EuroVis Workshop on Visual Analytics (2013)
M. Pohl and H. Schumann (Editors)

Visual Analysis of Tracts of Homozygosity in Human Genome

Sean Reber[1] and Ye Zhao[1] and Li Zhang[2] and Mohammed Orloff[2] and Charis Eng[2]

[1]Kent State University, USA [2]Cleveland Clinic Foundation, USA

Abstract

We propose a new visual analytics system designed for genetic researchers to study genome-wide homozygosity regions. Finding significant tracts of homozygosity (TOH) using single nucleotide polymorphisms (SNPs) from a large-scale genome data set can contribute to the discovery of genetic factors related to human diseases. Our system helps users to visually examine TOH clusters computed from the underlying patient data, lending itself a convenient and powerful tool for knowledge discovery. We illustrate the usability and performance of the system with a clinical data set of human cancers.

1. Introduction

Genomics - the study of DNA and related molecules, their functions, and their impact on human health - is a growing biomedical science highly reliant on computational methods. The expanding application of experimental high-throughput and high-resolution techniques in genomics is creating enormous challenges for the analysis of very large and complex genomic data sets. In particular, knowledge discovery of genetic factors associated with diseases is very important for early diagnosis and prevention.

In human chromosomes, germline homozygosity, a type of genomic variation, is a critical factor associated with an increased risk of human cancers and other diseases [HBPC*91, RRC*03, BDS02, ALPE08, OZBE12]. Studying homozygosity locus based on one of the most common genetic variants, single nucleotide polymorphisms (SNPs), is important [SSHN09]. However, the scale and complexity of genome-wide association study (GWAS) impose harsh challenge on computation and analysis: the traditional single SNP-association study requires up to millions of tests. Recent studies show that tracts of homozygosity (TOH) regions covering a sequence of SNPs may have a significant role in the genetics of complex diseases [BSW*08, LLD*08]. However, the existing tools for TOH analysis , e.g., Golden Helix [gol] and PLINK [PNTB*07], only provide naive functions and limited usability for biomedical researchers. A fast, easy-to-use, and interactive visual analytics system can greatly help researchers by facilitating a pleasant reasoning process of genomic SNP data, to overcome the translational barriers between clinical data and human understanding.

In this paper, we develop a visual analytics toolkit, aimed at identifying genetic risk factors related to the TOH clusters extracted from SNP data. Our visual system supports genetic researchers with new functions including:

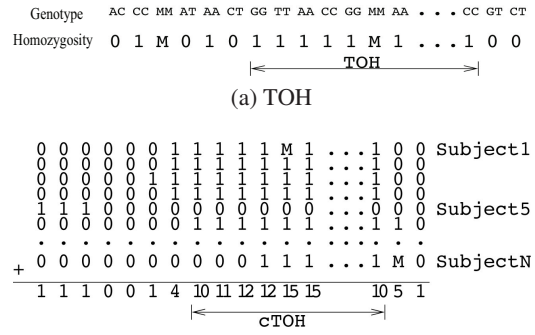
1. Defining similarity between TOHs and adapting a spectral clustering algorithm to discover TOH clusters with flexible user control of clustering parameters;
2. Proposing a new TOH cluster (TOHC) tree to hierarchically represent the clusters within common TOH regions;
3. Developing a visualization interface which supports examination of genome-wide TOH regions with (1) a visual cluster explorer, (2) navigation rings representing TOHC trees, and (3) an embedded NCBI Genome Map [map].
4. Incorporating statistical association study within the visualization system to investigate relationships between genotype and phenotype information;

This system is implemented with optimized performance and an easy-to-use graphics interface. It can be widely used by genetics scientists on studying SNP data in the identification of genomic regions associated with diseases. Our tool has been used by domain experts as a part in a genetics study whose biological background, methodology, and statistical analysis are published in [ZOR*13]. In this paper, we describe the details of the visualization approaches.

