Strategies for the intelligent integration of genetic variance information in multiscale models of neurodegenerative diseases

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)

der
Mathematisch-Naturwissenschaftlichen Fakultät

der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Mufssra Naz

aus
Lahore, Pakistan

Bonn 2019

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. rer. nat. Martin Hofmann-Apitius

2. Gutachter: Prof. Dr. Heiko Schoof

Tag der Promotion: 10. Juli 2019

Erscheinungsjahr: 2019

Abstract

A more complete understanding of the genetic architecture of complex traits and diseases can maximize the utility of human genetics in disease screening, diagnosis, prognosis, and therapy. Undoubtedly, the identification of genetic variants linked to polygenic and complex diseases is of supreme interest for clinicians, geneticists, patients, and the public. Furthermore, determining how genetic variants affect an individual's health and transmuting this knowledge into the development of new medicine can revolutionize the treatment of most common deleterious diseases. However, this requires the correlation of genetic variants with specific diseases, and accurate functional assessment of genetic variation in human DNA sequencing studies is still a nontrivial challenge in clinical genomics. Assigning functional consequences and clinical significances to genetic variants is an important step in human genome interpretation. The translation of the genetic variants into functional molecular mechanisms is essential in disease pathogenesis and, eventually in therapy design.

Although various statistical methods are helpful to short-list the genetic variants for fine-mapping investigation, demonstrating their role in molecular mechanism requires knowledge of functional consequences. This undoubtedly requires comprehensive investigation. Experimental interpretation of all the observed genetic variants is still impractical. Thus, the prediction of functional and regulatory consequences of the genetic variants using in-silico approaches is an important step in the discovery of clinically actionable knowledge. Since the interactions between phenotypes and genotypes are multi-layered and biologically complex. Such associations present

several challenges and simultaneously offer many opportunities to design new protocols for in-silico variant evaluation strategies.

This thesis presents a comprehensive protocol based on a causal reasoning algorithm that harvests and integrates multifaceted genetic and biomedical knowledge with various types of entities from several resources and repositories to understand how genetic variants perturb molecular interaction, and initiate a disease mechanism.

Firstly, as a case study of genetic susceptibility loci of Alzheimer's disease, I reviewed and summarized all the existing methodologies for Genome Wide Association Studies (GWAS) interpretation, currently available algorithms, and computable modelling approaches. In addition, I formulated a new approach for modelling and simulations of genetic regulatory networks as an extension of the syntax of the Biological Expression Language (OpenBEL). This could allow the representation of genetic variation information in cause-and-effect models to predict the functional consequences of disease-associated genetic variants. Secondly, by using the new syntax of OpenBEL, I generated an OpenBEL model for Alzheimer´s Disease (AD) together with genetic variants including their DNA, RNA or protein position, variant type and associated allele. To better understand the role of genetic variants in a disease context, I subsequently tried to predict the consequences of genetic variation based on the functional context provided by the network model. I further explained that how genetic variation information could help to identify candidate molecular mechanisms for aetiologically complex diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD). Though integration of genetic variation information can enhance the evidence base for shared pathophysiology pathways in complex diseases, I have addressed to one of the key

questions, namely the role of shared genetic variants to initiate shared molecular mechanisms between neurodegenerative diseases. I systematically analysed shared genetic variation information of AD and PD and mapped them to find shared molecular aetiology between neurodegenerative diseases.

My methodology highlighted that a comprehensive understanding of genetic variation needs integration and analysis of all omics data, in order to build a joint model to capture all datasets concurrently. Moreover genomic loci should be considered to investigate the effects of GWAS variants rather than an individual genetic variant, which is hard to predict in a biologically complex molecular mechanism, predominantly to investigate shared pathology.

Acknowledgement

"Praise be to Allah (God)"

Firstly, I would like to express my sincere gratitude to my supervisor Professor Dr. Martin Hofmann-Apitius, for the continuous support of my Ph.D. study and related research, for his patience, encouragement, and immense knowledge, without which I would not have reached here. His valuable guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D. study. Above all, thanks for improving my skills and making me the person that I am today.

Besides my supervisor, I would also like to thank Prof. Dr. Heiko Schoof for his valuable review and guidance. It is an honour to have him as the second supervisor for my thesis. My sincere gratitude also goes to Prof. Dr. Thomas Schultz and Prof. Dr. Diana Imhof for agreeing to be part of Ph.D. committee.

My special thanks goes to Christine A. Robinson for her valuable support in reviewing the PhD thesis. My sincere thanks also goes to all my colleagues from Fraunhofer SCAI for contributing to good scientific research work, thanks a lot for making me realize the value of working together as a team and giving me a new experience in a wonderful working environment, which challenges us every minute.

I am indebted to my parents for believing in me.

Last but not the least, I would like to thank my family and friends for always being supportive and encouraging throughout my studies.

Thanks for all your encouragement!

"*The laws of genetics apply even if you refuse to learn them*."

— Allison Plowden

Publications

1. **Naz M**, Hofmann-Apitius M.
   GWAS genetic variant data and their integration in the context of network biology.
   J Syst Integr Neurosci, 2016 Volume 2(4): 189-202
   doi: 10.15761/JSIN.1000135

2. **Naz M**; Kodamullil AT; Hofmann-Apitius M.
   Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases.
   Brief Bioinform. 2016 May; 17(3): 505-16.
   doi: 10.1093/bib/bbv063

3. **Naz M**, Younesi E, Hofmann-Apitius M.
   Systematic analysis of GWAS data reveals genomic hotspots for shared mechanisms between neurodegenerative diseases.
   J Alzheimers Dis Parkinsonism 2017, Vol 7(5): 368
   doi: 10.4172/2161-0460.1000368

4. Kodamullil AT, Younesi E, **Naz M**, Bagewadi S, Hofmann-Apitius M.
   Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis.
   Alzheimers Dement. 2015 Apr 4. pii:S1552-5260(15)00083-7.
   doi: 10.1016/j.jalz.2015.02.006.

5. Allen GI, Amoroso N, Anghel C, Balagurusamy V, Bare CJ, Beaton D, Bellotti R, Bennett DA, Boehme KL, Boutros PC, Caberlotto L, Caloian C, Campbell F, Chaibub Neto E, Chang YC, Chen B, Chen CY, Chien TY, Clark T, Das S, Davatzikos C, Deng J, Dillenberger D, Dobson RJ, Dong Q, Doshi J, Duma D, Errico R, Erus G, Everett E, Fardo DW, Friend SH, Fröhlich H, Gan J, St George-Hyslop P, Ghosh SS, Glaab E, Green RC, Guan Y, Hong MY, Huang C, Hwang J, Ibrahim J, Inglese P, Iyappan A, Jiang Q, Katsumata Y, Kauwe JS, Klein A, Kong D, Krause R, Lalonde E, Lauria M, Lee E, Lin X, Liu Z, Livingstone J, Logsdon BA, Lovestone S, Ma TW, Malhotra A, Mangravite LM, Maxwell TJ, Merrill E, Nagorski J, Namasivayam A, Narayan M, **Naz M**, Newhouse SJ, Norman TC, Nurtdinov RN, Oyang YJ, Pawitan Y, Peng S, Peters MA, Piccolo SR, Praveen P, Priami C, Sabelnykova VY, Senger P, Shen X, Simmons A, Sotiras A, Stolovitzky G, Tangaro S, Tateo A, Tung YA, Tustison NJ, Varol E, Vradenburg G, Weiner MW, Xiao G, Xie L, Xie Y, Xu J, Yang H, Zhan X, Zhou Y, Zhu F, Zhu H, Zhu S.
   Alzheimer's Disease Neuroimaging Initiative. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease.
   Alzheimers Dement. 2016 Jun;12(6):645-53.
   doi: 10.1016/j.jalz.2016.02.006.

6. Domingo-Fernández D, Kodamullil AT, Iyappan A, **Naz M**, Emon MA, Raschka T, Karki R, Springstubbe S, Ebeling C, Hofmann-Apitius M.
Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics. 2017Nov15;33(22):3679-3681.*
doi:10.1093/bioinformatics/btx399.

7. Kawalia SB, Raschka T, **Naz M**, Simoes RM, Senger P and Hofmann-Apitius M.
Analytical strategy to prioritize Alzheimer's disease candidate genes in gene regulatory networks using public expression data.
J Alzheimers Dis. 2017;59(4):1237-1254.
doi: 10.3233/JAD-170011.

Contents

List of Abbreviations

NDDs	Neurodegenerative Diseases
AD	Alzheimer's disease
PD	Parkinson's disease
SNP	Single-nucleotide polymorphism
SNVs	Single nucleotide variants
ESEs	Exon splicing enhancers
ESSs	Exon splicing silencers
TSS	Transcription start sites
eQTL	Expression quantitative trait loci
SBML	Systems Biology Markup Language
BioPAX	Pathway exchange language for Biological pathway data
BEL	Biological Expression Language
XML	Extensible markup language
HGNC	HUGO Gene Nomenclature Committee
RCR	Reverse Causal Reasoning
NPA	Network Perturbation Amplitude
ADRs	Adverse drug reactions
GWAS	Genome-wide association studies
WGS	Whole-genome sequencing
WES	Whole-exome sequencing
ROS	Religious Orders Study
MMSE	Mini-Mental State Examination
ADNI	Alzheimer's disease Neuroimaging Initiative
GERAD1	Genetic and Environmental Risk in AD Consortium 1
EADI	European Alzheimer's Disease Initiative sample
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology

Introduction

Human Genetic Variance Information

Genetic architecture describes the characteristics of genetic variation that are responsible for heritable phenotypic variability. Defining the genetic architecture of a complex trait or disease is central to the scientific and clinical goals of human genetics, which are to understand disease aetiology and aid in disease screening, diagnosis, prognosis, and therapy. Recent technological advances have enabled genome-wide association studies and emerging next-generation sequencing studies to begin to decipher the nature of the heritable contribution to traits and disease. Precisely, genetic architecture encompasses the number of variants participating in disease aetiology, the level of their functional impact on the disease, the variant's frequency in population and their interactions with each other and with environmental factors [1]. Thus, in contrast to the limited concept of heritability, which refers only to the impact of additive genetic effects on a complex disease [2], genetic architecture refers broadly to a comprehensive knowledge of all genetic contributions to a given phenotype or disease as well as understanding of the characteristics of this contribution [3].

Human genomic architectures can differ from one another at a single base position as single nucleotide variants (SNVs), or they can exhibit large structural modifications, like copy number variations, inversions and translocations [4]. In two randomly selected human genomes, 99.9% of the DNA sequence is identical. The remaining 0.1% of DNA sequence comprises variations. Single-nucleotide polymorphism (SNP) is the most common type of such variations. SNPs are less

mutable than other forms of variations, plentiful (high in frequency on the genome), and distributed throughout the genome, in coding regions as well as noncoding regions (promoter, intronic, intergenic and regulatory regions) (Figure 1). These variations are relevant to diversity in the population, individuality, susceptibility to diseases, and individual response to medicine. It has been proposed that genetic variants can be used for identification and mapping of complex and common diseases and also for homogeneity testing and pharmacogenetic studies.

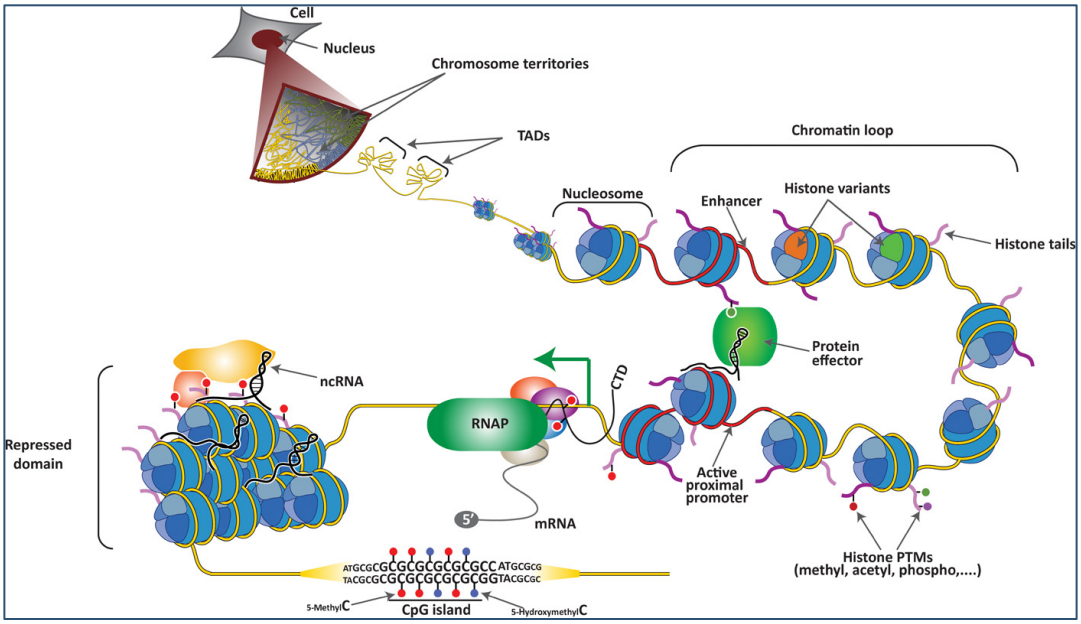


Figure 1: Hierarchical layers of chromatin organization Individual chromosomes cover a distinct region within the nucleus known as chromosome territory. At increasing resolution, chromosomes are composed of topologically associating domains (TADs), which are structural units defined by the high frequency of chromatin interactions between their loci that are partitioned by sharp boundaries. Within TADs, enhancer elements and active proximal promoters (both depicted in red) form chromatin loops, which are mediated and/or stabilized by protein effectors, noncoding RNAs (ncRNAs), and histone posttranslational modifications (PTMs). Enhancers and promoters are characterized by the presence of specific histone variants and PTMs on the histone tails. Upon transcription activation, elongating RNA polymerase II (RNAP, in green) is phosphorylated at Ser^5 and Ser^2 on its C-terminal domain (CTD) and begins to produce mRNA. Genomic regions that are transcriptionally silenced form repressed chromatin domains that are also stabilized by ncRNA and other repressive protein complexes. Finally, tracks of repetitive sequence are found in specific functional regions of the genome, including CpG islands (CGIs), in which cytosines can be modified (5-methylC and 5-hydroxymethylC). (taken from: Aranda S, et al. Sci Adv. 2015 Dec.)

Genetic variants may modify the encoded amino acids (non-synonymous) or can be silent (synonymous) or simply occur in the noncoding regions. They may influence gene regulation, promoter activity (gene expression), messenger RNA (mRNA) conformation (stability), and subcellular localization of mRNAs and/or proteins, regulation of miRNA and hence may produce disease. Therefore, identification of multiple variations in genes and analysis of their effects may lead to a better insight of their impact on gene function and health of an individual.

Genetics Architecture and Human Diseases

Genetic architecture is generally represented as monogenic, oligogenic or polygenic, meaning the contribution of one, few, or many genetic variants to phenotypic variability, respectively [5]. Moreover, a recent theoretical development in genetic architecture modelling has suggested that all complex diseases share a single 'omnigenic' architecture [6]. The 'omnigenic' model describes gene regulatory networks as adequately interconnected, to allow all expressed genes in a disease-relevant cell to participate in the disease progress. This model postulates that thousands of 'peripheral' or 'non-core' genes apply non-zero effects on essentially all downstream phenotypes [6]. This model can help to explain the complexity of genetic architecture, and it’s compatible with the polygenic or 'infinitesimal' model [7], in which all-genetic variants have a small but non-zero contribution in phenotypic variation. These broad yet comprehensive labels have been valuable in speculating on the nature of genetic architecture. Moreover modern technologies enable diverse data integration and mapping that can provide empirical evidence for the depiction of genetic architecture.

If, over an observable time frame, an alteration in DNA coding sequence always results in a specific disease expression, then the effect of the variant is considered to be highly penetrant. Moreover, that particular variant is referred to as a mutation if its allele frequency is less than 1%. During the previous century, various diseases have been identified as associated with relatively rare mutations, if either dominant (one) or recessive (two) variant copies are inherited.

In the current era, biomedical research is commencing to identify less rare gene variations that are linked with common diseases without direct causation. These are described as susceptibility polymorphisms. The association of the age of onset of Alzheimer's disease with APOE4 variant and the protective effect of the APOE2 variant are the most acknowledged and established examples of disease-susceptibility polymorphisms relationships [8,9].

Haplotypes (combinations of two or more variants) can also be linked with the expression of diseases. Many closely positioned polymorphisms, particularly within small (approximately 50–150 kb) DNA linkage disequilibrium (LD) regions, can distinguish a region of LD from other combinations. They can define the time in evolution when a new recombination event emerged in the vicinity of a disease susceptibility polymorphism. This hypothesis has already been established experimentally with the LD region of APOE4 polymorphism. By using a high-density SNP mapping analysis, APOE4 was re-discovered across a region of four million bases [10–12]. Identification of this small specific LD region limited the focus to only two genes, APOC-1 and APOE. Only the APOE4 variant at codon 112 was linked with the younger age of onset distribution of AD [13]. This outcome has been confirmed by multiple epidemiology studies in various populations with immune-

cytopathology of human brain, differences in control allele frequencies, protein expression in transgenic mice, analyses of neuronal APOE RNA and many others [14-16].

For many common diseases, rare mutations confer increased risk in heterozygous carriers. However, the vast majority of disease occurs without such mutations. Polygenic inheritance can also play a greater role than rare monogenic mutations, with multiple common genetic variants of small effect [7, 17-19].

Polygenic score is a quantitative metric based on the cumulative effect of multiple common polymorphisms, to measure a person's inherited risk. Weights are assigned to each genetic variant according to the strength of their association (effect estimate) with disease risk. An individual's score is measured based on how many risk alleles they have for each variant (i.e. 0, 1, or 2 alleles) in the polygenic score [20].

There are several computational algorithms in use to calculate the polygenic score of genetic variants. Bayesian approach based 'LDPred' is one of them, which computes a posterior mean effect size for every variant. In addition, the underlying Gaussian distribution determines the fraction of causal markers with a tuning parameter [20].

A second approach to calculate the polygenic score is 'pruning and thresholding'. It is built using a *p* value and a linkage disequilibrium-driven clustering procedure [21]. Concisely, the algorithm forms clusters around SNPs with associated *P* values within a provided threshold. Each cluster contains all SNPs in the region of 250 kb of the tag SNP as identified by a given pairwise correlation (r^2) threshold with the reference of linkage disequilibrium. The final result contains the most significantly

disease-associated SNP for each linkage disequilibrium-based cluster across the genome [20].

It is essential to acknowledge that the risk associated with a high polygenic score cannot reveal a single underlying mechanism, but rather may indicate the combined influence of multiple pathways [20].

Genetic Variants and Drug Therapy

Genetic factors also affect individual responses to drug therapy. Accordingly, DNA polymorphisms will be valuable in helping researchers determine and understand why persons are different in their abilities to absorb or clear certain drugs, furthermore to determine why a person may experience an adverse side effect to a specific drug. Consequently, the recent discovery of variants promises to revolutionize not only the process of disease progression but also the practice of preventative and curative medicine.

Pharmacogenomics is an analysis of the genome and genomic products like RNAs and proteins in the context of drug response, whilst pharmacogenetics is the study of variability in drug responses connected to genetic factors in different populations. For instance, gene expression profiling has enabled the demonstration of distinct gene clusters, which may be expressed differentially in healthy and disease tissues. These tissue specific gene expression profiles can assist to recognize treatment response at the genomic level. However, the expression profiles, which can predict responses to medicines in pharmacogenetics, are different from those profiles, which are used to estimate inherited differences in our genetic code to predict responses to medicines

[22, 23]. These genomic analytical techniques are essential for differential diagnosis of patients, specifically for heterogeneous diseases, which may have different molecular expression for similar clinical phenotypes.

In this technological advanced era, pharmacogenetics has promised to patients, that new prescribed medicines will be more effective and less likely to have adverse drug reactions (ADRs). This role of pharmacogenetics in the reduction of ADRs, has added further moral weight to the need to embolden this developing scientific venture. Variants may be associated with the absorbance and clearance of medicine. A drug verified effective in one patient maybe ineffective in other patients, or even cause an adverse reaction. The current practice of pharmaceutical companies is to develop only those therapeutic agents to which the "average" patient will respond. Therefore, drugs that might benefit a limited number of patients never make it to market.

Practically, the therapeutic response to conventional drugs (like multifactorial strategies, cholinesterase inhibitors) is genotype-specific [22]. Currently, chronic administration of anti-parkinson medication induces the phenomenon of "wearing-off", with further autonomic and psychomotor complications. In order to reduce these clinical complications, novel drugs and bioproducts should be developed, which can decrease premature neurodegeneration, instigate dopaminergic neuroprotection, and improve dopaminergic neurotransmission. Since therapeutic outcomes and biochemical changes are likely dependent on genomic profiles of patients, in order to optimize therapeutics, personalized treatments should rely on pharmacogenetic methodology. [24]

Genes involved in pharmacogenetic cascade can also be differentially expressed by epigenetic deviations (like histone modifications, DNA methylation, microRNA

dysregulation), which can induce abnormal drug processing and lead to diminish medicinal efficacy and safety [25,26]. Thus variations in the epigenetic machinery are accountable for defective tissue-specific expression of genes. Moreover epigenetic alterations in metabolic, transporter, and pathogenic genes can also develop drug resistance and toxicity [25-27].

Pharmacogenetic strategies comprises a series of steps in a multidisciplinary approach to develop new drug compounds, which includes: (a) genetic screening (genotyping) to find major gene targets; (b) genetic variation analysis to differentiate populations; (c) genomic structural and functional analyses by using genetic clusters and haplotypes; (d) genotype-phenotype correlations analysis to characterize key phenotypes as therapeutic targets associated with metabolic pathway genes; and (e) clinical and basic pharmacogenomic procedures implementation for drug development [28].

In the future, the most suitable drug for a patient could be determined by examining his genome profile, prior to his treatment. The ability to target a drug to specific patients most likely to benefit, is referred as "personalized medicine", and it would pursue medicine manufacturer to develop many more drugs and allow doctors to prescribe personalized therapies specific to a patient's needs.

Identification of Functional Genetic Variance –

Needles in a Haystack

Identification of functional genetic variants in the human genome is an intimidating prospect, but over the last 20 years, biomedical researchers have developed several powerful techniques. The underlying philosophy of each technique is a different method to compare selected portions of a DNA sequence obtained from different individuals who share a common disease. As it is currently difficult to measure and evaluate a variant's overall effect on disease aetiology, therefore, DNA sequence just refers to a person's genetic predisposition, or the potential of an individual to develop a disease based on genes and hereditary factors. The genetic factors involved in the intricate pathways of disease developmental progression are still not fully understood, so it is difficult to develop screening tests for a number of diseases and disorders. By studying genomic loci that have been found to harbour a genetic variant associated with a disease trait, researchers may reveal relevant genes associated with a disease. Defining and identifying the role of genetic factors in disease will also help researchers better evaluate the role of non-genetic factors on a disease, such as diet, behaviour, lifestyle, and physical activity.

Researchers have also identified a large number of genetic variants, which are not accountable for a disease state. Rather, they serve as biological markers for identifying a disease on the human genome map, because they are usually located near a gene associated with a certain disease.

Genome Wide Association Studies (GWASs)

GWASs implement a simple design to compare allele frequencies for hundreds of thousands of common variants distributed on the genome between large samples of disease cases and controls. If a variant influences the disease, even slightly, this should be apparent as greater allelic frequency in disease cases than controls, given a sufficient large sample size. This design has been applied to many different complex diseases, with varying levels of success. It is notable that these studies use a sample of SNPs to tag variation across the genome. A link between a single SNP and the disease cannot be used to conclude that SNP itself is involved. This association could be due to any of the other variants in linkage disequilibrium (LD) with it. These are referred to as tagged rare variants. Thus the goal of GWAS is not only to identify causal alleles but also to point out genomic loci that where they might be located [29] (Figure 2).

GWAS have been reasonably successful for many complex disorders. They have identified a number of candidate loci for many diseases. Some of these loci were previously involved as sites of known rare mutations that cause apparently Mendelian forms of the disease in question, whereas others are novel findings that connect new genes in disease progression. For some diseases, the findings congregate on particular bio-chemical processes or pathways. These studies have also discovered some shared genetic risks across multiple diseases, including neurodegenerative diseases**,** various autoimmune diseases and between schizophrenia and bipolar disorder [29] (Figure 3).

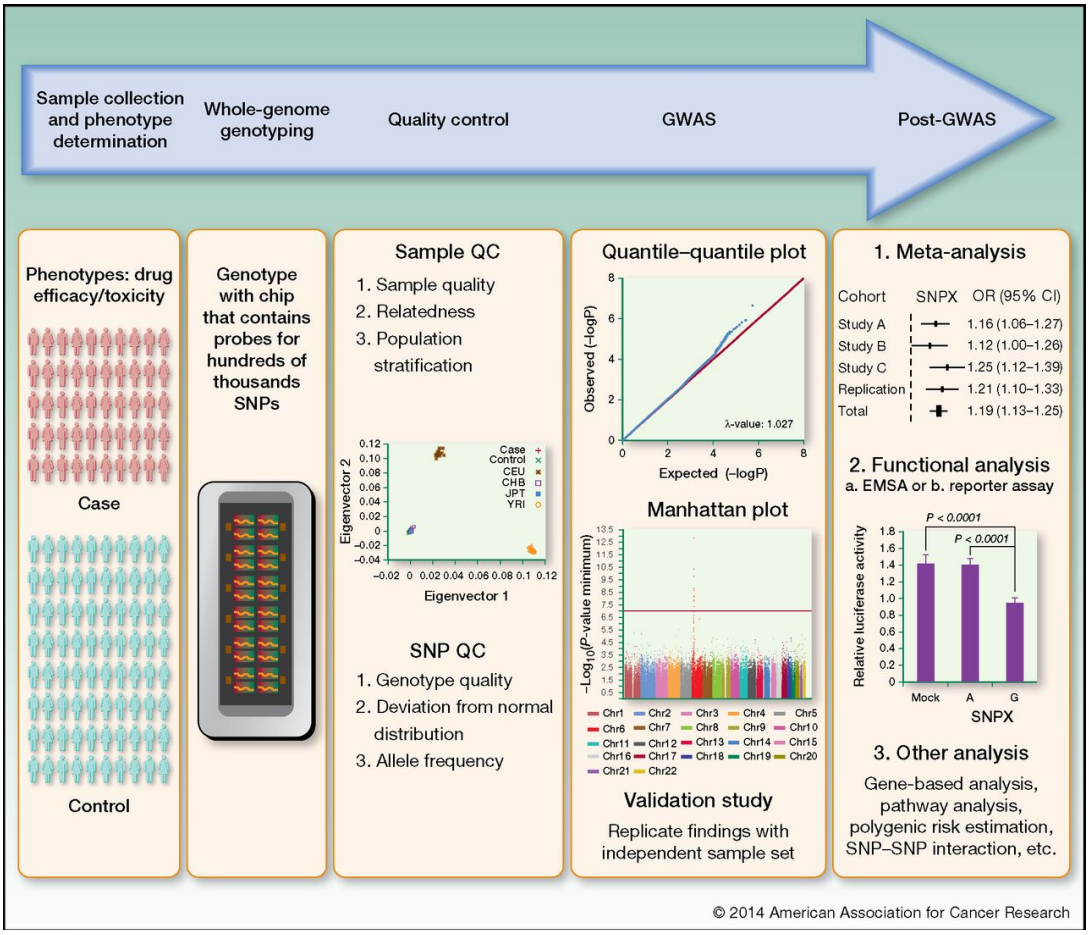


Figure 2: Summary workflow of GWAS. GWAS starts with the determination of phenotypes. In pharmacogenomics studies, cases are often the patients who do not respond or who develop severe adverse reactions, whereas controls are patients who respond to the treatment or who do not develop any adverse events after exposure to drug(s) treatment. All the samples are genotyped with chips that contain up to hundreds of thousands of SNPs. Quality control (QC) is a crucial step to ensure the association studies are performed with a good-quality sample and SNP set. Sample quality control usually includes (1) sample quality to exclude poorly genotyped samples, (2) identity-by-state analysis to exclude close relatedness samples, and (3) principal component analysis to evaluate population stratification of the sample sets to obtain a homogeneous sample set before performing the association study. SNPs are excluded if (1) they are of low genotype quality, (2) if they deviate from normal distribution by evaluating Hardy–Weinberg equilibrium in control samples, and (3) if they contain nonpolymorphic SNPs (minor allele frequency = 0). To evaluate the association distribution, quantile–quantile plots (Q–Q plot) of observed P value versus expected P value and genomic inflation factor (λ value) are evaluated to eliminate the possibility of population substructure. Manhattan plots of P value (−log10) versus chromosome loci are utilized to depict an overview of the GWAS, with each dot representing a SNP and each color representing a chromosome. The post-GWAS includes (1) a meta-analysis that combined multiple studies to identify significantly associated SNPs, and (2) functional analysis. Two of the most common functional analyses of the identified variants are (A) electrophoretic mobility shift assay (EMSA) to check the existence of proteins, mainly transcription factors, binding to SNP-contained DNA fragments and (B) luciferase reporter assay (comparison of relative luciferase activity) to assess the associated SNPs that could affect differential gene expression (as shown in figure). (3) Other analyses, including gene-based analysis, pathway analysis, polygenic risk estimation, SNP–SNP interaction, SNP–environment interaction, etc., could be carried out after GWAS. (taken from: Low SK, et al. Clin Cancer Res. 2014 May 15.)

Critically, the sample sizes, which are necessary to identify robust and replicable findings, are not obtainable be individual groups. Thus, collaborations have proliferated to enhance the statistical predictability of GWAS. Moreover, contribution of genetic factors also depends on a variant's allele frequency and effect size (Figure 4).

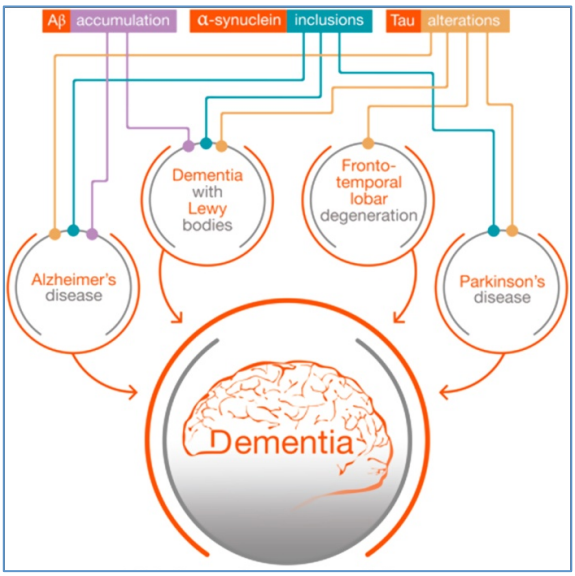


Figure 3: Molecular mechanism sharing in neurodegenerative diseases (taken from: Delgado-Morales R, et al. Mol Psychiatry. 2017 Apr)

Genetic and Gene Mapping Studies

Gene mapping studies are used to comprehend genetic architecture and to establish the association between DNA sequence variations and phenotypic variability. Undoubtedly the field of genetic studies has enjoyed great accomplishment over the past decade [29]. These association-mapping studies have gradually become genome-wide association studies (GWAS), whole-genome sequencing studies (WGS studies) and whole-exome sequencing studies (WES studies) [30]. GWAS that is a least expensive modern genome-wide gene mapping method have been successfully used

in large human populations to understand the direct association of common variants with complex diseases [29].

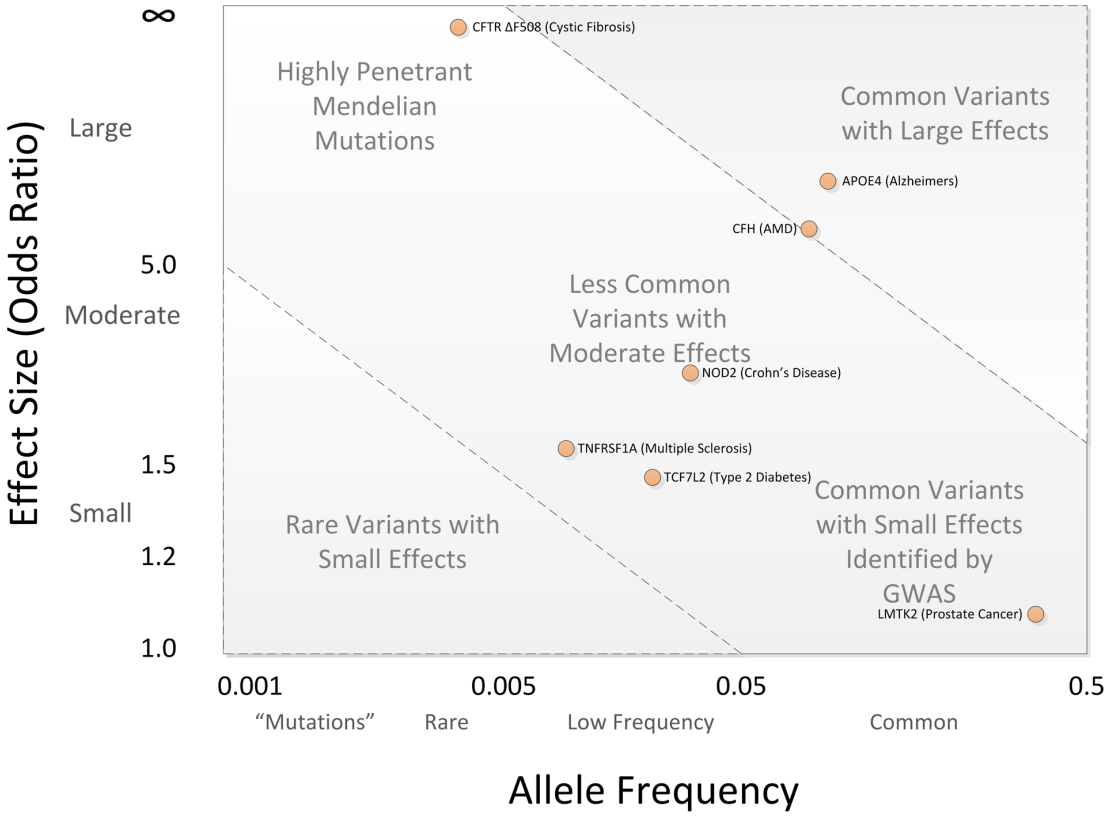


Figure 4. Spectrum of Disease Allele Effects. Disease associations are often conceptualized in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines. (taken from: Bush WS, et al. PLoS Comput Biol. 2012)

GWAS and Neurodegenerative Diseases

Many complex diseases, like heart disease, cancers, Alzheimer's disease, Parkinson's disease, and diabetes have a significant impact on the health of human populations. These diseases are associated with a combination of genetic and environmental factors, most of which have not yet been fully identified. The

contribution of genetic factors, specifically the links between genetic variants and diseases, is a long-established query in the study of complex diseases.

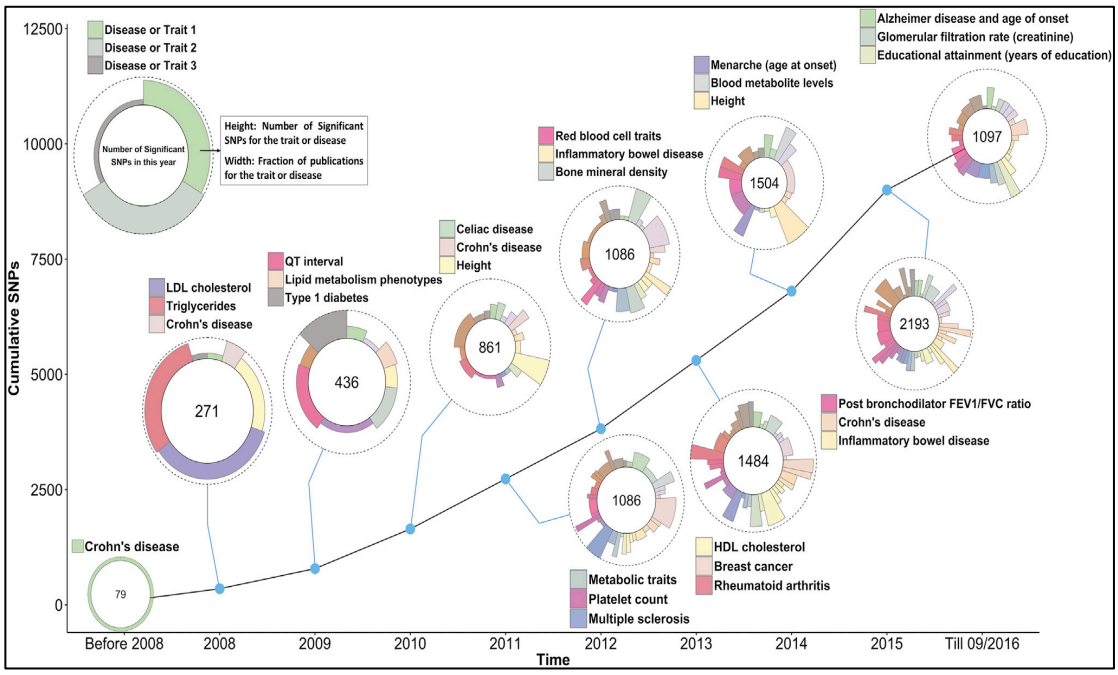


Figure 5: GWAS SNP-Trait Discovery Timeline (taken from: Visscher PM, et al. Am J Hum Genet. 2017 Jul 6)

Genome-wide association studies (GWAS) have revealed a large number of genetic loci associated with disease susceptibility in complex human diseases [31] (Figure 5). The majority of GWAS have identified that significantly associated genetic variants fall outside of DNA coding regions [32,33], which complicates our learning of how the specific variants intensify disease susceptibility. Therefore, our understanding of the functional impact of genetic variants in a complex disease remains limited. It is critical to further determine their role in molecular level biological functions [6].

Alzheimer's disease (AD):

Alzheimer's disease is the most common form of dementia. While AD is usually diagnosed among people aged 65 and older, it is not a normal part of aging. Heritability in AD is up to 76% but genetically it is very complex [34].

AD is characterized extracellular β-amyloid (Aβ) senile plaques and intracellular hyper-phosphorylated tau protein neurofibrillary tangles [34]. Early onset of AD is linked with mutations of the amyloid precursor protein (APP) protein and the presenilin1 (PSEN1) and presenilin2 (PSEN2) proteins. For late-onset form of AD, apolipoprotein E (APOE) is established explicitly as a susceptibility gene [34].

A GWAS of AD, using GERAD1 (Genetic and Environmental Risk in AD Consortium 1) sample published in 2009 [35] identified two susceptibility loci: Clusterin (CLU) and Phosphatidylinositol Binding Clathrin Assembly Protein (PICALM). Another independent AD GWAS presented significant evidence for AD association with CLU and Complement C3b/C4b Receptor 1 (CR1), and PICALM, by using EADI (European Alzheimer's Disease Initiative sample) data. Further studies with independent datasets have replicated the association of AD with CLU, PICALM and CR1 [36-39]. Moreover, Seshadri et al. [40] described a significant link between AD and Bridging Integrator 1 (BIN1) after merging GERAD1 and EADI with CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) data. More recently, an extended study by GERAD (GERAD+) and the American Alzheimer's Disease Genetic Consortium (ADGC), reported evidence for association at the ATP Binding Cassette Subfamily A Member 7 (ABCA7) and the Membrane Spanning 4-Domains A (MS4A) loci [41], and suggestive evidence for association with SNPs at the CD33 Molecule (CD33), CD2 Associated Protein (CD2AP), AT-

Rich Interaction Domain 5B (ARID5B) and EPH Receptor A1 (EPHA1) loci [42] (Figure 6).

