2005

Multipoint identity-by-descent computations for single-point polymorphism and microsatellite maps

Anthony L. Hinrichs
Washington University School of Medicine in St. Louis

Sarah Bertelsen
Washington University School of Medicine in St. Louis

Laura J. Bierut
Washington University School of Medicine in St. Louis

Gerald Dunn
Washington University School of Medicine in St. Louis

Carol H. Jin
Washington University School of Medicine in St. Louis

See next page for additional authors

BMC Genetics

BioMed Central

Proceedings

Open Access

Multipoint identity-by-descent computations for single-point polymorphism and microsatellite maps

Anthony L Hinrichs*[1], Sarah Bertelsen[1], Laura J Bierut[1], Gerald Dunn[1], Carol H Jin[1,2], John S Kauwe[3] and Brian K Suarez[1,2]

Address: [1]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, USA, [2]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA and [3]Division of Biology and Biomedical Sciences, Washington University School of Medicine, St. Louis, Missouri, USA

Email: Anthony L Hinrichs* - tony@silver.wustl.edu; Sarah Bertelsen - sarah@silver.wustl.edu; Laura J Bierut - bierutl@notes.wustl.edu; Gerald Dunn - dunnge@notes.wustl.edu; Carol H Jin - carolj@nackles.wustl.edu; John S Kauwe - keoni@icarus.wustl.edu; Brian K Suarez - bks@themfs.wustl.edu

* Corresponding author

from Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism
Noordwijkerhout, The Netherlands, 7-10 September 2004

Published: 30 December 2005

BMC Genetics 2005, **6**(Suppl 1):S34 doi:10.1186/1471-2156-6-S1-S34

Abstract

We used the LOKI software to generate multipoint identity-by-descent matrices for a microsatellite map (with 31 markers) and two single-nucleotide polymorphism (SNP) maps to examine information content across chromosome 7 in the Collaborative Study on the Genetics of Alcoholism dataset. Despite the lower information provided by a single SNP, SNP maps overall had higher and more uniform information content across the chromosome. The Affymetrix map (578 SNPs) and the Illumina map (271 SNPs) provided almost identical information. However, increased information has a computational cost: SNP maps require 100 times as many iterations as microsatellites to produce stable estimates.

Background

Traditionally, the mainstay of linkage has been use of highly polymorphic microsatellite markers. The ultimate goal would be completely polymorphic markers – each parent would have two uniquely occurring alleles. A highly polymorphic microsatellite provides a great deal of segregation information at a particular locus. At the other extreme, single-nucleotide polymorphisms (SNPs) usually have only two alleles (more alleles are possible but uncommon) and alone provide much less information for segregation. Because SNP typing is less expensive, and available at a finer density than microsatellites, the use of dense SNPs in the place of microsatellites for linkage analysis is being investigated using data from the Collaborative Study of the Genetics of Alcoholism (COGA). Because segregation is at the heart of any linkage analysis, we examined IBD (identity by descent) matrices to compare the information content of SNPs versus microsatellites for linkage. We used the LOKI software [1,2] to create the matrices after a set of preliminary tests to determine the appropriate number of iterations. However, due to time and computational constraints, we have restricted our attention to chromosome 7. Although the matrices are created irrespective of phenotype, published results from the COGA project have shown linkage with multiple phenotypes on this chromosome [3]. Other members of our group used the matrices to replicate some of these findings [4].

Methods

Sample

Approximately 1,300 individuals previously typed with a microsatellite screen were typed with 4,720 SNPs from the Illumina linkage panel and 11,120 SNPs from the Affyme-

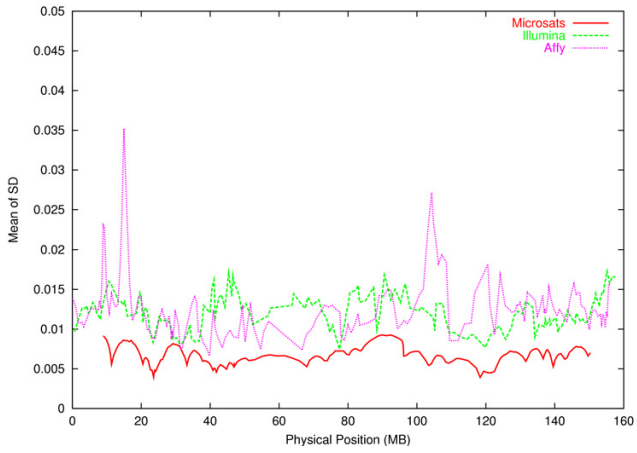

Figure 1
Mean of standard deviation of 2ϕ (10,000 iterations).

trix mapping array. These individuals were from 143 pedigrees with an average of 9.5 individuals typed per pedigree (range: 5–27). The sample was 77% White, 13% African American, and 10% from other ethnicities.

