



# Genome-wide scans of myopia in Pennsylvania Amish families reveal significant linkage to 12q15, 8q21.3 and 5p15.33

Anthony M. Musolf<sup>1</sup> · Claire L. Simpson<sup>1,2,3</sup> · Theresa A. Alexander<sup>1</sup> · Laura Portas<sup>1</sup> · Federico Murgia<sup>1</sup> · Elise B. Ciner<sup>4</sup> · Dwight Stambolian<sup>5</sup> · Joan E. Bailey-Wilson<sup>1</sup>

Received: 12 December 2018 / Accepted: 20 February 2019 / Published online: 2 March 2019

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019

## Abstract

Myopia is one of the most common ocular disorders in the world, yet the genetic etiology of the disease remains poorly understood. Specialized founder populations, such as the Pennsylvania Amish, provide the opportunity to utilize exclusive genomic architecture, like unique haplotypes, to better understand the genetic causes of myopia. We perform genetic linkage analysis on Pennsylvania Amish families that have a strong familial history of myopia to map any potential causal variants and genes for the disease. 293 individuals from 25 extended families were genotyped on the Illumina ExomePlus array and merged with previous microsatellite data. We coded myopia affection as a binary phenotype; myopia was defined as having a mean spherical equivalent (MSE) of less than or equal to  $-1$  D (diopters). Two-point and multipoint parametric linkage analyses were performed under an autosomal dominant model. When allowing for locus heterogeneity, we identified two novel genome-wide significantly linked variants at 12q15 (heterogeneity LOD, HLOD = 3.77) in *PTPRB* and at 8q21.3 (HLOD = 3.35) in *CNGB3*. We identified further three genome-wide significant variants within a single family. These three variants were located in exons of *SLC6A18* at 5p15.33 (LODs ranged from 3.51 to 3.37). Multipoint analysis confirmed the significant signal at 5p15.33 with six genome-wide significant variants (LODs ranged from 3.6 to 3.3). Further suggestive evidence of linkage was observed in several other regions of the genome. All three novel linked regions contain strong candidate genes, especially *CNGB3* on 8q21.3, which has been shown to affect photoreceptors and cause complete color blindness. Whole genome sequencing on these regions is planned to conclusively elucidate the causal variants.

## Introduction

Myopia, commonly called nearsightedness, is one of the most widespread eye disorders in the world. In myopic individuals, light focuses in front of the retina instead of directly on it. This causes distant images to be blurry and distorted. Myopia affects one quarter of Americans and prevalence

rates have been steadily rising since the 1970s (Vitale et al. 2009).

Myopia has been shown to be a complex phenotype with multiple genetic and environmental underpinnings (Stambolian 2013; Wojciechowski and Hysi 2013). Amount of time performing near work (e.g., reading) and education level have been shown to have deleterious effects, while outdoor activity has been shown to have protective effects (Stambolian 2013). On the genetic side, population-based genome-wide association studies (GWAS) have identified multiple risk loci for a negative refractive error (Fan et al. 2016; Kiefer et al. 2013; Miyake et al. 2015; Simpson et al. 2013, 2014; Stambolian et al. 2013; Verhoeven et al. 2012, 2013), although the vast majority of these associations are to common variants with a small to moderate effect. Family-based linkage studies have identified several linked regions that appear to harbor risk loci as well (Ciner et al. 2009; Guo et al. 2015; Musolf et al. 2017; Stambolian et al. 2004; Wojciechowski et al. 2006; Zhou et al. 2015), although the

---

Anthony M. Musolf and Claire L. Simpson are regarded as equal first authors.

---

Dwight Stambolian and Joan E. Bailey-Wilson are regarded as equal senior authors.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00439-019-01991-0>) contains supplementary material, which is available to authorized users.

---

✉ Joan E. Bailey-Wilson  
jebw@mail.nih.gov

Extended author information available on the last page of the article

actual causal variant(s) at the linked and/or associated loci mostly remain unknown.

Most complex traits are still analyzed using some form of population-based GWAS; the older linkage studies that were a hallmark of the pre-1990s genetic epidemiology studies are considerably less frequently used at present. Both the advent of inexpensive commercially available genotyping microarrays combined with the ease and low cost of collecting population-based samples when compared to collecting family-based samples are the main reasons GWAS have proliferated compared to linkage studies. However, the availability of genotype arrays with large numbers of rare variants and whole exome and whole genome sequencing have recently led to a resurgence in linkage studies, since these approaches can be powerful for detecting rare causal variants of moderate to large effect.

Both types of studies have advantages and limitations. GWAS are more effective at identifying common risk variants [minor allele frequency (MAF) > 0.05] but are severely underpowered for locating rare, highly penetrant variants. On the other hand, family-based linkage studies are better equipped to find these rare, highly penetrant variants than population-based studies. While a variant may be rare in a population at large, if a rare variant is present in a given family it will be much more common in family members than in the general population. Further, when analyzing genotype data that does not cover every variant in the genome, such as from genotyping chips, family-based linkage studies can use genomic architecture to their advantage better than population-based data for discovery of large regions harboring rare, causal variants. Individuals in a population are distantly related, and an extremely large number of meioses have broken up haplotypes into small blocks consisting of the variants that only exhibit the strongest linkage disequilibrium (LD). If an ungenotyped rare or common causal variant is not in high LD with the genotyped variants on the SNP genotyping chip, then the association methods will not exhibit high power to detect such a variant and when only one or two individuals in an entire GWAS study carry a rare variant, the power to detect its effect on the trait is minimal even if it is actually genotyped in the study. By contrast, in family-based studies, the haplotypes are determined by the haplotypes of the observed founders of each family; these haplotypes will be passed on to each subsequent generation. Since the majority of linkage studies are small (with families ranging from two to four generations), few recombinations will occur between variants that are moderately far apart on a chromosome and the end product is longer shared haplotypes across a genetic region among relatively close relatives. These shared haplotypes will boost power to find linkage between a trait and a causal variant along the haplotype even if the variant is not genotyped. In addition, founder populations, such as the Amish, may exhibit less

genetic heterogeneity for complex traits than in the general population.

Family-based linkage studies also have the advantage of observing linkage both within an individual family unit and across multiple families. In parametric linkage analyses, each family is analyzed as its own discrete unit (not as a group of cases and controls like population-based association studies) and each family receives its own LOD score. LOD scores are cumulative and can be added across families to get an overall LOD score. The cumulative LOD scores allow for more power to identify variants that are increasing disease risk in multiple families, whereas the individual family LOD scores allow for detection of a unique linkage only present in a single large family (genetic heterogeneity). This permits identification of any LOD scores that may be relatively large in a single family but are washed out or overwhelmed by multiple negative LOD scores in other families in the cumulative LODs. This is especially important in complex phenotypes that exhibit locus heterogeneity, where multiple families can reasonably be expected to have different causal variants and possibly different causal genes. This ability to identify individual risk variants within a family will become more and more important as we drive towards the age of personalized medicine. Greater emphasis will be placed upon our ability to quickly and accurately detect what is going on with a single patient and their family instead of at the population level.

Here, we present results from our family-based linkage study of myopia in the Myopia Family Study (MFS), using an exome-based array. The families studied here are Pennsylvania Amish, a founder population comprised primarily of German and Swiss immigrants. The Amish began settling in the Pennsylvania colony in the eighteenth century seeking religious tolerance and have become famous for their rural lifestyle and reluctance to adopt many aspects of modern technology. We have previously performed several genetic studies on Pennsylvania Amish families and refractive error/myopia, both of which have been shown to be heritable and strongly aggregate in these Amish families (Peet et al. 2007). All of our previous studies used microsatellite data. A previously published genome-wide linkage scan of a sparse map of microsatellite markers in these families found several suggestive linkages to myopia, including 3q26 (HLOD = 1.84), 11p13 (LOD = 1.54) and 8p23 with a LOD = 2.04 (Stambolian et al. 2005), which all independently confirmed Hammond et al.'s prior study's genome-wide significant evidence of linkage of refractive error to these regions in an English population (Hammond et al. 2004). A further fine-mapping study of the 1p34-36 region confirmed linkage to refractive error in our MFS Amish and Ashkenazi Jewish families (Wojciechowski et al. 2009) and a family-based association study found two significant variants on chromosomes 11 and 16 with both variants being located in matrix

metalloproteinase genes (Ciner et al. 2009). Here, we present for the first time genome-wide linkage results of these Pennsylvania Amish families using a dense exome microarray panel of marker genotypes, which has the advantage that a large number of moderately rare exonic variants as well as a backbone of common variants have been genotyped.

## Materials and methods

### Study design and patient recruitment

As part of the Myopia Family Study, 293 individuals from 25 extended Pennsylvania Amish families were recruited for this study. Details of the study recruitment are given in Stambolian et al. (2005). Briefly, these families were examined and enrolled in the study at the Amish Eye Clinic in Strasburg, Pennsylvania from January 1999 through August 2002. Eligibility in the study was limited to families that had at least three participating members and probands that had only one myopic parent and at least two myopic siblings. Any children under 5 years old were not permitted to participate. Both DNA samples and phenotype measurements were collected from all participants. The phenotype collection consisted of a comprehensive, dilated eye examination including refraction measurements. The specific measurement used in this study was mean spherical equivalent (MSE) in diopters (D). This value is calculated by adding the spherical refraction component to one-half of the cylindrical refraction component and averaging over both eyes. This study was approved by the institutional review boards of both the National Human Genome Research Institute and the University of Pennsylvania. All subjects provided informed consent and all procedures adhered to the Declaration of Helsinki.

### Genotyping and quality control

Previously extracted (Stambolian et al. 2005), stored DNA samples were genotyped at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University with the Illumina HumanExome-12v1-PlusCustomContent array. This array provided coverage of the exome and limited coverage of introns and intergenic regions. It is also highly enriched for rare variants; 6000 of the final cleaned variants had a MAF < 0.01 and 28,000 had a MAF < 0.2, making it an effective array for finding rare, highly penetrant variants of large effect on the phenotype. Individual families were genotyped on the same plate. Blind duplicates and HapMap controls were distributed across the plates for concordance checks. Gender mismatched samples or samples with unusual X and Y patterns (such as XXY individuals) were removed prior to release. SNPs with a genotype quality score

of less than 0.15 were recoded as missing; SNPs with cluster separation of less than 0.3 or heterozygote rate greater than 80% were dropped. SNPs were filtered to a call rate of 95% which resulted in a mean call rate of 99% per variant.

Seventy-two individuals who did not have genotype data were added to connect any disjointed pedigrees and insure proper relationships for the linkage analyses. These connecting individuals were known to exist from family history but for various reasons did not fully participate. In some cases, phenotype data were collected on these individuals. All these individuals were coded as having unknown genotypes at all variants and those without phenotype data were also coded as having an unknown phenotype. Identity-by-descent (IBD) values were calculated by PLINK (Purcell et al. 2007) to verify all familial relationships. Any variant with a single Mendelian inconsistency was dropped in that family; variants with multiple Mendelian errors were removed from the entire data set. PLINK was also used to detect any sex discrepancies and to remove all monomorphic markers.

