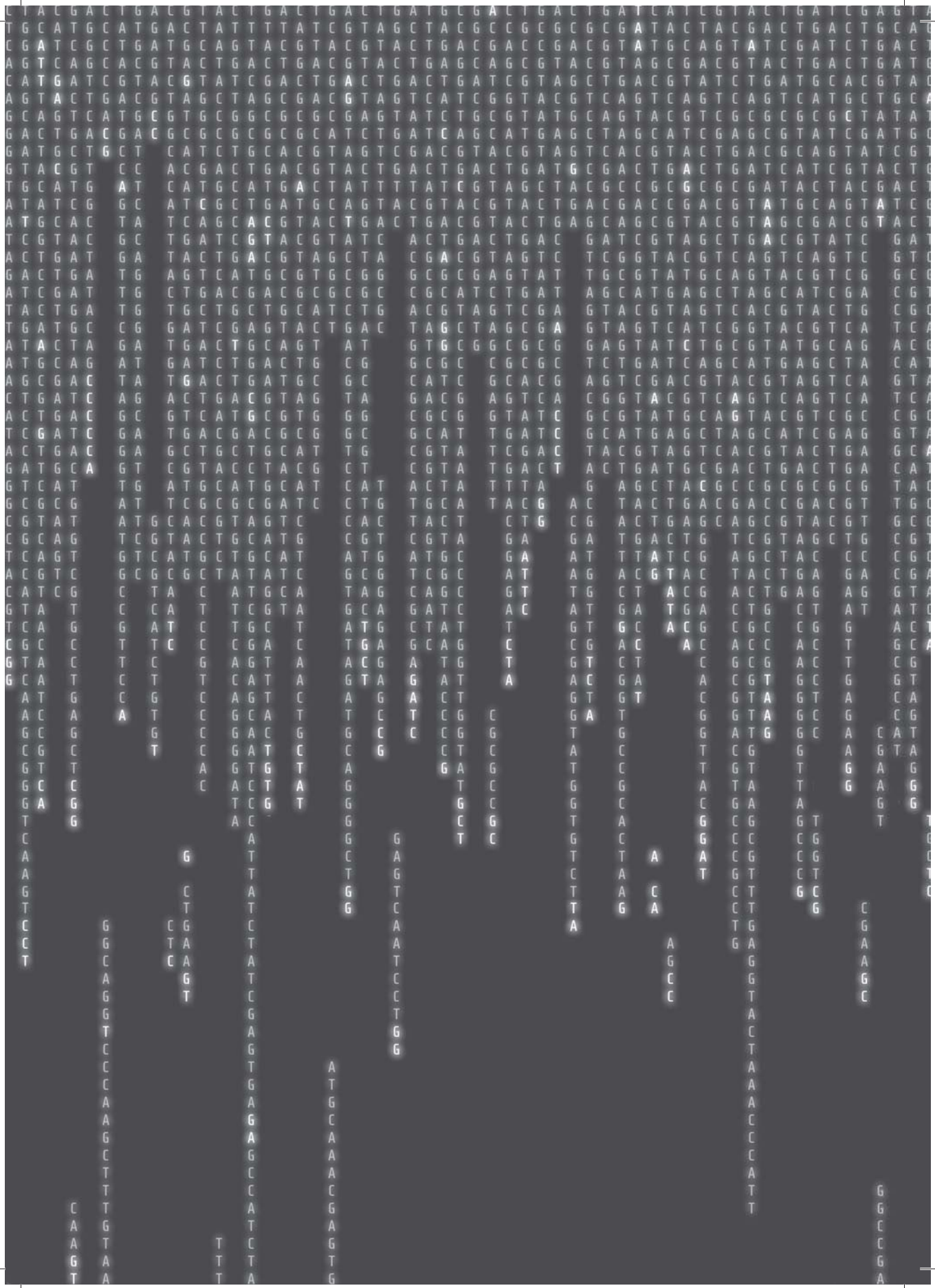
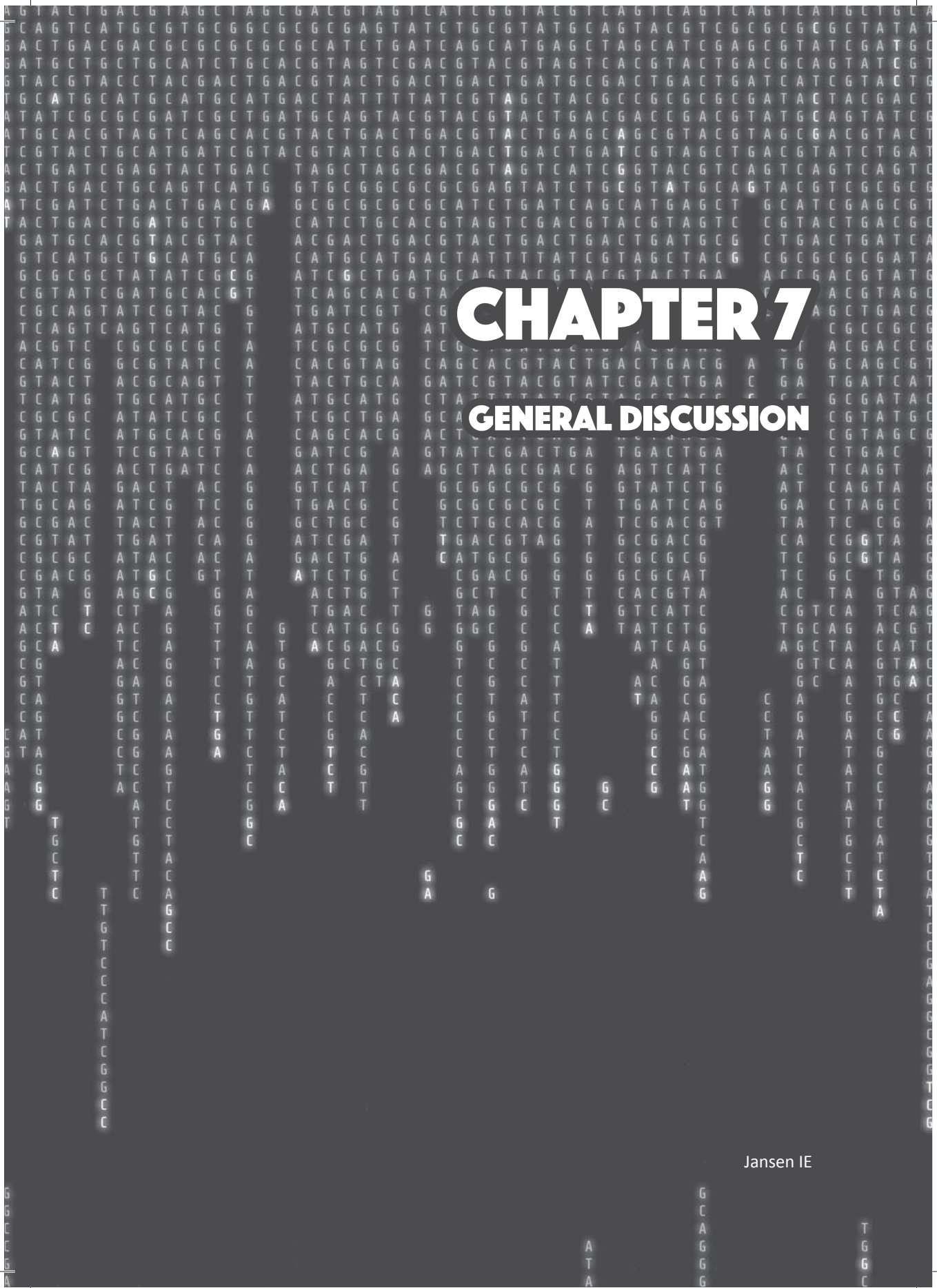
Cracking the code-ing sequence for Parkinson’s disease

