Seqeyes: A multi-scale interactive visualization tool for structural variations

Richard Park, Nils Gehlenborg, Peter J Park

Harvard Medical School and Boston University

ABSTRACT

Genomic structural variations are known to play an important role in cancer and other diseases. Next-generation sequencing is a key technology for the identification of such variations, but current data and algorithms yield many false positives and false negative predictions. We have created a prototype tool called Seqeyes to explore and interpret predicted structural variations to help guide experimental validations. Users can sort, filter, and aggregate samples based on clinical attributes, which facilitates the association of phenotypes with specific patterns of structural variation. Our tool provides both linear and circular representations. Two genome browser views show detail at multiple locations concurrently, while the circular ideogram view provides a global summary. Multiple molecular data types including copy number, gene expression, and methylation microarrays for each sample are integrated as additional genomic tracks. We leverage advanced open-source indexes available from Postgresql and Postbio to greatly enhance the speed and amount of data available to visualize. Seqeyes is a novel multi-scale visualization that can interactively navigate dozens of genomes down to individual sequencing reads within a web browser.

KEYWORDS: Sequencing, structural variations, cancer, visualization, database, indexing

1 INTRODUCTION

With advances in next-generation sequencing (NGS) technology, rapid improvements in speed, quality, and cost are enabling identification of structural variations at an unprecedented pace. NGS, in particular paired end sequencing, provides opportunities to discover structural variants that could not be detected on conventional microarray-based platforms, such as dosage-invariant chromosomal translocations and inversions. Traditional array based methods have difficulty in predicting exact boundaries due to their low coverage of the genome even with the highest density arrays, coverage resolution is still limited to 10-20 kb [1]. Structural variations (SV) are relevant for both in research and clinic. Algorithms make predictions that serve as initial hypotheses for experimental validation. But biologists want to see (A) supporting information, e.g. in from of reads in the genomic context (B) other supporting information, e.g. other data types, or (C) same predictions in other samples. We aim to explore these aspects dealing with 4 types of SV including deletions, amplifications, inversions, and chromosomal translocations by allowing users to view global SV activity and interactively allow users to zoom in detail for a specific SV in a single sample.

Traditional genome browsers such as the UCSC Genome Browser [2] were not designed with paired-end sequencing in mind. Common approaches to structural variation visualization include Circos [3] (and Gremlin [4]). Circos displays various types of genomic variations based on a circular ideogram layout, while most genome browsers display data in a linear view with multiple genomic tracks. Chromosomal translocations are typically drawn as arcs connecting between corresponding genomic regions. We aim to provide a combination of circular and linear views integrated with data types from multiple platforms of experimental data and improve overall performance by leveraging recent open source indexes available from Postgresql including multi-column B+, R+ trees, general index search trees (GIST).

2 PRELIMINARY RESULTS

The Seqeyes prototype is an Adobe Flash web based visualization connected to a Postgresql 9.0 database using open source genomic extensions provided by PostBio [5]. Database tables are heavily preindexed using a combination of B+, R+, GIST, and suffix trees. The use of advanced indexes has improved query speed and performance up to a 1000 fold in certain cases with up to 500 million reads in a single table. This preindexing significantly improves fast random retrieval of genomic alignments and reduces the memory overhead required by the visualization.

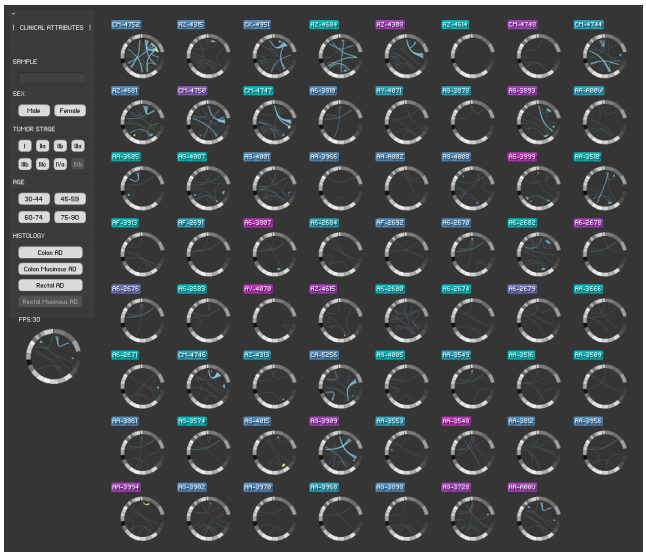

Figure 1. Global SV activity of 62 TCGA Colorectal tumors.

Presented are 62 colorectal cancer samples from The Cancer Genome Atlas (TCGA) with associated low pass DNA sequencing, methylation, copy number variation, and gene expression. Clinical variables including sex, age, cancer subtype, and stage are included as options to aggregate and filter at a multi-sample view. Seqeyes is a combination of linear and circular representations. Multiple samples are presented as individual circles with SV predictions denoted as arcs between genomic regions and colored based on its class (Figure 1). A menu is provided to allow the user to select a subset of the samples to

facilitate the association of phenotypes with specific patterns of structural variation. If greater detailed is required, each sample can be clicked and zoomed into an individual genome view. The single sample view allows greater inspection of the number and types of structural variation present for that individual. Multiple linear genome browsers are available depending if a single region or SV is clicked. When a specific SV is focused upon, clicking on the arc opens up linear genome views at each of the associated genomic locations. Within these linear views (Figure 2), integrated data from other experimental platforms are provided as additional tracks. If the specified SV is supported by discordant mate pairs, methylation, copy number variation, and/or gene expression, the integration of data gives confidence on the biologically likelihood of the predictions and data.

