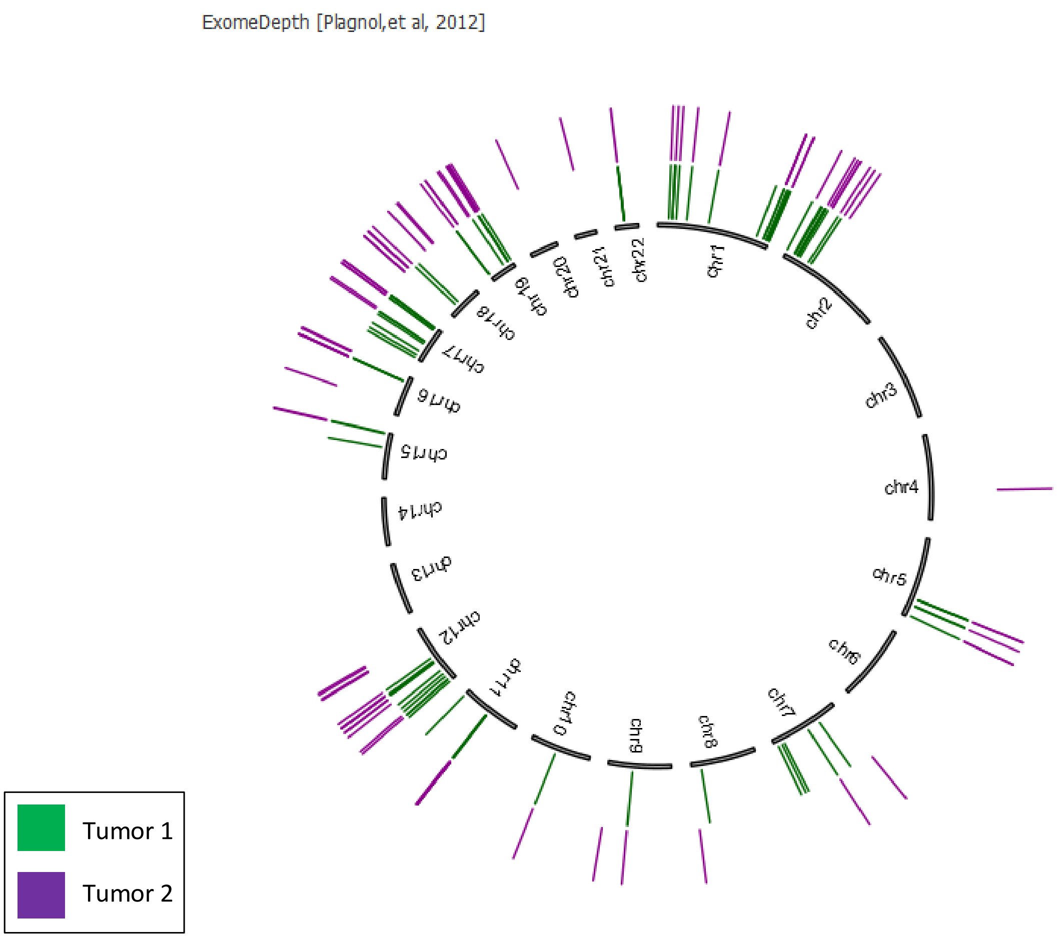
Equation 1: $$Reads\ Ratio = \frac{X/_{Nx}}{Y/_{Ny}}$$

Where X and Y denote the number of reads mapped within the exon or segment being observed in the case (tumor) and control (matched normal). Nx and Ny representing the total number of aligned reads in the case and control, respectively (11).

ExomeDepth calculates read count ratio as the ratio of observed counts over the expected counts from the statistical model0. Preliminary read count ratios are calculated with Equation 2, with Xrepresenting exonic read count for the test sample (tumor) and Y the exonic read count for the reference sample (normal).

Equation 2: $Read\ count\ ratio = \frac{X}{X+Y}$

There are a multitude of reasons for these contradictory calls. For example, the read counts calculated by ExomeDepth were filtered for reads with a mapping quality of 20 or higher, perhaps decreasing the read counts in specific low quality locations, which led to arbitrary skewing of the copy number at that region when compared to the normal sample. Despite this disparity the CNV callers identified 27 congruent regions which can be categorized as relatively confident regions of the copy number variation.



ExomeDepth [Plagnol,et al, 2012]

Figure 9. *Circos of Tumor 1 and Tumor 2 of CNV regions called by ExomeDepth. Variant region are unfiltered by reads ratio. Screenshot of R shiny application CNV_Sample_Comparison, focusing on comparison of paired samples and the consensus signal produced by different CNV detection methods.*

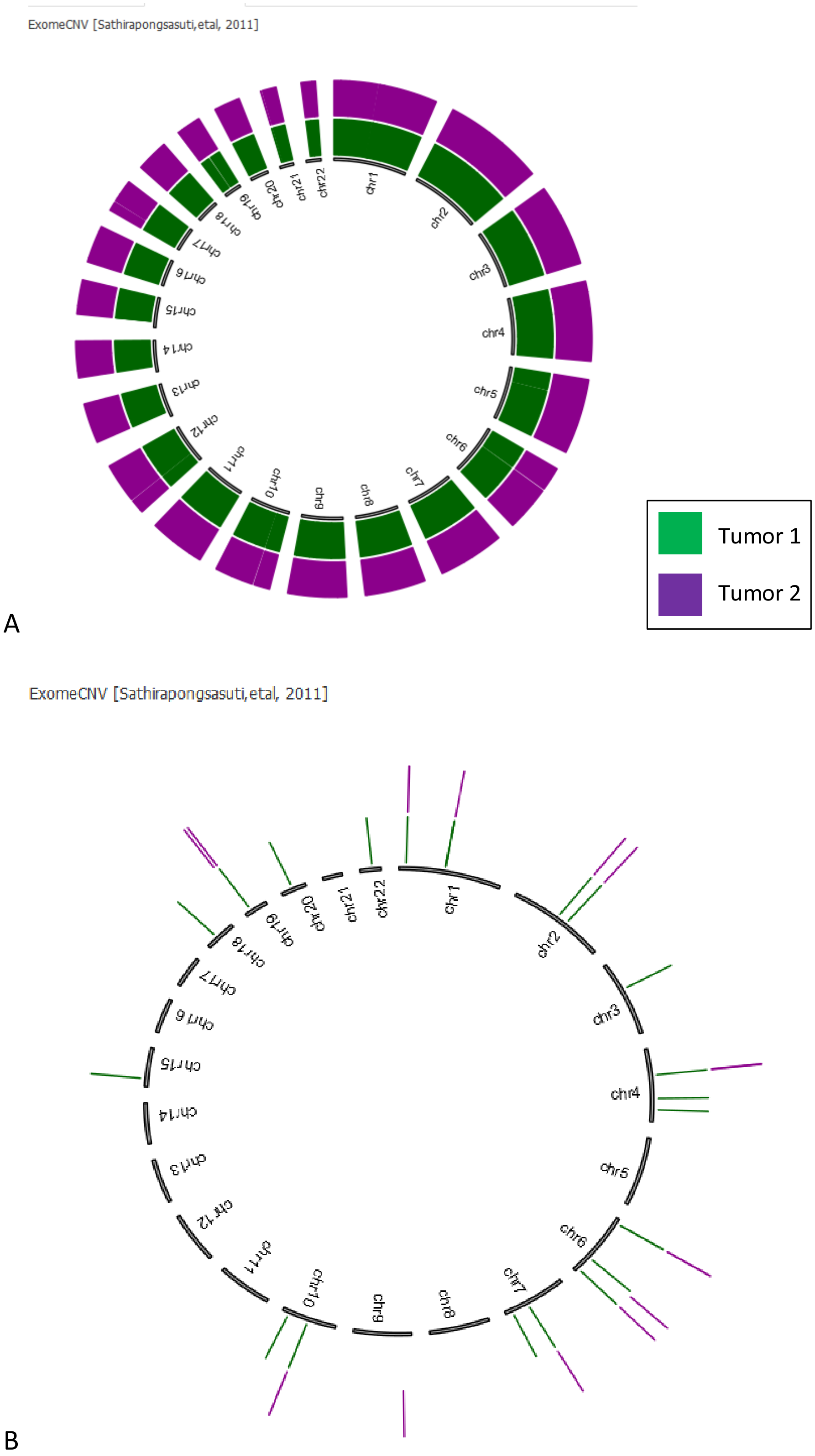
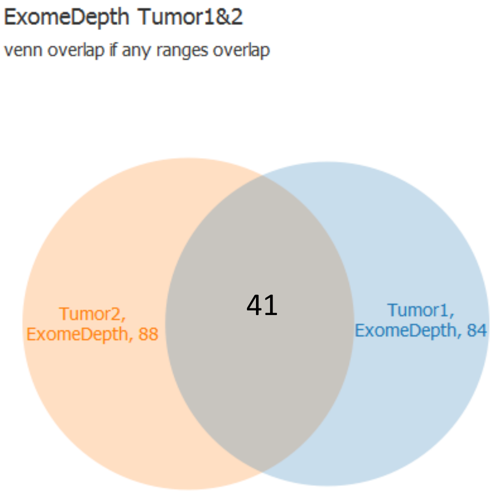
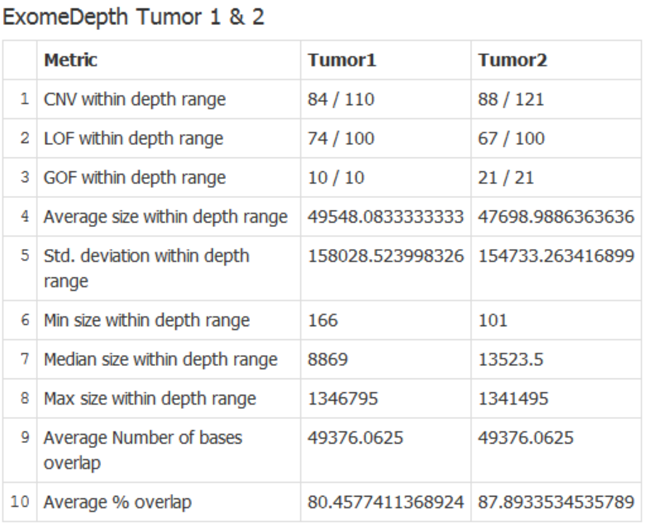


Figure 10. *Circos of Tumor 1 and Tumor 2 of CNV regions called by ExomeCNV. A) Unfiltered ExomeCNV regions provides copy number state for the entire genome. B) ExomeCNV regions*

filtered by variant regions, reads ratio of < 0.5 and > 1.5. Screenshot of R shiny application CNV_Sample_Comparison, focusing on comparison of paired samples and the consensus signal produced by different CNV detection methods.

The resulting context specific comparing the biological replicates depicts the consistency of the method. The differences in the two samples can be distinguished as unique to the particular tumor, and can either result from noise or inter-tumor variability. The consensus regions between these replicates could also help identify regions of high confidence and for calculation of the IDR. The variants detected are fairly consistent among biological replicates (**Figure 10, 11**), for both CNV methods.



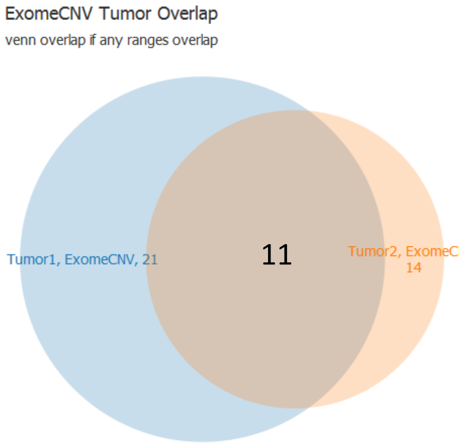


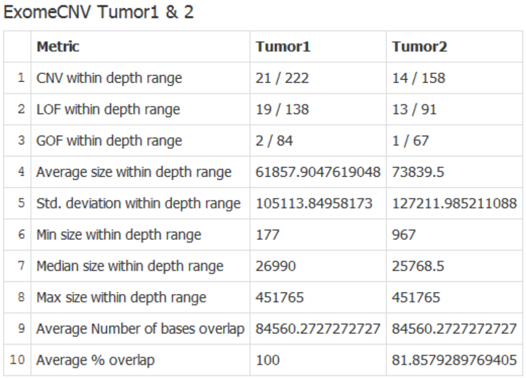
ExomeDepth Tumor 1 & 2

	Metric	Tumor1	Tumor2
1	CNV within depth range	84 / 110	88 / 121
2	LOF within depth range	74 / 100	67 / 100
3	GOF within depth range	10 / 10	21 / 21
4	Average size within depth range	49548.0833333333	47698.9886363636
5	Std. deviation within depth range	158028.523998326	154733.263416899
6	Min size within depth range	166	101
7	Median size within depth range	8869	13523.5
8	Max size within depth range	1346795	1341495
9	Average Number of bases overlap	49376.0625	49376.0625
10	Average % overlap	80.4577411368924	87.8933534535789