Jansen, I.E.

2017

document version
Publisher's PDF, also known as Version of record





CHAPTER 7

GENERAL DISCUSSION

Jansen IE

Parkinson's disease (PD) is a neurodegenerative disease that still awaits successful treatment slowing, altering or reversing the disease progression. The current available therapies only diminish the clinical symptoms and are not modifying the underlying disease pathology. Research studying the fundamental properties of the disease contributes to our understanding of the biological processes causing PD. Once the PD field comprehends the origin and progression of the disease on a detailed pathological and molecular biological level, it will enable the development of improved therapies as targets for drugs are more accurately defined. Genetics plays a major role in the onset of PD.[1] Dozens of common risk loci with moderate effect and rare variants with a high penetrance have been linked to PD. However, still only up to 50% of the genetic variation can be explained by these PD-related genetic factors.[2,3] The identification of novel genetic variants, and a more detailed understanding of PD-related genes, will help to improve our knowledge of PD pathogenesis. Research has shown that PD genes converge on multiple biological processes and novel genes will potentially point to additional dysfunctional molecular pathways.[4] As PD is a heterogeneous disease on a clinical, pathological and genetic level, it is anticipated that therapies, apart from personal medicine for individuals with monogenic PD, targeting common underlying biological pathways will probably be beneficial to treat multiple groups of PD patients with distinct underlying genetic causes.

To improve our understanding of the genetic factors contributing to PD, we used whole exome sequencing (WES), a next-generation sequencing (NGS) technique focusing on the coding regions of the genome. With the introduction of NGS in the first decade of the 21st century, it became a standard approach to focus on rare variants, anticipating the discovery of novel genetic variants that would elucidate the remaining unexplained heritability. Where GWAS had previously dominated the genetics research field with the identification of common risk factors, the NGS technology allowed the investigation of rare variants, thereby causing a shift towards an interest in the genetic influence of rare variants on diseases. The scientific work performed for this thesis was started at a time point when a large number of common risk loci for PD were established. Therefore, by using WES, and thereby exploring the rarer variants, it was expected that novel PD-related genetic factors would be revealed.

SUMMARY

Within this thesis we have established the contribution of known PD genes in more detail and searched for novel ones. To further explore the known genes, two IPDGC exome datasets were accessed for validation of the novel Mendelian PD gene *CHCHD2* and for determining the effect of rare genetic variants within the published PD risk loci. For *CHCHD2* (**chapter 2**), a gene involved in mitochondrial function which was first observed in a PD cohort from Japan,[5] no association signal of common variants was detected in our

European population. One identical common variant to a significant SNP in the original Asian study was not associated in our NeuroX dataset, suggesting Asian-specificity for that particular variant. Although we did not observe an association of common variants, 3 rare nonsynonymous variants (MAF ≤ 0.0008) were observed in 4 cases while not a single rare variant was observed in healthy controls. These results suggest that *CHCHD2* could also be a rare risk factor in the European population, but more extensive sequencing research involving larger datasets should be performed to enable any definitive conclusions.

Multiple genome-wide association studies have linked dozens of common loci to the risk of developing PD.[6-9] By using our exome datasets, we tested whether rare coding variants within these known risk loci might also influence disease risk (**chapter 3**). For the analysis of single rare variants large datasets with tens of thousands of individuals are required to reach sufficient statistical power. To increase this power we therefore aggregated rare variants per gene and gene-set to test for their joint effect of rare variants on PD. As GWAS risk loci indicate regions in which variants are in linkage disequilibrium (LD) that frequently encompass multiple genes of which the causal gene is undefined, we used a strategy (PrixFixe) that selected the most probable causal gene based on underlying functional similarities. We confirmed the association of the well-established *LRRK2* concerning both common and rare variant associations. Furthermore, *STBD1* and *SPATA19* show an association to PD in the WES and NeuroX datasets, respectively. The gene-set aggregation approach detected, as anticipated, a moderate association of common variants. However, there was no effect of aggregated rare variants when testing the gene-set. It is complicated to conclude whether this is due to an incorrect selection of causal genes by the PrixFixe approach or whether rare variants in PD risk loci are genuinely not contributing to PD risk. This chapter clearly shows the need for improved functional prioritization tools, besides PrixFixe, that are able to adequately call the most probably gene(s) while taking into account the biological features of the disease of interest.

