REVIEW PAPER

Revisiting Mendelian disorders through exome sequencing

Chee-Seng Ku · Nasheen Naidoo · Yudi Pawitan

Abstract Over the past several years, more focus has been placed on dissecting the genetic basis of complex diseases and traits through genome-wide association studies. In contrast, Mendelian disorders have received little attention mainly due to the lack of newer and more powerful methods to study these disorders. Linkage studies have previously been the main tool to elucidate the genetics of Mendelian disorders; however, extremely rare disorders or sporadic cases caused by de novo variants are not amendable to this study design. Exome sequencing has now become technically feasible and more cost-effective due to the recent advances in high-throughput sequence capture methods and next-generation sequencing technologies which have offered new opportunities for Mendelian disorder research. Exome sequencing has been swiftly applied to the discovery of new causal variants and candidate genes for a number of Mendelian disorders such as Kabuki syndrome, Miller syndrome and Fowler syndrome. In addition, de novo variants were also identified for sporadic cases, which would have not been possible without exome sequencing. Although exome sequencing has been proven to be a promising approach to study Mendelian disorders, several shortcomings of this method must be noted, such as the inability to capture regulatory or evolutionary conserved sequences in non-coding regions and the incomplete capturing of all exons.

Introduction

Over the past two decades, much progress has been made in identifying the causal variants or mutations and candidate genes for Mendelian (single gene or monogenic) disorders through mainly traditional linkage studies (Botstein and Risch 2003). The terms 'variant' and 'mutation' have been used interchangeably throughout the literature; however, 'variant' will be used consistently throughout this article. Mendelian or monogenic disorders encompass 'classical' disorders such as Freeman–Sheldon syndrome (Ng et al. 2009), Fowler syndrome (Lalonde et al. 2010) and the monogenic form of complex diseases such as autosomal-dominant amyotrophic lateral sclerosis (Johnson et al. 2010b) and hypercholesterolemia (Rios et al. 2010). Currently, causal variants for approximately 3,000 Mendelian disorders have been identified (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim).

Genome-wide linkage studies followed by positional cloning have been very successful in identifying causal variants for Mendelian disorders because of the perfect segregation pattern of the causal variant with the disorder according to Mendelian inheritance patterns (e.g. autosomal dominant, autosomal recessive and X-linked). This perfect segregation pattern is due to complete or almost-complete penetrance of the causal variant. In genome-wide linkage studies no prior hypothesis is needed as evenly distributed genetic markers, for example several hundred microsatellites or several thousand single polymorphisms

C.-S. Ku (✉) · N. Naidoo
Department of Epidemiology and Public Health,
Centre for Molecular Epidemiology,
Yong Loo Lin School of Medicine,
National University of Singapore,
Singapore, Singapore
e-mail: g0700040@nus.edu.sg

Y. Pawitan (✉)
Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden
e-mail: yudi.pawitan@ki.se

(SNPs) are sufficient to cover the whole genome. There are only a limited number of recombination events within a family or pedigree. The genetic markers will reveal genomic regions which are co-segregated in affected individuals. This could then be followed up by positional cloning to identify the causal variants and candidate genes within the genomic regions, which can be up to tens of centimorgans (cM). On the contrary, candidate-gene based linkage studies require a prior hypothesis and are not designed to reveal novel genomic regions for Mendelian disorders (Botstein and Risch 2003).

Classical linkage studies are the main tool for elucidating the genetics of Mendelian disorders; however, not all of these disorders are amendable to this study design. Homozygosity-mapping, on the other hand, is a more powerful and effective approach to study recessive disorders in consanguineous families (Harville et al. 2010; Pang et al. 2010; Iseri et al. 2010; Collin et al. 2010). For those disorders that are not amendable to these two conventional approaches, their causal variants remain elusive. These disorders include (a) 'extremely rare' Mendelian disorders where only a small number of cases are available, (b) unrelated cases from different families and (c) sporadic cases due to de novo variants. For some Mendelian disorders, cases can occur sporadically by a de novo or new variant arising during meiosis and which is undetected in the parents (Table 1). We use the term 'extremely rare' to distinguish those Mendelian disorders which cannot be investigated by linkage studies due to their low incidence in the population from 'rare' disorders where an adequate sample size can still be collected for linkage studies. For extremely rare disorders, usually only several affected siblings in one family or several unrelated cases from different families are available for investigation. However, exome (the collection of all exons in the human genome) sequencing now offers new opportunities to study extremely rare disorders and sporadic cases (Table 1) as well as complex diseases (Li et al. 2010b).

Two recent review papers on exome sequencing of Mendelian disorders focused on variant filtering strategies (Ng et al. 2010c) and novel genomic techniques (Kuhlenbäumer et al. 2011). However, we review this area in a broader context and focus on several topics which have not been comprehensively discussed previously. In this paper, we start by discussing the need for exome sequencing of Mendelian disorders and the technological developments leading to the feasibility of this approach. We also recall the importance and value of interrogating the genetics of Mendelian disorders which tend to have been given less emphasis in the era of genome-wide association studies (GWAS) and then further elaborate on the application of exome sequencing in elucidating the genetics of Mendelian disorders and the recent advances achieved in the field. The

pros and cons of currently employed variant filtering strategies will also be discussed. We also examine the advantages and challenges of exome sequencing in identifying causal variants for Mendelian disorders. Finally, as most of the known causal variants were found in exons (protein coding regions), we share our views on whether whole-genome sequencing is needed for Mendelian disorder research.

Why exome sequencing is needed

The linkage study design is unsuitable for extremely rare Mendelian disorders because of the difficulty in collection of an adequate number of affected individuals (of multi-generational pedigree) and families for a statistically powerful study. This approach is also not applicable for sporadic cases, for example Kabuki syndrome, an extremely rare autosomal-dominant Mendelian disorder with an estimated incidence of 1 in 32,000, where the majority of reported cases are sporadic (Ng et al. 2010a). As a result, the causal variant and candidate gene for Kabuki syndrome have remained unknown until recently. A total of 33 different causal variants in *MLL2* were identified by Ng et al. (2010a) in 35 of 53 individuals affected with Kabuki syndrome. Additionally, in 12 of these individuals whose parental samples were available, their variants in *MLL2* were found to have occurred de novo. Only ten of these individuals were investigated in the discovery study using exome sequencing to identify the causal variants in *MLL2,* and the exons of this gene were then screened in an additional 43 cases using Sanger sequencing (Ng et al. 2010a).

Similarly, most of the cases of Schinzel–Giedion syndrome have occurred sporadically suggesting that heterozygous de novo variants may have caused the disorder. This has now been further supported by identifying de novo causal variants in *SETBP1* in four individuals affected with this disorder through exome sequencing (Hoischen et al. 2010). These de novo causal variants would not have been otherwise identified without exome sequencing. In contrast, although none of the causal variants in *DHODH* appeared to have occurred de novo for Miller Syndrome, it is still an extremely rare disorder (Ng et al. 2010b). Therefore, these disorders are intractable to the linkage study design. Collectively, these studies have demonstrated the advantages of exome sequencing over the linkage study design in situations where a small number of unrelated samples or sporadic cases are available. Up to ten samples have been previously interrogated by exome sequencing in discovery studies (Table 1).

Furthermore, the linkage study design is also not robust enough for Mendelian disorders with genetic heterogeneity

Table 1 Summaries of exome and whole-genome sequencing studies of Mendelian disorders

Study, Mendelian disorder and sample	Variant filtering methodology and analysis strategy	Major results
(A) Exome sequencing of unrelated individuals		
Ng et al. 2009 Freeman–Sheldon syndrome Four unrelated individuals	• First investigated how many genes had one or more non-synonymous cSNPs, splice site disruptions or coding indels in one or several exomes • Then applied filters to remove presumably common variants, removing dbSNP-catalogued variants from consideration reduced the number of candidates considerably • The eight HapMap exomes provided a filter nearly equivalent to dbSNP • Combining the two catalogs had a synergistic effect, such that the candidate list could be narrowed to a single gene	• *MYH3* is the only gene where (1) at least one non-synonymous cSNP, splice-site disruption or coding indel is observed in all four individuals; (2) the mutations are neither in dbSNP, nor in the eight HapMap exomes
Ng et al. 2010a Kabuki syndrome Ten unrelated individuals	• Focused primarily on nonsynonymous variants, splice acceptor and donor site mutations and coding indels • Defined variants as previously unidentified if they were absent from all datasets used for comparison, including dbSNP129, the 1000 Genomes Project, exome data from 16 individuals and 10 exomes sequenced as part of the Environmental Genome Project (EGP) • To allow for a modest degree of genetic heterogeneity and/or missing data, conducted a less stringent analysis by looking for candidate genes shared among subsets of affected individuals • Searched for subsets of x out of 10 exomes having ≥ 1 previously unidentified variant in the same gene, with $x = 1$ to $x = 10$ • Genotypic and/or phenotypic stratification would facilitate the prioritization of candidate genes identified by subset analysis • Assigned a categorical rank to each individual with Kabuki syndrome based on a subjective assessment of the presence of, or similarity to, the canonical facial characteristics of Kabuki syndrome and the presence of developmental delay and/or major birth defects	• Identified a nonsense substitution or frameshift indel in *MLL2* in seven of the ten individuals with Kabuki syndrome • Further analyzed the three cases in which did not initially find a loss-of-function variant in *MLL2* • Sanger sequencing did identify frameshift indels in two of these three cases • Screened all 54 exons of *MLL2* in 43 additional cases by Sanger sequencing • Previously unidentified nonsynonymous, nonsense or frameshift mutations in *MLL2* were found in 26 of these 43 cases
Hoischen et al. 2010 Schinzel-Giedion syndrome Four unrelated individuals	• On average, 21,800 genetic variants were identified per individual, including 5,351 nonsynonymous changes • A comparison with the NCBI dbSNP build130 as well as with recently released SNP data from other groups and in-house SNP data showed that >95% of all variants investigated here were previously reported SNPs • Focused on the 12 genes for which all four individuals studied carried variants and found that only two genes showed variants at different genomic positions • One of these two candidate genes, *CTBP2*, was excluded from further analysis because it contained numerous variants found during different in-house exome sequencing experiments	• The second candidate was *SETBP1* • Validation of all four variants in this gene by Sanger sequencing confirmed that these variants were indeed present in a heterozygous state in all four affected individuals • Tested the DNA of the parents of the affected individuals, which showed that all mutations occurred de novo • Using Sanger sequencing, identified *SETBP1* mutations in eight out of nine additional individuals • For six of the eight follow-up cases, parental DNA was available, and the mutations present in the affected individuals were again shown to have occurred de novo

