bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

3 **[1] School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40292, USA**

4 **[2] Department of Neurology, University of California San Francisco, San Francisco, CA.**

5 **Corresponding Author:**

6 **Elizabeth O'Brien, Ph.D.**

7 **School of Public Health and Information Sciences**

8 **University of Louisville**

9 **Louisville, KY 40292**

10 **liz.obrien@louisville.edu**

11

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

13 **Abstract.**

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

30 **Introduction.**

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

53 not necessarily implicated in familial clusters and sporadic cases (without a family history), even in the same
54 population; and 3) familial cancers are relatively rare, and thus do not account for more than approximately 25% of all
55 cases in a population, or 20% of incident breast cancers [12]. For these reasons, much doubt has been expressed over
56 the last decade that family studies had much future utility for resolving complex genotypes for diseases like breast
57 cancer[13]. Instead, as genome-wide data rapidly became available, and with it an acute need for "high-throughput"
58 analyses, the focus of research quickly shifted to simpler association study designs to measure genetic differences
59 between phenotypic classes, such as cases and controls.

60 The genome-wide association study (GWAS) approach focuses on genotype-phenotype co-variation, usually for a
61 densely distributed set of SNPs over the genome. Positive associations occur where genotype differences correspond
62 to phenotype differences outside of what is expected under a null hypothesis, and their locations mark points in or
63 near genetic variants that cause disease or contribute to its risk. Numerous GWAS have been done in search of genes
64 that condition risk of breast cancer, and a list of genes and variants with modest effects on cancer risk has certainly
65 developed as a result [14] [15]. However, the small fraction of breast cancers attributable to these relatively common
66 but low-penetrance alleles suggests that a larger set of genetic factors, more of them reaching moderate effect, but
67 occurring with low frequency in a population, might account for such common cancer phenotypes. This "heritability
68 gap" has been considered a problem of statistical lack of power to resolve a potentially large number of genetic
69 variants, some of them low in frequency (rare), and of only moderate or low risk effect for common but deadly disease
70 phenotypes.

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

76 simply underpowered to sort high throughput data for a definitive set of genetic factors —unknown in number and
77 varying in frequency and effect size--responsible for a complex disease phenotype. As a result, there is much
78 uncertainty about what an association study really captures, and clarification is often sought by improving power and
79 reliability—by increasing genome coverage, sample size, or by meta-analyses. In this regard, today's study designs are
80 ambitious, involving huge numbers of cases and increasingly narrow definitions of the phenotype[16]. Even so, GWAS
81 of breast cancer have not resolved single genes of major effect comparable to *BRCA1* and *BRCA2*; neither have they
82 established a comprehensive predisposing genotype for the disease.

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

92

93 ## Subjects and Methods.

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

95 The Utah Population Database (UPDB) is a repository of longitudinal information originally constructed from
96 genealogical data pertaining to Utahans and their families [17]. Through successive record linking efforts, the database
97 integrates cancer registry data, medical records data, Utah State certified deaths and births, etc. Currently, the UPDB

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

98 captures information for approximately 7 million individuals, many of whom are members of extensive pedigree
99 networks 2 to 14 generations deep [18]. Pedigree information from the UPDB, and Utah's SEER cancer registry, were
100 used to establish diagnosed breast cancer cases clustered in large multigenerational families. We then compared
101 observed and expected breast cancer incidence in case families and recruited study subjects from "high risk" families,
102 i.e., those with excess incidence having a probability of less than 0.01 of occurring by chance [19]. However, we
103 excluded cases and families previously studied and known to be segregating *BRCA1* or *BRCA2* mutations as their
104 primary genetic risk factors for breast cancer.

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

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

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

118

119 ***Genotypes***

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

120 Genotyping was performed with Illumina 370 Duo and 610 Quad arrays at deCODE Genetics, Reykjavik, Iceland. SNPs
121 with low quality scores (GenCall[20] quality score < 0.15), and those with inconsistent allele frequencies between the
122 two arrays (any absolute difference in minor allele frequency > 0.05), were eliminated. All alleles were called on the
123 forward strand, and checked for consistency between arrays. After approximately 15% of the SNPs were eliminated by
124 these quality control criteria, a total of 285,630 genotypes per subject were retained. Mendelian consistency checks
125 were not performed because of the very small number of families with informative data.

126

127 ***Evaluation of genetic vs. genealogical relatedness***

128 We examined the degree to which relatedness assessed by genome-wide genetic similarity corresponded to
129 relatedness as reported in the UPDB genealogical data for pairs of relatives. For this study, we used genotypes on 429
130 individuals, including the 248 subjects in the linkage study, as well as 181 women from families with fewer than 3
131 genotyped breast cancer cases. A total of 91,806 pairs were evaluated, using coefficient of relatedness to characterize
132 the genealogical data, and the genetic relatedness matrix computed by GCTA[21] to characterize relatedness from SNP
133 data. To facilitate comparison, relatedness from each measure was grouped by rounding $-\log_2$(relatedness) to
134 correspond to degree of relationship.

135

136 ***Identity by Descent (IBD) estimation for linkage analysis***

137 Pairs formed from the sample set were used to generate Identity by Descent (IBD) matrices for linkage analysis. IBD
138 was computed using PEDIBD software developed by Li and colleagues[22]. Their method employs a Viterbi algorithm
139 [23] to find the most likely path of descent of an ancestral allele through a deep, but sparsely genotyped pedigree
140 structure, via hidden Markov models of inheritance and recombination. The method efficiently parses the high-density
141 genotype data of the Illumina arrays, permitting estimation of IBD matrices for 1,011 affected relative pairs at up to
142 285,630 loci in approximately 24 hours of CPU time on current equipment (substantially less for thinned data sets).