Figure 11. Venn Diagram of Overlap

Table 7. *Statistical summary of Overlap. Overlap of Tumor 1 and Tumor2 CNV regions detected by ExomeDepth Screenshot of R shiny application CNV_Sample_Comparison, focusing on comparison of paired samples and the consensus signal produced by different CNV detection methods.*





ExomeCNV Tumor1 & 2

	Metric	Tumor1	Tumor2
1	CNV within depth range	21 / 222	14 / 158
2	LOF within depth range	19 / 138	13 / 91
3	GOF within depth range	2 / 84	1 / 67
4	Average size within depth range	61857.9047619048	73839.5
5	Std. deviation within depth range	105113.84958173	127211.985211088
6	Min size within depth range	177	967
7	Median size within depth range	26990	25768.5
8	Max size within depth range	451765	451765
9	Average Number of bases overlap	84560.2727272727	84560.2727272727
10	Average % overlap	100	81.8579289769405

***Figure 12**. Venn of Overlaps*

***Table 8.** Statistical summary of CNV overlaps of biological duplicates Tumor 1 and Tumor2 GIST samples analyzed by ExomeCNV. Screenshot of R shiny application CNV_Sample_Comparison, focusing on comparison of paired samples and the consensus signal produced by different CNV detection methods.*

Irreproducible discovery rate (IDR)

The IDR approach provides a ranked list of significant values between two biological replicates (10). A measure of correspondence between the two replicates is calculated based on these ranked values. However, only one of the CNV detection methods reported a significance score for the copy number state of a genomic region. ExomeDepth reported a Bayes Factor value, described as the $\log_{10}$ likelihood ratio of a variant call divided by the normal copy number (11). The IDR was applied to Bayes factor values obtained from ExomeDepth, measuring the correspondence between the biological replicates, tumor 1 and tumor2 and generated the correspondence curves in **Figure 13**. The correspondence curve attempts to find the junction at which the IDR, a measure of the correspondence of the biological replicates no longer becomes reproducible. The transition can be easier to obtain and visualize from the derivative of the correspondence plot (10). The proportion of signal that was estimated as reproducible between the two biological replicates was 0.6 of the total reproducible ExomeDepth signal.

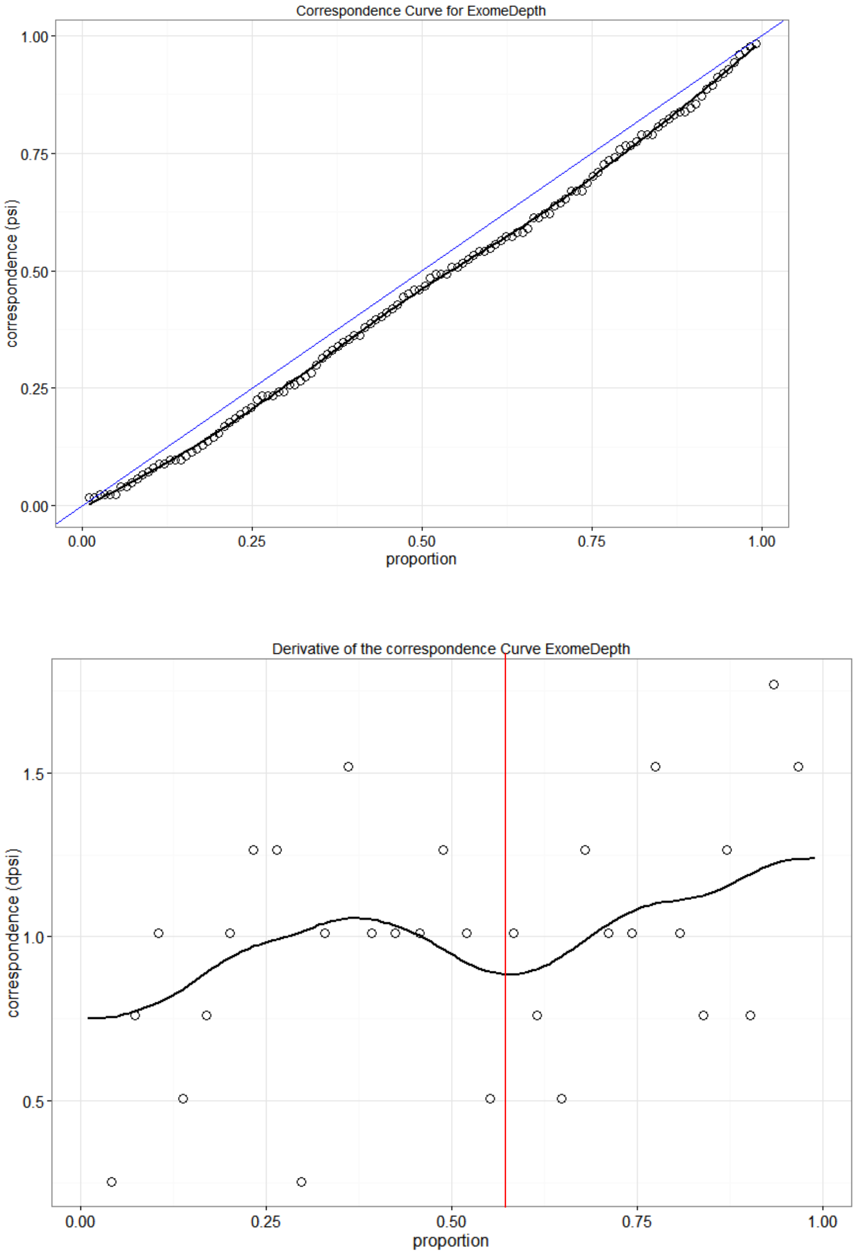


Figure 13. *Correspondence curves for the CNV detector ExomeDepth. Correspondence between copy number variants detected on Tumor 1 and Tumor2. (A) The correspondence curves of the IDR versus the ranked proportion. (B) The derivative of the correspondence cure. The red line indicates the threshold used as the threshold for reproducibility to calculate the*

IDR.

The next step in estimating the IDR for each reproducible output from the replicates is applying the IDR's statistical model and estimating the posterior probability for each observation (10). The model was fit with a reproducible proportion of 0.6 and 20 iterations, which was optimized based on the approximate proximity of the all the estimated parameters, i.e. mean, standard deviation, correlation, of the model to the calculated parameters of the reproducible component.

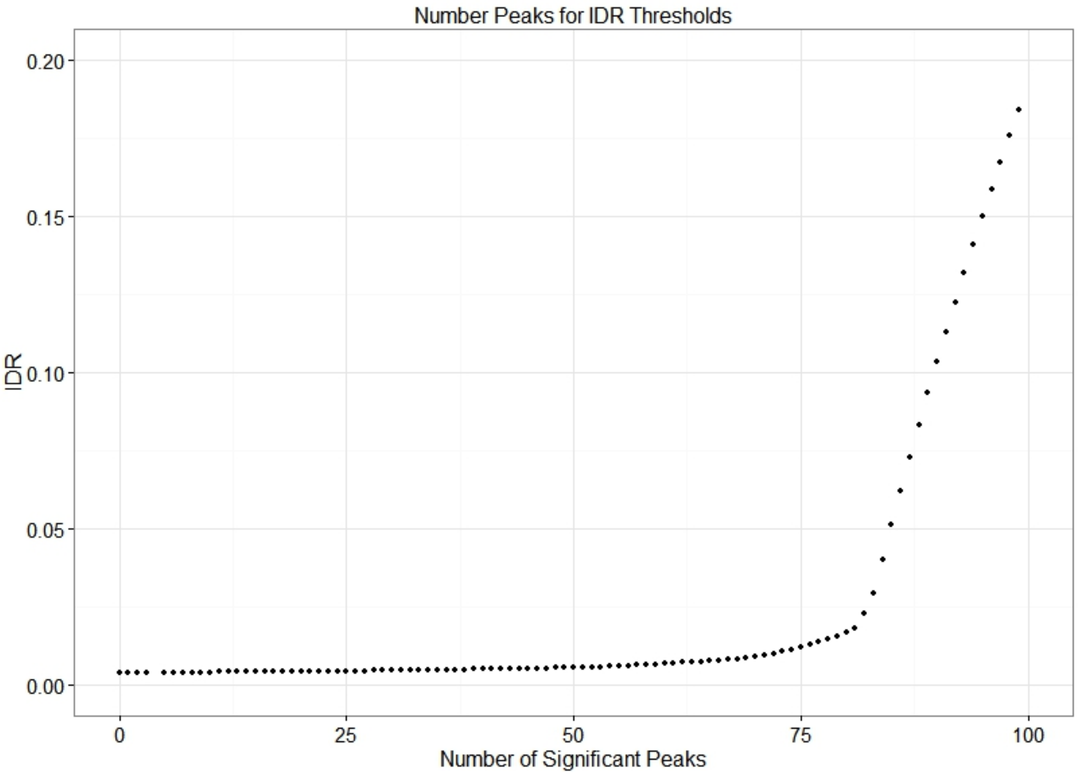


Figure 14. *Number of Peaks at IDR Thresholds. Amount of CNV regions included at varying IDR thresholds for ExomeDepth. The lower IDR represents increased correspondence between the biological replicates, Tumor 1 and Tumor2.*

A depiction of the number observations included at IDR thresholds (**Figure 14**) can help us identify the proportion of signal captures at various IDR thresholds. The threshold of irreproducibility is less apparent because of the limited signal observed by ExomeDepth. The IDR assumes that the ranked significance score consists of a significant amount of signal of both noise and truly biologically relevant signal (34). This method of CNV detection reported 110 variants for Tumor 1 and 121 CNVs in Tumor2, which is low compared to other CNV callers (11). Because of the lack of significance score reported for the signal in ExomeCNV, a measure of the IDR could not be performed. The output from ExomeCNV can instead by thresholded by the corresponding reads ratio. I must note, however, that this threshold is arbitrary and is does not have the statistical robustness of the IDR.

GC Enrichment Distributions

The percentage of guanine and cytosine (GC) bases in a genomic regions varies along the human genome, and has been found to effect the read coverage on next generation sequencing platforms, by affecting primer annealing. A study of sequencing platform biases found a positive correlation between read coverage and GC content when GC percentage was between 24 to 47% (35).

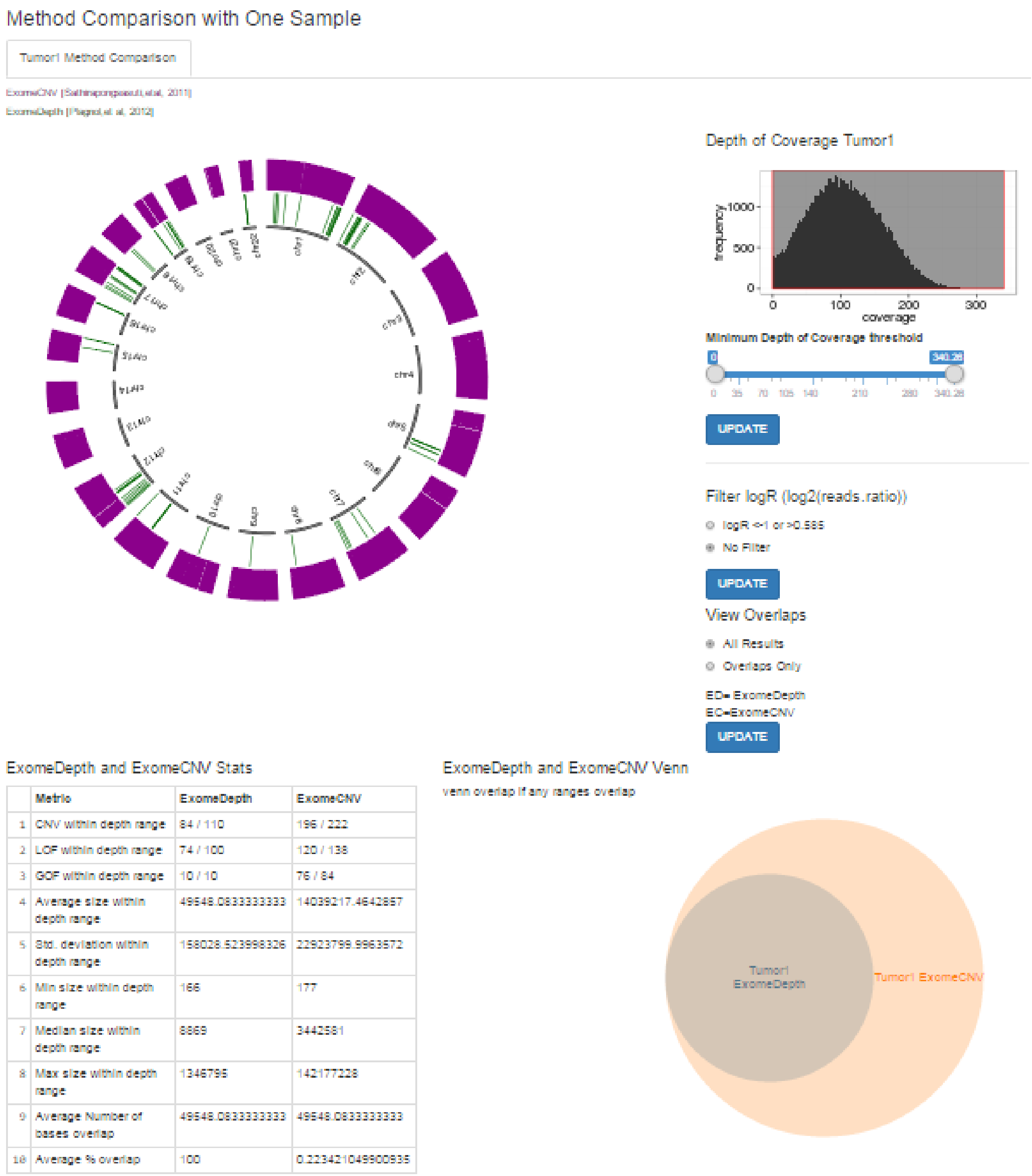
GC correction is performed by normalizing the read count for the GC percentage for each window of detection across the genome during segmentation analysis. This adjusted read count is then associated with the copy number to the corresponding genomic region (18).

