Genetics

Volume 2, Issue 3, 219-229.
DOI: 10.3934/genet.2015.3.219
Received date 29 April 2015,
Accepted date 16 September 2015,
Published date 22 September 2015

http://www.aimspress.com/

Review

Influence of technology in genetic epidemiology

Marcos Morey*, Ana Fernández-Marmiesse, Jose Angel Cocho and María L. Couce

Unit of Diagnosis and Treatment of Congenital Metabolic Diseases, Neonatology Service, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, CIBERER, Health Research Institute of Santiago de Compostela (IDIS), A Choupana, s/n, 15706 Santiago de Compostela, A Coruña, Spain

*** Correspondence:** Email: marcosmorey84@gmail.com Tel: +34 660-006-337.

Abstract: Genetic epidemiology is the study of genetic factors and their influence on health and disease. Traditionally, these studies have been based on familial aggregation, segregation, or linkage analysis, mainly allowing the study of monogenic disorders. Advances in biotechnology have made techniques such as genome-wide association studies and next-generation sequencing possible, allowing more complex studies. In addition to the completion of large consortia projects, such as the Human Genome Project, ENCODE, and the 1000 Genome Project, these techniques make it possible to explain a higher proportion of the heritability in polygenic disorders compared to previous techniques. Here, we provide an overview of approaches to genetic epidemiology and how technological improvements have influenced experimentation in this area. These improvements have led genetic epidemiology to unprecedented advances, being excellent tools for understanding the genetic variability underlying complex phenotypes.

Keywords: genetic epidemiology; linkage analysis; genome-wide assays; next generation sequencing

1. Introduction

Genetic epidemiology studies how genetic factors determine health and disease in families and populations and their interactions with the environment. Classical epidemiology usually studies disease patterns and factors associated with disease etiology, with a focus on prevention, whereas

molecular epidemiology measures the biological response to environmental factors by evaluating the response in the host (e.g., somatic mutations and gene expression) [1].

Interest in how the environment triggers a biological response started in the mid-nineteenth century, but approximately 100 years passed until epidemiologists and genetic epidemiologists had adequate analytical methods at their disposal to understand how genes and the environment interact [2]. The beginning of genetic epidemiology as a stand-alone discipline started with Morton in the 1980s with one of the most accepted definitions: "a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in a population" [3]. However, epidemiology is clearly a multidisciplinary area that examines the role of genetic factors and environmental contributors to disease. Equal attention has to be given to the differential impact of environmental agents (familial and non-familial) on different genetic backgrounds [4] to detect how the disease is inherited, and to determine related genetic factors.

With advances in molecular biology techniques in the last 15 years, our ability to survey the genome, give a functional meaning to the variants found, and compare it among individuals has increased dramatically [5]. Although there is still a long way to go to fully understanding rare diseases and how genetic variability influences phenotype, these technological advances allow more in depth biological knowledge of epidemiology [6] (Figure 1).

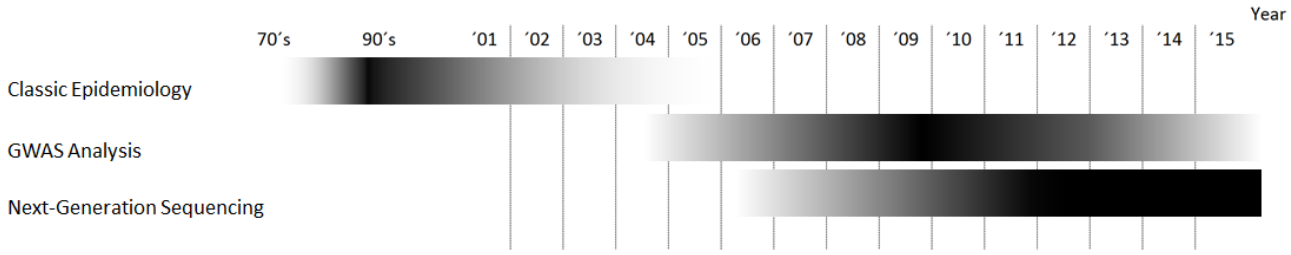

Figure 1. Relative importance of each methodology over time.

