

Whole-Genome Association Analysis to Identify Markers Associated with Recombination Rates Using Single Nucleotide Polymorphisms and Microsatellites

Song Huang¹, Shuang Wang², Nianjun Liu³, Liang Chen⁴, Cheongeun Oh³, Hongyu Zhao^{3, 5§}

¹Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

²Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

³Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520, USA

⁴Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520, USA

⁵Department of Genetics, Yale University, New Haven, CT 06520, USA

§Corresponding author

Email addresses:

SH: song.huang@yale.edu

SW: shuang.wang@columbia.edu

NL: nianjun.liu@yale.edu

LC: liang.chen@yale.edu

CO: cheongeun.oh@yale.edu

HZ: hongyu.zhao@yale.edu

Abstract

Recombination during meiosis is one of the most important biological processes and the level of recombination rates for a given individual is under genetic control. In this study, we conducted genome-wide association studies to identify chromosomal regions associated with recombination rates. We analyzed genotype data collected on the pedigrees in the Collaborative Study on the Genetics on Alcoholism data provided by GAW14. A total of 315 microsatellites and 10,081 SNPs from Affymetrix on 22 autosomal chromosomes were used in our association analysis. Genome-wide gender-specific recombination counts for family founders were inferred first and association analysis was performed using multiple linear regressions. We used the positive False Discovery Rate (pFDR) to account for multiple comparisons in the two genome-wide scans. Eight regions showed some evidence of association with recombination counts based on the SNP analysis after adjusting for multiple comparisons. However, no region was found to be significant using microsatellites.

Background

Recombination between two homologous chromosomes during meiosis generates novel gene combinations and creates genetic diversity among chromosomes. Furthermore, recombination is critical for proper segregation of homologous chromosomes, and is a major factor shaping linkage disequilibrium (LD) patterns in the genome [1]. Much research has been done recently to establish human genetic maps based on recombination and on estimating local recombination rates to augment LD studies and aid in LD study design and interpretation [1-8]. Kong et al. [2] found marked regional differences in recombination rates and concluded that DNA changes contributing to evolution may not be completely random, but more concentrated within specific regions. This difference may be driven by sequence features. In addition, recombination rate is under genetic control, as exemplified in the finding by Ji et al. [9] that maize meiotic mutant *desynaptic* is a recombination modifier that controls recombination rates. In this study, seeking to identify regions potentially affecting recombination rates, we conducted genome-wide association studies based on microsatellites and SNPs of the COGA data provided by GAW14. A total of 315 microsatellites and 10,081 SNPs from Affymetrix on 22 autosomal chromosomes were analyzed. We found eight regions/thirteen SNPs that showed some evidence of association with recombination counts. No region was found to be significant using microsatellites after adjusting for multiple comparisons based on the positive False Discovery Rate (pFDR) criterion.

Methods

Recombination Counts

The COGA data consist of 143 pedigrees with 1,614 individuals, including 1,109 male and female meioses. Genetic maps for microsatellites and SNPs were both provided by GAW14. Some of the distinct SNPs have the same genetic map position, which made inferring recombination events between these SNPs impossible. Therefore, we added $1.0E-06$ at these SNPs' genetic map positions to make them distinguishable. To estimate the number of both maternal and paternal recombination events for each female or male meiosis, we used the *Best* option in the Haplotyping analysis in MERLIN [10], which outputs the most likely haplotype as well as the most likely sites for recombination

through a pedigree. The total number of gender-specific recombination counts for each parent was obtained by averaging the numbers of recombination events of all the offspring, which was calculated as the total number of recombination events observed in the 22 autosomal chromosomes. For pedigrees with only two generations, i.e., the nuclear families, the inferred average total number of recombination events from each meiosis of the founders was then treated as a quantitative trait and genome-wide association tests were conducted to identify markers associated with this quantitative trait. For the pedigrees with three or more generations, only recombination information from the founders were extracted and considered in the association tests. We compared the results from the two scans using either microsatellites or SNPs.

Genotyping Error Detection

Because genotyping error may lead to double recombinations within a short distance, it can significantly affect the overall recombination counts. To minimize this impact, the error-checking algorithm implemented in MERLIN, which identifies unlikely genotypes based on double recombination events, was applied and the erroneous genotypes were excluded before applying Haplotyping analysis. We used the default parameter in MERLIN, where the erroneous genotypes with a likelihood ratio $p \leq 0.025$ were excluded [11]. The same procedure was applied to both SNPs and microsatellites.

