Genome-wide Linkage Analysis with Clustered SNP Markers

KAJA K. SELMER,[1,2] KRISTIN BRANDAL,[1] OLE K. OLSTAD,[3]
BÅRD BIRKENES,[1] DAG E. UNDLIEN,[1,2] and THORE EGELAND[2,4]

Single nucleotide polymorphisms (SNPs) have recently replaced microsatellites as the genetic markers of choice in linkage analysis, primarily because they are more abundant and the genotypes more amenable for automatic calling. One of the most recently launched linkage mapping sets (LMS) is the Applied Biosystems Human LMS 4K, which is a genome-wide linkage set based on the SNPlex™ technology and the use of clustered SNPs. In this article the authors report on their experience with this set and the associated genotyping software GeneMapper® version 4.0, which they have used for linkage analyses in 17 moderate to large families with assumed monogenic disease. For comparison of methods, they also performed a genome-wide linkage analysis in 1 of the 17 families using the Affymetrix GeneChip® Human Mapping 10K 2.0 array. The conclusion is that both methods performed technically well, with high call rates and comparable and low rates of Mendelian inconsistencies. However, genotyping is less automated in GeneMapper® version 4.0 than in the Affymetrix software and thus more time consuming. (*Journal of Biomolecular Screening* 2009:92-96)

Key words: genome wide, linkage analysis, single nucleotide polymorphisms, large pedigrees, software

INTRODUCTION

TECHNICAL DEVELOPMENT HAS MADE AUTOMATED AND ACCURATE GENOTYPING of the abundant single nucleotide polymorphisms (SNPs) possible, and studies have shown that the use of SNPs in a greater number yields a better coverage of the genome, with just as high or higher information content than the microsatellites. As a consequence of the obvious advantages of using SNPs in linkage analysis,[1,2] several different SNP typing platforms, techniques, and associated kits and arrays especially designed for linkage analysis are now available. One of these kits is the newly launched Applied Biosystems Human Linkage Mapping Set 4K (LMS 4K; http://www.appliedbiosystems.com/). This is a kit based on the SNPlex™ technology,[3] a multiplex technology where up to 48 SNPs may be genotyped simultaneously in each sample. A new feature of this set, compared with other linkage mapping sets, is the marker map structure, where the SNPs are organized in clusters. Each cluster is to be handled as a composite marker to increase information content per locus compared with single SNPs and to lighten the computational burden when doing the statistical analyses on the nearly 4000 markers. As 1 of the first users of this kit, genotyping 17 moderate to large families with assumed monogenic disease, we report our experience, focusing particularly on practical issues concerning statistical analyses. For comparison, we also performed linkage analysis in 1 of the 17 families using the well-established Affymetrix GeneChip® Human Mapping 10K 2.0 array.[4]

[1]Institute of Medical Genetics, University of Oslo, Oslo, Norway.
[2]Department of Medical Genetics, Ullevål University Hospital, Oslo, Norway.
[3]Department of Clinical Chemistry, Ullevål University Hospital, Oslo, Norway.
[4]Oslo University College, Oslo, Norway.

Received Jul 9, 2008, and in revised form Sep 4, 2008.. Accepted for publication Sep 28, 2008..

Journal of Biomolecular Screening 14(1); 2009
DOI: 10.1177/1087057108327327

MATERIALS AND METHODS

Families

Of the 17 families included in this project, 15 have assumed autosomal dominant epilepsy, and 2 have autosomal recessive mental retardation (MR) syndromes. The 15 families with epilepsy were ascertained from a population-based Norwegian twin registry[5] and the 2 families with MR syndromes were identified in the pediatric ward, Ullevål University Hospital, Oslo. From these families, 315 individuals have been genotyped. The study was approved by the regional ethics committee, and all participants gave informed consent.

DNA samples

DNA was extracted from venous blood by standard means, using the DNA Extractor Model 340A from Applied Biosystems (Foster City, CA) based on phenol/chlorophorm extraction

(285 samples) or the MagNA Pure LC System (https://www.roche-applied-science.com/sis/automated/magna_lc/index.jsp; 30 samples).

Genotyping

Genotyping of all 315 samples was performed using the SNPlex™ technology,[3] according to the suppliers wet-DNA protocol (www.appliedbiosystems.com). This method uses locus-specific multiplexed oligonucleotide ligation followed by multiplex PCR target amplification with universal primers. Capillary electrophoresis was performed using an ABI 3730 DNA Analyzer.

