

Whole-Genome Linkage Analysis in Mapping Alcoholism Genes Using Single Nucleotide Polymorphisms and Microsatellites

Shuang Wang¹, Song Huang², Nianjun Liu³, Liang Chen⁴, Cheongeun Oh³, Hongyu
Zhao^{3, 5§}

¹Department of Biostatistics, Mailman School of Public Health, Columbia University,
New York, NY 10032, USA

²Program of Computational Biology and Bioinformatics, Yale University, New Haven,
CT 06520, 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:

SW: shuang.wang@columbia.edu

SH: song.huang@yale.edu

NL: nianjun.liu@yale.edu

LC: liang.chen@yale.edu

CO: cheongeun.oh@yale.edu

HZ: hongyu.zhao@yale.edu

Abstract

There is currently a great interest in using SNPs in genetic linkage and association studies because of the abundance of SNPs as well as the availability of high-throughput genotyping technologies. In this study, we empirically compared the performance of whole-genome scans using SNPs versus microsatellites on 143 pedigrees from the Collaborative Studies on Genetics of Alcoholism provided by GAW14. A total of 315 microsatellites and 10,081 SNPs from Affymetrix on 22 autosomal chromosomes were used in our analyses. We found that the results from the two scans had good overall concordance. One region on chromosome 2 and two regions on chromosome 7 showed significant linkage signals (i.e., $NPL \geq 2$) with alcoholism from both SNP and microsatellite scans. The different results observed between the two scans may be explained by the difference observed in information content between SNPs and microsatellites.

Background

There is currently a great interest in using SNPs in genetic linkage and association studies because of the abundance of SNPs as well as the availability of high-throughput genotyping technologies. Kruglyak [1] predicted in a theoretical study that approximately twice or three times the density of SNPs with a heterogeneity of 0.5 would be equivalent to the current microsatellites. With current high-throughput SNP genotyping technologies, it is now feasible and affordable to collect genotype data from tens of thousands of SNPs. John et al. [2] described the first whole-genome scans with linkage analysis of a complex disease, rheumatoid arthritis (RA[MIM 180300]), to

directly compare SNPs with microsatellites. In this paper, using the data from the Collaborative Studies on Genetics of Alcoholism (COGA) provided by GAW14, we compared the results based on whole-genome scans of 143 pedigrees using 315 microsatellites and 10,081 SNPs from Affymetrix throughout 22 autosomal chromosomes.

Methods

Nonparametric Linkage Analysis

COGA data provided by GAW14 include 143 pedigrees with 1,614 individuals genotyped with both microsatellites and SNPs. In addition, genetic maps for both microsatellites and SNPs were provided. We used the nonparametric linkage analysis implemented in MERLIN [3] for linkage analysis. Individuals were defined as unaffected with alcoholism if they never drank alcohol or if they showed some alcohol related syndromes but did not meet the criteria for alcoholism [4]. Allele frequencies were estimated using all genotyped individuals and the Whittemore and Halpern “ALL” statistic [5] was applied for the scan procedure, where NPL scores based on all affected pedigree members were calculated. Both SNP scan and microsatellite scan were performed at each marker locus.

Genotyping Error Detection

To avoid potential bias caused by possible genotyping errors on linkage signals, the error-checking algorithm implemented in MERLIN was applied. This algorithm identifies unlikely genotypes based on double recombination events, when erroneous genotypes can

imply excessive and unlikely recombination events between tightly linked markers [3].

We used the default parameter in MERLIN, where the likelihood ratio of an erroneous genotype with $p \leq 0.025$ was excluded [2]. The two whole-genome scans were carried out both with and without those erroneous genotypes excluded to exam the effect of genotyping error on the scan results.

Information Content (IC)

Information content (IC) was calculated using MERLIN to compare microsatellites and SNPs in order to investigate factors contributing to the differences between the two scans. The microsatellites have an average of 13 cM spacing, whereas the SNPs have an average of 0.35 cM spacing. Besides the comparison between microsatellites and the full set of SNPs, to assess the effect of the reduced IC on the SNP scan, a 3,360-SNP map with an average of 1.0 cM spacing was randomly extracted from the full set of SNPs as a subset for a separate scan.

