Hum Genet (2014) 133:1331–1341
DOI 10.1007/s00439-014-1479-4

REVIEW PAPER

Using familial information for variant filtering in high-throughput sequencing studies

Melanie Bahlo · Rick Tankard · Vesna Lukic · Karen L. Oliver · Katherine R. Smith

Received: 1 April 2014 / Accepted: 7 August 2014 / Published online: 17 August 2014

Abstract High-throughput sequencing studies (HTS) have been highly successful in identifying the genetic causes of human disease, particularly those following Mendelian inheritance. Many HTS studies to date have been performed without utilizing available family relationships between samples. Here, we discuss the many merits and occasional pitfalls of using identity by descent information in conjunction with HTS studies. These methods are not only applicable to family studies but are also useful in cohorts of apparently unrelated, 'sporadic' cases and small families underpowered for linkage and allow inference of relationships between individuals. Incorporating familial/pedigree information not only provides powerful filtering options for the extensive variant lists that are usually produced by HTS but also allows valuable quality control checks, insights into the genetic model and the genotypic status of individuals of interest. In particular, these methods are valuable for challenging discovery scenarios in HTS analysis, such as in the study of populations poorly represented in variant databases typically used for filtering, and in the case of poor-quality HTS data.

M. Bahlo (✉) · R. Tankard · V. Lukic · K. R. Smith
The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
e-mail: bahlo@wehi.edu.au

M. Bahlo · R. Tankard · K. R. Smith
Department of Medical Biology, University of Melbourne, Melbourne, VIC 3010, Australia

M. Bahlo
Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia

K. L. Oliver
Epilepsy Research Centre, The University of Melbourne, Austin Health, Heidelberg, VIC 3084, Australia

Introduction

High-throughput sequencing (HTS) has proven to be highly successful at identifying causal variants for many genetic disorders with a variety of underlying genetic etiologies (Boycott et al. 2013). This includes genetic disorders segregating in clear dominant and recessive inheritance modes, but also disorders that are caused by less well-studied genetic models such as de novo mutations. HTS has also been successfully applied to small cohort studies, even in the presence of locus heterogeneity (Riviere et al. 2012). HTS can potentially identify all causal variants, hence cohorts of unrelated individuals with different causal alleles are very powerful for the implication of a causal gene since the accumulation of rare, likely pathogenic variants in a single gene, gene family or pathway, is highly unlikely, even for just a few affected individuals.

HTS has identified variants that cause germline Mendelian disorders (Ng et al. 2010a, b) segregating in a highly penetrant manner, expediting the discovery of many such variants (Heron et al. 2012). Such diseases often come to notice because they segregate in one or more families. Prior to the HTS era, linkage analysis, or other identity by descent (IBD) approaches were applied to identify genomic regions of interest where the affected status co-segregates with genotypic status. The application of HTS has resolved several such Mendelian disorders that had achieved promising linkage results, including highly significant linkage results (parametric LOD scores >3), and were awaiting causal gene discovery (Corbett et al. 2010, 2011; Koenekoop et al. 2012; Reversade et al. 2009). As such, identification of a causal gene for a genetic disorder before the advent of HTS usually required a two-step process, with the first step being a genome-wide, or chromosome-wide, localization step, utilizing IBD information inferred

through co-segregation of genetic markers such as microsatellites and single-nucleotide polymorphisms (SNPs) with disease (linkage analysis). The second step consisted of the actual identification of the causal variant with Sanger sequencing making this task manageable by focusing the search on a particular genomic region, usually only a few Megabases in length. Genome-wide significant linkage was usually required before traditional Sanger sequencing commenced, but without any obvious candidate genes this could then lead to the laborious sequencing of all genes (50–100, depending on the size and location of the linkage region) in the region, potentially taking years before a causal variant and gene were found. HTS has led to an expedition of the discovery of causal variants by circumventing the gene by gene sequencing step (Fig. 1). Furthermore, since the HTS approach is now most cost-effectively performed genome-wide, via approaches such as whole exome sequencing (WES), the first step of genomic localization has been seen as unnecessary and more often bypassed. Probands of small families may then be contributed to consortium studies where they are treated as singletons, ignoring the family information.

HTS has also been applied successfully to identify causal mutations in a number of different genetic models including those not previously amenable to analysis. This includes de novo germline mutations (Epi4K Consortium et al. 2013; Jiang et al. 2013; Lim et al. 2013). Sporadic mutations benefit from the application of the trio approach, where variants inherited from the parents are eliminated to identify the novel mutations in the proband. This has led to new insights into the genetic basis of complex disorders such as autism and other neuropsychiatric disorders, leading to further debates about the role of such variants in heritability calculations (Gratten et al. 2013). Sporadic

