Next Generation Sequencing: Applications for the Clinic

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Kendall Winn Cradic

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Claudia Neuhauser, Ph.D. Co-advisor
George Vasmatzis, Ph.D. Co-advisor

June, 2016

Acknowledgements

This work could not have been accomplished without the help of many people. My advisors, Dr. Vasmatzis and Dr. Neuhauser have been valued coaches through this project. In addition, Dr. Grebe has been a much appreciated support and mentor. Thanks also to my other committee members, Drs. Sosa and Pankratz. Thank you to Dr. Eberhardt for providing thyroid tumor datasets.

Dedication

This is for you, Betsy. I could never express my gratitude for your never ending support. You make me a better person.

Abstract

Genomic information from the patient is becoming increasingly important for diagnosis of many diseases. Next Generation Sequencing (NGS), while commonly used as a research tool, is steadily making its way into clinical labs. One advantage of NGS is found in the observations that can be made, in addition to primary sequence, by analyzing raw data. This project is focused on the development of three such applications that have diagnostic utility. The first is a method to determine the phase of compound heterozygotes; an important problem when recessive genes contain more than one mutation. The second is a process designed to identify and interpret chromosomal rearrangements that are related to disease. And finally, the third is a technique used to calculate the copy number of mitochondrial DNA. These methods were developed for use in the clinical lab and can have a practical role in diagnosing disease.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Genomic sequence and structure are becoming increasingly important for diagnosis and management of many human diseases. The root cause of many disorders can be traced to genetic mutations allowing treatment options to be tailored to the individual. Historically, diagnostic sequencing for clinical decision making has been limited to single genes, and in most cases, only exonic regions are inspected. This is primarily due to the practical limitations of Sanger sequencing technology, in which the maximum single span of observable sequence is about 1000 base pairs.

The invention of next generation sequencing (NGS) (1) has vastly increased the capacity for researchers and clinicians to analyze the genetics of individuals. Although there are several different methods, each variety of NGS exploits the advantage of sequencing enormous numbers of short DNA fragments in parallel. Those fragments of DNA sequence, also called reads, are then re-assembled *in silico* using a reference sequence as a scaffold to reconstruct the subject's genome. To detect variations, different computational solutions have been developed that compare the specimen-derived sequence to the reference. Using these methods, several types of genetic variation can be

observed through analysis of raw data from a single sequencing assay. Small insertions and deletions (indels), single nucleotide polymorphisms (SNPs), and even large structural rearrangements can all be detected. In sharp contrast to Sanger sequencing, NGS is capable of providing several strata of genomic information from a single experiment.

Given the capabilities of whole genome sequencing, it is easy to see the utility for this technology within the medical community. Stunning successes have been reported in which whole genome or total nucleic acid sequencing has unlocked very complicated clinical cases and helped guide physicians toward more effective treatment options (2-4). Increasingly, physicians are realizing that the application of genomic information to the clinical landscape can add a rich dimension to traditional interpretation of biochemical or single gene diagnostic tests.

In addition to numerous clinical advantages, NGS also offers incredible economic benefits. The per-megabase cost of parallel sequencing has been reported to be 4 to 5 orders of magnitude less expensive than traditional Sanger sequencing (5), and rates are expected to continue their decline. Although initial capital expenses are currently high, the efficiency and ability to sequence multiple samples in parallel reduces potential costs for diagnostic scale sequencing to very attractive levels. Several clinical labs are already offering NGS diagnostic assays consisting of tens to hundreds of genes for the same price as one or two individual genes sequenced on Sanger platforms. In fact,

although it is not yet widely available, patients at specialized clinics can have their complete genome sequenced at a cost of roughly $5,000 - $10,000.

Despite its obvious advantages, massively parallel sequencing has been relatively slow to make its way into the formal clinical space. There are several reasons for this, including high capital costs, unique infrastructure requirements, and the need for specialized staffing. However, as genomic studies and personalized medicine become more commonplace, specialized equipment and trained personnel should become readily available. Perhaps the more significant bottleneck stems from the immense amount of computation required to analyze NGS data. Proper processing and examination of raw sequence reads currently requires skilled IT and bioinformatics specialists. In many cases, because they are based on rudimentary algorithms, the processing pipelines available require a great deal of human oversight and decision making. Only when the analysis can be reduced to a very dependable black-box system can NGS be widely used diagnostically without the aid of specialized bioinformaticians.

Generally, computational tools used to process NGS data have been created by academic researchers for discovery applications. For this reason, many of them are not adequate for use in clinical diagnostic procedures because they have not been rigorously validated. Also, algorithms intended for use in discovery pipelines are often tuned to allow high false positive discovery rates, at least for first pass analysis. A few commercial entities have tried to address the need for clinical grade algorithms with suites of software ready out of the box (6). However, due to the rapidly developing nature of the technology it has been hard

for them to keep up with the shifting regulatory landscape and validation of their software.

In addition to the gaps in our ability to process data, we also have the challenge of understanding what we observe. Total nucleic acid sequencing opens up a Pandora's Box of complexity that we have only begun to comprehend. While we can easily find variations, it is much more difficult to understand the clinical significance of each of them in the context of the rest of the genome. Even when complexity is greatly reduced by limiting analysis to coding exons (~ 2% of the genome) there are still over twenty-thousand variants identified in the average individual (7). Much work is still required in order to understand the full implication of each one.

Although challenges still exist, genetic and genomic evidence are expected to become increasingly significant for clinical applications. Because of the potentially profound and life-altering nature of this type of information, it is critical that analysis and interpretation be of the highest accuracy and quality possible. Many medical interventions like surgery, chemotherapy, and transplantation are commonly influenced by the results of genetic tests. As clinical interpretation continues to push the boundaries beyond primary sequence analysis and into genomic structure and relationships (transcription, regulation, isoforms, etc.) the availability of good algorithms and informatics pipelines will need to be improved. New analysis methods developed with sound statistical foundations will be increasingly required as more and more complex information is extracted from datasets.

The overarching goal of this project is to develop several new methods for using NGS data for clinical purposes. The first addresses a specific need for clinicians to overcome the difficulties in diagnosing recessive diseases in patients with compound heterozygosity. The second is a method that uses mate pair sequencing for rapid analysis and interpretation of structural variations in individuals or cohorts of associated specimens. This method is intended for a clinical setting where speed and accuracy are crucial. Finally, the third application is a method to determine the clinical utility of mitochondrial copy numbers as a marker of disease. The contribution of these methods to clinical diagnosis will help physicians treat disease more rapidly and with greater precision.

Chapter 2: A simple method for gene phasing using mate pair sequencing

Authors: [1]Kendall W. Cradic, [2]Stephen J. Murphy, [3]Travis M. Drucker, [4]Robert A. Sikkink, [5,6]Norman L. Eberhardt, Claudia Neuhauser, [2]George Vasmatzis, [1]Stefan K.G. Grebe

Affiliations: [1]Department of Laboratory Medicine and Pathology, Mayo Clinic & Foundation, Rochester, MN 55905, [2]Department of Molecular Medicine, Mayo Clinic & Foundation, Rochester, MN 55905, [3]Information Technology, Mayo Clinic & Foundation, Rochester, MN 55905, [4]Adavanced Genomics Technology Center, Mayo Clinic & Foundation, Rochester, MN 55905, [5]Department of Medicine, Division of Endocrinology, [6]Department of Biochemistry and Molecular Biology, [7]Biomedical Informatics and Computational Biology, University of Minnesota Rochester, Rochester, 111 South Broadway, Suite 300, Rochester, MN 55904.

2.1: Synopsis

Compound heterozygosity is a significant problem for researchers and clinicians. Of particular concern are patients who have two heterozygous disease-causing mutations and could be diagnosed as affected (one mutation on each allele) or as phenotypically normal (both mutations on the same allele). Several methods are available to phase genes, however due to cost, complexity and/or low sensitivity they are not suitable for clinical purposes. We have developed a simple method utilizing massively parallel sequencing that is capable of resolving haplotypes. This method will simplify interpretation of complex clinical cases and eliminate additional workup, including lineage studies and allele-specific PCR.

2.2: Background

The use of diagnostic gene sequencing has dramatically increased during the last two decades. However, accurate interpretation of sequencing data remains a challenge, despite technical advances. One common problem is uncertainty about the *cis/trans* status, or phase, of heterozygous variations. Properly phased genomic information is frequently required for accurate diagnosis of recessive genetic diseases. The scale of this problem is considerable, as indicated by a recent query of the Online Mendelian Inheritance in Man (OMIM) database which revealed over 250 recessive genes known to be

associated with more than 1,100 disorders (8). Unfortunately, Sanger sequencing, the most widely used technique and current gold standard, is incapable of separating phases without allele-specific capture or allele-specific amplification.

While this problem has long been recognized, a simple and effective solution has remained elusive. Computational methods have been developed to estimate haplotype sequences based on the individual's genotype compared to a population (9), but they lack the resolution and accuracy needed for clinical use.

A more definitive approach for genetic phasing is based on manipulation of single chromosomes, either through cell hybrid systems, using conversion technology (10, 11), or by means of size-exclusion devices (12). While this strategy is perhaps the most reliable for generating accurate haplotype sequences, it is by far the most labor intensive approach. It is also error and failure prone, due to its lengthy, complex and technically difficult workflows.

More recently, the phasing problem has been tackled using massively scaled Next Generation Sequencing (NGS). Briefly, these methods depend on the creation of at least 100 libraries from each patient using techniques such as bacterial fosmid construction or multiple displacement amplification (13, 14). Libraries are indexed, pooled, sequenced and then computationally combined into two haplotype consensus sequences. While these methods are powerful for generating phased sequences for entire genomes, they are cumbersome, slow and currently expensive.

Since each of these approaches is in some way unsuitable for routine clinical use, current protocols for solving *cis/trans* questions typically involve testing of family members. This is a costly and time consuming undertaking that may still fail, if there is insufficient genetic diversity in the tested familial cohort. As an alternative, allele-specific PCR can be employed. However, the cost and effort required to design and validate assays makes this prohibitive in genes where there are many possible combinations of mutant positions.

Revisiting NGS techniques, with a view to creating a simpler solution than multiple indexed library sequencing, could provide an attractive solution to the phasing problem, in particular as NGS is now starting to replace Sanger sequencing in clinical applications. Because NGS methods are based on deriving sequences from a single molecule, one should be able to adapt the methodology for accurate phasing of genomic sequences. Most of the current platforms use a paired end (PE) protocol in which a string of sequence is read from either end of a larger DNA fragment. Since the reads come from opposite ends of the same fragment and are linked through a continuous strand of DNA, we refer to them as linked reads. Given their linked nature, any variations detected in the same fragment are *cis* to one another.

The current Illumina PE library sequencing protocol restricts library fragment size to 250-500 bp because longer fragments decrease the quality of data through overlapping and reduced density of clusters. Coverage of larger distances between nucleotide positions of interest can, however, be achieved through the mate paired (MP) library protocol. This protocol initially utilizes larger

genomic fragments of 2-5 kb that are self-ligated prior to a secondary fragmentation to the conventional PE library size centering on 500 bp (Figure 2-1). Biotinylation of the termini of larger fragments prior to circularization enables the isolation of DNA containing the ligated ends. Sequencing of these fragments containing junction points thus generates paired reads that are linked across much greater distances than in conventional PE libraries, at the expense of some loss in coverage for short inter-variant distances. A combination of PE (100 - 600 bp) and MP (500 – 5,000 bp) libraries over a defined gene region could therefore complement each other in terms of phased coverage and should allow accurate determination of *cis/trans* status of multiple sequence variants over a relatively large range of distances.

We tested this supposition using the *CYP21A2* gene as a model system. This gene is commonly sequenced during diagnosis of congenital adrenal hyperplasia (CAH). The combination of the modest length of this gene (~3400 bp), a rate of at least 10% compound heterozygosity for mutations or variants of unknown significance in patients, and availability of genetic family studies in most cases, make *CYP21A2* a suitable model system as a proof of principle test of our approach.

2.3: RESULTS

To demonstrate that a combined PE and MP sequencing strategy could

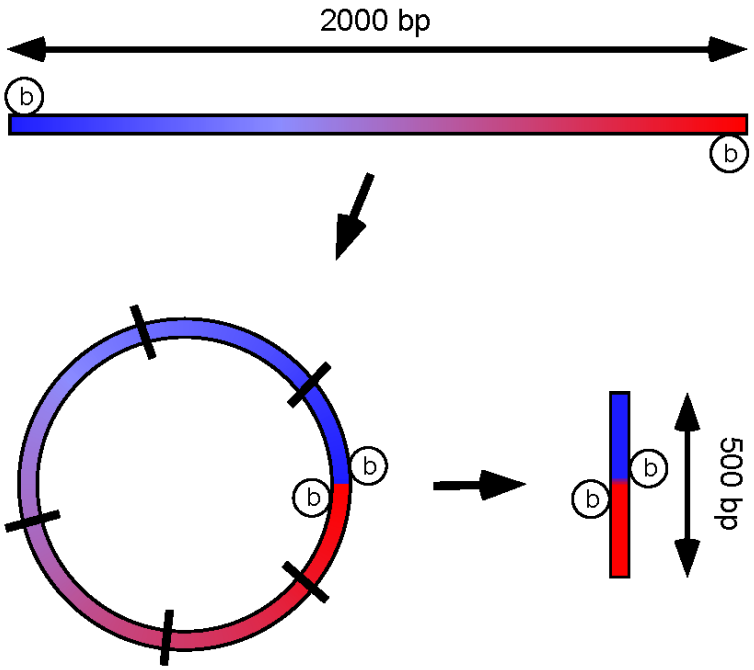

Figure 2-1: Mate Pair library preparation.

The MP protocol allows sequence information to be linked across greater distances than PE reads. Fragments averaging 2000 bp from a pool of sheared DNA are end-repaired using biotinylated nucleotides. Fragments are then self-ligated and all remaining linear DNA is removed by exonuclease treatment. Circularized DNA is fragmented again (black bars) to an average size of 500 bp and segments containing biotinylated junction points are isolated on streptavidin beads. In addition to fragments containing junction points, a portion of non-biotinylated DNA is co-purified and appears in the MP library as a subpopulation of PE reads. All fragments are end-repaired and indexed using TruSeq adapters followed by sequencing.

allow us to accurately phase compound heterozygous sequence variants over a significant genomic distance, we divided the problem into three components. First, we performed experiments to determine the necessary conditions for adequate sequence coverage and showed proof of principle of accurate variant phasing, using *CYP21A2* as a model system. Next, we demonstrated that the analysis can be extended to phase DNA fragments across distances that are much larger than those included in the MP library. Finally, we explored the principle sources of experimental error.

Phasing a single pair of heterozygote sequence variants

Confidence of NGS base calls is a function of coverage at a given position. Since our strategy requires accurate association of two heterozygous positions (a total of 4 base calls), high coverage is required throughout the target region. To this end, we designed a long range-PCR (lrPCR) for enrichment by amplification of the active gene *CYP21A2*, while excluding its highly homologous pseudogene, *CYP21A1P*.

While enrichment boosts coverage, it also increases the likelihood that two fragments of DNA from opposite *CYP21A2* alleles will be ligated together during MP library construction. This event would generate false *cis* associations between loci. We reasoned that we could reduce the probability of inter-allelic recombination by adding an excess of background genomic DNA to the gene specific lrPCR product, biasing any recombination towards non-target sequences. Libraries made with 10, 100, 500 and 1000 ng of lrPCR product

produced sequence coverages of 1,600X, 10,900X, 60,400X and 130,500X, respectively. By contrast, coverage by MP fragments outside of the amplified target region averaged slightly less than 2X.

Linked reads are of even greater importance for phasing than raw coverage is for accuracy. Any linked read method for phasing needs to generate an extended distribution of fragment sizes. This assures enough depth of coverage between any two points within a gene to accommodate a broad range of potential distances between heterozygous positions. To verify that we had achieved this goal we calculated the linked coverage in our NGS data as a function of distance Δ between base positions. For every position x in the amplicon, we counted the number of linked reads covering both x and $x + \Delta$, for Δs from 101 to 3000 bp, and then calculated and plotted the average linked coverage. Paired end libraries provided linked reads up to 500 bp while MP libraries produced a population of fragments ranging from about 200 bp to over 3000 bp (Figure 2-2).

Previous genotyping of the specimen tested here showed two heterozygous disease-causing mutations; however their phase was not clear from the Sanger sequences and required family studies. The first mutation, c.60G>A introduces a stop codon at amino acid position 20. The second, IVS2-13A>G is a common splice site mutation in intron 2. Both variants produce truncated proteins and are associated with the classical form of CAH.

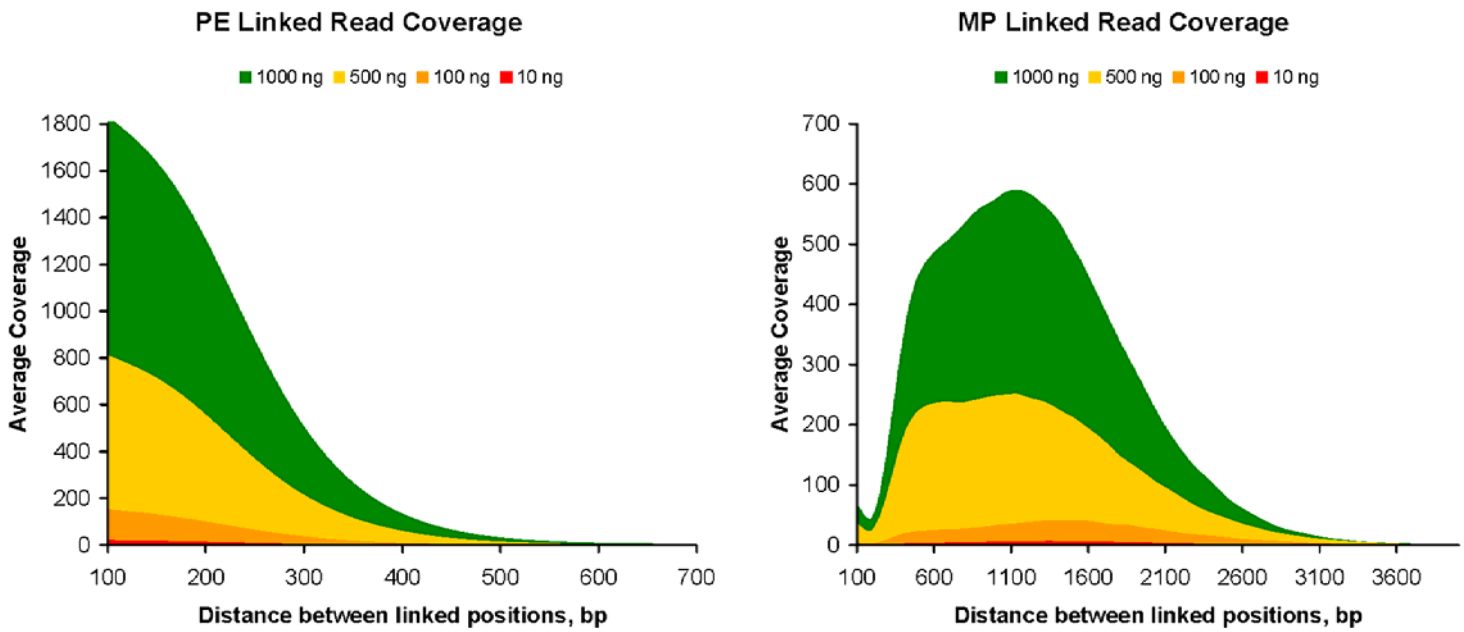

Figure 2-2: Average linked coverage by PE and MP reads from four libraries.

Linked read coverage is shown from PE and MP reads as a function of distance between linked positions for each of the enriched libraries (10, 100, 500 and 1,000 ng spiked lrPCR amplicon in WGA DNA).

After mapping all of the reads in the library, fragments were selected that covered both heterozygous positions with their pairs of sequence reads. The base calls at each heterozygous position from each fragment were observed and compared to establish the relationship between the two alleles. Using these base calls, an association matrix was constructed to measure the frequency of each association (Figure 2-3a). In each library, the wild-type G at position c.60 was most frequently associated with a mutant G in the IVS2-13 position. Conversely, the mutant A at position c.60 was most frequently associated with the wild-type A at IVS2-13. This indicated that the two mutants were on opposite alleles, a result that was congruent with the conventional phased genotype that had previously been established through allelic segregation studies of the proband's family.

The next step was to quantify more precisely how confident one could be that the *trans* phasing result was correct. We used bootstrapping for this, calculating 99% confidence intervals around the probability of each possible downstream base call. The width of the confidence intervals therefore, is related to the depth of linked coverage between the two mutant sites (Figure 2-3b). To clarify this relationship, we ran simulations for varying amounts of coverage using probabilities from a single dataset (500 ng amplicon spike) and calculated the average width of all resulting confidence intervals. Both the simulation and observed confidence intervals indicate that coverage above 500X provides diminishing returns in phasing confidence (Figure 2-3c).

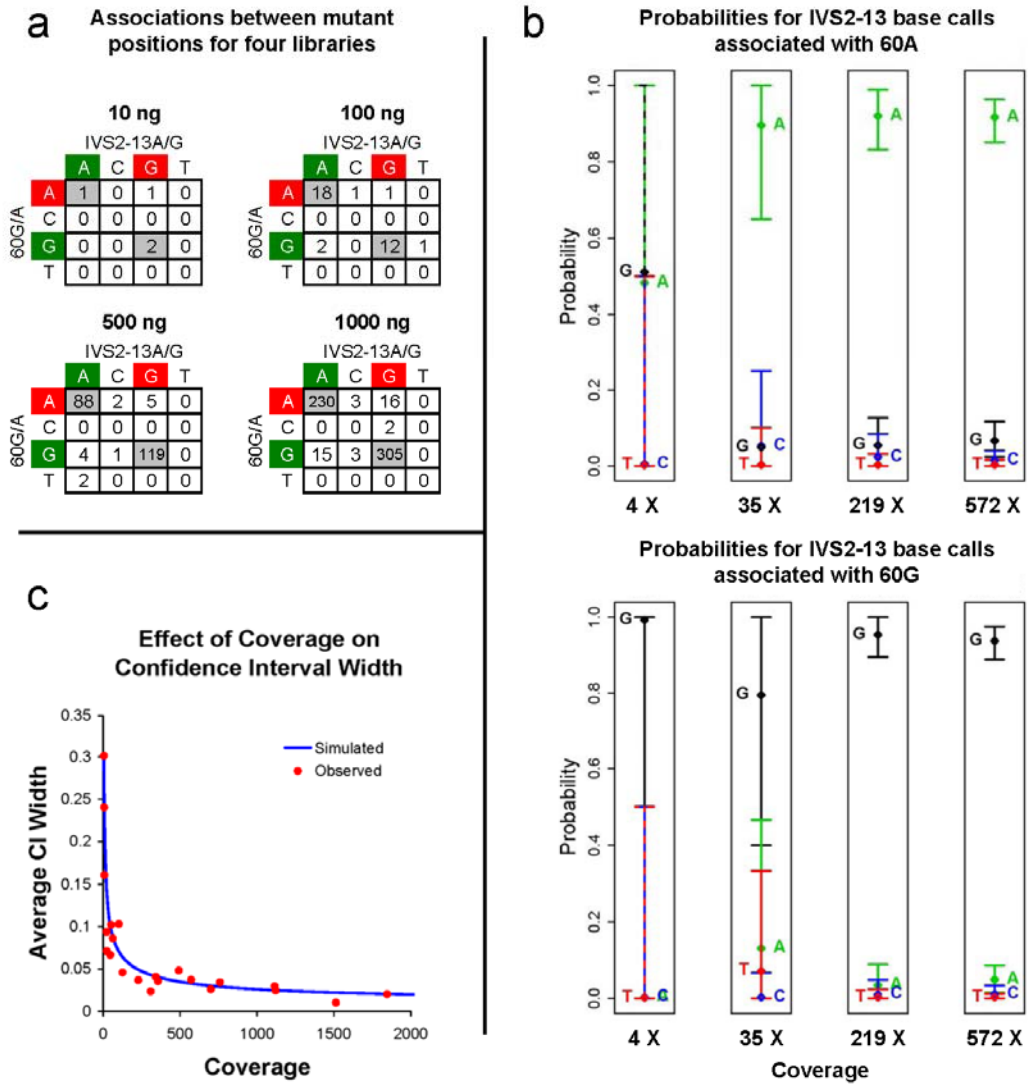

Figure 2-3: Confidence in phasing calls is dependent on coverage.

(a) Association matrices from each spiked library show the relationship between the two disease-causing heterozygous positions in this specimen. Highlighting of the bases at each locus indicates wild-type (green) and mutant (red). (b) The probabilities and 99% confidence intervals for all possible IVS2-13 base calls associated with each c.60G/A allele are shown for all four amplicon spiked libraries. Coverage increases linearly as spike concentration increases. (c) Average confidence interval (CI) width was calculated to evaluate the level of coverage required for confident heterozygote association. A simulation run to test varying coverage and observed data points both indicate that coverage beyond 500X provides diminishing returns in CI width.

