Current Genetic Medicine Reports
https://doi.org/10.1007/s40142-019-0161-5

NEUROGENETICS AND PSYCHIATRIC GENETICS (C CRUCHAGA AND C KARCH, SECTION EDITORS)

Benefits and Challenges of Rare Genetic Variation in Alzheimer's Disease

Detelina Grozeva[1] · Salha Saad[1] · Georgina E. Menzies[2] · Rebecca Sims[1,2]

Abstract

Purpose of Review It is well established that sporadic Alzheimer's disease (AD) is polygenic with common and rare genetic variation alongside environmental factors contributing to disease. Here, we review our current understanding of the genetic architecture of disease, paying specific attention to rare susceptibility variants, and explore some of the limitations in rare variant detection and analysis.

Recent Findings Rare variation has been shown to robustly associate with disease. These include potentially damaging and loss of function mutations that are easily modelled in silico, in vitro and in vivo, and represent potentially druggable targets. A number of risk genes, including *TREM2*, *SORL1* and *ABCA7* show multiple independent associations suggesting that they may influence disease via multiple mechanisms. With transcriptional regulation, inflammatory response and modification of protein production suggested to be of primary importance.

Summary We are at the beginning of our journey of rare variant detection in AD. Whole exome sequencing has been the predominant technology of choice. While fruitful, this has introduced a number of challenges with regard to data integration. Ultimately the future of disease-associated rare variant identification lies in whole genome sequencing projects that will allow the testing of the full range of genomic variation.

Keywords Alzheimer's disease · Genetics · Susceptibility · Rare variants

Introduction

Alzheimer's disease (AD) is a devastating and progressive neurodegenerative disease that is estimated to account for up to 80% of all dementia cases, meaning there are over 37 million AD sufferers world-wide. With an ageing population, the incidence of AD is expected to rise exponentially, and with no effective prevention or treatment, dementia is now one of the world's greatest public health issues.

It is well established that while AD has some symptoms common to all sufferers, aetiologically, there are at least two different forms of disease. A small number of individuals (~ 1%) carry a disease-causing mutation within the *APP*, *PSEN1* or *PSEN2* genes. These Mendelian forms of disease underpin the amyloid cascade hypothesis of AD. The hypothesis claims the misprocessing of β-amyloid (Aβ) and its deposition as the primary causal event in AD pathogenesis [1]. However, the failures of clinical trials focussed on Aβ pathology suggest that this hypothesis may only relate to Mendelian forms of disease. Conversely, these tested therapeutics may only be effective in the prodromal stages of AD, rather than the symptomatic phase when participants for clinical trials are recruited [2, 3]. Undoubtedly, the amyloid cascade is part of a complex interplay of influences on disease development that includes tau and immunity/inflammation.

Other forms of disease are seen to segregate, although not completely, in families where there are no fully penetrant causative mutations (heritability estimates of over 90%) [4]. Non-familial, sporadic AD is a highly heritable (estimates of 58–79%) [5] and polygenic disorder [6], with over 40 risk loci reliably established [3]. The majority of the identified loci are common (minor allele frequency > 1%) with small effect sizes and reside in non-coding parts of the genome. The identification of these common loci has greatly improved our understanding of the underlying biology of disease, highlighting

This article is part of the Topical Collection on *Neurogenetics and Psychiatric Genetics*

✉ Rebecca Sims
simsrc@cardiff.ac.uk

[1] Division of Psychological Medicine and Clinical Neuroscience, MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff, UK

[2] UK Dementia Research Institute at Cardiff, School of Medicine, Cardiff University, Cardiff, UK

Published online: 01 February 2019

immunity, endocytosis, cholesterol metabolism, protein ubiquitination and more recently Aβ processing [7]. However, the identified common variants do not pinpoint protein coding changes for direct modelling, nor do they explain the estimated disease heritability. In fact, the common variants known to associate with AD are thought to account for less than half of the genetic liability for disease [8]. In recent years, and as genotyping and sequencing technologies have advanced, the field has focused on the identification of rare variation for disease. Here, we will focus on these discoveries, the technologies that enabled their discovery and the challenges of working with big data for rare variant detection.

Identified Common Risk Loci

Risk for non-familial forms of AD is inferred by both environmental and genetic factors, with common and rare variation involved in non-Mendelian disease aetiology. *Apolipoprotein E* (*APOE*) on chromosome 19 was the first and remains the strongest genetic risk factor for AD [9]. *APOE* encodes three isoforms of the protein, ε2, ε3 and ε4. Disease risk is increased in carriers of the ε4 allele, in a dose-dependent manner, with a threefold increase in ε4 heterozygotes (ApoE ε3/ε4), and a 15-fold increase in ε4 homozygotes (ApoE ε4/ε4). The ε2 allele is thought to confer a small protective effect [10, 11].

Over the past 9 years, genome-wide association studies (GWAS) in case-control cohorts of tens of thousands of individuals have identified nearly 40 common genome-wide significant risk loci. The identified susceptibility variants generally have small effect sizes (odds ratios ~ 1.2) and are often found in intergenic or intronic regions (therefore termed locus rather than gene) making it difficult to pinpoint which gene has a functional effect. The *CLU*, *PICALM* and *CR1* loci [12, 13] were identified in 2009 in two back-to-back publications. Subsequent publications have identified *BIN1*, *EPHA1, MS4A, CD2AP*, *ABCA7*, *HLA-DRB5/HLA-DRB1*, *PTK2B*, *SORL1*, *SLC24A4-RIN3, INPP5D, MEF2C, NME8, ZCWPW1, CELF1, FERMT2* and *CASS4* as risk loci for sporadic AD [14–16]. This success can be largely attributed to the extensive collaboration across four genetic consortia; Genetic and Environmental Risk in AD (GERAD), European AD Initiative (EADI), Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) and AD Genetics Consortium (ADGC), badged together as the International Genomics of Alzheimer's Project (IGAP). In addition, genome-wide gene-wide analyses identified five novel genome-wide significant loci, *TP53INP1*, *IGHV1–67,* [17] *PPARGC1A*, *RORA* and *ZNF423* [18] using the IGAP dataset. Building upon the initial IGAP dataset, *TRIP4* [19], *ECHDC3* [7, 20], *IQCK*, *ACE*, *ADAM10* and *ADAMTS1* [7] have more recently been reported as genome-wide susceptibility loci.