Results

Nonparametric Linkage Analysis

We applied nonparametric linkage analysis to map regions linked to alcoholism for all 22 autosomal chromosomes. The results from the whole-genome scans using microsatellites and SNPs had good overall concordance. Six regions showed some evidence of increased allele sharing with a NPL cutoff value of 2 for either the SNP scan, or the microsatellite scan, or both. The results were summarized in Table 1, which also included those when erroneous genotypes were included in the analysis. The NPL scores

across the 22 autosomal chromosomes when erroneous genotypes were excluded were shown in Figure 1 for both microsatellites and SNPs. We noticed that, overall, the scan using SNPs gave stronger linkage signals than those using microsatellites. Except two regions on chromosomes 2 and 13 that showed significant linkage evidence using microsatellites but did not when using SNPs (there was no SNP genotyped at the region on chromosome 13.), SNP scan gave stronger linkage signal. Four regions on chromosomes 1, 2, 12, and 13 showed significant linkage evidence when using SNPs but did not when using microsatellites. Both the SNP and microsatellite scans indicated strong linkage signals on chromosome 7, and relatively strong linkage signals on chromosome 2. Results for these two chromosomes were plotted separately in Figure 2 and Figure 3 when erroneous genotypes were either excluded or included. We also presented one-LOD confidence intervals of these peaks in Figure 2. In general, the peaks were better defined for the SNPs, where peaks from SNPs had narrower 1-LOD interval than peaks from microsatellites (SNP 1-LOD interval was 20 cM, compared with a 40-cM 1-LOD interval in the microsatellites scan for the peak on chromosome 7 around 100 cM. SNP 1-LOD interval was 9 cM, compared with a 16-cM 1-LOD interval in the microsatellites scan for the peak on chromosome 7 around 60 cM. One-LOD intervals for the peaks on chromosome 2 around 10 cM have similar width for SNPs and microsatellites.) Comparison between SNP full set scan and SNP 1 cM subset scan showed that the NPL score decreased in the SNP 1cM scan for all but one region on chromosome 2, at about 18cM. With the NPL cutoff 2, several regions on chromosomes 2, 7, and 12 that were significant in the SNP full set scan were no longer significant in the SNP 1cM subset scan. We also noted that the effect of genotyping error on the results

from linkage analysis was small for this particular dataset, although potential genotyping errors seemed to increase the linkage signal slightly. This finding contradicted with the finding from John et al. [2], who suggested that removal of unlikely genotypes could increase the significance of nominal loci. The discrepancy may due to the different genotyping error rates in the two data sets. The COGA data had very small genotyping error rate according to MERLIN. There were 1295 microsatellites genotypes that were likely to be errors and were set missing with MERLIN's error checking algorithm. Among 1614 individuals and 315 microsatellites, there were in total 353015 genotypes, so the error rate for microsatellite marker sets was 0.367%. Similarly, there were 27338 SNP genotypes that were likely to be errors and were set missing with MERLIN's error checking algorithm. Among 1614 individuals and 10081 SNPs, there were 13395832 genotypes, so the error rate for SNP marker sets was 0.204%.

Information Content (IC)

The major advantage of using high density SNPs versus microsatellites is the increased IC. The IC calculated in MERLIN for all the pedigrees for 10,081 SNPs and those for 315 microsatellites across 22 autosomal chromosomes when erroneous genotypes were excluded were plotted in Figure 4. We can see that the IC for SNPs was significantly and uniformly higher than that for microsatellites. The mean genome-wide IC for microsatellites was 0.783 with an inter-quartile range of 0.134, and was 0.950 for SNPs with an inter-quartile range of 0.017. The mean IC for each individual chromosome for both SNPs and microsatellites was summarized in Table 2, where both results were shown when erroneous genotypes were included and when erroneous genotypes were

excluded. Again, the impact of genotyping errors was quite small on IC, although genotyping errors tended to reduce the values of IC slightly as expected. Again, this may be due to the small error rate of this data set.

SNP Subset

To further explore the possible effect of IC on the linkage scan results, we randomly extracted a subset of 3,360 SNPs from the full SNP set with an average of 1.0 cM spacing. We then re-did the whole-genome scan with this subset of SNPs. The mean genome-wide IC for the 1-cM SNP map was 0.905 with an inter-quartile range of 0.044 when erroneous genotypes were included. The mean genome-wide IC for the 1-cM SNP map was 0.910 with an inter-quartile range of 0.042 when erroneous genotypes were excluded. The mean IC results for each individual autosomal chromosome for the 1cM SNP subset were also shown in Table 2. There was a significant decrease in genome-wide IC for the SNP subset compared to the SNP full set. The multipoint NPL results based on the subset were also included in Table 1. For all previously identified regions that showed some evidence of linkage with SNP full set, except one region on chromosome 2, where NPL score increased in SNP 1cM subset scan, the NPL scores from the 1 cM SNP map scan all decreased. Several regions on chromosomes 2, 7, and 12 that were significant in the SNP full set scan were no longer significant in the SNP 1cM subset scan with the NPL cutoff 2. This suggested that the higher IC in the full set of 10,081 SNPs may partly explain the observed differences between the linkage results based on the full SNP set and SNP subset whole-genome scans, as well as those between the microsatellite and SNP whole-genome scans.