All SNP genotypes were then merged with a previously published set of 367 genotyped sequence tagged sites (STS) or microsatellite markers (Stambolian et al. 2005). The final genotype data set consisted of 365 individuals (293 genotyped) from 25 extended families and 52,035 genotyped markers (51,668 SNPs and 367 STS). All markers were mapped onto the Rutgers Genetic Map v3 using GRCh37 physical positions (Matise et al. 2007).

For the linkage analyses, allele frequencies were calculated for the data set using sib-pair (Duffy 2008). Estimating allele frequencies from the data set to be analyzed has been shown to properly control type I error rates (Mandal et al. 2000, 2001, 2006), since the allele frequencies in the founders of the families to be studied may differ substantially from allele frequencies in databases, especially for families from ethnic groups that are not well represented in these databases. This is especially important for a founder population such as the Amish, because their allele frequencies at certain variants may deviate from those in the general European population.

### Myopia affection classification

We used a binary coding scheme (i.e., an individual was either affected with myopia or not) based on MSE thresholds. All participants with a refraction of  $-1.00$  D were coded as being affected with myopia. The refraction score was based on an actual eye examination or if the person was not available for the exam, medical records or glasses prescriptions were used instead. Adult participants that had a refractive score of  $0.00$  D or greater were coded as unaffected and those with a refractive score in the range of  $-1.00$  to  $0.00$  D were coded as having a missing or unknown phenotype. Extra precaution was given to unaffected children,

as normal developmental changes during childhood could lead to potential misclassification. Therefore, stricter thresholds were used for unaffected status. Children in the 5–10 years old range were considered unaffected if their MSE was 2.00 D or higher and unknown if their refractive score was between  $-1.00$  and  $2.00$  D. Children in the 11–20 age group were considered unaffected if their refractive score was  $1.50$  D or higher and unknown if the refractive score was between  $-1.00$  and  $1.50$  D. This is a conservative approach based on lack of a good segregation model of age-dependent penetrance and appropriate genotype probabilities for young unaffected subjects. The final data set contained 165 affected individuals, 116 unaffected individuals, and 84 individuals with unknown phenotype. The average MSE was  $-1.60$  D with a standard deviation of 2.52. There were 25 pedigrees—12 two generation pedigrees, 12 three generation pedigrees, and 1 four generation pedigree. 14 of the pedigrees had 7–10 individuals, 6 had 11–17 individuals, 3 had 20–22 individuals, 1 had 30 individuals and 1 had 73 individuals (the four generation pedigree). All pedigrees had at least three affected individuals.

### Parametric linkage analysis

Parametric two-point linkage analysis was performed on these families using the TwoPointLods software (Thomas 2006). We assumed an autosomal dominant mode of inheritance for our model. This was clearly evident from examination of the pedigrees, as the phenotype appeared in all generations and did not affect males or females at a significantly different rate. The disease allele frequency (DAF) was 0.0133 and penetrance was set at 90% for carriers with no phenocopies. We selected a rare disease allele frequency even though myopia itself is a common trait, since our objective in this study was to find a rare, highly penetrant variant with a large effect on myopia. Also, these families were selected for extreme aggregation of myopia in an apparent autosomal dominant pattern. Since such families are not common in the general population, it is reasonable to believe that a high penetrant myopia risk variant would be rare in the general population. The allele frequencies, penetrances, and autosomal dominant model were further confirmed by previous sensitivity analysis of this data set (Ibay et al. 2004; Stambolian et al. 2005; Wojciechowski et al. 2009) using microsatellites that confirmed these parameters as the best fit for this data. Further sensitivity analyses on these data using DAF = 0.01 and 0.02 showed no significant difference in LOD scores from the preferred model of DAF = 0.0133. Linkage analysis was performed on all families individually and LOD scores were added across families to form a cumulative LOD score at each variant. Heterogeneity LOD (HLOD) scores were also calculated across families

to account for any heterogeneity across families at a given variant.

Parametric multipoint analysis was performed using SimWalk2 (Sobel and Lange 1996; Sobel et al. 2001, 2002). To prevent type I error inflation caused by inter-marker linkage disequilibrium in a dense marker map (such as in this exome chip), the variants were collapsed into 1 cM bins. The SNP with the highest MAF (and thus highest information content) in the bin was chosen to represent the bin in the multipoint analysis. Further pruning removed any SNPs with a  $r^2$  value greater than 0.2. 3285 variants were left for analysis after the pruning process. The multipoint analysis assumed the identical parameters as the two-point analysis—autosomal dominant mode of inheritance, DAF = 0.0133, and penetrance matrix of 90% for carriers and 0% for non-carriers.

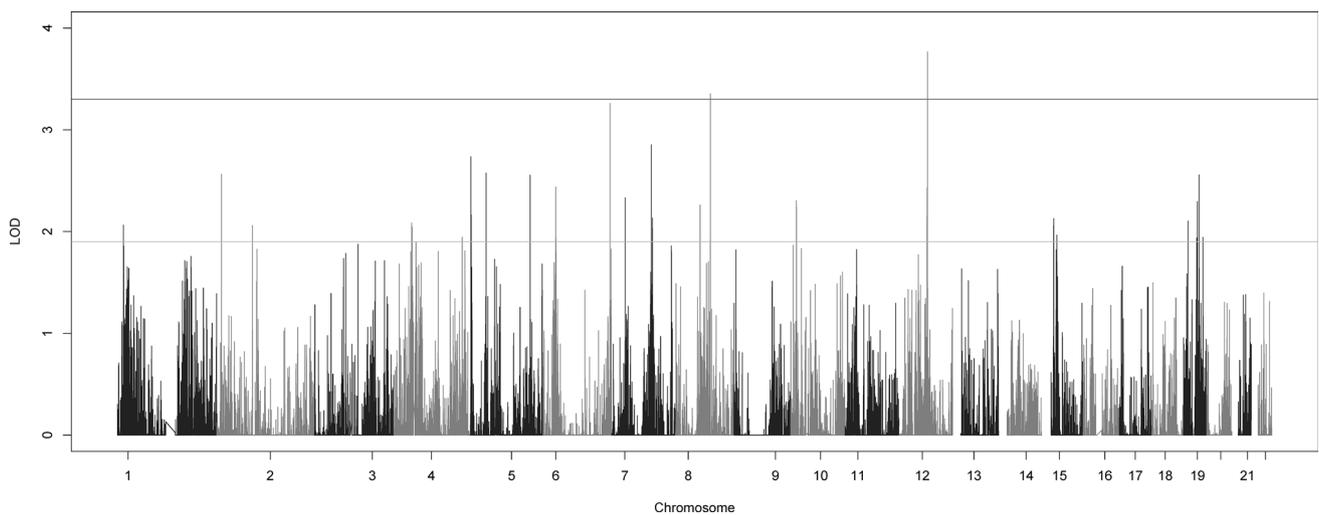
### Functional annotation of variants

All variants were annotated using wANNOVAR (Chang and Wang 2012; Wang et al. 2010), which provides information such as allele frequencies from 1000Genomes. It also provides protein predictions (i.e., whether a variant is predicted damaging) from multiple programs, including SIFT (Kumar et al. 2009; Ng and Henikoff 2003; Sim et al. 2012; Vaser et al. 2016; Zhao et al. 2013), PolyPhen2 (Adzhubei et al. 2013), MutationTaster (Schwarz et al. 2014), and REVEL (Ioannidis et al. 2016).

## Results

### Overall (across family) two-point linkage results

In analyses of all families combined, there were two genome-wide significantly linked variants located on chromosomes 12 and 8 and an additional 64 variants that showed suggestive linkage to myopia (Fig. 1) using two-point linkage. Here, we define genome-wide significant as a (H)LOD  $\geq 3.3$  and genome-wide suggestive as (H)LOD  $\geq 1.9$ , as recommended by Lander and Kruglyak (1995). The highest HLOD score (3.77) was located at rs2584021, an exonic SNP in the *PTPRB* gene at 12q15. The variant is non-synonymous and predicted benign by SIFT and PolyPhen2 but possibly damaging by MutationTaster. It is moderately rare in Europeans (MAF = 0.055 in 1000Genomes) and slightly enriched in our data set (MAF = 0.078). There was one other suggestive variant at 12q, rs1025016, an exonic non-synonymous SNP in *BEST3*. This variant was predicted damaging by multiple prediction algorithms including SIFT, PolyPhen2, and MutationTaster. It has a MAF of 0.08 in 1000Genomes and 0.05 in our data set. A LocusZoom style plot of the 12q15 signal with gene locations can be found in Supplemental Fig. 1 (Pruim et al. 2010).



**Fig. 1** Two-point genome-wide HLOD scores for all families. The genome-wide plot for the HLOD scores across all 25 families using two-point linkage analysis. The lines at 3.3 and 1.9 represent the thresholds for genome-wide significant and suggestive HLOD scores

The other significant variant was rs6471482 (HLOD = 3.35) a non-synonymous exonic SNP in *CNGB3* at 8q21.3. It is predicted benign by SIFT and PolyPhen2 but possibly damaging by MutationTaster. In 1000Genomes for Europeans, it has a MAF of 0.14 but is rare in other populations and only has an overall 1000Genomes MAF of 0.04. In our data set, it has a MAF of 0.08. Three other additional SNPs were suggestive at 8q21.3, including one in an intron of *CNGB3*. A zoomed plot of the 8q21.3 signal with gene locations can be found in Supplemental Fig. 2.

There was another SNP that exhibited linkage evidence that was just below genome-wide significance (HLOD = 3.26). This was rs2881194, an intronic SNP located in *RPS6KA2* at 6q27. The additional suggestive variants were spread among ten different chromosomes. There was a high concentration of suggestive linkage signals (28 variants) in the 6p21.32 region; however, this region is home to the human leukocyte antigen (HLA) complex of proteins. A list of significant and suggestive variants, along with selected annotations, can be found in Table 1.

### Individual two-point linkage family results

An advantage to having the large, multiplex families that comprised this study is that there is often sufficient power to observe significant linkage signals in an individual family. Indeed, we observed one family (Family 3316) that contained three genome-wide significant LOD scores (Fig. 2). Family 3316 (the family with the significant linkage peak on chromosome 5p) is a three generational pedigree with 30 people (10 affected, 9 unaffected, and 11 unknown) (Supplemental Fig. 3). All three variants were located within the *SLC6A18* gene at 5p15.33. rs74581452,

a synonymous exonic SNP, had the highest LOD score of 3.5. The SNP is moderately rare in 1000Genomes for Europeans (MAF = 0.03) and slightly enriched in our data set (MAF = 0.065). The other two variants were both non-synonymous exonic variants, rs7705355 (LOD = 3.47) and rs7728667 (LOD = 3.37). rs7705355 is moderately rare in Europeans (MAF = 0.04) and enriched in our data set (MAF = 0.08). rs7728667 is relatively common both in our data set (MAF = 0.14) and Europeans in general (MAF = 0.21). Both SNPs were predicted to be possibly damaging by MutationTaster. The three genome-wide significant SNPs are part of a larger linkage peak at 5p15.33 that contained an additional six suggestive SNPs (Fig. 2; Table 2, Supplemental Fig. 4).