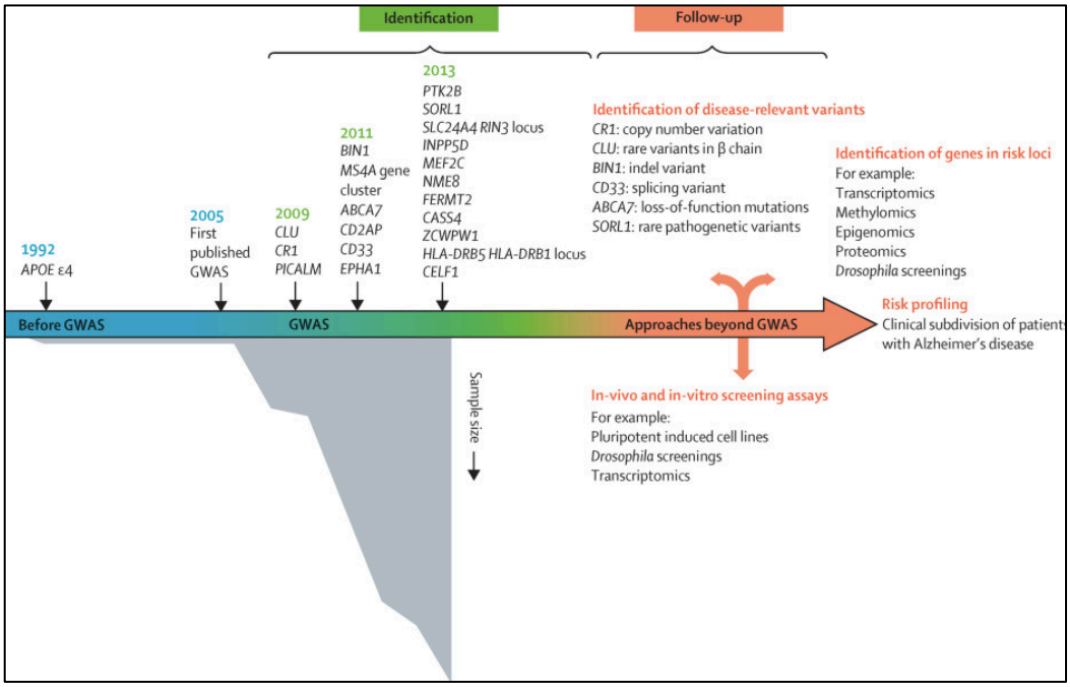


Figure 6: Genetic variations underlying Alzheimer's disease (taken from Cuyvers E, et al. Lancet Neurol. 2016 Jul)

In AD pathogenesis, psychotic symptoms are considerably more common than the general population, by affecting approximately 40% of cases [43]. Thus psychosis is suggested as a marker for a subtype of AD [44]. These symptoms are associated with more rapid cognitive [45] and functional decline [46]. It has been estimated in three independent cohorts that the heritability of AD and Psychiatry is approximately 61% [47-50].

Parkinson's disease (PD)

Parkinson's disease belongs to a group of conditions called motor system disorders. The four primary symptoms are tremor or trembling in hands, arms, legs, jaw, and face; rigidity or stiffness of the limbs and trunk; 'bradykinesia' or slowness of movement; and postural instability or impaired balance and coordination. PD is the most common form of Parkinsonism, the name for a group of disorders with similar features. These disorders share the four primary symptoms described above, and all are the result of the loss of dopamine producing brain cells [51].

Genetically, PD is a heterogeneous and complex disorder. Several GWAS for PD have been reported and three meta-analyses have been performed. All GWAS indicate a strong association to the Synuclein Alpha (SNCA) gene [52,53]. Most studies also confirm an association with the Microtubule Associated Protein Tau (MAPT) gene [54,55]. An early 2011, a meta-analysis of datasets from five PD GWAS identified eleven loci that exceeded the threshold for genome-wide significance. Six had been previously identified (MAPT (Microtubule Associated Protein Tau), SNCA (Synuclein Alpha), Major Histocompatibility Complex Class II - DR Beta 5 (HLA-DRB5), Bone Marrow Stromal Cell Antigen 1 (BST1), Cyclin G Associated Kinase (GAK) and Leucine Rich Repeat Kinase 2 (LRRK2)), whereas five were novel (Aminocarboxymuconate Semialdehyde Decarboxylase (ACMSD), Serine/Threonine Kinase 39 (STK39), Methylcrotonoyl-CoA Carboxylase 1 (MCCC1) / Lysosomal Associated Membrane Protein 3 (LAMP3), Synaptotagmin 11 (SYT11) and Coiled-Coil Domain Containing 62 (CCDC62) / Huntingtin Interacting Protein 1 Related (HIP1R)) [56]. A second meta-analysis revealed another five Parkinson disease risk loci (Parkinson Disease 16 (PARK16), Syntaxin 1B (STX1B), Fibroblast Growth

Factor 20 (FGF20), Starch Binding Domain 1 (STBD1) and Glycoprotein Nmb (GPNMB)) [57].

Recently our knowledge of the genetic architecture for PD has improved. Autosomal recessive mutations in PTEN Induced Putative Kinase 1 (PINK1), Parkinsonism Associated Deglycase (PARK7), and Parkin RBR E3 Ubiquitin Protein Ligase (PRKN), and dominant mutations in SNCA, LRRK2, and VPS35 Retromer Complex Component (VPS35) cause the disease with high penetrance. Therefore, approximately 5–10% of patients have a monogenic form of PD. Furthermore, autosomal recessive DnaJ Heat Shock Protein Family (Hsp40) Member C6 (DNAJC6) mutations are described in predominately atypical as well as typical PD [58] (Figure 7).

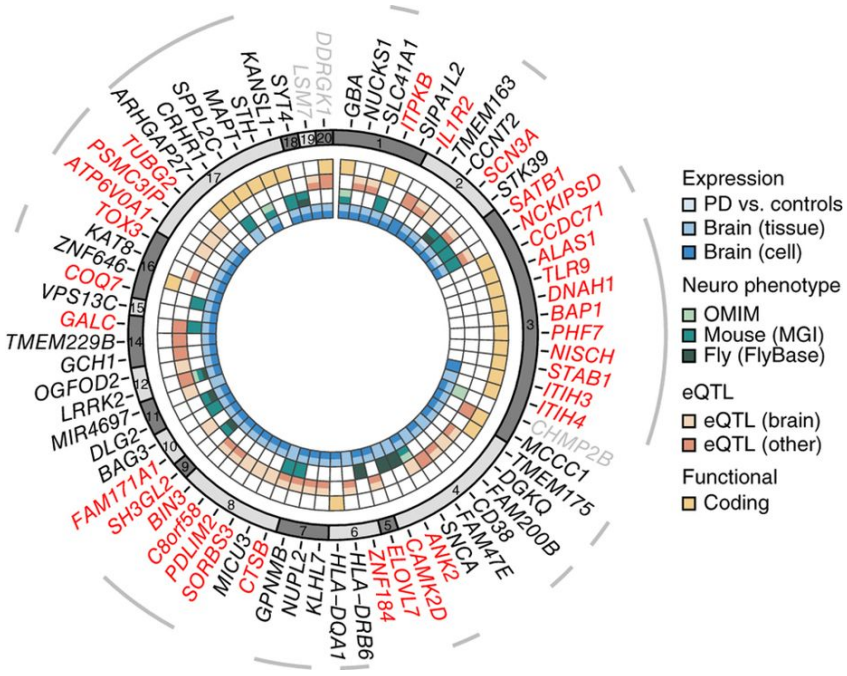


Figure 7: Genetics of Parkinson's disease (taken from: Chang D, et. al. Nature Genetics, 2017)

The latest and most comprehensive meta-analysis, using data from seven million polymorphisms, confirmed many previously reported risk loci. Evidence for a new

risk variant in the Integrin Subunit Alpha 8 (ITGA8) gene was also found. The risk factors identified by these studies provide clues to the basic molecular mechanisms involved and offer potential targets for novel treatments [59].

Psychiatric symptoms including depression and other visual hallucinations can be a prominent feature of PD. Around 25–50% of PD patients experience depression [60-62]. During the disease progression, dementia ultimately occurs in 20–40% of cases [63].

Functional Impact of Genetic Variants at Molecular Level

The functional impact of SNPs should be closely linked to their interference with (or modulation of) normal physiological functions. Some SNPs are very likely to directly interfere with bio-molecular functions of genes and genomic regions. By contrast, for other SNPs, the mechanism of disease susceptibility is unknown [64].

Protein Coding Variants have been most extensively studied due to their direct effect on the function of that encoded protein.

Non-Synonymous genetic variants change the amino acid sequence. They can modify amino acid composition, or truncate the protein sequences by causing an early translate on stop codon. Indels can also alter protein sequence; its effect on protein sequence depends on whether it is in-frame or frame-shifting. This substitution may affect protein folding, proper activity of binding or interaction sites, structure, stability or solubility of the protein [65].

Synonymous genetic variants do not alter the codon sequence. However, they can still impact protein function by modulating translation rates, with direct consequences to protein folding [66]. For instance, translation elongation rates are faster with higher codon adaptation to tRNA pools, along transcripts and slower with rare codons. Synonymous mutations have been revealed to have significant effects in the folding process of the emerging protein and can even modify substrate enzyme specificity [67]. This codon usage controls the speed of polypeptides vectorially translation from the ribosome and may impact protein-folding pathways [68]. It has also been shown that a fraction of codons specify not only an amino acid, but a transcription factor

binding site, providing an additional avenue through which synonymous polymorphisms may impart a functional effect [69].

Exonic Splicing Enhancers (ESEs) comprise specific hexamer sequences and an AG sequence at the intron-exon borderline. They instruct for the recruitment of the splicing complex to immature messengerRNA (pre-mRNA) and lead for intron excision and exon joining. SNPs may also present within exon splicing enhancers or silencers (ESEs/ESSs), resulting in deleterious intron retention or exon skipping [70 - 73]. SNPs and indels can also interrupt splicing sites to translate the protein isoform [74].

Regulatory Interactions at non-coding regions can take place over significant chromosomal distances up to an entire megabase (1MB) [75]. Genetic risk variants are very frequent on non-coding sequences [76]. Post-GWAS studies have revealed the capacity of these genetic risk variants to regulate gene expression by modulating cis-regulatory machineries through mechanisms involving DNA methylation [77], transcription factor binding [78], chromatin looping [79], or miRNA recruitment [80].

DNA Methylation refers to the addition of methyl groups to a cytosine nucleotide, which is basically part of a CpG dinucleotide. This DNA methylation is a heritable epigenetic event, which is involved in transcriptional regulation [81]. DNA hyper-methylation near transcription start sites (TSS) of tumour suppressor genes is associated with their silencing [81]. Across the genome, transcription factors bind to thousands of regulatory elements. These regulatory elements include promoters directly upstream of their target genes, and cis-regulatory elements such as enhancers, insulators and silencers [82]. ChIP-seq assays for transcription factors effectively annotate these cis-regulatory elements genome-wide. Analysis of these annotations

reveals that genetic risk variants commonly target cis-regulatory elements, mainly enhancers, in a disease- and tissue-specific manner [83-86] (Figure 8).

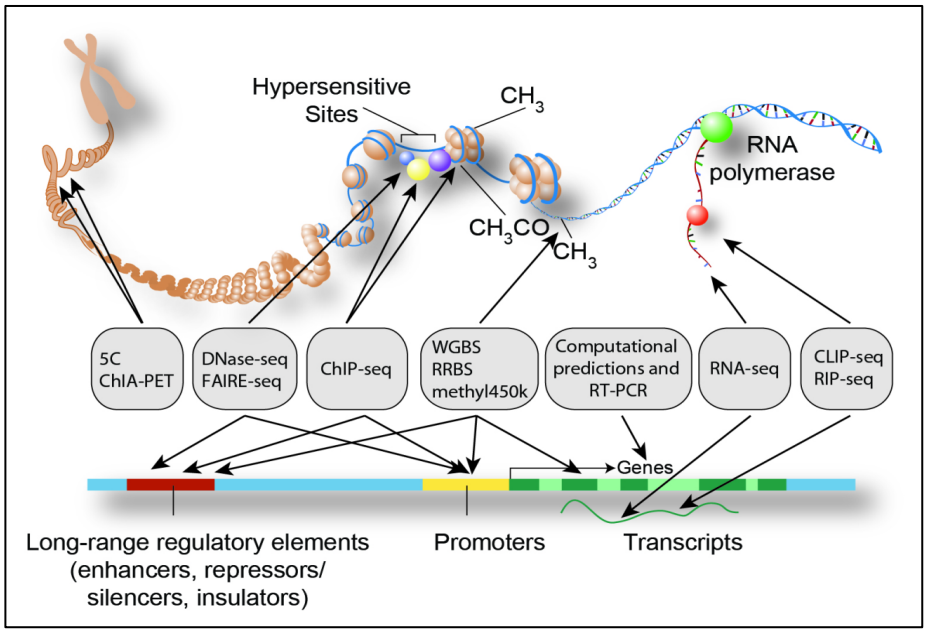


Figure 8: Functional genomic elements being identified by the ENCODE pilot phase. The indicated methods are being used to identify different types of functional elements in the human genome. (taken from: ENCODE Project Consortium. Science. 2004 Oct)

Genetic risk variants located within promoter regions can also change transcription factor binding to DNA, leading to differential target gene expression [87,88]. Enhancers are commonly targeted by those genetic variants of risk-associated loci that map to DNA recognition motifs, bound by transcription factors. These genetic variants can modulate the chromatin affinity for transcription factors and consequently gene expression [89–94]. Moreover, functional variants within a single risk locus can modulate multiple different enhancers. This multi-enhancer variant phenomenon was found to be a fundamental feature of many risk loci [95].

A genetic variant can modulate chromatin loop formation by altering the DNA affinity for looping factors. This can also result in allele-specific chromatin loop

formation. The human genome is structured in a three dimensional architecture which is thought to regulate a diverse set of DNA-templated processes [96-100]. This architecture facilitates the physical interaction of regulatory elements, like promoters and enhancers, through long-range chromatin loops or chromatin interactions [101,102] (Figure 9).

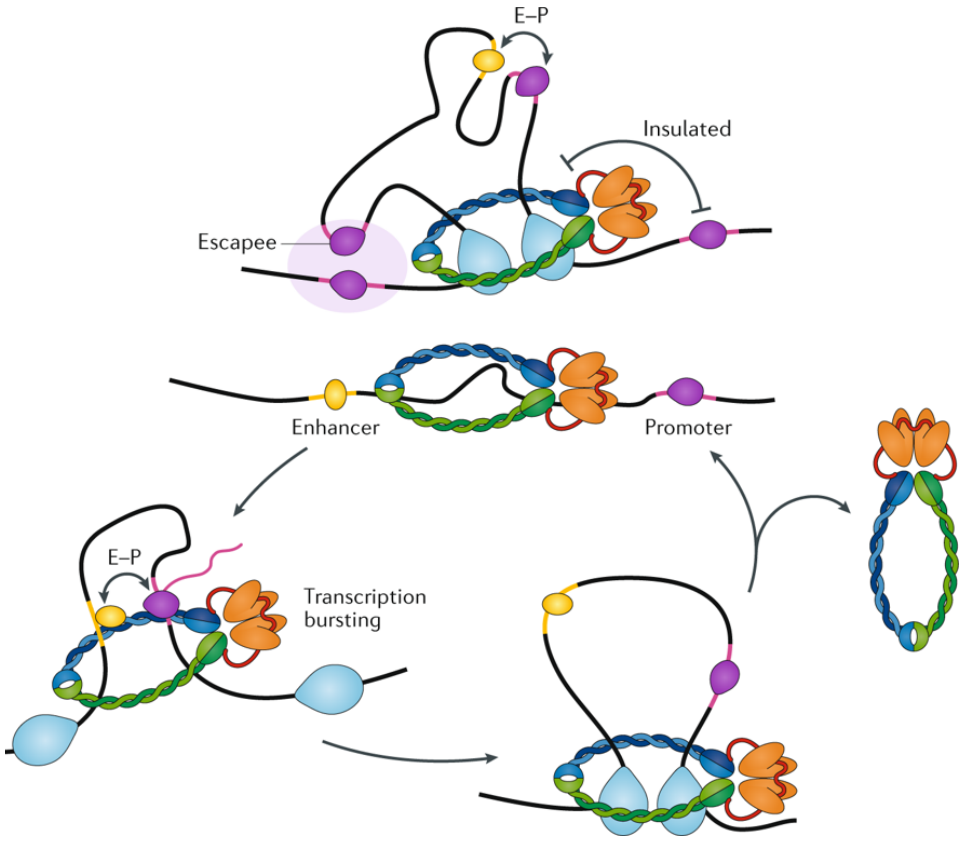


Figure 9: CTCF loops and enhancer–promoter interactions: CTCF (CCCTC-binding factor) loops establish domains in which sequences can interact more frequently. These contacts are thought to help promote enhancer–promoter (E–P) interactions when inside the domain but help insulate against those outside the domain. Enhancers are shown in yellow bound by a yellow transcription factor, and promoters are shown in pink bound by a purple RNA polymerase II (RNAPII). (taken from: Rowley MJ, et al. Nat Rev Genet. 2018 Oct 26)

If even a small fraction of these potential regulatory elements participate in chromatin looping, then most of the genomic interactions have to be characterized again, because many such loops appear to be tissue-specific [103-105]. These factors

contribute to the complexity of a systematic analysis of chromatin interactions [106] (Figure 10).

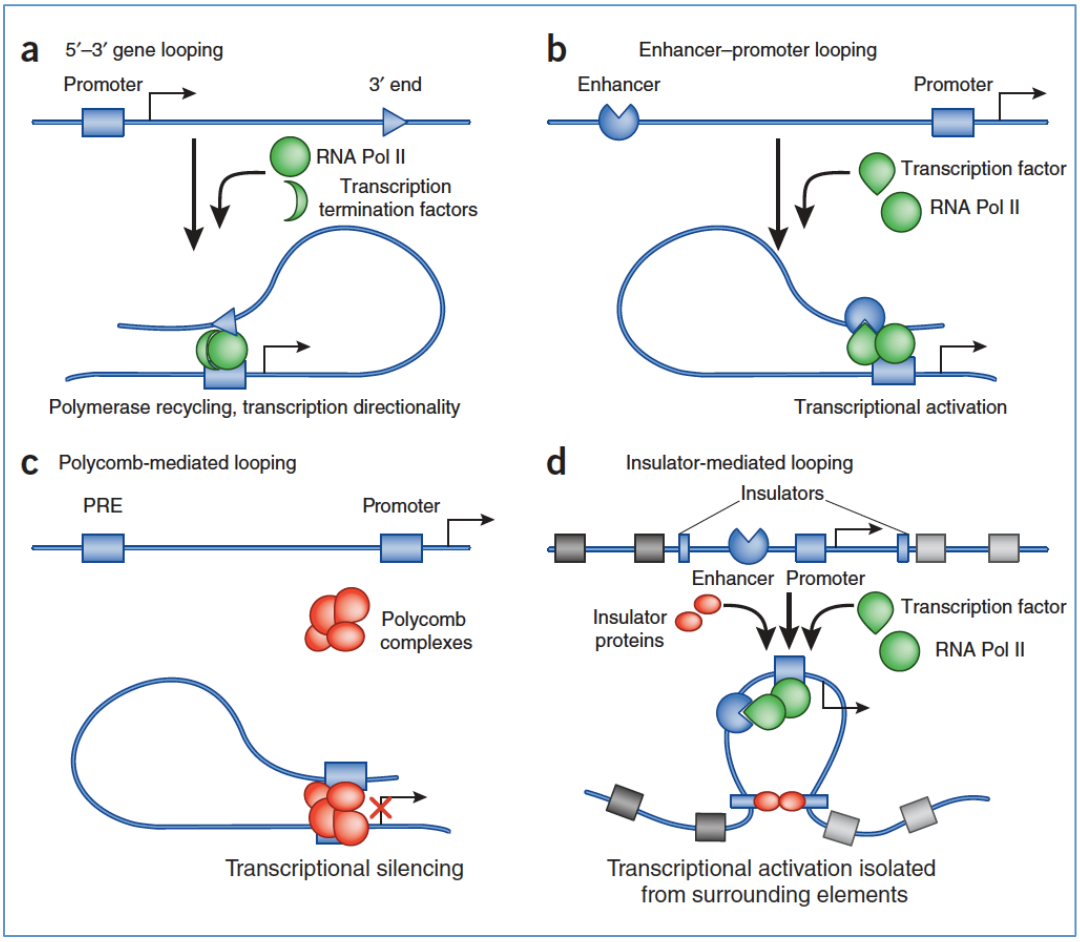


Figure 10: Four types of transcription regulatory chromatin loops. (**a**) Intragenic loops joining the 5′ and 3′ end of genes may allow recycling of RNA Pol II and facilitate maintenance of transcriptional directionality. (**b**) Enhancer-promoter loops—mediated by sequence-specific transcription factors, and possibly assisted by noncoding RNAs or by general DNA binding factors such as CTCF and cohesin—lead to transcriptional activation. (**c**) Loops between Polycomb-bound regions (PREs) and promoters prevent RNA Pol II recruitment and/or impair transcriptional elongation of promoter-bound RNA polymerases. (**d**) Insulator-mediated loops may segregate individual loci containing the coding part of the gene and its regulatory regions from the surrounding genome landscape with other regulatory elements. (Taken: Cavalli G, et al. Nat Struct Mol Biol. 2013 Mar)

MicroRNAs (miRNAs) target mRNAs by recognizing their complementary sequences mainly in 3' untranslated regions (3'UTRs). miRNAs largely function as post-transcriptional repressors. They can regulate the translation of hundreds of genes through sequence-specific binding to mRNA [107] (Figure 11).

lncRNAs are non-protein-coding transcripts which are found across intergenic regions of the human genome [108]. They can interact with chromatin regulators for their recruitment by chromatin [109,110]. Genetic variants can change lncRNA tertiary structures [111].

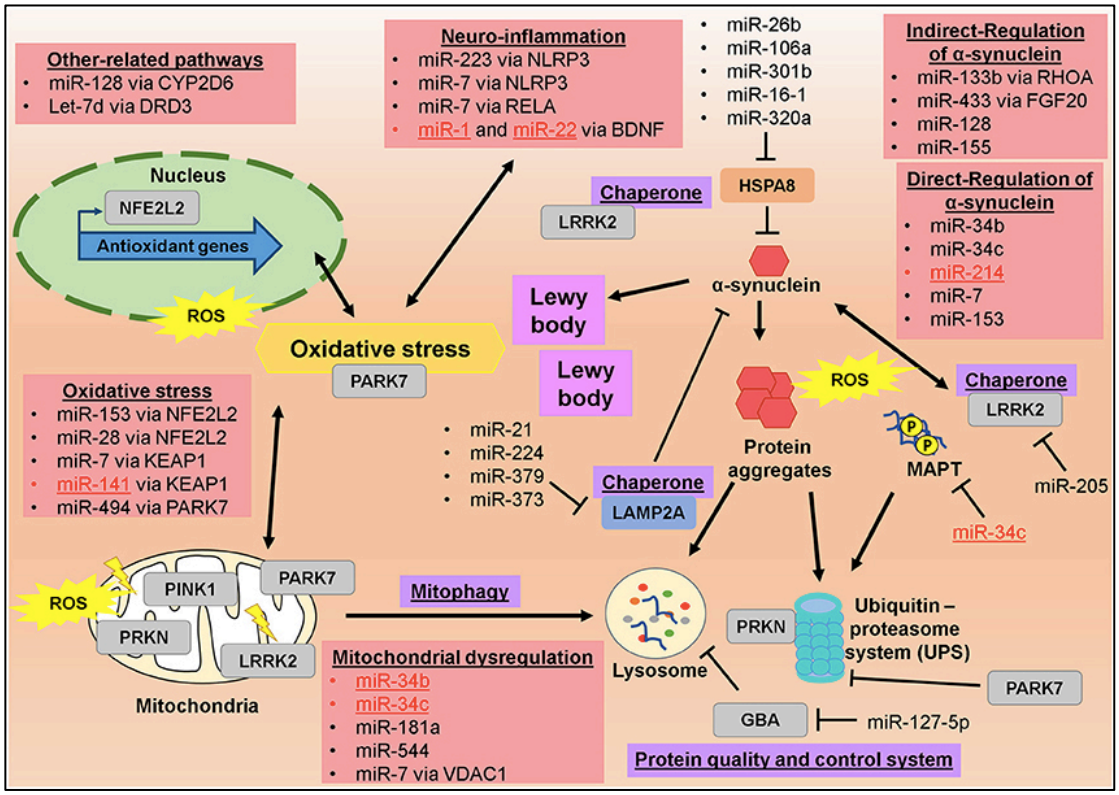


Figure 11. Illustrated regulatory network of altered microRNAs and their targeted genes in pathophysiology of Parkinson's disease: Altered microRNAs from EOPD and LOPD studies are significantly involved in the regulating various molecules in pathophysiological of PD, particularly in mitochondrial dysfunction, oxidative stress, neuro-inflammation, and toxic protein accumulation. Red-colored microRNAs are significantly altered in EOPD patients. (taken from: Arshad AR, et al. Front Mol Neurosci. 2017 Oct 31.)

Epistatic Interactions: There are three key categories of epistasis; functional, compositional, and statistical [112]. Functional epistasis ascertains the molecular interactions that genetic elements have with each another [113]. Compositional epistasis reveals the blocking effect on one allele by another allele at a different locus

[114]. Statistical epistasis expresses a quantitative way to detect how the genotype at one locus effects on the phenotype of another locus [115]; it measures deviation from the additive effects of two loci on the phenotype [112].

Strategies to determine the Causal Risk of Genetic Variants

The complexities of genetic variants are obscured at many different levels. Despite exhaustive effort, in most cases the interpretation of identified associations of genetic markers with diseases is still unclear. Intensive efforts in functional studies and fine mapping are required to more thoroughly understand the effects of genetic variants [116].

Current approaches to screen and identify the causal risk of variants at various non-coding loci have generally required multi-step experimental methodologies, like identifying allelic differences in both protein-DNA binding and transcriptional activity. In transcriptional activity, the allelic differences can be assessed using luciferase reporter assays [117]. In protein binding, they can be investigated by electrophoretic mobility shift assay and by ChIP to find upstream transcriptional regulators [118,119]. Moreover, genome editing techniques like CRISPR-Cas9 can be implemented, to determine the function of a specific variant. Then an assessment of cellular phenotypes and gene expression can be done [120]. Phenotypes can be compared between risk and non-risk alleles, in a carefully controlled experimental system. However, these methods are costly, slow and low-throughput. As GWAS have identified a large number of disease-associated variants, it is essential to acquire rapid, high-throughput data and knowledge-driven

methods to identify functional candidates linked to genetic risk alleles for the pathogenesis of complex diseases.

In the process of function investigation of genetic risk alleles, annotation of functional elements in the genome is a major goal. Annotations include determination of the inferred genomic elements where variants reside, like conserved elements, protein domains, transcription factor binding sites, promoters, exons, and introns. Additional annotations include prediction of the functional effect of variants on genomic components, like changes in the strength of transcription factor binding, microRNA binding, splicing efficiency, and protein function, annotations of biological and molecular processes to link variants across genes and genomic elements, and annotation of clinical and molecular characteristics of the gene or variant, like disease associations, eQTLs, population frequency, and pharmacogenetic variants.

In order to get a broad picture of the genomic functional components, the data and knowledge acquired through different methodologies needs to be combined and mapped using appropriate statistical and biological learning techniques. Likewise to keep up with these advance technologies, it is essential to design a framework for functional annotation. Such a framework can be attained through statistically justified and biologically motivated models [121].

Furthermore, functional genomic studies, at the RNA level, these consist of gene expression studies and at the protein level, these consist of proteomics studies, can provide useful complementary information to GWA studies. For instance, functional genomic studies can explain molecular and genetic mechanisms that influence disease development using critical information about gene regulation under different

conditions. GWAS, gene expression and proteomics studies independently have shown some successes in identifying genes associated with complex diseases (Figure 12).

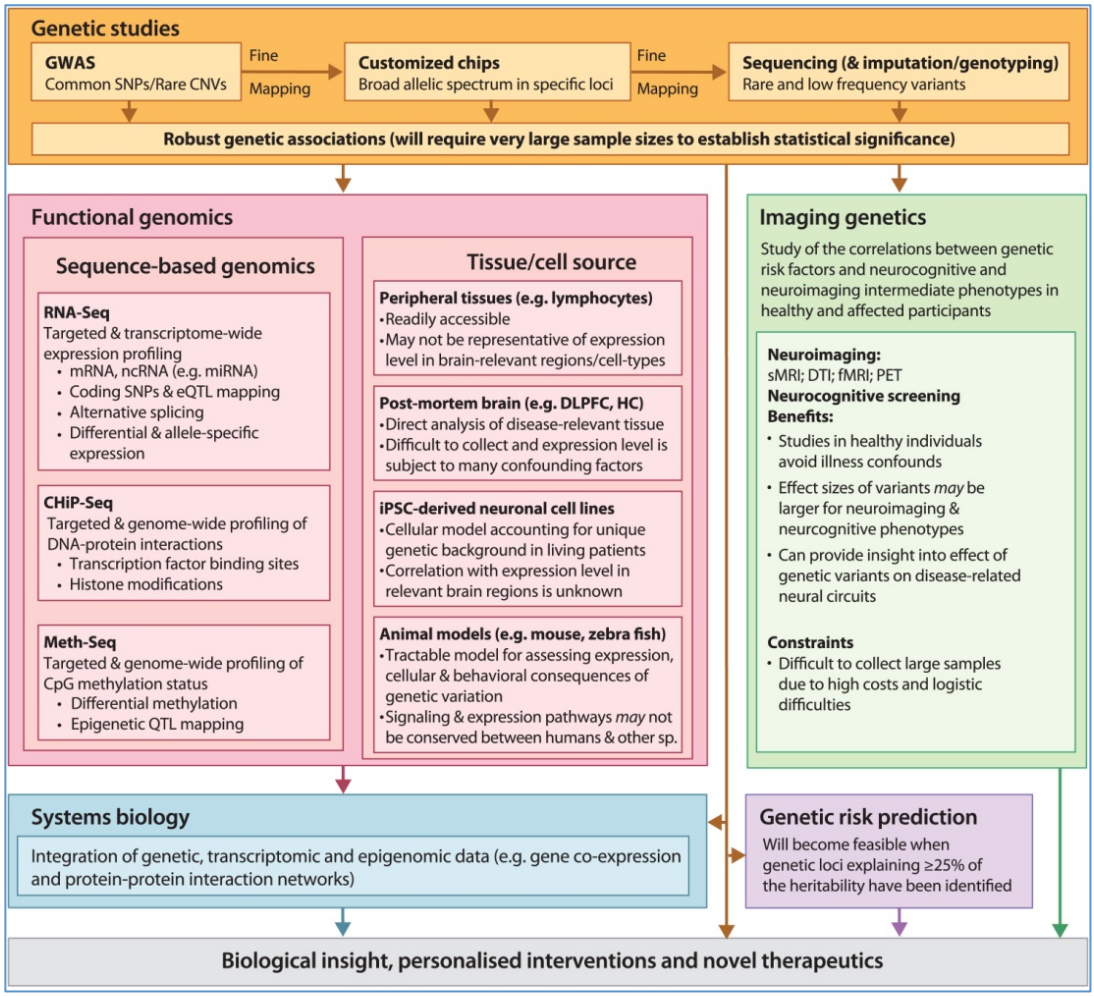


Figure 12: Roadmap for the identification and functional investigation of genetic risk alleles. In light of the growing evidence for a spectrum of allelic variation in disorders, diverse approaches, including GWAS, customized chips and sequencing, will all be important priorities for establishing novel statistical associations. These discoveries provide the foundation for functional investigations that aim to advance understanding of disease mechanisms at the molecular, cellular, synaptic and neural circuitry levels. (taken from: Mowry BJ, et al. Mol Psychiatry. 2013 Jan)

In conclusion, identification of causal genes for human complex disease is quite challenging. Replication is important and valuable, but it is difficult to achieve and maybe insufficient for validating GWA findings. Substantial information from other resources, such as genomic interactions and in detail molecular mechanistic studies,

may be useful for validating and clarifying the functional relevance of genes identified in GWA studies [122] (Figure 13).

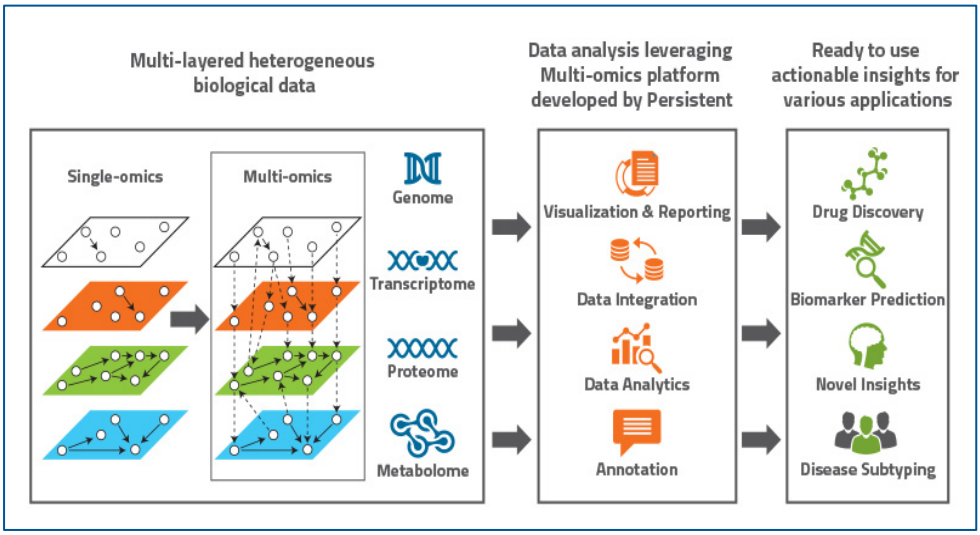


Figure 13: A schematic illustration for layers of data integration. (taken from: http://akross.info/?k=Genomic+Databases+Emerging+Tools+for+Molecular (Accessed on Nov. 2018-11-07))

One of the obvious questions is, which methodology can help in interpretation of GWAS data, in which most of the variants have small effects on disease susceptibility [123]. There is a lack of efficient and reliable algorithms as well as appropriate multi-scale modelling methodology, to evaluate the huge number of interdependent data from GWAS [124]. One way to reduce the combinatorial complexity of GWAS data is to reduce the dimensionality of genetic variation data by taking into account a priori knowledge about functional relationships between genes and proteins. Formalized knowledge about causal and correlative relationships in systems biology models provides a good starting point for that dimension reduction. So far, there have only been a few serious efforts to predict how these genetic variants would collectively be effective for specific phenotypes [125,126]. Thus there is a need for representation of data in standardized formats. Comparisons and evaluations of modern systems

biology modelling languages show that XML is superior format for systems biology information representation [127,128].

BEL (Biological Expression Language)

BEL (Biological Expression Language) represents knowledge in a computable form and is this suitable for use in modelling. It expresses the knowledge as BEL Statements that are stored in BEL documents. BEL is a highly expressive, triple-based language for the representation of knowledge about causal and correlative relationships [129]. BEL represents complex biological content as simplified, formalized, computable semantic triples that provide the ability to use and re-use experimental observations.

BEL is currently being applied to biological network analysis, disease modelling, understanding drug efficacy and toxicity, mechanisms for drug sensitivity and resistance, and other research and development related projects. A suite of software components called the BEL Framework provides tools that are required to create, compile, assemble and deliver computable knowledge models to BEL-aware applications [130]. BEL has the potential to impact scientific literature by introducing computable expressions in scientific publishing, that could be integrated efficiently into existing knowledge environments [131].

Integration of Genome Variance Information with Multi-scale Mechanistic Models

A key task in genetic variants interpretation lies in the ability to predict the molecular level mechanistic consequences of gene polymorphisms and mutations. As a consequence, systems biomedicine modelling approaches need to combine mechanistic information from various levels, including gene expression, miRNA expression, protein-protein interaction, genetic variation and pathway information.

1. Review of state-of-the-art approaches for functional interpretation of genetic variants

In this thesis, firstly I summarized the biomedical literature for assessment of the functional impact of genetic variation at molecular level. The literature review focuses on the functional interpretation of SNPs and mutations in a systems biology context with a strong link to network modelling approaches. The aim of that was to shed light on the perspectives for enhanced functional interpretation of complex SNP and mutation patterns in neurodegenerative disorders. I precisely summarized existing methodologies for genome wide association studies, currently available algorithms and computable modelling approaches. Moreover, I evaluated the required new approaches for modelling and simulations of genetic regulatory networks to predict the functional consequences of disease-associated genetic variants.

2. Developing predictive models to estimate cognitive decline and resilience in Alzheimer's disease:

Secondly, the next step was to collect publically available genetics data and to develop predictive models. These models estimate cognitive scores, predict discordance between cognitive ability and amyloid load, and/or predict diagnostic groups by identifying the most significant disease-associated genetic variants.

For this purpose, we participated in the Alzheimer's disease DREAM challenge [132], a crowdsourced computational project to benchmark the current state-of-the-art in predicting cognitive decline in Alzheimer's disease, by using high dimensional, publicly available genetic and structural imaging data (Figure 14).

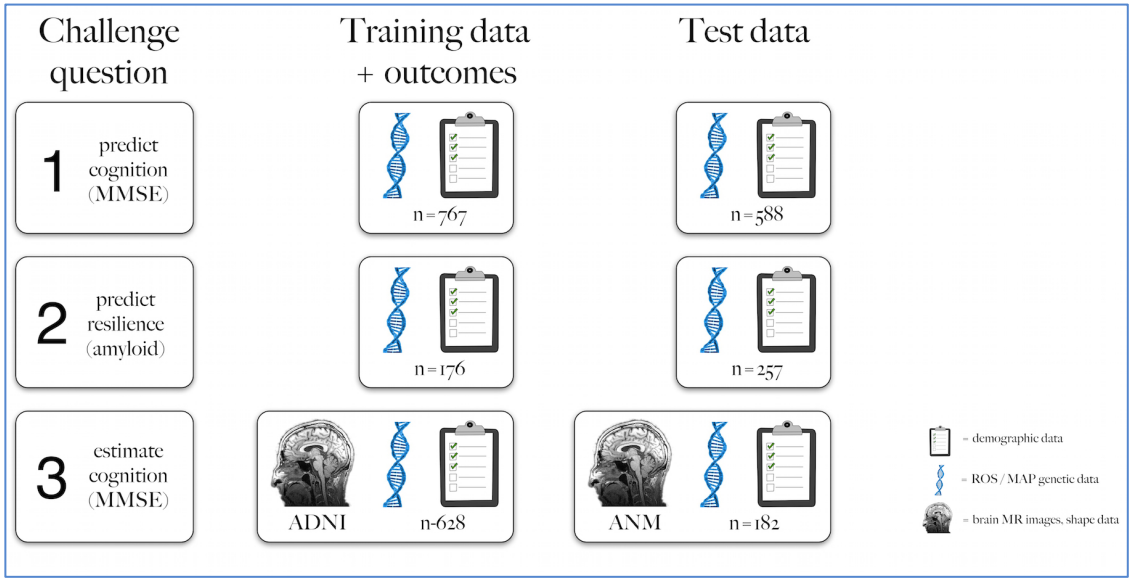


Figure 14: Dream Challenge overview: The top schematic summarizes the three challenge questions, the training and test data. (taken from: Alzheimer's Disease Neuroimaging Initiative. Alzheimers Dement. 2016 Jun; 12(6):645-53.)

Major collaborative efforts in the field are evaluating the association of genetic loci with AD to identify early biomarkers of diagnosis, but the utility of these approaches is not well established. Therefore to ensure that the genetic data were tested across a

broad spectrum of the state-of-the-art analytical approaches, the study was intended as a community-based challenge.

The ADNI (Alzheimer's disease Neuroimaging Initiative) [133] genetics, clinical and imaging data were used to train the predictive models. Moreover, data from the ROS (Religious Orders Study) [134] and MAP (Memory and Aging Project) [135] from the Rush Alzheimers Disease Center, and the European AddNeuroMed [136] study from InnoMed (the Innovative Medicines Initiative), were used to validate and test the predictive models.

Predictive models were designed to address the following three questions based on genetic data; 1: to predict 2-year cognitive decline and changes in MMSE scores. 2: to stratify individuals who exhibit resilience to AD pathology despite evidence of amyloid deposition. 3: to estimate MMSE scores using imaging data [137].

After pre-processing, filtering, and normalization of data, we prioritized SNPs based on their correlation with phenotype to stratify patient groups. Then we developed a generic ensemble method to integrate several independent models, which include linear regression and classification by using decision trees, SVMs, gaussian processes, and random forest. Afterwards, predictions were integrated via majority voting for classification models and by averaging for regression models (Figure 15).