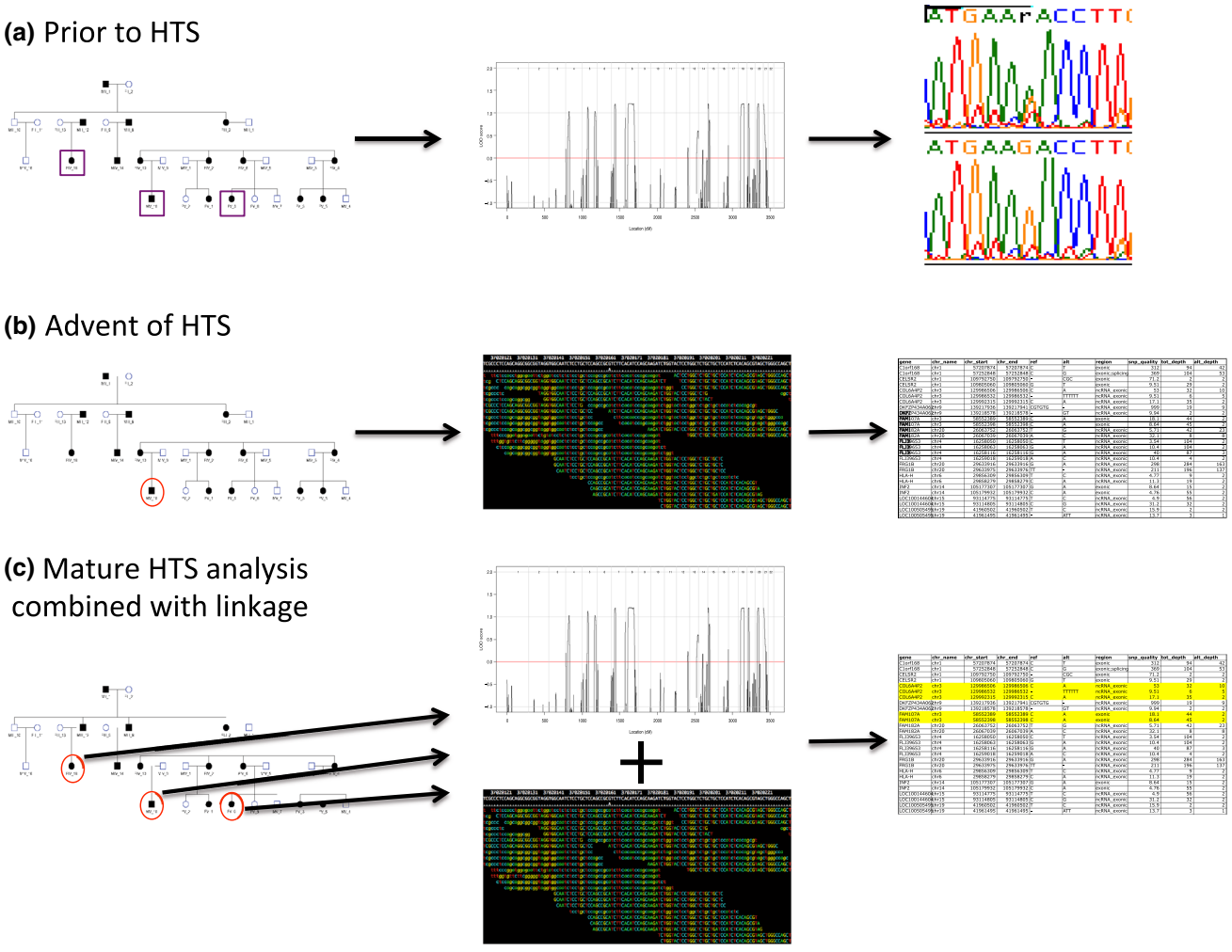

Fig. 1 Identification of a causal variant in a family study, **a** prior to the advent of HTS, using linkage and Sanger sequencing, **b** advent of HTS, using only a proband, or one or two individuals from a family, ignoring relationships between the individuals, leading to lengthy lists of putative causal variants, and **c** mature HTS analysis, where family information is incorporated using genotypes derived from HTS data to perform a linkage or IBD analysis, leading to a much reduced list of variants

somatic variants (Baek et al. 2013; Lee et al. 2012; Poduri et al. 2012) have also been identified. This is another genetic model that was intractable prior to HTS and highlights the possibility that somatic variants may play a hitherto underestimated role in causing genetic disease outside the cancer arena.

In germline Mendelian disorders HTS has been most successful for recessive disorders, in particular autozygous disorders, which are caused by single mutations. Causes of dominant disorders continue to be more resistant to discovery for several reasons. Heterozygous variants are harder to call as variants than homozygous alternative variants because the binomial variation of reads with particular alleles is higher for the heterozygote than for the two homozygous states. Furthermore, individuals have more heterozygous than homozygous novel rare variant alleles, making it more difficult to filter out novel benign polymorphisms. This is particularly problematic in populations that are not well represented in databases, where even large numbers of novel, nonsynonymous autozygous variants may remain (Azmanov et al. 2013) after filtering using standard databases. Proof that such discovered variants are not population polymorphisms can require the sequencing of many candidate variants in a relevant population sample which is time consuming and costly.

The advent of HTS saw the need for the development of pipelines to handle the tens of thousands of variants that are routinely called in human genomes. These pipelines are still mainly based on single sample/individual analysis, hence the incorporation of familial information is not routine. Even when familial information has been available for use in conjunction with HTS data it is not always used, despite the potential for IBD or linkage information to add a powerful filter for detected variants. Familial information is often included in a very indirect way by applying a genotyping filter, based on an inferred genetic model from the pedigree, with specific genotype assumptions for unaffected individuals, carriers, non-carriers and affected individuals.

HTS studies are highly reliant on ascribing likely pathogenicity according to mutational databases such as the Leiden Open Variation Database (LOVD, http://www.LOVD.nl/), The Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/) and ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). However, these databases are known to be incomplete, missing known mutations. Additionally, Bell et al. (2011) showed that 25 % of variants, described as causal in mutation databases, were low frequency polymorphisms and not causal, highlighting the need to assess candidate variants further, gathering additional evidence for likely causality. Early HTS studies were not always required to provide additional data to support variant discoveries, such as multiple mutations in the same gene or functional work or evidence of causality in an experimental model (Piton et al. 2013). Hence, there is a need to reconsider current HTS analysis methodology. One way forward is to make greater use of additional information such as that afforded by familial information to provide greater protection from false positive findings. Many more HTS studies remain unpublished because a causal variant has failed to be identified. These studies may benefit from the approaches discussed here, potentially highlighting a hitherto unpromising variant due to the current, perhaps limited, understanding of the gene's biological role, or at least highlighting regions in the genome that require greater scrutiny, perhaps with supplemental sequencing, or generation and analysis of a second attempt at WES, with a newer, more comprehensive exome capture platform, or performing WGS.

Here, we describe a set of observations for the analysis of both single and familial HTS data that can and should be used to help in the identification of the causal variant. Making use of familial and population-based information for HTS data helps by: (1) extending quality control (QC) steps, (2) identifying IBD regions, (3) validating the genetic model, and (4) validating haplotype segregation by making use of IBD information, with or without additional SNP chip data. Using this type of information can lead to the accumulation of evidence in favor of particular variants, determine the correct genetic model, lead to the elimination of irrelevant variants and can also narrow the search space for variants to particular genomic regions.

HTS data is highly non-uniform in coverage which is influenced by GC content of the underlying sequence where both very high and very low GC content DNA sequence is captured by fewer HTS reads (Dohm et al. 2008). In addition, this known effect can be exacerbated by technology such as exome capture, or by additional biological signals such as DNA source. An example of the latter is the DNA fragmentation present in cell-free DNA which displays DNA motif bias (Chandrananda et al. 2014). This bias leads to areas of the genome that are difficult to interrogate with HTS which may require higher coverage, multiple HTS applications or necessitates improvements in HTS technology to be able to be sequenced to a sufficient depth to detect variants. Thus, by identifying regions of the genome that have been detected as IBD researchers become aware of regions that deserve greater scrutiny, possibly also for more unusual forms of genetic variation.

Therefore, rather than being relegated to history, IBD methods should be recognized as providing a powerful ally for HTS to help identify causal variants. We wish to remind researchers of the availability of these approaches, outlining some resources, which could be applied to help solve their HTS study.