No other families contained any genome-wide significant variants, though many contained multiple genome-wide suggestive LOD scores. Two signals deserve closer examination. First is the 1p signal in Family 3333, a three generation family with 13 people (5 affected and 8 unaffected) (Supplemental Fig. 5). This signal consists of a long linked haplotype across 1p36.22–32.3 (approximately 57 cM) consisting of 249 variants with LOD scores of 1.9 (Fig. 3, Supplemental Fig. 6). The figure also clearly shows that there are several other haplotypes of lower magnitudes (e.g., haplotypes at LOD = 1.6 and 1.3) and little to no negative linkage signal in the region. This suggests a lower probability of this particular signal being a false positive.

The other interesting individual family signal was in Family 3074, a two generation pedigree with 13 individuals (6 affected, 4 unaffected, and 3 unknown) (Supplemental Fig. 7). This signal was a very small linked haplotype of five variants at 3q25.1 (Fig. 4, Supplemental Fig. 8). All variants had a LOD score of 2.22, with four of the five

**Table 1** All significant and suggestive SNPs across all families

CH	SNP	HLOD	GENE	FUNC	EXON	FRQ	SIFT	POLPH	MUTR
12	rs2584021	3.77	<i>PTPRB</i>	Exonic	nonsyn	0.06	T	B	P
8	rs6471482	3.35	<i>CNGB3</i>	Exonic	nonsyn	0.14	T	B	P
6	rs2881194	3.26	<i>RPS6KA2</i>	Intronic		0.08			
7	rs7792939	2.85	<i>TMEM225B</i>	Intronic		0.13			
5	rs74581452	2.73	<i>SLC6A18</i>	Exonic	syn	0.03			
5	rs13188259	2.57	<i>FYB</i>	Intronic		0.17			
5	rs7705355	2.57	<i>SLC6A18</i>	Exonic	nonsyn	0.04	T	B	P
2	rs76644468	2.56	<i>GREB1</i>	Exonic	nonsyn	0.05	T	B	N
19	rs17628	2.56	<i>RPS16</i>	Exonic	syn	0.38			
5	rs10058728	2.55	<i>CSNK1A1</i>	Intronic		0.47			
8	rs10504826	2.52	<i>CNGB3</i>	Intronic		0.28			
7	rs2572023	2.48	<i>OR2AE1</i>	Exonic	nonsyn	0.49	T	B	P
6	rs2239804	2.44	<i>HLA-DRA</i>	Intronic		0.46			
12	rs1025016	2.43	<i>BEST3</i>	Exonic	nonsyn	0.08	D	D	D
5	rs7728667	2.36	<i>SLC6A18</i>	Exonic	nonsyn	0.21	T	B	P
7	rs17822236	2.33	<i>BMPER</i>	Intronic		0.17			
10	rs4750568	2.30	<i>MEIG1</i>	Exonic	nonsyn	0.33	T	P	P
19	rs10775583	2.29	<i>SBSN</i>	Exonic	nonsyn	0.36	T	B	P
8	rs2882217	2.26	<i>LOC100130298, CLVS1</i>	Intergenic		0.40			
19	rs2304220	2.24	<i>TIMM50</i>	Exonic	nonsyn	0.17	D	B	P
5	rs17586674	2.16	<i>LOC100506858, LSINCT5</i>	Intergenic		0.44			
6	rs2894239	2.13	<i>NOTCH4, LOC101929163</i>	Intergenic		0.41			
6	rs440169	2.13	<i>NOTCH4, LOC101929163</i>	Intergenic		0.41			
6	rs3115573	2.13	<i>NOTCH4, LOC101929163</i>	Intergenic		0.41			
7	rs148981077	2.13	<i>LRWD1</i>	Exonic	nonsyn	0.04	T	B	N
15	rs2279482	2.13	<i>FAM189A1</i>	Exonic	nonsyn	0.08	T	B	P
15	rs2306933	2.13	<i>FAM189A1</i>	Exonic	nonsyn	0.07	T	B	P
19	rs8102258	2.10	<i>ZNF844</i>	Exonic	nonsyn	0.08	D	B	P
6	rs537757	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs4713518	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs742582	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs2143465	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs505274	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs1033500	2.10	<i>C6orf10</i>	Exonic	nonsyn	0.37	T	B	P
6	rs2073046	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs1559876	2.10	<i>NOTCH4, LOC101929163</i>	Intergenic		0.36			
6	rs531094	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs9268368	2.10	<i>C6orf10</i>	Exonic	nonsyn	0.37	T	B	P
6	rs9405090	2.10	<i>C6orf10</i>	Exonic	nonsyn	0.37	T		P
6	rs547261	2.10	<i>LOC101929163</i>	ncRNA		0.37			
6	rs477005	2.09	<i>LOC101929163</i>	ncRNA		0.37			
6	rs2076540	2.09	<i>LOC101929163</i>	ncRNA		0.37			
6	rs560505	2.09	<i>C6orf10</i>	Exonic	nonsyn	0.37	T	B	P
6	rs2395110	2.09	<i>NOTCH4, LOC101929163</i>	Intergenic		0.42			
6	rs6929776	2.09	<i>LOC101929163</i>	ncRNA		0.38			
6	rs547077	2.09	<i>LOC101929163</i>	ncRNA		0.38			
4	rs10009336	2.08	<i>KCTD8, YIPF7</i>	Intergenic		0.18			
1	rs10489962	2.06	<i>FHAD1</i>	Exonic	nonsyn	0.06	T	B	P
2	rs2241883	2.06	<i>FABP1</i>	Exonic	nonsyn	0.35	T	B	P
4	rs1497574	2.04	<i>GABRG1</i>	Intronic		0.45			

**Table 1** (continued)

CH	SNP	HLOD	GENE	FUNC	EXON	FRQ	SIFT	POLPH	MUTR
4	rs976156	2.04	<i>GABRG1</i>	Exonic	syn	0.45			
4	rs1391174	2.04	<i>GABRG1</i>	Intronic		0.45			
7	rs3800939	2.03	<i>LRRIC17</i>	Exonic	nonsyn	0.04	T	P	P
15	rs6495867	1.96	<i>C15orf41</i>	Intronic		0.14			
15	rs16963938	1.96	<i>C15orf41</i>	Intronic		0.13			
15	rs16963949	1.96	<i>C15orf41</i>	Intronic		0.13			
6	rs2187818	1.95	<i>LOC101929163, HLA-DRA</i>	Intergenic		0.38			
6	rs614549	1.95	<i>SLC44A4</i>	Intronic		0.40			
4	rs34540355	1.94	<i>NEK1</i>	Exonic	nonsyn	0.05	T	B	N
19	rs73058047	1.94	<i>PRR12</i>	Exonic	nonsyn	0.004	D	D	N
19	rs8110889	1.94	<i>CHST8</i>	Intronic		0.25			
6	rs6941112	1.93	<i>STK19</i>	Intronic		0.33			
6	rs4151657	1.93	<i>CFB</i>	Intronic		0.36			
6	rs181997	1.92	<i>LOC100294145, HLA-DMB</i>	Intergenic		0.34			
8	rs2976189	1.92	<i>SLC7A13</i>	Exonic	nonsyn	0.15	D	D	P
6	rs3130320	1.90	<i>LOC101929163</i>	ncRNA		0.35			

Table of all significant and suggestive SNPs from two-point linkage analysis

*CH* Chromosome, *SNP* rsID of SNP, *HLOD* heterogeneity LOD score, *GENE* gene location of SNP, *FUNC* SNP function (e.g., exonic, intronic), *EXON* exonic function (*nonsyn* non-synonymous, *syn* synonymous), *FRQ* 1000Genomes minor allele frequency for Europeans, *SIFT* SIFT prediction (*D* damaging, *T* tolerated), *POLPH* PolyPhen2 prediction (*B* benign, *D* damaging, *P* possibly damaging), *MUTR* MutationTaster prediction (*N* neutral, *D* damaging, *P* possibly damaging)

variants being located in or near antisense RNA genes. Like the larger linked haplotype in Family 3333, there are many lower magnitude linked variants underneath and almost no negative signal.

### Multipoint linkage results

The multipoint linkage results on the pruned marker set did not produce any genome-wide significant HLOD scores (Fig. 5). Eight genome-wide suggestive markers were identified, all located in the 5p15.33 region identified by the two-point analysis. Almost all the genome-wide significant and suggestive signals identified in the two-point analyses were severely depressed here in the multipoint analysis, likely due to the large loss of information in the pruning process and some stronger negative signal in a few families. For instance, the signal at 8q21.3 is still clearly present but has been diminished to a maximum of 0.87. Performing the multipoint analysis with the 20, 15, and 10 most informative families gradually restores the signal to genome-wide significance at 3.3 (Supplemental Fig. 9a). This is also true of the signal at 12q15, though even using the most highly linked families here only recovers the signal to 1.7; there is more information loss here even at the most highly linked families (Supplemental Fig. 9b).