Here, we present an overview of approaches in genetic epidemiology studies, ranging from classical family studies/segregation analysis and population studies to the more recent genome-wide association studies (GWAS) and next-generation sequencing (NGS), which have fueled research on this area by allowing more precise data to be obtained in less time.

2. Classical epidemiology

Genetic epidemiology was born in the 1960s as a combination of population genetics, statistics and classical epidemiology, and applied the methods of biological study available at that time. Generally, the studies included the following steps: establish genetic factor involvement in the disorder, measure the relative size of the contribution of the genetic factors in relation to other sources of variability (e.g., environmental, physical, chemical, or social factors), and identify the responsible genes/genomic areas. For that, family studies (e.g., segregation or linkage analysis) or population (association) studies are usually performed. Approaches include genetic risk studies to determine the relative contribution of the genetic basis and ambience by utilizing monozygotic and dizygotic twins [7]; segregation analyses to determine the inheritance model by studying family trees [8]; linkage studies

to determine the coordinates of the implicated gene(s) by studying its cosegregation; and association studies to determine the precise allele associated with the phenotype by using linkage disequilibrium analysis [9].

2.1. Genetic risk studies

Genetic risk studies require a family-based approach in order to evaluate the distribution of traits in families and identify the risk factors that cause a specific phenotype. Traditionally, twin studies have been used to estimate the influence of genetic factors underlying the phenotype by comparing monozygotic (sharing all of their genes) and dizygotic (sharing half of their genes) twins. In order to standardize the measurement of similarity, a concordance rate is used. Monozygotic twins generally being more similar than dizygotic twins is usually considered evidence of the importance of genetic factors in the final phenotype, but several studies have questioned this view [10]. Importantly, twin studies make some preliminary assumptions, such as random mating, in which all individuals in the population are potential partners, and that genetic or behavioral restrictions are absent, meaning that all recombinations are possible [11]. Twin studies also assume that the two types of twins share similar environmental experiences relevant to the phenotype being studied [12]. Concordance rates of less than 100% in monozygotic twins indicate the importance of environmental factors [13,14].

2.2. Segregation analysis

The objective of segregation analysis is to determine the method of inheritance of a given disease or phenotype. This approach can distinguish between Mendelian (i.e., autosomal or sex-linked, recessive or dominant) and non-Mendelian (no clear pattern [15]) inheritance patterns. For the non-Mendelian patterns, factors interfering with genotype-phenotype correlation, such as incomplete penetrance, variable expressivity and locus heterogeneity, and the variable effect of environmental factors can complicate the segregation analysis [16]. Thus, families with large pedigrees and many affected individuals can be particularly informative for these studies [17].

2.3. Linkage studies

Linkage studies aim to obtain the chromosomal location of the gene or genes involved in the phenotype of interest. Genetic Linkage was first used by William Bateson, Edith Rebecca Saunders, and Reginald Punnett, and later expanded by Thomas Hunt Morgan [18]. One of the main concepts in linkage studies is the recombination fraction, which is the fraction of births in which recombination occurred between the studied genetic marker and the putative gene associated with the disease. If the loci are far apart, segregation will be independent; the closer the loci, the higher the probability of cosegregation [19,20]. Classically, the percent of recombinants has been used to measure genetic distance: one centimorgan (cM), named after the geneticist Thomas Hunt Morgan, is equal to a 1% chance of recombination between two loci. With this information, linkage maps can be constructed. A linkage map is a genetic map of a species in which the relative positions of its genes or genetic markers are shown based on the frequencies of recombination between markers during the crossover of homologous chromosomes [21]. The more frequent the recombination, the farther both loci are. Linkage maps are not physical maps, but relative maps. Translating the measure into a physical unit

of distance, 1 cM is approximately 1 million bases [22].