A novel approach to AD risk variant detection has been to infer AD diagnosis based on reports of parental history of dementia in the UK Biobank dataset [21•, 22]. Although this approach lacks power and introduces diagnostic noise, it proved informative when combined with clinically defined AD cohorts. Combination of UK Biobank data with additional datasets identified association for AD risk with variants in *KAT8* [21•], *HESX1*, *CLNK/HS3ST1*, *CNTNAP2*, *APH1B*, *ALPK2* and *LOC388553* loci [22•].

It is estimated that a substantial proportion (approximately 60%) [23, 24] of the genetic variance of sporadic AD is not accounted for by *APOE* or the common genome-wide associated loci. GWAS in other complex traits suggest that more powerful GWAS will identify further additional associations [25]. However, a substantial proportion of the 'missing heritability' of AD is likely to be accounted for by rare and low frequency variation of small to moderate effect.

Identified Rare Risk Loci

A flurry of recent reports has confirmed that a proportion of AD is explained by rare variation with larger effect sizes than normally seen with common variation. Such variants point to the involvement of novel genes in the pathophysiology of disease and importantly highlight protein-coding changes that are novel therapeutic targets.

To date, four genes have been identified showing replicable genome-wide significant association with disease. In 2013, two groups identified p.Arg47His (R47H) within the triggering receptor expressed on myeloid cells 2 (*TREM2*) gene as a late-onset AD risk variant; the independent studies were published in back-to-back publications in the New England Journal of Medicine [26, 27]. Both studies utilised a multi-stage study design. Guerreiro and colleagues [27] undertook whole exome sequencing (WES) and Sanger sequencing in a discovery cohort of 1092 AD cases and 1107 controls, before confirming their identified association through additional independent genotyping and meta-analysis of imputed GWAS data. Simultaneously, Jonsson and colleagues [26] identified the same *TREM2* variant through whole genome sequencing (WGS) of 2261 Icelandic participants and reported that the R47H variant significantly associated with risk of AD in a largely Scandinavian population. Replication of their findings was achieved through additional genotyping of independent cohorts and imputed datasets from Europe and the USA, with meta-analysis of the datasets showing association at the genome-wide level (odds ratio, 2.90; 95% CI, 2.16 to 3.91; $P = 2.1 \times 10^{-12}$) [26]. The association at *TREM2* p.(Arg47His) and AD has been replicated in multiple populations of European descent [28–35, 36•, 37, 38], and although the risk-effect sizes vary by cohort, cumulatively, the results

suggest that *TREM2* R47H is the largest genetic effector of sporadic AD after *APOE* ε4.

A second *TREM2* variant that had previously shown suggestive evidence for association with disease [39, 40•, 41] was shown to increase risk for sporadic AD at the genome-wide significance level via an exome-chip microarray study [36•]. The Illumina exome-chip was designed as an intermediate experiment between current genotyping arrays, which focus on relatively common variants, and exome sequencing of very large numbers of samples. The array contains over a quarter of a million variants identified from WGS and WES data of over 12,000 individuals. We and others showed a rare coding mutation at *TREM2* p.Arg62His (R62H) increased risk for disease independently of the R47H *TREM2* mutation with an odds ratio of 1.67 [36•]. In the same publication, we identified novel association within two additional genes, *phospholipase C gamma 2* (*PLCG2*) and *ABI Family Member 3* (*ABI3*). The *PLCG2* variant p.Pro522Arg (P522R) shows a protective effect against disease, while the *ABI3* variant p.Ser209Phe (S209F) increased disease risk [36•]. AD risk inferred by *TREM2* does not appear to be population-specific with association reported in European [26, 27, 36•], African American [42] and Asian [43] populations. However, a number of studies show conflicting data with the *TREM2* R47H variant not significantly associated with AD risk in an African-American cohort [38], and four studies failing to detect the R47H variant in Chinese subjects [44–47]. In one study of Japanese subjects, the R47H variant was extremely rare (minor allele frequency < 0.006) and no association was found with AD.

In 2012, Jonsson and colleagues [39] showed, for the first time in sporadic AD, an association with the *amyloid precursor protein* (*APP*) gene that causes familial forms of AD. The identified protein-coding change p.Ala673Thr (A673T) was identified in WGS data from 1795 Icelanders and was shown to protect against disease, and cognitive decline in elderly non-diseased participants [39]. The protein change is thought to reduce β-cleavage of APP with approximately 40% reduction in the formation of amyloidogenic peptides in vitro.

Exome-wide significant association with sporadic disease was recently identified in the AD Sequencing Project (ADSP). Novel single nucleotide variant (SNV) association with disease was identified at the *IGHG3* (an immunoglobulin gene whose antibodies interact with β-amyloid) and *AC099552.4* (a long non-coding RNA) genes, while exome-wide gene-wide association was identified at the *ZNF655* (zinc-finger protein) gene [40•]. These newly discovered genes point to the important role of transcriptional regulation in AD pathogenesis and add further support to the role of inflammatory response and modification of protein production in disease biology [40•]. Analysis of the ADSP data to examine the contribution to disease across dementia genes and clinically diagnosed AD identified rare pathogenic variants within *ARSA*, *CSF1R* and *GRN*, along with candidate variants in *GRN* and *CHMP2B*. A further independent case-control study provided evidence of association between variants in *TREM2*, *APOE*, *ARSA*, *CSF1R*, *PSEN1* and *MAPT* and risk of AD [41]. Interestingly, the ADSP also identified a number of rare disease-associated variants within loci known to harbour common variants associated with sporadic AD, including *ABCA7* and *SORL1*. [48–51]. Bellenguez and colleagues [50] also observed that variants in *SORL1*, *ABCA7*, *TREM2* associated with AD. More specifically, the authors identified an exome-wide significant association between early onset AD risk and rare variants in all three genes. The authors estimated that the associated variants contributed equally to the heritability of early onset AD, and each explains between 1.1 and 1.5% of sporadic early onset AD heritability [50]. Further evidence for the role of *SORL1* in AD aetiology was provided by Holstege and colleagues [52], who observed that unique protein-truncating variants in *SORL1* occurred exclusively in a substantial proportion of AD cases. Variants in *ABCA7* have been identified to influence disease risk both across ethnic populations [15], and in an ethnic specific manner, with a frameshift deletion identified in African American and Caribbean Hispanic populations, but not a non-Hispanic White population [53].