Discussion

In this study, we have compared the genome wide linkage analyses based on microsatellites and SNPs. Software MERLIN was used to conduct nonparametric linkage analysis to map regions associated with alcoholism on 22 autosomal chromosomes. The results from the two scans had good concordance in general, although more significant signals were obtained using SNPs versus microsatellites. Both scans suggested strong linkage evidence on chromosome 2 and chromosome 7, where the two scans agreed especially well. Microsatellite scan had peak at marker locus D7S820 at 107.5 cM with the NPL score 2.56 on chromosome 7, and SNP scan had peak at marker locus tsc0046246 at 100.9 cM with the NPL score 2.81. For chromosome 2, microsatellite scan had peak at marker locus D2S1329 at 4.9 cM with the NPL score 2.13, and SNP scan had peak at marker locus tsc0056805 at 243.6 cM with the NPL score 2.80. The different results observed in the two scans were likely to be explained by the difference between the information content between microsatellites and SNPs. In fact, the higher IC is one major advantage of the high-density SNPs compared to the conventional microsatellite sets. The IC across the genome for the SNPs was uniformly higher than that for the microsatellites.

As expected, the analysis based on the SNP subset showed decreased IC, and reduced linkage signals compared to the SNP full set, which suggests that the difference in IC might be one key factor that contributes to the observed difference in the two scans. This was consistent with the conclusion from John et al. [2], who examined possible reasons for the observed difference between the scans using SNPs and microsatellites

comprehensively, including genotyping error of either SNPs or microsatellites, possible errors in the two maps used, presence of linkage disequilibrium (LD), and differences in IC. We have also investigated the possible effect of genotyping errors on the linkage results and IC. Our results suggested that the impact of genotyping errors was quite small for the COGA dataset, which may be due to the small genotyping error rate (0.37% for microsatellites and 0.20% for SNPs) for this specific data set and may not be generalized to other data sets. It is worth noting that for the full SNP set with an average of 0.35 cM spacing, it is highly possible that there is LD between SNPs, which may influence the linkage results from MERLIN as MERLIN assumes linkage equilibrium between all markers. John et al. [2] explored possible effect of LD on the two scans by keeping one SNP from a group of SNPs in LD, or by assigning haplotypes to individuals for clusters of SNPs in LD and treating them as multi-allelic markers. They found that for both cases, there were losses in IC, which made it difficult to assess whether the difference observed in the two scans was due to LD or to losses in IC. They concluded that overall the results were qualitatively similar when SNPs in LD are either included or excluded.

Finally, we noted that the scan with the 1 cM SNP subset was able to detect some regions detected by the scan with the full SNP set, and the 1 cM SNP subset has an average IC of 0.910 compared to the average IC of 0.950 for the 0.35 cM full SNP set. With the NPL cutoff 2, the 1 cM SNP subset scan resulted in loss of significance of several regions on chromosomes 2, 7, and 12.

Conclusions

We have identified two regions that showed some evidence of linkage with alcoholism on chromosome 2 and chromosome 7 from both microsatellite and SNP scans. For these regions, we had stronger linkage signals using SNPs than those using microsatellites. Although results from the two scans had good overall concordance, three regions of significant linkages were detected in the SNP scan but not in the microsatellite scan. Lastly, the difference in information content between SNPs and microsatellites might explain the different results observed in the two scans.

Authors' contributions

SW participated in the design of the study, performed the analysis, and drafted the manuscript. SH, 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.