All 10 members of 1 of the 17 families investigated were also genotyped using the Affymetrix GeneChip® Human Mapping 10K 2.0 array,[4] following the supplier's protocol. Washing and staining were performed with the Affymetrix Fluidics Station 450 and the signal intensities detected with the GeneChip Scanner 3000 7G.

The genotyping kits and arrays

There are 3922 SNP markers in the Applied Biosystems LMS 4K (https://products.appliedbiosystems.com), and 75% of these are localized in clusters of 2 to 4 SNPs within 200 kb. The mean distance between SNPs is 1.1 cM and between clusters 2.0 cM. Maximum intermarker distance is 10 cM. The SNPs are distributed across 95 pools; that is, 95 reactions are needed to genotype each sample for all 3922 SNPs. The supplier recommends an input of 74 ng of DNA in each reaction (http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042019.pdf); however, tests performed in our laboratory have shown that 120 ng was the optimal DNA amount for our samples, which varied in DNA quality. In total, 11.4 µg of DNA (7.0 µg according to Applied Biosystems' recommendations) is required for a whole genome scan for 1 individual.

The Affymetrix GeneChip® Human Mapping 10K 2.0 array comprises 10204 SNPs with an average intermarker distance of 0.31 cM or 210 Kb.[4] All SNPs are genotyped in the analysis of 1 array, requiring 250 ng of DNA per individual for a whole genome scan.

Genotype calling and data analysis

The genotype calling of the results from the Applied Biosystems LMS 4K was performed using the latest version of GeneMapper® Software, version 4.0. The results from the Affymetrix arrays were analyzed in GeneChip® Genotyping Analysis Software (GTYPE)[6] and GeneChip® Operating Software (GCOS). Data handling, Mendelian error control, and statistical analyses were done using Progeny Lab software (version 5; Progeny Software, LLC, South Bend, IN), PEDSTATS,[7] Merlin,[8] and MORGAN.[9]

Quality control

The Applied Biosystems LMS 4K kit provides positive hybridization controls, serving as quality control for the post-PCR steps, and a positive DNA control. For additional quality control, we also used CEPH 1347-02 (Centre d'Etude de Polymorphismes Humaines) as positive control and a nontemplate/negative control. Each of the 4 types of controls was dispensed in 8 wells of each 384-well plate, to ensure quality control for each injection of the 48 capillary electrophoresis.

Also the Affymetrix 10K set provides a DNA control sample. This was analyzed twice in parallel with patient samples, on separate arrays. These arrays served as control for the technical procedures and equipment, and could facilitate troubleshooting if results were poor. Affymetrix provides consensus genotypes for this control sample for the possibility of estimation of concordance rates.

Genotyping performance

Call rates, genotyping accuracy, and Mendelian error rates were used as measures of genotyping performance. For both the Applied Biosystems LMS 4K and the Affymetrix 10K array, call rates were calculated by dividing the number of produced genotypes on the maximum potential number of genotypes (SNPs × individuals). For the Applied Biosystems LMS 4K, SNPs with a call rate lower than 90% were omitted from analysis and from the call rate calculation. For the Affymetrix 10K arrays, arrays with a call rate below 90% were not accepted and the sample regenotyped. Genotype accuracy was measured in the Applied Biosystems LMS 4K by calculating concordance between the 8 genotype calls made for both the positive control and the CEPH control on each of the 95 plates. Concordance for the Affymetrix 10K arrays was measured by comparing genotype calls of the 2 positive controls with the reference genotypes provided by the supplier. Mendelian inconsistencies were detected in Progeny and Merlin,[8] and the error rate was calculated by dividing the total number of Mendelian inconsistencies on the total number of genotypes produced.

RESULTS AND DISCUSSION

Many high-throughput SNP genotyping systems have been evaluated and compared, both with other SNP genotyping methods[10] and with microsatellites.[2] In these comparisons, the technical performance measures such as call rates and genotyping error rates of the different SNP genotyping methods seem to be comparable. However, as pointed out by Wilcox et al.,[2] 1 of the main challenges in parametric linkage analysis in the future is the computational handling of the large amount of data produced from the analysis of large pedigrees. This was also the case in our project, and here we report our experiences with the newly launched applied Biosystems LMS 4K.