Another novel gene with rare coding variants observed to segregate in an autosomal-dominant way with AD is *UNC5C* [54]. Wetzel-Smith et al. proposed that the variants in the gene could contribute to developing the disease by increasing susceptibility to neuronal cell death in vulnerable regions of the brain in patients with AD. Based on a family-based study and further replication, Cruchaga et al. observed that low frequency variants in *PLD3* were enriched in individuals with AD compared to healthy controls. Furthermore, the authors also showed that *PLD3* was involved in amyloid-β precursor protein processing and was overexpressed in brain tissue from patients with AD [55]. However, at present, the statistical evidence for association with sporadic disease at these gene is not robust, with *PLD3* seeming to have a greater influence on familial forms of disease [32, 56–58].

Other notable studies identifying rare variant associations for sporadic AD include Jakobsdottir et al. who detected an association in the *TM2D3* gene. Work in Drosophila suggest that the damaging effect of this variant is through the β-amyloid cascade [59]. Kunkle et al. used WES in a mixed ethnic population to search for rare variants leading to sporadic early onset AD. The authors observed associations with missense variants in *PSD2*, *TCIRG1*, *RIN3* and *RUFY1*. Interestingly, these genes function in clearance of cellular debris and unwanted proteins, including Aβ, through the endolysosomal transport pathway [51]. Le Guennec et al. studied sporaric early onset AD and found a rare recurrent microduplication, affecting the 17q21.31 locus (including the *CRHR1*, *MAPT*, *STH* and *KANSL1* genes) in four cases but not in healthy controls. An increase in *MAPT* (encoding tau protein) expression was observed in the affected individuals, and neuroimaging and cerebrospinal fluid biomarker profiles suggest the primary role of *MAPT* in disease development [60].

The majority of studies testing rare variation in sporadic AD have focused on identifying association with disease development. Ridge and colleagues adopted an innovative, pedigree-based approach to identify genetic variation that segregate with AD resilience. They studied "AD-resilient" individuals who had the high-risk *APOE* ε4 allele and were above 75 years of age without any signs of cognitive decline assessed clinically and compared them with relatives who developed AD. The rs142787485 variant in *RAB10* was shown to significantly associate with "AD-resilience", which replicated in an independent sample. Furthermore, in cell models, the knockdown of *RAB10* led to a statistically significant decrease in Aβ42 and Aβ42/Aβ40 ratio [53].

It must be noted that these associations have largely been observed in Caucasian populations, However, the majority of non-Caucasian studies are underpowered [61]. The disparity of results in non-Caucasian studies of *TREM2* [26, 27, 36•, 38, 42–47], the work of Kunkle et al. [51] (detailed above), the cross-population and population-specific associations seen at *ABCA7* [62, 63] and the identification of putative association at the *AKAP9* gene in an African American cohort [64] emphasise the ethnic specific genetic aetiology of AD and the need for further research in this area.

Methods for Rare Variant Discovery

The gold standard for rare variant discovery remains WGS, assaying every base in the genome. WGS allows the analysis of the full range of genomic modifications including pathogenic variants, structural variants and variants in non-coding regulatory regions [65–68]. Additionally, WGS is the superior method for covering difficult genomic regions including those with high GC content due to its PCR-free sequencing protocol. However, given the nature of rare variation, potentially being seen once in less than 1000 individuals, the sample sizes required to achieve statistical power for association analysis make this method of data generation economically prohibitive. The majority of the known risk loci were identified by genome-wide genotyping microarrays, including low frequency variants *PLCG2* and *ABI3*. These arrays only assay known genetic variation and despite imputation accuracy down to MAF = 0.008 when using the latest reference panels, a large proportion of low and rare frequency variants do not genotype or impute well on such arrays.

An alternative, and intermediate experiment between WGS and GWAS is WES. This method assays bases in the protein coding regions of the genome (the exome), meaning that any identified associations are likely to have an understandable functional effect. The exome makes up about 1% of the human genome, making WES a cheaper and popular alternative to WGS for both exonic and splice-site [69] rare variant detection. It has been estimated that the exome harbours about 85% of mutations with large effects on disease-related traits [70]. Exome sequencing studies have brought to light the importance of rare coding variants in complex genetic traits that were undetectable by GWAS [71]. Being a comprehensive approach, exome sequencing also provides direct identification of the casual variants, both common and rare, without the use of linkage disequilibrium to impute genotypes, as routinely performed with GWAS data. Exome sequencing was especially successful in the identification of Mendelian disease genes. This is reflected in almost 2000 new entries in OMIM since 2008 describing the genetic basis of a certain phenotype [70]. Therefore, the three primary advantages of exome sequencing over other rare variant detection methods are; the high potential to identify genes responsible for complex traits, readily available functional annotation of coding variants and the cost-effectiveness of WES compared to the WGS.

Several platforms for human exome capture are on the market [69, 72, 73]. The question of which of these platforms is best for a given application remains unanswered, as with any experimental technique, there are both strengths and limitations. The major difference between these platforms corresponds to the number of genes targeted, the probe/bait lengths, probe/bait density and sequencing coverage. There is also some difference in capture efficacy performance (including specificity, uniformity and sensitivity), technological reproducibility, DNA input requirement and cost effectiveness of each platforms. A comprehensive comparison of all the commercially available human WES platforms is beyond the scope of this review and has been performed elsewhere [69, 72, 73]. It has to be said that none of the capture technologies are able to cover all of the exons of the Consensus CDS, RefSeq or Ensembl databases.

Challenges of Rare Variant Identification

Rare variant identification has proved challenging in common disease; there are a number of factors to explain this, including the method of variant detection and the methods of data merging. Microarrays assay only known variation meaning that a substantial proportion of rare risk loci are potentially missed. Additionally, calling and clustering of rare variants via standard GWAS techniques have proved to be problematic and labour intensive, with much of the Illumina exome chip content requiring visual inspection [36•]. WGS captures all bases in the genome, resulting in the generation of a large amount of data. This can prove both computationally challenging and difficult to interpret given a number of identified variants may lie outside regions of known functional relevance. Often WGS data are filtered to include only functionally relevant variation, such as protein-coding and splice site-specific variation, to reduce such burdens.

