BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 3 2014, pages 445–447
doi:10.1093/bioinformatics/btt686

Data and text mining

Advance Access publication December 3, 2013

HomSI: a homozygous stretch identifier from next-generation sequencing data

Zeliha Görmez[1], Burcu Bakir-Gungor[1,2,*] and Mahmut Şamil Sağıroğlu[1]

[1]Advanced Genomics and Bioinformatics Research Center, The Scientific and Technological Research Council of Turkey (TUBITAK-BILGEM), 41470 Gebze, Kocaeli, Turkey and [2]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Abdullah Gul University, 38039 Kayseri, Turkey

Associate Editor: Inanc Birol

ABSTRACT

Summary: In consanguineous families, as a result of inheriting the same genomic segments through both parents, the individuals have stretches of their genomes that are homozygous. This situation leads to the prevalence of recessive diseases among the members of these families. Homozygosity mapping is based on this observation, and in consanguineous families, several recessive disease genes have been discovered with the help of this technique. The researchers typically use single nucleotide polymorphism arrays to determine the homozygous regions and then search for the disease gene by sequencing the genes within this candidate disease loci. Recently, the advent of next-generation sequencing enables the concurrent identification of homozygous regions and the detection of mutations relevant for diagnosis, using data from a single sequencing experiment. In this respect, we have developed a novel tool that identifies homozygous regions using deep sequence data. Using *.vcf (variant call format) files as an input file, our program identifies the majority of homozygous regions found by microarray single nucleotide polymorphism genotype data.

Availability and implementation: HomSI software is freely available at www.igbam.bilgem.tubitak.gov.tr/softwares/HomSI, with an online manual.

Contact: bakirburcu@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 23, 2013; revised on October 28, 2013; accepted on November 19, 2013

1 INTRODUCTION

In communities with a high level of consanguineous marriage, the diagnosis of a recessive genetic disorder offers a unique advantage for positional cloning of rare diseases. In such an isolated inbred human population, several individuals may inherit a variation of an ancestor, and the offsprings born of consanguineous union will have a high probability of inheriting two copies of the mutated chromosomal segment and thus expressing the disease. In this respect, homozygosity mapping is an effective approach to identify potential disease loci. Because Lander and Botstein successfully used this technique in 1987 (Lander and Botstein, 1987), both the experimental and the computational methods that are used to generate and analyze relevant datasets have undergone diversification and refinement. Originally, to detect homozygous regions, individuals were genotyped with panels of highly polymorphic short tandem repeat (or microsatellite) markers, typically at genomic intervals of 10–12 cM. With the availability of single nucleotide polymorphism (SNP) microarrays, short tandem repeat microsatellites were later replaced by SNP arrays to survey the genome and to identify large stretches of homozygosity. The overlapping regions among the homozygous stretches of affected individuals are called runs of shared homozygosity (ROSHs), and these regions are expected to contain the disease gene. For this purpose, several computational tools have been developed (Amir el *et al.*, 2010; Carr *et al.*, 2006; Kayserili *et al.*, 2009; Papic *et al.*, 2011; Seelow *et al.*, 2009; Uz *et al.*, 2010; Zhang *et al.*, 2011).

Even though the identification of recessive disease loci (homozygous regions) is accomplished using the aforementioned SNP array-based techniques, the next step to detect pathogenic sequence variant is sequencing. Until recently, only the candidate genes in these loci are sequenced. If this disease locus is a large interval or if it includes several genes, the determination of disease causing mutations by Sanger sequencing becomes a back-breaking task. At this point, next-generation sequencing (NGS) platforms present an alternative solution with their capacity to sequence the entire genome. Owing to the moderate costs and tractable data amounts, exome sequencing is a promising approach to detect novel mutations of human monogenic disorders. Hence, it is now possible to concurrently identify homozygous regions and possibly deleterious sequence variants, using data from a single sequencing experiment. To the best of our knowledge, only two programs have been developed to detect homozygous regions from NGS data, i.e. AgileVariantMapper (Carr *et al.*, 2013) and HomozygosityMapper (Seelow and Schuelke, 2012). But, these programs do not take into account the distribution of the variants within the genomic coordinates. Here, we developed a novel sliding window-based methodology, which provides the advantage of scanning the genome in detail, but at the same time detecting the candidate homozygous regions easily. This type of analysis gives us the opportunity of converting the genotype information into a signal, where we can apply well-known signal processing techniques.

2 DESCRIPTION

HomSI was designed to define homozygous stretches in consanguineous families from NGS data. The overall analysis flow of HomSI is illustrated in Supplementary Figure S1.

*To whom correspondence should be addressed.

2.1 Processing NGS data

Variant call format (*.vcf) files are standard output files of variant identification programs that process NGS data. These files can be directly used as an input file to HomSI. Because the read depths for some regions might be low, we provide users different filtering options such as quality-based user-defined region or variant-based filtering. Details are available in the user manual.