2. Related Work

Biomedical data visualization has been widely studied to assist biologists and practitioners in understanding data and conveying information. It plays a significant role in many biological processes promoting knowledge discovery [OGG*10]. Genetics scientists and researchers thirst for effective and efficient visualization tools in understanding and analyzing big sequencing data. There are three major visual

DOI: 10.2312/PE.EuroVAST.EuroVA13.037-041



(a) TOH

(b) cTOH Region

Figure 1: *Illustration of TOH and cTOH.*

analytics tasks [NCD*10]: (1) presenting sequence data in the context of assembly and resequencing experiments; (2) browsing annotations and experimental data; (3) comparing sequences from different organisms or individuals. We refer the readers to the table of tools in [NCD*10] for the references of these methods.

Many computational tools have been developed for SNP and other types of genetic variation discovery [DB10, MSB09]. Together with the downstream data mining analysis programs, a genome browser can promote visual analysis for the biological validity. Visualization provides a natural and perceptual way to interpret and manipulate the data. A number of such browsers have been developed for visualizing genomic annotations and many other related information for various biological datasets, such as the successful UCSC [KSF*02] and Ensembl [PCB04] online browsers. One barrier that may curb researchers from intensive use of such general tools is: the visualization processes are disconnected from pertinent, computer-intensive analyses from different application requirements [NCD*10]. Identifying and analyzing TOHs is one of such particular applications. It demands special design and implementation of a visual analytics system that has not been previously addressed. In this paper we present such a system integrating computational methods (clustering), visualization techniques (cluster explorer and navigation rings), and statistical analysis (association study) into one toolkit. Moreover, our system also links to a gene browser to incorporate related information and promote user understanding.

3. Tracts of Homozygosity and Genetic Risks

SNPs are the most common genetic variation among people [Lis12]. SNPs occur normally throughout a person's DNA. On average one SNP appears every 300 nucleotides, which amounts to around 10 million SNPs in the human genome [Lis12]. SNPs are biomedical markers that may or may not be associated with genes related to disease risks. Finding the significant locus over human genomes linking to particular diseases has been a critical task for researchers. Given a group of human subjects, SNP genotyping is performed by SNP array, where hundreds of thousands of SNP probes are arrayed and interrogated simultaneously on a gene chip. Homozygosity is a genetic condition where the probe shows the same alleles for a particular SNP locus. It is then possible to identify specific homozygous alleles associated with clinical phenotypes through statistical association study. In this paper, we study the TOH, which was defined in [OZBE12] as a chromosomal segment that must meet the criteria of having at least L consecutive homozygous SNPs or a genetic distance of at least M kb on a single chromosome for a given subject. Here L and M are user defined parameters. Figure 1(a) illustrates a TOH residing in one chromosome of a probed subject. Some genotypes have the same nucleotide (TT, AA, CC, GG), i.e., homozygosity. The bottom row encodes homozygosity to 1 and heterozygosity to 0 (in this figure M represents experimentally unidentified values treated as homozygosity). A continuous tract of 1's constructs one TOH. The length of the rows of a whole chromosome is very long, e.g., human chromosome 1 has about 740,000 SNPs. Figure 1(b) shows an example TOH data set of N subjects (e.g., diseased and control cases attending a study). Lencz et al. [LLD*08] proposed a common TOH (cTOH) region for further investigation, which is defined as a window of at least k consecutive SNPs. Inside this region, the number of SNPs belonging to a TOH is at least n. Figure 1(b) shows one cTOH region where $n \geq 10$.

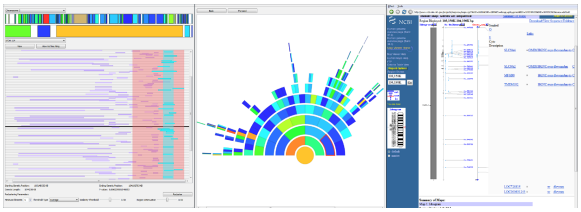
4. TOH Clustering and Exploration

Our system takes new measures to enhance the analysis of TOH data sets. It first groups a large set of TOHs to find their clusters, and then uses a TOHC tree structure to manage the identified clusters. This data structure provides a good medium for users to examine clusters with statistical association study and perform visual analysis.