However, study designs which use a case control or matched tumor and normal pair avoids the issue of GC enrichment. The associated read counts are differentiated from a paired tumor and normal sample are from the same patient and sequenced on the same platform. The effect of GC bias is cancelled out by directly comparing the genomes at each region (16, 18).

Initial filtering of all CNV variants from both ExomeDepth and ExomeCNV lead to considerable decrease in the number of CNV observed by ExomeCNV.

Conclusion

The evolution of the metrics development and context of visualization illustrated how a "hands-on" interactive interface can change the way we interact with the data and think about it, facilitating discovery and the transfer of knowledge. This can be clearly seen in the CNV example where three different context specific needs were identified after initial testing of the prototype: consistency in CNV calls across methods within a single tumor sample from a single patient (**Figures 3,4, 5, 7, 8, 15**), consistency across tumor samples within the same patient (**Figures 9, 10, 16**) and consistency in CNV calls across paired samples and cohort-level summaries (**Figures 6, 17**). The development of these interactive interfaces can communicate a multitude of metrics for the comparison of computational methods or samples, allowing the user to focus on the question asked and the data in front of them.



Method Comparison with One Sample

Tumor1 Method Comparison

ExomeCNV [Sathirapongsasuti,et.al, 2011]

ExomeDepth [Plagnol,et. al, 2012]

Depth of Coverage Tumor1

Minimum Depth of Coverage threshold

Filter logR (log2(reads.ratio))

○ logR <-1 or >0.585

● No Filter

View Overlaps

● All Results

○ Overlaps Only

ED= ExomeDepth

EC=ExomeCNV

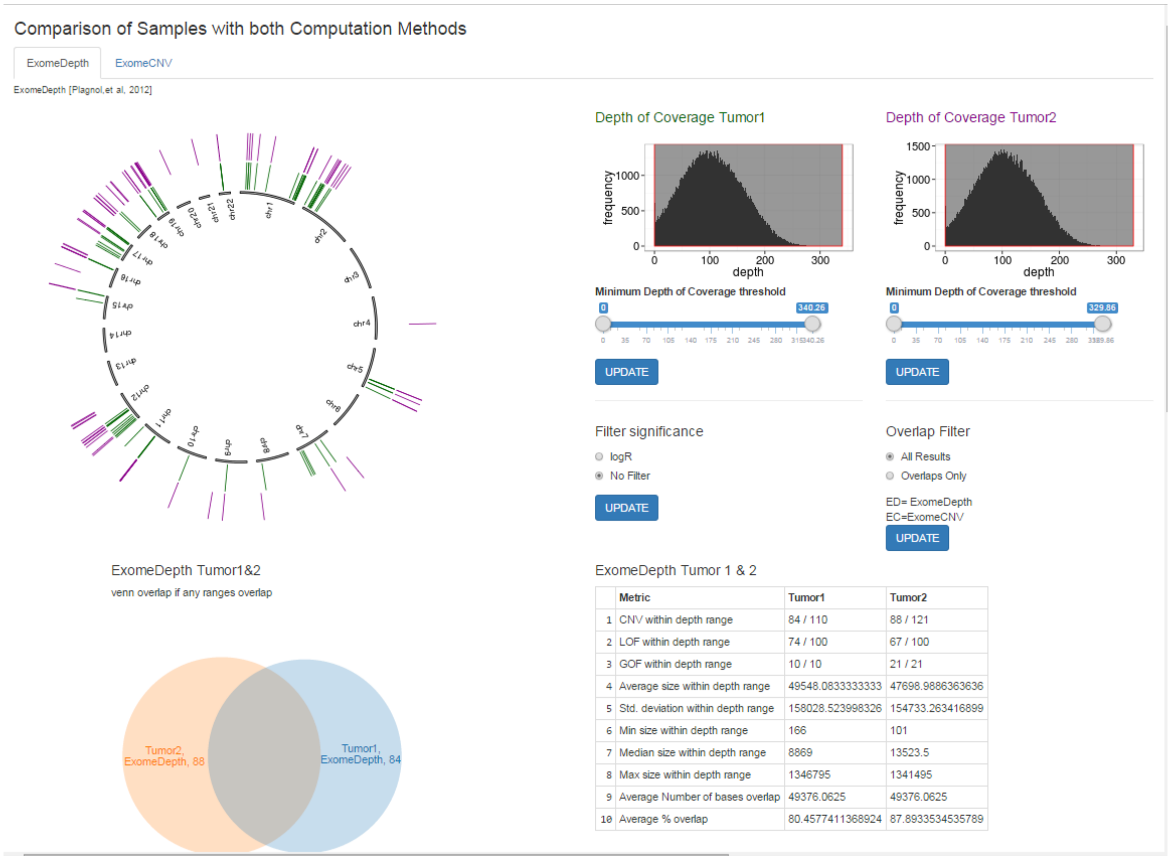
ExomeDepth and ExomeCNV Stats

	Metric	ExomeDepth	ExomeCNV
1	CNV within depth range	84 / 110	196 / 222
2	LOF within depth range	74 / 100	120 / 138
3	GOF within depth range	10 / 10	76 / 84
4	Average size within depth range	49548.0833333333	14039217.4642857
5	Std. deviation within depth range	158028.523998326	22923799.9963572
6	Min size within depth range	166	177
7	Median size within depth range	8869	3442581
8	Max size within depth range	1346795	142177228
9	Average Number of bases overlap	49548.0833333333	49548.0833333333
10	Average % overlap	100	0.223421049900935

ExomeDepth and ExomeCNV Venn

venn overlap if any ranges overlap

Figure 15. *Method comparison with one sample. This framework observes the consistency in CNV calls across methods within a single tumor sample from a single patient*



Comparison of Samples with both Computation Methods

ExomeDepth [Plagnol,et al, 2012]

Depth of Coverage Tumor1

Depth of Coverage Tumor2

Minimum Depth of Coverage threshold

Minimum Depth of Coverage threshold

Filter significance

○ logR
● No Filter

Overlap Filter

● All Results
○ Overlaps Only

ED= ExomeDepth
EC=ExomeCNV

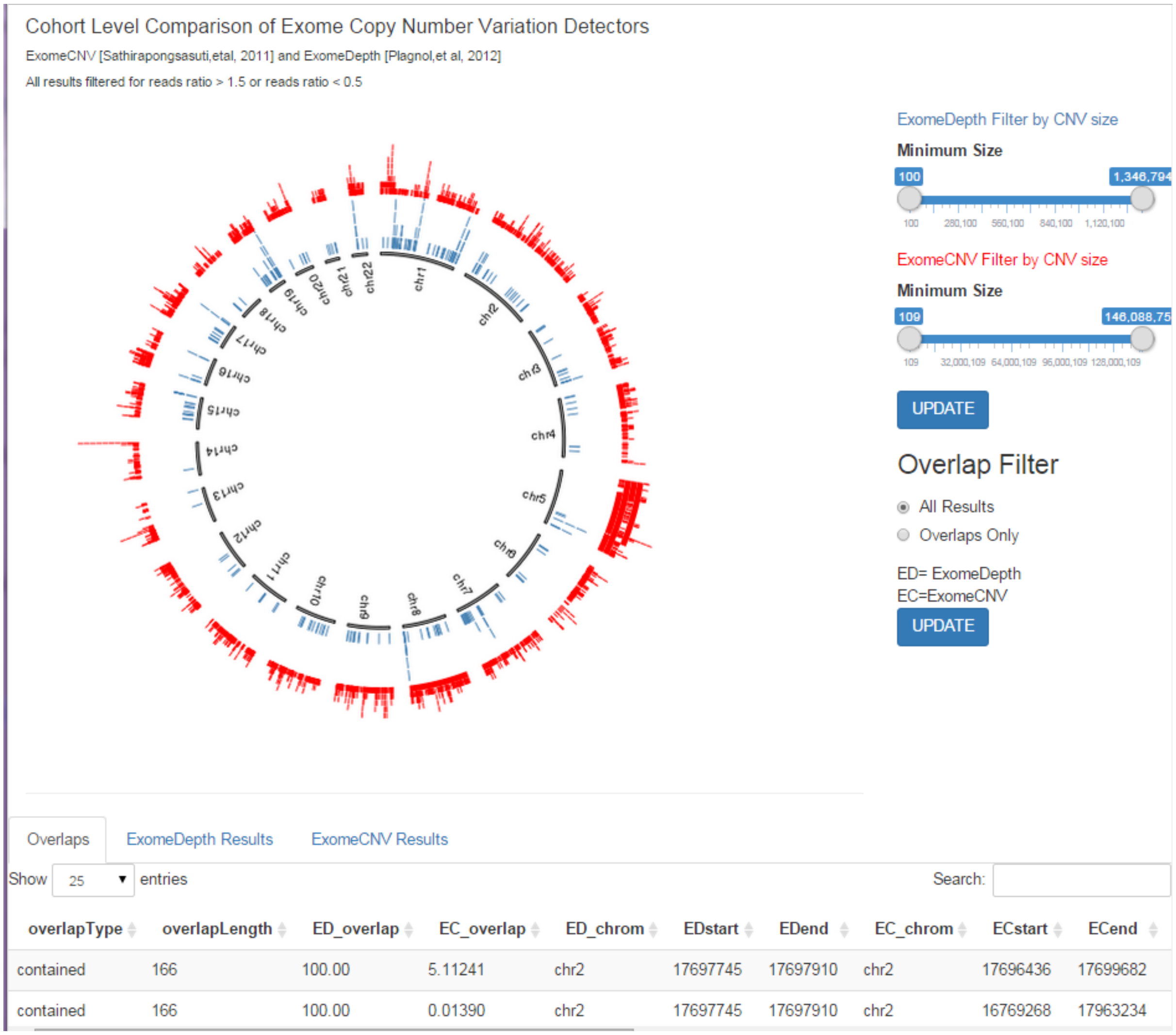
ExomeDepth Tumor1&2

venn overlap if any ranges overlap

ExomeDepth Tumor 1 & 2

	Metric	Tumor1	Tumor2
1	CNV within depth range	84 / 110	88 / 121
2	LOF within depth range	74 / 100	67 / 100
3	GOF within depth range	10 / 10	21 / 21
4	Average size within depth range	49548.0833333333	47698.9886363636
5	Std. deviation within depth range	158028.523998326	154733.263416899
6	Min size within depth range	166	101
7	Median size within depth range	8869	13523.5
8	Max size within depth range	1346795	1341495
9	Average Number of bases overlap	49376.0625	49376.0625
10	Average % overlap	80.4577411368924	87.8933534535789

Figure 16. *Sample comparison observing consistency between tumor samples within a single patient*



Cohort Level Comparison of Exome Copy Number Variation Detectors

ExomeCNV [Sathirapongsasuti,etal, 2011] and ExomeDepth [Plagnol,et al, 2012]

All results filtered for reads ratio > 1.5 or reads ratio < 0.5

ExomeDepth Filter by CNV size

Minimum Size

100 – 1,346,794

ExomeCNV Filter by CNV size

Minimum Size

109 – 146,088,75

Overlap Filter

● All Results

○ Overlaps Only

ED= ExomeDepth
EC=ExomeCNV

Show 25 ▾ entries

Search:

overlapType	overlapLength	ED_overlap	EC_overlap	ED_chrom	EDstart	EDend	EC_chrom	ECstart	ECend
contained	166	100.00	5.11241	chr2	17697745	17697910	chr2	17696436	17699682
contained	166	100.00	0.01390	chr2	17697745	17697910	chr2	16769268	17963234

Figure 17. *Cohort level comparison of CNV Detectors*

Chapter 3 Chromatin Immunoprecipitation (ChIP-seq)

Introduction

Nearly all aspects of genetic cellular activity involve the interaction of proteins with genetic material. The nature of these protein-DNA interactions is extremely diverse and has been shown to be involved in crucial pathways responsible for cellular structure, regulation and function (36). Understanding the associations of these protein-DNA interactions is crucial to understanding the biology that leads to genomic differentiation and disease. ChIP-seq provides methods of associating these regulatory and structural proteins to their genomic elements. The method identifies genome wide profiles of proteins, such as transcription factors, histone modifications, DNA methylation and nucleosome position. These protein profiles modify the composition of DNA and determine protein-DNA interaction sites (36). Due to the diverse range of ChIP-seq tools and its flexibility, ChIP-seq has an abundant amount of variability in its analytical methodologies. The multitude of methods available, over 31 open source analytical programs, adds to the diverse and vastly varied methods of evaluation used in ChIP-seq experiments (23, 37).

