Rare variant analysis in human neurodegenerative disorders guided by cellular models of proteotoxicity

Dina Zielinski

Sorbonne Université

École Doctorale Complexité du Vivant (ED515)

RARE VARIANT ANALYSIS IN HUMAN NEURODEGENERATIVE DISORDERS GUIDED BY CELLULAR MODELS OF PROTEOTOXICITY

Dina Zielinski

Présentée et soutenue publiquement le 9 juin 2021

Devant un jury composé de :

Stéphane Le Crom, PhD	Représentant ED515
Roselyne Friedenberg, PhD	Représentant Formation Continue
Emmanuelle Génin, PhD	Rapporteuse
Tuuli Lappalainen, PhD	Rapporteuse
Stéphanie Debette, MD/PhD	Examinatrice
Vikram Khurana, MD/PhD	Accompagnateur
Sophie Garnier, PhD	Accompagnatrice

ACKNOWLEDGMENTS

I would like to thank all members of my thesis committee, for generously agreeing to share both their time and expertise.

My mentor, Yaniv Erlich, for his brilliant creativity and enthusiastic leadership. His advice and support throughout my career and limitless confidence in my abilities pushed me beyond anything I thought I was capable of.

Team Erlich - Melissa Gymrek, Assaf Gordon, Thomas Willems, Joanna Kaplanis, Mona Sheikh, Barak Markus - who inspired me, taught me, and made me laugh, in equal amounts.

Sue Lindquist for having confidence in my ability to take on a significant role in this challenging project, for her pioneering spirit and for having contributed to making the Whitehead Institute a supportive and inspiring environment. It has been a privilege to continue the work that she started.

Vik Khurana for keeping this project going while starting his own lab and for supporting me in presenting this work as part of my PhD.

All Khurana lab members and collaborators at the Brigham and Harvard: Isabel Lam, Sumaiya Nazeen, Xinyuan Wang, Shamil Sunyaev and Jian Peng, without whose expertise this project would not have been possible.

Antony Cooper and all involved in the MGRB at the Garvan Institute for all their effort and for being a model of how genomic data sharing can and should be.

My colleagues at the New York Genome Center for their collaborative spirit and for creating a refuge in a city that never stops. Stephane, Pejman, Margot, Tomaz, and Kaja for adventures (and misadventures) throughout NYC and beyond.

My Curie girls (Laia, Camille, Shahad, and Mercè) for their unwavering support and much-needed humor.

Alena, Dónal (and Baby G) for convincing me to complete a PhD and for their advice over champagne and homecooked meals.

To all my friends and 'families' without whose love and support I would not have been able to accomplish my goals and for putting up with me when I spent way too much time in the lab.

I dedicate this work to all women scientists who persisted in pushing the boundaries of scientific knowledge in the face of gender discrimination, for breaking barriers to make science accessible and inclusive.

CONTENTS

I. PARCOURS PROFESSIONNEL

DINA ZIELINSKI | MOLECULAR & COMPUTATIONAL BIOLOGIST

genomics (rare/neurologic diseases, population genetics, cancer, developmental biology)

RESEARCH ACTIVITIES:

SENIOR SCIENTIST, INSERM, PARIS, FRANCE
SEPTEMBER 2019 - PRESENT

The Paris Transplant Group (Inserm U970) is a large multidisciplinary team of researchers and clinicians (paristransplantgroup.org).

- Contribute to the application of R&D advances for the development of a digital health platform in transplant medicine
- Implemented and evaluated synthetic data generation methods to permit sharing of sensitive data with R&D partners
- Lead the development of a tool to extract biological and histological data from PDFs and images
- Writing and proofing of scientific reports and technical translations (FDA, internal reports, grants)
- Provide expertise in bioinformatic data analysis methodology, descriptive/inferential statistics, and oversee collaborative research projects with R&D partners
- Application of machine learning methods to classify rejection in organ transplant recipients
- Project Lead: establish and oversee a consortium of researchers and clinicians in Canada, Europe, the US, and UK and manage collaborative research: evaluation and coordination of strategic partnerships, DTAs, SOPs, implement and drive scientific and organizational roadmap
- Integration of multi-modal patient data and application of machine learning methods to improve rejection diagnosis in organ transplant recipients
- Direct and contribute to the development of a web app (icdot.org) for sharing clinical, histological, and sequencing data to validate the application of tissue based molecular diagnosis in transplant patient care
- Participate in scientific communication through publications and presentations

RESEARCH AFFILIATE: HARVARD MEDICAL SCHOOL, BOSTON, MA, USA
2017 - PRESENT

Research affiliate of the Khurana lab in the Neurology Department at Brigham and Women's Hospital and Harvard Medical School (https://khuranalab.bwh.harvard.edu). The Khurana lab focuses on understanding the consequences of perturbed proteostasis in neurodegenerative disease and applies stem cell-based approaches to develop patient-specific therapies.

- Responsible for sample curation of DNA from alpha-synucleinopathy patients
- Performed targeted sequencing of human homologs of genes found to modify alpha-synuclein toxicity in a yeast model in more than 500 familial and sporadic synucleinopathy patients
- Variant calling, quality control, variant filtering and statistical analysis and integration with control data from several thousand healthy, aged individuals to uncover putative disease-causing mutations for validation in stem cell models
- Presented this work in a talk at Advances in Genome Biology and Technology (AGBT), considered the preeminent genome science and technology conference

BIOINFORMATICS SCIENTIST: INSTITUT CURIE, PARIS, FRANCE
MAY 2017 - SEPT 2019
Joint position in the bioinformatics platform (U900) and the Heard and Margueron labs in the Genetics and Developmental Biology Unit (U934/UMR3215).

- Co-lead author on a publication establishing a role in gene activation for BAP1, a chromatin modifier implicated in cancer, showing that the BAP1 complex promotes gene expression by opposing PRC1-mediated H2A ubiquitylation largely independently of antagonism with PRC2
- Performed all bioinformatic and statistical analyses of sequencing data from mammalian cell lines to assess the effects of transient disruption of PRC2 and the stable de-repression of a large subset of its target genes, demonstrating that negative feedback between Polycomb group proteins and transcription creates a heritable record of gene-autonomous switching of expression states
- Integration of allele-specific high-throughput expression, DNA-protein interaction, chromatin accessibility, and chromosome conformation sequencing data to understand the role of X chromosome inactivation and other chromatin modifications in cancer and rare disorders
- Developed a bioinformatic pipeline to integrate allele-specific expression and chromatin conformation sequencing data (RNA, ATAC, HiC, damID) to understand the interplay of nuclear localization and gene repression in stem cells (ESCs) and differentiated cells (NPCs)

SENIOR SCIENTIST: NEW YORK GENOME CENTER, NEW YORK, NY
JAN 2015 - APR 2017
Researcher in bioinformatics and molecular and cellular biology in a human genetics and genomics lab in the Computer Science Department at Columbia University and the New York Genome Center.

