Chapter 3
Integration of Linkage Analysis and Next-Generation Sequencing Data

Francesca Lantieri, Mark A. Levenstien, and Marcella Devoto

Abstract Genetic mapping by linkage analysis has been for many years the first step in the identification of genes responsible for rare Mendelian disorders. When the focus of genetic research shifted toward the study of the more complex common disorders, alternative approaches such as association studies were shown to be more successful in identifying common variants of small effect that are in part responsible for susceptibility to such conditions. Recent advances in technologies that make feasible the sequencing of whole exomes or genomes have renewed interest in the identification of rare variants, which are in principle amenable to being detected by linkage analysis. As a result, linkage analysis and family-based studies in general are being reexamined as an aid to filter and validate results of whole exome and whole genome sequencing experiments. This chapter will describe a few

F. Lantieri
Division of Human Genetics, The Children's Hospital of Philadelphia,
3615 Civic Center Blvd., ARC 1002, Philadelphia, PA, 19104, USA

Department of Health Science, Biostatistics Unit, University of Genoa, Genoa, Italy

M.A. Levenstien
Division of Human Genetics, The Children's Hospital of Philadelphia,
3615 Civic Center Blvd., ARC 1002, Philadelphia, PA, 19104, USA

M. Devoto (✉)
Division of Human Genetics, The Children's Hospital of Philadelphia,
3615 Civic Center Blvd., ARC 1002, Philadelphia, PA, 19104, USA

Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA

Department of Biostatistics and Epidemiology, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA, USA

Department of Molecular Medicine, University La Sapienza,Rome, Italy
e-mail: devoto@chop.edu

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_3, © Springer Science+Business Media Dordrecht 2012

representative papers that have incorporated linkage analysis and its results in the design, execution, and interpretation of whole genome or whole exome sequencing studies.

Keywords Linkage analysis • Rare variants • Family-based studies • Whole exome sequencing • Whole genome sequencing

3.1 Introduction

Linkage analysis and family-based tests have been a workhorse of genetic mapping for Mendelian disease gene identification. From the beginning of the 1980s, the combination of increasingly dense DNA marker maps and powerful software tools implementing such tests have led to the identification of the genes responsible for thousands of Mendelian disorders (Botstein and Risch 2003). When the focus of genetic research shifted from the rare, highly penetrant monogenic diseases to the more common, complex ones, it became evident that linkage analysis was underpowered to detect the common risk variants with small effects expected under the common disease/common variant hypothesis (Risch and Merikangas 1996). Instead, genome-wide association studies (GWAS) in case–control datasets have led to the identification of many such variants in a variety of different disorders and traits (http://www.genome.gov/gwastudies/). At the same time, it has become clear that common variants do not explain all the genetic susceptibility to such traits, and evidence has been accumulating that rare, possibly higher penetrant variants also underlie susceptibility to common complex traits (Manolio et al. 2009). In addition, many Mendelian disorders are too rare for the linkage analysis approach alone to work, and thus the corresponding genes still remain undetected.

The advent of massive parallel sequencing and the ability to sequence the whole exome or even genome of individuals at a relatively low cost has made the discovery of all variants present in an individual or family technically possible. These advances can lead to successful disease gene identification, as demonstrated initially in a few Mendelian disorders (Bamshad et al. 2011) and more recently in some complex ones (Zeggini 2011). However, analysis of whole exome or whole genome sequence (WES or WGS) data poses noticeable bioinformatics and statistical challenges, and the identification of the true risk variants among the many detected by such experiments has often been compared to finding the classic needle in a haystack. Every possible piece of information that can be used to facilitate such effort should be considered and incorporated into the analysis, and in this respect, analysis of the segregation of candidate risk variants in family members of affected individuals has been suggested as particularly useful (Cirulli and Goldstein 2010). In fact, linkage analysis can inform interpretation of WES or WGS data both by indicating regions of the genome with higher a priori chance of including the risk variants when results of linkage studies on the disease of interest are already available and a posteriori by limiting further evaluation of candidate variants detected through sequencing only to those that show co-segregation (i.e., linkage) to the disease or trait in families of affected individuals.

In this chapter, we will describe ways in which linkage analysis and family-based data have been incorporated into WES or WGS experiments that have led to the identification of new disease gene variants. We will also discuss methods that have been proposed for the specific purpose of integrating the use of family-based data in WES or WGS analysis.