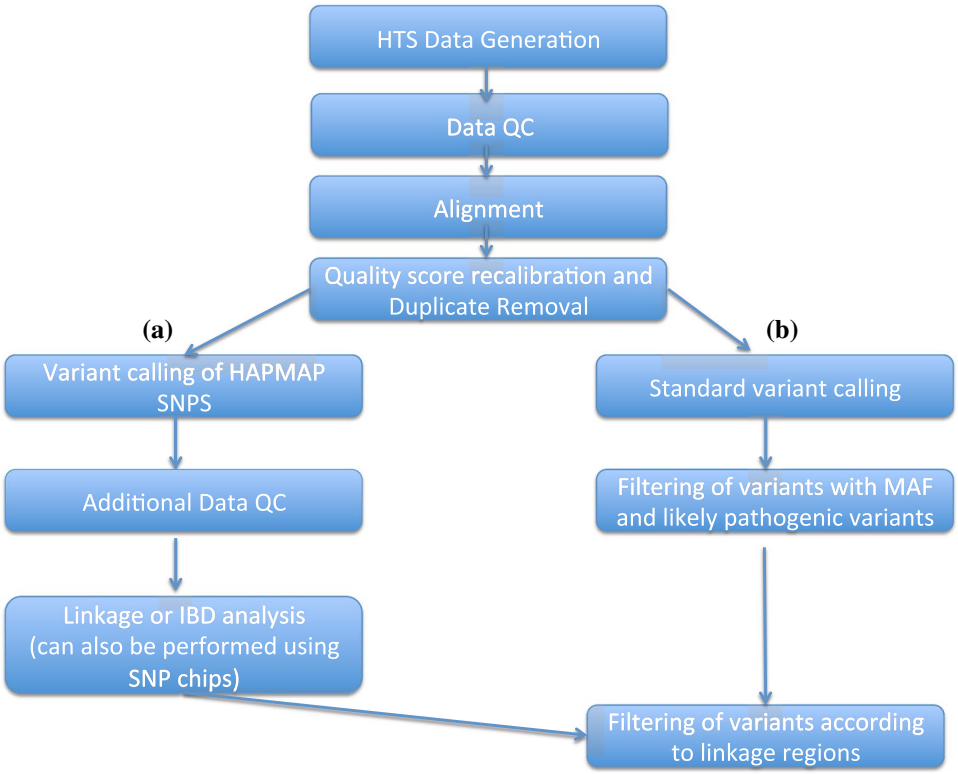

Fig. 2 Dual workflow for HTS data analysis that takes advantage of HapMap SNPs for linkage mapping (**a**), and then uses this information in the filtering steps in (**b**)

Extracting HapMap SNPs from HTS data

Many HTS studies are applied to multiple related individuals, with filtering of variants from all the individuals, ignoring familial relationships. One reason for this approach and its increasing popularity is the decreasing cost and bulk production of HTS data. This often sees researchers forgoing SNP chip data generation and traditional linkage mapping approaches, even when the family has sufficient power to produce genome-wide significant linkage. However, it is possible to extract SNPs from HTS data by instructing variant callers to genotype variants at HapMap SNP positions to create pseudo SNP chip data for all individuals that underwent HTS. Hence linkage analysis and other analyses can still be performed successfully.

Instructions on how to call HapMap SNPs from HTS data, including methods to process this data into linkage and IBD analysis ready files can be obtained via http://bioinf.wehi.edu.au/software/linkdatagen. The methods are described in Smith et al. (2011) and Bahlo and Bromhead (2009). Any genome-wide (or chromosome wide) HTS data suffices, as long as enough HapMap SNPs are covered sufficiently to allow genotype calling. The workflow required is more complicated but is not difficult to implement (Fig. 2).

WES by definition captures mainly exonic DNA. Exonic regions are known to have more SNPs that undergo balancing selection leading to lower MAFs on average (1000 Genomes Project Consortium et al. 2010), thus genotyping data from WES will not be as useful for IBD and linkage analysis as SNP chips or even genotyping data extracted from comparable median coverage WGS to the median coverage WES data (targeted areas only). Table 1 outlines the numbers of HapMap SNPs that are called with $\geq 10\times$ coverage in more than 50 % of samples using one WES capture platform in current use (Agilent V5 + UTR). This shows the availability of tens of thousands of SNPs for linkage and IBD analysis. Many IBD and linkage analyses make use of only ~10,000 SNPs due to the constraints of linkage equilibrium that are imposed to satisfy the first-order Markov models that underpin most of these methods. In Smith et al. (2011), we compare genome-wide LOD scores derived from SNPs genotyped from the Illumina TruSeq capture platform to those derived from a high-density SNP chip, showing very few differences in the linkage analysis results. Inbreeding inference utilizing HTS genotypes is also possible with FSuite (Gazal et al. 2014), showing good performance.

Additional quality control steps with familial HTS data

The bulk processing and high-throughput nature of HTS usually necessitates batching of samples, which is often accomplished through molecular barcoding. The generation

Table 1 HapMap SNPs available for linkage and IBD analysis based on ≥10× average coverage of the targeted regions in at least 50 % of samples

Population	Available SNPs	Covered SNPs	% of total SNPs	Number SNPs with Het >0.3	% of covered SNPs
CEU_2	3801563	153388	4.03	50231	32.75
CHB_2	3827537	153484	4.01	47498	30.95
JPT_2	3827726	153689	4.02	47351	30.81
YRI_2	3750555	152239	4.06	51241	33.66
CEU_3	1520715	88970	5.85	39174	44.03
ASW_3	1463106	83303	5.69	41279	49.55
CHB_3	1519591	88653	5.83	36806	41.52
CHD_3	1246085	69281	5.56	35887	51.80
GIH_3	1337706	74005	5.53	39370	53.20
JPT_3	1518437	88455	5.83	36720	41.51
LWK_3	1440446	83045	5.77	39271	47.29
MEX_3	1380212	79198	5.74	38910	49.13
MKK_3	1451099	81916	5.65	40321	49.22
TSI_3	1347642	74645	5.54	38875	52.08
YRI_3	1520811	89382	5.88	39684	44.40

The selected SNPs are over the four HapMap Phase II populations (designated by a _2) and the eleven HapMap Phase III populations (designated by _3). Data is based on the analysis of 20 Agilent V5 + UTR WES captured samples that were sequenced with Illumina HiSeq 2000 sequencing at the Australian Genome Research Facility, Melbourne, Australia

of input for HTS platforms is thus a source of potential error with sample mix-ups or mishaps in the sequencing library preparation step that can lead to contaminated samples. As such it is important to check sample veracity prior to analysis. The ability to extract genotype information allows an expanded repertoire of error checks that make use of the genotyping calls. Mendelian error detection, gender estimation and population-based checks, based on the genotyping data, can identify sample mix-ups, pedigree misspecifications and sample contamination. We now describe these in further detail.