CTCF (CCCTC-binding factor) is a highly conserved DNA binding protein (38). Evidence suggests that CTCF may be involved in regulating inter and intra chromosomal interactions, marking loci with a specific chromatin conformation (38, 39).

The following study uses ChIP-seq to identify the CTCF binding sites in one of the four gibbon genera, *Symphalangus syndactylus* (Siamang). Gibbons have accumulated a multitude of large-scale chromosomal rearrangements in comparison with the other hominoids. The abundance of these large rearrangements is the result of an accelerated rate of karyotype evolution (40). The purpose of this project was to explore whether the chromatin conformation marked by the localization of CTCF regions, might predispose regions of breakage in the gibbon genome. This association between the CTCF binding region and their association with human-gibbon synteny breakpoints could help specify the underlying mechanism of chromosomal rearrangements in gibbons and possibly other species.

The particular data set was chosen after having worked with it during a lab rotation, and finding that the two biological replicates of the Siamang genera had higher quality mappings and sequencing in the CTCF sequencing and the input control. The *Hylobates* genus, were obtained from two species *Hylobates moloch* and *Hylobates pileatus*, which may introduce additional variability not ideal for comparison of biological replicates. One of the three replicates sequenced from the *Hoolock* genus had drastically more peak regions identified by MACS14 (41) which was used during in a prior analysis to identify peak regions. This drastic increase was most likely caused by poor quality sequencing or inferior quality of the matched input control. The Siamang samples also had the most consistent signal between the two replicates when analyzed with MACS14. The affinity of these biological replicates was thought to increase the sensitivity when comparing these regions by different analysis methods.

Accurate filtering of binding sites is affected by a multitude of variables, from experimental variation to methods of analysis and detection. Extensive customization can be applied with the experimental design. For example, the methods can be varied depending on the specificity and sensitivity of the antibodies, the size of the protein, the binding sites being identified, and the potential inclusion of controls. Additionally, the methods and functions of ChIP-seq algorithms (i.e. peak callers) which computationally determine the genomic regions of protein binding, can vary significantly.

It would be an exhaustive and unrewarding endeavor to attempt to review all potential peak callers and to assess their performance and accuracy. Comparative assessments of smaller subsets of peak callers have already been performed (37, 42-44). The purpose of the analysis will be to present the results from these computational methods in a novel way that will enable us to truly distinguish the validity and quality of results from analysis of ChIP-seq data. The following analysis will be performed using two computational methods. The presentation of the metrics and distinct conclusions that we can draw with a quantifiable validity will enable us to distinguish the true biologically significant signal from the noise. Moreover, the development of an interactive interface to visualize and filter the results will provide a novel method communicating and transforming the results. The two types of peak callers used to identify CTCF regions in established lymphoblast cell lines from Siamang were, MACS2 (41) and BayesPeak (44).

A ChIP-seq experiment starts with the NA-protein crosslinking of the DNA to the protein from isolated tissue or cells. The DNA-protein complex is then sheared into small fragments through sonication and bound to an antibody specific for the targeted protein. Subsequently, antibody-protein-DNA compound is isolated through immunoprecipitation, often using streptavidincoated magnetic beads. The entire complex is then unbound and the DNA isolated is used to generate a next-generation sequencing library (31). Subsequent mapping and computation uses algorithms and statistical modeling to differentiate the sequencing reads that have accumulated at sites of protein binding. The binding regions are detected as “peaks” regions of the genome where multiple reads align that are indicative of areas where the protein was bound.

Experimental Factors contributing to Analysis

There are multiple potential sources of potential experimental artifacts in ChIP-seq experiments. First, antibody specificity and quality can determine the accuracy of identified binding regions. Second, uneven fragmentation of the DNA caused by open regions of the genome, corresponding to open chromatin, can lead to biased read lengths in some locations (13). Finally, regions of repetitive sequences in the genome can appear enriched for reads because of miss-mapped sequencing reads. These artefacts are often corrected using controls, to nullify regions of ambiguity. An input DNA control uses a portion of DNA sample prior to immunoprecipitation, but otherwise is treated as the DNA obtained from the ChIP-seq experiment. The mock IP DNA control consists of sequenced DNA obtained from immunoprecipitation without the addition of an antibody. DNA from a non-specific IP can also be used as a control in which a common antibody, for example immunoglobin G, is used to isolate DNA that is not binding the protein of interest, identifying indiscriminant binding sites and potential false positives (13).

The size of the binding site region being identified can have a significant factor in peak detection. The region in which enrichment occurs can vary from narrow punctate regions that cover a few hundred base pairs (bp), associated with most transcription factors, to some histone modifications that cover kilo bases. Further variation can be observed with proteins that bind RNA polymerases, where the range of binding regions can vary from narrow to broad regions of accumulated reads or peaks.

Computation: Peak Calling

A multitude of computational tools have been designed to analyze ChIP-seq data and locate protein binding sites from sequencing reads (44). These tools have been developed and optimized using different types of computational algorithms, different proteins, controls and conditions under different assumptions.

Two peak callers were used to analyze the ChIP-seq data in my project: MACS2, a model based analysis of ChIP-seq and BayesPeak. MACS2 uses a model based approach to identify regions of tag enrichment. Tags from ChIP-seq read are usually 20-50 bp sequences that represent the 5' (beginning) of the sheared DNA fragment. These tags exist in both forward and reverse strands resulting in bidirectional enrichment and a bimodal distribution for a single binding site (14). With MACS2, these tags, forward and reverse, are then shifted a half distance toward each other to identify a region of enrichment (45). However, if the binding site consists of a broad region spanning multiple kilo bases, assigning the summit as the average distance between the two distributions is no longer accurate. The symmetrical single peak profile cannot be assumed for broad peaks that can span several hundred kilo bases for RNA polymerases and some histone markers. These binding regions are rarely strongly localized and will consist of complex tag densities that may even contain more than one summit.

BayesPeak, on the other hand, computes the summit based on a priori probabilities, this model and other predictive algorithms are more accurate predictions of the binding sites of these broad regions (44).

After peak shifting, MACS2 identifies peaks modeling the tag distribution along the genome with a Poisson distribution. A local fold enrichment is calculated based on windows of the genome and false discovery rate is determined empirically from the number of peaks in the input control (45). However, over-dispersion of peaks, such as wide regions associated with some histone marks, would not be accurately modeled with a Poisson distribution, which assumes that the standard deviation equals the mean.

BayesPeak uses a hidden Markhov model to identify enriched locations in small (100-300 bp) genomic windows. First, the tag counts are assumed and modeled with the negative binomial statistical distribution. Then Bayesian methods, uses the posterior probability to identify significantly enriched binding regions. The disadvantages of the BayesPeak method, however, are its dependency on the prior probability and the finite window region that is evaluated along the genome without regard as to where the true junction may lie.

Irreproducible Discovery Rate

The comparison of these two computational peak calling methods would be difficult because of the fundamental differences in output of the statistical distribution. MACS2 outputs a significance score based on the FDR and is therefore represented as a p-value or q-value. Alternatively, BayesPeak measures significance as the posterior probability based on the Bayesian hidden Markov model that detected enriched locations, ranging from 0 to 1, with 1 representing enriched regions.

Li et. al. explain a solution to this inability to compare different statistical measurement by converting them to the rank-based method of the irreproducible discovery rate (IDR). The IDR uses biological replicates to compute a significance score based on the reproducibility of the peak. The model assumes that real peaks are reproducible and that noise is irreproducible among replicates. Therefore, each peak is assigned a probability of reproducibility. This is one of many statistical solutions being developed to compare the accuracy of methods with vastly diverse techniques. (10)

New computational methodologies require comparative techniques that can measure the accuracy, reproducibility and performance of these new methods. Care needs to be taken in the selection of the underlying algorithms as well as the parameters to provide the most accurate biological results. A comparative framework for analysis of peak calling algorithms can determine the robustness and appropriateness of a peak calling method. **Table 1** briefly explains metrics and visualizations that can be used to assess the downstream results from multiple peak callers. These metrics and visualizations help guide decisions regarding the validity and credibility of results from various peak callers.

Methods

MACS2

CTCF ChIP-sequencing was performed on lymphoblast cell lines established from two Siamang gibbon, male and female, using the protocol described in Schmidt et al (46). The CTCF-bound DNA was immunoprecipitated using an Anti-CTCF rabbit polyclonal antibody. The resulting DNA library was sequenced with 36 bp reads (Illumina Sequencing). Sequencing reads were aligned to the genome reference (Nleu3.0) using BWA MEM v 0.7.9a-r786. Duplicated sequencing reads were marked using Picard tools version 1.96. Samtools version 1.1 was used to retrieve uniquely mapped reads with a mapping quality greater than 20 and to sort and covert to aligned tagAlign files for peak calling (**Table 9**). Peak calling was performed using MACS version 2 (45), with an estimated gibbon genome size of 2.2 Gbp. MACS2 p-value threshold was set at p =1e-1, instead of the recommended p=1e-7 for analysis using the irreproducible discovery rate. Although no p-value value threshold could have also been used, the IDR can be sensitive to the amounts of signal observed. A comparison using relative equal amounts of peaks between biological duplicates with a significant proportion of true signal provides better results that are either not thresholded with a large proportion of captured noise. If, for example, no significance threshold was applied, the real signal can potentially be drowned out by an infinite amount of noise, rendering identification of a minute amount reproducible true signal difficult (34).

BayesPeak

BayesPeak version 1.20.0 (44) was used as the second peak callers. Bedtools v2.21.0 was used to convert bam alignment files from BWA MEM to bed files. Some of the chromosomes from

Nleu3.0 produced error outputs, hence each chromosome was split and evaluated separately, using default bin sizes of 100 bases.

Results and Discussion

	Sequences	Length	Reads Uniquely Mapped
Gibbon1 CTCF	30362590	36	12068703
Gibbon1 Input Control	27592805	36	14516517
Gibbon2 CTCF	27581495	36	13642550
Gibbon2 Input Control	44911572	36	26031195

Table 9. *General Metrics for Gibbon 1 and Gibbon 2, female and male, ChIP sequencing of CTCF binding regions.*

Seven metrics and visualizations were assessed and implemented (**Table 10**) for ChIP-seq.

These metrics and visualization will help guide decisions regarding the validity and credibility of results from various peak callers.

Visualizations and Metrics for Comparison ChIP-seq Analysis	
FRiP (Fraction of Reads in Peaks)	Significant peaks are biased by coverage (32, 47)
Cross-correlation analysis	Correlation between tags plotted against size of strand shift. Correlation between fragment length and read length can assess the signal to noise ratio. (32)
Number of peaks identified by peak calling method.	The number of regions identified can indicate the sensitivity and specificity of a particular peak calling method
Distribution of peak size or size of enrichment regions identified.	Assessment of the enrichment region size can define an algorithms intended target region and can reveal biases from over fitting.
Venn diagram of Peaks that overlap from different peak calling methods and algorithms.	Comparison of overlapping peaks, with an interactive scalable percentage overlap between 0% to 100% overlap
IDR Correspondence Curve for each analysis methods	Irreproducible discovery rate to determine empirical significance threshold of peaks based on reproducibility of biological replicates (10).
IDR threshold for each analysis methods **Number of peaks called at various IDR**	IDR threshold at which signal increases without reproducibility.

***Table 10.** Metrics and Visualizations for Comparison of the ChIP-seq Peak Calling Methods.*

For comparison of the two computational peak callers, MACS2 and BayesPeak, we compared the FRiP (Fraction of Reads in Peaks) score (**Table 11**), cross correlation peaks (**Figures 18-19 and Table 12**) , and the number of peaks called by each method (**Table 13**). Additionally, we considered distribution of size of enrichment regions (**Figure 20**), and utilized Venn diagrams for examining overlaps (**Figure 21-22**). Lastly, an assessment of the IDR applied to both of these peak calling methods (**Figure 23-25**).

Fraction of Reads in Peaks

The fraction of all reads that fall within an identified peak region (FRiP) can provide a general, first-call metric of the success of the immunoprecipitation and sequencing. The FRiP can vary depending on the targeted binding site, the antibody specificity and the sequencing depth. Some binding sites are rare and a low frequency of true peak enrichment is expected. CTCF, on the other hand, have high number of enrichment sites across the genome and can lead to FRiP scores that exceed expectations (32, 48). The ENCODE consortium guidelines scrutinize any ChIP-seq experiments with a FRiP score lower than 1% (32).

	Gibbon1	**FRiP Gibbon1**	**Gibbon2**	**FRiP Gibbon2**
Total Reads	30362590		27581495	
MACS2	3810192	0.125	3811112	0.138
BayesPeak	3737088	0.123	2152122	0.078

***Table 11.** Fraction of reads in peaks (FRiP) score for Gibbon 1 and 2, female and male, ChIP sequencing results analyzed using two*

computational peak callers MACS2 and BayesPeak.