3.2 Linkage Analysis in WES/WGS Studies and Identification of Disease Genes

Co-segregation of variants identified by WES/WGS and the disease phenotype in affected relatives is an obvious filter to impose on results of such experiments to reduce the number of candidate variants. As such, numerous studies have used this relatively simple strategy (Ng et al. 2010), which, however, does not take full advantage of the power of linkage analysis to identify candidate regions by modeling the disease mode of inheritance using allele frequency, reduced penetrance, and phenocopy rate. We will not review such studies, as their number is already large and increasing, and the approach relatively straightforward (i.e., remove from further consideration all variants that are not shared by affected relatives). Rather, we will review a few experimental studies that have integrated a formal linkage analysis with their sequencing experiments at various stages. As with all reviews, this list is necessarily limited, but we hope it will still illustrate different ways in which investigators are taking advantage of the power of linkage analysis in their sequencing experiments for the identification of disease genes.

A good proof of principle of the power of both the linkage and the whole exome sequencing approaches is the study of Bowne et al. (2011). These authors investigated an Irish family with autosomal dominant retinitis pigmentosa (adRP) by linkage analysis on 27 members and simultaneously analyzed one unaffected and three affected members by WES. No disease-causing mutations or copy-number variants had been identified by standard sequencing or multiplex ligation-dependent probe amplification (MLPA) of known candidate genes. Linkage and haplotype analyses, instead, mapped the disease locus to an 8.8-Mb region on chromosome 1p31, with a maximum multipoint LOD score of 3.6. The authors selected 11 candidate genes from the critical region using several criteria. Specifically, the candidates were associated or similar to genes associated with other types of inherited retinal degeneration, were included in the sensory cilium proteome or EyeSAGE data, or were highly expressed in the retina. Standard sequencing in two affected members and validation of variants found in both of them using the remainder of the family led to the identification of only one variant, located in *RPE65*. Meanwhile, WES in three affected individuals allowed the identification of 3,437 new variants, reduced to 1,373 after excluding variants that were synonymous or also found in one control DNA. Only three variants remained when restricting to those located in the linkage

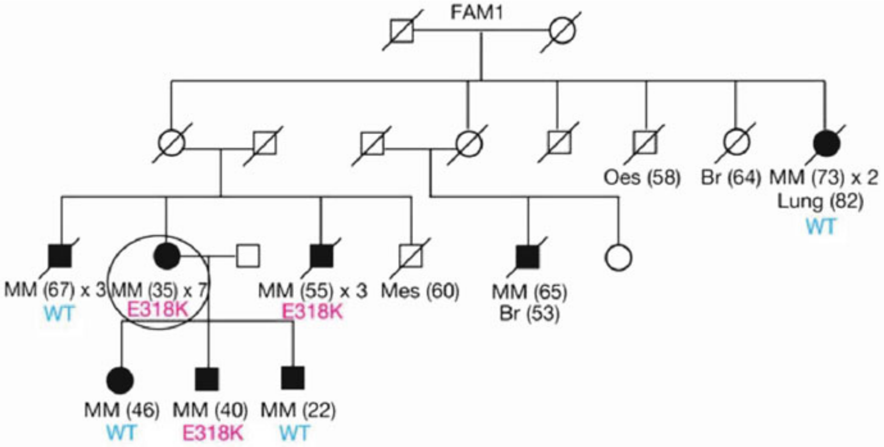

Fig. 3.1 Family 1 showing segregation of the *MITF* E318K variant in some, but not all, affected individuals (Yokoyama et al. 2011). The *circled* individual is the one in which the variant was identified by whole genome sequencing

region, and, of those three variants, the only one present in all three affected individuals was the same *RPE65* variant identified with the concurrent linkage-based approach. Evaluation of 12 Irish patients with a range of other inherited retinal degenerations revealed that one patient, as well as his two affected daughters, initially diagnosed with choroideremia but without a mutation in the *CHM* gene, had the *RPE65* mutation on the same haplotype as the extended adRP family. Combined linkage analysis of the two pedigrees yielded a maximum two-point LOD score of 5.3 at 0% recombination from the mutation. The authors commented that mutations in *RPE65* are a known cause of recessive RP and Leber congenital amaurosis but had never been associated with dominant disease. The less severe phenotype with reduced penetrance observed in the families studied by Bowne et al. (2011) was consistent with one mutant allele rather than two. The authors thus warned that carriers of “recessive” missense mutations in *RPE65* should be evaluated for subtle signs of disease. They also suggested that, given the co-occurrence of choroidal disease in the large adRP family and the diagnosis of choroideremia in the smaller family, mutations in *RPE65* may be the cause of choroideremia in families in which the typical X-linked gene, *CHM*, has been excluded.