Linkage analysis is based on the likelihood ratio, also called the logarithm of odds (LOD) score, which is the statistical estimate of whether two genes are likely to be located near each other on a chromosome and, therefore, their likelihood of being inherited together. This analysis can be either parametric (if the relationship between genotype and phenotype is assumed to be known) or non-parametric (if the relationship between phenotype and genotype is not established) [23].

2.4. Association studies

Association studies, which are frequently mixed up with linkage studies, focus on populations. This approach tests whether a locus differs between two groups of individuals with phenotypic differences. The loci are usually susceptibility markers that increase the probability of having the phenotype or disease, but for which there is not necessarily linkage, as it can be neither necessary nor sufficient for phenotype/disease expression [24]. Due to the increased number of individuals in such a study, the statistical power of this approach is greater than that of linkage analysis and more prone to detect genes with a low effect on the phenotype [25].

3. Molecular epidemiology

Although the current approaches to epidemiology studies rely on those discussed above, advances in biotechnology have brought significant changes to genetic epidemiology and how these studies are performed. Technological improvements have accelerated data gathering and interpretation [5], broadening our understanding of disease etiology.

3.1. GWAS analysis

In the last 10 years, GWAS have transformed the world of genetic epidemiology, with a large number of research studies and publications on complex diseases, allowing the identification of a great number of phenotype-associated genomic loci [26]. Typically, linkage studies in combination with information from family pedigrees are used to broadly estimate the position of the disease-associated loci [27]. With the advent and popularization of array technology, GWAS have become a widespread tool for genetic epidemiology studies. This approach allows the simultaneous and highly accurate interrogation of millions of genomic markers at a reasonable cost and speed. The first GWAS was published by Klein et al [28], and to date more than 2000 articles have been published based on this methodology [29]. These studies allow the determination of thousands of disease-associated genomic loci, which could serve as risk predictors if a large enough discovery sample size is provided [30]. In addition, these dense, genome-wide markers allow a reasonable approximation to understand narrow-sense heritability [31].

In order to find a genetic association with a given phenotype, GWAS need the effect of the variant(s) to be notorious and/or to have strong linkage disequilibrium with previously genotyped markers [32]. GWAS are mostly useful under the common-disease common-variant hypothesis [33]. Therefore, this approach may not be adequate for some common diseases for which rare variants with additive effects are the underlying mechanism [34].

GWAS have been useful for obtaining genomic information about the basis of several diseases,

but they have some limitations. First, as genetic markers are only being surveyed in this approach, it is difficult to interpret the results, partially due to our current lack of understanding of genomic function. The use of non-random associations of variants at different loci (i.e., linkage disequilibrium) as a correlation tool also impacts the interpretation of results [35]. GWAS identify blocks of variants, not necessarily the real functional variants [36]. Second, and related to the first point, we miss part of the heritability because of a gap between the variance explained by the significant single nucleotide polymorphisms (SNPs) identified and the estimated heritability [37]. This could be explained, at least partially, by the limited info obtained from the genome by GWAS. Small insertions and deletions, large structural variants, epigenetic factors, gene interactions, and gene by environment interactions could be playing a role in that [38-40].

3.2. Next-generation sequencing

In the last 8 years, the advent of NGS has helped fill the gap in understanding the genome. As with sequencing each individual base is interrogated, it may help in screening rare variants.

NGS promises great opportunities for finding the answers to questions raised by array technology, as it has the potential to provide additional biological insight into disease etiology. As we move into an era of personalized medicine and complex genomic databases, the demand for new and existing sequencing technologies is constant. Although it is not yet possible to routinely sequence an individual genome for $1000, novel approaches are reducing the cost per base and increasing throughput on a daily basis [41,42]. Moreover, advances in sequencing methodologies are changing the ways in which scientists analyze and understand genomes, whereas the results that they yield are being disseminated widely through science news magazines [43].

Advances in knowledge on the genetic basis of pathologies have changed the way in which such entities are understood. Thus, diseases have gone from being individual-specific to a familial phenomenon in which genetic alterations (mutations) can be genealogically traced to the molecular level.

