Familial heterogeneity in breast cancer predisposition: a study of 22 Utah families

Authors: Elizabeth O'Brien, Ph.D.[1], Richard A. Kerber, Ph.D.[1], Raymond L. White, Ph.D.[2]

[1] School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40292, USA

[2] Department of Neurology, University of California San Francisco, San Francisco, CA.

Corresponding Author:

Elizabeth O'Brien, Ph.D.

School of Public Health and Information Sciences

University of Louisville

Louisville, KY 40292

liz.obrien@louisville.edu

Keywords: missing heritability, breast cancer, linkage analysis, Utah

Abstract.

The problem of "missing heritability" in genome-wide analyses of complex diseases is thought to be attributable to some combination of: rare variants of moderate to large effect, common variants of very small effect, and epigenetic, epistatic, or shared environmental effects. Rare variants do not affect large numbers of people by definition, but identified genes and pathways frequently lead to important insights into pathogenesis, and become targets of chemoprevention or therapy. Family studies remain an efficient way to identify rare variants with sizable effects on disease risk. We present a genome-wide study of breast cancer in 22 large high-risk families including 154 women diagnosed with breast cancer. Appropriate marker spacing was achieved by simulation studies of founder haplotypes to reduce the chance that linkage disequilibrium produced spurious linkage peaks. For each family, we generated 100 simulations of null linkage genome-wide to estimate the probability that individual results were due to chance. We identified a total of 12 putative susceptibility regions with per-family genome-wide probability < 0.05. These regions were located on 10 chromosomes; 10 of the 22 families showed linkage at these locations; two or more families showed linkage to 6 regions on 5 chromosomes (4q, 5q, 6p, 14q, 18p, and 18q). These results indicate that there is considerable heterogeneity among families in genomic regions and thus variants predisposing to breast cancer. Moreover, they suggest that uncommon high- or medium-risk genetic variants remain to be found, and that family designs can be an efficient way to identify them.

Introduction.

The genetic dynamics of complex traits have concerned population scientists for more than a century, but the quantity of data streaming from genomic studies in recent decades has drawn new focus to the prospect of identifying genes underlying complex phenotypes. Especially important targets for genetic characterization are human disease phenotypes that commonly plague us and frequently kill us such as cancer.

Long before genome-wide data were available for complex trait analysis, family studies were the workhorses used to study the genetic basis of cancer because case clusters were originally observed in families. Examination of familial clusters of neoplastic disease led to the identification of the tumor suppressor role of *TP53* [1] in Li-Fraumeni syndrome, retinoblastoma, and the role of the *FANC* gene complex in Fanconi anemia (FA). Family studies of breast cancer also provided the first plausible evidence that a few genes of at least moderate effect might account for excess risk and observed case aggregation in families. This result was established for *BRCA1* and *BRCA2* mutations in familial breast and ovarian cancers [2] [3], and as a result the two genes were dubbed “most important” for breast cancer predisposition in high risk families [4].

Although breast cancer is not the most common of FA’s neoplastic effects, it has been demonstrated fairly recently that the products of the *FANC* gene complex function in congress with BRCA1 and BRCA2 in DNA repair pathways and provisionally explains their concordant effects on breast cancer predisposition in some families [5]. Mutations in *PTEN* and *STK11* may also exhibit relatively high penetrance effects [6-9] while other genes, such as *ATM*, *CHK2*, and *PALB1*, also account for excess breast cancer risk in some families with somewhat lower penetrance[10, 11]; however, families segregating these other mutations are rarer, and thus account for less of the total genetic risk estimated for large and heterogeneous case series. In fact, no other genes as commonly mutated, or of such high penetrance as *BRCA1* and *BRCA2*, have been identified yet through family studies of breast cancer. Therefore, it has generally been concluded from numerous studies of familial cancer risk (breast and other) in multiple populations, that: 1) the same genes do not account for cancer incidence in all families with elevated risks of the same cancer; 2) the same genes are

not necessarily implicated in familial clusters and sporadic cases (without a family history), even in the same population; and 3) familial cancers are relatively rare, and thus do not account for more than approximately 25% of all cases in a population, or 20% of incident breast cancers [12]. For these reasons, much doubt has been expressed over the last decade that family studies had much future utility for resolving complex genotypes for diseases like breast cancer[reference?]. Instead, as genome-wide data rapidly became available, and with it an acute need for "high-throughput" analyses, the focus of research quickly shifted to simpler association study designs to measure genetic differences between phenotypic classes, such as cases and controls.

The genome-wide association study (GWAS) approach focuses on genotype-phenotype co-variation, usually for a densely distributed set of SNPs over the genome. Positive associations occur where genotype differences correspond to phenotype differences outside of what is expected under a null hypothesis, and their locations mark points in or near genetic variants that cause disease or contribute to its risk. Numerous GWAS have been done in search of genes that condition risk of breast cancer, and a list of genes and variants with modest effects on cancer risk has certainly developed as a result [13] [14]. However, the small fraction of breast cancers attributable to these relatively common but low-penetrance alleles suggests that a larger set of genetic factors, more of them reaching moderate effect, but occurring with low frequency in a population, might account for such common cancer phenotypes.

This "heritability gap" has been considered a problem of statistical lack of power to resolve a potentially large number of genetic variants, some of them low in frequency (rare), and of only moderate or low risk effect for common but deadly disease phenotypes.

For complex diseases in general, GWAS have generated many significant associations between particular SNPs and disease phenotypes, but again, these are often inconsistent across studies, populations, designs, and samples. After more than a decade of modeling and measuring complex genotype-phenotype associations by GWAS, it remains difficult to value the contributed effects of particular genes to a disease phenotype by this method, and today it is widely appreciated that the approach has a critical shortcoming. For many individual studies the methodology is

simply underpowered to sort high throughput data for a definitive set of genetic factors —unknown in number and varying in frequency and effect size--responsible for a complex disease phenotype. As a result, there is much uncertainty about what an association study really captures, and clarification is often sought by improving power and reliability—by increasing genome coverage, sample size, or by meta-analyses. In this regard, today's study designs are ambitious, involving huge numbers of cases and increasingly narrow definitions of the phenotype[15]. Even so, GWAS of breast cancer have not resolved single genes of major effect comparable to *BRCA1* and *BRCA2*; neither have they established a comprehensive predisposing genotype for the disease.

Although it is now considerably easier and less expensive to collect genetic data for GWAS, it has remained elusive by association testing to capture enough genetic variants, or of sufficient effect, to account for what is manifestly familial and estimated as heritable. In this study we address the notion of "missing heritability" and compromised analytic power for detecting genetic factors contributive to breast cancer. In order to do this we have fashioned a "high-definition" approach to linkage analysis using deep pedigree data, albeit sparsely genotyped, and for pairs related over a range of relationships. The approach is not designed primarily to address the matter of heritability; more importantly, it is designed to advance the train of evidence leading to the identification of genetic variants that are potentially rare—i.e., found at low population frequency—of moderate effect on risk, and likely larger in number than the class of single genes of major effect, such as *BRCA1*.