While many studies have performed linkage analysis prior or in parallel to the WES/WGS experiments, as exemplified in the previous paper, others have used it to confirm co-segregation of the disease phenotype and variants usually identified in a small number of cases prior to further genetic studies and functional characterization. Yokoyama et al. (2011) looked for germ line mutations predisposing to melanoma starting from whole genome sequencing of a single individual in a family with eight affected relatives in three generations (Fig. 3.1) (Yokoyama et al. 2011). From 410 novel variants thus identified, a variant in *MITF*, a gene known for being somatically amplified or mutated in a subset of melanomas, was found to be present in three out of seven cases tested in the proband's family. Testing of additional patients from families with multiple melanoma cases eventually confirmed the presence of the *MITF* E318K variant in 31 unrelated cases with at least one first- or second-degree relative diagnosed with melanoma. A formal linkage analysis of melanoma with E318K was performed under a dominant model with reduced penetrance and a 5% phenocopy rate and produced a maximum LOD score of 2.7, a result consistent with E318K being an intermediate risk variant. Finally, the authors confirmed the role of E318K in melanoma by means of case–control association studies as well as expression profiling and analysis.

3.3 WES/WGS with Inconclusive Linkage Data

Many linkage studies have resulted in the identification of candidate regions that, however, have not led to the discovery of a specific disease gene. While some of these failures may be explained by false-positive findings (Ioannidis 2005), in other cases the size of the candidate region(s) may simply have prevented its full sequencing and therefore the disease gene identification. With the advent of next-generation sequencing (NGS) technologies, it is increasingly becoming more cost-effective to sequence the whole exome rather than a few target regions that may be relatively large in physical size and number of positional candidate genes contained.

A recent example of the use of this strategy reported by Louis-Dit-Picard et al. (2012) has led to the identification of *KLHL3* mutations in familial hyperkalemic hypertension (FHHt) (Louis-Dit-Picard et al. 2012). SNP-based linkage analysis in one informative family with five affected and seven unaffected individuals indicated six suggestive linkage regions (max LOD = 1.8), spanning a total of 35.6 Mb and containing 325 protein-coding genes. Given the number of positional candidate genes, the authors performed WES of one unaffected and three affected family members. A missense mutation in *KLHL3* was identified in one of the linkage regions on chromosome 5q31, and the same region was reported to be linked in a second family by microsatellite analysis (max LOD = 7.3). WES of three members of the second family also identified one missense mutation in *KLHL3*. Direct sequencing identified nonsynonymous *KLHL3* mutations in 11 out of 14 additional FHHt patients, including heterozygous as well as homozygous cases.

Sobreira et al. (2010) used WGS in a single individual combined with linkage analysis to identify the gene mutated in metachondromatosis (MC), another autosomal dominant disorder (Sobreira et al. 2010). Linkage analysis in seven members of a family segregating MC had identified six regions with positive LOD scores covering a total of 42 Mb, of which one reached the maximum possible in the small pedigree (7p14.1, LOD = 2.5) and two others were compatible with the presence of a single non-penetrant carrier (8q24.1 and 12q23, LOD = 1.8). Following WGS of a single proband, no variants unique to her and with a high likelihood of functional significance were found in five of these regions. However, one such variant was located in the 12q23 candidate region and was shown to be present in all affected individuals as well as the hypothetical non-penetrant carrier who, on more careful examination, showed symptoms of the disease, as well as did her daughter who had not previously been examined. This result clearly points out a possible explanation for the negative results of linkage studies that have been followed up by sequencing only the regions of maximum LOD scores as well as the importance, never overstated, of careful phenotyping. When linkage is inconclusive, like it was in this case, and more than one candidate region exists, it is now more efficient to perform a WES or even a WGS experiment rather than sequencing several candidate regions using more traditional approaches. Interestingly in this case, although the authors performed a WGS study, sequencing of the exome only would have been just as fruitful.

3.4 Homozygosity Mapping and WES/WGS Studies

Homozygosity mapping is a powerful approach for disease gene mapping in cases of rare recessive disease observed in consanguineous families (Lander and Botstein 1987). Homozygosity mapping is a variation on linkage analysis that exploits the fact that in rare autosomal recessive disorders, affected individuals, especially those born from consanguineous parents, are expected to be homozygous for alleles identical by descent (IBD) at the disease locus and at the marker loci tightly linked to it, a condition sometimes referred to as autozygosity. A search for regions of linkage can thus be achieved by looking for regions of IBD homozygosity in a few affected individuals. This approach has been applied successfully to several rare autosomal recessive disorders and is now being further revamped by pairing it with results of WES or WGS experiments.