The FRiP scores from the gibbon range from 7.8% to 13.8%, indicative of the high number of enrichment regions of CTCF. The FRiP is relatively consistent between the peak callers for gibbon 1, with 12.5% and 12.3% for MAC2 and BayesPeak, respectively. However, gibbon 2 shows a distinct variation in FRiP scores, with 13.8% for MACS2 and 7.8% for BayesPeak, indicating that the FRiP is variable and dependent on the computational method. The inverted relationship, with gibbon 1 having a lower FRiP than gibbon 2 for MACS2 but a higher FRiP than gibbon 2 for BayesPeak analysis makes it apparent that FRiP does not measure the relative success of the immunoprecipitation but the overall success versus failure of the experiment as a whole.

Cross-correlation analysis

Strand cross-correlation relies on the fact that reads from ChIP-seq experiments cluster around the locations specific to the enrichment site. The enrichment of the aligned reads will cluster a finite distance, depending on the fragment length, from the center or focal point of the binding site region (49). The clustering can then be analyzed by calculating the Pearson linear correlation between the adjacent clustering strands. Cross-correlation values will, therefore, increase representing the enrichment corresponding with the fragment length and a second location of enrichment, representing the “phantom”/noise peak that corresponds with the read length (32, 48).

The normalized cross-correlation coefficient (NSC), is calculated as the fragment length cross-correlation normalized by the background. The relative strand cross-correlations (RSC) is the cross-correlation at the fragment and the read length. These two ratios provide a method of quantifying the signal-to-noise ratio in the data.

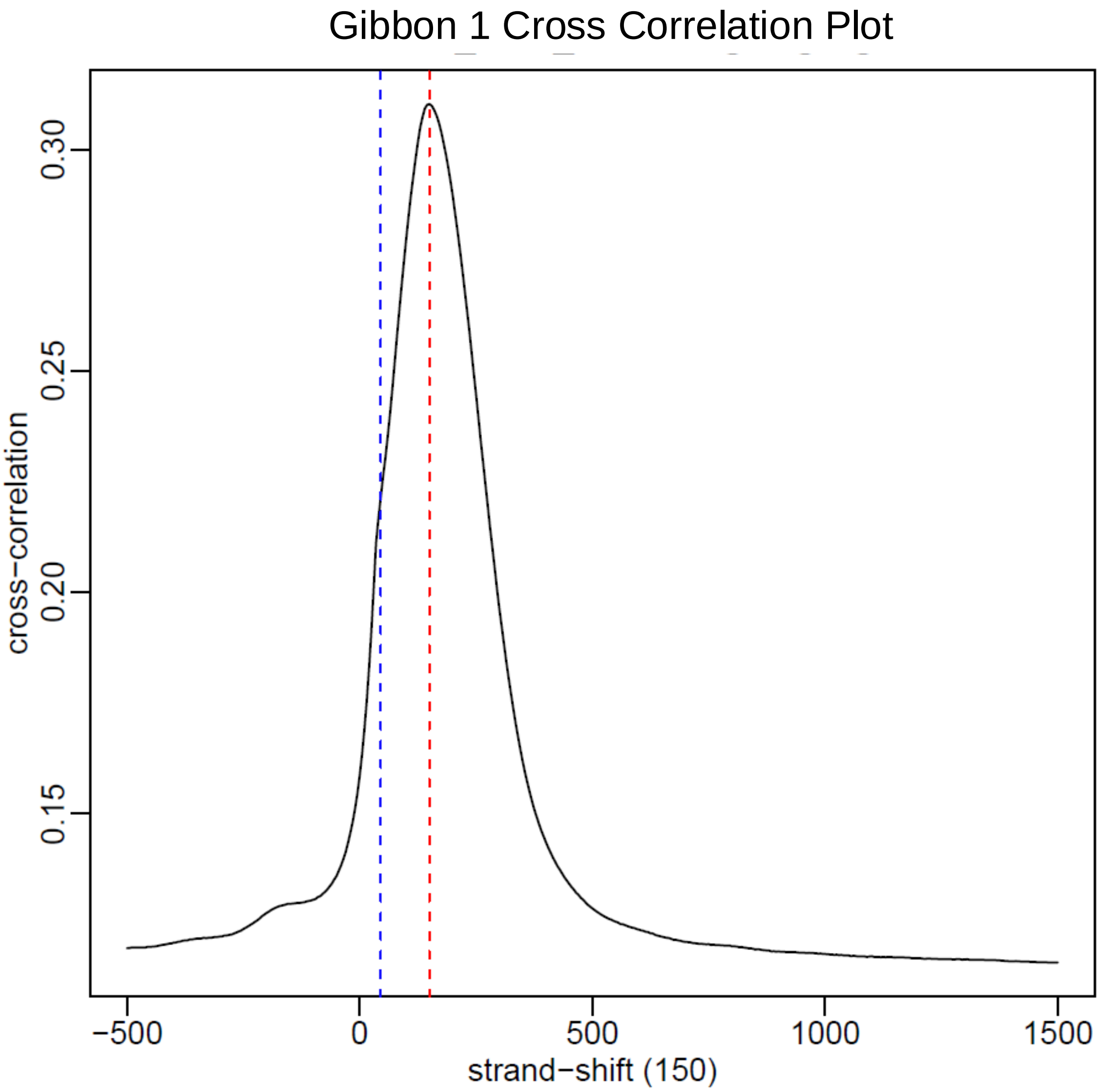


Figure 18. *Cross correlation peaks of lymphoblast cell line from Gibbon Sample 1 ChIP-sequencing of CTCF regions. Two cross-correlation peaks are usually observed in a ChIP experiment, one corresponding to the read length (''phantom'' peak) and one to the average fragment length of the library. X-axis: the strand shift, Y-axis: cross correlation value. Cross correlation analysis estimates of the fragment length is shown with the dashed red line and*

the phantom peak is shown with the blue line.

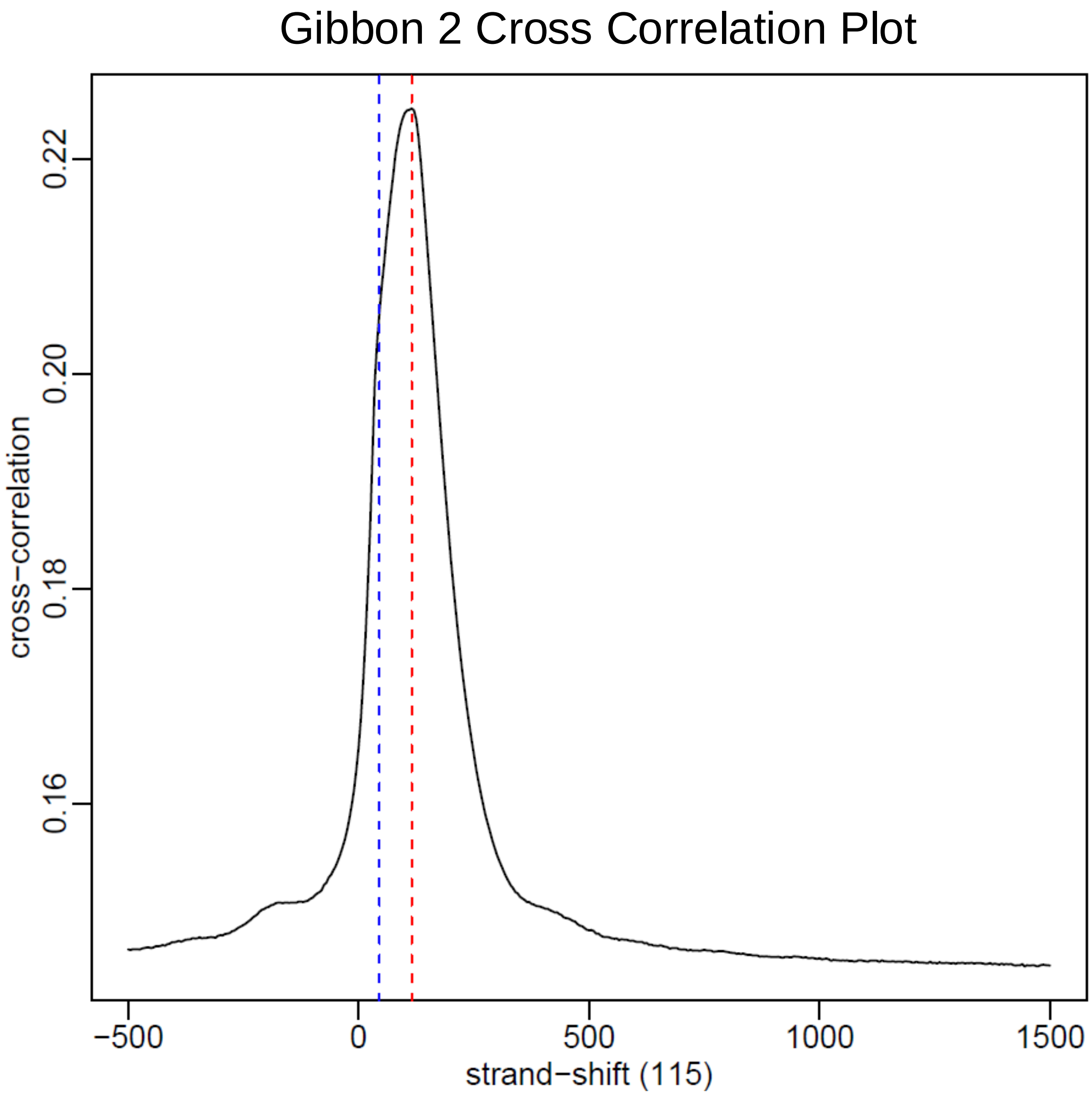


Figure 19. *Cross correlation peaks of lymphoblast cell line from Gibbon Sample 2 ChIP-sequencing of CTCF regions. Two cross-correlation peaks are usually observed in a ChIP experiment, one corresponding to the read length (''phantom'' peak) and one to the average fragment length of the library. X-axis: the strand shift, Y-axis: cross correlation value. Cross correlation analysis estimates of the fragment length is shown with the dashed red line and the phantom peak is shown with the blue line.*

Sample	Estimated Fragment Length	Correlation of Fragment Length	Phantom Peak Length	Correlation of Phantom Peak	NSC	RSC	Quality
Gibbon1	150	0.31	45	0.22	2.67	1.85	Very High
Gibbon1 Input Control	115	0.18	35	0.18	1.11	1.05	Medium
Gibbon2	115	0.22	45	0.21	1.55	1.31	High
Gibbon2 Input Control	95	0.26	35	0.27	1.04	0.73	Medium

Table 12. *Cross-correlation analysis of CTCF ChIP-sequencing of two gibbon lymphoblast cell lines. Estimated fragment length is the length strand cross-correlation with the highest correlation, that correlation is show in the adjacent column. The NSC is the Normalized strand cross-correlation coefficient, NSC=highest cross correlation of the fragment peak/minimum value of cross-correlation. The RSC is the Relative strand cross-correlation coefficient (RSC) = (highest cross correlation of the fragment peak-minimum value of cross-correlation) / (highest cross correlation of the phantom peak/minimum value of cross-correlation). The quality tag based on thresholded RSC (49).*

The standards established by Kharchenko et. al. threshold solely on the RSC, with higher RSC representing higher quality ChIP-seq (**Table 12**) (49). The quality analyzed by this metric is defined as the differentiation between the fragment clustering versus the clustering of the read length from the forward and reverse strands. **Table 12** also includes cross-correlation analysis on the control sequencing. The quality of the normal samples is lower than the CTCF sequencing regions indicating that the estimated fragment peaks have a closer correlation to the read length and the signal consists of a significant amount of random reads. Since this is observed and expected from our input control samples, this observation verifies the quality of the control and differentiates the CTCF sequencing as distinct from the not immunoprecipitated input control background regions.

Landt et. al. suggests that a successful ChIP experiment consists of an NSC of 1.05 and an RSC of 0.8 (32). Both CTCF regions are found to meet this standard, with NSC of 2.67 and 1.85 and RSC of 1.55 and 1.31 for gibbon sample 1 and sample 2, respectively.

Frequency of peaks identified by peak calling method

The basic metric of the number of peaks called by a peak caller can be a simple comparison of the performance of the computational method (Table 13).

	Number of Peaks	
	Gibbon 1	Gibbon 2
MACS2	304497	596976
BayesPeak	159069	186467

***Table 13**. Total signal or number of peaks called for both Gibbon 1 and Gibbon 2 for two peak callers MACS2 and BayesPeak*

This experiment found that the number of peaks called for MACS2 to is much larger than the number of peaks generated from BayesPeak. The increased number of enrichment regions identified by MACS2 may be caused by potential thresholding issues or other differences in the statistical model used for peak detection. The Poisson model assumes that the mean and the standard deviation are equal and is sensitive to over dispersion of reads. False positives increase with reads that have a larger variance than expected from the mean.