TOH Similarity and Clustering The cTOH method achieves its success as shown in [LLD*08], but the simple counting algorithm (Figure 1(b)) is not very effective. The TOHs within a cTOH region may not all overlap or may have distant boundaries. Therefore, TOH analysis can further be advanced by the identification of patterned clusters of TOHs. Clustering TOHs over the whole chromosome has a heavy load and is not necessary due to the nature of finding frequent horizontal overlaps among TOHs. We adopt a two-stage approach. First, cTOH regions are discovered in each chromosome across all subjects. Second, TOH clusters are generated from the TOHs within a cTOH region, where the similarity between two TOHs (e.g., T_1, T_2) is defined as

$$similarity(T_1,T_2) = min(\frac{len(OP(T_1,T_2))}{len(T_1)}, \frac{len(OP(T_1,T_2))}{len(T_2)}), \quad (1)$$

where $OP(T_1,T_2)$ is the overlap region between T_1 and T_2. $len()$ is the length of TOHs in terms of the number of SNPs. We adapt a normalized spectral clustering method [HKK07, NJW01] in creating clusters. The clustering algo-



(1) (2) (3)

Figure 2: *System interface overview.*

rithm relies on predefined parameters, such as the number of clusters k. In TOH study, users do not have such priori knowledge. They direly need a tool so that they can flexibly group TOHs, study the groups, and change parameters in their analysis. Our visual analytics system provides such guidance and control for users.

TOHC Tree We use hierarchical clustering and propose a new tree-based data structure, namely TOH cluster tree (TOHC tree), to accommodate TOH clusters and advocate user exploration. Within one cTOH region, a binary spectral clustering (i.e., setting the number of clusters $k = 2$) creates two clusters. The two clusters are further clustered by another binary clustering, respectively. Repeating this process, a TOHC tree is generated where each tree node represents one TOH cluster. Each cluster of the tree is potentially important for further analysis, not only the leaf node. So the TOHC tree preserves and manages all the clusters, and provides users a tool to navigate over them. The TOHC tree generation is controlled via three stopping criteria: (1) when the maximum depth of the tree is reached, which relates to the computational performance; (2) when the number of TOHs in a cluster is less than a given minimum, which determines the necessity of further clustering; (3) when statistical attributes are smaller than given thresholds.

Cluster Region Each TOH cluster node of the TOHC tree defines its starting and ending positions over a chromosome. A straightforward method is to use the leftmost and rightmost positions of all the belonging TOHs. Users can flexibly attenuate such a region: TOHs in a cluster are sorted by their start points in ascending order. For each point, we compute the number of those TOHs covering its position. If this number is more than a percentage (set as a user-controlled threshold θ) of all the TOHs, this point is kept. Otherwise, the point is removed. Then, the farthest left point remaining in the list is used as the starting position of this cluster region. The ending position is determined in a similar fashion.

Statistical Association Study Each TOH cluster presents a region which will be tested by statistical association study for its relationship with disease risks. By considering each such region as a genomic variant, a genome-wide case-control analysis was conducted. P-values are obtained to present the significance of the association. Such study is implemented through an embedded statistical computing component in our system.

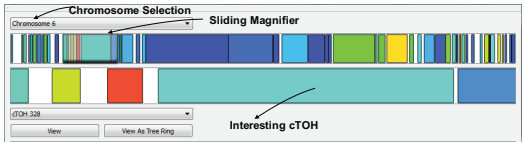


Figure 3: *Exploring and selecting chromosomes and cTOHs.*

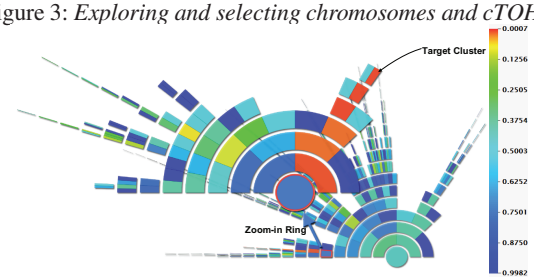


Figure 4: *Navigating TOHC tree.*

5. Visual Analytics System

Given a genome-wide SNP genotyping data set, our system performs preprocessing computation: (1) homozygosity detection, (2) TOH identification, (3) cTOH region generation, (4) clustering that creates one TOHC tree within each cTOH region, (5) statistical testing that produces P-values for each cluster. Then the corresponding results are loaded into the visualization system for user exploration. The first four stages are implemented with C++ and the fifth stage of statistical computing is implemented by the scripting language R [Gen08]. For a data set including thousands of subjects, the data processing runs in about ten minutes on a quad-core consumer PC. Figure 2 shows the visual interface supporting interactive exploration. It consists of three views:

