BMC Genomics

RESEARCH ARTICLE Open Access

Pairwise shared genomic segment analysis in three Utah high-risk breast cancer pedigrees

Zheng Cai[1], Alun Thomas[2], Craig Teerlink[2], James M Farnham[2], Lisa A Cannon-Albright[2,3†] and Nicola J Camp[2*†]

Abstract

Background: We applied a new weighted pairwise shared genomic segment (pSGS) analysis for susceptibility gene localization to high-density genomewide SNP data in three extended high-risk breast cancer pedigrees.

Results: Using this method, four genomewide suggestive regions were identified on chromosomes 2, 4, 7 and 8, and a borderline suggestive region on chromosome 14. Seven additional regions with at least nominal evidence were observed. Of particular note among these total twelve regions were three regions that were identified in two pedigrees each; chromosomes 4, 7 and 14. Follow-up two-pedigree pSGS analyses further indicated excessive genomic sharing across the pedigrees in all three regions, suggesting that the underlying susceptibility alleles in those regions may be shared in common. In general, the pSGS regions identified were quite large (average 32.2 Mb), however, the range was wide (0.3 – 88.2 Mb). Several of the regions identified overlapped with loci and genes that have been previously implicated in breast cancer risk, including *NBS1*, *BRCA1* and *RAD51L1*.

Conclusions: Our analyses have provided several loci of interest to pursue in these high-risk pedigrees and illustrate the utility of the weighted pSGS method and extended pedigrees for gene mapping in complex diseases. A focused sequencing effort across these loci in the sharing individuals is the natural next step to further map the critical underlying susceptibility variants in these regions.

Keywords: Breast cancer, High-risk pedigrees, Susceptibility, Germline, Genomic sharing

Background

Breast cancer (MIM #114480) is the most prevalent cancer among women in developed countries [1]. It is a common, complex disease, including substantial genetic heterogeneity with respect to both loci and alleles. To date, many germ-line variants in multiple genes have been confirmed to increase risk for breast cancer [2]. However, the majority of hereditary breast cancer remains unexplained and there are clearly more risk variants to identify. In particular, rare variants are likely to be a part of the missing heritability [3]. Pedigrees selected for excess disease (i.e. high-risk pedigrees) offer the potential for increased genetic homogeneity and enrichment for rare and more penetrant variants. Hence the high-risk pedigree design is advantageous for the detection of rare risk variants. However, although the complexity is arguably reduced, genetic heterogeneity may still remain and can pose a substantial challenge for conventional pedigree-based methods, such as linkage analysis. High-density single nucleotide polymorphism (SNP) data also provide challenges for conventional multi-point pedigree methods because of linkage disequilibrium (LD) between markers and because subtle non-Mendelian genotype errors or inaccuracies of physical position can confuse estimation of the inheritance vectors. Genomewide association is well-suited to high-density SNP arrays, however, the power for this approach lies with the existence of high LD between a SNP on the platform and the underlying risk variant; which is vastly reduced with rare risk variants leading to low power [4,5]. Identity-by-descent (IBD) mapping, such as shared genomic segment (SGS) analysis, in extended pedigrees have been developed precisely for use with high-density SNP platforms and have been suggested to be more powerful than association analysis and traditional linkage analysis for the identification of rare variants [3,6,7]. The probability of IBD is a challenge to

* Correspondence: nicola.camp@utah.edu
†Equal contributors
[2]Division of Genetic Epidemiology, University of Utah Medical School, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA
Full list of author information is available at the end of the article

calculate in large pedigrees. Conversely, identity-by-state (IBS) is easy to compute. Our SGS methods use excessive lengths of IBS to find regions of IBD. These IBS regions are assessed for significance empirically, conditional on a model for LD and a genetic model (for recombination in the pedigree). Our original SGS method [8] was designed to identify regions of excessive lengths of sharing across all, or all but 1 or 2, cases in a pedigree, which is powerful when the cases within pedigrees are reasonably genetically homogeneous [6]. For common diseases, large high-risk pedigrees may suffer from intra-familial heterogeneity, such as when more than one genetic locus segregates within the same family. In these situations, even at a true risk locus, a greater proportion of the cases may be non-sharers. Recently, we proposed an alternate weighted pairwise SGS (pSGS) method, which combines the sharing evidence across all possible pairs, which in simulated data indicated substantial increased robustness to intra-familial genetic heterogeneity and therefore is likely more useful for mapping common diseases [9].

We performed genomewide pSGS analysis in three Utah high-risk breast cancer pedigrees selected as unlikely to be due to *BRCA1* or *BRCA2*. Regions of excessive sharing in the cases of these pedigrees have good potential for harboring breast cancer susceptibility variants.

Methods

High-risk breast cancer pedigrees

Using existing mutation screening and microsatellite linkage data, pedigrees were selected to have low probability of being due to mutations in the genes *BRCA1* and *BRCA2*. Each met the following criteria [10]: (1) the pedigree did not contain any cases known to carry *BRCA1* or *BRCA2* mutations, and (2) the pedigree had no significant linkage to the *BRCA1* or *BRCA2* regions. Hence, a-priori these pedigrees have a low probability of segregating mutations in *BRCA1* or *BRCA2*.

The three extended, high-risk Utah pedigrees studied are shown in Figure 1. All pedigrees were descended from European founders. There are no known genealogical links between the pedigrees, as determined by the Utah Population Database (UPDB [11]) which contains up to eleven generations of genealogy. Pedigree 1 contains five cases connected by a total of 17 meioses. Pedigree 2 contains 9 cases connected by 20 meioses. Pedigree 3 consists of 10 cases connected by 33 meioses. Confirmation of cancer diagnoses was obtained from the Utah Cancer Registry (UCR). All other individuals were considered "unknown", and were not genotyped in this study. These pedigrees are defined as high risk because they contain significantly more female breast cancer than expected using cancer rates calculated from the UPDB (see [12]).