The majority of rare variant detection studies utilise, at least in part, WES. Assembly of WES data within an

experiment is now relatively standardised, with specifically designed software, quality control and analysis pipelines [74]. However, as evidenced from meta-analysis of GWAS [7, 75], much of the power of genetic analysis of complex traits is gained through combination of data from multiple independent experiments. This can prove problematic for WES experiments where different capture technologies have been utilised. None of the capture technologies available are able to cover all of the exons of the Consensus CDS, RefSeq or Ensembl databases. Of the four commercially available human capture kits on the market (NimbleGen, Agilent, TruSeq and Nextera), only 26.2 Mb of the total targeted bases overlap, equating to around a 1/3 of the total targeted bases per kit (NimbleGen targets 64.1 Mb, Agilent targets 51.1 Mb, TruSeq and Nextera target 62.08 Mb). Therefore, combination of data generated via differing capture kits can result in a significant loss of target bases and subsequent lack of analysis of potential disease-related mutations. An additional technical issue with combination of WES data is the differing base coverage achieved by each capture kit. For the 26 Mb target regions, common to all four technologies, Agilent detected the highest number of variants followed by TruSeq, Nextera and NimbleGen [72]. Additionally, the areas of optimal coverage differ by capture technology, with the largest number of Illumina variants detected in the untranslated regions compared to NimbleGen detecting the highest number of variants in the Ensembl regions [72]. These findings emphasise the importance of sequence capture uniformity and capture probe performance, which eventually determine the amount of raw sequence data available for downstream data analysis. Ideally, all studies would use the same capture technology to allow merging of raw data for optimal rare variant detection.

Another intricacy, specifically in the analysis of rare variants, is that by definition, rare variants are not frequent and therefore association tests of individual variants is challenging [76]. The typical GWAS of common variants strategy is analysis of one variant at a time. Such analysis will be largely underpowered for rare variant detection unless the variant effect size or the sample size of the cohort is very large. This is why a number of methods have been developed to analyse multiple rare variants collapsed together thus increasing the statistical power [77–80]. An exhaustive review of how to design and analyse data based on rare variants is beyond the scope of this manuscript and is specific to the study design and technology utilised. A number of conceptual frameworks for the design of rare variant association studies have been published [76, 81, 82]. Rare variant analyses, whether at the single variant or gene-wide level, require large sample sizes to provide the required statistical power for the genetic association analyses. Zuk et al. have shown that the analysis of common variant and rare variant studies requires similarly large sample collections. In particular, a well-powered rare variant association study should involve discovery sets with at least 25,000 cases, together with a substantial replication set [76].

Benefit of Rare Variant Discovery

The single nucleotide rare variant associations identified are protein coding, meaning that the effect of the amino acid change can be easily modelled in silico and via cellular and animal models. This allows for a much quicker translation from genetics to a functional outcome and can be utilised in efforts to validate therapeutic targets [83]. Molecular dynamic (MD) modelling is an in silico technique that is gaining momentum in its use to understand rare variants in many diseases [84–88]. One example of the use of both cellular and in silico models to further our understanding of the impact rare variants have on AD is that of the well-studied TREM2 coding changes which have been subject of a number of in vitro, in vivo and in silico models [89–91]. Homozygous mutations within TREM2 leading to complete loss of function are a known cause of Nasu Hakola syndrome [92], which includes symptoms of frontotemporal dementia [93]. The identified AD-risk variants are thought to result in partial loss of function. The publication of purified proteins including that of TREM2 [91, 94], allows for in silico mutational studies into the impact of the variants on the protein structure and by inference, its function. Indeed, we and others have successfully modelled the TREM2 rare variants in silico with interesting results [91, 95]. Using MD simulations, we were able to conclude that the binding differences observed in vitro between the two genome wide significant mutations, R47H and R62H, could be attributed to a different structural alteration in their binding loops (Fig. 1) [95]. TREM2 has been shown to bind to apolipoproteins, including both APOE and CLU/APOJ and subsequently is involved in the uptake of Aβ by microglia. This may indicate that the different changes to the binding loop can be attributed to the variants differing genetic risk, and the differing binding rates as shown in vitro [96].

Alternate Methods

It is becoming ever clearer that complex traits require more sophisticated data analysis methods to unpick the multifactorial aetiology of disease onset and progression. We know that AD is a polygenic disorder [6], simultaneous assessment of common and rare (i.e. polygenic and monogenic) models can be used to provide additional information about disease genetic architecture. This approach has been fruitful in studying blood lipid levels and neurodevelopmental disorders in large number of individuals [82, 97].

Instead of focusing on crude association studies, there are other innovative approaches that could provide additional

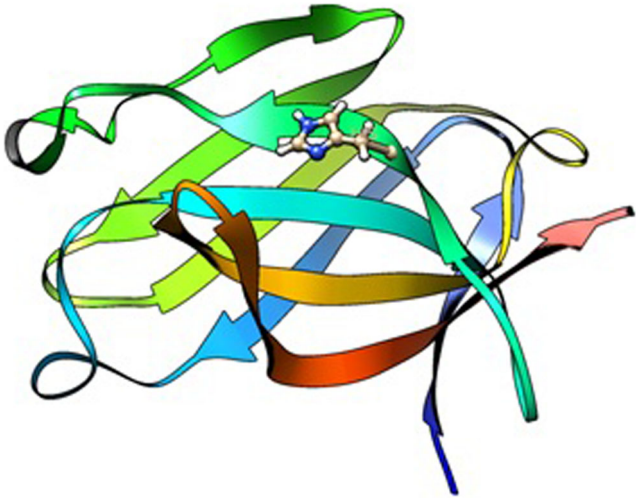

Fig. 1 The binding domain of the TREM2 protein, the R62H rare variant is seen in a stick all ball model format

information while studying rare variants. Some of the ways of exploring the data, include using other biological information, including gene expression (as reviewed by Verheijen and Sleegers [98]), methylation and biological pathways [99–101], in combination with genetic association data, to boost the statistical power of the analyses. To boost the statistical power of genetic association analyses, Ho et al. proposed a novel weight-adjustment approach to combine gene expression, methylation, transcriptional regulation and protein abundance information into rare variant analysis. Simulation studies have suggested that incorporating together such rich data can lead to substantial gain in statistical power. This integrative approach was applied successfully to find susceptibility variants in genes associated with blood pressure regulation [78, 102]. Furthermore, a number of studies have successfully used similar methods to integrate GWAS data with biological networks data (protein-protein interaction and co-expression networks) to predict causal genes at associated GWAS loci for various disorders [103–106]. Such integrative approaches, albeit currently focused more on analysis of common variants, have proved successful in studies of AD [107, 108].

To interrogate data from transcriptome-wide association studies (TWAS) studies, a TWAS hub was recently developed (http://twas-hub.org). The hub allows searchable access to TWAS results from hundreds of complex traits and dozens of expression studies based on the methodology described initially in Gusev et al. [109].

To better understand the pathobiology of disease, another way forward is to study a small number of carefully selected families with multiple affected individuals and with strong family history. This analysis is likely to be successful given the risk variants are likely to have larger effect sizes than GWAS loci. In addition, because such variants are likely to be coding, it is easier to subsequently functionally characterise, and to develop cellular and animal models. Such approaches have been reviewed previously in Lord et al. [110].

