Batch Effect Detection and Network Analysis in BCR/ABL-independent CML Imatinib Resistance

By

Adam Therneau

A Thesis

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

Master of Science
August 2017

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Adam T. Therneau

"Batch Effect Detection and Network Analysis in BCR/ABL-independent CML Imatinib Resistance"

Has been approved

_____

Thesis Advisor - Eilis Boudreau PhD/MD

_____

Committee Member - Shannon McWeeney PhD

_____

Committee Member - Cristina Tognon PhD

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my entire Thesis Advisory Committee, for their insightful feedback and invaluable help improving the quality of my work on this project. Specifically, I would like to thank my Thesis Advisor Dr. Eilis Boudreau for her generous guidance and help in seeing this project to its conclusion; Dr. Cristina Tognon for her generous editorial feedback and for sharing her extensive knowledge of cancer biology; and Dr. Shannon McWeeney for her critical feedback on the presentation of this work, which was invaluable for clarifying my thinking on several key aspects of this project. I would also like to thank Beth Wilmot for all of her contributions to this project, which were critical in both the study design and execution of my thesis.

Thanks as well to Dr. Guanming Wu for several very insightful and honest conversations discussing the HotNet2 algorithm and network analysis.

I would like to extend a heartfelt thanks to Dr. Jackie Wirz and Dr. Steven Bedrick for sitting in as informal members of my committee and for all of their generous feedback on this work.

A special thanks to my father for lending me his wealth of experience and knowledge in biostatistics and for being a patient sounding board for many of my ideas, both good and bad. Most of all I would like to thank my wife, Ildiko, for her unshakeable strength and support through the most challenging periods of the graduate school process, and our two beautiful girls, Cora and Remi, who helped remind me every day of the horizon I was aiming for. Lastly I would like to thank my mother, who taught me the golden rule and how to face the world with self-respect and courage.

1 Introduction

1.1 Background

Chronic myeloid leukemia (CML) is a hematological malignancy characterized by unregulated growth of myeloid cells in the bone marrow. The chronic phase of the disease is typified by an overproduction of haematopoietic stem cells (HSCs) which further develop into the various myeloid progenitor cell types and eventually lead to an excess accumulation of normally functioning and replicating myeloid cells in the blood stream. Progression from chronic phase to blast crisis leads to a rapid accumulation of primitive myeloid and lymphoid blast cells in the blood stream. The most prominent feature of CML cases is the famous "Philadelphia chromosome" translocation between chromosomes 9 and 22, present in nearly all cases and the most well known example in medicine of a genetic abnormality linked to disease[1–3]. This translocation results in the BCR/ABL fusion oncogene, which expresses a constitutively active tyrosine kinase. The aberrant tyrosine kinase resulting from this fusion has proven to be the primary driver of leukemogenesis in CML, acting through the activation of numerous pathways promoting cellular proliferation.

A major advance in the treatment of CML, and one of the first examples of direct molecular targeting for treatment of cancers, was achieved with the development of Imatinib mesylate, which inactivates the BCR/ABL fusion protein, inhibiting its spurious tyrosine phosphorylation. Imatinib has been uniquely successful in halting CML amongst patients in the chronic phase of the disease, with a 5 year overall survival rate of 85%[4]. Prior to the development of targeted BCR/ABL inhibition as a treatment approach for CML, the best available treatment options were either stem cell transplantation or drug intervention with interferon-α, which offered a highly variable extension in survival times and significant side effects. Despite this success a significant percentage of chronic phase patients acquire resistance to imatinib treatment over time, inevitably transitioning to more advanced stages with poor treatment options. The leading known cause of acquired imatinib resistance

is the acquisition of point mutations in the kinase domain(KD) of BCR/ABL, changing the steric conformation of the binding site for imatinib and thus disrupting its ability to inhibit phosphorylation[5]. Kinase domain mutations account for roughly 50-60% of resistance cases[6], and this avenue of aquisition for Imatinib resistance is generally classified as BCR/ABL-dependent. Second-generation tyrosine kinase inhibitors(TKIs) have been developed which are able to effectively block BCR/ABL activity even in the presence of select KD point mutations, as in the case of the E255V P-loop mutation salvaged by Dasatinib and the higly resistant T315I ATP-binding site mutation, which is effectively treated with Ponatinib[7]. Despite the fact that each of these second generation TKIs are susceptible to their own specific resistant point mutations, they offer a reasonable salvage therapy for patients exhibiting BCR/ABL-dependent resistance to imatinib.

No comparable second-line therapeutic options are available for the remaining half of imatinib resistant cases. For these "BCR/ABL-independent" cases clear molecular mechanisms have remained elusive. Amplification of the BCR/ABL gene, which would increase the necessary dosage of imatinib required for effective treatment, has been observed in select cases[8]. Mutations or amplifications leading to over or under expression of transmembrane transporter proteins such as ABCB1 and HOCT1 have been observed at high rates in non-BCR/ABL resistant cases. These may indicate mechanisms reducing the bioavailability of imatinib[9,10]. While the preceding examples are not due to disruption of imatinib with the BCR/ABL fusion protein, they could still be described as BCR/ABL-dependent because they act to reduce the dosing of the drug and interrupt effective targeting of BCR/ABL.

There are more cases in which imatinib (or other TKIs) binding efficiency to the BCR/ABL fusion protein is not inhibited and the genetic causes are not understood. Recent evidence has shown that these BCR/ABL-independent mechanisms of resistance may act through alternative activation of the same terminal downstream pathways seen in canonical BCR/ABL mediated CML progression, even as BCR/ABL expression is being effectively neutralized. One such study has identified putative imatinib "sensitizing genes", which when underexpressed lead to upregulation of PRKCH and activation of the c-RAF kinase, which

phosphorylates RAF and recapitulates the RAF/MEK/ERK pathway activation seen in CML cases prior to imatinib treatment[11].

There are numerous downstream pathways, some already known to play a central role in CML progression, which may be acting as mediators of CML relapse in BCR/ABL-independent TKI resistance. One of the most important canonical CML pathways is the aforementioned Ras/RAF/MEK/ERK cascade, which transmits cell-surface signalling from various growth factors (i.e. EGFR and FGF), to the nucleus of the cell for transcriptional regulation. Another key pathway involved in CML progression is the JAK/STAT receptor signalling cascade, which regulates transcription in response to cell-surface signalling[12]. There are also cell-intrinsic signalling pathways, such as the PI3K/AKT pathway, which play an important role in CML progression[13]. AKT has numerous downstream targets involved in CML and other cancers, such as mTOR, FOX0, and CREB. Many other cancer related signalling pathways have been implicated in CML progression, or in other related myeloid malignancies, and may also be involved in BCR/ABL-independent resistance. For example, Wnt/β-catenin signaling has been shown to be important for the growth of CML primary cells in in-vitro experiments[14].

The preceding examples are just some of the cancer associated signaling pathways which may act to re-trigger leukemia progression in BCR/ABL-independent cases. It is possible that many BCR/ABL-independent pathways are also active in BCR/ABL-dependent cases. Therefore, the identification of pathways involved in BCR/ABL-independent resistance is crucial for understanding the molecular mechanisms of CML relapse in all cases of TKI resistance. This knowledge will not only offer actionable targets for salvage treatment in Independent cases, but may also indicate possible compound therapies for more permanent treatment of dependent resistance cases exhibiting BCR/ABL point mutations.

1.2 Research Question

Due to the complexity of BCR/ABL-independent TKI resistance, a central question is whether relapse acts through the canonical BCR/ABL CML pathways seen in dependent cases, or through a different set of genes and pathways. To answer this question, both RNA-seq and Whole Exome Sequencing data was collected and analyzed as part of a large retrospective study of CML patients with acquired resistance to imatinib and other related TKI inhibitors. RNA-seq and Whole Exome sequencing are both next generation sequencing approaches which allow for the high-throughput, cost-effective characterization of an individual's entire mRNA transcript profile (RNAseq) or protein coding region of the genome (WES)[15]. This study was comprised of 244 samples (123 WES, 121 RNAseq) collected from seven different clinical locations across the United States and Europe. These samples represented 130 unique patients, 70 of which had samples collected for both data types, 36 for WES only, and 24 for RNAseq only.

This type of study, with data pooled from many different centers, has significant limitations. Due to variation in laboratory protocols, highly heterogenous clinical covariates (TKIs used, duration of treament, etc), and the inability to structure sample processing from the very beginning of the experimental pipeline in order to avoid aquisition of batch effects, the resulting dataset contain significant technical variation. As a result careful consideration must be taken to ensure that problematic samples are identified and if possible properly normalized.

As is often the case with studies based on data collected in clinical settings, the vast majority of the samples used for sequencing lacked matched normal germline samples. This creates a serious issue of scale in identified mutations and candidate genes that will be used as input for downstream analysis. Therefore, a strategy for filtering the variants/genes down to a more reasonable scale, based on a rational set of criteria, must be developed.

Aim1: Devise strategy for dealing with data shortcomings (retrospective/uneven protocols,

unmatched samples); identifying potential problem samples(or areas of bias) and developing a framework for filtering of mutations.

Aim2: Identify putative driver genes and functionally linked networks of genes important to BCR/ABL-Independent TKI resistance in CML using the HotNet2 network analysis algorithm.

2 Batch Effect Detection

2.1 Introduction

In recent years, issues with poor reproducibility have risen to the forefront of consciousness amongst scientists, with the results of numerous studies being called into question after re-analysis or attempted replication[16]. This has been of particular concern for biomedical scientists utilizing computational approaches and high-throughput datasets. In many cases these problematic studies can be attributed to mistakes in data-processing and/or failure to identify systemic effects introduced by experimental artifacts, or so-called "batch effects", unrelated to any biological variables of interest.

Even small errors in data processing, such as mislabeled sample tubes or 1-off excel table data entry errors, can have striking effects on downstream analyses and lead to mistakes not just in direction of future research but in treatment of patients within clinical trials, as in the famous Duke University Anil Potti example[17]. In that case, a column assignment error led to the mislabeling of several samples and eventually to the assignment of patients to incorrect arms of a clinical cancer trial, an example of an exceedingly simple error with very serious outcomes.

In the case of "batch effects", it has been known for some time that high-throughput sequencing based expression assays such as microarrarys are highly sensitive to a wide array of common experimental and extra-experimental factors; from the batch of reagents used, to the manufacturing lot of the chips, to the specific technician involved, to even atmospheric conditions on the day the hybridization was performed [18]. Though perhaps diminished, this same sensitivity has been demonstrated in the case of next-generation sequencing technologies such as RNAseq. These sorts of batch effects are often inadvertantly introduced due to a lack of careful planning around the timing of data collection or experimental procedures. With NGS assays requiring multiple steps in the sample preparation pipeline, from collection

of primary tissue samples to RNA extraction to sequencing library preparation, it is easy for one or multiple of these steps to be done on different days, potentially with different reagents and by a different technician, for different groups of samples. This can easily lead to the creation of unanticipated batch variation.

While many of these batch effects can be prevented by careful planning, in retrospective studies such as ours which attempt to pool samples from a large number of collaborating labs, certain steps in the experimental pipeline may be beyond our control to properly structure. Furthermore, data pertaining to these factors may not be uniformly available or even recorded in all cases. Even in the best of cases, the potential batch introducing factors are not always known in advance. For this reason, many of the tools for identifying batch effects are exploratory data visualizations, which allow for open-ended discovery of aggregate patterns in the data, unexpected under normal assumptions. One such method commonly used for this purpose is Principal Component Analysis (PCA), a dimensional reduction technique which can be used to identify the subset of genes or other features exhibiting the greatest variance across a complex multi-dimensional dataset. PCA has been used extensively as a statistical tool in microarray and NGS gene expression studies to identify and normalize for sources of noise[19], and in more recent years has been shown to have great utility as a visual tool for identification of strong batch effects which segregate samples when evaluated on the first several prinicipal components[20,21]. Many other visualization assays for NGS data have been developed for this sort of exploratory data analysis(EDA).

Though normalization techniques have become standard practice when dealing with NGS data, they have limited success in correcting for the systemic bias introduced by batch effects[22]. Since standard methods of normalization are typically inadequate in correcting for batch effects, more successful approaches such as ComBat[23] or Frozen SVA[24] explicitly incorporate any identified batch variables into their models in order to reduce the effect of the batch factor. Many of these approaches are dependent on the identification of batch variables prior to formal analysis. Additionally, the success of more sophisticated methods of batch correction are limited by the degree of statistical confounding between the batch variables

and the primary experimental variables of interest. One of the most common causes of batch effects is the tendency to collect experimental samples before controls, leading to correlation between changes in experimental components (new reagents, sequencing batches[25], etc) and experimental class. A high degree of confounding of this sort creates a scenario in which it is nearly impossible to distinguish differences between the groups of interest from differences introduced by the batch effect. For this reason, it is imperative to identify if any batch effects are present and if so determine the degree of confounding present in order to assess if the batch effects can be modeled appropriately.

2.2 Results and Discussion

IBS Clustering Plots

Identity-by-state (IBS) clustering[26] was performed on genotyping results from the Genome Analysis Toolkit (GATK)[27] for all the samples in our study cohort. IBS analysis calculates the degree of genetic similarity between all the possible pairings of samples, in this case by comparing all of the variant loci included in the genotyping output. Heirarchical clustering can then be performed on the IBS scores in order to infer the genetic similarity of the different samples. This is useful in checking for samples exhibiting similarity that would be expected, such as serial samples from the same patient, or for identifying unexpectedly similar sample pairs arising from other sources, such as mislabeled samples or batch effects.

An example of this is shown in **Figure 1** below, which shows the IBS clustering for the WES genotyping. While none of the serial samples in the WES group were mispaired, several unrelated samples clustered together, indicating potential cross-contamination. These three samples are shown in the central region of **Figure 1**, with dendrogram splines marked in green. These samples were subsequentely excluded from any eventual analyses.

IBS clustering was also performed on the genotyping data for the cohort's RNAseq samples,

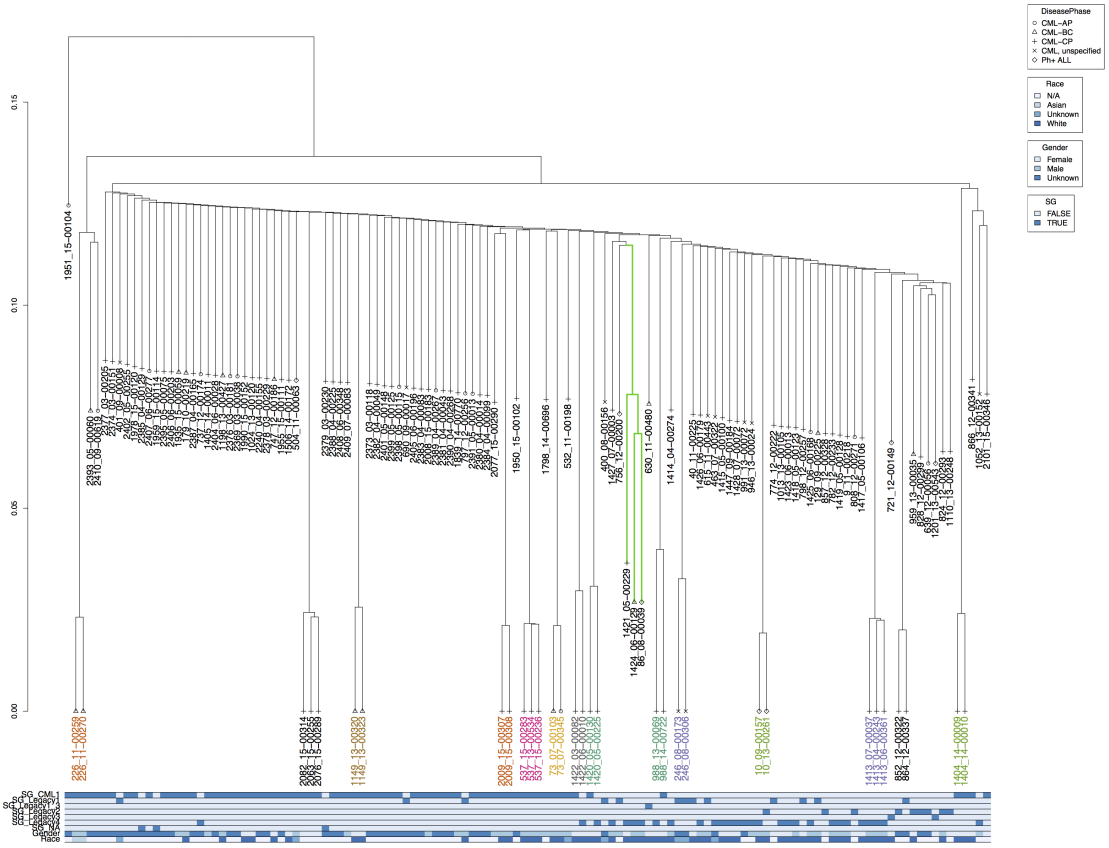

Figure 1: Identity-By-State relatedness clustering of WES samples. Samples with dendrograms highlighted in green show pattern of unexpectedly high relatedness, indicating potential cross-contamination issue.

shown in **Figure 2**. One example of mispaired serial samples from the same patient can be seen, with dendrogram splines highlighted in red. Additionally, when annotated with clinical variables, it was noticed that the genotypes for the majority of the RNAseq samples coming from the Fred Hutchinson Cancer Center clustered together. This group of clustered samples is marked with green dendrogram splines in **Figure 2**. While the clustering coefficient is not as strong as in the potential cross-contamination issue shown in **Figure 1**, the grouping by source was unexpected and was an indicator of some other possible batch effect influencing the sequence results for samples from this location.

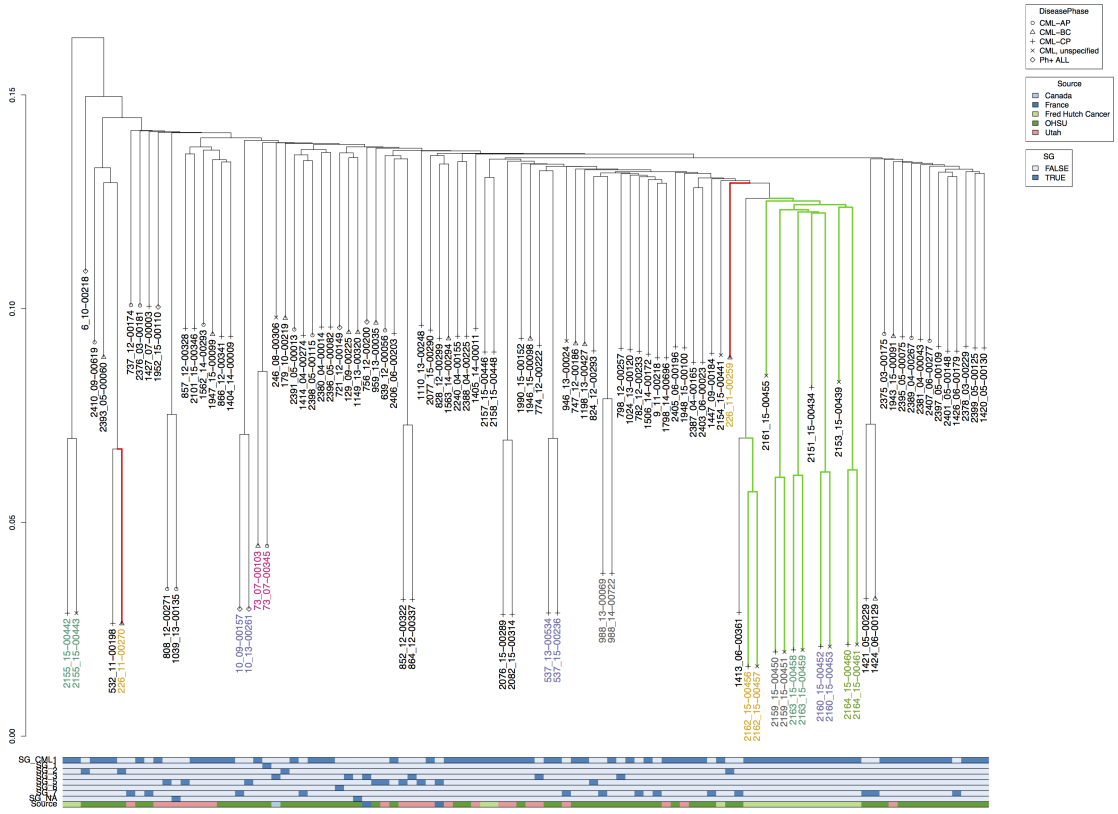

Figure 2: Identity-By-State relatedness clustering of RNA-seq samples. Samples with dendrograms highlighted in green show pattern of unexpected relatedness, indicating potential batch effect. 2 samples with dendrograms marked in red are sample pair from the same patient which failed to cluster together, indicating potential mislabeling issue.

Feature Count Boxplots

The EDAseq R Bioconductor package[28] contains a number of visualization functions useful for conducting exploratory data analysis on RNAseq data for the purpose of discovering potential sources of bias. The first thing that was examined was the overall distribution of gene expression for each sample, vizualized using boxplots in **Figure 3** below. While many individual genes may be expressed at different levels between samples, the overall distribution of gene counts should be similar for all samples. Contrary to that expectation, a number of samples exhibited a distinct pattern with extreme outlier genes and concomitantly lower interquartile ranges(IQR) for the remainder of their genes.

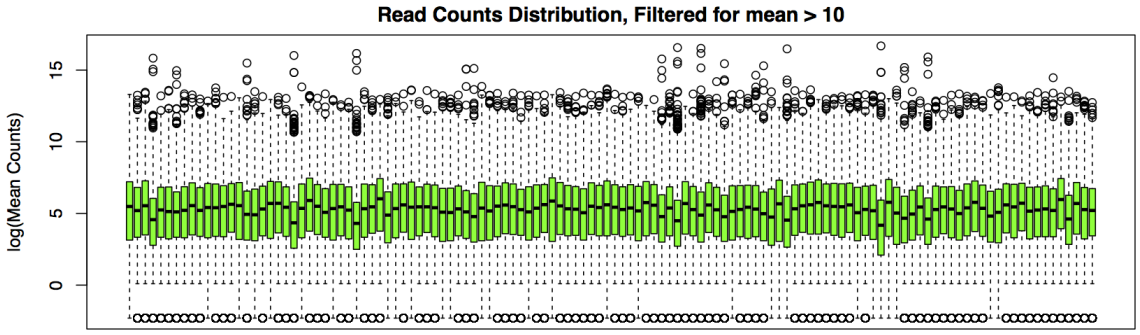

Figure 3: Unordered FeatureCounts BoxPlots - Distribution of log gene counts, for all genes with mean > 10 across entire sample cohort. A number of samples exhibit a distinct distributional pattern with extreme outliers and lowered IQRs.

Ordering the samples by clinical variables, we could examine whether any particular clinical covariates appeared to be strongly correlated with this expression pattern. When sample source is used as the covariate for grouping samples, shown in **Figure 4**, it is clear that almost all of the samples exhibiting this pattern came from the same source, indicating a potential batch effect. The source for the samples exhibiting this batch effect was the Fred Hutchinson Cancer Center, mirroring the pattern seen in the IBS clustering of **Figure 2**.

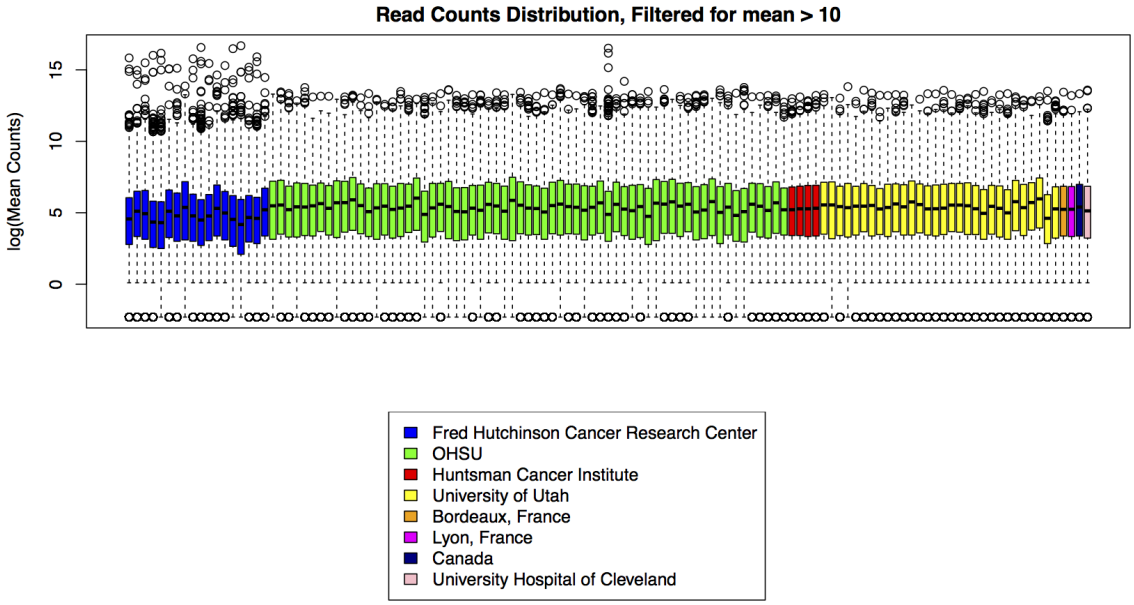

Figure 4: Ordered FeatureCounts BoxPlots - Distribution of log gene counts, for all genes with mean > 10 across entire sample cohort, ordered by sample source.
The expression pattern noted in Figure 3 is clearly correlated with sample source, indicating possible batch issue.

2.2.1 GC Issues

Several additional exploratory visualizations of interest were examined in order to see if any unusual patterns were present in our RNAseq samples. Lowess(locally-weighted polynomial) regression of log gene counts against both gene length and GC content (**Figure 5**) for all samples was performed, in order to determine if any bias was present. While there was no discernible bias indicated by the gene length plots, a number of samples exhibited an unusual trend between gene GC content and gene count, shown in the GC content plot.

The samples exhibiting an unusual relationship between log gene counts and GC content are isolated and shown below in **Figure 6**. The four samples exhibiting the steeper negative slope pattern are all samples from Fred Hutchinson Cancer Center, the same source as the samples exhibiting the batch effect identified in the boxplot graphs.

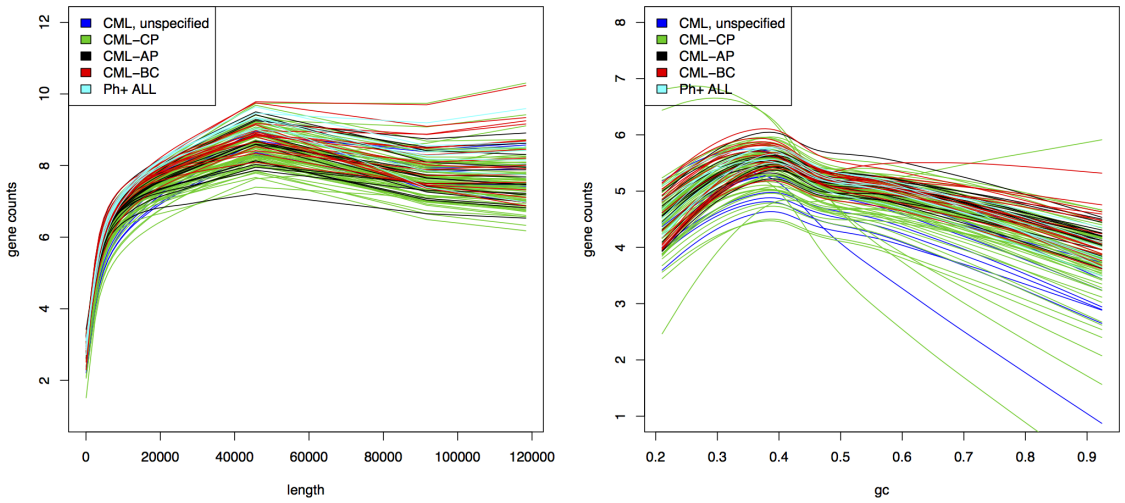

Figure 5: Bias Plots - Loess regression of log gene counts against gene length (left panel) and GC content (right panel).