- Developed an automated gating protocol for flow cytometry analysis of a reporter assay to characterize the role of short tandem repeats (STRs) on gene expression
- Project lead: oversaw sequencing library preparation and analyzed SNP array and exome sequencing data from rare disease patients
- Contributed to the development of tandem repeat detection in whole genome sequencing, including sequencing data analysis and experimental validation of repeat alleles
- Created a catalog of short tandem repeats for use in genetic and forensic studies
- Analyzed whole genomes for a public lecture and for Carl Zimmer's *Game of Genomes*, including summary statistics, local and global ancestry, and disease- and trait-associated SNPs
- Collaborated in the design and launch of a website to collect direct-to-consumer genetic data for use in biomedical research: dna.land

TECHNICAL ASSOCIATE: WHITEHEAD INSTITUTE/MIT, CAMBRIDGE, MA
JAN 2011 - DEC 2014
Research Associate in a genomics lab at the Whitehead Institute for Biomedical Research, an international leader in the fields of genetics and genomics that played a key role in the sequencing of the human genome.

- Oversaw all projects involving rare variant discovery, including sample handling, sequencing, and bioinformatics analysis
- Discovered a causal mutation for a rare craniofacial disorder through exome sequencing and microarray and performed molecular validation of the copy number variant
- Optimized a combinatorial pooling and gDNA hybrid selection and library preparation protocol
- Developed a web-based tool for sample handling (iPipet)
- Prepared sequencing libraries and performed variant calling and filtering to generate a catalog of risk alleles in the Ashkenazi Jewish population
- Accessioned and prepared several hundred Alzheimer's/Parkinson's patient samples for targeted sequencing to uncover putative risk/protective alleles

RESEARCH ASSOCIATE: GENETIC SERVICES INC, CAMBRIDGE, MA
JUN 2009 - DEC 2010

- Transgenic troubleshooting, contacting and assisting researchers, database management
- QC and prepare all DNA for injection; mutant screening projects, Drosophila embryo transgenesis

RESEARCH ASSISTANT: NYU SKIRBALL INSTITUTE, NEW YORK, NY
SEPT 2005 - MAY 2008

- Assisted with an RNAi screening project involving germline patterning in C. elegans

RESEARCH SKILLS AND EXPERTISE:

COMPUTATIONAL SKILLS:

Bash, R, Python, LaTeX, version control (Git/SVN), HTML/CSS, SQL, pipeline framework (Cromwell/WDL, Snakemake), containerization (Docker)

Data analysis: DNA-seq (exome/targeted/genome), RNA-seq, ChIP-seq, ATAC-seq, DNA/RNA microarray, *de novo* mutation discovery, summary statistics, variant filtering, extensive experience with high-throughput sequencing command-line tools (QC, alignment, variant calling, functional annotation), genome assembly, sequence homology detection, descriptive and inferential statistics, correlation and regression, ANOVA, dimension reduction (PCA/MDS/t-SNE), predictive modeling, machine learning classification, network analysis (molecular interaction), genotype imputation, variant phasing, ancestry analysis, GSEA, meta-analysis, linkage and genetic association, synthetic data generation

LAB SKILLS:

cell culture (bacterial/mammalian), flow cytometry, molecular cloning, bacterial transformation, mammalian transfection and transduction (lentivirus), genome editing (CRISPR), NGS sequencing library prep, single cell RNA-seq, hybrid capture, array genotyping (prep and analysis), Sanger sequencing, capillary electrophoresis, nucleic acid extraction/purification, fluorescent nucleic acid quantitation, FISH, PCR/qPCR/RT-PCR, extensive experience with Tecan-based liquid automation system (Tecan EVOware), sequencing library protocol/pipeline design, *Drosophila* embryo transgenesis

PUBLICATIONS:

Zielinski D*, Lam I*, Nazeen S, Wang X, Bras JT, Myers RH, Scherzer CR, Trojanowski JQ, Van Deerlin VM, Cooper AA, Lindquist SL, Erlich Y, Sunyaev SR, Peng J* and Khurana V. (2022) Pleiotropic effects of beta-amyloid and alpha-synuclein genetic networks in synucleinopathies. *In Preparation*

Holoch D*, Wassef M*, Lovkist C, **Zielinski D**, Aflaki S, Howard M, Margueron R. (2021) A PRC2-based double-negative feedback underlies transcriptional memory. *Nature Genetics* 53:1686–1697.

Marion-Poll L, Foret B, **Zielinski D**, Attia M, Carter AC, Syx L, Chang HY, Gendrel AV, Heard E. (2021) Locus specific epigenetic modalities of random allelic expression imbalance. *Nature Communications* 12:5330.

Aubert O*, Yoo D*, **Zielinski D***, Cozzi E* *et al.* (2021) COVID-19 pandemic and worldwide organ transplantation: a population-based study. *Lancet Public Health* S2468-2667(21)00200-0.

Ragazzini R*, Pérez-Palacios R*, Baymazt I, **Zielinski D**, Michaud A, Givelet M, Borsos M, Jansen P, Torres-Padilla ME, Bourc'his D, Fouchet P, Vermeulen M, and Margueron R. (2019) GPIF constrains Polycomb Repressive Complex 2 activity in germ cells. *Nature Communications* 10:3858.

Wassef M, Luscan A, Aflaki S, **Zielinski D**, Battistella A, Kersouani C, Servant N, Vermeulen M, Wallace MR, Vidaud M, Pasmant E and Margueron R. (2019) EZH1 and 2 function within canonical PRC2 and exhibit proliferation-dependent redundancy to shape mutational signatures in cancer. *PNAS* 116(13):6075-6080.

Zielinski D*, Campagne A*, Lee MK*, Michaud A, Le Corre S, Dingli F, Chen H, Vassilev I, Servant N, Loew D, Pasmant E, Postel-Vinay S, Wassef M, Margueron R. (2018) BAP1 complex promotes transcription by opposing PRC1-mediated H2A ubiquitylation. *Nature Communications* 10:348.

Yuan J, Gordon A, Speyer D, Aufrichtig R, **Zielinski D**, Pickrell J, Erlich Y. (2018) DNA.Land is a framework to collect genomes and phenomes in the era of abundant genetic information. *Nature Genetics* 50(2):160-165.

Willems T, **Zielinski D**, Gordon A, Gymrek M, Erlich Y. (2017) Genome-wide profiling of de novo STR variations. *Nature Methods* 14(6):590-592.

Erlich Y and **Zielinski D**. (2017) DNA Fountain enables a robust and efficient storage architecture. *Science* 355(6328):950-954.