Extending the method over longer genetic distances

In our test specimen, the two mutant positions were well covered by a subset of PE and MP fragments. However, it is likely that in some cases (or in different genes) heterozygous mutations will be separated by more than 2000 bases. For these situations, we have developed a computational method to chain together linked reads by constructing association matrices between pairs of several heterozygous sequence positions (normal sequence variants, VUSs. or mutations) in tandem through the length of the amplified region. Provided there are enough heterozygous positions in the specimen that fall within the limits of the combined MP and PE libraries, the entire amplified region can be phased using this iterative approach. Statistical analysis of the phase assignment across an entire chain of linked sequence variants is identical to the single association matrix, except that a cumulative probability and confidence interval is calculated between the two mutant positions to measure confidence in the data used to link the two. Since this cumulative measure is the product of all upstream probabilities in the chain, its value will decline in proportion to the amount of error in each association matrix. This diminishes the probability of the final overall phase-call in relation to the first. However, as long as there is full separation of the confidence limits of the final cumulative phase determination from all other possibilities a confident call can be made. It is thus possible to extend the phasing chain across the entire 10 kb lrPCR amplicon without any overlap of confidence intervals, indicating accurate phasing throughout (Figure 2-4).

Sources of error in the association matrix

Each association matrix contains a small percentage of incorrectly linked bases. The sources of error in NGS datasets have been previously explored and attributed principally to detection error during data acquisition, fluorescence spectral overlap and computational misalignment of reads in highly homologous regions (15, 16). Since our protocol includes lrPCR enrichment followed by MP library preparation, we also had to consider the contribution from *in vitro* recombination events that occur during amplification or circularization.

While a thorough investigation of this type of error is beyond the scope of this paper, we were able to quantify two types of inaccuracy by constructing association matrices between every pair of heterozygotes in the *CYP21A2* gene. Recombination events, i.e. MP reads that include a base from either allele, were the most common source of error. As a percentage of total reads, this type of error averaged about 7% and it proved to be constant across every combination of PCR product and background DNA mix (5.4%, 7.4%, 6.2% and 7.2% error for 10, 100, 500 and 1000 ng of amplicon input, respectively). In addition, there was no change in these error rates as a function of linked read length or coverage. This indicates that our initial assumption was incorrect; ligation of fragments from opposite alleles during MP library preparation did not prove to be a major contributor to erroneous base or phasing calls. Furthermore, because the percentage of recombination does not change with proportion to the amount of

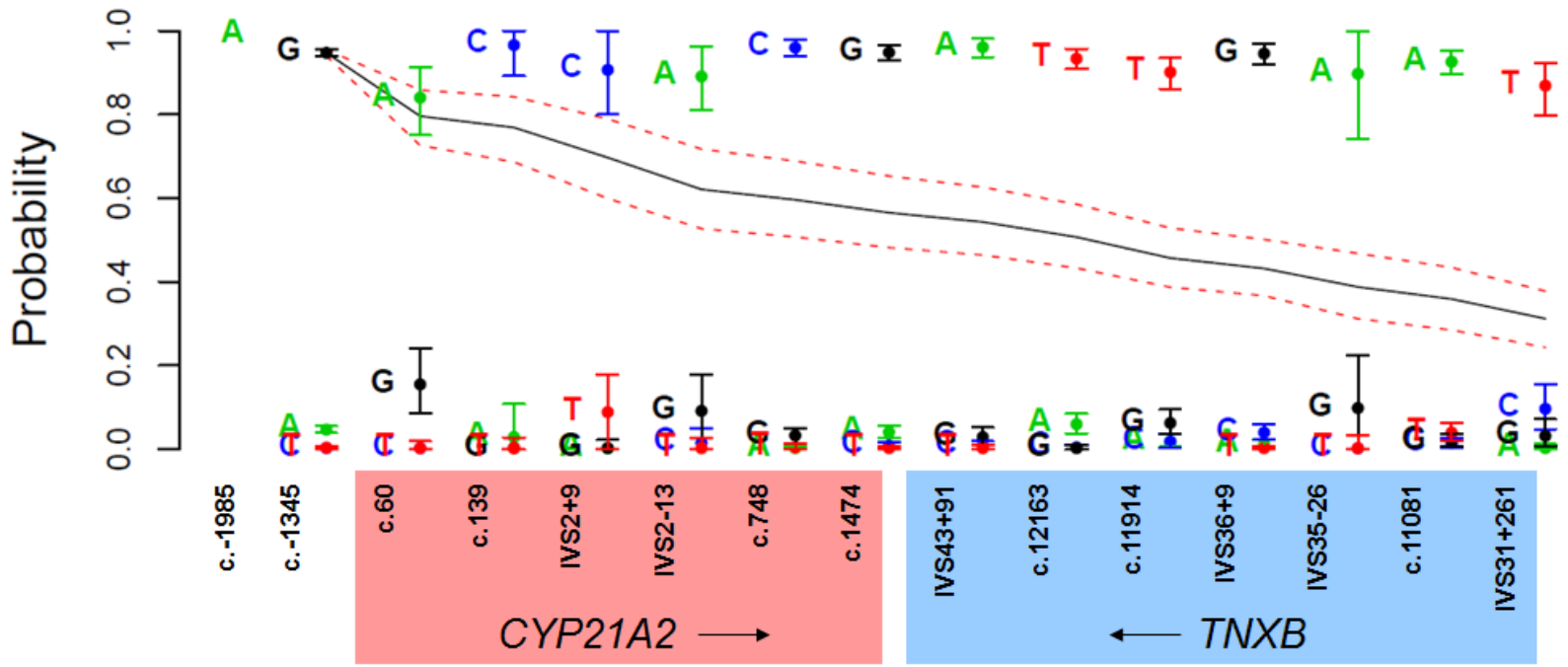

Figure 2-4: One phase of the entire 10 kb amplicon.
By beginning with the first heterozygote in the amplicon and sequentially moving through all downstream heterozygous positions, the phase of the entire 10 kb amplicon can be determined. Confidence intervals in the columns show the relationship of each base to the highest probability base call from the previous column. Lines showing the cumulative probability and confidence interval relate each downstream position to the very first in the chain. Cumulative probability diminishes in proportion to the quality of each association matrix in the chain (i.e. sufficient coverage and few errors). However, one can be sure of the accuracy of phasing so long as there is no overlap between the cumulative confidence interval and that of any rejected base.

amplicon spiked into each library, these events must occur prior to library creation, i.e. during lrPCR. These observations testify to the reliability of the MP library protocol and highlight the importance of high fidelity in PCR reactions.

Finally, some incorrect base and/or phasing call errors could not be attributed to recombination artefacts. Across all heterozygous pairs analyzed in our data, only 2% of the total reads fell into this category, a value that accords with other reported values for random error in NGS data (5, 17).

2.4: DISCUSSION

Using our MP library approach and subsequent computational analysis we have been able to successfully haplotype a specific region of interest in an individual who had two heterozygous disease-causing mutations. This method is an improvement over other available phasing protocols because of its simplicity and because of the statistical measure of assurance it provides. In regions where coverage is low or where recombination is present in the fragment library, erroneous phasing calls can easily be made by other methods. In addition to these advantages, our method provides the ability to phase heterozygous positions that are thousands of bases apart.

Performed as a single protocol, this method is capable of acquiring a completely phased genotype for an entire 10 kb lrPCR amplicon. Target regions of this size can be routinely amplified, and 20-30 kb amplicons are achievable in

many instances. In principle, this method could also work beyond 10-30 kb, if several overlapping lrPCR amplicons are used as starting material, and as long as sufficient overlapping MP fragments can be generated that share heterozygous positions. An average MP library size which exceeds the 2 kb observed in our study would be expected to improve the likelihood of finding an unbroken linked chain of polymorphisms, while simultaneously reducing the number association matrices needed for complete phasing of a region of interest, thereby improving the confidence in the accuracy of the overall haplotype. Since the Illumina MP protocol is optimized for initial fragmentation libraries of 2-5 kb, such improvements should be relatively easy to achieve.

In theory, there is no upper limit to the scalability of our approach and it could even be applied to whole genome sequencing, provided sequence coverage and linked coverage are high enough. Without regard to logistic or cost considerations, we speculate that this technique might actually be very successful in this setting, because the error attributable to inter-allelic MP ligation proved to be very low. Nevertheless, it is likely that one would have to break down the analysis of an entire genome into smaller haplotype units, in order to maintain high confidence of the phase calls. We would anticipate that the size of these units would be similar to what can be achieved by optimal combinations of lrPCR and MP protocols, as described above.

Two limitations that we foresee for accurate phasing are highly homologous genes and gene duplications or other copy number changes. In either of these cases, we would anticipate phasing errors to increase due to mis-

assignment of reads. In addition, increases in gene copy number would exponentially increase the number of possible phase combinations for any given combination of polymorphic positions, increasing computational requirements and decreasing ultimate haplotyping accuracy, and in some cases, phase assignments might be impossible.

2.5: CONCLUSIONS

In summary, compared with previous approaches, our MP NGS sequencing technique is a simple solution to the problem of accurately phased genotyping for many recessive diseases, and perhaps, many other genetic phasing problems. The method could be adapted to other NGS platforms since they are all based on deriving sequences by aligning large numbers of overlapping reads. As clinical molecular diagnosis rapidly approaches massively parallel sequencing as the preferred assay method, it could serve as a cost-effective way to obtain a completely resolved set of haplotypes for single genes, panels of related genes, or even significant portions of chromosomes.

2.6: MATERIALS AND METHODS

Long-range PCR

CYP21A2 is located in the HLA region on chromosome 6p2.13. An inactive yet highly homologous pseudogene (*CYP21A1P*) is located 30 kb upstream and has been known to confuse genotyping assays for *CYP21A2* (18, 19). To enrich our mate pair library with the active gene and eliminate the pseudogene we performed long-range PCR (lrPCR) using unique priming locations around *CYP21A2*. Priming sequences were 5'-AGTGGGGCTCTGAAGACTGA-3' for the forward position and 5'-CCCTCGGGAGATGATCTGTA-3' for the reverse to amplify a clean 10 kb product (Figure 2-5). LA Taq and associated buffers from TaKaRa were used in the reaction at their recommended concentrations. Approximately 150 ng of template DNA was used in the PCR reaction. Cycle conditions were as follows: 95 °C for 5 m; 10 cycles of (95 °C for 30 s, 60 °C for 30 s, 72 °C for 10 m); 20 cycles of (95 °C for 30 s, 55 °C for 30 s, 72 °C for 10 m); 72 °C for 20 m.

Whole Genome Amplified Background DNA

The MP protocol is driven towards intra-molecular circularization over inter-molecular ligation of two separate DNA fragments simply by spatial dilution. In order to minimize the complication of inter-fragment ligations from a limited sequence amplicon input we investigated spiking of the 10 kb amplicon at four

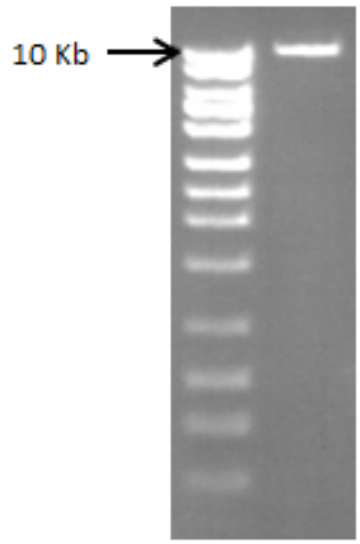

Figure 2-5: Clean 10 Kb amplicons.

A single, clean band at 10 Kb shows the specificity of our long-range amplification.

different concentrations into a background of whole genome amplified (WGA) DNA from a normal individual. WGA DNA was used due to its similar average fragment size to the 10 kb lrPCR amplicon, than conventional extracted genomic DNA preparations (~50 kb), making downstream fragmentation in the library prep protocol more predictable. This approach additionally enabled us to evaluate the role of amplicon concentration on inter-fragment ligation. Background WGA DNA was generated from genomic DNA using a Qiagen Repli-g midi kit according to recommended protocols. Background and amplified DNA concentrations were measured by fluorescence on a Qubit fluorometer (Invitrogen), and 10, 100, 500, and 1,000 ng aliquots of lrPCR product were spiked into WGA DNA to a total of 5 ug for each library preparation.

Library Preparation and Sequencing

MP libraries were prepared for each spiked pool of lrPCR product and WGA background DNA based on previously reported protocols (20). Each pool was fragmented on an M220 Focused-ultrasonicator (Covaris) to fragments ranging from 500 to 5000 bp with an average of 2000 bp. Following purification on Qiaex II beads, DNA fragment ends were repaired and biotinylated using a mixture of natural and biotinylated dNTPs. Excess reagents and by-products were removed using Qiaex II beads. Six-hundred ng of DNA from each pool were circularized in 16 hour ligation reactions at 30 °C prior to exonuclease treatment at 37 °C for 20 minutes to digest any remaining linear strands of DNA. The circularized DNA was then fragmented to 300 - 500 bp using the M220

Focused-ultrasonicator. Streptavidin beads were applied to isolate ligation junction fragments. End repair, blunt ending and adapter ligation were performed while fragments were bound to the beads. PCR was performed to produce bead-free fragments which were subsequently assembled into indexed MP libraries using TruSeq adapters (Illumina). While streptavidin beads provide good recovery of biotinylated DNA, they also co-purify a fraction of unlabeled fragments from other locations in the sheared, circularized DNA. We used this to our advantage by allowing these fragments into our libraries to provide PE reads covering positions 100 to 500 bp apart.

The four final indexed MP libraries were purified and analyzed on an Agilent Bioanalyzer DNA 1000 chip before equimolar pooling. The sample was loaded onto a single lane of an Illumina flow cell and sequenced to 101x2 paired-end reads on an Illumina HiSeq. Base calling was performed using Illumina Pipeline v1.5.

Sequence reads collected from the Illumina were demultiplexed and mapped to the hg19 assembly (21) using a custom mapping algorithm similar to the one used in previous publications (22, 23). To avoid the problem of reads from the amplified region erroneously mapping to the pseudogene, *CYP21A1P*, and/or homologous surrounding areas, the region from chromosome 6 between 31971000 and 31982000 was removed from the reference sequence.

Statistical Analysis

After mapping and alignment of linked reads covering two heterozygous positions a matrix was constructed to quantify the associations between every possible pair of base calls between the two positions. Confidence intervals for all base calls were calculated by bootstrapping based on the observed frequency of base calls in each association matrix. For each upstream base call (association matrix rows), a probability distribution was constructed for all possible downstream base calls (association matrix columns). Observed counts in each row were converted to probabilities and used for multinomial resampling with the total number of samples set to the sum of observations in the row. In addition to the observed probabilities, 1% was distributed across each row to simulate random error associated with NGS sequencing. Following every cycle of sampling, the counts for each base call were converted to probabilities and used to construct a set of distributions. After 1000 sampling iterations, confidence intervals were set for each possible downstream base call by ranking the resulting probabilities for that base and selecting the 1% and 99% values from the distribution.

For haplotyping regions longer than the span of PE or MP fragments several association matrices can be chained together. In this case, bootstrapping for each individual matrix was performed as described above. The linkage between pairs of heterozygous positions followed a Markov Chain model in that the probability of association between two base calls was unrelated to previous base calls in the chain. To begin the chain, association matrix A_1 was

constructed between two heterozygous positions, h_0 and h_1. One of the two bases was arbitrarily chosen from h_0, and probabilities and confidence intervals for each base at h_1 were calculated as described above. Next, association matrix A_2 was constructed between positions h_1 and h_2. The base with highest probability at h_1 from A_1 was selected and probabilities for association with this base at h_2 in A_2 were calculated. By iteration of this cycle, a chain of associated base calls can easily be made for one allele. To validate the results from one allele, the opposite allele can be phased by selecting the alternate base at h_0 and crosschecking the two resulting chains.

To quantify the confidence of association between two distant heterozygote calls, a cumulative probability was calculated as the product of all prior probabilities in the associated chain. Cumulative confidence intervals were also calculated from a distribution made from the products of each previously occurring bootstrap result. Using these measures, the limits to the length of chained phasing become apparent when the confidence intervals of rejected base calls begin to overlap with the cumulative interval.

Chapter 3: Clinical validation of a haplotyping method with next-generation sequencing

Authors: [1]Kendall W. Cradic, [2]Stephen J. Murphy, [3]Robert A. Sikkink, [4]Claudia Neuhauser, [2]George Vasmatzis, [1]Stefan K.G. Grebe

Affiliations: [1]Department of Laboratory Medicine and Pathology, Mayo Clinic & Foundation, Rochester, MN 55905, [2]Department of Molecular Medicine, Mayo Clinic & Foundation, Rochester, MN 55905, [3]Adavanced Genomics Technology Center, Mayo Clinic & Foundation, Rochester, MN 55905, [4]Informatics Institute, University of Minnesota Twin Cities, Minneapolis, MN 55455.

3.1: Synopsis

Background

Compound heterozygosity has been a significant problem for clinical interpretation of genetic results, particularly for recessive disorders. Traditionally, laboratories have relied on family studies or development of allele-specific PCR assays to determine *cis/trans* arrangements between compound heterozygous mutations. Next Generation Sequencing has provided the means to collect sequencing and phasing information from the same experiment. However, the haplotyping algorithms currently available are not suitable for clinical use.

Methods

We have developed a statistical method to phase compound heterozygote mutations based on a directed network and Markov Chain analysis. This method uses all observed data and provides a score indicating when sequencing quality is low, whether due to poor read coverage or erroneous sequence data. This is an improvement over other methods that either ignore errors or do not calculate a score to warn users of potential misinterpretation.

Results

We validated this method using a cohort of samples that had previously been tested for *CYP21A2* mutations and found to have at least two heterozygous variants. In every case, family members had been tested to provide

unambiguous haplotypes. The new method, using Next Generation Sequencing, was able to provide phased sequence results directly from the proband's sample, eliminating the need for additional workup.

Conclusions

Our method provides a useful yet simple solution to a practical problem faced by clinical sequencing labs. By using it, laboratories can obtain haplotyped sequence and provide confident clinical interpretation without allele-specific PCR or family studies.

3.2: Introduction

Clinical interpretation of genetic testing results can be difficult or impossible when compound heterozygosity is present, particularly for recessive disorders. Unfortunately, Sanger sequencing cannot determine the *cis/trans* relationship, or phase, between compound heterozygous variants. Historically, this problem has led to a multitude of secondary follow-up assays, such as custom designed allele specific priming PCR assays and family studies, putting additional strain on diagnostic laboratories and genetic counselors. In addition, even with extensive family studies it is occasionally difficult or impossible to resolve the phase of compound heterozygous sequence variants in a proband, in particular if the individual belongs to a community with reduced genetic variation.

Next Generation Sequencing (NGS) is being quickly adopted by clinical labs as a cost effective first-tier method for high throughput sequencing of gene panels or exomes. Current 'paired-end' NGS methods allow the concurrent sequencing of the distal ends of individual DNA fragments, which are then simultaneously mapped to a reference genome. By applying appropriate bioinformatics tools, the phase of multiple sequence variants can be detected from the paired sequencing reads.

Several classes of such phasing solutions exist, which have been developed to elucidate complete genomic structure (13, 24-26). Virtually all fall short of clinical needs and only provide purely qualitative phase assignments that

offer no quality metrics that can assess the validity (or questionable validity) of each *cis* or *trans* call.

We have recently developed a phasing method that addresses these problems and have shown proof of principle for accurate phasing of clinical samples over long distances with reliable confidence scores for each variant (27). We have since modified this approach to further improve certainty of phase calls by using a read-backed method using the paired sequences. We validated our method using a cohort of samples with compound heterozygosity in the *CYP21A2* gene.

Mutations in *CYP21A2* cause >90% of cases of congenital adrenal hyperplasia, a recessive disease which is one of the most common inborn errors of hormone metabolism (28). *CYP21A2* lies close to a largely identical pseudogene, *CYP21A1P*, with which it might undergo recombination or gene conversion events. These can introduce multiple, potentially deleterious sequence changes into *CYP21A2*. Roughly 10% of our clinical testing population of potentially affected cases is compound heterozygous for either two known mutant loci or one known mutant and one, or several, variants of unknown significance (VUS). These specimens nearly always require additional studies including testing of family members, often at the laboratory's expense. For this study, we show that for a cohort of 13 patients who had undergone sufficient family studies to allow unequivocal phasing by conventional methods, NGS data when combined with our phasing algorithm produces accurate *cis/trans* calls

directly from the proband's analysis, eliminating the need for additional sequencing of family members or custom-made allele-specific PCR assays.

3.3: Materials and Methods

SPECIMENS

We evaluated 13 specimens that had previously been tested, based on the presence of multiple heterozygous disease-causing mutations or VUSs, with indeterminate phase. Correlation with the results from biochemical analysis (steroid profiles) and known familial mutations, had allowed tentative phasing in about half of the patients. However, for the remainder of specimens reported, our initial clinical reports indicated that we could not distinguish the *cis/trans* relationship between mutations and that we recommended follow up testing to identify the correct gene structure and health related implications. In each such case, at least one family member was consented and tested. In addition, we had also employed other molecular techniques (copy number variation assessment or allele-specific PCR) in one case. The studies were considered IRB waived, as they were solely used for a phasing method comparison.

MATE PAIR SEQUENCING

Construction of mate pair libraries from long-range PCR amplified regions of chromosome 6 was accomplished using a Nextera Mate Pair Sample Preparation Kit (Illumina, San Diego, CA). Library assembly protocols and DNA

inputs for required coverage, as well as the need for a distribution of fragment lengths for concomitant coverage of all heterozygous positions in the gene were reported in detail in our previous publication (27).

Indexed libraries were pooled, loaded on two lanes of an Illumina flow cell and sequenced to 101x2 paired-end reads on an Illumina HiSeq. Base calling was performed using Pipeline v1.5 (Illumina). Sequence reads were collected and mapped to the hg19 assembly (21) using BIMA V3, a mapping algorithm designed especially for mate pair fragments (29).

COMPUTATIONAL ANALYSIS

Mapped reads were realigned using the Smith Waterman method (30) and heterozygous positions were determined. We then constructed a matrix in which each column represented a heterozygous position and each row represented a read that contained sequence calls in at least two of the columns (Figure 3-1). Since paired sequence reads come from a single DNA fragment, each row of this matrix is assumed to represent a *cis* relationship between two heterozygous positions. By counting the number of relationships between each combination of bases in the matrix we formed the Association Matrix A, a record of every read pair in the dataset that provides coverage to more than one heterozygous position (Figure 3-1).

Using this Association Matrix, probabilities of cis-relationships for each pair of loci can be calculated and a directed network can be constructed, where

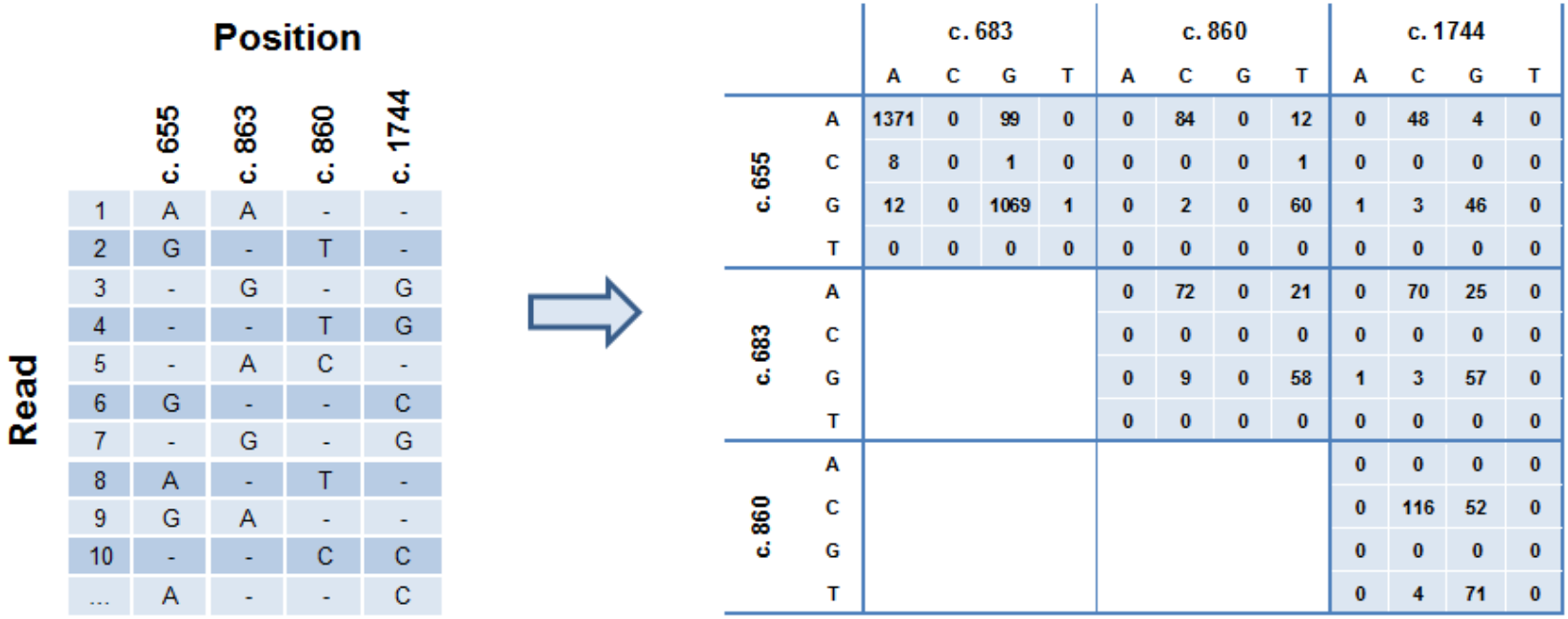

Figure 3-1: Data from all sequencing reads that cover at least two heterozygous positions are collected into a matrix following local realignment (left). Using this data, the Association Matrix A is constructed (right) by counting the base calls from different loci that come from the same DNA fragment. By definition these base calls are *cis* relative to each other.

each node represents a single base from the dataset. The edges connecting nodes represent the probability of a *cis* connection between the two. Bases that are not represented empirically in the Association Matrix are ignored in the directed network. Thus, all base calls from each heterozygous position are represented in the network and their relationships to other positions are weighted as a reflection of their proportion in the raw data (Figure 3-2).