Our search for novel PD genes was guided by focusing on groups of genes that share a functional property related to PD. *GBA* is a gene with a lysosomal function that has been related to autosomal recessive Gaucher disease, a rare lysosomal storage disorder (LSD), and has been shown to strongly increase PD risk. We therefore studied the involvement of all known LSD-related genes in PD susceptibility (**chapter 4**). By using two independent WES datasets and a third exome array dataset, we tested for the joint effect of rare variants in LSD genes with an emphasize on the effect of the total gene-set to increase statistical power to detect an association signal as the WES datasets had relatively low sample sizes. We identified a significant joint association of rare deleterious (predicted by CADD) variants within the total set of 54 LSD genes in the original IPDGC dataset, which we were able to replicate in the additional independent PPMI WES and IPDGC NeuroX datasets. The association signal remained in all datasets, when excluding *GBA.* Zooming in on the association of individual LSD genes, we observed significant

associations for *GBA*, which confirms the strength of our study design. Our results furthermore suggest that multiple variants in various LSD genes within a single individual, the so called oligogenic effect, might increase the risk for PD. Finally, *GBA, SMPD1, CTSD* and *SLC17A5* show independent significant effects on PD, of which the first 2 genes have already been discussed in the context of PD risk in literature. Overall, we conclude that additional non-*GBA* variants in LSD genes contribute to PD susceptibility where the effects might be consistent with a model of oligogenic risk factors. Additional sequencing in larger sample sizes and functional follow-up studies should further investigate the exact role of these LSD genes in PD pathogenesis.

A second approach based gene-sets of interest on biological processes that are affected in PD pathobiology (**chapter 5**), either based on literature or on gene-expression research.[4,10] The latter involved a genome-wide transcriptomics study that identified specific biological pathways based on differences in gene-expression levels at distinct levels of Braak stage α-synuclein pathology by using post-mortem brain samples. By specifically investigating the affected molecular processes of Braak stages 1 and 2, the pre-symptomatic stage, we aimed to determine whether these observed gene expression changes could be caused by genetic defects. Three biological pathways (*mitochondrial dysfuntion, caveolar-mediated endocytosis signaling* and *renin-angiotensin signaling*) showed an aggregated association signal in all three independent WES datasets. The literature-based pathway *mitochondrial dysfunction* showed significant association to PD in the IPDGC WES discovery dataset which was replicated in the smaller PPMI and Merck WES datasets. Although replication was not observed for individual genes within this pathway, multiple genes belonging to the NADH:ubiquinone oxidoreductase family, also called mitochondrial complex I subunits, were enriched for pathogenic variants in the distinct datasets. As mitochondrial complex I has an important function in ATP synthesis and a major contributor to the production of reactive oxygen species, these results imply that this specific molecular process might be affected in PD pathogenesis. Functional follow-up could specifically target these cellular events and potentially link them to PD-related phenotypes. The transcriptomics-based pathway *renin-angiotensin signaling* showed enrichment for common coding variants in PD cases compared to controls. This data suggests that the detected expression changes for genes belonging to these pathways at the beginning of PD pathology progression, are caused by genetic variation rather than by other intrinsic biological changes or environmental influences.

A last strategy to identify novel genetic risk factors for PD was applied in **chapter 6.** All coding regions of the genome were explored for high-impact loss of function variants showing a higher frequency in PD cases compared to healthy controls. Only homozygous and compound heterozygous variants were considered as the case sample set has a young average age of onset which has often been linked to an autosomal recessive disease inheritance pattern.[11] We reported 27 genes encompassing LoF variants with a higher

frequency in our PD cohort, of which *GPATCH2L, UHRF1BP1L, PTPRH, ARSB* and *VPS13C* genes show convincing PD-related phenotypes in functional experiments and genetic confirmation in independent PD datasets. The identification of the PD gene *VPS13C* emphasizes the strength of our study design. We strongly believe that these 5 genes are promising PD genes and encourage the PD research field to perform additional variant screening and functional validation studies to confirm their pathogenic contribution to PD.

DISCUSSION AND FUTURE PERSPECTIVES

Rare variant aggregation analyses

The main objective of this thesis was to determine the contribution of rare coding variants to PD. As the term implies, rare variants are infrequent in the general population and large sample sizes are required to reach sufficient statistical power for standard single-variant association tests. Besides the options to study rare variants in families or increase sample sizes to reach statistical power, a third approach aggregates rare variants in preselected units (e.g. genes or biological pathways), followed by comparison of the aggregated frequency statistics between cases and controls.[12] This strategy has been of great value for the identification of novel rare genetic factors in this thesis. The identification of associated rare variants in genes involved in lysosomal storage (chapter 4) and mitochondrial dysfunction (chapter 5) pathways, even if we do not consider established PD risk genes, was completely depending on the rare variant aggregation analysis. Our analyses demonstrated a difference between cases and controls in the enrichment of rare variants for genes belonging to these biological pathways. It emphasizes the need to focus future genetic and functional studies on the genes within these molecular processes.

Variant aggregation tests are accompanied by several challenges including the limited sample sizes of current sequencing studies (a general topic discussed in *Sample size* section), the selection of deleterious variants, interpretation of aggregated association signals, and confounding factors. The selection of deleterious variants is important as benign variants that do not contribute to PD risk will dilute potential association signals. A first strategy that was applied to cope with this problem was to exclude all coding synonymous variants or solely include variants that are predicted to have damaging effects by the bioinformatics algorithm CADD[13] in chapter 3 and 4. As an additional approach, we assigned a functional weight to every variant in chapter 5, which can be accomplished with the variant aggregation tool SKAT.[14] There is no gold standard for selecting either of these two strategies. The first approach has the advantage of applying weights to variants based on MAFs, thereby assigning more weight to rare variants. This is impossible for the second approach as it already assigns weights based on functionality. However, the strength of the second approach is that all variants below

a specified MAF threshold are considered, while the first approach forces the removal of variants that are not predicted to be deleterious, thereby potentially excluding true causal variants. We applied both approaches in this thesis, adding the second strategy in the last variant aggregation association chapter once it was introduced by SKAT. Significant rare variant associations of gene-sets were validated with a second independent genetic dataset for chapter 4 and 5, indicating that both pathogenic variant selection approaches could be applied successfully. As the genetic etiology of a disease is a priori unknown it is preferable to use both functional variant selection strategies.