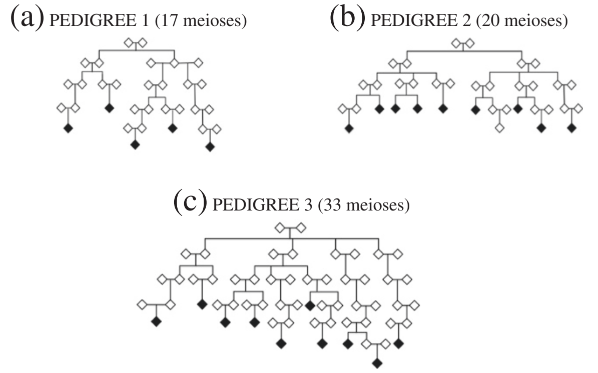
(a) PEDIGREE 1 (17 meioses)
(b) PEDIGREE 2 (20 meioses)
(c) PEDIGREE 3 (33 meioses)

Figure 1 Three extended high-risk breast cancer pedigrees.

Informed consent was obtained from all participants in this study. This study is approved by the Institutional Review Board at University of Utah.

Control Samples for estimation of LD

In SGS approaches, control samples are required to estimate genomewide LD structure that is used in the empirical assessment of significance. The primary set of controls used was ascertained locally via the UPDB resource (also the source of the pedigrees). These control individuals were known to be cancer-free and were self-declared Caucasian. These 224 'local controls' comprised 117 males and 107 females. To ensure robustness of our findings, for regions of interest identified from the genomewide pSGS analysis based on the LD model estimated from the local controls, we also assessed significance based on a set of genomically-matched controls. This second set of controls comprised individuals selected from the Illumina Genotype Control Database (iControlDB) (www.illumina.com). Principal components analysis was carried out on the set of all self-declared Caucasian samples in iControlDB with genotype data available for the 550K Illumina SNP array. We pruned the 550K Illumina SNPs to a set with $r^2<0.5$ and used *smartpca* [13] to extract the first two principal components and identified 1,490 iControlDB individuals who resided within 3 standard deviations of the centroid based on a bivariate normal distribution estimated from the cases. These 1,490 genomically matched controls comprised 949 female and 541 males. Figure 2 illustrates the 1,490 iControlDB individuals selected.

Genotypes

SNPs from the Illumina 550K array were used. SNPs with a significantly different missing data rate between cases and controls ($p < 10^{-5}$), those with a missing rate greater than 5%, a minor allele frequency (MAF) of less

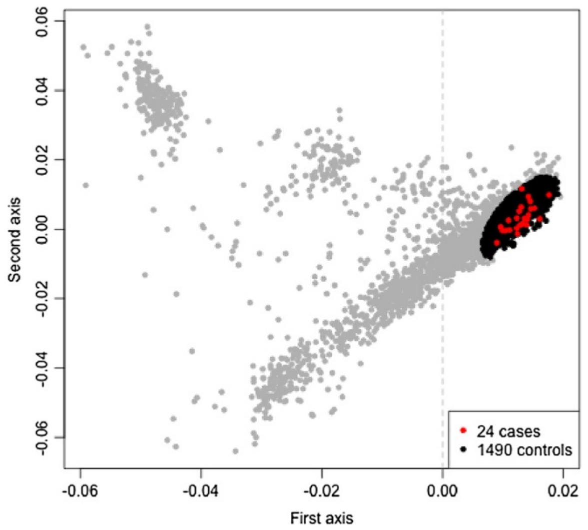

Figure 2 Principal Component Analysis for selection of iControlDB individuals. The first and second principal components The 24 breast cancer cases from the three pedigrees are shown in red, and the selected iControlDB individuals are shown in black.

than 1%, or significantly different from Hardy-Weinberg Equilibrium (p < 10^{-4}) were removed. This resulted in a total of 516,475 SNPs genomewide included in the SGS analyses.

Data analysis

Our primary analysis was a genomewide pSGS analysis. For regions of interest identified by pSGS with at least nominal evidence (p≤0.05) we also performed SGS and multipoint linkage analysis as secondary analyses, for comparison.

Shared genomic segment analysis

Thomas et al. [8] proposed a method of SGS analysis based on sharing among all cases in high-risk pedigrees. Assuming biallelic SNP loci with alleles 1 and 2, the three possible genotypes at each locus are 11, 12 and 22. Sharing is impossible between individuals with the two opposite homozygote types (11 and 22), otherwise IBS sharing exists. Therefore, the number of individuals sharing at a locus can be easily calculated on inspection of the number of homozygote individuals at each locus. We define S_i to be the number of cases sharing at least one allele IBS at SNP i.

$$S_i = N - \min\left(N_{11}^{i} N_{22}^{i}\right)$$

where N is the total number of cases in a pedigree, and N_{11}^i, N_{22}^i are the counts of cases homozygous 11 and 22, respectively. Missing genotypes are treated as heterozygotes.

We use R_i (t) to indicate the number of consecutive SNPs (which includes the ith SNP) with IBS sharing among at least t cases (also referred to as a “run length”), where t is usually the total number of individuals whose genotypes are in comparison (t = N). We recently introduced a new SGS test statistic, the weighted mean pairwise Shared Genomic Segment (pSGS) statistic [9]. It combines evidence from sharing in pairs of cases, weighted by their genetic distance and hence is less influenced by and has improved robustness to intrafamilial heterogeneity. Consider a pedigree with N cases, and denote d_{jk} as the number of meioses between cases j and k, and $R_i^{jk}(2)$ as the run length shared by the pair of cases j,k at locus i. The test statistic for the pSGS is defined as follows:

$$pSGS_i = \frac{1}{\binom{N}{2}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} d_{jk}} \sum_{j=1}^{N-1}\sum_{k=j+1}^{N} d_{jk} R_i^{jk}(2)$$

The significance is assessed empirically based on expected sharing under a model that includes LD as described in Thomas [14]. Our methodology is implemented in freely available java software (http://balance.med.utah.edu/wiki/index.php/Access_programs_by_name).