In a similar vein, another study approach that has proved successful in finding novel risk genes for AD is to focus on early onset sporadic AD rather than late onset sporadic AD exemplified by Kunkle et al. and Nicolas et al. [51]. These studies focus on extreme phenotypes, likely to be enriched for rarer variants with moderate effect sizes. As sample sizes grow, identification of disease-modifying genetics-utilising cohorts with deeply phenotyped data is likely to prove fruitful to understand more about the genetics of disease progression. Individuals with AD experience a range of non-cognitive symptoms that are distinct to each individual with disease [111].

Another approach recently utilised to identify common risk variation is to analyse large data sets such as the UK Biobank. Although most of the participants in these types of studies are too young to be diagnosed with AD, it is possible to study the disorder using family history data via a diagnosis by family history design [21•, 112•]. Currently, the UK Biobank data only include a limited number of accurate rare variant genotype data. However, there are plans for the UK Biobank cohort to be sequenced and the data to be made available to the scientific community [113]. Further initiatives such as the Genomics England project and studies based on data from electronic health records could provide further opportunities to mine large sample sets of data [114]. A recent review discusses the available resources and the statistical challenges with respect to analysing such data [115]. In the UK, the newly announced Digital Innovation Hub Programme by Health Data Research UK (with the help of MRC) aims to build towards a national hub to connect health-related data for research across populations of between three to five million people [116]. A word of caution with respect to using primary health and longitudinal cohort data is the potential overlap of sample sets across multiple studies, which could lead to false-positive observations, and the continued requirement for independent replication of new loci in similar sized cohorts.

Conclusions

Genetic heritability of sporadic AD is accounted for by both common and rare genetic variation. Here, we describe the established and putative rare AD risk variation identified by the field to date. We note that the rare variants identified contribute to disease susceptibility with larger effect sizes than generally seen with common risk variation and result in protein coding changes that can be easily modelled in silico, in vitro and in vivo. The identification of both common and rare disease-associated variants loci, including the *SORL1* and *ABCA7* genes, suggests that a number of the AD-associated genes may influence disease susceptibility via multiple mechanisms.

GWAS in other complex traits suggests that more powerful GWAS will identify further additional common and low frequency associations [25]. While, collaborative WES and WGS will undoubtedly unearth a significant number of rare variants that influence disease risk. As discussed, there are a number of issues that will need to be addressed to achieve this goal, primarily combining the differing sequence capture technologies, and adequately accessing the whole exome/genome. Further efforts by the IGAP, the European AD Biobank (EADB), AD European Sequencing (ADES) and ADSP among others are already underway. Ultimately, the future lies in WGS projects that will allow the detection and testing of the full range of genomic variation (including large structural alterations) with disease status, and this study design is being utilised for a range of rare diseases in projects such as Genomics England. Unfortunately, for complex traits that rely on large sample sizes to achieve the necessary statistical power, this is still beyond our reach.

The rare variants shown to associate with sporadic AD include potentially damaging and loss of function mutations, suggesting that careful assessment has to be considered for clinical practice and patient feedback along with the already established variation in *APOE*, *PSEN1*, *PSEN2* and *APP* [117]. Unquestionably, we are only at the beginning of our journey to identify rare protein-coding changes associated with disease. Already, the disease-associated protein-coded changes detected provide a greater understanding of the specific mechanism underlying disease risk as compared with common non-coding genetic risk factors and are likely to allow the expedited development of therapeutics.

Acknowledgements Cardiff University is supported by the UK DRI, MRC, ABBUK, ARUK, Welsh Government, Dementia Platform UK, Innovate UK and Moondance Foundation. Georgina Menzies is supported by the Ser Cymru II programme which is part funded by CardiffUniversity and the European Regional Development fund through the Welsh Government. Detelina Grozeva is funded by DPUK. We would like to thank our research participants and our extensive list of collaborators, without whom genetic discovery would not be achievable.

Compliance with Ethical Standards

Conflict of Interest Detelina Grozeva, Salha Saad, Georgina E. Menzies and Rebecca Sims each declare no potential conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- Of importance

1. Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science. 2002;297(5580):353–6.
2. Jagust W. Imaging the evolution and pathophysiology of Alzheimer disease. Nat Rev Neurosci. 2018;19:687–700.
3. Williams J, Hill M, Sims R. Decoding Alzheimer's disease. Nat Neurosci Rev. Manuscript submitted for publication.
4. Wingo TS, Lah JJ, Levey AI, Cutler DJ. Autosomal recessive causes likely in early-onset Alzheimer disease. Arch Neurol. 2012;69(1):59–64.
5. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168–74.
6. Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain. 2015;138(Pt 12):3673–84.
7. Kunkle BW, Grenier-Boley B, Sims R, Bis J, Naj AC, Boland A, et al. Meta-analysis of genetic association with diagnosed Alzheimer's disease identifies novel risk loci and implicates Abeta, Tau, immunity and lipid processing [preprint]. BioRxiv. https://www.biorxiv.org/content/early/2018/04/05/294629. Accessed 23 Jan 2019
8. Lee SH, Harold D, Nyholt DR, Goddard ME, Zondervan KT, Williams J, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet. 2013;22(4):832–41.
9. Saunders AM, Schmader K, Breitner JC, Benson MD, Brown WT, Goldfarb L, et al. Apolipoprotein E epsilon 4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases. Lancet. 1993;342(8873):710–1.
10. Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC, et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat Genet. 1994;7(2):180–4.
11. Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci U S A. 1993;90(5):1977–81.
12. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009;41(10):1088–93.
13. Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 2009;41(10):1094–9.
14. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA. 2010;303(18):1832–40.
15. Naj AC, Jun G, Reitz C, Kunkle BW, Perry W, Park YS, et al. Effects of multiple genetic loci on age at onset in late-onset Alzheimer disease: a genome-wide association study. JAMA Neurol. 2014;71(11):1394–404.

16. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet. 2011;43(5):429–35.
17. Escott-Price V, Bellenguez C, Wang LS, Choi SH, Harold D, Jones L, et al. Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease. PLoS One. 2014;9(6):e94661.
18. Baker E, Sims R, Leonenko G, Frizzati A, Harwood J, Grozeva D, et al. Gene based analysis in HRC imputed genome wide association data identifies three novel genes for Alzheimer's disease [preprint]. BioRxiv. 2018. https://www.biorxiv.org/content/early/2018/07/23/374876. Accessed 23 Jan 2019.
19. Ruiz A, Heilmann S, Becker T, Hernández I, Wagner H, Thelen M, et al. Follow-up of loci from the International Genomics of Alzheimer's Disease Project identifies TRIP4 as a novel susceptibility gene. Transl Psychiatry. 2014;4:e358.
20. Jun GR, Chung J, Mez J, Barber R, Beecham GW, Bennett DA, et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. Alzheimers Dement. 2017;13(7):727–38.
21.• Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry. 2018;8(1):99 **This manuscript is one of three to identify novel risk loci for AD using an innovative study design of family history by proxy in the large UK biobank cohort. While this does introduce noise with regards to misdiagnosis it also increases power. This study design could be utilized in other large population cohorts**.
22.• Jansen I, Savage J, Watanabe K, Bryois J, Williams D, Steinberg S, et al. Genetic meta-analysis identifies 9 novel loci and functional pathways for Alzheimers disease risk [preprint]. BioRxiv. https://www.biorxiv.org/content/early/2018/02/22/258533. Accessed 23 Jan 2019. **This manuscript is one of three to identify novel risk loci for AD using an innovative study design of family history by proxy in the large UK biobank cohort. While this does introduce noise with regards to misdiagnosis it also increases power. This study design could be utilized in other large population cohorts.**
23. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol. 2011;35(5):310–7.
24. Ridge PG, Mukherjee S, Crane PK, Kauwe JS, AsDG C. Alzheimer's disease: analyzing the missing heritability. PLoS One. 2013;8(11):e79771.
25. Gormley P, Kurki MI, Hiekkala ME, Veerapen K, Häppölä P, Mitchell AA, et al. Common variant burden contributes to the familial aggregation of migraine in 1,589 families. Neuron. 2018;98(4):743–53.e4.
26. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. N Engl J Med. 2013;368(2):107–16.
27. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013;368(2):117–27.
28. Benitez BA, Cooper B, Pastor P, Jin SC, Lorenzo E, Cervantes S, et al. TREM2 is associated with the risk of Alzheimer's disease in Spanish population. Neurobiol Aging. 2013;34(6):1711.e15–7.
29. Finelli D, Rollinson S, Harris J, Jones M, Richardson A, Gerhard A, et al. TREM2 analysis and increased risk of Alzheimer's disease. Neurobiol Aging. 2015;36(1):546.e9–13.
30. Ghani M, Sato C, Kakhki EG, Gibbs JR, Traynor B, St George-Hyslop P, et al. Mutation analysis of the MS4A and TREM gene clusters in a case-control Alzheimer's disease data set. Neurobiol Aging. 2016;42:217.e7–e13.
31. Gonzalez Murcia JD, Schmutz C, Munger C, Perkes A, Gustin A, Peterson M, et al. Assessment of TREM2 rs75932628 association with Alzheimer's disease in a population-based sample: the Cache County Study. Neurobiol Aging. 2013;34(12):2889.e11–3.
32. Hooli BV, Lill CM, Mullin K, Qiao D, Lange C, Bertram L, et al. PLD3 gene variants and Alzheimer's disease. Nature. 2015;520(7545):E7–8.
33. Pottier C, Wallon D, Rousseau S, Rovelet-Lecrux A, Richard AC, Rollin-Sillaire A, et al. TREM2 R47H variant as a risk factor for early-onset Alzheimer's disease. J Alzheimers Dis. 2013;35(1):45–9.
34. Rosenthal SL, Bamne MN, Wang X, Berman S, Snitz BE, Klunk WE, et al. More evidence for association of a rare TREM2 mutation (R47H) with Alzheimer's disease risk. Neurobiol Aging. 2015;36(8):2443.e21–6.
35. Ruiz A, Dols-Icardo O, Bullido MJ, Pastor P, Rodríguez-Rodríguez E, López de Munain A, et al. Assessing the role of the TREM2 p.R47H variant as a risk factor for Alzheimer's disease and frontotemporal dementia. Neurobiol Aging. 2014;35(2):444.e1–4.
36.• Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet. 2017;49(9):1373–84 **This manuscript used a genome-wide association study design to identify rare susceptibility genes for disease and successfully identified new therapeutically targetable protein-coding changes**.
37. Slattery CF, Beck JA, Harper L, Adamson G, Abdi Z, Uphill J, et al. R47H TREM2 variant increases risk of typical early-onset Alzheimer's disease but not of prion or frontotemporal dementia. Alzheimers Dement. 2014;10(6):602–8.e4.
38. Jin SC, Carrasquillo MM, Benitez BA, Skorupa T, Carrell D, Patel D, et al. TREM2 is associated with increased risk for Alzheimer's disease in African Americans. Mol Neurodegener. 2015;10:19.
39. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, et al. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature. 2012;488(7409):96–9.
40.• Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2018. https://doi.org/10.1038/s41380-018-0112-7. **This manuscript is the largest next generation sequencing experiment to date in AD and identifies novel susceptibility variants for disease and provides further evidence for the role of immunity in AD.**
41. Blue EE, Bis JC, Dorschner MO, Tsuang DW, Barral SM, Beecham G, et al. Genetic variation in genes underlying diverse dementias may explain a small proportion of cases in the Alzheimer's disease sequencing project. Dement Geriatr Cogn Disord. 2018;45(1–2):1–17.
42. Jin SC, Benitez BA, Karch CM, Cooper B, Skorupa T, Carrell D, et al. Coding variants in TREM2 increase risk for Alzheimer's disease. Hum Mol Genet. 2014;23(21):5838–46.
43. Jiang T, Tan L, Chen Q, Tan MS, Zhou JS, Zhu XC, et al. A rare coding variant in TREM2 increases risk for Alzheimer's disease in Han Chinese. Neurobiol Aging. 2016;42:217.e1–3.
44. Jiao B, Liu X, Tang B, Hou L, Zhou L, Zhang F, et al. Investigation of TREM2, PLD3, and UNC5C variants in patients with Alzheimer's disease from mainland China. Neurobiol Aging. 2014;35(10):2422.e9–e11.
45. Ma J, Zhou Y, Xu J, Liu X, Wang Y, Deng Y, et al. Association study of TREM2 polymorphism rs75932628 with late-onset Alzheimer's disease in Chinese Han population. Neurol Res. 2014;36(10):894–6.
46. Wang P, Guo Q, Zhou Y, Chen K, Xu Y, Ding D, et al. Lack of association between triggering receptor expressed on myeloid cells 2 polymorphism rs75932628 and late-onset Alzheimer's disease in a Chinese Han population. Psychiatr Genet. 2018;28(1):16–8.
47. Yu JT, Jiang T, Wang YL, Wang HF, Zhang W, Hu N, et al. Triggering receptor expressed on myeloid cells 2 variant is rare