In this challenge, more than 500 bioinformaticians worldwide demonstrated the viability of crowdsourced approaches in AD research. Our predictive model secured 9th position in the performance evaluation result (Figure 16). Unfortunately, top-scoring algorithms were unable to detect meaningful genetic biomarker of either cognitive decline or resilience, as the algorithms with the best predictive performances did not contain any genetic features beyond APOE haplotype. The

failure of this meta-analysis challenge suggested that alternate approaches should be considered for prediction of cognitive performance. It is also possible that the data were simply inadequate to estimate the cognition decline [137].

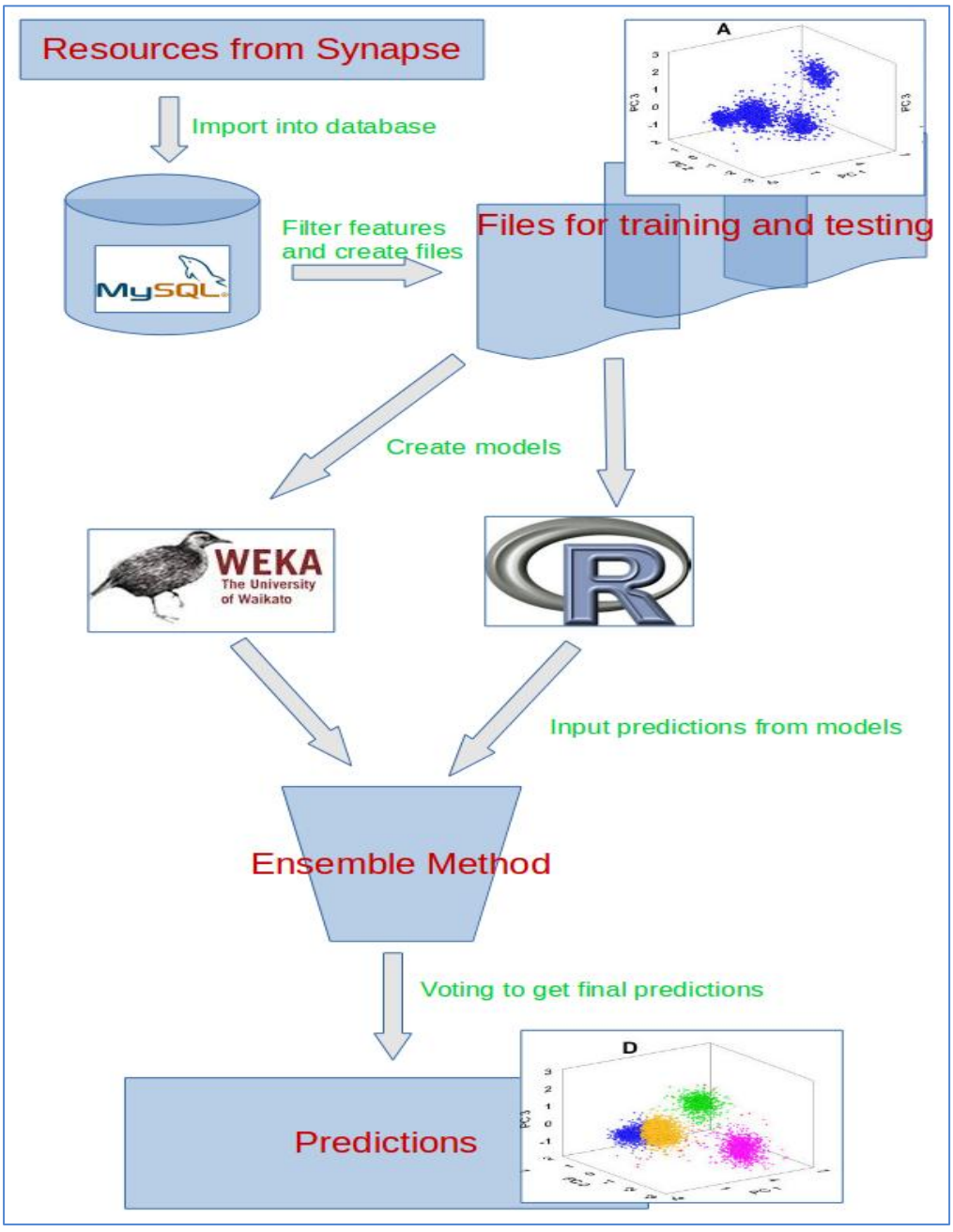


Figure 15: A Workflow for Ensemble learning with different classifier/regression models to estimate cognitive decline and resilience in Alzheimer's disease

While the DREAM challenge failed to detect a substantial genetic contribution to onset and progression of AD, linkage studies and heritability estimates have demonstrated that there is such a contribution [138-140]. Indeed, overall genetic contribution is the accumulated compilation of a large number of genetic loci with small epistatic or independent effects [141]. Historically, this type of signal is quite challenging to capture in predictive models and implausible to be useful in a diagnostic setting [142].

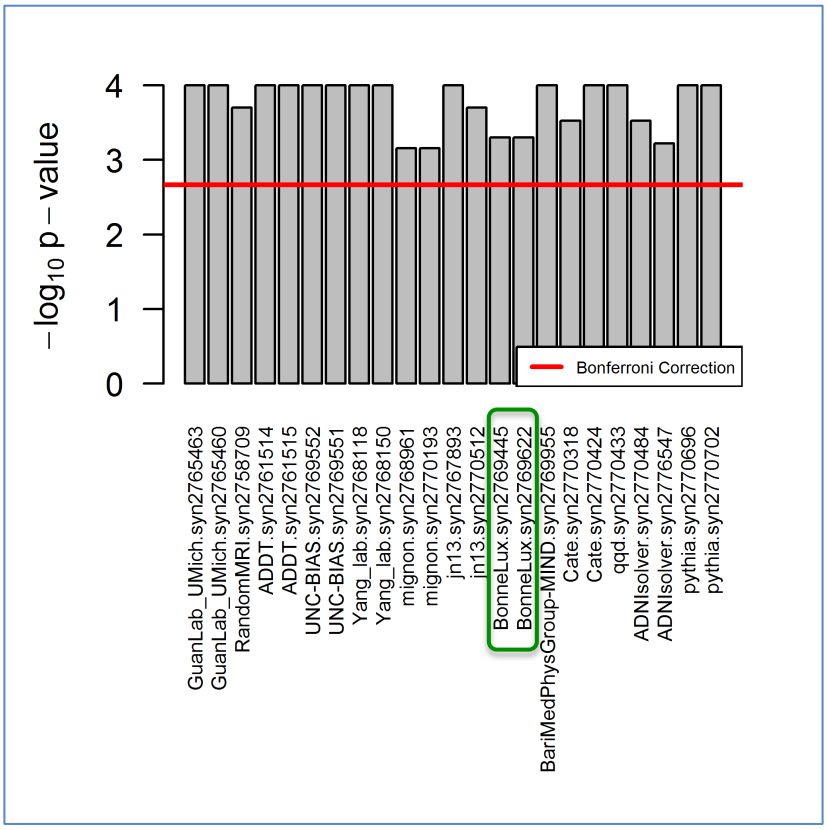


Figure 16. Performance evaluation results: P values (in negative log 10 scale) for intersection union tests investigating, which teams performed better than random. Explicitly, we tested the null hypothesis that Pearson's correlation (COR) or Lin's concordance correlation coefficient (CCC) are equal to zero, against the alternative that both COR and CCC are larger than zero. Adopting a 0.05 significance level, after Bonferroni correction. (taken from: Alzheimer's Disease Neuroimaging Initiative. Alzheimers Dement. 2016 Jun; 12(6):645-53.)

This suggests the need to develop new integrative approaches for biomedical heterogeneous data. Alternatively, in smaller scale analyses, prioritization of phenotypic depth over sample size may offer a more refined view of disease

associated molecular mechanisms. Most likely, successful identification of clinically relevant biomarkers of cognition will require the integration of multiple data sources and methods that represent greater phenotypic complexity.

3. **Designing a mechanistic approach for the interpretation of disease-associated risk variants in complex systems biomedicine models:**

Thirdly, I suggested a mechanistic approach for the interpretation of disease-associated risk variants in complex systems biomedicine models. I categorized and annotated genetic variants according to their predicted functional impact. Variants identified by GWA and eQTL studies were integrated with causal mechanisms through a multi-scale interconnection network (including genome – transcriptome – proteome) of epigenetic and genetic alterations located within significantly influential genomic regions. Associated verification evidences were derived from clinical or experimental outcomes.

For the integration of genetic variants into systems biomedicine models, OpenBEL (Open Biological Expression Language) [143], the open source version of BEL, has been extended. The BEL 2.0 syntax provides a representation for different genetic variant types, by introducing new variant functions for DNA, RNA and protein levels. This new variant function can be used as an argument within a gene(), rna(), microRNA(), or protein() procedure to indicate a sequence variant of the specified level. The variant function takes a HGVS (Human Genome Variation Society) [144] variant description expression, e.g., for a substitution, insertion, or deletion variant (Figure 17).

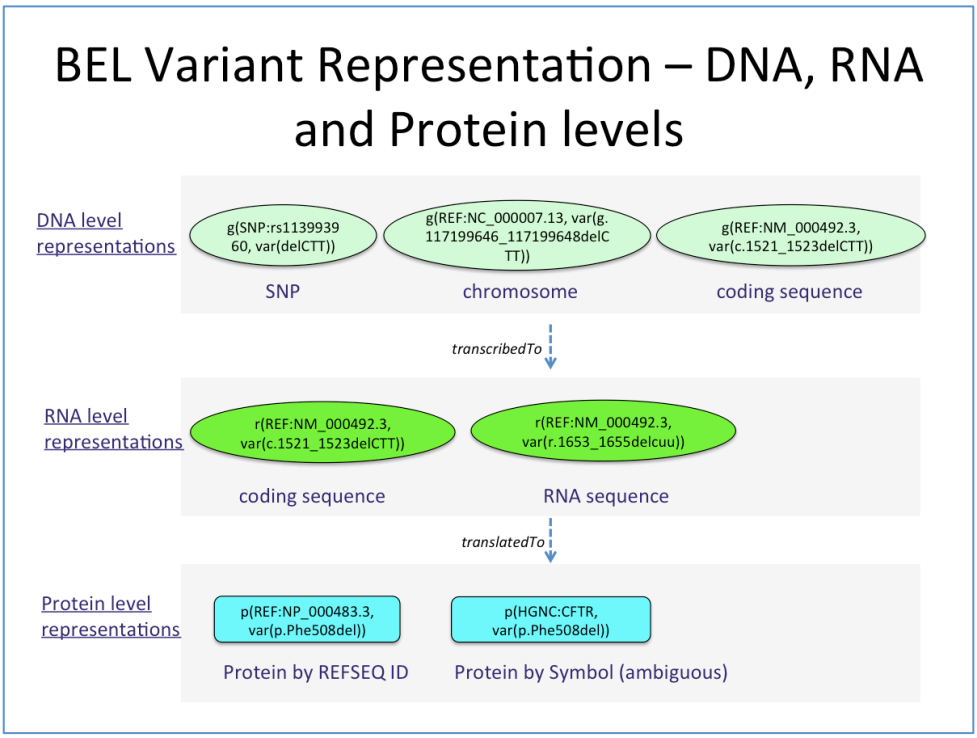


Figure 17: BEL variant representation at DNA, RNA and protein levels. Proposed syntax for the integration of genetic variants in cause-and-effect disease BEL model

The extended OpenBEL syntax is intended to support reasoning over cause-effect models that include genetic variation information. Accordingly, these variants were mapped to a disease model with both well established and novel disease associated genes. This is encoded in BEL to represent scientific findings by capturing causal and correlative relationships in context.

Using OpenBEL, I demonstrated how candidate mechanisms for early dysregulation events in AD can be identified. My integrative mining approach identifies "chains of causation" by comprising genetic information in BEL disease models. Through the annotation of disease models with genetic variants ranked according to the functional relevance-scoring scheme, I obtained an enhanced interpretation of the functional consequences of genetic variants in a mechanistic context.

4. **Implementing an integrative approach that starts with a data-driven approach to identify indicators in GWAS data, and allows for new insights into complex mechanisms through knowledge-driven context enrichment:**

Finally, I established an integrative approach that starts with a data-driven approach, identifies indicators in GWAS data, and allows for new insights into complex mechanisms, through knowledge-driven context enrichment. This approach integrates multi-model heterogeneous information and knowledge in biologically meaningful, computable graph models (Figure 18).

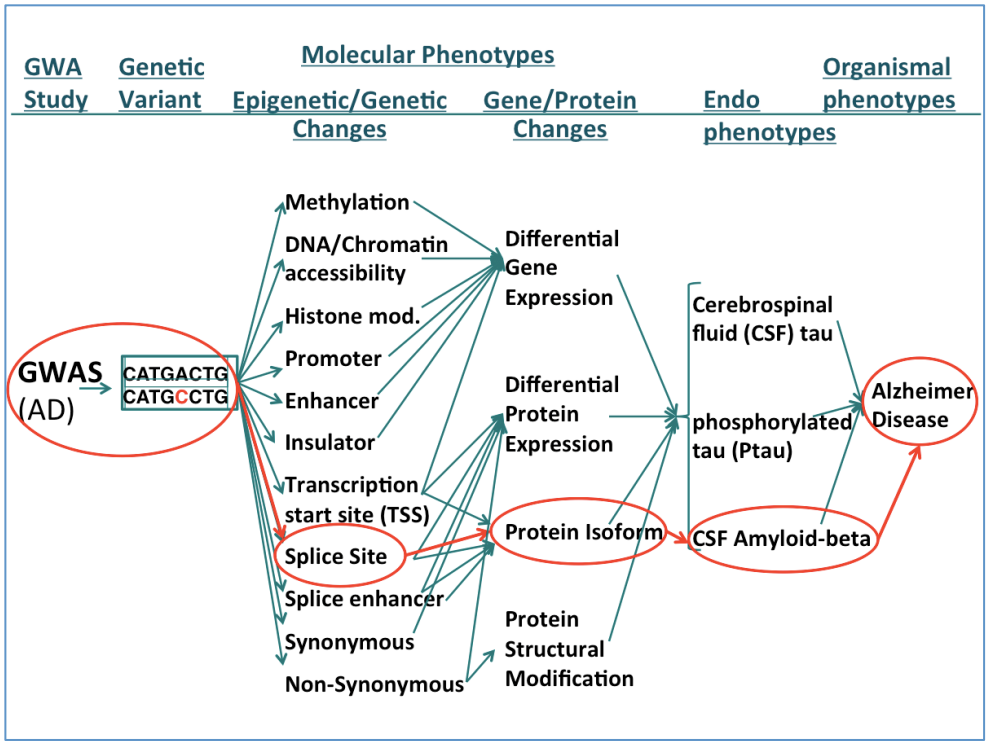


Figure 18: Integration scheme for GWAS data into systems biology models: Linking of genetic variants with Endo-phenotypes and phenotypes by using their functional consequences.

Such computable cause-and-effect models can be very helpful to identify possible molecular level perturbation mechanisms that contribute to disease pathology. As such, computable mechanistic models are essential to integrate diverse types of data

as well as relationships between the nodes. They can help to discover unknown links to illustrate the possible mechanism of dysregulation. Integrative models based on causal relationships span over multiple levels and scales and establish links e.g. from genetics to imaging features in one single, computable graph model.

I have tried to identify shared mechanisms underlying neurodegenerative diseases by a "shared genetics" approach. My strategy of indicator "enrichment via cause-and-effect modelling" is a novel contribution to shared genetics. It bears great potential for the mechanistic interpretation of the biological impact of genetic variation.

Finaly, I have also analysed GWAS data in various neurological conditions related to neurodegeneration, identified "shared loci" and came up with a short list of interesting "genomic hotspots" enriched for genetic variation with relevance for two major neurodegenerative diseases, Alzheimer´s Disease (AD) and Parkinsonism.

References

1. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat. Neurosci. 17, 782–790 (2014).
2. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. Nat. Rev. Genet. 9, 255–266 (2008).
3. Hansen, T. F. The evolution of genetic architecture. Annu. Rev. Ecol. Evol. Syst. 37, 123–157 (2006).
4. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. 10, 241–251 (2009).
5. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. Nat. Rev. Genet. 3, 779–789 (2002).
6. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. Cell 169, 1177–1186 (2017).
7. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. Proc. Roy. Soc. Edinburgh 52, 99–433 (1918).
8. Saunders, A.M., Strittmatter, W.J., Schmechel, D., George-Hyslop, P.H., Pericak-Vance, M.A., Joo, S.H., Rosi, B.L., Gusella, J.F., Crapper-MacLachlan, D.R., Alberts, M.J. et al. (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43, 1467–1472.
9. Corder, E.H., Saunders, A.M., Risch, N.J., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C.,Jr, Rimmler, J.B., Locke, P.A., Conneally, P.M., Schmader, K.E. et al. (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat. Genet.* , 7, 180–184.
10. Lai, E., Riley, J., Purvis, I. and Roses, A.D. (1998) A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics*, 54, 31-38.
11. Martin, E.R., Lai, E.H., Gilbert, J.R., Rogala, A.R., Afshari, A.J., Riley, J., Finch, K.L., Stevens, J.F., Livak, K.J., Slotterbeck, B.D. et al. (2000) SNPing away at complex diseases: analysis of single nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* , 67, 383–394.
12. Roses, A.D. (2000) Pharmacogenetics and the practice of medicine. *Nature*, 405, 857–865.

13. Roses, A.D. (1998) Genetic associations. *Lancet*, 351, 916.
14. Relkin, N.R., Tanzi, R., Breitner, J., Farrer, L., Gandy, S., Haines, J., Hyman, B., Mullan, M., Poirer, J., Strittmatter, W. et al. (1996) Apolipoprotein E genotyping in Alzheimer's disease (consensus statement of the National Institute of Aging/Alzheimer's Association Working Group). *Lancet*, 347, 1091–1095.
15. Roses, A.D. (1997) Apolipoprotein E, a complex gene with biological interactions in the aging brain. *Neurobiol. Dis.*, 4, 170–185.
16. Xu, P. -T., Schmechel, D., Rothrock-Christian, T., Burkhart, D.S., Qiu, H.-L., Popko, B., Sullivan, P., Maeda, N., Saunders, A.M., Roses, A.D. and Gilbert, J.R. (1996) Human apolipoprotein E2, E3 and E4 isoform specific transgenic mice: human-like pattern of neuronal immunoreactivity in central nervous system not observed in wild type mice. *Neurobiol. Dis.* , 3, 229–245.
17. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
18. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl Acad. Sci. USA* **111**, E5272–E5281 (2014).
19. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
20. Khera, A. V. & Kathiresan, S. Is coronary atherosclerosis one disease or many? Setting realistic expectations for precision medicine. *Circulation* **135**, 1005–1007 (2017).
21. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
22. Cacabelos R, Cacabelos P, Torrellas C, Tellado I, Carril JC. Pharmacogenomics of Alzheimer's disease: novel therapeutic strategies for drug development. Methods Mol Biol. 2014; 1175:323-556.
23. Allen D. Roses; Pharmacogenetics, *Human Molecular Genetics*, Volume 10, Issue 20, 1 October 2001, Pages 2261–2267
24. Cacabelos R. Parkinson's Disease: From Pathogenesis to Pharmacogenomics. Int J Mol Sci. 2017 Mar 4; 18(3). pii: E551.
25. Cacabelos R, Torrellas C (2015) Epigenetics of aging and Alzheimer's disease: Implications for pharmacogenomics and drug response. International Journal of Molecular Sciences16: 30483-30543.

26. Cacabelos R, Torrellas C, Aliev G (2016) Epigenetics-related drug efficacy and safety: The path to pharmacoepigenomics. Current Genomics.
27. Cacabelos R, Torrellas C (2014) Epigenetic drug discovery for Alzheimer's disease. Expert Opin Drug Discov 9:1059-1086.
28. Cacabelos R (2015) Impact of genomic medicine on the future of Neuropsychopharmacology. J Neuropsychopharmacol Mental Health 1:1.
29. Visscher, P. M. et al. 10 years of GWAS Discovery: biology, function, and translation. Am. J. Hum. Genet. 101, 5–22 (2017).
30. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2018 Feb; 19(2): 110-124.
31. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010 Jul 8; 363(2): 166-76.
32. Hindorff, L. a et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl Acad. Sci. USA 106, 9362–9367 (2009).
33. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013 Nov 7;93(5):779-97. doi: 10.1016/j.ajhg.2013.10.012.
34. Hollingworth P, Harold D, Jones L, Owen MJ, Williams J. Alzheimer's disease genetics: current knowledge and future challenges. Int J Geriatr Psychiatry. 2010
35. 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):1088–1093.
36. Corneveaux JJ, Myers AJ, Allen AN, Pruzin JJ, Ramirez M, Engel A, et al. Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. Hum Mol Genet. 2010; 19(16): 3295–3301.
37. Zhang Q, Yu JT, Zhu QX, Zhang W, Wu ZC, Miao D, et al. Complement receptor 1 polymorphisms and risk of late onset Alzheimer's disease. Brain Res. 2010
38. Carrasquillo MM, Belbin O, Hunter TA, Ma L, Bisceglio GD, Zou F, et al. Replication of CLU, CR1, and PICALM associations with alzheimer disease. Arch Neurol. 2010; 67(8): 961–964.

39. Jun G, Naj AC, Beecham GW, Wang LS, Buros J, Gallins PJ, et al. Meta-analysis Confirms CR1, CLU, and PICALM as Alzheimer Disease Risk Loci and Reveals Interactions With APOE Genotypes. Arch Neurol. 2010
40. 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): 1832–1840.
41. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, et al. Common variants in ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet. 2011 In Press.
42. Naj AC. Common variants in MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011 In Press.
43. Ropacki SA, Jeste DV. Epidemiology of and risk factors for psychosis of Alzheimer's disease: a review of 55 studies published from 1990 to 2003. Am J Psychiatry. 2005; 162(11): 2022–2030.
44. Sweet RA, Nimgaonkar VL, Devlin B, Jeste DV. Psychotic symptoms in Alzheimer disease: evidence for a distinct phenotype. Mol Psychiatry. 2003;8(4): 383–392.
45. Wilkosz PA, Seltman HJ, Devlin B, Weamer EA, Lopez OL, Dekosky ST, et al. Trajectories of cognitive decline in Alzheimer's disease. Int Psychogeriatr. 2009:1–10.
46. Lopez OL, Wisniewski SR, Becker JT, Boller F, DeKosky ST. Psychiatric medication and abnormal behavior as predictors of progression in probable Alzheimer disease. Arch Neurol. 1999; 56(10): 1266–1272.
47. Hollingworth P, Hamshere ML, Holmans PA, O'Donovan MC, Sims R, Powell J, et al. Increased familial risk and genomewide significant linkage for Alzheimer's disease with psychosis. Am J Med Genet B Neuropsychiatr Genet. 2007; 144B(7): 841–848.
48. Sweet RA, Nimgaonkar VL, Devlin B, Lopez OL, DeKosky ST. Increased familial risk of the psychotic phenotype of Alzheimer disease. Neurology. 2002; 58(6): 907–911.
49. Bacanu SA, Devlin B, Chowdari KV, DeKosky ST, Nimgaonkar VL, Sweet RA. Heritability of psychosis in Alzheimer disease. Am J Geriatr Psychiatry. 2005; 13(7): 624–627.

50. Sweet RA, Bennett DA, Graff-Radford NR, Mayeux R. Assessment and familial aggregation of psychosis in Alzheimer's disease from the National Institute on Aging Late Onset Alzheimer's Disease Family Study. Brain. 2010; 133(Pt 4):1155–1162.
51. https://www.healingwell.com/library/parkinsons/info1.aspx (Accessed on date: 8 August 2018)
52. Bostantjopoulou S, Katsarou Z, Papadimitriou A, Veletza V, et al.Clinical features of parkinsonian patients with the alpha-synuclein (G209A) mutation. *Mov Disord* 2001; **16**: 1007–1013.
53. Gasser T. Genetics of Parkinson's disease. *J Neurol* 2001; **248**: 833–840.
54. Rademakers R, Cruts M, van Broeckhoven C. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. *Hum Mutat* 2004; **24**: 277–295.
55. Giasson BI, Forman MS, Higuchi M, Golbe LI, et al. Initiation and synergistic fibrillization of tau and alpha-synuclein. *Science* 2003; **300**: 636–640.
56. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet. 2011 Feb 19; 377(9766): 641-9.
57. International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2). A two-stage meta-analysis identifies several new loci for Parkinson's disease. PLoS Genet. 2011 Jun; 7(6): e1002142.
58. Lill CM. Genetics of Parkinson's disease. Mol Cell Probes. 2016 Dec; 30(6): 386-396.
59. Kumar KR, Djarmati-Westenberger A, Grünewald A. Genetics of Parkinson's disease. Semin Neurol. 2011 Nov; 31(5): 433-40.
60. Dooneief G, Mirabello E, Bell K, Marder K, et al. An estimate of the incidence of depression in idiopathic Parkinson's disease. Arch Neurol 1992; 49: 305–307.
61. Starkstein SE, Mayberg HS, Leiguarda R, Preziosi TJ, et al. A prospective longitudinal study of depression, cognitive decline, and physical impairments in

patients with Parkinson's disease. J Neurol Neurosurg Psychiatry 1992; 55: 377–382.

62. Tandberg E, Larsen JP, Aarsland D, Cummings JL. The occurrence of depression in Parkinson's disease. A community-based study. Arch Neurol 1996; 53: 175–179.
63. Emre M. Dementia associated with Parkinson's disease. Lancet Neurol 2003; 2: 229–237.
64. ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, *et al.* A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. 9(4): e1001046 (2011)
65. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet. 38(6): 617-9(2006)
66. Zhang X, Bailey SD, Lupien M. Laying a solid foundation for Manhattan – 'setting the functional basis for the post-GWAS era'. Trends Genet. 30(4): 140-9 (2014)
67. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, *et al.* A 'silent' polymorphism in the MDR1 gene changes substrate specificity. Science. 315(5811): 525-8 (2007)
68. Komar AA. Genetics. SNPs, silent but not invisible. Science. 315(5811): 466-7 (2007)
69. Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. Science. 342(6164): 1367-72 (2013)
70. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 6(5): 386-98 (2005)
71. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci. 25(3): 106-10 (2000)
72. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 3(4): 285-98 (2002)
73. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 297(5583): 1007-13 (2002)

74. Gregory AP, Dendrou CA, Attfield KE, Haghikia A, Xifara DK, Butter F, *et al.* TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature. 488(7412): 508-11 (2012)
75. Holwerda S, de Laat W. Chromatin loops, gene positioning, and gene expression. Front Genet. 3:217 (2012)
76. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 10(4): 241-51(2009)
77. Docherty SJ, Davis OS, Haworth CM, Plomin R, D'Souza U, Mill J. A genetic association study of DNA methylation levels in the DRD4 gene region finds associations with nearby SNPs. Behav Brain Funct. 12; 8:31(2012)
78. Sribudiani Y, Metzger M, Osinga J, Rey A, Burns AJ, Thapar N, *et al.* Variants in RET associated with Hirschsprung's disease affect binding of transcription factors and gene expression. Gastroenterology. 140(2): 572-582.e2 (2011)
79. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol. 30(6): 1411-20 (2010)
80. Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, *et al.* A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. 43(3): 242-5 (2011)
81. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 128(4): 683-9(2007)
82. Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 12(4): 283-93 (2011)
83. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res 22(9): 1748-59 (2012)
84. Cowper-Sal lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet. 44(11): 1191-8 (2012)
85. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, *et al.* Super-enhancers in the control of cell identity and disease. Cell. 155(4): 934-47 (2013)
86. Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, *et al.* Chromatin stretch enhancer states drive cell- specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A. 110(44): 17921-6 (2013)

87. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science. 312(5777): 1215-7(2006)
88. Huang Y, Yang H, Borg BB, Su X, Rhodes SL, Yang K, et al. A functional SNP of interferon-gamma gene is important for interferon-alpha-induced and spontaneous recovery from hepatitis C virus infection. Proc Natl Acad Sci U S A. 104(3): 985-90 (2007)
89. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 470(7333): 264-8 (2011)
90. Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science. 342(6155): 253-7 (2013)
91. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 466(7307): 714-9 (2010)
92. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 41(8): 885-90 (2009)
93. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 41(8): 882-4 (2009)
94. Zhang X, Cowper-Sal lari R, Bailey SD, Moore JH, Lupien M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. 22(8):1437-46 (2012)
95. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 24(1): 1-13 (2014)
96. Bickmore WA. The spatial organization of the human genome. Annu Rev Genomics Hum Genet. 14:67-84 (2013)
97. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 447(7143): 413-7 (2007)

98. Gibcus JH, Dekker J. The hierarchy of the 3D genome. Mol Cell. 49(5): 773-82 (2013)
99. Misteli T. Beyond the sequence: cellular organization of genome function. Cell. 128(4): 787-800 (2007)
100. Roix JJ, McQueen PG, Munson PJ, Parada LA, Misteli T. Spatial proximity of translocation-prone gene loci in human lymphomas. Nat Genet. 34(3): 287-91 (2003)
101. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 489(7414): 109-13 (2012)
102. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 148(1-2): 84-98 (2012)
103. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 10(6): 1453-65(2002)
104. Lanzuolo C, Roure V, Dekker J, Bantignies F, Orlando V. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. Nat Cell Biol. 9(10): 1167-74(2007)
105. Spilianakis CG, Flavell RA. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nat Immunol. 5(10): 1017-27(2004)
106. Sexton T, Bantignies F, Cavalli G. Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. Semin Cell Dev Biol. 20(7): 849-55(2009)
107. Bartel B. MicroRNAs directing siRNA biogenesis. Nat Struct Mol Biol. 2005 Jul; 12(7): 569-71.
108. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 458(7235): 223-7(2009)
109. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 129(7): 1311-23 (2007)
110. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 329(5992): 689-93 (2010)

111. Shen LX, Basilion JP, Stanton VP Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci U S A. 96(14): 7871-6 (1999)

112. Phillips PC. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet.Nov;9(11):855-67(2008)

113. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nat Rev Genet. 8(6): 437-49 (2007)

114. Bateson W. Mendel's Principles of Heredity. Cambridge Univ. Press; Cambridge: (1909)

115. Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinb.52: 399–433 (1918)

116. Moutsianas L. et al The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 4, e1005165 (2015).

117. Parker, S. C. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci USA 110, 17921–17926 (2013).

118. Kulzer, J. R. et al. A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. Am J Hum Genet 94, 186–197 (2014).

119. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. PLoS Genet 10, e1004633 (2014).

120. Claussnitzer, M. et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med 373, 895–907 (2015).

121. Li MJ, Liu Z, Wang P, Wong MP, Nelson MR, Kocher JP, Yeager M, Sham PC, Chanock SJ, Xia Z, Wang J. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 2016 Jan 4; 44(D1):D869-76.

122. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12): e1002822.

123. Bader JS. The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics 2001; 2: 11–24.

124. Zhao N, Han JG, Shyu CR, et al. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. PLoS Comput Biol. 2014 May 1; 10(5):e1003592.

125. Burga, A., and Lehner, B. (2013). Predicting phenotypic variation from genotypes phenotypes and a combination of the two. Curr. Opin. Biotechnol. 24, 803–809.

126. Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. Nat. Rev. Genet. 14, 168–178.

127. Achard F., Vaysseix G, Barillot E. XML, bioinformatics and data integration. Bioinformatics 2001; 17:115-125.

128. McEntire R., Karp P, Abernethy N, et al. An evaluation of ontology exchange languages for bioinformatics. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2000; 8:239-250.

129. http://www.openbel.org/bel-expression-language

130. Catlett NL, Bargnesi AJ, Ungerer S, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics. 2013 Nov 23; 14:340.

131. Slater T. Recent advances in modeling languages for pathway maps and computable biological networks. Drug Discov Today. 2014 Feb; 19(2):193-8.

132. http://dx. doi.org/10.7303/syn2290704 (Accessed on 10-12-2018)

133. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. Neuroimaging clinics of North America 2005; 15:869–77. xi-xii.

134. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Current Alzheimer research 2012; 9:628–45.

135. Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS. Overview and findings from the rush Memory and Aging Project. Current Alzheimer research 2012; 9:646–63.

136. Lovestone S, Francis P, Kloszewska I, Mecocci P, Simmons A, Soininen H, et al. AddNeuroMed–the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Annals of the New York Academy of Sciences 2009; 1180:36–46.

137. Alzheimer's Disease Neuroimaging Initiative. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. Alzheimers Dement. 2016 Jun; 12(6):645-53.

138. Ridge PG, Mukherjee S, Crane PK, Kauwe JS, Alzheimer's Disease Genetics Consortium. Alzheimer's disease: analyzing the missing heritability. PloS one 2013; 8:e79771.

139. 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 : a journal of neurology 2015; 138(Pt 12):3673–84.

140. Lee SH, Harold D, Nyholt DR, ANZGene Consortium, International Endogene Consortium, Genetic and Environmental Risk for Alzheimer's disease Consortium, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet 2013; 22:832–41.

141. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature genetics 2013; 45:400–5. 5e1–3.

142. Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet 2013; 14:549–58.

143. http://openbel.org

144. http://www.hgvs.org

Goal of the thesis

The goal of this thesis is to design an integrative approach to interpret GWAS identified variants for neurodegenerative diseases based on their functional consequences. This work is limited to the most frequently occurring neurodegenerative diseases, AD and PD. This knowledge based integrative mapping and interpretation approach can facilitate the discovery novel links between genetic mutation and complex genetic diseases. Such links are indispensable to truly understand the genetic effect in the biological pathways of neurodegenerative diseases.

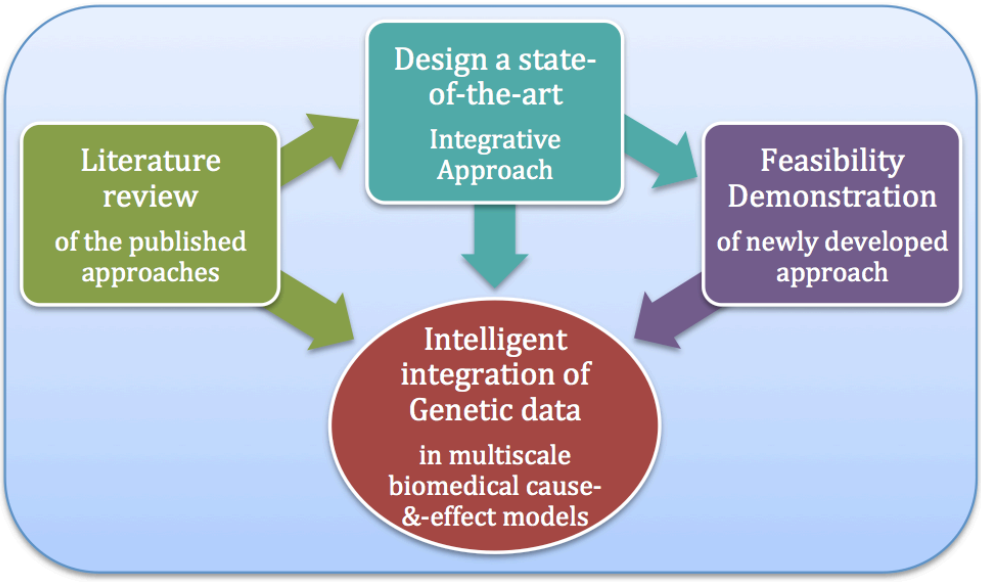


Figure 19: Workflow for the intelligent functional interpretation of genetic variance information in multi-scale biomedical cause-and-effect models

1. **Literature review to evaluate a network modelling approaches for integration of genetic variants in a systems biomedicine context**

   The aim of the systematic literature review is to provide a recapitulation of all known functional consequences of genetic variants and an exhaustive assessment of their perspectives. This review facilitated the identification of an appropriate methodology for modelling and simulations of genetic regulatory networks, which can predict the functional consequences of disease-associated genomic loci.

2. **Design an extension for evaluated systems biomedicine modelling language to integrate genetic variance information**

   The second major goal was to design an extension for evaluated systems biomedicine modelling language. This extension allows for the integration of genetic variation information with mechanistic molecular level knowledge. Moreover, it supports

scientific findings by capturing causal and correlative relationships in context and applying reasoning over cause-and-effect models.

3. **Establish an integrative approach that allows for new insights into complex mechanisms through data and knowledge-driven context enrichment**

   The subsequent task was to establish an integrative data-driven approach that finds indicators from genetic data and allows for new understanding of complex mechanisms through knowledge-driven context enrichment. This approach can take multimodal information into account and integrates heterogeneous biomedicine knowledge into computable graph models.

Chapter 1

Genetic variant data and their functional interpretation in biological context

Introduction

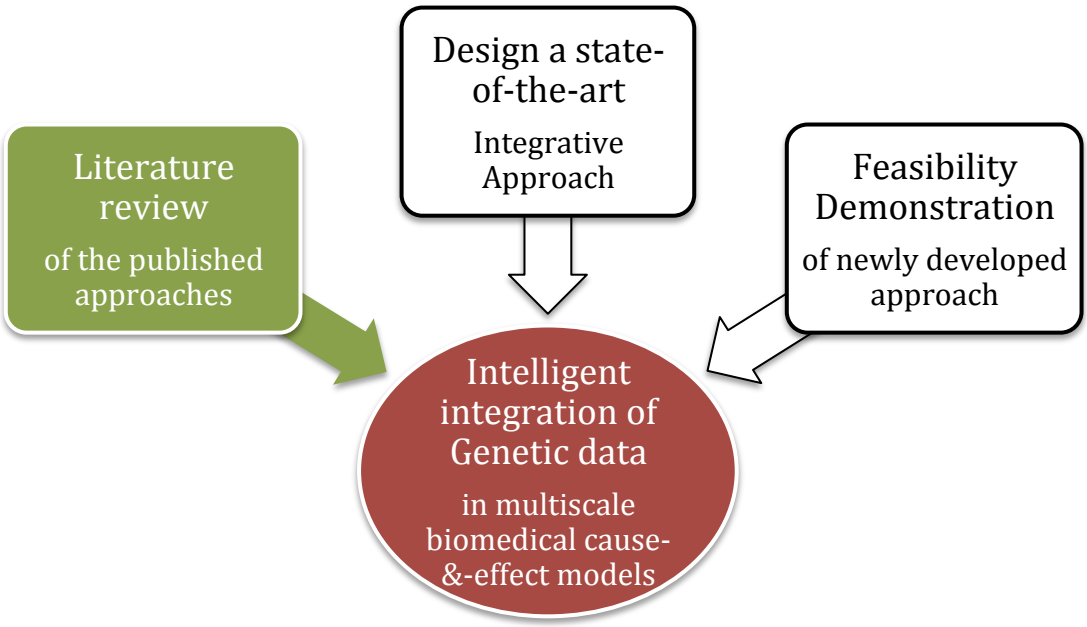


Figure 20: Part –I: Workflow for the intelligent functional interpretation of genetic variance information in multi-scale biomedical cause-and-effect models

Genome-wide association studies (GWAS) have established an important approach in human genetics. In total, GWAS are possibly the largest biological investigations of humans ever conducted. The total number of people who are genotyped with a GWAS array is difficult to know, but undoubtedly exceeds 1 million. Major findings from these studies are include the following: a) many common diseases have a polygenic architecture, b) the genetic effect sizes of common SNP variants are small,

c) the identification of the involvement of genes and biological processes not previously suspected, and d) the association of some loci with different diseases.

The ultimate goal for the post-GWAS era is to highlight those specific genetic variants from a risk-associated locus, which account for phenotypic differences based on the functional biology it modulates. Even for coding region variants, it is often not clear whether they are functional, due to the presence of several closely linked variants. Many statistical methods have been proposed to prioritize GWAS signals by incorporating diverse functional evidence. GWAS-identified variants can be prioritized at both the SNP level and gene level, depending on the biological features considered and the input signals.

In this review, biomedical literature is summarized for assessment of the functional impact of genetic variation at molecular level. The review focuses on the interpretation of SNPs and mutations in a systems biology context with a strong link to network modelling approaches.