Association Analysis to Identify Markers Associated with Recombination Rates

We used multiple linear regressions to evaluate the relation between recombination counts and markers across 22 autosomal chromosomes with adjustments for age and gender for both SNPs and microsatellites. Analysis was carried out based on Whites only to reduce potential confounding factors related to ethnic differences. To account for the multiple comparison problem in the two whole genome scans, we used pFDR through q-values [12], where a cutoff point of 5% is chosen. The q-value is a measure of significance in terms of the pFDR, and it is defined to be the minimum pFDR at which the statistic can be called significant. A pFDR of 5% means that among all the features that are called significant, 5% of them may correspond to the true null hypotheses on average. To get the q-value for each marker, we used the software Qvalue [12] on the p-values obtained from the multiple regressions.

Results and Discussion

Recombination Counts

The genome-wide gender-specific recombination counts of the founders were obtained through averaging recombination counts in all meioses leading to his/her offspring. For SNPs, we inferred the founders' genome-wide recombination counts from the gametes of 1,334 offspring from 130 nuclear families. For microsatellites, we inferred the founders' genome-wide recombination counts from the gametes of 1,409 offspring from 111 nuclear families. This resulted in 189 founders (121 females and 68 males) who were Whites with information for SNPs and 199 founders (129 females and 70 males) who were Whites with information for microsatellites. Among these founders, 14 of them missed all microsatellite genotype information and 23 of them missed all SNP genotype information. They had no contribution in the multiple regression analysis. Therefore, the distributions of the founders' gender-specific recombination counts plotted in Figure1 did

not include these founders. We now had 166 founders (106 females and 60 males) that were Whites with information for SNPs and 185 founders (120 females and 65 males) that were Whites with information for microsatellites. The scatter plots of the inferred recombination counts using SNPs and microsatellites for Whites only showed a higher correlation for males than for female. The recombination counts were much higher for females than for males, a well-known biological fact [2, 4]. We also noted that the recombination counts were higher using SNPs than those using microsatellites, which may be due to the fact that the SNPs were more dense than the microsatellites, allowing for the capture of recombination events missed by the microsatellites. The mean and median genome-wide gender-specific recombination counts are summarized in Table 1. From the scatter plot for the females, we noted that there were two female founders who had very high inferred recombination counts using the SNPs, 105.4 and 77.8, respectively. In our analysis, we removed the female founder with the average recombination count of 105.4. The above analysis was conducted after removing the possible erroneous genotypes. There were 1,295 microsatellite genotypes that were likely to be erroneous and were set missing with the MERLIN's error checking algorithm, making the estimated genotyping error rate for the microsatellite to be 0.367% as among the 1,614 individuals and the 315 microsatellites there were a total of 353,015 genotypes. Similarly, there were 27,338 SNP genotypes that were likely to be erroneous and were set missing with the MERLIN's error checking algorithm. This led to the estimated genotyping error rate for the SNPs to be 0.204% from among the 1,614 individuals and the 10,081 SNPs genotyped. There were a total of 13,395,832 genotypes examined.

We noted that our inferred female and male genome-wide recombination counts were slightly lower than that from previous studies [4]. One reason may be that the 10,081 SNPs did not cover the entire 22 autosomes as the updated SNP data from Affymetrix were not included in the analysis. Another possible reason was that some portion of the corrected genotypes was excluded as erroneous genotypes from the genotyping error detection algorithm.

Markers Associated with Recombination Counts

Multiple linear regressions with adjustments for age and gender generated p-values for each marker, which were not adjusted for multiple comparisons. The corresponding q-values based on the pFDR were calculated using the software Qvalue. We applied the 0.05 q-value cutoff, which gave us eight regions/thirteen SNPs that showed some evidence of association with recombination counts. The positions of those regions together with the raw p-values and q-values were summarized in Table 2. The 0.05 q-value cutoff suggested that one out of these thirteen SNPs may not be associated with recombination counts. For microsatellites, no region was found to be significant after adjusting for multiple comparisons using pFDR.