in late-onset Alzheimer's disease in Han Chinese individuals. Neurobiol Aging. 2014;35(4):937.e1–3.
48. Louwersheimer E, Ramirez A, Cruchaga C, Becker T, Kornhuber J, Peters O, et al. Influence of genetic variants in SORL1 gene on the manifestation of Alzheimer's disease. Neurobiol Aging. 2015;36(3):1605.e13–20.
49. Steinberg S, Stefansson H, Jonsson T, Johannsdottir H, Ingason A, Helgason H, et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat Genet. 2015;47(5):445–7.
50. Bellenguez C, Charbonnier C, Grenier-Boley B, Quenez O, Le Guennec K, Nicolas G, et al. Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. Neurobiol Aging. 2017;59:220.e1–9.
51. Kunkle BW, Vardarajan BN, Naj AC, Whitehead PL, Rolati S, Slifer S, et al. Early-onset Alzheimer disease and candidate risk genes involved in Endolysosomal transport. JAMA Neurol. 2017;74(9):1113–22.
52. Holstege H, van der Lee SJ, Hulsman M, Wong TH, van Rooij JG, Weiss M, et al. Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy. Eur J Hum Genet. 2017;25(8):973–81.
53. Ridge PG, Karch CM, Hsu S, Arano I, Teerlink CC, Ebbert MTW, et al. Linkage, whole genome sequence, and biological data implicate variants in RAB10 in Alzheimer's disease resilience. Genome Med. 2017;9(1):100.
54. Wetzel-Smith MK, Hunkapiller J, Bhangale TR, Srinivasan K, Maloney JA, Atwal JK, et al. A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death. Nat Med. 2014;20(12):1452–7.
55. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature. 2014;505(7484):550–4.
56. Heilmann S, Drichel D, Clarimon J, Fernández V, Lacour A, Wagner H, et al. PLD3 in non-familial Alzheimer's disease. Nature. 2015;520(7545):E3–5.
57. Lambert JC, Grenier-Boley B, Bellenguez C, Pasquier F, Campion D, Dartigues JF, et al. PLD3 and sporadic Alzheimer's disease risk. Nature. 2015;520(7545):E1.
58. van der Lee SJ, Holstege H, Wong TH, Jakobsdottir J, Bis JC, Chouraki V, et al. PLD3 variants in population studies. Nature. 2015;520(7545):E2–3.
59. Jakobsdottir J, van der Lee SJ, Bis JC, Chouraki V, Li-Kroeger D, Yamamoto S, et al. Rare functional variant in TM2D3 is associated with late-onset Alzheimer's disease. PLoS Genet. 2016;12(10):e1006327.
60. Le Guennec K, Quenez O, Nicolas G, Wallon D, Rousseau S, Richard AC, et al. 17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression. Mol Psychiatry. 2017;22(8):1119–25.
61. Miyashita A, Wen Y, Kitamura N, Matsubara E, Kawarabayashi T, Shoji M, et al. Lack of genetic association between TREM2 and late-onset Alzheimer's disease in a Japanese population. J Alzheimers Dis. 2014;41(4):1031–8.
62. Cukier HN, Kunkle BW, Vardarajan BN, Rolati S, Hamilton-Nelson KL, Kohli MA, et al. ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. Neurol Genet. 2016;2(3):e79.
63. Reitz C, Jun G, Naj A, Rajbhandary R, Vardarajan BN, Wang LS, et al. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E ϵ4, and the risk of late-onset Alzheimer disease in African Americans. JAMA. 2013;309(14):1483–92.
64. Logue MW, Schu M, Vardarajan BN, Farrell J, Bennett DA, Buxbaum JD, et al. Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. Alzheimers Dement. 2014;10(6):609–18.e11.
65. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72.
66. Carss KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, et al. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. Am J Hum Genet. 2017;100(1):75–90.
67. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. Hum Mutat. 2015;36(8):815–22.
68. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am J Hum Genet. 2016;98(1):58–74.
69. Asan XY, Jiang H, Tyler-Smith C, Xue Y, Jiang T, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol. 2011;12(9):R95.
70. Majewski J, Wang Z, Lopez I, Al Humaid S, Ren H, Racine J, et al. A new ocular phenotype associated with an unexpected but known systemic disorder and mutation: novel use of genomic diagnostics and exome sequencing. J Med Genet. 2011;48(9):593–6.
71. Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, et al. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. 2013;9(2):e1003301.
72. Chilamakuri CS, Lorenz S, Madoui MA, Vodák D, Sun J, Hovig E, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics. 2014;15:449.
73. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. Genome Biol. 2011;12(9):R97.
74. Naj AC, Lin H, Vardarajan BN, White S, Lancour D, Ma Y, et al. Quality control and integration of genotypes from two calling pipelines for whole genome sequence data in the Alzheimer's disease sequencing project. Genomics. 2018. https://doi.org/10.1016/j.ygeno.2018.05.004.
75. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452–8.
76. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111(4):E455–64.
77. Santorico SA, Hendricks AE. Progress in methods for rare variant association. BMC Genet. 2016;17 Suppl 2:6.
78. Ho YY, Guan W, O'Connell M, Basu S. Powerful association test combining rare variant and gene expression using family data from genetic analysis workshop 19. BMC Proc. 2016;10(Suppl 7):251–5. https://doi.org/10.1186/s12919-016-0039-4
79. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009;33(6):497–507.
80. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6):841–53.
81. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44(6):623–30.
82. Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat Commun. 2018;9(1):3391.
83. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18(1):77.