A good example of this approach is the study of Wang et al. (2011), aimed at identifying novel disease alleles or genes involved in Leber congenital amaurosis (LCA) by combining genetic mapping with WES (Wang et al. 2011). LCA is a genetically heterogeneous eye dystrophy that most often presents as a recessive disease. Standard Sanger sequencing of the coding exons from all 15 known LCA disease genes in one affected member from a consanguineous family from Saudi Arabia had failed to find the causative homozygous mutation. The authors then performed homozygosity mapping by genotyping three affected members using the Illumina 370 K SNP array and identified a single novel region of homozygosity spanning 11.2 Mb on chromosome 15 shared by all three affected members. Due to the high gene density in this region, direct Sanger sequencing of all coding exons was unfeasible. By WES of a single affected individual, they found a total of 370,000 SNPs and in/dels. After filtering out common variants and variants that did not affect protein-coding or splicing regions, they were left with 352 candidate variants. The only homozygous missense change in the critical region was located in *BBS4*, a gene known to cause Bardet-Biedl syndrome (BBS). BBS is a rare human genetic disorder that, similarly to LCA, presents ocular phenotypes as a common clinical feature. The authors confirmed the presence of the mutation with Sanger sequencing and that it segregated with the disorder in the family by direct genotyping of all the other members. Moreover, they excluded the presence of this variant in 200 normal matching controls, including 96 from Saudi Arabia, and confirmed its pathological role in a zebrafish model.

Using a similar approach, Schrader et al. (2011) investigated an extended family that presented with autosomal recessive spondyloepiphyseal dysplasia (SED), retinitis pigmentosa (RP), and a high incidence of corneal abnormalities among affected individuals (Schrader et al. 2011). Given the geographical isolate from which the family originated, the known consanguinity, and the autosomal recessive inheritance pattern of the disease, the authors hypothesized that the causative mutation would be novel and would lie within an extended block of linkage that was homozygous in the affected individuals and heterozygous in the unaffected obligate carriers. For this reason, they performed WES in three affected individuals and one unaffected obligate carrier from the family and in parallel applied SNP chip genotyping to the same individuals to rule out homozygous microdeletions and to identify blocks of linkage surrounding candidate novel variants. Among the variants detected by WES, only two uncommon ones were homozygous in all three affected individuals and heterozygous in the obligate carrier: a nonsynonymous variant in *RPL3L* and a 6-bp deletion in *GNPTG*. These variants were validated by Sanger sequencing and found to co-segregate with the disease in the other 14 family members. Furthermore, both variants were located within a 3.5-Mb region of linkage defined by homozygosity in affected individuals, containing 202 UCSC genes. The authors focused their analysis on the mutation in *GNPTG,* a gene associated with mucolipidosis type IIIγ (MLIIIg), an autosomal recessive lysosomal storage disorder with a broad phenotypic spectrum including progressive destruction of the hip joint, increased lysosomal enzyme levels in serum, and reduced lysosomal enzyme levels in cultured fibroblasts. Elevated lysosomal enzyme activity was confirmed in the serum of affected individuals, and histochemical analysis of a section of the femoral head of one member of the family revealed microvesicular changes in the chondrocytes. Thus, their approach eventually led to a molecular diagnosis of MLIIIg and to a further broadening of the phenotypic spectrum of MLIII. These authors compared the traditional linkage mapping, homozygosity mapping, and whole exome sequencing approaches and concluded that the latter should be sufficient to identify causal mutations in most Mendelian disorders. However, they did recommend SNP array genotyping in at least one individual to rule out homozygous deletions and duplications that could be missed otherwise.

The study of Puffenberger et al. (2012) in the Amish and Mennonite populations of Pennsylvania represents perhaps the best example of the power of the combined homozygosity mapping/WES approach (Puffenberger et al. 2012). Taking advantage of the characteristics of these populations, including relative isolation and high inbreeding, the same authors had previously identified the loci for 28 genetic disorders by homozygosity mapping. For 11 of these, however, the corresponding gene could not be found, and the authors cited the large size of the candidate regions and large number of genes there contained as the main obstacles to achieving this goal. The authors looked at seven such diseases where gene mapping had been achieved by SNP genotyping using either a 10K or 50K SNP microarray. In six cases, only one candidate region had been identified using either two or more affected individuals from a single family or multiple cases from different families; in the remaining case, analysis of two affected siblings, their parents, and six unaffected siblings resulted in the identification of 12 candidate genomic regions each greater than 5 Mb. Even when a single genomic region was consistent with linkage, the average size of the candidate regions was 4.4 Mb (range 1.6–8.4 Mb), and the average number of genes included in them was 79 (range 22–187). Sequencing a number of candidate genes included between 2 and 45 for each condition failed to identify the causative variants. In contrast, WES of a number of patient samples included between one and five (for a total of 15 cases for all disorders) and subsequent filtering of candidate variants led to the identification of a single causal mutation in all seven diseases, five of which located in genes that had not previously been associated with these conditions. Criteria for disease variant identification included homozygosity in the affected patients, localization in the regions of linkage, and absence from dbSNP 129 and 1000 Genomes Project. All putative disease variants were confirmed by Sanger sequencing in the cases and their available relatives; their frequency in the population was further evaluated in more than 400 chromosomes, and no homozygous controls were identified. In some instances, the presence of the same mutation was confirmed in independent cases with the same phenotype, or pathogenicity was supported by high PolyPhen2 scores. Finally, *in vitro* studies supported the causal relationship between some of the candidate variants and the respective disease phenotype.