Conclusions

In summary, we have identified several candidate SNPs likely associated with recombination events, and further studies on these genes may help us gain valuable,

knowledge on recombination, better understand LD patterns, and lead to more efficient methods to map disease genes.

Authors' contributions

SH participated in the design of the study, performed the analysis, and drafted the manuscript. SW helped to obtain recombination counts and preparation of the manuscript. SW, NL, LC, and CO participated in the design and the discussion of the study. HZ conceived the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Supported in part by NIH grant R01 GM59507 and NSF grant DMS 0241160.

References

1. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *Am J Hum Genet* 2001, **69**: 1-14.
2. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al.: **A high-resolution recombination map of the human genome.** *Nat Genet* 2002, **31**: 241-247.
3. Yu AD, Zhao CF, Fan Y, Jang WH, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KL, et al.: **Comparison of human genetic and sequence-based physical maps.** *Nature* 2001, **409**: 951-953.
4. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: **Comprehensive human genetic maps: individual and sex-specific variation in recombination.** *Am J Hum Genet* 1998, **63**: 861-869.
5. Yu J, Lazzeroni L, Qin J, Huang MM, Navidi W, Erlich H, Arnheim N: **Individual variation in recombination among human males.** *Am J Hum Genet* 1996, **59**: 1186-1192.
6. Hudson RR: **Two-locus sampling distributions and their application.** *Genetics* 2001, **159**: 1805-1817.
7. McVean GA, Awasalla P, Fearnhead P: **A coalescent-based method for detecting and estimating recombination from gene sequences.** *Genetics* 2002, **160**: 1231-1241.
8. Stumpf MPH, McVean GAT: **Estimating recombination rates from population-genetic data.** *Nat Reviews* 2003, **4**: 959-968.
9. Ji YF, Stelly DM, Donato MD, Goodman MM, Williams CG: **A candidate recombination modifier gene for Zea mays L.** *Genetics* 1999, **151**: 821-830.
10. Abecasis G, Cherny SS, Cookson WO, Crdon LR: **Merlin-rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30**: 97-101
URL: <http://www.sph.umich.edu/csg/abecasis/Merlin/>.
11. John S, Shephard N, Liu GY, Zehhini E, Cao MQ, Chen WW, Vasavda N, Mills T, Barton A, Hinks A, et al.: **Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites.** *Am J Hum Genet* 2004, **75**: 54-64.

12. Story JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100**: 9440-9445
 URL: <http://faculty.washington.edu/~jstorey/qvalue/>.

Figure legends

Figure 1 – Distribution of the gender-specific recombination counts

Results shown are for Whites only when erroneous genotypes are excluded and when founders with all genotype information missing are excluded as well.

Tables

Table1 – Mean and median genome-wide gender-specific recombination counts using SNPs and microsatellites

Results shown are after removing erroneous genotypes and removing founders that have all genotype information missing for Whites only.

		SNP		Microsatellite	
		Mean (S.D.)	Median	Mean (S.D.)	Median
Female	With outlier	35.83 (9.35)	35.1	25.20 (3.74)	25.38
	Without outlier	35.17 (6.42)	35.0		
Male		19.57 (2.34)	19.6	17.06 (2.23)	17.0

Table 2. Significant results (q-value < 0.05) for genome-wide association analysis for recombination rates

Chr	Using SNP		
	Position (cM) (Marker)	Unadjusted p-value	q-value
1q	129.761 tsc0831812	3.99E-05	0.032874
	173.005 tsc1229896	3.64E-06	0.005141
	*176.25 tsc1687896	3.74E-05	0.032874
2q	*136.452 tsc0333128	1.01E-05	0.012482
	167.814 tsc0045403	2.00E-08	0.000198
	167.817 tsc1108827	1.21E-06	0.001994
3q	70.968 tsc0753329	4.95E-05	0.037662
4q	94.916 tsc0056600	8.00E-08	0.000264
8q	103.504 tsc1305199	6.00E-08	0.000264
10q	45.654 tsc0615240	2.60E-07	0.000514
	74.381 tsc0046577	1.45E-05	0.015691
13q	50.514 tsc0616973	2.40E-07	0.000514
14q	*85.42 tsc1112831	1.59E-05	0.015691

- No significant results are found with microsatellites using the pFDR 0.05 cutoff
- * indicates that the marker identified was from the markers that had the same map position