To identify and visualize the homozygous stretches, HomSI processes each variant and generates several graphs. Here, we illustrate HomSI outputs using autosomal recessive Klippel Feil Syndrome dataset (Bayrakli *et al.*, 2013). In Figure 1(a–b), the genotype information is plotted for different individuals, where the *x*-axis shows the samples and the *y*-axis indicates genomic coordinates. The variants of an index case are colored in blue for homozygous and in yellow for heterozygous cases. For other individuals, although a homozygous variant is indicated with blue, the contrasting homozygous is represented with white and the heterozygous with orange. For all individuals, gray indicates no call and yellow indicates non-informative SNPs, which are heterozygous for index case (see Supplementary Table S1 for details). HomSI starts with an unbiased approach and generates homozygosity maps for whole genome (Fig. 1a). Once a continuous blue line, which is shared among affected individuals, but not in unaffected ones, is recognized as in Figure 1a, the user can zoom in this region on the chromosome-based representation (Fig. 1b).

To be able to identify the ROSHs, we normalized the number of homozygous and heterozygous variants within each window separately (named as Hom and Het signals) and plotted these signals for each affected (Fig. 1c) and unaffected (Fig. 1d) individual. Using the differences between the Hom and Het signals, which were plotted for each individual in Figure 1e, a prediction signal was generated in Figure 1f. This final signal indicates the