Subjects and Methods.

Study Sample: breast cancer cases from high risk families in Utah

The Utah Population Database (UPDB) is a repository of longitudinal information originally constructed from genealogical data pertaining to Utahans and their families [16]. Through successive record linking efforts, the database integrates cancer registry data, medical records data, Utah State certified deaths and births, etc. Currently, the UPDB

captures information for approximately 7 million individuals, many of whom are members of extensive pedigree networks 2 to 14 generations deep [17]. Pedigree information from the UPDB, and Utah's SEER cancer registry, were used to establish diagnosed breast cancer cases clustered in large multigenerational families. We then compared observed and expected breast cancer incidence in case families and recruited study subjects from "high risk" families, i.e., those with excess incidence having a probability of less than 0.01 of occurring by chance [18]. However, we excluded cases and families previously studied and known to be segregating *BRCA1* or *BRCA2* mutations as their primary genetic risk factors for breast cancer.

Female members of high risk families who were diagnosed with breast cancer and alive at the start of the study were invited to join, as were unaffected women drawn from the same large families. Study participants were home visited, at which time individual and family health histories were documented and blood samples collected (by venous puncture) as the source of DNA for genome- wide SNP analyses.

The genotyped study sample consisted of 154 women diagnosed with breast cancer, and 94 unaffected relatives. "Families" were defined after recruitment as the largest set of genotyped subjects, including a minimum of 3 cases, all descended from a common ancestor. By this method all participants (n=248) are members of 22 large families with evident excess risk of breast cancer. Cases (n=154) collectively form 1,011 affected relative (AA) pairs for linkage analysis; genotypes of unaffected subjects (94) were used to estimate allele frequencies and identity by descent probabilities. The families included in this study are pictured schematically in Figure 1.

The University of Utah Health Sciences Institutional Review Board and the University of Louisville Biomedical Institutional Review Board approved the study protocol; all recruited subjects provided their written consent to be included in this study.

Genotypes

Genotyping was performed with Illumina 370 Duo and 610 Quad arrays at deCODE Genetics, Reykjavik, Iceland. SNPs with low quality scores (GenCall[19] quality score < 0.15), and those with inconsistent allele frequencies between the two arrays (any absolute difference in minor allele frequency > 0.05), were eliminated. All alleles were called on the forward strand, and checked for consistency between arrays. After approximately 15% of the SNPs were eliminated by these quality control criteria, a total of 285,630 genotypes per subject were retained. Mendelian consistency checks were not performed because of the very small number of families with informative data.

Evaluation of genetic vs. genealogical relatedness

We examined the degree to which relatedness assessed by genome-wide genetic similarity corresponded to relatedness as reported in the UPDB genealogical data for pairs of relatives. For this study, we used genotypes on 429 individuals, including the 248 subjects in the linkage study, as well as 181 women from families with fewer than 3 genotyped breast cancer cases. A total of 91,806 pairs were evaluated, using coefficient of relatedness to characterize the genealogical data, and the genetic relatedness matrix computed by GCTA[20] to characterize relatedness from SNP data. To facilitate comparison, relatedness from each measure was grouped by rounding $-\log_2$(relatedness) to correspond to degree of relationship.

Identity by Descent (IBD) estimation for linkage analysis

Pairs formed from the sample set were used to generate Identity by Descent (IBD) matrices for linkage analysis. IBD was computed using PEDIBD software developed by Li and colleagues[21]. Their method employs a Viterbi algorithm [22] to find the most likely path of descent of an ancestral allele through a deep, but sparsely genotyped pedigree structure, via hidden Markov models of inheritance and recombination. The method efficiently parses the high-density genotype data of the Illumina arrays, permitting estimation of IBD matrices for 1,011 affected relative pairs at up to 285,630 loci in approximately 24 hours of CPU time on current equipment (substantially less for thinned data sets). Allele frequencies were estimated by simple counting among all genotyped individuals, affected or unaffected. As

noted by Boehnke[23] and others, simple counting among family members does not introduce any systematic bias in the absence of allelic association, and any association would introduce a conservative bias as it would lead to overestimation of the frequency of a disease-associated allele.

Test statistic for linkage

We employed the IBDREG quasi-likelihood approach described by Schaid, et al. [24] to test for concordant pair (affected only) linkage without covariates. IBDREG has an important advantage in comparison to competing methods as it appropriately adjusts for between-pair covariance when multiple relative pairs are drawn from the same pedigree structure. Because the families studied vary considerably in size, and some have only a few affected members, the distributional properties (and hence the asymptotic p-values) of the test statistic were uncertain. Therefore, we used simulation to compute p-values and family-wise error rates. The approach is described below.

Simulation of Identity by Descent in the Absence of Linkage, but the Presence of Linkage DisequilibriumWe performed 100 full-genome simulations of identity by descent using all 285,630 autosomal markers and all 22 families for three reasons: 1) to allow accurate estimation of error rates for IBD estimates across all family structures and all autosomes; 2) to give a reference against which different thinning strategies could be evaluated for their effects on both IBD accuracy and the distribution of the linkage test statistic; and 3) to provide distributions of the test statistic under the null hypothesis.

Estimation of error rates

It is well known that linkage analysis based on high-density SNP arrays is subject to potentially severe bias away from the null because of linkage disequilibrium (LD). LD between nearby markers will cause overestimation of the probability that two related individuals share marker alleles that are identical by descent (IBD) [25, 26]. Although in

principle, simultaneous modeling of LD among founders and IBD among descendants would be the most powerful approach to using all the genotype data at our disposal [27], the computational burden of such modeling in complex multigenerational families is not readily surmountable at present.

Marker thinning intervals

Marker thinning effectively varies the strength of LD by setting maximum R^2 between SNPs at various thresholds (0.6, 0.4, and0.2 here). At each threshold SNPs were thinned by recursively finding the midpoint of a block of SNPs mutually correlated at R^2 > the current threshold, then dropping all but the midpoint SNP, so that the maximum pairwise correlation could not exceed the selected level. Thinned marker sets were run against simulated (null) genotype data for chromosome 7 to establish error rates in the IBD estimates and thus, the contribution to false positive linkage scores for varying strengths of LD structure.