pSGS and SGS: LD model

We used FitGMLD to obtain a LD model based on the 224 local control samples using default parameters [15]. This program applies graphical models to estimate a general finite multivariate distribution for allelic association between genetic loci in each autosomal chromosome. In the model, the variables are alleles at each SNP loci, which are indicated using nodes. Edges connect loci that are in LD with each other and SNPs in a chromosome are modeled using a Markov graph. The program iteratively performs phase imputation and estimation of LD model from genotype dataset of unrelated individuals. The method incorporates an error model for genotyping. The program takes computation time in the magnitude of $O(nm)$, given n individuals with m genotyped markers [15].

pSGS and SGS: Significance assessment

We estimated nominal p-values for each locus using Monte Carlo procedures, by comparing the observed lengths to expected lengths under the null. Sharing under the null was achieved using a gene-dropping procedure assuming random mating, Mendelian inheritance and a genetic map for recombination. Founder haplotypes in the pedigree were generated using the estimated LD model. These were segregated through the known pedigree structure using random Mendelian inheritance to generate genotypes for each descendant in the pedigree. Recombinant events were based on an established genetic map [14]. Simulated genotypes were only retained for the studied cases in each pedigree and SGS statistics were calculated using the null data configurations to generate a distribution of lengths shared under the null for each pedigree.

The simulation procedure was implemented using a parallel Java program to improve computational efficiency.

pSGS and SGS: Genomewide thresholds

Genomewide thresholds provide a correction for the multiplicity of tests performed across the genome. For SGS methods, the multiple testing corresponds to the number of SGS segments across the genome, and this depends on the pedigree structure (number of meioses between the studied cases) and the sharing statistic considered (pSGS or SGS). Hence, we estimated genomewide thresholds empirically for each pedigree for both pSGS and SGS. A genomewide significant threshold was defined as the level of significance that would be achieved at a rate of 0.05 times per genome under the null (false positive rate per genome, μ=0.05). A genomewide suggestive threshold was defined as the level of significance achieved at a rate of 1 per genome under the null (μ=1.0). To estimate these thresholds we generated 1,000 null genome configurations for each pedigree (matched to the real genetic data for LD and recombination model), performed SGS and pSGS, identified the shared segments and their respective p-values (with p-values estimated based on an empirical distribution of up to 1,000,000 null values). For each pedigree and statistic, the p-values for all segments across all 1,000 genomes were ranked. We identified the 50th ranked p-value across all 1,000 genomes (50/1,000 = 0.05 per genome) to determine the level for the significant threshold; and the 1,000th ranked p-value (1,000/1,000 = 1 per genome) to determine the suggestive threshold.

Linkage analysis

We also performed multipoint linkage analysis on each pedigree. In order to eliminate inflation of linkage statistics due to LD, a pruned set of SNPs (n=26,177) were used for the linkage analysis. This set of "LD-free" SNPs had a minimum spacing of 0.1 cM, a minimum heterozygosity of 0.3 (to maintain good information content), and a maximum r^2 of 0.16 over a sliding 500 kb window in the public available HapMap CEU data, and exceeded an individual call rate of 98% of genotyped subjects. We used an established genetic map [16], plus linearly interpolated SNPs from Human Genome Build 35.1. Allele frequencies were estimated from all genotyped individuals at each SNP. The multipoint linkage analysis was performed using MCLINK, a multipoint Markov chain Monte Carlo (MCMC) linkage method that can analyze extended pedigrees [17]. A cases-only parametric analysis was performed based on a general dominant model.

Results

For all three pedigrees, nominal evidence was considered to be $p\leq0.05$. For pedigree 1, the empirical genomewide suggestive and significant thresholds for pSGS were $p=6.5\times10^{-3}$ and $p=3.0\times10^{-4}$, respectively, and the genomewide suggestive and significant thresholds for SGS were $p=1.3\times10^{-4}$ and $p<1.0\times10^{-6}$. For pedigree 2, no results surpassed the nominal threshold therefore empirical genomewide thresholds were not determined. For pedigree 3, the empirical genomewide suggestive and significance thresholds for pSGS were $p=5.0\times10^{-3}$ and $p=2.5\times10^{-4}$, respectively, and the suggestive and significant thresholds for SGS were estimated as $p=3.8\times10^{-5}$ and $p<1.0\times10^{-6}$. Genomewide thresholds for suggestive and significant linkage signals have been previously established to be LODs of 1.86 ($p=1.7\times10^{-3}$) and 3.30 ($p=4.9\times10^{-5}$), respectively [18].

Figure 3 shows the genomewide pSGS results for each pedigree based on a LD model estimated from the local controls. Table 1 illustrates all pSGS regions containing at least nominal evidence. For each of these regions, Table 1 also summarizes the best SGS p-value in the region (all *N* cases sharing) and the multipoint LOD score from linkage analysis. Table 2 shows a comparison between the pSGS p-values attained based on the local controls LD model and those based on the iControlDB individuals LD model and indicates that our results are extremely robust.

Four genomewide suggestive pSGS regions were identified, with an additional two borderline. One of these regions was also genomewide suggestive in the SGS analysis, and one was genomewide suggestive using a dominant linkage analysis (Table 1). Three of the genomewide suggestive pSGS results were found in pedigree 1 on chromosome 4 (37.5-54.6 Mb; p=0.006; μ=0.98, indicating that a finding this extreme would be expected 0.98 times per genome under the null), chromosome 7 (16.7-31.2 Mb; p=0.005; μ=0.82) and chromosome 8 (38.3-122.6 Mb; p=0.0025; μ=0.48). Hence for pedigree 1, three genomewide suggestive regions were observed, compared to less than 1 expected under the null. One genomewide suggestive region was identified in pedigree 3 on chromosome 2 (74.8-163.0 Mb; p=0.004; μ=0.92), in addition two borderline suggestive findings were also identified on chromosome 7 (11.4-96.7 Mb; p=0.0065; μ=1.08, an overlap with the genomewide suggestive region in pedigree 1) and chromosome 14 (56.9-99.3 Mb; p=0.007; μ=1.25, an overlap with a nominal region in pedigree 1). Hence, for pedigree 3, three regions were found compared to a