GWAS genetic variant data and their integration in the context of network biology

Mufassra Naz[1, 2], Martin Hofmann-Apitius [1, 2 *]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53754 and [2]Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for IT, Dahlmannstrasse 2, 53113 Bonn, Germany

Corresponding author: Martin Hofmann-Apitius. Head of the Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany. E-mail: martin.hofmann-apitius@scai.fraunhofer.de

J Syst Integr Neurosci, 2016 Volume 2(4): 189-202
doi: 10.15761/JSIN.1000135

Key words: GWAS, genetic variants, SNP, network biology, variant's functional consequences

Abstract

Regardless of the success of Genome Wide Association Studies (GWAS) to identify genetic variants associated with human diseases, investigating the molecular mechanisms and disease-associated genes linked to those genetic variants, is a very complex task. Specifically, where intergenic genetic variants are linked to the adjacent neighbouring genes. Consequently, the inference for the mechanistic connection between diseases and its susceptible genetic variants becomes more challenging.

Functional genomics studies can support to reveal the significance of variants via intermediate molecular traits. Moreover, approaches like computational and bioinformatics predictions based on the variants location and its sequence attributes can assist to propose the candidate genes. As, the spectrum of potential functional consequences of variants is much broader; it still requires new methodologies to predict any molecular level perturbation. Thus, specialized algorithms and computable modelling approaches are essential, for the modelling and simulation of genetic regulatory networks.

In this review, we are briefly summarizing all the existing methodologies for genome wide association studies, currently available algorithms and computable modelling approaches; moreover also emphasizing the required new approaches for modelling

and simulations of genetic regulatory networks to predict the functional consequences of disease-associated genetic variants.

Key words: GWAS, Genetic variants, SNP, Network biology, Variant's functional consequences, Alzheimer's disease genetics

Introduction

Genome-wide association studies (GWAS) are well established in human genetics. In total, GWAS are possibly the largest molecular biology investigations of human beings ever conducted. The total number of people, who have been genotyped in GWAS studies, exceeds 1 Million. Major insights have been possible based on GWAS studies:

a. Many common diseases have a polygenic architecture,
b. The genetic effect sizes of common Single nucleotide polymorphism (SNP) variants are small,
c. The identification of the involvement of genes and biological processes not previously suspected, and
d. The association of some loci with different diseases.

GWAS have identified thousands of SNPs, known as lead-SNPs, which are associated with hundreds of human traits and diseases [1, 2]. These lead-SNPs capture the variation present at risk-associated loci, but do not necessarily represent causal genetic variants that underlie the molecular mechanism of the association [1]. With the original lead-SNP, a collection of genetic variants at each risk-associated locus, all putatively causal, are in linkage disequilibrium (LD) according to the initial design of the GWAS studies [3, 4]. Those genetic variants, which are within a risk-associated locus and in strong LD with the lead-SNP could account for the observed difference in phenotype associated with that locus.

The ultimate goal for the post-GWAS era is to highlight those specific genetic variants identified within a risk-associated locus that account for phenotypic differences based on the functional biology they modulate. However, more than 88% of disease-associated variants fall into non-coding regions of the genome [1], which makes it extremely challenging to generate testable hypotheses about the functional

involvement of neighbouring genes. Even for SNPs in genic regions, it remains often unclear, whether they are functional due to the presence of several closely linked variants. A variety of statistical methods have been proposed to prioritize GWAS signals by incorporating diverse functional evidence [5]. GWAS identified variants can be prioritized at both, the SNP level and gene level, depending on the biological features considered and the input signals available.

Until recently, the functional characterization of risk-associated loci was limited by the incomplete annotation of non-coding sequences in the human genome. Population-based studies have revealed that non-coding genetic variants are linked with gene expression [6–9], RNA splicing [10], transcription factor binding [11], chromatin openness measured by DNase I hypersensitivity [12], DNA methylation [13], and histone modifications [14–16]. Additionally, SNPs are more commonly linked with a particular phenotype if they fall within a DNase I hypersensitive region from a disease relevant cell type [17].

Likewise, with the integration of other data informative about trait association (like gene expression, expression quantitative trait loci (eQTL) and others), the prioritized genes/loci are more likely to be truly associated with a trait. For instance, there is accumulating evidence that trait-associated loci are more intense in regions with certain genomic features, such as protein coding regions and eQTL [5].

A series of large-scale genomics projects, including the Encyclopedia of DNA Elements (ENCODE) [18, 19], the International Human Epigenome Consortium (IHEC) [20], the Roadmap Epigenomics [21] and the Functional Annotation of the Mammalian Genome (FANTOM) [22] projects, as well as independent labs have undertaken significant effort to systematically annotate non-coding regions of the human genome in several different cell and tissue types and across several developmental stages.

These large-scale studies have profited from advances in next generation sequencing technologies to generate genome-wide maps of functional elements, such as origins of replication, transcripts and regulatory elements. RNA-sequencing (RNA-seq) and cap analysis of gene expression sequencing (CAGE-seq) approaches led to the identification and annotation of known as well as novel transcripts such as long non-

coding RNA (lncRNA) and enhancer RNA (eRNA) [23–25]. Whole-genome epigenetic mapping (WGEM) for histone modifications through chromatin immune-precipitation sequencing (ChIP-seq) identifies regulatory elements including promoters, enhancers, and insulators [26–30].

Moreover, inter-species evolutionarily conserved DNA sequences can complement these maps by predicting potential functional DNA elements [31,32]. Taken together, such biological information, across the human genome, assist as the foundation for post-GWAS functional studies.

Genetic Variants and their detection Power

Genome wide association studies (GWAS): Over the last years, GWAS have established as popular approaches for the identification of genetic variants that are associated with disease risk loci. In a standard GWAS study design; a case control comparison to assess the association between each individual genotyped SNP and disease risk is performed. Very often, a discovery phase in which an initial set of promising susceptibility loci is identified, is followed by a confirmation stage in which the SNPs identified in the initial stage are replicated in a separate study cohort [33]. The standard methodology for analyzing GWAS in the discovery phase consist of individual SNP analysis, then SNPs are ranked on the basis of their individual p-values and a threshold is set such that all SNPs with p-value less than that threshold will be validated further.

However, with this individual-SNP analysis, reproducibility is very limited, since multiple high-ranked SNPs in the discovery phases are false positives and cannot be verified [34]. Besides, the true causal SNP (if it exists at all) is rarely genotyped; instead, other typed SNPs which are in linkage disequilibrium (LD) with the causal SNP, are being measured and these "related SNPs" may show only moderate effects at mechanistic level and – as a consequence – moderate association with the disease phenotype. Therefore, a locus-centric analysis could be beneficial to consider the joint effect of multiple SNPs in analysis as it is likely that several of these markers are in LD with the causal SNP and could show the true effect more effectively [35]. Additionally, individual SNP analysis only considers the marginal effect of each SNP

and cannot detect epistatic effects. Epistatic interactions between SNPs can contribute to disease susceptibility [36].

The statistical power of a GWAS is a function of sample size, effect size, causal allele frequency, and marker allele frequency and its correlation with the causal variant [37]. Because GWASs are underpowered to detect associations of modest effect sizes (odds ratio (OR) = 1.1–1.5) [38][39][40], large population samples are required to detect variants of even moderate effect (OR = 1.5–2). Meta-analyses of independent GWASs for a trait reap the full benefit of GWASs that have already been performed, greatly increasing sample size and statistical power. When different GWASs use different genotyping platforms, only a minority of the SNPs are in common to all platforms. Imputation methods have been developed to infer genotypes at un-typed SNPs using a reference panel of more densely genotyped samples [41]. After imputation, GWAS results can be combined across multiple studies [42].

For **meta-analysis**, it would be ideal to include the raw data as a covariate for all studies contributing to the analysis, but meta-analysis could also be done without the use of the raw genotypes. It calculates the effect size that each study attributes to the genetic variant and weighted according to the relevant study size. In such analysis, small studies contribute less than large studies because they are likely to give less accurate effect-size estimates [43]. The significance of any given effect size can be determined by the size of the sample studied. The simple equation is:

Significance Test = Effect Size x Study Size

As an alternative, **a natural grouping strategy** has been proposed. This approach is based on the grouping of SNPs into SNP sets based on proximity to genomic features such as genes or haplotype blocks; it can significantly reduce the number of multiple comparisons [34]. An extension of gene-based SNP set analysis is to group SNPs based on whether they are located within a pathway represented in Kyoto Encyclopedia of Genes and Genomes (KEGG) [44] or a Gene Ontology Consortium functional category [45]. Even though, making inference on a pathway further reduces the number of multiple comparisons, but it still allows inference on a biologically meaningful unit [34]. It is noteworthy in this context, that the functional context represented by pathways (e.g. in KEGG) can be expanded towards entire computable disease models (e.g. in Biological Expression Language (BEL) [46].

Functional impact of Genetic Variants at Molecular level

The functional impact of SNPs should be closely linked to their interference with (or modulation of) normal physiological functions. As, some SNPs are very likely to directly interfere with bio-molecular functions of genes and genomic regions whereas other SNPs can only convey susceptibility of human diseases by yet unknown mechanisms [47].

Following section describes the different functional categories that can be articulated as "mode-of-SNP-action" classes.

Genetic variants on coding regions

Protein Coding SNPs have been most extensively studied due to their direct effect on the function of that encoded protein.

1. **Non-synonymous genetic variants:** Proteins have a unique sequence of amino acids specified by the coding DNA, and a modification to its sequence can significantly impact its function [48]. The risk associated with non-synonymous genetic variants (nonsense or missense) can easily be translated into a change in protein structure or function due to change in amino acid sequence. Non-synonymous SNPs can modify amino acid composition, or truncate the protein sequence by causing an early codon [49]. Indels (insertion or deletion of nucleotide base(s)) can also alter protein sequence with varying consequence depending on whether the indel is in-frame or frame-shifting, and this substitution may affect protein folding, proper activity of binding or interaction sites, structure, stability or solubility of the protein. For example, the rs1990760 SNP associated with type 1 diabetes (T1D), is an example of a non-synonymous genetic risk variant of IFIH1 (interferon induced with helicase C domain 1) gene, causing an alanine to threonine substitution at position 946 (A946T) of the IFIH1/MDA5 protein [49].

2. **Synonymous genetic variants:** Synonymous genetic variants do not alter the codon sequence and consequently cannot encode any change in protein sequences. However, synonymous genetic risk variants can still impact protein function by

modulating translation rates with direct consequences to protein folding [50]. As an example, we will discuss here the rs1045642 SNP that maps to the MDR1 (Multidrug Resistant-1) gene [51,52]. The MDR1 gene (ABCB1 - relevant human gene) encodes a cell membrane transporter protein involved in drug trafficking [53] and the rs1045642 SNP changes the drug substrate specificity of MDR1 but does not influence the sequence or the expression of the MDR1 protein [52]. Due to the rs1045642 SNP, the frequent isoleucine (Ile) codon ATC replaces by the rare Ile codon ATT [52]. It has been suggested that this alteration slows down the rate of translation of the MDR1 mRNA, and this impacts protein folding [54], and that the subsequent altered MDR1 conformation decreases its drug substrate specificity [51–53]. It has also been shown that a fraction of codons specify not only an amino acid, but a transcription factor binding site, providing an additional avenue through which synonymous polymorphisms may impart a functional effect [55].

3. **Splice Site Genetic Variants:** Splicing is a process, in which introns are excised and exons are joined, at RNA sequence level [56]. Exonic splicing enhancers (ESEs) comprise specific hexamer sequences and an AG sequence at the intron-exon borderline, that instruct for the recruitment of the splicing complex to immature RNA (pre-mRNA) and lead for intron excision and exon joining. SNPs may also present within exon splicing enhancers or silencers (ESEs/ESSs). ESEs and ESSs are typically 6 to 8 consecutive nucleotide sequences in an exon region. Similar to the SNPs occurring in splice sites, SNPs within ESEs or ESSs can also result in deleterious intron retention or exon skipping [56–59]. SNPs and indels can also interrupt splicing sites to translate the protein isoform. A mechanistic insight, how a SNP can affect splicing, is provided through the rs1800693 SNP example. This SNP is located at the edge of exon/intron of the TNFRSF1A (tumour necrosis factor receptor superfamily member 1A) gene and is associated with multiple sclerosis. The SNP affects the splicing of the TNFRSF1A mRNA and leading to translate an isoform [60].

Genetic Variants on Non-coding regions

Mammalian regulatory interactions can take place over significant chromosomal distances up to an entire megabase (1MB) [61]. Genetic risk variants are very

frequent on non-coding sequences [62]. Post-GWAS studies have revealed the capacity of these genetic risk variants to regulate gene expression by modulating cis-regulatory machineries through mechanisms involving DNA methylation [63], transcription factor binding [64], chromatin looping [65], or miRNA recruitment [66]. Databases that provide information of experimentally verified transcriptional regulatory regions can be used to identify SNPs that can alter gene expression like HTRIdb (http://www.lbbc.ibb.unesp.br/htri/) [67].

1. **DNA methylation and Genetic variants at promoters:** DNA methylation means addition of methyl groups to a cytosine nucleotide, which is basically part of a CpG dinucleotide. This DNA methylation is a heritable epigenetic event, which is involved in transcriptional regulation [68]. DNA hyper-methylation near transcription start sites (TSS) of tumour suppressor genes associates with their silencing [68]. For instance, the HNF1B (hepatocyte nuclear factor 1 homeo-box B) gene is silenced by DNA methylation in serous ovarian tumours. The rs7405776 SNP defines a risk locus for intrusive serous ovarian cancer that is located within the promoter region of the HNF1B gene. This risk-associated locus, at the HNF1B gene promoter region, is located in a CpG island and is associated with higher DNA methylation levels [10].

2. **Transcription factor binding to the chromatin and Genetic variants:** Across the genome, transcription factors bind to thousands of regulatory elements, including promoters directly upstream of their target genes and cis-regulatory elements such as enhancers, insulators and silencers [69]. ChIP-seq assays for transcription factors effectively annotate these cis-regulatory elements genome-wide. Analysis of these annotations reveals that genetic risk variants commonly target cis-regulatory elements, mainly enhancers, in a disease- and tissue-specific manner [17, 27, 70-73]. For example, loci associated with erythrocyte phenotypes commonly harbour enhancers that are functional in K562 erythrocyte leukemia cells, but not enhancers that are functional in other cell types [27].

Genetic risk variants located within promoter regions can also change transcription factor binding to DNA, leading to differential target gene expression [74, 75]. For example, expression of the a-globin gene locus is affected by a genetic variant associated with the a-thalassemia blood disorder [74]. That genetic variant creates a

GATA1 motif at a promoter-like region that down-regulates the expression of the downstream a-globin genes [74]. Down-regulation of a-globin genes promotes a-thalassemia [76].

Enhancers are commonly targeted by those genetic variants of risk-associated loci that map to DNA recognition motifs, bound by transcription factors. These genetic variants can modulate the chromatin affinity for transcription factors and consequently gene expression [77–82]. One example for this type of functional impact is the rs1427407 SNP, which is associated with fetal hemoglobin level. It decreases the recruitment of the GATA1 (GATA binding protein 1)/TAL1 (T cell acute lymphocytic leukemia 1) nuclear complex to the enhancer region, and results in lower levels of expression for the BCL11A (B cell CLL/lymphoma 11A) gene, a repressor of the fetal hemoglobin level [78]. Likewise, the rs12740374 SNP, which is associated with a lower level of plasma low-density lipoprotein cholesterol (LDL-C), shows higher expression level of the SORT1 (sortilin 1) gene by increasing the binding affinity of the C/EBP (CCAAT enhancer-binding protein) transcription factor to chromatin [79]. Over-expression of SORT1 leads to a lower LDL-C level in livers [79]. Moreover, functional variants within a single risk locus can modulate multiple different enhancers. This multi-enhancer variant phenomenon was found to be a fundamental feature of many risk loci [83].

3. **Chromatin loop formation bridging enhancers and promoters and Genetic variants:** Genetic risk variant can modulate chromatin loop formation, it can alter the DNA affinity for looping factors, which can also result in allele-specific chromatin loop formation. The human genome is structured in a three dimensional architecture which is thought to regulate a diverse set of DNA-templated processes [84–88]. This facilitates regulatory elements, like promoters and enhancers, to interact physically through long-range chromatin loops, or chromatin interactions, to regulate gene expression [89, 90]. This has been shown for the rs12913832 SNP, which resides in an enhancer 21 kb upstream of the OCA2 (Oculocutaneous albinism II) pigment gene. This particular SNP is a human pigmentation-associated SNP, which interferes with (modulates) allele-specific chromatin loop formation [91].

Recent studies have analyzed CTCF (CCCTC binding factor) [92] and cohesin [93,94] binding sites, DNase-hypersensitive sites [95] and putative enhancers [96] on

a genome-wide scale. If a minor fraction of these potential regulatory elements participate in chromatin looping, then most of the genomic interactions have yet to be characterized again, because many such loops appear to be tissue-specific [97-99], which makes their comprehensive analysis appear even more disconcerting [100].

4. **Genetic variants and miRNAs:** MicroRNAs (miRNAs) target mRNAs by recognizing their complementary sequences mainly in 3' untranslated regions (3'UTRs). miRNAs largely function as post-transcriptional repressors. They recruit RNA-induced silencing complex (RISC) to their target mRNAs, leading to mRNA degradation or translation repression [101]. They can regulate the translation of hundreds of genes through sequence-specific binding to mRNA [102]. Abelson et al. showed that SNPs linked to miRNA can affect disease phenotype, they identified a mutation, residing in the 'miR-189' binding site of gene SLITRK1 (SLIT and NTRK-like protein 1) that was associated with Tourette's syndrome [103].

SNP variants, linked with miRNAs, can affect gene functionality with three different ways: 1) by transcription of primary transcript, 2) by pri-microRNA and pre-microRNA processing and 3) by effecting the microRNA- microRNA interaction [104]. For instance, SNPs, reside in the pri regions of let-7e and mir-16, reduce the levels of mature micRNA [105, 106]. Thus, SNPs located in miRNA binding site of target mRNAs can interrupt miRNA-dependent regulation and eventually effect gene expression in cancer, like a miRNA from let-7 family binds to 3'UTR region of the gene RAS and regulates its expression level [107]. For example, the rs100672, a Crohn's disease-associated SNP, lies within the 3' UTR of the IRGM (immunity-related GTPase M) gene and this risk allele alters the complementary target sequence of miRNA-196 [78]. This reduces miRNA-196 binding to the IRGM mRNA increasing the stability of the IRGM mRNA and protein levels [78, 108].

Tools such as RegRNA 2.0 and miRBase (the microRNA database) can predict how genetic variants impact miRNA target specificity [78, 109].

5. **Genetic variants and Long non-coding RNAs (lncRNAs):** lncRNAs are non-protein-coding transcripts which could be longer than 200 nucleotides in length. lncRNAs are found across intergenic regions of the human genome [23]. They can interact with chromatin regulators for their recruitment by chromatin [110, 111], a

process, which relies on a highly conserved lncRNA tertiary structure. Though, lncRNA tertiary structures can be changed by genetic risk variants [112]. The 9q21.3 (coronary artery disease) and 22q12.1 (myocardial infarction) risk loci have SNPs associated with the ANRIL and MIAT (myocardial infarction associated transcript) lncRNAs, respectively [113,114]. The risk SNP rs35955962 is located in the MIAT lncRNA, that increases its affinity for nuclear proteins [114].

The fundamental question about the effective distance between influential regulatory elements and target genes has not yet been answered. However, regulatory elements (like enhancers) necessary for tissue-specific gene expression have been identified at megabase (1MB) distances from their target genes, and have been shown to physically interact with them [115, 116].

Integrative functional post-GWAS methodologies

Bioinformatics tools/methodologies and integrative functional genomics that combine GWAS data, linkage disequilibrium, and whole-genome functional annotations can provide a means to identify the targets of risk-associated loci [17,27,70,71]. Such tools can be employed to predict the biological impact of genetic risk variants and identify putative causal genetic variant responsible for risk loci.

1. **Protein Deleteriousness Predictions:** Many computational tools have been developed to predict "deleteriousness" of SNPs and indels [117, 118]. These methods generally take features like biochemical property of the altered amino acid, conservation and sequence homology as input, and use machine-learning technique to train a classifier. The most extreme case of protein function interruption is the loss of function mutation. However, genome-sequencing studies found that all human carry loss of function mutations without obvious phenotypic effect, and such common loss of function variants were depleted in polymorphisms associated with complex disease like Crohn's disease and rheumatoid arthritis [119]. The results indicate that the "deleteriousness" feature should be interpreted with caution, since disruption of protein function does not necessarily have a phenotypic effect. In this regard, the "residual variance intolerance score" has been defined quantitatively measure the tolerance of a protein to mutations [120]. Numerous tools have been developed to predict the putative deleterious effects of non-synonymous SNPs that cause an amino

acid change in a translated protein including SIFT (http://sift.jcvi.org/) [110], *PolyPhen-2* (Polymorphism Phenotyping v2) (http://genetics.bwh.harvard.edu/pph2/) [111]. Tools like, PolyPhen and MuTIP predict changes in protein structure imposed by genetic risk variants mapping to coding regions [118,121].

2. **DNA recognition motifs to modulate transcription factor binding:** Motif-prediction tools, such as HaploReg, RegulomeDB, FunSeq, and SnpEff, identify genetic variants that significantly alter DNA recognition motifs to modulate transcription factor binding [122–125]. The intra-genomic replicates (IGR) method provides an alternative and can predict changes in chromatin-binding affinity of transcription factors caused by risk variants without the use of position-weighted matrices (PWM) [71].

3. **DNase I Hypersensitive Sites:** DNase I hypersensitive sites (DHSs) are markers of accessible chromatin, which indicate regulatory roles in the transcription process. DHS have been mapped in 349 cell and tissue samples genome-wide by next-generation sequencing [126]. Enrichment analysis showed that trait-associated SNPs are more concentrated within DHS regions, excluding confounding factors such as allele frequency and distance from the nearest transcriptional start site [17].

4. **DNA methylation:** Epigenome data in disease states are valuable for understanding disease and prioritize disease susceptible loci. However, more efforts are needed in disease-specific epigenome mapping studies and the implementation of databases to make such data publicly available. For DNA methylation alone, one database exists, (DiseaseMeth), which has incorporated methylation data for 72 human diseases [127].

5. **Gene expression:** Studying the association between genetic variation and gene expression offers a straightforward way to begin the complicated task of connecting risk variants to their putative target genes. Networks created using gene expression data from patient samples can also model the underlying molecular machinery [128] and can be exploited to bridge GWAS results with an underlying disease mechanism, as exemplified in the autism spectrum disorder [129]. Chen R [130] analysed 476 expression datasets available from Gene Expression Omnibus

(GEO), and calculated the frequency that a gene was differentially expressed in these datasets, which they called "differential expression ratio." They found that differential expression ratio is positively correlated with the likelihood that a gene harbours disease-associated variants, where the list of disease-associated genes was created by combining information from the Genetic Association Database (GAD; [131] and Human Gene Mutation Database (HGMD; [132]). In addition, they found that among the genes discovered in the initial scan of the WTCCC type 1 diabetes mellitus GWAS dataset, the differential expression ratio was higher in genes that were replicable than those not replicable in follow-up studies. These authors have developed an online server, FitSNPs, to incorporate this feature (http://fitsnps.stanford.edu/index.php).

6. **The Encyclopedia of DNA Elements:** There are many more genomic features collected and annotated in large community projects, such as the Encyclopedia of DNA Elements (ENCODE) [47], which are potentially valuable for SNP prioritization. Kindt [133] examined enrichment or depletion of trait-associated SNPs in 58 genomic features. The features investigated covered genic and regulatory features, conservation features, and chromatin state features (see Table 1 in [133]). Among those features, genomic regions annotated as "heterochromatin" and "low expression signals" are depleted of trait-associated SNPs, while eQTLs and "strong enhancer" showed the highest level of enrichment [70].

Genetic Risk Variants' Analyses

1. **Expression Quantitative Trait Loci:** Genetic variation associated with gene expression, known as expression quantitative trait loci (eQTL), can identify the target genes of risk loci [6–9, 134]. Polymorphism situated in DNA regulatory elements can alter the gene transcript frequency. Thus, as a quantitative trait locus, gene transcript frequency can be determined with substantial power [135, 136]. Brem et al. [137] published the first genome-wide study of gene expression in 2002. eQTLs that link locally to adjacent genes, are denoted as cis-eQTLs. Whereas, those that are connected to genes at a distance either on the same or different non-homologous chromosome, are denoted as trans-eQTLs [138]. In most studies, 'cis' (local) has often been defined as being within 1 Mb of the variant under consideration [139].

Typically, cis-eQTLs are more abundant near transcription start sites (TSS) and transcription end sites (TES), and may map with low frequency more than 20kb away from gene [140]. Sometimes, exonic SNPs can also act as cis-eQTLs [140]. Even though, some cis-eQTLs are identified as shared or common eQTLs in different tissue types, trans-eQTLs are mostly dynamic and tissue-dependent [141]. In humans, the effects of cis-eQTLs are usually stronger than those of trans-eQTLs [125, 126].

An analysis of Lymphoblastic Cell Line (LCL) eQTLs has revealed that GWAS identified SNPs, strongly associated with Crohen's disease and these variants have been demonstrated to impact on PTGER4 (prostaglandin receptor 4) expression; a gene located around 270 kb away from the variant region [142].

In recent years, a number of eQTL studies have been executed, to explore the effects of cis and trans-acting variants in human tissues of liver [143], adipose fat [144], [145] and brain [146]. The Genotype-Tissue expression (GTex) project (http://gtexportal.org), proposed and initiated by National Institutes of Health (NIH) (http://www.nih.gov/), promises to make available eQTL information derived from 30 sets of 1000 samples each, representing 30 different tissues for disease genetics [147].

Online tools such as SCAN and the eQTL browser are publicly available to query eQTL data [12,134] and several reviews regarding the application of eQTL studies are available [148,149].-VarySysDB is a public eQTL database that covers around 36,000 loci holding 190,000 annotated mRNA transcripts. Besides SNPs, VarySysDB also includes indel (deletion/insertion) variants from dbSNP, copy number variants (CNVs) from Genomic Variants Database, short tandem repeats and single amino acid repeats from H-InvDB and linkage disequilibrium regions from D-HaploDB [150].

eQTL analysis can also complement pathway-based association approaches that apply prior biological knowledge of genes and pathways to the interpretation of GWAS data [151–155]. Pathway-based tools, such as 'Gene Relationships Among Implicated Loci' (GRAIL), can also identify candidate target genes by identifying genes that are part of a pathway(s) that is enriched within multiple risk-associated loci identified for the same disease [156]. However, pathways are constantly evolving and adapting in parallel with our knowledge of them.

2. **Variant Set Enrichment (VSE) Analysis:** The variant set enrichment (VSE) approach is among a set of first-generation integrative tools that have been developed [71]. It is a permutation-based method that compares the enrichment of genetic risk variant sets within any functional genomic element to randomly generated matched genetic risk variant sets [71,157]. In essence, it is a statistical test that assays for non-randomness. Similar methodologies have associated genetic risk variants from various diseases with specific chromatin states defined by WGEM [27] and regions of open chromatin [17,70]. However, Weng et al. [158] is suggested a SNP Set Enrichment Analysis (SSEA), based on 'Adaptive rank truncated product method', to assign at least one indicative SNP for each gene [158].

3. **Gene Set Enrichment (GSE) Analysis:** In order to prioritize the set of genes mapped with selected SNPs, a Gene set Enrichment analysis could be implemented either on the bases of a gene relevant SNP count or functional scores associated with SNPs or with their connotation with Gene Ontology (GO) biological process [159]. GSE analysis needs multiple data sources, like gene expression, association and linkage studies, literature search, and biological pathways for a list of genes.

WebGestalt is a gene prioritization method, which visualizes and categorizes gene sets in multiple biological contexts, like chromosome distribution, GO, tissue expression pattern, protein domain information, signaling and metabolic pathways and research literature [131]. Another method, Bayesian gene-set analysis (BGSA), is suggested by Shahbaba et al. [160], to evaluate the statistical significance of a specific pathway, based on the posterior distribution of its parallel hyper-parameter. It is a hierarchical Bayesian model, which combines data at the gene level by merging significance measures of SNPs linked with each gene, as well as at the pathway level by linking significance measures of genes relevant to each pathway [161].

4. **Pathway Enrichment Analysis:** Likewise, various methods are implemented to evaluate pathway-based analyses for GWAS data, by taking gene set enrichment from transcriptomic studies into account [162-164], which have been extensively reviewed in the literature [151, 155, 158, 165-172]. These methods could be used to test whether a group of genes in a biological pathway are jointly linked with a disease and different from selective statistics of genes and pathways. For instance, while

using the GSEA framework, to evaluate the statistical significance for permutation and correction in multiple testing, Wang et al. allocated the highest statistic value as the statistic value of the gene, among all SNPs linked to a gene [155]. Another, related method, GSEA-SNP is recommended by Holden et al. [173], which computes all SNPs annotated to a pathway without evaluation of summary statistic at gene-level. While, Chen et al. [168] proposed another approach based on principal component, to identify ‘‘eigenSNPs’’ for each gene to measure their joint association of multiple SNPs. Segre et al. proposed another protocol named as MAGENTA (Meta-Analysis Gene-set Enrichment of variaNT Associations), which can be used for both hypothesis testing and hypotheses generating analyses. By using GWAS results, it tests for genetic association enrichment in a group of functionally related genes or predefined biological processes [153].

ALIGATOR (Association LIst Go AnnoTatOR) method is suggested by Jones et al. [174]. It can be used to check for the overrepresentation of biological pathways, in lists of significant SNPs from GWA studies by using gene-ontology terms as index [169]. Likewise, Zhang et al. [175] developed an analytical framework named as ICSNPathway (Identify candidate Causal SNPs and Pathways) [175], to generate hypothesis of SNP, gene and pathway(s) to reveal the disease mechanism.

5. **Co-expression network:** Undirected and weighted gene networks that characterise the correlation among gene expression levels are known as co-expression networks. In a co-expression network, genes (or probes) are represented by vertices, which measure the expression levels of gene transcripts. While an edge, between two vertices, indicates statistically significant correlation, moreover it is weighted by the correlation coefficient value [176].
Co-expression network can be employed to identify the functional annotation of undefined genes. Integration of eQTL analysis with co-expression network is such an application that is used successfully for this purpose. One key benefit of it is that without prior knowledge, regulatory insights can achieve [177].

6. **Protein-Protein Interaction (PPI) network and Interactome:** Gene set enrichment analysis (GSEA) could be improved by performing on protein-protein interaction network data, which can provide a better way to evaluate GWAS data by measuring the combined effects of multiple markers/genes, while individually that

may have very weak to moderate association effects [178]. In biological functions, like biochemical reactions, signal transduction systems, transcriptional regulation and cytoskeletal structures, binding affinity between proteins is very important; which, can be measured by different high- throughput experimental techniques, like affinity purification-mass spectrometry and two-hybrid system [176].

New analytical approaches are well recognized, in which different data resources are integrated to get their maximum predicting power. Bakir-Gungor et al. proposed a procedure to select functionally significant KEGG pathways by identifying genes within these pathways, where these genes are short-listed through SNP analysis, by initiating with a list of SNPs associated with selective phenotype in GWAS [179]. dmGWAS 2.0, proposed by Jia et al [178], is based on a Dense Module Searching (DMS) methodology [191]. It can annotate relevant genes or sub-network region for complex diseases, by mapping association signals from GWAS datasets into the human PPI network. Particularly for low p-value genes in GWAS data, this DMS method systematically explores the most relevant sub-networks [178].

Moreover, to reveal the most relevant sub-networks for the disease, Liu Y et al. [180] has suggested two discrete approaches and the integration of both approaches is used to discover well-known as well as novel disease relevant genes or biological pathways [180]. PANOGA (Pathway And Network-Oriented GWAS Analysis) is another method proposed by Bakir-Gungor et al. [181] The method sum-ups p-values of GWAS SNPs and aggregates the functional score of SNPs from predictions produced by the SPOT [181] and F-SNP (The Functional Single Nucleotide Polymorphism) web-servers [182]; the resulting score is labelled as 'p_w-values' [179]. PANOGA identifies the SNP associated with the gene that shows the most important functional effect, from all known SNP/gene transcript designations [179].

Iyappan et al. proposed an integrative approach, which takes benefit of the renowned and well-accepted RDF technology to incorporate data from different resources. That approach can be used to complement major heterogeneous resources (like, omics and gene expression data, and literature), to generate hypotheses for causal disease mechanisms. This approach not only can help to tackle the ever-growing data; but also it can support to integrate new data resources without changing the overall frame-work [183].

7. **Epistatic interactions:** Systematically, there are three key categories of epistasis; functional, Compositional and Statistical [184]. Functional epistasis ascertains the molecular interactions that genetic elements have with each another [185]. Compositional epistasis reveals the blocking effect on one allele by another allele at a different locus [186]. Statistical epistasis expresses a quantitative way to detect how the genotype at one locus effects on the phenotype of another locus [187]; it measures deviation from the additive effects of two loci on the phenotype [184]. In the literature, for a pair of SNPs, there are two fundamental tests of epistasis. First one is the 'two-locus interaction test' and the other is 'two-locus association test allowing for interaction' [188].

Mao et al. [189] identified four types of epistasis effects of two candidate gene SNPs with linkage disequilibrium (LD) and Hardy-Weinberg disequilibrium (HWD), i.e. additive × additive, additive × dominance, dominance × additive, and dominance × dominance [190]. Zhang et al. proposed another algorithm, TEAM (Tree-based Epistasis Association Mapping), which is exhaustive (i.e. check all epistatic interaction). The TEAM algorithm uses the MST (minimum spanning tree) structure; and without perusing all individuals, it updates the contingency tables on incremental bases for epistatic tests. [191]. Emily et al. [192] proposed a statistic method, named as IndOR (independence-based odds ratio), based on the biologically functional epistasis.

Piriyapongsa et al. presented iLOCi (Interacting Loci), a SNP interaction prioritization algorithm. iLOCi identifies marker dependencies discretely for case and control groups and ranked them by calculating the difference in marker dependencies for all possible pairs of case and control groups [193]. Arkin et al. [194] presented an algorithm named as EPIQ (EPIstasis detection for Quantitative GWAS) for the detection of epistasis in quantitative GWAS data. EPIQ discovers SNPs with epistatic effect, without exhaustively testing all pairs of SNPs [194].

Case Study: GWAS and Alzheimer's disease

Over the past few years, in the field of Alzheimer's disease like many other complex and genetically heterogeneous diseases; the application of GWA screening to reveal novel susceptibility genes has attained substantial momentum. Beyond the well-

known APOE association, more than two-dozen novel susceptibility loci are identified by these GWA studies [195].

Familial and Sporadic Alzheimer's disease

Alzheimer's disease is the most common form of dementia and it is linked with 'complex' and multifactorial genetic characteristics. AD can be categorized into two major genetic etiologies, the familial AD form and the sporadic form. Familial AD typically exhibits an early age of onset (50-65 years) and follows a mendelian way of disease transmission; while sporadic AD shows no evident familial aggregation and typically it is associated with relatively late-onset age (beyond 65 years). The familial form of AD is usually caused by rare and highly penetrant mutation in the genes of APP, PSEN1 and PSEN2. GWA studies have identified more than 200 mutations within these three genes [196,197]. These genes are linked with the dysfunctioning in amyloid-β peptide (Aβ) production, that is a key element of β-amyloid in senile plaques [198]. Indubitably numerous other potential disease-causing genes still need to be discovered for familial AD, however this type of genetically determined AD accounts only for less than five percent of all AD cases [199,200].

More than 95% of all cases belong to the so-called sporadic form of AD. The genetics of sporadic AD is much less well established. Generally, it is believed that sporadic AD is likely to be determined by a number of common risk alleles with low-penetrance, across several distinct loci. Currently, these loci are rather imprecise. However, genes located on these loci affect several pathways, many of which are supposed to be linked with the production, accumulation and elimination ("clearance") of Aβ. Moreover, there is rational evidence to suggest that collectively, these genes have a significant impact on disease susceptibility and age of onset [195,201].

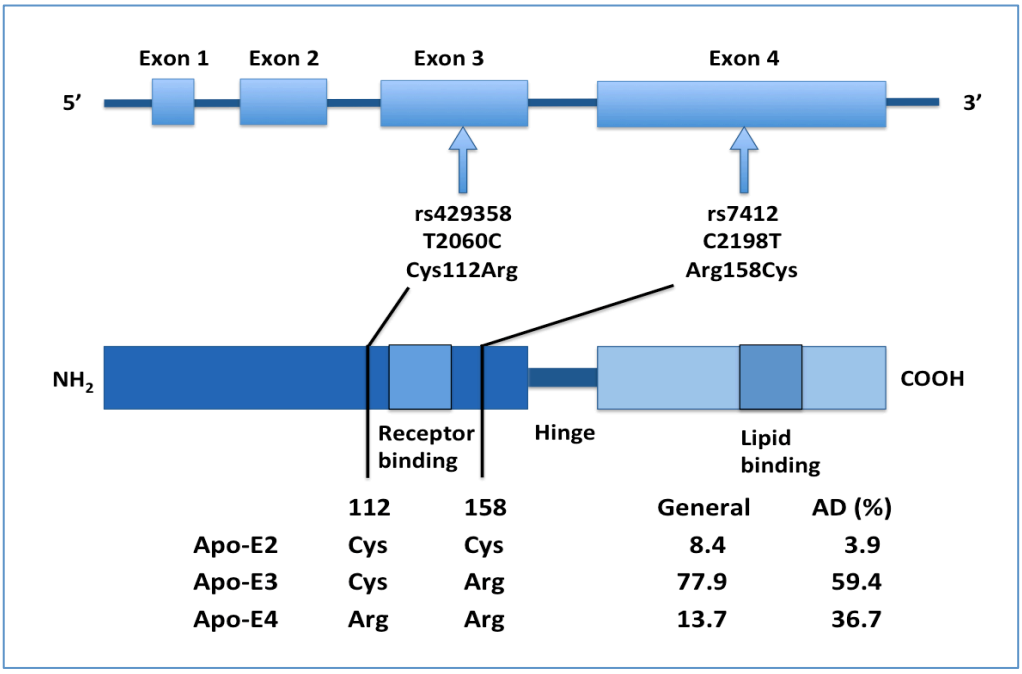
APOE alleles and Alzheimer's disease

In account of late-onset AD (LOAD), a number of candidate gene studies dedicatedly focused on those potential genes and proteins that play a specific role in Aβ production.

Linkage studies have identified apolipoprotein E (APOE), a gene located on chromosome 19q13, as a candidate gene with the epsilon allele showing strong association to the disease [202]. The APOE gene has three risk alleles (i.e. the ε2, ε3, and ε4). However, out of them, the ε4 allele has a 4-fold greater risk for late-onset AD than the ε3 allele [203]. In contrast, the ε2 allele is relatively less common and has

some protective effect with longevity [204,205] [Figure 1].

Hence, some extensive studies suggest that only the ε4 allele of APOE does not describe all of the genetic risk of this region of chromosome 19, for AD. There are two other potential gene candidates: the first one is TOMM40 (encoding translocase of outer mitochondrial membrane 40 homolog) (207,208] and the second one is EXOC3L2 (exocyst complex component 3-like-2) [209]. These genes, located in close proximity to APOE on chromosome 19, have been proposed to also increase disease susceptibility. Involvement of these genes suggests that other biological mechanisms, like mitochondrial dysfunction may play a role in disease progression [210].



	112	158	General	AD (%)
Apo-E2	Cys	Cys	8.4	3.9
Apo-E3	Cys	Arg	77.9	59.4
Apo-E4	Arg	Arg	13.7	36.7

Figure 1: Schematic representation of the APOE SNPs and genotypes [206]: *Two SNPs (rs429358 and rs7412) are in strong linkage disequilibrium and result in three APOE alleles (E2, E3 and E4). APOE ε4 is a major genetic risk factor for AD. The Apo-E2, -E3 and -E4 isoforms, which are encoded by the ε2, ε3 and ε4 alleles of the APOE gene, respectively, differ from each other at amino acid residues 112 and/or 158. Apo-E has two structural domains: the N-terminal domain, which contains the receptor-binding region (residues 136–150), and the C-terminal domain, which contains the lipid-binding region (residues 244–272); a hinge region joins the two domains. A meta-analysis demonstrated a significant association between the ε4 allele of APOE and AD (Adapted from: Ref. Gene. 2014 Jul 25;545(2):185-93. doi: 10.1016/j.gene.2014.05.031)*

GWAS and Other Susceptibility loci for Alzheimer's disease

Correspondingly, the largest GWA study for AD to date that included up to around 75,000 individuals, were performed with European ancestry subjects. These association studies identified BIN1, CR1, EPHA1, CD2AP, MS4A6A, CLU, ABCA7, PICALM, PTK2B, HLA-DRB5/HLA-DRB1, SLC24A4/RIN3, SORL1, MEF2C, INPP5D, ZCWPW1, NME8, FERMT2, CELF1, CD33, CASS4 and EPHA1 as susceptibility loci for AD [209,211-215]. Most of these genes congregate into three pathways: immune and inflammation response, endocytosis/intracellular trafficking and lipid metabolism [216].