Interestingly, these authors noted that when multiple cases were available, the use of WES coupled with the assumption of mutation homozygosity in the patients but not in unaffected individuals would have been sufficient to identify the disease gene mutations even in the absence of mapping data. In fact, in each of these cases, only one variant was identified that satisfied these conditions. Even when only a single case was sequenced, the number of potentially pathogenic homozygous variants was relatively small, perhaps surprisingly given the high inbreeding coefficients of these populations, and only six variants were not homozygous in unaffected controls. In conclusion, the authors suggested that a strategy based on WES and a search for homozygous or compound heterozygous novel variants in the same gene in multiple affected individuals has a high chance of being successful even in outbred populations.

However, a cautionary note on the use of WES comes from the study of Bloch-Zupan et al. (2011) of two first-degree cousins affected with major dental developmental defects (Bloch-Zupan et al. 2011). Because of the high consanguinity in the family, the authors used homozygosity mapping to identify a critical region for the disease gene located on 6q27-ter and spanning 3 Mb. Sequencing of two candidate genes in the critical region led to the identification of a splice site mutation in the *SMOC2* gene that was present in the homozygous state in the two children and in the heterozygous state in the children's carrier parents. To confirm that no other mutations were present in the children that may explain the phenotype, Bloch-Zupan et al. performed WES in one of the two patients. Interestingly, they found out that 6.6 Kb of the 3-Mb critical region identified by homozygosity mapping were not sufficiently covered by the WES data, and specifically the mutation in *SMOC2* identified by traditional sequencing was not detected by WES. Analysis of the genomic region containing the mutation showed it to be GC rich, and independent sequencing experiments from other projects confirmed the deficit in sequence coverage. The authors concluded that had they only applied the exome-capture approach, they would have missed the causative mutation in their patients.

3.5 Linkage and WES/WGS in Quantitative Trait Analysis

Integration of linkage analysis and sequencing studies can also be used successfully for the identification of the molecular basis underlying a quantitative trait locus (QTL), as exemplified by the study of Bowden et al. (2010) on adiponectin plasma levels (Bowden et al. 2010). In this case, variance-component linkage analysis, a popular approach for QTL mapping, had identified a strong linkage signal in a single genomic region (3q, LOD = 8.02). The linkage critical region contained an ideal candidate for variation in adiponectin plasma level, the adiponectin protein-coding gene *ADIPOQ*. However, association to common variants in this gene did not explain the linkage signal. The authors cleverly used the linkage results to select individuals for sequencing by prioritizing families with a higher individual LOD score in the critical region. WES of three individuals with values of adiponectin plasma level in the tails of its distribution (one high, two low) from two of these families led to the identification of a single variant not previously reported and present in the two low-adiponectin samples. Through conventional sequencing and additional genotyping, the same rare variant was shown to co-segregate with the plasma adiponectin trait in the linkage families and to account for most of the 3q linkage signal. While this study may be considered just a proof of principle given the presence of a very strong candidate gene in the linkage critical region, it showed that the combination of QTL linkage mapping and sequencing of individuals with extreme values of the quantitative trait is a potentially valuable approach for the identification of rare variants, as it has been recently advocated (Cirulli and Goldstein 2010).