Finding a significant gene or gene-set association is encouraging. However, apart from indicating that the gene(-set) is potentially involved in PD pathogenesis, the signal is not informative on the exact route through which the association occurs. The results generated by variant aggregation tools, such as SKAT[14], lack any reports on specific variants, therefore rare causal variants are not pinpointed. Furthermore, SKAT tests for a two-sided model and therefore has the disadvantage that the directionality of the aggregated effect is undefined, meaning it could be either protective or damaging. However, defining the causal variants is essential for improved knowledge on underlying molecular processes and follow-up functional experiments. Therefore, methods that identify individual causal variants among the aggregated rare variant association are required. Two such methods have been proposed, of which the first one chooses causal rare variants based on backwards elimination (BE).[15] Variants are defined as non-causal if the overall gene association becomes stronger after exclusion of these variants, which is performed on a one by one basis. The second method, called adaptive combination of *P*-values method (ADA)[16], performs a similar rare variant aggregation test like SKAT but is built upon single-marker *p*-values, therefore allowing to simultaneously perform an aggregated variant test and pinpointing the causal variants. This latter strategy eliminates individual variants with higher *p*-values during the association tests, which could be more powerful in comparison to SKAT that includes all variants regardless the individual variant significance levels.

As the implementation of the second method on our data was not feasible within the time-frame of this thesis and we were not convinced by the first method, we manually scanned the variants of interest for their individual *p*-values and directionality of effect (MAF cases vs. controls). Besides the fact that the sample size of the datasets might not be sufficient for gene-based association tests, as shown for the power calculations in chapter 4, we report in chapter 3 individual variant interpretations to indicate the difficulties to move from a gene based association towards causal variant detection. We rather inspect the variants manually than using the BE method for different reasons: with BE, singletons are prone to be eliminated due to their small contributions to the aggregated effect. The authors of the method propose to use functional predictions to put more weight on these extremely rare variants. However, these algorithms might incorrectly predict the variants to be benign, which would lead to exclusion of a causal variant. Furthermore, BE

allows the inclusion of only one covariate. Bearing in mind the heterogeneous state of our WES data (which will be discussed in the next paragraph), this would seem a simplistic model resulting in potential prioritization of associations that are driven by confounding factors. However, the two discussed methods that aim to surpass the identification of an overall gene-based association by additionally establishing the true causal variants, are valuable initiatives to deal with the challenge to interpret the joint association results. Building further on these concepts will likely improve the current rare variant aggregation bioinformatics tools.

A last concern regarding rare variant association studies encompasses the influence of confounding factors. GWAS, testing the association of common variants, are prone to population differences and the gold standard is to correct for potential population stratification by the inclusion of principle components.[17] Rare variant association studies are even more sensitive to confounding factors. More subtle systematic ancestry differences could result in spurious outcomes and differences in library capture kits are a concern. Rare variants tend to be more population specific than common variation.[18,19] One convincing study showed elevated rates of spurious outcomes for 9 rare variant association methods as a result of fine-scale population structure.[20] A small country with a highly condensed population, such as the Netherlands, already shows differences in genomic background, which are clearly visualized plotting the first 3 principle components, thereby mimicking the geographic location of the subpopulations.[21] The second putative confounding factor is technical differences. The IPDGC WES dataset has been generated over a time-frame of 4 years starting in 2010, a time period where the WES library capture kits have evolved at a rapid pace. As a result, multiple capture kits were used, which adds an additional confounding factor to the model. Although all capture kits aim to cover the coding regions of the genome, differences in targeted regions exist which results in capture kit specific sequencing errors. To deal with these confounding factors we applied a few strategies. Firstly, we only selected the variants that were located in regions covered by all capture kits. Furthermore, 20 multi-dimensional scaling components (MDS) were calculated to include in the variants aggregation analysis as covariates. Plotting the first 3 MDS components visualized differences in capture kit and population, implying the MDS components to capture the most important variation caused by these two factors. However, MDS components are based on common variants (MAF > 1%) and it is unclear whether they have the ability to correct for structural rare variant differences as it is unknown whether the underlying population structure of common variants is correlated to the population background of rare variants.[22] Therefore, when exploring our significant association results, we consistently inspect variants of interest on individual genotype level to determine that signals originate from diverse samples instead of individuals with a specific European population or sequenced with a particular library exome capture kit.

Finally, association signals were only considered genuine after replication in at least one additional independent genetic dataset.

In sum, with the use of rare variant aggregation analysis this thesis detected novel association signals for the mitochondrial- and lysosomal-related pathways. These genetic discoveries were exclusively depending on rare variant aggregation studies, stressing the importance of testing the joint effect of rare variants rather than testing for single variant associations. The discussed variant aggregation challenges emphasize how crucial it is to design studies adequately, especially when it concerns sequencing projects. As it is frequently inevitable to merge different cohorts due to the requirement of large sample sizes for rare variant association studies, we acknowledge that subpopulations often have to be mixed. When large merged cohorts will be sequenced for a single project it is important to properly assign samples to different batches, generating subgroups that are matched for affection status and population.

Sample size

The majority of the analyses of this thesis were performed on the IPDGC WES and NeuroX datasets, currently the largest genetic datasets enriched for coding variants in the PD research field. I was therefore in the privileged position to investigate the genetic background of PD with novel genetic models at genomic positions that were previously unexplored. My analyses have resulted in the confirmation and identification of relevant genes and biological mechanisms. For future studies, the usage of larger sample sizes will presumably enable the discovery of additional important genetic factors. Previous research has estimated minimal sample sizes of ten thousands of cases[23,24] to be required for sufficient power to detect genome-wide and gene-based rare variant association signals. As there is a huge gap between our included sample sizes (1,167 and 6,801 PD cases for IPDGC and NeuroX datasets, respectively) and the, by literature proposed, ideal sample size, it is currently not feasible to overcome the potential power issue by increasing the number of samples with in-house additional sequencing experiments, especially considering the relatively high costs of sequencing studies.