Although BayesPeak does not require individual and separate chromosome files, the implementation resulted in output errors from the combined file. Chromosomes that were not found in both ChIP-seq samples with corresponding chromosomes within the input control alignment files would cease with errors. These references from Nleu3.0 were constructs identified as either as mitochondrial references, or unlocalized and unplaced scaffolds. Moreover, the shorter constructs also output errors because of the bin sizes and the state assignment for the hidden Markhov chain Monte Carlo algorithm used by BayesPeak. These constructs were excluded from the peak calling analysis from BayesPeak. Although, no error was observed with MACS2, this may be one source of the discrepancy in the total of called peaks from both models and should be noted.

Distribution of the size of enrichment region called by algorithms

The size of the enrichment region identified can be indicative of specific biases MACS2 identifies slightly smaller regions (Figure 20). This may be because of Poisson distribution observed by MACS, which a much more liberal statistical model than the binomial distribution.

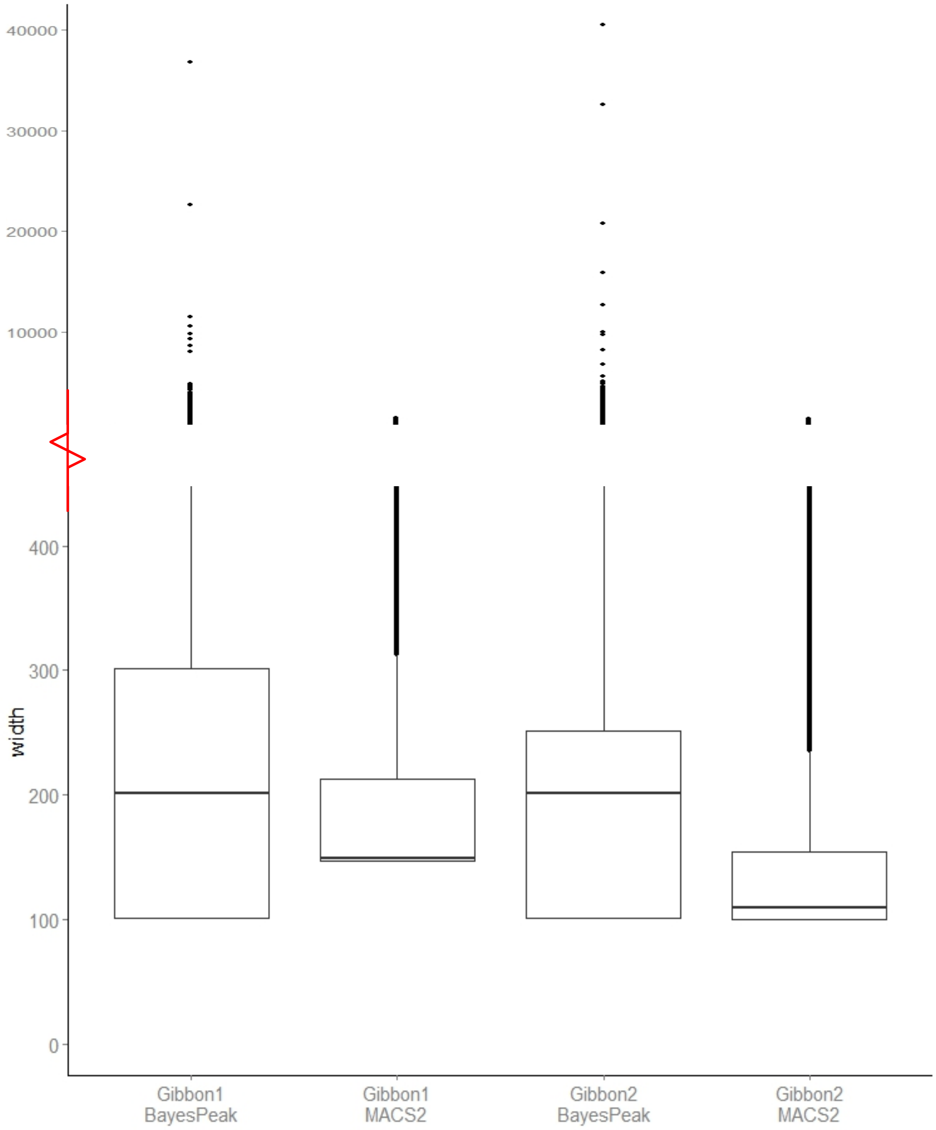
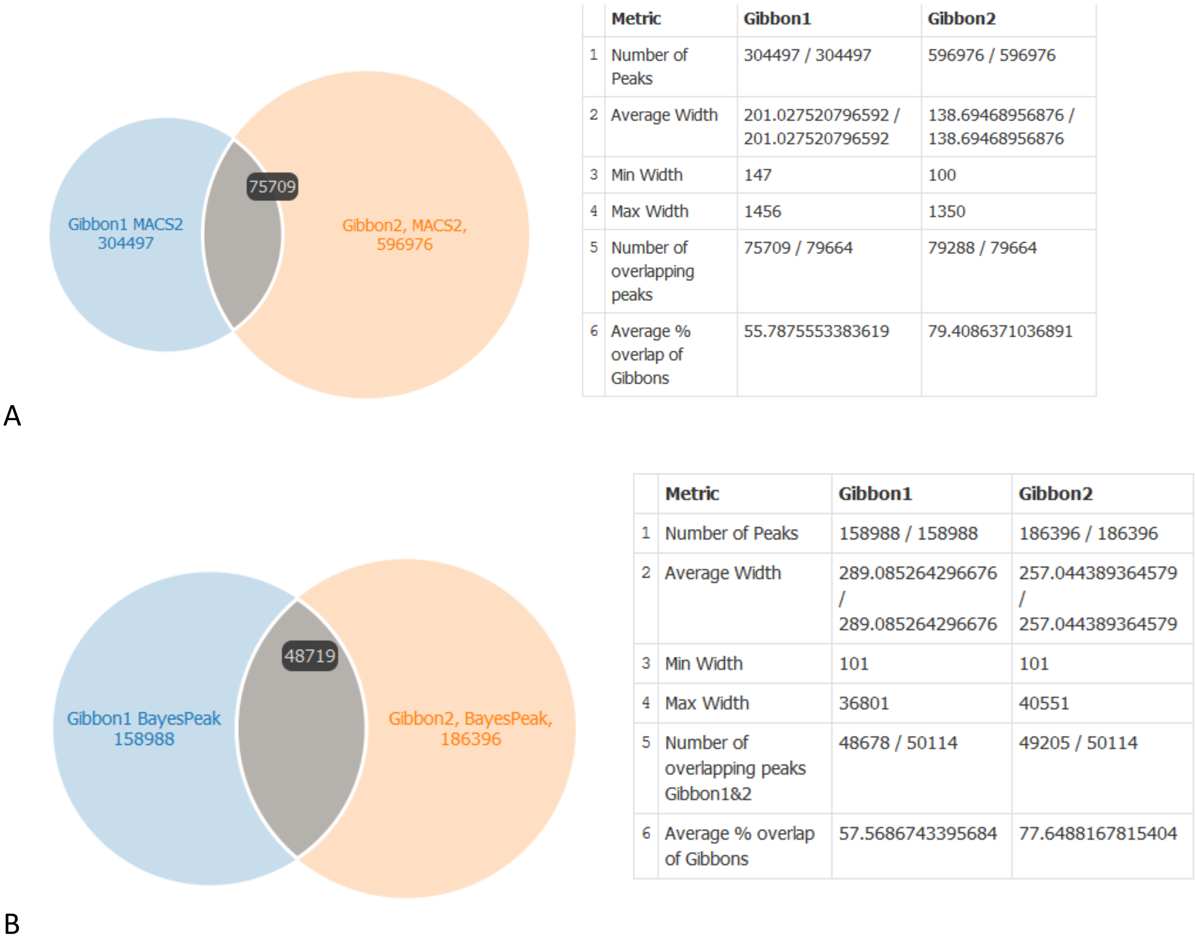


Figure 20. Distribution of CTCF enrichment regions identified for two gibbons lymphoblast cell lines, analyzed using two peak callers BayesPeak and MACS2. Boxplot including peaks with widths from 0 to 400 bases, the truncated signal showing the larger outlier regions identified by both peak callers are also shown. X-axis: Category of gibbon and computational analysis peak calling method used. Y-axis: width or size, in bases, of the regions identified.

Venn diagram of Peaks that overlaps

The interactive framework provides a context for to visualize the various types of overlap between the two samples and the computational methods.



	Metric	Gibbon1	Gibbon2
1	Number of Peaks	304497 / 304497	596976 / 596976
2	Average Width	201.027520796592 / 201.027520796592	138.69468956876 / 138.69468956876
3	Min Width	147	100
4	Max Width	1456	1350
5	Number of overlapping peaks	75709 / 79664	79288 / 79664
6	Average % overlap of Gibbons	55.7875553383619	79.4086371036891

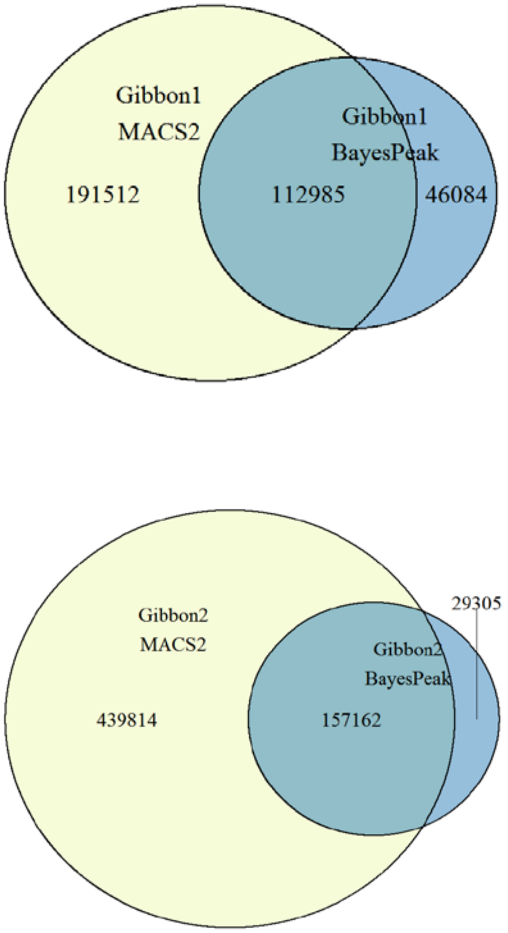
	Metric	Gibbon1	Gibbon2
1	Number of Peaks	158988 / 158988	186396 / 186396
2	Average Width	289.085264296676 / 289.085264296676	257.044389364579 / 257.044389364579
3	Min Width	101	101
4	Max Width	36801	40551
5	Number of overlapping peaks Gibbon1&2	48678 / 50114	49205 / 50114
6	Average % overlap of Gibbons	57.5686743395684	77.6488167815404

***Figure 21**. Overlaps and statistical summaries between ChIP-sequencing of CTCF regions for gibbon lymphoblat cell lines. (A) MACS2 peak calling method for ChIP-seq, overlap between both gibbon samples. (B) BayesPeak peak calling method for ChIP-seq, overlap between both gibbon samples. . Screenshot of R shiny application Sample_Comparison_chipseq, focusing on sample comparisons within a peak detection method.*

The observation of the sample variability within a computation method can be visualized in **Figure 21**. The number overlapping peaks identified with MACS2 is much larger than the overlaps located by BayesPeak. MACS identified 75709 overlapping peak regions versus the 48719 peak regions found by BayesPeak.

Additionally, the widths of the identified regions varied with the computational method, MACS2 calls narrower peak regions, with an average width of 170 bases, while BayesPeak had an average enrichment region of 273 bases. The expected peak size of a CTCF region is approximately 200-300 bp, and BayesPeak is clearly the more accurate when comparing the average size of the region. Bayespeak identifies larger enrichment regions, a potential cause of this discrepancy may be caused by the read shifting that MACS2 performs. MACS2 initially shifts forward and reverse strands of enrichment toward each other, to account for tag enrichment from 5’ and 3’ ends of the reads. Read shifting also helps to find the summit of reads, but may change the approximation of the width of the enrichment region. Bayespeak on the other hand does not rely on read shifting identify enrichment regions. The discrepancy between the sizes of the enrichment regions, although slight, may be slight because of the narrowness in the expected region, which requires less read shifting from MACS2. Therefore, a significant amount error would be expected from a targeted protein with a larger enrichment region, for example in cases where enrichment regions are kilo bases long with multiple summits.

Gibbon sample 1 was found to have a lower proportion of overlaps. MACS had 20% overlap with gibbon 2, while gibbon 2 had 77% of the identified peaks overlap with the gibbon1. This discrepancy in the overlaps indicates a higher proportion of unique reads were identified for gibbon 1. However, the proportion of shared peaks in BayesPeak consists of 23% for gibbon 1 and 21% for gibbon2, indicating a nearly equal amount of shared and unique proportion of peaks identified by for both biological replicates. The differences in these proportions may be caused by MACS2, statistical model, which is sensitive to over dispersion. A potential increased variance of reads observed in the gibbon sample 1, would cause an increase in the number of regions called for that sample.