The SORL1 (Sortilin-Related Receptor, L (DLR Class) A Repeats Containing) gene had been established to regulate managing of APP in a candidate gene approach and intracellular trafficking [217, 218]. CLU (Clusterin) is a lipoprotein that highly expressed in both the brain and the periphery [219]. Like APOE gene, it is also involved in lipid transport [220]. It is also hypothesized that CLU acts as an extracellular chaperone that regulates receptor-mediated Aβ clearance and Aβ-aggregation by endocytosis [219].

BIN1 (Bridging Integrator 1) is a part of the Bin1/amphiphysin/RVS167 (BAR) family that are associated with various cellular processes, including membrane trafficking, actin dynamics and clathrin-mediated endocytosis [221], which also influence Aβ production, APP processing and Aβ clearance from the brain. The PICALM (Phosphatidylinositol Binding Clathrin Assembly Protein) gene is associated with clathrin-mediated endocytosis in translocation of adaptor protein complex 2 and clathrin to sites of vesicle assembly [222].

The CD33 gene encodes a transmembrane protein of type-I that is linked to mediating cell-cell interactions and sialic acid-binding immunoglobulin-like lectins. In human brain, it is expressed in microglial cells; while increased expression of CD33 and CD33-positive microglia are observed in AD brains relative to controls. Contrariwise, a protective minor allele of the CD33, SNP rs3865444, leads to reductions in both CD33 expression in microglial cells and number of insoluble Aβ42 in AD brain. Additionally, the level of CD33-immunoreactive microglia positively correlates with the level of both insoluble Aβ42 and the amyloid plaque in AD cases. [223,224].

CR1 (Complement receptor type 1) is a cell-surface receptor and member of the complement system that is associated with clearance of immune-complexes including C3b and C4b. Hence, C3b can bind Aβ oligomers; and in this way CR1 may be

potentially involved in Aβ clearance. CR1 may also play a role in neuroinflammatory processes relevant for AD [225]. During this process, the CLU gene may get involved as an inhibitor [226].

The MS4A4A/MS4A4E/MS4A6E (Membrane-Spanning 4-Domains, Subfamily-A: Members 4A, 4E and 6E) locus maps to chromosome 11 and is a member of a group of 15 MS4A genes. As CD33, MS4A4A is also expressed on monocytes and myeloid cells, which suggests that it is involved in an immune-related function.

EPHA1 (EPH Receptor A1) is a member of the protein-tyrosine kinase family and the ephrin receptor subfamily. Members of this family are cell surface receptors, which binds with ephrin ligands on contiguous cells to regulate synapse formation, axon guidance, cell adhesion, migration and plasticity. EPHA1 also regulates cell motility and morphology [227]. In humans, besides expression in intestinal epithelium and colon epithelium, EPHA1 can be detected also in monocytes [228] and CD4-positive T lymphocytes [229]. This may imply that the basis for the genetic association of EPHA1 and AD lies in its putative function in the immune system.

CD2AP (CD2-Associated Protein) produces a scaffolding protein that binds to nephrin, actin and other proteins associated with cytoskeletal organization [230]. CD2AP is also involved in membrane trafficking and dynamic actin remodelling that occurs during receptor cytokinesis and endocytosis, whereas in the immune system, it is essential for synapse formation [231].

ABCA7 (ATP-Binding Cassette, Sub-Family A (ABC1), Member 7) is a member of the ATP-binding cassette (ABC) transporter superfamily. ABC family members involve in transportation of several molecules across intra- and extra- cellular membranes, including amyloid precursor protein [232] that is involved in host defence by influencing the phagocytosis of apoptotic cells by macrophages [233]. In addition, ABCA7 interacts with APOA-I and plays a role in cholesterol efflux and apolipoprotein-mediated phospholipid uptake from cells [232]. An independent GWA study, performed in African Americans, also confirmed that the ABCA7 gene is a susceptibility locus for AD [234].

Ridge et al. projected the phenotypic variance in Alzheimer's disease case-control status concentrating on the 11 known AD markers. By using the HapMap imputed ADGC dataset with 2,042,116 SNPs, they anticipated that common variants identified in GWAS genes for Alzheimer's disease, only elucidate 33% of the total phenotypic variance; within that APOE alone explicate 6% and other well-known 9 known high

frequency SNPs 2%, whereas more than 25% of phenotypic variance are still need to be identified [216,235].

A rare mutation of TREM2 gene linked to Alzheimer's disease

A whole genome sequencing study performed by Jonsson et al. based on 2261 Icelandic individuals, discovered a rare mutation of rs75932628-T (R47H) located on TREM2 (Triggering Receptor Expressed On Myeloid Cells 2) gene with a frequency of 0.63%. This rare mutation is identified as a new promising genetic risk marker associated with AD, with the odds ratio of 2.92. Afterwards, this rare variant was confirmed in a replication study with the cohorts from Germany, Norway, Spain, the Netherlands and the USA [236,237]. Alongside, the link between the R47H variant and LOAD confirmed by Guerreiro et al. with a meta-analysis of three independent imputed data sets of GWA studies (i.e. EADI, GERAD and ANM) [238].

Six additional variants Q33X, Y38C, T66M, D87D, R98W and H157Y were also identified as associated with affected cases, which might be linked to AD pathology. Out of those three variants Q33X, Y38C and T66M, in the homozygous state, had been already identified in relation with a frontotemporal dementia like syndrome [239]. The TREM2 gene is linked with inflammatory responses; it is also involved in immunological pathways in AD. Microglial cells interact with β-amyloid plaques and produce high levels of pro-inflammatory cytokines and reactive oxygen species, which may exhibit an alteration in morphology [240].

TREM2 is the only gene to be recognized with an adequate risk effect in AD since the establishment of the ε4 allele of APOE for AD [236,239].

Conclusion

Even though, GWAS is very successful in revealing genetic loci associated with human diseases and traits, reconnoitring the disease associated genes and molecular mechanisms underlying the identified genetic variants is not a trivial task. As more than 80% of disease/trait-associated SNPs are located in outside the coding regions, and only 12% are located in or close to protein-coding regions of genes, and within that even only <5% are non-synonymous SNPs. Thus mostly genetic variants have to link to the adjacent (such as 500 kb distant) genes, to nominate them as candidate genes. Consequently, the inference for the mechanistic connection between diseases and its susceptible genetic loci is more challenging than ever supposed.

Functional genomics studies can support to reveal the functional significances of

variants on intermediate molecular traits like protein products, alternative splicing, and gene expression. Thus subsequently, approaches, like computational and bioinformatics predictions based on the variants location and its sequence properties, can assist to propose the candidate genes. However, the range of potential functional consequences of variants is much broader, and therefore, new methodology is required to predict alteration in gene function. Furthermore, generally algorithms can only estimate variant effects on single proteins; likewise machine-learning approaches, that are being used to assess the effect of deleterious SNPs, have limitations.

Substantial knowledge about candidate genes in disease context are required to reveal the functional consequences at the molecular level, such as expression data at RNA and protein levels with time and space dimensions (such as at what time, in which tissue and in which organ). Furthermore, gene regulatory networks consists of many components linked to each other by multiple positive and negative feedback interactions, thus a deterministic understanding of their context is hard to achieve owing to rapidly growing complexity. Therefore, specialized algorithms and computable modelling approaches are essential, for the modelling and simulation of genetic regulatory networks.

Funding:

The research leading to these results has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY grant agreement n°115568, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

References:

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362-7 (2009)
2. Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, *et al.* A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [July 4, 2016].
3. McClellan J, King MC. Genetic heterogeneity in human disease. Cell 141(2):210-7 (2010)
4. Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. Cell 147(1):57-69 (2011)

5. Hou L, Zhao H. A review of post-GWAS prioritization approaches, Front. Genet. 4:280(2013)
6. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325(5945):1246-50
7. Grisanzio C, Werner L, Takeda D, Awoyemi BC, Pomerantz MM, Yamada H, *et al.* (2012) Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. Proc Natl Acad Sci U S A. 109(28):11252-7 (2012)
8. Li Q, Seo JH, Stranger B, McKenna A, Pe'er I, Laframboise T, *et al*. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152(3):633–641 (2013)
9. Pomerantz MM, Shrestha Y, Flavin RJ, Regan MM, Penney KL, Mucci LA, *et al.* Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. PLoS Genet 6(11): e1001204 (2010)
10. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, *et al.* Genome-wide analysis of transcript isoform variation in humans. Nat. Genet. 40(2):225-31 (2008)
11. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, *et al.* Variation in transcription factor binding among humans. Science 328(5975):232-5 (2010)
12. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482(7385):390-4 (2012)
13. Shen H, Fridley BL, Song H, Lawrenson K, Cunningham JM, Ramus SJ, *et al.* Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. Nat. Commun. 4:1628 (2013)
14. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, *et al.* Identification of genetic variants that affect histone modifications in human cells. Science 342(6159):747-9 (2013)
15. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, *et al.* Extensive variation in chromatin states across humans. Science 342(6159):750-2 (2013)
16. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science 342(6159):744-7(2013)
17. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, *et al.* Systematic localization of common disease-associated variation in regulatory DNA. Science 337(6099):1190-5(2012)
18. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447(7146):799-816(2007)

19. ENCODE Project Consortium,Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, *et al.* An integrated encyclopedia of DNA elements in the human genome. Nature. 489(7414):57-74(2012)
20. Yan J, Enge M, Whitington T, Dave K, Liu J, Sur I, *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell 154(4):801-13(2013)
21. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, *et al.* The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. 28(10):1045-8(2010)
22. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140(5):744-52(2010)
23. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 458(7235):223-7(2009)
24. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, *et al.* Landscape of transcription in human cells. Nature 489(7414):101-8(2012)
25. Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat. Methods 7(7):528-34(2010)
26. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. 28(8):817-25(2010)
27. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473(7345):43-9 (2011)
28. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459(7243):108-12(2009)
29. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. 39(3):311-8 (2007)
30. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132(6):958-70 (2008)
31. Kolaczkowski B, Kern AD. On the power of comparative genomics: Does Conservation Imply Function? In (Eds: Bell MA, Futuyma DJ, Eanes WF, Levinton JS). Evolution Since Darwin: The First 150 Years. Sinauer Associates pp. 151– 168 (2010)
32. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15(8):1034-50 (2005)
33. Kraft P, Cox DG. Study designs for genome-wide association studies. Adv Genet. 60:465-504. (2008)

34. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, *et al.* Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 86(6):929-42 (2010)
35. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 70(2):425-34 (2002)
36. Hunter DJ, Kraft P. Drinking from the fire hose–statistical issues in Genome-wide association studies. N Engl J Med. 357(5):436-9 (2007)
37. Stranger BE, Stahl EA, Raj T. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics. 187(2):367-83 (2011)
38. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 273(5281):1516-7(1996)
39. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5(5):e1000477(2009)
40. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet.42(6):508-14(2010)
41. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet.10:387-406 (2009)
42. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 17(R2):R122-8 (2008)
43. Munafò MR, Flint J. Jonathan Flint Meta-analysis of genetic association studies. Trends Genet. 20(9):439-44(2004)
44. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1):27-30(2000)
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* Gene Ontology: tool for the unification of biology. Nat Genet. 25(1):25-9 (2000)
46. Biological Expression Language (BEL): http://openbel.org/# (accessed on 30 June, 2016)
47. ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, *et al.* A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. 9(4):e1001046 (2011)
48. Nelson DL, Cox MM. In: Lehninger's Principles of Biochemistry. (4th Ed.) New York: W.H. Freeman Publishers (2005)
49. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet. 38(6):617-9(2006)

50. Zhang X, Bailey SD, Lupien M. Laying a solid foundation for Manhattan – 'setting the functional basis for the post-GWAS era'. Trends Genet. 30(4):140-9 (2014)
51. Hoffmeyer S, Burk O, von Richter O, Arnold HP, Brockmöller J, Johne A, *et al.* Functional polymorphisms of the human multidrug-resistance gene: multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. Proc Natl Acad Sci U S A. 97(7):3473-8 (2000)
52. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, *et al.* A 'silent' polymorphism in the MDR1 gene changes substrate specificity. Science. 315(5811):525-8 (2007)
53. Fung KL, Gottesman MM. A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. Biochim Biophys Acta. 1794(5):860-71(2009)
54. Komar AA. Genetics. SNPs, silent but not invisible. Science. 315(5811):466-7 (2007)
55. Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. Science. 342(6164):1367-72 (2013)
56. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 6(5):386-98 (2005)
57. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci. 25(3):106-10 (2000)
58. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 3(4):285-98 (2002)
59. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 297(5583):1007-13 (2002)
60. Gregory AP, Dendrou CA, Attfield KE, Haghikia A, Xifara DK, Butter F, *et al.* TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature. 488(7412):508-11 (2012)
61. Holwerda S, de Laat W. Chromatin loops, gene positioning, and gene expression. Front Genet. 3:217 (2012)
62. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 10(4):241-51(2009)
63. Docherty SJ, Davis OS, Haworth CM, Plomin R, D'Souza U, Mill J. A genetic association study of DNA methylation levels in the DRD4 gene region finds associations with nearby SNPs. Behav Brain Funct. 12;8:31(2012)
64. Sribudiani Y, Metzger M, Osinga J, Rey A, Burns AJ, Thapar N, *et al.* Variants in RET associated with Hirschsprung's disease affect binding of transcription factors and gene expression. Gastroenterology. 140(2):572-582.e2 (2011)

65. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol. 30(6):1411-20 (2010)
66. Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, *et al.* A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. 43(3):242-5 (2011)
67. Bovolenta LA, Acencio ML, Lemke N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genomics. 13:405(2012)
68. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 128(4):683-9(2007)
69. Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 12(4):283-93 (2011)
70. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res 22(9):1748-59 (2012)
71. Cowper-Sal lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet. 44(11):1191-8 (2012)
72. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, *et al.* Super-enhancers in the control of cell identity and disease. Cell. 155(4):934-47 (2013)
73. Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, *et al.* Chromatin stretch enhancer states drive cell- specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A. 110(44):17921-6 (2013)
74. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, *et al.* A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science. 312(5777):1215-7(2006)
75. Huang Y, Yang H, Borg BB, Su X, Rhodes SL, Yang K, *et al.* A functional SNP of interferon-gamma gene is important for interferon-alpha-induced and spontaneous recovery from hepatitis C virus infection. Proc Natl Acad Sci U S A. 104(3):985-90 (2007)
76. Higgs DR, Vickers MA, Wilkie AO, Pretorius IM, Jarman AP, Weatherall DJ. A review of the molecular genetics of the human alpha-globin gene cluster. Blood. 73(5):1081-104 (1989)
77. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 470(7333):264-8 (2011)
78. Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science. 342(6155):253-7 (2013)

79. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 466(7307):714-9 (2010)
80. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 41(8):885-90 (2009)
81. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 41(8):882-4 (2009)
82. Zhang X, Cowper-Sal lari R, Bailey SD, Moore JH, Lupien M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. 22(8):1437-46 (2012)
83. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 24(1):1-13 (2014)
84. Bickmore WA. The spatial organization of the human genome. Annu Rev Genomics Hum Genet. 14:67-84 (2013)
85. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 447(7143):413-7 (2007)
86. Gibcus JH, Dekker J. The hierarchy of the 3D genome. Mol Cell. 49(5):773-82 (2013)
87. Misteli T. Beyond the sequence: cellular organization of genome function. Cell. 128(4):787-800 (2007)
88. Roix JJ, McQueen PG, Munson PJ, Parada LA, Misteli T. Spatial proximity of translocation-prone gene loci in human lymphomas. Nat Genet. 34(3):287-91 (2003)
89. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 489(7414):109-13 (2012)
90. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 148(1-2):84-98 (2012)
91. Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. 22(3):446-55 (2012)
92. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 128(6):1231-45 (2007)
93. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell. 132(3):422-33 (2008)

94. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature. 451(7180):796-801(2008)
95. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, *et al.* Highresolution mapping and characterization of open chromatin across the genome. Cell. 132(2):311-22(2008)
96. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 39(3):311-8(2007)
97. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 10(6):1453-65(2002)
98. Lanzuolo C, Roure V, Dekker J, Bantignies F, Orlando V. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. Nat Cell Biol. 9(10):1167-74(2007)
99. Spilianakis CG, Flavell RA. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nat Immunol. 5(10):1017-27(2004)
100. Sexton T, Bantignies F, Cavalli G. Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. Semin Cell Dev Biol. 20(7):849-55(2009)
101. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 136(2):215-33 (2009)
102. Bartel B. MicroRNAs directing siRNA biogenesis. Nat Struct Mol Biol. 2005 Jul;12(7):569-71.
103. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science. 2005 Oct 14; 310(5746):317-20.
104. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer. 2010 Jun;10(6):389-402.
105. Wu M, Jolicoeur N, Li Z, Zhang L, Fortin Y, L'Abbe D, *et al.* Genetic variations of microRNAs in human cancer and their effects on the expression of miRNAs. Carcinogenesis. 2008 Sep;29(9):1710-6.
106. Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, *et al.* A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med. 2005 Oct 27;353(17):1793-801.
107. Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, *et al.* RAS is regulated by the let-7 microRNA family. Cell. 2005 Mar 11;120(5):635-47.
108. Singh SB, Davis AS, Taylor GA, Deretic V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. Science. 313(5792):1438-41 (2006)

109. Huang HY, Chien CH, Jen KH, Huang HD. RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. Nucleic Acids Res. 34(Web Server issue):W429-34 (2006)
110. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 129(7):1311-23 (2007)
111. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. Science. 329(5992):689-93 (2010)
112. Shen LX, Basilion JP, Stanton VP Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci U S A. 96(14):7871-6 (1999)
113. Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H, Green F, *et al.* Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Hum Mol Genet. 17(6):806-14 (2008)
114. Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A, *et al.* Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J Hum Genet. 51(12):1087-99 (2006)
115. Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev Cell. 16(1):47-57(2009)
116. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 12(14):1725-35(2003)
117. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31(13):3812-4(2003)
118. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al.* A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248-9 (2010)
119. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. Science. 335(6070):823-8 (2012)
120. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9(8):e1003709 (2013)
121. Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, *et al.* MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum Genet. 132(11):1235-43(2013)
122. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40(Database issue):D930-4 (2012)

123. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, *et al.* Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22(9):1790-7 (2012)
124. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 342(6154):1235587 (2013)
125. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 6(2):80-92 (2012)
126. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, *et al.* The accessible chromatin landscape of the human genome. Nature. 489(7414):75-82 (2012)
127. Lv J, Liu H, Su J, Wu X, Liu H, Li B, et al. DiseaseMeth: a human disease methylation database. Nucleic Acids Res. 40(Database issue):D1030-5 (2012)
128. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 44(8):841-7 (2012)
129. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 474(7351):380-4 (2011)
130. Chen R, Morgan AA, Dudley J, Deshpande T, Li L, Kodama K, *et al.* FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. Genome Biol. 9(12):R170 (2008)
131. Becker KG, Barnes KC, Bright TJ, Wang SA. The Genetic Association Database. Nat Genet. 36(5):431-2 (2004)
132. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, *et al.* Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 21(6):577-81(2003)
133. Kindt AS, Navarro P, Semple CA, Haley CS. The genomic signature of trait-associated variants. BMC Genomics. 14:108 (2013)
134. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 6(4):e1000888(2010)
135. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, *et al.* Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003 Mar 20;422(6929):297-302
136. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, *et al.* Genetic analysis of genome-wide variation in human gene expression. Nature. 2004 Aug 12;430(7001):743-7.
137. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002 Apr 26;296(5568):752-5.

138. Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eQTL). Methods. 2009 Jul;48(3):265-76.
139. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 43(6):513-8 (2011)
140. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008 Oct;4(10):e1000214. doi:10.1371/journal.pgen.1000214.
141. Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, *et al.* Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet. 2009 Oct;5(10):e1000692.
142. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. 2007 Apr 20;3(4):e58.
143. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, *et al.* Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008 May 6;6(5):e107.
144. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, *et al.* Genetics of gene expression and its effect on disease. Nature. 2008 Mar 27;452(7186):423-8.
145. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, *et al.* Variations in DNA elucidate molecular networks that cause disease. Nature. 2008 Mar 27;452(7186):429-35.
146. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, *et al.* A survey of genetic human cortical gene expression. Nat Genet. 2007 Dec;39(12):1494-9.
147. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580-5.
148. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 24(8):408-15 (2008)
149. Montgomery SB, Dermitzakis ET. From expression QTLs to personalized transcriptomics. Nat Rev Genet. 12(4):277-82 (2011)
150. Shimada MK, Matsumoto R, Hayakawa Y, Sanbonmatsu R, Gough C, Yamaguchi-Kabata Y, *et al.* VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. Nucleic Acids Res. 2009 Jan;37(Database issue):D810-5.
151. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 11(12):843-54 (2010)
152. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 44(8):841-7 (2012)

153. Segrè AV; DIAGRAM Consortium; MAGIC investigators, Groop L, Mootha VK, Daly MJ, *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6(8). pii: e1001058 (2010)
154. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics. 92(5):265-72 (2008)
155. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 81(6):1278-83 (2007)
156. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC; International Schizophrenia Consortium, Purcell SM, *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 5(6):e1000534 (2009)
157. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, *et al.* Epigenomic enhancer profiling defines a signature of colon cancer. Science. 336(6082):736-9 (2012)
158. Weng L, Macciardi F, Subramanian A, Guffanti G, Potkin SG, Yu Z, et al. SNP-based pathway enrichment analysis for genome-wide association studies. BMC Bioinformatics12:99 (2011)
159. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, *et al.* The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32(Database issue):D258-61 (2004)
160. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 33(Web Server issue):W741-8 (2005)
161. Shahbaba B, Shachaf CM, Yu Z. A pathway analysis method for genome-wide association studies. Stat Med. 31(10):988-1000 (2012)
162. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat.1:107–129 (2007)
163. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat. 2007;1:85–106.
164. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA.102:15545–15550 (2005)
165. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 85(1):13-24 (2009)
166. Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet. 125:63–79 (2009)
167. Kraft P, Raychaudhuri S. Complex diseases, complex genes: keeping pathways on the right track. Epidemiology. 20:508–511 (2009)

168. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, *et al.* Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet. 86:860–871 (2010)
169. Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. Adv Genet.72:141–179 (2010)
170. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. Eur J Hum Genet.18:1045–1053(2010)
171. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res.38:W90–W95(2010)
172. Schaid DJ, Sinnwell JP, Jenkins GD, McDonnell SK, Ingle JN, Kubo M, Goss PE, Costantino JP, Wickerham DL, Weinshilboum RM. Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. Genet Epidemiol.36:3–16(2012)
173. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24:2784
174. Jones L, Holmans PA, Hamshere ML, Harold D, Moskvina V, Ivanov D, et al. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. PLoS One. 5(11):e13950(2010)
175. Zhang K, Chang S, Cui S, Guo L, Zhang L, Wang J. ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W437-43
176. Sun YV. Integration of biological networks and pathways with genetic association studies. Hum Genet. 131(10):1677-86(2012)
177. Serin EA, Nijveen H, Hilhorst HW, Ligterink W. Learning from Co-expression Networks: Possibilities and Challenges. Front Plant Sci.7:444 (2016)
178. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics. 27(1):95-102 (2011)
179. Bakir-Gungor B, Egemen E, Sezerman OU. PANOGA: a web server for identification of SNP-targeted pathways from genome-wide association study data. Bioinformatics. 30(9):1287-9 (2014)
180. Liu Y, Patel S, Nibbe R, Maxwell S, Chowdhury SA, Koyuturk M, Zhu X, Larkin EK, Buxbaum SG, Punjabi NM, Gharib SA, Redline S, Chance MR. Systems biology analyses of gene expression and genome wide association study data in obstructive sleep apnea. Pac Symp Biocomput. 2011:14-25
181. Saccone SF, Bolze R, Thomas P, Quan J, Mehta G, Deelman E, Tischfield JA, Rice JP. SPOT: a web-based tool for using biological databases to

prioritize SNPs after a genome-wide association study. Nucleic Acids Res. 38(Web Server issue):W201-9(2010)
182. Lee PH, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res. 36(Database issue):D820-4 (2008)
183. Iyappan A, Bagewadi S, Page M, Hofmann-Apitius M, and Senger P. NeuroRDF: Semantic Data Integration Strategies for Modeling Neurodegenerative Diseases. In Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM2014). Aveiro, Portugal, (2014)
184. Phillips PC. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet.Nov;9(11):855-67(2008)
185. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nat Rev Genet. 8(6):437-49 (2007)
186. Bateson W. Mendel's Principles of Heredity. Cambridge Univ. Press; Cambridge: (1909)
187. Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinb.52:399–433 (1918)
188. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 10(6):392-404 (2009)
189. KEMPTHORNE O. The correlation between relatives in a random mating population. Proc R Soc Lond B Biol Sci. 143(910):102-13 (1954)
190. Mao Y, London NR, Ma L, Dvorkin D, Da Y. Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. Physiol Genomics. 28(1):46-52 (2006)
191. Zhang X, Huang S, Zou F, Wang W. TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics. 26(12):i217-27(2010)
192. Emily M. IndOR: a new statistical procedure to test for SNP-SNP epistasis in genome-wide association studies. Stat Med. 31(21):2359-73 (2012)
193. Piriyapongsa J, Ngamphiw C, Intarapanich A, Kulawonganunchai S, Assawamakin A, Bootchai C, Shaw PJ, Tongsima S. iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. BMC Genomics.13 Suppl 7:S2(2012)
194. Arkin Y, Rahmani E, Kleber ME, Laaksonen R, März W, Halperin E. EPIQ-efficient detection of SNP-SNP epistatic interactions for quantitative traits. Bioinformatics. 30(12):i19-25 (2014)
195. Bertram L, Tanzi RE. Genome-wide association studies in Alzheimer's disease. Human Molecular Genetics. 18(R2):R137-R145 (2009)
196. Glenner GG, Wong CW. Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. Biochem Biophys Res Commun.120(3):885-90(1984)
197. Ryan NS, Rossor MN. Correlating familial Alzheimer's disease gene mutations with clinical phenotype. Biomark Med.4(1):99-112 (2010)
198. Tanzi RE, Bertram L. Twenty years of the Alzheimer's disease amyloid

hypothesis: a genetic perspective. Cell. 120(4):545-55(2005)
199. Raux G, Guyant-Maréchal L, Martin C, Bou J, Penet C, Brice A, Hannequin D, Frebourg T, Campion D. Molecular diagnosis of autosomal dominant early onset Alzheimer's disease: an update. J Med Genet. 42(10):793-5(2005)
200. Janssen JC, Beck JA, Campbell TA, Dickinson A, Fox NC, Harvey RJ, Houlden H, Rossor MN, Collinge J. Early onset familial Alzheimer's disease: Mutation frequency in 31 families. Neurology. 60(2):235-9(2003)
201. Bertram L, Tanzi RE. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. Nat Rev Neurosci.9(10):768-78(2008)
202. Pericak-Vance MA, Bebout JL, Gaskell PC Jr, Yamaoka LH, Hung WY, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA, et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am J Hum Genet. 48(6):1034-50(1991)
203. Genin E, Hannequin D, Wallon D, Sleegers K, Hiltunen M, Combarros O, *et al*. APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol Psychiatry.16(9):903-7(2011)
204. Olesen OF, Mikkelsen JD, Gerdes C, Jensen PH. Isoform-specific binding of human apolipoprotein E to the non-amyloid beta component of Alzheimer's disease amyloid. Brain Res Mol Brain Res. 44(1):105-12(1997)
205. Deelen J, Beekman M, Uh HW, Broer L, Ayers KL, Tan Q, *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. Hum Mol Genet.23(16):4420-32(2014)
206. Kim DH, Yeo SH, Park JM, Choi JY, Lee TH, Park SY, *et al.* Genetic markers for diagnosis and pathogenesis of Alzheimer's disease. Gene.545(2):185-93(2014)
207. Roses AD. An inherited variable poly-T repeat genotype in TOMM40 in Alzheimer disease. Arch Neurol.67(5):536-41(2010)
208. Roses AD, Lutz MW, Amrine-Madsen H, Saunders AM, Crenshaw DG, Sundseth SS, *et al.* A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. Pharmacogenomics J. 10(5):375-84(2010)
209. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, *et al.* CHARGE Consortium; GERAD1 Consortium; EADI1 Consortium. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA.303(18):1832-40(2010)
210. Ferencz B, Karlsson S, Kalpouzos G. Promising Genetic Biomarkers of Preclinical Alzheimer's Disease: The Influence of APOE and TOMM40 on Brain Integrity. Int J Alzheimers Dis.2012:421452(2012)
211. 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.41(10):1088-93(2009)
212. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet.43(5):436-41(2011)
213. 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.41(10):1094-9(2009)
214. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez

C, DeStafano AL, *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 45(12):1452-8(2013)
215. 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. 43(5):429-35(2011)
216. Reitz C. Genetic diagnosis and prognosis of Alzheimer's disease: challenges and opportunities. Expert Rev Mol Diagn. 15(3):339-48(2015)
217. Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F. et al. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat Genet.39(2):168-77(2007)
218. Reitz C, Mayeux R. Use of genetic variation as biomarkers for mild cognitive impairment and progression of mild cognitive impairment to dementia. J Alzheimers Dis.19(1):229-51(2010)
219. Nuutinen T, Suuronen T, Kauppinen A, Salminen A. Clusterin: a forgotten player in Alzheimer's disease. Brain Res Rev.61(2):89-104(2009)
220. Wollmer MA, Sleegers K, Ingelsson M, Zekanowski C, Brouwers N, Maruszak A. et al. Association study of cholesterol-related genes in Alzheimer's disease. Neurogenetics.8(3):179-88(2007)
221. Pant S, Sharma M, Patel K, Caplan S, Carr CM, Grant BD. AMPH-1/Amphiphysin/Bin1 functions with RME-1/Ehd1 in endocytic recycling. Nat Cell Biol. 11(12):1399-410 (2009)
222. Tebar F, Bohlander SK, Sorkin A. Clathrin assembly lymphoid myeloid leukemia (CALM) protein: localization in endocytic-coated pits, interactions with clathrin, and the impact of overexpression on clathrin-mediated traffic. Mol Biol Cell. 10(8):2687-702(1999)
223. Griciuc A, Serrano-Pozo A, Parrado AR, Lesinski AN, Asselin CN, Mullin K, Hooli B, Choi SH, Hyman BT, Tanzi RE. Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. Neuron. 78(4):631-43(2013)
224. Guerreiro R, Hardy J. Genetics of Alzheimer's disease. Neurotherapeutics. 11(4):732-7(2014)
225. Crehan H, Holton P, Wray S, Pocock J, Guerreiro R, Hardy J. Complement receptor 1 (CR1) and Alzheimer's disease. Immunobiology.217(2):244-50(2012)
226. McGeer PL, Kawamata T, Walker DG. Distribution of clusterin in Alzheimer brain tissue. Brain Res.579(2):337-41(1992)
227. Yamazaki T, Masuda J, Omori T, Usui R, Akiyama H, Maru Y. EphA1 interacts with integrin-linked kinase and regulates cell morphology and motility. J Cell Sci.122(Pt 2):243-55(2009)
228. Sakamoto A, Sugamoto Y, Tokunaga Y, Yoshimuta T, Hayashi K, *et al*. Expression profiling of the ephrin (EFN) and Eph receptor (EPH) family of genes in atherosclerosis-related human cells. J Int Med Res.39(2):522-7(2011)
229. Holen HL, Nustad K, Aasheim HC. Activation of EphA receptors on CD4+CD45RO+memory cells stimulates migration. J Leukoc Biol.87(6):1059-68(2010)
230. Lehtonen S, Zhao F, Lehtonen E. CD2-associated protein directly interacts with the actin cytoskeleton. Am J Physiol Renal Physiol.283(4):F734-43(2002)
231. Dustin ML, Olszowy MW, Holdorf AD, Li J, Bromley S, Desai N, Widder P, Rosenberger F, van der Merwe PA, Allen PM, Shaw AS. A novel adaptor

protein orchestrates receptor patterning and cytoskeletal polarity in T-cell contacts. Cell.94(5):667-77(1998)
232. Chan SL, Kim WS, Kwok JB, Hill AF, Cappai R, Rye KA, Garner B. ATP-binding cassette transporter A7 regulates processing of amyloid precursor protein in vitro. J Neurochem.106(2):793-804(2008)
233. Tanaka N, Abe-Dohmae S, Iwamoto N, Yokoyama S. Roles of ATP-binding cassette transporter A7 in cholesterol homeostasis and host defense system. J Atheroscler Thromb.18(4):274-81(2011)
234. Reitz C, Jun G, Naj A, Rajbhandary R, Vardarajan BN, Wang LS, *et al*. Alzheimer Disease Genetics Consortium. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E ϵ4,and the risk of late-onset Alzheimer disease in African Americans. JAMA.309(14):1483-92(2013)
235. Ridge PG, Mukherjee S, Crane PK, Kauwe JS; Alzheimer's Disease Genetics Consortium. Alzheimer's disease: analyzing the missing heritability. PLoS One.8(11):e79771(2013)
236. Jonsson T, Stefansson K. TREM2 and eurodegenerative disease. N Engl J Med.369(16):1568-9(2013)
237. Yaghmoor F, Noorsaeed A, Alsaggaf S, Aljohani W, Scholtzova H, Boutajangout A, Wisniewski T. The Role of TREM2 in Alzheimer's Disease and Other Neurological Disorders. J Alzheimers Dis Parkinsonism. 4(5). pii: 160(2014)
238. Guerreiro R, Hardy J. TREM2 and neurodegenerative disease. N Engl J Med. 369(16):1569-70(2013)
239. Guerreiro RJ, Lohmann E, Brás JM, Gibbs JR, Rohrer JD, Gurunlian N, *et al*. Using exome sequencing to reveal mutations in TREM2 presenting as a frontotemporal dementia-like syndrome without bone involvement. JAMA Neurol.70(1):78-84(2013)
240. Krabbe G, Halle A, Matyash V, Rinnenthal JL, Eom GD, Bernhardt U, *et al*. Functional impairment of microglia coincides with Beta-amyloid deposition in mice with Alzheimer-like pathology. PLoS One.8(4):e60921(2013)

Summary

The keep purpose of genetic studies is to highlight genetic variants from a risk-associated locus, which account for phenotypic differences. GWAS-identified variants can be prioritized at the molecular level, on the basis of their functional consequences.

In this article, I have reviewed and summarized biomedical literature to assess the diverse functional impacts of genetic variation at the genomic as well as molecular level. I have primarily focused on the interpretation of genetic variants and mutations in a systems biology context, by recapitulating numerous computational and bioinformatics methodologies which are used to assist in identifying the candidate genes. They predict functional consequences of genome based on the disease-associated variant's location and their sequence characterization. However, the spectrum of potential functional consequences of variants is much broader, and therefore, new methodologies are required to predict alteration of gene function. Furthermore, most of the algorithms can only estimate variant effects on single proteins. Machine-learning approaches used to assess the effect of deleterious SNPs have similar limitations.

In this article, I conclude that gene regulatory networks are comprised of many components linked to each other by multiple positive and negative feedback interactions. This rapidly-growing complexity makes a deterministic understanding of their context hard to achieve. Consequently, dedicated specialized algorithms and computable modelling approaches are needed, for the modelling and simulation of integrated genetic and molecular level networks.

Chapter 2

Genetic Variance Information in Cause-and-Effect Models

Introduction

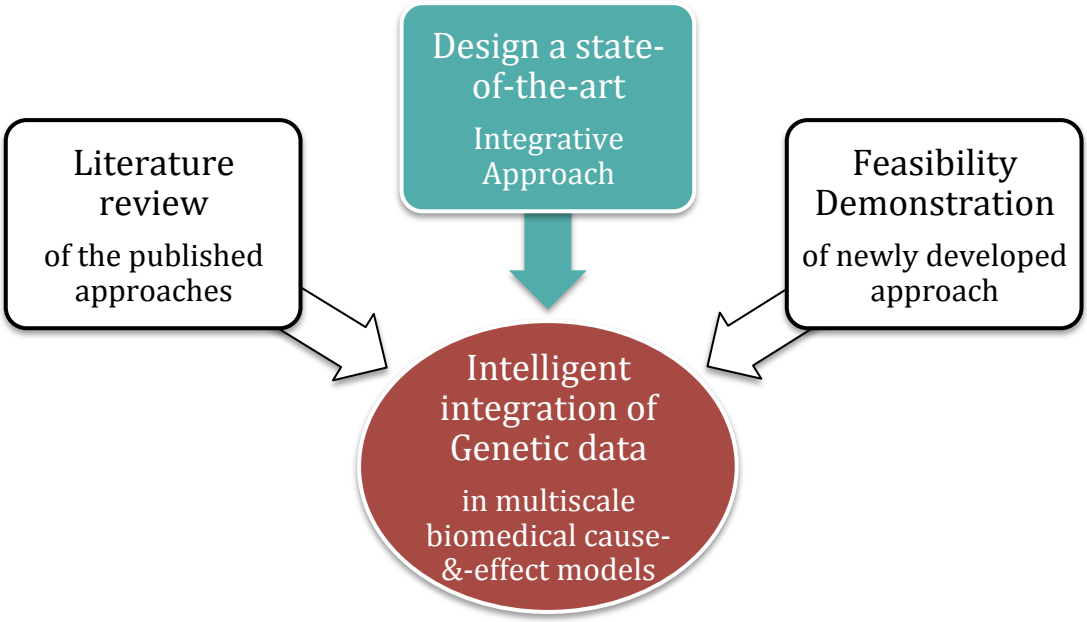


Figure 21: Part –II: Workflow for the intelligent functional interpretation of genetic variance information in multi-scale biomedical cause-and-effect models

Until recently, GWAS data associated with complex diseases or quantitative traits could not be annotated effectively in systems biomedicine models, due to unknown mechanism of action by which these variants influence disease or quantitative traits. Genetic variants located in non-coding regions of the genome have proven especially challenging. Contemporarily, a series of large-scale genomics projects (including ENCODE, IHEC, FANTOM etc.) have established new approaches to reveal functional characterization of these genetic variants. Now the greatest challenge in

this computationally advanced era, is the interpretation and mapping of GWAS data into biological networks with an evidential reasoning approach to annotate the variants' functional consequences.

I am proposing here a rational approach to map genetic data into biomedicine models to interpret biological insights into clinical benefits. Biological Expression Language (BEL) is designed to represent biological knowledge in a computable form by capturing causal and correlative relationship in context, where information about biological system, reference citation and process of curation could be included. For the next version of BEL, I have developed a representation of genetic variant types, including substitutions, insertions, deletions, fusions, unspecified mutations and variants affecting intergenic regions. These genetic variations could be propagated between DNA, RNA and protein levels by using HGVS mutation nomenclature.

"Reasoning over Genetic Variance Information in Cause-and-Effect Models of Neurodegenerative Diseases"

Mufassra Naz[1,2], Alpha Tom Kodamullil[1,2] and Martin Hofmann-Apitius[1,2]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53754 and [2]Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for IT, Dahlmannstrasse 2, 53113 Bonn, Germany

Brief Bioinform. 2016 May; 17(3): 505-16.
doi: 10.1093/bib/bbv063

Abstract

The work we present here is based on the recent extension of the syntax of the Biological Expression Language (BEL); which now allows for the representation of genetic variation information in cause-and-effect models. In our paper, we describe, how genetic variation information can be used to identify candidate disease mechanisms in diseases with complex aetiology such as Alzheimer´s Disease and Parkinson's Disease. In those diseases, we have to assume that many genetic variants contribute moderately to the overall dysregulation that in the case of neurodegenerative diseases has such a long incubation time until the first clinical symptoms are detectable. Due to the multi-level nature of dysregulation events, systems biomedicine modelling approaches need to combine mechanistic information from various levels, including gene expression, miRNA expression, protein-protein interaction, genetic variation and pathway. OpenBEL, the open

source version of BEL, has recently been extended to match this requirement and we demonstrate in our paper, how candidate mechanisms for early dysregulation events in Alzheimer´s Disease can be identified based on an integrative mining approach that identifies "chains of causation" that include SNP information in BEL models.