Table 1 continued

Study, Mendelian disorder and sample	Variant filtering methodology and analysis strategy	Major results
Gilissen et al. 2010 Sensenbrenner syndrome Two unrelated individuals with remarkably similar phenotypes of Sensenbrenner syndrome	• On average, 12,736 genetic variants were identified per patient in the coding regions or the canonical dinucleotide of the splice sites, including 5,657 nonsynonymous changes • A prioritization scheme was applied to identify the pathogenic mutation in each patient separately, similar to a recent study by Hoischen et al. (2010)	• Identified *WDR35* as the most likely candidate disease gene in both patients • In patient 1, identified a canonical splice-site mutation 2 bp upstream of exon 2 and a missense mutation in exon 17 • In patient 2, a deletion of a C nucleotide in exon 25 predicts a frameshift and a premature stop and a substitution in exon 23 • Performed mutation analysis in six additional patients • Did not find any causative mutations in *WDR35* in these patients
Pierce et al. 2010 Perrault syndrome One individual	• DNA variants were filtered against dbSNP131 and Phase 3 of the 1000 Genomes project, then classified by predicted function to include all missense, nonsense, frameshift, or splice-site alleles • Given that dbSNP includes both disease-associated and benign alleles, included in the analysis known SNPs identified by dbSNP as clinically associated • Surveyed all genes harboring the 207 rare nonsense, missense, frameshift, or splice variants for any genes with two such alleles in this patient	• Only one gene, *HSD17B4* on chromosome 5q23.1 • The patient's DNA included mutations at: c.650A > G (p.Y217C) in exon 9 c.1704T > A (p.Y568X) in exon 20 • Both the proband and her sister were compound heterozygotes for both mutations
Lalonde et al. 2010 Fowler syndrome Two affected fetuses from two distinct French Canadian and French non-consanguineous families (labeled as F1 and F3)	• Focused on nonsynonymous (NS) variants, splice acceptor, and donor site mutations (SS), and short coding insertions or deletions (indels; I) • A novel variant was defined as one that did not exist in the databases used for comparison, namely, dbSNP and 1000 Genomes • Searched for genes with a unique homozygous mutation in both samples, genes with distinct homozygous mutations in either sample, and genes for which both individuals are compound heterozygotes • Searched unsuccessfully for a homozygous mutation present in both probands • No evidence was found for a single gene harboring different homozygous mutations in the two samples • Proceeded to search for potentially damaging compound heterozygous variants	• 83 and 92 candidate genes were identified in F1 and F3, respectively • Taking the intersection of these two lists revealed only one candidate gene, *FLVCR2* • F1 is a compound heterozygote for two novel mutations c.997C > T, exon 4 c.1341 + 2T > C, 5′ splice site downstream of exon 7 • F3 is also a compound heterozygote c.1192C > G, exon 6 c.329_334del6 deletion, exon 1
Choi et al. 2009 Congenital chloride diarrhea One patient referred with a suspected diagnosis of Bartter syndrome	• In this subject, 462 Mb of the genome were homozygous by descent, including 5.3 Mb of the exome, comprising 2,459 genes • Anticipated finding a homozygous disease-causing mutation within a segment that is homozygous by consanguineous descent • Sought novel and rare mutations that alter the encoded protein within the 5.3 M base pairs and 2,495 genes in these segments • Identified 2,405 homozygous variations, including 1,493 homozygous cSNVs (coding single nucleotide variants) • These cSNVs included 668 non-synonymous substitutions (29 of which were novel) • Ranked the novel missense variants by conservation scores to identify the most likely functional mutations using the phyloP conservation score	• Ten of the homozygous variants were at highly conserved positions • Among the novel homozygous mutations at extremely conserved positions, one strongly stood out: a single-base substitution that introduced a missense variant in *SLC26A3* • This mutation introduces a D652N substitution at a position that is completely conserved among all invertebrates and vertebrates studied

Table 1 continued

Study, Mendelian disorder and sample	Variant filtering methodology and analysis strategy	Major results
(B) Exome sequencing of unrelated individuals and individuals in families		
Ng et al. 2010b Miller syndrome Two siblings (kindred one) and two additional unrelated affected individuals (kindreds two and three), totaling four affected individuals in three independent kindreds	• Focused only on nonsynonymous (NS) variants, splice acceptor and donor site mutations (SS), and short coding insertions or deletions (indels; I) • A new variant was defined as one that did not exist in the datasets used for comparison, namely dbSNP129, exome data from eight HapMap individuals, and both groups combined • In recessive model, each sibling was required to have at least two new NS/SS/I variants in the same gene • Both siblings were predicted to share the causal variant for Miller syndrome, so considered candidate genes shared between them • Under recessive model, the number of candidate genes was reduced to 9 • Compared the candidate genes from both siblings in kindred 1 to those in two unrelated individuals with Miller syndrome (kindreds 2 and 3)	• Identified *DHODH* as the sole candidate gene for Miller syndrome under recessive model • Screened three additional unrelated kindreds (three simplex cases) and an affected sibling in kindred 2 by directed Sanger sequencing • All four individuals were found to be compound heterozygotes for missense mutations in *DHODH* • Collectively, 11 different mutations in 6 kindreds with Miller syndrome were identified in *DHODH* • None of the mutations appeared to have arisen de novo • Ten of these mutations were missense mutations, and one was a 1-bp indel that is predicted to cause a frameshift resulting in a termination codon seven amino acids downstream
Johnson et al. 2010a Brown-Vialetto-van Laere syndrome Two siblings and two unrelated patients	• All identified variants were filtered against dbSNP and 1000 Genomes, removing all previously reported variants • Restricting the variants to 937 nonsynonymous alterations detected in at least one of the four samples • Selected only those changes that were homozygous, or when a variant was heterozygous, an additional heterozygous variant was present in the same gene, in the same sample • This resulted in a sample-specific list of 37 genes with 39 homozygous nonsynonymous mutations and 48 genes with 140 compound-heterozygous nonsynonymous changes. • Compared these gene lists across samples and identified genes present in the affected siblings and at least one of the two other patients	• Resulted in a list of eight genes, three of which contained a large number of variants, suggesting that these variants are likely an artifact of the methodology • The remaining five genes were primary candidates for Brown-Vialetto-van Laere syndrome • c.639C > G (p.Y213X) and c.211G > A (p.E71K) in *C20orf54* were present in both affected siblings as heterozygous changes
(C) Exome sequencing of individuals in families with/without linkage analysis		
Wang et al. 2010 Autosomal-dominant spinocerebellar ataxias Four affected individuals in one four-generation Chinese family (Family CS)	• Focused only on non-synonymous (NS) variants, splice acceptor and donor site mutations (SS), and short, frame-shift coding insertions or deletions (indel) • A total of 23,067 NS/SS/Indel variants were detected in ∼5,700 candidate genes in at least one of the affected individuals sequenced in Family CS • Compared against dbSNP129, eight previously exome-sequenced HapMap samples (HapMap 8), and the SNP release of the 1000 Genome Project (20100208 release), and removed the shared SNPs • After filtering, the candidate gene pool was reduced ∼20-fold • Used PolyPhen to assess the non-synonymous variants for a likely functional impact • Because spinocerebellar ataxias in this family is an autosomal-dominant inherited disease, all affected individuals should share the same causal variant	• Comparison of exome data from four samples to identify the shared variant were sufficient to identify *TGM6* as the sole candidate gene, which had a variant c.1550T > G transition (L517W) in exon 10 • All nine patients were heterozygous for this mutation and no clinically unaffected family members carried this variant, except for two clinically unknown (presymptomatic) individuals who were also heterozygous • The mutation in *TGM6* completely cosegregated with the spinocerebellar ataxias phenotype within this family • The missense variant identified in Family CS via exome sequencing was also located in the associated region identified using linkage analysis

Table 1 continued

Study, Mendelian disorder and sample	Variant filtering methodology and analysis strategy	Major results
Johnson et al. 2010b Autosomal-dominant amyotrophic lateral sclerosis Two affected individuals in a four-generation Italian family (ITALS#1) in which four individuals had been diagnosed with amyotrophic lateral sclerosis	• SNPs identified in the 1000 Genomes project or in dbSNP (Build 131) were removed • Excluded variants that were not shared by both patients • Synonymous changes were identified and filtered from the variant list using SIFT software • Sanger sequencing was performed to determine the presence of the remaining variants in the other affected member of the ITALS#1 family • Variants and indels detected in the ITALS#1 family were filtered against exome data generated for 200 neurologically normal control subjects • After filtering, there were 75 heterozygous coding variants and 13 heterozygous coding indels • Sanger sequencing in an additional affected member reduced this list to 24 variants and nine indels	• Linkage analysis identified 18 regions across the autosomes with a LOD score greater than zero • Only four variants were within the 18 genomic regions • Each of these variants segregated with disease within the family, and therefore were plausible candidates for the genetic defect responsible for disease • Within this list, identified a c.961G > A nucleotide change that resulted in a p.R191Q amino acid change in the VCP gene
Musunuru et al. 2010 Familial combined hypolipidemia Two siblings	• For each participant, approximately 270 new single-nucleotide variants were found, of which approximately 150 were missense mutations and approximately 6 were nonsense (premature termination) mutations • Hypothesized that the phenotype of combined hypolipidemia was inherited in an autosomal recessive fashion • Sought genes harboring novel variants in both alleles of both participants • Eliminated variants that were not shared between the two probands • Eliminated any variant sites that were present in the dbSNP (build 130), in the genomes sequenced by the 1000 Genomes Project, or in the 60 control exomes	• Only one gene harbored novel variants in both alleles in both siblings: ANGPTL3 • ANGPTL3 harbored two nonsense variants: a single-nucleotide variant (GAA > TAA) and a double-nucleotide variant (TCC > TGA) • Both mutations are located in the first exon of ANGPTL3 • Performed linkage analyses, using LDL cholesterol and HDL cholesterol as quantitative traits • The highest lod scores for both traits occurred on chromosome 1p33-31.1, the region harboring ANGPTL3
Krawitz et al. 2010 Hyperphosphatasia mental retardation syndrome Three siblings of non-consanguineous parents	• Called variants were filtered to exclude variants not found in all affected persons as well as common variants identified in the dbSNP130 or HapMap databases, which left 14 candidate genes on multiple chromosomes • Developed a statistical model to infer regions that are identical by descent (IBD) from the exome sequences of only the affected children of a family in which an autosomal recessive disorder segregates • Reduced the number of candidate genes with mutations present in all three siblings from 14 to 2	• The two mutations, c.859G > A in SLC9A1 and c.1022C > A in PIGV, were located within a 13-Mb homozygous block that was part of a larger 35-Mb IBD = 2 block • Further homozygous and compound heterozygous mutations were detected in PIGV in individuals from three additional families • All of these missense mutations affect evolutionarily highly conserved residues of PIGV