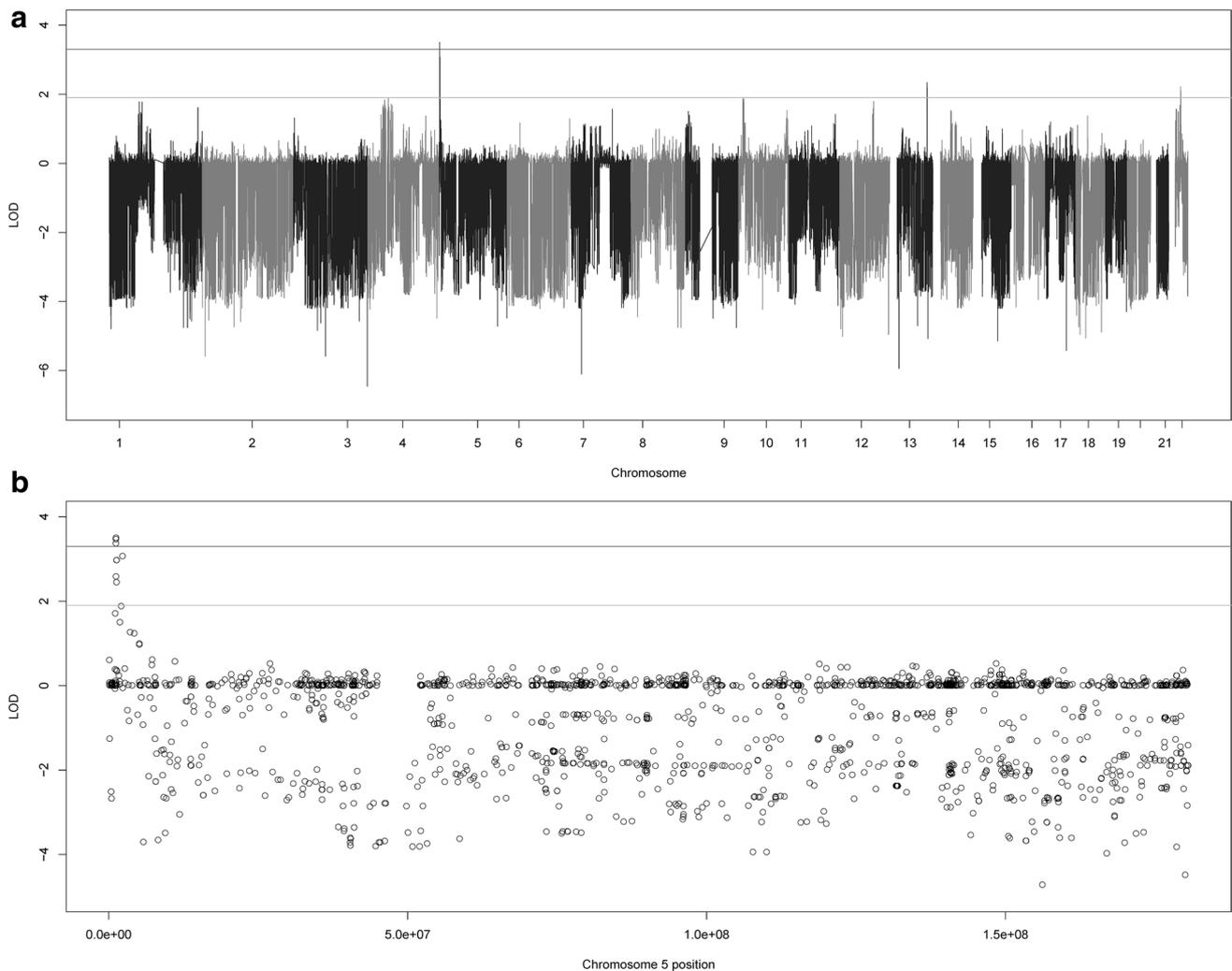
Unlike the across family results, the individual family multipoint results closely mirrored the two-point results. Family 3316 displayed six genome-wide significant LOD

scores at 5p15.33, ranging from LOD=3.62–3.33 (Fig. 6; Table 3) and was the driving force in this signal being the only genome-wide suggestive signal across families. Additional two genome-wide suggestive results were found at the same region; no other genome-wide suggestive results were found anywhere else in the family. Again, the highest variant was rs74581452 in *SLC6A18*. All other significant or suggestive variants were intronic or intergenic.

The variants at 5p15.33 in Family 3316 were the only genome-wide significant variants in any of the individual families; though there were additional genome-wide suggestive variants. Specifically, the suggestive signal at 1p36.22 in Family 3333 was recapitulated in the multipoint (Supplemental Fig. 10). Here, the signal started still started at 1p36.22 but ended slightly earlier at 1p34.2. All variants within this haplotype had LOD scores of ~2. The linked haplotype on 3q25.1 was also replicated in the multipoint, again with five tightly clustered variants with LOD scores ranging from 2.23 to 1.8 (Supplemental Fig. 11).

### Discussion

In this parametric linkage study using multiplex Pennsylvania Amish families, we have identified three genome-wide significant linkage signals to myopia at 12q15, 8q21.3, and 5p15.33. The 12q and 8q signals were each due to a single variant that exhibited a genome-wide significant



**Fig. 2** Family 3316 two-point genome-wide and chromosome 5 LOD scores. **a** The genome-wide two-point LOD score plot for family 3316. **b** The chromosome 5 two-point LOD score plot for family

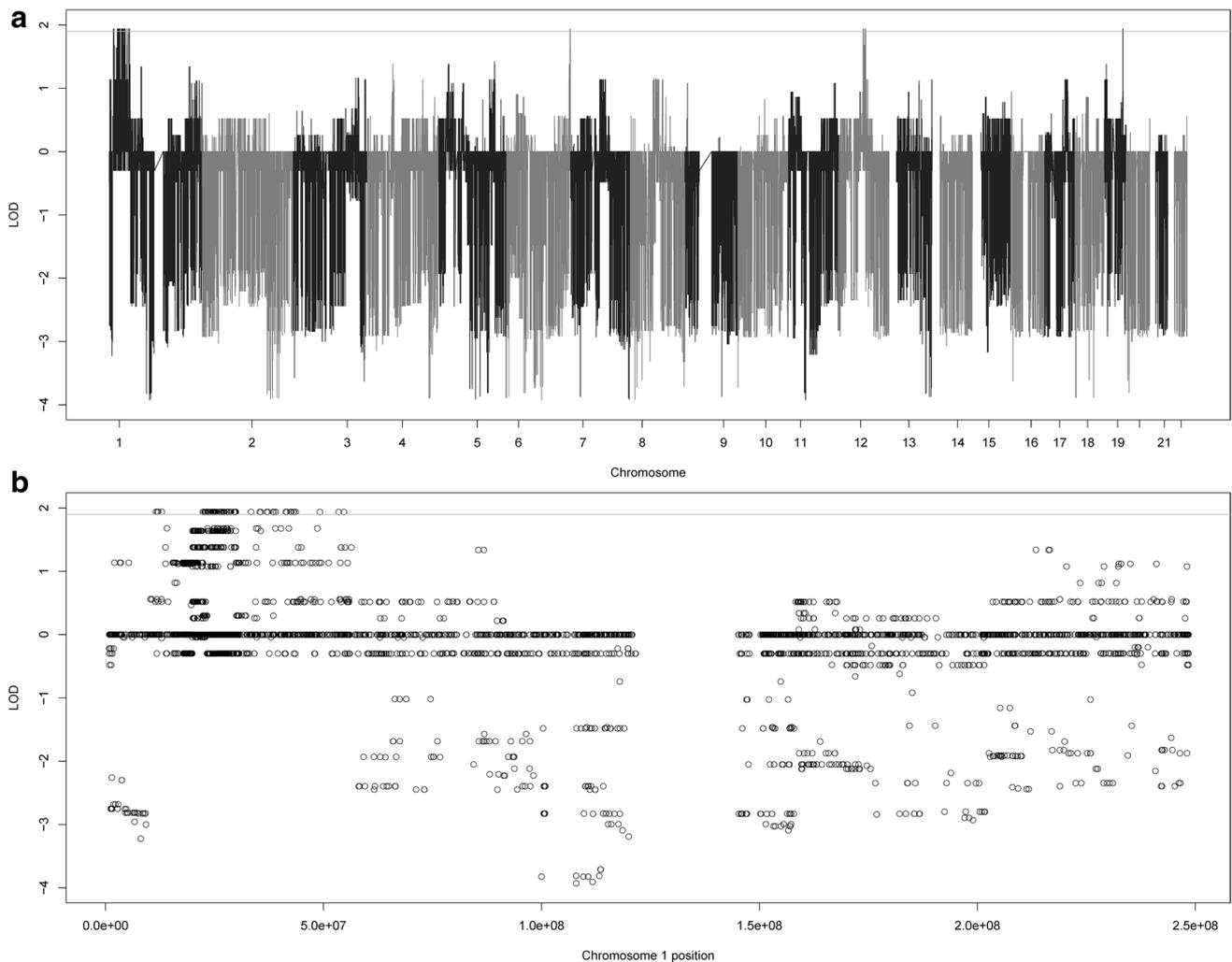
3316, showing a clearer look at the significant linkage signal at 5p15. The lines at 3.3 and 1.9 represent the thresholds for genome-wide significant and suggestive LOD scores

**Table 2** Significant and suggestive SNPs at 5p15.33 in family 3316

CHR	SNP	LOD	GENE	FUNC	EXON	FRQ	SIFT	POLPH	MUTR
5	rs74581452	3.51	<i>SLC6A18</i>	Exonic	syn	0.03			
5	rs7705355	3.47	<i>SLC6A18</i>	Exonic	nonsyn	0.04	T	B	P
5	rs7728667	3.37	<i>SLC6A18</i>	Exonic	nonsyn	0.21	T	B	P
5	rs17586674	3.07	<i>LOC100506858, IRX2</i>	Intergenic		0.44			
5	rs31489	2.98	<i>CLPTMIL</i>	Intronic		0.41			
5	rs401681	2.97	<i>CLPTMIL</i>	Intronic		0.44			
5	rs4975616	2.97	<i>MIR4457, CLPTMIL</i>	Intergenic		0.42			
5	rs4073918	2.59	<i>SLC6A18</i>	Exonic	nonsyn	0.33	T	B	P
5	rs402710	2.45	<i>CLPTMIL</i>	Intronic		0.33			

Table of all significant and suggestive SNPs from two-point linkage analysis

*CHR* Chromosome, *SNP* rsID of SNP, *LOD* LOD score in family 3316, *GENE* gene location of SNP, *FUNC* SNP function (e.g., exonic, intronic), *EXON* exonic function (*nonsyn* non-synonymous, *syn* synonymous), *FRQ* 1000Genomes minor allele frequency for Europeans, *SIFT* SIFT prediction (*D* damaging, *T* tolerated), *POLPH* PolyPhen2 prediction (*B* benign, *D* damaging, *P* possibly damaging), *MUTR* MutationTaster prediction (*N* neutral, *D* damaging, *P* possibly damaging)



**Fig. 3** Family 3333 two-point genome-wide and chromosome 1 LOD scores. **a** The genome-wide two-point LOD score plot for family 3333. **b** The chromosome 1 two-point LOD score plot for family