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

143 Allele frequencies were estimated by simple counting among all genotyped individuals, affected or unaffected. As
144 noted by Boehnke[24] and others, simple counting among family members does not introduce any systematic bias in
145 the absence of allelic association, and any association would introduce a conservative bias as it would lead to
146 overestimation of the frequency of a disease-associated allele.

147

148 ***Test statistic for linkage***

149 We employed the IBDREG quasi-likelihood approach described by Schaid, et al. [25] to test for concordant pair
150 (affected only) linkage without covariates. IBDREG has an important advantage in comparison to competing methods
151 as it appropriately adjusts for between-pair covariance when multiple relative pairs are drawn from the same pedigree
152 structure. Because the families studied vary considerably in size, and some have only a few affected members, the
153 distributional properties (and hence the asymptotic p-values) of the test statistic were uncertain. Therefore, we used
154 simulation to compute p-values and family-wise error rates. The approach is described below.

155

156 ***Simulation of Identity by Descent in the Absence of Linkage, but the Presence of Linkage Disequilibrium***

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

162

163 ***Estimation of error rates***

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

164 It is well known that linkage analysis based on high-density SNP arrays is subject to potentially severe bias away from
165 the null because of linkage disequilibrium (LD). LD between nearby markers will cause overestimation of the
166 probability that two related individuals share marker alleles that are identical by descent (IBD) [26, 27]. Although in
167 principle, simultaneous modeling of LD among founders and IBD among descendants would be the most powerful
168 approach to using all the genotype data at our disposal [28], the computational burden of such modeling in complex
169 multigenerational families is not readily surmountable at present.

170

171 ***Marker thinning intervals***

172 Marker thinning effectively varies the strength of LD by setting maximum R^2 between SNPs at various thresholds (0.6,
173 0.4, and0.2 here). At each threshold, SNPs were thinned by recursively finding the midpoint of a block of SNPs
174 mutually correlated at $R^2 >$ the current threshold, then dropping all but the midpoint SNP, so that the maximum
175 pairwise correlation could not exceed the selected level. Thinned marker sets were run against simulated (null)
176 genotype data for chromosome 7 to establish error rates in the IBD estimates and thus, the contribution to false
177 positive linkage scores for varying strengths of LD structure.

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

185 HapMap's estimated recombination fractions. Recombination between markers was estimated by cubic spline
186 interpolation using R[32].

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

190

191 ***Distribution of test statistics under the null***

192 The IBD estimates generated by PEDIBD were input to IBDREG to calculate linkage statistics. All simulated IBD states
193 and marker allele data were generated under the null hypothesis of no linkage between marker loci and disease
194 predisposition. Thus, the distribution of test statistics for each marker locus within each family can be taken to
195 represent a sample from the null distribution for a whole genome scan of that family. In addition to the per-locus
196 asymptotic p-value computed by IBDREG, we report a family-specific per-locus Monte Carlo p-value, a family-specific
197 per-genome Monte Carlo p-value, and a Monte Carlo composite false discovery rate (FDR) controlling for the whole-
198 genome analysis of 22 families[33].

199

200 ***Identification of linkage peaks and boundaries***

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

205

206 **Results.**

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

207 An initial check for correspondence between coefficients of relatedness estimated from pedigree information and
208 from SNP genotypes was made for all possible pairs of study subjects (see Methods). This information is plotted in
209 **Figure 2** for pairs of related individuals**.** The most distantly related pairs in the genealogical data were 13th degree
210 relatives, so pairs unrelated by genealogy and pairs estimated to be genetically more distant than 13th degree were
211 plotted as though they were 14th degree relatives on either scale. There was generally very good agreement between
212 genealogical and genetic distance up to about the 6th degree, and a gradual loss of precision past that point in this
213 population.

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

219 Simulations were done to depict the inflationary effect of LD on IBD and false positive linkage scores (see **Methods**).
220 These results are shown in **Figure 3** and **Table 3**. In order to control for this effect, and reduce false positive linkage
221 signal, SNPs were thinned to various thresholds of correlation between them. At the threshold $R^2 \leq 0.4$, IBD over-
222 estimation due to LD was controlled fairly well, but positive linkage peaks still occurred. At $R^2 \leq 0.2$, spurious linkage
223 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
224 thinned set of 19,609 SNPs.

225 **Table 4** gives linkage results for 1,011 affected relative pairs generated from a total of 154 genotyped breast cancer
226 cases. The analysis identified 19 distinct peaks with asymptotic unadjusted within-family $p < 0.001$. More realistic
227 estimates of the probability of these results under the null hypothesis are derived from the 100 per-family genome-
228 wide simulations, and presented in Table 4 as well. Monte Carlo per-locus p-values are generally considerably larger
229 than the asymptotic p-values, particularly for smaller families. After further adjustment for genome-wide comparisons

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

230 within families, 11 regions retained adjusted p-values below 0.05, and 17 regions retained adjusted p-values below
231 0.1. However, when we adjusted for simultaneous whole-genome search across all 22 families, only the 3 peaks with
232 the highest scores were large enough that a single random result under the null would not have been expected to
233 exceed them 100% of the time.