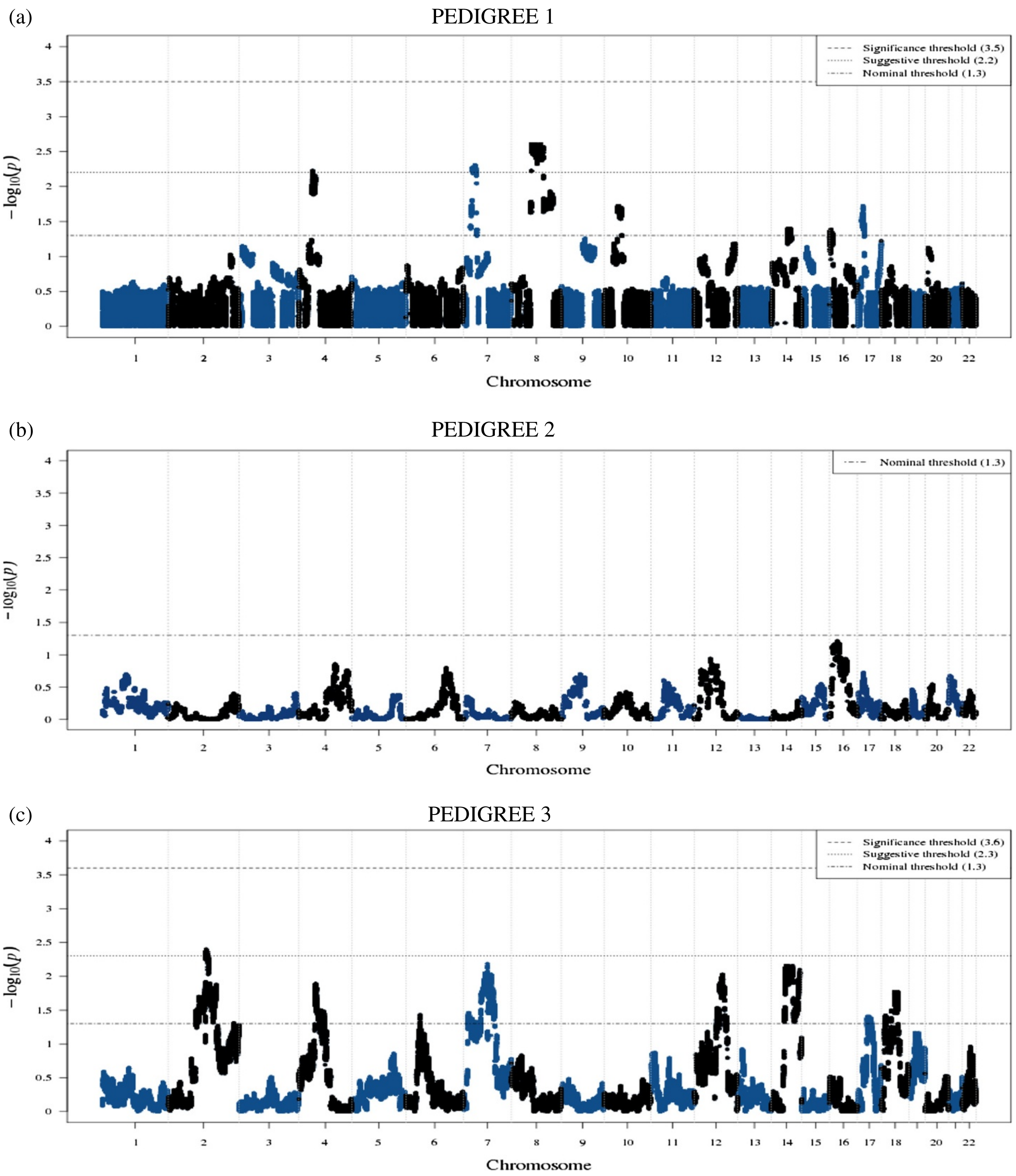

Figure 3 Genomewide pSGS results. For pedigree 1, genomewide suggestive and significance pSGS thresholds were estimated to be p=6.5×10^{-3} (–log$_{10}$p=2.2) and p=3.0×10^{-4} (–log$_{10}$p=3.5), respectively. For pedigree 3, genomewide suggestive and significance pSGS thresholds were estimated to be p=5.0×10^{-3} (–log$_{10}$p=2.3) and p=2.5×10^{-4} (–log$_{10}$p=3.6), respectively.

Table 1 Regions with at least nominal evidence in pSGS (p≤0.05)

Pedigree (number cases; min. possible sharing)[†]	Chromosome	Region[€]	Length (Mb)	pSGS p-value[¥]	Average cases sharing (range)	SGS	Linkage
1	**4***	**37,542,764 - 54,575,432**	**17.0**	**0.0060**	**4.78 (3–5)**	0.00017	1.19
(n=5; min=3)	**7***	**16,704,212 - 31,213,647**	**14.5**	**0.0050**	**4.83 (3–5)**	0.00010	**2.62**
	8	**38,344,499 - 122,638,989**	**83.3**	**0.0025**	**4.78 (3–5)**	**0.000018**	1.55
	10	28,738,098 - 49,576,878	20.8	0.019	4.78 (3–5)	0.0020	1.25
	14*	66,272,834 - 77,581,481	11.3	0.040	4.78 (3–5)	0.029	1.19
	16	359,567 - 8,197,462	7.8	0.041	4.76 (3–5)	0.009	1.03
	17	10,784,088 - 16,883,680	6.1	0.019	4.81 (3–5)	0.0015	1.30
3	**2**	**74,758,934 - 162,960,873**	**88.2**	**0.0040**	**9.41 (6–10)**	0.00014	1.66
(n=10; min=5)	4*	47,003,076 - 88,807,556	41.8	0.013	9.44 (5–10)	0.00033	0.89
	6	31,320,810 - 31,628,733	0.3	0.037	9.60 (7–10)	0.0062	0.19
	7*	***11,358,235 - 96,674,424***	***85.3***	***0.0065***	***9.40 (6–10)***	0.00025	1.04
	12	67,987,630 - 101,376,241	33.4	0.0095	9.44 (6–10)	0.00020	1.27
	14*	***56,883,760 - 99,254,712***	***42.4***	***0.0070***	***9.43 (6–10)***	0.00050	0.57
	17	32,760,735 - 51,072,912	18.3	0.039	9.33 (6–10)	0.0014	0.03
	18	8,247,249 - 50,460,551	42.2	0.017	9.43 (6–10)	0.0011	1.32