NGS can be used to identify several types of alterations in the genome, the most common of which are SNPs, structural variants, and epigenetic variations on very large regions of the genome [44-48]. Because of the capacity of NGS to detect many types of genomic and epigenetic variations on a genome scale in a hypothesis-free manner with great coverage and accuracy, it is starting to explain the missing heritability gap left by GWAS [37,49]. With these tools, it is currently possible to obtain a more comprehensive view of how phenotypic variance works in genetic epidemiology.

NGS allows researchers to study all of the SNPs in each individual directly [50]. This is a large amount of information, which requires large data analysis resources. In whole genome and whole exome analysis, the number of rare variants that is revealed can be overwhelmingly large. Most of these variants have no known functional relevance. Therefore, it is not yet easy or straightforward to filter and identify the causal variants, even after accurate variant calling has been performed. Targeted resequencing of candidate genes could be a feasible option for avoiding the high number of variants obtained by whole genome and whole exome sequencing in cases in which there is already a strong knowledge basis regarding phenotype etiology, but the number of genes is still large for traditional Sanger sequencing [51-53]. This type of study significantly reduces analysis costs, as samples could be multiplexed for the analysis, and simultaneously reduces the number of variants found in the regions of interest. Therefore, the analysis will be comparatively easier and the amount of information given

per individual less, though more focused on targeted genes. Another option would be to sequence family trios in order to allow filtering of shared variants and speed up the identification of *de novo* mutations on the affected individual [54-57].

Thus, NGS can be applied to the study of both rare and common diseases. For rare monogenic diseases, genes can be directly sequenced and variants identified with a small sample size [58-60]. Depending on the genetic heterogeneity, finding the involved allele could still be challenging. Rare diseases are usually identified by symptoms, which could be shared by completely different diseases, as the mechanisms underlying the phenotype could be different. This is one of the most difficult points when analyzing rare diseases with genetic heterogeneity. For these cases, larger sample sizes are usually required in order to find the genomic loci implicated in the phenotype etiology [6,55,61-64]. Time of appearance and disease severity are often ruled by the residual enzymatic activity of mutated proteins and the influence of the individual genomic background. Therefore, the type of causal variants could be diverse (e.g., coding, splicing, non-coding, missense, epigenetic alterations), as well as the influence on final protein activity. To make it even more complex, those alterations could be shared between individuals with different phenotypes depending on the penetrance of the variant, background, or environment [65].

3.3. Functional annotation

As advances in technology imply generating a larger amount of data, genomic annotation is crucial for variant prioritization and the interpretation of results. With the use of adequate tools, random and systematic noise, false positives, and false negatives can be reduced, easing the final analysis. Study design can also influence the analysis, as it is a compromise between the amount of data to be generated and the scope of the study; whole genome sequencing is expected to provide hundreds of thousands of variants, most with yet unknown significance, in intronic or non-coding regions. Whole exome sequencing will still result in a large number of variants, but the annotation of exonic regions is much more curated than that of intronic regions. In the case of a gene-panel targeted study, the list of variants could be reduced to several hundred, depending on the number of genes included, making the analysis and filtering easier, but the data will be limited to the previously selected genes.

The Human Reference Genome established in 2001 [66,67] and the achievements of large sequencing projects such as the 1000 Genome Project [68] are catalyzing advances in human genetics. Large samples obtained with these projects allow adequate statistical power to shed light into rare variant effects [6,64,69] and empower the usage of analysis tools for automatic variant annotation.

Methods for variant analysis and effect prediction have been developed in order to speed up this process. A complete list of software and tools is available online [70]. These methods focus mostly on coding regions in the human genome. Although 98% of the human genome is non-coding [71], these regions are less well known [72]. Thus, there are annotation tools extending the scope to the non-coding and regulatory areas, such as HaploReg [73], RegulomeDB [74], CADD [75], VariantDB [76], GWAVA [77], and ANNOVAR [78], among others [79]. However, the final judgment regarding potential variants is in the hands of the user.

Large consortia, such as the ENCODE project, have generated a large amount of information on the human genome [80], including information on transcriptional binding sites, histone modifications, and DNA methylation, in order to explain the influence on overall phenotype.