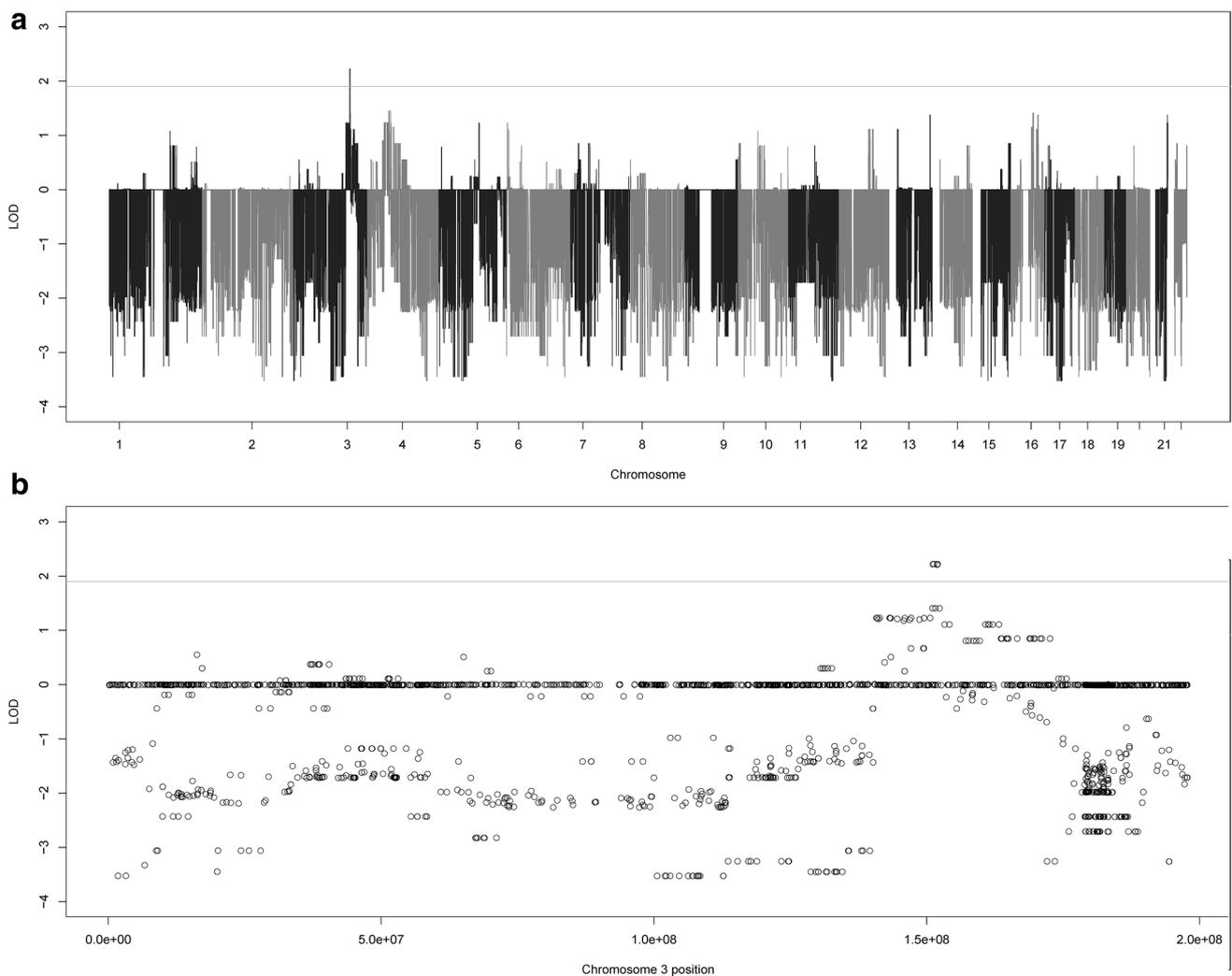
3333, showing a clearer look at the suggestive long linked haplotype at 1p36.22–32.22. The line at 1.9 represents the threshold for genome-wide suggestive LOD scores

HLOD score across families, while the 5p15 linkage consisted of three genome-wide significant SNPs in a single family. Multipoint analysis confirmed the final signal with six genome-wide significant variants at 5p15.33 in the same single family.

The 12q15 signal had the highest HLOD score (3.77) in the two-point analysis. Several previous studies had found linkage upstream of this region at 12q21–24 (Farbrother et al. 2004; Hawthorne et al. 2013; Klein et al. 2011; Young et al. 1998) and downstream at 12q13.11 (Metlapally et al. 2009) but neither overlaps with this signal on 12q15. The signal centers on the protein tyrosine phosphatase receptor type B (*PTPRB*). The linked variant and gene do not have any prior evidence of involvement in causation of myopia or eye disease in general. However, other protein tyrosine phosphatase receptors have been associated or linked with myopia including *PTPRR*, *PTPRF*, and *PTPRJ* (Hawthorne et al.

2013; Musolf et al. 2017). Both *PTPRR* and *PTPRF* have been put forth as possible candidate genes for the *MYP3* locus at 12q21–24. A suggestive rare non-synonymous variant that was predicted highly damaging by multiple databases was found at *BEST3*, also at 12q15. While *BEST3* has not been implicated in eye disease, another member of the bestrophin family, *BEST1*, is known to be the causal gene for vitelliform macular dystrophy (also known as Best’s macular dystrophy or Best’s disease) (Marchant et al. 2001; Marquardt et al. 1998; Petrukhin et al. 1998). Given that multiple independent linkage studies have implicated this region, the 12q13.1–12q24 region should be considered an excellent candidate region for one or more genes involved in myopia causation.

The significant linkage signal at 8q21.3 is novel and centered on the cyclic nucleotide gated channel beta 3 gene (*CNGB3*). A previous suggestive linkage had been located



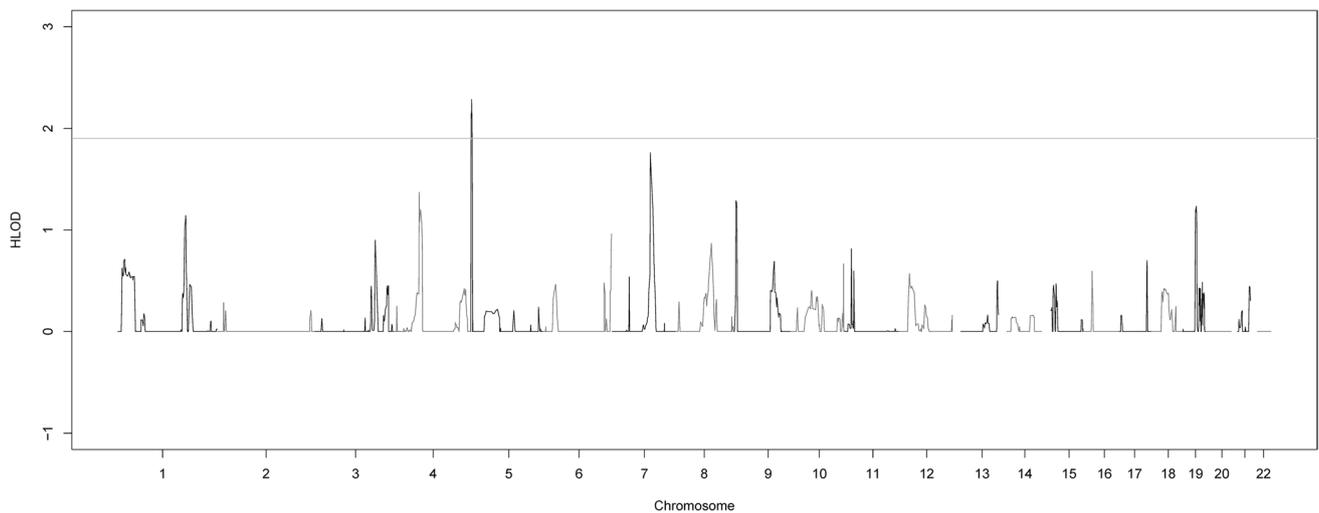
**Fig. 4** Family 3074 two-point genome-wide and chromosome 3 LOD scores. **a** The genome-wide two-point LOD score plot for family 3074. **b** The chromosome 3 two-point LOD score plot for family

3074, showing a clearer look at the suggestive long linked haplotype at 3p25.1. The line at 1.9 represents the threshold for genome-wide suggestive LOD scores

on chromosome 8 using microsatellites in these Amish families, however, this was located on the opposite arm at 8p23 (Stambolian et al. 2005). *CNGB3* has been found to affect photoreceptors and has been found to be causal in eye diseases, especially achromatopsia (Kohl et al. 2000, 2005) and progressive cone dystrophy (Michaelides et al. 2004). Achromatopsia symptoms include total color blindness (i.e., only seeing in shades of gray), reduced visual acuity, and light sensitivity. Progressive cone dystrophy results in reduced color vision and visual acuity with symptoms worsening over time. This is the first time *CNGB3* has been reported to be linked or associated with myopia.

The 5p15.33 linkage was driven by a single family and the linked region contained three genome-wide significant variants in the solute carrier family 6 member 18 (*SLC6A18*) gene in the two-point analysis. Six genome-wide significant

variants (one repeated from the two-point analysis) were found in the same region in the same family. One variant was again in *SLC6A18*, one was in solute carrier family 6 member 3 (*SLC6A3*), and the rest were in intergenic regions. *SLC6A18* is a sodium cotransporter that is not known to have any implication in myopia or eye disease in general, though SLC transporters have been shown to be expressed in the retina (Nishimura and Naito 2008), including *SLC6A3* (Andre et al. 2012), identified as significant in the multipoint analysis. There are other good candidate genes in the region, including iroquois homeobox 2 (*IRX2*) and iroquois homeobox 1 (*IRX1*). Intergenic regions between both these genes were found to be significant in the multipoint analysis. The iroquois homeobox genes are critical transcription factors in development and have been shown to regulate retinogenesis (Choy et al. 2010).



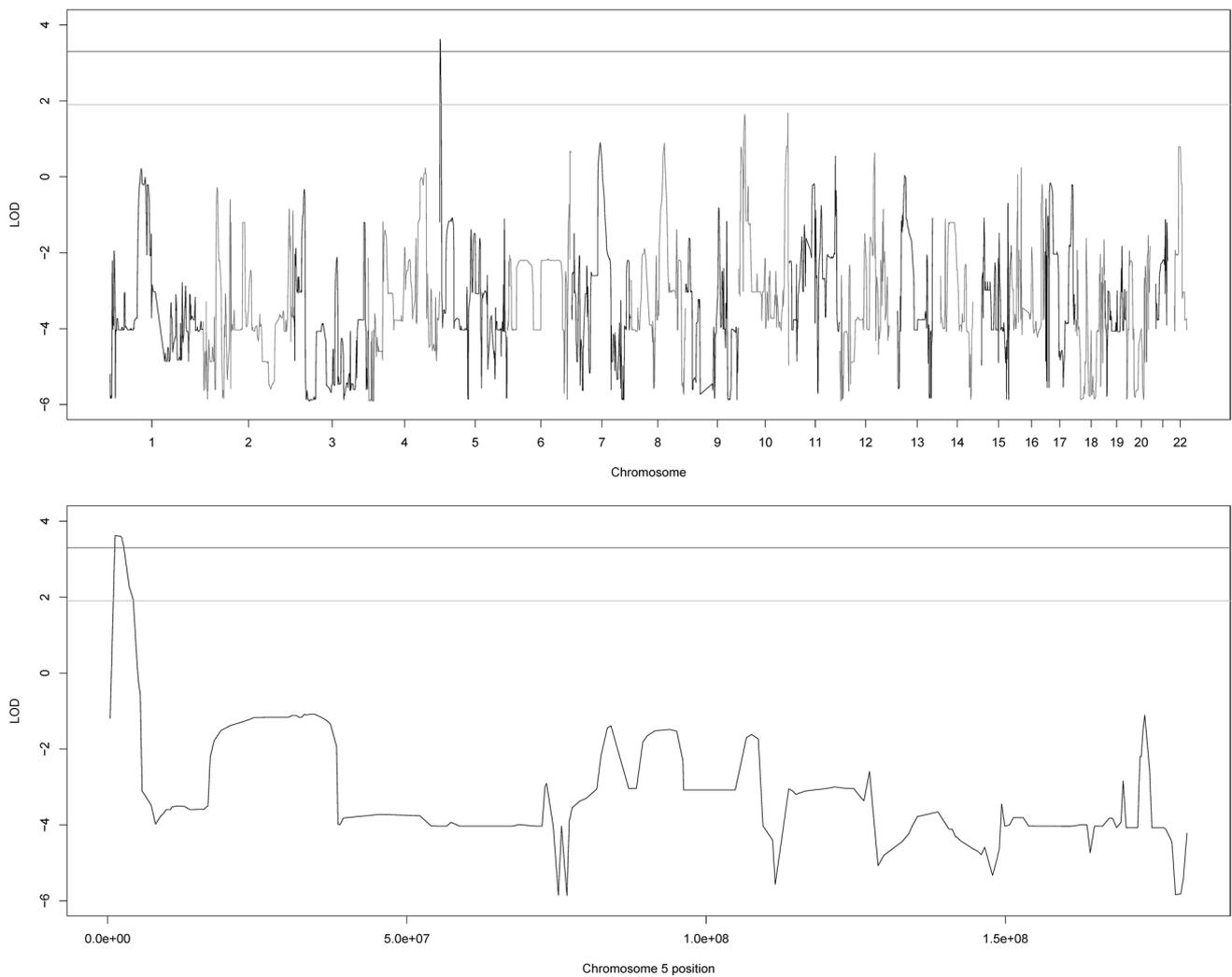
**Fig. 5** Multipoint genome-wide HLOD scores for all families. The genome-wide plot for the HLOD scores across all 25 families using multipoint linkage analysis. The line at 1.9 represents the threshold for genome-wide suggestive HLOD scores