234 **Supplementary Table S1** lists all breast cancer- associated genes from DisGeNET
235 [http://www.disgenet.org/web/DisGeNET/v2.1], TCGA [34] and Cancer Resource[35] located within peaks defined by a
236 10-fold increase in asymptotic p-value. The large peak on chromosome 6 for family 1 includes multiple genes that have
237 been associated with breast cancer risk and/or tumorigenesis, including members of the *HLA* complex, *NOTCH4*, and
238 *TNF*, among others. Also noteworthy is that the chromosome 13 peak for family 10 includes *BRCA2*, while no family
239 exhibited linkage to the *TP53* or *BRCA1* regions on chromosome 17. **Figure 4** shows the relative locations and
240 amplitudes of the linkage peaks by family.

241

242 **Discussion.**

243 It is low-frequency variants that are difficult to find in convincing association with a disease phenotype from genome-
244 wide association tests[13]. However, if we are to resolve this low frequency, moderate risk class of variants, then
245 population-wide sampling from whole undifferentiated, or minimally structured populations, is perhaps not the most
246 strategic sampling approach to use. Variants of this class occur de novo, are replicated and transmitted to
247 descendants. For this reason, they will reach their highest frequencies within family lineages[36], the larger the better,
248 while remaining at low frequency (rare) in any usual population sample, whether n = 100s or 100,000s. The moderate
249 risk nature of this class of variants is likely due to the fact that their risk effects depend on participation in larger gene
250 networks to account for increased cancer risk in particular families. In this sense, variants of smaller effect can alter
251 disease risk in the context of gene networks that regulate the functional pathways involved in the onset and/or
252 progression of the disease.

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

253 The family study approach does not rest on anticipating "a new breast cancer genotype", nor a "comprehensive
254 genotype" to account for breast cancer risk in this population and by the usual purview of linkage analysis. Instead,
255 we tried to capture evidence of low frequency variants at the population level, but enriched at the level of very large
256 high-risk families. Our approach yielded 17 genomic regions possibly linked (per-family per-genome Monte Carlo $p \leq$
257 0.1) to breast cancer for the 22 families studied, but with considerable variation among families: 15 of the 22 families
258 (68%) showed possible linkage to at least 1 region by the criteria used here; 1 family showed possible linkage to 4
259 regions; 1 family to 3 regions; 5 families to 2 regions; and 8 families showed linkage to 1 region. It is noteworthy that
260 linkage to the *BRCA2* region on chromosome 13 was observed in only one family (10), while no family exhibited
261 linkage to the *BRCA1* region on chromosome 17.

262 The availability of high-density marker sets, efficient algorithms for estimating IBD in large families, and substantial
263 computational resources permitted simulation of 100 null genome-wide results for each family. The simulation results
264 then allowed us to compare Monte Carlo p-values with asymptotic p-values based on large sample theory. In Table 4 it
265 is shown that the asymptotic estimates are frequently smaller than the Monte Carlo p-values by an order of
266 magnitude or more. The genome-wide results for each family represent an appropriate basis for comparison to other
267 published results based on linkage studies of one or a few families. Adjusting linkage estimates for all 22 families
268 simultaneously, we find no linkage scores, or peaks, that could not have occurred by chance: the observed Z-score of
269 6.21 for family 1 on 6p21-22 was exceeded in 90% of null simulations—for at least one family at some location over
270 the genome. However, in the simulated data, anomalously high Z-scores were much more commonly observed in
271 small families—and never in family 1—so the adjustment across families is in some part size dependent, and
272 therefore, less than perfect. For this reason, it might be more appropriate to consider the simulated probability of
273 observing the result in 22 families just like family 1, which would be approximately $1-(1-0.0018)^{22} = 0.039$. Moreover,
274 the existence of a possible linkage peak for family 5 in precisely the same location as family 1 on 6p21-22 strengthens
275 the case for a susceptibility locus in this region.

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

276 For heuristic purposes, we can combine multiple lines of evidence to rank the various linked regions by priority. First,
277 regions that overlap across multiple families (e.g. 6p22-21, families 5 and 1; 18p11, families 5 and 16; 18q21-22,
278 families 20 and 21) likely indicate either a shared disease-predisposing haplotype inherited from an unknown common
279 ancestor, or multiple predisposing variants in the same gene in truly unrelated, or variably related, families. Next, per-
280 family, per-genome Monte Carlo p-values well below 0.05 (e.g. 3p11-13, family 3; 12q21-24, family 22) warrant further
281 investigation. Finally, linked regions in individual families (13q12-14, family 10) that overlap with known breast cancer-
282 predisposing loci, i.e. BRCA2, have the potential to greatly simplify mapping variants associated with specific diseases.
283 In addition, the other regions suggestive of linkage without known breast cancer associated variants, might provide
284 useful new clues about the location of genetic variants that increase the risk of breast cancer in members of these
285 families, and serve as evidence of residual heterogeneity in genomic regions responsible for familial cancer
286 susceptibility.

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

300 differences in sequence sharing per exome through the linked regions. By using the linkage-partitioned information
301 thus far, sequencing should reveal more specifically the locations of rarer variants likely relevant to the disease.
302 Linkage and sequencing techniques together should do much to clarify the genetic architecture of breast cancer in this
303 population, its heterogeneity among families[37], and importantly, give us a deeper understanding of the role of rare
304 variants in conditioning risk differently among groups.