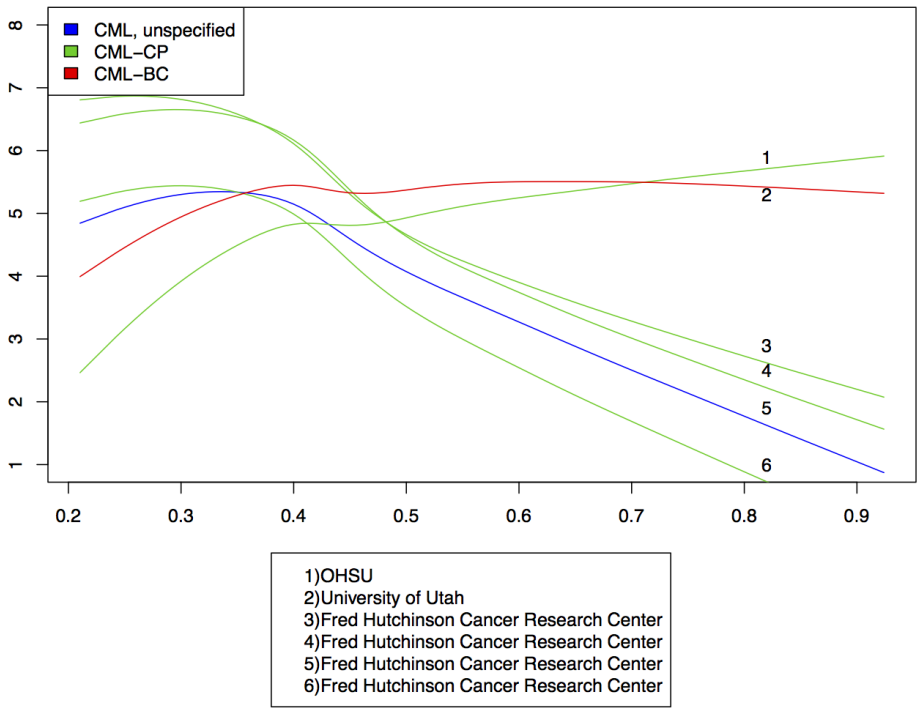

Figure 6: GC Bias Plot Outliers - Loess regression of log gene counts against GC content, highlighting outlier samples with unusual GC content/expression relationship.

2.2.2 FASTQC Plots

In followup to the observation that several of the Fred Hutchinson batch effect samples exhibited an unusual relationship between overall GC content and gene counts, I examined potential GC-bias in more detail by looking at the GC content of raw reads rather than gene-level summary data. The FASTQC set of quality control tools[29] provides a number of measures for evaluating raw RNAseq read files. One of the included analysis modules is Per Sequence GC Content (**PSGCC**), which graphs the observed distribution of read counts for different values of mean read GC content. This observed distribution is overlayed with the expected theoretical normal distribution which would be observed in the case of an unbiased sequencing library. Sharp, unexpected peaks in the observed distribution may be indicative of specific contaminants or strongly overrepresented sequences, and FASTQC will issue a failure flag if more than 30% of the reads deviate from the theoretical normal distribution.

Among the batch effect (Fred Hutchinson) samples, 17 of the 18 (94.4%) samples were flagged for failure on the **PSGCC** Content module, whereas only 31 of 103 (30.1%) normal samples (those not exhibiting the batch effect distribution pattern in the FeatureCount boxplot graphs) were flagged. Examining the **PSGCC** graphs more closely, the samples from the batch effect group all exhibit a strongly abherrant distribution of mean GC content, with numerous sharp peaks. In contrast, for normal samples, the mean GC content distributions closely match the expected theoretical distribution, even in the case of the normal samples which were flagged for failure. Examples of **PSGCC** Content graphs for both normal and batch effect samples are shown in **Figure 7**.

This abherrant GC distribution almost perfectly matches the pattern shown in the FeatureCount boxplot graphs. In fact, the two samples which do not fit the batch effect sample/source pattern shown in **Figure 2** do conform with this observed GC-bias pattern. The one batch sample which did not fail the FASTQC Per Sequence GC Content module, 15-00446, is also the only sample from Fred Hutchinson that did not display the FeatureCount boxplot motif of extreme outlier genes. Conversely, the only sample which displayed this

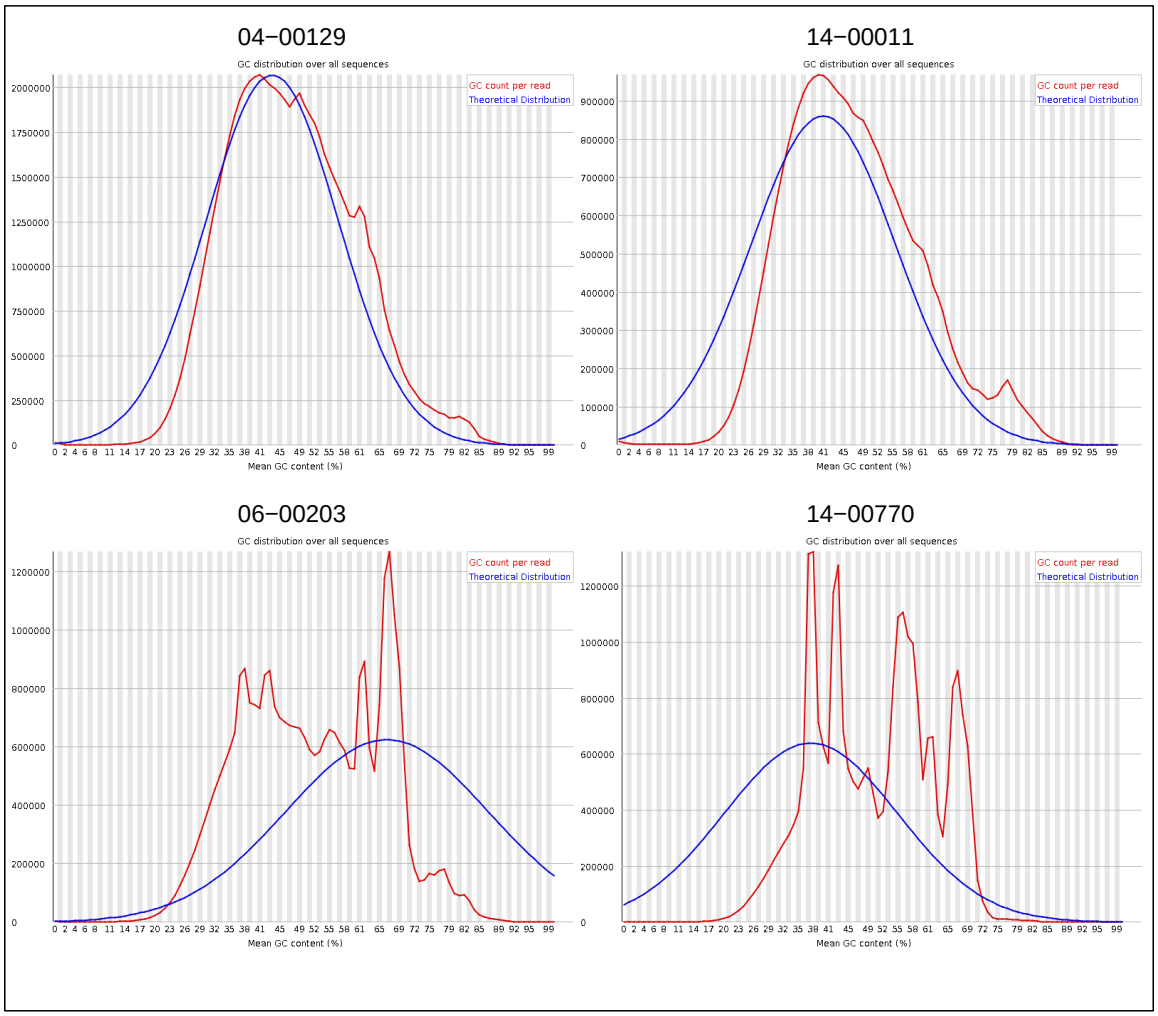

Figure 7: FASTQC Per Sequence GC Content (PSGCC) graphs for a selection of normal (non-batch) and batch samples. The observed distribution of read counts for different values of mean read GC content is shown in red. The expected theoretical normal distribution which would be observed in the case of an unbiased sequencing library is overlayed in blue. Top panel are normal samples, bottom panel are batch samples. Note: Sample 14-00011 failed PSGCC module

pattern but was not from the Fred Hutchinson group, 05-00225, also displayed the unusual distribution of mean GC content. FeatureCount boxplots highlighting these two samples and their **PSGCC** graphs are shown in **Figure 8** below.

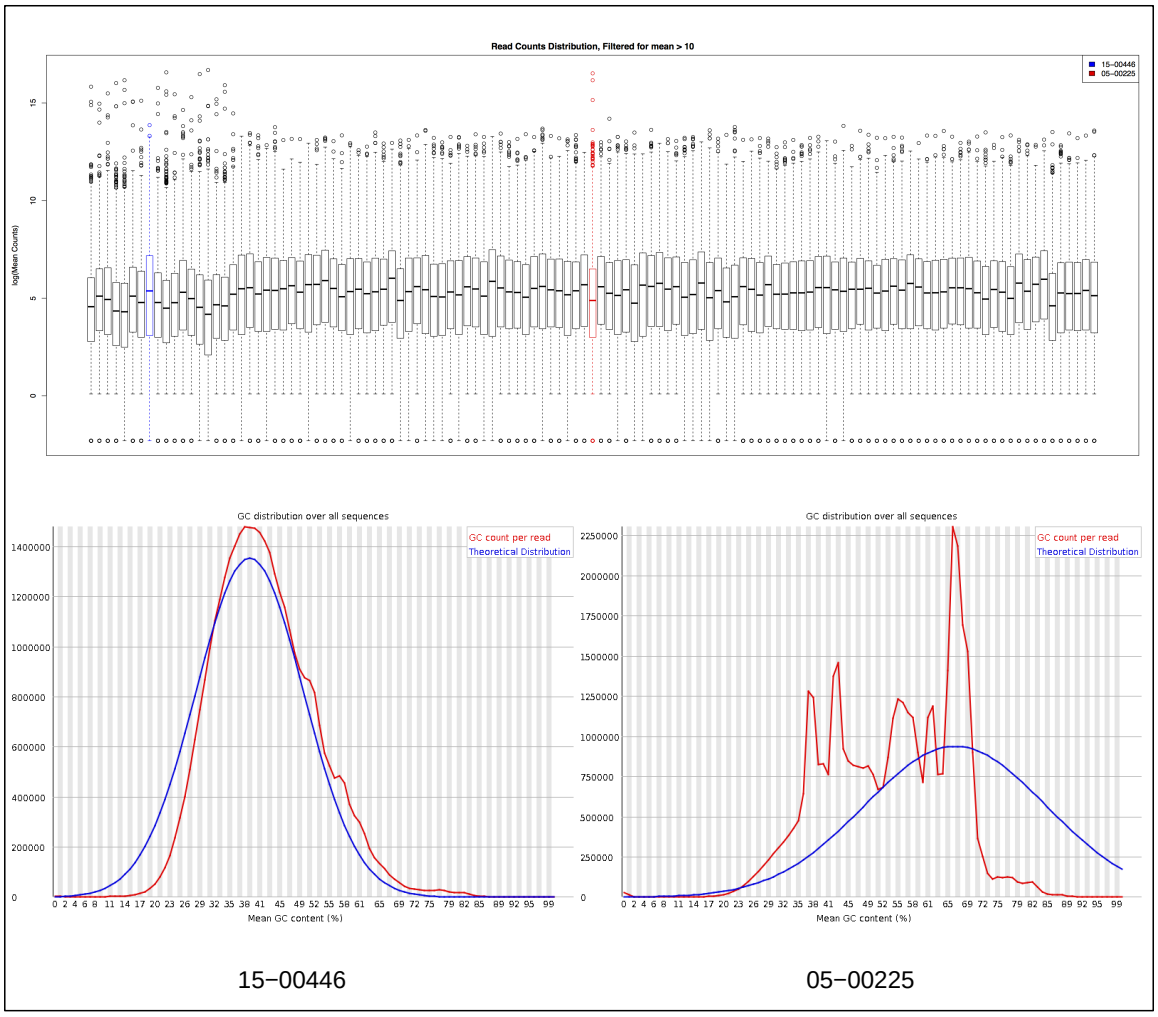

Figure 8: FeatureCounts Boxplots and FASTQC Per Sequence GC Content (PSGCC) graphs for two samples lacking distribution pattern/source batch correlation. Single sample from Fred Hutchinson(FH) lacking boxplot outlier pattern (15-00446) highlighted in blue, with matched PSGCC graph shown in bottom left panel. Single non-FH sample displaying boxplot outlier pattern (05-00225) is highlighted in red, with matched PSGCC graph shown in bottom right panel.

Looking at the complete FASTQC reports for the samples, a majority of the batch effect samples were also flagged by FASTQC's Overrepresented Sequences module. The overrepresented sequences from the batch samples were pooled and NCBI's Nucleotide Blast tool[30] was used to search for potential source transcripts. The top BLAST results returned for

the overrepresented sequences, which were present in all of the batch samples (excluding sample 15-00446), are shown in **Supplemental Table 5** in the Supplemental section. Excluding matches to cloning vector sequences, all of the hits were for sequences related to the hemoglobin beta subunit.

2.2.3 Globin Genes

Top outlier genes were then extracted for all of the batch samples, and searched for those annotated to globin genes. This returned nine globin outlier genes, shown in **Table 1** below.

Table 1: Fred Hutchinson Outlier Globin Genes

Ensemble Gene ID	HUGO ID	Description
ENSG00000206172	HBA1	hemoglobin, alpha 1
ENSG00000257017	HP	haptoglobin
ENSG00000244734	HBB	hemoglobin, beta
ENSG00000188536	HBA2	hemoglobin, alpha 2
ENSG00000206177	HBM	hemoglobin, mu
ENSG00000213934	HBG1	hemoglobin, gamma A
ENSG00000196565	HBG2	hemoglobin, gamma G
ENSG00000223609	HBD	hemoglobin, delta
ENSG00000169877	AHSP	alpha hemoglobin stabilizing protein

Of these, Hemoglobin Beta and the two Hemoglobin Alpha Subunit genes were present as extreme outliers in nearly all of the batch samples. Highlighting these three genes in **Figure 9** below, the correlation between expression of these three globin genes and the FeatureCounts boxplot pattern is clear.

In fact, the expression pattern for these three globin genes nicely matches the two obvious exceptions to the Fred Hutchinson batch pattern. As noted earlier, sample 15-00446 from

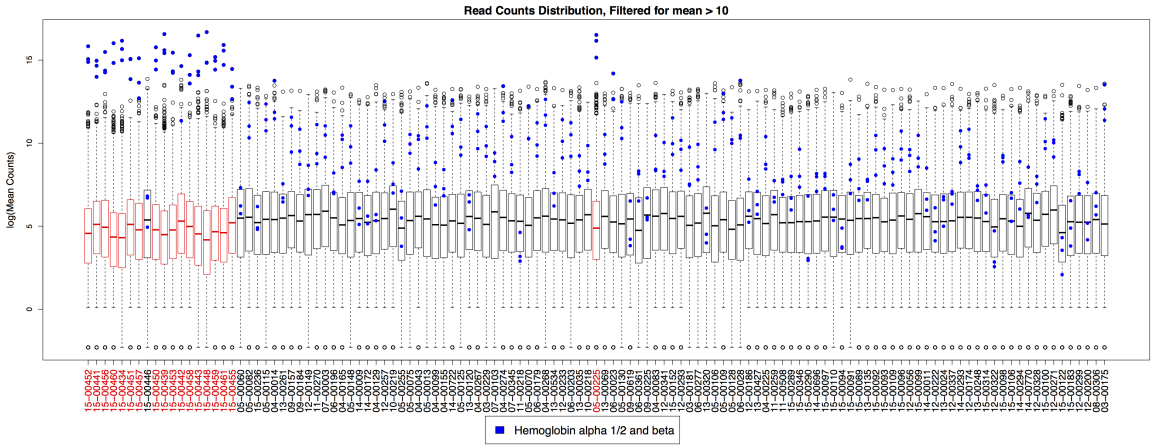

Figure 9: FeatureCounts BoxPlots - Select globin genes highlighted. The samples exhibiting the batch/outlier pattern (Fred Hutch, with 2 noted exceptions) are highlighted in red. The globin alpha and beta 1/2 genes in all samples are highlighted in blue. This clearly shows correlation between batch/outlier pattern and abundance of transcripts for these 3 globin genes.

the batch group was the one sample which did not fail the FastQC Per Sequence GC Content module, and it also was the only batch sample lacking the extreme outlier pattern in the FeatureCount boxplots. As can be seen in **Figure 9**, this was the single batch sample with expression of the three flagged globin genes within the main distribution of gene counts. The other exception to the batch pattern was sample 05-00225, which both displayed the extreme outlier gene pattern in the FeatureCount boxplots but also displayed the abherrant Per Sequence GC Content distribution. In this case, the expression level for the three flagged globin genes matches the extreme outliers and explains both patterns. Based on these observations, it was clear that overrepresentation of these globin genes was the source of the observed distributional pattern, which strongly correlates with the sample source from Fred Hutchinson Cancer Center.

The identification of highly expressed globin genes as a common feature of this batch effect group offered a potential experimental source of the observed pattern. Depletion of globin transcripts has become a standard step in preparation of peripheral blood samples collected for NGS studies, in order to prevent these highly abundant transcripts from saturating the available sequencing primers and being disproportionately represented in downstream

sequencing data[31].

Followup examination of the clinical data confirmed that the RNA samples from Fred Hutchinson were isolated from whole blood cells, unlike the peripheral mononuclear blood cells harvested in the samples from the other sites. As a result the red blood cells were not removed prior to cell lysis during sample prep. This could account for a higher level of initial globin transcripts present in these samples. However, while the difference in cell sample isolate correlates perfectly with the sample source, it does not explain the two samples (15-00446 and 05-00225) which do not fit the batch effect/source pattern. Whatever the source, excess globin transcripts are clearly the cause of the batch effect initially observed in **Figure 3**, introducing potential technical bias to the samples affected, primarily from Fred Hutchinson Cancer Center.

2.2.4 Dropped Genes

The presence of excess globin transcripts during sample prep for this batch of samples may have had a profound effect on the level of sequencing across the entire set of transcribed genes, beyond the presence of these highly expressed globin genes. This is evident in the noticeably lower IQRs for many of the featureCount boxplots of the batch samples, shown in **Figure 4**. It has been noted that without proper globin depletion, globin transcripts in peripheral blood samples can comprise as much as 50-75% of total mRNA present, and lead to a proportional monopolization of the resultant sequencing reads[31]. One potential effect of this is the biased dropping of low expression genes[32] from the batch samples, which have not been properly globin depleted. Selective dropping of genes from one group in our primary variable of interest (resistance type) could seriously confound any potential analysis between the two groups, such as differential expression or pathway enrichment analysis.

Despite the greater dynamic range afforded by RNAseq technology for detecting transcripts in comparison with microarrays, low expression genes are disproportionately effected by the

random sampling process in NGS[33] and thus exhibit less accurate observed gene counts. In many cases, these noisy low expression genes may fall below the threshold of detection, resulting in the appearance of no expression. Therefore, it is expected that across our entire sample cohort we would observe a number of samples with zero read counts for particularly low expression genes.

Given that fact, a gene with zero read counts for all or nearly all of the samples in a specified group (such as samples from one source) may simply be the result of random sampling of all the zero count samples for that gene. However, the dominance of the globin gene transcripts during sequencing may shift the expression levels of all the remaining genes downward, pushing more genes close to the threshold of detection and artefactually increasing the number of samples exhibiting zero read counts.

In order to identify whether genes are being "dropped" from the batch group that one would not expect to occur by random sampling, the binomial probability of each individual gene (with zero gene counts for at least one sample within the batch group) being absent through chance was calculated, seperately for each sample source. Due to the differences in size between the different groups, the dropped genes from groups with fewer samples do not have the possibility of binomial probilities as small as for groups with more samples, because of the smaller n in $PrBin(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$). This means that a comparison of the binomial probabilities of "dropped genes" between two source groups with unequal number, is not perfectly even. For this reason, the best comparison of dropped genes is done by specifying the single group of interest and then randomly subdividing the remaining samples into equally sized groups. This is shown for the batch samples (those exhibiting the globin outlier and GC bias patterns) in **Figure 10**. The large number of dropped genes with very low binomial probability of random selection for the batch group is striking when compared with the remaining samples randomly subdivided into six equal groups.

This visualization of dropped genes supports the notion that this batch effect, linked to the presence of the three Globin outlier genes, does in fact lead to potentially biased elimination

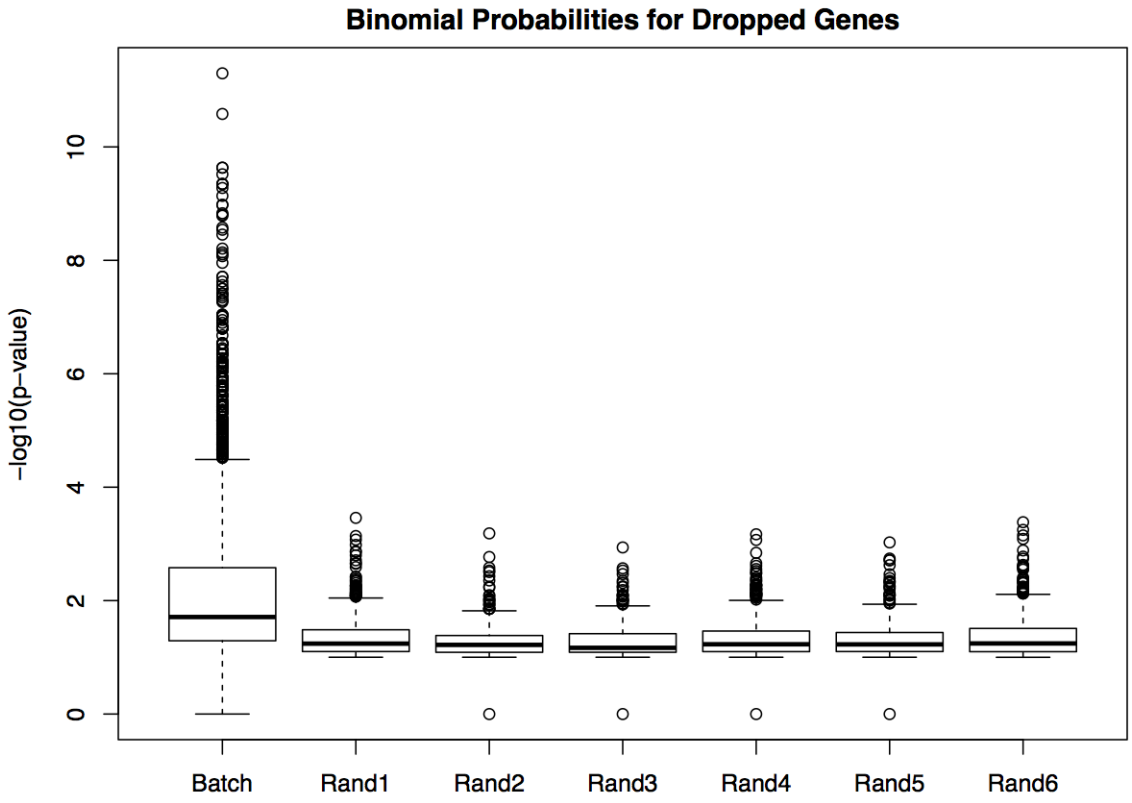

Figure 10: Dropped Genes - Log binomial probabilities of a gene having k samples identified as having zero read counts for gene across designated sample group(Fred Hutchinson), with remaining samples randomly assigned to equally sized groups.

of low-expression genes from the batch group samples.

2.2.5 Differential Expression

Having identified globin transcript contamination during sample prep as the likely culprit of the batch effect, and a clear pattern of bias between batch and normal samples for both the three outlier globin genes and the low expression dropped genes, the full distribution of genes was examined more closely to assess whether removal of the genes identified as biased could allow for normalization of the batch effect. It is common practice to filter low expression genes out before performing Differential Expression analysis[34]. If the bias introduced to the batch samples is primarily focused in the low expression genes which will be eliminated by this filtering step, it may be possible to use standard methods of normalization to adjust global expression levels of the remaining genes.

First, Principal Component Analysis was used to plot all of the samples using their gene expression in the first and second principal components, in order to visualize the clustering of the samples both with and without the genes biased by the batch effect. The principal components visualizations are shown in **Figure 11** below. The top row left panel shows the PCA with all genes included, and the top row middle panel shows the same visualization with the three globin outlier genes and the "dropped" genes removed. As a control comparison, the top row right panel shows the PCA with samples colored by Gender rather than batch status. As can be clearly seen, the samples seperate fairly distictly into two primary groups, indicating that the batch variable is exerting a strong influence on the set of genes which display the most variance across the entire dataset. Furthermore, this effect is virtually unchanged after elimation of the globin outlier and dropped genes, suggesting that the bias introduced by the batch effect is present across the entire distribution of genes.

The next natural step is to evaluate whether standard normalization procedures can globally adjust the expression levels in a way that corrects for this apparent bias, if in fact it is simply

an issue of scaling for the remaining genes. PCA graphs after the use of three standard normalization scaling methods[28], which adjust the global expression values in order to equalize across samples using the median, upper quantile, and full quantile as equalizing measures, respectively, are shown in the bottom row of **Figure 11**. Even after normalization the batch and normal samples are clearly segregated by both the first and second principle components, indicating a significant bias in the expression data for the batch group, even with the globin outlier and dropped genes removed.