This network arrangement allows us to utilize a Markov Chain strategy (31) to traverse the network from any given starting position and determine the probability of landing on any other downstream position. In taking this approach, we build the network such that the starting node is the wild-type allele of the upstream locus. The rest of the network is constructed in layers representing heterozygous alleles between the two positions in question (Figure 3-2). A process that starts at the top of the network and moves from node to node based on edge probabilities will terminate in the lowest level of the network on a node that is predicted to be in *cis* with the starting base. However, since error can be introduced into the sequencing data during library prep and amplification, there will be some chance that the terminal node is incorrect. To determine the confidence level of the terminal base call, the network can be traversed many times to establish an average outcome. Then, the network can be resampled and traversed again. Repeating this process many times results in a

set of probability distributions, one for each node represented in the lowest

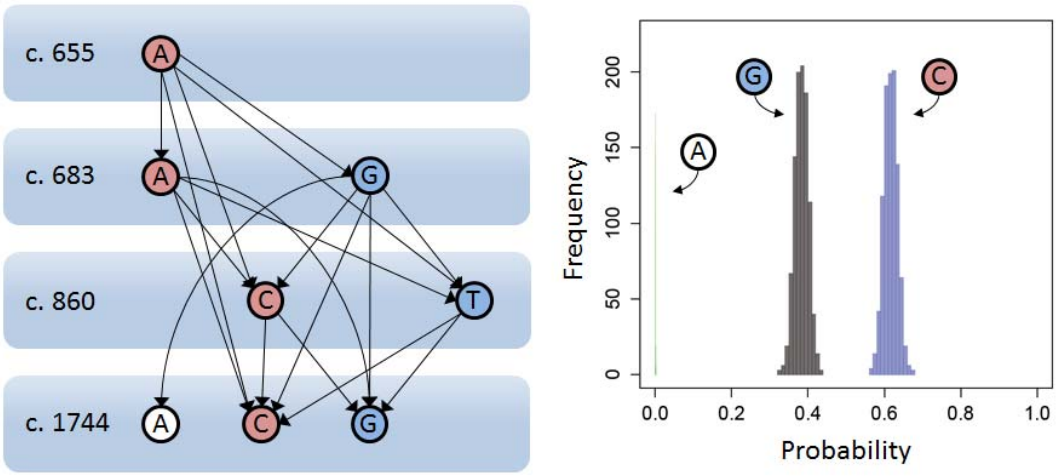

Figure 3-2: A directed network is constructed using probabilities calculated from the Association Matrix. Each node represents one allele and each edge is the probability of a *cis* connection. The network shown is constructed from a string of four heterozygous loci as represented by the four layers. Bases (nodes) from one chromosome are shown in red and the opposite in blue. Markov Chain analysis begins with the wild-type allele in the top layer and traverses the network based on the probabilities leaving each node. The process terminates on a single node in the lowest level of the network. After each traversal of the network, the matrix is resampled and the process repeated up to 1000 times to generate probability distributions (right). These results indicate that the A allele at c.655 is in *cis* with C at c.1744.

level of the network (Figure 3-2). Each distribution represents the probability of one particular base existing in *cis* with the initial allele from the top of the network. If there is no overlap on the highest probability distribution, we can be confident that this base is on the same chromosome as the initial allele.

PROBABILITY DISTRIBUTIONS

In an errorless dataset, the probability distributions generated from the directed network would be completely contained within the points of 100% and 0% for *cis* and *trans* respectively. However, as error increases through recombination events or misalignment of reads, the distributions will both eventually converge on a mean of 50%. In addition, as coverage decreases, the distributions become broader. Both of these conditions increase the likelihood of overlap between probability distributions, making an unambiguous phase call difficult.

To quantify the confidence in a phasing call based on probability distributions, we required a statistical measure of the degree of correspondence between the two distributions. In this situation, the Bhattacharyya coefficient, *B*, is a calculation well suited to provide such a measure (32) (Equation 1).

Equation 3.1 $$B = \sum_{i=1}^{n} \sqrt{(c_i \cdot t_i)}$$

The Bhattacharyya coefficient is a coarse integration of the overlapping area between two histograms that share n bins where c_i and t_i are the respective frequencies in bin i. The value of B ranges from 0 to 1, with 1 indicating that the two histograms are identical and 0 indicating that there is no overlap between the two histograms.

MATRIX BASED CALCULATIONS

While the Markov Chain method of traversing the network is functional, it is computationally demanding and slow. Fortunately, the same result can be calculated directly by matrix multiplication. Beginning with the complete association matrix, the dataset is broken down into sub-matrices according to connections between two heterozygous loci (Equation 2). The values in matrix A can be converted to probabilities by dividing each cell by the sum of its row to create the probability matrix P. Each sub-matrix P_{jk} is the 4 x 4 matrix of probabilities for a *cis* association between every possible pair of bases at the heterozygous positions j and k.

Equation 3.2 $$A = \begin{bmatrix} A_{12} & A_{13} & \cdots & A_{1n} \\ & A_{23} & \cdots & \cdots \\ & & \cdots & \cdots \\ & & & A_{(n-1)n} \end{bmatrix}$$

The overall probability of a *cis* relationship between a base at position 1 and another at position n can be calculated by adding together the probabilities from every possible path between the two bases in the directed network. The

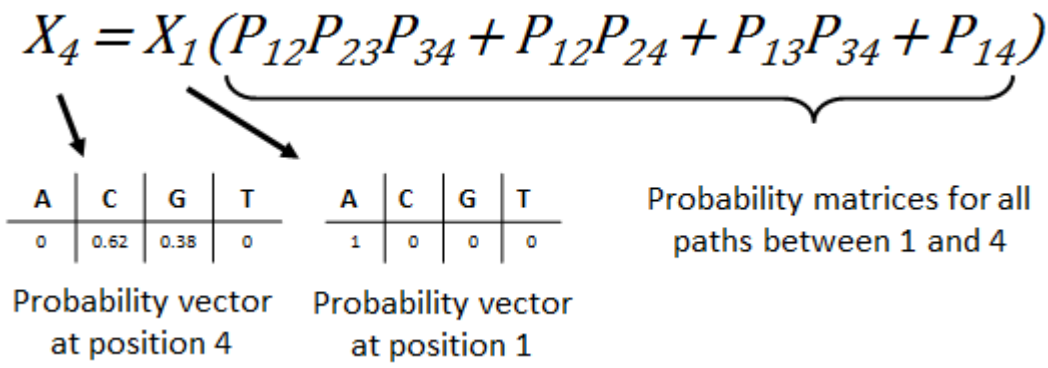

Figure 3-3: Multiplication of probability matrices generated from the Association Matrix produces the probability vector X_4 for all alleles in the terminal level of the directed network. The vector X_1 is a set of probabilities indicating the initial wild-type allele, in this case, A. Every possible path between levels 1 and 4 is represented as a term in the equation.

probability of each path is calculated by multiplication of the sub-matrices representing the path. Figure 3-3 demonstrates the specific case of a four-level network as described in Figure 3-2.

3.4: Results

Table 1 displays the phasing results from each of the specimens. The NGS phasing results agreed in all cases with the results of the conventional phasing methods described in the Methods section. In 12 cases, complete separation of the probability distributions was achieved (B = 0), while case number 45 resulted in a small amount of overlap (B = 0.64%) indicating some problem with the quality, or sequence-coverage, in this sample. Further investigation revealed that the total coverage linking these two positions was low (a total of 64 read pairs), and that this specimen contained no other heterozygous positions between the two sequence variants, which could have added supporting data (Figure 3-4). Finally, there were a significant number of recombinant or chimeric read pairs found in the data (16%). These factors resulted in broadened probability distributions that had overlapping tails, reducing the confidence of the final call (Figure 3-4). While deduction of the proper phase in this case is still obvious, with a chance of an erroneous assignment of phase of only 0.64%, the B-score acts as a warning signal, alerting the user to the need for further investigation of the data.

Specimen	Common Name	Allele 1	Genomic position (hg19)	Allele 2	Phenotype	Phase	*B*
44	P30L	C	32006291	T	NC	*trans*	0
	P453S	T	32008783	C	NC		
45	I-2 splice	A	32006858	G	SW	*trans*	0.0064
	S301F	T	32007948	C	VUS		
46	I172N	T	32007203	A	SV		
	R426C	T	32008702	C	SV	*trans*	0
	R479L	G	32008862	T	SV	*cis*	0
47	I-2 splice	C	32006858	G	SW	*trans*	0
	I172N	A	32007203	T	SV		
48	I-2 splice	A	32006858	G	SW	*trans*	0
	R426P	C	32008703	G	SW		
49	I-2 splice	A	32006858	G	SW	*cis*	0
	I172N	T	32007203	A	SV		
52	I-2 splice	A	32006858	G	SW	*cis*	0
	P453S	C	32008783	T	NC		
55	P30L	C	32006291	T	NC	*cis*	0
	V281L	G	32007887	T	NC		
56	I-2 splice	C	32006858	G	SW	*trans*	0
	I172N	A	32007203	T	SV		
57	H62L	A	32006387	T	NC		
	I236N	A	32007584	T	SW	*trans*	0
	V237E	A	32007587	T	SW	*trans*	0
	M239K	A	32007593	T	SW	*trans*	0
	P453S	C	32008783	T	NC	*cis*	0
59	I236N	T	32007584	A	SW		0
	V237E	T	32007587	A	SW	*cis*	0
	M239K	T	32007593	A	SW	*cis*	0
60	I-2 splice	A	32006858	G	SW	*trans*	0
	R124H	A	32006952	G	NC		
61	I-2 splice	C	32006858	G	SW	*trans*	0
	R356Q	A	32008313	G	SV		

Table 3.1: Haplotyping results from our cohort of 13 patients who were found to have compound heterozygous mutations. In each case, family members were tested to unequivocally determine the phase of each mutation. Mutant alleles are shown with a shaded background. In cases with more than one heterozygous allele, phase calls for each row denote the relationship of that allele to the first wild-type allele in that specimen. Phenotypes for congenital adrenal hyperplasia associated with mutations are non-classical (NC), salt-wasting (SW), simple virilizing (SV), and variant of unknown significance (VUS).

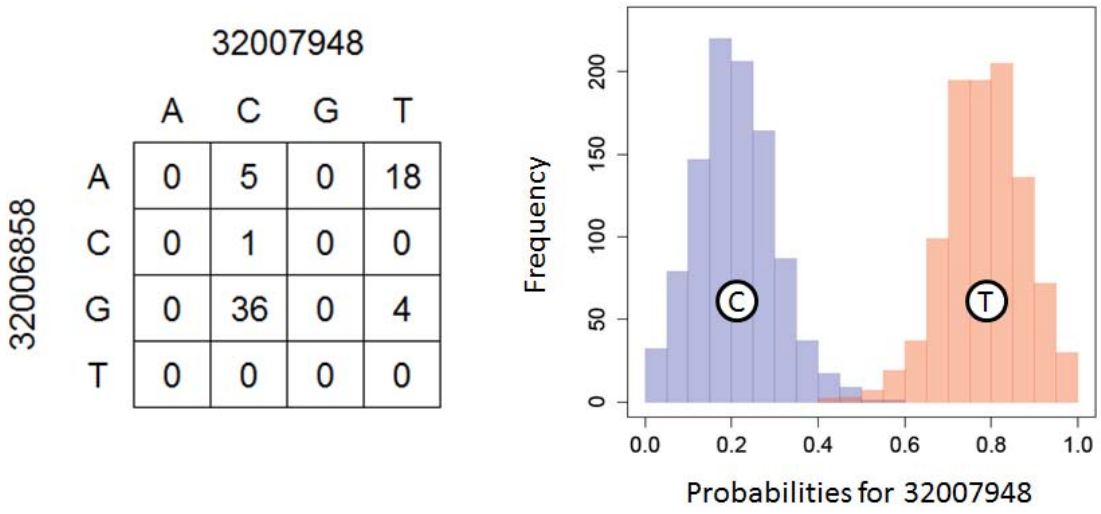

Figure 3-4: Calculation of the B-score for sample 45 produced a small but positive result (0.64%), indicating overlap between the probability distributions of two alleles (right). This was caused by low coverage and high error in the association matrix (left). The B-score acts as a gauge of uncertainty, warning the user of potential error leading to misinterpretation.

To better understand the practical limitations of the B-score we performed simulations using a variety of values for depth of coverage and error. In every test, both coverage and error were split as evenly as possible between both chromosomes. By keeping coverage constant and increasing error in the simulated data, we were able to estimate a transition curve where B becomes positive (Figure 3-5). This boundary represents practical limitations for coverage and accuracy that must be met for confident haplotyping.

3.5: Discussion

With the rapid adoption of NGS in the clinical testing arena our ability to collect raw data has far outpaced our capacity to analyze and interpret results. While there have been numerous computational tools developed for these purposes most of them have been designed for research purposes and lack the proper statistical underpinnings that are required at the level of clinical interpretation.

Phasing is a good example for an unmet clinical need. Accurate phasing of multiple heterozygous alleles is an important and common problem for many clinical sequencing labs. Our method offers a simple yet robust solution that is capable of providing a statically relevant score to help laboratory personnel properly analyze and interpret haplotypes in the presence of experimental error. In our testing scenario, it showed complete concordance with much more laborious conventional phasing methods. The method is in principle applicable to

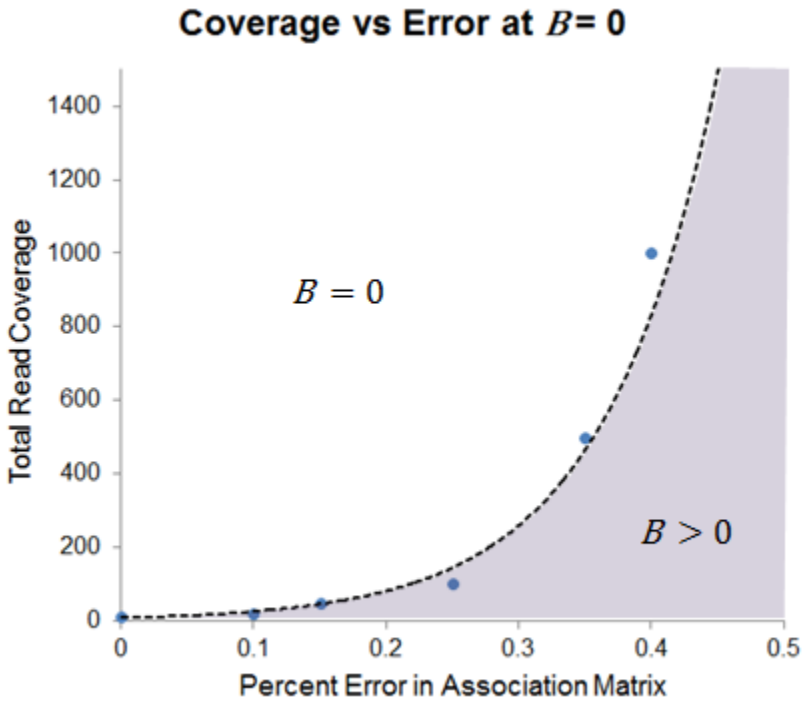

Figure 3-5: Simulations were run to approximate the read depth and error parameters required for the Bhattacharyya coefficient, *B* to remain at 0. A system with four heterozygous loci was modeled for these simulations with error distributed as equally as possible across both chromosomes. The shaded area shows conditions where probability distributions are expected to overlap, reducing confidence in the phase call.

any gene/gene locus and capable of phasing across significant genetic distances. When we evaluated our previous method for phasing (27), we found in simulations that 10-30 kb of phasing distance appeared feasible.

In terms of everyday reliability and applicability to clinical testing, the method offers a statistical basis for making reliable phase calls and provides an alert through the B-score for loci that have low coverage or an abnormal amount of error. Using this scoring system, users can be sure that the phase call made is correct, even with a dataset that contains erroneous read pairs. This is important because almost all types of parallel sequencing will include steps that introduce errors. Promiscuous adaptor ligation is the primary cause of these introduced errors in most methods. Where PCR is used to amplify regions of interest prior to NGS, however, many chimeras are also formed via *in vitro* recombination. Similarly, capture probe based approaches will also lead to a low percentage of incorrect captures of homologous genetic material, which may be ineffectively filtered out by many existing alignment algorithms. Finally, mapping and alignment methods can misplace highly homologous sequence reads, thus also potentially introducing incorrect base calls.

Because our method is based on resampling the association matrix, the position and width of each probability distribution is influenced most by the coverage at each locus and the gross amount of error represented in the matrix. Low sequence coverage could result in wider distributions, while erroneous chimeric fragments resulting from library preparations could influence the

distributions that converge on mean probabilities of 50%. Either case leads to overlap in the probability distributions and reduces phase call confidence accordingly.

While our method is successful in almost all cases, it has limitations in situations where large chromosomal rearrangements are present and with some small deletions. However, the method could be expanded and modified to deal with these conditions. Finally, like other variant analysis algorithms, this method depends on the quality of mapping and alignment of reads, highlighting the importance of these processing steps. If the aligning tools used before phasing analysis introduce too much error to the association matrix phasing results may be rendered invalid.

As clinical labs continue to adopt NGS as a standard application platform the need for high quality analysis methods will become increasingly important. Our method is one step towards the goal of having a complete suite of clinical grade algorithms that are suitable for medical use. The method eliminates the need for additional workup in all but the most complicated cases, offering reductions in labor, cost, and complexity for the lab. It provides phasing results directly from the proband, which greatly simplifies analysis, in particular for closely related families where low genetic diversity may cause familial studies to fail. The method should be applicable to all recessive diseases, but will be most useful for those that show delayed clinical penetrance or mild phenotypes.

Chapter 4: Analysis of structural variation in clinical specimens using mate pair sequencing

Manuscript in preparation: Structural characterization reveals differences between follicular thyroid carcinomas and follicular adenomas.

4.1: Introduction

Genomic rearrangements are defined as structural changes to chromosomes including translocations, inversions, and large (>1000 base pairs) deletions or duplications. Only in the last decade have these structural variations become widely associated with diseases such as cancer. Although there were early discoveries of the link between rearrangements and disease, they were thought to be rare exceptions. More recent studies however, have identified important rearrangements that cause expressed fusion proteins such as *TMPRSS2* and *ERG* in prostate cancer (33), and *EML4* with *ALK* in lung, breast, and colon cancers (34, 35). Findings like these have inspired research in the arena of genomic structure and opened our eyes to the importance of this class of variation. While the mechanism of disease onset or progression is not well understood for each rearrangement identified, a growing body of evidence shows the significance of structural variation as a cause of disease. In fact, the

Catalogue of Somatic Mutations in Cancer (COSMIC) now lists over 17,000 identified fusion genes from cancer specimens in its v76 release (36).

The reasons for slow uptake on this class of mutation are mostly technical. Rearrangements were traditionally identified by visual observation in tumor cell culture metaphases and followed up with fluorescent in situ hybridization (FISH) assays and/or PCR. Now, Next Generation Sequencing (NGS) technology has provided the means to observe structural variation with greater resolution and at a more rapid rate than has previously been possible.

While the diagnostic and prognostic value of structural variation is clear, there is not a well-developed method to quickly identify and evaluate effected genes in biopsy tissue or other clinical specimens. Several techniques are currently in use, however they suffer from relatively low resolution (cytogenetic and array-based methods, for example), or genome coverage limitations (such as PCR and Sanger sequencing methods). Next Generation Sequencing provides the ability to overcome both of these problems with its coverage of the entire genome in a single experiment, and its capability to observe the genome at single base resolution.

Detection of rearrangements in genomic DNA using NGS has been previously reported (37, 38), however, there is much room for improvement of the mapping process, both in accuracy and speed. In order to make the technique suitable for use in a diagnostic workflow we have made improvements to increase the speed and accuracy of analyzing NGS data for structural variation

and propose a method to quickly evaluate the results for interpretation. These improvements can be used to rapidly analyze clinical specimens as we demonstrate with a dataset collected from thyroid tumor tissue.

4.2: Methods

One of the most informative and cost effective ways to study genomic structure is by mate pair sequencing. The technology is well suited to detect most types of rearrangements across the entire genome, it is relatively inexpensive, and it can provide breakpoint information at, or near, sequence-level resolution. This method detects translocations, inversions, deletions, and other rearrangements by locating regions of the genome that contain junction points. While the concept is simple, the chemistry is relatively complex and sequencing products pose unique challenges to downstream computation and analysis. In order to appreciate the difficulties encountered and the solutions developed, one must first have a practical understanding of the process.

Chromosomal rearrangements can be detected by shearing DNA into short fragments and identifying pieces that contain genomic junction points. This is done by sequencing both ends of each fragment and, after mapping them to a reference, searching for fragments whose sequence reads align to discordant locations in the genome. This strategy can be used with virtually any paired end sequencing technique, however not all library construction methods are optimal

for locating breakpoints. Clearly, the likelihood that any given piece of DNA will span a genomic breakpoint improves as the length of that fragment increases. Therefore, a method's ability to detect breakpoints increases when genomic DNA is sheared into longer fragments. Using fewer and longer DNA fragments for junction point detection gives rise to the need for an appropriate evaluation of genomic coverage for each dataset. One very suitable measurement of coverage for mate pair sequencing is bridged coverage. Defined as the number of *fragments* that cover a given locus, this measurement includes any reads that are mapped directly to the site as well as any inferred DNA that is between two reads. Using this gauge of coverage yields a much more realistic estimate of the number of long fragments that span a given region of the genome.

Unfortunately, the Illumina platform was developed to sequence DNA fragments up to 500 base pairs in length; a limitation of the solid-phase bridge amplification used to generate clusters of DNA from a single fragment on the flow cell surface. If longer fragments are used, during amplification the clusters expand and merge with one another, effectively making them all unusable for sequencing. Much like bacterial colonies growing to confluence on a plate, resolution between amplified clusters is lost and the signal from any individual is impossible to pick out of the mixture.

In response to the need for longer sequencing fragments, the mate pair (MP) library preparation method was developed. This technique produces sequence reads that are 5 to 10 times farther apart than traditional paired end reads. This is possible because regions of DNA that are located several

thousand base pairs apart are physically conjoined through ring-forming self-ligation. Ligated regions can then be isolated and sequenced after the rings are broken into fragment lengths appropriate for bridge amplification and sequencing.

While the long span paired reads produced by MP sequencing are valuable for analysis, the chemistry involved in creating them produces artifacts that cause unique challenges for downstream read mapping algorithms. After the self-ligation step, DNA rings are randomly sheared to make fragments sufficiently short for paired end sequencing. Therefore, genomic junction points (if any exist in the fragment) or ligation junction points may be positioned near either end of one of the sequenced fragments. Sequences that are made of these mixed reads will be difficult, if not impossible, for most mapping algorithms to process. Because these reads are made of two conjoined segments of DNA, they are by definition hybrids and will not accurately match any portion of the reference sequence. Most mapping algorithms, after failing to find a matching location in the reference, simply eliminate them from further analysis. This is a severe loss of data because these hybrids make up roughly 20 percent of a typical MP dataset. In addition, some of these hybrid reads will actually contain a genomic junction point, providing direct evidence of structural variation. If these reads can be accurately mapped to their respective locations, there is great potential to make more reads usable and directly observe genomic breakpoints. To address the specific challenges of mapping mate paired reads, the Biomarker Discovery Group at Mayo Clinic developed the Binary Indexing Mapping and Alignment algorithm (BIMA) (20, 22, 39). Our third release of the algorithm

(BIMA v3) was highly optimized for speed and accuracy while dealing with unique and problematic artifacts related to mate pair library construction (29).

The BIMA v3 algorithm converts DNA sequence into three separately encoded binary sequences. By reducing genomic sequence to a binary format, the speed and efficiency at which alignment calculations can be made is dramatically increased. The algorithm also aligns reads in small, 32-base increments, increasing the method's ability to detect junction points. These two characteristics of BIMA v3 are important factors that make the algorithm exceptionally fast and accurate for mapping mate pair data.

After completion of mapping, the reads can be analyzed for structural variation. A suite of algorithms collectively called Structural Variant Analysis (SVA) has been developed to quickly locate discordant reads in the dataset and evaluate them as evidence for junction points. These tools were developed to use mate pair sequence data to its fullest potential in terms of structural variation detection.

BioSystems Database:

Although BIMA v3 and SVA are powerful and effective tools, they are limited to the search for rearrangements and their respective junction points. Understanding the practical effects of disrupted genes requires a good deal more user input and analysis. Information regarding the functions and pathways associated with individual genes is still only sporadically available and, in most cases, incomplete. Fortunately, there are several reliable databases that offer

pieces of information available for almost every human gene. Resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (40), the Gene Ontology (GO) (41), WikiPathways (42), and others are continually being populated with the latest data available.