(A)

(B)

***Figure 22**. Number of total overlaps between enriched regions identified by two different peak calling methods for the same a single lymphoblast gibbon cell line. Consensus exists if there is any overlap between the ranges of the peak region. (A) Gibbon sample 1 cell line comparison of the overlaps between both peak calling methods. (B) Gibbon sample 2 cell line comparison* *of the overlaps between both peak calling methods.*

The percentage overlap from **Figure 22**, shows that there are significantly more regions identified by MACS2 than there are for BayesPeak. BayesPeak has a higher proportion of overlapping peaks between the two methods, with approximately 20% and 84%, of BayesPeak signal also contained in MACS2, for gibbon 1 and 2 respectively. MACS2, on the other hand, has a 6% and 26% overlap with the peak region identified by BayesPeak, indicating an increased sensitivity in the algorithm and statistical model used by MACS2 versus BayesPeak.

IDR threshold for each analysis methods and IDR Correspondence Curves

The irreproducible discovery rate was applied for both computational methodologies of MACS2 and BayesPeak. The correspondence between the biological replicates, gibbon 1 and gibbon 2 can be obtained from the correspondence curves in **Figure 23A and 24A**. The junction at which the slope of this correspondence curves converts from a line with a slope of 1 to a x^2 line represents the ranked proportion at which signal no longer becomes reproducible. This transition can be easier to extract from the derivative of the correspondence plot (**Figure 23B and 24B**).

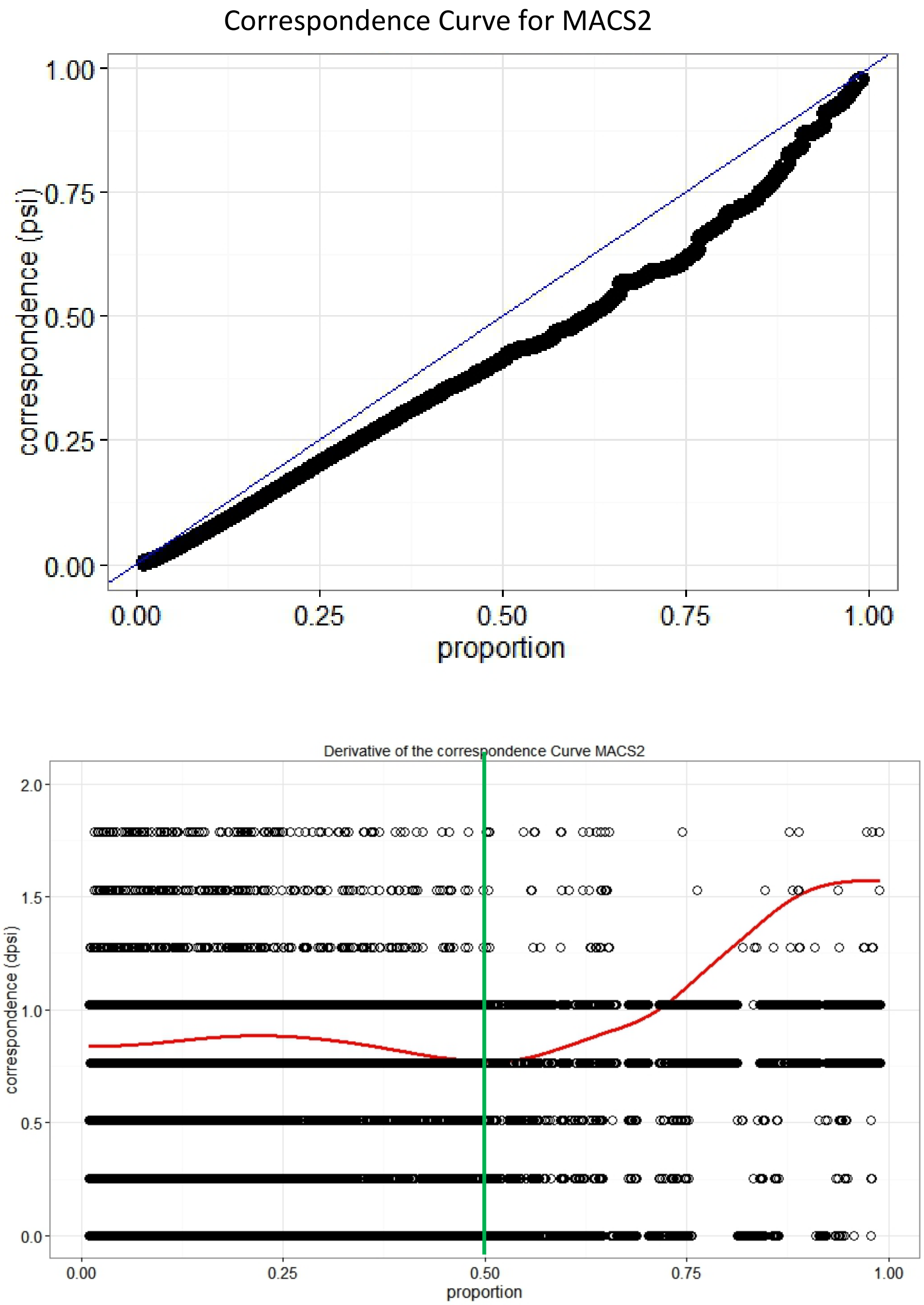


Figure 23. *Correspondence curves for the ChIP-seq peak caller, MACS2. Correspondence measured between gibbon 1 and gibbon2 . (A) The correspondence curves of the IDR versus the ranked proportion. (B) The derivative of the correspondence cure. The green line indicates the threshold for the proportion of signal that was reproducible to calculate the IDR.*

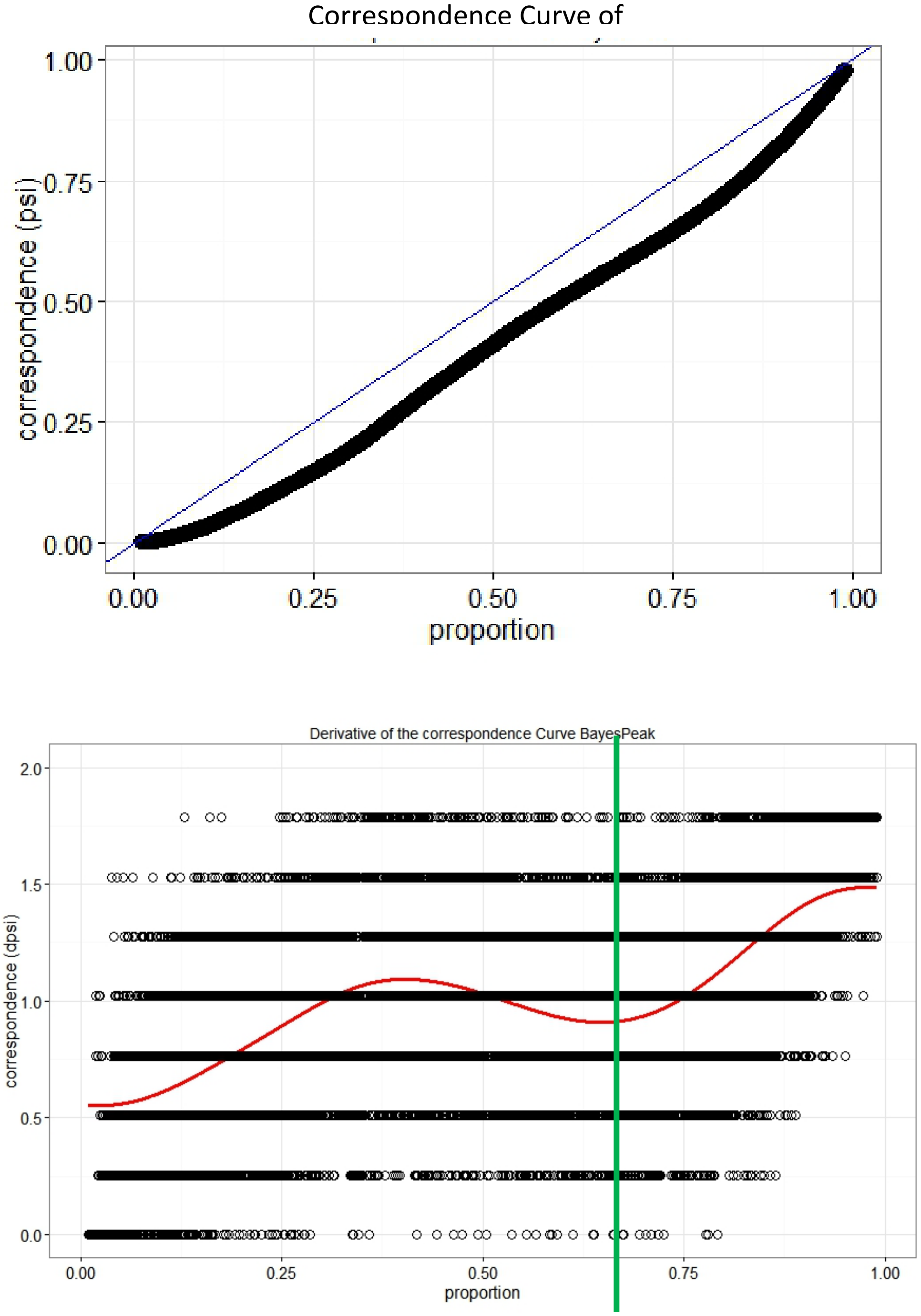


***Figure 24**. Correspondence curves for the ChIP-seq peak caller, BayesPeak. Correspondence CTCF binding regions identified from ChIP-sequencing of two gibbon lymphoblast cell lines. (A) The correspondence curves of the ranked proportion of the called enrichment regions significance score to the measured correspondence between the samples. The blue line indicates perfect correspondence. X-axis: The proportional ranked list of peaks. Y-axis: a measure of correspondence. (B) The derivative of the correspondence curve representing*

the change of correspondence along the decreasing order of significance. X-axis: The proportional ranked list of peaks. Y-axis: a measure of the change in correspondence.

The proportion that is reproducible was 0.5 for the MACS2 peaks and 0.6 for the BayesPeak peaks. The higher transition point for MACS2 indicates that it has a higher reproducibility with a later occurrence of the transition from reproducible signal to non-reproducible signal.

The rank proportion of the total signal was used to determine the IDR for each observation by applying the IDR's statistical model and estimating the posterior probability (10). The model was fit with the reproducible proportion and 20 iterations for both peak callers. The iterations were optimized based on the proximity of the fitted model's parameters, i.e. mean, standard deviation, correlation, to the calculated parameters of the reproducible component.

The IDR for each observation provides a measure of the irreproducibility of that component. The IDR assumes that reproducible output consists of the true signal while non-reproducible consists only of noise a threshold of the IDR would represent would ideally represent that transition (10). Therefore, the amount of signal captured at a particular IDR threshold can be indicative of the sensitivity of a particular peak caller. **Figure 25** highlights that MACS2 produces more reproducible signal for all IDR thresholds.