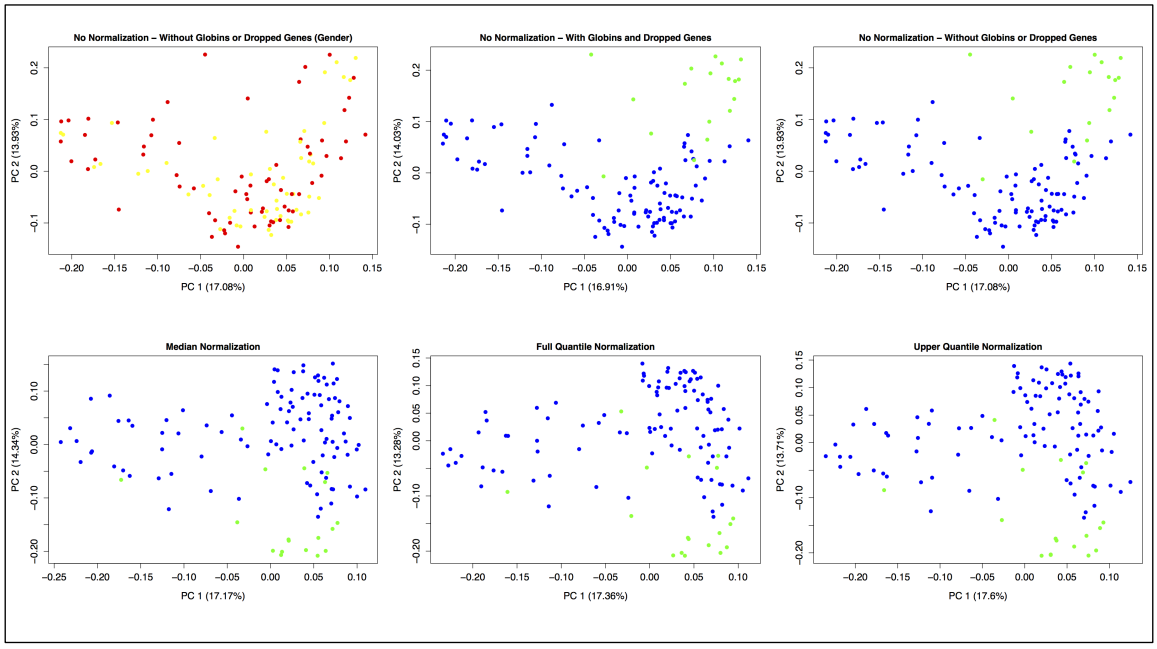

Figure 11: Principle Components Visualization - Unnormalized data shown in top row. Samples colored by Gender in top-left PCA plot, male (red) and female (yellow). All other PCA plots colored by batch status, batch (green) and normal (blue). Top-center plot shows PCA with all genes retained. Top-right plot shows PCA with globin outlier and dropped genes removed. Bottom row shows PCA plots for data normalized using Median, Full Quartile, and Upper Quartile normalization, all with globin outliers and dropped genes removed.

Simple evaluation using principal components visualization seemed to suggest that the gene expression of samples identified in the batch group was biased, even with the globin outlier genes and "dropped" genes removed, and that this bias is not easily corrected by standard methods of normalization. In order to more quantitatively assess this bias, I performed differential expression analysis comparing the batch and normal groups using the edgeR package[34]. Differential expression between sample groups divided by an unimportant

variable should result in little or no significantly differentially expressed genes, provided the sample groups are reasonably even in size. For this reason patient gender, which also showed no clear pattern of seperation in the PCA plots, was selected as a control differential expression variable. The entire table of DE results is shown in **Table 2** below.

When comparing the batch and normal groups, the number of differtially expressed genes identified is extremely high, from 50-52% of the total genes depending on the normalization method used. The full dataset contains a large degree of heterogeneity for various clinical variables and is quite imbalanced in terms of sample size (batch(n=18) vs normal(n=105)). Another comparison was made restricting the samples used to only include those from the more clinically relevant Chronic Phase of the disease, as these are the most likely samples to be used in any two-group comparisons. The samples used were also restricted to those from the BCR/ABL-independent Imatinib resistant group in order to ensure that any differential expression seen was not due to this variable rather than batch status. Even with this comparison, in which sample group size is more even and other major disease variables have been restricted to a single class, the total number of DE genes is still approximately 15% of the total genes. By comparison, for the control differtial expression analysis, comparing groups based on Gender, the number of DE genes is negligibly small (0.2-1.2%).

Table 2: edgeR Differential Expression Totals

	Batch vs Normal			
	No Norm	Upper Quant	Full Quant	Median
All Samples				
Total	19176			
DE(-1)	3643	4778	4012	3860
DE(+1)	5986	4832	5867	6071
DE(0)	9547	9566	9297	9245
Ind/CP Only				
Total	19176			
DE(-1)	1284	1619	—	—
DE(+1)	1400	1319	—	—
DE(0)	16492	16238	—	—
	Male vs Female			
	No Norm	Upper Quant	Full Quant	Median
All Samples				
Total	19176			
DE(-1)	58	52	—	—
DE(+1)	164	142	—	—
DE(0)	18954	18982	—	—
Ind/CP Only				
Total	19176			
DE(-1)	10	18	—	—
DE(+1)	36	22	—	—
DE(0)	19136	19136	—	—

In addition to a very large number of significant DE genes, the batch vs normal DE comparison also includes a large number of outliers with extremely small p-values. These highly significant DE genes represent genes expressed across the distribution of gene expression read counts. This is shown clearly in **Figure 12**, which highlights the top 50 most significant DE genes identified for the batch vs normal comparison. The presence of this large number of highly significant DE genes, in genes across the distribution of expression levels, demonstrates the presence of a strong, systematic bias introduced by the batch effect.

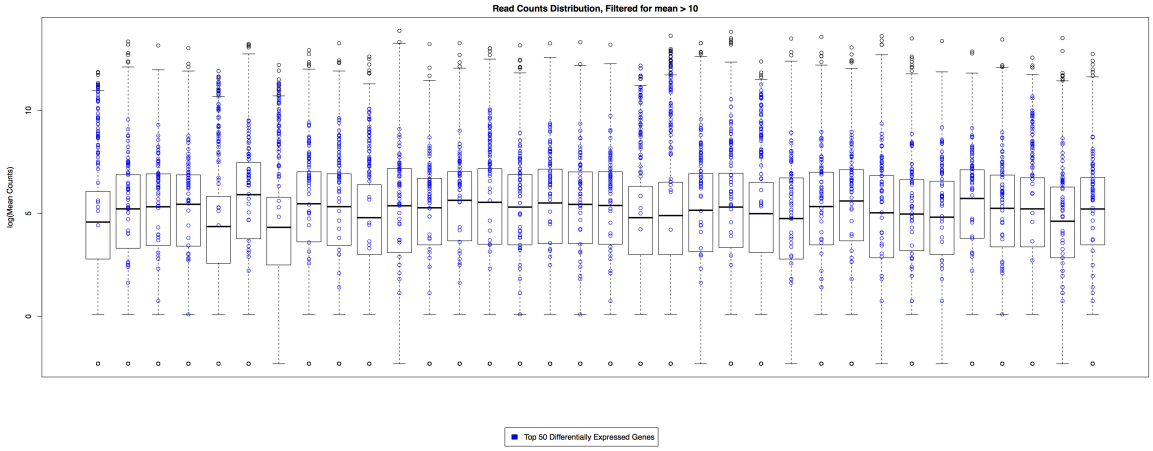

Figure 12: FeatureCounts Boxplots - Top 50 differentially expressed genes between batch and normal samples highlighted in blue. DE analysis with edgeR restricted to BCR-ABL Independent and Chronic Phase samples only.

Given the systematic bias introduced by the batch effect, the most important issue is one of statistical confounding with our primary variable of interest, Imatinib resistance. The batch pattern is almost entirely confounded with our resistance variable, with 16/18 batch samples in the BCR/ABL-independent group. While there are more sophisticated methods of normalization, which allow for explicit modeling of technical factors like batch variables, this level of confounding between an experimental variable and batch effect variable is fundamentally uncorrectable and can lead to erroneous analysis, as has been shown in several studies[35].

2.3 Conclusion

The gene count distributional pattern exhibiting extreme outliers shown in **Figure 4**, and it's correlation with sample source, revealed a serious batch effect attributable to excess globin transcripts during sample preparation for sequencing. This overabundance of globin sequences created a bias in the batch samples, leading not only to the dropping of many low expression genes in the batch group below the threshold of detection, but also to the appearance of strong differential expression in a large proportion of genes surveyed, across the range of expression levels.

This clearly illustrates the way in which systemic bias can be introduced by minor differences in sample preparation protocol, in this case inconsistent handling of globin transcript depletion. Furthermore, when these sources of bias, or batch effects, are not identified by careful exploratory analysis at the onset of a project, it is possible for any downstream analysis to be strongly skewed as a result. While some analysis methodologies allow for normalization of technical variation without explicit pre-specification of the causitive variables, these methods are still severely limited if the batch variable is highly confounded with the primary variable of study, as in this case. The batch effect identified here is a perfect illustration of the need for careful exploratory analysis in NGS studies, particularly when pooling samples where every step of sample generation cannot be made uniform.

3 Genotyping Variant Filtering and Dimensional Reduction

3.1 Introduction

With the exception of certain well known inherited mutations such as the BRCA1/2 genes strongly linked to breast and ovarian cancer[36,37], cancer-causing mutations are largely considered to arise from new mutations to somatic tissues as opposed to mutations inherited from germline tissues. For this reason, most genomic studies of cancers and methods developed for the analysis of cancer genomes have focused on somatic variants. The most widely used method for distinguishing germline from somatic variants in NGS experiments is the collection and analysis in parallel of a matched normal tissue sample from the same patient. Unfortunately, matched normal samples are rarely aquired in the clinical settings under which many cancer study samples are initially collected, due to reasons ranging from the need for additional consent paperwork in aquiring germline samples, to increased lab and sequencing costs[38]. Additionally, the issue of collecting appropriate "matched normal" samples for blood cancers is further complicated by the circulating nature of the tumor cells whereby contamination of normal tissues by leukemic cells can be an issue.[39]. As a result for studies such as this one, which pool samples from leukemia patients collected in a clinical setting, matching normal samples are largely unavailable and the ability to differentiate somatic from germline mutations is curtailed. A further issue created by the inability to distinquish somatic and germline mutations is the size of the resulting variant lists. Since much bioinformatic software for analysis of cancer variants has been designed on datasets with matched normal samples to allow somatic calls, these methods are subsequently tested and validated on dramatically smaller variant/gene lists, raising potential issues of performance and validity on larger-scale data.

Apart from using match normal samples, one of the best remaining metrics that can be used to filter out potential germline variants is the use of single-nucleotide polymorphisms (SNPs) common to the general population that are annotated in various databases. The

assumptions behind this approach are twofold; both that the human reference genome is a somewhat inaccurate and static snapshot of the normal human gene sequence, due to its haploid nature and the small number of individuals the reference was built from, and that there is diversity across the human population at numerous loci sequences, which are benign and play no role in disease. For these reasons, a number of resources have been developed to pool data from published research and across different human populations, in order to compile population catalogs of commonly seen variants and if possible provide an estimate of their frequency in the populace. Two of the most commonly used population catalogs, NCBI's dbSNP[40] and the 1000 Genomes Project[41], have been used together extensively for filtering of germline variants. While allele frequencies (AF) are included in dbSNP based on their prevalence in the original 1000 Genomes cohort, these AFs only offer a very loose estimate of true frequency in the population and are of limited use for filtering. Furthermore, the dbSNP and 1000G databases pool data from a number of sources and sequencing methodologies, and it understood that some poor quality variants and somatic variants are present in these datasets, raising the rate of false negative variants which will be erroneously filtered out. A far more powerful population catalog released in the last few years is the Exome Aggregation Consortium (ExAC) dataset[42], which provides variant data from 60,706 whole exome sequenced samples, expanding both the number of variants included and number of samples used to estimate frequency by an order of magnitude. This provides a powerful tool for fine-tune filtering of potential germline variants based on population frequency.

Additional criteria for filtering of variants is often based on inclusion in certain more specifically curated variant databases. Several examples of these are the clinVar[43] and COSMIC[44] databases, which catalog previously identified pathogenic variants and cancer-causing variants, respectively. Functional annotation of variants can also be used to decide which variants to retain, for example restricting inclusion to those variants predicted to have specific downstream effects according to tools such as Ensemble's Variant Effect Predictor[45]. While the use of these annotations allows for significant reduction in the number of variants retained and therefore a focus on variants more likely to be relevant for various reasons,

there is an inherent tradeoff to such filtering choices which somewhat undercuts the powerful open-ended discovery framework of NGS studies. For example, in restricting subsequent analysis to only variants with previously known phenotypic or disease etiologic effects, or to those with a narrow range of predicted functional impacts, the ability to discover and utilize truly novel gene mutations is sharply limited.

In addition to the issue of distinguishing germline versus somatic variants, there are numerous sources of artefactual variant calls which arise in NGS studies. Examples of these are miscalled bases on the 3' ends of reads and in homopolymeric regions, alignment errors in regions of low mappability, and uneven coverage due to GC bias or primer border regions in PCR library preps[46]. Many of these issues can be dealt with by filtering based on minimum mapping quality, read depth, and presence in regions of poor mappability, filters which are in fact standard components of common genotyping pipelines such as the Genome Analysis Toolkit (GATK)[27] used in this study. However, further post-genotyping filtering based on some of these metrics may be necessary in order to make certain that variants are not introducing inaccuracy or bias when evaluated in the context of cohort or group-level mutational frequencies, rather than individual samples alone. For example, while a variant may be retained in the genotyping output if it passes these minimum quality metrics in at least one sample, it may be lacking good quality or valid genotypes at the same loci in all the other samples, rendering any cohort-level measure such as mutational frequency meaningless.

Given this complexity selecting variants in cancer studies lacking matched normal samples, careful strategies must be devised for filtering variants in order to eliminate as many potentially artifactual variants as possible, increase the proportion of somatic variants remaining, and reduce the total variant/gene totals down to levels appropriate for the desired analysis methods.

3.2 Results and Discussion

3.2.1 Filtering Totals

The genotyping results (SNVs and Indels) for the CML Imatinib Resistance project cohort contain approximately 1 million variants, called using GATK's genotyping pipeline[27]. Without matched normal samples to distinguish germline from somatic variants, we need alternative criteria to reduce the number of variants/genes included in our downstream analysis. The total variants undoubtedly include both a large number of false positives arising due to technical artifacts, and also a mix of relatively few important "driver" mutations and many symptomatic "passenger" mutations. In both cases, a filtering scheme to minimize these potential sources of noise would be advantageous. Additionally HotNet2, which deals with gene-level mutation data, was originally run on data covering approximately 12K genes, and filtering our dataset down to a comparable scale was deemed desirable.

Filtering Steps

In order to reduce the degree of symptomatic background genetic heterogeneity and focus on samples more relevant to the aquisition of imatinib resistance and secondary leukemia progression, the sample cohort was restricted to those designated as Chronic Phase, and to those clearly categorized with either a BCR/ABL-independent or BCR/ABL-dependent resistance subtype. This reduced the set of WES samples selected from the total genotyping data set to a total of 41 samples(BCR/ABL-independt n=26, BCR/ABL-dependent n=15). Variants were then filtered out based on several criteria at the cohort level.

1. Variants with all samples within the cohort having overall read depth at variant locus less than 10 reads were filtered out.
2. Variants present in the ExAc database at an allele frequency greater than 0.1% were

filtered out.

The total number of variants filtered out at the cohort level for either uniformly insufficient read depth and for presence in the ExAc database at several potential allele frequency(AF) thresholds is shown in **Table 3** below. An ExAc database AF threshold of 0.01 was chosen.

Table 3: Cohort Level Filtering Totals

	Filtered Out	Remaining
Total		966108
Filtered out variants for which all samples have read depth < 10	286142	679966
Filtered out for ExAc AF > 0.05	119693	592074
Filtered out for ExAc AF > 0.01	87892	560273
Filtered out for ExAc AF > 0.001	147251	532715

The cohort was then split into BCR-ABL-independent and dependent groups. Only those variants with at least one valid, non-reference genotype call in the group were retained.

Further **secondary** filters were applied, either to the entire cohort or seperately to the groups, in order to eliminate variants with missing genotypes or insufficient read depth, which may be an indication of variants called at sites with poor coverage and which can complicate the accurate calculation of variant frequency within the groups.

1. Variants with any missing genotypes were filtered out.
2. Variants with overall read depth less than 10 reads for any sample in the group at the variant locus were filtered out.

Filter Scheme

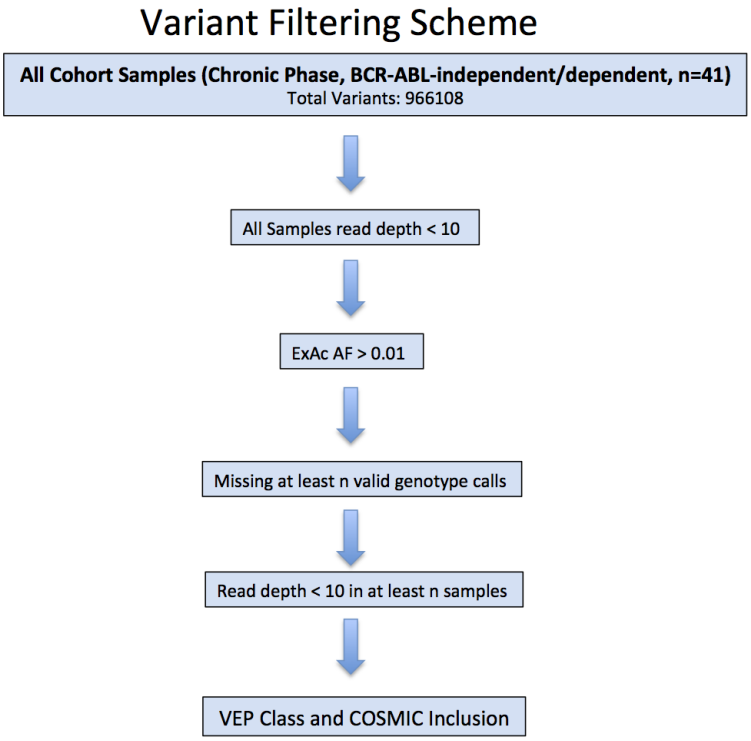

Figure 13: Overall filtering scheme. Primary steps are: 1) Minimal read depth across cohort 2) Either absence from ExAc database or presence below designated allele frequency threshold of 0.01 3) Missing Genotype filter 4) Read Depth filter 5) Selection based on Variant Effect Predictor(VEP) annotation and/or presence in COSMIC database.

Variant Classes

Variants were annotated with Ensemble's Variant Effect Predictor(VEP)[45] and a single "most deleterious" transcript and corresponding VEP functional consequence category were determined using the vcf2maf perl script. Variants were further grouped into three classes of VEP functional consequence, for further selection after filtering.

Class 1 (Severe protein coding consequence)

- frameshift variant
- stop lost
- stop gained
- start lost

Class 2 (Further coding variants)

- missense variant
- splice region variant
- splice donor variant
- splice acceptor variant
- coding sequence variant
- stop retained variant
- incomplete terminal codon variant
- initiator codon variant
- inframe deletion
- inframe insertion
- protein altering variant

Class 3 (Regulatory Variants)

- Nonsense-mediated decay(NMD) transcript variant
- 3'prime UTR variant
- 5'prime UTR variant
- TF binding-site variant
- non-coding transcript exon variant
- non-coding transcript variant
- mature miRNA variant

3.2.2 Evaluating Secondary Filtering Step Thresholds

The purpose of the secondary filters (missing genotypes (**MG**) and read depth < 10 (**RD**)) is to eliminate variants with missing or low confidence data which will add noise or complicate calculation of gene-level mutational frequency. However the use of a hard cutoff (only one sample having a missing genotype or read depth less than 10 at the variant locus) is potentially too strict and may filter out many variants with nearly complete genotype calls and/or sufficient read depth across most samples, which we may wish to keep. In total, 450,401 variants were filtered out when using the strict cutoffs for these filters (> 0 samples with missing genotypes or read depth < 10 at a particular locus). Additionally, if relaxed secondary filtering thresholds are used we need to evaluate whether those filters should be applied at the level of the entire sample cohort or at the group level after splitting the samples into BCR-ABL Independent and Dependent groups.

Filtering at Group or Cohort Level

One question about the secondary filtering steps is whether to filter at the group or cohort level. In order to assess this, the variants were filtered both ways and the distributions of

filtered variants were plotted, with the variants grouped into those present in both groups or only in one. If filtering is done at the group level and a large number of variants are filtered for one group but not both, this could create the appearance of a difference in mutational frequency between the two resistance groups for a particular gene which is an artifact of these secondary filters. If present, this could introduce bias to our downstream network analysis.

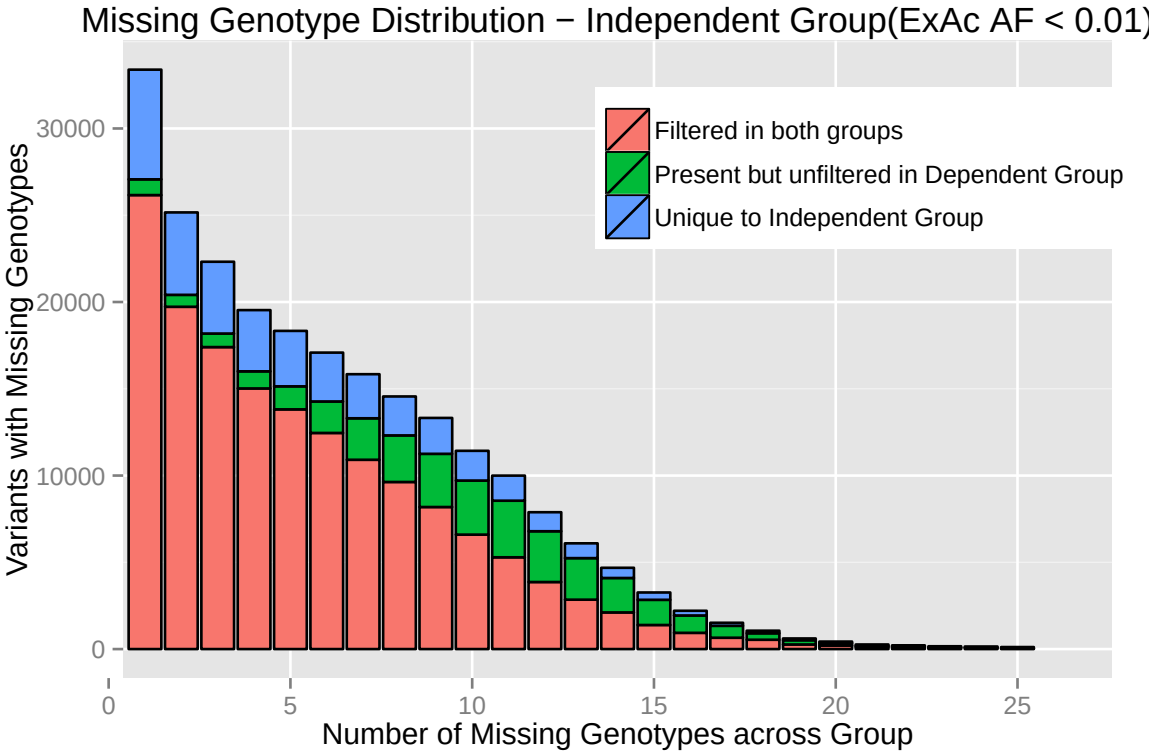

Figure 14: MG Filter Distribution-Independent Group: This shows the number of variants filtered out using an MG filter cutoff of zero, binned by how many missing genotypes each variant has across the Group. Resistance group membership is shown using bar color; variants filtered out from both groups (red), variants present but unfiltered in the opposite group (green) and variants unique to the group(blue).

In the distribution of variants filtered for missing genotypes(**MG**) in the BCR-ABL Independent group, shown in **Figure 14**, there are a sizable number of variants filtered out which are also present in the opposite (Dependent) group, but which are not filtered in that group (shown in **green**). Furthermore, in the distribution of variants filtered for low read depth(**RD**), shown in **Figure 15**, there again are a sizable number of variants filtered out which are present in the Dependent group, but which are not filtered in that group.

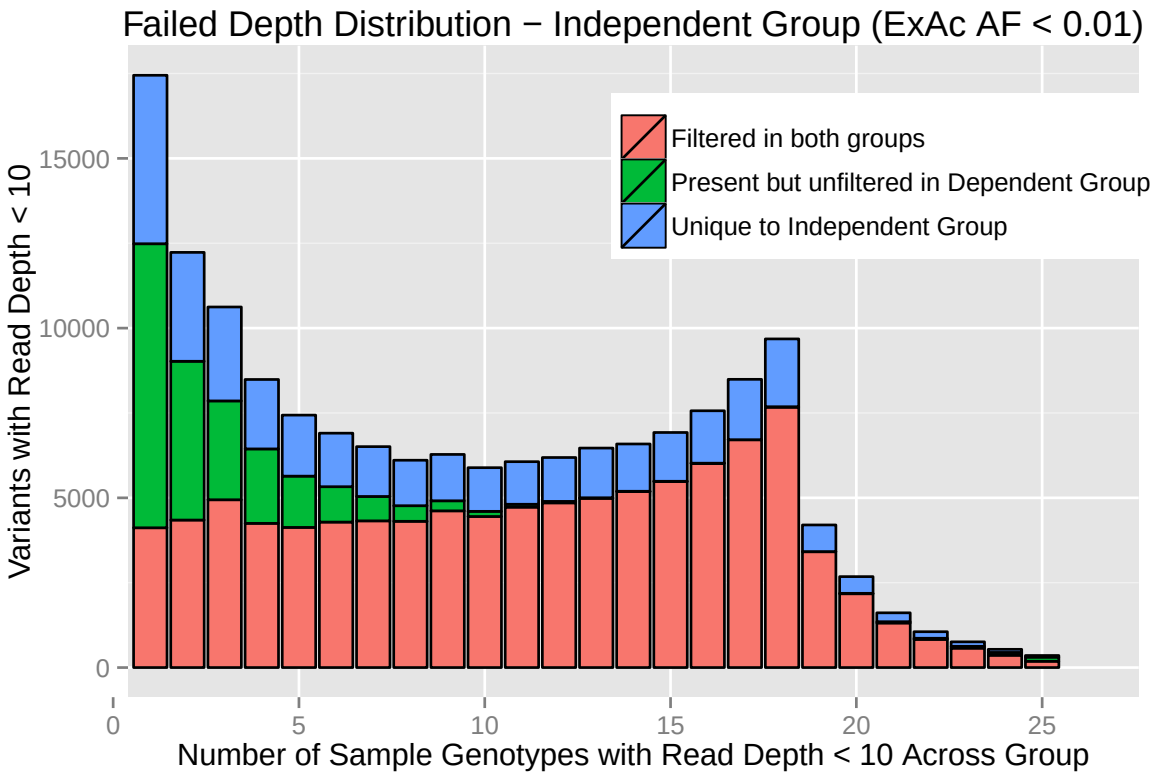

Figure 15: RD Filter Distribution-Independent Group: This shows the number of variants filtered out using an RD filter cutoff of zero, binned by how many sample genotypes each variant has with read depth < 10 across the Group. Resistance group membership is shown using bar color; variants filtered out from both groups (red), variants present but unfiltered in the opposite group (green) and variants unique to the group(blue).

Similarly, in the distribution of variants filtered for missing genotypes in the BCR-ABL Dependent group, shown in **Figure 16**, there are a number of variants filtered out which are also present in the opposite (Independent) group, but which are not filtered in that group (shown in **green**). In the distribution of variants filtered for low read depth, shown in **Figure 17**, there are a small number of variants filtered out which are also present in the Independent group, but which are not filtered in that group.

These group-level filtered variant distributions show a substantial number of variants which are present in both groups, but only filtered from one group (**green**). This could lead to issues of bias between the two groups in the lists of mutated genes retained, particularly in those variants filtered for low read depth. Furthermore, due to the difference in sample size between the two groups, it will be difficult to pick a relaxed threshold for the filters that would be consistent between the two groups. For these reasons, filtering at the cohort level is more appropriate. The downside is that this approach will be slightly more conservative, since it eliminates the green variants from both groups.