For our simulations and analyses, we imputed an LD structure descending from founders by adapting the HapMap3, Phase 2 observed LD structure for 234 independent haplotypes estimated from 117 CEU subjects [28]. The HapMap sample series is appropriate as a reference set for this study because it too is a Utah family series [29]. HAPGEN2 software [30] was used to generate 4000 random haplotypes with the desired LD characteristics for all 285,630 autosomal loci. For each pedigree founder, two random haplotypes were chosen, from the 4000 randomly generated, by sampling with replacement. Alleles for each SNP marker were randomly generated in proportion to each marker's allele frequencies. Haplotypes were descended through the study pedigrees, resetting random segregation indicators according to HapMap's estimated recombination fractions. Recombination between markers was estimated by cubic spline interpolation using R [31].

Pedigree information and simulated marker data were input to PEDIBD to obtain a full matrix of IBD estimates for all affected pairs. IBD estimates generated by PEDIBD were compared to the simulated "true" IBD states (0, 1, or 2 alleles known to be shared for each pair) to determine error rates.

Distribution of test statistics under the null

The IBD estimates generated by PEDIBD were input to IBDREG to calculate linkage statisticsAll simulated IBD states and marker allele data were generated under the null hypothesis of no linkage between marker loci and disease predisposition. Thus, the distribution of test statistics for each marker locus within each family can be taken to represent a sample from the null distribution for a whole genome scan of that family. In addition to the per-locus asymptotic p-value computed by IBDREG, we report a family-specific per-locus Monte Carlo p-value, a family-specific per-genome Monte Carlo p-value, and a Monte Carlo composite false discovery rate (FDR) controlling for the whole-genome analysis of 22 families[32].

Identification of linkage peaks and boundaries

We defined a putative linkage peak as the chromosomal location of the smallest p-value over a run of consecutive SNPs with asymptotic p-values less than 0.001. The extent of the linked "peak" region was identified from the focal SNP (smallest p-value) to the nearest SNP either side with a p-value tenfold greater than the focal SNP, thus establishing the boundary maximum *p*. Overlapping peaks across multiple families were counted as a single peak.

Results.

An initial check for correspondence between coefficients of relatedness estimated from pedigree information and from SNP genotypes was made for all possible pairs of study subjects (see Methods). This information is plotted in **Figure 2** for pairs of related individuals**.** The most distantly related pairs in the genealogical data were 13th degree relatives, so pairs unrelated by genealogy and pairs estimated to be genetically more distant than 13th degree were

plotted as though they were 14th degree relatives on either scale. There was generally very good agreement between genealogical and genetic distance up to about the 6th degree, and a gradual loss of precision past that point in this population.

It is common that some members of large Utah families overlap in family membership in descending generations, and **Table 1** gives counts of these individuals. Note that most subjects are members of only one family, and the majority of those who overlap in family membership do so as pedigree members, rather than genotyped study subjects. This is shown in **Table 1.** where counts are given for the number of individuals with membership in >1 of the 22 families, by disease status. Counts of individuals and affected pairs by family are given in **Table 2**.

Simulations were done to depict the inflationary effect of LD on IBD and false positive linkage scores (see **Methods**). These results are shown in **Figure 3** and **Table 3**. In order to control for this effect, and reduce false positive linkage signal, SNPs were thinned to various thresholds of correlation between them. At the threshold $R^2 \leq 0.4$, IBD over-estimation due to LD was controlled fairly well, but positive linkage peaks still occurred. At $R^2 \leq 0.2$, spurious linkage peaks disappeared. The results given in **Figure 4** and **Table 4** are based on the inter-marker threshold $R^2 \leq 0.2$ for the thinned set of 19,609 SNPs.

Table 4 gives linkage results for 1,011 affected relative pairs generated from a total of 154 genotyped breast cancer cases. The analysis identified 19 distinct peaks with asymptotic unadjusted within-family $p < 0.001$. More realistic estimates of the probability of these results under the null hypothesis are derived from the 100 per-family genome-wide simulations, and presented in Table 4 as well. Monte Carlo per-locus p-values are generally considerably larger than the asymptotic p-values, particularly for smaller families. After further adjustment for genome-wide comparisons within families, 11 regions retained adjusted p-values below 0.05, and 17 regions retained adjusted p-values below 0.1. However, when we adjusted for simultaneous whole-genome search across all 22 families, only the 3 peaks with the highest scores were large enough that a single random result under the null would not have been expected to exceed them 100% of the time.

Supplementary Table 1 lists all breast cancer- associated genes from DisGeNET [http://www.disgenet.org/web/DisGeNET/v2.1], TCGA [39] and Cancer Resource[40] located within peaks defined by a 10-fold increase in asymptotic p-value. The large peak on chromosome 6 for family 1 includes multiple genes that have been associated with breast cancer risk and/or tumorigenesis, including members of the *HLA* complex, *NOTCH4*, and *TNF*, among others. Also noteworthy is that the chromosome 13 peak for family 10 includes *BRCA2*, while no family exhibited linkage to the *TP53* or *BRCA1* regions on chromosome 17. **Figure 4** shows the relative locations and amplitudes of the linkage peaks by family.

Discussion:

It is low-frequency variants that are difficult to find in convincing association with a disease phenotype from genome-wide association tests[41]. However, if we are to resolve this low frequency, moderate risk class of variants, then population-wide sampling from whole undifferentiated, or minimally structured populations, is perhaps not the most strategic sampling approach to use. Variants of this class occur de novo, are replicated and transmitted to descendants. For this reason, they will reach their highest frequencies within family lineages[42], the larger the better, while remaining at low frequency (rare) in any usual population sample, whether n = 100s or 100,000s. The moderate risk nature of this class of variants is likely due to the fact that their risk effects depend on participation in larger gene networks to account for increased cancer risk in particular families. In this sense, variants of smaller effect can alter disease risk in the context of gene networks that regulate the functional pathways involved in the onset and/or progression of the disease.

The family study approach does not rest on anticipating “a new breast cancer genotype”, nor a “comprehensive genotype” to account for breast cancer risk in this population and by the usual purview of linkage analysis. Instead, we tried to capture evidence of low frequency variants at the population level, but enriched at the level of very large high-risk families. Our approach yielded 17 genomic regions possibly linked (per-family per-genome Monte Carlo p $\leq$

0.1) to breast cancer for the 22 families studied, but with considerable variation among families: 15 of the 22 families (68%) showed possible linkage to at least 1 region by the criteria used here; 1 family showed possible linkage to 4 regions; 1 family to 3 regions; 5 families to 2 regions; and 8 families showed linkage to 1 region. It is noteworthy that linkage to the *BRCA2* region on chromosome 13 was observed in only one family (10), while no family exhibited linkage to the *BRCA1* region on chromosome 17.