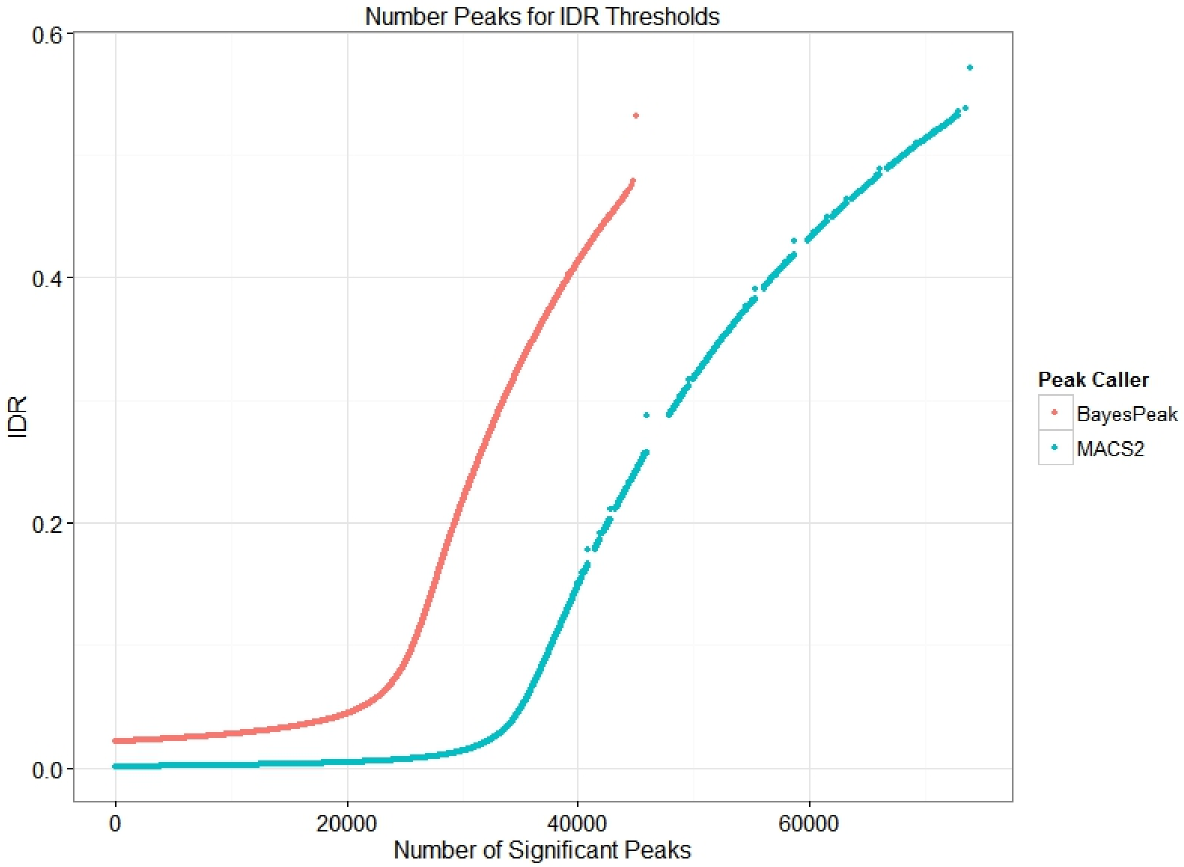
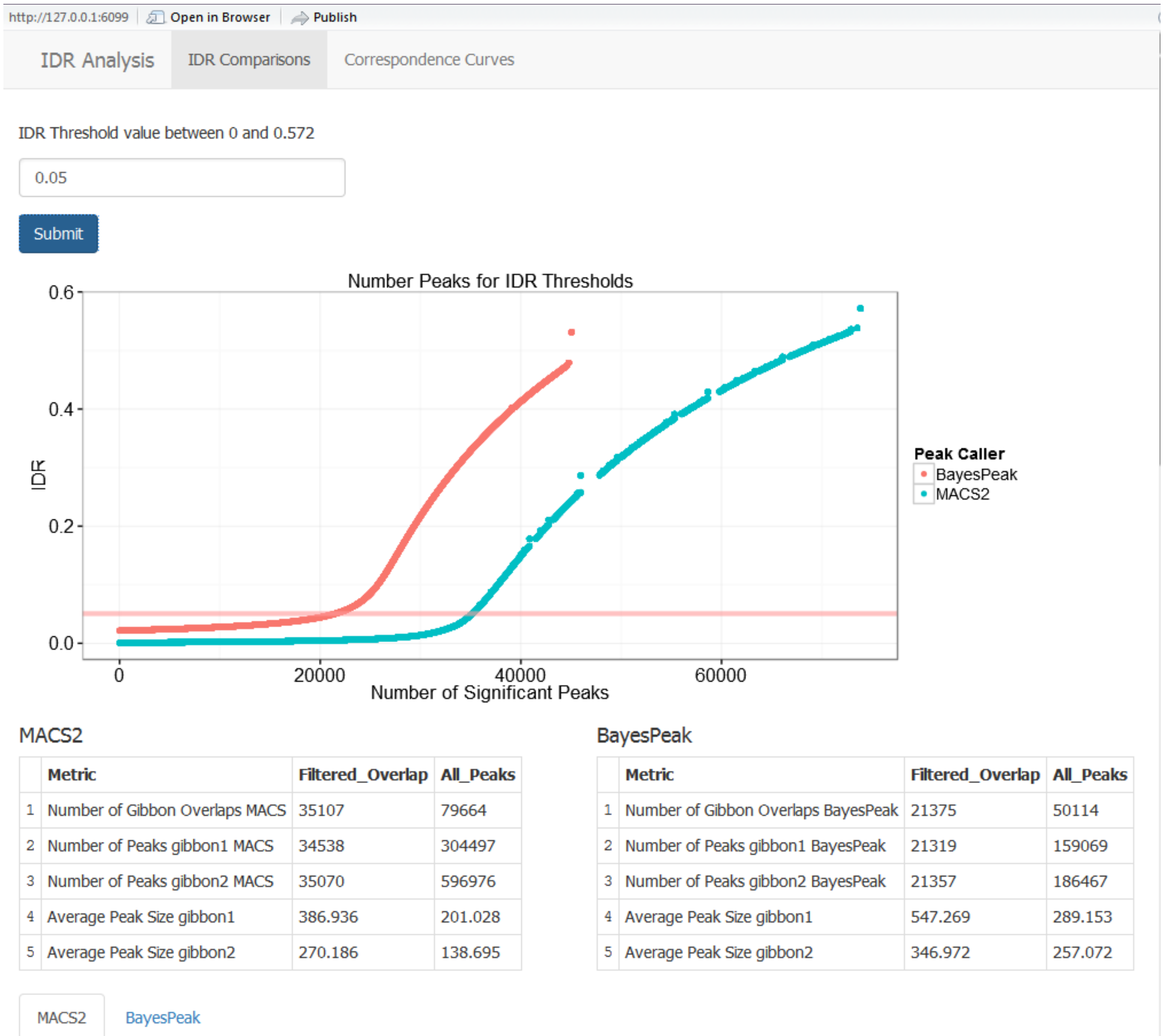


***Figure 25**. Irreproducible discovery rate (IDR) at different number of included peaks, plotted at various IDR thresholds for BayesPeak and MACS2 peak callers analyzed for CTCF ChIP-seq from two gibbon lymphoblast cell lines. Peaks are selected using the IDR for each observation. X-axis: The rank list of peaks, ranked by the IDR, Y -axis: Irreproducible discovery rate (IDR).*

The interactive framework developed for the IDR filters the peak observation of the reproducible components by their corresponding calculated IDR (**Figure 26**). The framework allows users to select the IDR threshold by observation and peaks identified. The potential to filter based on the maximum of overlap of the methods could also be a useful component, reinforcing the identification of any known enrichment region with confidence in the remaining overlapping regions.



IDR Threshold value between 0 and 0.572

0.05

MACS2

	Metric	Filtered_Overlap	All_Peaks
1	Number of Gibbon Overlaps MACS	35107	79664
2	Number of Peaks gibbon1 MACS	34538	304497
3	Number of Peaks gibbon2 MACS	35070	596976
4	Average Peak Size gibbon1	386.936	201.028
5	Average Peak Size gibbon2	270.186	138.695

BayesPeak

	Metric	Filtered_Overlap	All_Peaks
1	Number of Gibbon Overlaps BayesPeak	21375	50114
2	Number of Peaks gibbon1 BayesPeak	21319	159069
3	Number of Peaks gibbon2 BayesPeak	21357	186467
4	Average Peak Size gibbon1	547.269	289.153
5	Average Peak Size gibbon2	346.972	257.072

Figure 26. *Screenshot of R shiny application IDR_chipseq, focusing on peak list filtering by IDR threshold.*

The multitude of factors involved in ChIP-seq experiments, from the experimental parameters to the variety of peak calling methods, highlights the need for a these types of metrics to allow researchers to identify high confidence peaks and understand how different factors influence each method. It is critical to note that this type of consensus approach and metrics would not be a replacement for experimental follow-up and validation.

Chapter 4 Final Conclusions

The evolution of the metrics of metrics development and context for the visualization illustrated how a “hands-on” interactive interface can change the way we interact with the data and think about it, facilitating discovery and transfer knowledge. This can be clearly seen in the CNV example where three different context specific needs were identified after initial testing of the prototype.

The development of the tools and methods to increase data transparency can help increase scientific reproducibility. This interactive prototype framework is an important first step in providing a measurable and consistent check of data validity at the final step in analysis. These metrics and guidelines would allow scientist to check the validity of their methods and ensure the accuracy of the data. It is important to emphasize that the interactive visualizations allowed identification of issues or characteristics of the data that would have been missed in traditional, static plots. The type of framework provide efficient, robust and transparent access to the data, allowing an investigator to interact with and query the data. This is critical for quality control as well as considerations for rigor and reproducibility. It also emphasizes the need for a greater focus on human-data interaction.

References

1. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data & Society. 2014 SAGE Publications;1(1).

2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011 print;10(9):712-.

3. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012 03/29;483(7391):531-3.

4. Unreliable research: Trouble at the lab. The Economist. 2013 Oct 19, 2013.

5. Hawkins RD, Hon GC, Ren B. Next-Generation Genomics: an Integrative Approach. Nature Reviews.Genetics. 2010 07;11(7):476-86.

6. Kitchin R. Big data and human geography: Opportunities, challenges and risks. Dialogues in Human Geography. 2013 November 01;3(3):262-7.

7. FastQC v0.11.3 [Computer Software] [Internet].; 2015 [updated March 25, 2015;]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

8. Heer J. Data Visualizaton From Data to Discovery: Art Center + Caltech + JPL. Interactive Data Analysis; May 23, 2013; Caltech, Pasadena, CA, USA: Jeffrey Heer; 2013.

9. Heer J, Shneiderman B. Interactive Dynamics for Visual Analysis. Queue. 2012 feb;10(2):30:30,30:55.

10. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics. 2011;5(3):1752 <last_page> 1779.

11. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012 08/21;28(21):2747-54.

12. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011 Oct 1;27(19):2648-54.

13. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009 Oct;10(10):669-80.

14. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat Meth. 2009 print;6(11):S22-32.

15. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R. 2015;R package version 0.12.2:http://CRAN.R-project.org/package=shiny.

16. Collas P, editor. Chromatin Immunoprecipitation Assays : Methods and Protocols. Humana Press; 2009.

17. Wilbanks EG, Facciotti MT. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PLoS ONE. 2010 07/08;5(7):e11471.

18. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012 Apr 11;485(7398):376-80.

19. Merkenschlager M, Odom DT. CTCF and cohesin: linking gene regulatory elements with their targets. Cell. 2013 Mar 14;152(6):1285-97.

20. Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014 09/11;513(7517):195-201.

21. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS). Genome Biol. 2008;9:R137.

22. Laajala T, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo L. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC Genomics. 2009;10(1):618.

23. Rye MB, Saetrom P, Drablos F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Res. 2011 Mar;39(4):e25.

24. Spyrou C, Stark R, Lynch A, Tavare S. BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics. 2009;10(1):299.

25. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M. Modeling ChIP sequencing in silico with applications. PLoS Comput Biol. 2008;4:e1000158.

26. Schmidt D, Schwalie P, Wilson M, Ballester B, Gonçalves Â, Kutter C, et al. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell. 2012 1/20;148(1–2):335-48.

27. (2012) ENCODE: TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework [Internet].; 2013 []. Available from: https://sites.google.com/site/anshulkundaje/projects/idr.

28. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 2008 Nov;26(11):1293-300.

29. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31.

30. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol. 2013 11/14;9(11):e1003326.

31. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008;26:1351-9.

32. Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genetics 2014/12;206(12):432-40.

33. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. Human Mutation. 2014 07/01;35(7):899-907.

34. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14 Suppl 11:S1,2105-14-S11-S1. Epub 2013 Sep 13.

35. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget. 2013 11/14;4(11):1868-81.

36. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 04/28;12(4):R41-.

37. Krumm N, Sudmant PH, Ko A, O'Roak B,J., Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012 05/11;22(8):1525-32.

38. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc.; 2002.

39. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. PLoS ONE. 2011 01/31;6(1):e16327.

40. Nilsson B, Bümming P, Meis-Kindblom JM, Odén A, Dortok A, Gustavsson B, et al. Gastrointestinal stromal tumors: The incidence, prevalence, clinical course, and prognostication in the preimatinib mesylate era. Cancer. 2005;103(4):821-9.

41. Corless CL, Barnett CM, Heinrich MC. Gastrointestinal stromal tumours: origin and molecular oncology. Nat Rev Cancer. 2011 print;11(12):865-78.

42. Hirota S, Isozaki K, Moriyama Y, Hashimoto K, Nishida T, Ishiguro S, et al. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. Science. 1998 Jan 23;279(5350):577-80.

43. Dematteo RP, Heinrich MC, El-Rifai W, Demetri G. Clinical management of gastrointestinal stromal tumors: Before and after STI-571. Hum Pathol. 2002 /5;33(5):466-77.

44. Duensing A, Medeiros F, McConarty B, Joseph NE, Panigrahy D, Singer S, et al. Mechanisms of oncogenic KIT signal transduction in primary gastrointestinal stromal tumors (GISTs). Oncogene. 2004 03/08;23(22):3999-4006.

45. Janeway KA, Albritton KH, Van Den Abbeele AD, D'Amato GZ, Pedrazzoli P, Siena S, et al. Sunitinib treatment in pediatric patients with advanced GIST following failure of imatinib. Pediatr Blood Cancer. 2009 Jul;52(7):767-71.

46. Antonescu CR, Besmer P, Guo T, Arkun K, Hom G, Koryotowski B, et al. Acquired resistance to imatinib in gastrointestinal stromal tumor occurs through secondary gene mutation. Clin Cancer Res. 2005 Jun 1;11(11):4182-90.

47. Liegl B, Kepten I, Le C, Zhu M, Demetri G, Heinrich M, et al. Heterogeneity of kinase inhibitor resistance mechanisms in GIST. J Pathol. 2008;216(1):64-74.

48. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015 print;16(3):172-83.

49. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 2011 11/08;12(11):R112-.