The consensus for genetic studies is to replicate the discovery findings in at least one other independent dataset with distinct individuals.[25] For GWAS this has proven to be fruitful by setting a genome-wide significance level for the large discovery dataset, while maintaining a lower significance threshold (up to a nominal *p*-value of 0.05) for a second smaller replication dataset. The reason for this success can be explained by the common nature of the variants tested with GWAS, allowing replication in controllable sizes of sample sets. Replication of individual rare variants is more complicated. An example reflecting the difficulties around rare variant replication is *CHCHD2* which we tried to replicate in our Caucasian IPDGC WES dataset in chapter 2. In the original paper, both familial and sporadic PD patients of Japanese descent presented rare *CHCHD2* variants.[5]

These variants were not present within the IPDGC WES dataset. We did find 3 other rare variants in *CHCHD2,* all present in cases and not a single coding variant in controls, implying that this might be a rare genetic risk factor for the European population too. Exploring other populations (Asian vs. European) complicates replication as rare variants are often population specific. Chapter 2 shows that with only 3 identified variants it is impossible to replicate *CHCHD2* by rare variant aggregation test as the variant frequencies are too low to reach significance. Overall, replication of rare variants requires large datasets which is often inapplicable due to insufficient numbers of patients or financial resources.

Besides the replication of a gene that was discovered by another research group in chapter 2, we experienced comparable replication obstacles for our in-house explorative analyses. As we were lacking a second independent WES dataset similar to the size of the original one (~ 1,200 cases and ~1,700 controls), we attempted to replicate our findings with a variety of other genetic exome datasets. In chapter 3, where we established the role of rare variants in common PD risk loci, we used the NeuroX dataset as a second dataset. As the NeuroX array contains a custom-array, it was enriched for variants in PD GWAS loci. However, no genetic replication was observed. For chapter 4, focusing on the effect of rare variants in LSD genes, we were more successful. By using a small second independent WES dataset of the public available PPMI project (approximately 1/5th in size),[26] we were able to replicate the joint effect of rare deleterious variants when considering all LSD genes simultaneously. And even though the NeuroX array was not specifically enriched for variants with LSD genes, we also replicated an aggregated rare variant signal for the total gene-set in NeuroX. NeuroX has the advantage to consist of a relative large sample size (almost 7,000 cases and 6,000 controls), therefore boosting power. When establishing the aggregated effect of coding variants in transcriptomics-based deregulated pathways in chapter 5, we aimed to replicate our findings with a larger WES dataset as we in the meantime obtained access to ~1,000 additional PD exomes through a novel collaboration. By meta-analyzing this novel dataset with the PPMI dataset we were able to generate a larger replication dataset. Thus, we managed to deal with the sample size related replication issues for all our analyses.

Chapter 6 proposes a study design, to overcome the problem arising from small replication datasets, that implements functional screens. Instead of aspiring significant genetic replication, we aimed to detect highly promising PD genes by linking PD-related phenotypes in multiple system models to our initial list of 27 candidate genes. Using mitochondrial dysfunction and α-synuclein toxicity assays, 5 of these candidate genes are prioritized for future studies. Combining suggestive genetic replication with functional work could be a favorable solution to pinpoint potential causal genes. Similar functional screening approaches might furthermore be beneficial for identification of true causal genes linked to the PD GWAS loci.

The current replication challenges of rare variant association signals, emphasizes the need for researchers to collaborate. When joining forces (funding resources and subject cohorts) it possibly increases the number of cases to ten thousands of individuals. The IPDGC is an excellent start but even larger initiatives are required. Ideally, all generated sequencing data would be publically shared on repositories, such as the database of Genotypes and Phenotypes (dbGaP) or the European Genome-phenome Archive (EGA).[27,28] It seems the genetic world is slowly moving towards this situation as many scientific journals are promoting data sharing by demanding the public release of the analyzed data in a repository.

Gene-set analysis

Chapter 3, 4 and 5 performed variant aggregation association tests in predefined gene-sets that were based on biological pathways or other disease-related properties. Similarly to the issue observed in significant gene-based association results, where it is impossible to distinguish genuine causal variants from the benign ones, it is for significant gene-set associations complicated to assess which specific genes are generating the signal. A second challenge, which is specific to the gene-set association analysis, is the selection of correct sets of genes that are associated to PD. As we observed significant and replicable gene-set findings for the LSD gene-set in chapter 4, and the mitochondrial dysfunction and caveolar-mediated endocytosis gene-sets in chapter 5, we assume to have chosen a relatively close representation of truly associated pathways. However, the gene-set based on the previously published PD GWAS risk loci of chapter 3 is more complex. There is no significant joint effect of rare variants for this gene-set, but it is still undecided whether rare variants within common PD risk loci are truly not associated to PD or that the gene-set is not including the true causal genes underneath the common loci.

Chapter 3 shows just how essential correct definition of a gene-set is to be able to draw definitive conclusions. Subsequent to the extensive analyses of this chapter, we are currently still unable to argue in favor or against a genuine influence of rare variants in common PD risk loci. As discussed in the chapter 3, basing a gene-set on common risk loci is challenging as often multiple genes are located within risk loci, all equally likely candidates to be the causal ones. Although we have attempted to establish the true causal genes based on underlying functional similarities between the different GWAS peaks, there is a chance that non-causal genes were selected.

Besides the fact that this shows the importance of adequate gene-set selections, it furthermore emphasizes the need for improved methods to functionally interpret GWAS results. The PrixFixe tool[29] that was used for this purpose in chapter 3 is an improvement over the traditional selection of genes based on their closest physical position to the highest associated SNP. However, the assumption that causal genes from different loci are likely to be functionally connected, thereby prioritizing genes that have similar

functions, might result in false negatives if many distinct molecular processes are affected in a disease etiology. Furthermore, PrixFixe has a disease-unbiased approach meaning it will not consider disease-related features. Yet it might be helpful to build underlying functional networks on for example transcriptomics datasets of body tissues that are affected in the disease pathogenesis. Lastly, one should consider the possibility that an associated locus (represented by a linkage disequilibrium (LD) block) harbors the causal variant in a regulatory element (e.g. an enhancer) which influences the expression of a gene beyond the borders of the LD block. eQTL analysis is able to detect such signals.[30] The development of a bioinformatics tool that reduces the number of candidate genes per locus by integration of these aspects would be an instructive guide for functional follow-up experiments aiming to validate the pathogenic effects of the selected genes on the disease of interest.