The availability of high-density marker sets, efficient algorithms for estimating IBD in large families, and substantial computational resources permitted simulation of 100 null genome-wide results for each family. The simulation results then allowed us to compare Monte Carlo p-values with asymptotic p-values based on large sample theory. In Table 4 it is shown that the asymptotic estimates are frequently smaller than the Monte Carlo p-values by an order of magnitude or more. The genome-wide results for each family represent an appropriate basis for comparison to other published results based on linkage studies of one or a few families. Adjusting linkage estimates for all 22 families simultaneously, we find no linkage scores, or peaks, that could not have occurred by chance: the Z-score of 6.21 observed for family 1 on 6p21-22 was exceeded in 90% of null simulations—for at least one family at some location over the genome. However, in the simulated data, anomalously high Z-scores were much more commonly observed in small families—and never in family 1—so the adjustment across families is in some part size dependent, and therefore, less than perfect. For this reason, it might be more appropriate to consider the simulated probability of observing the result in 22 families just like family 1, which would be approximately 1-(1-0.0018)^22 = 0.039. Moreover, the existence of a possible linkage peak for family 5 in precisely the same location as family 1 on 6p21-22 strengthens the case for a susceptibility locus in this region.

For heuristic purposes, we can combine multiple lines of evidence to rank the various linked regions by priority. First, regions that overlap across multiple families (e.g. 6p22-21, families 5 and 1; 18p11, families 5 and 16; 18q21-22, families 20 and 21) likely indicate either a shared disease-predisposing haplotype inherited from an unknown common ancestor, or multiple predisposing variants in the same gene in truly unrelated, or variably related, families. Next, per-

family, per-genome Monte Carlo p-values well below 0.05 (e.g. 3p11-13, family 3; 12q21-24, family 22) warrant further investigation. Finally, linked regions in individual families (13q12-14, family 10) that overlap with known breast cancer-predisposing loci, i.e. BRCA2, have the potential to greatly simplify mapping variants associated with specific diseases. In addition, the other regions suggestive of linkage without known breast cancer associated variants, might provide useful new clues about the location of genetic variants that increase the risk of breast cancer in members of these families, and serve as evidence of residual heterogeneity in genomic regions responsible for familial cancer susceptibility.

Although this linkage analysis was meant to identify regions of the genome that include putative genetic contributors to disease, there is still considerable distance between regions identified by linkage, and discovery of whether or what variants within them contribute to breast cancer risk ("true positives"). Having identified segments of the genome smaller than the whole, there are still at least six to ten segments to consider, each spanning many genes, a large amount of information and a lot of variation. Depending upon the definition of peak region—whether 1Mb-5Mb surrounding the focal SNP, or the larger regions bounded by a tenfold change in p-value—many genes that have been associated with breast cancer risk in other studies are captured in the linkage regions (Table S1), including BRCA2. From functional annotation, the linked regions we have identified encompass many genes that "look good" as candidates for further analysis. However, in order to identify specific variants relevant to breast cancer phenotypes in this study, especially those that are rare and of obscure overall effect, it remains to further interrogate the linkage regions by sequencing. An efficient approach might begin with whole exome sequencing to address functional variants first. Regional, functional, family and pair-specific information can all be used to direct targeted evaluations of concordance between expected linkage (SNP-based probabilities of sharing IBD) generated by our model, and differences in sequence sharing per exome through the linked regions. By using the linkage-partitioned information thus far, sequencing should reveal more specifically the locations of rarer variants likely relevant to the disease. Linkage and sequencing techniques together should do much to clarify the genetic architecture of breast cancer in this

population, its heterogeneity among families[43], and importantly, give us a deeper understanding of the role of rare variants in conditioning risk differently among groups.

For any novel variants that might be established, or candidate genes that might be confirmed with sequencing, the hope is that the information will advance knowledge of the genetic pathways involved and their interacting factors so essential to personalized therapies, management, and outcomes in the clinical setting. The developing field of Molecular Epidemiology and its unique integrative approach to medical research only begins with the address of large and growing quantities of data for translation to improved risk prediction. Studies of this type, that inform us about what specific genomic variation underlies risk variation in a population, lead to the identification of risk subgroups, and most importantly, high-risk families and individuals. New and abundant genetic information will no doubt lead us to understand important features of how genes—common, rare, and in multiplicity—contribute to disease spectra, from well to mortal, and the intermediate.

Supplemental Table and Figures.

1. Figure S1. Manhattan plots for each family. Families are labeled as per Figure 1 in upper right corner of each plot.
2. Figure S2. Locations of genes within linkage peaks with unadjusted p-value < 0.001. Within peaks, cyan lines indicate genes, red lines indicate genes mutated in TCGA breast cancer specimens, black lines indicate boundaries of overlapping peaks. Coding strand is indicated by placement within box: genes coded on the forward strand are drawn above the midline, while genes coded on the reverse strand are drawn below.
3. Table S1. Table of all genes in linked regions, ordered by bioinformatics resource scores (see text for references): tcga.mut = number of mutations observed in TCGA breast tumors; cr = cancer resource breast cancer associated (1) vs not associated (0); dg = disGeNet breast cancer association score; sum = sum((tcga.mut > 0)+(cr > 0)+(dg > 0)); tcga = TCGA breast tumor mutations/bp.

Acknowledgments.

Data analysis and genotyping was supported by Susan G. Komen Foundation grant KG-071514. Partial support for all datasets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute and

326 the Huntsman Cancer Institute Cancer Center Support grant, P30-CA2014 from the National Cancer Institute. Support
327 for the Utah Cancer Registry is provided by Contract No HHSN2612013000171 from the National Cancer Institute with
328 additional support from the Utah Department of Health. We thank Debbie Shea, Matt Hoffman, Terah Young, and
329 Loren Budge for subject recruiting and data collection; Geri Mineau, Alison Fraser and Janice Conrads at the Huntsman
330 Cancer Institute's Pedigree and Population Resource for linking subject data to the Utah Population Database; Xin Li
331 for consultation on PEDIBD modifications for application to our data; Dan Schaid, Duncan Thomas, Steve Schrodi, and
332 Ken Boucher for valuable discussions of this work.