Looking then at the distributions of filtered variants when doing the secondary filtering at the Cohort level, it is possible to distinguish those variants present in both groups from those present in only one. In this case, those variants only filtered from one group may be of interest since we are interested in comparing the two resistance groups, and variants which were present in only one group may be of particular interest. In the distribution of variants filtered for missing genotypes at the cohort level, shown in **Figure 18**, the majority of variants are present in both groups, but there are quite a few present in only one group(shown in **green** or **blue**). In the distribution of variants filtered for low read depth at the cohort level, shown in **Figure 19**, the majority of variants are present in both groups, but there are quite a few present in only one group(shown in **green** or **blue**).

Looking at both the distributions in **Figure 18** and **Figure 19**, it is clear that the number of variants filtered out for each value of the filter cutoff decreases as the cutoff is raised (loosened). This indicates that filtering with the strict **MG/RD** cutoffs of zero, which

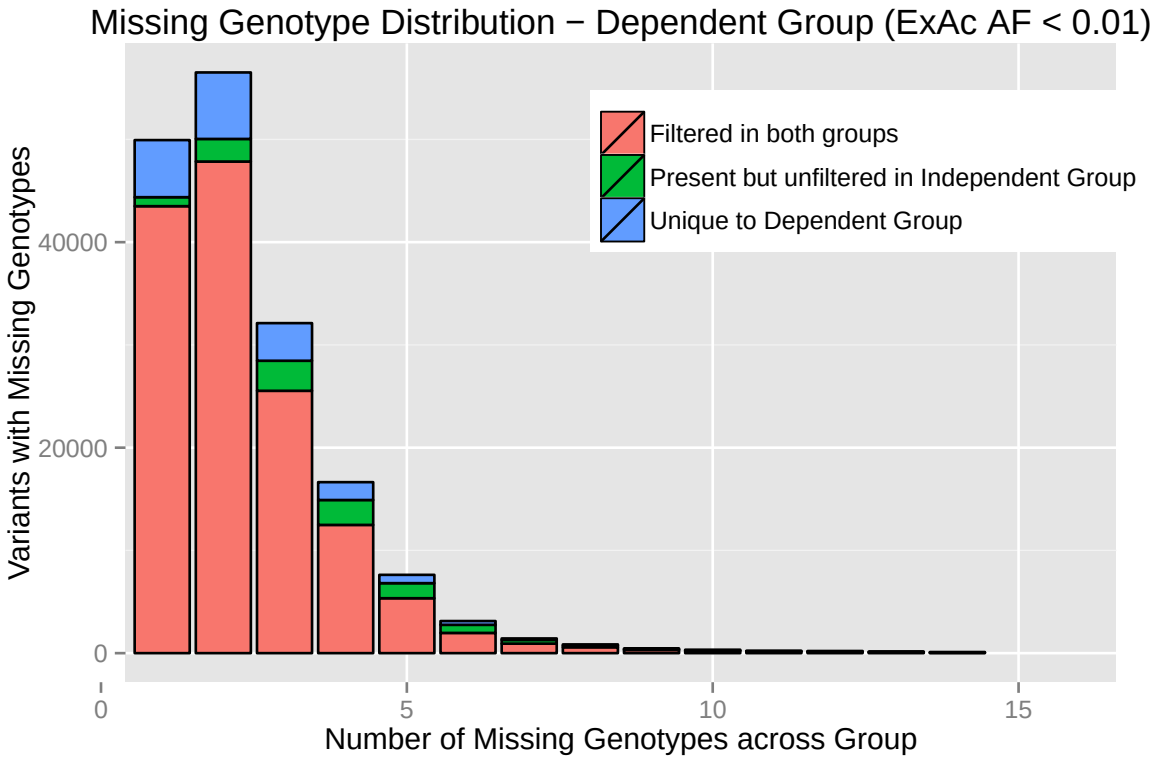

Figure 16: MG Filter Distribution-Dependent Group: This shows the number of variants filtered out using an MG filter cutoff of zero, binned by how many missing genotypes each variant has across the Group. Resistance group membership is shown using bar color; variants filtered out from both groups (red), variants present but unfiltered in the opposite group (green) and variants unique to the group(blue).

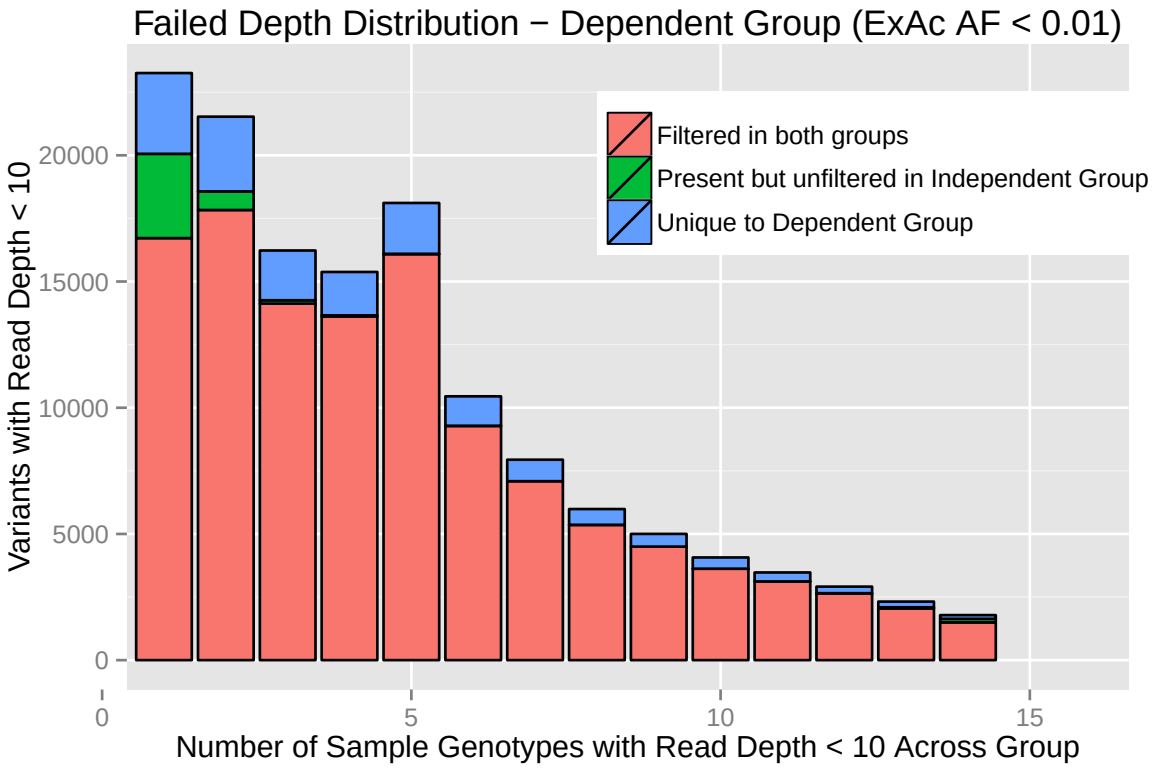

Figure 17: RD Filter Distribution-Dependent Group: This shows the number of variants filtered out using an RD filter cutoff of zero, binned by how many sample genotypes each variant has with read depth < 10 across the Group. Resistance group membership is shown using bar color; variants filtered out from both groups (red), variants present but unfiltered in the opposite group (green) and variants unique to the group(blue).

removes a total of 450,401 variants, filters out a very large number of variants that would pass a somewhat loosened criteria. Furthermore, while there are a large number of variants present in only one group which are filtered out, the number decreases significantly as the cutoffs are raised, indicating that the downsides to filtering at the cohort level would be minimized at loosened filter cutoffs.

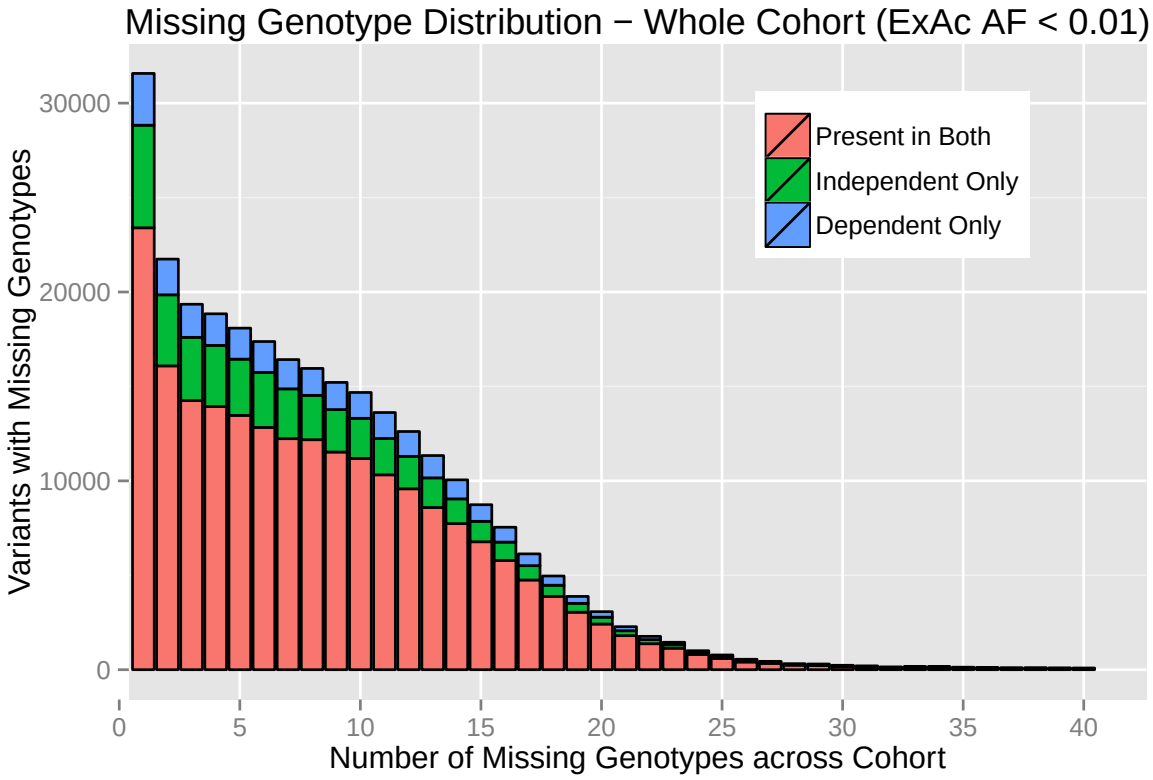

Figure 18: MG Filter Distribution-Whole Cohort: This shows the number of variants filtered out using an MG filter cutoff of zero, binned by how many missing genotypes each variant has across the Cohort. Resistance group membership is shown using bar color; variants filtered out from both groups (red), variants which were only present in the Independent group (green) and variants which were only present in the Dependent group(blue).

Relaxed Filter Cutoffs

With gene-level mutational frequencies (**MF**) used as input for the HotNet2 network analysis, the effect of relaxing the MG and RD filter thresholds should be evaluated based on the change to the resulting number of mutated genes and their respective MFs. In order to evaluate the gene-level impact of using **MG/RD** thresholds with varying strictness, the

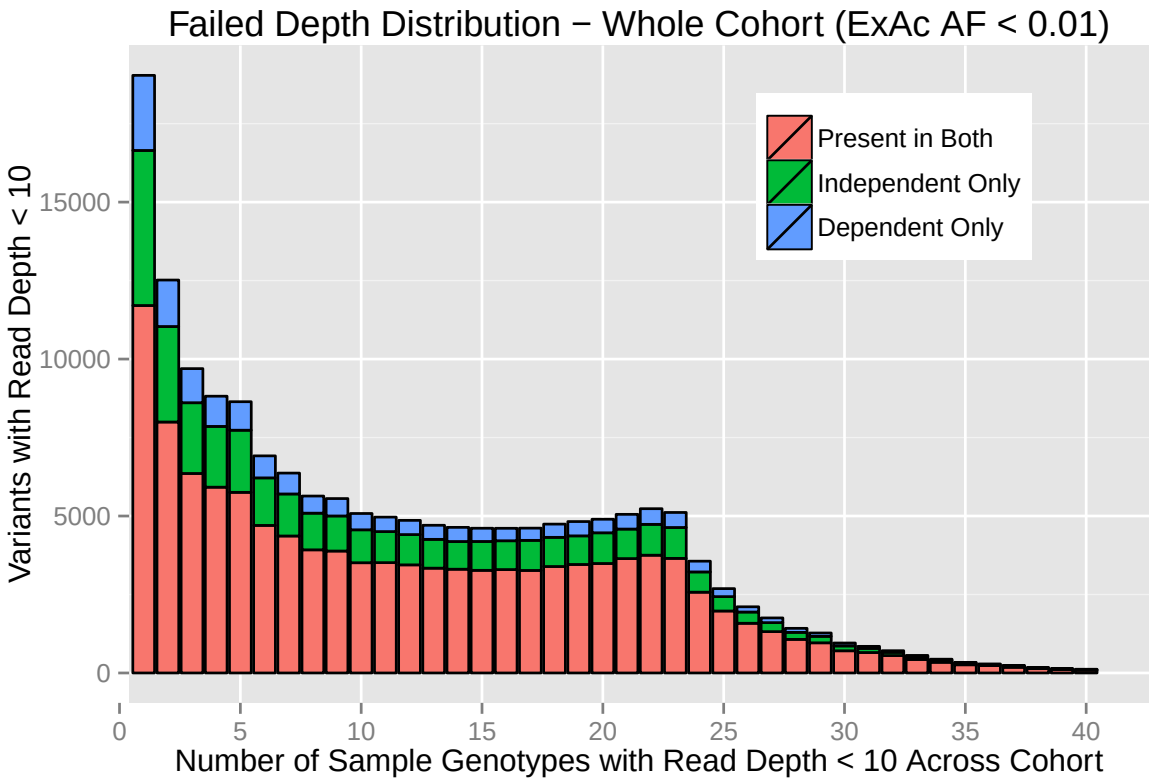

Figure 19: RD Filter Distribution-Whole Cohort: This shows the number of variants filtered out using an RD filter cutoff of zero, binned by how many sample genotypes each variant has with read depth < 10 across the Cohort. Resistance group membership is shown using bar color; variants filtered out from both groups (red), variants which were only present in the Independent group (green) and variants which were only present in the Dependent group(blue).

change in the number of retained genes when varying the threshold from 1-41 (with **RD** threshold varied from 1-41 as well), when compared with the most strict **MG** threshold value of zero missing genotypes was calculated.

The two changes of interest are the overall number of genes retained, and the number of genes with altered mutational frequency. The total number of genes retained and the total number of genes with changed MFs, both in all VEP categories (**Figure 20**) and those annotated to the first two classes of VEP consequence categories (**Figure 21**), are shown below.

It is clear looking at **Figure 20** that the overall number of mutated genes retained (~19-25K), if all VEP categories are included, is considerably higher than the number of genes used in the HotNet2 paper (~12K).

Looking at the total number of mutated genes and the number of genes with changed MF for variants only within our top two designated VEP classes, shown in **Figure 21**, the range of total mutated genes(~7-9K) is within a more reasonable range for the HotNet2 algorithm.

Due to the way in which gene-level mutational frequency is calculated, as the missing genotype filter threshold is relaxed and more variants are retained, the mutational frequency for genes already present will only increase. Given this fact, the mutational frequencies of genes with variants filtered out using the missing genotype or read depth criteria are potentially underestimating the MF.

Under that assumption, relaxing the **MG/RD** filter cutoffs would potentially improve the accuracy of our MF estimates for genes already retained when using **MG** and **RD** cutoffs of zero. However, the other effect of loosening the cutoffs is the inclusion of new genes not present using the strict threshold. While many of these genes may be important, once the **MG** threshold is lowered considerably this means that the number of valid genotypes in the total Cohort becomes small and any mutational frequency calculation becomes less

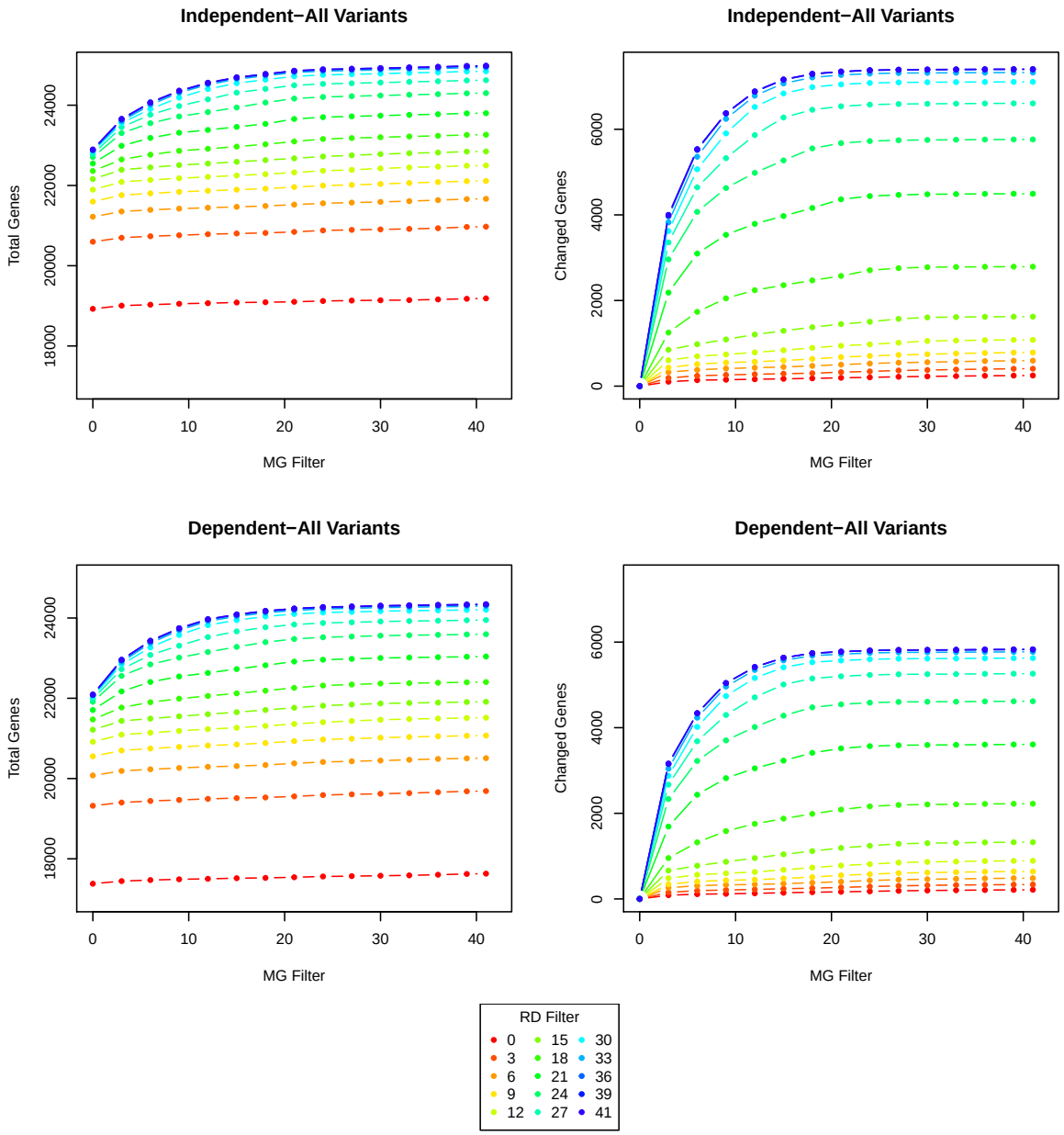

Figure 20: Mutated Genes or Different Filter Cutoff Combinations (All VEP Categories) - This shows the total number of genes retained after filtering, and the number of genes present for RD=0 with changed mutational frequency when the MG/RD cutoffs are varied.

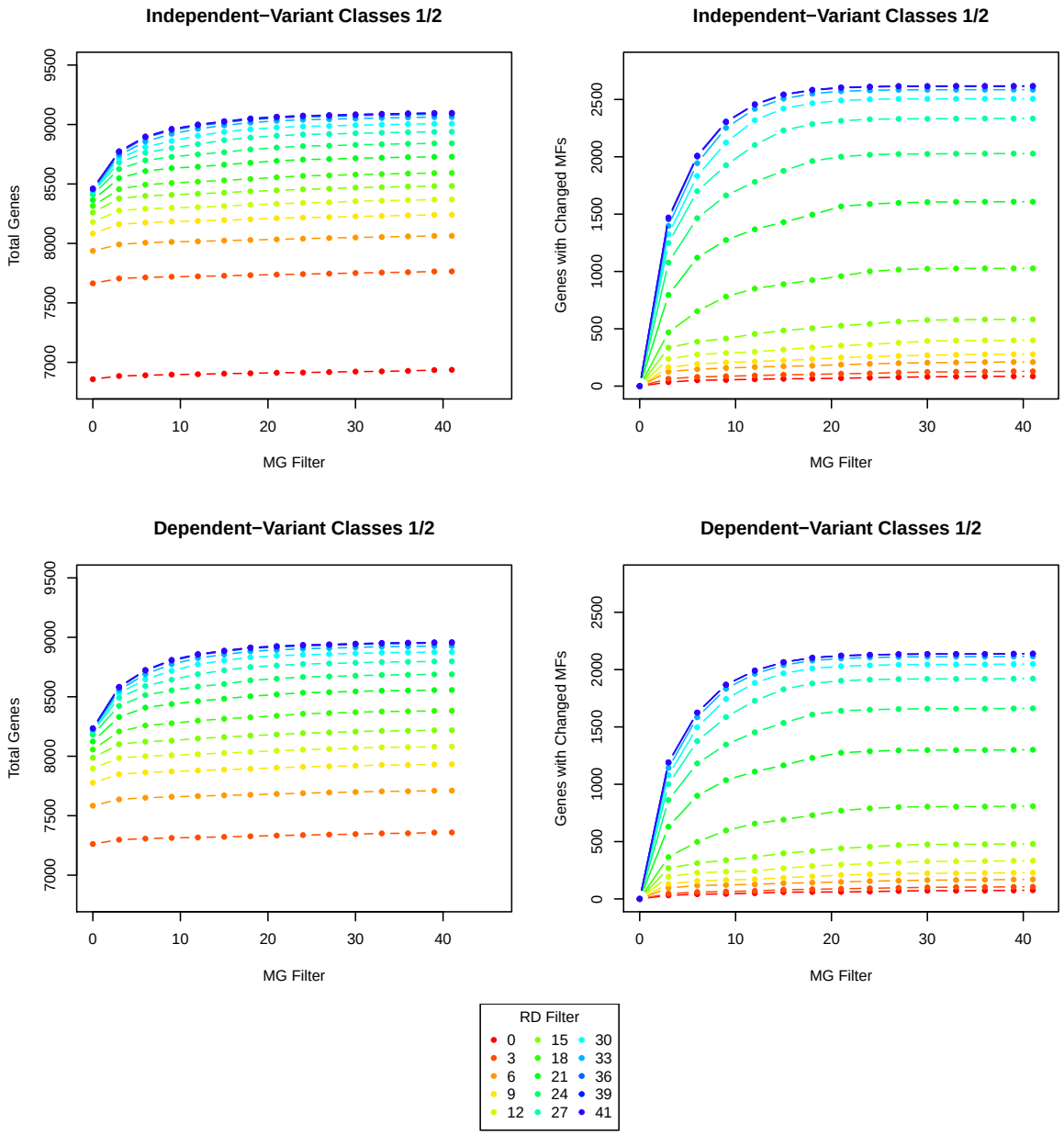

Figure 21: Mutated Genes or Different Filter Cutoff Combinations (Variant Classes 1/2) - This shows the total number of genes retained after filtering, and the number of genes present for RD=0 with changed mutational frequency when the MG/RD cutoffs are varied.

meaningful. Therefore it seems that an **MG** threshold that improves the estimates of existing gene MFs and increases the number of genes retained, without relaxing so far that the MFs of new genes added are too imprecise, is desirable. Looking at **Figure 20** and **Figure 21**, there is a clear saturation point around **MG** threshold of 20 at which almost no more genes have changed MFs, and very few new genes are added. Based on this data an **MG** threshold of 20 seems a good choice for achieving the desired tradeoff.

Furthermore, looking at the combined effect of varying the two filters, it is clear that at more strict (ie lower) RD filter cutoffs, the changes due to varying the MG filter is minimized, with an almost flat relationship between total genes and MD cutoff at RD cutoff of zero. One explanation for this is a high degree of overlap between the variants filtered by each criteria seperately. In fact, if the variants are filtered under two cases, one with MG:0/RD:41 and the other with MG:41/RD:0, in order to get all the variants filtered out at the most strict cutoff by each filter, we find that of the 281652 variants filtered out for MG of zero, 278977(99.05%) are also filtered out by the RD filter if no MG filter is applied. This indicates that while there is not a clear relationship between the number of missing genotypes and mean read depth in the remaining samples, there is a more general relationship between the presence of missing genotypes and the presence of sample genotypes with low read depth. This means that if the MG filter cutoff is relaxed, most of the variants with missing genotypes retained will be filtered out by the RD filter. Thus the full distribution of variants filtered out for different **RD** cutoffs and the corresponding mean read depth amongst the remaining samples is more accurately shown in **Figure 22**.

Given this, evaluation of a suitable relaxed RD filter cutoff should be done with the proposed MG cutoff of 20, shown in **Figure 23**.

Looking at the distribution closely, one can see that at an **RD** threshold of 30, the number of outlier variants with mean read depth (among the remaining sample genotypes) above 30 drops off considerably. This therefore seems like a good alternative filtering threshold for the **RD** filter.

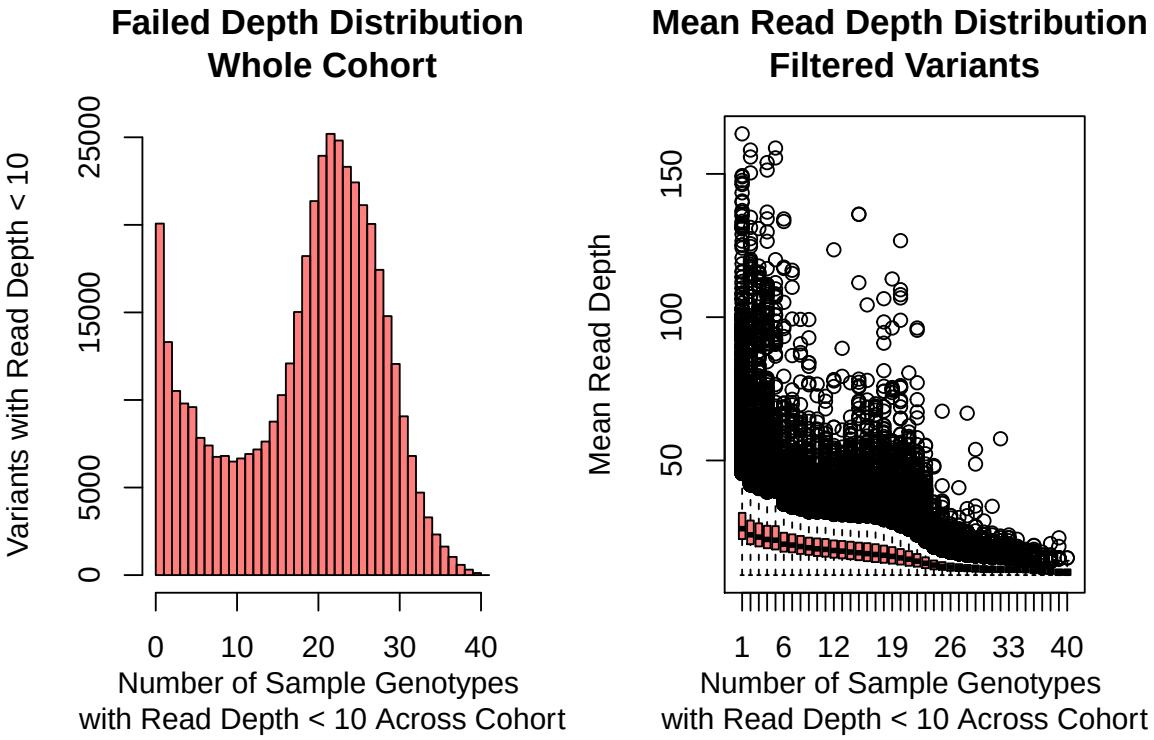

Figure 22: Read Depth Filter and Mean Read Depth Distribution (MG=41, RD=0) - Left Panel: The number of variants filtered out using an RD filter cutoff of zero, binned by how many sample genotypes each variant has with read depth < 10 across the Cohort. Right Panel: Boxplots of the mean read depths for the variants are shown, for each bin of variants shown in left panel figure.