[†] For each pedigree the total number of cases in the pedigree, and the minimum possible number of cases that can share (i.e. the number of cases sharing cannot go below this value) are shown.
[€]Coordinates are based on GRCh37/hg19.
[¥]Significance based on a LD map from 224 local controls.
*Overlapping regions are indicated by an asterisk.
For pedigree 1, genomewide pSGS thresholds are p=6.5×10^{-3} and p=3.0×10^{-4}, for suggestive and significance, respectively, and corresponding SGS thresholds are p=1.3×10^{-4} and p<1.0×10^{-6}.
For pedigree 3, genomewide pSGS thresholds are p=5.0×10^{-3} and p=2.5×10^{-4}, for suggestive and significance, respectively, and corresponding SGS thresholds are p=3.8×10^{-5} and p<1.0×10^{-6}.
For linkage, genomewide suggestive and significance LOD thresholds are 1.86 and 3.30, respectively, corresponding to p-values of 1.7×10^{-3} and 4.9×10^{-5}.
Genomewide suggestive signals are indicated in **bold**. Borderline genomewide suggestive signals are ***bold and italicized***.

false positive rate of 1.25 or less per genome (3 observed, 1.25 expected). Even accounting for the multiple testing of analyzing three independent pedigrees, we identified 6 signals (5 distinct chromosomal regions) with μ≤1.25, which is greater than the 3.75 would be expected by chance at this significance level.

For all 15 regions shown in Table 1, the average number of cases sharing across the regions was high (> N-1 cases), although the range of the number of cases sharing was wide; generally spanning the total range possible. The size of the shared regions also varied quite widely; for the six regions of interest 14.5 – 88.2 Mb (Table 1). Shared regions were defined as the segment of contiguous loci remaining above nominal statistical evidence. Under the null hypothesis (no disease locus) and the assumption that recombinations at each meiosis occur as independent Poisson processes, the expected length of a shared IBD segment is Exponentially distributed with mean $1/d$ Morgans, where d is the number of meiosis separating the individuals. For example, a cousin-pair (d=4) will share segments of size 25 cM on average. Under the alternate hypothesis that a disease locus exists, the length follows a Gamma distribution with mean $2/d$. In the cousin-pair example, the average segment length surrounding a shared disease locus would be 50 cM. In our pedigrees, breast cancer pairs ranged from siblings (d=2), to pairs separated by 11 meioses (Figure 1). Importantly, it should be noted that our SGS analysis, by design, identifies IBS segments (a less stringent criteria than IBD), with an aim is to identify excessively long regions that are therefore likely to be IBD. For the above reasons, the region lengths we identify may be longer than expected by chance for the given relationships.

Three regions on chromosomes 4, 7 and 14 showed overlapping evidence in pedigrees 1 and 3 (Table 1). To investigate these three regions for evidence of common sharing across pedigrees, we performed two-pedigree pSGS analyses across all cases in pedigrees 1 and 3. Because there were no known genealogical links between these pedigrees, the pSGS statistic in these two-pedigree analyses could not be weighted by the number of meioses between cases, so an un-weighted paired average method was used. Table 3 illustrates the results of these two-pedigree analyses. All regions remained at least nominally significant indicating that the underlying risk variants could be the same in the two pedigrees. Of particular note was the 7.6 Mb region on chromosome 4 that increased in significance, despite the potential loss of power due to our inability to weight the sharing by meioses in the two-pedigree analysis.

Table 2 Comparison of pSGS p-values: local controls and iControlDB individuals

Pedigree	Chromosome	Region	Length (Mb)	Local Controls (n=224)	iControlDB (n=1,490)
				pSGS	pSGS
1	4*	37,542,764 - 54,575,432	17.0	0.0060	0.015
	7*	16,704,212 - 31,213,647	14.5	0.0050	0.0060
	8	38,344,499 - 122,638,989	83.3	0.0025	0.0013
	10	28,738,098 - 49,576,878	20.8	0.019	0.014
	14*	66,272,834 - 77,581,481	11.3	0.040	0.041
	16	359,567 - 8,197,462	7.8	0.041	0.044
	17	10,784,088 - 16,883,680	6.1	0.019	0.017
3	2	74,758,934 - 162,960,873	88.2	0.0040	0.0033
	4*	47,003,076 - 88,807,556	41.8	0.013	0.013
	6	31,320,810 - 31,628,733	0.3	0.037	0.035
	7*	11,358,235 - 96,674,424	85.3	0.0065	0.0055
	12	67,987,630 - 101,376,241	33.4	0.0095	0.0095
	14*	56,883,760 - 99,254,712	42.4	0.0070	0.0055
	17	32,760,735 - 51,072,912	18.3	0.039	0.032
	18	8,247,249 - 50,460,551	42.2	0.017	0.012

Discussion

We investigated three extended Utah high-risk breast cancer pedigrees using weighted pSGS analysis to identify regions of excessive sharing that could potentially harbor breast cancer susceptibility loci. Five regions of interest were identified on chromosomes 2, 4, 7, 8 and 14. Three of these regions (chromosomes 4, 7 and 14) showed evidence for excessive sharing in two pedigrees (pedigrees 1 and 3), with chromosome 4 being perhaps of particular interest because the region gained significance in the two-pedigree analysis. All five of these regions have either been identified previously in genomewide searches or candidate susceptibility genes reside in them. Our region on chromosome 4 is supported by evidence from two previous genomewide linkage studies of families not attributable to *BRCA1* or *BRCA2*. A large international multi-center linkage study of 149 breast cancer families identified the chromosome 4 region as the best linkage across the genome (LOD=1.8) [19]. This location on chromosome 4 was also reported as one of the top candidate regions in another genomewide linkage scan (LOD=1.3) [20]. In addition, our region includes cytogenetic band 4q12 which has previously been proposed as a location potentially harboring genes important in breast cancer development because of observed loss of heterozygosity at 4q12 in both *BRCA1/2* and sporadic breast cancer tumors [21,22]. Furthermore, there has been recent interest in two candidate genes in this region, with increased gene copy number for genes *KIT* and *VEGFR2* found in triple negative breast cancer, an aggressive and difficult to treat form of the disease [23]. Our region on chromosome 14q includes the breast cancer candidate gene *RAD51L1*, which contains one of the two most significant associations reported in a multi-stage genomewide association study of 9,770 cases and 10,799 controls [24]. Our region on chromosome 7 contains the *AHR* gene that has been associated with breast cancer risk [25,26], and *IL6* [27], which contains a marker associated with increased risk for breast carcinoma [28]. Our 88.2 Mb region on chromosome 2 includes the gene *ZEB2* that is involved in RAS pathway that has been proposed as involved in clinical breast cancer progression [29]. There are also two SNPs (rs17188434 and rs12472911) that have been associated with age at menarche in this region [30], and early menarche is suggested to be a risk factor for breast cancer. The large 83.3 Mb region on chromosome 8 encompasses multiple possible breast cancer candidate