In an attempt to unify as much available information as possible, the BioSystems database was recently created. Its stated purpose is to "centralize access to existing biosystems databases", "connect biosystems records with associated literature, molecular, and chemical data", and "facilitate computation on biosystems data" (43). This tool is meant to be used as a clearinghouse for the most useful information concerning biological pathways and genetics. It provides the most current and complete collection of information for researchers and allows data from multiple databases to be combined and compared. In addition, the database is not limited to a single species. Whenever possible, data from homologous genes in other organisms is presented. This allows a user to take a cross-section of knowledge from biological spectrum and apply it intelligently to their research.

The second half of this characterization method utilizes the BioSystems database as a search engine for connectivity between effected genes and the pathways that lead to malignant behavior. This important segment in the process links clinical observations with currently available knowledge about how genes and their products function in living cells. In addition, it shows how genes and their pathways are interconnected with other cellular functions, thus allowing the user to extrapolate findings into possible downstream effects.

The strategy for analysis of genes affected by structural variation is straightforward. Beginning with the entire genome of ~25,000 genes, the goal is to successively filter out genes that are irrelevant to the phenotype or disease under investigation until a few remaining candidates are left (Figure 4-1). If initial assumptions are accurate and the filtering process is performed correctly, any genes left at the end of the procedure should have a high probability of disease involvement. These genes should then be considered for additional investigation, including *in vivo* or biochemical studies.

The first filter in the process removes all genes not directly affected by breakpoints. Next, each affected gene is explored in the BioSystems database to collect whatever information is known about the gene. When a user searches the BioSystems database for a gene of interest, by default the results are sorted and presented within the context of three basic categories. First, known metabolic and/or signaling pathway associations are shown for the gene or its products. Next, the physical location of the gene product is reported in terms of the structural or cellular complexes in which it has been found. Finally, functional attributes of the product are grouped and reported. These subsets of information are sorted and then searched for user-defined keywords that are relevant to the disease being studied. Those genes that have annotations containing a keyword are said to be “associated” to the disease and are retained, while all others are removed from consideration. BioSystems data for associated genes is further analyzed to determine the specific pathways, structures, and functions in which

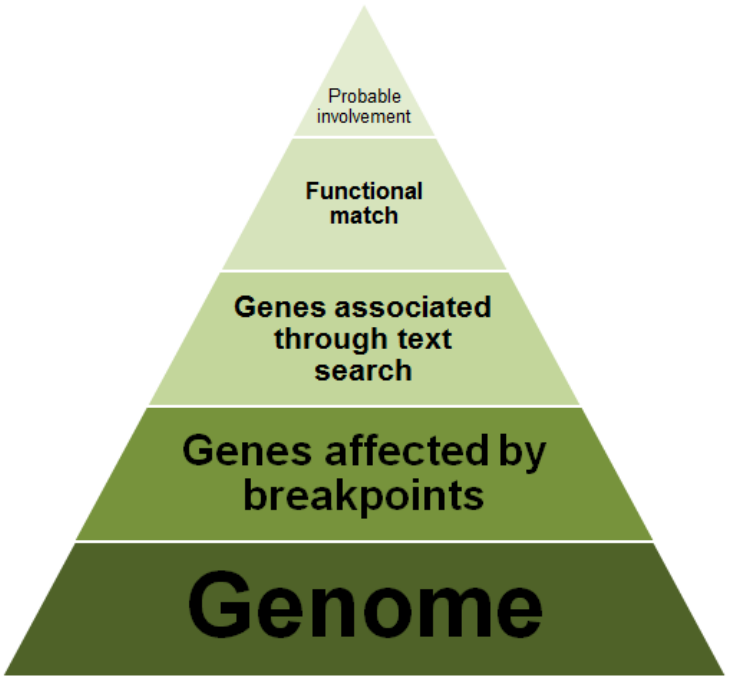

Figure 4-1: Analysis of structural variation in diseased tissue begins with examination of the entire genome via mate pair sequencing. Filter #1 removes genes that are not directly affected by genomic breakpoints. Remaining genes are entered as a search query in the BioSystems database. Filter #2 retains any genes that are found to have user defined keywords in their annotations. Genes that have database annotations containing keywords are said to be "associated" with the disease. Associated genes are then processed by filter #3; a manual determination of the functional relevance of the association. Filter #4 is a manual process of analysis of the junction point in question. Potential expression products are considered in the context of the disease. Any genes remaining after this process have a high likelihood of disease involvement and warrant additional studies.

the gene or its products are involved. The final filter selects genes with functional or physical attributes that are, or could be, relevant to the disease. These genes will then be inspected for structural mutations or fusion products that were introduced by the genomic rearrangement identified by BIMA v3 and SVA.

Structural variation in follicular thyroid tumors

Using the method described above, we performed analysis of the genomic structure of a common, yet problematic subset of thyroid tumors. Our primary goal was to observe and compare the genomic rearrangements present in two different types of follicular thyroid tumors and possibly identify a molecular marker that could be used to distinguish them in biopsy tissue.

It has been estimated that up to half of the US population could have some type of thyroid lesion (44). Fortunately, most are benign and require no medical intervention. Of the nodules which are malignant, 80% - 90% are diagnosed as papillary thyroid carcinoma (PTC). These can be readily diagnosed via fine needle aspiration cytology in most cases. The second most common type of thyroid malignancy is follicular thyroid carcinoma (FTC). These nodes are much more difficult to diagnose using cytology techniques because of their similarity to benign follicular adenoma (FA). These two follicular neoplasms share very similar architecture and morphology, which results in a high percentage of indeterminate cytology results. When a follicular type nodule is identified, the patient will typically undergo partial or total thyroidectomy to obtain a definitive diagnosis by histology. Reports estimate that roughly 80% of the

tumors removed in these cases were, in fact, benign adenomas (45-47). Therefore, the majority of patients with follicular neoplasms assume the needless risk and expense of an unnecessary surgery, and if the entire thyroid is removed, the patient is left without a vital organ for which they will need treatment and management for the rest of their lives. Clearly, there is a need for better methods to distinguish malignant follicular neoplasms from their non-malignant counterparts.

There has, in fact, been much work done toward finding molecular markers for FTC in the last decade. A large number of studies have been reported in which point mutations, expression patterns, microRNA fingerprints, and oncogene panels have been explored as a possible means to differentiate FTC from FA (48, 49). Many of these reports have provided useful information, but none has yet identified a definitive molecular signature of the disease. As yet, there has not been a comparative study reported on the occurrence of chromosomal rearrangements in both types of follicular nodules. Several genomic rearrangements have been observed in thyroid cancers; most notably, the t(2,3) PAX8/PPARγ (PPFP) translocation that is most often associated with FTC. This mutation has become an important diagnostic marker when used in conjunction with several other genetic variants (50). Given the importance of this variant, it is surprising that more attention has not been given to the search for additional genomic rearrangements.

Working with physicians in the Department of Endocrinology at Mayo Clinic, we obtained sixteen fresh frozen thyroid tumors that were surgically

resected for diagnosis. Each nodule was characterized histologically as FTC (n=10) or FA (n=6). Genomic DNA was extracted from each specimen and used to create mate pair libraries for sequencing. Mate pair sequencing data was collected on an Illumina HiSeq instrument and reads were mapped using BIMA3. After mapping and alignment were complete, SVA analysis was performed to identify rearrangements in each specimen. Structural variations were identified for each specimen, and those that were reported with high or medium confidence were used for analysis using the BioSystems database.

Current practice for diagnosis of follicular thyroid tumors relies on observation of capsular or vascular invasion as the sole histological characteristic differentiating FTC from FA. Therefore, this study was focused on the mechanisms and pathways involved in the interactions between follicular cells and vascular or connective tissue. Based on our postulation that genes that play a role in these pathways may be compromised in carcinomas and not in adenomas, we searched results from the Biosystems database for:

1. Any association with cellular adhesion or communication – especially with vascular or connective tissue
2. Any known association with thyroid cancer
3. Any known association with other cancers

A list of keywords that were used to filter BioSystems results can be found in Table 4-1. Findings for each specimen can be found in Appendix A.

Table 4-1: Keywords used for association of genes

Keywords associated with cellular adhesion or communication
Junction, adhesion, communication, signaling, cell-to-cell, tight junction, vascular, capsule, connective, cell surface, intercellular
Keywords associated with thyroid cancer
thyroid, thyroglobulin, thyroxin, iodine, PAX8, PPARg, RET, PTC, NRAS, KRAS, HRAS, BRAF
Keywords associated with other cancers
Cell cycle regulation, cell cycle, apoptosis, ABL1, AKT1, ALK, APC, ATM, CDH1, CDKN2A, CSF1R, CTNNB1, EGFR, ERBB2, ERBB4, EZH2, FBXW7, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MET, MLH1, MPL, NOTCH1, NPM1, PDGFRA, PIK3CA, PTEN, PTPN11, RB1, SMAD4, SMARCB1, SMO, SRC, STK11, TP53, VHL

4.3: Results

In our study of structural variation in 16 follicular thyroid tumors we found a total of 79 events where genes were affected by a genomic junction point. Every affected gene was considered for a possible role in malignant behavior by using the BioSystems database to determine if the gene had any known associations with thyroid cancer, associations with pathways involved in other cancer types, and/or associations with intercellular adhesion or communication functions.

Of the 79 events, 24 were associated to disease via 45 individual keyword matches. Next, 6 of the 24 were removed by filter #3. These were obvious context and/or functional mis-associations. Of the remaining 18 events, 4 were removed because of an unlikely functional match to disease.

Of the 14 events that were left after the filtering process, there were several gene replicates. In addition, two genes were eliminated from the process due to a lack of keyword association that should have been included. A final list including 12 genes were included at the Functional Match stage and those were retained for additional analysis and determination if there was a possibility of disease involvement. Annotations in the BioSystems database as well as other sources were searched manually to learn relevant details about the purpose and function of each candidate gene and how it might relate to the disease state of thyroid tissue. The following synopsis is a summary of the known function of each candidate that was considered to be relevant in our dataset.

Table 4-2: Genes affected by junction points and associated to follicular thyroid carcinoma (FTC) and follicular adenoma (FA) by the BioSystems database.

Genes associated by cellular adhesion or communication keywords		
Specimen	**Gene**	**Description**
MP-1 (FTC)	*NRXN3*	Neurexin 3
MP-2 (FTC)	*CDH4* *CNTN1* *PTPRT*	Cadherin 4 Contactin 1 Protein Tyrosine Phosphatase, Receptor Type, T
MP-5 (FTC)	*GOLPH3* *PDZD2*	Golgi Phosphoprotein 3 (Coat-Protein) PDZ Domain Containing 2
MP-9 (FA)	*CTNNA3*	Catenin (Cadherin-Associated Protein), Alpha 3
MP-10 (FTC)	*SORBS1*	Sorbin And SH3 Domain Containing 1
Genes associated by thyroid cancer keywords		
Specimen	**Gene**	**Description**
MP-2 (FTC)	*TG*	Thyroglobulin
MP-6 (FA)	*PAX8* *PPARg*	Paired Box 8 Peroxisome Proliferator-Activated Receptor Gamma
MP-13 (FTC)	*PAX8* *PPARg*	Paired Box 8 Peroxisome Proliferator-Activated Receptor Gamma
MP-14 (FA)	*PAX8* *PPARg*	Paired Box 8 Peroxisome Proliferator-Activated Receptor Gamma
Genes associated by other cancers keywords		
Specimen	**Gene**	**Description**
MP-1 (FTC)	*FAM134B* *LIN52*	Family With Sequence Similarity 134, Member B Lin-52 DREAM MuvB Core Complex Component
MP-2 (FTC)	*NFATC2* *DOK5* *HNF4A*	Nuclear Factor Of Activated T-Cells Docking Protein 5 Hepatocyte Nuclear Factor 4, Alpha
MP-4 (FTC)	*BRD4* *RBL1* *COLCA2*	Bromodomain Containing 4 Retinoblastoma-Like 1 Colorectal Cancer Associated 2
MP-16 (FA)	*RBL1*	Retinoblastoma-Like 1

Genes with associations to cellular adhesion or communication

Seven different genes were identified in this category, of which, all but one were exclusive to carcinoma samples. This is striking since the most reliable feature distinguishing FTC from FA is direct observation of neoplastic cells invading vascular space or breaking through their containment capsule. Some of the following genes, if disrupted, may be responsible for that behavior.

- **Cadherin 4 (*CDH4*):** Cadherin 4 is well known to be involved in cell-to-cell adhesion, especially in neuronal tissues (51). It is found anchored in the cell membrane and contains a highly conserved cytoplasmic tail that binds to a complex of proteins including actin filaments making up the cytoskeleton of the cell.

- **Catenin (Cadherin-Associated Protein), Alpha 3 (*CTNNA3*):** The protein is involved in the binding of cytoskeletal actin filaments with cadherin complexes (52). This is a critical component of cell-cell adhesion, however, the literature is somewhat lacking in this area.

- **Contactin 1 (*CNTN1*):** Contactin 1 is a member of the immunoglobulin superfamily and is anchored to the extracellular side of the plasma membrane. Mostly studied in relation to its known function of neuronal development, it has only recently been associated with a variety of cancers (53-55). One very recent study suggests that contactin overexpression may promote invasion in thyroid tumors (56).

- **Protein Tyrosine Phosphatase, Receptor Type, T (*PTPRT*):** This protein, also known as PTPrho, is part of a signaling complex that spans the plasma membrane. It is known to interact with the catenin and cadherin families of adhesion proteins. It has been reported as a critical component of homophilic cell-cell aggregation where mutations to the extracellular domain result in a loss of cellular adhesion (57).

Genes with associations to thyroid cancer

- **PAX8/PPARg fusion protein (PPFP):** This translocation is a well-established marker of follicular nodules, both malignant and benign (58, 59). We were not surprised to see these mutations in three specimens (MP-6, MP-13, and MP-14) because each had previously been tested for the translocation using a traditional PCR-based assay.

- **Thyroglobulin (*TG*):** Thyroglobulin is not directly related to thyroid cancer as an oncogene or tumor suppressor, however, it is expressed specifically in thyroid tissue and a fusion product under the control of its promotor could be germane to a diseased phenotype.

Genes with associations to other cancers

- **Bromodomain Containing 4 (*BRD4*):** The *BRD4* gene is well known to be associated with various cancer types. Most commonly, transfusion events have been found between *BRD4* and several nuclear protein,

testes (NUT) family members in aggressive midline NUT carcinomas (MNC) (60). Bromodomains in the *BRD4* protein bind to acylated histones and recruit transcription factors to the chromosome. The importance of bromodomain containing proteins in cancer is demonstrated by the recent development of several bromodomain-blockers that are currently being tested as therapeutic agents (61).

- **Colorectal Cancer Associated 2 (*COLCA2*):** This expression product has been strongly associated with colon cancer; however, the mechanisms of its involvement are not clear. It was first observed in comparative studies of the 11q23 locus in which the single nucleotide polymorphism rs3802842 was identified in GWAS studies of colorectal cancer (62) and was shown to be expressed in the cytoplasm of many cell types.

- **Family With Sequence Similarity 134, Member B (*FAM134B*):** Little is known about the *FAM134B* gene product, also known as *JK1*, in relation to cancer. However, recent studies show convincing evidence that the protein may be important by repressing migration in colorectal cancer cells (63, 64).

- **Nuclear Factor Of Activated T-Cells (*NFATC2*):** This gene has been most studied in cases of breast cancer; specifically, in conjunction with

highly metastatic tumors (65). It is generally understood that *NFATC2* increases cell motility when overexpressed and is considered to be a pro-invasive gene. In addition to breast tumors, fusion products between NFATC2 and Ewing sarcoma breakpoint region 1 (*EWSR1*) have been associated with spread in bone tumors due to loss of regulatory elements that lead to overexpression (66).

- **Retinoblastoma-Like 1 (*RBL1*):** *RBL1* is similar in sequence to the reteinoblastoma 1 gene (*RB1*) which is a known tumor suppressor. Although no anti-tumor activity for *RBL1* has ever been identified in humans, its murine homolog, p107, has been shown to regulate cell cycle transitions, possibly by repressing the transcription factor *MYC* (67).

Because each of the genes listed above have known functions that could plausibly implicate them for involvement in disease, they were all retained for the last step in the filtering process. For each gene, processed read data was examined to determine the probable outcome of the structural variation disrupting the gene. Based on the rearrangement and the presumed expression product (if any), an assessment was made as to whether the gene was likely to be involved in disease. Figures of the junction plots from each rearrangement can be found in Appendix B. Summary information is shown in Table 3-3.

Table 3-3: Summary of gene rearrangements

Genes with associations to cellular adhesion or communication		
Gene	Rearrangement	Probable expression product(s)
Cadherin 4 (*CDH4*)	Complex rearrangement with 3 inversions and gain	At least 3 copies of the *CDH4* gene are present in this sample and each of those is affected by inversions. **None of the detected rearrangements is expected to produce an expression product.** It is not clear whether these events are part of a bi-allelic system, or if there are multiple clones in the tissue. There does appear to be at least one normal copy of the gene present.
Catenin, Alpha 3 (*CTNNA3*)	Deletion	This heterozygous deletion **removes exons 8 and 9 from CTNNA3 on one chromosome**.
Contactin 1 (*CNTN1*)	Deletion	A deletion results in a **fusion product joining exon 1 of contactin with exons 3 – 9 of *C12orf54*.** Nothing is currently known about the presumed expression product of *C12orf54*.
Protein Tyrosine Phosphatase, Receptor Type, T (*PTPRT*)	Complex rearrangement with 2 inversions and gain	A pair of inversions renders **two copies of *PTPRT* non-functional; however, at least one viable copy remains intact.** It is not clear whether these events are part of a bi-allelic system, or if there are multiple clones in the tissue.
Genes with associations to thyroid cancer		
Gene	Rearrangement	Probable expression product(s)
PPFP (*PAX8/PPARg*)	Translocation	**This well-known translocation results in a fusion protein that has been associated with follicular thyroid tumors.** Previous genotyping of these specimens had found the rearrangement in all 3 specimens.
Thyroglobulin (*TG*)	Translocation and deletion	A translocation event between chromosomes 8 and 9 results in **one truncated thyroglobulin protein**. **A deletion removes exons 35 – 41 of a second copy of *TG*.** It is not likely that this abbreviated product would be expressed as a functional protein. It is unclear whether there is an unaffected copy of *TG* in this tissue.

Table 3-3: Summary of gene rearrangements (Continued)

Genes with associations to other cancers		
Gene	Rearrangement	Probable expression product(s)
Bromodomain Containing 4 / Colon Cancer Associated 2 (*BRD4/COLCA2*)	Balanced translocation	This balanced translocation results in two fusion products. First, **BRD4 exons 1 – 15 are fused with COLCA2 exons 2 – 7.** The second fusion product is **COLCA2 exon 1 fused with BRD4 exons 16 – 20.**
Family With Sequence Similarity 134, Member B (*FAM134B*)	Tandem duplication	This gene has a **single, large intron repeated in tandem within the gene.** No exons appear to be affected; however, there is one alternate splice site in the repeated region.
Nuclear Factor Of Activated T-Cells (*NFATC2*)	Inversion	This inversion results in a **fusion product between DOK5 exon 1 and NFATC2 exons 2 – 11.** This event appears to have occurred only on one chromosome.
Retinoblastoma-Like 1 (*RBL1*)	No event	This was a **false positive caused by a transposon** in intron 1 of RBL1

4.4: Discussion

One of the most critical problems with the clinical use of NGS is interpretation of data. While it is relatively easy to collect and process data, the problem of sorting out disease-causing mutations from irrelevant passenger mutations has remained stubbornly difficult. Here we demonstrate a method to quickly locate and classify one type of mutation, namely, structural rearrangements. We have developed the tools to rapidly map NGS data from mate pair libraries and detect structural variants using the BIMA v3 and SVA algorithms. Additionally, we propose here a method for filtering out affected genes based on known functional attributes of each gene and its products. This method has shown its utility in the analysis of a set of follicular thyroid tumors that differ only in the histological characteristics of vascular or capsular invasion. The method was able to show that genes involved in cell-cell adhesion were affected more often in FTC tissues where invasion had been observed.

Of the 79 structural events observed in the data, 12 genes were ultimately filtered out as potentially relevant to the disease state by annotations in the BioSystems database. This filtering step reduced the number of candidate genes to 15% of the initial list, providing significant time-savings for the task of interpretation. Much of this work was automated, significantly reducing the laborious work of searching literature and other information sources.

As mentioned previously, there were two false negative events that were erroneously filtered out using this method. Neither the Bromodomain Containing

4 (*BRD4*) nor the Colon Cancer Associated 2 (*COLCA2*) genes were associated to disease by a keyword. This is surprising since both genes have been well connected to cancer in the literature. This occurrence serves to illustrate the importance of choosing keywords that find a balance between too general and too specific to be useful.

This method produces a list of candidate genes with potential for involvement in disease. These results, however, should be used with caution and with the realization that in most cases additional workup will need to be performed for a full understanding of the affected genes. In our study, multiple regions of chromosomal gain made results unclear as to the number of affected and unaffected copies of some genes. Follow up molecular assays such as qPCR or digital PCR will be extremely useful for sorting out these quantification issues.

As with any analysis method, this one has its limitations. First, the results can only be as good as the data provided by the BioSystems database. Inaccuracies and omissions will be continually problematic, however as the database matures, improvements will follow. Additionally, information is currently provided without a standardized format or nomenclature. This problem ranges from annoying to mildly difficult, and is generally overcome by a quick literature search when questions arise. Again, as the system matures, there will be opportunity for more consistency in data and nomenclature which will only improve the method.

While analysis of data with BIMA v3 and SVA is completely automated, the steps following are currently more manual. Future versions should incorporate more automation, including logic to help understand the context in which key words are used in database annotations.

Chapter 5: Mitochondrial DNA copy number as a biomarker

5.1: Background

Mitochondrial DNA (mtDNA), while tiny in comparison to the nuclear genome, is critical to the function of the organelle, and by extension, the entire cell. This circular plasmid resides in the mitochondrial matrix and, in humans, codes for 13 proteins, 2 ribosomal RNAs (rRNA), and 22 transfer RNAs (tRNA). The expressed proteins are components of the oxidative phosphorylation (OXPHOS) pathway and can be found in the membranes of mitochondrial cristae. Ribosomes and tRNAs are transcribed and utilized exclusively inside the mitochondrial matrix. Because of high turnover rates for OXPHOS proteins, mitochondria contain multiple copies of mtDNA. And since each cell contains multiple mitochondria, there can be thousands of copies of this circular chromosome in a single cell depending on its type and metabolic rate.

The impact on human health from variations in mtDNA is only beginning to be understood. In fact, the first diseases definitively attributed to mitochondrial mutations were muscle mitochondrial myopathy and Lebers optic neuropathy, both in 1988 (68, 69). Since then, a handful of disorders have been assigned to specific mutations, and our collective understanding of the complex relationships

between mitochondrial and nuclear genetics has continued to grow (70). Because all of the genes in mtDNA code for OXPHOS-specific proteins or RNA, mutations disrupt the cell's ability to produce energy. Although the actual mechanisms of disease are often not clear, most mitochondrial disorders are caused by tissues with high energy requirements (like nerve and muscle) being starved for ATP or other electron transport chain components.

From a diagnostic perspective, most of the focus on mtDNA has been on primary sequence. Nucleotide substitutions and insertions/deletions are relatively easy to identify and use in diagnostic workflows. And because these types of variation have concrete implications on expression products, it is generally straightforward to understand and predict outcomes. There is another aspect of mtDNA, however, that may serve as a clinically useful marker of disease. It has been suggested that the copy number of mitochondria could reflect the disease status in some tissues like malignant tumors. Each cell contains multiple copies of mtDNA, and over time that number can change. Periods of oxidative stress and high levels of metabolic activity can stimulate replication of mitochondria as well as the mtDNA contained in each (71). It is still unclear how well the amount of mtDNA in a tissue serves as a diagnostic or prognostic marker; however it is clear that some diseases, like certain types of cancer, have localized regions of high metabolic activity and the mtDNA copy number could serve as an indicator of disease status.

5.2: Methods

The method used to calculate mtDNA copy number is important. The most common way to quantify DNA in Next Generation Sequencing (NGS) data is to use some measure of the number of sequenced fragments at a given position in the genome. There are a number of ways to measure this, including read coverage and bridge coverage. Read coverage is calculated by simply counting the number of reads that cover a single base in the sequence. Bridged coverage, on the other hand, is a measure of how many fragments cover each position, including reads and the inferred space between them (Figure 5-1). Using this calculation offers the advantage of effectively including each entire DNA fragment rather than limiting coverage to the sequence reads. Since mate pair fragments generally have an average length of several thousand base pairs, using this method allows a greater portion of the genome to be represented by each library of sequence reads.