A final point for improvement for gene-set association analysis includes the development of gene-set variant aggregation tools that are specific for rare variants. Multiple gene-set software packages exist that are designed to test for gene-set association by using common variants generated with genotype arrays.[31] A commonly used gene-set tool is MAGMA that first calculates gene-based association signals, which are used to determine the gene-set association.[32] This differs from the rare variant aggregation analysis, for instance tested with SKAT, which measures for overall gene-set association by aggregation of the variants in all genes without an in-between step generating per gene associations. However, calculating the gene association is preferred as it allows equal contribution of every gene in the subsequent gene-set analysis. However, in the case of SKAT a large gene could contribute more to the test statistic since in general they encompass more variants. In practice this could mean that a true association signal of a small gene is missed in the presence of large genes that are non-causal. Another advantage of MAGMA is that it takes into account linkage-disequilibrium structures. Although rare variants are in less strong LD than common variants, not accounting for LD could generate false positives. In the current analyses of this thesis, first a MAF threshold to distinguish common and rare variants is set, which is followed by LD-pruning of the common variants. Establishing this MAF cutoff is relatively arbitrary, meaning that variants defined as rare could still be rather common and in LD with other variants. The genetic research field focusing on rare variant aggregation analyses would therefore benefit greatly from a software tool such as MAGMA but then specifically designed to analyze rare variants.

Functional interpretation: from genes to therapy

The identification of causal rare coding variants could reveal the corresponding affected genes, which serves as a guide to establish deregulated biological processes in PD. Such genetic discoveries are essential for our knowledge on PD pathogenesis which is the basis for the development of improved treatment. The identification of the variants and genes

within the research covered by this thesis puts emphasis on the molecular processes involved in mitochondrial health (chapter 2, 5 and 6), lysosomal storage (chapter 4 and 6), caveolar-mediated endocytosis (chapter 5) and renin-angiotensin signaling (chapter 5). Targeting these deregulated pathways with drugs might become a fruitful approach to treat PD. Especially considering the heterogeneous etiology of the disease, which suggests that a focus on converging deregulated pathways could be beneficial for a larger group of PD patients.

The strategy to view diseases as health conditions where molecular processes are deregulated, thereby affecting multiple levels of functional units, is a theory that is gaining popularity.[33] This system-based approach relies for a large part on genetic studies, showing the importance to perform WES research such as ours. However, cellular processes are dynamic, temporary and interactive, thereby making them complex. They constitute multiple levels of functional networks which are interconnected. It is therefore preferable to study genetics while putting it in perspective of gene expression, protein synthesis or energy consumption. The system-based approach therefore encompasses multiple levels of biology systems, represented by the different levels of omics data, which should be viewed as intertwined subjects rather than addressing them as distinct topics. An exceptional study on genetic factors of hereditary spastic paraplegias (HSP), a motor neuron disease, serves as a good example on omics integration leading to more in-depth genetic findings.[34] Based on publically available protein interaction data they generated a proteome network of proteins and genes that interact with known HSP genes. This 'HSPome' was their guide to investigate a WES dataset consisting of 55 families with autosomal recessive HSP and resulted in the identification of 18 novel genetic factors. Therefore, the combination of different omics levels is a fruitful approach, including the integration of in-house omics dataset as well as public datasets, such as ENCODE[35] and FANTOM[36].

In chapter 5 we have connected transcriptomics with genomics data by testing the genetic involvement of biological pathways that were previously determined through gene-expression profiling research. Similarly, in chapter 6 we used publically available resources and UKBEC analyses covering gene (co-)expression outcomes to functionally interpret our genetic results. These two studies utilize different omics data levels in a complementary fashion. Evidence originating from different functional levels increases the likelihood of identifying a genuine finding. Another preferred approach to integrate multiple omics data is called meta-dimensional analysis,[33] where raw data, transformed data or the underlying models are combined in a simultaneous analysis. Although convergence of the different types of data generates a more complete biological picture of the underlying etiology of a disease, it comes with many challenges. Software tools such as ATHENA,[37] Glmpath[38] and Weka3[39] use different approaches to integrate multiple omics levels, while aiming to adequately take into account analytical issues like missing

data patterns, background noise, sample sizes and data normalization that differ between the distinct data types.

Of special interest is the generation of multiple data levels for the same individuals. For our studies, distinct PD cohorts were used for the different omics projects. However, an ultimate design would encompass genomics, epigenomics, transcriptomics, proteomics and metabolomics data from the same individuals. It is anticipated that such approaches are more sensitive to detect disease-related signals as it allows for observation of direct effects of affected functional units because the different omics data types can be linked on individual level. Although system biology approaches matching samples on individual level is the preferred design, the omics research field still has to overcome many obstacles to perform it properly. Often it is difficult to obtain sufficient tissue to perform multiple omics experiments. And although integration of distinct omics data types for the same individual would be possible with bioinformatics tools such as ATHENA, the data integration methods have only just started to develop and currently no general consensus exists. One of the major issues to address is to choose the appropriate underlying model to test for the disease of interest. As system-based approach studies are relatively novel, we lack adequate published examples for PD therefore predicting the model is uncertain. Combining the distinct data integration tools will presumably supply the most comprehensive analysis method guiding us to an improved understanding of PD pathogenesis by visualizing the affected complex biology system.

Post WES era: future of PD genetics

The discovery of rare variants with non-conventional WES studies in cohorts using unrelated subjects, besides the traditional linkage analysis within families, has been successful in a few neurologic disorders.[40,41] However, in general the genetic research field had higher expectations from WES at the beginning of the NGS era, as it was anticipated that many novel causal variants would be easily revealed with simple analysis methods. For PD, conservative calculations predict that only 10% of the heritability is explained by known genetic causes.[42] This suggests that the focus of the PD genetics research field should be switched to a different type of genetic defect or we simply are lacking the means to generate datasets of sufficient sample size. So how to uncover the remaining 90%?

The major point of improvement that will increase the success rate for rare variant discovery is the required growth in sample size to ten thousands of individuals. Furthermore, many points to improve WES analysis have been considered in this discussion, such as rare variant aggregation perspectives and the putative essential step of integration with large-scale functional work. The most important aspect that requires emphasize is the system-biology approach. With the development of more comprehensive omics integration tools and increase in publically available omics dataset, it will be increasingly

achievable to combine WES datasets with other PD or brain-related datasets that will lead to novel genetic insights.