Table 3 pSGS results for two-pedigree analyses including pedigrees 1&3 in the overlapping regions

Chr	Region	Length (Mb)	pSGS pedigree 1	pSGS pedigree 3	Two-ped-pSGS (local controls)	Av. sharers in 2-ped analysis (range)
4	47,003,076 - 54,575,432	7.6	0.0060	0.013	0.0025	14.14 (10–15)
7	16,704,212 - 31,213,647	14.5	0.0050	0.0065	0.0076	14.00 (9–15)
14	66,272,834 - 77,581,481	11.3	0.040	0.0070	0.011	14.04 (9–15)

genes: for example, *POLB* [31] and *NBS1* (NBN) [32,33] have previously been implicated in heritable susceptibility to breast cancer; *EBAG9* has been suggested to be involved in early stage breast cancer [34].

Seven nominal regions were also identified in our analyses. Notably two of these regions are on chromosome 17. One of the chromosome 17 regions in pedigree 1 (10.8-19.9 Mb) contains the candidate gene, *ELAC2*, which was previously proposed as a susceptibility gene for prostate cancer using Utah high-risk pedigrees [35]. Genes that increase susceptibility to both breast and prostate cancer have been observed previously; for example, *BRCA2* [36]. The second chromosome 17 region was found in pedigree 3 (32.8-51.1 Mb) and contains the high-risk breast cancer gene, *BRCA1*. It is perhaps surprising that a region containing *BRCA1* would arise, given our aim to screen out families with known BRCA mutations. In agreement with our selection criteria, we show no linkage at this locus (LOD=0.03). Nonetheless, it is possible that *BRCA1* remains a potential factor for risk in this pedigree.

We selected weighted pairwise SGS as our primary analysis specifically because the original SGS method will lose power quickly with intra-familial heterogeneity, and breast cancer is known to be a complex and very heterogeneous disease. In line with this assumption, only one of our genomewide suggestive pSGS regions also showed genomewide suggestive evidence using the original SGS algorithm. Furthermore, while the number of sharers across our regions of interest remained high, the range was wide and often reduced to the minimum possible number sharing. Hence, it appears that the pairwise algorithm may have been successful at providing more robustness to noise from heterogeneity, in addition to any residual genotyping error. One of our most significant pSGS regions (chromosomes 2) also showed genomewide suggestive evidence using multipoint linkage analysis with a dominant model.

An advantage of a pedigree design for gene identification is that a small number of cases and a well-delimited region can be easily defined and increases the efficiency of downstream experiments. Sequencing multiple cases selected for their high likelihood of sharing the underlying susceptibility variant provides an additional and powerful filter that can be used to parse findings from sequencing efforts. Hence, follow-up regionally-focused sequencing of the most compelling of these regions is a cost-effective and logical next step to identify the critical underlying risk variant at these loci.

Conclusions

Our pSGS analyses have highlighted several regions that have the potential to harbor susceptibility variants for breast cancer, some of which confirm loci previously proposed by others. Three of our most significant regions (chromosomes 4, 7 and 14) were observed in two pedigrees and show evidence for shared risk variants across those pedigrees. Arguably, these three regions in pedigrees 1 and 3 are particularly good candidates to pursue using regionally-focused sequencing to identify novel breast cancer risk variants. In addition, and more broadly, this study has illustrated the potential utility of our new weighted pSGS method and extended pedigrees for gene mapping in complex diseases.

Abbreviations

IBD: Identity-By-Descent; IBS: Identity-By-State; iControlDB: Illumina Genotype Control Database; LD: Linkage Disequilibrium; MAF: Minor Allele Frequency; MCMC: Markov chain Monte Carlo; pSGS: pairwise Shared Genomic Segment; SGS: Shared Genomic Segment; SNP: Single Nucleotide Polymorphism; UCR: Utah Cancer Registry; UPDB: Utah Population Database.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZC performed the statistical analyses, drafted the manuscript and participated in algorithm and parallelization software development. AT participated in algorithm development, study design and statistical oversight. CT performed statistical analyses. JF contributed to data coordination and processing. LACA conceived of the study. NJC had oversight for study design, algorithm development, statistical analyses and manuscript writing. All authors read and approved the final manuscript.

Acknowledgements

Data collection for the high-risk breast cancer pedigrees was made possible by the UPDB and the UCR. Partial support for all datasets within the UPDB was provided by the University of Utah Huntsman Cancer Institute (HCI) and the HCI Cancer Center Support grant, P30 CA42014 from the NCI. The UCR is funded by contract HHSN261201000026C from the NCI SEER program with additional support from the Utah State Department of Health and the University of Utah. The genotyping and analysis was supported by funding from R01GM81417 (to AT), R01CA134674, R01 CA163353 and R21CA152336 (to NJC). LC-A acknowledges support from the Huntsman Cancer Foundation. The project was supported by Award Number P30CA042014 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We also thank University of Utah Center for High Performance and Dr. Julio Facelli for Parallel computing support, and Alex Bigelow for data processing support.