333

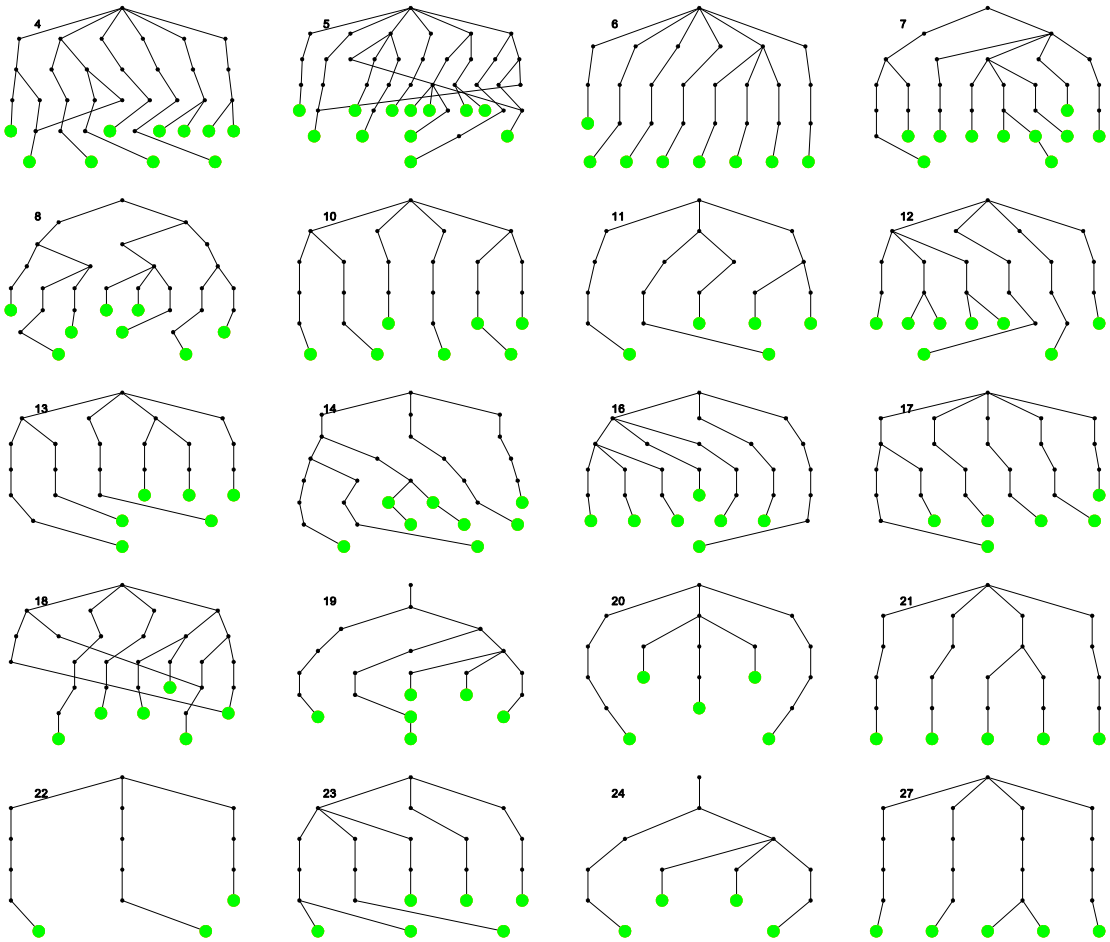
References

1. Malkin D, Li FP, Strong LC, Fraumeni J, Nelson CE, Kim DH, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science. 1990;250(4985):1233-8.
2. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990;250(4988):1684-9.
3. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266(5182):66-71.
4. Ford D, Easton D, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The American Journal of Human Genetics. 1998;62(3):676-89.
5. Seal S, Barfoot R, Jayatilake H, Smith P, Renwick A, Bascombe L, et al. Evaluation of Fanconi Anemia genes in familial breast cancer predisposition. Cancer research. 2003;63(24):8596-9.
6. Nieuwenhuis MH, Kets CM, Murphy-Ryan M, Yntema HG, Evans DG, Colas C, et al. Cancer risk and genotype–phenotype correlations in PTEN hamartoma tumor syndrome. Familial cancer. 2014;13(1):57-63.
7. Tan M-H, Mester JL, Ngeow J, Rybicki LA, Orloff MS, Eng C. Lifetime cancer risks in individuals with germline PTEN mutations. Clinical Cancer Research. 2012;18(2):400-7.
8. Riegert-Johnson DL, Gleeson FC, Roberts M, Tholen K, Youngborg L, Bullock M, et al. Research Cancer and Lhermitte-Duclos disease are common in Cowden syndrome patients. 2010.
9. Hearle N, Schumacher V, Menko FH, Olschwang S, Boardman LA, Gille JJ, et al. Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. Clin Cancer Res. 2006;12(10):3209-15. doi: 10.1158/1078-0432.CCR-06-0083. PubMed PMID: 16707622.
10. Tung N, Battelli C, Allen B, Kaldate R, Bhatnagar S, Bowles K, et al. Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. Cancer. 2015;121(1):25-33. doi: 10.1002/cncr.29010. PubMed PMID: 25186627.
11. Walsh T, King MC. Ten genes for inherited breast cancer. Cancer Cell. 2007;11(2):103-5. doi: 10.1016/j.ccr.2007.01.010. PubMed PMID: 17292821.
12. Kerber RA, O'Brien E. A cohort study of cancer risk in relation to family histories of cancer in the Utah population database. Cancer. 2005;103(9):1906-15. Epub 2005/03/22. doi: 10.1002/cncr.20989. PubMed PMID: 15779016.
13. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007;447(7148):1087-93.
14. Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, Newman B, et al. A genome wide linkage search for breast cancer susceptibility genes. Genes, Chromosomes and Cancer. 2006;45(7):646-55.
15. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013;45(4):353-61.
16. Skolnick M. The Utah Genealogical Database: a Resource for Genetic Epidemiology. In: Cairns JJ, Lyon JL, Skolnick M, editors. Cancer Incidence in Defined Populations. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1980. p. 285-96.
17. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol. 2003;21(3):113-6. PubMed PMID: 12628367.

18. Kerber RA. Method for calculating risk associated with family history of a disease. Genet Epidemiol. 1995;12(3):291-301. Epub 1995/01/01. doi: 10.1002/gepi.1370120306. PubMed PMID: 7557350.
19. Kermani BG. Artificial intelligence and global normalization methods for genotyping. Google Patents; 2006.
20. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011;88(1):76-82.
21. Li X, Yin X, Li J. Efficient identification of identical-by-descent status in pedigrees with many untyped individuals. Bioinformatics. 2010;26(12):i191-8. Epub 2010/06/10. doi: 10.1093/bioinformatics/btq222. PubMed PMID: 20529905; PubMed Central PMCID: PMC2881406.
22. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes: The Art of Scientific Computing. 3 ed. Cambridge: Cambridge University Press; 2007.
23. Boehnke M. Allele frequency estimation from data on relatives. American journal of human genetics. 1991;48(1):22.
24. Schaid DJ, Sinnwell JP, Thibodeau SN. Testing genetic linkage with relative pairs and covariates by quasi-likelihood score statistics. Human heredity. 2007;64(4):220-33. Epub 2007/06/15. doi: 10.1159/000103751. PubMed PMID: 17565225; PubMed Central PMCID: PMC2880728.
25. Huang Q, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. The American Journal of Human Genetics. 2004;75(6):1106-12.
26. Webb EL, Sellick GS, Houlston RS. SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal. Bioinformatics. 2005;21(13):3060-1.
27. Thomas A. Towards linkage analysis with markers in linkage disequilibrium by graphical modelling. Human heredity. 2007;64(1):16-26.
28. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8. doi: http://www.nature.com/nature/journal/v467/n7311/abs/nature09298.html#supplementary-information.
29. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics. 1990;6(3):575-7. PubMed PMID: 2184120.
30. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011;27(16):2304-5. doi: 10.1093/bioinformatics/btr341.
31. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.
33. Basu S, Di Y, Thompson E. Exact Trait-Model-Free Tests for Linkage Detection in Pedigrees. Annals of human genetics. 2008;72(5):676-82.
34. Thompson E, Basu S. Genome sharing in large pedigrees: multiple imputation of ibd for linkage detection. Human heredity. 2003;56(1-3):119-25.
35. Thompson EA. Statistical Inferences from Genetic Data on Pedigrees. Beachwood, OH: IMS; 2000.
36. Thompson EA. MCMC in the Analysis of Genetic Data on Pedigrees. In: Liang F, Wang J-S, Kendall W, editors. Lecture Note Series of the IMS, National University of Singapore. Singapore: World Scientific So Pte Ltd; 2005. p. 183-216.