Analyses

In the presence of non-genotyped founders, allele frequency estimates are of paramount importance for IBD estimation. The large differences in allele frequencies between Whites and African Americans for microsatellites have been well established. Our group identified similar differences for allele frequencies of SNPs [5]. Because of this all of our analyses were restricted to 112 of the 143 pedigrees in which the entire pedigree was unambiguously White (as determined by self report and STRUCTURE) [5]. These pedigrees had an average of 9.3 individuals typed per pedigree (range: 5–27). Maximum likelihood estimators (MLE) for allele frequencies for both SNPs and microsatellites were computed using the 'freq mle' option in SOLAR [6].

The generation of IBD matrices with microsatellite markers has been addressed through a number of different techniques. We are using multipoint IBD (mIDB) as our standard because of the low information provided by a single SNP. Although a new version of GENEHUNTER [7] has recently been released to deal with SNP markers, the large size of some of these pedigrees required extensive trimming because the basic algorithm still computes all possible inheritance vectors. We therefore decided to use the LOKI software to generate IBD matrices.

LOKI uses Markov chain Monte Carlo (MCMC) methodology to repeatedly sample possible segregation patterns. However, determining appropriate run length (number of iterations) is important for the accuracy of the IBD estimates. Using three separate sets of markers: (microsatellites, Affymetrix, and Illumina) we compared the average standard deviation of "phi2" (twice the kinship coefficient) between each pair of individuals in each pedigree at each centimorgan position on chromosome 7 for 10 replicates (from 10 different starting seeds) with 10,000, 100,000 and 1,000,000 iterations per replicate. To compare this information on different maps, we translated the genetic positions of the markers to the physical position based on NCBI build 34.3 (Figures 1 and 2). Ultimately, we used 1,000,000 iterations to compute IBD estimates for each White pedigree for the SNP maps. These computations were performed on a Beowulf-class computer cluster consisting of 60 dual processor nodes (25 dual Pentium II 350 MHz, 8 dual Pentium II 550 MHz, 18 dual Pentium III 800 MHz, 9 dual Pentium III 1,000 MHz), each with 512 MB of RAM; this provides an effective 18 GFLOP/s capacity (based on the Linpack benchmark [8]).

Using the resulting matrices, information was first computed on sibships (regardless of phenotype) using the method presented in Kruglyak and Lander [9]: for N sibling pairs (i, j) at position x, we compute

$$I(x) = 1 - \frac{1}{N}\sum_{i,j}\frac{\sigma^2_{i,j}(x)}{\sigma^2_{i,j}}.$$

At each chromosome position for each sibling pair, the variance in IBD 0, IBD 1, and IBD 2 estimates is divided by the variance in the absence of marker information (for siblings, this is 0.5). The mean of this measure is subtracted from 1. If the posterior IBD status is known with certainty for all pairs, the variance is 0 and the information is 1. This measure was then computed on all relative pairs except for parent-offspring, where the prior variance is 0 (since Pr(IBD = 1) = 1.0), notwithstanding a new mutation. The results for the sibling pairs are presented in Figure 3.

Finally, we note that LOKI assumes the markers are in linkage equilibrium (LD). However, substantial pair-wise LD exists between the SNPs, especially in the Affymetrix map. Considering all pairs of Affymetrix SNPs on chromosome 7, 59 pairs have a correlation greater than 0.9. For Illumina, 13 pairs have a correlation greater than 0.9. Recent work [10] suggests that when parental genotypes are unavailable, this linkage disequilibrium between SNPs can artificially inflate IBD estimates and may also inflate estimates of information content because the inflated IBD estimate has an artificially high precision. Although many of the COGA pedigrees have parental genotypes, we conjectured that a less dense SNP map might still contain most of the information. The Illumina map began with lower LD so we chose to reduce it (rather than