3.6 Linkage Analysis as an Aid in Designing WES/WGS Experiments

Bowden et al. (2010) have shown that results from linkage analysis can be utilized in WES/WGS projects to optimize family selection (Bowden et al. 2010). The GAW17 dataset provided an opportunity to investigate the efficacy and cost efficiency of various strategies for next-generation sequencing sample selection as a follow-up to linkage analysis (http://www.gaworkshop.org). The GAW17 dataset included genome-wide genotype data as well as exome sequencing for approximately 3,000 genes for eight simulated pedigrees. In addition, risk factors including age, smoking status, and three quantitative trait variables were provided for each individual. Allen-Brady et al. (2011) performed linkage analyses on these families and compared nine approaches for selecting subjects for subsequent partial exome sequencing (Allen-Brady et al. 2011). For this study, the authors split the original 8 pedigrees into 23 smaller pedigrees and used 10 of the 200 replicate datasets provided by the GAW17 organizers. Using the results from a logistic regression model which incorporated the five risk factors, Allen-Brady et al. (2011) classified all individuals as either high-covariate subjects whose nongenetic risk factors are highly predictive of their affection status or low-covariate subjects whose nongenetic risk factors are poorly predictive of their affection status. They found that selecting for exome sequencing all affected individuals classified as low-covariate and possessing a linked haplotype identified in the linkage analysis was the most reliable strategy across both recessive and dominant models. Furthermore, selecting the youngest affected individuals may provide a satisfactory alternative in cases where the major nongenetic risk factors are unknown.

Starting from the GAW17 pedigrees, Cai et al. (2011) defined as high risk those with at least 15 total meioses between case subjects and a statistical excess of disease ($p < 0.01$) over all 200 replicates, thus identifying 18 pedigrees (Cai et al. 2011). They then performed a linkage analysis using the shared genomic segment (SGS) method (Thomas et al. 2008), modified in order to examine sharing among all pairs of cases instead of all subjects, and assessed the test statistic against an empirical distribution. Following this approach, they successfully identified at least one region containing one true causal variant in 13 out of the 18 high-risk pedigrees. Additional causative genes would have been identified at lower significance thresholds ($p \leq 0.01$), but this would also have increased false-positive findings. The inability to detect the other rare causative variants was ascribed to the small sample size and high heterogeneity. Of note, this method considered only pairs of relatives and did not take into account the specific relationships between them. The authors claimed that when they incorporated the specific relationships into the analysis, they did not see substantial improvement in the results.

Gagnon et al. (2011) analyzed the GAW17 data to select which families should be sequenced in order to identify rare variants that have large effects on quantitative trait variance (Gagnon et al. 2011). They hypothesized that rare functional variants segregating with a quantitative phenotype are more likely to be present in families with more quantitative trait loci (QTLs) than the other families. For this reason, they estimated the mean number of QTLs in each family by segregation analysis assuming an oligogenic linear model and selected one family with more QTLs than the average. They then tested this family for linkage using a variance-component oligogenic approach. Sequencing data from regions surrounding loci with at least modest evidence of linkage (LOD ≥ 0.6) were investigated for the presence of rare functional variants, and the variants thus detected were analyzed for association to the quantitative traits. By this approach, they identified one region with a maximum LOD of 5.3 ($p = 4 \times 10^{-7}$) for one trait and two regions with maximum LOD of 2.02 ($p = 0.001$) for another trait for a total of 216 and 85 variants that were thus tested for association with the same traits in all the families. They correctly identified two rare functional variants, including one private to the family selected for sequencing. Both variants were located in regions with a combined LOD score in all families greater than 4. All other variants identified in the family selected for linkage analysis were false positives, and all had LOD scores below 2, confirming once again the importance of linkage evidence in discovering the actual causative variants. The authors claimed that prioritizing the sequencing of carefully selected extended families is a simple and cost-efficient design strategy to identify rare functional variants that explain a significant proportion of the trait variance, especially for variants that are unlikely to segregate in more than a few families. However, they noted that the use of large, multigenerational families remains crucial, and other complementary designs are still needed to further decrease type I error, including parallel analysis of large samples of unrelated individuals.

Other research utilizing the GAW17 dataset focused on the potential of linkage studies to guide deep sequencing efforts by narrowing the search space to genomic regions under linkage peaks. Choi et al. (2011) compared the effectiveness of two mapping strategies: (1) performing association tests which adjust for familial relationships on variants identified by whole exome sequencing and (2) performing a linkage analysis followed by targeted sequencing of regions beneath the linkage peaks and family-based association on variants identified in those regions (Choi et al. 2011). They found that both mapping strategies demonstrate a limited ability to detect association for variants of small effect sizes. In addition, both strategies only found the same two loci with a reasonable amount of power (>70%). However, the linkage-guided strategy on average required sequencing of only 2.5% of the whole exome and found 52% of the associated loci identified by the whole exome sequencing strategy. Choi et al. concluded that while the whole exome sequencing strategy appears more powerful, targeted sequencing under linkage peaks still offers a viable and cost-effective alternative.