4. Conclusion

Technological advances are playing a crucial role in the evolution of genetic epidemiology as a discipline, as they allow us to address more complex biological questions. The spread and popularization of NGS due to its reduction on the cost per sequenced base is democratizing access to these technologies, allowing researchers to continue on the path opened by previous tools, such as GWAS. This has been observed by the increasing number of research groups and publications using these technologies.

Currently, NGS has the potential to move genetic epidemiology forward, as it allows the assessment of common and rare SNPs, as well as other diverse types of genomic and epigenetic variations using a hypothesis-free whole genome analysis. The elucidation of genome variability for increasing our understanding of living systems is crucial.

Nonetheless, advances would not be possible without the appropriate mathematical algorithms to transform the sequences into meaningful information or without databases to annotate the identified variants. To fill this gap in information, large programs have been established (1000 Genomes Project consortium [81] and the NHGRI Genome Sequencing Program (GSP) [82]) to provide annotation data on the variations in the human genome.

Overall, new technologies such as GWAS and NGS constitute an opportunity for researchers to understand the genetic variability underlying complex phenotypes and provide unprecedented tools in their investigation.

Conflict of Interest

The authors declare that they have no competing interest.

References

1. Morton NE (1997) Genetic epidemiology. *Ann Hum Genet* 61: 1-13.
2. Morton NE (1994) Fundamentals of genetic epidemiology. *Genet Epidemiol* 11: 389-390.
3. Morton NE (1982) Outline of genetic epidemiology. S. Karger AG (Switzerland), 252.
4. Cohen BH (1980) Chronic obstructive pulmonary disease: A challenge in genetic epidemiology. *Am J Epidemiol* 112: 274-288.
5. Morey M, Fernández-Marmiesse A, Castiñeiras D, et al. (2013) A glimpse into past, present, and future DNA sequencing. *Mol Genet Metab* 110: 3-24.
6. Matullo G, Gaetano CD, Guarrera S (2013) Next generation sequencing and rare genetic variants: From human population studies to medical genetics. *Environ Mol Mutagen* 54: 518-532.
7. IJzerman RG, Stehouwer CDA, Boomsma DI (2000) Evidence for genetic factors explaining the birth Weight–Blood pressure relation: Analysis in twins. *Hypertension* 36: 1008-1012.
8. Ostern R, Fagerheim T, Hjellnes H, et al. (2014) Segregation analysis in families with charcot-marie-tooth disease allows reclassification of putative disease causing mutations. *BMC Med Genet* 15: 12.
9. Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Res* 10: 1435-1444.