(D) Exome sequencing coupled with homozygosity mapping

Bolze et al. 2010 A condition affecting at least four members of an extended consanguineous kindred of Pakistani origin with biological features of autoimmune lymphoproliferative syndrome (ALPS), no clinical features of ALPS, a complex infectious phenotype with both viral and bacterial infections, and congenital cardiovascular malformations. This clinical syndrome has never been described One affected individual	• Hypothesized that this syndrome would segregate as an autosomal-recessive trait • Two regions (>1 Mb) were homozygous in the three affected patients: a 8-Mb region on chromosome 11 and a 9-Mb region on chromosome 18 • Identified 23,146 variations from exome sequencing • Of these, 67 variants were found in the chromosome 11 candidate region and 14 variants were found in the chromosome 18 candidate region • Comparisons with the NCBI dbSNP build 129, the 1000 Genomes Project database, and in-house database (composed of 70 exomes) identified only one nonsynonymous variant that had not been reported before and mapped to the chromosome 11 region	• The variant identified was a missense homozygous c.315T > G in exon 2 of FADD that changes cysteine at amino acid position 105 to tryptophan, p.C105W • Cysteine at amino acid position 105 is highly conserved throughout evolution • This variant segregated with the disease status in all family members examined

Table 1 continued

Study, Mendelian disorder and sample	Variant filtering methodology and analysis strategy	Major results
Walsh et al. 2010 Nonsyndromic hearing loss *DFNB82* One individual from a consanguineous Palestinian family (family CG)	• Applied whole exome sequencing to identify the gene responsible for nonsyndromic hearing loss *DFNB82* in a Palestinian family • The *DFNB82* region spans 3.1 Mb on chromosome 1p13.1 • Focused on exonic and flanking intronic variants within the linkage region • In the *DFNB82* region, 80 variants, all single nucleotide substitutions, passed quality thresholds • Of the 80 substitutions, 73 appeared in dbSNP130 and thus were not considered further and 7 were previously unidentified • None of these seven variants were present in the sequenced exomes of three unrelated Palestinian individuals	• The seven variants included five synonymous substitutions, one missense substitution (MYBPHL p.I65T), and one nonsense mutation (GPSM2 p.R127X) • The frequency of MYBPHL p.I65T was approximately 1.0% both in controls and in unrelated persons with hearing loss from the Palestinian population, suggesting that it is a benign polymorphism • GPSM2 p.R127X was not present in controls or in unrelated cases • GPSM2 p.R127X mutation segregated as expected with deafness in family CG, based on PCR and Sanger sequencing of all family members
Caliskan et al. 2011 Autosomal recessive nonsyndromic mental retardation Both parents in a consanguineous family (Family G with five adult children with nonsyndromic mental retardation)	• Performed whole-genome genotyping of the affected individuals and identified a shared homozygous segment that spans >2 Mb on chromosome 19p13 • Included both obligate carrier parents in an exome sequencing study • Identified 71 variants (33 missense, 37 synonymous, 1 non-protein coding) in the critical region in the parents • Assumed that the nonsyndromic mental retardation in this family was due to a fully penetrant autosomal recessive mutation • Considered 18 of the 71 variants that were heterozygous in both parents as candidates for the nonsyndromic mental retardation mutation • Seventeen of these 18 variants were present in the dbSNP (v131) and were, therefore, further excluded	• The remaining novel mutation occurred in exon eight of the *TECR* gene, causing a Pro to Leu substitution at amino acid 182 in the TECR protein • Using Sanger sequencing, confirmed that the mutation in the *TECR* gene segregated with nonsyndromic mental retardation in Family G • The non-reference allele was homozygous in the five affected siblings, heterozygous in the parents and seven unaffected siblings, and absent in one unaffected sibling
(E) Whole-genome sequencing		
Sobreira et al. 2010 Metachondromatosis Whole-genome sequencing of a single patient from Pedigree 1 (a metachondromatosis family)	• Performed linkage analysis using multiple individuals from Pedigree 1 and found six regions with positive LOD scores • Considered these six regions as showing suggestive evidence for the presence of a causal variant	• In the 12q23 region (105 RefSeq genes), identified one frameshifting indel, an 11-bp deletion in exon 4 of *PTPNII* (c.514_524del11)
Lupski et al. 2010 Charcot–Marie–Tooth disease Whole-genome sequencing of the proband. The study family consisted of four affected siblings, four unaffected siblings, and an unaffected mother and father	• A total of 2,255,102 of the SNPs were in extragenic regions and 1,165,204 SNPs were within gene regions, including introns, promoters, 3′ and 5′ untranslated regions, and splice sites • Of the intragenic SNPs, 9,069 were nonredundant SNPs predicted to result in nonsynonymous codon changes, and 121 of the 9,069 were nonsense mutations • Of the more than 3.4 million SNPs, 561,719 were novel • Examined the putative mutations in 40 genes known to cause or be linked to neuropathic or related conditions • This exercise led to closer examination of 3,148 putative SNPs, including 54 coding SNPs	• Of these 54, 2 were at the *SH3TC2* locus—1 missense mutation and 1 nonsense mutation • Mutations in this locus have previously been found to be associated with Charcot–Marie–Tooth type 4C disease • The R954X nonsense mutation has previously been implicated in Charcot–Marie–Tooth disease; the missense mutation (A > G, occurring on chromosome 5 at nucleotide 148,402,474 and corresponding to the amino acid mutation Tyr169His [Y169H]) is novel

Table 1 continued

Study, Mendelian disorder and sample	Variant filtering methodology and analysis strategy	Major results
Rios et al. 2010 Severe hypercholesterolemia Whole-genome sequencing of an infant with severe hypercholesterolemia of unknown etiology	• A total of 3,797,207 deviations from the reference sequence • Restricted to nonsynonymous variants and splice junction variants that were not present in the public repository of sequence variants (dbSNP) or in 16 exomes and 5 genomes from 21 individuals who did not have hypercholesterolemia • Of the 9,726 nonsynonymous single-nucleotide variants identified in the proband, 699 novel nonsynonymous SNPs that were predicted to change an amino acid or a consensus splice site in 604 genes • Two or more novel nonsynonymous variants were present in 42 genes • Queried whether there were any nonsense mutations among the 23 single-copy genes that were found to contain two or more novel, nonsynonymous variants	• A single gene, *ABCG5*, contained two different nonsense mutations: Q16X and R446X • Sanger sequencing confirmed both mutations in the proband and that her mother and father were heterozygotes for the Q16X and R446X mutations, respectively • Both these mutations are incompatible with the expression of a functional protein.

The exome and whole genome sequencing studies of Mendelian disorders summarized in this table are not a complete list. We have included as many studies as possible published until January 2011 to provide the most up-to-date list. However, we are unable to include all the studies due to the length constraint of the article. Therefore, we apologize to those whose work we were unable to summarize in this table. This table has briefly summarized the variant filtering methodology and analysis strategy and major results with the aim to provide a glimpse of the published studies. However, readers are encouraged to refer to the full papers for more detailed information. The information in this table was derived from the original papers with minimal editing and modification of the sentences. In this paper, the terminology 'variant' was used consistently throughout the article. However, in this table, as we only summarized the information from published studies, we retained the various terminologies used in each study. This table is divided into five categories according to the different study designs

(i.e. the causal variants are present in different genes) and phenotypic heterogeneity (i.e. diverse clinical or phenotypic manifestations leading to uncertainty in diagnosis of the disorder or ambiguity in phenotype). Similarly, these problems are well depicted in Kabuki syndrome which is likely a genetically heterogeneous disorder because not all the affected individuals have causal variants in the single candidate gene (*MLL2*) (Ng et al. 2010a; Paulussen et al. 2010). Nevertheless, causal variants in different genes have not yet been found to further support its genetic heterogeneity. Exome sequencing is more robust for disorders with a presumably genetic heterogeneity background. Kabuki syndrome is also characterized by phenotypic heterogeneity. To account for this, investigators have performed additional phenotypic stratification and ranking steps (Ng et al. 2010a). Initially, this study failed to identify a compelling candidate gene harboring causal variants in all the ten investigated individuals. However, by accounting for the genetic and phenotypic heterogeneity, the investigators successfully identified causal variants in *MLL2* in a subset of individuals. This illustrates the additional challenges present in studying disorders with genetic or phenotypic heterogeneity. Other Mendelian disorders summarized in Table 1 also demonstrate varying degrees of genetic or phenotypic heterogeneity.

High-throughput sequence capture and sequencing technologies

The high-throughput sequence capture methods are able to isolate the collection of exons in a more efficient and cost-effective way than traditional PCR-based methods. Without these sequence capture methods, the approximately 180,000 exons in the human genome would require designing an equivalent or larger number of PCR primer sets to isolate and amplify (Ng et al. 2009), and it would therefore be costlier and time consuming to study the exome using PCR-based isolation methods. These high-throughput sequence capture methods are commercially marketed, for example the NimbleGen Sequence Capture technology (http://www.nimblegen.com/) and Agilent SureSelect Target Enrichment technology (http://www.home.agilent.com). These sequence capture methods allow researchers to target custom genomic regions of interest in the human genome for up to tens of megabases and also enable enrichment of the exome in a single experiment. This development coupled with the high-throughput sequencing data produced by next-generation sequencing (NGS) technologies ensures an adequate depth of sequencing coverage to accurately detect the variants in the exome or targeted regions (Mamanova et al. 2010; Turner et al. 2010; Koboldt et al. 2010; Metzker 2010; Shendure and Ji 2008).