Nida H, Blum S, **Zielinski D**, Srivastava D, Elbaum R, Xin Z, Erlich Y, Fridman E, Shental N. (2016) Highly efficient de novo mutant identification in a Sorghum bicolor TILLING population using the ComSeq approach. *The Plant Journal* 86(4):349-59.

Zielinski D, Gilbert A, Ngo H, Rudra A, Thierry-Mieg N, Wootters M, Erlich Y and Zuk O. (2015) Biological screens from linear codes: theory and tools. *bioRxiv*

Zielinski D, Gordon A, Zaks B, Erlich Y. (2014) iPipet: sample handling using a tablet. *Nature Methods* 11(8):784-5.

Zielinski D, Markus B, Sheikh M, Gymrek M, Chu C, et al. (2014) OTX2 Duplication Is Implicated in Hemifacial Microsomia. *PLoS ONE* 9(5): e96788.

Zielinski D, Gymrek M, Erlich Y. (2012) Back to the Family: A Renewed Approach to Rare Variant Studies. Genome Med. 4(12):97.

PRESENTATIONS:

European Society of Transplantation (2021). Milan, Italy. **Talk:** *Precision diagnostics: world-wide integration of transcriptomics data.*

Morals & Machines (2021). Dresden, Germany. **Talk:** *Emerging Infrastructures of the Future.*

Kavli Frontiers of Science Symposium (2020). Seattle, WA, USA. **Talk:** *DNA Data Storage.*

Data Science Congress (2020). Mumbai, India. **Talk:** *A future for digital data in DNA.*

Woodstock Bio (2020). Tel Aviv University, Tel Aviv, Israel. **Talk:** *Hard lessons from rare variant analysis in neurological disease.*

Possible Futures (2019) Casa Firjan, Rio, Brazil. **Talk:** *A future for digital data in DNA.*

Web à Québec (2019) Québec, CA. **Talk:** *A future for digital data in nature's oldest storage device.* https://webaquebec.org/ [highest rated talk of the conference]

Health, Education, and Research Talks (2019) Cluj-Napoca, Romania. **Talk:** *Bringing data to life: from the bench to the command line.* https://heartcluj.com

DNA for Digital Storage (2019) Banbury Center at CSHL, Huntington, NY, USA. **Talk:** *Communicating Advances in DNA Storage Technology.*

Advances in Genome Biology and Technology (2019) Marco Island, FL, USA. **Talk:** *Rare variant analysis of human neurodegenerative disorders guided by high-throughput yeast screens.*

TEDxVienna (2017) Vienna, Austria. **Talk:** *All the world's data in DNA.* [published on TED.com]

American Society of Human Genetics (2016) Vancouver, BC, CAN. **Poster**: *Rare variant analysis of human neurodegenerative disorders guided by high-throughput yeast screens for alpha-synuclein toxicity.*

DNA.Land User Group Meeting (2016) New York Genome Center, New York, NY, USA. **Talk:** *How global data sharing helped uncover the cause of a rare genetic disorder.*

Jewish Genomics Evening (2016) New York Genome Center, New York, NY, USA. **Analysis (for Carl Zimmer):** https://zimmerome.gersteinlab.org/

RootsTech Conference (2016) Salt Lake City, UT, USA. **Talk**: *DNA.Land: Rule Your Genome, Democratize Health.*

American Society of Human Genetics (2015) Baltimore, MD, USA. **Poster**: *Characterizing the Role of STRs in Gene Regulation.*

American Society of Human Genetics (2014) San Diego, CA, USA. **Poster**: *Harnessing Genome Engineering to Characterize the Role of STRs in Gene Regulation.*

American Society of Human Genetics (2013) Boston, MA, USA. **Poster**: *A Genetic Snapshot of Risk Alleles in the Ashkenazi Jewish Population.*

Genetic Research and Discovery in Jewish Populations (2013) Albert Einstein College of Medicine, Bronx, NY, USA. **Poster**: *A Genetic Snapshot of Risk Alleles in the Ashkenazi Jewish Population.*

American Society of Human Genetics (2012) San Francisco, CA, USA. **Poster**: *DNA Sudoku: Mining Rare Genetic Variations Using Combinatorial Pooling.*

Group Testing Designs, Algorithms, and Applications to Biology (2012) IMA - University of Minnesota, MN, USA. **Talk**: *Technical challenges of automating group testing designs in the lab.*

Personal Genomes (2011) CSHL, Huntington, NY, USA. **Poster**: *DNA Sudoku: Mining Rare Genetic Variations Using Combinatorial Pooling.*

EDUCATION:

Université de Paris | Paris, France, 2018
MS | Bioinformatics (highest honors)

Brandeis University, Rabb School | Waltham, MA, Sept 2015-Apr 2017
Bioinformatics Master's program
Coursework: Mathematical Modeling, Statistical Genetics, R for Biomedical Informatics, Biological Sequence Analysis, Molecular Profiling and Biomarker Discovery

Harvard University, Extension School | Cambridge, MA, 2012-2014
Graduate coursework: Biology of Neurodegenerative Diseases, Genomics, Biotech: Tools for Business

New York University, College of Arts and Science | New York, NY, 2008
BA | Biology; French

HONORS/AWARDS:

- NAS Kavli Frontiers of Science Fellow 2020
- AAUW Career Development Grant 2016

OTHER ACTIVITIES:

- Global STEM Alliance Mentor, NY Academy of Sciences (2016-present)
- NYU Alumni mentor to undergraduates (2016-present)
- New York Genome Center mentor to new hires (2015-2016)
- MIT Triathlon; club officer (2011-2014)
- Whitehead Institute Partner to local high school science teachers (Sept 2011-Dec 2014)
- Volunteer, Leukemia and Lymphoma Society's Team in Training (Jan 2010-June 2010)
- Triathlon (5/207 F30-34, NYC Triathlon 2016); top 15% age group average all races

LINKS:

sites.google.com/view/dinazielinski

linkedin.com/in/dinazielinski

@dinazielinski

gist.github.com/dinovski & github.com/dinovski/

2. RAPPORT D'ACTIVITES DE RECHERCHE

Les missions diverses et les développements réalisés au sein des laboratoires multidisciplinaires, détaillées ci-dessous de façon chronologique, montrent ma capacité à mobiliser mes compétences et à appliquer des méthodologies appropriées pour répondre à des questions pertinentes dans plusieurs domaines.

2011 - 2014 : WHITEHEAD INSTITUTE FOR BIOMEDICAL RESEARCH/MIT

J'ai commencé ma carrière en tant que biologiste moléculaire à l'Institut Whitehead/MIT, dans une équipe axée sur la génomique. Dans ce poste, j'ai fait des manipulations en biologie moléculaire et cellulaire et analyses en génomique pour établir les liens entre les stratégies moléculaires et informatiques en génétique humaine.