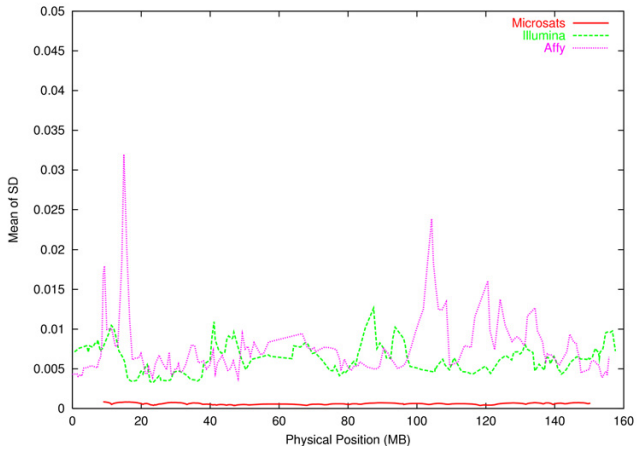

Figure 2
Mean of standard deviation of 2ϕ (1,000,000 iterations).

the Affymetrix map) because fewer deletions would be required. We constructed a subset of 166 markers from the Illumina dataset on chromosome 7. Markers were deleted from the dataset if they had a D' value greater than 0.1 with nearby markers. When possible, we retained markers with the highest possible minor allele frequency. The results are presented in Figure 3.

Results

Due to the large size of some of these pedigrees, software which computes all possible inheritance vectors (such as GENEHUNTER or Merlin [11]) has excessive memory requirements unless the pedigrees are pruned. Because all genotyped individuals can be used for quantitative trait analysis, this was deemed unacceptable.

In general, MCMC software trades these memory requirements for substantially greater CPU usage. We thus chose LOKI for IBD matrix generation, but this software requires a choice of run length (number of iterations). The initial tests to determine an appropriate number of iterations for the generation of IBD matrices with LOKI shows that the higher density of the SNP maps greatly slows the MCMC process. In particular, while 10,000 iterations for microsatellites shows only slightly higher average variance of phi2 than one 100,000 iterations, the variance for SNPs is still quite high with 100,000 iterations. We ultimately used 1,000,000 iterations for the SNP datasets. Figures 1 and 2 show that 1,000,000 iterations for the SNP map produces about the same variance of phi2 as 10,000 iterations of the microsatellite map. However, there are still peaks of high variance, especially in the Affymetrix map.

The information content results for microsatellites show substantial dips between markers. Both the Affymetrix and Illumina map provide a higher and much more uniform level of information. There are several sharp dips in the Affymetrix map, but this may be due to the IBD generation failing to converge; the locations of reduced information correspond to the locations of higher variance in Figures 1 and 2. Although the Affymetrix map has more than twice as many SNPs as the Illumina map, information is not significantly higher ($p = 0.3$). We also note that the "sparse" Illumina map (with an average intermarker spacing of 1.1 cM) contains substantially more information for sibships than the microsatellites and nearly as much information as the full Illumina map (with an average intermarker spacing of 0.69 cM). The information for relative pairs is uniformly lower than sibling pairs for all four map sets (results not shown). This may be due to the greater number of missing founders when considering the extended pedigrees as opposed to sibships.

Discussion and Conclusions

Our results show that the information provided by dense SNP maps is generally higher and more uniformly distributed than with standard microsatellite panels composed of about 400 markers. This increased information comes at a cost of increased computational complexity. At least 100 times as many iterations are required and each iteration took 10–20 times longer for the SNP maps as for the microsatellite. For example, 100,000 iterations took 3.4 hours for the microsatellites, 30 hours for the Illumina SNPs, and 68 hours for the Affymetrix SNPs. While the increased time for each iteration is likely due to the increased number of markers, the increase in required iterations may be due to the reduced information of the SNP markers. This could be tested by comparing convergence with a dense microsatellite map.

The Affymetrix map contains regions of reduced information, corresponding to the same locations where variance of phi2 is high. We examined the Affymetrix SNPs around the largest peak (at 15 Mb) and found that they were not significantly different from other Affymetrix SNPs on chromosome 7 in terms of density, heterozygosity, LD, or missing data rate. Other possible explanations include an increase in Mendelian-compatible genotyping errors or incorrect maps (either spacing or marker order). These possibilities could be tested in additional datasets to see if convergence problems existed at the same location.