10. Guo SW (2001) Does higher concordance in monozygotic twins than in dizygotic twins suggest a genetic component?. *Hum Hered* 51: 121-132.
11. King RC, Mulligan P, Stansfield W (2013) A dictionary of genetics. Oxford University Press, 641.
12. Wong AHC, Gottesman II, Petronis A (2005) Phenotypic differences in genetically identical organisms: The epigenetic perspective. *Hum Mol Genet* 14: R11-18.
13. Chaganti RSK, Miller DR, Meyers PA, et al. (1979) Cytogenetic evidence of the intrauterine origin of acute leukemia in monozygotic twins. *N Engl J Med* 300: 1032-1034.
14. Bell JT, Saffery R (2012) The value of twins in epigenetic epidemiology. *Int J Epidemiol* 41: 140-150.
15. Elston RC (1981) Segregation analysis. In: Harris H and Hirschhorn K, eds. Springer US, 63-120.
16. Jarvik GP (1998) Complex segregation analyses: Uses and limitations. *Am J Hum Genet* 63: 942-946.
17. Terwilliger JD, Goring HH (2000) Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Hum Biol* 72: 63-132.
18. Bateson W, Waunders ER, Punnett RC (1909) Experimental studies in the physiology of heredity. *Zeitschrift fÃ¼r Induktive Abstammungs- Und Vererbungslehre* 2: 17-19.
19. Stevens WL (1939) Tables of the recombination fraction estimated from the product ratio. *J Genet* 39: 171-180.
20. Tan YD, Fu YX (2007) A new strategy for estimating recombination fractions between dominant markers from an F2 population. *Genetics* 175: 923-931.
21. Botstein D, White RL, Skolnick M, et al. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32: 314-331.
22. Stocker AJ, Rusuwa BB, Blacket MJ, et al. (2012) Physical and linkage maps for drosophila serrata, a model species for studies of clinal adaptation and sexual selection. *G3 (Bethesda)* 2: 287-297.
23. Bailey-Wilson JE (2005) Parametric versus nonparametric and two-point versus multipoint: Controversies in gene mapping. In: Anonymous Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. John Wiley & Sons, Ltd.
24. Hirschhorn JN, Lohmueller K, Byrne E, et al. (2002) A comprehensive review of genetic association studies. *Genet Med* 4: 45-61.
25. Cordell HJ, Clayton DG (2005) Genetic association studies. *Lancet* 366: 1121-1131.
26. McCarthy MI, Abecasis GR, Cardon LR, et al. (2008) Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet* 9: 356-369.
27. St George-Hyslop PH, Haines JL, Farrer LA, et al. (1990) Genetic linkage studies suggest that alzheimer's disease is not a single homogeneous disorder. *Nature* 347: 194-197.
28. Klein RJ, Zeiss C, Chew EY, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308: 385-389.
29. Welter D, MacArthur J, Morales J, et al. (2013) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42: D1001-1006.
30. Kooperberg C, LeBlanc M, Obenchain V (2010) Risk prediction using genome-wide association studies. *Genet Epidemiol* 34: 643-652.

31. Gusev A, Bhatia G, Zaitlen N, et al. (2013) Quantifying missing heritability at known GWAS loci. *PLoS Genet* 9: e1003993.
32. Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187: 367-383.
33. Visscher P, Brown M, McCarthy M, et al. (2012) Five years of GWAS discovery. *Am J Hum Genet* 90: 7-24.
34. Gibson G (2012) Rare and common variants: Twenty arguments. *Nat Rev Genet* 13: 135-145.
35. Slatkin M (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9: 477-485.
36. Cooper GM, Shendure J (2011) Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 12: 628-640.
37. Manolio TA, Collins FS, Cox NJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461: 747-753.
38. Frazer KA, Murray SS, Schork NJ, et al. (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 10: 241-251.
39. Johnson DS, Mortazavi A, Myers RM, et al. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497-1502.
40. Shen P, Wang W, Krishnakumar S, et al. (2011) High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci U S A* 108: 6549-6554.
41. Service RF (2006) The race for the $1000 genome. *Science* 311: 1544-1546.
42. Wetterstrand KA, DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. 2015. Available from: www.genome.gov/sequencingcosts
43. Broadwith P (2012) Sequencing in the fast lane. *Chem World* 9: 54-58.
44. Feldman AL, Dogan A, Smith DI, et al. (2010) Massively parallel mate pair DNA library sequencing for translocation discovery: Recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas. *ASH Annual Meeting Abstracts* 116: 633.
45. Green R, Malaspinas A, Krause J, et al. (2008) A complete neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134: 416-426.
46. Durbin RM, Altshuler DL, Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073.
47. Peters BA, Kermani BG, Sparks AB, et al. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487: 190-195.
48. Butler J, MacCallum I, Kleber M, et al. (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res* 18: 810-820.
49. Furlotte NA, Heckerman D, Lippert C (2014) Quantifying the uncertainty in heritability. *J Hum Genet* 59: 269-275.
50. Majewski J, Schwartzentruber J, Lalonde E, et al. (2011) What can exome sequencing do for you?. *J Med Genet* 48: 580-589.
51. Wooderchak-Donahue W, O'Fallon B, Furtado L, et al. (2012) A direct comparison of next generation sequencing enrichment methods using an aortopathy gene panel- clinical diagnostics perspective. *BMC Medical Genomics* 5: 1-10.
52. Kalender Atak Z, De Keersmaecker K, Gianfelici V, et al. (2012) High accuracy mutation detection in leukemia on a selected panel of cancer genes. *PLoS One* 7: e38463.