Keywords

BEL Model, Alzheimer's Disease, Genetic Variants, GWAS, Causal Reasoning, Cause-and-Effect

Systems biology models and genetic variation: two separate worlds

Barabási *et al*. assert "given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network" [1].

Genome-wide genetic association studies (GWAS) have become a very useful and frequently used tool for discovering genetic variants as a disease risk [2]. However, for complex traits and phenotypes, interpretation of association data largely benefits from available prior biological and environmental knowledge, spanning over multiple scientific disciplines [3].

In human genetics, several strategies were developed and implemented to determine the effect of SNPs, particularly, for the analysis of genotyping data. The limitation of many of these algorithms is that they can predict only either to

have no effect or to have negative effect on clinical readouts and endpoints. However, the spectrum of possible biological effects caused by genetic variants is much wider, thus methods are required to predict also potential gain, loss or even modification of gene function [4]. Moreover, most of the algorithms can only predict variant effects on individual proteins [5], and machine-learning supervised and semi-supervised approaches are being used to predict the effect of deleterious SNPs [4]. Generally, GWA studies are used to establish links between genotypes and phenotypes through identifying the differences (and commonalities) between thousands of individuals. These approaches work as black boxes and make use of statistical and machine-learning approaches that require huge datasets.

In order to reveal the functional context at the molecular level, substantial knowledge about the genes involved, their expression at RNA and protein level, the time when they are expressed and in which tissue and in which organ, is required. Regulation of gene expression is mediated through genetic regulatory systems, which are controlled by complex interaction networks involving DNA, RNA, proteins, and small molecules. These regulatory networks involve many components linked to each other by positive and negative feedback loops and a deterministic understanding of their dynamics is hard to attain due to rapidly increasing complexity. Therefore, specialized methods and computer software are essential, for the modelling and simulation of genetic regulatory networks [6].

Systems biology is the systemic contextual representation and modelling of a plurality of discrete observations. In systems biology, modelling is a

representation of disease high-level concepts in a unified and comprehensive network that can help to identify the differential sub-networks by comparing it with a network representing the healthy state [1,7-10]. Building blocks of systems biology models, such as signalling pathways, metabolic systems, and gene regulation networks are already widely used in computational biology. Comprehensive disease models, however, are going way beyond these comparably well-understood functional modules. One of the explicit goals of systems biomedicine is to use a generalized model of disease to assess the parameters from high throughput data of a single patient, in order to generate 'personalized models' that predict disease progression and treatment responses [11,12].

In systems biology, there are several "entry points" to generate initial networks: protein–protein interaction, metabolic networks, and signalling pathways have been widely used to model biological processes [7]. In the last decade, however, new modelling approaches have been developed [13]. For pharmacogenomics, these networks represent complex relationships between drugs and targets. The diseasome [14] is a disease – gene and drug – target (protein) network, where disease information is associated with a gene and drugs are linked to proteins by drug-target associations [15].

Despite the complexity of regulatory networks, attempts at unravelling the impact of genetic variation on regulatory networks has been addressed by a number of groups. Leiserson et al. [16], Carter et al. [17] and Atias et al. [18] have worked on network approaches to scrutinize the genetic risks for human disease. They have developed methodology that allows to detect causal genes

within disease-associated loci by network analysis, and to ascertain causal paths from allele to disease through intermediate molecular phenotypes [16, 17, 18]. Trynka et al. [19] proposed new approaches on the interpretation of transcriptional regulation effects to estimate the involvement of variant alleles in common diseases. They suggested that most of the causal complex trait variants have regulatory roles with cell type specificity, by interconnecting GWAS data with genome-wide chromatin assays results. They emphasized the importance of cell-type specific regulatory context and highlighted the value of the inclusion of epigenomics information [19].

Sahni et al. [20] questioned the strong bias in the literature towards coding variant effects on protein–DNA, protein–RNA and protein–protein interactions. He proposed to put more emphasis on effects outside the protein centric scope of functional assessment, to understand the impact of genetic variants on specific interactions; for instance, mechanisms safeguarding protein folding and stability [20].

Types of genetic variation information relevant for systems biomedicine

If genetic variation is to be included in a systems biology model of disease, we need to assess the biological impact of a single nucleotide polymorphism or a mutation. Dependent on the way (the "mode-of-action") how a SNP or a mutation exerts its biological impact, we can distinguish several classes ("types") of SNPs. In this section, we identify and discuss the different functional

categories that can be distinguished as "mode-of-SNP-action" classes, (see Table 1).

Table 1: Types of genetic variation information relevant for systems biomedicine: DNA regions with functional categories and consequences

Types of genetic variation information relevant for systems biomedicine		
DNA Regions	**Functional Categories**	**Functional Consequences**
1. Coding regions	1. Non-synonymous genetic variants	Change in protein structure or function due to a change in the amino acid sequence or protein sequence truncation
	2. Synonymous genetic variants	Modulating translation rates with direct consequences to protein folding
	3. Exon Splicing Enhancers or Silencers	Translate the protein isoform by deleterious intron retention or exon skipping
2. Non-coding regions	1. DNA methylation	Associates with genes silencing
	2. Transcription factor binding to regulatory elements	Can change transcription factor binding to DNA that leads to differential target gene expression
	3. Chromatin loop bridging the enhancers and promoters	Can alter the DNA affinity for looping factors and chromatin interactions, which regulates gene expression
	4. MicroRNAs	Can affect gene functionality: 1) by transcription of primary transcript, 2) by primary microRNA (pri-microRNA) and precursor microRNA (pre-microRNA) processing and 3) by effecting microRNA-microRNA interaction
	5. Long non-coding RNAs (lncRNAs)	Can modify highly conserved lncRNA tertiary structure that can affect chromatin regulator's interactions

Genetic variants on coding regions

The risk associated with **non-synonymous genetic variants** can be easily translated into a change in protein structure or function due to a change in the amino acid sequence. It can modify amino acid composition, or truncate the protein sequence by causing an early stop codon [21]. **Synonymous genetic**

variants do not alter the codon sequence. However, synonymous genetic risk variants can still impact protein function by modulating translation rates with direct consequences to protein folding [22]. For example, rs1045642 SNP slows down the rate of translation of the MDR1 mRNA and impacts protein folding [23]. **Exon Splicing Enhancers or Silencers** (ESEs/ESSs) are typically 6 to 8 consecutive nucleotide sequences in an exon region. Where, SNP can also result in deleterious intron retention or exon skipping, and translate the protein isoform [24–27]. For example, rs1800693 SNP affects the splicing of the TNFRSF1A mRNA and leading to translate an isoform [28].

Genetic Variants on Non-coding regions:

Model gene system studies have revealed that local DNA interactions between regulatory sites and genes are important for transcriptional control. Such regulatory interactions, in mammals, can take place over significant chromosomal distances up to an entire mega-base (1MB) [29]. Genetic risk variants are very frequent on non-coding sequences [30]. Post-GWAS studies have revealed the capacity of these genetic risk variants to regulate gene expression by modulating cis-regulatory machineries through mechanisms involving DNA methylation [31], transcription factor binding [32], chromatin looping [33], or miRNA recruitment [34]. If SNPs occur within transcriptional regulatory regions, like transcription factor binding sites, CpG islands, and microRNAs, they may modify the binding affinity of the regions, remove

recognition sites, or create new binding sites for other regulatory proteins. All of these modifications can lead to alterations in the level, timing, and localization of gene expression [35].

DNA methylation: DNA methylation means addition of methyl groups to a cytosine nucleotide, which is basically part of a CpG dinucleotide [63]. DNA hyper-methylation near transcription start sites (TSS) of tumor suppressor genes associates with their silencing [36].

Transcription factor binding to regulatory elements: Across the genome, transcription factors bind to thousands of regulatory elements, including promoters (directly upstream of their target genes) and cis-regulatory elements such as enhancers, insulators and silencers [37]. Genetic risk variants located within promoter regions can also change transcription factor binding to DNA, leading to differential target gene expression [38, 39]. For example, expression of the α-globin gene locus is affected by a genetic variant associated with the a-thalassemia blood disorder [38]. Enhancers are commonly targeted by those genetic variants of risk-associated loci that map to DNA recognition motifs, bound by transcription factors. These genetic variants can modulate the chromatin affinity for transcription factors and consequently gene expression [40–47]. For example, the rs12740374 SNP, which is associated with a lower level of plasma low-density lipoprotein cholesterol (LDL-C), increases the expression level of the SORT1 (Sortilin 1) gene by increasing the binding affinity of the C/EBP (CCAAT enhancer-binding protein) transcription factor to chromatin [46].

Chromatin loop bridging the enhancers and promoters: Genetic risk variants can modulate chromatin loop formation; it can alter the DNA affinity for looping factors, which results in allele-specific chromatin loop formation. The human genome is structured in a three-dimensional architecture, which is thought to regulate a diverse set of DNA-template processes [47–52]. This facilitates regulatory elements, like promoters and enhancers, to interact physically through long-range chromatin loops, or chromatin interactions, to regulate gene expression [53, 54]. This has been shown for the rs12913832 SNP, which resides in an enhancer 21 KB upstream of the OCA2 (Oculocutaneous albinism II) pigment gene [55]. Over the last decade, the development of chromosome conformation capture (3C) technology has initiated several 3D studies on regulatory chromatin loops, but what has been done until now is far from exhaustive. If a minor fraction of these potential regulatory elements participate in chromatin looping, then most of the genomic interactions have yet to be characterized again, because many such loops appear to be tissue-specific [56-58], which makes their comprehensive analysis appear even more disconcerting [59].

MicroRNAs: MicroRNAs (miRNAs) target mRNAs by recognizing their complementary sequences mainly in 3' untranslated regions (3'UTRs). miRNAs largely function as post-transcriptional repressors. They recruit RNA-induced silencing complex (RISC) to their target mRNAs, leading to mRNA degradation or translation repression [60]. They can regulate the translation of hundreds of genes through sequence-specific binding to mRNA [61]. SNP variants, linked with miRNAs, can affect gene functionality with three different ways: 1) by

transcription of primary transcript, 2) by primary microRNA (pri-microRNA) and precursor microRNA (pre-microRNA) processing and 3) by effecting microRNA- microRNA interaction [62]. For example, rs10065172, a Crohn's disease-associated SNP, lies within the 3' UTR of the IRGM (immunity-related GTPase M) gene and alters the complementary target sequence of miRNA-196 [63].

Long non-coding RNAs (lncRNAs): lncRNAs are found across intergenic regions of the human genome [64]. They can interact with chromatin regulators for their recruitment by chromatin [65,66], a process, which relies on a highly conserved lncRNA tertiary structure; that can be changed by genetic risk variants [67]. Kim et al. [68] described enhancer RNAs (eRNAs), a new class of non-coding RNAs (ncRNAs), form from polymerase II-bound enhancers. The level of expression of eRNAs positively correlated with the expression of neighboring coding genes [68] Genetic variants in enhancer sequences can modify TF binding, resulting in 'improper' gene expression and eventually susceptibility to diseases [69, 70]. The micropeptides, called small pri-peptides, are also expressed from the lncRNA-pri and direct the proteolytic cleavage or other modifications of target proteins or transcription factors [71].

Expression quantitative trait loci (eQTL): Studying the association between genetic variation and gene expression offers a straightforward way to begin the complicated task of connecting risk variants to their putative target genes [72]. Networks created using gene expression data from patient samples can be exploited to bridge GWAS results with an underlying disease mechanism, as exemplified in the autism spectrum disorder [73]. Genetic variation associated

with gene expression, known as expression quantitative trait loci (eQTL), can identify the target genes of risk loci [74–78]. Polymorphism situated in DNA regulatory elements can alter the gene transcript frequency. Thus, as a quantitative trait locus, gene transcript frequency can be determined with substantial power [79, 80]. Brem et al. published the first genome-wide study of gene expression in 2002 [81]. Stranger and Raj reviewed the genetics of human variation and diversity in eQTLs. These eQTL data are very dynamic with great specificity for different tissues and environmental perturbations [82].

The ENCODE project: Identification of genomic functional elements:

The ENCODE project has delivered an incredible compilation of genetic functional elements of the human genome [83]. As most of the SNPs detected in GWAS data belong to non-coding regions of the genome; usage of ENCODE regulatory elements to reinterpret GWAS data sets, might be a valuable approach [84]. Undoubtedly, structural genomic variation are more influential and systemic than the smaller scale variations, however any framework or methodology employed to predict genetic variant effects needs to contribute for both small and large-scale variations [13]. If possible, it should be able to predict the level in which coding or non-coding genetic variants individually or collectively have a functional impact on biology, ranging from relevant protein function or expression to the perturbation of entire networks. It can help us to annotate the massive amount of re-sequencing data meaningfully without having to test the effects of all variants experimentally [13].

Thus, now it is the time to move ahead from merely bio-statistical approaches for GWAS data interpretations to a more comprehensive approach that can be acquainted with gene–gene and gene–environment interactions, along with the complexity of the relationship between genotype and phenotype [85].

The need to integrate genetic variant information in systems biomedicine models

Currently, GWAS variance data interpretation has become a bottleneck in the progression of mapping and exploring complex diseases. For example, multiple genes have been associated with Amyotrophic lateral sclerosis (ALS) in GWAS data, but there is no clear perspective of involved pathways and mechanisms that would emerge from the available high throughput data, by taking multiple rare variants into account [86].

Substantial research for several complex diseases has been conducted to unravel causal mechanisms underlying their disease aetiology. Often this type of research is multidisciplinary, using research studies spreading over a wide range of time and length scales. Consequently, a disease model representing disease aetiology may have many modules and interactions. Such a disease model would provide a nice template for the interpretation of the functional consequences of genetic variation [87].

One of the obvious questions is of course, which methodology can help in interpretation of GWAS data, when most of the SNPs have small effects on

disease susceptibility [88]. There is lack of efficient and reliable algorithms as well as appropriate multi-scale modelling methodology, to evaluate the huge number of interdependent data from GWAS [5]. One way to reduce the combinatorial complexity of GWAS data is, to reduce the dimensionality of genetic variation data by taking a priori knowledge about functional relationships between genes and proteins into account. Formalised knowledge about causal and correlative relationships in systems biology models provides a good starting point for that dimension reduction. So far, there have been only few serious efforts to predict how these genetic variants would collectively be effective for specific phenotypes [89, 90].

Systems biology modelling language syntax adaptions

A massive amount of data for molecular interactions and pathways are stored in online databases. Moreover experimental data is accumulating very rapidly and correspondingly the demand for exchange of data to allow analysis and comparison of larger datasets is intensifying. Thus there is a need for representation of data in standardized formats. Comparisons and evaluations of modern systems biology modelling languages show [91,92] that XML is a remarkable and easy-to-use format for systems biology information representation. Here, we compare the recent updates to the standard XML-based representation formats for exchange of data.

The Resource Description Framework (RDF) model [93] is based upon the idea of making statements about resources. A RDF statement, also called a triple in RDF terminology is an association of the form (subject, predicate, object). RDF

Schema (RDFS) [94] and the Web Ontology Language (OWL) [95] are used to explicitly represent the meanings of the resources described on the Web and how they are related. These specifications, called ontologies, describe the semantics of classes and properties used in Web documents. These ontologies should be linked to a top-level ontology in order to enable knowledge sharing and reuse [95]. Unfortunately, each bio-ontology seems to be built as an independent piece of information, that does not enable the sharing and reuse of knowledge and complicates data integration [96]. Moreover, various sources of biological data must be combined in order to obtain a full picture and to build new knowledge. However, a large majority of current databases does not use a uniform way to name biological entities. As a result, a same biomedical object is frequently associated with different names.

Systems Biology Markup Language (SBML) [97-100] was designed by the Systems Biology Workbench Development group. The purpose of SBML is to model biochemical reaction networks, comprising cell signalling, gene regulation and metabolic pathways. In SBML 'Species' is used as a notation to represent the interactors, while reaction, modelling a transformation, transport or binding to represent interaction. Each reaction is allowed to interact with three predefined interactors i.e. reactant, product and modifier [101]. An SBML model encodes a reaction network as pathway. Mathematical relations are also available for reactions. References to other sources and extra information can be added only in the annotation field. Currently, the representation of parts of molecules is not possible [102].

The Proteomics Standards Initiative Molecular Interaction XML format (PSI MI) [103] is designed by the Proteomics Standards Initiative that is an initiative of the Human Proteome Organization (HUPO). The main purpose of the initiative is to standardize proteomics data representation to facilitate data exchange, comparison and verification. The format is projected for exchange of protein-protein interaction data [103]. PSI MI is structured around an entry. It is not anticipated to be a pathway [102]. Links to publications and databases are possible, but a representation of relationships through mathematical equations and an inheritance is not available [102].

The Biological Pathway Exchange (BioPAX) format is designed by the BioPAX working group [104, 105]. The main purpose of this standard is to introduce a unified framework for sharing pathway information. BioPAX offers more explicit use of relations between concepts than SBML and PSI MI. It is defined as ontology of concepts with attributes [105]. However, reasoning and integration of data increases its computational complexity [102]. A specific data type is available for pathway representation, but mathematical equations underlying the relations are not possible.

CellML Model Repository [106] contains biochemical pathway models that have been published in peer-reviewed articles or expressed in SBML [107]. CellML [108] and the CellML Model Repository are part of the IUPS Physiome Project [109]. The CellML Model Repository contains models describing a wide range of biological processes [110]. It uses mathematical descriptions of biological systems and adds semantic meaning by annotating elements by ontologies and constrained vocabularies [110]. It is also very precise, thus the association

between dependent and independent species is implicit rather than explicit. However due to this generality and explicit nature, complexity is increased, especially for software developers, consequently, there are a very few tools which can read and write CellML [111].

Biological Expression Language (BEL):

BEL is a highly expressive, triple-based knowledge representation language for the representation of knowledge about causal and correlative relationships [112]. Several groups in academia, and pharma are already applying BEL in various areas including biological network analysis, disease modeling, understanding drug efficacy and toxicity, mechanisms for drug sensitivity and resistance, and other research and development related projects. A suite of software components called the BEL Framework provides tools that are required to create, compile, assemble and deliver computable knowledge models to BEL-aware applications [112].

BEL represents complex biological content as simplified, formalized, computable semantic triples that provide the ability to use and re-use experimental observations. BEL can also be used for next-generation sequencing applications, like gene expression profiling and genome annotation data, by using Reverse Causal Reasoning (RCR) algorithm to get mechanistic insights into the high throughput data, which could be complementary to the result of analysis using pathway gene set. BEL has many utility tools such as a dedicated Cytoscape plug-in for network visualization, algorithms of causal reasoning (RCR) for understanding disease mechanism by identifying up-stream and down-stream

controllers, electronic workbook integration, BEL-to-RDF translation, text mining in BEL, and nano-publication concepts [113]. BEL has the potential to impact scientific literature, by introducing computable expressions in scientific publishing, that could be integrated efficiently into existing knowledge environment [114]. Moreover, these causal-reasoning models can provide a valuable addition to the biologists to interpret the gene expression data [115]. By using these models, Huang CL et al. [116] has proposed a data-driven method, Correlation Set Analysis (CSA), to detect active regulators in disease by integrating co-expression analysis and literature-derived causal relationships [116].

Reasoning over genetic variance information integrated in disease networks: concepts and strategies

A key task in genetic variants interpretation, to understand the phenotypic consequences, lies in the ability to predict the molecular level mechanistic consequences of gene polymorphisms and mutations.

As a consequence, systems biomedicine modelling approaches need to combine mechanistic information from various levels, including gene expression, miRNA expression, protein-protein interaction, genetic variation and pathway information. OpenBEL, the open source version of BEL, has recently been extended to match this requirement. With the extended syntax, the new version of BEL 2.0 is now enabled for encoding genetic variants in biomedical models. The last release of the BEL syntax proposes a representation for different genetic

variant types, for example, <substitution>, <insertion>, <deletion> and <intergenic>; by introducing new variant functions for DNA, RNA and protein levels.

In this version, the variant (<expression>) function can be used as an argument within a gene(), rna(), microRNA(), or protein() to indicate a sequence variant of the specified level. The variant() function takes HGVS variant description expression, e.g., for a substitution, insertion, or deletion variants. The extended BEL syntax is supposed to support reasoning over cause-effect models that include genetic variation information

Representation of variant at Proteins level: Effects of genetic variants located on coding region or splice site, if express at protein level, they can be represented through protein level functions. Protein level variants representation is purposed to see the genetic variants with their relevancy to protein, like their location on the protein sequence, effect on the protein structure (see Table 2).

Representation of variant across DNA/RNA: To see the genetic variants impact at DNA/RNA level, protein level variants can also be expressed by DNA/RNA level functions. Whereas, genetic variants located on non-coding regions (like, intergenic or intronic) can only be represented through DNA/RNA level functions, which are designed to see the genetic variants with their relevancy to genome or gene expression (see Table 3).

Table 2: Representation of different genetic variant categories with variant functions at Proteins level in Biological Expression Language (2.0V):

Variant Categories	Variant() function in protein
Reference allele	p(HGNC:CFTR, var(=))
Unspecified variant	p(HGNC:CFTR, var(?))
Substitution variant	p(REF:NP_000483.3, var(p.Gly576Ala))
Deletion variant	p(REF:NP_000483.3, var(p.Phe508del))
Frameshift variant (HGVS short description)	p(REF:NP_000483.3, var(p.Thr1220Lysfs))
Frameshift variant (HGVS long description)	p(REF:NP_000483.3, var(p.Thr1220Lysfs*7))

Table 3: Representation of genetic variants across DNA/RNA with the reference of chromosomal or mRNA position in Biological Expression Language (2.0V):

Level categories	var() function at different genetic levels
DNA - SNP	g(SNP:rs113993960, var(delCTT))
DNA - chromosome	g(REF:NC_000007.13, var(g.117199646_117199648delCTT))
DNA - coding sequence	g(REF:NM_000492.3, var(c.1521_1523delCTT))
RNA - coding sequence	r(REF:NM_000492.3, var(c.1521_1523delCTT))
RNA - RNA sequence	r(REF:NM_000492.3, var(r.1653_1655delcuu))

Integration of genetic variation information in BEL models of Alzheimer´s Disease: enhanced functional interpretation of complex SNP patterns

As a support of this review, here we demonstrate an example to highlight this promising approach, by integrating genetic variant information into an Alzheimer's disease (AD) BEL model.

We have recently published the AD BEL model [117]. This model has 4,052 nodes and 9,926 edges, it was generated by extracting relevant knowledge from the specific biomedical literature. The AD BEL model comprises disease-associated genes, protein-protein interactions, miRNAs, bioprocesses and pathways. To integrate disease specific genetic variant information into AD BEL model, genetic data is retrieved from GWAS databases and the biomedical literature using text-mining methods. The AD BEL model was enriched with AD-SNP associated data, after annotating functional impact of these genetic variants using the ENSEMBL variant database.

Subsequently, these genetics variants were prioritized, according to their functional consequences. Then we mapped them to the AD BEL model to identify sub-networks with SNPs that display a substantial biological impact. To complete the functional impact assessment for these variants, we have excavated the biomedical literature to analyze the role of these SNPs in the context of age of onset of AD and specifically in the endocytosis pathway.

The early endosome is the first vacuolar compartment in the context of EP, where enlarged early endosomes are identified as the earliest neuro-pathologic features to develop in the early onset of AD. In sporadic AD, endosomal enlargement adds to an average 2.5-fold larger total endosomal volume per neuron, suggesting a significant increase in endocytic activity. It is the site of internalization and initial processing of amyloid precursor protein (APP) and apolipoprotein E (ApoE), two significant proteins in AD aetiology [118-120]. Here we focus on the internalization of APP based on the functional role of SNPs.

AD is mainly characterized by the deposition of insoluble amyloid beta peptides 42 (Aβ42) in the brain, which cannot be easily removed through the blood brain barrier. In healthy brain, APP is processed by ADAM10, which produces soluble amyloid beta peptide 40 (Aβ40), whereas in the non-amyloidogenic pathway, APP is proteolytically processed by BACE and γ-secretase to generate Aβ42 peptides. A SNP rs514049, linked to the ADAM10 gene, may perturb the normal processing of APP to produce soluble Aβ40, as rs514049 is associated with lower level of CSF APPα in AD [121]. BACE1 and BACE2 associated with γ-secretase complex proteins. Moreover, a SNP rs3754048, with allele G, in the promoter of APH1A gene, might alter the binding ability of YY1 transcription factor, resulting in an increased level of APH1A and γ-secretase activity to facilitate Aβ42 generation [122].

All these players in the non-amyloidogenic pathway are trans-membrane proteins, that traffic through the endocytic pathway [123], where these proteins are internalized from the plasma membrane and recycled back to the surface (as in early endosomes and recycling endosomes), or, alternatively, sorted to degradation (as in late endosomes and lysosomes [124, 125]. However, BACE1 is a genetically very significant gene with a number of high ranked AD-associated SNPs. It is also evident that APP and BACE1 are up-regulated in AD. Moreover, experimental evidences suggested that at the cell surface, APP and BACE1 strongly interact and co-localize and are being internalized together into early endosomes, where both proteins remain co-localized and produce amyloid-β. This evidence confirms that endocytosis may be an important step for amyloid-β

production [126]. This can be again supported by the association of genetic variants linked with the trafficking proteins in the EP.

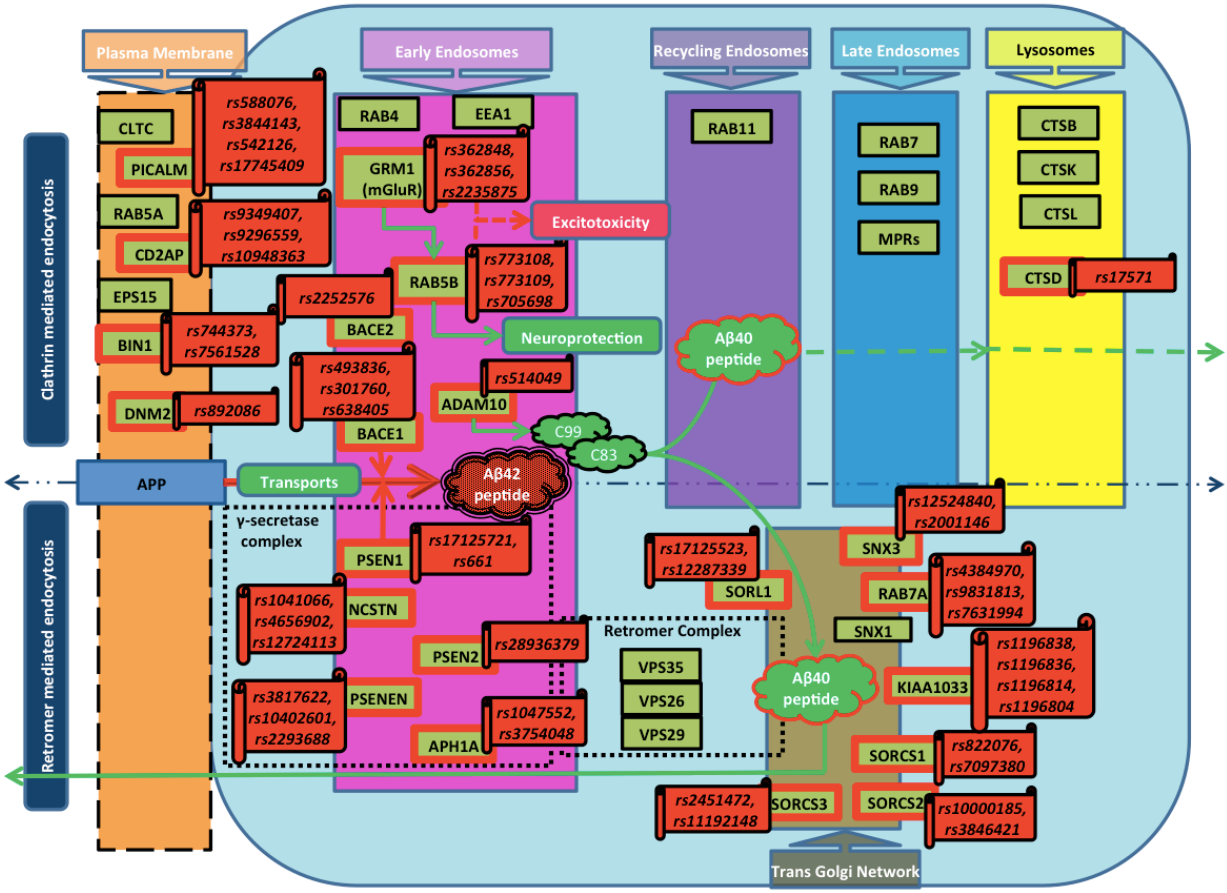


Figure 1: *In this diagram, we present a flowchart that depicts an abstracted BEL sub-network derived from the original AD BEL Model. This flowchart represents causal relationships between genes and genetic variants for the EP components. Gene symbols written in the textboxes with red outline are showing association with the GWAS identified SNPs for AD.*

As shown in Figure 1, there are two branches of EP: firstly, clathrin-mediated endocytosis (CME) and secondly, retromer-mediated endocytosis (RME). In the CME pathway, various proteins such as CLTC, PICALM, DNM2, EPS15, and BIN1 modulate APP transport for its further internalization, subsequent Aβ generation

and further processing in lysosomes, which is required for neurotransmission and signal transduction. Clathrin (CLTC) is a major protein component of coated vesicles and coated pits in CME pathway [127]. These specialized organelles are involved in the intracellular trafficking of receptors and endocytosis of a variety of macromolecules including APP with the help of additional accessory proteins such as PICALM, EPS1, DNM2, EGF and its substrate EPS15. PICALM encodes a clathrin assembly protein, which recruits CLTC and AP2, and regulates the size of the clathrin vesicle at neuromuscular junction, whereas an intronic PICALM SNP, rs588076, is associated with allelic expression of a PICALM isoform [128]. Stable DNM2 recruitment during CME correlates well with CLTC lifetime [129], while a risk allele at rs892086 associated with reduced expression of DNM2 mRNA in the hippocampus in AD patients compared to non-demented controls [130].

On the other hand, the EP is also regulated by retromer which transports APP from early endosomes to trans-Golgi network (TGN) and released outside cell mainly by retromer complex (VPS35, VPS29, VPS26), SORL1, SNX3, SNX1, WASH complex (KIAA1033) and so on [131]. SORL1 protein belongs to type-I trans-membrane, which is expressed in neurons and plays a critical role in the intracellular transport and in APP processing. SORL1 binds to the retromer complex and works as an adaptor protein for APP trafficking from endosomes to TGN. It is observed that SORL1 levels are reduced in AD diseased brain; while, overexpression of it redistributes APP to the Golgi apparatus, thus the placement and interaction time of APP and BACE1 is reduced in the early endosomes, which will reduce the amount of Aβ42. SNX3 mediates recruitment of cargo selective retromer complex in association with VPS35 [131]. Recent studies have shown

that SNX3 and RAB7A are also required for proper recruitment of the cargo-selective complex. Constitutively active RAB7A Q67L mutant is overexpressed, resulting in displacement of the cargo-selective complex [132]. The cargo-selective retromer subcomplex (VPS35– VPS29– VPS26) recruits the WASH complex (KIAA1033), which mediates the production of branched actin networks on the surface of endosomes. The cargo-selective retromer complex together with SNX27 and the WASH complex operate in the endosome-to-cell surface recycling of receptors and proteins.

Integration of genetic variation information enhances the evidence base for shared pathophysiology pathways in neurodegenerative diseases

Parkinson´s Disease (PD) and AD may share pathophysiological mechanisms and – as a consequence – may actually share some of their molecular aetiology. In order to identify evidences that would speak for shared pathophysiology between AD and PD, we systematically analyzed genetic variation information that is common between AD and PD and that can be mapped to putatively shared pathways. We have selected the common SNPs from AD and PD GWAS data, and mapped them to gene annotation. Then we searched diseased BEL models to identify the functional impacts of these genes on AD, PD or neurodegenerative diseases (see Table 4).

Table 4: A list of common SNPs/genes in AD and PD with their possible role in the disease context specifically for AD and PD and generally for Neurodegenerative diseases (NDD):

Common SNPs in AD & PD	Gene	AD	PD	NDD
rs931977 (Intronic)	ERG2	- EGR2 targeted by mAChRs (muscarinic acetylcholine receptors), which is associated with cognitive functions, synaptic plasticity and memory - EGR2 also associated with apoptosis	-	- EGR2 is involved in myelination of peripheral nerves
rs2672893 (Intronic)	RPTOR	- RPTOR is downstream of MTOR and is expressed highly in AD hippocampus - RPTOR activates of PI3K-Akt pathway	Alpha-synuclein reduced the activation of AMPK target RPTOR	-
rs6488270 (Intergenic)	Downstream_variant for: TMEM52B Upstream_variant for: GABARAPL1	-	GABARAPL1 plays role in development and homeostasis of the mouse brain	- GABARAPL1 presents a regulated tissue expression and is the most highly expressed gene among the family in the central nervous system
rs4742095 (Intergenic)	Upstream_variant for: CD274 PLGRKT	- PD1/PD-L1 (CD274) pathway have role in neuroinflammation of AD - PD1/PD-L1 (CD274) pathway is associated with IL-10 production	-	-PLGRKT is regulating plasminogen activation which plays a key role in regulating catecholaminergic neurosecretory cell function -PLGRKT is also involved in macrophage recruitment in the inflammatory response - PLGRKT is believed to have role in plasminogen binding and cell migration
rs1984129 (Intergenic)	Downstream_variant for: GBP6 Upstream_variant for: LRRC8B	-	-	- LRRC8B is implicated in proliferation and activation of lymphocytes and monocytes
rs10515758 (Intergenic)	Downstream_variant for: EBF1 Upstream_variant for:	-	-	- EBF1 have role in axonal pathfinding -CLINT1 interacts with clathrin, the adapter protein AP-1 and phosphoinositides. This

	CLINT1			protein may be involved in the formation of clathrin coated vesicles and trafficking between the trans-Golgi network and endosomes
rs6810871 (Intergenic)	Downstream_variant for: FAM114A1, TMEM156 Upstream_variant for: KLHL5, TLR6	-	-	- FAM114A1 plays a role in neuronal cell development - FAM114A1 expressed in dentate gyrus, the hippocampus, the cerebellum and the olfactory bulb

Conclusion:

Given the complexity of neurodegenerative diseases and the limited accessibility to experimental tissues of brain, we need new strategies to integrate data driven and knowledge driven approaches to unravel the mechanism behind these complex diseases. Disease networks based on the systems biology models, comprising of various interacting molecules such as genes, proteins, bioprocesses etc., succeeded to integrate most of the available data. In this review, we tried to recapitulate all the major breakthroughs, which demonstrated the collective capturing of disease-related knowledge, modelling it as a system. In addition, we have revisited the major studies around identification of genetic variants and prioritizing these variants based on statistical analysis.

So far, disease networks could not easily accommodate information on genetic variation. We have introduced a novel methodology based on BEL, which enables us to integrate genetic variation information into a disease network. We developed a strategy to analyze the functional consequences of SNPs based on their location in the genome and an interpretation of their putative role in a

network model. Currently using the capabilities of extended BEL version, we have developed the AD BEL models together with genetic variants with their DNA, RNA or protein position, variant type and associated allele; which can be used to better understand the role of SNPs in a disease context and tried to predict its consequences based on the functional context provided by the network model.

Although BEL provides certain powerful algorithms like reverse causal reasoning (RCR), which allows identifying upstream controllers of an observed effect, there are still limitations to overcome in order to enable reasoning over genetic variants. It is obvious, that we need to develop more sophisticated algorithms for reasoning over genetic variant information in network models, by integrating the functional impact of genetic variants on genes in the disease context. One route to go to refine that algorithm is based on machine-learning approaches to train a model with the established knowledge of functionally identified genetic variants for different complex diseases. That model will then be applied to neurodegenerative diseases to overcome the deficiency of genetic variant evidential data in this area.

Summary (Key points):

1. Systems biomedicine modelling approaches need to combine various types of mechanistic details to address multi-level nature of disease dysregulation processes
2. This work represents genetic variation information integration in cause-and-effect models to identify candidate disease mechanisms in diseases with complex aetiology

3. It is an integrative mining approach that identifies "chains of causation" with reasoning over genetic information in BEL models.
4. It exemplifies a new strategy to integrate data driven and knowledge driven approaches to unravel the mechanism of complex diseases

Short bio of Authors

Mufassra Naz is a researcher at Fraunhofer SCAI and PhD student at the University of Bonn. Her main research interests are functional interpretation of genetic variants, causal algorithms and neurodegenerative diseases.

Alpha Tom Kodamullil is a researcher at Fraunhofer SCAI and PhD student at the University of Bonn. Her research work focuses on automatic expansion of computable disease models by developing reasoners.

Martin Hofmann-Apitius is Professor for Applied Life Science Informatics at the University of Bonn. He is also the Head of the Department of Bioinformatics at Fraunhofer Institute SCAI.

References

1. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011; 12: 56–68.
2. Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449:851–861.
3. Williamson R. The molecular genetics of complex inherited diseases. Br J Cancer Suppl 1988; 9:14–16.
4. Cardoso JG, Andersen MR, Herrgård MJ, et al. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. Front Bioeng Biotechnol. 2015 Feb 16;3:13.
5. Zhao N, Han JG, Shyu CR, et al. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. PLoS Comput Biol. 2014 May 1;10(5):e1003592.
6. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol. 2002;9(1):67-103.
7. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. Cell 2011; 144: 986–998.
8. Sonnenschein N, Golib Dzib JF, Lesne A, et al. A network perspective on metabolic inconsistency. BMC Syst Biol 2012; 6: 41.
9. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol 2012; 8: 565.
10. Hütt MT. Understanding genetic variation - the value of systems biology. Br J Clin Pharmacol. 2014 Apr;77(4):597-605.

11. Kuepfer L. Towards whole-body systems physiology. Mol Syst Biol 2010; 6: 409.
12. Kühn A, Lehrach H. The 'Virtual Patient' system: modeling cancer using deep sequencing technologies for personalized cancer treatment. J Verbr Lebensm 2012; 7: 55–62.
13. Goh KI, Cusick ME, Valle D, et al. The human disease network. Proc Natl Acad Sci U S A 2007; 104: 8685–8690.
14. Wysocki K, Ritter L. Diseasome: an approach to understanding gene-disease interactions. Annu Rev Nurs Res. 2011;29:55-72. Review. PubMed PMID: 22891498.
15. Yildirim MA, Goh KI, Cusick ME, et al. Drug-target network. Nat Biotechnol 2007; 25: 1119–1126.
16. Leiserson MD, Eldridge JV, Ramachandran S, et al. Network analysis of GWAS data. Curr Opin Genet Dev. 2013 Dec;23(6):602-10.
17. Carter H, Hofree M, Ideker T. Genotype to phenotype via network analysis. Curr Opin Genet Dev. 2013 Dec;23(6):611-21.
18. Atias N, Istrail S, Sharan R. Pathway-based analysis of genomic variation data. Curr Opin Genet Dev. 2013 Dec;23(6):622-6.
19. Trynka G, Raychaudhuri S. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. Curr Opin Genet Dev. 2013 Dec;23(6):635-41.
20. Sahni N, Yi S, Zhong Q, et al. Edgotype: a fundamental link between genotype and phenotype. Curr Opin Genet Dev. 2013 Dec;23(6):649-57.