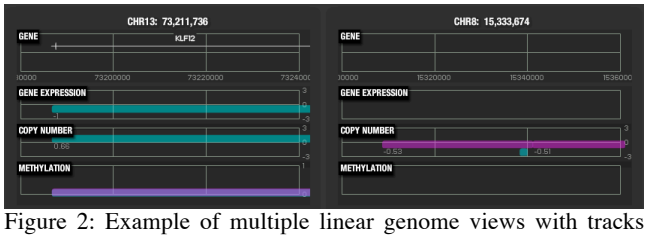

Figure 2: Example of multiple linear genome views with tracks corresponding to genes expression, copy number variation, and methylation for each predicted structural variation.

3 METHODS

The Cancer Genome Atlas (TCGA) is a comprehensive effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. Twenty cancer types are being studied, the goal to collect 500 tumors per cancer type, with multidimensional data sets being created for each: DNA sequencing (WGS and WES), mRNA expression and miR expression, methylation, CNV (SNP arrays), and clinical attributes.

3.1 Data Types

Paired-end DNA sequencing data were obtained for 62 primary tumor samples from patients diagnosed with colorectal carcinoma. The samples were processed through an Illumina HiSeq at coverage of 3-5X with an insert size of 200 bp – 300 bp. Array based data for methylation, copy number variation, and gene expression were also obtained for each sample. Methylation data for each sample were produced on the Illumina Infinium HumanMethylation27 BeadChip array. Copy number changes were identified using Affymetrix Genome Wide SNP Array 6.0, while mRNA gene expression measurements were provided by the Agilent G4502A platform.

3.2 Software

NGS raw reads were mapped to the human genome using BWA [7]. Copy number variations were calculated using BIC-seq [8] and structural variation predictions were produced by CAPS [9] algorithm specifically calling deletions, insertions, inversions, and intra and interchromosomal translocations. Postgresql 9.0 was used as the database backend and Adobe Flash was used to create the browser-based front end.

3.3 Indexes

Postgresql 9.0 database and Postbio genomic extensions were used as the backend database for the application. PostBio enables extremely fast genomic range queries and use of complex indexes including multi-column B-tree, R-tree, and generalized index search trees (GIST) were created on the tables stored in the database. Other indexes include stree, which is a serialized version of suffix tree implementation in MUMer [10]. As well as fmindex, a compressed suffix array based on the Burrow-Wheeler transformation allowing for fast exact matching of short sequences.

4 DISCUSSION

Visualization is used in this context to build evidence for or against hypotheses. Seqeyes provides an overview of all SV activity for dozens of cancer samples and allows the samples to be filtered and ordered based on relevant clinical attributes. We provide details on demand including multiple data types and supporting discordant mate pairs for each SV prediction. Typical visual representations of the genome include circular and linear views with arcs and blocks denoting genomic variations [6]. Circular views are useful for global summaries, while linear views are better suited to more detailed genomic windows. Most tools choose either a circular or linear view of the genome. We believe that incorporating both views at various scales and contexts improves the overall ability of the user to interpret the data. Seqeyes strengths include the ability to deal with multiple samples, providing multiple genome browser views, integration of experimental platforms, and its ability to sort, filter, and aggregate samples based on a clinical phenotype.

5 CONCLUSION

Seqeyes is a novel interactive multi-scale visualization, which provides the ability to view dozens of genomes down to individual sequencing reads and genomic tracks. The tool provides a combination of linked linear and circular genomic views leveraging the advantages of each representation and allows users to filter and group samples based on clinical attributes. Multiple genome browsers and integration of multiple experimental types help biologists evaluate predictions of SV algorithms to improve successful experimental validation.

REFERENCES

[1] Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 2007;39:S16-21.

[2] D. Karolchik, G. Bejerano, A. S. Hinrichs, R. M. Kuhn, W. Miller, K. R. Rosenbloom, A. S. Zweig, D. Haussler, and W. J. Kent. Comparative genomic analysis using the UCSC genome browser. 395, November 2007.

[3] Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. *Genome Res* (2009) 19:1639-1645

[4] O'Brien et al. Gremlin: an interactive visualization model for analyzing genomic rearrangements. *IEEE Trans Vis Comput Graph*. 2010 Nov-Dec;16(6):918-26.

[5] Carvalho L. Postbio: Bioinformatic extensions for Postgresql. http://postbio.projects.postgresql.org/

[6] Meyer M, et al MizBee: a multiscale synteny browser. *IEEE Trans. Vis. Comput. Graph. 2009*;15:897-904

[7] Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

[8] Xi R et al, BIC-seq (currently in submission)

[9] Xi R et al. CAPS (unpublished)

[10] Delcher A, Kasif S, Fleischmann R, Salzberg SL. Alignment of Whole Genomes *Nucleic Acids Research*, 27:11 (1999), 2369-2376.