3.7 New Methods of Linkage Analysis and WES/WGS Studies

One of the challenges of linkage analysis has always been the analysis of large pedigrees, which, however, can also provide very valuable information. In place of traditional linkage analysis as a filter for whole exome sequence data, some groups have attempted to find equally effective strategies that are less computationally intense. Markello et al. (2012) advocated using high-density genotyping panels and Boolean logic for recombination mapping efforts (Markello et al. 2012). High-density genotyping panels provide a relatively cost-effective option that covers introns as well as intergenic regions as compared to a strategy that extracts SNPs directly from WES. However, double-crossover events in between contiguous informative markers are problematic for this approach. Markello et al. (2012) showed that the risk of double-crossover events between informative SNPs was extremely low through simulation and a literature search to identify the smallest reported interval containing a double crossover. In addition, these authors compared their recombination mapping method to a traditional multipoint linkage analysis utilizing microsatellite markers and demonstrated that the identified regions were identical. Finally, they incorporated the method into a filtering scheme on variants identified by whole exome sequencing of a single pedigree and identified two potential disease-causing mutations in a candidate gene for progressive myoclonic epilepsy type 3.

Marchani and Wijsman (2011) proposed a method to test for linkage that enjoys computational advantages because of how it records and groups the inheritance patterns (Marchani and Wijsman 2011). Their method can employ Markov chain Monte Carlo (MCMC) techniques to handle large pedigrees and can also be applied to common disorders where Mendelian inheritance is not strictly followed and genetic heterogeneity is present. Visualization of IBD sharing allows the investigator to observe which affected pedigree members share a genetic segment within a region tied to a linkage signal. This knowledge allows efficient selection of individuals for deep sequencing. The strategy advocated by Marchani and Wijsman (2011) is to select the most distantly related affected individuals who share such a DNA segment.

Other research has focused on the utility of gleaning variants directly from next-generation sequencing for use in linkage studies. Smith et al. (2011) investigated the efficacy of using genotypes generated from WES as a surrogate for array-based genotypes in linkage studies (Smith et al. 2011). Limitations of the WES approach include coverage gaps in non-exonic regions, higher genotyping error rates, and markers with lower heterozygosity. Smith et al. (2011) performed linkage analyses on three pedigrees with different Mendelian neurological disorders employing both array-based markers and HapMap phase II SNP genotypes derived from WES. They found (1) a substantial number of WES-derived SNPs resided outside of coding regions due to a technical artifact of the sequencing method, (2) almost a 100% concordance rate for genotypes derived from either of the two methods, signifying an acceptable error rate for WES-derived SNPs, and (3) the resulting LOD scores for the analyses using genotypes derived from WES closely resembled those for the analyses using genotypes acquired by array-based technology at the positions of linkage peaks. Smith et al. (2011) concluded that, while SNP arrays are preferable for linkage studies due to better coverage and marker informativeness, generating genotypes for linkage studies directly from WES data is a viable option.

3.8 Conclusions

The advent of NGS technologies and the ability to sequence whole exomes or genomes have generated a new interest in the analysis of family data and thus in genetic linkage analysis (Bailey-Wilson and Wilson 2011). Linkage analysis is ideal for identifying the location of rare disease-causing variants, such as those that are the object of analysis in most sequencing studies. Candidate loci identified by linkage studies can now be examined more extensively than ever before, leading to a new wave of gene discoveries, particularly for Mendelian disorders (Ng et al. 2010).

Whether the same success will occur in the analysis of complex traits remains to be demonstrated. Nonetheless, it is especially important that all available approaches are considered when tackling a difficult question such as the identification of the genetic basis of complex disease, and we recommend that linkage analysis should be considered whenever families are available to investigators.

References

Allen-Brady K, Farnham J, Cannon-Albright L. Strategies for selection of subjects for sequencing after detection of a linkage peak. BMC Proc. 2011;5 Suppl 9:S77.

Bailey-Wilson JE, Wilson AF. Linkage analysis in the next-generation sequencing era. Hum Hered. 2011;72(4):228–36.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745–55.

Bloch-Zupan A, Jamet X, Etard C, Laugel V, Muller J, Geoffroy V, Strauss JP, Pelletier V, Marion V, Poch O, Strahle U, Stoetzel C, Dollfus H. Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. Am J Hum Genet. 2011;89(6):773–81.

Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl):228–37.

Bowden DW, An SS, Palmer ND, Brown WM, Norris JM, Haffner SM, Hawkins GA, Guo X, Rotter JI, Chen YD, Wagenknecht LE, Langefeld CD. Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS family study. Hum Mol Genet. 2010;19(20):4112–20.

Bowne SJ, Humphries MM, Sullivan LS, Kenna PF, Tam LC, Kiang AS, Campbell M, Weinstock GM, Koboldt DC, Ding L, Fulton RS, Sodergren EJ, Allman D, Millington-Ward S, Palfi A, McKee A, Blanton SH, Slifer S, Konidari I, Farrar GJ, Daiger SP, Humphries P. A dominant mutation in RPE65 identified by whole-exome sequencing causes retinitis pigmentosa with choroidal involvement. Eur J Hum Genet. 2011;19(10):1074–81.