Chromosome and TOH cluster explorer (Figure 2(1)): all cTOH regions inside a user-selected human chromosome are shown as colored bars on the top panel. The colors (red to blue) are mapped from the P-values (low to high) of the cTOH region. Users can drag a sliding magnifier to fully review all the cTOH regions. Below the panel, a close-up view displays the magnified TOHs for users to clearly observe them.

Navigation rings of clusters (Figure 2(2)): As an interactive visualization tool of the TOHC tree, the root node corresponds to the given cTOH region, and outward rings are the offspring of the inner rings. This approach adapts the sunburst visualization [SZ00] by using only the upper half circle, which is easy to read and gains user satisfaction in analysis. Each cluster is represented by a wedge-shaped bar whose color manifests its P-value, so that users can identify interesting clusters easily. Hovering over a cluster will highlight the bar and all its children, and clicking it will trigger a zoom-in view.

Genome map of a chromosome region (Figure 2(3)): Genome region related to a TOH cluster involves a large set of attributes of existing biomedical and genetic knowledge such as gene symbols, gene names, protein, regulation, etc. We integrate a genome browser, the widely used NIH NCBI

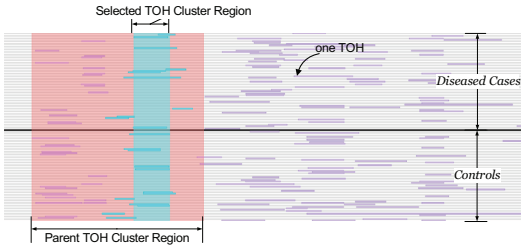


Figure 5: *Exploring one selected target cluster.*

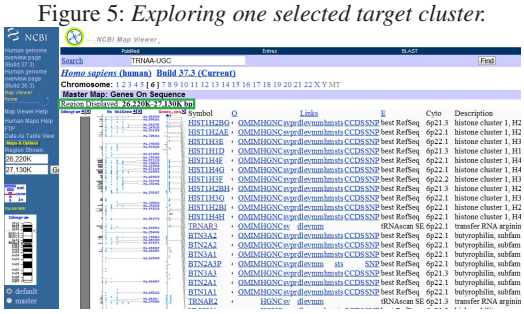


Figure 6: *Genome map of a target cluster region.*

Map Viewer, to plot genome maps. When users click on an interesting TOH cluster, the genome map will display associated cytogenetic, genetic, physical, and other information.

6. Visual Exploration of Lung Cancer Genome Data

Clinical Data: We utilize a data set of lung cancer project in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial [PAB*08], designed and sponsored by the National Cancer Institute (NCI). It includes DNA sequencing samples of 1618 subjects which have 788 lung cancer cases and 830 controls of European Americans. After genotyping, 514,355 autosomal SNPs were available for subsequent TOH analyses. The data set is processed following the five stages in Section 5 for the preparation of visual analysis. TOHC trees for the cTOH regions are obtained using the following criteria: (1) 5 minimum components in a cluster; (2) a maximum tree depth of 100, and (3) the average of the lower quartile of similarity values smaller than 0.75 . The cluster region attenuation threshold is set as $\theta = 0.50$.

Exploring Chromosome cTOHs: Figure 3 depicts Chromosome 6 selected for investigation. The whole chromosome is shown as cTOH segments with colors representing the P-values. Using the sliding magnifier, a cTOH region 328 (shown in teal) attracts the user's attention and is selected for further investigation.