Whole genome sequencing (WGS) is gradually becoming the gold standard to study variants in a genome-wide fashion. The WGS technology field is evolving by developing methods to sequence longer reads, decreasing sequencing errors and generating protocols for single cells.[43,44] Of special interest is single-cell sequencing as it allows to study mosaicism, the existence of two or more genetically distinct cell populations within a single organism. In 2013, a study on mosaicism in neurons revealed that at least 13% of neuronal cells have de novo CNVs.[45] In perspective of PD, sequencing single-cells from the substantia nigra would potentially reveal de novo mutations, putatively explaining the vulnerability for neuronal cell loss of this specific region. Not specific to single-cell WGS but applicable to WGS in general, is the search for structural and non-coding variation. Structural variation might affect large regions of the genome, potentially affecting the function of multiple genes.[46,47] WES is not ideal to be utilized for calling structural variants, since large parts of the genome are missing, and the non-coding regions are not covered. As structural variants have been shown to cause both Mendelian as complex diseases,[48] and PD has been associated to multiple CNVs in *SNCA, parkin, PINK1* and *DJ-1*,[49] it is plausible that more undiscovered structural variation causes PD. The implementation of third-generation WGS techniques, such as long read sequencing that facilitates unique alignment of the reads,[50,51] would enhance the identification of structural variants and the establishment of repetitive regions. Finally, non-coding regions are mostly unexplored. Considering that approximately 1% of the genome consists of coding sequences,[52] the regions that are targeted with WES, it suggests there is still 99% of the genome that has been studied poorly in relation to PD. As a recent study revealed the functional connection between a non-coding variant in an *SNCA* enhancer and PD,[53] it is anticipated that comparable unidentified non-coding causal variants will be detected by WGS.

Although WGS is replacing GWAS, the latter remains a good study design to establish the contribution of common loci to PD etiology. Although the PD field has not established the functional route through which the previously identified risk loci are contributing to PD, it is important to continue detecting novel risk loci, as identification of more genes tends to fill the gaps of knowledge on the functional network underlying PD pathogenesis. Furthermore, initiatives such as the Haplotype Reference Consortium (HRC) are improving the imputation process substantially.[54] HRC allows the imputation of variants with MAFs of minimal 0.1%, thereby increasing the number of SNPs that can be tested.

Two different GWAS strategies will determine novel PD risk loci. One obvious solution is to increase sample size. As the costs for genotyping are dropping to ~$30- per sample, it becomes feasible to aim for hundred thousands of individuals. Financial expenses will not be the limited factor, but the collection of large PD cohorts is. The

IPDGC is the largest PD genetics consortium and their sources are close to fully exhausted with a total of ~25,000 cases. It is questionable if extending collaborations will increase sample sizes substantially to reach sufficient power for detection of novel rare genetic factors. One solution might be to collect genetic PD data across ethnically different populations. A second strategy encompasses the re-examination of the existing GWAS data by focusing on semi-significant hits in combination with functional annotation data, performing a system-based approach as discussed previously. GenoSkyline is a genetic tool that integrates GWAS data with tissue-specific functional predictions.[55] The functional information is based on epigenomic data and application of the tool could potentially reveal novel PD risk loci.

Besides the identification of novel PD risk loci, it is essential to continue to explore the known PD risk loci to understand how these associated LD blocks are associated to PD. Resequencing of all LD blocks would be a valuable strategy to pinpoint the causal variant and gene. Successful studies have followed this strategy and identified causal variants for inflammatory bowel disease and schizophrenia.[56-58] For PD, such a study is currently ongoing and will putatively explain true causal variants for a few loci. An alternative approach to comprehend the underlying molecular mechanisms of the PD risk loci, is to functionally follow-up all genes underneath the GWAS peaks. Although such a project would require dedication due to its elaborate amount of work, it is feasible with the recent developments in high-throughput functional screening experiments. In this regard, an inspiring study has suggested *RAB29* and *GAK* as causal genes of 2 PD risk loci by performing a large functional screening effort searching for interactors of *LRRK2*.[59] Combining their interactors with PD GWAS data exposed the 2 genes. However, one should be aware that such functional approaches could still miss the PD genes that are located outside the associated LD block of which expression is regulated by a causal variant within the associated locus.

Automated high-throughput functional screening efforts have the potential to uncover genetic and biological insights which lies beyond the limits of pure genetic studies. The functional screening research on HSP,[34] the abovementioned *LRRK2* interactor study,[59] but also our own study in chapter 6, show the ingenuity of these approaches to overcome the obstacles we experience when interpreting genome-wide results, by using broad-scale functional experiments. Furthermore, automating differentiation protocols of induced pluripotent stem cells and CRISPR technology,[60,61] mimicking the disease state (e.g. dopaminergic neuronal cells for PD) and the exact mutation of interest, will be promising methods to elucidate the causal mechanism positioned between the mutation and phenotype.

In sum, PD genetics in the near future will be characterized by the usage of various technologies (GWAS, WES and WGS) on large datasets, which will be followed by domination of WGS in the more distant future as costs will continue to decrease.

Approaching PD pathogenesis as a complex system of multiple functional levels, by integrating different omics data, is likely to disclose novel genetic factors. Functional screening methods are crucial for identification and validation of causal variants and genes. Linking the variant to a biological mechanism is a critical step that must be taken to enable the development of treatment that slows or alter the progression of PD pathogenesis.