Because of the inherent variability of mtDNA concentrations, normalization techniques are critical. A simple, yet effective method for normalization is to calculate the ratio of coverage between nuclear and mitochondrial DNA. This assures that comparison of mtDNA between different specimens or tissues is standardized back to a consistent, bi-allelic reference point. Since read depth in NGS data is known to vary widely across the genome, it is important to consider coverage in the context of the entire genome and use an appropriate average; thereby eliminating regional effects from the calculation. The Structural Variant Analysis algorithms (SVA) developed by the Biomarker Discovery Group at Mayo

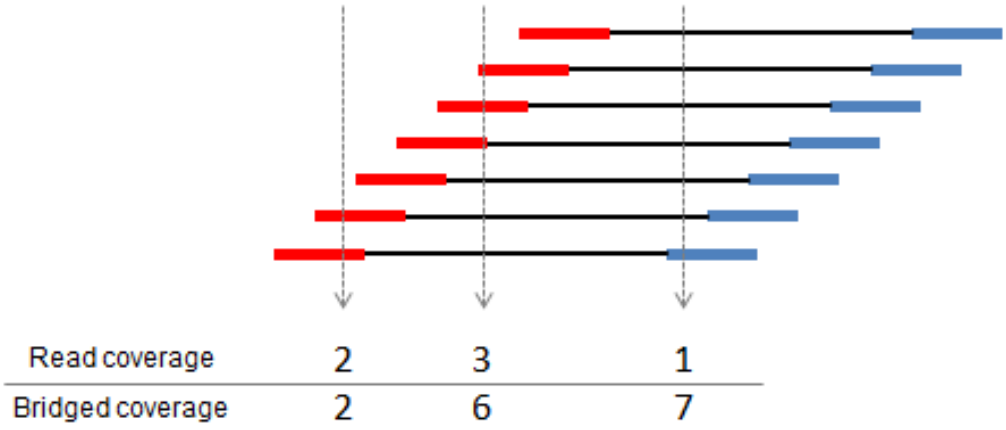

Figure 5-1: Coverage as calculated by two different methods is shown at three different points in the sequence. Bridged coverage includes the span of DNA between reads that is not sequenced.

Clinic calculate the average bi-allelic coverage for NGS data based on distributions of bridged read coverage using data from the entire nuclear genome. Briefly, coverage is calculated in small windows across the entire genome, then, regions with gain or loss are removed so that an accurate average can be calculated for bi-allelic regions. This method provides a more truthful value for bridged coverage by eliminating the influence of deletions or amplifications.

To explore the relationship between mtDNA copy number and disease, we used NGS data collected from normal and malignant samples of lung and prostate tissue. In addition, we analyzed data collected from follicular thyroid carcinoma (FTC) and follicular adenoma (FA) to compare mtDNA copy numbers in each. All specimens were processed for mate pair library construction and sequenced on various Illumina parallel sequencing platforms. Reads were mapped to the reference genome (GRCh37 or GRCh38) using the Binary Indexing Mapping and Alignment algorithm (BIMA) (22, 29, 39).

5.3: Results

Our initial analysis of mtDNA copy number was done using a collection of follicular thyroid tumors. Follicular thyroid cancer (FTC) is nearly impossible to distinguish from follicular adenoma (FA) without histological analysis. In our search for molecular markers distinguishing the two tissue types we performed whole genome sequencing using mate pair libraries on DNA extracted from 16

different specimens (10 FTC, 6 FA). From this data, we quantified mtDNA to determine if copy number could be used as a differentiating factor. Initial examination suggested that mtDNA copy number may be elevated in certain cases of FTC; however, statistical analysis indicated that two high values were likely outliers (Figure 5-2). One of the two samples was available for additional assessment and was found to be a Hurthle cell carcinoma, a rare thyroid tumor very similar in structure to follicular tumors and easily mistaken. Hurthle cells are also known to contain high levels of mitochondria, although no quantitative report could be found. Given this finding, it is likely that the other elevated result also belongs to a Hurthle cell carcinoma that was previously mis-diagnosed. The t-statistic calculated for these two samples indicates that there is not an appreciable difference in mtDNA copy number in these two tumor types.

Also available for mtDNA analysis was a dataset from a study of prostate cancer. Because prostate tumors are staged on the Gleason system we simplified the nominal ranking of tumors to low grade (G3 through G5) and high grade (G6 – G9). Normal tissue was also tested; however, it had been processed using an alternate method of mate pair construction making results unreliable for direct comparison. Results from those specimens were removed from our analysis. A correlation is clear between increasing mtDNA copy number and progression of disease (Figure 5-3). Because of the gross categorization of our data it remains unclear whether this trend might be consistent through each stage of disease. Additional specimens from every category would be needed to fully explore this question.

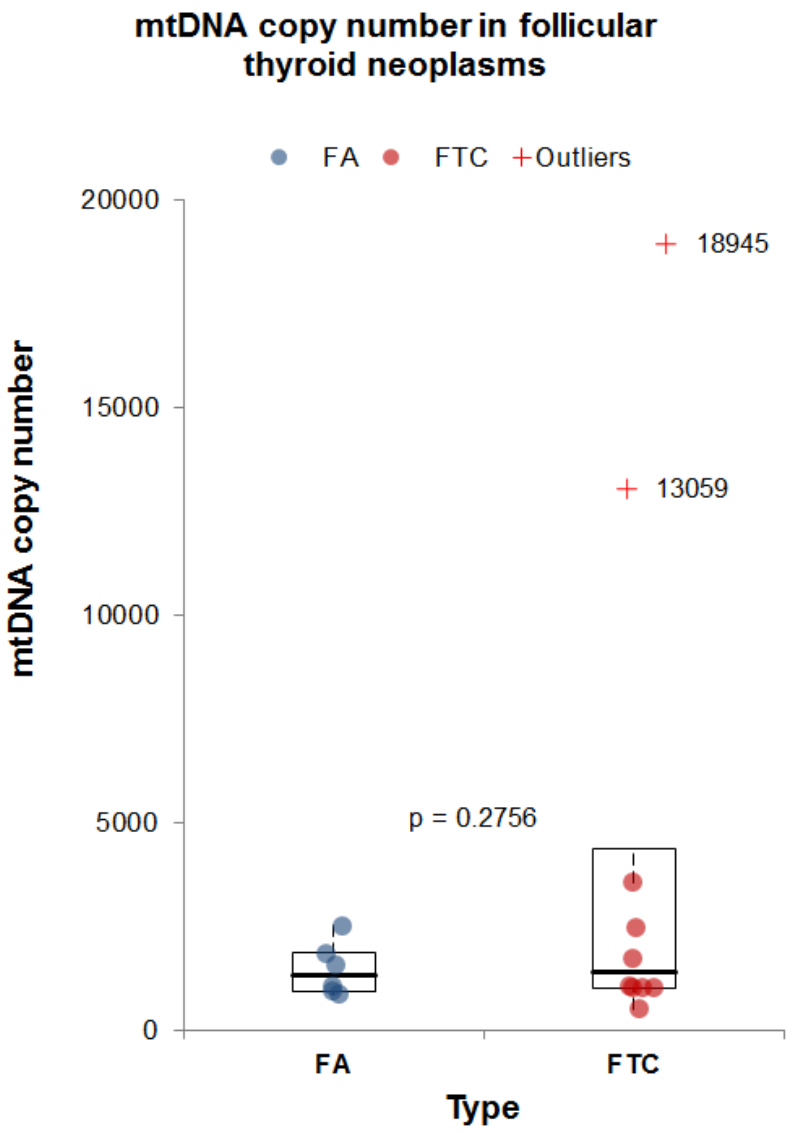

Figure 5-2: Mitochondrial DNA copy number was calculated for 6 follicular thyroid adenomas (FA) and 10 follicular thyroid carcinomas (FTC). Two outliers originated from Hurthle cell tumors that had been mistaken as FTC.

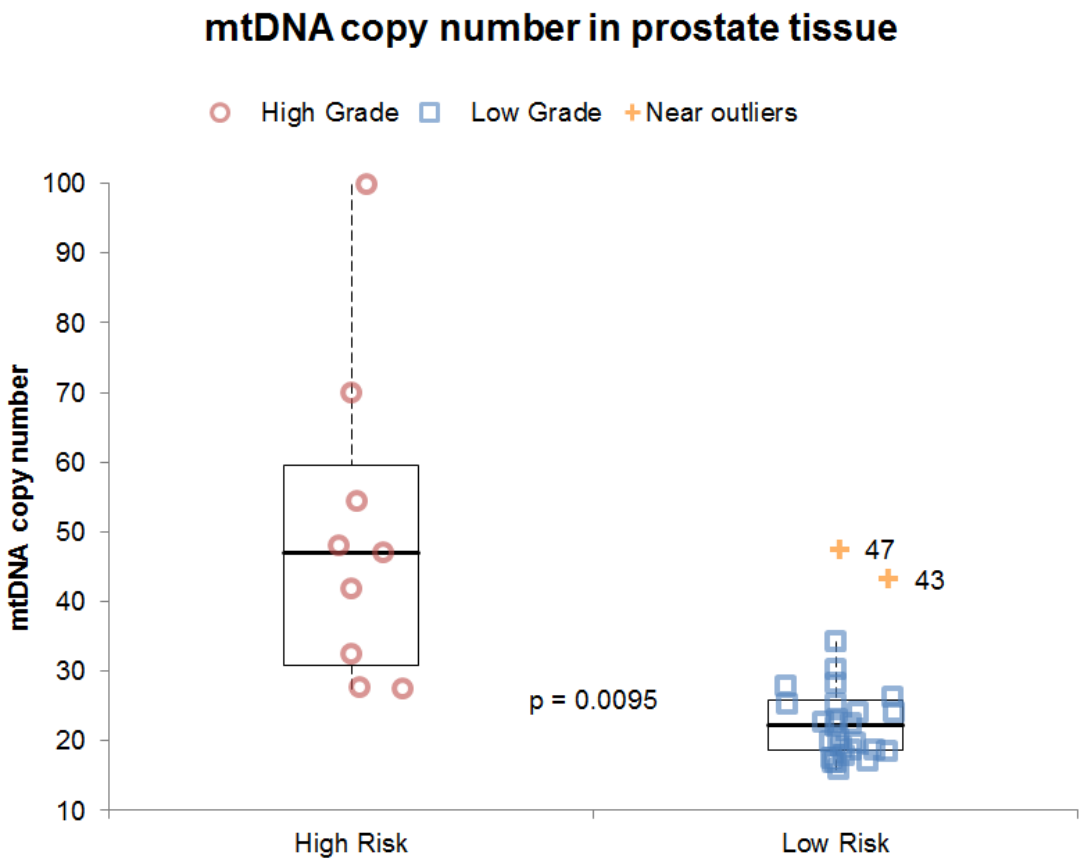

Figure 5-3: Mitochondrial DNA copy number was calculated for samples of low-risk and high-risk prostate tumors.

In addition to thyroid and prostate tissues we were also able to evaluate mitochondrial DNA copy number in a few samples of normal and diseased lung tissue (2 normal, 3 adenocarcinoma, and 5 carcinoid tumors). These samples displayed the clearest separation of copy number between tissue type, although the sample size was extremely small (Figure 5-4). Because of this limitation, we performed the Freeman-Halton extension of the Fisher exact probability test (72) for 3 x 3 contingency tables to evaluate the data.

5.4: Discussion

Using our method of calculating mtDNA copy number based on normalized bridged coverage we were able to detect differences in the amount of mtDNA in different tissues. In follicular thyroid tissue we did not see appreciable differences in copy number between FTC and FA, however, we were able to detect Hurthle cell carcinoma in two cases that had initially been mis-diagnosed. In our study of prostate tumors, we observed differences between the copy number of high- and low-risk tumors, and we suspect that if sample sizes were large enough we would be able to divide the groups more precisely to determine the trend of copy number increase throughout the disease spectrum. The differences between tissue types in the lung study were the most clearly defined of those we observed.

In each of these categories, larger sample sizes are needed to further explore the correlation between mtDNA concentration and disease state. This problem brings to light the difficulty in obtaining tissue specimens (especially

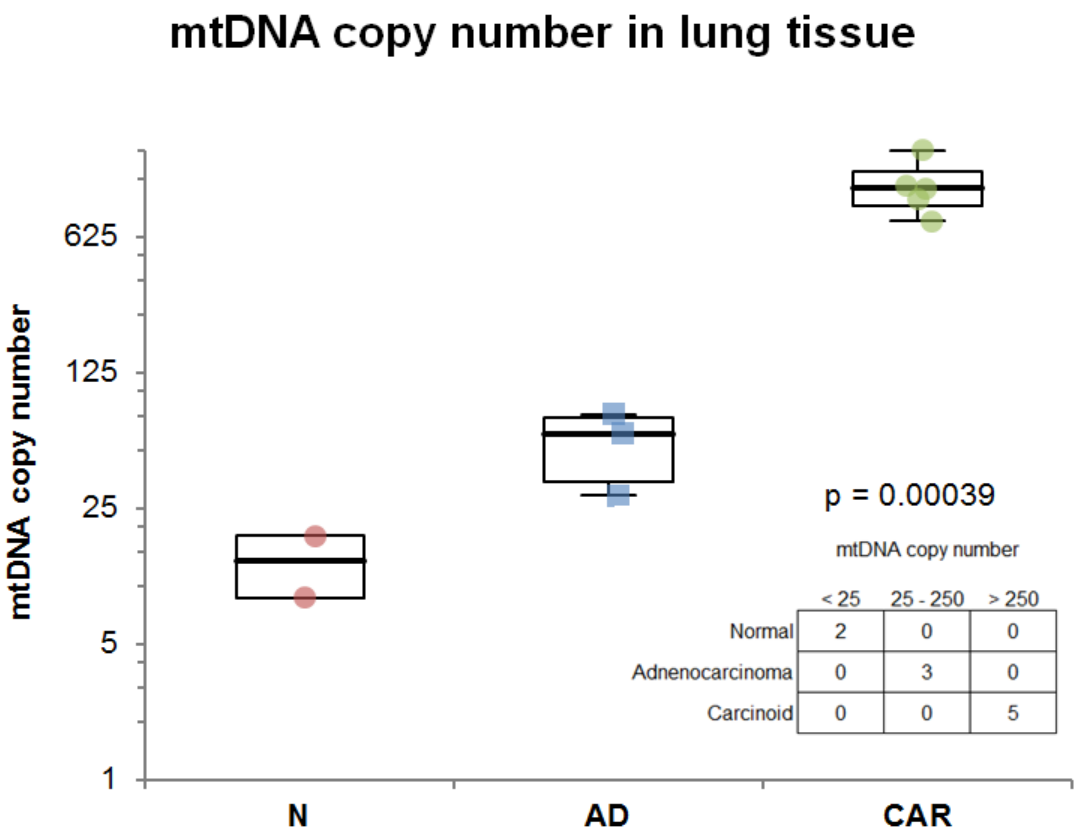

Figure 5-4: Mitochondrial DNA copy number for normal lung tissue (N), adenocarcinoma (AD) and carcinoid tumors (CAR). The p-value was calculated using the Freeman-Halton extension of the Fisher exact probability test.

normal tissue), as well as the expense of obtaining sequencing data. A less expensive method may be advantageous in this area. Quantitative methods such as digital droplet PCR have been shown to be extremely accurate for detecting relative copy numbers of genes, and could easily be extended to mtDNA. This would reduce the cost by several orders of magnitude, and allow testing to be done on extremely small samples such as residual needle biopsy washes.

The question of clinical utility for mtDNA copy number remains to be answered. While it is clear that mtDNA does indeed correlate with some disease states, it is also evident that this is not true for others. Further, as in the case of the thyroid study, our observations may reflect the difference between cell types rather than disease stage. While this may prove useful for cases where different tissue types are difficult to distinguish, it highlights the caution that that is necessary when tissue is initially categorized. Additional study is needed with carefully staged specimens to determine the true utility of mtDNA copy number in diseased tissue. However, if initial results can be expanded, there may be wide applicability for quantification of mitochondrial DNA in diagnostic workups.

To incorporate this method into a routine workflow for patient testing, a few additional considerations would need to be made. First and foremost, an evaluation of the specimen type would need to be completed. For routine clinical testing, most biopsy specimens are fine needle cores or aspirates. It will be important to understand the influence of contaminating normal cells on the

calculation and to know the acceptable limits. While this concept is true in any type of tissue collection and testing, it is especially true in fine needle biopsies where the percentage of malignant tissue in the sample will be variable but the user may not have any means to assess the concentration and correct for it.

Second, a robust evaluation of the average bridged coverage calculation would need to be made. Understanding the sample to sample variability in the calculation is important since many malignant specimens may contain significant regions of aneuploidy. If the influence of chromosomal gain or loss is large enough, calculations will be skewed and false results might be made. However, since chromosomal gain and loss can also be detected by mate pair sequencing, these factors can be observed and perhaps corrected for before calculation of mtDNA copy number.

Third, an orthogonal method would be extremely useful in the process of validating the clinical utility of mtDNA copy number. As previously mentioned, it is feasible that a quantitative PCR method could be created to accurately estimate the mtDNA concentration directly. Using a method of this nature might help to identify errors in the computation since the PCR process does not rely on accurate mapping and computation of an average bridged coverage.

Assuming we can understand the limitations of calculating mtDNA copy number there seems to be very good potential for its use as a clinical marker, either for disease, or to differentiate certain cell types. Even in the extremely limited population of specimens shown here, it seems clear that measurable

differences in mtDNA copy number do exist and can be measured using mate pair sequencing. These differences should be studied and understood so that they can be used clinically to improve human health.

Chapter 6: Summary

Advances in sequencing technology have transformed our ability to make observations of human genetics. However, it is the interpretation of what we observe that will translate knowledge into clinical utility. The path forward in clinical decision-making based on individual genomics will require a great deal of effort and the development of new methods for analysis. The applications developed herein represent a step in that direction. The first application addresses an immediate need for accurate diagnosis of recessive diseases in which compound heterozygosity has been identified. The second offers a method for rapid evaluation of genomic structural variation that can give clinicians insight into how rearrangements may influence disease. And finally, the third provides a means for evaluating mitochondrial DNA copy number as a potential marker of disease. All of these applications are made possible by the fusion of massive NGS datasets and computational power. They demonstrate unique ways in which NGS will continue to become a powerful tool for the clinic.

Bibliography

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.
2. Welch JS, Westervelt P, Ding L, et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 2011;305:1577-84.
3. Link DC, Schuettpelz LG, Shen D, et al. Identification of a novel tp53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related aml. JAMA 2011;305:1568-76.
4. Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I, Connolly S, et al. Clinical diagnosis by whole-genome sequencing of a prenatal sample. New England Journal of Medicine 2012;367:2226-32.
5. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour 2011;11:759-69.
6. Gullapalli RR, Lyons-Weiler M, Petrosko P, Dhir R, Becich MJ, LaFramboise WA. Clinical integration of next-generation sequencing technology. Clin Lab Med 2012;32:585-+.
7. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for mendelian disease gene discovery. Nat Rev Genet 2011;12:745-55.
8. Online mendelian inheritance in man, omim®. World Wide Web URL: http://omim.org/ (Accessed 1/23/2013 2013).
9. Salem R, Wessel J, Schork N. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Human Genomics 2005;2:39 - 66.
10. Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, et al. Conversion of diploidy to haploidy - individuals susceptible to multigene disorders may now be spotted more easily. Nature 2000;403:723-4.
11. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 2001;28:361-4.
12. Fan HC, Wang J, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. Nat Biotechnol 2011;29:51-7.
13. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, et al. Haplotype-resolved genome sequencing of a gujarati indian individual. Nat Biotechnol 2011;29:59-63.
14. Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, Bibikova M, et al. Whole-genome haplotyping by dilution, amplification, and sequencing. Proceedings of the National Academy of Sciences 2013.
15. Kircher M, Stenzel U, Kelso J. Improved base calling for the illumina genome analyzer using machine learning strategies. Genome Biology 2009;10.

16. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of illumina sequencers. Nucleic Acids Res 2011;39:e90.
17. Luo CW, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. Plos One 2012;7.
18. Tsai LP, Cheng CF, Chuang SH, Lee HH. Analysis of the cyp21a1p pseudogene: Indication of mutational diversity and cyp21a2-like and duplicated cyp21a2 genes. Anal Biochem 2011;413:133-41.
19. Concolino P, Mello E, Zuppi C, Capoluongo E. Molecular diagnosis of congenital adrenal hyperplasia due to 21-hydroxylase deficiency: An update of new cyp21a2 mutations. Clin Chem Lab Med 2010;48:1057-62.
20. Murphy SJ, Cheville JC, Zarei S, Johnson SH, Sikkink RA, Kosari F, et al. Mate pair sequencing of whole-genome-amplified DNA following laser capture microdissection of prostate cancer. DNA Res 2012;19:395-406.
21. Lander ES, Consortium IHGS, Linton LM, Birren B, Nusbaum C, Zody MC, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
22. Vasmatzis G, Johnson SH, Knudson RA, Ketterling RP, Braggio E, Fonseca R, et al. Genome-wide analysis reveals recurrent structural abnormalities of tp63 and other p53-related genes in peripheral t-cell lymphomas. Blood 2012;120:2280-9.
23. Feldman AL, Dogan A, Smith DI, Law ME, Ansell SM, Johnson SH, et al. Massively parallel mate pair DNA library sequencing for translocation discovery: Recurrent t(6;7)(p25.3;q32.3) translocations in alk-negative anaplastic large cell lymphomas. Blood 2010;116:278-.
24. Salem RM, Wessel J, Schork NJ. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2005;2:39-66.
25. Xie M, Wang J, Jiang T. A fast and accurate algorithm for single individual haplotyping. Bmc Syst Biol 2012;6 Suppl 2:S8.
26. Bansal V, Bafna V. Hapcut: An efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 2008;24:i153-9.
27. Cradic K, Murphy S, Drucker T, Sikkink R, Eberhardt N, Neuhauser C, et al. A simple method for gene phasing using mate pair sequencing. BMC Medical Genetics 2014;15:19.
28. Lee HH. Variants of the cyp21a2 and cyp21a1p genes in congenital adrenal hyperplasia. Clin Chim Acta 2013;418:37-44.
29. Drucker TM, Johnson SH, Murphy SJ, Cradic KW, Therneau TM, Vasmatzis G. Bima v3: An aligner customized for mate pair library sequencing. Bioinformatics 2014.
30. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195-7.
31. Markov AA. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. In: Howard R, ed. Dynamic probabilistic systems, volume 1: Markov chains, Vol. 1: John Wiley and Sons, 1971.

32. Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distributions. Bull Calcutta Math Soc 1943;35:99-109.
33. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. Science 2005;310:644-8.
34. Lin E, Li L, Guan YH, Soriano R, Rivers CS, Mohan S, et al. Exon array profiling detects eml4-alk fusion in breast, colorectal, and non-small cell lung cancers. Mol Cancer Res 2009;7:1466-76.
35. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming eml4-alk fusion gene in non-small-cell lung cancer. Nature 2007;448:561-U3.
36. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. Cosmic: Exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805-11.
37. Grossmann V, Kohlmann A, Klein HU, Schindela S, Schnittger S, Dicker F, et al. Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure. Leukemia 2011;25:671-80.
38. Wang QG, Xia JF, Jia PL, Pao W, Zhao ZM. Application of next generation sequencing to human gene fusion detection: Computational tools, features and perspectives. Briefings in Bioinformatics 2013;14:506-19.
39. Feldman AL, Dogan A, Smith DI, Law ME, Ansell SM, Johnson SH, et al. Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in alk-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. Blood 2011;117:915-9.
40. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010;38:D355-D60.
41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25:25-9.
42. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. Wikipathways: Pathway editing for the people. PLoS Biol 2008;6:e184.
43. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, et al. The ncbi biosystems database. Nucleic Acids Res 2010;38:D492-6.
44. Mortensen JD, Woolner LB, Bennett WA. Gross and microscopic findings in clinically normal thyroid glands. J Clin Endocrinol Metab 1955;15:1270-80.
45. Yang J, Schnadig V, Logrono R, Wasserman PG. Fine-needle aspiration of thyroid nodules: A study of 4703 patients with histologic and clinical correlations. Cancer Cytopathol 2007;111:306-15.