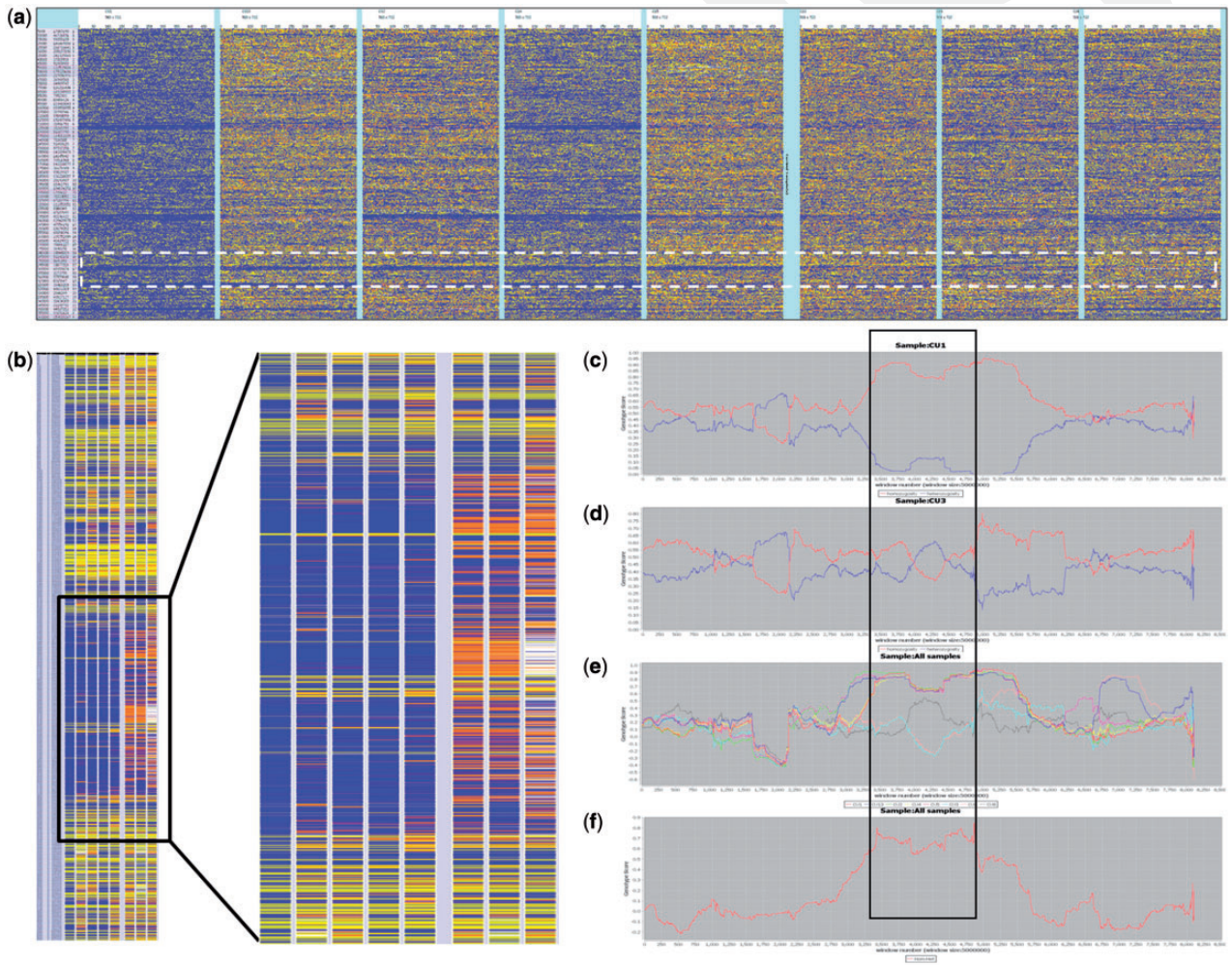

Fig. 1. A subset of the graphs produced by HomSI. (**a**) Graphical representation of genome-wide homozygosity map for affected (five columns in left) and unaffected (three columns in right) individuals. Dashed box indicates ROSH among affected individuals. (**b**) Homozygosity map of chromosome 17. Shared homozygous region by the affected individuals is surrounded with a black rectangle. First column shows index case, second to fifth columns show the other four affected individuals and sixth to eight columns show the control individuals. (**c** and **d**) Hom signal with red and Het signal with blue for an (c) affected and (d) unaffected individual. (**e**) Differences between the Hom and Het signals are plotted for all individuals. (**f**) Prediction signal of chromosome 17

ROSH for this consanguineous family. The detailed results (Supplementary Figs S2 and S7) and calculations of all signals are presented in Supplementary Material.

2.2 Comparative evaluation

We evaluated HomSI using both a simulated dataset generated by (Seelow and Schuelke, 2012) and a real dataset of three disease genes within the homozygous regions, which have been previously identified using a combination of exome and SNP microarray data (Aldahmesh *et al.*, 2012; Shamseldin *et al.*, 2012). On the simulated dataset, HomSI successfully detected the homozygous region on chromosome 15, as shown in Supplementary Figures S8 and S9. On the real dataset, we investigated whether the exome sequence data could identify a homozygous region, without prior SNP array analysis. For all three patients, HomSI detects the homozygous regions, as shown in Supplementary Figures S10 and S12.

3 CONCLUSIONS

We have developed a state-of-the-art method to identify homozygous stretches using NGS data and created a user-friendly tool. Because the usage of HomSI is so simple and intuitive, the clinicians or researchers can immediately start their data analysis without consulting to any dedicated information technology specialist. We provide a step-by-step tutorial and a detailed documentation on our Web site.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Dominik Seelow, Dr Ian M. Carr, Dr Fowzan S. Alkuraya and Dr Fatih Bayraklı for sharing their datasets with them to test our program.

Funding: This work is supported by the Turkish State Planning Organization Research Grants, Grant Number: 108S420 and by UEKAE, BİLGEM, The Scientific and Technology Research Council of Turkey (TUBITAK), Grant Number: K030-T439.

Conflict of Interest: none declared.

REFERENCES

Aldahmesh,M.A. *et al.* (2012) Identification of a truncation mutation of acylglycerol kinase (AGK) gene in a novel autosomal recessive cataract locus. *Hum. Mutat.*, **33**, 960–962.

Amir el,A.D. *et al.* (2010) KinSNP software for homozygosity mapping of disease genes using SNP microarrays. *Hum. Genomics*, **4**, 394–401.

Bayrakli,F. *et al.* (2013) Mutation in MEOX1 gene causes a recessive Klippel-Feil syndrome subtype. *BMC Genet.*, **14**, 95.

Carr,I.M. *et al.* (2006) Interactive visual analysis of SNP data for rapid autozygosity mapping in consanguineous families. *Hum. Mutat.*, **27**, 1041–1046.

Carr,I.M. *et al.* (2013) Autozygosity mapping with exome sequence data. *Hum. Mutat.*, **34**, 50–56.

Kayserili,H. *et al.* (2009) ALX4 dysfunction disrupts craniofacial and epidermal development. *Hum. Mol. Genet.*, **18**, 4357–4366.

Lander,E.S. and Botstein,D. (1987) Homozygosity mapping—a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Papic,L. *et al.* (2011) SNP-array based whole genome homozygosity mapping: a quick and powerful tool to achieve an accurate diagnosis in LGMD2 patients. *Eur. J. Med. Genet.*, **54**, 214–219.

Seelow,D. and Schuelke,M. (2012) HomozygosityMapper2012-bridging the gap between homozygosity mapping and deep sequencing. *Nucleic Acids Res.*, **40**, W516–W520.

Seelow,D. *et al.* (2009) HomozygosityMapper-an interactive approach to homozygosity mapping. *Nucleic Acids Res.*, **37**, W593–W599.

Shamseldin,H.E. *et al.* (2012) Genomic analysis of mitochondrial diseases in a consanguineous population reveals novel candidate disease genes. *J. Med. Genet.*, **49**, 234–241.

Uz,E. *et al.* (2010) Disruption of ALX1 causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive ALX-related frontonasal dysplasia. *Am. J. Hum. Genet.*, **86**, 789–796.

Zhang,L. *et al.* (2011) Homozygosity mapping on a single patient-identification of homozygous regions of recent common ancestry by using population data. *Hum. Mutat.*, **32**, 345–353.