The total size of the human exome is approximately 30 Mb which comprises approximately 1% of the entire human genome. Therefore, exome sequencing requires many-fold lesser amounts of sequencing data to achieve the desired depth of sequencing coverage for variants detection compared to whole-genome sequencing. As a result, exome sequencing has emerged as a more popular approach to study Mendelian disorders (Table 1). Although many recent studies are labeled as 'exome sequencing', the sequence capture methods employed are unable to completely isolate all the exons experimentally, i.e. a fraction of exons will be missed. Furthermore, the probes in sequence capture methods are designed based on the sequence information from gene annotation databases such as the consensus coding sequence (CCDS) database and RefSeq database; therefore, unknown or yet-to-annotate exons cannot be captured. Regions that are poorly mapped with short sequence reads due to paralogous sequences elsewhere in the genome have to be excluded as well (Ng et al. 2009). As such, the exome capture is not complete. The incomplete capture of the exome can create additional problems in identifying the causal variants and candidate genes for Mendelian disorders.

The exome sequencing studies have focused primarily on the approximately '30Mb sequences' encompassing exons and splice sites using commercially available sequence capture methods. As such, these sequence capture methods have limited or no coverage of other important regulatory sequences such as promoters, enhancers, microRNAs and other annotated regulatory elements and evolutionary conserved non-coding sequences. For example, the Agilent SureSelect Human All Exon Kit covers 38 Mb of sequences corresponding to the exons and flanking intronic regions of 23,739 genes in the CCDS database (September 2009 release) and also encompasses 700 microRNAs from the Sanger v13 database and 300 non-coding RNAs (Walsh et al. 2010). Although this 'all exon kit' has expanded the coverage beyond the exome, the coverage of regulatory sequences is not complete and also raises a further question of why evolutionary conserved non-coding sequences are not included.

Some researchers may consider the limited coverage of important regulatory and evolutionary conserved sequences as one limitation of exome sequencing; hence, there is now an increasing demand to include these regions in future exome sequencing studies. Undoubtedly, it is advantageous to include as many annotated regulatory and evolutionary conserved sequences as possible where the causal variants might be found, but this will then add to the cost of sequence capture methods and sequencing as more sequences will need to be isolated and sequenced. Recently, the introduction of the Illumina TruSeq Exome Enrichment Kit has doubled the size of targeted regions to 62 Mb with more than 90% coverage of the exons or genes in the latest version of the CCDS and RefSeq database (http://www.illumina.com/products/truseq_exome_enrichment_kit.ilmn). However, in this scenario where there is a continuous demand to increase the coverage or size of targeted regions beyond the exome, whole-genome sequencing is probably a more viable option that has been adopted in some studies (Sobreira et al. 2010; Lupski et al. 2010; Rios et al. 2010). Ultimately, it will be more efficient and cost-effective to subtract the sequence reads in 'unwanted' regions by bioinformatic analysis after whole-genome sequencing than including the 'wanted' regions during the sequence capturing stages if the coverage of targeted regions continues to expand.

All the NGS technologies have higher base calling error rates than Sanger sequencing, although this can be remedied to some extent by increasing the depth of sequencing coverage to ensure minimal errors (Koboldt et al. 2010). An adequate depth of sequencing coverage is also critical for identifying heterozygotes such as de novo variants or heterozygous variants causing dominant Mendelian disorders or compound heterozygotes causing recessive disorders. Gilissen et al. (2010) used the Agilent SureSelect human exome kit in combination with ABI SOLiD sequencing to generate 3.6 and 3.4 gigabases of mappable sequence data for two patients with Sensenbrenner syndrome and achieved an average sequencing coverage of $67\times$ and $59\times$ for the exomes (Gilissen et al. 2010), while Wang et al. (2010) obtained an average coverage of $65\times$ for four exomes affected with autosomal-dominant spinocerebellar ataxias and reported that approximately 97% of the targeted bases were covered sufficiently to pass their thresholds for variant calling (Wang et al. 2010). Thus, this depth of sequencing coverage was deemed sufficient for accurate detection of variants. This is critical for subsequent downstream analysis because base calling errors could mistakenly be thought of as rare variants. These artifacts will make the searching for causal variants and candidate genes more difficult if not properly accounted for.

The barcoding method allows multiplexing of up to tens of samples to be sequenced per instrument run and offers a cost advantage. The levels of multiplexing depend on the size of the targeted regions to be sequenced and the depth of sequencing coverage to be achieved. Given the continuous increase in the throughput of sequencing data generated by NGS technologies, where several hundred gigabases of data are generated per instrument run, barcoding of the samples will be more cost-effective and avoid over-sequencing of samples. Over-sequencing of samples would result in diminishing returns in accuracy gains in variants detection (Craig et al. 2008; Szelinger et al. 2011).

To conclude, technological developments have made exome sequencing more practical and affordable: from

several samples needed for Mendelian disorders to hundreds of samples for complex diseases (Li et al. 2010b). These technological developments have been one of the main driving forces of the exome sequencing era with more than 20 studies being published on the subject in 2010 (Table 1) (Bilgüvar et al. 2010; Roach et al. 2010; Byun et al. 2010; Haack et al.2010; Bonnefond et al. 2010; Worthey et al. 2010). In addition, these technologies have also accelerated efforts in sequencing the previously identified linkage regions (Brkanac et al. 2009; Nikopoulos et al. 2010; Volpi et al. 2010; Rehman et al. 2010). Brkanac et al. (2009) applied sequence capture and NGS methods to sequence all the genes in a previously identified linkage region, chromosome 7q22-q32, for autosomal-dominant sensory/motor neuropathy with ataxia and identified a nonsynonymous variant in *IFRD1* causing the disorder. Without these technologies, interrogating the linkage regions of several centimorgans using PCR and Sanger sequencing methods would be a daunting task. Therefore, targeted sequence capture followed by NGS should be performed to investigate the established linkage regions from previous studies.

The rise of complex disease research

Over the past 5 years, the genetics research community has focused studies mainly on dissecting the genetic basis of complex (non-Mendelian, polygenic or multifactorial) diseases and traits. Prior to this, studies of complex phenotypes have met with limited success using candidate-gene association and linkage study designs (Hirschhorn et al. 2002; Hirschhorn 2005). Although linkage studies have identified causal variants for thousands of Mendelian disorders, this approach is ineffective and unsuitable for complex diseases caused by complex interactions of multiple genetic and environmental factors. Nevertheless, significant progress has been achieved since 2005 through GWAS (Altshuler et al. 2008; Hindorff et al. 2009; Ku et al. 2010). Presently, more than 4,000 SNPs have been reported to be associated with various human complex diseases and traits (A Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/26525384).

However, due to the indirect study design of GWAS being reliant on linkage disequilibrium, the causal variants remain elusive in most of the GWAS-detected loci. It is also more difficult to identify the causal variants for complex diseases resulting from multiple genetic variants of low penetrance. This is in contrast to Mendelian disorders which are caused by variants with complete (or nearly complete) penetrance showing a strong genotype–phenotype relationship. Therefore, despite the success of GWAS in unraveling thousands of statistically robust SNP associations, the causal variants and candidate genes for most complex diseases have not been convincingly identified (Altshuler et al. 2008; Hindorff et al. 2009; Ku et al. 2010).

In comparison with complex diseases, relatively slower progress has been made in identifying causal variants for Mendelian disorders during the peak period of GWAS research (2006–2009) until the first proof-of-principle study demonstrated the feasibility of exome sequencing to identify a known candidate gene for Freeman–Sheldon syndrome (Ng et al. 2009). Several reasons have been cited for Mendelian disorders receiving little attention in recent years. First, most of the Mendelian disorders with their causal variants and candidate genes which can be investigated by linkage studies have already been identified (Amberger et al. 2009). Other disorders are too rare to be investigated by linkage studies. Second, a powerful method to study extremely rare disorders or cases caused by de novo variants is not previously available. Although NGS technologies have been available since 2005, exome sequencing was not technically feasible and efficient until the advent of high-throughput sequence capture methods to isolate the exome (Mamanova et al. 2010; Turner et al. 2010). These problems are related to the Mendelian disorders themselves; however, other factors are more in favor towards complex diseases research and will be discussed further.

Third, it is due to the increased enthusiasm of researchers in pursuing complex diseases research after the notable success in the GWAS of age-related macular degeneration (Klein et al. 2005). The completion of the International HapMap Project, the advent of high-resolution genotyping microarrays, the collection of large sample sizes, and the development of powerful statistical analysis methods have led to the rapid increase in publications of GWAS since 2005 (Seng and Seng 2008). Furthermore, delineating the genetics of complex diseases such as metabolic, cardiovascular, autoimmune and chronic inflammatory and infectious diseases is believed to be more important from the public health perspective as these diseases affect a much larger fraction of the population than Mendelian disorders (McCarthy 2010; Musunuru and Kathiresan 2010; Baranzini 2009). Collectively, these factors have gradually attracted more attention towards complex diseases.

Why study Mendelian disorders

Research into the previously unexplained Mendelian disorders (i.e. where causal variants have not been identified) should become a priority now and in the near future for several reasons. First, Mendelian disorders as a collective make up approximately 7,000 known or suspected disorders and contribute significantly to the disease burden in society, even though they have been labeled as rare or extremely rare disorders compared with the more common

complex diseases (Ropers 2007; Ropers 2010; Antonarakis and Beckmann 2006; Antonarakis et al. 2010). We further discuss the importance of studying Mendelian disorders from three aspects: (a) revealing genes for complex diseases and traits, (b) providing new biological insights and (c) identifying drug targets.

Studying Mendelian disorders can reveal genes and biological pathways that are associated with the development of complex diseases. This was illustrated in the identification of SNPs in *WFS1* and *TCF2* associated with the polygenic form of type-2 diabetes (Sandhu et al. 2007; Winckler et al. 2007). The *WFS1* gene was prioritized as a candidate to be interrogated in candidate gene association studies because the rare variants in *WFS1* cause a monogenic form of diabetes (Wolfram syndrome). Thus, *WFS1* becomes a biologically plausible gene for polygenic type-2 diabetes (Sandhu et al. 2007). Similarly, rare variants in the *TCF2* gene cause maturity-onset diabetes of the young (MODY) (Winckler et al. 2007).