37. McPeek MS. Optimal allele-sharing statistics for genetic mapping using affected relatives. Genetic epidemiology. 1999;16(3):225-49.
38. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PloS one. 2011;6(6):e20284.
39. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113-20.
40. Ahmed J, Meinel T, Dunkel M, Murgueitio MS, Adams R, Blasse C, et al. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. Nucleic acids research. 2011;39(suppl 1):D960-D7.
41. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53. doi: 10.1038/nature08494. PubMed PMID: 19812666; PubMed Central PMCID: PMC2831613.
42. Wijsman EM. The role of large pedigrees in an era of high-throughput sequencing. Human genetics. 2012;131(10):1555-63. doi: 10.1007/s00439-012-1190-2. PubMed PMID: 22714655; PubMed Central PMCID: PMC3638020.
43. Lynch H, Wen H, Kim YC, Snyder C, Kinarsky Y, Chen PX, et al. Can unknown predisposition in familial breast cancer be family-specific? Breast J. 2013;19(5):520-8. doi: 10.1111/tbj.12145. PubMed PMID: 23800003.

442 **Figures.**



443

444 **Figure 1. Schematic pedigrees of the 22 families studied.**

445 Affected subjects are indicated with enlarged green dots. Only lines of descent from common ancestors are
446 shown. Crossing lines indicate inbreeding, although only one affected subject was herself inbred (family 18).
447 Family numbers 2, 9, 15, 25 and 26 were assigned to families not used for this analysis, either because of overlap
448 with another family (2), or insufficient number of usable samples from breast cancer cases.

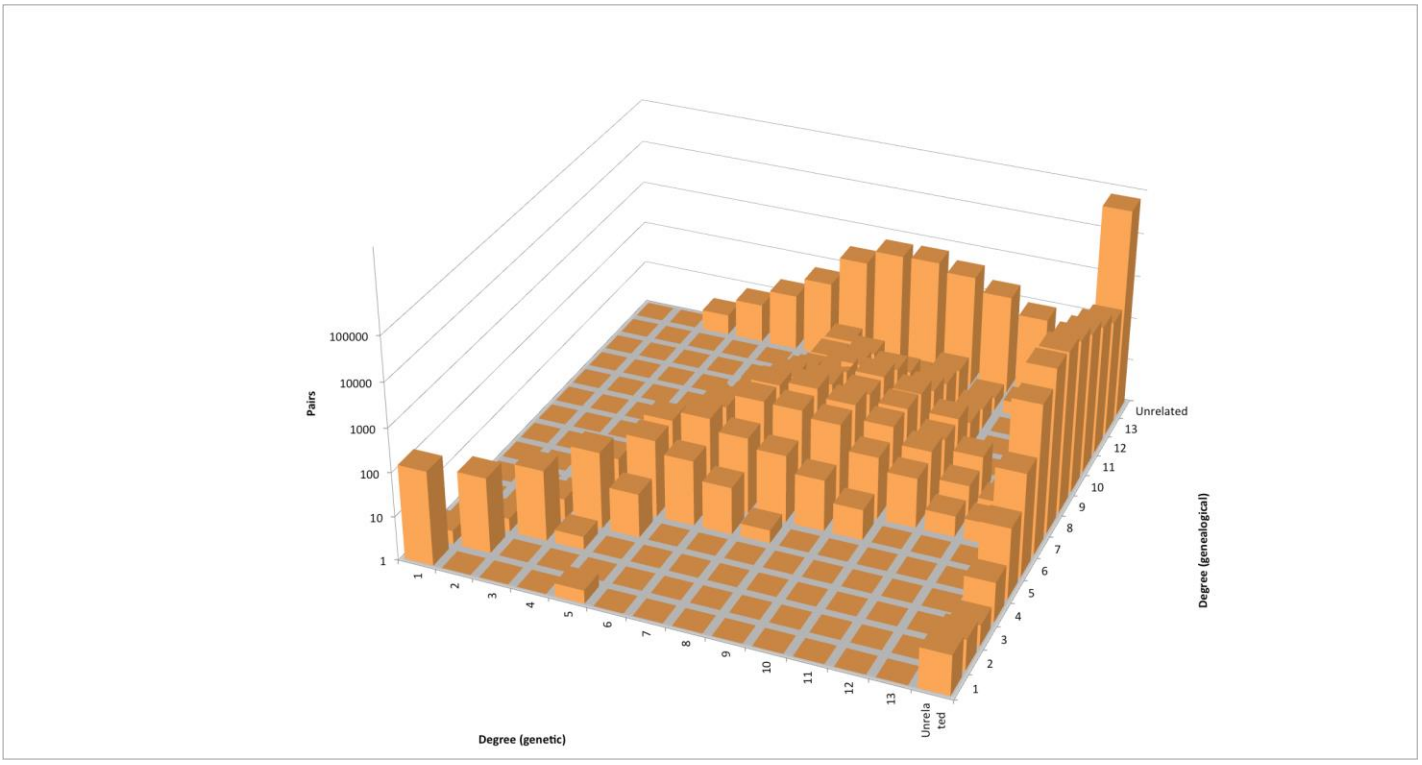


Figure 2. Genetic vs. genealogical relatedness.

Relatedness estimated as global IBD from genetic data (SNPs) compared to genealogical relatedness (from pedigrees) for all possible pairs of study subjects (affected and unaffected). Red dots indicate pairs with substantial mismatch between genealogical and genetic distances; these pairs were dropped from the analysis by inspection and removal of one or both subjects from pedigree data.