Variant calling and Mendelian error checking

HTS analysis usually proceeds through standard pipelines such as those described in Altmann et al. (2012), which are predicated on single samples. We recommend calling samples to be analyzed as a genetic entity (single or multiple families) jointly, as this can increase power to identify rare variants. Furthermore, using a family-oriented HTS analysis pipeline can also produce Mendelian consistent genotype data from HTS data by making use of variant callers that incorporate familial information. FamSeq (Peng et al. 2013) and PolyMutt (Li et al. 2012) improve variant calling by making use of familial information while MATE-CLEVER (Marschall et al. 2013) performs family aware indel detection and calling. Once HapMap SNP genotypes are called, Mendelian error checking can be performed via LINKDATAGEN (nuclear family Mendelian error checks) or through even more sensitive Mendelian error detection programs such as PEDCHECK (O'Connell and Weeks 1998) and sophisticated pedigree misspecification analysis such as PREST (Sun et al. 2002) where Mendelian errors are utilized to check specified pedigree relationships. LINKDATAGEN also includes a population-based test for samples that assesses the fit of the three classes of genotype calls (homozygous reference, heterozygote, homozygous non-reference allele) against the expected values for all known HapMap populations. This test not only highlights whether an assumed population is correct but can also detect sample contamination that can occur with HTS data.

Further error checking is performed by specific linkage analysis and IBD sharing programs that can detect errors or incorporate errors into analysis. This is more important for HTS-generated genotypes since genotyping errors can be much higher than for SNP chip-derived genotyping data. The error rates depend on the thresholds placed on called variants that have to be set taking into account the median coverages achieved in the family study and the robustness of the following analyses to genotyping errors. Further error checks using probabilistic approaches make use of expected versus observed recombination events, such as those implemented in MERLIN (Abecasis et al. 2002), and are able to detect additional genotyping errors that can also be removed.

One important benefit of linkage analyses in families is that more missingness and error can be tolerated than in pairwise IBD analysis since much information can be often

imputed, with little uncertainty. We performed an analysis on a pedigree of 15 WES samples with a high level of missingness (median of medians of coverage 3, range [2, 9]) due to poor WES data generation, yet we were still able to detect a linkage peak and rapidly identify the causal variant despite the inability to identify the variant in many samples due to low coverage (Eggers et al. 2014).

IBD and linkage analyses

IBD analyses are usually performed pairwise, either with known or unknown relationships between the two individuals. IBD analysis is often performed agnostic to the disease status, although a typical application is to assess IBD sharing between affected individuals. In the case of a known relationship between a pair of individuals IBD calculations simply calculate the probability of the observed genotyping data given the relationship, where the number of meioses separating the two individuals determines the rate of change of the underlying (hidden) IBD process. IBD can be calculated as a genome-wide summary, or at specified locations in the genome. In contrast, linkage analysis is a likelihood ratio test statistic that compares the probability of the genotyping data, for possibly more than two individuals, under two different hypotheses. The null hypothesis calculates the probability of the data assuming no linkage between the phenotype (modeled as another genetic marker) and the genotyping data from many genetic markers, and compares it to the alternative hypothesis of linkage between the phenotype and the set of genetic markers. This function of the likelihood ratio is known as the LOD score and can be calculated at any position in the genome.

Linkage and IBD analyses are powerful approaches for the identification of genomic regions harboring the causal variants for Mendelian diseases. Thus, both linkage and IBD methods are useful tools for filtering detected variants in cohorts of related individuals with HTS data. To implement this, one performs IBD or linkage analysis as described above, with or without SNP chip data, prior to variant filtering. These regions are then used as an additional filter for variants that remain to be considered. Other authors have designed genetic model-specific hidden Markov models for IBD detection (Eggers et al. 2014; Krawitz et al. 2010; Roach et al. 2010; Rodelsperger et al. 2011). For autozygosity mapping some authors have made use of simple homozygous run searches, focusing their search in these regions (Bilguvar et al. 2010). Extracting HapMap SNPs and their genotypes from HTS data allows a much broader usage of SNP genotypes derived from these studies and enables the use of well-understood statistical models and existing algorithms and software for IBD and linkage analyses (Tsoi et al. 2014).

Linkage analysis not only identifies linkage regions but allows post hoc analysis of the best inferred haplotypes, generated by programs such as MERLIN. This is achieved with graphical software such as HaploPainter (Thiele and Nurnberg 2005), by examining regions of interest. Individuals that violate the assumed genetic model can be identified, such as phenocopies or incompletely penetrant individuals. Once a susceptibility haplotype is identified, it is then possible to filter the variants based on matching of the genotypes called for the variant to the susceptibility haplotype status, rather than the underlying genotypes proposed by the original phenotype. An example of where such an approach is valuable is a dominantly inherited phenotype with age-dependent penetrance, but fully penetrant otherwise. Consider a younger, unaffected individual. This individual is identified as having the susceptibility haplotype, based on the linkage analysis. Thus, if this individual is sequenced it is appropriate to seek variants that are heterozygous for this individual. However, for most rare diseases the phenotype and susceptibility haplotype carrier status will match the hypothesized genetic model. If the WES data is poor, resulting in high levels of missingness, it is important to acknowledge that the inferred haplotypes conditional on the data will likely be non-unique, thus sampling from the posterior probability and examining a few sets of haplotypes is a useful supplementary approach. This can be achieved using the --sample option in MERLIN. We used this approach in Eggers et al. (2014) to verify the robustness, or lack thereof, of the inferred haplotypes. The uncertainty in the haplotypes does not have as much impact on the calculation of the LOD scores since these are summed over all possible haplotypic states that fit the data.

With the advent of dense SNP genotyping data, such as that observable in SNP chip or HTS data, it is possible to detect relatedness between individuals (Purcell et al. 2007), without prior genealogical information, including the presence of consanguinity (Albrechtsen et al. 2009; Leutenegger et al. 2006). This relatedness inference can then be included in subsequent linkage or IBD analysis (Guergueltcheva et al. 2012; Ivanov et al. 2014). For a review of methods that are applicable for this problem see Browning and Browning (2012).

We, and others, have successfully applied this approach to HTS data, detecting inbreeding and thus making powerful use of small families (Browning and Browning 2012; Guergueltcheva et al. 2012; Smith et al. 2012, 2013). The identification of autozygous regions, even after discovery of rare or novel homozygous variants, help to provide further evidence that the variants are in fact causal with variants that are merely homozygous by state and not homozygous by descent (autozygous) being able to be filtered out.