In cases where defective *MC4R* was the leading cause of monogenic severe childhood-onset obesity, it was also found that common SNPs near *MC4R* were associated with fat mass, weight and risk of obesity and other metabolic-related traits (Loos et al. 2008; Chambers et al. 2008). Numerous GWAS-identified common SNPs which are associated with triglycerides, high-density lipoprotein (HDL) cholesterol and low-density lipoprotein (LDL) cholesterol levels were also found in the candidate genes causing the monogenic form of these lipid metabolism disorders (Kathiresan et al. 2008; Hegele 2009). The convergence of genes identified for Mendelian and polygenic diseases were also seen in other diseases such as Parkinson's disease (Gasser 2009; Lesage and Brice 2009). Recently, the identification of *TECR* for non-syndromic mental retardation through exome sequencing also suggests that this gene should be further studied in patients with neurological and psychiatric diseases such as schizophrenia and autism. This study has implicated a potential candidate gene to be investigated in other diseases which may then provide important information for revealing common molecular pathways underlying the development of these diseases (Caliskan et al. 2011).

The discovery of causal variants and candidate genes responsible for Mendelian disorders will also help in understanding their biological function. For example, the discovery of causal variants in *DHODH* (which encodes the enzyme dihydroorotate dehydrogenase) for Miller syndrome has provided new insights into the role of pyrimidine metabolism in craniofacial and limb development (Ng et al. 2010b). The discovery of causal variants in *TMG6* for spinocerebellar ataxias also provides further evidence to suggest its involvement in the pathogenesis of neurodegenerative diseases (Wang et al. 2010).

Much of the molecular biological research on amyotrophic lateral sclerosis is based on the discovery of causal variants in genes such as *SOD1*, *TDP-43* and *FUS* that are responsible for the familial or monogenic form of this disease. Given its value in providing new biological insights, investigating the genetic basis of the monogenic form of complex diseases has received recent attention. A causal variant in a previously unreported gene has been identified for familial amyotrophic lateral sclerosis (Johnson et al. 2010b), and a nonsynonymous variant in *VCP* was identified through exome sequencing of two affected individuals in a family. This discovery provided new insights into the investigation and understanding of the molecular biology and pathogenesis of amyotrophic lateral sclerosis. The finding of causal variants in *VCP* for familial amyotrophic lateral sclerosis implicates defects in the ubiquitination/protein degradation pathway in motor neuron degeneration.

The potential discovery of new drug targets through studying the genetics of Mendelian disorders should also be emphasized. The discovery of drugs targeting *PPARγ* and *KCNJ11* as a treatment for type-2 diabetes strongly supports this potential. The drugs used to lower cholesterol levels by inhibiting the enzyme HMG-CoA reductase (i.e. statins) were also discovered through studying familial hypercholesterolaemia (Brinkman et al. 2006). Finally, it also contributes to our understanding of human physiology, e.g. studying the Mendelian forms of hypertension have improved our knowledge of blood pressure and volume regulation (Luft 2003).

Currently, the revisiting of Mendelian disorders is mainly due to the 'attraction' of the exome sequencing approach and the 'distraction' of the disappointing GWAS results that explain only a small fraction of the heritability of complex diseases and traits (Manolio et al. 2009). Nevertheless, studying complex diseases should not be abandoned, as GWAS have also revealed new biological insights, such as unraveling the autophagy and interleukin-23 receptor pathways for Crohn's disease (Mathew 2008; Cho 2008). A balance between Mendelian disorders and complex diseases research is needed, as research in one cannot be substituted by the other. The knowledge gained from studying Mendelian disorders and complex diseases will eventually complement each other and synergistically enhance our understanding of genotype-phenotype relationships.

A balance for Mendelian disorders and complex diseases

Over the past few years, enormous resources have been invested in research on complex diseases and traits where hundreds of GWAS projects were funded and many huge

consortia established to tackle the genetics of these phenotypes (Voight et al. 2010; Teslovich et al. 2010). Several international projects such as the International HapMap Project and 1000 Genomes Project were initiated with the aim of providing useful resources for elucidating the genetics of complex diseases (International HapMap 3 Consortium 2010; 1000 Genomes Project Consortium 2010). Biobanks were also established to properly collect and store hundreds of thousands of biological samples for future investigation of complex diseases (Palmer 2007; Nakamura 2007; Fan et al. 2008). Fortunately, the desire and endeavor to study the genetics of complex diseases has also driven unprecedented developments in microarray (genotyping and sequence capture) and sequencing technologies. These developments have eventually enabled exome sequencing or whole-genome sequencing to be applied to Mendelian disorders.

Despite Mendelian disorder research being considered successful, more than half of the approximately 7,000 known or suspected Mendelian disorders identified based on clinical features have not yet been linked to their candidate genes harboring causal variants. New efforts to improve this include a recent initiative by the National Human Genome Research Institute (USA) to establish 'A Center for Mendelian Disorders' whose mission will be to take on the sequencing of Mendelian disorders. This center would be expected to solve the molecular basis of 40–50 disorders per year. In addition, this center will also coordinate the collection and distribution of samples for all remaining unexplained Mendelian disorders, for example by identifying samples within the community and obtaining commitments from the investigators who have samples for distribution to other groups who are able to do exome sequencing. This will facilitate and accelerate the effort to identify causal variants for as many of these disorders as possible (NHGRI Large-Scale Sequencing Program May 2010, http://www.genome.gov/).

Exome sequencing of Mendelian disorders

Sequencing of unrelated individuals

The advent of the exome sequencing approach has immediately overcome the major obstacles in studying extremely rare Mendelian disorders and de novo variants. This proof-of-concept was demonstrated by Ng et al. (2009) in Freeman–Sheldon syndrome. Only four unrelated cases were subjected to exome sequencing and *MYH3* was identified as the single candidate gene harboring at least one nonsynonymous variant, splice-site disruption or coding indel in all cases. The causal variants identified in *MYH3* were previously unidentified, i.e. neither cataloged in dbSNP nor

present in exome sequencing data of eight HapMap samples. Although *MYH3* is a known candidate gene for Freeman–Sheldon syndrome, this study showed the feasibility of applying exome sequencing to identify the candidate gene for a Mendelian disorder despite a small number of unrelated cases (Ng et al. 2009).

As tens of thousands of single nucleotide variants and short indels have been detected in the sequencing of the human exome, multiple robust filtering criteria needed to be applied to discern the causal variants. These filters included first identifying the genes with one or more nonsynonymous variants, splice-site disruptions or coding indels in the exomes of the four individuals with Freeman–Sheldon syndrome (which assumed no genetic heterogeneity among the cases) investigated by Ng et al. (2009) and excluding those common variants as they were less likely to be causative (Table 1) (Ng et al. 2009). These filters have proven effective in identifying the known candidate gene for Freeman–Sheldon syndrome.

Nonetheless, for other Mendelian disorders, the identification of the causal variant or candidate gene is not as straightforward as demonstrated in Freeman–Sheldon syndrome. The exomes of ten unrelated individuals affected with Kabuki syndrome were also sequenced by the same group of researchers (Ng et al. 2010a). However, after applying the same filtering strategies, the study failed to identify a compelling candidate gene whose previously unidentified variants were seen in all the individuals. This result suggests the presence of genetic and phenotypic heterogeneity underlying the disorder. To account for genetic heterogeneity, a less stringent strategy was applied by looking for candidate genes shared among subsets of affected individuals. Additionally, various ranking and stratifying steps were also taken into account for phenotypic heterogeneity. These additional strategies finally led to the identification of causal variants in the *MLL2* gene (Table 1) (Ng et al. 2010a).

Similar to other disorders such as Sensenbrenner syndrome, only two out of eight individuals had causal variants in *WDR35*. These causal variants were only identified in two unrelated cases with a strikingly similar phenotype. No causal variant in the gene was identified in the other six patients presenting with additional clinical phenotypes, and these patients did not show the striking phenotypic similarity as observed between the first two patients in the discovery study (Gilissen et al. 2010). This highlights the complexity of genetic and phenotypic heterogeneity and implies that classifying the phenotypic heterogeneity (by focusing on a very similar phenotype) helps in identifying the causal variants.

Further studies have also identified a number of novel candidate genes harboring causal variants for disorders such as Miller syndrome (Ng et al. 2010b), Fowler
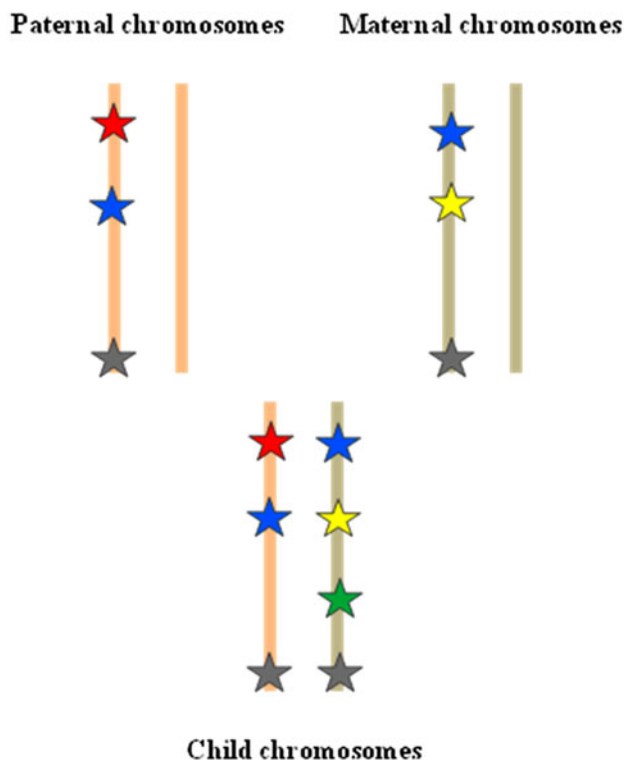
syndrome (Lalonde et al. 2010), Perrault Syndrome (Pierce et al. 2010) and Schinzel-Giedion syndrome (Hoischen et al. 2010) (Table 1). Of particular interest is the candidate gene identified for Fowler syndrome. A compound heterozygote of two variants in *FLVCR2* was identified for each of the two cases, and thus a total of four different variants were identified in this gene (Lalonde et al. 2010). Compound heterozygotes in *HSD17B4* and *DHODH* was also found for a number of individuals with Perrault syndrome (Pierce et al. 2010) and Miller syndrome (Ng et al. 2010b), respectively. Compound heterozygote refers to the presence of two different heterozygous variants occurring in two distinct positions in the homologous chromosomes (i.e. one variant in maternal chromosome and the other variant in paternal chromosome). Therefore, these

Paternal chromosomes   **Maternal chromosomes**

Child chromosomes

Fig. 1 This schematic diagram illustrates the concepts of 'compound heterozygote' and 'de novo variant'. The *red* and *yellow stars* represent two deleterious variants in two different positions in a gene. The *red* variant is passed on to the child from the father and the *yellow* variant from the mother. Each deleterious variant is in a heterozygote state. In the child, both copies of the gene are defective due to the presence of *red* and *yellow* variants. Therefore, even in a heterozygous state, these deleterious variants can result in a recessive disorder. This is different from a homozygote variant (*gray star*) causing a recessive disorder. Therefore, recessive disorders can be caused by (1) two different heterozygous variants (compound heterozygote) and (2) two similar variants (homozygote). The *green star* represents a de novo variant which is absent in paternal and maternal chromosomes

deleterious variants can result in a recessive disorder even in the heterozygote state (Fig. 1). The exome sequencing studies have applied various strategies to identify the causal variants for different disorders, and some studies have integrated exome sequencing data with linkage and homozygosity analysis (Table 1).