46. Yassa L, Cibas ES, Benson CB, Frates MC, Doubilet PM, Gawande AA, et al. Long-term assessment of a multidisciplinary approach to thyroid nodule diagnostic evaluation. Cancer Cytopathol 2007;111:508-16.
47. Baloch ZW, LiVolsi VA, Asa SL, Rosai J, Merino MJ, Randolph G, et al. Diagnostic terminology and morphologic criteria for cytologic diagnosis of thyroid lesions: A synopsis of the national cancer institute thyroid fine-needle aspiration state of the science conference. Diagn Cytopathol 2008;36:425-37.
48. Lemoine NR, Mayall ES, Wyllie FS, Farr CJ, Hughes D, Padua RA, et al. Activated ras oncogenes in human thyroid cancers. Cancer Res 1988;48:4459-63.
49. Marques AR, Espadinha C, Catarino AL, Moniz S, Pereira T, Sobrinho LG, Leite V. Expression of pax8-ppar gamma 1 rearrangements in both follicular thyroid carcinomas and adenomas. J Clin Endocr Metab 2002;87:3947-52.
50. Witt RL, Ferris RL, Pribitkin EA, Sherman SI, Steward DL, Nikiforov YE. Diagnosis and management of differentiated thyroid cancer using molecular biology. Laryngoscope 2013;123:1059-64.
51. Harris TJ, Tepass U. Adherens junctions: From molecules to morphogenesis. Nat Rev Mol Cell Biol 2010;11:502-14.
52. Janssens B, Goossens S, Staes K, Gilbert B, van Hengel J, Colpaert C, et al. Alphat-catenin: A novel tissue-specific beta-catenin-binding protein mediating strong cell-cell adhesion. J Cell Sci 2001;114:3177-88.
53. Zhang R, Yao W, Qian P, Li Y, Jiang C, Ao Z, et al. Increased sensitivity of human lung adenocarcinoma cells to cisplatin associated with downregulated contactin-1. Biomed Pharmacother 2015;71:172-84.
54. Liu PF, Chen SM, Wu WT, Liu BT, Shen WD, Wang FJ, et al. Contactin-1 (cntn-1) overexpression is correlated with advanced clinical stage and lymph node metastasis in oesophageal squamous cell carcinomas. Jpn J Clin Oncol 2012;42:612-8.
55. Chen DH, Yu JW, Wu JG, Wang SL, Jiang BJ. Significances of contactin-1 expression in human gastric cancer and knockdown of contactin-1 expression inhibits invasion and metastasis of mkn45 gastric cancer cells. J Cancer Res Clin 2015;141:2109-20.
56. Shi KY, Xu D, Yang C, Wang LP, Pan WY, Zheng CM, Fan LY. Contactin 1 as a potential biomarker promotes cell proliferation and invasion in thyroid cancer. Int J Clin Exp Pathol 2015;8:12473-81.
57. Yu J, Becka S, Zhang P, Zhang X, Brady-Kalnay SM, Wang Z. Tumor-derived extracellular mutations of ptprt /ptprho are defective in cell adhesion. Molecular cancer research : MCR 2008;6:1106-13.
58. Raman P, Koenig RJ. Pax-8-ppar-gamma fusion protein in thyroid carcinoma. Nature reviews Endocrinology 2014;10:616-23.
59. Eberhardt NL, Grebe SKG, McIver B, Reddi HV. The role of the pax8/ppar gamma fusion oncogene in the pathogenesis of follicular thyroid cancer. Mol Cell Endocrinol 2010;321:50-6.

60. French CA. Demystified molecular pathology of nut midline carcinomas. J Clin Pathol 2010;63:492-6.
61. Shi J, Vakoc CR. The mechanisms behind the therapeutic activity of bet bromodomain inhibition. Mol Cell 2014;54:728-36.
62. Peltekova VD, Lemire M, Qazi AM, Zaidi SHE, Trinh QM, Bielecki R, et al. Identification of genes expressed by immune cells of the colon that are regulated by colorectal cancer- associated variants. International Journal of Cancer 2014;134:2330-41.
63. Kasem K, Sullivan E, Gopalan V, Salajegheh A, Smith RA, Lam AK. Jk1 (fam134b) represses cell migration in colon cancer: A functional study of a novel gene. Exp Mol Pathol 2014;97:99-104.
64. Kasem K, Gopalan V, Salajegheh A, Lu CT, Smith RA, Lam AK. The roles of jk-1 (fam134b) expressions in colorectal cancer. Exp Cell Res 2014;326:166-73.
65. Jauliac S, Lopez-Rodriguez C, Shaw LM, Brown LF, Rao A, Toker A. The role of nfat transcription factors in integrin-mediated carcinoma invasion. Nat Cell Biol 2002;4:540-4.
66. Szuhai K, Ijszenga M, de Jong D, Karseladze A, Tanke HJ, Hogendoorn PCW. The nfatc2 gene is involved in a novel cloned translocation in a ewing sarcoma variant that couples its function in immunology to oncology. Clin Cancer Res 2009;15:2259-68.
67. Chen CR, Kang YB, Siegel PM, Massague J. E2f4/5 and p107 as smad cofactors linking the tgf beta receptor to c-myc repression. Cell 2002;110:19-32.
68. Holt IJ, Harding AE, Morganhughes JA. Deletions of muscle mitochondrial-DNA in patients with mitochondrial myopathies. Nature 1988;331:717-9.
69. Wallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AMS, et al. Mitochondrial-DNA mutation associated with lebers hereditary optic neuropathy. Science 1988;242:1427-30.
70. Brandon MC, Lott MT, Nguyen KC, Spolim S, Navathe SB, Baldi P, Wallace DC. Mitomap: A human mitochondrial genome database - 2004 update. Nucleic Acids Res 2005;33:D611-D3.
71. Lee HC, Wei YH. Mitochondrial biogenesis and mitochondrial DNA maintenance of mammalian cells under oxidative stress. Int J Biochem Cell B 2005;37:822-34.
72. Freeman GH, Halton JH. Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 1951;38:141-9.

Appendix A: Tables of genes associated to disease by keywords

For each specimen tested, a table contains summary information on the BioSystems database annotations for every gene affected by a breakpoint.

Legend

Filter 1 - Genes were found to be affected by a genomic junction point

Filter 2 - Genes were removed because they were not associated to disease by a keyword

Filter 3 – Genes were removed because they did not have a functional match. These are typically contextual errors

Filter 4 – Genes were removed because the presumed product was deemed irrelevant to disease.

N/A – Genes that passed all filters and were retained for additional study.

NR – No record was available in the BioSystems database.

* Gene was falsely eliminated by text search.

Sample 1 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
CCDC91	1	Protein transport - through membrane	NR	NR
CHCHD6	1	Mitochondrial critae formation and organization	NR	NR
FAM134B		reticulophagy; negative regulation of neuron apoptosis	NR	NR
FAM172A	1	NR	Endoplasmic reticulum	NR
LIN52	1	Involved in G0 - G1 events	Part of the DRM complex (a transcriptional repressor)	NR
LINC00871	1	NR	NR	NR
NLRC3	1	Negative regulation of interleukin production	Intercellular component	Binds nucleotides - ATP
NRXN3	2	Cell signaling in the intersynaptic space. Protein sits on the pre-synaptic cell	presynaptic membrane complex	binds with neuroligin protein family
PDS5A	1	Cohesin loading on chromatin - meiotic replication of chromosomes	chromosome, centromeric region	NR
PIGQ	2	synthesis of GPI-anchored proteins	glycosylphosphatidylinositol-N-acetylglucosaminyltransferase (GPI-GnT) complex; ER membrane component	phosphatidylinositol N-acetylglucosaminyltransferase activity
SLX4	1	Replication fork repair; initiation of telomeric circle formation.	rDNA (replication) complexes	3' and 5' DNA flap endonuclease activity
TMCC1	1	NR	Endoplasmic reticulum	NR
WDR90	1	NR	NR	NR

Sample 2 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
CDH4	N/A	Cell adhesion and cell-to-cell junction points	Part of a complex embedded in the plasma membrane	Binds calcium
CNTN1	N/A	Contactin-1 - Involved in NOTCH1 and NOTCH2 signaling.	Part of an anchored membrane complex	Binds glycoproteins and/or sugars
DLGAP4	1	Signaling - very generic	Part of dendrites, neurons, and/or synapse	Binds proteins
DOK5	3	Ret signaling pathway	NR	Receptor signaling
FRMD3	1	None associated	Cytoskeleton	cytoskeleton binding
GNAS-AS1	1	GNAS is associated with many pathways	NR	Silencing of GNAS
HNF4A	2	Negative regulation of JAK2 kinase; Negative regulation of STAT5 phosphorylation	Part of a nuclear transcription factor	promotor specific binding
LARP4	1	Cell morphogenesis - and cytoskeletal organization	Membrane component?	Binds RNA - polyA
LINC00494	1	NR	NR	NR
MIR4457	1	NR	NR	NR
MUC19	1	O-glycosylation; involved in many pathways	Golgi -> extracellular membrane	NR
NFATC2	2	NFAT activation	Nuclear pore	NFAT binding
PLTP	1	Regulation of high-density lipoprotein particle formation	Extracellular space	Lipid binding and transport
PREX1	1	Positive regulation of several signaling pathways	found in projecting regions of the cell - growth cone, flagella, etc. Often found in neuronal cells	GTPase activity
PTPN9	1	Several tyrosine dephosphorylation pathways	cytoplasmic	phosphatase activity

Sample 2 – Follicular Carcinoma (continued)				
PTPRT	N/A	cell-to-cell adhesion and intercellular signaling	cell surface	Actinin binding; catenin binding; cadherin binding
SALL4	1	FGF receptor pathway regulation	Transcription factor complex	DNA binding
SERINC3	2	Pathways that depend on serine	Golgi -> extracellular membrane	Serine transport between cells
SLA	1	Sulfoquinovose pathways	MHC transmembrane complex	NR
TG	N/A	Thyroid hormone biosynthesis	Extracellular	Thyroxine synthesis
TSHZ2	1	CR	CR	CR

Sample 3 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
DNAH2	1	Microtubule based movement	Complex with microtubule structures	Movement along microtubules; ATPase activity
GRID2	1	Cerebellar structure formation; Long-term depression	Post-synaptic membrane; ionotropic glutamate receptor complex	extracellular-glutamate-gated ion channel activity

Sample 4 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
BRD4	1*	Histone acetylation; DNA damage checkpoint; Signal transduction resulting in cell cycle arrest	Part of a nuclear complex that is responsible for transcription elongation.	Binds p53; Binds chromatin; Binds histones
COLCA2	1*	NR	Cytoplasmic protein	NR
HIATL1	1	Transmembrane transport	Integral membrane part	NR
RBL1	3	Regulation of cell cycle (murine homolog p107)	Part of a cytoplasmic/nuclear repressor complex that regulates MYC expression	Cell signaling
VPS13D	1	Protein association with cell organelles	Part of membranes of extracellular vesicles.	Binding activity
ZNF169	1	Generic transcription pathways	Nuclear transcription factor	Metal binding; Nucleic acid binding

Sample 5 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
PDZD2	1	Cellular adhesion; Signal transduction	Postsynaptic junction	Protein binding
GOLPH3	3	Seems to be associated with golgi function for production of proteins that end up on the outside of the cell and help communicate with or adhere to other cells.	found in the golgi membrane. Also found in the mitochondria	binds phosphatidylinositols

Sample 6 – Follicular Adenoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
AKAP3	1	Cell surface receptor signaling via g proteins; Fertilization	Sperm fibrous sheath	Protein kinase a binding
KLHL4	1	Involved in ubiquitination of proteins	Ubiquitin ligase complex	Actin binding; ubiquitin transferase activity
PAX8	N/A	Thyroid hormone secretion	NR	NR
PPARG	N/A	Regulation of adipocyte differentiation	Chromatin binding structure	DNA binding
SARNP	1	mRNA transport from the nucleus	Transcription export complex	RS binding activity; Some transcriptional repressor activity
SATB2	1	Development of neurons	Nuclear; histone deacetylase	Sequence specific DNA binding

Sample 7 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
TANGO2	1	Secretion from the cell	Golgi	NR

Sample 8 – Follicular Adenoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
FBXL5	1	Protein folding	Chaperone folding complex	NR
PTGIS	1	Nicotinomide salvage	Plasma membrane raft forming caveola	Prostaglandin synthesis
TRMT10C	1	Mitochondrial tRNA processing	Mitochondrial matrix	Methyltransferase activity
TTC28	1	Functions in nuclear division during mitosis	Part of the mitotic spindle complex of microtubules that form between DNA	NR

Sample 9 – Follicular Adenoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
ASMT	1	Seratonin and melatonin biosynthesis	Cytoplasmic	Forms a homodimer; Acetylserotonin O-methyltransferase activity
CNST	1	Regulation of golgi protein transport mechanisms	Golgi -> cell membrane	Binds to connexin in gap junction complexes
CTNNA3	N/A	Involved in heterotypic cell adhesion and communication	Golgi network and plasma membrane	Found in facia adherens; Adherens junctions

Sample 10 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
MIR512-2	1	NR	NR	NR
SORBS1	2	Involved in smooth muscle contraction; Involved in cell adhesion	Found in the insulin receptor complex; Found in the flotillin complex of the plasma membrane; Associated with actomyosin complexes	Seems to have various binding capabilities via SH2 and SH3 domain interactions
ZNF667	1	Generic gene expression pathways	Nucleus	Transcription factor - binding DNA and metal

Sample 11 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
AGPAT3	1	Triacylglycerol synthesis	Endoplasmic reticulum and the outer nuclear envelope	Acyltransferase activity
BAGE	1	NR	Secreted protein	NR
NFIA	1	RNA polymerase III transcription	Found in the nucleus	Transcription factor
NOL4	1	NR	Found in the nucleus - part of the nucleolus	Binds to nucleic acid and protein

Sample 12 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
LINC01194		NR	NR	NR

Sample 13 – Follicular Carcinoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
GAPDH	1	Glycolysis and gluconeogenesis	Cytoplasmic	Phosphorylation
GPBP1	1	Positive regulation of transcription	nucleolus and nucleus	transcription factor and DNA binding activity
KCCAT333	1	NR	NR	NR
NOL4L	1	NR	nucleolus and nucleus	binds to nucleic acid and protein
PAX8	N/A	Thyroid hormone secretion	NR	NR
PPARG	N/A	Regulation of adipocyte differentiation	Chromatin binding structure	DNA binding

Sample 14 – Follicular Adenoma				
Gene	Filter	Pathway	Structure	Function
PAX8	N/A	Thyroid hormone secretion	NR	NR
PPARG	N/A	Regulation of adipocyte differentiation	Chromatin binding structure	DNA binding

Sample 15 – Follicular Adenoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
PIP4K2A	1	Synthesis of phosphatidylinositol phosphates	Cytoplasmic	Phosphatidylinositol phosphate kinase activity
ZNF100	1	Generic transcription pathways	Nuclear	DNA binding

Sample 16 – Follicular Adenoma				
Gene	**Filter**	**Pathway**	**Structure**	**Function**
MAP3K7	1	Many signaling pathways	Various protein complexes	Kinase activity; Receptor signaling
RBL1	3	Regulation of cell cycle (murine homolog p107) **Retroposon mapping**	Part of a cytoplasmic/nuclear repressor complex that regulates *MYC* expression	Cell signaling
VPS13D	1	Protein association with cell organelles	Part of membranes of extracellular vesicles.	Binding activity

Appendix B: Junction plots and region plots of structural rearrangements found in follicular thyroid tumors.

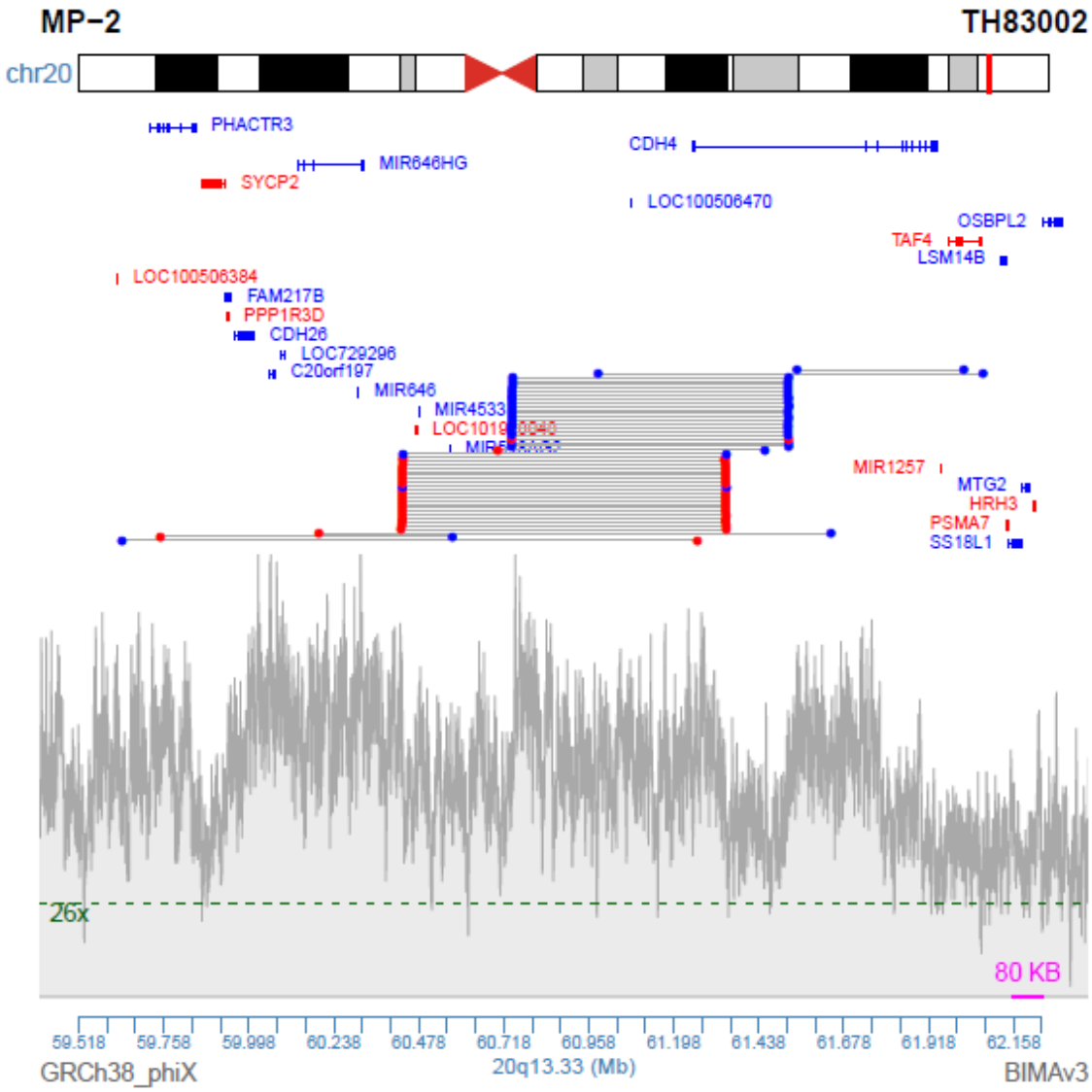

Figure B-1: Region plot showing two inversion events that disrupt the CDH4 gene. Both events leave the gene non-functional.

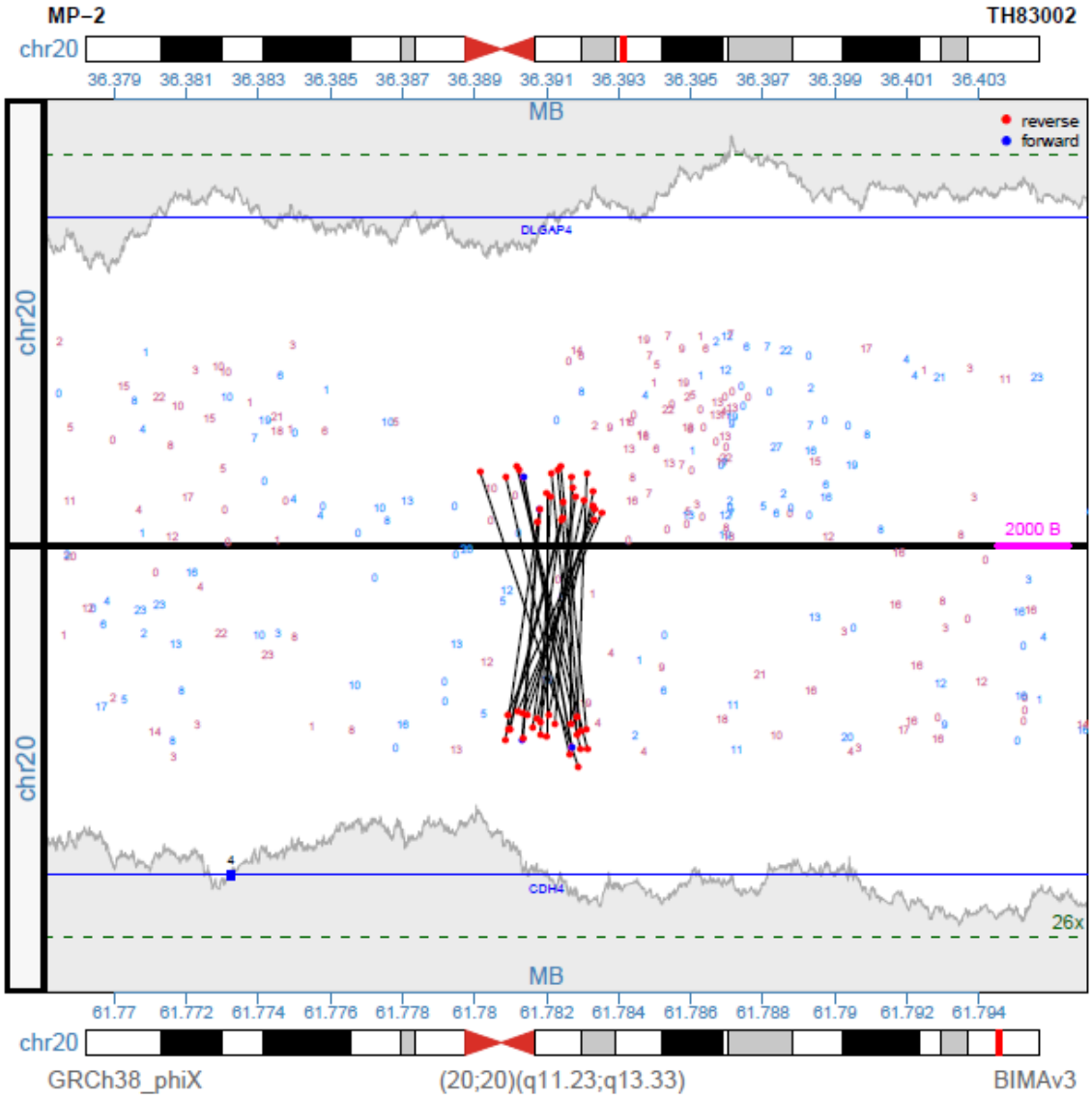

Figure B-2: Junction plot showing the third inversion affecting the CDH4 gene. This event leaves the gene non-functional.

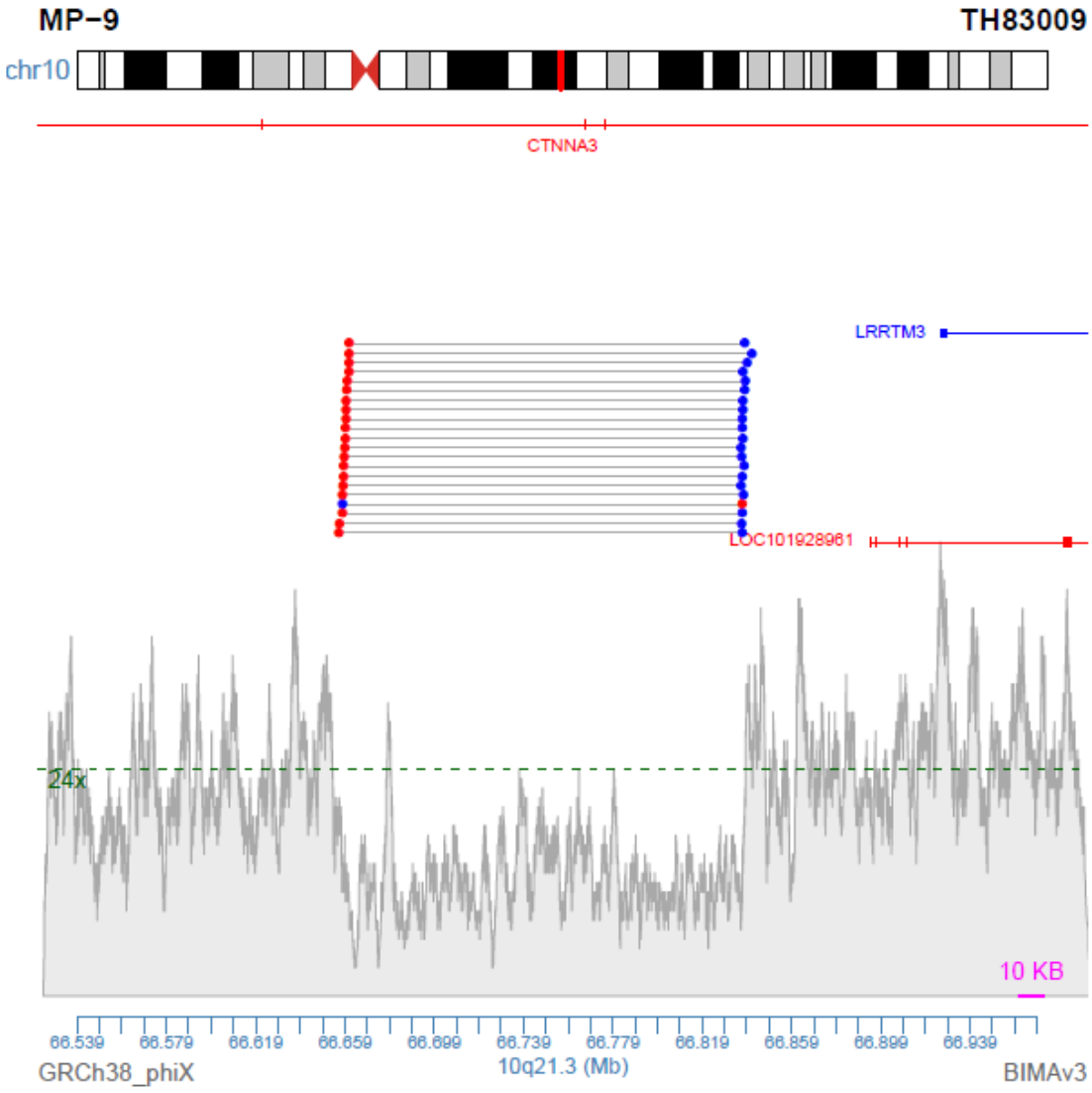

Figure B-3: Region plot showing a heterozygous deletion affecting CTNNA3. Exons 8 and 9 are deleted.