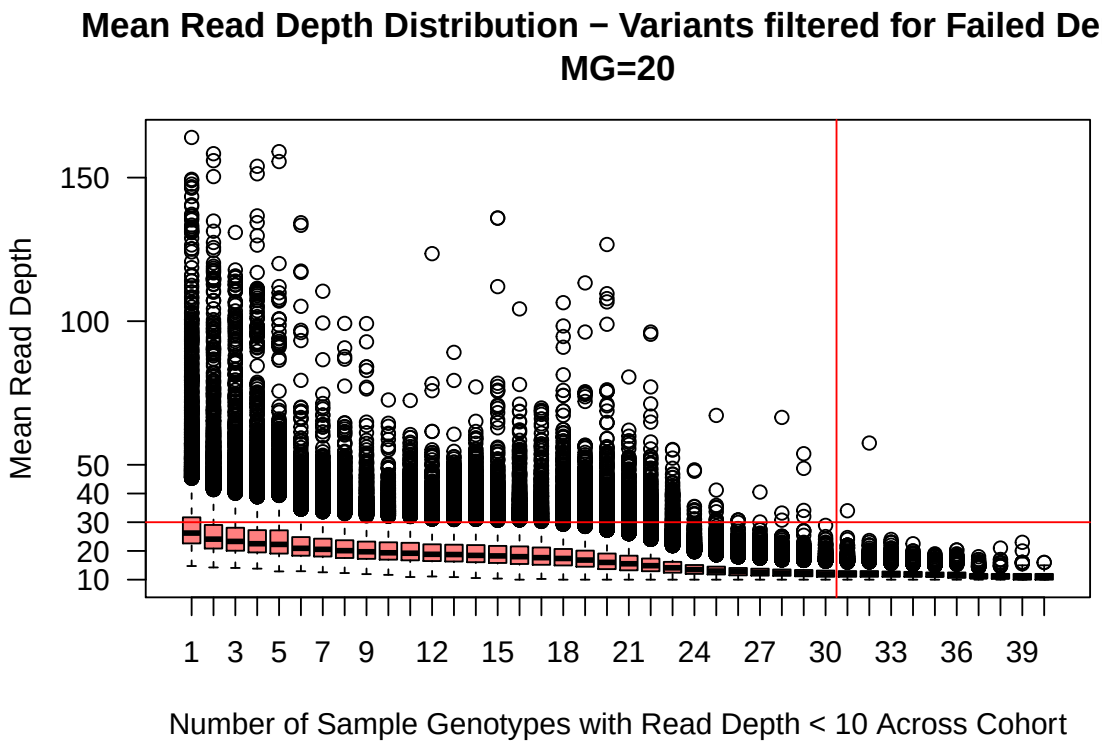

Figure 23: Mean Read Depth Distribution (MG=20, RD=0) - Boxplots of the variant mean read depths, for variants binned by number of sample genotypes with read depth < 10 across the Cohort. Cutoffs for mean read depth of 30 and RD cutoff of 25 shown in red.

3.2.3 Final Filtering Results

Bias in Filtered Variants

Utilizing the relaxed secondary filter cutoffs or **MG**=20 and **RD**=30, one final point of concern is whether filtering in this way biases our analysis by disproportionately filtering out variants which are only present in one BCR-ABL Imatinib resistance group. In order to address this, I plotted the distributions of variants filtered for the **MG/RD** filters, shown below.

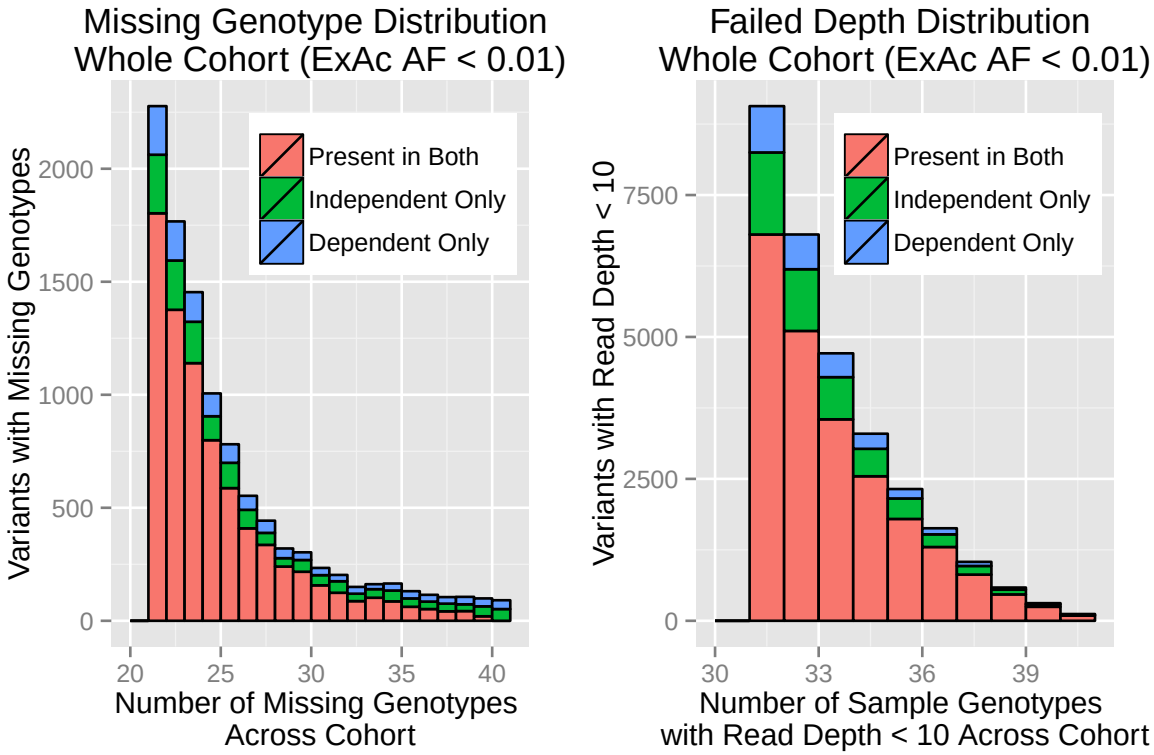

Figure 24: Missing Genotype Filter and Read Depth Filter Distributions (MG=20, RD=30) Left Panel: The number of variants filtered out using an MG filter cutoff of zero, binned by how many missing genotypes each variant has across the Cohort. Right Panel: The number of variants filtered out using an RD filter cutoff of zero, binned by how many sample genotypes each variant has with read depth < 10 across the Cohort.

The total number of variants filtered within each category, either both groups or individually, is shown in **Figure 25** below.

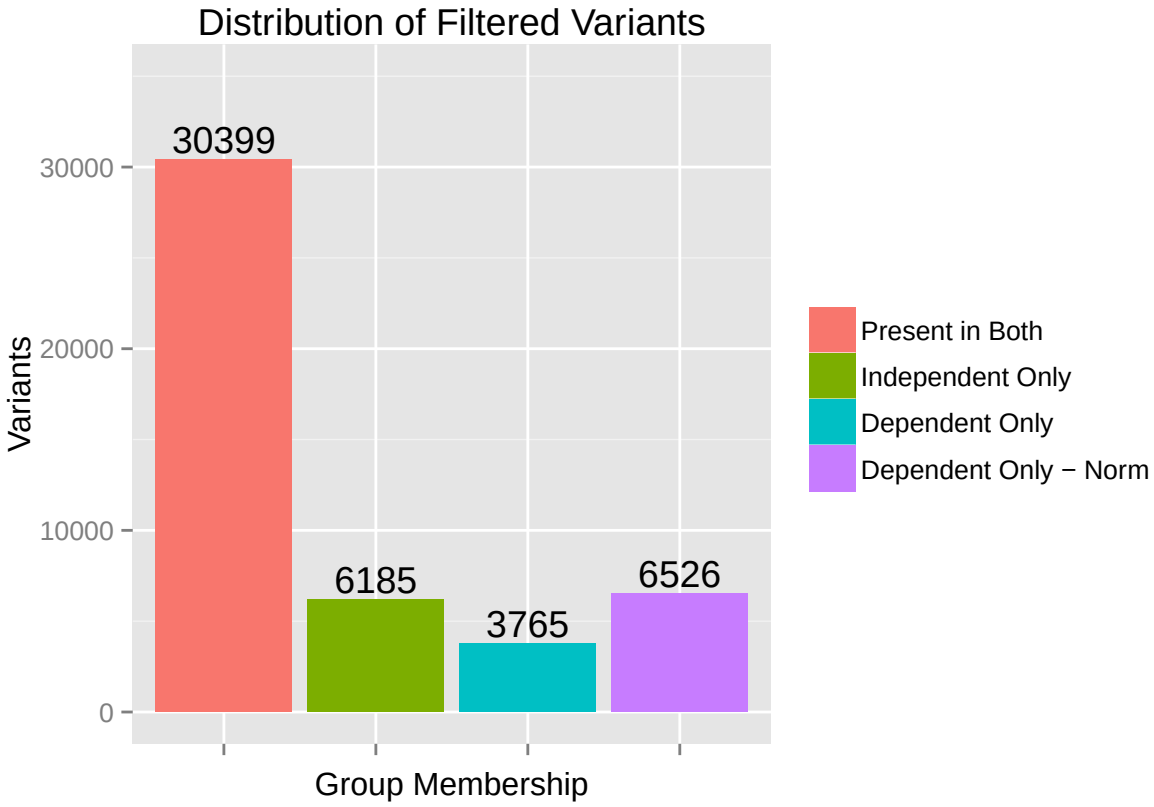

Figure 25: Final Filtered Variants BCR-ABL Resistance Groups - Total numbers of variants filtered out using filter cutoffs of MG=20 and RD=30, subsetted by resistance group membership; variants filtered out from both groups (red), variants present in only the Independent Group (green), present only in the Dependent Group (blue). The variants present only in the Dependent Group, normalized for sample size, are shown in Purple.

Looking at **Figure 25**, it is clear that the number of filtered variants present only in the Independent or Dependent groups are proportional and nearly the same when the number of variants in the Dependent group is normalized to account for the difference in sample size (n=26 vs n=15). This shows that there is no discernible bias between the resistance groups in the variants filtered out using these criteria.

Gene-Level Totals

With the filtering choices of **MG**=20 and **RD**=30, the total number of mutated genes retained are shown in **Table 4** below.

Table 4: Gene Totals after Filtering

BCR-ABL Independent	
Total Genes	22232
Top 2 VEP Classes	9356
Present in COSMIC	2436
BCR-ABL Dependent	
Total Genes	21892
Top 2 VEP Classes	7172
Present in COSMIC	1841

3.3 Conclusion

Lacking matched normal samples, reducing the number of variants to a more manageable number and attempting to eliminate as many germline variants as possible is an inherently difficult task. Filtering out variants based on allele frequency in the ExAC database was the best approach for removing potential common germline variants, even though some somatic variants were no doubt inadvertantly filtered out as well. Aside from this, cutoffs for the

number of samples within the sample cohort either missing a valid genotype call or lacking sufficient read depth were chosen in order to strike a balance between inclusion of as many variants/genes as possible and ensuring that the resulting gene-level mutational frequencies were of reasonable quality. Finally, the variants were filtered two different ways based on functional and phenotypic annotation, allowing for variants from either only the top two designated tiers of VEP functional classifications, or from all VEP classes but restricted to only those variants present in the COSMIC database of known somatic cancer variants.

While the choices for filtering made here are necessarily ad hoc, they highlight the importance of considering the context of how the data (genotypes) will be used in the proposed analysis (cohort-wide measures, 2 group) in order to carefully make decisions about how to select variants. Many of the quality metrics which are built into standard pipelines such as the GATK genotyping pipeline used here, are evaluated on a per-sample basis. In the case of genotyping data, a single sample passing all the quality metrics with a non-WT genotype at a particular locus is enough for that variant to be included in the dataset. If however, the final value of interest is measured across the cohort or group, as in gene-level mutational frequency, it needs to be ensured that all samples meet these standards. In doing so, one needs to consider whether using a strict cutoff will filter out too many variants or bias the variants filtered out between groups if the cohort is being split into several analysis groups, as in this case. These factors highlight the complexity of making rational choices for filtering, and the need for careful consideration of how those filtering choices will map onto the proposed study design.

4 HotNet Functional Network Analysis

4.1 Introduction

The advent of massively parallel next-generation sequencing (NGS) technologies has radically lowered the cost of speculative sequencing projects, leading to a huge expansion in cancer genome characterization. With the ability to characterize whole genomes or exomes across a cohort of patients, it is now possible to identify somatic variants important to cancers in a de novo fashion, in contrast to prior approaches with microarray or targeted sequencing. NGS approaches are accompanied with their own sets of analytic challenges, however. One limitation for using traditional biostatistical methods on these datasets is the disproportionate ratio between the number of parameters or variables being investigated (commonly genes or transcripts) and the number of samples available. This P>>n inversion requires great care for devising tests of statistical significance and avoiding issues of multiple testing. Another common pitfall is that of correlation, between genes or pathways. For example, many competitive gene set tests such as the widely used Gene Set Enrichment Analysis (GSEA) algorithm, which identifies pre-defined sets of genes with significant differences in gene expression between two disease states, assume independence of genes[47]. These approaches require resampling procedures to adjust significance, which are not appropriate when applied to studies with small sample numbers[48].

Most cancers are ultimately the result of somatic mutations acquired over the course of a lifetime, either through replication errors or as a result of environmental and other epigenetic factors. This makes it difficult to distinguish genetic mutations with an important role in the etiology of a particular cancer, i.e. “driver mutations”, from the numerous benign somatic mutations acquired concurrently. Additionally the acquisition of “passenger mutations”, those arising after the onset of cancer as a byproduct of the increased mutational rate seen in unregulated cell division and proliferation, further muddies the water. Instead of a few important genes mutated at high frequency, most cancers exhibit a high degree of

mutational heterogeneity, with numerous genes mutated at low frequency[49]. This so-called "long-tail" phenomenon poses a serious challenge to identification of important "driver genes", which may be perturbed through myriad disjoint mutations when compared across a cohort of patients. Methods that ignore frequency of specific mutations in favor of identifying the parsimonious collection of genes perturbed by these mutations offer a better approach. One such approach is to identify groups of genes that are significantly mutated across patient samples when evaluated as a group rather than individually. A method utilizing this approach, the HotNet2 algorithm[50], was used in this study. HotNet2 identifies significantly mutated groups of genes with related downstream function by creating a network of variants. Network nodes represent the variants and edges representing putative interactions identified by a pre-specified protein interaction network. Gene scores for weighting of the nodes can either use across sample variant frequency or other metrics of significance. A heat diffusion algorithm then iteratively transfers "heat" between the nodes. Groups of genes that have both high frequency of mutations and/or interacting protein products will form highly significant gene subnetworks, which may represent groups of genes acting in a common regulatory pathway perturbed in the cancer. The extrapolation of a gene set testing approach onto a protein interaction network adds an additional dimension that is typically absent in curated gene set lists. The representation as a network allows for methods of discovering important sets or "subnetworks", which avoid the multiple testing issues that plague a naïve gene set testing approach.

4.2 Results and Discussion

4.2.1 Heat Diffusion Parameter β

The B heat diffusion parameter was determined for each of the three protein-protein interaction (PPI) networks (STRING, iRefindex, and consensusPathDB) using the procedure described in the **Methods: B Determination** section. The maximum influence cutoff β value was determined for 100 vertice genes for each PPI, and the distribution of all 100 β

values is shown in **Figure 26** below.

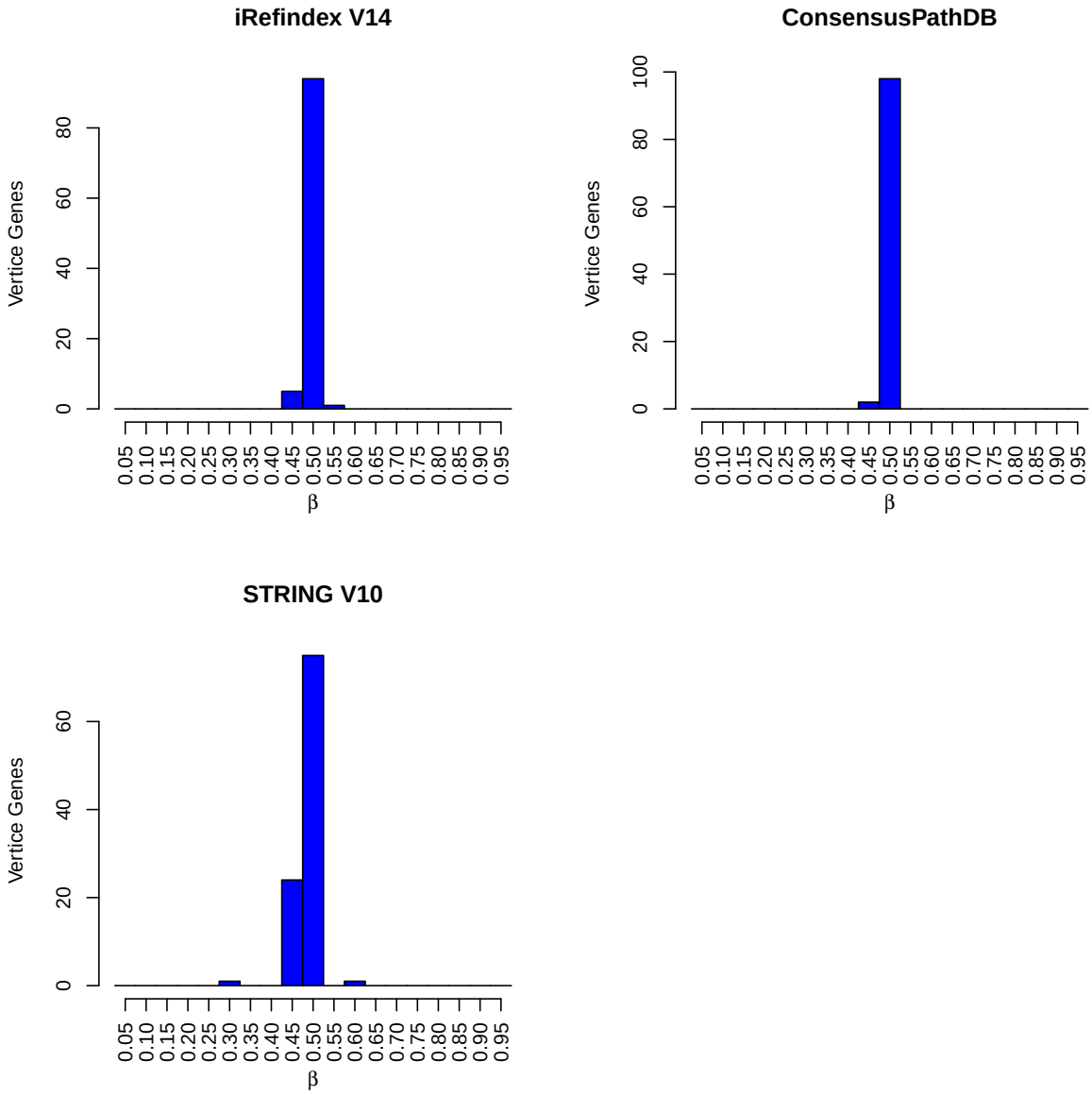

Figure 26: Distribution of Inflection Point Maximum Influence Cutoff Beta Values - The maximum inflection points for the selected vertice genes were predominately 0.50, with a small number at 0.45.

4.2.2 Minimum Edge Weight Parameter δ

The ideal δ edge weight parameter value was determined for each protein-protein interaction network (PPI), by calculating delta values which are unlikely to return large connected subnetworks when running our dataset on permuted versions of the PPIs, as described in the

Methods: Delta Determination section. The selected delta values for all the HotNet runs is shown below in **Table 5**

Table 5: Delta edge weight parameter values chosen

	AllVars	COSMIC
BCR-ABL Independent		
iRefindex	0.00637	0.00216
STRING	0.00508	0.00152
consensusPath	0.01065	0.00216
BCR-ABL Dependent		
iRefindex	0.00612	0.00200
STRING	0.00526	0.00149
consensusPath	None	0.00212

4.2.3 Significant Subnetworks

Subnetworks of mutated genes were identified using the HotNet2 algorithm, separately for the BCR/ABL-independent and BCR/ABL-dependent groups, for each of the three pre-selected protein-protein interaction networks (STRING, iRefindex, and consensusPathDB). This was repeated on the smaller variant subset identified as present in the COSMIC database, for a total of 12 HotNet2 runs. All significant subnetworks are listed in **Supplementary Tables 1-4**. Overlap between significant subnetworks for HotNet2 runs on different PPIs were identified using the procedure described in **Methods: Cross-PPI Consensus Identification**.

There are numerous genes identified in these results which exhibit pathway memberships and functional annotations rich in cancer-related hallmarks informative to our context. Out of the total subnetworks identified, eight of the significant subnetworks for the independent group

are highlighted here, six of which comprise two larger overlap networks. The highlighted independent subnetworks are shown in **Table 6** and **Table 7**.

Table 6: Significant Subnetworks of Note - All Variants. Genes discussed in text highlighted in red.

	PPI	Genes	Pathways	pval
Net1	iRefindex	CYP3A5 DHRS7 *ETV5* FAIM3 HOXD9	ERK RhoA FGF8 IL12/Stat4 FoxM1	0.01
Net2	STRING	AFF1 *ASXL2* ATP6V0A4 ATP6V1D ATP6V1E1 ATP6V1H BPTF CASP5 CASZ1 CBFA2T2 *CCNT2* CD34 CDHR1 CECR2 DAP DMXL1 EIF4G2 GPR87 HOXD9 *KAT6B* MAP1S MEIS3 MLLT1 MLLT3 *MST1* PPA2 PROM1 *RASSF1* RHCG *TET2* TM4SF5 TSHZ2 VAX1 ZNF462	p53 SMAD2/3/4 Wnt FoxO Integrin RhoA	0.00
Net3	iRefindex	AFF1 *CCNT2* CHD1 MLLT1 MLLT3 RFX5	p53 SMAD2/3/4	0.02
Net4	STRING	*ADAMTS9* ATAD2 DARS DARS2 DCLK2 DSPP FBLN7 KARS KCNK1 *NFIC* PLK5 RBMS1 RCN1 RFX1 RFX5 RRBP1 SENP1 SENP7 *SP100* SPATA4 SULT1A1 THADA TSPAN8 UBA2 *USP34* WARS	wnt FOXA1 o-glycosylation SUMO/SUMOylation Interferon-gamma	0.01
Net5	iRefindex	GABRB2 PCSK9 RCN1 *RHOT1* *RHOT2* TRAK1 TRAK2	Rho GTPase	0.04
Net6	STRING	*B3GNT3* BCAM FUT2 FUT3 *GALNT1* ISM2 MSLN MUC12 MUC16 MUC17 MUC20 MUC21 *MUC4* MUC5B MUC6 MUC7 PRDM16 *SETBP1* UCP1	O-linked glycosylation	0.00

Table 7: Significant Subnetworks of Note - COSMIC Variants. Genes discussed in text highlighted in red.

	PPI	Genes	Pathways	pval
Cos1	STRING	BRD2 *CARD11* DCHS2 DKC1 DNAH17 DYNC1LI2 EP400 ERAP1 FBXO2 FYB GPAM GPD2 HLA-A HLA-B HLA-C HLA-DQB1 HLA-DRB1 HLA-DRB5 IRF2 JUP RUVBL1 SH2B3 SIRPA	Interferon alpha/beta/gamma TCR signalling EGFR1 IL12/STAT4 Integrin c-myc beta-catenin/TCF complex IL1/IL6 BCR NF-kappaB IKK Wnt	0.02
Cos2	consPath	ATG12 ATN1 AUTS2 *CARD11* CEP170B CNOT1 CTBP2 DDX20 EHMT1 EPSTI1 ETV3 FCGBP FYB GIGYF2 MIER2 NEB NOL4 PCGF6 POTEF PRDM6 PRRC2A PSPH RB1CC1 RERE RING1 RPS29 SAFB2 SRA1 STYXL1 TFAP2A TGIF1 TNRC6B TNXB TRIM39 TRIM5 ULK2 USP2 WBP11 WBP2NL ZBTB33	mTOR p53 PDGF Integrin signaling SUMOylation MAPK6/4 PIP3 AKT signaling Wnt Interferon NF-kappaB TNFAlpha SMAD2/3/4 TGF-beta EGFR1	0.02
Cos3	STRING	ACAN *ADAMTS12* *ADAMTS7* *B3GNT3* COMP FUT3 *GALNT1* MSLN MUC12 MUC16 MUC17 MUC20 MUC21 *MUC4* MUC5B MUC6 MUC7	o-glycosylation Integrin signalling	0.02

The first of the multimeric overlap networks, comprised of subnetworks Net1-5 in **Table 6**, is shown in **Figure 28**.