Table 1. Mendelian Error Rates and Call Rates

	Individuals	*Genotypes*	*Mendelian Errors*	*Mendelian Error Rate (%)*	*Call Rate (%)*
Applied Biosystems LMS 4K set	315	1,022,685	213	0.02	96.4
Subset	10	36,449	9	0.02	95.6
Affymetrix 10K	10	97,796	21	0.02	96.7

Technical performance

Having employed the SNPlex™ method for SNP genotyping for years, our lab was 1 of the first to try out the Applied Biosystems LMS 4K for this technology. For comparison, 1 of the families (10 samples) was also analyzed with the Affymetrix 10K array. Our results show that overall call rates are fairly high (**Table 1**) and do not diverge much from the results reported in the original articles presenting the methods.[4,11] The somewhat lower call rate in the Affymetrix 10K array samples in our study (96.7% vs. 98.6% in the original article) might be due to our limited experience with the method or to differences in DNA quality. In fact, we could observe increasingly higher call rates when comparing the first samples with the last samples analyzed (from 91.1% on the first array, to 99.0% on the last), which could reflect a "training effect," because these procedures were new to the technician. To see whether differences in DNA quality of the sample sets were affecting our results, the Applied Biosystems LMS 4K results of the 10 samples were drawn out as a subset. The call rate of this subset is somewhat lower than for the whole sample and we might speculate that the call rate for the Affymetrix 10K arrays is somewhat underestimated, due to a lower DNA quality of these 10 samples.

The Mendelian error rates are low in both sets (**Table 1**), indicating low genotyping error rates, which is also supported by the high concordance rates (**Table 2**) found in the positive controls genotyped.

Inasmuch as previous studies have shown that both coverage and performance of SNPs on chromosome 19 have been problematic with the Affymetrix 10K array,[12] we also compared chromosome 19 results in particular. Our results (**Table 3**) show that the call rate for this chromosome is nearly unchanged compared with the genomic call rate (**Table 1**) in the Applied Biosystems LMS 4K, whereas it is somewhat reduced in the results from the Affymetrix 10K array. Thus, the Applied Biosystems LMS 4K has improved the relative performance of chromosome 19 SNPs; however, the total number of SNPs with a call rate >75% is still higher in the Affymetrix 10K array.

Associated software

Genotype calling was performed by GeneMapper® version 4.0 in the Applied Biosystems LMS 4K and with GCOS and GTYPE in the Affymetrix 10K arrays. The results (**Table 1**) show that both algorithms performed well. In GeneMapper® version 4.0, 2 algorithms are available, the "Rules Algorithm" and the "Model Algorithm," where the latter is recommended by the supplier when genotyping more than 40 samples (http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042040.pdf). We experienced that this algorithm failed to cluster, and hence automatically genotype, a large proportion of our samples, and we therefore felt the need for visual inspecting and manually calling all plates analyzed to lose as few genotypes as possible. This is not recommended by the supplier, but because controls of rates of Mendelian error and concordance remained at an acceptable level, we still chose to do this. This approach could bias our results to a higher call rate, but also to a higher Mendelian error rate and a lower concordance rate. The approach was time consuming; however, even the recommended approach in the GeneMapper® version 4.0 protocol requires visual inspection of different quality control steps, such as inspection of graphs and results of ladders, size standards, positive hybridization controls, assay quality controls, and SNP quality. All quality control steps have default values in the "Model Algorithm," but if any of them fail, one can change cutoffs and reanalyze. Some alternative cutoffs are recommended in the protocol, whereas

Table 2. Concordance Rates

	Controls	*Genotypes*	*Discrepancies*	*Concordance Rate (%)*
Applied Biosystems LMS 4K	16	61,808	23	99.96
Affymetrix 10K	2	20,184	7	99.97

Table 3. Results for Chromosome 19

	# of SNPs	*#SNPs with a Call Rate > 75%*	*Concordance Rate (%)*	*Call Rate (%)*
Applied Biosystems LMS 4K	130	125 (96.2%)	100	96.7
Affymetrix 10K	149	139 (93.3%)	100	95.9

SNPs, single nucleotide polymorphisms.

others have no guidelines. Because the SNPlex™ method is very dependent on DNA quality, the user may tailor the algorithm to the particular sample set being analyzed. Although flexible, the adjustment options make the analysis process prone to maladjustments making the overall genotyping results potentially less reliable, dependent on the experience and expertise of the particular user.