Sequencing of family members

In addition to the previously unexplained Mendelian disorders, exome sequencing has also identified novel causal variants and candidate genes for disorders which have been studied previously, for example autosomal-dominant spinocerebellar ataxias. To date, causal variants in 19 genes have been identified for this disorder. Recently, a causal variant in an additional gene (*TGM6*) was revealed through exome sequencing (Wang et al. 2010). However, instead of sequencing unrelated individuals from different families as demonstrated in other studies (Table 1), the investigators performed exome sequencing in four affected individuals in one four-generation Chinese family with autosomal-dominant spinocerebellar ataxias. Although this study also applied almost similar variant filtering strategies as with other studies of unrelated cases, comparison of the exome data among the four cases to find the shared variant was sufficient to identify *TGM6* as the sole candidate gene containing a new nonsynonymous variant in exon 10 of this gene.

This study highlighted the advantage of sequencing multiple affected individuals from one family, because it allowed the investigators to hypothesize that all affected individuals should share the same causal variant, as spinocerebellar ataxia was inherited in an autosomal-dominant pattern in this family. The finding from exome sequencing was also supported by linkage analysis where the causal variant was found in a region revealed by linkage analysis. Spinocerebellar ataxias are also characterized by clinical and genetic heterogeneity which would benefit from exome sequencing. The sequencing of affected individuals in one family would offer further advantage to the study design (Wang et al. 2010) as unrelated cases from different families are likely to have causal variants in different genes. This study highlighted the advantage of exome sequencing in affected family members, as compared with unrelated individuals, in identifying causal variants for clinically and genetically heterogeneous disorders. Other studies have also performed exome sequencing in multiple siblings and identified causal variants and candidate genes for disorders such as autosomal-dominant amyotrophic lateral sclerosis (Johnson et al. 2010b), familial combined hypolipidemia (Musunuru et al. 2010) and hyperphosphatasia mental retardation syndrome (Krawitz et al. 2010).

Integration with homozygosity mapping

Exome sequencing has also been swiftly integrated with homozygosity mapping to accelerate the investigation of recessive disorders in consanguineous families (Walsh et al. 2010; Anastasio et al. 2010; Sirmaci et al. 2010; Bolze et al. 2010). Bolze et al. (2010) have demonstrated the advantages of integrating both approaches in identifying causal variants for a clinical syndrome that has never been described previously (Table 1). Homozygosity mapping was performed in three patients and their parents, which identified two homozygosity regions in chromosome 11 and 18, respectively. In parallel, the exome of one patient was sequenced identifying 23,146 variants; however, only 67 variants and 14 variants were found in the homozygosity region in chromosome 11 and 18, respectively. The availability of homozygosity data has allowed the investigators to substantially narrow down the search space to less than 100 variants from the exome data. The subsequent comparisons with SNP databases identified only one nonsynonymous variant that was previously unreported in the homozygosity region in chromosome 11 and was located in exon 2 of *FADD*. The filtering and identifying of causal variants have been greatly facilitated by integration with homozygosity mapping data (Bolze et al. 2010).

Diagnostic application

Exome sequencing is also a useful tool for diagnostic application. The genetic diagnosis of congenital chloride diarrhea in a patient was made through exome sequencing revealing a homozygous missense variant in *SLC26A3*. The position of this variant is completely conserved from invertebrates to humans (Choi et al. 2009). However, other studies have adopted whole-genome sequencing as a diagnostic application. For example, it was applied to an 11-month-old patient with severe hypercholesterolemia and identified approximately 3.8 million variants where only 9,726 were nonsynonymous variants and of which 699 were new. The defective gene *ABCG5* was identified because it had two nonsense variants (Rios et al. 2010). The diagnostic application was further illustrated by Lupski et al. (2010) through whole-genome sequencing of a proband with Charcot–Marie–Tooth disease. However, this study only focused on those genes known to cause the neuropathic condition. One missense variant and one nonsense variant were detected in *SH3TC2*, and all affected individuals in the family of the proband were found to be compound heterozygotes for these variants (Lupski et al. 2010). Although whole-genome sequencing was done in some studies (Rios et al. 2010; Lupski et al. 2010), exome

sequencing would have been sufficient to identify the causal variants and genes for severe hypercholesterolemia and Charcot–Marie–Tooth disease. The cost of a diagnostic test would be an important factor to consider for clinical utility. Exome sequencing is anticipated to be used increasingly in molecular diagnosis (Bonnefond et al. 2010; Worthey et al. 2010; Montenegro et al. 2011).

Pros and cons of variant filtering strategies

There are two important assumptions underlying the variant filtering strategies of these exome sequencing studies: (a) causal variants for Mendelian disorders would be rare and therefore likely to be previously unidentified in public databases or control sequencing data and (b) synonymous variants would be far less likely to be causative. However, several caveats must be noted.

The filtering of common variants in the exome by comparison with public databases such as the dbSNP, the HapMap Project, the 1000 Genomes Project and other exome sequencing data is of benefit. This has proven effective in removing a substantial number of less likely causal variants (Table 1) as the causal variants for extremely rare Mendelian disorders should be 'very rare'. In addition, de novo variants are also rare, occurring in a heterozygote state for dominant disorders. This simple assumption and filtering strategy offers an advantage to quickly sift through the exome data for promising causal variants. However, the removal of common variants by comparison to the dbSNP has a weakness due to the considerable fraction of false-positive errors in the dbSNP. Currently, more than 17 million SNPs in the human genome have been documented in the dbSNP with a false-positive rate of 15–17% estimated for the database (Day 2010). Therefore, some important variants in the exome may be discarded. This problem is likely to be overcome by a more accurate database upon the completion of the 1000 Genomes Project. However, with the continuous cataloging of rarer variants in the human genome, an 'optimal' cutoff of frequency needs to be imposed to distinguish between what constitutes the 'common variants' that are less likely to be causative compared to 'rare variants' that need to be retained for analysis.

The exome sequencing studies have focused on nonsynonymous and nonsense variants, splice-site variants and frameshift indels (collectively known as deleterious variants) and ignored synonymous variants which are far less likely to be deleterious (Table 1). By discarding synonymous variants, the number of variants is substantially reduced for downstream analysis. However, in the event that some cases are unexplained by deleterious variants, it is not immediately clear whether synonymous variants are causative for the unexplained cases. Nonetheless, it is currently

unclear how best to incorporate the synonymous variants into an analysis with deleterious variants robustly and efficiently to identify candidate genes for Mendelian disorders.

Exome sequencing versus whole-genome sequencing

In this section we discuss the advantages and pitfalls of exome sequencing in comparison to whole-genome sequencing. Since the exome constitutes only approximately 1% of the human genome, it requires a lesser amount of sequencing data to achieve the desired depth of sequencing coverage to accurately detect variants compared with whole-genome sequencing. For example, 138 gigabases of mappable sequence data were generated and achieved an average coverage of $49\times$ in the whole-genome sequencing of a patient with severe hypercholesterolemia (Rios et al. 2010). In contrast, <4 gigabases of mappable sequence data were generated and achieved an average coverage of $59\times$ and $67\times$ in the exome sequencing of two patients with Sensenbrenner syndrome (Gilissen et al. 2010).

Furthermore, most of the known causal variants for Mendelian disorders were found in exons. The exome has been the focus of studies for Mendelian disorders because nonsynonymous variants leading to amino acid changes affect the function of the protein and nonsense variants producing truncated protein are significantly deleterious to cause Mendelian disorders. In addition, small indels in exons can adversely affect the amino acid sequence through frameshift reading of the codons. Variants in splice sites can affect the mRNA stability and alternative splicing. Although whole-genome sequencing was performed in some studies (Rios et al. 2010; Lupski et al. 2010), these analyses still focused on the variants in exons.

However, whole-genome sequencing offers an advantage to study other genetic variants besides deleterious variants in exons. The paired-end sequence reads generated by whole-genome sequencing are useful for the detection of various structural variants or chromosomal rearrangements in the genome which collectively become the second source of genetic abnormalities responsible for Mendelian disorders (Lupski and Stankiewicz 2005; Chen et al. 2010a). Several structural variant detection methods such as paired-end mapping and depth-of-coverage are developed by leveraging on the high-density short sequence read data generated by NGS technologies (Korbel et al. 2007; Yoon et al. 2009; Medvedev et al. 2009). Preparation of several DNA fragment libraries with different sizes coupled with these sequencing-based detection methods have demonstrated to be powerful enough to detect different structural variants of varying sizes. The application of the mate-pair sequencing method to identify copy number

variants was demonstrated in the whole-genome sequencing study of Charcot-Marie-Tooth disease. In parallel, the study also used a comparative genomic hybridization (CGH)-based array and identified a total of 234 copy number variants. However, none of the copy number variants affecting genes was known to be involved in Charcot-Marie-Tooth disease (Lupski et al. 2010). Although high-resolution oligonucleotide CGH or SNP microarrays can be used to supplement exome sequencing, these microarray-based methods are only able to detect copy number changes, whereas inversions, translocations and other more complex chromosomal rearrangements are beyond their detection (Carter 2007). The use of these microarrays will also add to the cost of the exome sequencing study.