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

314 **Supplemental Table and Figures.**

315

316 1. Figure S1. Manhattan plots for each family. Families are labeled as per Figure 1 in upper right corner
317 of each plot.
318 2. Figure S2. Locations of genes within linkage peaks with unadjusted p-value < 0.001. Within peaks,
319 cyan lines indicate genes, red lines indicate genes mutated in TCGA breast cancer specimens, black
320 lines indicate boundaries of overlapping peaks. Coding strand is indicated by placement within box:
321 genes coded on the forward strand are drawn above the midline, while genes coded on the reverse
322 strand are drawn below.
323 3. Table S1. Table of all genes in linked regions, ordered by bioinformatics resource scores (see text for
324 references): tcga.mut = number of mutations observed in TCGA breast tumors; cr = cancer resource
325 breast cancer associated (1) vs not associated (0); dg = disGeNet breast cancer association score; sum
326 = sum((tcga.mut > 0)+(cr > 0)+(dg > 0)); tcga = TCGA breast tumor mutations/bp.
327

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

328 **Acknowledgments.**

329 Data analysis and genotyping was supported by Susan G. Komen Foundation grant KG-071514. Partial support for all
330 datasets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute and
331 the Huntsman Cancer Institute Cancer Center Support grant, P30-CA2014 from the National Cancer Institute. Support
332 for the Utah Cancer Registry is provided by Contract No HHSN2612013000171 from the National Cancer Institute with
333 additional support from the Utah Department of Health. We thank Debbie Shea, Matt Hoffman, Terah Young, and
334 Loren Budge for subject recruiting and data collection; Geri Mineau, Alison Fraser and Janice Conrads at the Huntsman
335 Cancer Institute's Pedigree and Population Resource for linking subject data to the Utah Population Database; Xin Li
336 for consultation on PEDIBD modifications for application to our data; Dan Schaid, Duncan Thomas, Steve Schrodi, and
337 Ken Boucher for valuable discussions of this work.

338

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

339 **References**

340 1. Malkin D, Li FP, Strong LC, Fraumeni J, Nelson CE, Kim DH, et al. Germ line p53 mutations in a
341 familial syndrome of breast cancer, sarcomas, and other neoplasms. Science. 1990;250(4985):1233-8.
342 2. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset
343 familial breast cancer to chromosome 17q21. Science. 1990;250(4988):1684-9.
344 3. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong
345 candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266(5182):66-71.
346 4. Ford D, Easton D, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and
347 penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The American Journal of
348 Human Genetics. 1998;62(3):676-89.
349 5. Seal S, Barfoot R, Jayatilake H, Smith P, Renwick A, Bascombe L, et al. Evaluation of Fanconi
350 Anemia genes in familial breast cancer predisposition. Cancer research. 2003;63(24):8596-9.
351 6. Nieuwenhuis MH, Kets CM, Murphy-Ryan M, Yntema HG, Evans DG, Colas C, et al. Cancer risk and
352 genotype–phenotype correlations in PTEN hamartoma tumor syndrome. Familial cancer.
353 2014;13(1):57-63.
354 7. Tan M-H, Mester JL, Ngeow J, Rybicki LA, Orloff MS, Eng C. Lifetime cancer risks in individuals
355 with germline PTEN mutations. Clinical Cancer Research. 2012;18(2):400-7.
356 8. Riegert-Johnson DL, Gleeson FC, Roberts M, Tholen K, Youngborg L, Bullock M, et al. Research
357 Cancer and Lhermitte-Duclos disease are common in Cowden syndrome patients. 2010.
358 9. Hearle N, Schumacher V, Menko FH, Olschwang S, Boardman LA, Gille JJ, et al. Frequency and
359 spectrum of cancers in the Peutz-Jeghers syndrome. Clin Cancer Res. 2006;12(10):3209-15. doi:
360 10.1158/1078-0432.CCR-06-0083. PubMed PMID: 16707622.
361 10. Tung N, Battelli C, Allen B, Kaldate R, Bhatnagar S, Bowles K, et al. Frequency of mutations in
362 individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing
363 with a 25-gene panel. Cancer. 2015;121(1):25-33. doi: 10.1002/cncr.29010. PubMed PMID: 25186627.
364 11. Walsh T, King MC. Ten genes for inherited breast cancer. Cancer Cell. 2007;11(2):103-5. doi:
365 10.1016/j.ccr.2007.01.010. PubMed PMID: 17292821.
366 12. Kerber RA, O'Brien E. A cohort study of cancer risk in relation to family histories of cancer in the
367 Utah population database. Cancer. 2005;103(9):1906-15. Epub 2005/03/22. doi: 10.1002/cncr.20989.
368 PubMed PMID: 15779016.
369 13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing
370 heritability of complex diseases. Nature. 2009;461(7265):747-53. doi: 10.1038/nature08494. PubMed
371 PMID: 19812666; PubMed Central PMCID: PMC2831613.
372 14. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide
373 association study identifies novel breast cancer susceptibility loci. Nature. 2007;447(7148):1087-93.
374 15. Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, Newman B, et al. A genome wide linkage
375 search for breast cancer susceptibility genes. Genes, Chromosomes and Cancer. 2006;45(7):646-55.
376 16. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale
377 genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013;45(4):353-
378 61.
379 17. Skolnick M. The Utah Genealogical Database: a Resource for Genetic Epidemiology. In: Cairns JJ,
380 Lyon JL, Skolnick M, editors. Cancer Incidence in Defined Populations. Cold Spring Harbor, NY: Cold
381 Spring Harbor Laboratory; 1980. p. 285-96.