References

1. Kruglyak L: **The use of a genetic map of biallelic markers in linkage studies.** *Nat Genet* 1997, **17**: 21-24
2. John S, Shepard N, Liu GY, Zeggini ZE, 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
3. Abecasis GR, Cherny SS, Cookson WO, Cardon 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/>
4. Sun FZ, Cheng R, Flanders WD, Yang QH, Khoury MJ: **Whole genome association studies for genes affecting alcohol dependence.** *Genet Epidemiol* 1999, **17**: Suppl 1:S337-42
5. Whittemore AS, Halpern J: **A class of tests for linkage using affected pedigree members.** *Biometrics* 1994, **50**: 118-127

Figure legends

Figure 1 – Multipoint NPL scores for the whole cohort for SNPs and microsatellites across 22 autosomes (erroneous genotypes excluded).

Blue solid line stands for microsatellites and red dashed line stands for SNPs. Results shown are when erroneous genotypes are excluded.

Figure 2 – Multipoint NPL scores and IC from 1cM SNP scan for chromosomes 2, 7 (erroneous genotypes excluded).

Blue solid line stands for microsatellites and red dashed line stands for SNPs. Vertical lines stand for 1-LOD intervals. IC stands for information contents. Results shown are when erroneous genotypes are excluded.

Figure 3 – Multipoint NPL scores and IC from 1cM SNP scan for chromosomes 2, 7 (erroneous genotypes included).

Blue solid line stands for microsatellites and red dashed line stands for SNPs. IC stands for information contents. Results shown are when erroneous genotypes are included.

Figure 4 – IC for the whole cohort for SNPs and microsatellites across 22 autosomes (erroneous genotypes excluded).

Blue solid line stands for microsatellites and red dashed line stands for SNPs. IC stands for information contents. Results shown are when erroneous genotypes are excluded.

Tables

Table 1 – Regions that show some evidence of increased allele sharing.

Results shown are when NPL scores are greater or equal to 2.0 on either SNP scan or microsatellite scan or both with and without erroneous genotypes.

Chr.	Position (cM)	Microsatellites (excluding erroneous genotypes) (n = 315)		SNPs from Affymetrix (excluding erroneous genotypes) (n = 10,081)		SNPs 1cM subset (excluding erroneous genotypes) (n = 3,360)		Microsatellites (with erroneous genotypes) (n=315)		SNP (with erroneous genotypes) (n=10,081)		SNPs 1cM subset (with erroneous genotypes) (n = 3,360)	
		NPL score	P	NPL score	P	NPL score	P	NPL Score	P	NPL Score	P	NPL score	P
1q	77	0.24	0.4	1.80	0.04	1.64	0.05	0.29	0.4	2.06	0.02	2.10	0.02
	146	0.85	0.2	1.97*	0.03	1.88*	0.03	0.82	0.2	2.06	0.02	1.63	0.05
2q	5	2.13*	0.02	1.14	0.13	0.52	0.3	2.19	0.014	1.29	0.10	0.27	0.4
	18	2.08	0.02	2.16	0.02	2.35 ^c	0.009	2.08	0.02	2.19	0.014	2.29 ^c	0.01
	118	0.49	0.30	2.24	0.013	1.88	0.03	0.54	0.30	2.00	0.02	2.13 ^f	0.02
	135	0.59	0.30	2.15	0.02	1.83	0.03	0.29	0.40	2.03	0.02	1.67	0.05
	244	0.90	0.20	2.80*	0.003	2.46*	0.007	0.92	0.20	3.04	0.0012	2.46	0.007
7q	14	1.21 ^a	0.11	2.30	0.011	1.77	0.04	1.64 ^a	0.05	2.30	0.011	1.66	0.05
	32	1.56	0.06	2.69	0.004	2.36	0.009	1.83	0.03	2.73	0.003	2.32	0.01
	60	2.37 ^b	0.009	2.10	0.02	1.32	0.09	2.83 ^b	0.002	2.02	0.02	1.30	0.1
	94	1.90	0.03	2.20	0.014	2.01	0.02	2.28	0.011	2.20	0.014	1.92	0.03
	101	1.97	0.02	2.81*	0.002	2.51*	0.006	2.10	0.02	2.88	0.002	2.47	0.007
	106	2.56*	0.005	2.32	0.01	1.94	0.03	2.45	0.007	2.32	0.01	2.07	0.02
11q	107	1.32	0.09	2.24* [#]	0.012	2.14*	0.02	1.32	0.09	2.23	0.013	2.15	0.02
	120	2.61*	0.004	NA	NA	NA	NA	2.60	0.005	NA	NA	NA	NA
12q	122	1.02 ^c	0.2	2.02*	0.02	1.87*	0.03	1.02 ^c	0.2	1.95	0.03	1.81	0.04
13q	86	1.15 ^d	0.13	2.63*	0.004	2.61*	0.005	1.31 ^d	0.10	2.56	0.005	2.55	0.005

a. At position 21

b. At position 57

c. At position 117

d. At position 90

e. At position 20

f. At position 120

• stands for peak location on that chromosome

at position 108 with NPL score 2.40

NA stands for not available

cM stands for centi-Morgan

Table 2 – Mean information content and its standard deviations across 22 autosomes

Results shown are for SNP full set, 1cM SNP subset, and microsatellites.