Mission : trouver la mutation causale d'une maladie rare

Résumé

La microsomie hémifaciale (HFM) est la deuxième anomalie faciale la plus courante après la fente labio-palatine. Le phénotype est très variable et la plupart des cas sont sporadiques. Nous avons étudié cette maladie sur un large échantillon de cinq individus. Nos résultats suggèrent un rôle de la sensibilité au dosage de l'OTX2 dans le développement craniofacial humain et soulèvent la possibilité d'une étiologie commune entre un sous-type de microsomie hémifaciale et le médulloblastome.

Résultats

La découverte d'une duplication du gène OTX2 dans le syndrome de Goldenhar (Zielinski et al. 2014)

Méthodologie

- J'ai recruté dans l'étude la plus grande famille de l'époque dont 5 individus présentant les symptômes d'une maladie craniofaciale (le syndrome de Goldenhar ou dysplasie oculo-auriculo-vertébrale)
- Préparation des libraires pour faire le séquençage ciblé (*whole exome*)
- Exclusion des mutations synonymes (qui ne changent pas l'acide aminé) et celles d'une fréquence supérieure à 0,1 % dans des bases de données publiques des projets de séquençage à grande échelle
- J'ai regardé les mutations dans les régions IBD (segments d'ADN en commun entre tous les membres de la famille)
- J'ai effectué le génotypage du génome entier utilisant une puce de génotypage afin de détecter des mutations structurelles
- Analyse des données génotypes et identification d'une variation structurelle (une duplication de 8 gènes) dans tous les individus qui ne se trouvaient pas dans les bases de données des échantillons de contrôle
- Validation de la duplication (CNV) par qPCR

- Afin de prédire le gène étiologique de la duplication, j'ai donné la priorité à la similarité des signatures moléculaires avec les gènes étiologiques connus d'autres malformations faciales. Deux algorithmes ont indépendamment classé OTX2 comme le gène ayant la signature moléculaire la plus proche.

Mission : la mise en place d'un protocole de *combinatorial pooling*

Résumé

Les cribles biologiques à grande échelle nécessitent l'analyse d'un nombre important d'échantillons à la recherche d'un résultat particulièrement rare. Les modèles combinatoires, dans lesquels les groupes plutôt que chaque spécimen sont examinés, permettent l'identification des quelques spécimens à l'origine du résultat rare sans qu'il soit nécessaire de tester chacun d'entre eux individuellement. Dans cette approche, les spécimens sont d'abord regroupés dans un petit nombre de « pools » en utilisant des règles mathématiques (codes linéaires Reed-Solomon). Les conceptions précédentes ont été développées par des mathématiciens qui ont principalement mis l'accent sur l'élégance algébrique plutôt que sur les contraintes expérimentales, ce qui limite l'applicabilité de cette approche dans le laboratoire.

Résultats

- Développement d'un outil web pour adapter une tablette électronique qui facilite le transfert des échantillons manuellement, appelé *iPipet* qui a également été un sujet de publication (Zielinski et al. 2014)
- Développement d'un protocole pour effectuer des études des mutations rares (bioRxiv 2015)
- Présentation de la méthode lors d'une conférence à l'*Institute for Mathematics and its Applications* (IMA) en 2012 (www.ima.umn.edu/2011-2012/W2.13-17.12/11763)
- L'application de ce protocole pour effectuer des expériences en collaboration avec l'équipe de Noam Shental de l'Open University of Israel (Nida et al. 2016)

Méthodologie

- Application de la théorie du codage pour améliorer l'efficacité des criblages à grande échelle tout en abordant les aspects pratiques l'applicabilité des modèles combinatoires
- Optimisation d'un protocole de *combinatorial pooling* et la programmation d'un robot pour mélanger des échantillons d'ADN selon des règles mathématiques pour des dépistages à grande échelle afin d'augmenter la puissance de la découverte de variations génétiques rares dans de grandes cohortes
- Réalisation de deux criblages combinatoires en utilisant la stratégie Reed-Solomon
- Analyse des contraintes et démonstration que les codes Reed-Solomon satisfont les propriétés souhaitées (les règles mathématiques correspondent aux exigences à la paillasse, (par exemple les plaques 96 puits) pour la réalisation de criblages groupés

Mission : découvrir des mutations rares dans les maladies neurodégénératives

Résumé

Les variantes de risques rares constituent un défi majeur pour les études d'association de conditions humaines complexes. Les recherches sans hypothèse nécessitent un nombre important d'individus pour atteindre la puissance statistique nécessaire à l'association de ces variantes. Nous avons adopté une approche radicalement différente pour la recherche de mutations rares dans les maladies neurodégénératives. Dans nos études précédentes, nous avons établi un modèle de levure alpha-synucléine qui récapitule les pathologies moléculaires de la maladie de Parkinson. En utilisant ce modèle, nous avons effectué des criblages à haut débit pour rechercher des modificateurs de la toxicité de l'alpha-synucléine dans la levure. Nous avons ensuite mené une étude exploratoire de séquençage ciblant les homologues humains de ces gènes ainsi que les gènes de risque connus de la maladie de Parkinson chez plus de 500 patients atteints de synucléinopathie familiale et sporadique.

Résultats

- Identification de l'enrichissement de multiples variations qui étaient partagées entre différentes synucléinopathies et enrichies en types de cellules du SNC distincts. Nous avons redécouvert de rares (MAF<1%) mutations pathogènes connus dans les gènes de risque de PD (GBA et LRRK2), et avons récupéré de multiples nouvelles mutations (Zielinski et al. *In Preparation*).

Méthodologie

- La collecte des échantillons, la préparation de librairies de séquençages pour le re-séquençage ciblé des homologues humains des gènes qui ont modifié la toxicité du gène impliqué dans Parkinson (*alpha-synuclein*) par un dépistage de génome entier dans la levure *S. cerevisiae*
- Contrôle de qualité des données brutes et vérification des identités des échantillons et des effets *batch*
- Intégration des données de patients avec celles issues de ~2500 témoins pour faire la détection de variantes de l'ensemble des échantillons
- Des analyses statistiques pour prioriser les mutations chez les patients
- La mise en place une chaîne de traitement des données, y compris l'annotation et filtration de mutations. Application de la méthode aux données issues de trois grandes cohortes de la maladie Parkinson et MSA (Multiple System Atrophy) pour trouver des mutations significatives de la même façon
- La mutation la plus significative issue de mon analyse était une mutation déjà associée avec la maladie de Parkinson validant ainsi l'approche statistique

Autres activités

- Préparation des échantillons et détection des mutations et la filtration pour générer un catalogue des allèles de risque dans la population juive ashkénaze : teamerlich.org/riskcatalog

2015 - 2017 : NEW YORK GENOME CENTER/COLUMBIA UNIVERSITY

À la Columbia University et New York Genome Center, j'ai travaillé au sein d'une équipe multidisciplinaire associant des coordinateurs d'études cliniques, des bioinformaticiens, des scientifiques de laboratoire et des chercheurs postdoctoraux pour comprendre les bases génétiques des traits et maladies complexes. Ce poste a nécessité une solide connaissance théorique de la génétique humaine pour contribuer à faire avancer des projets divers, principalement sur des maladies génétiques rares, comprenant des parties expérimentales et des analyses computationnelles.