Although the Affymetrix map consists of more than twice as many SNPs as the Illumina map, increased density of SNPs in the Affymetrix map does not appear to provide more information. However, many of the SNPs in the Affymetrix map have fairly low heterozygosity [5]. We also observed that a subset of the Illumina map provided nearly as much information as the full map. Although the best solution for markers in LD is probably to modify existing software to haplotypic information, it appears

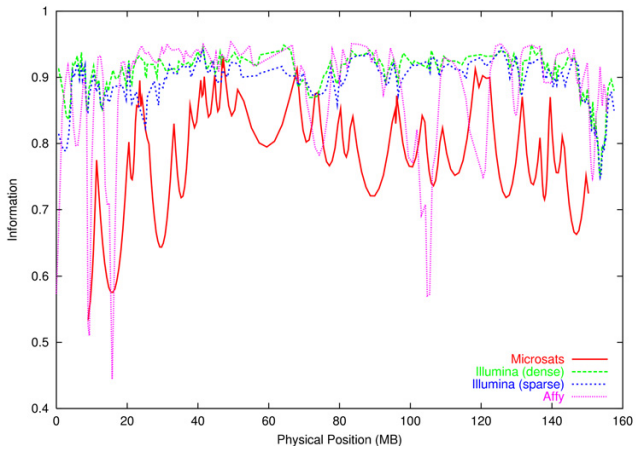

Figure 3
Information for sibling pairs.

that simply removing SNPs may be a useful interim procedure.

These data suggest that SNPs are a cost effective and informative replacement for microsatellites for linkage analysis. Although the computational burden is substantially greater for IBD computations, the resulting information is higher and more uniform. Although estimates of IBD and information content may be elevated when markers are in linkage disequilibrium and parents are untyped, further tests also suggest that a less dense map would provide nearly the same level of information.

Abbreviations

COGA: Collaborative Study on the Genetics of Alcoholism

IBD: Identity by descent

LD: Linkage disequilibrium

MCMC: Markov chain Monte Carlo

MLE: Maximum likelihood estimators

SNP: Single-nucleotide polymorphism

Authors' contributions

ALH formatted the data files, analyzed the data, and drafted the manuscript. JSKK provided analyses identifying White subgroups. SB, LJB, GD, CHJ, and BKS participated in the design and coordination of the study. All authors read and approved the final manuscript.

References

1. Heath SC: **Markov chain Monte Carlo segregation and linkage analysis for oligogenic models.** *Am J Hum Genet* 1997, **61:**748-760.
2. Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM: **MCMC segregation and linkage analysis.** *Genet Epidemiol* 1997, **14:**1011-1015.
3. Foroud T, Edenberg HJ, Goate A, Rice J, Flury L, Koller DL, Bierut LJ, Conneally PM, Nurnberger JI, Bucholz KK, Li TK, Hesselbrock V, Crowe R, Schuckit M, Porjesz B, Begleiter H, Reich T: **Alcoholism susceptibility loci: confirmation studies in a replicate sample and further mapping.** *Alcohol Clin Exp Res* 2000, **24:**933-45.
4. Dunn G, Hinrichs AL, Bertelsen S, Jin CH, Kauwe JSK, Suarez B, Bierut LJ: **Microsatellites versus single-nucleotide polymorphisms in linkage analysis for qualitative and quantitative measures.** *BMC Genet* 2005, **6(Suppl 1):**S122.
5. Kauwe JSK, Bertelsen S, Bierut LJ, Dunn G, Hinrichs AL, Jin CH, Suarez BK: **The efficacy of short tandem repeat polymorphisms versus single-nucleotide polymorphisms for resolving population structure.** *BMC Genet* 2005, **6(Suppl 1):**S84.
6. Almasy L, Blangero J: **Multipoint quantitative trait linkage analysis in general pedigrees.** *Am J Hum Genet* 1998, **62:**1198-1211.
7. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: **Parametric and nonparametric linkage analysis: a unified multipoint approach.** *Am J Hum Genet* 1996, **58:**1347-1363.
8. Jack Dongarra: **Linear algebra libraries for high-performance computers: a personal perspective.** *IEEE Parallel Distributed Technology* 1993, **1:**17-24.
9. Kruglyak L, Lander ES: **Complete multipoint sib-pair analysis of qualitative and quantitative traits.** *Am J Hum Genet* 1995, **57:**439-454.
10. Huang Q, Shete S, Swartz M, Amos CI: **Examining the effect of linkage disequilibrium on multipoint linkage analysis.** *BMC Genet* 2005, **6(Suppl 1):**S83.
11. 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.