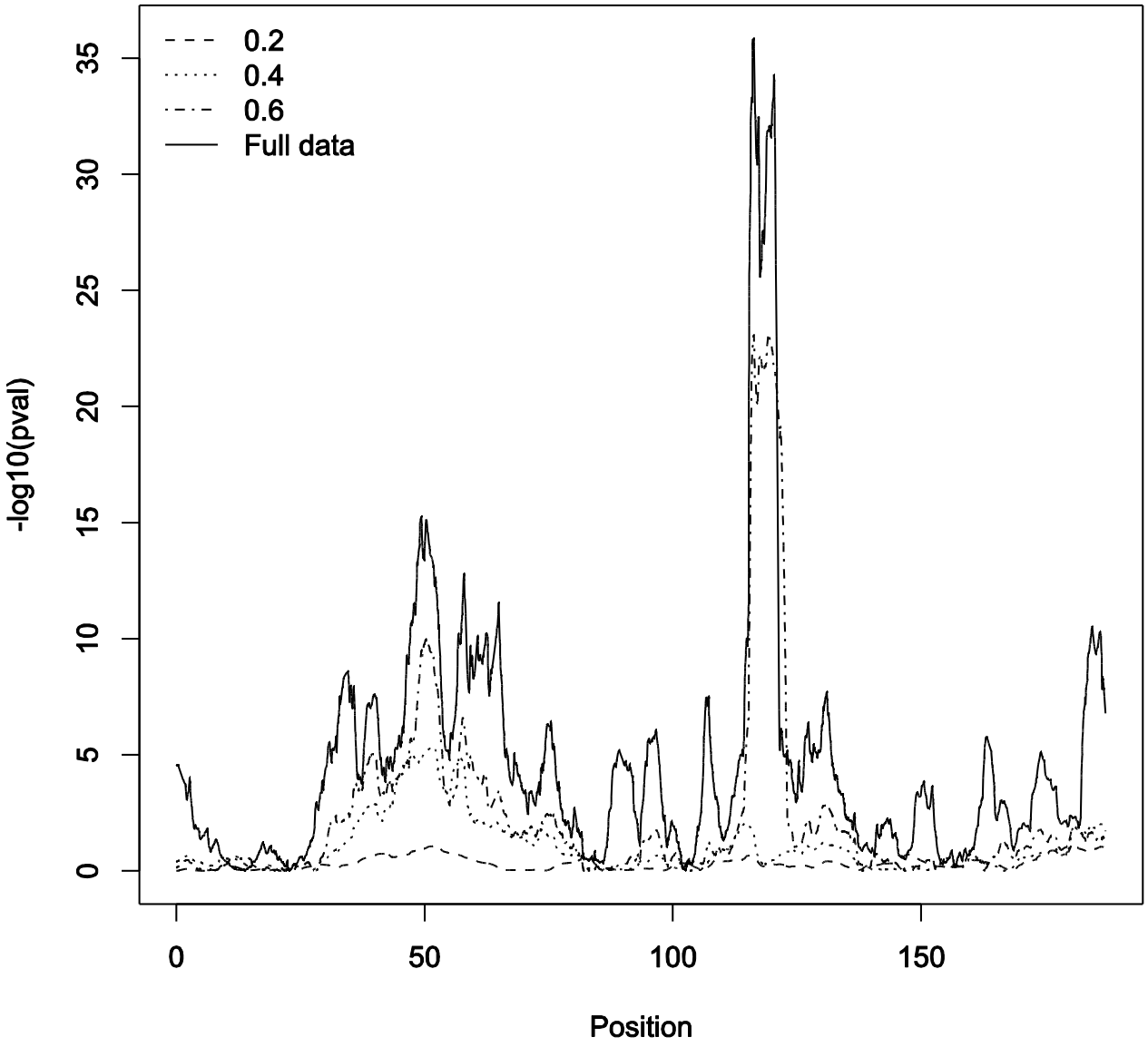


Figure 3. Null simulation results for chromosome 7 markers.

False positive linkage peaks from simulation of null linkage at varying LD thinning thresholds on chromosome 7.