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

382 18. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends
383 Biotechnol. 2003;21(3):113-6. PubMed PMID: 12628367.
384 19. Kerber RA. Method for calculating risk associated with family history of a disease. Genet
385 Epidemiol. 1995;12(3):291-301. Epub 1995/01/01. doi: 10.1002/gepi.1370120306. PubMed PMID:
386 7557350.
387 20. Kermani BG. Artificial intelligence and global normalization methods for genotyping. Google
388 Patents; 2006.
389 21. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis.
390 The American Journal of Human Genetics. 2011;88(1):76-82.
391 22. Li X, Yin X, Li J. Efficient identification of identical-by-descent status in pedigrees with many
392 untyped individuals. Bioinformatics. 2010;26(12):i191-8. Epub 2010/06/10. doi:
393 10.1093/bioinformatics/btq222. PubMed PMID: 20529905; PubMed Central PMCID: PMC2881406.
394 23. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes: The Art of Scientific
395 Computing. 3 ed. Cambridge: Cambridge University Press; 2007.
396 24. Boehnke M. Allele frequency estimation from data on relatives. American journal of human
397 genetics. 1991;48(1):22.
398 25. Schaid DJ, Sinnwell JP, Thibodeau SN. Testing genetic linkage with relative pairs and covariates
399 by quasi-likelihood score statistics. Human heredity. 2007;64(4):220-33. Epub 2007/06/15. doi:
400 10.1159/000103751. PubMed PMID: 17565225; PubMed Central PMCID: PMC2880728.
401 26. Huang Q, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers
402 induces false-positive evidence of linkage for affected sib pair analysis. The American Journal of Human
403 Genetics. 2004;75(6):1106-12.
404 27. Webb EL, Sellick GS, Houlston RS. SNPLINK: multipoint linkage analysis of densely distributed
405 SNP data incorporating automated linkage disequilibrium removal. Bioinformatics. 2005;21(13):3060-
406 1.
407 28. Thomas A. Towards linkage analysis with markers in linkage disequilibrium by graphical
408 modelling. Human heredity. 2007;64(1):16-26.
409 29. The International HapMap 3 Consortium. Integrating common and rare genetic variation in
410 diverse human populations. Nature. 2010;467(7311):52-8. doi:
411 http://www.nature.com/nature/journal/v467/n7311/abs/nature09298.html#supplementary-
412 information.
413 30. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'etude du polymorphisme
414 humain (CEPH): collaborative genetic mapping of the human genome. Genomics. 1990;6(3):575-7.
415 PubMed PMID: 2184120.
416 31. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics.
417 2011;27(16):2304-5. doi: 10.1093/bioinformatics/btr341.
418 32. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R
419 Foundation for Statistical Computing; 2013.
420 33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach
421 to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.
422 34. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer
423 genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113-20.
424 35. Ahmed J, Meinel T, Dunkel M, Murgueitio MS, Adams R, Blasse C, et al. CancerResource: a
425 comprehensive database of cancer-relevant proteins and compound interactions supported by
426 experimental knowledge. Nucleic acids research. 2011;39(suppl 1):D960-D7.

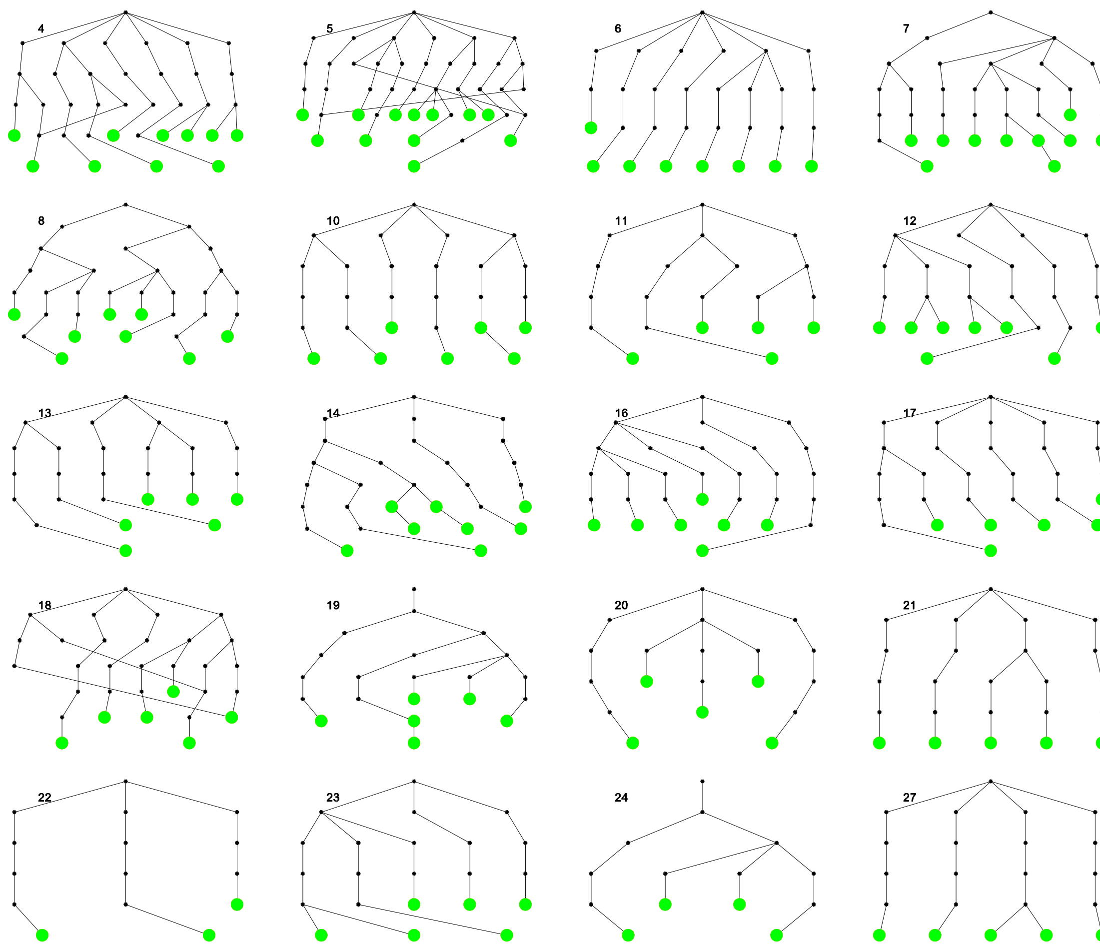
bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