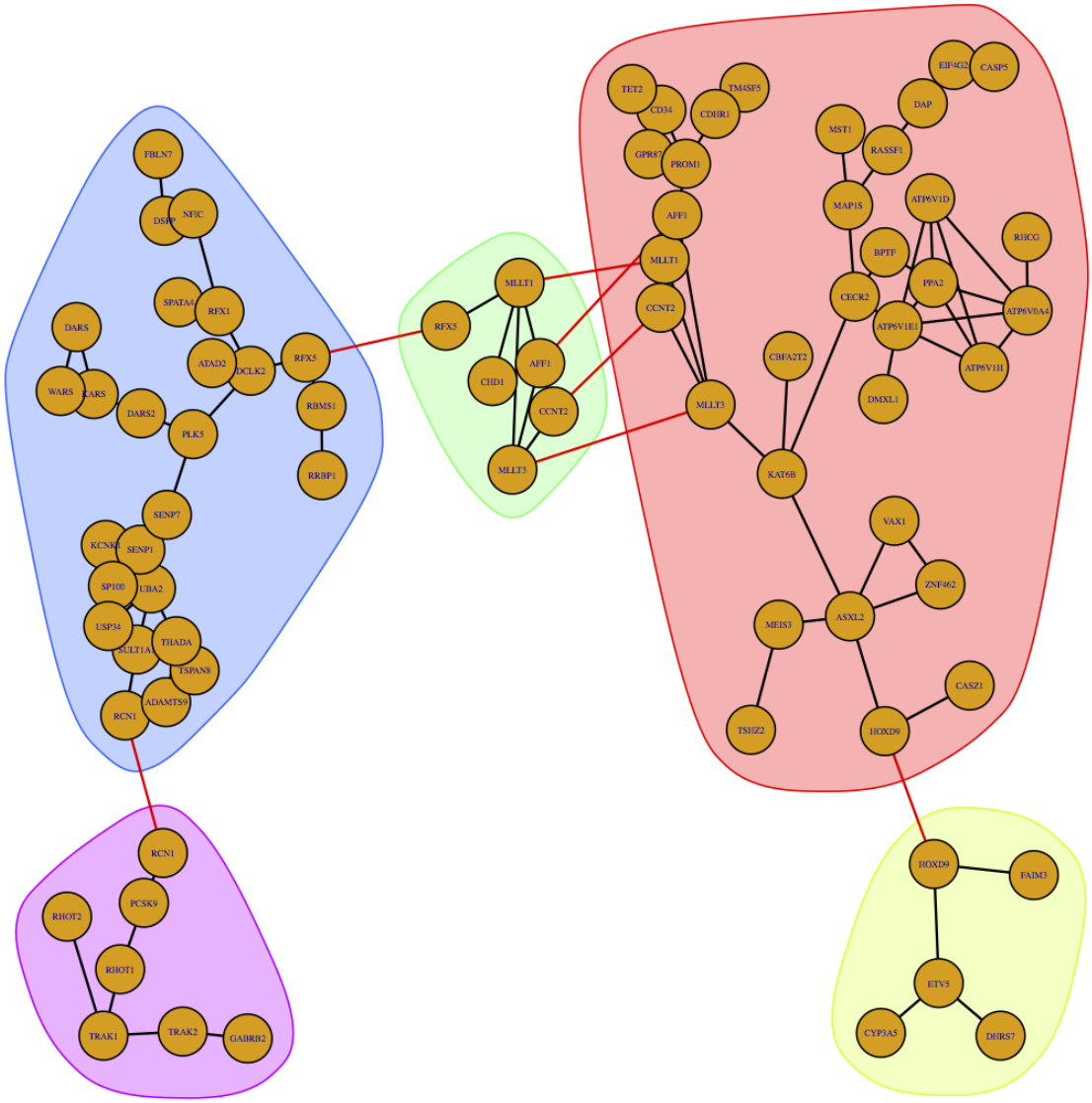

Figure 28: Master Network 1 - Subnetworks with overlapping gene membership. iRefindex subnetwork Net1 shown in gold. STRING subnetwork Net2 shown in red. iRefindex subnetwork Net3 shown in green. STRING subnetwork Net4 shown in blue. iRefindex subnetwork Net5 shown in purple. Red graph edges denote overlap.

Of particular note in Net1 is the ETV5 gene, which encodes a transcription factor involved in downregulation of the MEK/ERK pathway central to CML progression[51]. It is also involved in regulation of the RhoA, FGF8, IL12, Stat4, and FoxM1 signalling pathways, all of which are involved in cancer regulation/progression. RhoA and the STAT family of transcription factors in particular also play important known roles in CML[52,53].

Net2 contained several genes with relevant known molecular roles. RASSF1, KAT6B and CCNT2 are all members of the p53 pathway. RASSF1 is a tumor suppressor that inhibits RAS[54]. Silencing by methylation of the RASSF1A isoform has been linked to tumorigenesis in numerous cancers[55]. MST1 has links to Integrin, RhoA, and FoxO signaling. FoxO in particular acts as a hub for many other signaling pathways, including the aforementioned RAS/RAF/MEK/ERK pathway. In fact, the MST1 kinase is a downstream target of the RASSF1A isoform, after RASSF1A's interaction with RAS, stimulating the apoptotic cascade[55]. Furthermore, mutations involving KAT6B, MST1, and CCNT2 have all been shown to play roles in acute myeloid leukemia (AML). CCNT2 has been shown to act as an inhibitor of myeloid differentiation in AML, via post-transcriptional targeting by the miR-29 family of micro-RNAs[56]. TET2 is a dioxegynase involved in DNA demethylation and acts as a tumor suppresor. It is commonly mutated in myeloid malignancies such as AML and chronic myelomonocytic leukemia (CMML). TET2 has also been shown in previous studies to be preferentially mutated in BCR/ABL-independent imatinib resistant cases of CML[57,58]. In this study both TET2 and RASSF1 exhibited markedly higher mutational frequencies in the BCR/ABL-independent group compared with the BCR/ABL-dependent group (TET2 - Ind: 6/26 Dep: 1/15, RASSF1 - Ind: 18/26 Dep: 5/15). Net3 also contains the CCNT2 gene.

Net4 also contained a number of interesting genes, USP34, SP100, NFIC, and ADAMTS9. USP34 plays a role in regulation of the Wnt/β-catenin signaling pathway, through activation of the β-catenin destruction complex. SP100 is a nuclear antigen which acts as a tumor suppresor and activator of the the Interferon-γ signaling pathway[59]. Interferon-γ induces phosphorylation and activation of the JAK/STAT cascade and has been shown to attenuate TKI sensitivity in CML cells[60]. NFIC is a nuclear transcription factor, a member of the FOXA1 transcription factor network which play an important regulatory role in breast and prostate cancers. ADAMTS9 is a metalloproteinase that acts as a tumor suppressor in a variety of cancers, primarily through epigenetic regulation. In gastric cancer, ADAMTS9 has been shown to inhibit the AKT/mTor pathway[61]. ADAMTS9 has also been shown to act as a tumor suppressor in multiple myeloma cell lines, with cell proliferation directly

linked to ADAMTS9's promoter methylation status[62]. ADAMTS9 exhibited a markedly higher mutational frequency in the Independent group compared with the Dependent group (Ind: 7/26 Dep: 2/15).

Net5 identified RHOT1 and RHOT2, which encode members of the mitochondrial Rho-GTPase family. These GTPases are similar to the RhoA and Ras Rho-GTPases which play a prominent role in CML progression. Rhot1 has been shown to promote proliferation in pancreatic cancer via suppression of SMAD4, which is a key mediator of the TGF-β pathway and Wnt/β-catenin signaling[63].

The second highlighted subnetwork is shown in **Figure 29**.

Net6 identified B3GNT3 and GALNT1. Both are involved in O-linked glycosylation of MUCINS. MUCINS 4-21 are also present in the subnetwork. Several Mucins are known to play roles in many varieties of cancer. Mucin4 has been demonstrated to be a target of TGF-β in pancreatic cancer, and has been linked to the MapKinase/ERK and RAF/ERK pathways in epithelial carcinomas[64]. Overexpression of the related Mucin1 has been found in several myeloid cancers such as multiple myeloma (MM), AML, and blast phase CML[65].

Another gene of particular interest in Net6 is SETBP1. SETBP1 has been identified as an important oncogene in myeloid cancers, but its function remains poorly understood. Consistent mutations in the SETBP1 gene were found in a study of atypical Chronic Myeloid Leukemia (aCML)[66]. This type of CML is particularly interesting because it lacks the BCR/ABL fusion gene and is a logical target for candidate mechanisms of BCR/ABL-independent TKI resistance. SETBP1 mutations have also been found in other leukemias such as chronic myelomonocytic leukemia, secondary acute myeloid leukemia, and juvenile myelomonocytic leukemia. In all of these myeloid cancers, SETBP1 appears to play a role in triggering secondary leukemogenesis, and is usually preceded by the presence of mutations in a small subset of other genes (ASXL1, SRFS2, CBL, RUNX1, TET2)[67]. Of these top co-occuring mutated genes, TET2 was present in Net2. ASXL2, a closely related paralog of

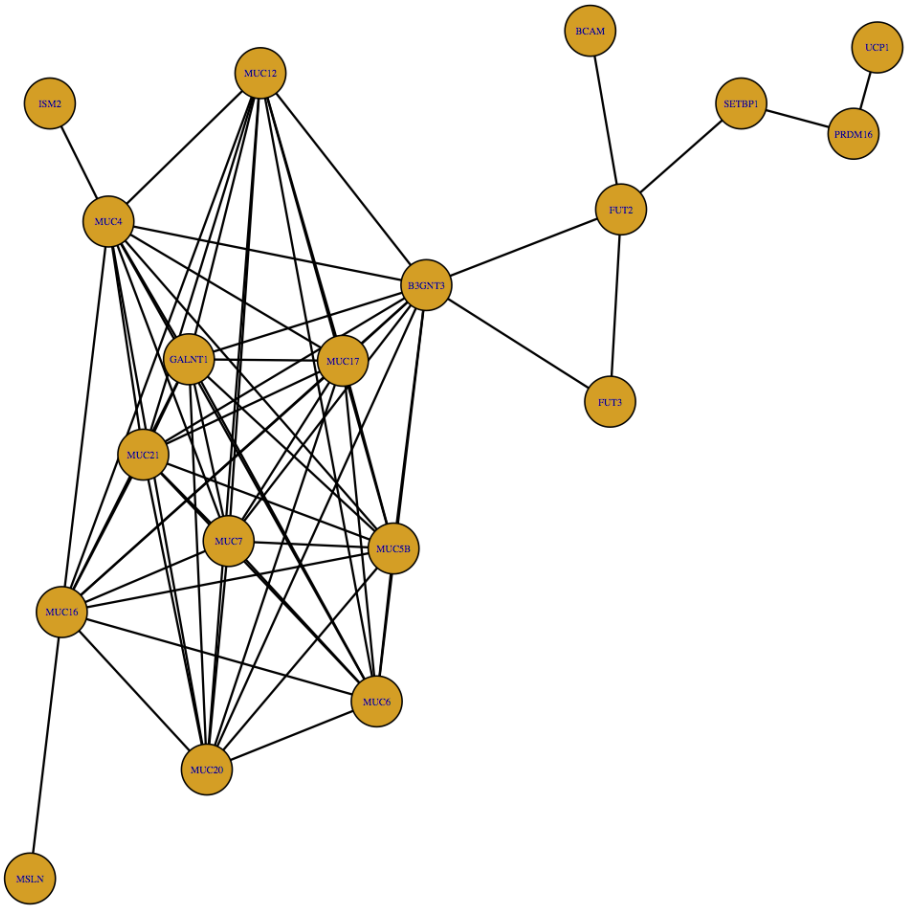

Figure 29: STRING subnetwork(Net6) containing genes with numerous roles in O-linked glycosylation, and known leukemia oncogene SETBP1

ASXL1, was also present in Net2. SETBP1, similarly to TET2, is mutated almost exclusively in the Independent group.

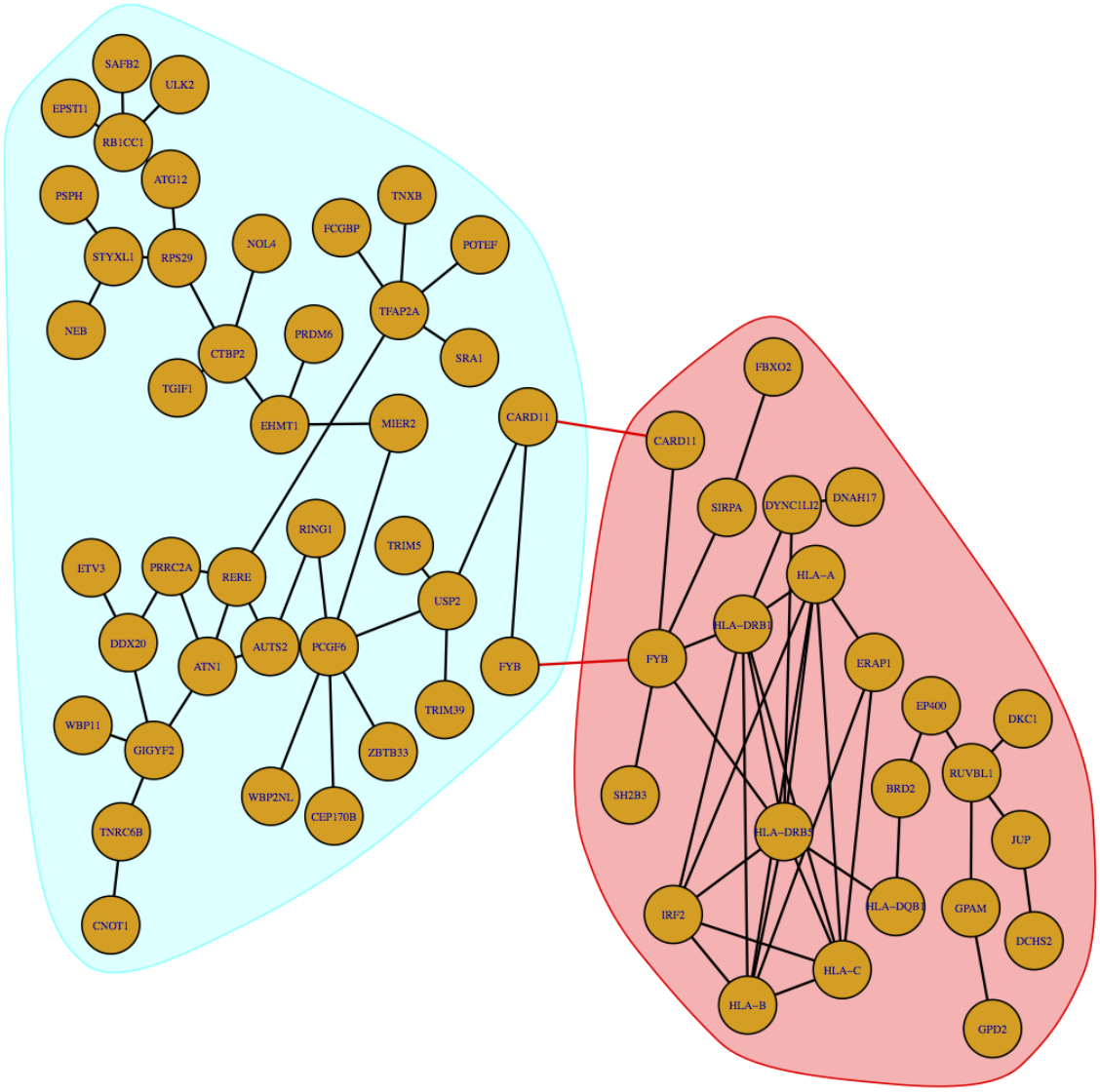

Figure 30: Master Network 2 - COSMIC subnetworks(Cos1-2) with overlapping gene membership. ConsensusPathDB network shown in blue, STRING network in red. Red graph edges denote overlap.

A very large number of the genes in the subnetworks Cos1-2 (**Table 7** and **Figure 30**) play interesting roles in cancer pathways either central to or implicated in CML progression, such as Interferon-$\alpha/\beta/\gamma$, TCR, EGFR, IL/STAT, Wnt, NF-κB, and β-catenin signaling pathways. Interestingly, CARD11 is shared by both subnetworks. CARD11 is an intermediate component of BCR signaling and downstream activation of NF-κB, PI-3 kinase, and the

ERK-MAP kinase cascades, all of which are involved in canonical CML tumorigenesis[68]. CARD11 mutations constitutively activate NF-κB signaling in the activated B cell-like (ABC) subtype of diffuse large B cell lymphomas (DLBCL)[69]. CARD11 has also been shown to act as a mediator of tumor suppresion by NF-κB in CML.

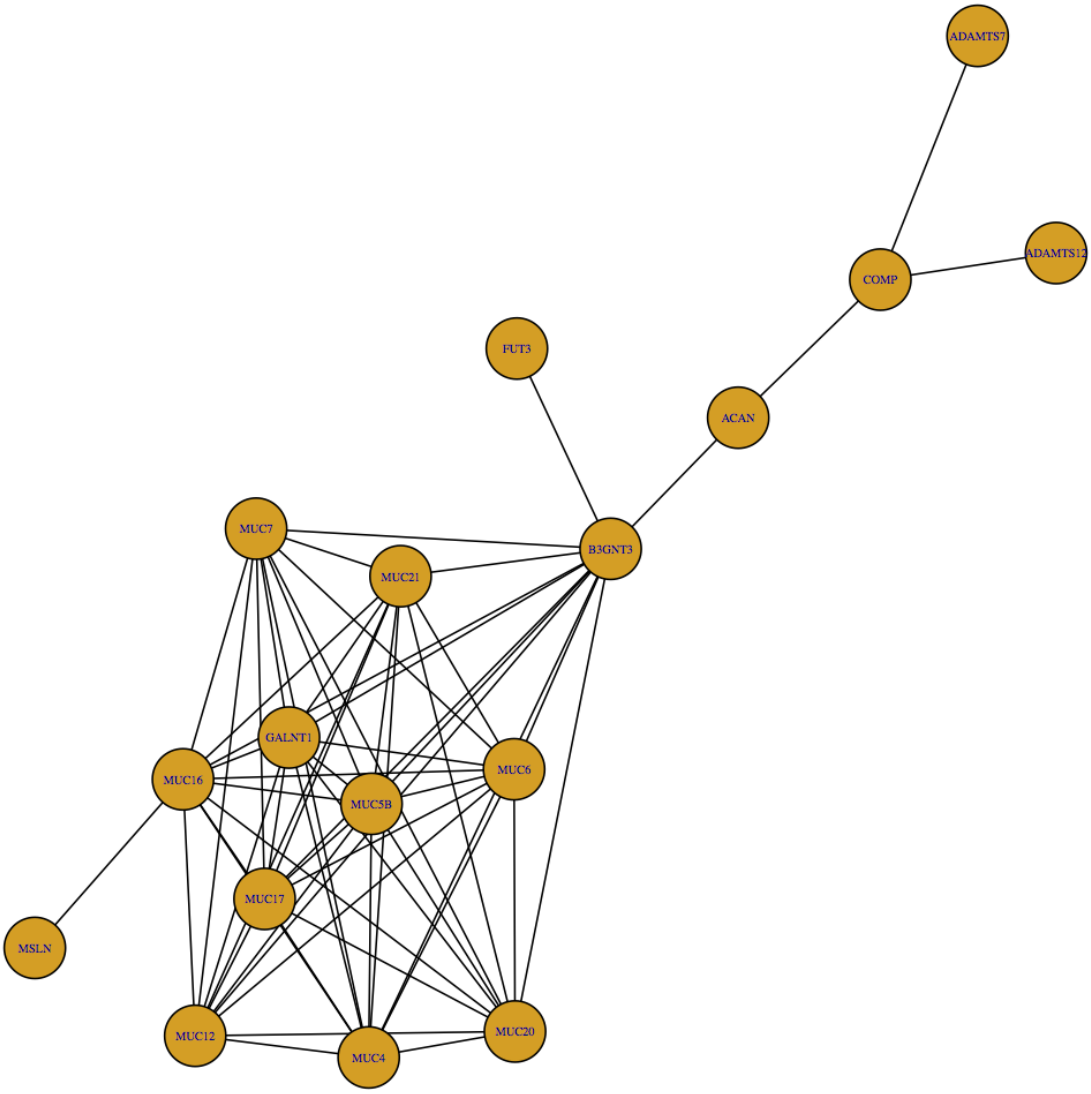

Figure 31: COSMIC STRING subnetwork(Cos3) containing genes with numerous roles in O-linked glycosylation.

Subnetwork Cos3, shown in **Figure 31** contained the genes ADAMTS12, ADAMTS7. These ADAMTS genes encode metalloproteinases which have been implicated in various cancers. ADAMTS12 is involved in colorectal cancer via epigenetic silencing, similarly to ADAMTS9 in multiple myeloma[70]. ADAMTS7 has been found in the urine of prostate and bladder

cancer patients and is proposed to be a biomarker of disease progression[71]. ADAMTS12 acs as a tumor suppressor, by inhibiting the RAS-ERK signaling pathway[70]. This subnetwork also contains the B3GNT3, GALNT1, and MUCIN genes from Net6.

The highlighted subnetworks contained many genes from signaling pathways that are known drivers of CML tumorigenesis and progression. Of these, several genes in particular stand out.

TET2, RASSF1, and CARD11 are all known tumor suppressor genes, and SETBP1 is an important myeloid oncogene. SETBP1 mutations have been implicated as a triggering mechanism of secondary leukemia in several myeloid cancers, including atypical CML. TET2 complements SETBP1, and occur almost exclusively in BCR/ABL-Independent Imatinib resistant CML patients. A close paralog of one of the other co-occuring genes, ASXL2 was also present in the same subnetwork as TET2 (Net2). RASSF1 and the associated MST1 (also Net2) are key components of proliferative and apoptotic pathways involved in CML progression. CARD11 is a regulator of NF-κB, and plays a role in mediating NF-κB signaling's tumor suppressing activity in CML. TET2, SETBP1, and RASSF1 are also mutated at a higher rate in the BCR/ABL-independent group in this study.

The subnetworks Cos3 and Net6 contained a number of genes with roles in O-linked glycosylation, either as substrates or as key mediators of glycosylation. Glycosylation states regulate signaling pathways in many cancers, including myeloid varieties. For example abherrent regulation of the Fucosyltransferase 7 enzyme, which terminally caps glycan chains, has been been linked to adult T-cell leukemia[72].

Truncated O-linked glycan structures have been found in numerous cancers and are commonly associated with Mucin genes. Mucins act as regulators of signaling pathways by interacting with cell surface kinases and other receptors. They are also defined by a tandem repeat domain enriched with serine, proline, and threonine residues. This "mucin-domain" acts as a substrate for abundant O-linked glycosylation. MUC4 glycan truncation has been proposed to

stimulate downstream pathways such as ERK and PI3K by allowing interaction between the Erb-B2 Kinase and EGF-like domains, which are normally blocked by competitive interaction with MUC4[73]. ADAMTS12 and ADAMTS7, which were identified in subnetwork Cos3, are the only two known ADAMTS proteins with "mucin-like" domains. These domains are a likely target of extensive O-glycosylation[74,75].

Several genes were present in both subnetworks Cos3 and Net6 which play important roles as mediators of proper O-glycan formation. GALNT1 is a GalNAc-transferase enzyme and initiates the the initial glycan structure by adding *N*-acetylgalactosamine (GalNAc) to serine and threonine residues in the mucin domain. B3GNT3 encodes a *B*-1,3-GlcNAc-transferase which extends these by adding *N*-acetylglucosamine to GalNAc. The FUT2 and FUT3 genes encode fucosyltransferase enzymes that terminate extension of the glycan chain.

In summary these subnetworks contain two families of proteins, the MUCINs and ADAMTSs, with similar protein domains heavily modified by O-linked glycosylation. These glycosylation target genes also have been linked to regulation of the RAS/RAF/ERK pathway central to CML progression. These subnetworks also contain a group of genes which are key components of the O-glycan synthesis machinery. These results suggests a possible role for dysregulated glycosolation as a potential driver of CML reactivation in BCR/ABL-independent TKI resistance.

4.3 Conclusion

The results of this HotNet analysis identified a number of significant subnetworks containing genes involved in key CML signaling pathways. They identified several functionally connected genes of particular relevance in our study, a number of which were mutated almost exclusively in the BCR/ABL-Independent patients. TET2, SETBP1, RASSF1, MST1, ASXL2 and CARD11 are all known tumor suppressor genes (or closely associated genes) with prominent roles in other related myeloid disorders. These represent candidate driver genes worthy of

further study in BCR/ABL-Independent CML.

A novel set of genes connected to glycosylation pathways in BCR/ABL-independent CML progression was also identified. These genes (GALNT1, B3GNT3, FUT2, FUT3, MUC4, ADAMTS7, ADAMTS12), representing both targets and key mediators of O-linked glycosylation, are potential driver genes in BCR/ABL-Independent CML imatinib resistance and warrant additional study.

5 Methods and Materials

5.1 Sequencing

A total of 127 Whole Exome Sequencing(WES) samples were sequenced at OHSU and Oregon State University(OSU) sequencing centers. Seventy-two of the samples were sequenced at OSU Sequencing Core(Sample Group CML1). Of these, 69 samples were single end sequenced as three FlowCells - three samples on each lane. Three samples (04-00129, 03-00230, 15-00114) initially failed (Low Library Size) as part of FlowCell1 and these were re-sequenced paired end but only the R1 data - single end was used for downstream processing. There were 55 Legacy CML samples that were sequenced by OHSU Sequencing core and these were single end sequenced. For each flowcell and each sample, the FASTQ files were aggregated into single files for reads one and two, and trimmed by three bases on the 5' end and five bases on the 3' end. The Nimblegen Seqcap Target Enrichment Kit was used for sequence capture on all samples.

Table 8: WES Sequencing Totals

Sample Group	Sample Total	Chemistry	Instrument	Sequencing Core
3 (Legacy)	14	Nimblegen	HiSeq 2000	OHSU
5 (Legacy)	9	Nimblegen	HiSeq 2000	OHSU
6 (Legacy)	3	Nimblegen	HiSeq 2000	OHSU
7 (Legacy)	29	Nimblegen	HiSeq 2000	OHSU
CML1	72	Nimblegen	HiSeq 3000	OHSU

A total of 124 RNAseq samples were sequenced at the New York Genome Sequencing center (NYGC). Note that NYGC core provided the data in two batches. The first batch is referred to as CML1, the second batch is called CML2. There are only CML tumor samples. Paired end 125 cycle reads were generated. The NYGC core used the KAPA Total RNASeq Strand-Specific RNA Library Preparation Kit for sequence capture on all samples.

Table 9: RNASeq Sequencing Totals

Sample Group	Sample Total	Chemistry	Instrument	Sequencing Core
CML1	103	KAPA Total RNASeq	N/A	NYGC
CML2	21	KAPA Total RNASeq	N/A	NYGC

5.2 Alignments/Post-process/Read Summarization

5.2.1 WES

BWA MEM version 0.7.10-r789[76] was used to align read pairs for each sample-lane FASTQ file. The Genome Analysis Toolkit v3.3 and bundled Picard v1.120.1579[27] were used for alignment post-processing. The files contained within the Broad's bundle 2.8 were used including their version of the build 37 human genome (These files were downloaded from: ftp://ftp.broadinstitute.org/bundle/2.8/b37/). **Note**: Nimblegen Intervals available as Nimblegen_SeqCap_EZ_v3.bed were used for all the steps below. The following steps were performed per sample-lane SAM file generated by BWA:

- The SAM files were sorted and converted to BAM via SortSam
- CollectMultipleMetrics was used to obtain Alignment Metrics for each Sample.
- MarkDuplicates was run, marking both lane level standard and optical duplicates
- The reads were realigned around indels from the reads–RealignerTargetCreator/IndelRealigner.
- Base Quality Score Recalibration

The resulting BAM files were then aggregated by sample and an additional round of MarkDuplicates was carried out at the sample level.

5.2.2 RNAseq

Full alignments of reads was perfomed using Subjunc aligner(1.5.0-p1)[77]. BAM files obtained from Subjunc were used as inputs into featureCounts(1.5.0-p1)[78] and reads summarization was performed. **Note**: subjunc was run with following options: Reference used: human_g1k_v37.fasta (Source: ftp://ftp.broadinstitute.org/bundle/2.8/b37/) Note: During this process these reads were trimmed by 3 on the 5' end and 5 on the 3' end * Reference indexing: subread-buildindex -o human_g1k_v37 human_g1k_v37.fasta * Alignments: subjunc -i path/to/reference/ -u -r fastq1 -R fastq2 –gzFASTQinput -o outputBAMFilename -I 5 -n 10 -T

featureCounts was run with following options: Annotation File used: Homo_sapiens.GRCh37.75.gtf (Source: ftp://ftp.ensembl.org/pub/release-75/gtf/homo featureCounts -a annotation_file -o output -F GTF -t exon -g gene_id -s 2 -C -T 10 -p -B BAM_files –> Subjunc alignments have uniquely mapped reads.