Author details

[1]Department of Biomedical Informatics, University of Utah Medical School, Health Sciences Education Building, Salt Lake City, UT 84112, USA. [2]Division of Genetic Epidemiology, University of Utah Medical School, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA. [3]George E. Wahlen Department of Veterans Affairs Medical Center, 500 Foothill Drive, Salt Lake City, UT 84148, USA.

Received: 4 May 2012 Accepted: 23 November 2012
Published: 28 November 2012

References

1. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, Barlow WE, Geller BM, Kerlikowske K, Edwards BK, Lynch CF, Urban N, Chrvala CA, Key CR, Poplack SP, Worden JK, Kessler LG: **Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database.** *AJR Am J Roentgenol* 1997, **169**(4):1001–108. PubMed PMID: 9308451.
2. McClellan J, King MC: **Genetic heterogeneity in human disease.** *Cell* 2010, **141**(2):210–217. PubMed PMID: 20403315.

3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**(7265):747–753. Review. PubMed PMID: 19812666; PubMed Central PMCID: PMC2831613.
4. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *Am J Hum Genet* 2001, **69**(1):1–14. Epub 2001 Jun 14. Review. PubMed PMID:11410837; PubMed Central PMCID: PMC1226024.
5. Iles MM: **What can genome-wide association studies tell us about the genetics of common disease?** *PLoS Genet* 2008, **4**((2):e33. PubMed PMID: 18454206; PubMed Central PMCID: PMC2323402.
6. Knight S, Abo RP, Abel HJ, Neklason DW, Tuohy TM, Burt RW, Thomas A, Camp NJ: **Shared Genomic Segment Analysis: The Power to Find Rare Disease Variants.** *Ann Hum Genet* 2012, doi:10.1111/j.1469-1809.2012.00728.x. [Epub ahead of print] PubMed PMID: 22989048.
7. Wijsman EM, Amos CI: **Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions.** *Genet Epidemiol* 1997, **14**(6):719–735. Review. PubMed PMID: 9433569.
8. Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA: **Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays.** *Ann Hum Genet* 2008, **72**(Pt 2):279–87. Epub 2007 Dec 18. PubMed PMID: 18093282; PubMed Central PMCID: PMC2964273.
9. Cai Z, Knight S, Thomas A, Camp NJ: **Pairwise shared genomic segment analysis in high-risk pedigrees: application to Genetic Analysis Workshop 17 exome-sequencing SNP data.** *BMC Proc* 2011, **9**(5):S9. [Epub ahead of print] PubMed PMID: 22373081; PubMed Central PMCID: PMC3287931.
10. Allen-Brady K, Camp NJ: **Characterization of the linkage disequilibrium structure and identification of tagging-SNPs in five DNA repair genes.** *BMC Cancer* 2005, **5**:99. PubMed PMID: 16091150; PubMed Central PMCID: PMC1208870.
11. Skolnick M: **The Utah genealogical database: a resource for genetic epidemiology.** *Banbury report* 1980, **4**:285–287.
12. Cannon Albright LA: **Utah family-based analysis: past, present and future.** *Hum Hered* 2008, **65**(4):209–220. Epub 2007 Jan 11. Review. PubMed PMID: 18073491.
13. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**(8):904–909. Epub 2006 Jul 23. PubMed PMID: 16862161.
14. Thomas A: **Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium.** *Genet Epidemiol* 2010, **34**(2):119–124. PubMed PMID:19582786; PubMed Central PMCID: PMC2811755.
15. Abel HJ, Thomas A: **Accuracy and computational efficiency of a graphical modeling approach to linkage disequilibrium estimation.** *Stat Appl Genet Mol Biol* 2011, **0**(1):5. Epub 2011 Jan 6. PubMed PMID: 21291415; PubMed Central PMCID: PMC3045084.
16. Duffy DL: **An integrated genetic map for linkage analysis.** *Behav Genet* 2006, **36**(1):4–6. Epub 2006 Mar 8. PubMed PMID: 16523245.
17. Thomas A, Gutin A, Abkevich V, Bansal A: **Multilocus linkage analysis by blocked Gibbs sampling.** *Statistics and Computing* 2000, **10**(3):259–269.
18. Lander E, Kruglyak L: **Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.** *Nat Genet* 1995, **11**(3):241–7. PubMed PMID: 7581446.
19. Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, Newman B, Chenevix-Trench G, kConFab Investigators, Szabo C, Southey M, Renard H, Odefrey F, Lynch H, Stoppa-Lyonnet D, Couch F, Hopper JL, Giles GG, McCredie MR, Buys S, Andrulis I, Senie R, BCFS, BRCAX Collaborators Group, Goldgar DE, Oldenburg R, Kroeze-Jansema K, Kraan J, Meijers-Heijboer H, Klijn JG, van Asperen C, *et al*: **reast Cancer Susceptibility Collaboration (UK), Rahman N, Stratton MR. A genome wide linkage search for breast cancer susceptibility genes.** *Genes Chromosomes Cancer* 2006, **45**(7):646–55. ubMed PMID: 16575876; PubMed Central PMCID: PMC2714969.
20. Gonzalez-Neira A, Rosa-Rosa JM, Osorio A, Gonzalez E, Southey M, Sinilnikova O, Lynch H, Oldenburg RA, Van Asperen CJ, Hoogerbrugge N, Pita G, Devilee P, Goldgar D, Benitez J: **Genomewide high-density SNP linkage analysis of non-BRCA1/2 breast cancer families identifies various candidate regions and has greater power than microsatellite studies.** *BMC Genomics* 2007, **8**:299. PubMed PMID: 17760956; PubMed Central PMCID: PMC2072960.