53. Ni T, Wu H, Song S, et al. (2009) Selective gene amplification for high-throughput sequencing. *Recent Pat DNA Gene Seq* 3: 29-38.
54. Gaugler T, Klei L, Sanders SJ, et al. (2014) Most genetic risk for autism resides with common variation. *Nat Genet* 46: 881-885.
55. Muona M, Berkovic SF, Dibbens LM, et al. (2015) A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nat Genet* 47: 39-46.
56. Xu B, Ionita-Laza I, Roos JL, et al. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 44: 1365-1369.
57. Cardinale CJ, Kelsen JR, Baldassano RN, et al. (2013) Impact of exome sequencing in inflammatory bowel disease. *World J Gastroenterol* 19: 6721-6729.
58. Gilissen C, Arts HH, Hoischen A, et al. (2010) Exome sequencing identifies WDR35 variants involved in sensenbrenner syndrome. *Am J Hum Genet* 87: 418-423.
59. Boycott KM, Vanstone MR, Bulman DE, et al. (2013) Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat Rev Genet* 14: 681-691.
60. Roach JC, Glusman G, Smit AF, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328: 636-639.
61. Fernandez-Marmiesse A, Morey M, Pineda M, et al. (2014) Assessment of a targeted resequencing assay as a support tool in the diagnosis of lysosomal storage disorders. *Orphanet J Rare Dis* 9: 59.
62. Audo I, Bujakowska KM, Leveillard T, et al. (2012) Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases. *Orphanet J Rare Dis* 7: 8.
63. Mardis ER (2009) New strategies and emerging technologies for massively parallel sequencing: Applications in medical research. *Genome Med* 1: 40.
64. Wendl MC, Wilson RK (2009) The theory of discovering rare variants via DNA sequencing. *BMC Genomics* 10: 485.
65. Cooper DN, Krawczak M, Polychronakos C, et al. (2013) Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* 132: 1077-1130.
66. Venter JC, Adams MD, Myers EW, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
67. Lander ES, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
68. Durbin R, Altshuler D, Durbin R, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073.
69. Panoutsopoulou K, Tachmazidou I, Zeggini E (2013) In search of low-frequency and rare variants affecting complex traits. *Hum Mol Genet* 22: R16-21.
70. Li J, Schmieder R, Ward RM, et al. (2012) SEQanswers: An open access community for collaboratively decoding genomes. *Bioinformatics* 28: 1272-1273.
71. Elgar G, Vavouri T (2008) Tuning in to the signals: Noncoding sequence conservation in vertebrate genomes. *Trends Genet* 24: 344-352.
72. Eisenberger T, Neuhaus C, Khan AO, et al. (2013) Increasing the yield in targeted next-generation sequencing by implicating CNV analysis, non-coding exons and the overall variant load: The example of retinal dystrophies. *PLoS One* 8: e78496.

73. Ward LD, Kellis M (2011) HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40: D930-934.
74. Boyle AP, Hong EL, Hariharan M, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22: 1790-1797.
75. Kircher M, Witten DM, Jain P, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46: 310-315.
76. Vandeweyer G, Van Laer L, Loeys B, et al. (2014) VariantDB: A flexible annotation and filtering portal for next generation sequencing data. *Genome Med* 6: 74.
77. Ritchie GRS, Dunham I, Zeggini E, et al. (2014) Functional annotation of noncoding sequence variants. *Nat Meth* 11: 294-296.
78. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 38: e164-.
79. Henry VJ, Bandrowski AE, Pepin A, et al. (2014) OMICtools: An informative directory for multi-omic data analysis. *Database (Oxford)* bau069.
80. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.
81. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.
82. McEwen JE, Boyer JT, Sun KY, et al. (2014) The ethical, legal, and social implications program of the national human genome research institute: Reflections on an ongoing experiment. *Annu Rev Genomics Hum Genet* 15: 481-505.

AIMS Press