Chr	Mean Information Content (S.D.)					
	SNP (Exclude erroneous genotypes)	SNP (Include erroneous genotypes)	Microsatellites (Exclude erroneous genotypes)	Microsatellite s (Include erroneous genotypes)	SNP 1cM subset (Exclude erroneous genotypes)	SNP 1cM subset (Include erroneous genotypes)
1	0.951 (0.020)	0.947 (0.020)	0.821 (0.073)	0.821 (0.072)	0.914 (0.040)	0.909 (0.045)
2	0.955 (0.017)	0.952 (0.017)	0.831 (0.082)	0.830 (0.081)	0.918 (0.040)	0.914 (0.041)
3	0.954 (0.016)	0.951 (0.016)	0.745 (0.086)	0.745 (0.086)	0.917 (0.042)	0.914 (0.044)
4	0.953 (0.028)	0.949 (0.028)	0.763 (0.036)	0.761 (0.036)	0.921 (0.046)	0.916 (0.055)
5	0.954 (0.018)	0.951 (0.019)	0.791 (0.078)	0.789 (0.078)	0.916 (0.043)	0.913 (0.046)
6	0.954 (0.017)	0.951 (0.016)	0.784 (0.080)	0.782 (0.079)	0.921 (0.031)	0.918 (0.033)
7	0.953 (0.014)	0.950 (0.015)	0.879 (0.062)	0.878 (0.061)	0.916 (0.031)	0.912 (0.033)
8	0.954 (0.016)	0.950 (0.017)	0.801 (0.079)	0.800 (0.078)	0.916 (0.033)	0.911 (0.039)
9	0.956 (0.020)	0.951 (0.020)	0.763 (0.088)	0.764 (0.089)	0.916 (0.042)	0.911 (0.045)
10	0.953 (0.015)	0.950 (0.015)	0.662 (0.084)	0.662 (0.084)	0.917 (0.036)	0.914 (0.040)
11	0.936 (0.017)	0.933 (0.019)	0.699 (0.054)	0.698 (0.055)	0.910 (0.035)	0.903 (0.046)
12	0.952 (0.021)	0.947 (0.023)	0.807 (0.086)	0.806 (0.086)	0.906 (0.062)	0.899 (0.066)
13	0.951 (0.027)	0.948 (0.027)	0.791 (0.092)	0.792 (0.092)	0.916 (0.056)	0.913 (0.057)
14	0.947 (0.032)	0.942 (0.032)	0.762 (0.048)	0.761 (0.047)	0.909 (0.050)	0.902 (0.054)
15	0.949 (0.017)	0.945 (0.017)	0.801 (0.032)	0.801 (0.033)	0.907 (0.038)	0.903 (0.038)
16	0.937 (0.045)	0.933 (0.045)	0.738 (0.089)	0.737 (0.088)	0.877 (0.077)	0.867 (0.093)
17	0.930 (0.050)	0.926 (0.049)	0.705 (0.082)	0.704 (0.082)	0.859 (0.061)	0.851 (0.073)
18	0.949 (0.017)	0.945 (0.017)	0.678 (0.040)	0.678 (0.039)	0.895 (0.049)	0.889 (0.050)
19	0.902 (0.074)	0.899 (0.073)	0.709 (0.041)	0.710 (0.042)	0.759 (0.142)	0.747 (0.149)
20	0.941 (0.028)	0.936 (0.029)	0.750 (0.115)	0.750 (0.115)	0.879 (0.062)	0.873 (0.067)
21	0.948 (0.026)	0.945 (0.026)	0.780 (0.074)	0.780 (0.075)	0.899 (0.043)	0.892 (0.053)
22	0.904 (0.046)	0.901 (0.046)	0.644 (0.131)	0.644 (0.130)	0.794 (0.101)	0.786 (0.102)
Overall	0.9500 (0.025)	0.9465 (0.025)	0.7828 (0.092)	0.7822 (0.092)	0.910 (0.051)	0.9046 (0.056)