21. Burger AM, Zhang X, Li H, Ostrowski JL, Beatty B, Venanzoni M, Papas T, Seth A: **Down-regulation of T1A12/mac25, a novel insulin-like growth factor binding protein related gene, is associated with disease progression in breast carcinomas.** *Oncogene* 1998, **16**(19):2459–67. PubMed PMID: 9627112.
22. Johannsdottir HK, Johannesdottir G, Agnarsson BA, Eerola H, Arason A, Johannsson OT, Heikkilä P, Egilsson V, Olsson H, Borg A, Nevanlinna H, Barkardottir RB: **Deletions on chromosome 4 in sporadic and BRCA mutated tumors and association with pathological variables.** *Anticancer Res* 2004, **24**(5A):2681–2687. PubMed PMID: 15521105.
23. Johansson I, Aaltonen KE, Ebbesson A, Grabau D, Wigerup C, Hedenfalk I, Rydén L: **Increased gene copy number of KIT and VEGFR2 at 4q12 in primary breast cancer is related to an aggressive phenotype and impaired prognosis.** *Genes Chromosomes Cancer* 2012, **51**(4):375–83. doi:10.1002/gcc.21922. Epub 2011 Dec 14 PubMed PMID: 22170730.
24. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, *et al*: **A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1).** *Nat Genet* 2009, **4**(5):579–584. Epub 2009 Mar 29. PubMed PMID: 19330030; PubMed Central PMCID: PMC2928646.
25. Dialyna IA, Arvanitis DA, Spandidos DA: **Genetic polymorphisms and transcriptional pattern analysis of CYP1A1, AhR, GSTM1, GSTP1 and GSTT1 genes in breast cancer.** *Int J Mol Med* 2001, **8**(1):79–87. PubMed PMID: 11408954.
26. Long JR, Egan KM, Dunning L, Shu XO, Cai Q, Cai H, Dai Q, Holtzman J, Gao YT, Zheng W: **Population-based case–control study of AhR (aryl hydrocarbon receptor)and CYP1A2 polymorphisms and breast cancer risk.** *Pharmacogenet Genomics* 2006, **16**(4):237–243. PubMed PMID: 16538170.
27. Knüpfer H, Preiss R: **Significance of interleukin-6 (IL-6) in breast cancer (review).** *Breast Cancer Res Treat* 2007, **102**(2):129–135. Epub 2006 Aug 23. Review. PubMed PMID: 16927176.
28. Snoussi K, Strosberg AD, Bouaouina N, Ben Ahmed S, Chouchane L: **Genetic variation in pro-inflammatory cytokines (interleukin-1beta, interleukin-1alpha and interleukin-6) associated with the aggressive forms, survival, and relapse prediction of breast carcinoma.** *Eur Cytokine Netw* 2005, **16**(4):253–260. PubMed PMID: 16464738.
29. Stinson S, Lackner MR, Adai AT, Yu N, Kim HJ, O'Brien C, Spoerke J, Jhunjhunwala S, Boyd Z, Januario T, Newman RJ, Yue P, Bourgon R, Modrusan Z, Stern HM, Warming S, de Sauvage FJ, Amler L, Yeh RF, Dornan D: **miR-221/222 targeting of trichorhinophalangeal 1 (TRPS1) promotes epithelial-to-mesenchymal transition in breast cancer.** *Sci Signal* 2011, **4**(186):pt5. Epub 2011 Aug 9. PubMed PMID: 21868360.
30. Elks CE, Perry JR, Sulem P, Chasman DI, Franceschini N, He C, Lunetta KL, Visser JA, Byrne EM, Cousminer DL, Gudbjartsson DF, Esko T, Feenstra B, Hottenga JJ, Koller DL, Kutalik Z, Lin P, Mangino M, Marongiu M, McArdle PF, Smith AV, Stolk L, Van Wingerden SH, Zhao JH, Albrecht E, Corre T, Ingelsson E, Hayward C, Magnusson PK, Smith EN, *et al*: **Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies.** *Nat Genet* 2010, **42**(12):1077–1085. PubMed PMID: 21102462; PubMed Central PMCID: PMC3140055.
31. Sliwinski T, Ziemba P, Morawiec Z, Kowalski M, Zadrozny M, Blasiak J: **Polymorphisms of the DNA polymerase beta gene in breast cancer.** *Breast Cancer Res Treat* 2007, **103**(2):161–166. Epub 2006 Nov 28. PubMed PMID: 17131038.
32. Roznowski K, Januszkiewicz-Lewandowska D, Mosor M, Pernak M, Litwiniuk M, Nowak J: **I171V germline mutation in the NBS1 gene significantly increases risk of breast cancer.** *Breast Cancer Res Treat* 2008, **110**(2):343–348. Epub 2007 Sep 26. PubMed PMID: 17899368.
33. Lu M, Lu J, Yang X, Yang M, Tan H, Yun B, Shi L: **Association between the NBS1 E185Q polymorphism and cancer risk: a meta-analysis.** *BMC Cancer* 2009, **9**:124. PubMed PMID: 19393077; PubMed Central PMCID: PMC2680905.
34. Tsuneizumi M, Emi M, Nagai H, Harada H, Sakamoto G, Kasumi F, Inoue S, Kazui T, Nakamura Y: **Overrepresentation of the EBAG9 gene at 8q23 associated with early-stage breast cancers.** *Clin Cancer Res* 2001, **7**(11):3526–3532. PubMed PMID: 11705872.

35. Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, Tranchant M, Woodland AM, Labrie F, Skolnick MH, Neuhausen S, Rommens J, Cannon-Albright LA: **A candidate prostate cancer susceptibility gene at chromosome 17p.** *Nat Genet* 2001, **27**(2):172–180. PubMed PMID: 11175785.

36. Kote-Jarai Z, Leongamornlert D, Saunders E, Tymrakiewicz M, Castro E, Mahmud N, Guy M, Edwards S, O'Brien L, Sawyer E, Hall A, Wilkinson R, Dadaev T, Goh C, Easton D, Collaborators UKGPCS, Goldgar D, Eeles R: **BRCA2 is a moderate penetrance gene contributing to young-onset prostate cancer: implications for genetic testing in prostate cancer patients.** *Br J Cancer* 2011, **105**(8):1230–1234. doi:10.1038/bjc.2011.383. Epub 2011 Sep 27. PubMed PMID: 21952622; PubMed Central PMCID: PMC3208504.

doi:10.1186/1471-2164-13-676
Cite this article as: Cai *et al.*: Pairwise shared genomic segment analysis in three Utah high-risk breast cancer pedigrees. *BMC Genomics* 2012 13:676.