The methods outlined here use IBD analysis to lead to a discretized, independent filter of variants. Ideally, one

would like to utilize the IBD information as weights rather than this discretized approach. Two recently published methods are the first to describe such approaches (Koboldt et al. 2014; Santoni et al. 2014). This will be useful in families where the genetic model is uncertain due to incomplete penetrance or possible phenocopies and vital for studies and pedigrees where strict Mendelian inheritance does not apply.

Compound heterozygote filtering in recessive diseases

Once a region of interest has been identified and the haplotypic status of the sequenced affected and unaffected individuals ascertained, it may be possible to further interrogate variants under certain genetic models. One such model is the recessive compound heterozygote model where affected individuals have inherited one defective allele from each parent; one located on a maternally derived haplotype, and the other defective allele located on a paternally inherited haplotype. Thus, any two causal alleles need to be in 'trans', that is, on different haplotypes within the same gene. Genes that only contain rare or novel variants that are all in 'cis' may then be filtered out. This can be assessed in two ways.

The first method is contingent on the availability of parental genotypes and allows the phasing of each variant using the trio approach where one checks that each of the two variants being interrogated is from one parent precisely. This can be done using a simple SNP by SNP approach or can be performed using inference of inherited haplotypes to infer where recombinations between grandparental haplotypes have taken place. This requires a hidden Markov Model that allows for linkage disequilibrium and has been implemented using a population-based approach by Delaneau et al. (2013) for HTS data and in a family-based approach in MERLIN (Abecasis and Wigginton 2005), for genotyping data processed with LINKDATAGEN. These two methods can be used even in the absence of parental data. The Delaneau et al. (2013) method uses a property of HTS data: each read and read-pair fragment is a mini haplotype and these mini-haplotypes can be strung together (assembled), similar to de novo assembly, to inform whether variants are in the desired 'trans' or incorrect 'cis' state. This method is useful when parental information is absent, or only partially available, and thus uninformative for haplotype of origin based on the SNP by SNP phasing in trios method already described. To implement the method one can use the ReadBackedPhasing tool from the GATK toolkit (DePristo et al. 2011), HapCUT (Bansal and Bafna 2008) or HapCompass (Aguiar and Istrail 2012). The likelihood of homozygous parents is small for rare variants but is likely to also be more important when the individuals are from populations underrepresented in variant databases, where more of these types of variants will be observed.

More unusual sources of genetic variation

Standard HTS pipelines call single-nucleotide variants (SNVs) and small indels only (those that can be detected with a simple HTS analysis pipeline). The identification of moderate-sized indels require additional analysis of the HTS data with specialized algorithms such as Pindel (Ye et al. 2009). Analysis of larger CNVs also require nonstandard pipeline analysis, since they depend either solely or at least partially on depth of coverage (Alkan et al. 2011). The detection of structural variation other than CNVs, for example translocations, is also possible (Chen et al. 2009). All of these variants can be heritable and causal for genetic disorders. The identification of IBD regions or linkage regions can indicate where standard pipelines have probably failed or need to be enhanced by looking for these alternative, less common, sources of disease causing variants. Many of these detection methods are still under development or prone to false positives and negatives, thus will benefit from fine tuning such as that afforded by being able to focus on specific genomic regions.

Caveats of family-based filtering

While rare, it is possible that the assumed genetic model, under which the linkage or IBD analysis has been performed, may not be correct. This would lead to an incorrect filter. This can occur and is of particular concern for consanguineous pedigrees where it is usual to infer an autozygous recessive model. However, the inbreeding may be a coincidence, with the true mutation being most likely an incompletely penetrant dominant mutation or two compound heterozygous recessive mutations. The deeper the inbreeding loop, the less likely it is that an affected individual will still share a segment of an ancestor that is homozygous by descent, as indicated by an increasing LOD score for such regions in affected individuals, as parents are more distantly related. Extensive inbreeding may also make mapping results difficult to interpret, especially in the presence of locus heterogeneity (Markus et al. 2012).

The trio design for HTS projects is an extremely attractive approach for detecting hypothesized de novo variants in a cohort of sporadic cases where DNA from both parents is available. Without additional evidence for inbreeding in the population of origin, which suggests the possibilities of an autozygously inherited variant, the de novo genetic model is the most likely genetic model for

singleton affected individuals and provides a powerful filter for variants. The trio design for de novo analysis is another family-based study design but it does not make use of linkage information, instead using the family data to eliminate inherited variants. Again, however, the success of this approach relies heavily on the hypothesized inheritance model being correct. In Taft et al. (2013), the initially hypothesized de novo genetic model was proven incorrect, with the discovery of two causal variants constituting a compound recessive model. However, if a de novo model is true, a linkage-based strategy would remove the causal variant from consideration as it constitutes a Mendelian error. This is the danger of using sporadics or even small families: occasionally the proposed genetic model may be incorrect. Thus, we propose that if de novo discovery fails both an autozygous mapping strategy and a compound heterozygote filtering model are applied. In addition, a linkage analysis can be performed with unaffected siblings. One unaffected sibling in a linkage analysis with a recessive model will remove ¼ of the genome from consideration, in the absence of inbreeding.

Conclusions

Here, we have shown how IBD sharing analysis can be used in tandem with HTS to ensure data veracity, to identify appropriate genetic models that facilitate genotype filtering and to make inferences about which individuals are likely to share the ancestral susceptibility haplotype according to the best-fitting genetic model at each locus.

HTS is increasingly being applied to high morbidity, rare, Mendelian disorders where only sporadic cases are observed due to individuals being unlikely, or unable, to reproduce. These types of cohorts usually lack familial information for linkage or IBD studies available, unless the affected individuals are the product of a consanguineous union, in which case a powerful linkage analysis is possible. Other studies can only avail themselves to small, underpowered pedigrees for linkage, however, there is no longer a requirement for a priori statistically significant linkage, since WES and WGS are genome-wide variant identification methods, with no additional costs involved to examine further regions in the genome. A filter based on using multiple regions identified by either linkage or IBD analysis can still be a powerful method for the reduction of the number of candidate variants.

Many of the approaches we have outlined will become even more powerful as WGS supersedes WES and other targeted sequencing approaches. WGS is known to have less bias overall than WES and thus may identify causal variants that have been missed with WES (Bamshad et al. 2011). WGS will enable the genotype calling for many of the HapMap SNPs that can be further supplemented with 1000 Genome SNPs and SNVs leading towards the ultimate SNP chip. This will also lead to an increased ability to uniformly sample SNPs derived from HTS data, further improving linkage and IBD approaches.