5.3 Genotyping

GATK's Unified Genotyper[79] was used for generation of both WES and RNAseq genotypes, seperately. Prior to genotyping, an additional round of Indel realignment was carried out. VCF files, one for SNVs, another for Indels were obtained at the end of genotyping.

5.4 Relatedness Analysis/Clustering

Identity-by-State(IBS) relatedness analysis was performed on WES and RNAseq samples, using the SNPrelate Bioconductor R package[26]. SNV calls from the genotyping results for WES and RNAseq cohorts, obtained using GATK's Unified Genotyper, were used as input. Indels were not included. The SNPrelate package's clustering function uses the UPGMA

unweighted heirarchical method.

5.5 RNAseq Exploratory Analysis

Exploratory data vizualizations of gene-level expression data, using the featureCounts gene counts summaries for the RNAseq samples as input, was done using a custom script EDASeq_Ext.r. This script extends the functionality of a number of plotting functions included in the EDASeq Bioconductor R package[28].

5.5.1 Dropped Genes

Also included in the EDASeq_Ext.r script are two functions for visualizing unusual patterns of dropped(absent) genes within a particular group of samples. These functions are entirely of my own design and the concept is as follows:

Due to the noisiness of absolute gene counts from NGS for low expression genes, we would expect genes expressed at low levels across our entire sample cohort to display a pattern of absentness amongst many samples due to the expression in those samples dropping below the level of detection. Therefore, a gene expressed at very low levels will have a mix of low gene counts and zero read counts across the entire sample cohort. If a subset of samples is randomly selected, they may all (or nearly all) by chance have zero read counts for a particular gene.

Given a specified group of samples, say all the samples with a particular value for one specified clinical covariate (referred to here as **batch group**), genes with no read counts in a large number of samples across the entire **batch group** may be a result of random chance due to sampling or instead be the result of a systematic shift downward in the sequencing coverage of low expression genes.

In order to identify whether genes are being dropped from the **batch group** that wouldn't be expected randomly, these functions calculate the binomial probability of each individual gene with samples having zero read counts within the batch group being absent through chance, and then visualizes these dropped genes and their probabilities. For a given gene we can calculate the chance that the specified group contains k zeros entirely by chance as follows:

1. The probability of any single randomly selected sample having zero read counts by chance, is simply the total number of samples with zero read counts (z) divided by the total number of samples (N):

$$p = z/N$$

2. The probability of having k zeros across the specified sample group of size n is equal to the binomial probability:

$$PrBin(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Plotting the binomial probability of dropped genes with p-values above a specified threshold in this way allows one to compare the number and improbability of dropped genes across different groups given a specified clinical covariate, in order to discern any groups exhibiting any obvious batch effect of unusual dropped genes.

Due to the differences in size between the different groups, the dropped genes from groups with fewer samples do not have the possibility of binomial probilities as small as for groups with more samples, because of the smaller n in $PrBin(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$). This means that a comparison of the binomial probabilities of "dropped genes" between two groups with different sample sizes is not perfectly even. Due to this, a good followup comparison of dropped genes is done by specifying the single group of interest and then randomly subdividing the remaining samples into equally sized groups. Both versions of this dropped gene visualization function are included in the EDASeq_Ext.r script.

5.5.2 FastQC/BLAST

FastQC v0.11.4[29] was used to generate Quality Control reports on FASTQ raw read output files for all RNAseq samples. The Per Sequence GC Content and Overrepresented Sequence analysis module results were extracted for closer examination. The overrepresented sequences were pooled and submitted for BLAST sequence matching using NCBI's Nucleotide Blast tool[30].

5.5.3 Differential Expression

Differential expression (DE) analysis was performed using the edgeR Bioconductor package[34]. Median, Upper Quantile, and Full Quantile normalization of gene counts was performed using the EDASeq package, prior to DE analysis. For each type of normalization, within-sample normalization of gene counts by GC content was performed, followed by between-sample normalization of gene counts. The normalized gene count values were then passed as an offset to edgeR's generalized linear model. Raw p-values were adjusted to account for multiple-testing, holding the false discovery rate (FDR) at 5%.

5.6 Genotype Variant Filtering

5.6.1 Variant Effect Predictor

Variants were annotated functional consequence predictions using Version 78 of Ensemble's Variant Effect Predictor[45]. Version 1.6.6 of the vcf2maf.pl annotation tool was used to create a Mutation Annotation Format(MAF) file, with predicted "most deleterious" transcript, ExAc version 0.3.1, and Cosmic version 80 annotations.

5.7 HotNet2 Analysis

5.7.1 Beta Determination

The B heat diffusion parameter for each of the three selected PPIs(STRING, iRefindex, consensusPathDB) was determined using the description of parameter selection given in Supplemental Section 1.4.1 of the HotNet2 paper[50]. A total of 100 genes were selected as a representative sample for each connectivity vertice point (representing maximum, 75th percentile, median, and 25th percentile values for their **betweenness connectivity** in the network). For each gene the influence on all other nodes in the network after diffusion was calculated, for 20 values of β from 0.05 to 0.95. The influence drop off point was then chosen by finding the maximum influence value that nearly all (97.5 percent) of the first-order nodes meet. This was done for all the values of beta, and the inflection point beta value having the highest influence cutoff drop point was chosen.

5.7.2 Delta Determination

The second major parameter needed for the HotNet2 algorithm, aside from the β diffusion parameter, is the δ, or minimum edge weight parameter, which determines which edges will be cropped out of the network after the diffusion step is performed. Determination of an ideal δ parameter value is done for each PPI, by calculating δ values which are unlikely to return large connected subnetworks when running our data on the permuted versions of the PPIs. So for each resistance group dataset (BCR/ABL-independent and BCR/ABL-dependent) and each PPI (STRING, iRefindex, consensusPathDB), the HotNet2 delta estimation calculates the δ threshold at which all strongly connected components identified are smaller than or equal to a specified maximum size. This was done for four sizes: 5, 10, 15, and 20, and then repeated for all 100 permuted PPI networks, following the recommendations of the HotNet2 paper[50].

The heuristic used in the HotNet2 paper to determine one final value of δ for the HotNet2 run used is as follows:

1. Determine the median δ value for each max connencted component size, from all 100 permuted networks.
2. Run HotNet2 on all four median δ values.
3. Choose the smallest δ value with the largest number of statistically significant ($P < 0.05$) network sizes k.

5.7.3 Protein-Protein Interaction Networks

Three protein-protein interaction networks were used as the influence networks. In all three PPIs, self-edges were excluded. Only proteins with corresponding HUGO(**hgnc**) nomenclature gene names were retained.

iRefindex

Version 14 of iRefindex was used. This was significantly more current than the version used in the HotNet2 paper (v9). In order to mirror the selections made in the HotNet2 paper, certain interactions were excluded; (MI:0403(colocalization), MI:0208(genetic interaction),and MI:0914(association).

STRING

Version 10.0 of the STRING Consortium protein interaction database was used. High confidence interactions were selected by filtering for entries with combined score of $Score_{total} > 0.7$.

ConsensusPathDB

Version 31 of the ConsensusPathDB database of human protein-protein interactions was used. The biomaRt Bioconductor package was used to retrieve matching **hgnc** gene names for the **uniprot_swissprot** protein names provided by consensusPathDB.

Table 10: Protein-protein interaction network gene and edge totals.

PPI	Total Genes	Total Edges
iRefindex	15495	160065
STRING	14337	292301
consPathDB	15240	195900

5.7.4 Subnetwork Significance Calculation

Statistical significance was calculated using HotNet's findComponents.py module. The statistical significance calculation is not for the specific gene networks output by the HotNet2 algorithm, but is an estimate of the probability of seeing the n clusters of size k observed, using random permutations of the heat input data.

5.7.5 Cross-PPI Consensus Identification

Overlap in gene membership between significant subnetworks for HotNet2 runs on different PPIs, but with the same resistance type (BCR/ABL-independent, BCR/ABL-dependent) and filter schemes (COSMIC-only, all Variants), were identified in order to highlight subnetworks with consensus across HotNet runs.

In addition to pairwise matches representing shared gene membership between two subnetworks, multimeric networks of overlapping subnetworks can occur. Multimeric overlap

networks represent more than two subnetworks connected by direct and indirect links. An example of a multimeric network with three members would be: subnetwork A overlaps with subnetwork B, and subnetwork B overlaps with subnetwork C, but with disjoint overlap genes.

In order to account for multimeric overlap networks, and identify the most parsimonious set of overlap networks (both pairwise and/or multimeric) to account for the relationship in gene membership between the subnetworks, the NetworkX network analysis toolset was used.

The procedure is as follows:

1. All of the pairwise matches (gene overlap) between subnetworks from different PPI runs were identified.
2. An undirected NetworkX network object was created, with the matches between subnetworks designated as edges, and the subnetworks as nodes.
3. The NetworkX connected_component_subgraphs() module was used to identify all of the subgraphs on the graph.

5.8 Work Contributions

Dan Bottomly and Sashi Challa made critical contributions to this work; in performing the alignments, read summarization, and genotyping of the raw sequencing data for this study, and for generously sharing valuable analysis code. The McWeeney laboratory provided insight and assistance throughout, including input on the filtering criteria workflow. In the Druker lab, Christopher Eide spearheading this project and sharing his unparalleled knowledge of CML, and Samantha Savage for all of her work gathering clinical information from so many disparate sources. All other work described was performed by Adam Therneau.

Supplemental

Supplemental Table 1: Significant HotNet Subnetworks: BCR/ABL-Independent Group (All Vars)

PPI	**Genes**	**pval**
consPath	CDHR2 FIGN LANCL1 MYO7A MYO7B PCDH15 USH1C USH2A	0.01
consPath	ASPH CCDC180 CNDP1 COX7A2 HRC MANEA OLFM1 PDGFRL SCD5 TRDN	0.04
iRef	GALNT1 PPFIA2 PPFIA3 PTPRF SEC23A SEC24D	0.02
iRef	CYP3A5 DHRS7 ETV5 FAIM3 HOXD9	0.01
iRef	ANK3 AP3B1 ARHGEF10L CENPU PTPRN2 SCN2A SPTBN4	0.04
iRef	KRT20 KRT80 PLEKHA5 PLEKHA6 PROM1	0.01
iRef	POLA1 POLE RAD17 RBMS1	0.04
iRef	ACAN MMP19 MMP20 MMP8 TNFAIP6 UMOD	0.02
iRef	CSHL1 MAML2 MAML3 NOTCH4 RAI1 SSX3 ZNF496	0.04
iRef	BAIAP2L1 GRID2IP IQSEC2 SHANK3	0.04
iRef	ADAMTS12 ADAMTS7 COL9A1 COMP	0.04
iRef	ASPH CASQ2 HRC PDGFRL TRDN	0.01
iRef	COL17A1 KAZN LAD1 PPL	0.04
iRef	CLEC4M ITGAM LRP1B MMP12 PLAUR SRPX2	0.02
iRef	CNOT1 CNOT2 CNOT6L TNRC6A TNRC6B	0.01
iRef	AP1G1 AP1S1 ASB10 KIF13A	0.04

iRef	ABCA9 ATP6V1E1 ATP6V1H DEDD	0.04
iRef	HLA-B HLA-C KIR3DL1 LILRB1 TAP2	0.01
iRef	MUL1 USP32 USP6 VPS35	0.04
iRef	CRCP INTS4 KIAA0513 NBL1 SCAMP1 ZMIZ2	0.02
iRef	PDS5A PDS5B STAG1 STAG2	0.04
iRef	ABLIM1 APBA1 CNTNAP3 KCNJ12 LIN7A	0.01
iRef	AFF1 CCNT2 CHD1 MLLT1 MLLT3 RFX5	0.02
iRef	DEFA5 PRSS1 PTPN4 SPINK5 TST	0.01
iRef	FOXE1 PLCB3 TRPM6 TRPM7	0.04
iRef	GABRB2 PCSK9 RCN1 RHOT1 RHOT2 TRAK1 TRAK2	0.04
iRef	CNTRL MAP2K3 MAP3K4 TAOK1	0.04
STRING	CPSF3L CSTF2T INTS4 INTS7 KIAA0513 PAPOLG RALGAPA1 RBBP6 SON WDR33 ZC3H4 ZC3H6 ZNF292	0.00
STRING	CCDC39 CCDC40 DNAH1 DNAH12 DNAH14 DNAH17 DNAH5 DNAH7 DNAL1 DYNC1LI1 DYNC1LI2 DYNC2H1 OSBPL1A	0.00
STRING	ARMC4 ATG12 ATG2A ATG2B ATG3 ATG4C C9orf72 RB1CC1 SLC22A18 SMCR8 TBC1D32 ULK2 ZDHHC15 ZDHHC20	0.00

STRING	AFF1 ASXL2 ATP6V0A4 ATP6V1D ATP6V1E1 ATP6V1H BPTF CASP5 CASZ1 CBFA2T2 CCNT2 CD34 CDHR1 CECR2 DAP DMXL1 EIF4G2 GPR87 HOXD9 KAT6B MAP1S MEIS3 MLLT1 MLLT3 MST1 PPA2 PROM1 RASSF1 RHCG TET2 TM4SF5 TSHZ2 VAX1 ZNF462	0.00
STRING	COL11A1 COL18A1 COL20A1 COL21A1 COL22A1 COL24A1 COL25A1 COL4A4 COL4A5	0.02
STRING	APBA1 APBA2 APPBP2 ASPM BAIAP2L1 C16orf70 CARD10 CDH23 CIT EPS8 FGF12 FGFRL1 FUBP1 IMMT KIF14 KIF17 KIF20B KIF25 KIFC2 KLC1 LAIR1 LIN7A MAPK8IP2 MPP1 MYO15A MYO7A NCAPD3 NCAPG2 OTOG PCDH15 PDZD7 SMC4 TECTA TUBD1 USH1C USH2A VEZT WDR62	0.04
STRING	ADAMTS9 ATAD2 DARS DARS2 DCLK2 DSPP FBLN7 KARS KCNK1 NFIC PLK5 RBMS1 RCN1 RFX1 RFX5 RRBP1 SENP1 SENP7 SP100 SPATA4 SULT1A1 THADA TSPAN8 UBA2 USP34 WARS	0.01
STRING	ALG2 ANXA11 CC2D1A CHMP5 CHMP6 IST1 LYST MVB12A PDCD6IP RHPN2 SGSM3 UEVLD USP2 VPS37C VTA1 ZFYVE1	0.00
STRING	ANKRD28 ANKRD33 ATXN1 ATXN1L ATXN3 C10orf2 CACNA1A CIC MPV17 PLEKHG4 POLG PPP6R3 SCN9A TNIP3 USP36	0.00
STRING	ACE ACE2 ACSS2 ALDH3B2 APOL4 BCKDHA CNDP1 COMT DBH DBT DLAT HAL LDHAL6B LDHB METTL2B MMADHC MUT OGDH PCCA PDHA1 SLC36A1 SLC6A18 TRIM11	0.01

STRING	ACIN1 AGMO CLASRP CLK4 CRLS1 CYP20A1 CYP4F11 CYP4F2 CYP4F3 CYP4Z1 CYP7B1 DGKB LBR NDUFAF4 NDUFAF5 NDUFAF6 NDUFAF7 PHLDB2 RBM23 SLC25A41 SRPK1 SRSF10	0.00
STRING	ANK3 CNTN3 CNTNAP2 LGI1 NRCAM PTPRN2 SCN2A SPTBN4 SPTBN5	0.02
STRING	B3GNT3 BCAM FUT2 FUT3 GALNT1 ISM2 MSLN MUC12 MUC16 MUC17 MUC20 MUC21 MUC4 MUC5B MUC6 MUC7 PRDM16 SETBP1 UCP1	0.00
STRING	AGAP3 AP1G1 AP1S1 AP3B1 AP3S1 AP4B1 ATF5 CMYA5 COPB1 DTNBP1 ECD ERN1 MIA3 SAR1B SEC23A SEC24D SEC31A SGIP1 SLC2A8	0.00
STRING	CUL2 DTL GLMN KLHL13 KLHL3 LHX6 OXSR1 SLC22A13 SLC22A14 SLC47A2 WNK1 WNK4 ZMYM3 ZMYM4 ZMYM6 ZNF280C ZNF280D	0.00
STRING	CTBP2 ELAC2 EXOG KANK1 PPP1R15A SEPSECS TGIF1 ZBTB14 ZFYVE9 ZNF217	0.01
STRING	ACTR8 BRD8 CHD8 EP400 EPC1 HMGA1 INO80C INO80D KANSL1 KAT8 MGA	0.00

Supplemental Table 2: Significant HotNet Subnetworks: BCR/ABL-Independent Group (COSMIC Vars)

PPI	**Genes**	**pval**
STRING	ATAD5 CEP164 CRB2 DNA2 ERCC8 HUS1 MRE11A MSH3 ORC2 ORC4 PLAA PMS2 POLE RAD17 SETD1B SLC13A2 SMURF2 STAM2 TOP1MT TP53BP1 UBXN11 USP8 WDR90	0.02

STRING	ACAN ADAMTS12 ADAMTS7 B3GNT3 COMP FUT3 GALNT1 MSLN MUC12 MUC16 MUC17 MUC20 MUC21 MUC4 MUC5B MUC6 MUC7	0.02
STRING	BRD2 CARD11 DCHS2 DKC1 DNAH17 DYNC1LI2 EP400 ERAP1 FBXO2 FYB GPAM GPD2 HLA-A HLA-B HLA-C HLA-DQB1 HLA-DRB1 HLA-DRB5 IRF2 JUP RUVBL1 SH2B3 SIRPA	0.02
consPath	A2ML1 CELA1 IVL L2HGDH LOR LPA MMP12 SPRR3	0.04
consPath	BCL6B GPRIN2 HOXA1 KRTAP10-1 KRTAP12-4 KRTAP4-7 LCE4A PCSK5	0.04
consPath	ATG12 ATN1 AUTS2 CARD11 CEP170B CNOT1 CTBP2 DDX20 EHMT1 EPSTI1 ETV3 FCGBP FYB GIGYF2 MIER2 NEB NOL4 PCGF6 POTEF PRDM6 PRRC2A PSPH RB1CC1 RERE RING1 RPS29 SAFB2 SRA1 STYXL1 TFAP2A TGIF1 TNRC6B TNXB TRIM39 TRIM5 ULK2 USP2 WBP11 WBP2NL ZBTB33	0.02
iRef	CELA1 IVL LOR LPA MMP12 PRSS3 SPRR3	0.04
iRef	ATG12 CTBP2 DDX6 EHMT1 EPSTI1 NEB NOL4 PRDM6 PSPH RB1CC1 RPS29 SAFB2 SOX13 STYXL1 TFAP4 TFCP2L1 TGIF1 ULK2	0.02
iRef	CCDC150 DTNBP1 ISCU NUP153 NUP62 P4HA3 SSC5D	0.04

Supplemental Table 3: Significant HotNet Subnetworks: BCR/ABL-Dependent Group (All Vars)

PPI	Genes	pval
STRING	ACTR8 AFF1 ANKRD17 ASXL2 ATP6V0A4 ATP6V0D1 ATP6V1D ATP6V1E1 ATP6V1G2 ATP6V1H BPTF BRD8 CASZ1 CBFA2T2 CD34 CECR2 CHD8 DMXL1 ENKUR EP400 FAN1 GPR87 HOXD9 INO80C INO80D KANSL1 KAT6B KAT8 KMT2C KMT2E MADCAM1 MAP1S MEIS3 MGA MLLT1 MLLT3 MST1 MTMR10 NKX2-3 PMS1 PMS2 PPA2 PROM1 RHCG SELL SETD1B TRPC1 TRPM1 TRPM3 TRPM7 VAX1 WDR90 ZNF462	0.01
STRING	JPH3 LIPN PCSK5 PCSK7 PCSK9 VPS13A VPS13B VPS13C	0.02
STRING	ANXA5 ARL13B IFT88 PKD1 PKHD1 RPGR RPGRIP1 TBCD	0.02
STRING	AGAP3 COPB1 ECD MIA3 SAR1B SEC16A SEC23A SEC24D SEC31A	0.03
STRING	ACE ACE2 ALDH3B2 CNDP1 COMT DBH HAL METTL2B UROC1	0.03

STRING	ABCD3 AGFG2 AHNAK ANXA2R AP1G1 AP1S1 AP3B1 AP3S1 AP5B1 APPBP2 BAIAP2 BAIAP2L1 CARD10 CASP5 CDH23 CMYA5 DAP DCP1B DDHD1 DEDD DIAPH1 DIAPH3 DNAJC6 DNMBP DOCK4 DTNBP1 EIF4G2 ELMO2 ELMO3 ENAH EPS15L1 EPS8 ESRP1 ESRP2 FLG FMN1 FMN2 FUBP1 HDLBP HRNR IMMT INF2 IVL KAZN KIAA0319 KIF13A KIF17 KIF25 KIFC2 KLC1 KLK5 KRT10 KRT18 KRT2 KRT77 KRT8 KRT9 LIN7A LOR LRRC3C LRRC4B MAL MN1 MPV17 MUL1 MYO15A MYO7A NR6A1 NUMBL OTOG OTUD4 PCDH15 PLXNB3 POLG PPFIA2 PPFIA3 PPFIBP1 PPFIBP2 PPHLN1 PPL PTPRF RAB36 RAB38 ROBO1 ROM1 RPTN S100A10 SCN9A SEC63 SH3D19 SH3YL1 SLC5A7 SNAP91 SORL1 SPG11 SPIRE2 SPRR3 SYTL3 TBC1D24 TCHH TNFRSF11A TRAF5 TUBD1 TXNDC5 UBXN11 USH1C USH2A USP32 USP4 USP48 USP53 USP6 VPS35 YTHDF2 ZFC3H1 ZFYVE26	0.01
STRING	ASPM ATAD5 ESCO1 KIF20B NCAPD3 NCAPG2 SMC1A SMC4 WDR62	0.03
STRING	CPSF3L INTS4 INTS7 KIAA0513 PAPOLG RALGAPA1 RBBP6 SON ZC3H6	0.03
STRING	COG3 COG5 COG6 COG7 DPY19L2 GOLGA2 GOLGA3 RUSC2 SPATA16	0.03
STRING	RCN1 SENP1 SENP7 SULT1A1 THADA TSPAN8 UBA2 USP34	0.02

STRING	ATAD2 DCLK2 HIVEP2 RBMS1 RFX1 RFX5 RRBP1 SPATA4	0.02
STRING	CDON FNDC1 IGDCC4 IQSEC2 LRRC4C NEO1 NTNG1 PRTG	0.02
iRef	HLA-A HLA-B HLA-C LILRB1 TAP2	0.04
iRef	CP DMTN DYNC1LI2 PROC PROCR	0.04
iRef	CUX1 ELAC2 GOLGA5 RECQL4 RECQL5	0.04
iRef	ADAM10 SH3D19 SH3YL1 SOS2 TSPAN33	0.04
iRef	CSHL1 MAML2 MAML3 NOTCH4 RAI1	0.04
iRef	AP1G1 AP1S1 ASB10 EBLN2 KIF13A	0.04
iRef	CLEC4M ITGAM LRP1B MMP12 PLAUR	0.04

Supplemental Table 4: Significant HotNet Subnetworks: BCR/ABL-Dependent Group (COSMIC Vars)

PPI	**Genes**	**pval**
consPath	AKNA ATG12 ATN1 AUTS2 BCL6B CENPJ CEP170B CNOT1 CTBP2 DAZAP1 DDX20 DDX6 EHMT1 EPSTI1 ESRRA GIGYF2 GPRIN2 HOXA1 KRTAP10-1 KRTAP12-4 KRTAP4-7 LCE4A LYST MEGF8 MIER2 NEB NOL4 PCGF6 PCSK5 PRRC2A PSPH RB1CC1 RC3H1 RPS29 SAFB2 SAP130 STYXL1 TFAP4 TGIF1 TNRC6B ULK2 WBP11 ZBTB33 ZNF462 ZNF609	0.04

iRef	ATG12 ATN1 BCL6B CENPJ CNOT1 CTBP2 DDX20 DDX6 DHX57 EHMT1 EIF4E EPSTI1 ESRRA FUBP1 GIGYF1 GIGYF2 GPRIN2 HOXA1 IMMT KAT6B KRTAP10-1 KRTAP12-4 KRTAP4-7 LCE4A LYST MEGF8 MKNK2 NEB NOL4 PABPC1 PCSK5 PRRC2A PSPH RB1CC1 RPS29 SAFB2 SAP130 STYXL1 TFAP4 TGIF1 TNRC6B UBXN11 ULK2 WBP11 WDFY3 ZNF462	0.02
STRING	BRD2 BRIX1 CARD11 DCHS2 DKC1 DNAH17 DYNC1LI2 EP400 ERAP1 FBXO2 FYB GPAM GPD2 HLA-A HLA-B HLA-C HLA-DQB1 HLA-DRB1 HLA-DRB5 IRF2 JUP KCTD19 NOP9 PA2G4 PRMT8 RPF1 RUVBL1 SH2B3 SIRPA SOX4	0.04

Supplemental Table 5: Top BLAST Hits for Overrepresented Sequences Present in All Batch Samples

	BLAST Hit	E-Score
1	Homo sapiens beta-globin (HBB) gene, complete cds	5.68487e-16
2	Cloning vector pTT-PB-hTERT-puro, complete sequence	5.68487e-16
3	Cloning vector pTT-PB-SOKM-puro, complete sequence	5.68487e-16
4	Homo sapiens hb, hbb gene for beta globin, complete cds, note: HbHofu 126(GTG>GAG)	5.68487e-16
5	Homo sapiens clone BT009B hemoglobin beta chain (HBB) gene, partial cds	5.68487e-16
6	PREDICTED: Nomascus leucogenys hemoglobin subunit beta (LOC100580975), transcript variant X3, mRNA	5.68487e-16

7	PREDICTED: Nomascus leucogenys hemoglobin subunit beta (LOC100580975), transcript variant X2, mRNA	5.68487e-16
8	PREDICTED: Nomascus leucogenys hemoglobin subunit beta (LOC100580975), transcript variant X1, mRNA	5.68487e-16
9	Cloning vector pTT-PB-SOKMLNpuro, complete sequence	5.68487e-16
10	Homo sapiens isolate BT048B beta globin (HBB) gene, partial cds	5.68487e-16
11	Homo sapiens isolate BT047B beta globin (HBB) gene, partial cds	5.68487e-16
12	Homo sapiens isolate BT046B beta globin (HBB) gene, partial cds	5.68487e-16
13	Homo sapiens isolate BT045B beta globin (HBB) gene, partial cds	5.68487e-16
14	Homo sapiens isolate BT033B beta globin (HBB) gene, partial cds	5.68487e-16
15	Homo sapiens isolate BT031B beta globin (HBB) gene, partial cds	5.68487e-16
16	Homo sapiens isolate BT011B beta globin (HBB) gene, partial cds	5.68487e-16
17	Homo sapiens isolate BT010B beta globin (HBB) gene, partial cds	5.68487e-16
18	Homo sapiens isolate HC2B beta globin (HBB) gene, partial cds	5.68487e-16
19	Homo sapiens beta globin (HBB) gene, partial sequence	5.68487e-16
20	PREDICTED: Pan troglodytes hemoglobin, beta (HBB), mRNA	5.68487e-16