Mission : évaluation du rôle des microsatellites dans les traits et maladies complexes

Résumé

L'étude des éléments répétitifs est un défi tant sur le plan expérimental que sur le plan des analyses computationnelles. Cependant, ils constituent plus de la moitié du génome humain et jouent un rôle important dans les traits et maladies complexes. Les microsatellites (une classe d'éléments répétitifs) sont témoins d'un dysfonctionnement cellulaire au cours de la réplication. Des erreurs se traduisant par un nombre de répétitions des nucléotides non conservé, soit réduit, soit plus élevé par rapport à la séquence originale (le nombre de répétitions) du microsatellite considéré. L'approche « gold standard » est le génotypage par électrophorèse capillaires, une méthode à bas débit. J'ai participé à la conception des plans expérimentaux pour améliorer la détection des STRs à partir des technologies de séquençage à haut débit et caractérisation de leur rôle sur l'expression des gènes. Outre leur rôle dans les maladies génétiques, l'analyse des STRs est une méthode largement utilisée en médecine légale.

Résultats

- Validation d'une méthode de la détection des STRs depuis des données de séquençage haut débit (Willems et al. 2017)
- Création d'un catalogue de polymorphismes pour améliorer l'empreinte génétique (identification de l'ADN basé sur la répétition des microsatellites des données *NGS*) : teamerlich.org/refstr

Méthodologie

- Concevoir et réaliser des expériences avec des lignées cellulaires à l'aide du système CRISPR/Cas9 pour tester l'architecture génétique de traits complexes
- La caractérisation des STRs à partir du séquençage à haut débit (NGS)
- La mise en place du typage de marqueurs répétitifs (STRs) d'échantillons publics utilisant deux méthodes différentes en parallèle pour améliorer les algorithmes de détection des éléments répétitifs dans le cadre d'études médico-légale
- La validation de la longueur de tous les STRs que nous avons a détectés depuis des données NGS par le clonage et séquençage *Sanger*. J'ai créé un catalogue de polymorphismes pour améliorer l'empreinte génétique (identification de l'ADN basées sur la répétition des microsatellites des données *NGS*), pour partager en tant que ressource publique.

- L'application d'un gène rapporteur pour quantifier le niveau d'expression des microsatellites présentant un nombre de répétitions différent. Pour cette démarche, j'ai conçu le plan expérimental, fait les manipulations, y compris les transfections et le tri des cellules.
- Afin d'extraire l'importance biologique et améliorer la reproductibilité j'ai développé des protocoles biologiques et informatiques pour automatiser l'analyse des données de cytométrie en flux.
- Pour l'analyse des données issues des expériences du tri, j'ai transformé des fichiers issus d'un instrument FACS, classé les cellules utilisant la démarche d'apprentissage non-supervisé *k-means* et comparer la pente de la régression entre toutes les conditions des manipulations. Ces différences montrent l'effet de la longueur d'un STR particulier sur l'expression du gène en proximité.

Mission : analyses généalogiques et de génétique médicale

Résumé

Les analyses bioinformatiques nécessitent d'être calibrées et relancées à de multiples reprises, notamment pour comparer des approches et/ou paramètres différents. Dans le cadre de l'analyse de données issues des génomes entiers, chaque analyse peut prendre un temps considérable. L'analyse de ce type de données demande donc d'optimiser ses développements et de paralléliser les traitements informatiques. J'ai été responsable des analyses généalogiques et de génétique médicale à partir de données *NGS* à l'échelle des génomes entiers séquençage jusqu'au rendu final des résultats pour une conférence au New York Genome Center.

Résultats

- J'ai fourni tous les résultats aux experts en génomique, en génétique des populations, et en généalogie et le grand public à une conférence au New York Genome Center
- J'ai mené cette analyse en collaboration avec un journaliste scientifique (Carl Zimmer) qui paraît dans son roman (*She Has Her Mother's Laugh*) ainsi que son projet « Game of Genomes » [nature.com/articles/d41586-018-05211-z]

Méthodologie

- J'ai fait des contrôles de qualité techniques et génétiques (évaluation de la couverture moyenne du génome, le rapport *transition/transversion* (Ti/Tv), le rapport entre les substitutions non-synonymes/synonymes, les variantes hétérozygotes et homozygotes, et la vérification du sexe de l'individu (homozygotie entre les marqueurs du chromosome X).
- Après avoir validé les données, j'ai déterminé l'ascendance de chaque échantillon en exploitant des données des populations de références pour trouver les différences de fréquence des allèles dans les données pour affecter les individus à des sous-populations en fonction de l'analyse des probabilités.
- J'ai effectué les analyses généalogiques, dont l'ascendance génétique au niveau local et global, les marqueurs des lignées maternelles et paternelles, ainsi que les polymorphismes associés aux traits et maladies génétiques depuis des génomes entiers.

Mission : améliorer la méthodologie du stockage des données sur l'ADN

Résumé

L'ADN comme outil de stockage de données numériques est non seulement incroyablement dense, mais il ne deviendra jamais obsolète. Aucun média artificiel n'est fiable après 10 à 20 ans. Par contre, l'ADN peut durer des dizaines de milliers d'années. L'idée d'exploiter l'ADN pour le stockage des données numériques n'est pas récente, datant des années 1950 avec la première mise en œuvre réussie dans les années 1980. Les principales avancées ont eu lieu au cours des dix dernières années, mais la plupart des méthodes n'ont pas réussi à récupérer toutes les données sans erreur.