To ensure a more thorough interrogation of both deleterious single nucleotide variants in exome and structural rearrangements, the cost of an 'exome sequencing study' will be comprised of spending on sequence capture methods, exome sequencing and CGH or SNP microarrays. This also means that three laboratory experiments are needed and two sets of data (sequencing and microarray data) will be generated. Opting for exome sequencing is mainly driven by the cost advantage. However, given the decreasing cost of whole-genome sequencing, the price gap between the two approaches is becoming smaller. Despite whole-genome sequencing being more costly, it has greater value in that data of the whole-genome are obtained compared with 1% of the genome from exome sequencing. Furthermore, only a few samples are usually studied in exome sequencing for Mendelian disorders, unlike complex diseases that require hundreds to thousands of samples. Thus, the difference in cost between the two sequencing approaches will only be multiplied by several samples.

Exome sequencing studies have, without a doubt, identified causal variants and candidate genes for a number of Mendelian disorders (Table 1); nevertheless, a subset of cases for some of the disorders remain unexplained. There are several reasons for this. The capture of the entire collection of exons in the human genome using the available sequence capture methods is by no means complete; thus, variants in the missing exons cannot be studied. Furthermore, non-coding regions (introns and intergenic regions) are not considered in exome sequencing. It is still unclear whether synonymous variants or variants in non-coding regions or deleterious variants in other genes are responsible for the unexplained cases (Cooper et al. 2010; Chen et al. 2010b). In contrast, whole-genome sequencing studies do not have the problem of 'missing exons' as a result of incomplete capture. Furthermore, the variants in highly evolutionary conserved non-coding regions can be readily explored for unexplained cases (Dermitzakis et al. 2005). Many causal variants identified for Mendelian disorders

were located in protein sequences which are highly conserved throughout evolution (Table 1). This could also hint at the importance of investigating variants in evolutionary conserved non-coding regions where important functional elements were found (Dermitzakis et al. 2005; Alexander et al. 2010).

It is anticipated that the >3 million single nucleotide variants detected in whole-genome sequencing will create additional challenges to identifying causal variants and thus more robust filtering strategies are needed. However, most of the common and less likely causal variants should be removed efficiently with the data from the full completion of the 1000 Genomes Project. Although the 'whole-genome data' are generated, investigators can still focus on and prioritize the variants in the exome for first-tier analysis. The remaining variants can be used in subsequent tiers of analysis. This strategy was also applied in whole-genome cancer sequencing (Ley et al. 2008). If it appears that those variants in evolutionary conserved non-coding regions or regulatory sequences are also causative or acting as modifiers affecting the severity of disorders, then the exome-sequenced samples may need to be resequenced at the whole-genome level. Identifying the variants acting as modifiers will help in better understanding of phenotypic heterogeneity, but this will be challenging (Génin et al. 2008).

Currently, the sequencing data for Mendelian disorders is still rudimentary; it is difficult to be convinced that the variants in the remaining 99% of the genome are not 'important' to these disorders either as causative variants or modifiers (Cooper et al. 2010; Chen et al. 2010b; Dermitzakis et al. 2005). It was previously believed that 99% of the genome consisted of 'junk DNA' because these regions did not encode proteins. The functional importance of the 'junk DNA' was eventually discovered (Castillo-Davis 2005; Alexander et al. 2010). We hope that more knowledge and understanding will be gained through exploration of variants in the whole genome.

Summary and future direction

In summary, exome sequencing has now been applied in multiple situations where (a) several affected siblings in a family, (b) several unrelated cases and (c) sporadic cases are available for analysis where the causal variants for a number of Mendelian disorders have been successfully identified (Table 1). In addition, exome sequencing has also been shown to be more robust to study disorders with genetic and phenotypic heterogeneity. It has also proved viable to study Mendelian disorders if only a single case is available. In addition, de novo causal variants have also been successfully identified for sporadic cases. Exome sequencing has

also been demonstrated as a powerful tool in diagnostic application. Integration with linkage and homozygosity data has greatly facilitated the discovery of causal variants and candidate genes for Mendelian disorders.

The number of causal variants and candidate genes identified for Mendelian disorders is anticipated to grow rapidly through individual researchers and large-scale collaborative efforts. The cost of exome sequencing studies is now more affordable and only a few exomes need to be sequenced. Although exome sequencing studies have provided compelling evidence that the identified variants are causative for Mendelian disorders, mutagenesis and animal model studies will still be needed to lend further support to the causality and to demonstrate the effect that causal variants have on the phenotypic level.

Exome sequencing with sufficient depth of coverage has generated high-quality data for single nucleotide variant detection. However, it is difficult to detect indels with short sequence reads generated by NGS technologies. For example, frameshift indels in two individuals with Kabuki syndrome were undetected by exome sequencing, but were successfully identified using Sanger sequencing (Ng et al. 2010a). Furthermore, exome sequencing is unable to detect structural variants or chromosomal rearrangements which are believed to be important for Mendelian disorders as well. These, together with the problem of incomplete exome capture and the potential reward from interrogating non-coding regions, especially the highly evolutionary conserved regions, have sparked a debate on whether whole-genome sequencing is needed. However, this will likely be a non-issue in the next few years when the cost of whole-genome sequencing becomes cheaper. Similar to other fields such as cancer genome sequencing (Ley et al. 2008; Pleasance et al. 2010; Lee et al. 2010) and studies of human genetic variants (Bentley et al. 2008; Wheeler et al. 2008; Wang et al. 2008), research on Mendelian disorders will also benefit tremendously from de novo genome assembly when it becomes feasible with better assembly algorithms and longer sequence reads generated by third-generation sequencing technologies (Li et al. 2010a; Schadt et al. 2010).

References

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, 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

Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB (2010) Annotating non-coding regions of the genome. Nat Rev Genet 11:559–571

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) Nucleic Acids Res 37:D793–D796

Anastasio N, Ben-Omran T, Teebi A, Ha KC, Lalonde E, Ali R, Almureikhi M, Der Kaloustian VM, Liu J, Rosenblatt DS, Majewski J, Jerome-Majewska LA (2010) Mutations in SCARF2 are responsible for Van Den Ende–Gupta syndrome. Am J Hum Genet 87:553–559

Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet 7:277–282

Antonarakis SE, Chakravarti A, Cohen JC, Hardy J (2010) Mendelian disorders and multifactorial traits: the big divide or one for all? Nat Rev Genet 11:380–384

Baranzini SE (2009) The genetics of autoimmune diseases: a networked perspective. Curr Opin Immunol 21:596–605

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bilgüvar K, Oztürk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoğlu D, Tüysüz B, Cağlayan AO, Gökben S, Kaymakçalan H, Barak T, Bakircioğlu M, Yasuno K, Ho W, Sanders S, Zhu Y, Yilmaz S, Dinçer A, Johnson MH, Bronen RA, Koçer N, Per H, Mane S, Pamir MN, Yalçinkaya C, Kumandaş S, Topçu M, Ozmen M, Sestan N, Lifton RP, State MW, Günel M (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467:207–210

Bolze A, Byun M, McDonald D, Morgan NV, Abhyankar A, Premkumar L, Puel A, Bacon CM, Rieux-Laucat F, Pang K, Britland A, Abel L, Cant A, Maher ER, Riedl SJ, Hambleton S, Casanova JL (2010) Whole-exome-sequencing-based discovery of human FADD deficiency. Am J Hum Genet 87:873–881

Bonnefond A, Durand E, Sand O, De Graeve F, Gallina S, Busiah K, Lobbens S, Simon A, Bellanné-Chantelot C, Létourneau L, Scharfmann R, Delplanque J, Sladek R, Polak M, Vaxillaire M, Froguel P (2010) Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. PLoS One 5:e13630

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33:228–237

Brinkman RR, Dubé MP, Rouleau GA, Orr AC, Samuels ME (2006) Human monogenic disorders—a source of novel drug targets. Nat Rev Genet 7:249–260

Brkanac Z, Spencer D, Shendure J, Robertson PD, Matsushita M, Vu T, Bird TD, Olson MV, Raskind WH (2009) IFRD1 is a candidate gene for SMNA on chromosome 7q22-q23. Am J Hum Genet 84:692–697

Byun M, Abhyankar A, Lelarge V, Plancoulaine S, Palanduz A, Telhan L, Boisson B, Picard C, Dewell S, Zhao C, Jouanguy E, Feske S, Abel L, Casanova JL (2010) Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. J Exp Med 207:2307–2312

Caliskan M, Chong JX, Uricchio L, Anderson R, Chen P, Sougnez C, Garimella K, Gabriel SB, Depristo MA, Shakir K, Matern D, Das S, Waggoner D, Nicolae DL, Ober C (2011) Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. Hum Mol Genet (Epub ahead of print)

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–S21

Castillo-Davis CI (2005) The evolution of noncoding DNA: how much junk, how much func? Trends Genet 21:533–536

Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40:716–718

Chen JM, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP (2010a) Genomic rearrangements in inherited disease and cancer. Semin Cancer Biol 20:222–233

Chen JM, Férec C, Cooper DN (2010b) Revealing the human mutome. Clin Genet 78:310–320

Cho JH (2008) The genetics and immunopathogenesis of inflammatory bowel disease. Nat Rev Immunol 8:458–466

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096–19101

Collin RW, Safieh C, Littink KW, Shalev SA, Garzozi HJ, Rizel L, Abbasi AH, Cremers FP, den Hollander AI, Klevering BJ, Ben-Yosef T (2010) Mutations in C2ORF71 cause autosomal-recessive retinitis pigmentosa. Am J Hum Genet 86:783–788

Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum Mutat 31:631–655

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5:887–893

Day IN (2010) dbSNP in the detail and copy number complexities. Hum Mutat 31:2–4

Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences—an unexpected feature of mammalian genomes. Nat Rev Genet 6:151–157

Fan CT, Lin JC, Lee CH (2008) Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical island. Pharmacogenomics 9:235–246

Gasser T (2009) Mendelian forms of Parkinson's disease. Biochim Biophys Acta 1792:587–596

Génin E, Feingold J, Clerget-Darpoux F (2008) Identifying modifier genes of monogenic disease: strategies and difficulties. Hum Genet 124:357–368

Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, Roepman R, Knoers NV, Veltman JA, Brunner HG (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87:418–423

Haack TB, Danhauser K, Haberberger B, Hoser J, Strecker V, Boehm D, Uziel G, Lamantea E, Invernizzi F, Poulton J, Rolinski B, Iuso A, Biskup S, Schmidt T, Mewes HW, Wittig I, Meitinger T, Zeviani M, Prokisch H (2010) Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. Nat Genet 42:1131–1134

Harville HM, Held S, Diaz-Font A, Davis EE, Diplas BH, Lewis RA, Borochowitz ZU, Zhou W, Chaki M, MacDonald J, Kayserili H, Beales PL, Katsanis N, Otto E, Hildebrandt F (2010) Identification of 11 novel mutations in eight BBS genes by high-resolution homozygosity mapping. J Med Genet 47:262–267

Hegele RA (2009) Plasma lipoproteins: genetic influences and clinical implications. Nat Rev Genet 10:109–121

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hirschhorn JN (2005) Genetic approaches to studying common diseases and complex traits. Pediatr Res 57:74R–77R

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61

Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, Devriendt

K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, Veltman JA (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42:483–485

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

Iseri SU, Wyatt AW, Nürnberg G, Kluck C, Nürnberg P, Holder GE, Blair E, Salt A, Ragge NK (2010) Use of genome-wide SNP homozygosity mapping in small pedigrees to identify new mutations in VSX2 causing recessive microphthalmia and a semidominant inner retinal dystrophy. Hum Genet 128:51–60

Johnson JO, Gibbs JR, Van Maldergem L, Houlden H, Singleton AB (2010a) Exome sequencing in Brown-Vialetto-van Laere syndrome. Am J Hum Genet 87:567–569

Johnson JO, Mandrioli J, Benatar M, Abramzon Y, Van Deerlin VM, Trojanowski JQ, Gibbs JR, Brunetti M, Gronka S, Wuu J, Ding J, McCluskey L, Martinez-Lage M, Falcone D, Hernandez DG, Arepalli S, Chong S, Schymick JC, Rothstein J, Landi F, Wang YD, Calvo A, Mora G, Sabatelli M, Monsurrò MR, Battistini S, Salvi F, Spataro R, Sola P, Borghero G; ITALSGEN Consortium, Galassi G, Scholz SW, Taylor JP, Restagno G, Chiò A, Traynor BJ (2010b) Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron 68:857–864

Kathiresan S, Musunuru K, Orho-Melander M (2008) Defining the spectrum of alleles that contribute to blood lipid concentrations in humans. Curr Opin Lipidol 19:122–127

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

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

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Krawitz PM, Schweiger MR, Rödelsperger C, Marcelis C, Kölsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Köhler S, Jäger M, Grünhagen 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

Ku CS, Loy EY, Pawitan Y, Chia KS (2010) The pursuit of genome-wide association studies: where are we now? J Hum Genet 55:195–206

Kuhlenbäumer G, Hullmann J, Appenzeller S (2011) Novel genomic techniques open new avenues in the analysis of monogenic disorders. Hum Mutat 32:144–151

Lalonde E, Albrecht S, Ha KC, Jacob K, Bolduc N, Polychronakos C, Dechelotte P, Majewski J, Jabado N (2010) Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum Mutat 31:918–923

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z (2010) The mutation spectrum

revealed by paired genome sequences from a lung cancer patient. Nature 465:473–477

Lesage S, Brice A (2009) Parkinson's disease: from monogenic forms to genetic susceptibility factors. Hum Mol Genet 18:R48–R59

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66–72

Li Y, Hu Y, Bolund L, Wang J (2010a) State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics 4:271–277

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, Grarup N, Guo Y, Hellman I, Jin X, Li Q, Liu J, Liu X, Sparsø T, Tang M, Wu H, Wu R, Yu C, Zheng H, Astrup A, Bolund L, Holmkvist J, Jørgensen T, Kristiansen K, Schmitz O, Schwartz TW, Zhang X, Li R, Yang H, Wang J, Hansen T, Pedersen O, Nielsen R, Wang J (2010b) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42:969–972

Loos RJ, Lindgren CM, Li S et al (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40:768–775

Luft FC (2003) Mendelian forms of human hypertension and mechanisms of disease. Clin Med Res 1:291–300

Lupski JR, Stankiewicz P (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 1:e49

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N Engl J Med 362:1181–1191

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118

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 (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mathew CG (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat Rev Genet 9:9–14

McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. N Engl J Med 363:2339–2350

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6:S13–S20

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Montenegro G, Powell E, Huang J, Speziani F, Edwards YJ, Beecham G, Hulme W, Siskind C, Vance J, Shy M, Züchner S (2011)

Exome sequencing allows for rapid gene identification in a Charcot–Marie–Tooth family. Ann Neurol (Epub ahead of print)

Musunuru K, Kathiresan S (2010) Genetics of coronary artery disease. Annu Rev Genomics Hum Genet 11:91–108

Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ, Fennell T, Banks E, Ambrogio L, Cibulskis K, Kernytsky A, Gonzalez E, Rudzicz N, Engert JC, DePristo MA, Daly MJ, Cohen JC, Hobbs HH, Altshuler D, Schonfeld G, Gabriel SB, Yue P, Kathiresan S (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med 363:2220–2227

Nakamura Y (2007) The BioBank Japan Project. Clin Adv Hematol Oncol 5:696–697

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

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

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

Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010c) Massively parallel sequencing and rare disease. Hum Mol Genet 19:R119–R124

Nikopoulos K, Gilissen C, Hoischen A, van Nouhuys CE, Boonstra FN, Blokland EA, Arts P, Wieskamp N, Strom TM, Ayuso C, Tilanus MA, Bouwhuis S, Mukhopadhyay A, Scheffer H, Hoefsloot LH, Veltman JA, Cremers FP, Collin RW (2010) Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreo-retinopathy. Am J Hum Genet 86:240–247

Palmer LJ (2007) UK Biobank: bank on it. Lancet 369:1980–1982

Pang J, Zhang S, Yang P, Hawkins-Lee B, Zhong J, Zhang Y, Ochoa B, Agundez JA, Voelckel MA, Fisher RB, Gu W, Xiong WC, Mei L, She JX, Wang CY (2010) Loss-of-function mutations in HPSE2 cause the autosomal recessive urofacial syndrome. Am J Hum Genet 86:957–962

Paulussen AD, Stegmann AP, Blok MJ, Tserpelis D, Posma-Velter C, Detisch Y, Smeets EE, Wagemans A, Schrander JJ, van den Boogaard MJ, van der Smagt J, van Haeringen A, Stolte-Dijkstra I, Kerstjens-Frederikse WS, Mancini GM, Wessels MW, Hennekam RC, Vreeburg M, Geraedts J, de Ravel T, Fryns JP, Smeets HJ, Devriendt K, Schrander-Stumpel CT (2010) MLL2 mutation spectrum in 45 patients with Kabuki syndrome. Hum Mutat (Epub ahead of print)

Pierce SB, Walsh T, Chisholm KM, Lee MK, Thornton AM, Fiumara A, Opitz JM, Levy-Lahad E, Klevit RE, King MC (2010) Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. Am J Hum Genet 87:282–288

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordoñez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463:184–190

Rehman AU, Morell RJ, Belyantseva IA, Khan SY, Boger ET, Shahzad M, Ahmed ZM, Riazuddin S, Khan SN, Riazuddin S, Friedman TB (2010) Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. Am J Hum Genet 86:378–388

Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC (2010) Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum Mol Genet 19:4313–4318

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

Ropers HH (2007) New perspectives for the elucidation of genetic disorders. Am J Hum Genet 81:199–207

Ropers HH (2010) Single gene disorders come into focus again. Dialogues Clin Neurosci 12:95–102

Sandhu MS, Weedon MN, Fawcett KA, Wasson J, Debenham SL, Daly A, Lango H, Frayling TM, Neumann RJ, Sherva R, Blech I, Pharoah PD, Palmer CN, Kimber C, Tavendale R, Morris AD, McCarthy MI, Walker M, Hitman G, Glaser B, Permutt MA, Hattersley AT, Wareham NJ, Barroso I (2007) Common variants in WFS1 confer risk of type 2 diabetes. Nat Genet 39:951–953

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240

Seng KC, Seng CK (2008) The success of the genome-wide association approach: a brief story of a long struggle. Eur J Hum Genet 16:554–564

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Sirmaci A, Walsh T, Akay H, Spiliopoulos M, Sakalar YB, Hasanefendioğlu-Bayrak A, Duman D, Farooq A, King MC, Tekin M (2010) MASP1 mutations in patients with facial, umbilical, coccygeal, and auditory findings of Carnevale, Malpuech, OSA, and Michels syndromes. Am J Hum Genet 87:679–686

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 (2010) Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet 6:e1000991

Szelinger S, Kurdoglu A, Craig DW (2011) Bar-coded, multiplexed sequencing of targeted DNA regions using the Illumina genome analyzer. Methods Mol Biol 700:89–104

Teslovich TM, Musunuru K, Smith AV et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713

Turner EH, Ng SB, Nickerson DA, Shendure J (2010) Methods for genomic partitioning Annu Rev Genomics Hum Genet 10:263–284

Voight BF, Scott LJ, Steinthorsdottir V et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589

Volpi L, Roversi G, Colombo EA, Leijsten N, Concolino D, Calabria A, Mencarelli MA, Fimiani M, Macciardi F, Pfundt R, Schoenmakers EF, Larizza L (2010) Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am J Hum Genet 86:72–76

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M (2010) Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause

of nonsyndromic hearing loss DFNB82. Am J Hum Genet 87:90–94

Wang J, Wang W, Li R et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, Jiang H, Zhang P, Shen L, Guo JF, Li N, Li YR, Lei LF, Zhou J, Du J, Zhou YF, Pan Q, Wang J, Wang J, Li RQ, Tang BS (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133:3510–3518

Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Winckler W, Weedon MN, Graham RR, McCarroll SA, Purcell S, Almgren P, Tuomi T, Gaudet D, Boström KB, Walker M, Hitman G, Hattersley AT, McCarthy MI, Ardlie KG, Hirschhorn JN, Daly MJ, Frayling TM, Groop L, Altshuler D (2007) Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. Diabetes 56:685–693

Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP (2010) Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med (Epub ahead of print)

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592