427 36. Wijsman EM. The role of large pedigrees in an era of high-throughput sequencing. Human
428 genetics. 2012;131(10):1555-63. doi: 10.1007/s00439-012-1190-2. PubMed PMID: 22714655; PubMed
429 Central PMCID: PMC3638020.
430 37. Lynch H, Wen H, Kim YC, Snyder C, Kinarsky Y, Chen PX, et al. Can unknown predisposition in
431 familial breast cancer be family-specific? Breast J. 2013;19(5):520-8. doi: 10.1111/tbj.12145. PubMed
432 PMID: 23800003.

433

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

434 **Figures.**

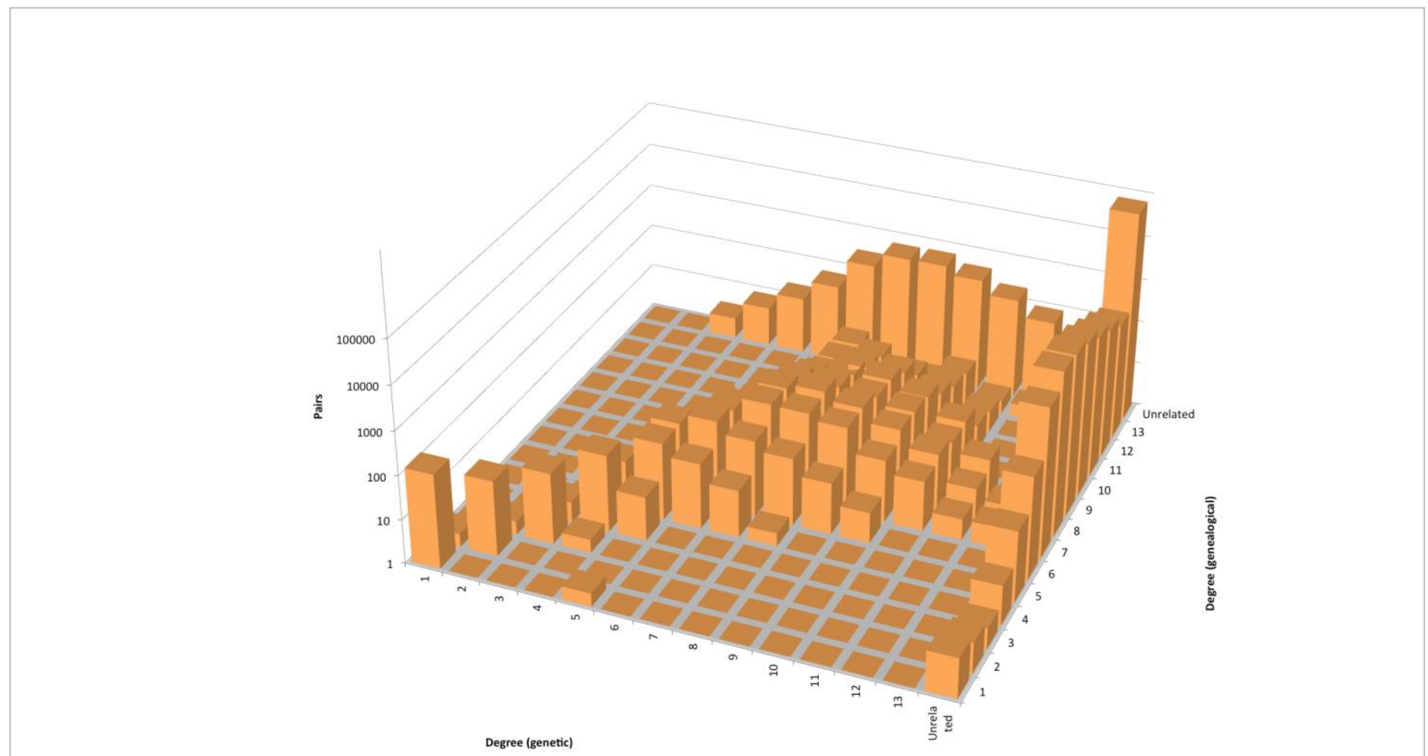


435

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

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

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

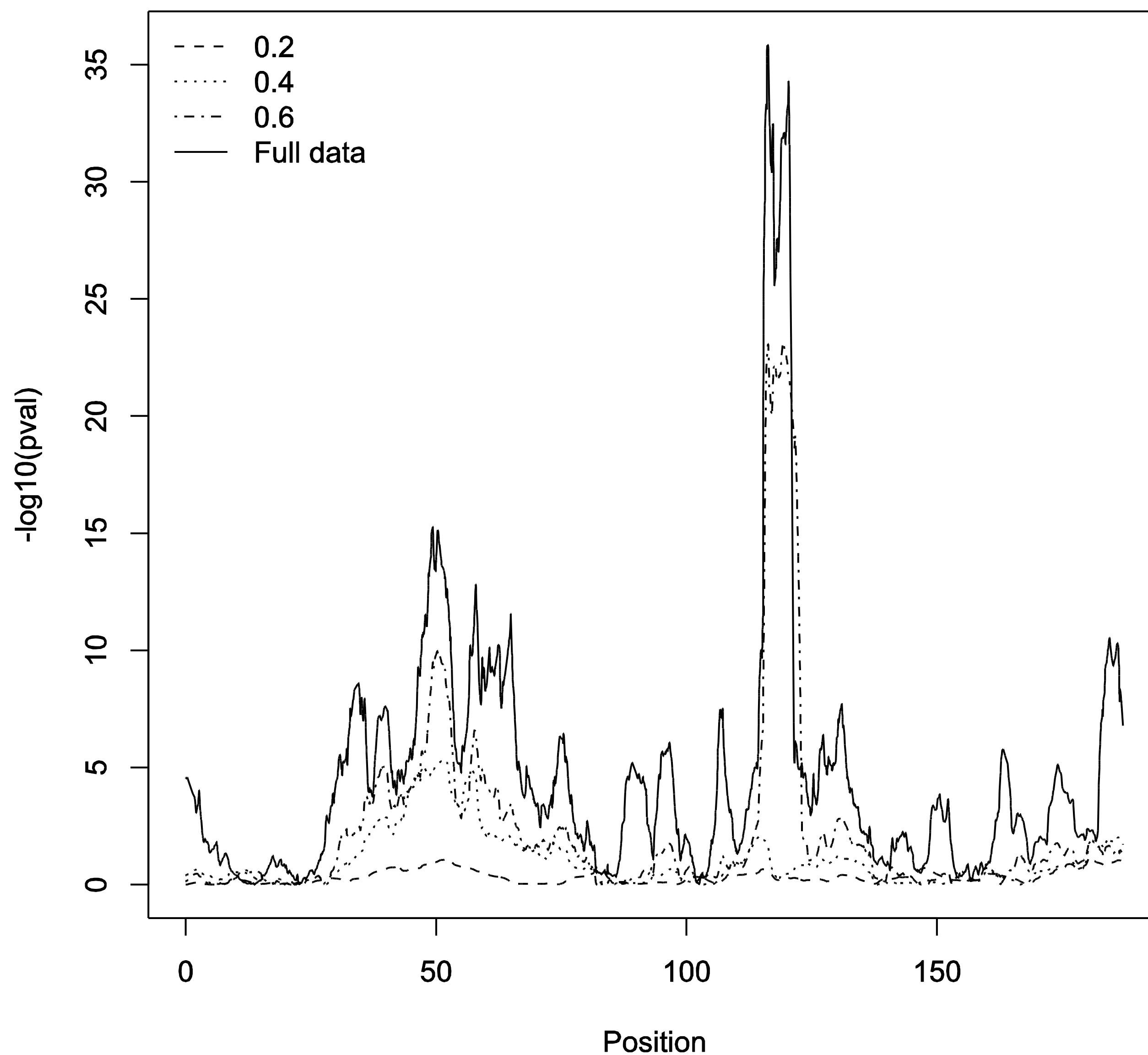
441


442 **Figure 2. Genetic vs. genealogical relatedness.**

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