Résultats

- Publication d'une méthode qui permet la récupération des informations à partir d'une densité physique d'environ 215 PB (215 000 000 GB) par gramme d'ADN (Erlich and Zielinski 2017)
- Démonstration de la faisabilité de stocker des données dans l'ADN à l'échelle 60 % de mieux que tout autre groupe avant et 85 % de la limite théorique
- J'ai été invitée à prendre la parole de cette recherche à plusieurs conférences, notamment TEDxVienna, Cold Spring Harbor Laboratory (NY, USA), Web à Québec, Casa Firjan (Rio, Brésil), Kavli Frontiers of Science (Washington, USA), Data Science Congress (Mumbai, Inde), et Morals & Machines (Dresden, Allemagne)

Méthodologie

- Nous avons exploité un algorithme très robuste qui s'appelle *fountain codes* pour surmonter des contraintes biochimiques qui limitent la densité par nucléotide
- Nous prétraitons le fichier binaire en une série de segments qui ne se chevauchent pas, regroupons les données dans un nombre quelconque de messages courts, appelés *droplets* (gouttelettes), en sélectionnant un sous-ensemble aléatoire de segments du fichier à l'aide d'une distribution spéciale (afin d'assurer que la plupart des *droplets* sont créés avec un petit nombre de segments inputs ou un nombre intermédiaire fixe de segments) et nous avons ajouté bit par bit sous un champ binaire
- Traduction du *droplet* du binaire en ADN et exclusions des séquences qui ne répondent pas aux contraintes biochimiques pour réaliser le potentiel de codage de chaque nucléotide
- Pour évaluer la fiabilité de notre approche, nous avons aussi effectué des tests. Nous avons inséré des erreurs ou encore supprimé aléatoirement (*downsampling*) des séquences d'ADN. Nous avons également exploré la densité physique maximale réalisable : en diluant l'ADN pour chercher le point de rupture (information illisible).

Autres activités

- J'ai participé au développement d'une méthode de collecte des données génétiques déjà existantes (dna.land) grâce à des personnes désireuses de partager leurs données pour découvrir la base génétique des maladies complexes de traits nécessitant des quantités importantes de données génomiques. Afin d'impliquer les utilisateurs, nous avons analysé

l'ascendance génétique, les scores de risques polygéniques, et des cousins éloignés par l'analyse des segments d'ADN partagés (Yuan et al. 2018)

- Préparation des libraires et l'intégration de ces données et des données puces afin d'identifier des mutations dans une cohorte de patients atteint d'une maladie rare (*Multiple Symmetric Lipomatosis*)

2017 - 2019 : INSTITUT CURIE

J'ai été engagée à l'institut Curie en 2017, partagée entre la plateforme de bioinformatique et deux équipes de l'unité de biologie du développement (dirigés par le Pr. Edith Heard et le Dr. Raphaël Margueron, respectivement). Les deux équipes sont spécialisés en épigénétique et se focalisent sur le rôle de l'expression des gènes et leur association avec les modifications de la chromatine. Mes missions comprenaient l'intégration de données sur l'expression à haut débit spécifique des allèles, l'interaction ADN-protéine, l'accessibilité de la chromatine et le séquençage de la conformation des chromosomes pour comprendre le rôle de l'inactivation du chromosome X et d'autres modifications de la chromatine dans le cancer et les maladies rares.

Mission : Évaluer le rôle du gène BAP1 sur l'expression des gènes

Résumé

L'activité de la famille PcG (*Polycomb-group Proteins*) est nécessaire au développement normal et au maintien des schémas d'expression génétique. Le complexe Polycomb PR-DUB, composé des protéines BAP1 et ASXL1/2, est responsable de la déubiquitylation de l'histone H2A, une activité impliquée dans le maintien de la répression transcriptionnelle des gènes cibles du complexe. BAP1 est un suppresseur de tumeurs qui interagit avec de nombreuses autres protéines, dont des facteurs de transcription et des modificateurs de la chromatine, rendant plus complexe son analyse. Pour comprendre si le complexe BAP1 oppose l'action du PRC2 (*polycomb repressive complex*) qui dépose H3K27me3, j'ai évalué l'expression des gènes dans des lignées cellulaires portant une délétion de BAP1 ou EZH2 (l'enzyme catalytique du PRC2), ou une délétion des 2 gènes. Voyant que la plupart des gènes sous-régulés dans le BAP1 KO restent silencieux dans le BAP1/EZH2 KO, nous pouvons dire que BAP1 régule principalement l'expression des gènes non régulés par les protéines *Polycomb*, soutenant son rôle d'activateur plutôt que son anti-répresseur. Sur la base de ces analyses, nous avons pu conclure que BAP1 limite l'activité répressive de la PRC1 pour promouvoir la transcription des gènes cibles.

Résultats

La démonstration que BAP1, un modificateur de la chromatine impliqué dans le cancer, agit comme un activateur transcriptionnel en montrant que le complexe BAP1 favorise l'expression des gènes en s'opposant à l'ubiquitylation de l'H2A médiée par la PRC1, de manière largement indépendante de l'antagonisme avec la PRC2 (Zielinski, Campagne, Lee et al. 2019).

Méthodologie

- J'ai intégré des données issues de ChIP-seq afin de détecter des régions significatives enrichies en marques épigénétiques répressives ou actives

- Évaluation de l'impact au niveau transcriptomique de la délétion de *BAP1*, *ASXL1* et *ASXL2* dans la lignée cellulaire HAP1 (une lignée cellulaire quasi-haploïde dérivée d'un patient atteint de leucémie myéloïde chronique) en comparant ces lignées KO (*knockout*) à une lignée sauvage et à une lignée présentant une délétion de la protéine Polycomb *EZH2*
- J'ai réalisé l'analyse de l'expression différentielle à partir de données RNA-seq pour mieux comprendre le rôle de BAP1 sur l'expression des gènes. J'ai trouvé que la plupart des gènes étaient sous-exprimés lors de la suppression de BAP1 par rapport à des lignées sauvages, suggérant que BAP1 agit comme un activateur transcriptionnel
- J'ai comparé le taux de gènes sur- et sous-exprimés après la suppression de BAP1 avec ceux issus de la suppression de deux autres activateurs transcriptionnels (SMARCB1 et CREBBP)
- J'ai identifié l'association des gènes significativement sous régulés aux marques d'histone permissives ou répressives de la transcription en analysant la présence des marques H3K4me3 et H3K27me3 autour du TSS (*Transcription Start Site*) des gènes sous-exprimés. J'ai trouvé que tous les gènes sous-régulés en l'absence de BAP1 sont enrichis de la marque répressive H3K27me3, ce qui suggère qu'un tel gain est une conséquence plutôt qu'une cause de la sous-régulation transcriptionnel

Mission : Évaluer des interactions entre la chromatine et la lamina nucléaire

Résumé

La lamina nucléaire joue un rôle important dans la régulation de l'expression des gènes par des interactions avec la chromatine. Les régions de la chromatine qui s'associent aux lamina « LADs » constituent environ un tiers des génomes de la souris et de l'humain. Le but de ce projet était de comprendre des interactions entre la chromatine et la lamina nucléaire au niveau allèle spécifique pour éclaircir le rôle de la délétion de DXZ4, qui joue un rôle dans l'échappement à l'inactivation du chromosome X, et l'association avec la lamina nucléaire des gènes qui échappent à l'inactivation du chromosome X. Pour ce faire, nous avons exploité des données issues des mêmes lignées cellulaires des souris, dont deux lignées portant une mutation pour comprendre les effets allèle spécifique sur l'expression des gènes et la conformation de la chromatine lors de la suppression d'une région sur le chromosome X.