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;5 Suppl 9:S9.

Choi SH, Liu C, Dupuis J, Logue MW, Jun G. Using linkage analysis of large pedigrees to guide association analyses. BMC Proc. 2011;5 Suppl 9:S79.

Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.

Gagnon F, Roslin NM, Lemire M. Successful identification of rare variants using oligogenic segregation analysis as a prioritizing tool for whole-exome sequencing studies. BMC Proc. 2011;5 Suppl 9:S11.

Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.

Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236(4808):1567–70.

Louis-Dit-Picard H, Barc J, Trujillano D, Miserey-Lenkei S, Bouatia-Naji N, Pylypenko O, Beaurain G, Bonnefond A, Sand O, Simian C, Vidal-Petiot E, Soukaseum C, Mandet C, Broux F, Chabre O, Delahousse M, Esnault V, Fiquet B, Houillier P, Bagnis CI, Koenig J, Konrad M, Landais P, Mourani C, Niaudet P, Probst V, Thauvin C, Unwin RJ, Soroka SD, Ehret G, Ossowski S, Caulfield M, Bruneval P, Estivill X, Froguel P, Hadchouel J, Schott JJ, Jeunemaitre X. KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. Nat Genet. 2012;44(5):609.

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):747–53.

Marchani EE, Wijsman EM. Estimation and visualization of identity-by-descent within pedigrees simplifies interpretation of complex trait analysis. Hum Hered. 2011;72(4):289–97.

Markello TC, Han T, Carlson-Donohoe H, Ahaghotu C, Harper U, Jones M, Chandrasekharappa S, Anikster Y, Adams DR, Gahl WA, Boerkoel CF, Program NCS. Recombination mapping using Boolean logic and high-density SNP genotyping for exome sequence filtering. Mol Genet Metab. 2012;105(3):382–9.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010;42(1):30–5.

Puffenberger EG, Jinks RN, Sougnez C, Cibulskis K, Willert RA, Achilly NP, Cassidy RP, Fiorentini CJ, Heiken KF, Lawrence JJ, Mahoney MH, Miller CJ, Nair DT, Politi KA, Worcester KN, Setton RA, Dipiazza R, Sherman EA, Eastman JT, Francklyn C, Robey-Bond S, Rider NL, Gabriel S, Morton DH, Strauss KA. Genetic mapping and exome sequencing identify variants associated with five novel diseases. PLoS One. 2012;7(1):e28936.

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273(5281):1516–17.

Schrader KA, Heravi-Moussavi A, Waters PJ, Senz J, Whelan J, Ha G, Eydoux P, Nielsen T, Gallagher B, Oloumi A, Boyd N, Fernandez BA, Young TL, Jones SJ, Hirst M, Shah SP, Marra MA, Green J, Huntsman DG. Using next-generation sequencing for the diagnosis of rare disorders: a family with retinitis pigmentosa and skeletal abnormalities. J Pathol. 2011;225(1):12–8.

Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M. Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol. 2011;12(9):R85.

Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, Gumbs CE, Pevsner J, Thomas G, Valle D, Hoover-Fong JE, Goldstein DB. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet. 2010;6(6):e1000991.

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.

Wang H, Chen X, Dudinsky L, Patenia C, Chen Y, Li Y, Wei Y, Abboud EB, Al-Rajhi AA, Lewis RA, Lupski JR, Mardon G, Gibbs RA, Perkins BD, Chen R. Exome capture sequencing identifies a novel mutation in BBS4. Mol Vis. 2011;17:3529–40.

Yokoyama S, Woods SL, Boyle GM, Aoude LG, MacGregor S, Zismann V, Gartside M, Cust AE, Haq R, Harland M, Taylor JC, Duffy DL, Holohan K, Dutton-Regester K, Palmer JM, Bonazzi V, Stark MS, Symmons J, Law MH, Schmidt C, Lanagan C, O'Connor L, Holland EA, Schmid H, Maskiell JA, Jetann J, Ferguson M, Jenkins MA, Kefford RF, Giles GG, Armstrong BK, Aitken JF, Hopper JL, Whiteman DC, Pharoah PD, Easton DF, Dunning AM, Newton-Bishop JA, Montgomery GW, Martin NG, Mann GJ, Bishop DT, Tsao H, Trent JM, Fisher DE, Hayward NK, Brown KM. A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. Nature. 2011;480(7375):99–103.

Zeggini E. Next-generation association studies for complex traits. Nat Genet. 2011;43(4):287–8.