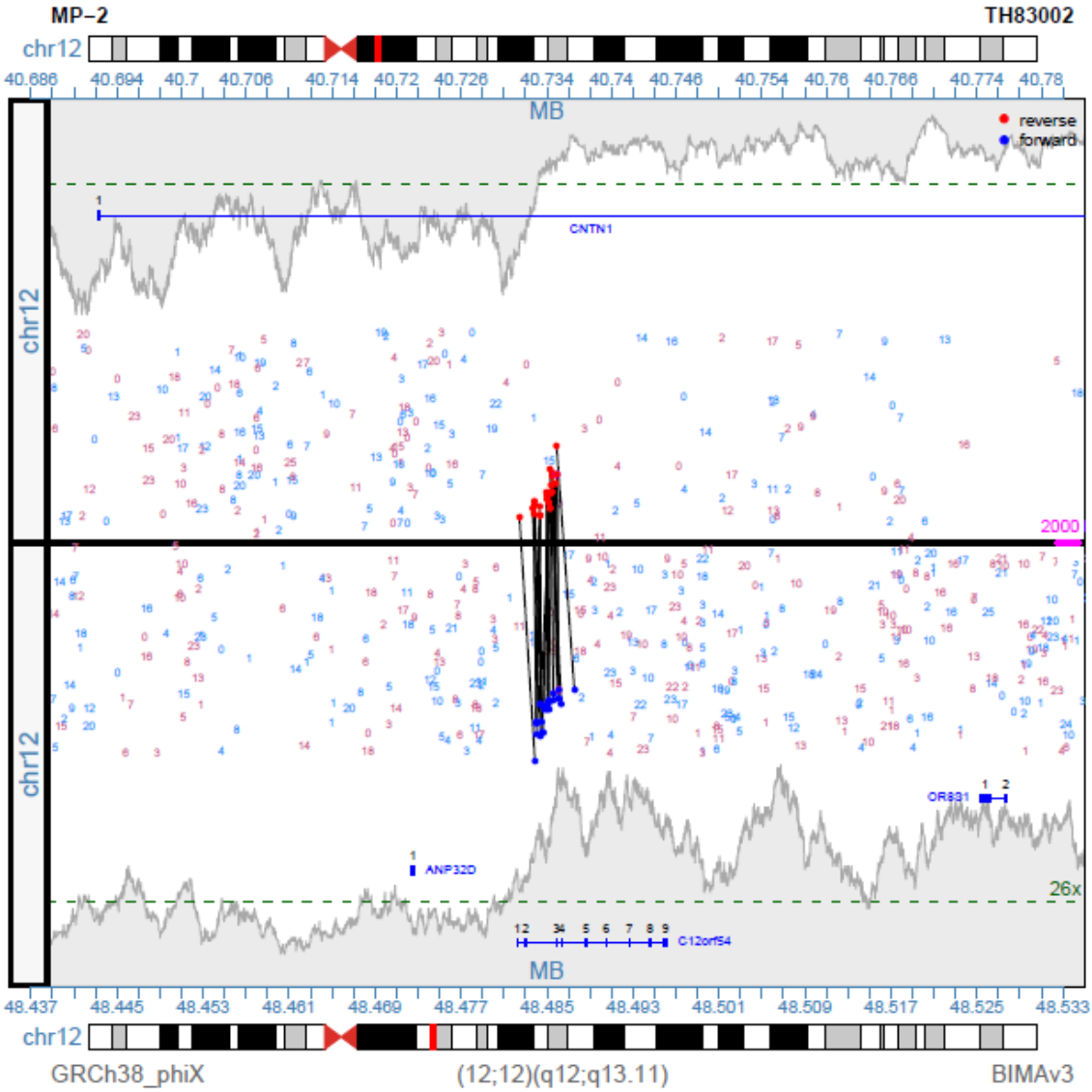

Figure B-4: Region plot showing a deletion involving CNTN1 that results in a fusion protein between CNTN1 and C12orf54. This deletion appears to be heterozygous.

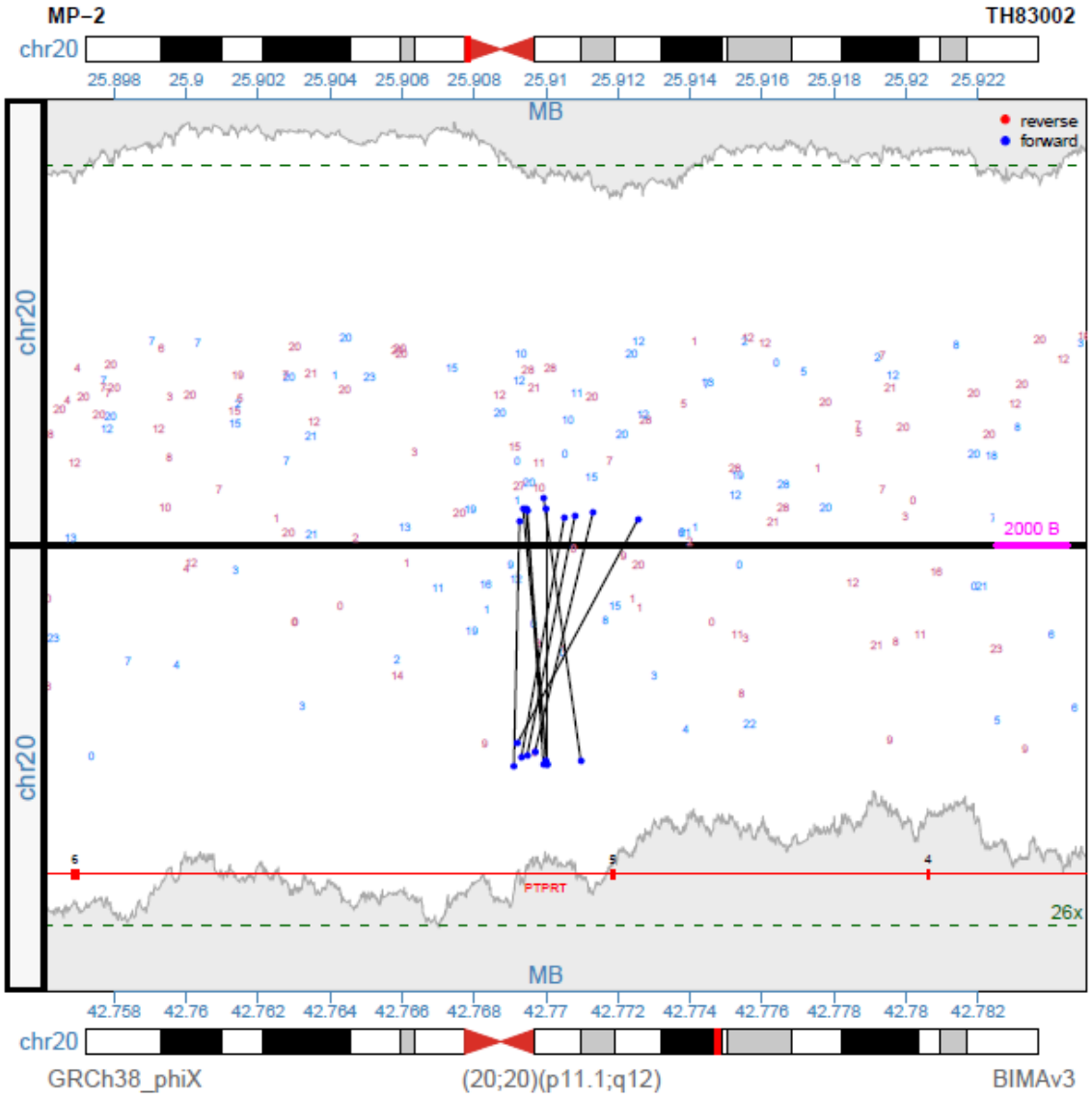

Figure B-5: Junction plot showing the first of three inversion events involving PTPRT. The affected gene will not be expressed.

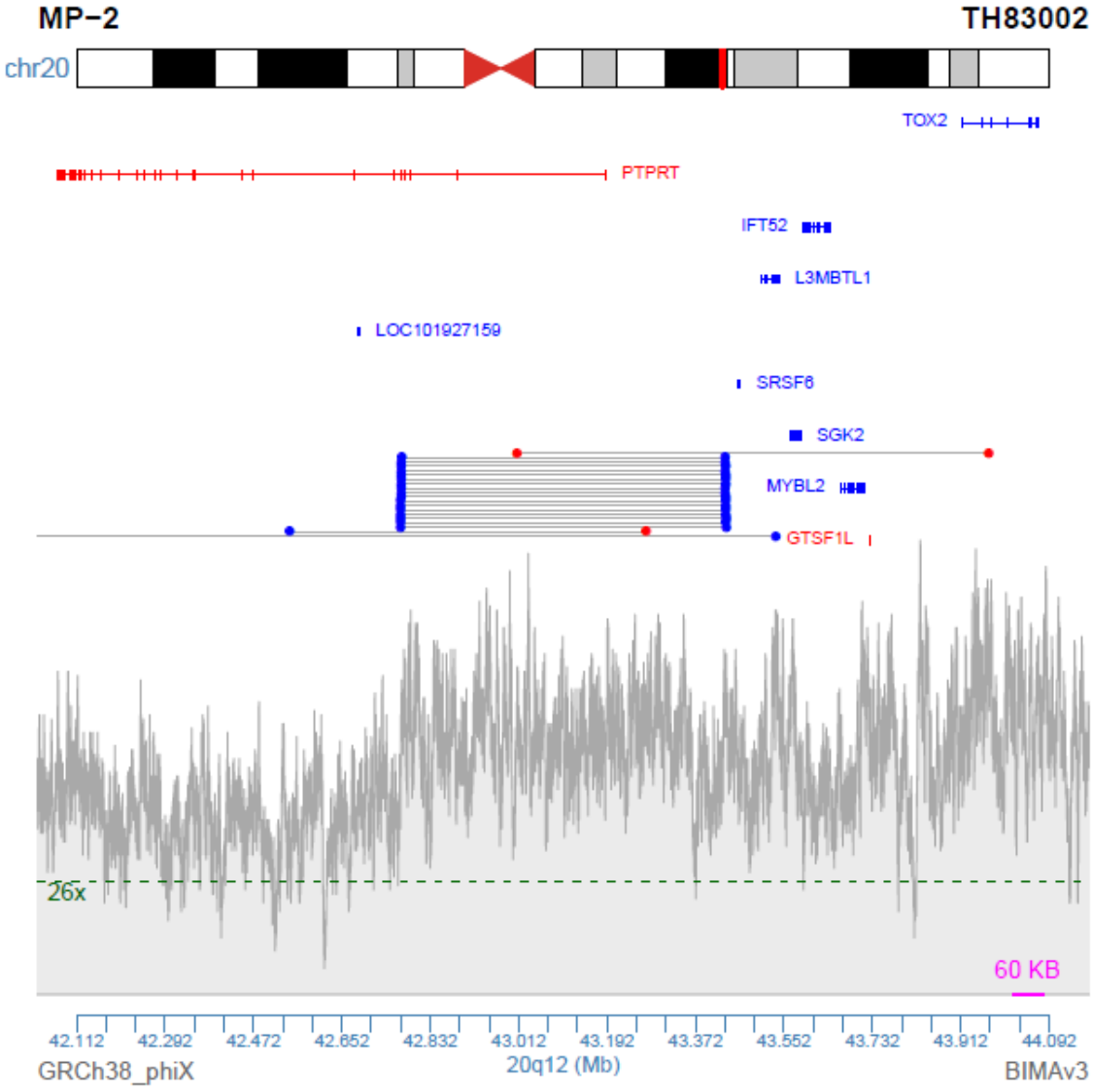

Figure B-6: Region plot showing the second of three inversions that affect the PTPRT gene. The gene affected by this inversion will not be expressed.

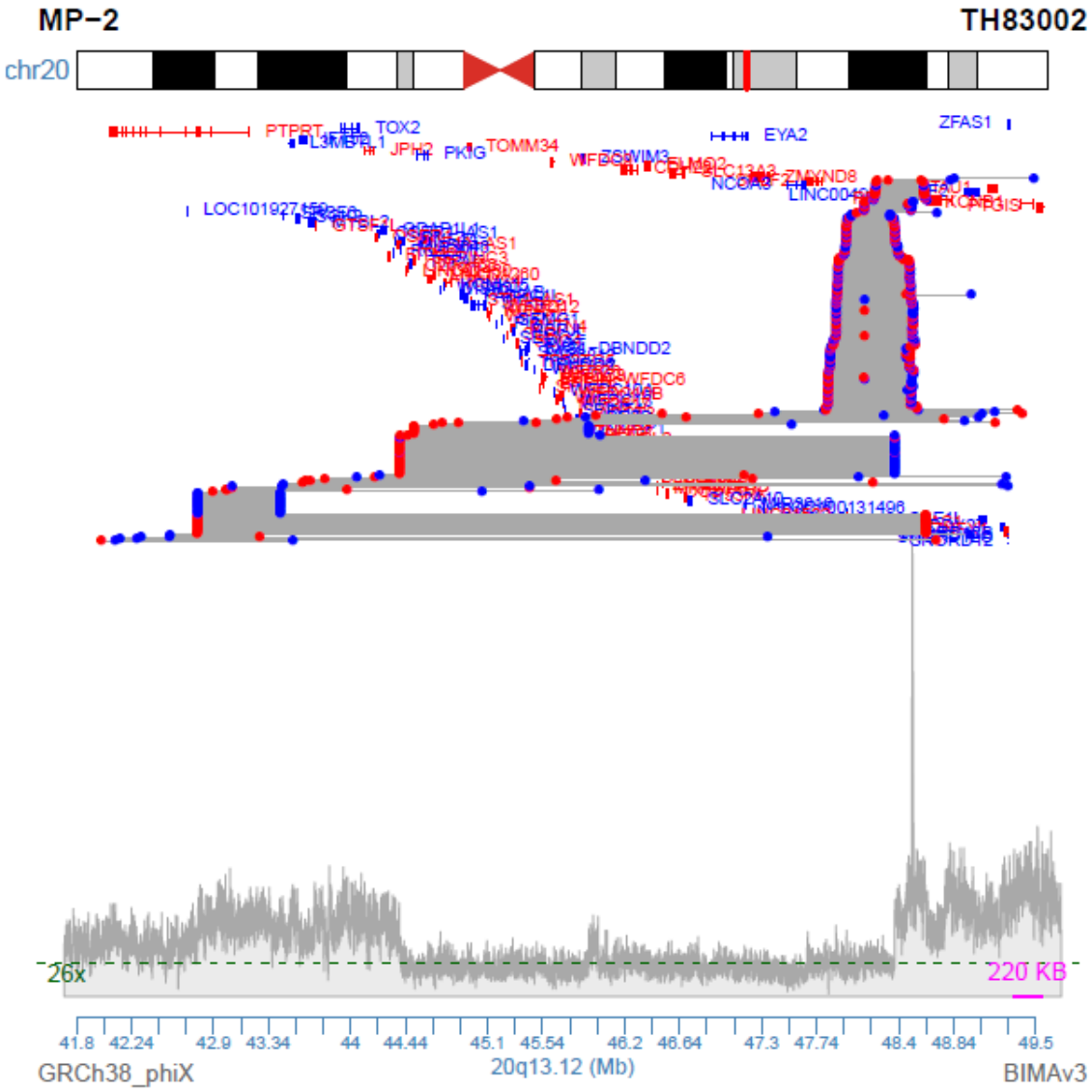

Figure B-7: Region plot showing the second and third of three inversion events that affect PTPRT. Both events eliminate expression of the PTPRT gene.

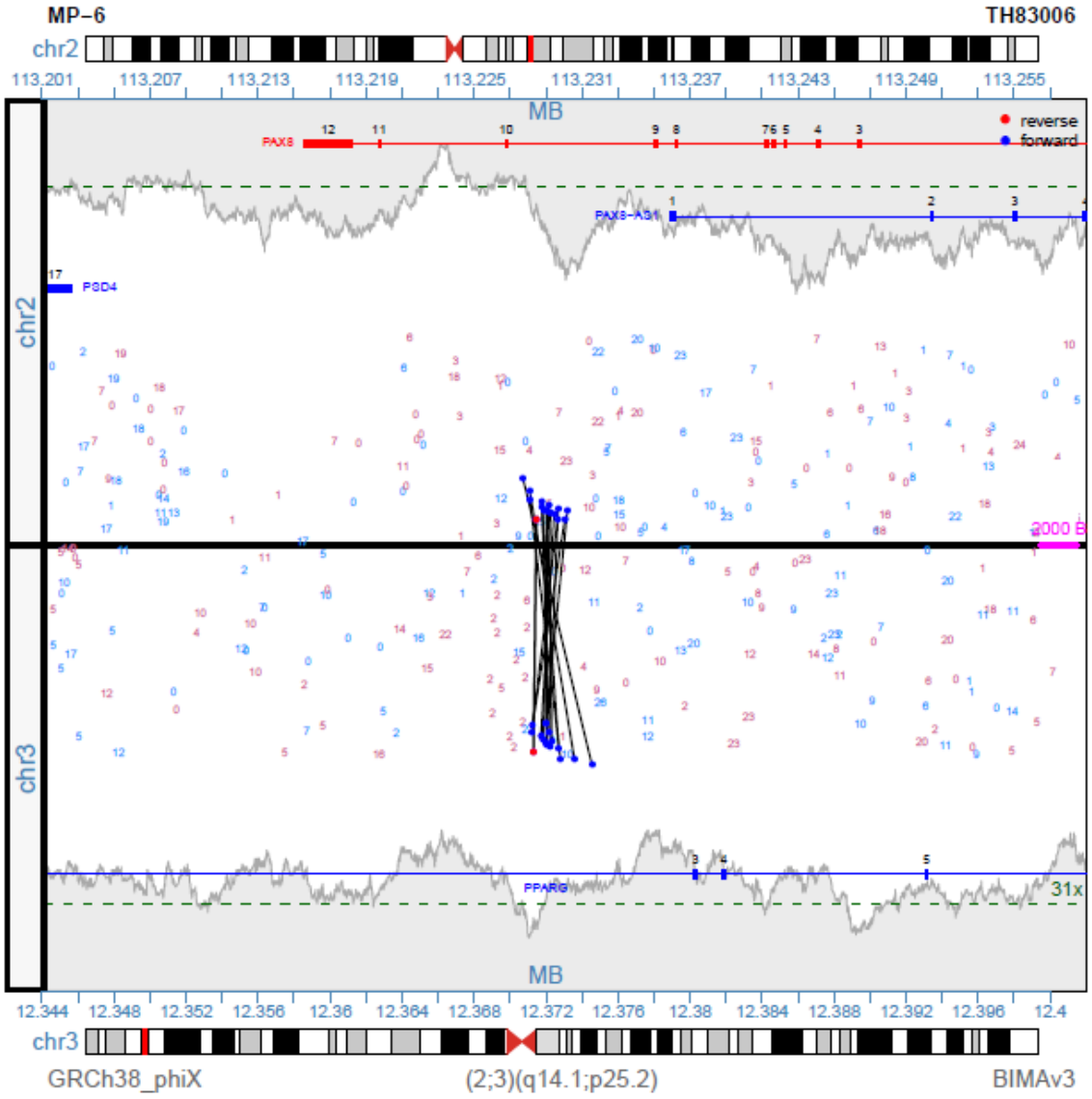

Figure B-8: Junction plot showing the first of three *PAX8/PPARg* translocations. This event results in a fusion protein.

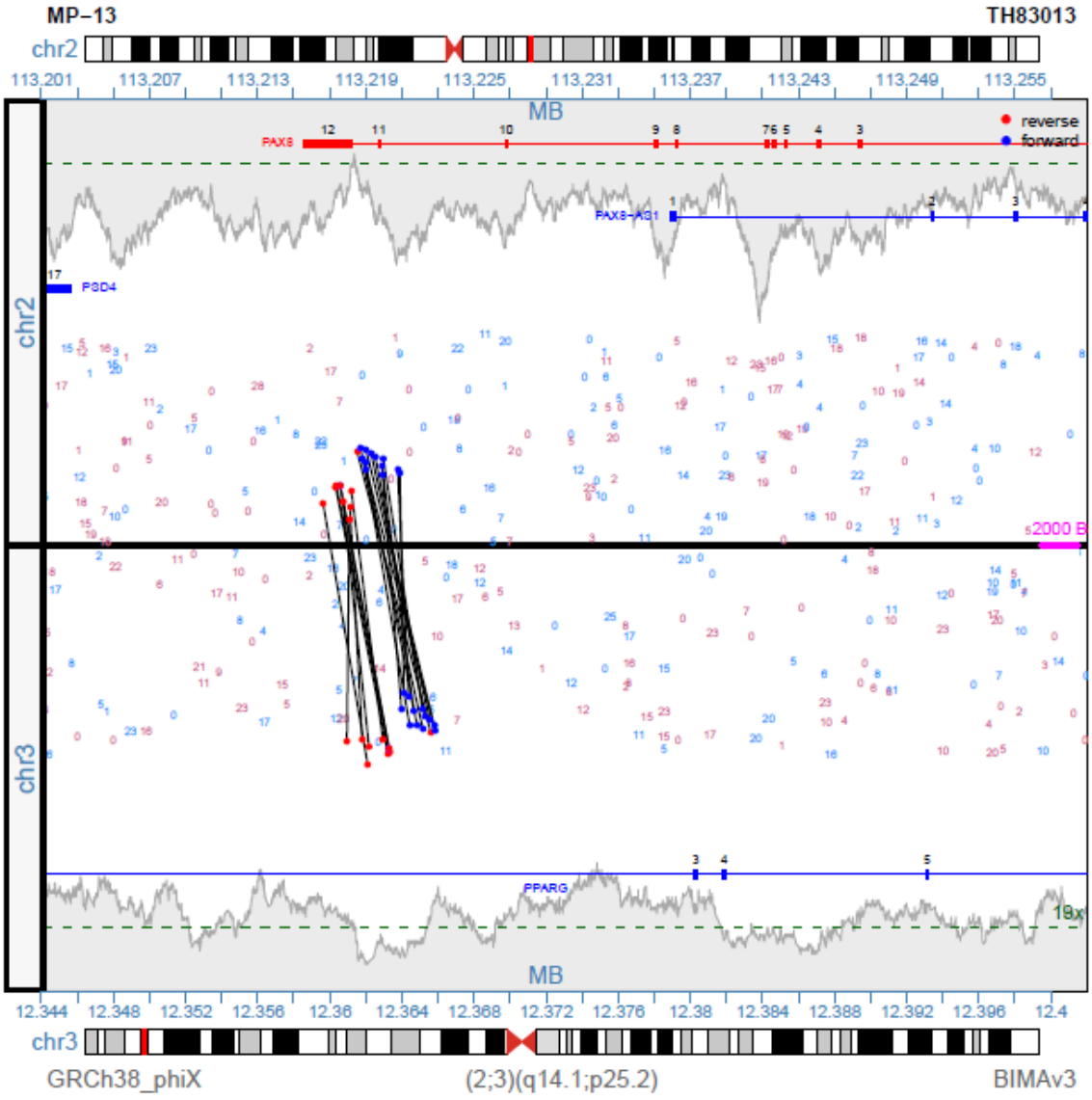

Figure B-9: Junction plot showing the second of three *PAX8/PPARg* translocations. This event results in a fusion protein.