We identified multiple suggestive signals as well, including an apparent long linked haplotype from 1p36.22–32.2 that was identified by both two-point and multipoint analysis within a single family. This region contains the known myopia risk locus *MYP14* at 1p36, first identified in Ashkenazi Jewish subjects using quantitative measures of refraction (Wojciechowski et al. 2006) and later replicated in other, larger-scale linkage studies (Abbott et al. 2012; Klein et al. 2011; Li et al. 2009). This is the first time the locus has been identified in the Amish population, or within a single family (all other linkage signals were cumulative across multiple families). Despite multiple replications, the causal gene of *MYP14* remains unknown. The fact that our signal is confined to a single family gives us the unique opportunity to perform targeted sequencing of the 1p haplotype to try to identify the causal locus.

Another very small suggestive linked haplotype was found at 3q25.1, again identified in a single family by both two-point and multipoint analyses. This region does contain a few interesting genes, specifically muscleblind-like splicing regulator 1 (*MBNL1*) and its antisense RNA *MBNL1-AS1*. *MBNL1* is known to cause myotonic dystrophy and eye abnormalities (Kanadia et al. 2003).

The two-point and multipoint results were in clear agreement for the individual family results, both identified multiple genome-wide significant variants at 5p15.33 in Family 3316, and genome-wide suggestive signals at 1p36–34 in Family 3333 and 3q25.1 in Family 3074. There were some discrepancies in the across family HLOD scores; the multipoint cumulative HLOD scores were consistently lower than the two-point scores, including at 8q21.3 and 12q15. This is almost certainly due to the loss of information from the pruning process. The exome-based chip used here is not

well suited to multipoint analyses. It provided very good coverage of the exome but also created a very dense marker map in those regions with high LD and a very sparse one in intergenic regions. Thus, pruning was required and we were left with only 3285 variants for multipoint analysis, a very large loss of information. For example, in the 5 cM flanking regions around the 8q and 12q signals, 88 and 108 variants, respectively, were removed by pruning. Many of these SNPs had informative HLOD scores > 1 and led to a loss of power at these sites. Further, this exome array is not a SNP linkage panel (where the variants are chosen to be highly informative). It is enriched with rare variants, with over 6000 of the variants having a MAF < 0.01 and over half the variants (approximately 28,000) had a MAF < 0.2. Even after pruning for the most informative markers at each 1 cM bin, there were over 1000 markers with MAF < 0.3. Pruning the markers resulted in situations where certain families were less informative or uninformative at a given marker, resulting in wholesale lower HLOD scores for all signals than the two-point analysis. This is why we observed a wholesale decrease in across family HLOD scores in the multipoint. The signals at 8q and 12q, which were of small, cumulative effect and centered on variants with a MAF < 0.15, are at particular risk for losing information added across the families. We were able to demonstrate that using the most highly linked families for each of the regions, the signal could be gradually recovered; in the case of 8q21.3 it was recovered to being genome-wide significant (Supplemental Fig. 9). The information problem was not as much of an effect on the individual family signals like 5p15.33, which are to a single family, is of larger effect, and is located on a longer, intra-family haplotype. Variants only needed to be informative in that family, not in cumulative fashion across the families.



**Fig. 6** Multipoint genome-wide and chromosome 5 LOD scores for Family 3316. **a** The genome-wide multipoint LOD score plot for Family 3316. **b** The chromosome 5 multipoint point LOD score plot

for Family 3316, showing a clearer look at the significant linkage signal at 5p15. The lines at 3.3 and 1.9 represent the thresholds for genome-wide significant and suggestive LOD scores

**Table 3** Significant and suggestive SNPs at 5p15.33 in Family 3316 for multipoint linkage analysis

CH	SNP	LOD	GENE	FUNC	EXON	FRQ
5	rs74581452	3.62	<i>SLC6A18</i>	Exonic	syn	0.03
5	rs27047	3.62	<i>SLC6A3</i>	Intronic		0.31
5	rs2354124	3.61	<i>MIR4277, MRPL36</i>	Intergenic		0.45
5	rs4527649	3.60	<i>MIR548BA, LOC100506858</i>	Intergenic		0.48
5	rs17586674	3.59	<i>LOC100506858, IRX2</i>	Intergenic		0.44
5	rs2962595	3.33	<i>LOC100506858, IRX2</i>	Intergenic		0.45
5	rs1651478	2.28	<i>IRX1, LOC10192915</i>	Intergenic		0.46
5	rs492478	2.11	<i>IRX1, LOC10192915</i>	Intergenic		0.075
5	rs12110273	1.93	<i>IRX1, LOC10192915</i>	Intergenic		0.47

A table of all significant and suggestive SNPs from multipoint linkage analysis

*CHR* chromosome, *SNP* rsID of SNP, *LOD* LOD score in Family 3316, *GENE* gene location of SNP, *FUNC* SNP function (e.g., exonic, intronic), *EXON* exonic function (*nonsyn* non-synonymous, *syn* synonymous), *FRQ* 1000Genomes minor allele frequency for Europeans

In conclusion, we have found significant linkage for myopia in Amish families to 12q15, 8q21.3, and 5p15.33. Further suggestive signals were found, including a replication of the known *MYP14* locus at 1p36 (localized to a single family). The 5p15.33 signal is particularly promising, as it was identified in both the two-point and multipoint linkage analyses. Since it is confined to a single family, it is likely of large effect on myopia within that family. The 8p21.3 and 12q15 signals are cumulative across families and of small effect on the phenotype. They were not present in the multipoint analysis, likely due to loss of information in the pruning process. Many of the linked loci have very promising candidate genes, particularly *CNGB3* at 8q21.3. *SLC6A18* at 5p15.33 is also very interesting as even though it has not been previously implicated in any eye disease, multiple exonic variants were found to be genome-wide significant within a single family. We plan on performing future functional analysis with *SLC6A18* in an attempt to observe any possible eye phenotypes. The two *IRX* genes at 5p15.33, which are known to be involved in retinogenesis, are also very good candidate genes. However, we note any implications of causality are still speculative at this point. We used a limited exome-based microarray in this study, thus it is possible that the causal variant was not genotyped on this chip and the significant variants here are simply tagging the true causal variants located along the same linked haplotype. To further elucidate the causal variants, targeted sequencing is planned. We plan to perform targeted sequencing on our most informative families at 12q15 and 8q21.3 as well as all of family 3316 at 5p15. Sequencing may also be performed at some of the suggestive linked haplotypes in individual families, particularly at 1p36.22–32.22 in family 3333.

**Acknowledgements** The authors thank all study participants and their families. This work was funded in part by the National Eye Institute Grant R01 EY020483 and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