Another important feature of these programs is the data export option. Export of custom-made tables is available in both programs; however, in GTYPE direct export of files in formats compatible with the most common software in statistical genetical analyses such as Merlin, GENEHUNTER, and Haploview is also possible. This is both time and work saving, but, more importantly, it might prevent errors made from manually formatting files. Another useful option within GTYPE is the possibility to do Mendelian inheritance checks directly.

Other analyses, such as testing for copy number alterations or detecting loss of heterozygosity (LOH), may be done on the Affymetrix 10K results, using a separate software called the CNAT (Copy Number Analysis Tool). Software for similar analyses is not yet available for the Applied Biosystems LMS 4K; however, a study shows that small modifications in the design of probes might make this feature available also for this kit in the future (http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_042602.pdf).

Marker map and statistical analyses

The marker map structure of the Applied Biosystems LMS 4K is based on clusters of SNPs being considered as composite markers, where the SNP haplotypes of each cluster represent the composite marker alleles (http://marketing.appliedbiosystems.com/mk/get/SNP_PRODUCT_LITERATURE). The SNPs in a cluster are selected to yield a high haplotype heterozygosity, which should lead to low linkage disequilibrium (LD) between them.

Several articles have been positive to a marker panel design with clustered SNPs.[13,14] This map structure may potentially decrease the computational burden of analyzing large pedigrees by treating each cluster as a composite marker. In general there are several approaches to analyzing SNPs in clusters. The use of single-point analysis (regarding each cluster as a composite marker) is 1 possibility, although this will reduce the information extraction of this map considerably, because 25% of the SNPs are singletons and these would yield very low information with this approach. Another option is multipoint analysis, providing that the intermarker distances within a cluster are set to low, nonzero values. The results may then be analyzed with the Lander-Green algorithm, available in a number of different statistical programs such as GENEHUNTER, Allegro, and Merlin. However, this approach also includes pitfalls affecting results: Intra-cluster LD may inflate the logarithm of the odds (LOD) scores as the Lander-Green algorithm assumes linkage equilibrium, of particular importance when founders genotypes are missing.[14] Also, misspecification of the order of markers in clusters might bias the results.[15] Another possible approach is the use of multipoint analysis, where each cluster represents a composite marker and the different haplotypes of the cluster the marker alleles. Modeling of LD between SNPs in a cluster is available in Merlin[8] and Aladin,[16] and in both programs is done by estimating haplotype frequencies of the clusters, rather than allele frequencies of the separate markers. This method assumes no recombination between markers within a cluster and no LD between clusters. If recombination within a cluster occurs, this cluster will be treated as missing in the analysis. This could potentially reduce the power; however, the frequency of recombinations within clusters in a moderate size pedigree is expected to be negligible when the clustered markers are genetically close,[13] which is the case in the Applied Biosystems LMS 4K.

Because Merlin may only handle small to moderate size pedigrees, only 4 of the 17 families in our study could be analyzed in Merlin. These families were analyzed using parametric multipoint analysis, first by ignoring LD, then by modeling LD between SNPs in clusters. Modeling LD did not change the results much, because almost all founders were genotyped and hence the benefit of modeling the LD was negligible, as is consistent with results based on simulated data.[14] A comparison of genome-wide linkage analysis of the family genotyped by both methods shows that the results are fairly consistent between the 2 methods (**Fig. 1**). Both detect 2 linkage peaks, which coincide with the maximum estimated LOD score, in the same chromosomal regions. Haplotype analyses of the markers in the 2 linkage peaks limit the regions to a total of 31 Mb for the Applied Biosystems LMS 4K, and 22 Mb for the Affymetrix 10K. The denser map of the Affymetrix 10K array produces smaller intervals, which again alleviates the next step of fine mapping with an extended set of markers. For analysis of the 13 larger families, only a few programs, for example, Simwalk2, MORGAN, and Aladin, are able to handle their complexity. The latter has implemented the possibility of modeling LD between markers; however, the current version is only a test release and is not yet able to reconstruct haplotypes. Linkage analysis was therefore done in the more established program MORGAN. The analysis of many SNPs in large pedigrees is time consuming, and our approach to this problem was to remove markers in close proximity in the first analyzing stage, which reduces both information content and the power to detect linkage,[14] but makes the computation feasible.