21. Smyth DJ, Cooper JD, Bailey R, et al. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet. 2006;38(6):617-9
22. Zhang X, Bailey SD, Lupien M. Laying a solid foundation for Manhattan – 'setting the functional basis for the post-GWAS era'. Trends Genet. 2014;30(4):140-9
23. Komar AA. Genetics. SNPs, silent but not invisible. Science. 2007;315(5811):466-7
24. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6(5):386-98
25. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci. 2000;25(3):106-10
26. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 2002;3(4):285-98
27. Fairbrother WG, Yeh RF, Sharp PA, et al. Predictive identification of exonic splicing enhancers in human genes. Science. 2002;297(5583):1007-13
28. Gregory AP, Dendrou CA, Attfield KE, et al. TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature. 2012;488(7412):508-11
29. Holwerda S, de Laat W. Chromatin loops, gene positioning, and gene expression. Front Genet. 2012;3:217
30. Frazer KA, Murray SS, Schork NJ, et al. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10(4):241-51

31. Docherty SJ, Davis OS, Haworth CM, et al. A genetic association study of DNA methylation levels in the DRD4 gene region finds associations with nearby SNPs. Behav Brain Funct. 2012;12;8:31
32. Sribudiani Y, Metzger M, Osinga J, et al. Variants in RET associated with Hirschsprung's disease affect binding of transcription factors and gene expression. Gastroenterology. 2011;140(2):572-582.e2
33. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol. 2010;30(6):1411-20
34. Brest P, Lapaquette P, Souidi M, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. 2011;43(3):242-5
35. Bandele OJ, Wang X, Campbell MR, et al. Human single-nucleotide polymorphisms alter p53 sequence-specific binding at gene regulatory elements. Nucleic Acids Res. 2011;39(1):178-89
36. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683-9
37. Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12(4):283-93
38. De Gobbi M, Viprakasit V, Hughes JR, et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science. 2006;312(5777):1215-7
39. Huang Y, Yang H, Borg BB, et al. A functional SNP of interferon-gamma gene is important for interferon-alpha-induced and spontaneous recovery

from hepatitis C virus infection. Proc Natl Acad Sci U S A. 2007;104(3):985-90

40. Cowper-Sal Lari R, Zhang X, Wright JB, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat. Genet. 2012;44, 1191–1198
41. Harismendy O, Notani D, Song X, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 2011;470(7333):264-8
42. Bauer DE, Kamran SC, Lessard S, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science. 2013;342(6155):253-7
43. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010;466(7307):714-9
44. Tuupanen S, Turunen M, Lehtonen R, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 2009;41(8):885-90
45. Pomerantz MM, Ahmadiyeh N, Jia L, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 2009;41(8):882-4
46. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol. 2010;30(6):1411-20

47. Zhang X, Cowper-Sal lari R, Bailey SD, et al. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. 2012;22(8):1437-46
48. Bickmore WA. The spatial organization of the human genome. Annu Rev Genomics Hum Genet. 2013;14:67-84
49. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007;447(7143):413-7
50. Gibcus JH, Dekker J. The hierarchy of the 3D genome. Mol Cell. 49(5):773-82 (2013)
51. Misteli T. Beyond the sequence: cellular organization of genome function. Cell. 2007;128(4):787-800
52. Roix JJ, McQueen PG, Munson PJ, et al. Spatial proximity of translocation-prone gene loci in human lymphomas. Nat Genet. 2003;34(3):287-91
53. Sanyal A, Lajoie BR, Jain G, et al. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109-13
54. Li G, Ruan X, Auerbach RK, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148(1-2):84-98
55. Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. 2012;22(3):446-55
56. Tolhuis B, Palstra RJ, Splinter E, et al. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 2002;10(6):1453-65

57. Lanzuolo C, Roure V, Dekker J, et al. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. Nat Cell Biol. 2007;9(10):1167-74
58. Spilianakis CG, Flavell RA. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nat Immunol. 2004;5(10):1017-27
59. Sexton T, Bantignies F, Cavalli G. Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. Semin Cell Dev Biol. 2009;20(7):849-55
60. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215-33
61. Bartel B. MicroRNAs directing siRNA biogenesis. Nat Struct Mol Biol. 2005 Jul;12(7):569-71.
62. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer. 2010 Jun;10(6):389-402.
63. Brest P, Lapaquette P, Souidi M, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. 2011;43(3):242-5
64. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223-7
65. Rinn JL, Kertesz M, Wang JK, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007;129(7):1311-23

66. Tsai MC, Manor O, Wan Y, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010;329(5992):689-93
67. Shen LX, Basilion JP, Stanton VP Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci U S A. 1999;96(14):7871-6
68. Kim TK, Hemberg M, Gray JM, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010. pp. 182–187.
69. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010;466:714–719.
70. Harismendy O, Notani D, Song X, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature. 2011;470:264–268.
71. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y. Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science. 2010 Jul 16;329(5989):336-9.
72. Califano A, Butte AJ, Friend S, et al. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 2012;44(8):841-7
73. Voineagu I, Wang X, Johnston P, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380-4

74. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 2009;325(5945):1246-50
75. Grisanzio C, Werner L, Takeda D, et al. Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. Proc Natl Acad Sci U S A. 2012;109(28):11252-7
76. Li Q, Seo JH, Stranger B, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 2013;152(3):633–641
77. Pomerantz MM, Shrestha Y, Flavin RJ, et al. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. PLoS Genet 2010;6(11): e1001204
78. Nicolae DL, Gamazon E, Zhang W, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6(4):e1000888
79. Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003 Mar 20;422(6929):297-302
80. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004 Aug 12;430(7001):743-7.
81. Brem RB, Yvert G, Clinton R, et al. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002 Apr 26;296(5568):752-5.
82. Stranger BE, Raj T. Genetics of human gene expression. Curr Opin Genet Dev. 2013 Dec;23(6):627-34.

83. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012; 488: 57–74.
84. Ecker JR, Bickmore WA, Barroso I, et al. Genomics: ENCODE explained. Nature 2012; 489: 52–55.
85. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics 2010; 26: 445–455.
86. Kiernan MC, Vucic S, Cheah BC, et al. Amyotrophic lateral sclerosis. Lancet 2011; 377: 942–955.
87. Thomas PD, Mi H, Swan GE, et al. Pharmacogenetics of Nicotine Addiction and Treatment Consortium. A systems biology network model for genetic association studies of nicotine addiction and treatment. Pharmacogenet Genomics. 2009 Jul;19(7):538-51.
88. Bader JS. The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics 2001; 2: 11–24.
89. Burga, A., and Lehner, B. (2013). Predicting phenotypic variation from genotypes phenotypes and a combination of the two. *Curr. Opin. Biotechnol.* 24, 803–809.
90. Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* 14, 168–178.
91. Achard F., Vaysseix G, Barillot E. XML, bioinformatics and data integration. Bioinformatics 2001;17:115-125.
92. McEntire R., Karp P, Abernethy N, et al. An evaluation of ontology exchange languages for bioinformatics. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2000;8:239-250.

93. S.R. Bratt, Toward a Web of Data and Programs, in IEEE Symposium on Global Data Interoperability - Challenges and Technologies, 2005.
94. J. Dupré, The Disorder of Things: Metaphysical Foundations of the Disunity of Science, Harvard University Press ed, 1993.
95. O. Bodenreider and R. Stevens, Bio-ontologies: current trends and future directions, Briefings in Bioinformatics, vol. 7, pp. 256-274, 2006.
96. L.N. Soldatova and R.D. King, Are the current ontologies in biology good ontologies?, Nature Biotechnology, vol. 23, pp. 1095-1098, 2005.
97. http://sbml.org/
98. Finney A.2004. Systems biology markup language (SBML) Level 3: proposal: multi-component species features. Proposal manuscript. March 2004 (April 2004).
99. Finney A., Hucka M. Systems biology markup language (SBML) Level 2: structures and facilities for model definitions. 2003. June 28, 2003 (April 2004).
100. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003;19:524-531.
101. Strömbäck L, Lambrix P. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. Bioinformatics. 2005 Dec 15;21(24):4401-7
102. Ruebenacker O, Moraru II, Schaff JC, et al. Integrating BioPAX pathway knowledge with SBML models. IET Syst Biol. 2009 Sep;3(5):317-28.

103. Hermjakob H., et al. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat. Biotechnol. 2004a;22:177-183.

104. http://www.biopax.org

105. BioPAX working group. BioPAX—biological pathways exchange language. 2004. Level 1, Version 1.0 Documentation.

106. http://www.cellml.org/models

107. Le Novere N, Bornstein B, Broicher A, et al. BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. 2006;34:D689-D691

108. Lloyd C.M., Halstead MD, Nielsen PF. CellML: its future, present and past. Prog. Biophys. Mol. Biol. 2004;85:433-450.

109. Hunter P, Nielsen P. A strategy for integrative computational physiology. Physiology (Bethesda) 2005;20:316-325.

110. Lloyd CM, Lawson JR, Hunter PJ, et al. The CellML Model Repository. Bioinformatics. 2008 Sep 15;24(18):2122-3.

111. Sauro HM, Bergmann FT. Standards and ontologies in computational systems biology. Essays Biochem. 2008;45:211-22.

112. http://www.openbel.org/bel-expression-language

113. Catlett NL, Bargnesi AJ, Ungerer S, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics. 2013 Nov 23;14:340.

114. Slater T. Recent advances in modeling languages for pathway maps and computable biological networks. Drug Discov Today. 2014 Feb;19(2):193-8.

115. Chindelevitch L, Ziemek D, Enayetallah A, et al. Causal reasoning on biological networks: interpreting transcriptional changes. Bioinformatics. 2012 Apr 15;28(8):1114-21.

116. Huang CL, Lamb J, Chindelevitch L, et al. Correlation set analysis: detecting active regulators in disease populations using prior causal knowledge. BMC Bioinformatics. 2012 Mar 23;13:46.

117. Kodamullil AT, Younesi E, Naz M, Bagewadi S, Hofmann-Apitius M. Computable

cause-and-effect models of healthy and Alzheimer's disease states and their

mechanistic differential analysis. Alzheimers Dement. 2015 Apr 4. pii: S1552-5260(15)00083-7.

118. Nixon RA. Endosome function and dysfunction in Alzheimer's disease and other neurodegenerative diseases. Neurobiol Aging. 2005 Mar;26(3):373-82.

119. Cataldo AM, Barnett JL, Pieroni C, et al. Increased neuronal endocytosis and protease delivery to early endosomes in sporadic Alzheimer's disease: neuropathologic evidence for a mechanism of increased beta-amyloidogenesis. J Neurosci. 1997 Aug 15;17(16):6142-51.

120. Cataldo AM, Peterhoff CM, Troncoso JC, et al. Endocytic pathway abnormalities precede amyloid beta deposition in sporadic Alzheimer's disease and Down syndrome: differential effects of APOE genotype and presenilin mutations. Am J Pathol. 2000 Jul;157(1):277-86.

121. Bekris LM, Galloway NM, Millard S, Lockhart D, Li G, Galasko DR, Farlow MR,Clark CM, Quinn JF, Kaye JA, Schellenberg GD, Leverenz JB, Seubert P,

Tsuang DW, Peskind ER, Yu CE. Amyloid precursor protein (APP) processing genes and cerebrospinal fluid APP cleavage product levels in Alzheimer's disease. Neurobiol Aging. 2011 Mar;32(3):556.e13-23.

122. Qin W, Jia L, Zhou A, et al. The -980C/G polymorphism in APH-1A promoter confers risk of Alzheimer's disease. Aging Cell. 2011 Aug;10(4):711-9.

123. Choy RW, Cheng Z, Schekman R. Amyloid precursor protein (APP) traffics from the cell surface via endosomes for amyloid β (Aβ) production in the trans-Golgi network. Proc Natl Acad Sci U S A. 2012 Jul 24;109(30):E2077-82.

124. Decourt B, Mobley W, Reiman E, et al. Recent Perspectives on APP, Secretases, Endosomal Pathways and How they Influence Alzheimer's Related Pathological Changes in Down Syndrome. J Alzheimers Dis Parkinsonism. 2013 Mar 20;Suppl 7:002.

125. Grant BD, Donaldson JG. Pathways and mechanisms of endocytic recycling. Nat Rev Mol Cell Biol. 2009 Sep;10(9):597-608.

126. Kinoshita A, Fukumoto H, Shah T, et al. Demonstration by FRET of BACE interaction with the amyloid precursor protein at the cell surface and in early endosomes. J Cell Sci. 2003 Aug 15;116(Pt 16):3339-46.

127. McMahon HT, Boucrot E. Molecular mechanism and physiological functions of clathrin-mediated endocytosis. Nat Rev Mol Cell Biol. 2011 Jul 22;12(8):517-33.

128. Parikh I, Medway C, Younkin S, et al. An intronic PICALM polymorphism, rs588076, is associated with allelic expression of a PICALM isoform. Mol Neurodegener. 2014 Aug 29;9:32.

129. Grassart A, Cheng AT, Hong SH, et al. Actin and dynamin2 dynamics and interplay during clathrin-mediated endocytosis. J Cell Biol. 2014 Jun 9;205(5):721-35.

130. Aidaralieva NJ, Kamino K, Kimura R, et al. Dynamin 2 gene is a novel susceptibility gene for late-onset Alzheimer disease in non-APOE-epsilon4 carriers. J Hum Genet. 2008;53(4):296-302.

131. Seaman MN. The retromer complex - endosomal protein recycling and beyond. J Cell Sci. 2012 Oct 15;125(Pt 20):4693-702. doi: 10.1242/jcs.103440. Epub 2012 Nov 12.

132. Vardarajan BN, Bruesegem SY, Harbour ME, et al. Identification of Alzheimer disease-associated variants in genes that regulate retromer function. Neurobiol Aging. 2012 Sep;33(9):2231.e15-2231.e30.

Summary

As the previous version of BEL did not support the representation of genetic variation information, I have proposed an extended syntax for OpenBEL that enables the coding of genetic variants in biomedicine models. I have proposed a representation for different genetic variant types: substitution, insertion-deletion and intergenic, by introducing new variant functions: g(SNP:rs113993960,var(delCTT)), r(REF:NM_000492.3,var(r.1653_1655delcuu)), and p(HGNC:CFTR,var(p.Phe508del)) for DNA, RNA and protein levels respectively.

The extended BEL syntax is designed to support reasoning over cause-effect models that include genetic variation information. This work therefore aims to develop a rationale for the mapping of genetic variants to nodes in a BEL network, and to generate a rule-set that supports reasoning over a BEL model enriched with genetic variance information.

In this article, I have suggested a mechanistic approach for the interpretation of disease-associated risk variants in complex systems biomedicine models. I have categorized genetic variants according to their predicted functional impact. Variants identified by GWA and eQTL studies are integrated with causal mechanisms through a multi-scale interconnection network (including genome – transcriptome – proteome) of epigenetic and genetic alterations, located within significantly influential genomic regions. Associated validating indicators (“supportive evidences”) are derived from clinical or experimental outcomes. Other potential validation signals at the molecular level may include protein abundance variation, protein isoform detection, mRNA splicing deviation and differential gene expression. Subsequently, these SNPs are

mapped with both well-established and novel disease-associated genes, to a disease model encoded in BEL.

Through the annotation of disease models with SNPs ranked according to the functional relevance-scoring scheme, we enable an enhanced interpretation of the functional consequences of SNPs in a mechanistic context. We can ask questions such as “if this SNP modifies the effect of an upstream controller identified by reverse causal reasoning, can this result in a stronger dysregulation event downstream of the chain of causation”? This approach will allow us to assess the functional consequences of entire SNP panels in a given mechanistic context and to estimate the contribution of complex SNP patterns to dysregulation processes at the systems level.

Chapter 3

Systematic analysis of GWAS data to identify genomic hotspots for shared disease mechanisms

Introduction

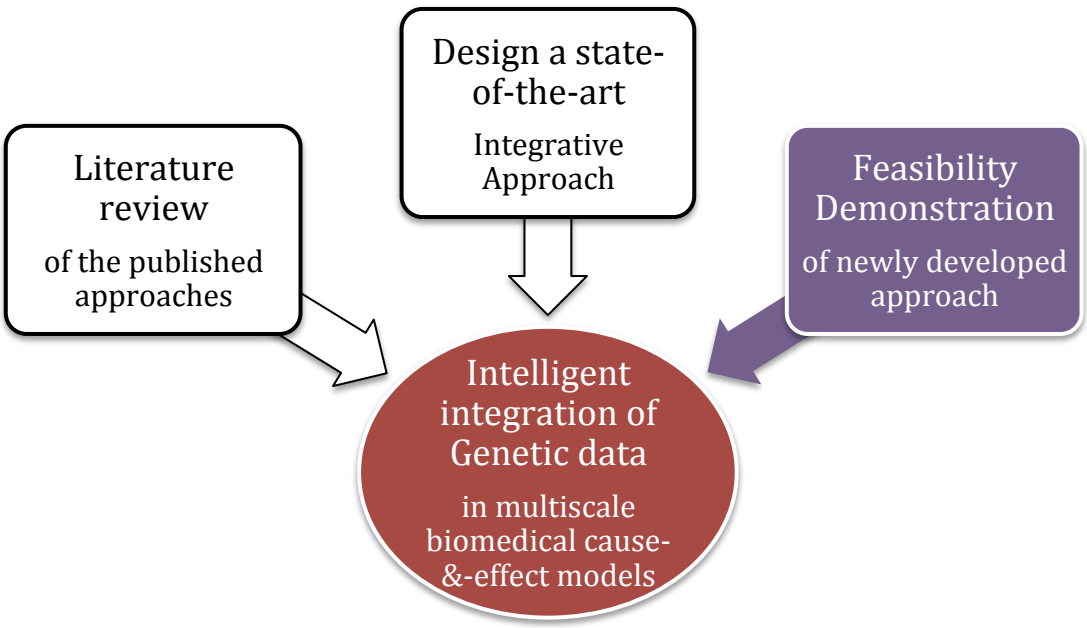


Figure 22: Part –III: Workflow for the intelligent functional interpretation of genetic variance information in multi-scale biomedical cause-and-effect models

Genome wide association studies (GWAS) have delivered a large set of genome loci inducing "risk loci" for many common diseases. Such association studies normally investigate one specific trait in a population in a discrete way without considering the potential genetic and phenotypic correlation between diseases. However, along with literature and clinical evidences, GWAS data can be resourcefully used to annotate overlapping loci with analogous or contrasting effects on different diseases.

I described here a systematic approach to interpret GWAS data and emphasised the analysis of shared genetic variants across these diseases. Moreover, I conducted a comprehensive literature search to identify shared genes and extract shared genomic loci from GWAS for disease pairs, to assess overlapping stretches of the genome shared between them.

The observation revealed that a high fraction of the SNPs studied in this work are associated with related diseases, which provides suggestive evidence that the molecular mechanisms influencing aetiology and progression of selected neurodegenerative diseases are partly interrelated. Genetic overlap between these diseases also suggests the significance of the affected entities in the specific pathogenic pathways, these should be investigated experimentally.

Systematic analysis of GWAS data reveals genomic hotspots for shared mechanisms between neurodegenerative diseases

Mufassra Naz[1,2], Erfan Younesi[1], Martin Hofmann-Apitius[1,2 *]

[1]*Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin and* [2]*Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for IT, Dahlmannstrasse 2, 53113 Bonn, Germany*

J Alzheimers Dis Parkinsonism 2017, Vol 7(5): 368
doi: 10.4172/2161-0460.1000368

ABSTRACT:

Objective: Genome wide association studies (GWAS) have delivered a large set of genome loci inducing "risk loci" for many common diseases. Such association studies normally investigate one specific trait in a population in a discrete way without considering the potential genetic and phenotypic correlation between diseases. However, along with literature and clinical evidences, GWA data can be resourcefully used to annotate overlapping loci with analogous or contrasting effects on different diseases.

Methods: In the work presented here, we describe a systematic approach to interpret GWA data and related Linkage Disequilibrium (LD) variants for neurodegenerative diseases. We emphasize on the analysis of shared genetic variants across these diseases. A comprehensive literature search was conducted to extract a list of shared genes for disease pairs; this list of shared genes was subsequently compared to a list

of genes identified by a systematic analysis of GWAS genetic variants, to identify overlaps between them. Overlapping genome stretches comprising “shared loci” of variants were used to create a candidate gene list for genes potentially involved in disease progression mechanisms. In the course of a “functional enrichment process”, we mapped that gene list from shared loci to pathway information to identify shared molecular level perturbation in the pathophysiology of related diseases.

Results: The observation revealed that a high fraction of the SNPs studied in this work are associated interlacing with related diseases, which provides suggestive evidence that the molecular mechanisms influencing aetiology and progression of selective neurodegenerative diseases are at least partly interrelated.

Conclusion: Genetic overlaps between these diseases suggest the significance of the affected entities in the specific shared pathogenic pathways.

KEY WORDS: GWAS, LD (Linkage Disequilibrium), Shared Genetic Loci, Genetic Variants, Shared Pathology, Neurodegenerative Diseases

BACKGROUND:

Genome wide association studies have been very useful for the identification of genetic variants as disease risk markers; however, the impact of these genetic variants in disease aetiology remains largely unclear. In this study, we tried to unravel molecular mechanisms underlying “shared genetic variants” and developed a strategy to identify candidate mechanisms for shared aetiology of diseases that display similar patterns of genetic variation organized in shared genomic hotspots. We demonstrate, how this approach leads to new insights that help to uncover biological relationships between quantitative traits or related neurodegenerative diseases.

Many traits or diseases have been shown to share genetic architecture [1,2]. This phenomenon, that a genetic variant affects multiple phenotypes, is often called 'pleiotropy' [3-5]. Such pleiotropic variants are particularly interesting, as the functional impact of a SNP on one or several genes may provide clues about the underlying molecular mechanism. For example, a significant overlap of shared genetic variants and pathways has been detected in immune-mediated diseases, suggesting extensive pleiotropic effects [6-8]. These shared genetics variants linked to pathways are ideally suited to identify candidate mechanisms underlying a "shared aetiology" of different diseases.

So far, various studies have been implemented across the genome, mostly on those groups of diseases, which are already well recognized or hypothesized to be interconnected [6, 8-10] or by investigating influence of individual genetic variant on a wide range of diverse diseases [11-13].

Biologically, a genetic variant can influence different traits fundamentally in two different ways; firstly, it can influence two distinct phenotypes through two independent physiological mechanisms, while secondly, its effect on the second trait can be mediated through its effect on the first one.

Apparent genetic similarities in a pair of distinct diseases may be indicative for potential overlaps in the underlying disease mechanisms. Thus investigating common factors and network modules shared within a pair of distinct, but related diseases, may point at shared mechanisms. Rather than studying individual diseases separately,

investigation and analysis of common dysregulated pathways or dysfunctional proteins of a pair of related diseases can be expected to reveal deeper comprehensive knowledge about pathophysiological processes.

Correspondingly, computing of shared molecular level mechanisms of related disorders can not only assist understanding of the etiology of a disease; but also such associations between shared pathways, and correlation with biological processes can accelerate drug discovery efforts by suggesting promising treatment candidates for already approved drugs (known as drug repositioning) [14].

In the work presented here, we performed a systematic and comprehensive analysis of shared genomic loci likely to represent genomic hotspots with genes functionally involved in the aetiology of neurodegenerative diseases. We go way beyond classical meta-analysis of GWAS data by performing a 'functional context enrichment' that is tailored to embed candidate genes in these genomic hotspots in a mechanistic context. We demonstrate that this functional enrichment can lead to the identification of new candidate mechanisms for shared aetiology of Alzheimer´s Disease and Parkinsonism.

RESULTS:

In an initial step, we selected five different brain diseases including Alzheimer's disease (AD), Parkinson disease (PD), Schizophrenia, Multiple sclerosis (MS) and Type 2 diabetes mellitus (T2DM).

Spatial analysis, after mapping of GWAS disease-associated intronic SNPs to the genes, they belong to; and intergenic SNPs to the most likely, nearby genes; reveals that most of the GWAS SNPs are located around specific genome loci ("genomic

hotspots"). Our assumption is, that the genes existing in the vicinity of these genome loci may play a role in the dysregulation of disease-associated pathways.

Moreover, we computed pair-wise analysis for shared genetic variants to see the relevancy between each pair of diseases. However, enumerating of pair-wise shared GWAS SNPs before LD SNPs enrichment revealed that only a very limited number of individual SNPs are shared in a pair of diseases, while after LD analysis, most of the disease pairs showed a substantial count of shared variants; which also signify the genetically linkage between SNPs located on these specific genomic loci around GWAS SNPs. Thus it can be explained that these pairs of disease may share disease-associated genomic loci rather than individual genetic variants [Supplementary File].

Pair-wise analysis also revealed that AD and PD have the largest number of shared disease-associated loci. There is no doubt, that this is reflecting the bias that comes with the higher number of GWAS studies and available data around these two diseases. But it also may indicate an overlap of the genetics relevant for pathophysiology mechanisms shared between AD and PD. Other disease pairs, for instance the AD – T2DM pair, did also show a promising number of shared genetic markers and genomic loci. Successively, pairs of AD – Schizophrenia and AD – MS also presented a reasonable number of shared SNPs and genomic loci [Table 1].

Shared SNPs and Genes Count for 10 pairs of 5 Diseases		
Disease Pair	**Shared SNP Count**	**Shared Gene Count**
AD – PD	35958	1793
AD – T2DM	2187	103
AD – Schizophrenia	867	46
AD – MS	771	62
PD – Schizophrenia	701	24
PD – T2DM	463	21
MS – Schizophrenia	421	28
T2DM – MS	250	17
PD – MS	246	19
T2DM – Schizophrenia	223	18

Table 1: List of disease pairs with GWAS associated shared genetic variants and genes count.

The analysis of specific overlapping genome stretches between 'loci identified for shared GWAS-LD genetic variants' and 'loci of already established disease-associated genes in the literature' revealed that there was a quite significant overlap between GWAS loci and literature based disease-associated gene loci [Supplementary File], which provides suggestive evidence for an association between genetic variants and disease pathology.

Analysis of putative shared pathways was done by mapping genes in genomic hotspots to pathways using MsigDB. Shared pathways – as a functional layer on top of shared genetics - are indicative for putative pathology mechanisms shared between pairs of diseases. The analysis workflow thus identifies disease pairs that do display a high number of shared genomic hotspots, a significant number of putative shared pathways and – as a consequence - may have significantly shared molecular level mechanisms, that – when perturbed - may contribute to disease etiology.

To explore the pathophysiology of putative shared mechanisms in detail, we selected the pair of AD and PD for a mechanistic case study. Amongst their shared genomic

loci, we selected the well-known Tau locus, located on chromosome 17, to explore further detailed molecular mechanisms, as it showed top ranked "functional consequences" scores, based on ENCODE data, the Roadmap Epigenome Consortium data, DNase footprinting analysis, and DNA Methylation data. Apparently, selecting the tau locus seems to add nothing new and novel, as the tau locus is already well known and has been studied in detail. However, this locus has never been studied in a comprehensive way by "embedding" all the affected genes in that locus into a functional context. In our analysis, we expand the mechanistic context associated with the genes in the tau locus, by collecting and assembling all genetic, molecular and statistical evidences from the literature, from patents, from gene expression studies and from knock-out experiments in one comprehensive mechanistic model. In the following, we are presenting this locus in a very novel and unique perspective of stress induced shared pathology of AD and PD.

This genomic hotspot around tau is highlighted in many association studies for multiple statistically significant SNPs (references). The hotspot covers approximately 1 MB of a chromosomal region characterized by linkage disequilibrium region, that contains a large number of genetic variants.

Three genes are prominent in this locus: MAPT (Microtubule-Associated Protein Tau), the CRHR1 receptor-1 (Corticotropin Releasing Hormone Receptor 1) and the CRHR1-IT1 gene (CRHR1 Intronic Transcript 1). These genes are linked to several disease-associated genetic markers mapping to both, coding and non-coding regions. Moreover, disease-associated intergenic and intronic SNPs of this locus have several eQTL links with neighboring genes [Supplementary File].

In the course of our investigation of this shared genomic locus, we identified that, other than AD and PD, it also has well-established associations with Stress and Depression phenotypes. We searched for potential genetic, molecular and statistical evidences from the scientific literature and collected additional evidences from patents, gene expression studies and knock-out experiments, that all support the notion of a shared molecular mechanism linking Stress, AD and PD [Figure 3].

To enrich the genetics-driven identification of candidate genes with functional context and to identify potential mechanisms that bear explanatory potential for the presumed shared etiology linked to this particular locus on chromosome 17, we performed a systematic literature analysis using our literature mining environment SCAIView [15]. Contextual information relevant to the previously identified, disease-associated tau locus and being specific for the context of AD and PD was systematically identified and harvested. The extracted information comprises cause-and-effect relationships representing protein-protein interactions, protein inhibitory and activating patterns, protein-complex formation, insights from disease animal model studies, patterns from knockout and gene expression studies, other genetic associations; from gene mapping (fine-mapping) and GWAS meta-analysis studies, and from drug effects; all with high specificity for either AD or PD or both. The vast amount of information extracted was subsequently encoded using the OpenBEL (Open Biological Expression Language) syntax to construct a cause-and-effect computable model [16]. Models were developed separately for human and mouse. The resulting, comprehensive BEL models represent the state of published knowledge in the context of the genes under investigation in the context of AD and PD; the models are then visualized by Cytoscape_v2.8.3 [17], and queried for disease

associated molecular mechanisms to unravel mechanistic context that link molecular level perturbation with the disease etiology.

The mouse model reveals that repeated stress induces the expression of the wild-type CRH (Corticotropin Releasing Hormone) gene, while inactivation of the CRHR1 gene, which is the receptor of CRH, not only inhibits the complex of CRH+CRHR1 but also causes a reduction of MAPT phosphorylation and a reduction of Amyloid beta (Aβ) peptide concentration [18]. CRHR2, which is another receptor of CRH, inhibits MAPT phosphorylation; moreover, it has been shown to down-regulate the expression of CDK5, ERK, GRK and JNK genes [19]. As further supportive evidence for the functional antagonism between CRHR1 and CRHR2, the CRHR1 inhibitor 'Antalarmin' blocks CRHR1 and its complex with CRH; the inhibitor also causes a reduction of MAPT phosphorylation [19] [Figure 1].

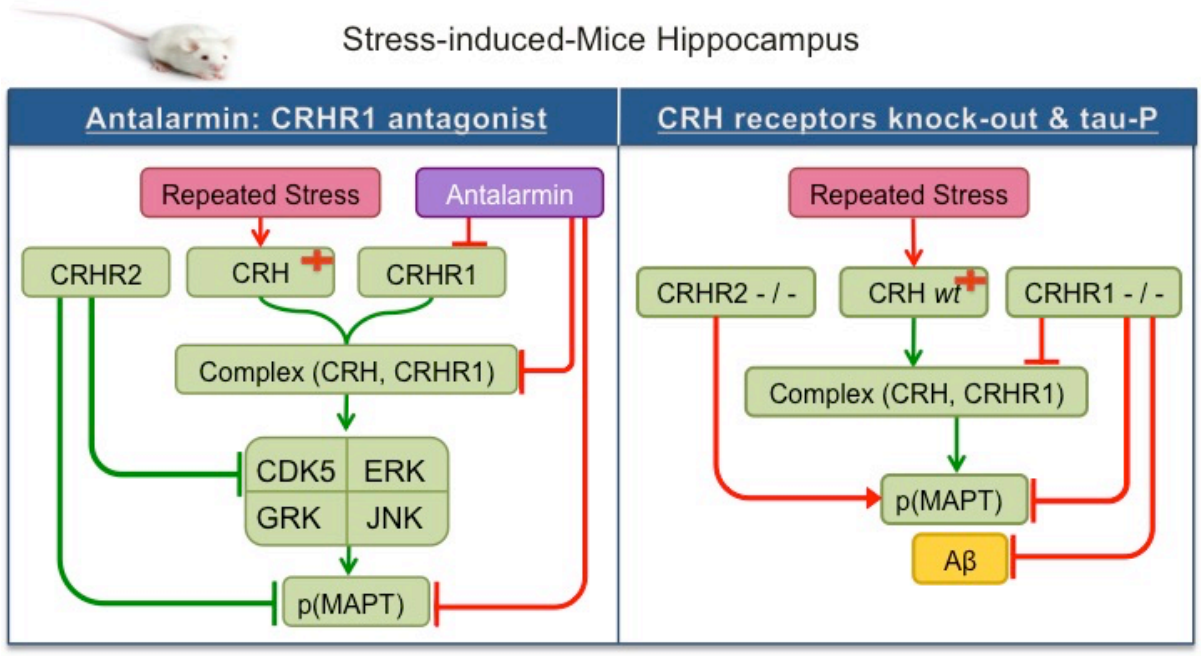


Figure 1: Experimental evidences for CRH gene and its receptors CRHR1 and CRHR2 and CRHR1 antagonist 'Antalarmin': *Experimental evidences for CRH gene and its receptors CRHR1 and CRHR2 and CRHR1 antagonist 'Antalarmin',*

collected from mouse model knock-out experiments to identify their role in the stress induced mice Hippocampus: Repeated stress induces the expression of the wild-type CRH gene, while inactivation of the CRHR1 gene, which is the receptor of CRH, not only inhibits the complex of CRH+CRHR1 but also causes a reduction of MAPT phosphorylation and a reduction of Amyloid beta (Aβ) peptide concentration. CRHR2, which is another receptor of CRH, inhibits MAPT phosphorylation; moreover, it has been shown to down-regulate the expression of CDK5, ERK, GRK and JNK genes. As further supportive evidence for the functional antagonism between CRHR1 and CRHR2, the CRHR1 inhibitor 'Antalarmin' blocks CRHR1 and its complex with CRH; the inhibitor also causes a reduction of MAPT phosphorylation.

The contextual BEL model specific for human pathophysiology demonstrates that a genetic variant rs1800547, located on the intron region of the MAPT gene, is positioned on the Haplotype-1 region of chromosome 17, which is associated with PD [20, 21]. Its 'A' allele is associated with 'Dementia in PD patients' and its 'G' allele with 'familial FTD' [22]. This genetic variant is linked with the expression of host gene MAPT and also expression of a neighboring gene CRHR1; thus, its 'A' allele is associated with an up-regulation of MAPT and a concomitant down-regulation of the CRHR1 gene, while the 'G' allele is associated with an up-regulation of the CRHR1 gene [22]. In addition, the 'A' allele containing SNP rs1800547 is linked to a neuro-imaging readout, the reduction of gray matter volume [22]. Likewise, the 'A' allele of another SNP named rs393152, which is located near the CRHR1 gene, is associated with the up-regulation of the MAPT gene [23]. Moreover, the rs393152-A allele has been associated with AD and PD, and seems to be linked to a reduction of gray matter volume as well as atrophy of the hippocampus and entorhinal cortex [23][Figure 2].

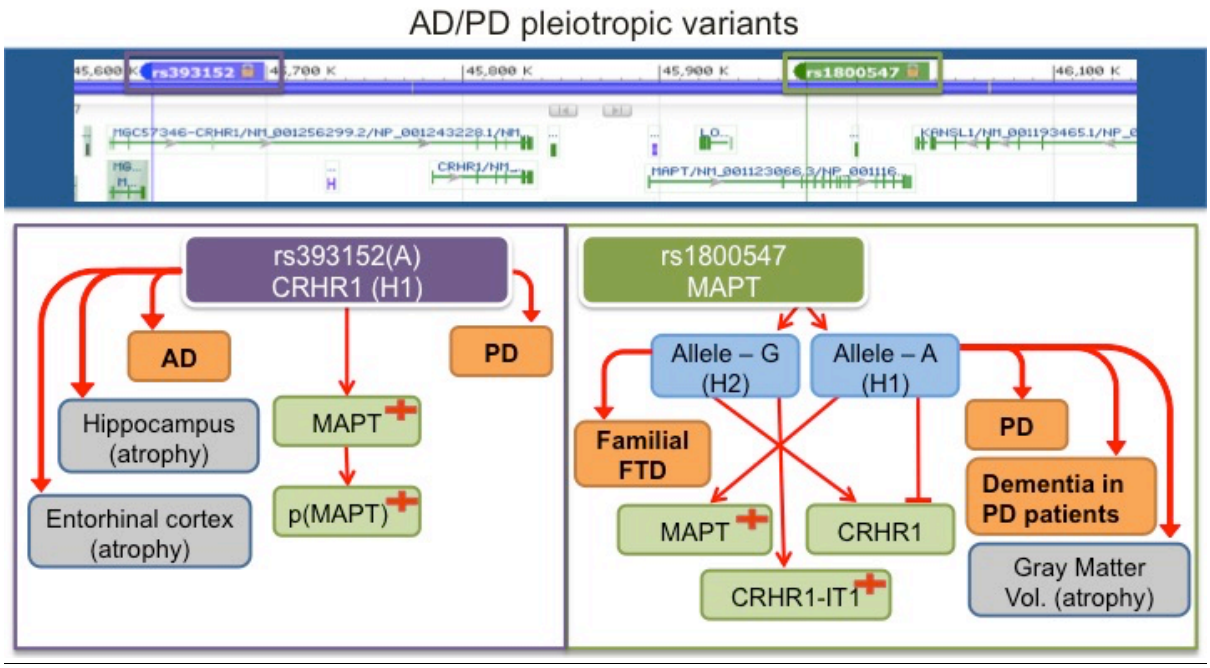


Figure 2: Experimental evidences for AD/PD pleiotropic variants, to identify the functional consequences for these variants: *A genetic variant rs1800547, located on the intron region of the MAPT gene, is positioned on the Haplotype-1 region of chromosome 17, which is associated with PD. Its 'A' allele is associated with 'Dementia in PD patients' and its 'G' allele with 'familial FTD'. This genetic variant is linked with the expression of host gene MAPT and also expression of a neighboring gene CRHR1; thus, its 'A' allele is associated with an up-regulation of the MAPT and a down-regulation of the CRHR1 gene, while the 'G' allele is associated with an up-regulation of the CRHR1 gene. Moreover, the 'A' allele containing SNP rs1800547 is also linked to a neuro-imaging readout, the reduction of gray matter volume. Likewise, the 'A' allele of another SNP named rs393152, which is located near the CRHR1 gene, is associated with the up-regulation of the MAPT gene. Moreover, the rs393152-A allele is associated with AD and PD, and also linked with a reduction of gray matter volume as well as atrophy of the hippocampus and entorhinal cortex.*

BEL models are excellent tools to represent complex physiology; the representation in BEL bears great explanatory potential on how complex physiology works across scales. Our contextual BEL models representing complex physiology of genes in the Tau locus provide a mechanistic explanation, how excessive and repeated stress may modulate pathophysiology. Repeated stress induces the expression of the CRH gene in the hippocampal area [24], while under AD conditions, reduced CRH immune-reactivity is observed [25]. CRH interacts with its receptor, the CRHR1 protein; the CRHR1 gene is highly expressed in hippocampus and the complex between the hormone and its receptor (CRH+CRHR1) can be detected in that brain region [26]. In addition, the CRHR1 protein also interacts with γ-secretase, which is associated with Aβ accumulation, one of the hallmarks of AD pathophysiology [27].

The hormone receptor protein complex (CRH+CRHR1) is further linked to the up-regulation of GSK3β and the phosphorylation of essential elements of the ERK1/2/MAPK pathway [19, 28]. Up-regulation of GSK3β is associated with MAPT hyper-phosphorylation [28, 29]; in addition, phosphorylated MAPT and ERK1/2/MAPK pathway up-regulate Neurofilament phosphorylation, which has been associated with AD [19, 28]. The complex physiology is even increased through the interaction of the 'CRH+CRHR1' protein complex with the BDNF protein; this interaction has already been associated with AD pathology [30]. The complex also enhances neuronal activity by interacting with adenylate cyclase, cAMP, act(PAK), Ca2 signaling pathways [30]. The resulting enhanced neuronal activity has been shown to further accumulate interstitial fluid amyloid beta (ISF Aβ), while this accumulation of ISF Aβ is also linked with up-regulation of CRH gene expression [31], effectively establishing a feedback loop that can enhance negative dysregulation

events. MAPT hyper-phosphorylation also increases its dissociation from microtubules, a process that has been linked to lewy-bodies and Parkinsonism, in the PD context [32].

Finally, the CRHR1 antagonist 'Antalarmin', which is used in response of chronic stress, has been shown to reduce Aβ accumulation in brain [33][Figure 3], adding further meaningful, supportive evidence in context.