Résultats

Des biais dans les données suggèrent que les lignées cellulaires avaient accumulé des mutations suite à un plus grand nombre de passages en culture dans certains clones. La suite de ce projet nécessiterait un re-séquençage.

Méthodologie

- L'analyse des données d'expression, (RNA-seq), ATAC-seq (régions ouvertes d'ADN), l'interaction entre la chromatine la protéine LaminB1 (DamID-seq) et les interactions de la chromatine (HiC)
- Le séquençage des souches de souris hybrides et l'alignement utilisant un génome masqué qui tient compte des mutations spécifiques à la souche pour permettre l'analyse des allèles particuliers

- Contrôle de qualité pour assurer une profondeur suffisante afin d'attribuer les lectures aux souches parentales pour classifier l'expression ou la conformation de manière allèle-spécifique
- Développement d'une chaine de traitement des données allèle spécifique issues du séquençage *DamID* (DNA adenine methyltransferase Identification)
- Veille scientifique sur le sujet et la technologie
- Intégration des données et interprétation des résultats afin de comprendre l'association de certains gènes avec la lamina

Mission : Évaluer le rôle du *Polycomb group* dans la transmission épigénétique

Résumé

L'héritage épigénétique des états d'expression des gènes permet à un seul génome de générer des identités cellulaires distinctes et est donc essentiel pour le développement d'organismes multicellulaires. Il est estimé que les protéines du *Polycomb group* sont les médiateurs d'une mémoire épigénétique à base de chromatine dans les cellules de mammifères, mais cette hypothèse n'a pas encore été directement vérifiée.

Résultats

Nous avons montré que l'inactivation transitoire du *Polycomb Repressive Complex 2* (PRC2) dans les cellules somatiques de mammifères entraîne une dé-répression stable de grands sous-ensembles de ses gènes cibles. Ces cas de commutation épigénétique irréversible suggèrent que le circuit de régulation transcriptionnelle de ces gènes est structuré pour convertir des signaux de courte durée en une mémoire d'expression à long terme (Holoch et Wassef et al. 2021).

Méthodologie

- Analyse différentielle entre des lignées sauvages (MEFs et NPCs), PRC2 KO et *rescue* (réexpression) pour détecter des régions enrichies entre les conditions différentes
- Caractérisation des gènes « PRC2 responsive » comme ceux avec un taux de marques H3K27me3 autour du TSS (*Transcription Start Site*) détecté par ChIP-seq qui sont aussi significativement sur régulés lors la suppression de PRC2 (EZH1/EZH2)
- Classification de la réversibilité de l'expression des gènes ciblés par PRC2 selon l'expression dans des lignées différentes

Autres activités

- Toutes les analyses bioinformatiques (RNA et ChIP-seq) pour évaluer la redondance dépendent de la prolifération des sous-unités EZH1 et EZH2 de la PRC2 dans des tumeurs malignes de la gaine nerveuse périphérique (Wassef et al. 2019)
- Analyses des données *single cell* RNA-seq issues des cellules dérivées de la lignée germinale et du soma de la souris pour étudier la réactivation du chromosome X paternel en cours de développement

- Évaluation de certains gènes autosomiques qui ont été classifiés comme monoallélique aléatoire et montrent une expression monoallélique persistante et stable (Marion-Poll et al. 2020)
- Analyse allèle spécifique de l'expression des allèles des cellules gliales et des neurones des souris montrant que le gène Mecp2 (impliqué dans les maladies neurologiques) s'échappe ou non de l'inactivation du chromosome X
- Toutes les analyses bioinformatiques (RNA et ChIP-seq) pour comprendre le rôle du gène EZHIP dans la régulation de la chromatine et l'expression des gènes dans les gamètes (Ragazzini et al. 2019)
- La mise en place des chaînes de traitement bioinformatique
- Animer des cours en bioinformatique pour les biologistes

2019 - PRESENT : INSERM

J'ai été recrutée en tant que chercheuse en bioinformatique en septembre 2019 pour assurer la liaison entre une start-up médicale (Cibiltech) et une équipe de recherche Inserm (U970 : Paris Transplant Group). J'entretiens un dialogue permanent avec les membres de l'équipe de recherche et mène le développement des stratégies d'analyse pour catégoriser les rejets parmi les receveurs de greffes d'organes afin d'appliquer l'évaluation moléculaire des biopsies aux soins standard et de surmonter les limites des systèmes de diagnostic classiques.

Mission : Exploitation des données moléculaires issues des biopsies dans le diagnostic et l'évaluation des allogreffes