CONCLUSIONS

In comparison with the well-established Affymetrix 10K array, the Applied Biosystems LMS 4K did technically well with comparable genotyping performance. However, the analyses are

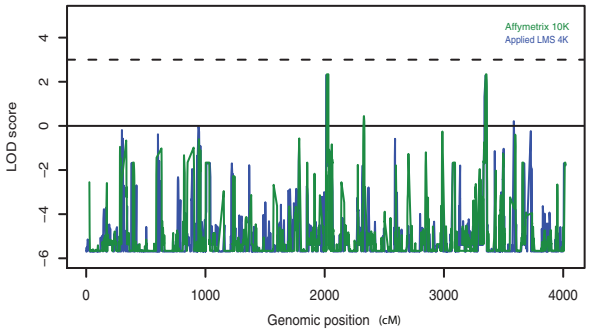

FIG. 1. Results from genome-wide linkage analysis of a family with autosomal recessive mental retardation: Affymetrix 10K in green and Applied LMS 4K in blue. Both sets identified 2 linkage peaks, which coincide with the maximum estimated LOD score of 2.3 in 2 different locations.

not as automated, and thus are more time consuming, in GeneMapper® version 4.0 as in the associated Affymetrix software GCOS and GTYPE. The Affymetrix software also provides more data export options and the possibility of doing Mendelian error checks directly.

The Applied Biosystems LMS 4K requires considerably more DNA per genome-wide screen per sample; however, both methods can also use whole genome-amplified DNA.

The clustered SNP map design yields more information per locus (considering each cluster as a locus) but has a lower resolution. It could potentially ease the computation in linkage analysis of many SNPs in large pedigrees, but appropriate and complete statistical software for the analyses is not yet available. As of today, the design of the Applied Biosystems LMS 4K is probably best suited for linkage analysis in small- to moderate-sized pedigrees, where the possibility of considering LD between markers may prevent inflated LOD scores in families where founders are not genotyped.

ACKNOWLEDGMENT

This study was financially supported by Eastern Norway Regional Health Authority.

REFERENCES

1. Kruglyak L: The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* 1997;17:21-24.
2. Wilcox MA, Pugh EW, Zhang H, Zhong X, Levinson DF, Kennedy GC, et al: Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation Groups 1, 2, and 3. *Genet Epidemiol* 2005;29(Suppl 1):S7-S28.
3. Tobler AR, Short S, Andersen MR, Paner TM, Briggs JC, Lambert SM, et al: The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech* 2005;16:398-406.
4. Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, et al: Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 2004;14:414-425.
5. Kjeldsen MJ, Corey LA, Solaas MH, Friis ML, Harris JR, Kyvik KO, et al: Genetic factors in seizures: a population-based study of 47,626 US, Norwegian and Danish twin pairs. *Twin Res Hum Genet* 2005;8:138-147.
6. Liu WM, Di X, Yang G, Matsuzaki H, Huang J, Mei R, et al: Algorithms for large-scale genotyping microarrays. *Bioinformatics* 2003;19:2397-2403.
7. Wigginton JE, Abecasis GR: PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 2005;21: 3445-3447.
8. 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.
9. George AW, Thompson EA: Discovering disease genes: multipoint linkage analysis via a new Markov Chain Monte Carlo approach. *Stat Sci* 2003;18:515-531.
10. Sawcer SJ, Maranian M, Singlehurst S, Yeo T, Compston A, Daly MJ, et al: Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Hum Mol Genet* 2004;13:1943-1949.
11. De la Vega F, Lazaruk KD, Rhodes MD, Wenz MH: Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP genotyping assays and the SNPlex genotyping system. *Mutat Res* 2005;573: 111-135.
12. Huentelman MJ, Craig DW, Shieh AD, Corneveaux JJ, Hu-Lince D, Pearson JV. et al: SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. *BMC Genomics* 2005;6:149.
13. Browning BL, Brashear DL, Butler AA, Cyr DD, Harris EC, Nelsen AJ, et al: Linkage analysis using single nucleotide polymorphisms. *Hum Hered* 2004;57:220-227.
14. Abecasis GR, Wigginton JE: Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet* 2005;77:754-767.
15. Daw EW, Thompson EA, Wijsman EM: Bias in multipoint linkage analysis arising from map misspecification. *Genet Epidemiol* 2000;19:366-380.
16. Albers CA, Stankovich J, Thomson R, Bahlo M, Kappen HJ: Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. *Am J Hum Genet* 2008;82:607-622.

Address correspondence to:
Kaja K. Selmer, M.D.
Dept. of Medical Genetics
Ullevål University Hospital
0407 Oslo, Norway

E-mail: k.k.selmer@medisin.uio.no