Navigating Clustered TOH Groups: The corresponding navigation rings are shown in Figure 4. The user finds some cluster nodes in the outer layers of the ring (i.e., deeper levels of the TOHC tree). The bright red colors indicate small P-values of statistical importance (the smaller a P-value is, the more significant it is). By clicking on a parent node of those potentially important clusters, the user highlights the region and a zoom-in ring is created for clearer details. One target TOH cluster at a leaf is selected. This cluster refers to

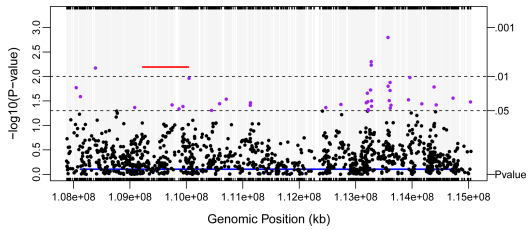


Figure 7: *Plot of P-values of the target region.*

a special region over this chromosome. Users can browse it in the cluster explorer by scrolling up and down. A snapshot of the TOH view shows part of this cluster in Figure 5.

Linking Genome Maps: To further study the region of this target cluster, the NCBI Map Viewer is loaded automatically by setting the physical starting and ending position of this region (26,220k to 27,130k basepair (bp) in Human Chromosome 6), shown in Figure 6 as the green highlighted box. The regional genome map is shown, which allows the user to further examine this region with associated information, e.g., many known genes on this region are presented for further investigation.

Analysis After statistical association study, Figure 7 plots the P-value of this target region, as well as P-values for individual SNPs inside the cTOH region containing this target region. Single-SNP association analysis was performed independently of TOH analysis and compared. X-axis is the genomic position, and Y-axis represents $-log10$(P-value) computed from P-values. With this function, small P-values (more significant) are shown higher than large P-values (less significant). Black dots represent SNPs at P-value ≥ 0.05 based on single SNP analysis, and purple dots represent SNPs at P-value< 0.05 based on single SNP analysis. The blue line depicts the P-value computed for the whole cTOH region. The red line shows the P-value for the selected target TOH cluster region. This figure illustrates that the target TOH cluster region is more significant compared with the whole cTOH region, as well as most single SNPs. The domain researchers among our authors performed intensive study over the genome region with many others, to gain insights of its link to human cancers. The detailed statistical analysis and discussion can be found at [ZOR*13].

7. Conclusion

We have proposed a new visual analytics system to compute and study clusters of SNPs with extended homozygosity. It supports investigation of the characteristics of TOHs. This genome-wide visualization system, incorporating statistical measures, allowing intuitive and interactive exploration for critical analysis.

Acknowledgements

This work is partially supported by Ohio Board of Regents and NSF IIS-0916131.

References

[ALPE08] ASSIE G., LAFRAMBOISE T., PLATZER P., ENG C.: High frequency of germline genomic homozygosity associated with cancer cases. *JAMA 299*, 12 (2008), 1437–1445. 1

[BDS02] BHATTACHARYA P., DUTTAGUPTA C., SENGUPTA S.: Proline homozygosity in codon 72 of p53: a risk genotype for human papillomavirus related cervical cancer in indian women. *Cancer Lett 188*, 1-2 (2002), 207–11. 1

[BSW*08] BACOLOD M., SCHEMMANN G., WANG S., SHATTOCK R., GIARDINA S., ET AL.: The signatures of autozygosity among patients with colorectal cancer. *Cancer Res 68* (2008), 2610–2621. 1

[DB10] DALCA A. V., BRUDNO M.: Genome variation discovery with high-throughput sequencing data. *Briefings in Bioinformatics 11*, 1 (Jan. 2010), 3–14. 2

[Gen08] GENTLEMAN R.: *R Programming for Bioinformatics*. Computer Science & Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, 2008. 3

[gol] http://www.goldenhelix.com. Golden Helix, Inc. 1

[HBPC*91] HENRY I., BONAITI-PELLIE C., CHEHENSSE V., BELDJORD C., SCHWARTZ C., UTERMANN G., JUNIEN C.: Uniparental paternal disomy in a genetic cancer-predisposing syndrome. *Nature 351*, 6328 (1991), 665–667. 1

[HKK07] HIGHAM D. J., KALNA G., KIBBLE M.: Spectral clustering and its use in bioinformatics. *J. Comput. Appl. Math. 204*, 1 (July 2007), 25–37. 2