## References

- Abbott D, Li YJ, Guggenheim JA, Metlapally R, Malecaze F, Calvas P, Rosenberg T, Paget S, Zayats T, Mackey DA, Feng S, Young TL (2012) An international collaborative family-based whole genome quantitative trait linkage scan for myopic refractive error. *Mol Vis* 18:720–729
- Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet Chap* 7:7–20. <https://doi.org/10.1002/0471142905.hg0720s76>
- Andre P, Saubamea B, Cochois-Guegan V, Marie-Claire C, Cattelotte J, Smirnova M, Schinkel AH, Scherrmann JM, Cisternino S (2012) Transport of biogenic amine neurotransmitters at the mouse blood-retina and blood-brain barriers by uptake1 and uptake2. *J Cereb Blood Flow Metab* 32:1989–2001. <https://doi.org/10.1038/jcbfm.2012.109>
- Chang X, Wang K (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 49:433–436. <https://doi.org/10.1136/jmedgenet-2012-100918>
- Choy SW, Cheng CW, Lee ST, Li VW, Hui MN, Hui CC, Liu D, Cheng SH (2010) A cascade of *irx1a* and *irx2a* controls *shh* expression during retinogenesis. *Dev Dyn* 239:3204–3214. <https://doi.org/10.1002/dvdy.22462>
- Ciner E, Ibay G, Wojciechowski R, Dana D, Holmes TN, Bailey-Wilson JE, Stambolian D (2009) Genome-wide scan of African-American and white families for linkage to myopia. *Am J Ophthalmol* 147:512–517 e2. <https://doi.org/10.1016/j.ajo.2008.09.004>
- Duffy D (2008) SIB-PAIR: a program for simple genetic analysis v1.00.beta. Queensland Institute of Medical Research. <https://genepi.qimr.edu.au/staff/davidD/Sib-pair/Documents/sib-pair.html>. Accessed Jan 2018
- Fan Q, Verhoeven VJ, Wojciechowski R, Barathi VA, Hysi PG, Guggenheim JA, Hohn R, Vitart V, Khawaja AP, Yamashiro K, Hosseini SM, Lehtimäki T, Lu Y, Haller T, Xie J, Delcourt C, Pirastu M, Wedenoja J, Gharahkhani P, Venturini C, Miyake M, Hewitt AW, Guo X, Mazur J, Huffman JE, Williams KM, Polasek O, Campbell H, Rudan I, Vataavuk Z, Wilson JF, Joshi PK, McMahon G, St Pourcain B, Evans DM, Simpson CL, Schwantes-An TH, Igo RP, Mirshahi A, Cougnard-Gregoire A, Bellenguez C, Blettner M, Raitakari O, Kahonen M, Seppala I, Zeller T, Meitinger T, Consortium for Refractive E Myopia, Ried JS, Gieger C, Portas L, van Leeuwen EM, Amin N, Uitterlinden AG, Rivadeneira F, Hofman A, Vingerling JR, Wang YX, Wang X, Tai-Hui Boh E, Ikram MK, Sabanayagam C, Gupta P, Tan V, Zhou L, Ho CE, Lim W, Beuerman RW, Siantar R, Tai ES, Vithana E, Mihailov E, Khor CC, Hayward C, Luben RN, Foster PJ, Klein BE, Klein R, Wong HS, Mitchell P, Metspalu A, Aung T, Young TL, He M, Parssinen O, van Duijn CM, Jin Wang J, Williams C, Jonas JB, Teo YY, Mackey DA, Oexle K, Yoshimura N, Paterson AD, Pfeiffer N, Wong TY, Baird PN, Stambolian D, Wilson JE et al (2016) Meta-analysis of gene-environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nat Commun* 7:11008. <https://doi.org/10.1038/ncomms11008>
- Farbrother JE, Kirov G, Owen MJ, Pong-Wong R, Haley CS, Guggenheim JA (2004) Linkage analysis of the genetic loci for high myopia on 18p, 12q, and 17q in 51 UK families. *Invest Ophthalmol Vis Sci* 45:2879–2885. <https://doi.org/10.1167/iovs.03-1156>
- Guo H, Tong P, Liu Y, Xia L, Wang T, Tian Q, Li Y, Hu Y, Zheng Y, Jin X, Li Y, Xiong W, Tang B, Feng Y, Li J, Pan Q, Hu Z, Xia K (2015) Mutations of P4HA2 encoding prolyl 4-hydroxylase 2 are associated with non-syndromic high myopia. *Genet Med* 17:300–306. <https://doi.org/10.1038/gim.2015.28>
- Hammond CJ, Andrew T, Mak YT, Spector TD (2004) A susceptibility locus for myopia in the normal population is linked to the PAX6 gene region on chromosome 11: a genomewide scan of dizygotic twins. *Am J Hum Genet* 75:294–304. <https://doi.org/10.1086/423148>
- Hawthorne F, Feng S, Metlapally R, Li YJ, Tran-Viet KN, Guggenheim JA, Malecaze F, Calvas P, Rosenberg T, Mackey DA, Venturini C, Hysi PG, Hammond CJ, Young TL (2013) Association mapping of the high-grade myopia MYP3 locus reveals novel candidates UHRF1BP1L, PTPRR, and PPFIA2. *Invest Ophthalmol Vis Sci* 54:2076–2086. <https://doi.org/10.1167/iovs.12-11102>
- Ibay G, Doan B, Reider L, Dana D, Schliifka M, Hu H, Holmes T, O'Neill J, Owens R, Ciner E, Bailey-Wilson JE, Stambolian D (2004) Candidate high myopia loci on chromosomes 18p and 12q do not play a major role in susceptibility to common myopia. *BMC Med Genet* 5:20. <https://doi.org/10.1186/1471-2350-5-20>

- Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 99:877–885. <https://doi.org/10.1016/j.ajhg.2016.08.016>
- Kanadia RN, Johnstone KA, Mankodi A, Lungu C, Thornton CA, Esson D, Timmers AM, Hauswirth WW, Swanson MS (2003) A muscleblind knockout model for myotonic dystrophy. *Science* 302:1978–1980. <https://doi.org/10.1126/science.1088583>
- Kiefer AK, Tung JY, Do CB, Hinds DA, Mountain JL, Francke U, Eriksson N (2013) Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet* 9:e1003299. <https://doi.org/10.1371/journal.pgen.1003299>
- Klein AP, Duggal P, Lee KE, Cheng CY, Klein R, Bailey-Wilson JE, Klein BE (2011) Linkage analysis of quantitative refraction and refractive errors in the Beaver Dam Eye Study. *Invest Ophthalmol Vis Sci* 52:5220–5225. <https://doi.org/10.1167/iovs.10-7096>
- Kohl S, Baumann B, Broghammer M, Jagle H, Sieving P, Kellner U, Spegal R, Anastasi M, Zrenner E, Sharpe LT, Wissinger B (2000) Mutations in the CNGB3 gene encoding the beta-subunit of the cone photoreceptor cGMP-gated channel are responsible for achromatopsia (ACHM3) linked to chromosome 8q21. *Hum Mol Genet* 9:2107–2116
- Kohl S, Varsanyi B, Antunes GA, Baumann B, Hoyng CB, Jagle H, Rosenberg T, Kellner U, Lorenz B, Salati R, Jurklics B, Farkas A, Andreasson S, Weleber RG, Jacobson SG, Rudolph G, Castellani C, Dollfus H, Legius E, Anastasi M, Bitoun P, Lev D, Sieving PA, Munier FL, Zrenner E, Sharpe LT, Cremers FP, Wissinger B (2005) CNGB3 mutations account for 50% of all cases with autosomal recessive achromatopsia. *Eur J Hum Genet* 13:302–308. <https://doi.org/10.1038/sj.ejhg.5201269>
- Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–1081. <https://doi.org/10.1038/nprot.2009.86>
- Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11:241–247. <https://doi.org/10.1038/ng1195-241>
- Li YJ, Guggenheim JA, Bulusu A, Metlapally R, Abbott D, Malecaze F, Calvas P, Rosenberg T, Paget S, Creer RC, Kirov G, Owen MJ, Zhao B, White T, Mackey DA, Young TL (2009) An international collaborative family-based whole-genome linkage scan for high-grade myopia. *Invest Ophthalmol Vis Sci* 50:3116–3127. <https://doi.org/10.1167/iovs.08-2781>
- Mandal DM, Wilson AF, Elston RC, Weissbecker K, Keats BJ, Bailey-Wilson JE (2000) Effects of misspecification of allele frequencies on the type I error rate of model-free linkage analysis. *Hum Hered* 50:126–132. <https://doi.org/10.1159/000022900>
- Mandal DM, Wilson AF, Bailey-Wilson JE (2001) Effects of misspecification of allele frequencies on the power of Haseman–Elston sib-pair linkage method for quantitative traits. *Am J Med Genet* 103:308–313
- Mandal DM, Sorant AJ, Atwood LD, Wilson AF, Bailey-Wilson JE (2006) Allele frequency misspecification: effect on power and Type I error of model-dependent linkage analysis of quantitative traits under random ascertainment. *BMC Genet* 7:21. <https://doi.org/10.1186/1471-2156-7-21>
- Marchant D, Gogat K, Boutboul S, Pequignot M, Sternberg C, Dureau P, Roche O, Uteza Y, Hache JC, Puech B, Puech V, Dumur V, Mouillon M, Munier FL, Schorderet DF, Marsac C, Dufier JL, Abitbol M (2001) Identification of novel VMD2 gene mutations in patients with best vitelliform macular dystrophy. *Hum Mutat* 17:235. <https://doi.org/10.1002/humu.9>
- Marquardt A, Stohr H, Passmore LA, Kramer F, Rivera A, Weber BH (1998) Mutations in a novel gene, VMD2, encoding a protein of unknown properties cause juvenile-onset vitelliform macular dystrophy (Best's disease). *Hum Mol Genet* 7:1517–1525
- Matise TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, Hyland FC, Kennedy GC, Kong X, Murray SS, Ziegler JS, Stewart WC, Buyske S (2007) A second-generation combined linkage physical map of the human genome. *Genome Res* 17:1783–1786. <https://doi.org/10.1101/gr.7156307>
- Metlapally R, Li YJ, Tran-Viet KN, Abbott D, Czaja GR, Malecaze F, Calvas P, Mackey D, Rosenberg T, Paget S, Zayats T, Owen MJ, Guggenheim JA, Young TL (2009) COL1A1 and COL2A1 genes and myopia susceptibility: evidence of association and suggestive linkage to the COL2A1 locus. *Invest Ophthalmol Vis Sci* 50:4080–4086. <https://doi.org/10.1167/iovs.08-3346>
- Michaelides M, Aligianis IA, Ainsworth JR, Good P, Mollon JD, Maher ER, Moore AT, Hunt DM (2004) Progressive cone dystrophy associated with mutation in CNGB3. *Invest Ophthalmol Vis Sci* 45:1975–1982
- Miyake M, Yamashiro K, Tabara Y, Suda K, Morooka S, Nakanishi H, Khor CC, Chen P, Qiao F, Nakata I, Akagi-Kurashige Y, Gotoh N, Tsujikawa A, Meguro A, Kusuhara S, Polasek O, Hayward C, Wright AF, Campbell H, Richardson AJ, Schache M, Takeuchi M, Mackey DA, Hewitt AW, Cuellar G, Shi Y, Huang L, Yang Z, Leung KH, Kao PY, Yap MK, Yip SP, Moriyama M, Ohno-Matsui K, Mizuki N, MacGregor S, Vitart V, Aung T, Saw SM, Tai ES, Wong TY, Cheng CY, Baird PN, Yamada R, Matsuda F, Nagahama Study G, Yoshimura N (2015) Identification of myopia-associated WNT7B polymorphisms provides insights into the mechanism underlying the development of myopia. *Nat Commun* 6:6689. <https://doi.org/10.1038/ncomms7689>
- Musolf AM, Simpson CL, Moiz BA, Long KA, Portas L, Murgia F, Ciner EB, Stambolian D, Bailey-Wilson JE (2017) Caucasian families exhibit significant linkage of myopia to chromosome 11p. *Invest Ophthalmol Vis Sci* 58:3547–3554. <https://doi.org/10.1167/iovs.16-21271>
- Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814
- Nishimura M, Naito S (2008) Tissue-specific mRNA expression profiles of human solute carrier transporter superfamilies. *Drug Metab Pharmacokin* 23:22–44
- Peet JA, Cotch MF, Wojciechowski R, Bailey-Wilson JE, Stambolian D (2007) Heritability and familial aggregation of refractive error in the Old Order Amish. *Invest Ophthalmol Vis Sci* 48:4002–4006. <https://doi.org/10.1167/iovs.06-1388>
- Petrukhin K, Koisti MJ, Bakall B, Li W, Xie G, Marknell T, Sandgren O, Forsman K, Holmgren G, Andreasson S, Vujic M, Bergen AA, McGarty-Dugan V, Figueroa D, Austin CP, Metzker ML, Caskey CT, Wadelius C (1998) Identification of the gene responsible for best macular dystrophy. *Nat Genet* 19:241–247. <https://doi.org/10.1038/915>
- Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–2337. <https://doi.org/10.1093/bioinformatics/btq419>
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and