HTS is not fail-safe. True causal variants may fail to be detected due to insufficient coverage or erroneous data (Koboldt et al. 2010). Current practice now sees coverage assessments for known genes post-WES and targeted capture screens of known disease causing genes (Carvill et al. 2013). Different laboratories use different criteria by which to ascribe success in a gene hunting effort. Hence, it is very difficult to know the failure rate of current studies with varying reports of success and even more difficult to determine false positive rates. It is important to delineate between the discoveries of novel genes versus novel variants in known genes. The latter solves a family but does not constitute a new basic research finding of high impact, yet it is an important clinical diagnostic success. Researchers searching for novel genes for disorders in cohorts of patients that have already undergone unsuccessful targeted sequencing for known genes have a much lower chance of success in finding a causal variant but are more likely to make a high impact research finding once the genetic cause of disease is identified. Thus, IBD and linkage approaches are likely to play a more crucial role in novel gene discovery, however, even targeted approaches will benefit from QC methodology that allows inference on relatedness and the detection of sample swaps and contamination.

The forthcoming improvements in HTS will lead to increased read length. These will improve our capacity to infer CNVs, microsatellites, other structural variations and facilitate de novo assembly and in silico haplotyping, which we have described above to also aid variant filtering. Detection of these types of variants in HTS data will also benefit from the use of family-based information. PennCNV (Wang et al. 2007) utilizes trio and quartet pedigree information to improve CNV calls for Illumina and Affymetrix SNP chip data and is currently being extended to HTS. Similarly, FamSeq (Peng et al. 2013) and PolyMutt (Li et al. 2012) improve variant calling by making use of familial information.

With the decreasing costs of HTS larger cohorts of affected individuals will be sequenced and some of these individuals will be related, either cryptically or by a known pedigree. Despite the undoubted success of HTS studies, the incorporation of familial information has the potential to enhance our efforts in this fast-moving field.

Acknowledgments This work was supported by the Australian Government National Health and Medical Research Council Program Grant (490037 to M.B.) and the Independent Research Institute Infrastructure Support Scheme and the Victorian State Government Operational Infrastructure Program (to all authors); The Australian

Research Council (FT100100764 to M.B. and an Australian Postgraduate Award to R.T.); and the Pratt Foundation (to K.R.S). The authors would like to thank Dr Thomas Scerri for useful discussions.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. doi:10.1038/nature09534

Abecasis GR, Wigginton JE (2005) Handling marker–marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754–767. doi:10.1086/497345

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101. doi:10.1038/ng786

Aguiar D, Istrail S (2012) HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. J Comput Biol 19:577–590. doi:10.1089/cmb.2012.0084

Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet Epidemiol 33:266–274. doi:10.1002/gepi.20378

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376. doi:10.1038/nrg2958

Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131:1541–1554. doi:10.1007/s00439-012-1213-z

Azmanov DN, Chamova T, Tankard R, Gelev V, Bynevelt M, Florez L, Tzoneva D, Zlatareva D, Guergueltcheva V, Bahlo M, Tournev I, Kalaydjieva L (2013) Challenges of diagnostic exome sequencing in an inbred founder population. Mol Genet Genomic Med 1:71–76. doi:10.1002/mgg3.7

Baek ST, Gibbs EM, Gleeson JG, Mathern GW (2013) Hemimegalencephaly, a paradigm for somatic postzygotic neurodevelopmental disorders. Curr Opin Neurol 26:122–127. doi:10.1097/WCO.0b013e32835ef373

Bahlo M, Bromhead CJ (2009) Generating linkage mapping files from Affymetrix SNP chip data. Bioinformatics 25:1961–1962. doi:10.1093/bioinformatics/btp313

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755. doi:10.1038/nrg3031

Bansal V, Bafna V (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24:i153–i159. doi:10.1093/bioinformatics/btn298

Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3:65ra4. doi:10.1126/scitranslmed.3001756

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, Kaymakcalan H, Barak T, Bakircioglu M, Yasuno K, Ho W, Sanders S, Zhu Y, Yilmaz S, Dincer A, Johnson MH, Bronen RA, Kocer N, Per H, Mane S, Pamir MN, Yalcinkaya C, Kumandas S, Topcu M, Ozmen M, Sestan N, Lifton RP, State MW, Gunel M (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467:207–210. doi:10.1038/nature09327

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14:681–691. doi:10.1038/nrg3555

Browning SR, Browning BL (2012) Identity by descent between distant relatives: detection and applications. Annu Rev Genet 46:617–633. doi:10.1146/annurev-genet-110711-155534

Carvill GL, Heavin SB, Yendle SC, McMahon JM, O'Roak BJ, Cook J, Khan A, Dorschner MO, Weaver M, Calvert S, Malone S, Wallace G, Stanley T, Bye AM, Bleasel A, Howell KB, Kivity S, Mackay MT, Rodriguez-Casero V, Webster R, Korczyn A, Afawi Z, Zelnick N, Lerman-Sagie T, Lev D, Moller RS, Gill D, Andrade DM, Freeman JL, Sadleir LG, Shendure J, Berkovic SF, Scheffer IE, Mefford HC (2013) Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nat Genet 45:825–830. doi:10.1038/ng.2646

Chandrananda D, Thorne NP, Ganesamoorthy D, Bruno DL, Benjamini Y, Speed TP, Slater HR, Bahlo M (2014) Investigating and correcting plasma DNA sequencing coverage bias to enhance aneuploidy discovery. PLoS One 9:e86993. doi:10.1371/journal.pone.0086993

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677–681. doi:10.1038/nmeth.1363

Corbett MA, Bahlo M, Jolly L, Afawi Z, Gardner AE, Oliver KL, Tan S, Coffey A, Mulley JC, Dibbens LM, Simri W, Shalata A, Kivity S, Jackson GD, Berkovic SF, Gecz J (2010) A focal epilepsy and intellectual disability syndrome is due to a mutation in TBC1D24. Am J Hum Genet 87:371–375. doi:10.1016/j.ajhg.2010.08.001

Corbett MA, Schwake M, Bahlo M, Dibbens LM, Lin M, Gandolfo LC, Vears DF, O'Sullivan JD, Robertson T, Bayly MA, Gardner AE, Vlaar AM, Korenke GC, Bloem BR, de Coo IF, Verhagen JM, Lehesjoki AE, Gecz J, Berkovic SF (2011) A mutation in the Golgi Qb-SNARE gene GOSR2 causes progressive myoclonus epilepsy with early ataxia. Am J Hum Genet 88:657–663. doi:10.1016/j.ajhg.2011.04.011

Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J (2013) Haplotype estimation using sequencing reads. Am J Hum Genet 93:687–696. doi:10.1016/j.ajhg.2013.09.002