[KSF*02] KENT W., SUGNET C., FUREY T., ROSKIN K., PRINGLE T., ZAHLER A., HAUSSLER D.: The human genome browser at ucsc. *Genome Res. 12*, 6 (2002), 996–1006. 2

[Lis12] LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS, U.S. NATIONAL LIBRARY OF MEDICINE, NIH: *Genetics Home Reference, Handbook, Help Me Understand Genetics*. http://ghr.nlm.nih.gov/handbook, 2012. 2

[LLD*08] LENCZ T., LAMBERT C., DEROSSE P., BURDICK K. E., MORGAN V. T., KANE J. M., ET AL.: Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America 104* (2008), 19942–19947. 1, 2

[map] http://www.ncbi.nlm.nih.gov/projects/mapview/. National Center for Biotechnology Information, US National Library of Medicine, National Institutes of Health. 1

[MSB09] MEDVEDEV P., STANCIU M., BRUDNO M.: Computational methods for discovering structural variation with next-generation sequencing. *Nature methods 6*, 11 Suppl (Nov. 2009), S13–S20. 2

[NCD*10] NIELSEN C. B., CANTOR M., DUBCHAK I., GORDON D., WANG T.: Visualizing genomes: techniques and challenges. *Nature methods 7*, 3 Suppl (Mar. 2010). 2

[NJW01] NG A., JORDAN M., WEISS Y.: On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (2001), Dietterich T., Becker S., Ghahramani Z., (Eds.), MIT Press, pp. 849–856. 2

[OGG*10] O'DONOGHUE S. I., GAVIN A.-C. C., GEHLENBORG N., GOODSELL D. S., HÉRICHÉ J.-K. K., NIELSEN C. B., NORTH C., OLSON A. J., PROCTER J. B., SHATTUCK D. W., WALTER T., WONG B.: Visualizing biological data-now and in the future. *Nature methods 7*, 3 Suppl (Mar. 2010), S2–S4. 1

[OZBE12] ORLOFF M. S., ZHANG L., BEBEK G., ENG C.: Integrative genomic analysis reveals extended germline homozygosity with lung cancer risk in the plco cohort. *PLoS One 7*, 2 (2012), e31975. 1, 2

[PAB*08] PROROK P. C., ANDRIOLE G. L., BRESALIER R. S., BUYS S. S., ET AL.: Design of the prostate, lung, colorectal and ovarian (plco) cancer screening trial. *Controlled clinical trials 21*, 6 (2008), 273S–309S. 4

[PCB04] PROCTOR G., CLAMP M., BIRNEY E.: The ensembl core software libraries. *Genome Research 14*, 5 (2004), 929ÍC933. 2

[PNTB*07] PURCELL S., NEALE B., TODD-BROWN K., THOMAS L., FERREIRA M. A., BENDER D., MALLER J., SKLAR P., DE BAKKER P. I., DALY M. J., SHAM P. C.: PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics 81*, 3 (Sept. 2007), 559–575. 1

[RRC*03] RUDAN I., RUDAN D., CAMPBELL H., CAROTHERS A., WRIGHT A., SMOLEJ-NARANCIC N., JANICIJEVIC B., JIN L., CHAKRABORTY R., DEKA R., RUDAN P.: Inbreeding and risk of late onset complex disease. *J Med Genet 40*, 12 (Dec. 2003), 925–932. 1

[SSHN09] SEELOW D., SCHUELKE M., HILDEBRANDT F., NÚRNBERG P.: Homozygositymapper–an interactive approach to homozygosity mapping. *Nucleic Acids Res. 37* (2009), W593–9. 1

[SZ00] STASKO J., ZHANG E.: Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the IEEE Symposium on Information Vizualization 2000* (2000), INFOVIS '00, p. 57. 3

[ZOR*13] ZHANG L., ORLOFF M. S., REBER S., LI S., ZHAO Y., ENG. C.: cgaTOH: Extended approach for identifying tracts of homozygosity,. *PLOS One 8*, 3 (2013), e57772. 1, 4