447

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

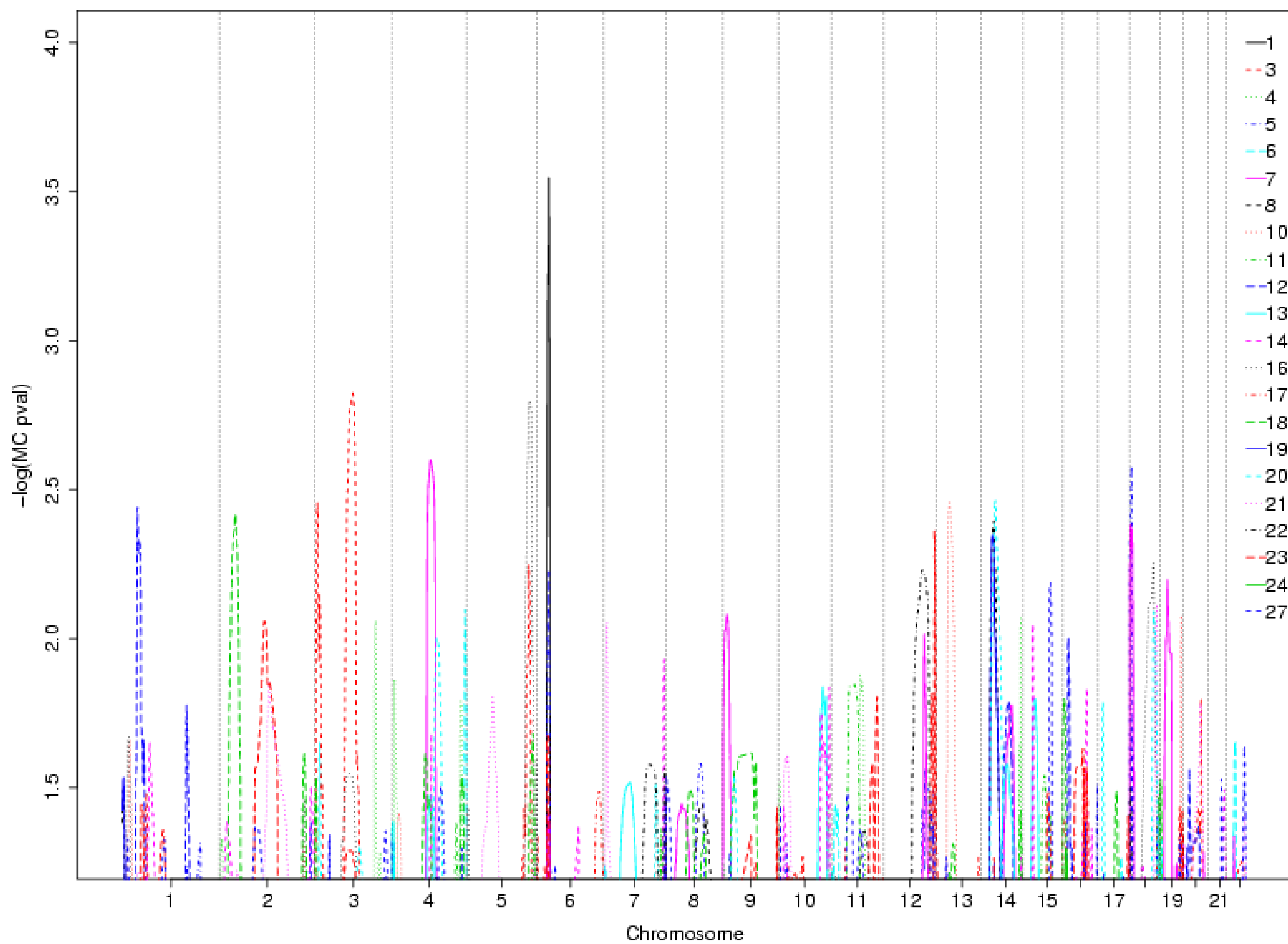


448
449 **Figure 3. Null simulation results for chromosome 7 markers.**

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

451

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



452
453 **Figure 4.** Linkage peaks by chromosome and family.

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

455

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

456 **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

457

458

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

459 **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	

460

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

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

463

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

465