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. doi:10.1038/ng.806

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36:e105. doi:10.1093/nar/gkn425

Eggers S, Smith KR, Bahlo M, Looijenga LH, Drop SL, Juniarto ZA, Harley VR, Koopman P, Faradz SM, Sinclair AH (2014) Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. Eur J Hum Genet. doi:10.1038/ejhg.2014.130

Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, Heinzen EL, Hitomi Y, Howell KB, Johnson MR, Kuzniecky R, Lowenstein DH, Lu YF, Madou MR, Marson AG, Mefford HC, Esmaeeli

Nieh S, O'Brien TJ, Ottman R, Petrovski S, Poduri A, Ruzzo EK, Scheffer IE, Sherr EH, Yuskaitis CJ, Abou-Khalil B, Alldredge BK, Bautista JF, Berkovic SF, Boro A, Cascino GD, Consalvo D, Crumrine P, Devinsky O, Dlugos D, Epstein MP, Fiol M, Fountain NB, French J, Friedman D, Geller EB, Glauser T, Glynn S, Haut SR, Hayward J, Helmers SL, Joshi S, Kanner A, Kirsch HE, Knowlton RC, Kossoff EH, Kuperman R, Kuzniecky R, Lowenstein DH, McGuire SM, Motika PV, Novotny EJ, Ottman R, Paolicchi JM, Parent JM, Park K, Poduri A, Scheffer IE, Shellhaas RA, Sherr EH, Shih JJ, Singh R, Sirven J, Smith MC, Sullivan J, Lin Thio L, Venkat A, Vining EP, Von Allmen GK, Weisenberg JL, Widdess-Walsh P, Winawer MR (2013) De novo mutations in epileptic encephalopathies. Nature 501:217–221. doi:10.1038/nature12439

Gazal S, Sahbatou M, Babron MC, Genin E, Leutenegger AL (2014) FSuite: exploiting inbreeding in dense SNP chip and exome data. Bioinformatics 30:1940–1941. doi:10.1093/bioinformatics/btu149

Gratten J, Visscher PM, Mowry BJ, Wray NR (2013) Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. Nat Genet 45:234–238. doi:10.1038/ng.2555

Guergueltcheva V, Azmanov DN, Angelicheva D, Smith KR, Chamova T, Florez L, Bynevelt M, Nguyen T, Cherninkova S, Bojinova V, Kaprelyan A, Angelova L, Morar B, Chandler D, Kaneva R, Bahlo M, Tournev I, Kalaydjieva L (2012) Autosomal-recessive congenital cerebellar ataxia is caused by mutations in metabotropic glutamate receptor 1. Am J Hum Genet 91:553–564. doi:10.1016/j.ajhg.2012.07.019

Heron SE, Smith KR, Bahlo M, Nobili L, Kahana E, Licchetta L, Oliver KL, Mazarib A, Afawi Z, Korczyn A, Plazzi G, Petrou S, Berkovic SF, Scheffer IE, Dibbens LM (2012) Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. Nat Genet 44:1188–1190. doi:10.1038/ng.2440

Ivanov IS, Azmanov DN, Ivanova MB, Chamova T, Pacheva IH, Panova MV, Song S, Morar B, Yordanova RV, Galabova FK, Sotkova IG, Linev AJ, Bitchev S, Shearwood A-MJ, Kancheva D, Gabrikova D, Karcagi V, Guergueltcheva V, Geneva IE, Bozhinova V, Stoyanova VK, Kremensky I, Jordanova A, Savov A, Horvath R, Brown MA, Tournev I, Filipovska A, Kalaydjieva L (2014) Founder p.Arg 446* mutation in the PDHX gene explains over half of cases with congenital lactic acidosis in Roma children. Mol Genet Metab. doi:10.1016/j.ymgme.2014.07.017

Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, Wang G, Liang J, Wang Z, Cao D, Carter MT, Chrysler C, Drmic IE, Howe JL, Lau L, Marshall CR, Merico D, Nalpathamkalam T, Thiruvahindrapuram B, Thompson A, Uddin M, Walker S, Luo J, Anagnostou E, Zwaigenbaum L, Ring RH, Wang J, Lajonchere C, Wang J, Shih A, Szatmari P, Yang H, Dawson G, Li Y, Scherer SW (2013) Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. Am J Hum Genet 93:249–263. doi:10.1016/j.ajhg.2013.06.012

Koboldt DC, Ding L, Mardis ER, Wilson RK (2010) Challenges of sequencing human genomes. Brief Bioinform 11:484–498. doi:10.1093/bib/bbq016

Koboldt DC, Larson DE, Sullivan LS, Bowne SJ, Steinberg KM, Churchill JD, Buhr AC, Nutter N, Pierce EA, Blanton SH, Weinstock GM, Wilson RK, Daiger SP (2014) Exome-based mapping and variant prioritization for inherited Mendelian disorders. Am J Hum Genet 94:373–384. doi:10.1016/j.ajhg.2014.01.016

Koenekoop RK, Wang H, Majewski J, Wang X, Lopez I, Ren H, Chen Y, Li Y, Fishman GA, Genead M, Schwartzentruber J, Solanki N, Traboulsi EI, Cheng J, Logan CV, McKibbin M, Hayward BE, Parry DA, Johnson CA, Nageeb M, Finding of Rare Disease Genes Canada C, Poulter JA, Mohamed MD, Jafri H, Rashid Y, Taylor GR, Keser V, Mardon G, Xu H, Inglehearn CF, Fu Q, Toomes C, Chen R (2012) Mutations in NMNAT1 cause Leber congenital amaurosis and identify a new disease pathway for retinal degeneration. Nat Genet 44:1035–1039. doi:10.1038/ng.2356

Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Kohler S, Jager M, Grunhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet 42:827–829. doi:10.1038/ng.653

Lee JH, Huynh M, Silhavy JL, Kim S, Dixon-Salazar T, Heiberg A, Scott E, Bafna V, Hill KJ, Collazo A, Funari V, Russ C, Gabriel SB, Mathern GW, Gleeson JG (2012) De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet 44:941–945. doi:10.1038/ng.2329

Leutenegger AL, Labalme A, Genin E, Toutain A, Steichen E, Clerget-Darpoux F, Edery P (2006) Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. Am J Hum Genet 79:62–66. doi:10.1086/504640

Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, Cucca F, Kang HM, Abecasis GR (2012) A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet 8:e1002944. doi:10.1371/journal.pgen.1002944

Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, Neale BM, Kirby A, Ruderfer DM, Fromer M, Lek M, Liu L, Flannick J, Ripke S, Nagaswamy U, Muzny D, Reid JG, Hawes A, Newsham I, Wu Y, Lewis L, Dinh H, Gross S, Wang LS, Lin CF, Valladares O, Gabriel SB, dePristo M, Altshuler DM, Purcell SM, Project NES, State MW, Boerwinkle E, Buxbaum JD, Cook EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Daly MJ (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 77:235–242. doi:10.1016/j.neuron.2012.12.029

Markus B, Narkis G, Landau D, Birk RZ, Cohen I, Birk OS (2012) Autosomal recessive lethal congenital contractural syndrome type 4 (LCCS4) caused by a mutation in MYBPC1. Hum Mutat 33:1435–1438. doi:10.1002/humu.22122

Marschall T, Hajirasouliha I, Schonhuth A (2013) MATE-CLEVER: mendelian-inheritance-aware discovery and genotyping of mid-size and long indels. Bioinformatics 29:3143–3150. doi:10.1093/bioinformatics/btt556

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J (2010a) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42:790–793. doi:10.1038/ng.646

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010b) Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42:30–35. doi:10.1038/ng.499

O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266. doi:10.1086/301904

Peng G, Fan Y, Palculict TB, Shen P, Ruteshouser EC, Chi AK, Davis RW, Huff V, Scharfe C, Wang W (2013) Rare variant detection using family-based sequencing analysis. Proc Natl Acad Sci USA 110:3985–3990. doi:10.1073/pnas.1222158110

Piton A, Redin C, Mandel JL (2013) XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. Am J Hum Genet 93:368–383. doi:10.1016/j.ajhg.2013.06.013

Poduri A, Evrony GD, Cai X, Elhosary PC, Beroukhim R, Lehtinen MK, Hills LB, Heinzen EL, Hill A, Hill RS, Barry BJ, Bourgeois BF, Riviello JJ, Barkovich AJ, Black PM, Ligon KL, Walsh CA (2012) Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 74:41–48. doi:10.1016/j.neuron.2012.03.010

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. doi:10.1086/519795

Reversade B, Escande-Beillard N, Dimopoulou A, Fischer B, Chng SC, Li Y, Shboul M, Tham PY, Kayserili H, Al-Gazali L, Shahwan M, Brancati F, Lee H, O'Connor BD, Schmidt-von Kegler M, Merriman B, Nelson SF, Masri A, Alkazaleh F, Guerra D, Ferrari P, Nanda A, Rajab A, Markie D, Gray M, Nelson J, Grix A, Sommer A, Savarirayan R, Janecke AR, Steichen E, Sillence D, Hausser I, Budde B, Nurnberg G, Nurnberg P, Seemann P, Kunkel D, Zambruno G, Dallapiccola B, Schuelke M, Robertson S, Hamamy H, Wollnik B, Van Maldergem L, Mundlos S, Kornak U (2009) Mutations in PYCR1 cause cutis laxa with progeroid features. Nat Genet 41:1016–1021. doi:10.1038/ng.413

Riviere JB, van Bon BW, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, Gijsen S, Sullivan CT, Christian SL, Abdul-Rahman OA, Atkin JF, Chassaing N, Drouin-Garraud V, Fry AE, Fryns JP, Gripp KW, Kempers M, Kleefstra T, Mancini GM, Nowaczyk MJ, van Ravenswaaij-Arts CM, Roscioli T, Marble M, Rosenfeld JA, Siu VM, de Vries BB, Shendure J, Verloes A, Veltman JA, Brunner HG, Ross ME, Pilz DT, Dobyns WB (2012) De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser–Winter syndrome. Nat Genet 44(440–4):S1–S2. doi:10.1038/ng.1091

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639. doi:10.1126/science.1186802

Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, Bamshad M, de Condor BJ, Schweiger MR, Robinson PN (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 27:829–836. doi:10.1093/bioinformatics/btr022

Santoni FA, Makrythanasis P, Nikolaev S, Guipponi M, Robyr D, Bottani A, Antonarakis SE (2014) Simultaneous identification and prioritization of variants in familial, de novo, and somatic genetic disorders with VariantMaster. Genome Res 24:349–355. doi:10.1101/gr.163832.113

Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M (2011) Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol 12:R85. doi:10.1186/gb-2011-12-9-r85

Smith KR, Damiano J, Franceschetti S, Carpenter S, Canafoglia L, Morbin M, Rossi G, Pareyson D, Mole SE, Staropoli JF, Sims KB, Lewis J, Lin WL, Dickson DW, Dahl HH, Bahlo M, Berkovic SF (2012) Strikingly different clinicopathological phenotypes determined by progranulin-mutation dosage. Am J Hum Genet 90:1102–1107. doi:10.1016/j.ajhg.2012.04.021

Smith KR, Dahl HH, Canafoglia L, Andermann E, Damiano J, Morbin M, Bruni AC, Giaccone G, Cossette P, Saftig P, Grotzinger J, Schwake M, Andermann F, Staropoli JF, Sims KB, Mole SE, Franceschetti S, Alexander NA, Cooper JD, Chapman HA, Carpenter S, Berkovic SF, Bahlo M (2013) Cathepsin F mutations cause Type B Kufs disease, an adult-onset neuronal ceroid lipofuscinosis. Hum Mol Genet 22:1417–1423. doi:10.1093/hmg/dds558

Sun L, Wilder K, McPeek MS (2002) Enhanced pedigree error detection. Hum Hered 54:99–110 (67666)

Taft RJ, Vanderver A, Leventer RJ, Damiani SA, Simons C, Grimmond SM, Miller D, Schmidt J, Lockhart PJ, Pope K, Ru K, Crawford J, Rosser T, de Coo IF, Juneja M, Verma IC, Prabhakar P, Blaser S, Raiman J, Pouwels PJ, Bevova MR, Abbink TE, van der Knaap MS, Wolf NI (2013) Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. Am J Hum Genet 92:774–780. doi:10.1016/j.ajhg.2013.04.006

Thiele H, Nurnberg P (2005) HaploPainter: a tool for drawing pedigrees with complex haplotypes. Bioinformatics 21:1730–1732. doi:10.1093/bioinformatics/bth488

Tsoi H, Yu AC, Chen ZS, Ng NK, Chan AY, Yuen LY, Abrigo JM, Tsang SY, Tsui SK, Tong TM, Lo IF, Lam ST, Mok VC, Wong LK, Ngo JC, Lau KF, Chan TF, Chan HY (2014) A novel missense mutation in CCDC88C activates the JNK pathway and causes a dominant form of spinocerebellar ataxia. J Med Genet. doi:10.1136/jmedgenet-2014-102333

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665–1674. doi:10.1101/gr.6861907

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871. doi:10.1093/bioinformatics/btp394