- population-based linkage analyses. *Am J Hum Genet* 81:559–575. <https://doi.org/10.1086/519795>
- Schwarz JM, Cooper DN, Schuelke M, Seelow D (2014) Mutation-Taster2: mutation prediction for the deep-sequencing age. *Nat Methods* 11:361–362. <https://doi.org/10.1038/nmeth.2890>
- Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 40:W452–W457. <https://doi.org/10.1093/nar/gks539>
- Simpson CL, Wojciechowski R, Yee SS, Soni P, Bailey-Wilson JE, Stambolian D (2013) Regional replication of association with refractive error on 15q14 and 15q25 in the Age-Related Eye Disease Study cohort. *Mol Vis* 19:2173–2186
- Simpson CL, Wojciechowski R, Oexle K, Murgia F, Portas L, Li X, Verhoeven VJ, Vitart V, Schache M, Hosseini SM, Hysi PG, Raffel LJ, Cotch MF, Chew E, Klein BE, Klein R, Wong TY, van Duijn CM, Mitchell P, Saw SM, Fossarello M, Wang JJ, Group DER, Polasek O, Campbell H, Rudan I, Oostra BA, Uitterlinden AG, Hofman A, Rivadeneira F, Amin N, Karssen LC, Vingerling JR, Doring A, Bettecken T, Bencic G, Gieger C, Wichmann HE, Wilson JF, Venturini C, Fleck B, Cumberland PM, Rahi JS, Hammond CJ, Hayward C, Wright AF, Paterson AD, Baird PN, Klaver CC, Rotter JJ, Pirastu M, Meitinger T, Bailey-Wilson JE, Stambolian D (2014) Genome-wide meta-analysis of myopia and hyperopia provides evidence for replication of 11 loci. *PLoS One* 9:e107110. <https://doi.org/10.1371/journal.pone.0107110>
- Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 58:1323–1337
- Sobel E, Sengul H, Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered* 52:121–131. <https://doi.org/10.1159/000053366>
- Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 70:496–508. <https://doi.org/10.1086/338920>
- Stambolian D (2013) Genetic susceptibility and mechanisms for refractive error. *Clin Genet* 84:102–108. <https://doi.org/10.1111/cge.12180>
- Stambolian D, Ibay G, Reider L, Dana D, Moy C, Schlifka M, Holmes T, Ciner E, Bailey-Wilson JE (2004) Genomewide linkage scan for myopia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 22q12. *Am J Hum Genet* 75:448–459. <https://doi.org/10.1086/423789>
- Stambolian D, Ciner EB, Reider LC, Moy C, Dana D, Owens R, Schlifka M, Holmes T, Ibay G, Bailey-Wilson JE (2005) Genome-wide scan for myopia in the Old Order Amish. *Am J Ophthalmol* 140:469–476. <https://doi.org/10.1016/j.ajo.2005.04.014>
- Stambolian D, Wojciechowski R, Oexle K, Pirastu M, Li X, Raffel LJ, Cotch MF, Chew EY, Klein B, Klein R, Wong TY, Simpson CL, Klaver CC, van Duijn CM, Verhoeven VJ, Baird PN, Vitart V, Paterson AD, Mitchell P, Saw SM, Fossarello M, Kazmierkiewicz K, Murgia F, Portas L, Schache M, Richardson A, Xie J, Wang JJ, Rohtchina E, Group DER, Viswanathan AC, Hayward C, Wright AF, Polasek O, Campbell H, Rudan I, Oostra BA, Uitterlinden AG, Hofman A, Rivadeneira F, Amin N, Karssen LC, Vingerling JR, Hosseini SM, Doring A, Bettecken T, Vatauvuk Z, Gieger C, Wichmann HE, Wilson JF, Fleck B, Foster PJ, Topouzis F, McGuffin P, Sim X, Inouye M, Holliday EG, Attia J, Scott RJ, Rotter JJ, Meitinger T, Bailey-Wilson JE (2013) Meta-analysis of genome-wide association studies in five cohorts reveals common variants in RBFOX1, a regulator of tissue-specific splicing, associated with refractive error. *Hum Mol Genet* 22:2754–2764. <https://doi.org/10.1093/hmg/ddt116>
- Thomas A (2006) TwoPointLods. University of Utah. <http://www.genepi.med.utah.edu/~alun/software/>. Accessed Jan 2018
- Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC (2016) SIFT missense predictions for genomes. *Nat Protoc* 11:1–9. <https://doi.org/10.1038/nprot.2015.123>
- Verhoeven VJ, Hysi PG, Saw SM, Vitart V, Mirshahi A, Guggenheim JA, Cotch MF, Yamashiro K, Baird PN, Mackey DA, Wojciechowski R, Ikram MK, Hewitt AW, Duggal P, Janmahasatian S, Khor CC, Fan Q, Zhou X, Young TL, Tai ES, Goh LK, Li YJ, Aung T, Vithana E, Teo YY, Tay W, Sim X, Rudan I, Hayward C, Wright AF, Polasek O, Campbell H, Wilson JF, Fleck BW, Nakata I, Yoshimura N, Yamada R, Matsuda F, Ohno-Matsui K, Nag A, McMahon G, St Pourcain B, Lu Y, Rahi JS, Cumberland PM, Bhattacharya S, Simpson CL, Atwood LD, Li X, Raffel LJ, Murgia F, Portas L, Despriet DD, van Koolwijk LM, Wolfram C, Lackner KJ, Tonjes A, Magi R, Lehtimäki T, Kahonen M, Esko T, Metspalu A, Rantanen T, Parssinen O, Klein BE, Meitinger T, Spector TD, Oostra BA, Smith AV, de Jong PT, Hofman A, Amin N, Karssen LC, Rivadeneira F, Vingerling JR, Eiriksdottir G, Gudnason V, Doring A, Bettecken T, Uitterlinden AG, Williams C, Zeller T, Castagne R, Oexle K, van Duijn CM, Iyengar SK, Mitchell P, Wang JJ, Hohn R, Pfeiffer N, Bailey-Wilson JE, Stambolian D, Wong TY, Hammond CJ, Klaver CC (2012) Large scale international replication and meta-analysis study confirms association of the 15q14 locus with myopia. The CREAM consortium. *Hum Genet* 131:1467–1480. <https://doi.org/10.1007/s00439-012-1176-0>
- Verhoeven VJ, Hysi PG, Wojciechowski R, Fan Q, Guggenheim JA, Hohn R, MacGregor S, Hewitt AW, Nag A, Cheng CY, Yonova-Doing E, Zhou X, Ikram MK, Buitendijk GH, McMahon G, Kemp JP, Pourcain BS, Simpson CL, Makela KM, Lehtimäki T, Kahonen M, Paterson AD, Hosseini SM, Wong HS, Xu L, Jonas JB, Parssinen O, Wedenoja J, Yip SP, Ho DW, Pang CP, Chen LJ, Burdon KP, Craig JE, Klein BE, Klein R, Haller T, Metspalu A, Khor CC, Tai ES, Aung T, Vithana E, Tay WT, Barathi VA, Consortium for Refractive E, Myopia, Chen P, Li R, Liao J, Zheng Y, Ong RT, Doring A, Diabetes C, Complications Trial/Epidemiology of Diabetes I, Complications Research G, Evans DM, Timpson NJ, Verkerk AJ, Meitinger T, Raitakari O, Hawthorne F, Spector TD, Karssen LC, Pirastu M, Murgia F, Ang W, Wellcome Trust Case Control C, Mishra A, Montgomery GW, Pennell CE, Cumberland PM, Cotlarciuc I, Mitchell P, Wang JJ, Schache M, Janmahasatian S, Igo RP Jr, Lass JH, Chew E, Iyengar SK, Fuchs' Genetics Multi-Center Study, Gorgels TG, Rudan I, Hayward C, Wright AF, Polasek O, Vatauvuk Z, Wilson JF, Fleck B, Zeller T, Mirshahi A, Muller C, Uitterlinden AG, Rivadeneira F, Vingerling JR, Hofman A, Oostra BA, Amin N, Bergen AA, Teo YY et al (2013) Genome-wide meta-analyses of multiethnicity cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* 45:314–318. <https://doi.org/10.1038/ng.2554>
- Vitale S, Sperduto RD, Ferris FL 3rd (2009) Increased prevalence of myopia in the United States between 1971–1972 and 1999–2004. *Arch Ophthalmol* 127:1632–1639. <https://doi.org/10.1001/archophthalmol.2009.303>
- Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164. <https://doi.org/10.1093/nar/gkq603>
- Wojciechowski R, Hysi PG (2013) Focusing in on the complex genetics of myopia. *PLoS Genet* 9:e1003442. <https://doi.org/10.1371/journal.pgen.1003442>
- Wojciechowski R, Moy C, Ciner E, Ibay G, Reider L, Bailey-Wilson JE, Stambolian D (2006) Genomewide scan in Ashkenazi Jewish families demonstrates evidence of linkage of ocular refraction to a QTL on chromosome 1p36. *Hum Genet* 119:389–399. <https://doi.org/10.1007/s00439-006-0153-x>
- Wojciechowski R, Bailey-Wilson JE, Stambolian D (2009) Fine-mapping of candidate region in Amish and Ashkenazi families

confirms linkage of refractive error to a QTL on 1p34-p36. *Mol Vis* 15:1398–1406

Young TL, Ronan SM, Alvear AB, Wildenberg SC, Oetting WS, Atwood LD, Wilkin DJ, King RA (1998) A second locus for familial high myopia maps to chromosome 12q. *Am J Hum Genet* 63:1419–1424. <https://doi.org/10.1086/302111>

Zhao Y, Zhao F, Zong L, Zhang P, Guan L, Zhang J, Wang D, Wang J, Chai W, Lan L, Li Q, Han B, Yang L, Jin X, Yang W, Hu X, Wang X, Li N, Li Y, Petit C, Wang J, Wang HY, Wang Q (2013) Exome sequencing and linkage analysis identified tenascin-C (TNC) as

a novel causative gene in non-syndromic hearing loss. *PLoS One* 8:e69549. <https://doi.org/10.1371/journal.pone.0069549>

Zhou L, Li T, Song X, Li Y, Li H, Dan H (2015) NYX mutations in four families with high myopia with or without CSNB1. *Mol Vis* 21:213–223

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Anthony M. Musolf<sup>1</sup> · Claire L. Simpson<sup>1,2,3</sup> · Theresa A. Alexander<sup>1</sup> · Laura Portas<sup>1</sup> · Federico Murgia<sup>1</sup> · Elise B. Ciner<sup>4</sup> · Dwight Stambolian<sup>5</sup> · Joan E. Bailey-Wilson<sup>1</sup> 

<sup>1</sup> Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, 333 Cassell Dr, Suite 1200, Baltimore, MD 21224, USA

<sup>2</sup> Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

<sup>3</sup> Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, TN, USA

<sup>4</sup> The Pennsylvania College of Optometry at Salus University, Elkins Park, PA, USA

<sup>5</sup> Department of Ophthalmology, University of Pennsylvania, Philadelphia, PA, USA