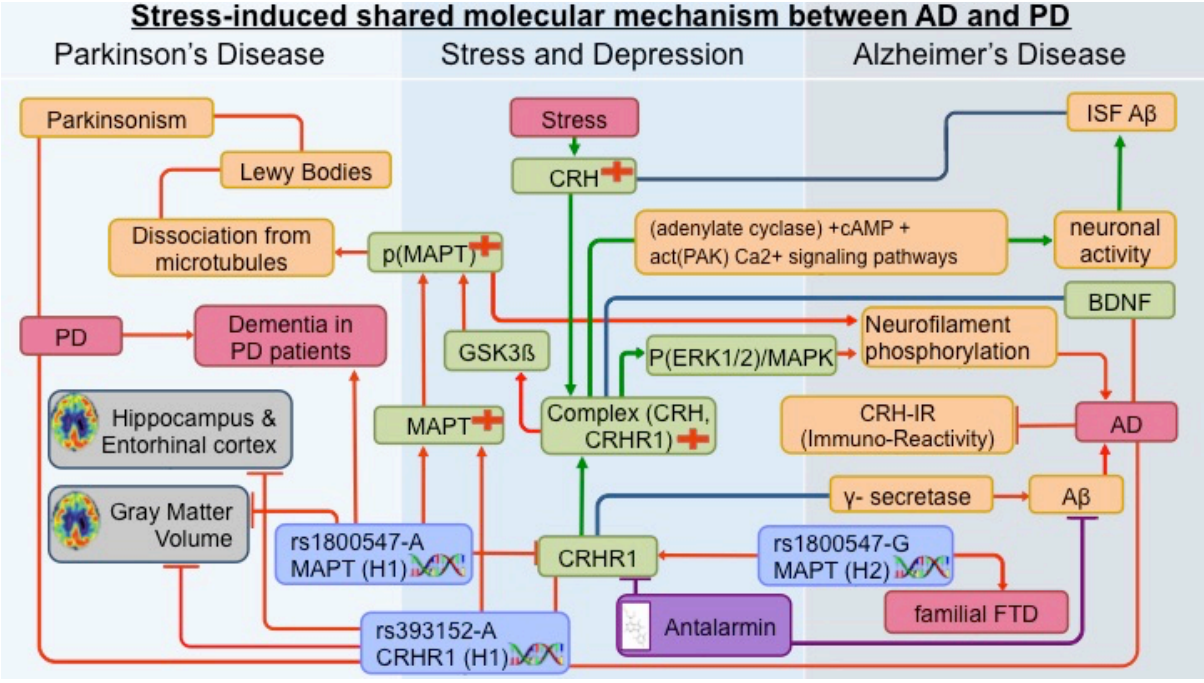


Figure 3: Stress induced comorbidity association of AD and PD by genetic variants of Tau locus genes: *Stress up regulate CRH gene expression, which interacts with its receptor, the CRHR1 protein; the CRHR1 gene is highly expressed in hippocampus and the complex between the hormone and its receptor (CRH+CRHR1) can be detected in that brain region. In addition, the CRHR1 protein also interacts with γ-secretase, which is associated with Aβ accumulation, one of the hallmarks of AD pathophysiology. The hormone receptor protein complex*

(CRH+CRHR1) is further linked to the up-regulation of GSK3β and the phosphorylation of essential elements of the ERK1/2/MAPK pathway. Up-regulation of GSK3β is associated with MAPT hyper-phosphorylation; in addition, phosphorylated MAPT and ERK1/2/MAPK pathway up-regulate Neurofilament phosphorylation, which has been associated with AD. The complex physiology is even increased through the interaction of the 'CRH+CRHR1' protein complex with the BDNF protein; this interaction has already been associated with AD pathology. The complex also enhances neuronal activity by interacting with adenylate cyclase, cAMP, act(PAK), Ca2 signaling pathways. The resulting enhanced neuronal activity has been shown to further accumulate interstitial fluid amyloid beta (ISF Aβ), while this accumulation of ISF Aβ is also linked with up-regulation of CRH gene expression, effectively establishing a feedback loop that can enhance negative dysregulation events. MAPT hyper-phosphorylation also increases its dissociation from microtubules, a process that has been linked to lewy-bodies and Parkinsonism, in the PD context. Finally, the CRHR1 antagonist 'Antalarmin', which is used in response of chronic stress, has been shown to reduce Aβ accumulation in brain, adding further meaningful, supportive evidence in context.

Additionally, exhaustive analysis of relevant patent literature revealed that an antagonist against the CRH receptor is effective as a prophylactic or therapeutic agent for diseases, like, anxiety, depression, AD and PD [34]. Patents also describe multiple lines of evidence that suggest the significant role of CRHR1 on neuropsychiatric disorders, and MAPT gene as a well-studied candidate gene for neuropsychiatric disorders [35]. Moreover, it is also patented that although the two receptors CRHR1 and CRHR2 share 70% sequence identity, they differ substantially in ligand binding

affinity, and the CRH gene itself has a much higher affinity for CRHR1 rather than CRHR2 [36]. Another patent describes that the accumulation of hyperphosphorylated tau protein in the central nervous system, may be reduced through the administration of CRHR1 selective antagonists and/or CRHR2 selective agonists. Patent [EP2522351 A1] indeed describes methods for the prevention of the onset of Alzheimer's disease by the administration of CRHR1 selective antagonists [37] [Supplementary File].

DISCUSSION:

In the work presented here, we established an integrative approach that starts with a data-driven approach, identifies signals in GWAS data, and gains explanatory potential and allows for new insights into putative complex mechanisms through knowledge-driven context enrichment. Our approach goes way beyond classical "pathway enrichment" approaches, as it takes multimodal information into account and integrates heterogeneous information and knowledge in biologically meaningful, computable graph models. Data, a priori knowledge and inferred insights are combined in a seamless fashion. Meaningful cause-and-effect relationships are established and the signals originally identified are made interpretable in a rational modeling and mining approach.

Our workflow is tailored towards the identification of novel shared mechanisms. It starts with comparative GWAS analysis tailored to identify shared genetic variants; puts the enriched SNPs in contigs (based on linkage disequilibrium); identifies those genes belonging to the "shared" LD loci, and establishes compelling evidence for

shared molecular mechanisms and biological pathways associated with those genes, for a given pair of disease. The workflow was applied to a comprehensive set of related diseases and allowed us to investigate shared molecular level mechanisms between a pair of diseases, based on both, data driven and knowledge driven strategies.

We would like to emphasize that genomic loci (genomic hotspots) should be considered to investigate the effects of GWAS variants rather than individual genetic variants, particularly to investigate shared pathology. Whereas the biological impact of single SNPs is often hard to predict, the association of several SNPs in a disease-associated LD block provides evidence for a much stronger association that may affect an entire locus with several genes. As a consequence, a set of SNPs in a genomic hotspot may contribute to dysregulation events involving several genes.

Modeling the functional context of these genes in computable cause-and-effect models can be very helpful to identify possible molecular level perturbation mechanisms that contribute to disease pathology. As such, computable mechanistic models are essential to integrate diverse types of data as well as relationships between the nodes; they can help to discover unknown links to illustrate the possible mechanism of dysregulation.

At this point we would like to stress that we are not talking about pathways when we talk about mechanisms. Although pathways are abstractions of biological functional context that is shared by many cell types and often conserved across species boundaries, the pathway concept as it was established over the last 30 years is not

taking into account genetic variation information and is not well-suited to take into account the specifics of cell-cell-interactions. “Chains of causation” as we find them in the BEL model graphs may as well exist in pathways, but in pathways they are confined to one type (one “mode” or “level”) of information. Integrative models based on causal relationships, however, span over multiple levels and scales and establish links e.g. from SNPs to imaging features in one single, computable graph model. We would encourage the community to clearly distinguish between pathways (representing canonical information) and mechanisms (representing causes and effects associated with a disease context). Mechanistic modeling allows us to be highly specific with respect to the available knowledge in a given context, without restricting us to make use of canonical knowledge if we wish to include that type of common information.

The mechanistic hypothesis generated from our ‘tau locus BEL model’ establishes a rational, how stress could cause deficits in memory [38-43]. We may actually have established a functional context that puts a “sensor” for environmental and life style into a pathophysiology mechanism that could play a significant role in the etiology of Alzheimer’s disease. Our model provides also mechanistic clue, how hippocampal atrophy may be linked to the pathophysiology of stress [38, 40-43]. The stress-related HPA axis activation (linked to the CRH-CRHR1 complex) may thus represent a pathophysiological initiation of memory loss [41]. Likewise, it is reported that the decline in CRH Immuno-Reactivity (CRH-IR) in AD is due to the reciprocal accumulation of CRH receptors in affected cortical areas [26]. The alteration in pre- and postsynaptic indicators for CRH is significantly correlated with decline in ChAT (choline acetyltransferase) activity [26].

The H1 haplotype of MAPT extends towards the 5′ region and includes the contiguous gene CRHR1. Linkage Disequilibrium (LD) of this region is substantially associated with PD patients [44]. Strikingly, the oldest and most extensively case-control studies for PD demonstrated the greatest evidence for MAPT and H1 haplotype association. By genotyping H1 haplotype SNPs within the CRHR1-MAPT interval, we can hypothesize that the CRHR1 gene may be responsible for at least part of the disease association of this locus due to the genetic variability and could become a good biomarker candidate, since it is significantly involved in both, immune and nervous systems physiology [44]. Missense and splicing genetic variants in MAPT were first uncovered in 'frontotemporal dementia with parkinsonism' associated with chromosome 17 (FTDP-17) [20].

Thus, forgoing studies have already specified associative links between stress, CRH-CRHR1, and tau pathology mediated by CRH-CRHR1 dependent activation of tau kinases induced by stress [19,30,45]. On the other side, the H1 haplotype is associated with the accumulation of hyperphosphorylated Tau in neuronal cell bodies, which has always been associated with neurodegenerative diseases [46, 47]

Though AD and PD are likely to have different mechanisms underlying their etiology, and may affect different brain regions, and display different clinical features, still they have a significant overlap in the progression of neurodegenerative processes. A recent study has been investigating AD and PD GWAS SNPs to identify AD-PD pleiotropic genetic variants/loci. They found, that the 'A' allele of rs393152, within the CRHR1 and CRHR1-IT1 region (MAPT locus) on chromosome 17, with a MAF (minor allele

frequency) value of 23.1%, significantly increased AD and PD risk, additionally, that allele is linked to the up-regulation of MAPT expression [23]. With APOE-stratified GWAS, another study revealed that genetic variants in the chromosome 17q21.31 region are associated with AD [48].

Besides all the genetic evidences that support our mechanistic model, there is also evidence from pharmacology that adds to the plausibility of the pathophysiological context we have established. Rissman et. al. described that the selective CRHR1 blocker "antalarmin" blocks stress-induced escalations in tau phosphorylation. This points at a direct function for CRF-dependent signaling in the stress response [19]. Fully in line with this observation is the finding, that CRHR1 antagonist antalarmin is able to suppress amyloid beta accumulation associated with AD pathology, in mice [49].

It is also notable that CRHR1 has a vital role in inflammation [50, 51], and the CRHR1 antagonists that are used to treat depression [52, 53], also control peripheral inflammation [54-56]. Similarly, those antidepressants, which are known to modulate inflammatory responses, also confer protection against cytokine-induced depressive-like behavioral and biological modifications [57-61].

In a clinical study with MCI (Mild Cognitive Impairment), AD and control groups, Arsenault-Lapierre et al. [62] couldn't find group differences in cortisol levels. This contradicts several previous studies that found different cortisol levels between normal and AD [63-67], and between normal and MCI groups [67, 68]. Contrarily, it supports the literature that found no correlation between cortisol level and perceived

stress in different populations [69-71]. Whereas, in this study more MCIs were diabetic, and diabetic patients have been found to secrete higher cortisol levels [72-74], likewise more AD patients were on sedatives or antidepressants. These medications may affect levels of cortisol and perceived stress measured in the patients [62]. However, these findings support our hypotheses that with the mechanism introduced here and the link between HPA axis, genetics and major determinants of AD and PD pathophysiology, we see a source for the highly stochastic nature of sporadic NDDs.

Repetitively, in a recent publication, Park HJ stated that stress response meditated by CRH-CRHR1 mechanism could also contribute to AD pathogenesis [27]. But he also described that under some circumstances, CRHR1 antagonism does not achieve required results against acute stress-induced Aβ production, rather he suggested that either direct targeting of CRH or G protein-biased CRHR1 agonist that could suppress β-arrestin recruitment to CRHR1 might be required to effectively target associated pathway for therapeutic benefit in AD. [27]

METHODS:

GWAS disease-associated variants are identified throughout the entire genome. In order to reveal shared genomic hotspots, that could have been comprised candidate genes for shared molecular mechanisms between two or multiple neurodegenerative and related diseases, genetic variants were collected from GWAS catalog [75] with the threshold of p-values $< 1.0 \times 10^{-5}$ for five diseases; including Alzheimer's disease, Parkinson disease, Schizophrenia, Multiple sclerosis and Type 2 diabetes mellitus. These collected genetic variants were belonged to multiple disease association studies

and each association study was conducted with different sample sizes. Thus according to basic principle of meta-analysis, we combined the evidence for association from individual studies, with the implementation of appropriate weights, by using a whole genome association analysis toolset Metal [76], and normalized them for their different sample sizes.

Afterwards, Linkage Disequilibrium (LD) analysis was conducted separately for each disease by using Haploreg DB V.4.0 [77]. Next, shared genetic variants were queried by pair-wise analysis for ten pairs of disease of these five diseases.

Subsequently, we made use of the ENSEMBL variant database [78-80] as a reference database to map the SNPs with their relevant chromosome, location, gene, allele and potential functional features (intergenic SNPs were mapped to the nearest gene on the chromosome). Additionally, these shared SNPs were interpreted with the characteristics of predicted functional consequences by using RegulomeDB V.1.1 [81] to get annotation from current ENCODE data (updated with recent ENCODE releases: [82, 83]), Chromatin States data from the Roadmap Epigenome Consortium and updated data for DNase footprinting, PWMs, and DNA Methylation, and finally ranked the variant lists according to predicted functional consequences attributes.

Most of the GWAS identified genetic variants are located on the non-coding regions of the genome. In order to investigate, whether there are any overlapping genome stretches between the ‘loci of shared GWAS and LD genetic variants’ and ‘loci of the well-established disease-associated genes in the literature’; in addition to the data-driven approaches described above, a comprehensive knowledge driven approach was

also conduction, by searching systematically from literature with the help of a literature mining environment – SCAIView. [84].

To extract shared genes for a pair of disease from literature, we were queried via SCAIView for those genes, which were studied for both diseases comprised in a pair (i.e. for AD and T2DM disease pair: {(([MeSH Disease:"Alzheimer Disease"]) AND [MeSH Disease:"Diabetes Mellitus Type 2"]) AND [Human Genes / Proteins]}). This literature search was conducted in a pair-wise analysis of genes for all of the ten pairs of diseases. The extracted list of "shared genes for a pair of disease from literature" (represented in the workflow as 'List: A') from SCAIView, was then used to pinpoint overlaps by comparing it with the list of "genes mapped with shared GWAS-LD genetic variants for a pair of disease" (represented in the workflow as 'List: B'); and resulting file had 'shared genes for a pair of disease' common in GWAS-LD and Literature [Figure 4].

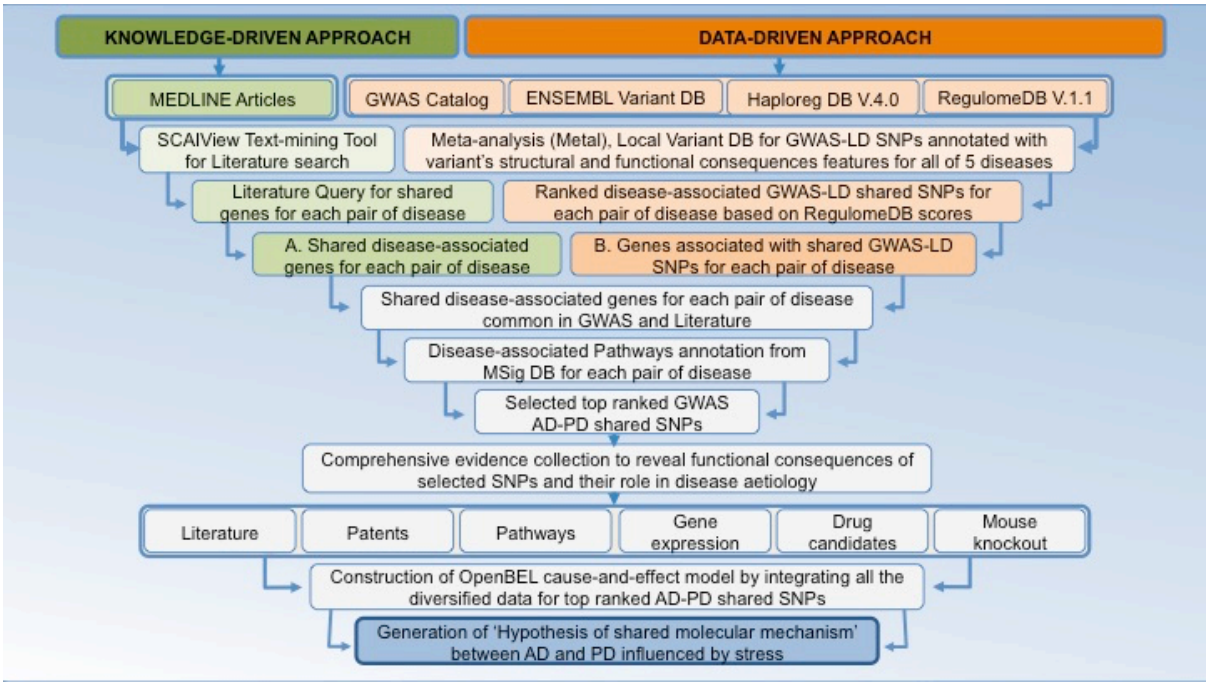


Figure 4: Flowchart for data analysis steps: *Flowchart for data analysis steps starting from GWAS Data collection to integration, mapping, annotation, filtering,*

Modeling and finalizing by Hypothesis generation.

Afterwards, we mapped these shared genes list to biological pathways by using MsigDB [85], to identify common pathways for each pair of disease. To demonstrate the potential of the approach, we did an exploratory study on one putative shared mechanism relevant for AD and PD. The genomics locus investigated maps to chromosome 17; to a region that displays highest scores for functional consequences in RegulomeDB, and one of high ranked shared pathway between AD and PD from MsigDB result table, that is 'KEGG_LONG_TERM_DEPRESSION' [Supplementary File]. The high-resolution analysis of that shared genomic locus for its potential role in the aetiology of the disease pair AD and PD includes – besides the identification of the candidate locus and the candidate genes within - the collection of evidences from gene expression studies, patents, knock-out studies and other literature, ultimately resulting in a comprehensive knowledge-driven approach towards the enrichment with supportive evidence.

Declarations

Acknowledgements

Not applicable.

Funding

This work was supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY grant agreement n°115568, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

Availability of data and materials

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Authors' contributions

EY proposed the idea; MN, EY and MHA designed data analysis process; MN analysed the data, generated the results and wrote the manuscript. MHA read, corrected and improved the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

REFERENCES:

1. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet. 89(5):607-18.
2. Rzhetsky A, Wajngurt D, Park N, Zheng T (2007) Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci U S A. 104(28):11694-9.
3. Stearns FW (2010) One hundred years of pleiotropy: a retrospective. Genetics. 186(3):767-73.
4. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW (2013) Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 14(7):483-95.
5. Paaby AB, Rockman MV (2013) The many faces of pleiotropy. Trends Genet. 29(2):66-73.
6. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 7(8):e1002254.
7. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, et al. (2016) Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet. 48(5):510-8.
8. Parkes M, Cortes A, van Heel DA, Brown MA (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. 14(9):661-73.
9. Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 381(9875):1371-9.

10. Fortune MD, Guo H, Burren O, Schofield E, Walker NM, et al. (2015) Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. Nat Genet. 47(7):839-46.
11. Li L, Ruau DJ, Patel CJ, Weber SC, Chen R, et al. (2014) Disease risk factors identified through shared genetic architecture and electronic medical records. Sci Transl Med. 6(234):234ra57.
12. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 31(12):1102-10.
13. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, et al. (2016) Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 48(7):709-17.
14. Li P, Nie Y, Yu J (2015) An Effective Method to Identify Shared Pathways and Common Factors among Neurodegenerative Diseases. PLoS One. 10(11):e0143045.
15. Hofmann-Apitius M, Fluck J, Furlong L, Fornes O, Kolárik C, et al. (2008) Knowledge environments representing molecular entities for the virtual physiological human. Philos Trans A Math Phys Eng Sci. 366(1878):3091-110.
16. Kodamullil AT, Younesi E, Naz M, Bagewadi S, Hofmann-Apitius M (2015) Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. Alzheimers Dement. 11(11):1329-39.

17. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498-504.
18. Campbell SN, Zhang C, Roe AD, Lee N, Lao KU, et al. (2015) Impact of CRFR1 Ablation on Amyloid-β Production and Accumulation in a Mouse Model of Alzheimer's Disease. J Alzheimers Dis. 45(4):1175-84.
19. Rissman RA, Lee K-F, Vale W, Sawchenko PE (2007) Corticotropin-releasing factor receptors differentially regulate stress-induced tau phosphorylation. J. Neurosci. 27:6552-6562.
20. Hutton M (2001) Missense and splice site mutations in tau associated with FTDP-17:multiple pathogenic mechanisms. Neurology. 56(11 Suppl 4):S21-5.
21. Skipper L, Wilkes K, Toft M, Baker M, Lincoln S, et al. (2004) Linkage isequilibrium and association of MAPT H1 in Parkinson disease. Am J Hum Genet. 75(4):669-77.
22. de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, et al. (2012) Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. BMC Genomics. 13:458.
23. Desikan RS, Schork AJ, Wang Y, Witoelar A, Sharma M, et al. (2015) Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. Mol Psychiatry. 20(12):1588-95.
24. Brunson KL, Avishai-Eliner S, Hatalski CG, Baram TZ (2001) Neurobiology of the stress response early in life: evolution of a concept and the role of corticotropin releasing hormone. Mol Psychiatry. 6(6):647-56.

25. Behan DP, Heinrichs SC, Troncoso JC, Liu XJ, Kawas CH, et al. (1995) Displacement of corticotropin releasing factor from its binding protein as a possible treatment for Alzheimer's disease. Nature. 378(6554):284-7.
26. De Souza EB, Whitehouse PJ, Price DL, Vale WW (1987) Abnormalities in corticotropin-releasing hormone (CRH) in Alzheimer's disease and other human disorders. Ann N Y Acad Sci. 512:237-47.
27. Park HJ, Ran Y, Jung JI, Holmes O, Price AR, et al. (2015) The stress response neuropeptide CRF increases amyloid-β production by regulating γ-secretase activity. EMBO J. 34(12):1674-86.
28. Refojo D, Echenique C, Müller MB, Reul JM, Deussing JM, et al. (2005) Corticotropin-releasing hormone activates ERK1/2 MAPK in specific brain areas. Proc Natl Acad Sci U S A. 102(17):6183-8.
29. Gong CX, Iqbal K (2008) Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for Alzheimer disease. Curr Med Chem. 15(23):2321-8.
30. Le MH, Weissmiller AM, Monte L, Lin PH, Hexom TC, et al. (2016) Functional Impact of Corticotropin-Releasing Factor Exposure on Tau Phosphorylation and Axon Transport. PLoS ONE 11(1): e0147250.
31. Kang JE, Cirrito JR, Dong H, Csernansky JG, Holtzman DM (2007) Acute stress increases interstitial fluid amyloid-beta via corticotropin-releasing factor and neuronal activity. Proc Natl Acad Sci U S A. 104(25):10673-8.
32. Mietelska-Porowska A, Wasik U, Goras M, Filipek A, Niewiadomska G (2014) Tau protein modifications and interactions: their role in function and dysfunction. Int J Mol Sci. 15(3):4671-713.

33. Lane RF, Dacks PA, Shineman DW, Fillit HM (2013) Diverse therapeutic targets and biomarkers for Alzheimer's disease and related dementias: report on the Alzheimer's Drug Discovery Foundation 2012 International Conference on Alzheimer's Drug Discovery. Alzheimers Res Ther. 5(1):5.
34. Nakazato A, Okubo T, Nozawa D, Yamaguchi M, Tamita T, et al. (2004) Pyrrolopyrimidine and pyrrolopyridine derivatives substituted with cyclic amino group. [WO/2004/058767]
35. Scherer S, Uddin M (2015) Method of determining disease causality of genome mutations. [WO/2015/054777]
36. Hsu SY, Hsueh AJW (2005) Stresscopins and their uses. [WO/2002/034934]
37. Rissman RA, Lee KF, Vale WW, Sawchenko PE (2012) Methods for treatment and prevention of tauopathies and amyloid beta amyloidosis by modulating CRF receptor signaling. [EP2522351 A1]
38. Mizoguchi K, Kunishita T, Chui DH, Tabira T (1992) Stress induces neuronal death in the hippocampus of castrated rats. Neurosci Lett. 138(1):157-60.
39. Rachal Pugh C, Fleshner M, Watkins LR, Maier SF, Rudy JW (2001) The immune system and memory consolidation: a role for the cytokine IL-1beta. Neurosci Biobehav Rev. 25(1):29-41.
40. McEwen BS (1999) Stress and hippocampal plasticity. Annu Rev Neurosci. 22:105-22.
41. McEwen BS, Sapolsky RM (1995) Stress and cognitive function. Curr Opin Neurobiol. 5(2):205-16.
42. Garcia R (2001) Stress, hippocampal plasticity, and spatial learning. Synapse. 40(3):180-3.

43. Joseph R (1999) The neurology of traumatic "dissociative" amnesia: commentary and literature review. Child Abuse Negl. 23(8):715-27.
44. Webster EL, Torpy DJ, Elenkov IJ, Chrousos GP (1998) Corticotropin-releasing hormone and inflammation. Ann N Y Acad Sci. 840:21-32.
45. Rissman RA, Staup MA, Lee AR, Justice NJ, Rice KC, et al. (2012) Corticotropin-releasing factor receptor-dependent effects of repeated stress on tau phosphorylation, solubility, and aggregation. Proc Natl Acad Sci U S A. 109(16):6277-82.
46. Webb A, Miller B, Bonasera S, Boxer A, Karydas A, et al. (2008) Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders. Arch Neurol. 65(11):1473-8.
47. Stoothoff WH, Johnson GV (2005) Tau phosphorylation: physiological and pathological consequences. Biochim Biophys Acta. 1739(2-3):280-97.
48. Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, et al. (2016) A novel Alzheimer disease locus located near the gene encoding tau protein. Mol Psychiatry. 21(1):108-17.
49. Lee KW, Kim JB, Seo JS, Kim TK, Im JY, et al. (2009) Behavioral stress accelerates plaque pathogenesis in the brain of Tg2576 mice via generation of metabolic oxidative stress. J Neurochem. 108(1):165-75.
50. Sternberg EM, Hill JM, Chrousos GP, Kamilaris T, Listwak SJ, et al. (1989) Inflammatory mediator-induced hypothalamic-pituitary-adrenal axis activation is defective in streptococcal cell wall arthritis-susceptible Lewis rats. Proc Natl Acad Sci USA 86: 2374–2378.
51. Chrousos GP (1995) The hypothalamic-pituitary-adrenal axis and immune-mediated inflammation. N Engl J Med. 332: 1351–1362.

52. Holsboer F (2003) Corticotropin-releasing hormone modulators and depression. Curr Opin Invest Drugs 4: 46–50.

53. Zobel AW, Nickel T, Künzel HE, Ackl N, Sonntag A, et al. (2000) Effects of the high-affinity corticotropin-releasing hormone receptor 1 antagonist R121919 in major depression: the first 20 patients treated. J Psychiatr Res. 34: 171–181.

54. Webster EL, Lewis DB, Torpy DJ, Zachman EK, Rice KC, et al. (1996) In vivo and in vitro characterization of antalarmin, a nonpeptide corticotropin-releasing hormone (CRH) receptor antagonist: suppression of pituitary ACTH release and peripheral inflammation. Endocrinology 137: 5450–5747.

55. Webster EL, Barrientos RM, Contoreggi C, Isaac MG, Ligier S, et al. (2002) Corticotropin releasing hormone (CRH) antagonist attenuates adjuvant induced arthritis: role of CRH in peripheral inflammation. J Rheumatol 29: 1252–1261.

56. Wlk M, Wang CC, Venihaki M, Liu J, Zhao D, et al. (2002) Corticotropin-releasing hormone antagonists possess anti-inflammatory effects in the mouse ileum. Gastroenterology 123: 505–515.

57. Dantzer R, Wollman E, Vitkovic L, Yirmiya R (1999) Cytokines and depression: fortuitous or causative association? Mol Psychiatry 4: 328–332.

58. Licinio J, Wong ML (1999) The role of inflammatory mediators in the biology of major depression: central nervous system cytokines modulate the biological substrate of depressive symptoms, regulate stress-responsive systems, and contribute to neurotoxicity and neuroprotection. Mol. Psychiatry 4: 317–327.

59. Leonard BE (2001) The immune system, depression and the action of antidepressants. Prog Neuropsychopharmacol Biol. Psychiatry 25: 767–780.

60. Connor TJ, Harkin A, Kelly JP, Leonard BE (2000) Olfactory bulbectomy provokes a suppression of interleukin-1beta and tumour necrosis factor-alpha production in response to an in vivo challenge with lipopolysaccharide: effect of chronic desipramine treatment. Neuroimmunomodulation 7: 27–35.

61. Castanon N, Leonard BE, Neveu PJ, Yirmiya R (2002) Effects of antidepressants on cytokine production and actions. Brain Behav. Immun. 16: 569–574.

62. Arsenault-Lapierre G, Whitehead V, Lupien S, Chertkow H (2012) Effects of anosognosia on perceived stress and cortisol levels in Alzheimer's disease. Int J Alzheimers Dis. 2012:209570.

63. Davis KL, Davis BM, Greenwald BS, Mohs RC, Mathé AA, et al. (1986) Cortisol and Alzheimer's disease, I: Basal studies. Am J Psychiatry. 143(3):300-5.

64. Masugi F, Ogihara T, Sakaguchi K, Otsuka A, Tsuchiya Y, et al. (1989) High plasma levels of cortisol in patients with senile dementia of the Alzheimer's type. Methods Find Exp Clin Pharmacol. 11(11):707-10.

65. Maeda K, Tanimoto K, Terada T, Shintani T, Kakigi T (1989) Elevated urinary free cortisol in patients with dementia. Neurobiol Aging. 12(2):161-3.

66. Giubilei F, Patacchioli FR, Antonini G, Sepe Monti M, Tisei P, et al. (2001) Altered circadian cortisol secretion in Alzheimer's disease: clinical and neuroradiological aspects. J Neurosci Res. 66(2):262-5.

67. Arsenault-Lapierre G, Chertkow H, Lupien S (2010) Seasonal effects on cortisol secretion in normal aging, mild cognitive impairment and Alzheimer's disease. Neurobiol Aging. 31(6):1051-4.

68. Lind K, Edman A, Nordlund A, Olsson T, Wallin A (2007) Increased saliva cortisol awakening response in patients with mild cognitive impairment. Dement Geriatr Cogn Disord. 24(5):389-95.

69. Van Eck M, Berkhof H, Nicolson N, Sulon J (1996) The effects of perceived stress,

traits, mood states, and stressful daily events on salivary cortisol. Psychosom Med. 58(5):447-58.

70. Murphy L, Denis R, Ward CP, Tartar JL (2010) Academic stress differentially influences perceived stress, salivary cortisol, and immunoglobulin-A in undergraduate students. Stress. 13(4):365-70.
71. Garner B, Phassouliotis C, Phillips LJ, Markulev C, Butselaar F, et al. (2011) Cortisol and dehydroepiandrosterone-sulphate levels correlate with symptom severity in first-episode psychosis. J Psychiatr Res. 45(2):249-55.
72. Lee ZS, Chan JC, Yeung VT, Chow CC, Lau MS, et al. (1999) Plasma insulin, growth hormone, cortisol, and central obesity among young Chinese type 2 diabetic patients. Diabetes Care. 22(9):1450-7.
73. Chiodini I, Di Lembo S, Morelli V, Epaminonda P, Coletti F, et al. (2006) Hypothalamic-pituitary-adrenal activity in type 2 diabetes mellitus: role of autonomic imbalance. Metabolism. 55(8):1135-40.
74. Bruehl H, Rueger M, Dziobek I, Sweat V, Tirsi A, et al. (2007) Hypothalamic-pituitary-adrenal axis dysregulation and memory impairments in type 2 diabetes. J Clin Endocrinol Metab. 92(7):2439-45.
75. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42(Database issue):D1001-6.
76. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 26(17):2190-1.
77. Ward LD, Kellis M (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40(Database issue):D930-4.
78. Yates A, Akanni W, Amode MR, Barrell D, Billis K, et al. (2016) Ensembl 2016. Nucleic Acids Res. 44(Database issue): D710–D716.

79. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, et al. (2016) The Ensembl Variant Effect Predictor. Genome Biol. 17: 122.

80. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, et al. (2016) The Ensembl gene annotation system. Database (Oxford) 2016: baw093.

81. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22(9):1790-7.

82. Xie D, Boyle AP, Wu L, Zhai J, Kawli T, et al. (2013) Dynamic trans-acting factor colocalization in human cells. Cell. 155(3):713-24.

83. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, et al. (2014) Comparative analysis of regulatory information and circuits across distant species. Nature. 512(7515):453-6.

84. Younesi E, Toldo L, Müller B, Friedrich CM, Novac N, et al. (2012) Mining biomarker information in biomedical literature. BMC Med Inform Decis Mak. 12: 148.

85. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 102(43):15545-50.

Summary

In this work, I have tried to identify shared mechanisms underlying neurodegenerative diseases by a “shared genetics” approach. This paradigm is not novel however, the route I have taken to map and integrate biomarkers by means of cause-and-effect modelling approaches, is really new. It holds great potential for the mechanistic interpretation of the biological impact of genetic variation.

I have systematically analysed data from GWAS experiments in various neurological conditions related to neurodegeneration. “Shared SNPs” identified “shared loci” and I came up with significant “genomic hotspots” enriched for genetic variation with relevance for two major neurodegenerative diseases, Alzheimer´s Disease (AD) and Parkinsonism (PD).

Evading the established route of performing “gene set enrichment” or any other “pathway association” method usually applied to signals in omics-data. I rather tried to reconstruct the putative pathophysiology processes associated with genes in “genomic hotspots” by integrating them in cause-and-effect models. The resulting systems pathophysiology models have great explanatory potential for the biological impact of the “shared genetics”. Moreover, they permit the integration of other evidences supporting the notion of a pathophysiology mechanism. The methodology I have developed has led us to a very interesting new “shared mechanism” that has relevance to both AD and PD. It also explains how environmental factors and stress can influence neurodegenerative disease risk and pathophysiology at a mechanistic level.

Chapter 4
Discussion and Future outlook

Discussion

A major challenge in the translation of GWAS evidences into disease mechanism is to reveal which gene or set of genes at or near disease-associated genetic loci is makes etiological contributions to disease. GWASs are expected to continue to bear fruit performing GWAS keep. Carefully designed meta-analyses of GWAS can detect novel small-effect genomic regions associated with disease, and fine-mapping approaches, as well as studies in ethnic subgroups, can refine existing ones. In parallel, there is a demanding need to translate genetic markers of complex traits and diseases into molecular mechanisms, through both global meta-analysis of multiple GWAS intervals and in-depth mechanistic studies of transcription, chromatin structure and DNA methylation at individual GWAS intervals. This functional translation is crucial for the identification of novel “druggable” components and pathogenic pathways. This in turn has the potential to empower clinical care through, for example, improved risk prediction, biomarker identification, disease sub-classification, drug development and dosing.

Network analysis can play an increasingly important role in prioritizing candidate causal variants for further experimental validation. Ultimately, the combination of computational and experimental approaches will yield mechanistic insights into the process by which a genetic variant, or a combination of variants, affect a complex phenotype. Given the complexity of neurodegenerative diseases and the limited

accessibility to experimental tissues of brain, we need new strategies to integrate data driven and knowledge driven approaches to reveal the mechanism behind these complex diseases. Disease networks based on systems biology models, comprising various interacting molecules and bioprocesses, were successful in integrating most of the available data. This methodology demonstrates how genetic and genome wide association data can be systematically analysed to gain knowledge and finally how knowledge can be quantified to guide decisions that improve success for target and biomarker discovery.

In the work presented here, I established an integrative approach that starts with a data-driven approach, identifies risk variants in GWAS data, and allows for better understanding of putative complex mechanisms through knowledge-driven context enrichment. This approach goes far beyond classical "pathway enrichment" approaches, as it takes multiscale and multimodal information into account and integrates heterogeneous information and knowledge in biologically meaningful, computable graph models. Data, a priori knowledge and inferred insights are combined in a seamless fashion. Meaningful cause-and-effect relationships are established and the signals originally identified are made interpretable in a rational modelling and mining approach.

Modelling the functional context of genes associated with genetic variants in computable cause-and-effect models can be very helpful to identify possible molecular level perturbation mechanisms that contribute to disease pathology. Mechanistic modelling allows us to be highly specific with respect to the available knowledge in a given context.

The proposed extension for genetic variance information into computable modelling

is a valuable contribution to the genetic research community. It can be further extended, and can be adopted for analysis using different algorithms. The developed protocol served as one of the keystones for the IMI funded European project; AETIONOMY [https://www.aetionomy.eu]. The construction and simulation of computable cause-and-effect models of disease pathology may therefore be able to predict the response to a drug or improve the design of clinical experiments.

The mechanistic hypothesis generated in the work, from 'tau locus BEL model' establishes a rationale, how stress could cause deficits in memory. It may actually have established a functional context that puts a "sensor" for environmental and life style into a pathophysiology mechanism that could play a significant role in the aetiology of Alzheimer's disease. Comprehensive genetic knowledge related to aetiology and treatment of neurodegenerative diseases is of paramount importance, as this field has always been enigmatic for scientists and clinicians. In that regard, this study shows how systems biology modelling can be used to link data from disparate domains, which can allow crosswalking through them and hence could be used to map diagnostic and treatment options.

However, validation of the established mechanisms or hypotheses in wet labs still remains to be done. In the future, it is also needed to connect the mechanistic model with available clinical data to do more advanced analysis and ensure the completeness of the model. More research is needed in these areas to bridge the gap between the molecular and genomic level data to generate the mechanistic knowledge. Lastly, there is an immense need to continue to facilitate automatic updating and enrichment of the knowledge model as new insights are gathered in the research field.

Future outlook

The genetic research will benefit from this effort, as the approach developed in the course of this work has led to the discovery of novel knowledge in the genetic and biomedical domain and can be further applied to topics beyond the scope of this work. In the future, mechanistic modelling could be further used for extraction of common pathways and mechanisms shared by different diseases existing in the knowledge space. This protocol can be adapted to other scientific domains apart from genetics. It can also be used to systematically identify and rank new hypotheses for different diseases.

The study confers enhanced interpretation power for screening and prioritization of the most suitable genetic candidates with the aim of advancing genetic and drug discovery and development efforts. Hence, the predicted genes and genetic candidates enriched with supporting evidences may guide future validation efforts in experimental clinical research and molecular biology laboratories.

In view of the rapid pace of genome studies, this thesis proposes a strategy that facilitates mining and prioritization of interesting candidates in a way that could shift research new and rewarding directions. Taken together, the work presented here demonstrates the development of new knowledge discovery techniques that enable the collection, curation, annotation, interpretation and discovery of a broad spectrum of knowledge needed for efficient and systematic biomedicine research.

In the future, by using such longitudinal data, integrated networks could be generated and the inter-connectivity of such networks can divulge both known and novel candidate biomarkers. Moreover, analysis of such longitudinal data can identify

network perturbations associated with common diseases. Recently a wellness study (*Price ND, et al. 2017* [1]) has been conducted by collecting dynamic, dense, and personal data for 108 persons over 9 months, including clinical tests, whole genome sequence, proteomes, metabolomes, and microbiomes at three different time points.

However, the design of such diagnostics to reveal early disease transitions, and other interventions to reverse the disease process, is still at its very earliest state. Nevertheless, such dynamic and diversified personal data integration will take place at a crossroads for the emerging arena of scientific wellness, which may initiate the preventative, predictive, personalized and participatory (P4) medicine (like Precision Medicine) of the 21st century.

1. Price ND, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol. 2017;35(8):747-756.