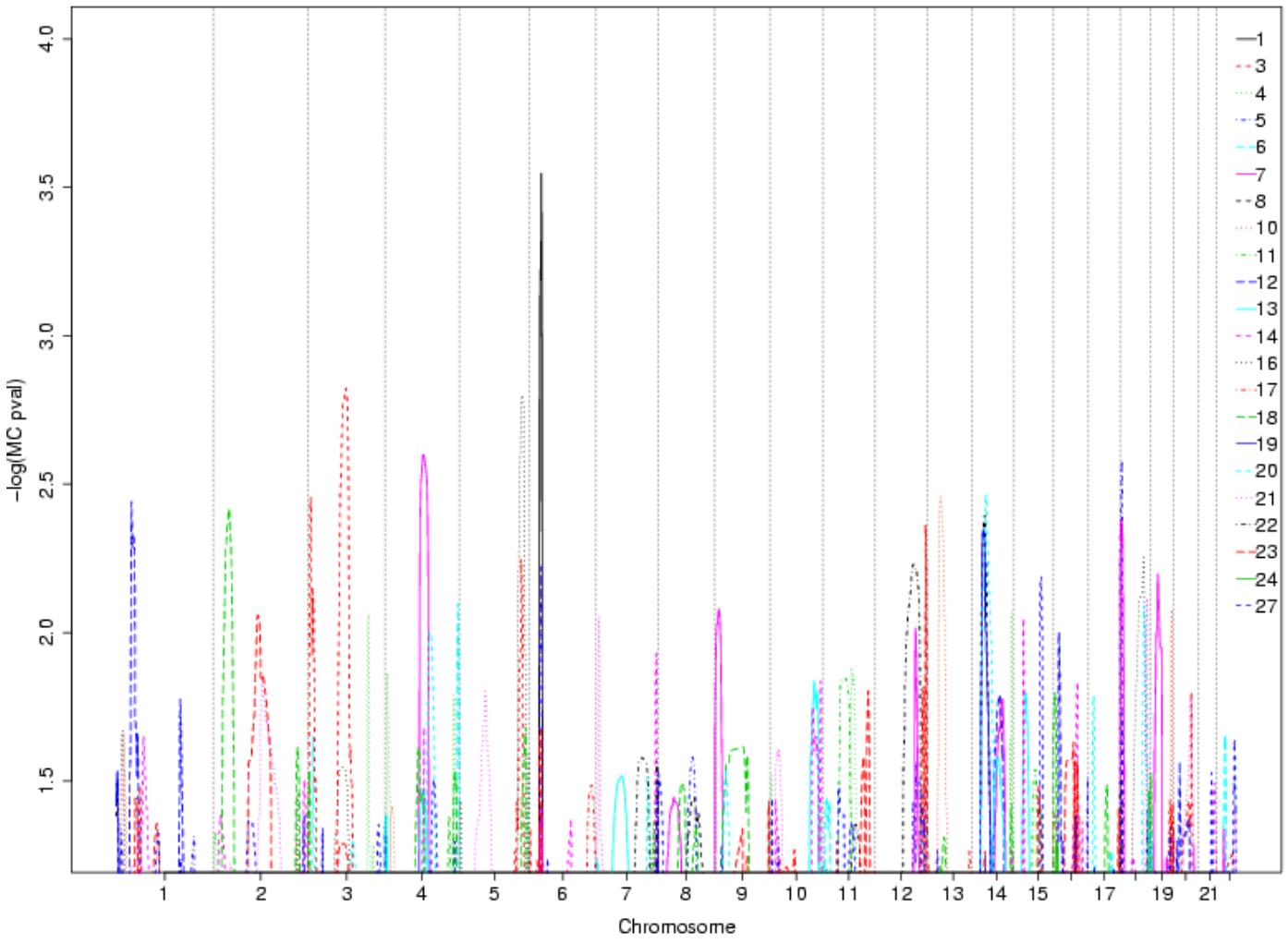


Figure 4. Linkage peaks by chromosome and family.

Only unadjusted p-values < 0.1 are displayed. Family numbers (legend) correspond to those shown in Figure 1.

464 **Table 1. Number of individuals with membership in ≥1 of the 22 family groups, by disease status.**

Status	How many Families?		
	1	2	3
Pedigree member only	1618	125	8
Unaffected subject	76	17	1
Affected subject	128	25	1
Total	1822	167	10

465

466

Table 2. Total number of affected study subjects per family, and number of pairs per family for linkage analysis.

Family[a]	Affected Individuals	Pairs
1	31	465
3	16	120
4	11	55
5	11	55
6	8	28
7	10	45
8	8	28
10	7	21
11	5	10
12	8	28
13	6	15
14	7	21
16	7	21
17	6	15
18	6	15
19	6	15
20	5	10
21	5	10
22	3	3
23	6	15
24	4	6
27	5	10
Total	181	1011
Count[b]	154	

[a]Families are numbered to 27, but 2, 9, 15, 25, and 26 were not included in the study; total families = 22.

[b]The total number of distinct individuals. Some subjects were members of more than one family (see Table 1).

472 **Table 3**. Summary of null simulation results for chromosome 7 at various thinning intervals.

473

	Max $R^{2[a]}$				
	0.2	0.4	0.6	Full (1.0)	No LD
Markers[b]	951	3031	6112	14008	14008
Bias[c]	0.007	0.013	0.025	0.028	0.007
MSE[d]	0.024	0.017	0.038	0.024	0.011
FP rate[e]	0.142	0.147	0.271	0.274	0.044
FN rate[f]	0.012	0.002	0.013	0.0003	0.001
Called pos[g]	118.6	150.9	120.0	188.7	132.7
True pos[h]	135.2	135.5	123.2	137.3	128.4
-log10(min(p))[i]	1.17	5.32	23.1	35.8	

474

475 [a]Max R^2 : maximum allowed pairwise R^2 between adjacent SNPs (as thinning threshold).

476 [b]Markers: number of SNPs in map.

477 [c]Bias: average difference between estimated IBD state and true IBD state.

478 [d]MSE: mean-squared error of estimated IBD probability.

479 [e]FP rate: false positive IBD rate, assuming estimates of probability $\geq$ 0.5 to be positive calls.

480 [f]FN rate: false negative IBD rate, assuming estimates < 0.5 to be negative calls.

481 [g]Called pos: mean number of pairs called IBD at a given locus.

482 [h]True pos: mean number of pairs simulated as IBD at a given locus.

483 [i]-log10(min(p)): smallest linkage p-value across all markers.

484

Table 4. Linkage peaks with asymptotic $p < 10^{-3}$.

Region	Chromosome	Family	cM	Mb	Z	per-locus asymptotic	per-locus Monte Carlo	per-family per-genome Monte Carlo	across families, per-genome
1p36.13-p36.11	1	10	43.52	20.39	3.10	0.000961	0.0220	0.1386	1
1p34.3-p33	1	12	68.39	40.18	3.53	0.000210	0.0036	0.0464	1
2p23.2-p21	2	18	65.84	40.09	4.07	0.000023	0.0038	0.0327	1
3p11.2-q13.11	3	3	109.99	97.11	4.60	0.000002	0.0015	0.0155	1
4q22.1-q28.1	4	7	110.81	98.89	3.76	0.000087	0.0025	0.0291	1
	4	20	125.57	115.29	3.36	0.000390	0.0099	0.0855	1
4q35.1-q35.2	4	6	205.52	186.97	3.66	0.000127	0.0080	0.0736	1
5q33.2-q34	5	3	165.04	159.59	3.55	0.000196	0.0057	0.0668	1
	5	16	167.31	161.86	5.20	0.000000	0.0016	0.0173	0.99
6p22.2-p21.32	6	5	48.67	30.01	3.23	0.000616	0.0060	0.0809	1
	6	1	48.67	30.04	6.21	0.000000	0.0003	0.0018	0.9
7p22.2-p21.3	7	21	13.78	7.79	3.20	0.000699	0.0088	0.0832	1
9p24.3-p22.2	9	7	23.37	10.02	3.09	0.000985	0.0083	0.0973	1
10q24.31-q26.13	10	13	137.44	114.22	3.23	0.000625	0.0145	0.0964	1
12q21.33-q24.11	12	22	114.02	97.68	5.81	0.000000	0.0059	0.0291	0.95
13q12.3-q14.11	13	10	30.72	34.28	4.75	0.000001	0.0035	0.0286	1
14q11.2-q22.1	14	19	21.97	29.89	3.52	0.000218	0.0045	0.0523	1
	14	8	26.80	33.25	3.54	0.000197	0.0040	0.0514	1
	14	20	34.51	36.73	3.59	0.000165	0.0034	0.0409	1
15q11.2-q14	15	13	34.32	29.69	3.18	0.000729	0.0159	0.1023	1
18p11.32-p11.23	18	7	5.97	2.31	3.45	0.000278	0.0041	0.0495	1
	18	5	11.56	3.99	3.80	0.000072	0.0026	0.0405	1
18q21.1-q22.3	18	16	83.12	60.03	4.32	0.000008	0.0056	0.0432	1
	18	20	85.31	61.23	3.44	0.000287	0.0081	0.0736	1
	18	21	96.32	68.83	3.26	0.000550	0.0077	0.0727	1
19p13.2-q12	19	7	45.08	18.47	3.21	0.000667	0.0063	0.0741	1
19q13.41-q13.42	19	10	92.39	53.91	3.76	0.000084	0.0083	0.0682	1