REFERENCES

1. Bras J, Guerreiro R, Hardy J. SnapShot: Genetics of Parkinson's disease. *Cell* 2015; 160(3): 570-.e1.
2. Keller MF, Saad M, Bras J, et al. Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Human molecular genetics* 2012; 21(22): 4996-5009.
3. Hamza TH, Payami H. The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors. *Journal of human genetics* 2010; 55(4): 241-3.
4. Trinh J, Farrer M. Advances in the genetics of Parkinson disease. *Nature reviews Neurology* 2013; 9(8): 445-54.
5. Funayama M, Ohe K, Amo T, et al. CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease: a genome-wide linkage and sequencing study. *Lancet neurology* 2015.
6. Simon-Sanchez J, Schulte C, Bras JM, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature genetics* 2009; 41(12): 1308-12.
7. International Parkinson's Disease Genomics Consortium, Wellcome Trust Case Control Consortium2. A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS genetics* 2011; 7(6): e1002142.
8. Nalls MA, Plagnol V, Hernandez DG, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 2011; 377(9766): 641-9.
9. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature genetics* 2014; 46(9): 989-93.
10. Dijkstra AA, Ingrassia A, de Menezes RX, et al. Evidence for Immune Response, Axonal Dysfunction and Reduced Endocytosis in the Substantia Nigra in Early Stage Parkinson's Disease. *PloS one* 2015; 10(6): e0128651.
11. Klein C, Westenberger A. Genetics of Parkinson's disease. *Cold Spring Harbor perspectives in medicine* 2012; 2(1): a008888.
12. Kiezun A, Garimella K, Do R, et al. Exome sequencing and the genetic basis of complex traits. *Nature genetics* 2012; 44(6): 623-30.
13. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM. A general framework for estimating the relative pathogenicity of human genetic variants. 2014; 46(3): 310-5.
14. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* 2011; 89(1): 82-93.
15. Ionita-Laza I, Capanu M, De Rubeis S, McCallum K, Buxbaum JD. Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. *PLoS genetics* 2014; 10(12): e1004729.
16. Lin WY. Beyond Rare-Variant Association Testing: Pinpointing Rare Causal Variants in Case-Control Sequencing Study. *Scientific reports* 2016; 6: 21824.
17. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Human molecular genetics* 2008; 17(R2): R143-50.
18. Nelson MR, Wegmann D, Ehm MG, et al.

An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, NY)* 2012; 337(6090): 100-4.
19. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, NY)* 2012; 337(6090): 64-9.
20. O'Connor TD, Kiezun A, Bamshad M, et al. Fine-scale patterns of population stratification confound rare variant association tests. *PloS one* 2013; 8(7): e65834.
21. Boomsma DI, Wijmenga C, Slagboom EP, et al. The Genome of the Netherlands: design, and project goals. *European journal of human genetics : EJHG* 2014; 22(2): 221-7.
22. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics* 2012; 44(3): 243-6.
23. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* 2014; 111(4): E455-64.
24. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics* 2014; 95(1): 5-23.
25. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS computational biology* 2012; 8(12).
26. Parkinson Progression Marker Initiative The Parkinson Progression Marker Initiative (PPMI). *Progress in neurobiology* 2011; 95(4): 629-35.
27. Tryka KA, Hao L, Sturcke A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic acids research* 2014; 42(Database issue): D975-9.
28. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nature genetics* 2015; 47(7): 692-5.
29. Tasan M, Musso G, Hao T, Vidal M, MacRae CA, Roth FP. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature methods* 2015; 12(2): 154-9.
30. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nature reviews Genetics* 2015; 16(4): 197-212.
31. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. *Nature reviews Genetics* 2016; 17(6): 353-64.
32. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology* 2015; 11(4): e1004219.
33. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews Genetics* 2015; 16(2): 85-97.
34. Novarino G, Fenstermaker AG, Zaki MS, et al. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science (New York, NY)* 2014; 343(6170): 506-11.
35. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489(7414): 57-74.
36. Lizio M, Harshbarger J, Shimoji H, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* 2015; 16: 22.
37. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. ATHENA: the analysis tool for heritable and

environmental network associations. *Bioinformatics (Oxford, England)* 2014; 30(5): 698-705.

38. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PloS one* 2011; 6(11): e24709.
39. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. *Cell* 2010; 143(6): 1005-17.
40. Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science (New York, NY)* 2015; 347(6229): 1436-41.
41. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 2014; 506(7487): 185-90.
42. Singleton A, Hardy J. The Evolution of Genetics: Alzheimer's and Parkinson's Diseases. *Neuron* 2016; 90(6): 1154-63.
43. Kuleshov V, Xie D, Chen R, et al. Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology* 2014; 32(3): 261-6.
44. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nature reviews Genetics* 2016; 17(3): 175-88.
45. McConnell MJ, Lindberg MR, Brennand KJ, et al. Mosaic copy number variation in human neurons. *Science (New York, NY)* 2013; 342(6158): 632-7.
46. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nature reviews Genetics* 2006; 7(2): 85-97.
47. Guan P, Sung WK. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods (San Diego, Calif)* 2016; 102: 36-49.
48. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature reviews Genetics* 2013; 14(2): 125-38.
49. Toft M, Ross OA. Copy number variation in Parkinson's disease. *Genome Medicine* 2010; 2(9): 62.
50. Wang Y, Yang Q, Wang Z. The evolution of nanopore sequencing. *Frontiers in genetics* 2014; 5: 449.
51. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics* 2015; 13(5): 278-89.
52. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009; 461(7261): 272-6.
53. Soldner F, Stelzer Y, Shivalila CS, et al. Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature* 2016; 533(7601): 95-9.
54. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* 2016; 48(10): 1279-83.
55. Lu Q, Powles RL, Wang Q, He BJ, Zhao H. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS genetics* 2016; 12(4): e1005947.
56. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics* 2011; 43(11): 1066-73.
57. Beaudoin M, Goyette P, Boucher G, et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis.

PLoS genetics 2013; 9(9): e1003723.
58. Gonzalez-Penas J, Amigo J, Santome L, et al. Targeted resequencing of regulatory regions at schizophrenia risk loci: Role of rare functional variants at chromatin repressive states. *Schizophrenia research* 2016; 174(1-3): 10-6.
59. Beilina A, Rudenko IN, Kaganovich A, et al. Unbiased screen for interactors of leucine-rich repeat kinase 2 supports a common pathway for sporadic and familial Parkinson disease. *Proceedings of the National Academy of Sciences of the United States of America* 2014; 111(7): 2626-31.
60. Yamanaka S. Induced pluripotent stem cells: past, present, and future. *Cell stem cell* 2012; 10(6): 678-84.
61. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 2014; 157(6): 1262-78.