21	PREDICTED: Pongo abelii hemoglobin, beta (HBB), transcript variant X2, mRNA	5.68487e-16
22	PREDICTED: Pongo abelii hemoglobin, beta (HBB), transcript variant X1, mRNA	5.68487e-16
23	PREDICTED: Pan paniscus hemoglobin subunit beta (LOC100976465), mRNA	5.68487e-16
24	Expression vector pFUSE-HEAVY, complete sequence	5.68487e-16
25	Expression vector pFUSE-LIGHT, complete sequence	5.68487e-16
26	Homo sapiens hemoglobin, beta (HBB) gene, complete cds	5.68487e-16
27	Homo sapiens beta globin gene, exon 3 and partial cds	5.68487e-16
28	Expression vector pFUSE-rFc2-adapt-scFv, complete sequence	5.68487e-16
29	Expression vector pFUSE-mFc2-adapt-scFv, complete sequence	5.68487e-16
30	Expression vector pFUSE-hFc2-adapt-scFv, complete sequence	5.68487e-16
31	Cloning vector pnlslacZ-ACN, complete sequence	5.68487e-16
32	Cloning vector pAP-ACN, complete sequence	5.68487e-16
33	Homo sapiens beta-globin Showa Yakushiji variant (HBB) gene, HBB-Showa Yakushiji allele, exon 3 and partial cds	5.68487e-16
34	Homo sapiens beta globin region (HBB); and beta globin locus transcript 3 (non-protein coding) (BGLT3); and hemoglobin subunit beta (HBB); and hemoglobin subunit delta (HBD); and hemoglobin subunit epsilon 1 (HBE1); and hemoglobin subunit gamma 1 (HBG1); and hemoglobin subunit gamma 2 (HBG2), RefSeqGene on chromosome 11	5.68487e-16

35	Pongo abelii BAC clone CH276-201O10 from chromosome unknown, complete sequence	5.68487e-16
36	Homo sapiens beta globin chain (HBB) gene, complete cds	5.68487e-16
37	Cloning vector pAAV-EF1alpha-hFAH.AOS2, complete sequence	5.68487e-16
38	Leontopithecus chrysomelas beta-globin gene, intron 2 and 3' flanking region	5.68487e-16
39	Homo sapiens isolate HbC-Dgn83 beta globin (HBB) gene, complete cds	5.68487e-16
40	Homo sapiens isolate HbC-Dgn99 beta globin (HBB) gene, complete cds	5.68487e-16
41	Homo sapiens isolate HbC-Dgn66 beta globin (HBB) gene, complete cds	5.68487e-16
42	Homo sapiens isolate HbC-Ghn117 beta globin (HBB) gene, complete cds >gi|71727260|gb|DQ126320.1| Homo sapiens isolate HbC-Ghn133 beta globin (HBB) gene, complete cds >gi|71727262|gb|DQ126321.1| Homo sapiens isolate HbC-Ghn40 beta globin (HBB) gene, complete cds >gi|71727264|gb|DQ126322.1| Homo sapiens isolate HbC-S782 beta globin (HBB) gene, complete cds	5.68487e-16
43	Homo sapiens isolate HbC-Ghn195 beta globin (HBB) gene, complete cds	5.68487e-16

44	Homo sapiens isolate HbC-Dgn06 beta globin (HBB) gene, complete cds >gi|71727232|gb|DQ126306.1| Homo sapiens isolate HbC-Dgn31a beta globin (HBB) gene, complete cds >gi|71727234|gb|DQ126307.1| Homo sapiens isolate HbC-Dgn31b beta globin (HBB) gene, complete cds >gi|71727236|gb|DQ126308.1	5.68487e-16
45	Homo sapiens isolate HbA-Dgn99 beta globin (HBB) gene, complete cds	5.68487e-16
46	Homo sapiens isolate HbA-Ivc18 beta globin (HBB) gene, complete cds	5.68487e-16
47	Homo sapiens isolate HbA-Dgn58 beta globin (HBB) gene, complete cds	5.68487e-16
48	Homo sapiens isolate HbA-G37 beta globin (HBB) gene, complete cds	5.68487e-16
49	Homo sapiens isolate HbA-Dgn66 beta globin (HBB) gene, complete cds	5.68487e-16
50	Homo sapiens isolate HbA-Cmn087 beta globin (HBB) gene, complete cds	5.68487e-16
51	Homo sapiens isolate HbA-Ghn117 beta globin (HBB) gene, complete cds	5.68487e-16
52	Homo sapiens isolate HbA-Ivc04 beta globin (HBB) gene, complete cds	5.68487e-16
53	Homo sapiens isolate HbA-Ivc16 beta globin (HBB) gene, complete cds	5.68487e-16
54	Homo sapiens isolate HbA-G08 beta globin (HBB) gene, complete cds	5.68487e-16
55	Homo sapiens isolate HbA-Dgn52 beta globin (HBB) gene, complete cds	5.68487e-16

56	Homo sapiens isolate HbA-Sen10 beta globin (HBB) gene, complete cds	5.68487e-16
57	Homo sapiens isolate HbA-G25 beta globin (HBB) gene, complete cds	5.68487e-16
58	Homo sapiens isolate HbA-Ivc11 beta globin (HBB) gene, complete cds	5.68487e-16
59	Homo sapiens isolate HbA-Ghn009 beta globin (HBB) gene, complete cds	5.68487e-16
60	Homo sapiens isolate HbA-Ghn017 beta globin (HBB) gene, complete cds	5.68487e-16
61	Homo sapiens isolate HbA-Nov24 beta globin (HBB) gene, complete cds	5.68487e-16
62	Homo sapiens isolate HbA-JK1033 beta globin (HBB) gene, complete cds	5.68487e-16
63	Homo sapiens isolate HbA-Cmn097 beta globin (HBB) gene, complete cds	5.68487e-16
64	Homo sapiens isolate HbA-Dgn67 beta globin (HBB) gene, complete cds	5.68487e-16
65	Homo sapiens isolate HbA-Dgn06 beta globin (HBB) gene, complete cds	5.68487e-16
66	Homo sapiens isolate HbA-Sen50 beta globin (HBB) gene, complete cds	5.68487e-16
67	Homo sapiens isolate HbA-Cmr15 beta globin (HBB) gene, complete cds >gi|71727182|gb|DQ126281.1| Homo sapiens isolate HbA-Gna27 beta globin (HBB) gene, complete cds >gi|71727184|gb|DQ126282.1| Homo sapiens isolate HbA-S782 beta globin (HBB) gene, complete cds	5.68487e-16

68	Homo sapiens isolate HbA-Dgn83 beta globin (HBB) gene, complete cds	5.68487e-16
69	Homo sapiens isolate HbA-Sen31 beta globin (HBB) gene, complete cds	5.68487e-16
70	Homo sapiens isolate HbA-Ghn133 beta globin (HBB) gene, complete cds	5.68487e-16
71	Homo sapiens isolate HbA-Ghn195 beta globin (HBB) gene, complete cds	5.68487e-16
72	Homo sapiens isolate HbA-Dgn37 beta globin (HBB) gene, complete cds >gi|71727168|gb|DQ126274.1| Homo sapiens isolate HbA-Ghn023 beta globin (HBB) gene, complete cds >gi|71727170|gb|DQ126275.1| Homo sapiens isolate HbA-Ghn40 beta globin (HBB) gene, complete cds	5.68487e-16
73	Homo sapiens isolate HbA-Sen42 beta globin (HBB) gene, complete cds	5.68487e-16
74	Homo sapiens isolate HbA-Cmr13 beta globin (HBB) gene, complete cds	5.68487e-16
75	Homo sapiens isolate HbA-Ivc05 beta globin (HBB) gene, complete cds	5.68487e-16
76	Homo sapiens hemoglobin beta (HBB) gene, HBB-Hinsdale allele, exon 3 and partial cds	5.68487e-16
77	Homo sapiens hemoglobin beta mRNA, complete cds	5.68487e-16
78	Homo sapiens hemoglobin beta chain (HBB) gene, HBB-HB Camden allele, partial cds	5.68487e-16
79	Homo sapiens beta-globin beta thalassemia variant (HBB) gene, complete sequence	6.92559e-15
80	Homo sapiens beta globin chain gene, complete cds	5.68487e-16

81	Homo sapiens beta globin mutant (HBB) gene, complete cds	5.68487e-16
82	Homo sapiens partial HBB gene for hemoglobin beta chain, hyperunstable truncated variant, exon 3, isolate 0523762	5.68487e-16
83	Homo sapiens hemoglobin subunit beta (HBB), mRNA	5.68487e-16
84	Homo sapiens chromosome 11, clone CTD-2643I7, complete sequence	5.68487e-16
85	Homo sapiens beta globin chain variant (HBB) gene, HBB-O-Arab allele, complete cds	5.68487e-16
86	Homo sapiens chromosome 11, clone RP11-1205H24, complete sequence	5.68487e-16
87	Human messenger RNA for beta-globin	5.68487e-16
88	Pithecia pithecia beta globin gene, complete cds	5.68487e-16
89	Pongo pygmaeus beta-globin gene, exon 3 and partial cds	5.68487e-16
90	Homo sapiens beta-globin gene, complete cds	5.68487e-16
91	Homo sapiens beta-globin (HBB) gene, with a 1 bp (g) insertion mutation FS21G at base 279 resulting in beta-thalassemia, (J00179 bases 61971-63802)	5.68487e-16
92	Homo sapiens beta-globin (HBB) gene, with a 1 bp (t) insertion mutation FS10C at base 246 resulting in beta-thalassemia (J00179 bases 61971-63802)	5.68487e-16
93	Homo sapiens beta-globin (HBB) gene, with a to c allele 28 bp 5' to exon 1, (J00179 bases 61971-63802)	5.68487e-16
94	Homo sapiens beta-globin (HBB) gene, with c to a mutation 88 bp 5' to exon 1, (J00179 bases 61971-63802)	5.68487e-16

95	Homo sapiens beta-globin (HBB) gene, with c to t allele 90 bp 5' exon 1, (J00179 bases 61971-63802)	5.68487e-16
96	Homo sapiens beta-globin (HBB) gene, with t to c mutation L114P resulting in dominant beta-thalassemia intermedia, (J00179 bases 61971-63802)	5.68487e-16
97	Homo sapiens beta-globin (HBB) gene, with a one bp (c) deletion mutation FS45L at base 481, (J00179 bases 61971-63802)	5.68487e-16
98	Homo sapiens beta-globin (HBB) gene, with g to a mutation W37X resulting in premature stop and beta-thalassemia (J00179 bases 61971-63802)	5.68487e-16

References

1. Ren, R. (2005). Mechanisms of bCR-aBL in the pathogenesis of chronic myelogenous leukaemia. Nature reviews.Cancer *5*, 172–183.

2. Druker, B.J., Guilhot, F., O'Brien, S.G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M.W., Silver, R.T., Goldman, J.M., and Stone, R.M. *et al.* (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. The New England journal of medicine *355*, 2408–2417.

3. O'Hare, T., Zabriskie, M.S., Eiring, A.M., and Deininger, M.W. (2012). Pushing the limits of targeted therapy in chronic myeloid leukaemia. Nature reviews.Cancer *12*, 513–526.

4. Bixby, D., and Talpaz, M. (2011). Seeking the causes and solutions to imatinib-resistance

in chronic myeloid leukemia. Leukemia *25*, 7–22.

5. Quintas-Cardama, A., Kantarjian, H.M., and Cortes, J.E. (2009). Mechanisms of primary and secondary resistance to imatinib in chronic myeloid leukemia. Cancer control : journal of the Moffitt Cancer Center *16*, 122–131.

6. Donato, N.J., Wu, J.Y., Stapley, J., Lin, H., Arlinghaus, R., Aggarwal, B.B., Shishodia, S., Albitar, M., Hayes, K., and Kantarjian, H. *et al.* (2004). Imatinib mesylate resistance through bCR-aBL independence in chronic myelogenous leukemia. Cancer research *64*, 672–677.

7. Patel, A.B., O'Hare, T., and Deininger, M.W. (2017). Mechanisms of resistance to aBL kinase inhibition in chronic myeloid leukemia and the development of next generation aBL kinase inhibitors. Hematology/oncology clinics of North America *31*, 589–612.

8. Gorre, M.E., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P.N., and Sawyers, C.L. (2001). Clinical resistance to sTI-571 cancer therapy caused by bCR-aBL gene mutation or amplification. Science (New York, N.Y.) *293*, 876–880.

9. Mahon, F.X., Deininger, M.W., Schultheis, B., Chabrol, J., Reiffers, J., Goldman, J.M., and Melo, J.V. (2000). Selection and characterization of bCR-aBL positive cell lines with differential sensitivity to the tyrosine kinase inhibitor sTI571: Diverse mechanisms of resistance. Blood *96*, 1070–1079.

10. Thomas, J., Wang, L., Clark, R.E., and Pirmohamed, M. (2004). Active transport of imatinib into and out of cells: Implications for drug resistance. Blood *104*, 3739–3745.

11. Ma, L., Shan, Y., Bai, R., Xue, L., Eide, C.A., Ou, J., Zhu, L.J., Hutchinson, L., Cerny, J., and Khoury, H.J. *et al.* (2014). A therapeutically targetable mechanism of bCR-aBL-independent imatinib resistance in chronic myeloid leukemia. Science translational

medicine *6*, 252ra121.

12. Wang, Y., Cai, D., Brendel, C., Barett, C., Erben, P., Manley, P.W., Hochhaus, A., Neubauer, A., and Burchert, A. (2007). Adaptive secretion of granulocyte-macrophage colony-stimulating factor (gM-cSF) mediates imatinib and nilotinib resistance in bCR/ABL+ progenitors via jAK-2/STAT-5 pathway activation. Blood *109*, 2147–2155.

13. Steelman, L.S., Pohnert, S.C., Shelton, J.G., Franklin, R.A., Bertrand, F.E., and McCubrey, J.A. (2004). JAK/STAT, raf/MEK/ERK, pI3K/Akt and bCR-aBL in cell cycle progression and leukemogenesis. Leukemia *18*, 189–218.

14. Jamieson, C.H., Ailles, L.E., Dylla, S.J., Muijtjens, M., Jones, C., Zehnder, J.L., Gotlib, J., Li, K., Manz, M.G., and Keating, A. *et al.* (2004). Granulocyte-macrophage progenitors as candidate leukemic stem cells in blast-crisis cML. The New England journal of medicine *351*, 657–667.

15. Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. Molecular cell *58*, 586–597.

16. Ioannidis, J.P., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., and Jurman, G. *et al.* (2009). Repeatability of published microarray gene expression analyses. Nature genetics *41*, 149–155.

17. Baggerly, K.A., and Coombes, K.R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Ann. Appl. Stat. *3*, 1309–1334. Available at: http://dx.doi.org/10.1214/09-AOAS291.

18. Fare, T.L., Coffey, E.M., Dai, H., He, Y.D., Kessler, D.A., Kilian, K.A., Koch, J.E., LeProust, E., Marton, M.J., and Meyer, M.R. *et al.* (2003). Effects of atmospheric ozone on

microarray data quality. Analytical Chemistry *75*, 4672–4675.

19. Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences of the United States of America *97*, 10101–10106.

20. HOLMES, S., ALEKSEYENKO, A., TIMME, A., NELSON, T., PASRICHA, P.J., and SPORMANN, A. (2011). Visualization and statistical comparisons of microbial communities using r packages on phylochip data. Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing, 142–153.

21. Yang, H., Harrington, C.A., Vartanian, K., Coldren, C.D., Hall, R., and Churchill, G.A. (2008). Randomization in laboratory procedure is key to obtaining reproducible microarray results. PLoS ONE *3*, e3724. doi:10.1371/journal.pone.0003724.

22. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature reviews.Genetics *11*, 733–739.

23. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics (Oxford, England) *8*, 118–127.

24. Parker, H.S., Corrada Bravo, H., and Leek, J.T. (2014). Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ *2*, e561.

25. Baggerly, K.A., Edmonson, S.R., Morris, J.S., and Coombes, K.R. (2004). High-resolution serum proteomic patterns for ovarian cancer detection. Endocrine-related cancer *11*, 583–4; author reply 585–7.

26. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A

high-performance computing toolset for relatedness and principal component analysis of sNP data. Bioinformatics (Oxford, England) *28*, 3326–3328.

27. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. *et al.* (2010). The genome analysis toolkit: A mapReduce framework for analyzing next-generation dNA sequencing data. Genome research *20*, 1297–1303.

28. Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for rNA-seq data. BMC bioinformatics *12*, 480–2105–12–480.

29. Andrews, S. FastQC a quality control tool for high throughput sequence data.

30. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI bLAST: A better web interface. Nucleic acids research *36*, W5–9.

31. Mastrokolias, A., Dunnen, J.T. den, Ommen, G.B. van, Hoen, P.A. 't, and Roon-Mom, W.M. van (2012). Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood rNA. BMC genomics *13*, 28–2164–13–28.

32. Shin, H., Shannon, C.P., Fishbane, N., Ruan, J., Zhou, M., Balshaw, R., Wilson-McManus, J.E., Ng, R.T., McManus, B.M., and Tebbutt, S.J. *et al.* (2014). Variation in rNA-seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. PloS one *9*, e91041.

33. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. Nature reviews.Genetics *10*, 57–63.

34. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics

(Oxford, England) *26*, 139–140.

35. Akey, M., Joshua, Biswas, S., Leek, T., Jeffrey, and Storey, D., John On the design and analysis of gene expression studies in human populations.

36. Rahman, N., and Stratton, M.R. (1998). The genetics of breast cancer susceptibility. Annual Review of Genetics *32*, 95–121.

37. Pal, T., Permuth-Wey, J., Betts, J.A., Krischer, J.P., Fiorica, J., Arango, H., LaPolla, J., Hoffman, M., Martino, M.A., and Wakeley, K. *et al.* (2005). BRCA1 and bRCA2 mutations account for a large proportion of ovarian carcinoma cases. Cancer *104*, 2807–2816.

38. Jones, S., Anagnostou, V., Lytle, K., Parpart-Li, S., Nesselbush, M., Riley, D.R., Shukla, M., Chesnick, B., Kadan, M., and Papp, E. *et al.* (2015). Personalized genomic analyses for cancer mutation discovery and interpretation. Science translational medicine *7*, 283ra53.

39. Dvinge, H., Ries, R.E., Ilagan, J.O., Stirewalt, D.L., Meshinchi, S., and Bradley, R.K. (2014). Sample processing obscures cancer-specific alterations in leukemic transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *111*, 16802–16807.

40. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). DbSNP: The nCBI database of genetic variation. Nucleic acids research *29*, 308–311.

41. Consortium, 1.G.P., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., and McVean, G.A. *et al.* (2015). A global reference for human genetic variation. Nature *526*, 68–74.

42. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T.,

O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., and Cummings, B.B. *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

43. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., and Hoover, J. *et al.* (2016). ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic acids research *44*, D862–8.

44. Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., and Stratton, M.R. *et al.* (2004). The cOSMIC (catalogue of somatic mutations in cancer) database and website. British journal of cancer *91*, 355–358.

45. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome biology *17*, 122–016–0974–4.

46. Mwenifumbo, J.C., and Marra, M.A. (2013). Cancer genome-sequencing study design. Nature reviews.Genetics *14*, 321–332.

47. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., and Lander, E.S. *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545–15550.

48. Wu, D., and Smyth, G.K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. Nucleic acids research *40*, e133.

49. Garraway, L.A., and Lander, E.S. (2013). Lessons from the cancer genome. Cell *153*,

17–37.

50. Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., and McLellan, M. *et al.* (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics *47*, 106–114.

51. Steelman, L.S., Franklin, R.A., Abrams, S.L., Chappell, W., Kempf, C.R., Basecke, J., Stivala, F., Donia, M., Fagone, P., and Nicoletti, F. *et al.* (2011). Roles of the ras/Raf/MEK/ERK pathway in leukemia therapy. Leukemia *25*, 1080–1094.

52. Molli, P.R., Pradhan, M.B., Advani, S.H., and Naik, N.R. (2012). RhoA: A therapeutic target for chronic myeloid leukemia. Molecular cancer *11*, 16–4598–11–16.

53. Deininger, M.W., Goldman, J.M., and Melo, J.V. (2000). The molecular biology of chronic myeloid leukemia. Blood *96*, 3343–3356.

54. Vos, M.D., Ellis, C.A., Bell, A., Birrer, M.J., and Clark, G.J. (2000). Ras uses the novel tumor suppressor rASSF1 as an effector to mediate apoptosis. The Journal of biological chemistry *275*, 35669–35672.

55. Weyden, L. van der, and Adams, D.J. (2007). The ras-association domain family (rASSF) members and their role in human tumourigenesis. Biochimica et biophysica acta *1776*, 58–85.

56. Makishima, H., Jankowska, A.M., McDevitt, M.A., O'Keefe, C., Dujardin, S., Cazzolli, H., Przychodzen, B., Prince, C., Nicoll, J., and Siddaiah, H. *et al.* (2011). CBL, cBLB, tET2, aSXL1, and iDH1/2 mutations and additional chromosomal aberrations constitute

molecular events in chronic myelogenous leukemia. Blood *117*, e198–206.

57. Grossmann, V., Kohlmann, A., Zenger, M., Schindela, S., Eder, C., Weissmann, S., Schnittger, S., Kern, W., Muller, M.C., and Hochhaus, A. *et al.* (2011). A deep-sequencing study of chronic myeloid leukemia patients in blast crisis (bC-cML) detects mutations in 76.9. Leukemia *25*, 557–560.

58. Makishima, H., Jankowska, A.M., McDevitt, M.A., O'Keefe, C., Dujardin, S., Cazzolli, H., Przychodzen, B., Prince, C., Nicoll, J., and Siddaiah, H. *et al.* (2011). CBL, cBLB, tET2, aSXL1, and iDH1/2 mutations and additional chromosomal aberrations constitute molecular events in chronic myelogenous leukemia. Blood *117*, e198–206.

59. Grotzinger, T., Jensen, K., and Will, H. (1996). The interferon (iFN)-stimulated gene sp100 promoter contains an iFN-gamma activation site and an imperfect iFN-stimulated response element which mediate type i iFN inducibility. The Journal of biological chemistry *271*, 25253–25260.

60. Held, S.A., Heine, A., Kesper, A.R., Schonberg, K., Beckers, A., Wolf, D., and Brossart, P. (2015). Interferon gamma modulates sensitivity of cML cells to tyrosine kinase inhibitors. Oncoimmunology *5*, e1065368.

61. Du, W., Wang, S., Zhou, Q., Li, X., Chu, J., Chang, Z., Tao, Q., Ng, E.K., Fang, J., and Sung, J.J. *et al.* (2013). ADAMTS9 is a functional tumor suppressor through inhibiting aKT/mTOR pathway and associated with poor survival in gastric cancer. Oncogene *32*, 3319–3328.

62. Peng, L., Yang, Z., Tan, C., Ren, G., and Chen, J. (2013). Epigenetic inactivation of aDAMTS9 via promoter methylation in multiple myeloma. Molecular medicine reports *7*,

1055–1061.

63. Li, Q., Yao, L., Wei, Y., Geng, S., He, C., and Jiang, H. (2015). Role of rHOT1 on migration and proliferation of pancreatic cancer. American journal of cancer research *5*, 1460–1470.

64. Jonckheere, N., Perrais, M., Mariette, C., Batra, S.K., Aubert, J.P., Pigny, P., and Van Seuningen, I. (2004). A role for human mUC4 mucin gene, the erbB2 ligand, as a target of tGF-beta in pancreatic carcinogenesis. Oncogene *23*, 5729–5738.

65. Kufe, D.W. (2009). Mucins in cancer: Function, prognosis and therapy. Nature reviews.Cancer *9*, 874–885.

66. Piazza, R., Valletta, S., Winkelmann, N., Redaelli, S., Spinelli, R., Pirola, A., Antolini, L., Mologni, L., Donadoni, C., and Papaemmanuil, E. *et al.* (2013). Recurrent sETBP1 mutations in atypical chronic myeloid leukemia. Nature genetics *45*, 18–24.

67. Makishima, H. (2017). Somatic sETBP1 mutations in myeloid neoplasms. International journal of hematology *105*, 732–742.

68. Davis, R.E., Ngo, V.N., Lenz, G., Tolar, P., Young, R.M., Romesser, P.B., Kohlhammer, H., Lamy, L., Zhao, H., and Yang, Y. *et al.* (2010). Chronic active b-cell-receptor signalling in diffuse large b-cell lymphoma. Nature *463*, 88–92.

69. Lenz, G., Davis, R.E., Ngo, V.N., Lam, L., George, T.C., Wright, G.W., Dave, S.S., Zhao, H., Xu, W., and Rosenwald, A. *et al.* (2008). Oncogenic cARD11 mutations in human diffuse large b cell lymphoma. Science (New York, N.Y.) *319*, 1676–1679.

70. Moncada-Pazos, A., Obaya, A.J., Fraga, M.F., Viloria, C.G., Capella, G., Gausachs, M., Esteller, M., Lopez-Otin, C., and Cal, S. (2009). The aDAMTS12 metalloprotease gene is

epigenetically silenced in tumor cells and transcriptionally activated in the stroma during progression of colon cancer. Journal of cell science *122*, 2906–2913.

71. Roy, R., Louis, G., Loughlin, K.R., Wiederschain, D., Kilroy, S.M., Lamb, C.C., Zurakowski, D., and Moses, M.A. (2008). Tumor-specific urinary matrix metalloproteinase fingerprinting: Identification of high molecular weight urinary matrix metalloproteinase species. Clinical cancer research : an official journal of the American Association for Cancer Research *14*, 6610–6617.

72. Hiraiwa, N., Yabuta, T., Yoritomi, K., Hiraiwa, M., Tanaka, Y., Suzuki, T., Yoshida, M., and Kannagi, R. (2003). Transactivation of the fucosyltransferase vII gene by human t-cell leukemia virus type 1 tax through a variant cAMP-responsive element. Blood *101*, 3615–3621.

73. Hanson, R.L., and Hollingsworth, M.A. (2016). Functional consequences of differential o-glycosylation of mUC1, mUC4, and mUC16 (downstream effects on signaling). Biomolecules *6*, 10.3390/biom6030034.

74. Kelwick, R., Desanlis, I., Wheeler, G.N., and Edwards, D.R. (2015). The aDAMTS (a disintegrin and metalloproteinase with thrombospondin motifs) family. Genome biology *16*, 113–015–0676–3.

75. Somerville, R.P., Longpre, J.M., Apel, E.D., Lewis, R.M., Wang, L.W., Sanes, J.R., Leduc, R., and Apte, S.S. (2004). ADAMTS7B, the full-length product of the aDAMTS7 gene, is a chondroitin sulfate proteoglycan containing a mucin domain. The Journal of biological chemistry *279*, 35159–35175.

76. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-

wheeler transform. Bioinformatics (Oxford, England) *25*, 1754–1760.

77. Liao, Y., Smyth, G.K., and Shi, W. (2013). The subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research *41*, e108.

78. Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England) *30*, 923–930.

79. Auwera, G.A. Van der, Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. *et al.* (2013). From fastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. Current protocols in bioinformatics *43*, 11.10.1–33.