	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	

466

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

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

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

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

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

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

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

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

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

476

bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

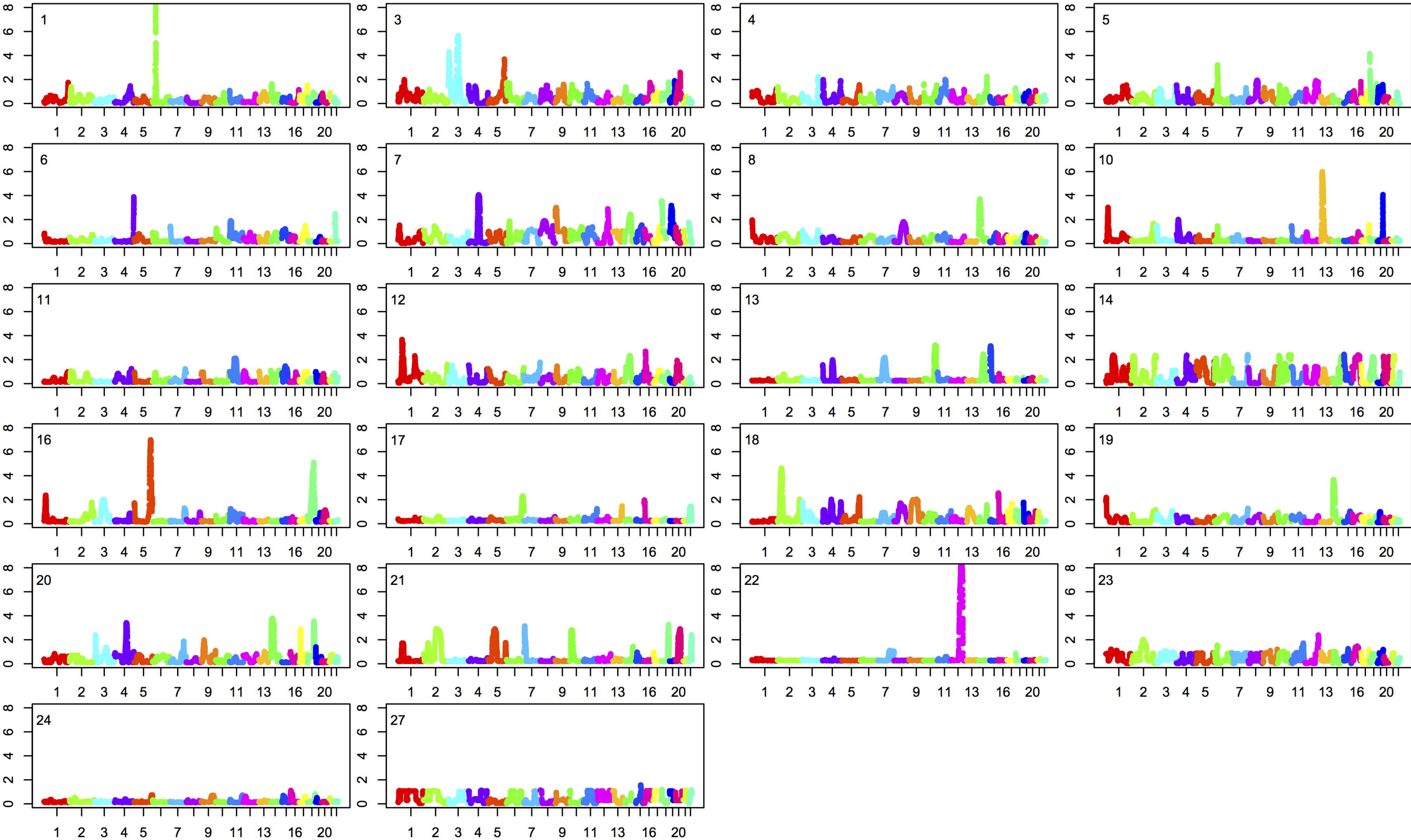
477 **Table 4. Linkage peaks with asymptotic $p < 10^{-3}$.**

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

478

479

480



bioRxiv preprint doi: https://doi.org/10.1101/092841; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.