84. Cubellis MV, Baaden M, Andreotti G. Taming molecular flexibility to tackle rare diseases. Biochimie. 2015;113:54–8.
85. Zimmermann MT, Urrutia R, Oliver GR, Blackburn PR, Cousin MA, Bozeck NJ, et al. Molecular modeling and molecular dynamic simulation of the effects of variants in the TGFBR2 kinase domain as a paradigm for interpretation of variants obtained by next generation sequencing. PLoS One. 2017;12(2):e0170822.
86. Jatana N, Thukral L, Latha N. Structural signatures of DRD4 mutants revealed using molecular dynamics simulations: implications for drug targeting. J Mol Model. 2016;22(1):14.
87. Singh G, MSK J, Sharma R, Bhat B, Madhusudhan C, Singh A. Structural, functional and molecular dynamics analysis of cathepsin B gene SNPs associated with tropical calcific pancreatitis, a rare disease of tropics.
88. Padhi A, Gomes J. A molecular dynamics based investigation reveals the role of rare ribonuclease 4 variants in amyotrophic lateral sclerosis susceptibility. Mutat Res. 2019;813:1–12. https://doi.org/10.1016/j.mrfmmm.2018.11.002
89. Song W, Hooli B, Mullin K, Jin SC, Cella M, Ulland TK, et al. Alzheimer's disease-associated TREM2 variants exhibit either decreased or increased ligand-dependent activation. Alzheimers Dement. 2017;13(4):381–7.
90. Wang Y, Cella M, Mallinson K, Ulrich JD, Young KL, Robinette ML, et al. TREM2 lipid sensing sustains the microglial response in an Alzheimer's disease model. Cell. 2015;160(6):1061–71.
91. Kober DL, Alexander-Brett JM, Karch CM, Cruchaga C, Colonna M, Holtzman MJ, Brett TJ Neurodegenerative disease mutations in TREM2 reveal a functional surface and distinct loss-of-function mechanisms. Elife. 2016;5.
92. Dardiotis E, Siokas V, Pantazi E, Dardioti M, Rikos D, Xiromerisiou G, et al. A novel mutation in TREM2 gene causing Nasu-Hakola disease and review of the literature. Neurobiol Aging. 2017;53:194.e13–22. https://doi.org/10.1016/j.neurobiolaging.2017.01.015
93. Cuyvers E, Bettens K, Philtjens S, Van Langenhove T, Gijselinck I, van der Zee J, et al. Investigating the role of rare heterozygous TREM2 variants in Alzheimer's disease and frontotemporal dementia. Neurobiol Aging. 2014;35(3):726.e11–9.
94. Kober DL, Wanhainen KM, Johnson BM, Randolph DT, Holtzman MJ, Brett TJ. Preparation, crystallization, and preliminary crystallographic analysis of wild-type and mutant human TREM-2 ectodomains linked to neurodegenerative and inflammatory diseases. Protein Expr Purif. 2014;96:32–8.
95. Menzies G, Sims R, Williams J. Molecular dynamics simulations of Alzheimer's variants, R47H and R62H, in TREM2 provide evidence for structural alterations behind functional changes. 2018. Manuscript submitted for publication.
96. Yeh FL, Wang Y, Tom I, Gonzalez LC, Sheng M. TREM2 binds to apolipoproteins, including APOE and CLU/APOJ, and thereby facilitates uptake of amyloid-Beta by microglia. Neuron. 2016;91(2):328–40.
97. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;562(7726):268–71.
98. Verheijen J, Sleegers K. Understanding Alzheimer disease at the interface between genetics and transcriptomics. Trends Genet. 2018;34(6):434–47.
99. Richardson TG, Timpson NJ, Campbell C, Gaunt TR. A pathway-centric approach to rare variant association analysis. Eur J Hum Genet. 2016;25(1):123–9.
100. Kao PY, Leung KH, Chan LW, Yip SP, Yap MK. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. Biochim Biophys Acta Gen Subj. 2017;1861(2):335–53.
101. Petersen A, Sitarik A, Luedtke A, Powers S, Bekmetjev A, Tintle NL. Evaluating methods for combining rare variant data in pathway-based tests of genetic association. BMC Proc. 2011;5 Suppl 9:S48.
102. Ho YY, Baechler EC, Ortmann W, Behrens TW, Graham RR, Bhangale TR, et al. Using gene expression to improve the power of genome-wide association analysis. Hum Hered. 2014;78(2):94–103.
103. Jia P, Zhao Z. Network assisted analysis to prioritize GWAS results: principles, methods and perspectives. Hum Genet. 2014;133(2):125–38.
104. Huan T, Meng Q, Saleh MA, Norlander AE, Joehanes R, Zhu J, et al. Integrative network analysis reveals molecular mechanisms of blood pressure regulation. Mol Syst Biol. 2015;11(1):799.
105. Gustafsson M, Gawel DR, Alfredsson L, Baranzini S, Björkander J, Blomgran R, et al. A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. Sci Transl Med. 2015;7(313):313ra178.
106. Calabrese GM, Mesner LD, Stains JP, Tommasini SM, Horowitz MC, Rosen CJ, et al. Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. Cell Syst. 2017;4(1):46–59.e4.
107. Hu YS, Xin J, Hu Y, Zhang L, Wang J. Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach. Alzheimers Res Ther. 2017;9(1):29.
108. Tansey KE, Cameron D, Hill MJ. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. Genome Med. 2018;10(1):14.
109. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
110. Lord J, Lu AJ, Cruchaga C. Identification of rare variants in Alzheimer's disease. Front Genet. 2014;5:369.
111. Hollingworth P, Hamshere ML, Moskvina V, Dowzell K, Moore PJ, Foy C, et al. Four components describe behavioral symptoms in 1,120 individuals with late-onset Alzheimer's disease. J Am Geriatr Soc. 2006;54(9):1348–54.
112.• Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. Nat Genet. 2017;49(3):325–31 **This manuscript is one of three to identify novel risk loci for AD using an innovative study design of family history by proxy in the large UK biobank cohort. While this this does introduce noise with regards to misdiagnosis it also increases power. This study design could be utilized in other large population cohorts**.
113. UK Biobank. Whole genome sequencing will 'transform the research landscape for a wide range of diseases' 2018 [Available from: https://www.ukbiobank.ac.uk/2018/04/whole-genome-sequencing-will-transform-the-research-landscape-for-a-wide-range-of-diseases/. Accessed 23 Jan 2019.
114. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017;106(1):1–9.
115. Beesley L, Salvatore M, Fritsche L, Pandit A, Rao A, Brummett C, et al. The Emerging Landscape of Epidemiological Research Based on Biobanks Linked to Electronic Health Records: Existing Resources, Analytic Challenges and Potential Opportunities. 2018.
116. Medical Research Council. Industrial Strategy Challenge Fund Digital Innovation Hub Sprint Exemplar Innovation Projects [Available from: https://mrc.ukri.org/funding/browse/iscf-dih/digital-innovation-hub-exemplar-projects/?utm_medium=email&utm_source=govdelivery. Accessed 23 Jan 2019
117. Hsu S, Gordon BA, Hornbeck R, Norton JB, Levitch D, Louden A, et al. Discovery and validation of autosomal dominant Alzheimer's disease mutations. Alzheimers Res Ther. 2018;10(1):67.