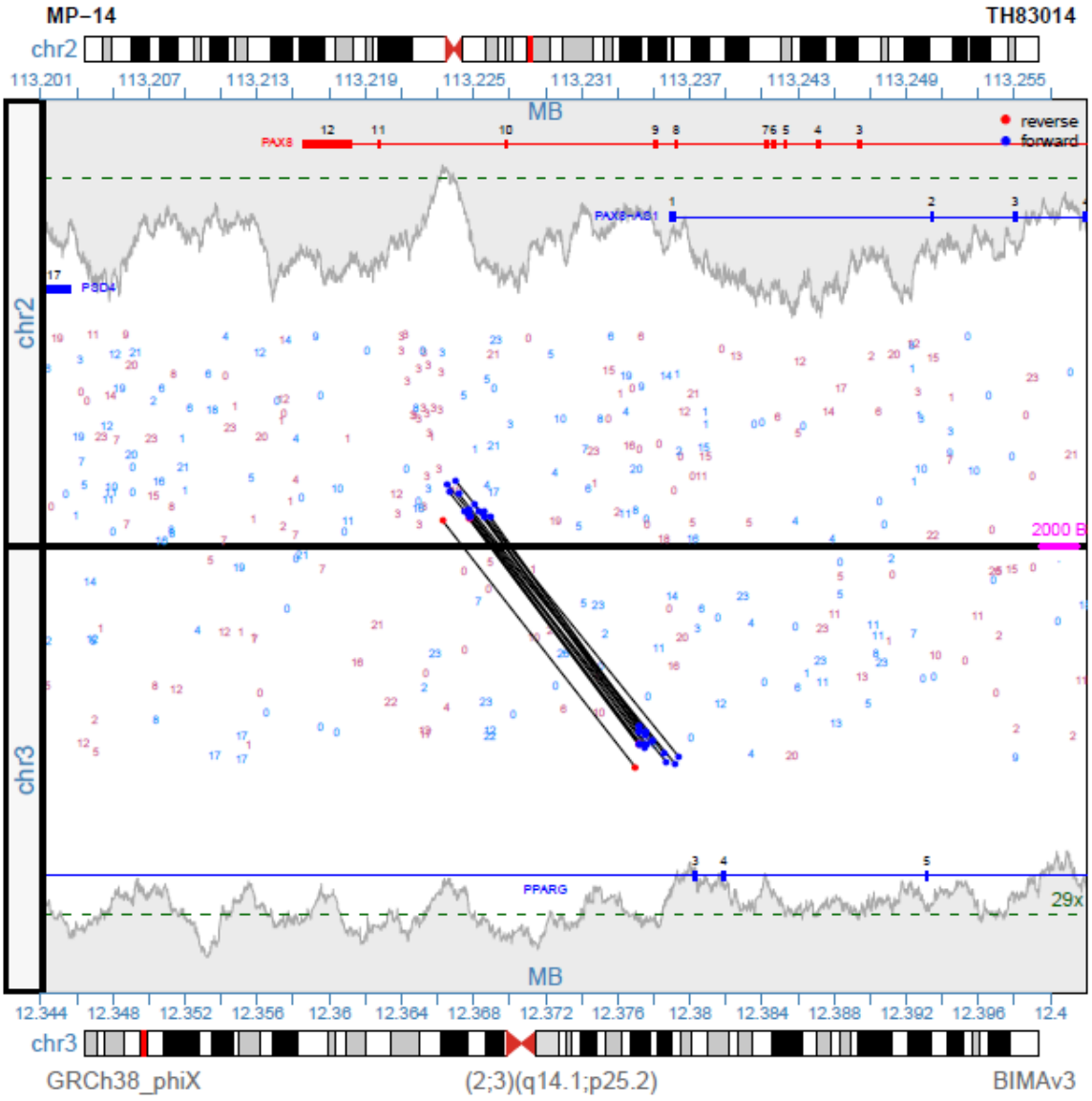

Figure B-10: Junction plot showing the third of three *PAX8/PPARg* translocations. This event results in a fusion protein.

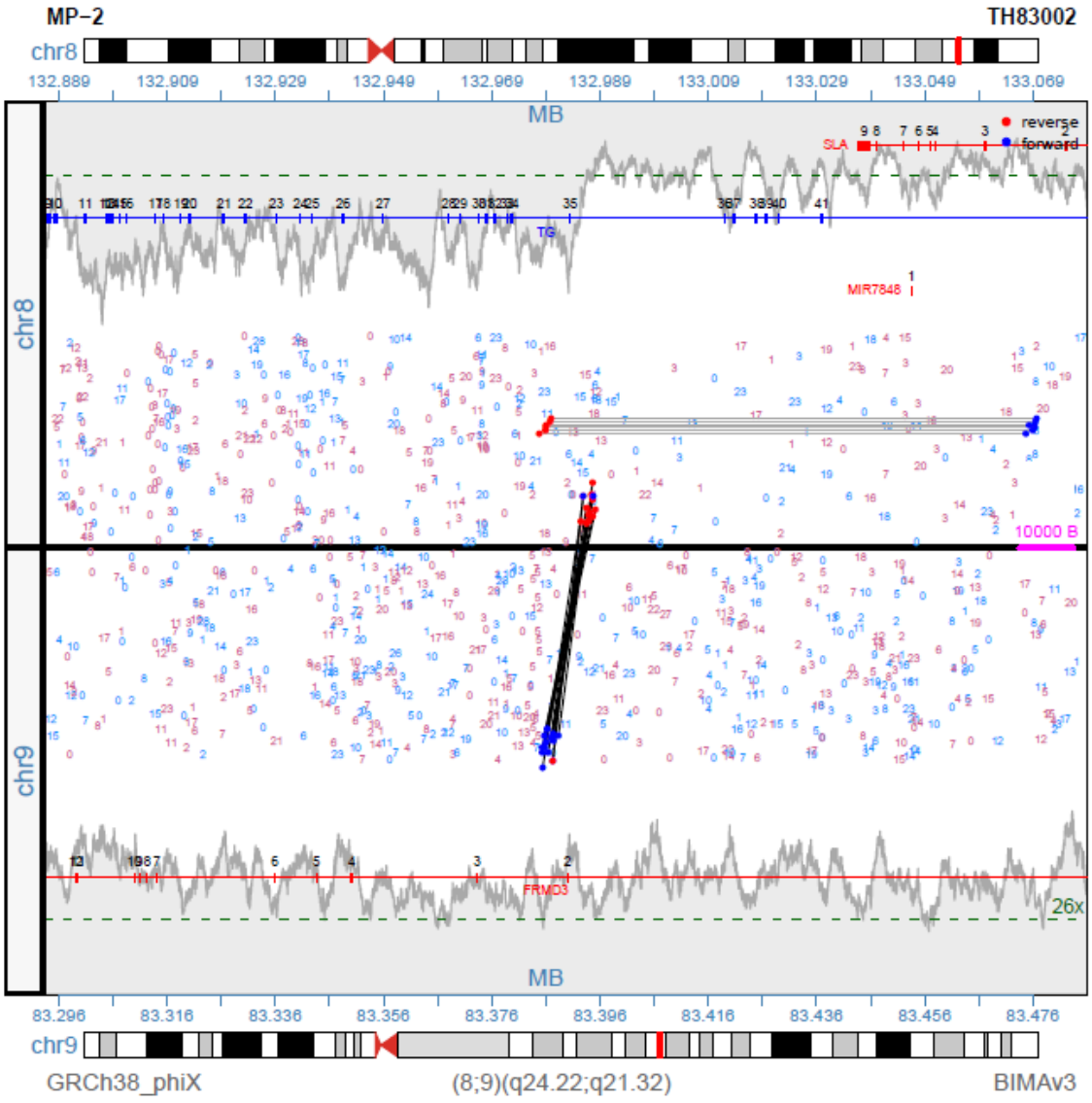

Figure B-11: Junction plot showing two rearrangements in the TG gene. One event produces a truncated protein while the other is a deletion of exons 35 – 41.

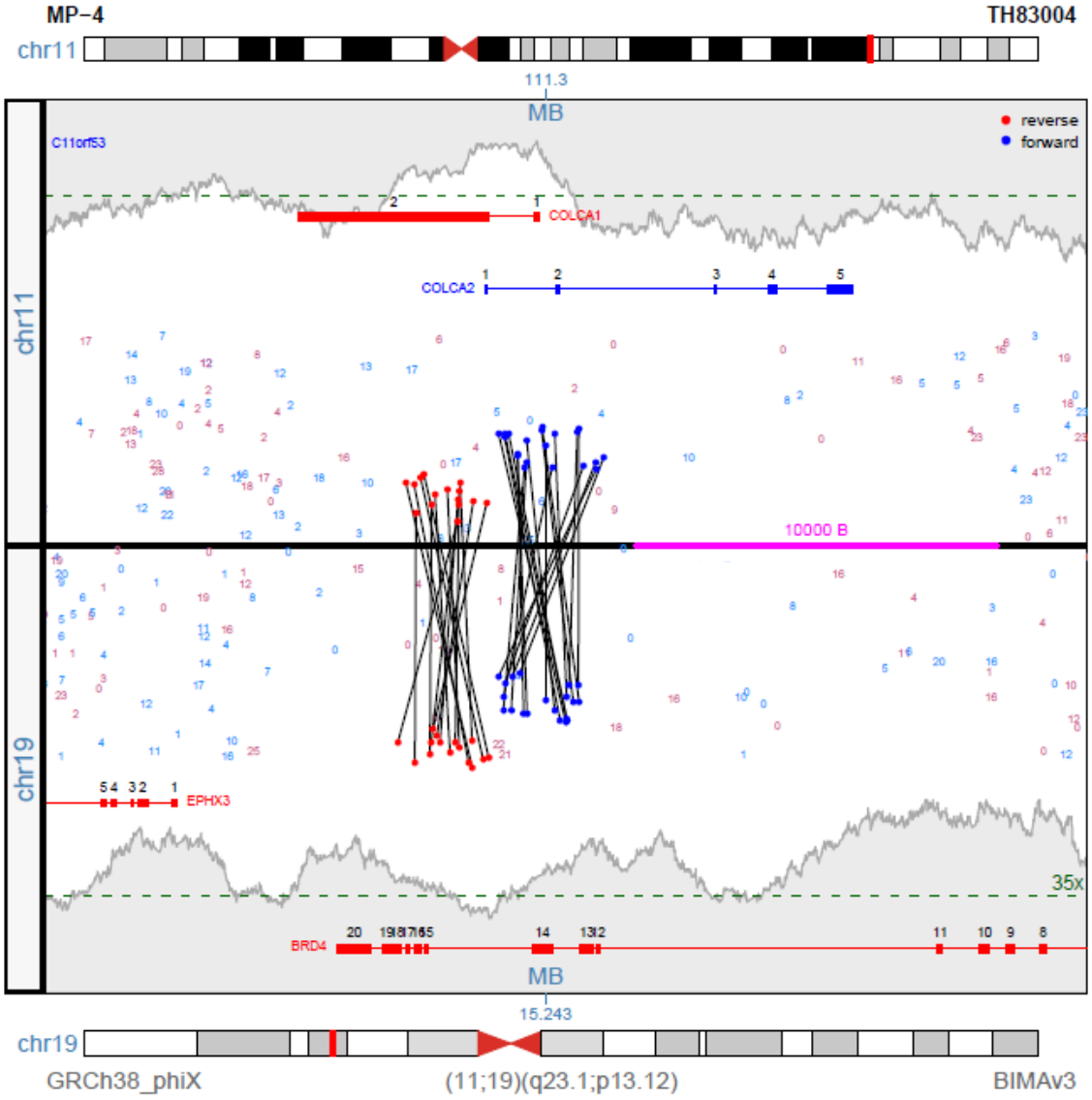

Figure B-12: Junction plot showing the balanced translocations between BRD4 and COLCA2.

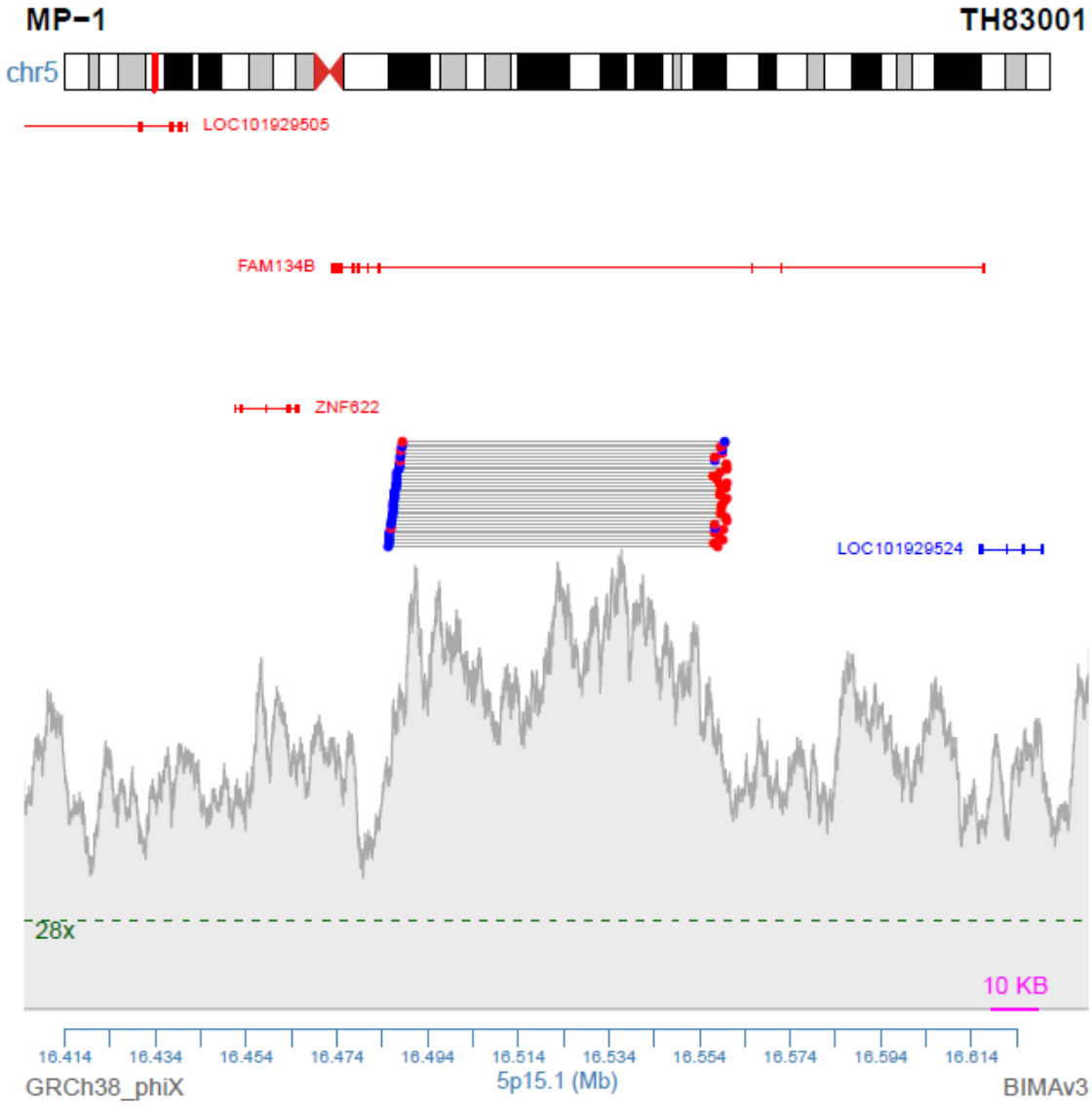

Figure B-13: Region plot showing the tandem insertion of a large intron in the FAM134 gene.

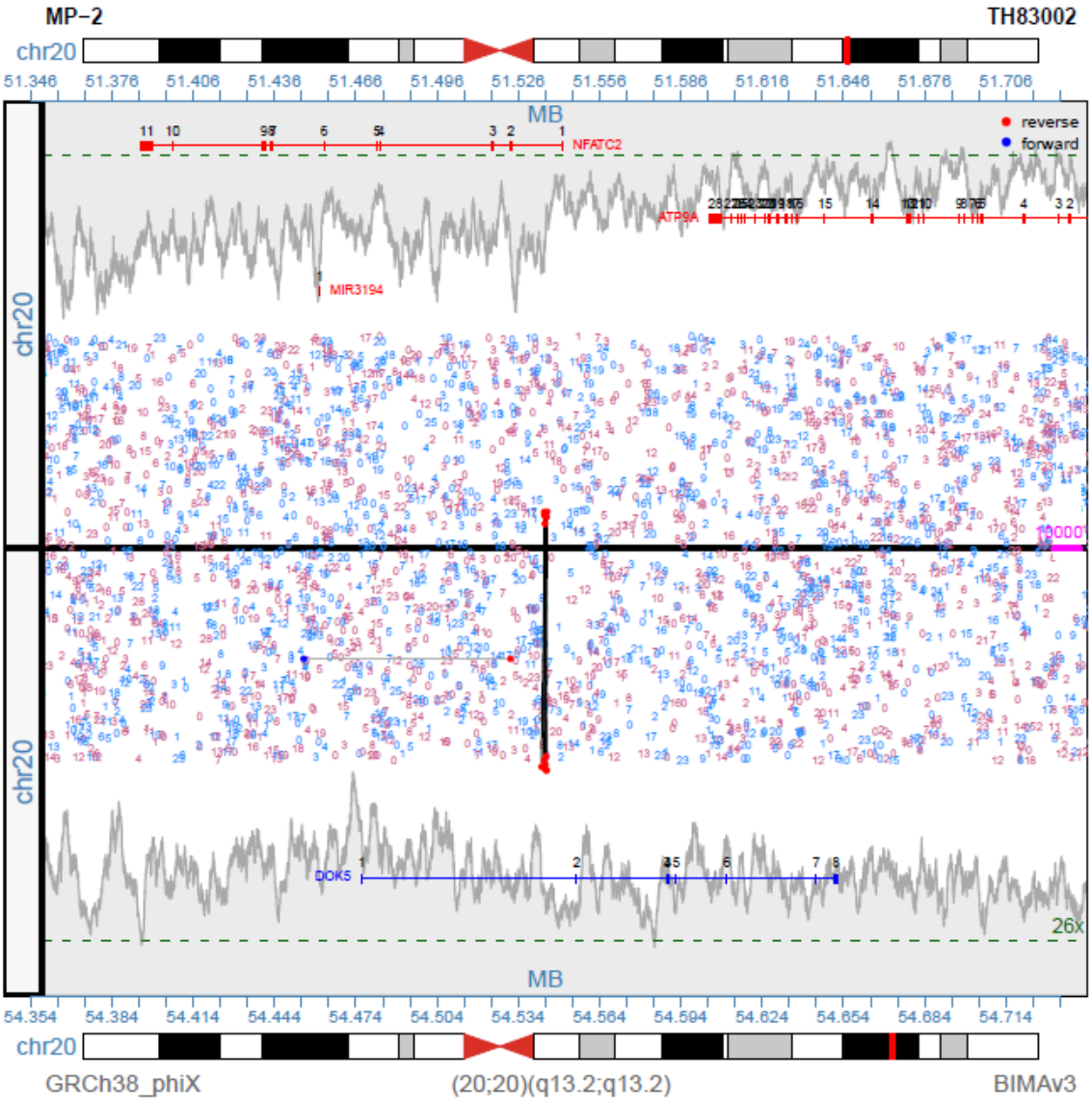

Figure B-14: Junction plot showing a translocation resulting in a fusion protein between DOK5 and NFATC2.

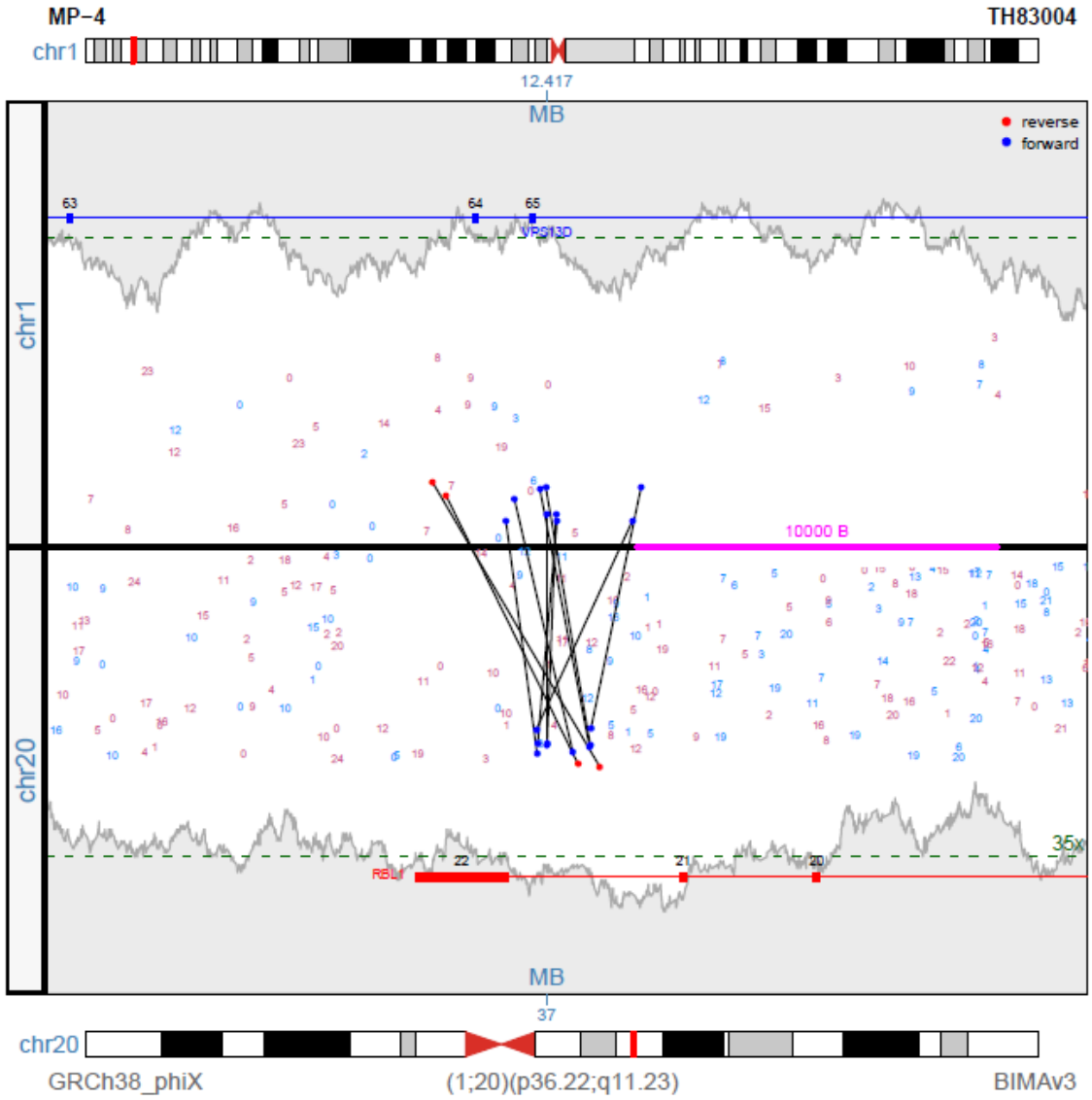

Figure B-15: Junction plot showing reads mapped to a pair of transposons resulting in a false positive event.

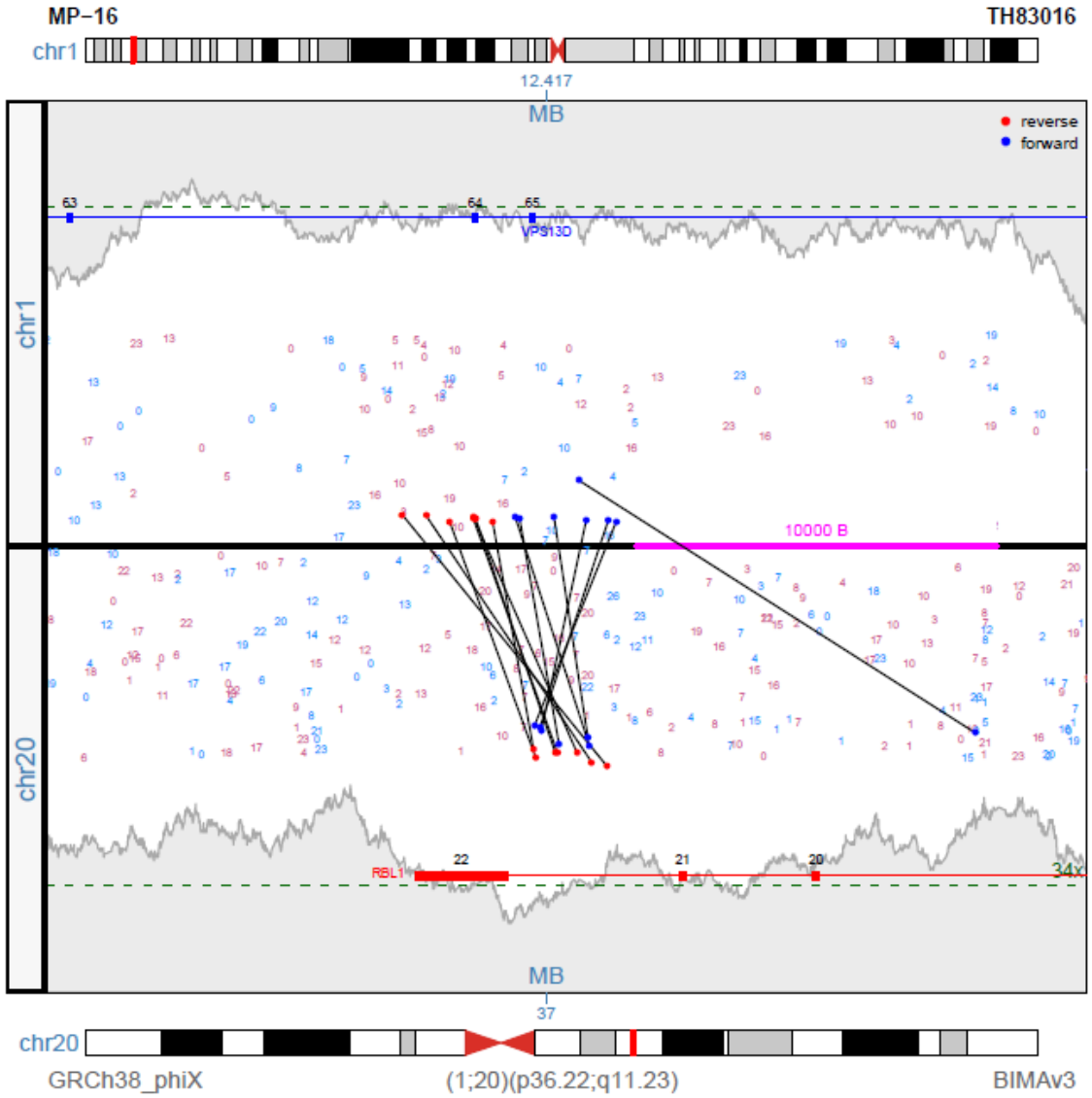

Figure B-16: The second of two junction plots showing reads mapped to a pair of transposons resulting in a false positive event.