Résumé

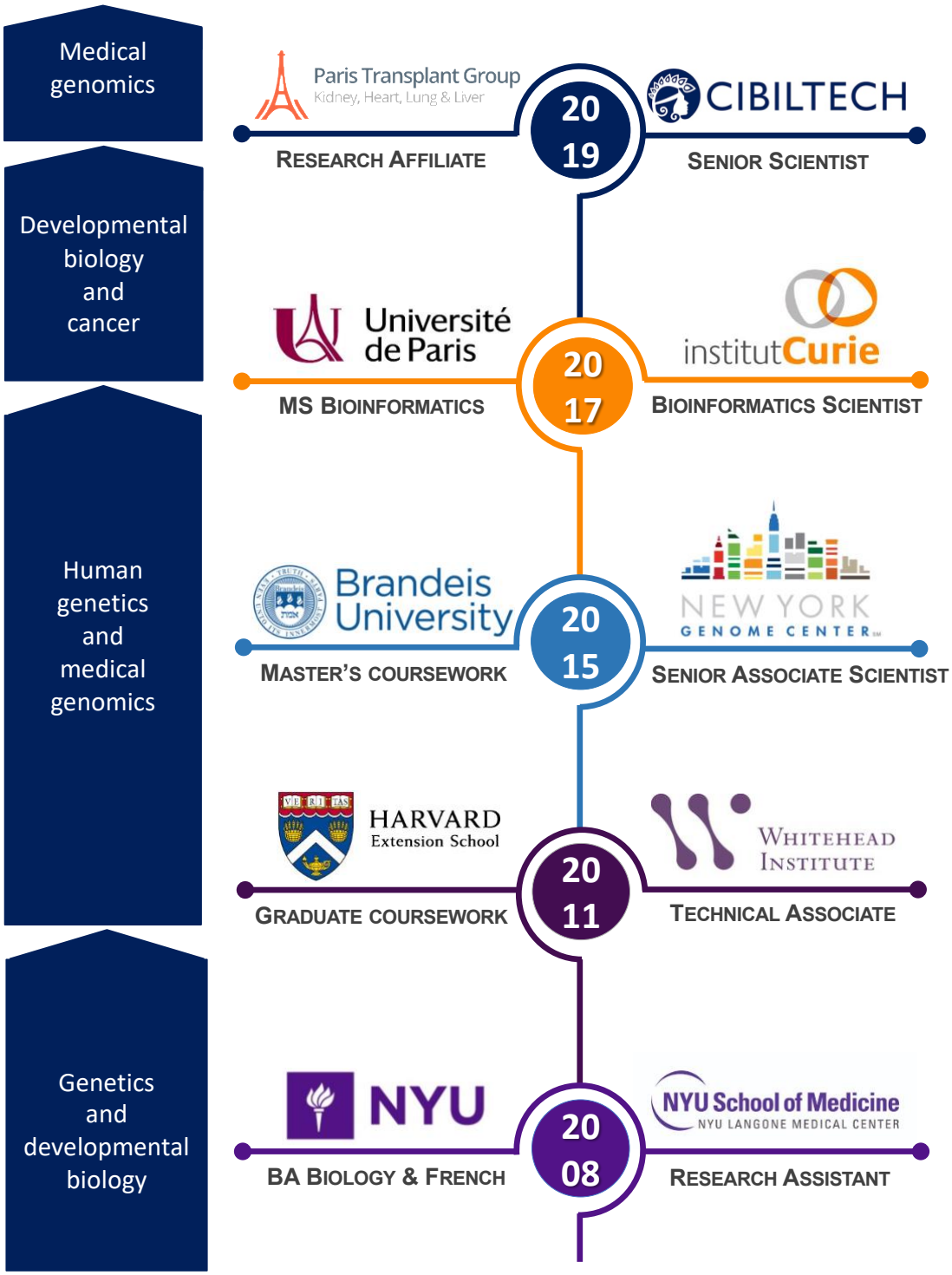
Un diagnostic précis du rejet d'une greffe est crucial pour un traitement approprié et pour mener des essais cliniques afin d'améliorer la survie à long terme des allogreffes. Les lésions histologiques associées à un rejet sont classées à l'aide d'une notation semi-quantitative basée sur des avis d'experts plutôt que sur des données. L'expression des gènes fournit une mesure plus objective et quantitative que l'histologie.

Résultats

La mise en place et direction d'un consortium international d'experts en transplantation pour centraliser et analyser des données moléculaires afin de faciliter le diagnostic du rejet : icdot.org.

Méthodologie

- Harmonisation des données moléculaires, cliniques, et histologiques et statistiques descriptives et inférentielles
- Application des méthodes d'apprentissage supervisé pour faciliter la classification du rejet depuis les données d'expression d'un panel de gènes critiques de la réponse immunitaire, des lésions tissulaires et des mécanismes d'action pour les médicaments immunosuppresseurs dans les allogreffes prédominantes (rein, cœur, poumon, et foie)


Medical genomics
Developmental biology and cancer
Human genetics and medical genomics
Genetics and developmental biology
Paris Transplant Group
Kidney, Heart, Lung & Liver
RESEARCH AFFILIATE
2019
CIBILTECH
SENIOR SCIENTIST
Université de Paris
MS BIOINFORMATICS
2017
institutCurie
BIOINFORMATICS SCIENTIST
Brandeis University
MASTER'S COURSEWORK
2015
NEW YORK GENOME CENTER
SENIOR ASSOCIATE SCIENTIST
HARVARD Extension School
GRADUATE COURSEWORK
2011
WHITEHEAD INSTITUTE
TECHNICAL ASSOCIATE
NYU
BA BIOLOGY & FRENCH
2008
NYU School of Medicine
NYU LANGONE MEDICAL CENTER
RESEARCH ASSISTANT

RETOUR REFLEXIF

J'ai toujours été attirée par les activités intellectuelles, notamment celles qui impliquent notre compréhension du monde naturel. Toutefois, étant la première personne de me famille à accéder à des études supérieures, je manquais de confiance en ma propre capacité à réussir. Pour contrebalancer ce manque de confiance, j'ai toujours fait preuve de détermination dans la poursuite de mes objectifs et je me suis toujours poussée au-delà de ce que je pensais être capable de réaliser. La curiosité et la poursuite des études n'étaient pas encouragées pendant mon enfance mais l'école m'a fournie la détermination nécessaire. Lorsque j'ai entamé mes études à New York University, j'étais avide d'apprendre. J'ai décidé de me spécialiser en biologie, et en parallèle d'approfondir le français, les deux sujets qui m'avaient passionnée le plus au lycée. J'étais intéressée par les processus biologiques du vivant, sans forcément avoir l'idée d'être chercheuse et je souhaitais assimiler la culture et la langue françaises, mais je n'avais pas encore l'intention de m'installer en France. Les étudiants en biologie n'étaient pas obligés d'acquérir une expérience pratique de la recherche mais, dès la première année, j'ai décidé de m'intégrer dans un laboratoire, spécialisé dans l'étude du contrôle physiologique et développemental de l'établissement des cellules souches germinales dans les nématodes. J'ai choisi ce laboratoire parce que j'avais participé à un programme de recherche pendant l'été avant la terminale, à l'Institut Waksman de Rutgers University, où j'avais contribué à des recherches similaires sur les nématodes. J'ai été séduite par le chaos organisé du laboratoire. J'ai savouré l'épuisement de perdre la notion du temps en dépistant les vers minuscules qui se tortillaient sous le microscope, tard dans la nuit, et j'étais complètement captivée par ce que je faisais.

Quand j'ai obtenu mon diplôme pendant la crise économique en 2008, je savais que je voulais continuer à travailler dans un laboratoire et je cherchais désespérément un poste. Au bout de quelques mois, j'ai décidé de m'installer en France, puisque c'était mon objectif lorsque j'ai débuté le français à l'âge de 13 ans. C'est ainsi que j'ai mis les pieds en France pour la première fois, et j'ai emménagé dans une famille que je n'avais jamais rencontrée (qui est maintenant devenue ma famille française) dans la banlieue Parisienne, en tant que jeune fille au pair. Pendant ces six mois, j'ai continué à dévorer des livres et des publications scientifiques, tout en améliorant mon français. Peu après mon retour aux États-Unis, j'ai accepté un poste à Cambridge au Massachusetts, dans une petite entreprise de recherche génétique qui utilisait l'organisme modèle *Drosophile*. Chacun était chargé d'un projet ou une tâche spécifique, mais après plusieurs mois, j'ai été capable d'effectuer toutes les techniques liées à la recherche, y compris le clonage et l'extraction de l'ADN, les injections dans les embryons, le dépistage des mutants, et les croisements génétiques. Je posais souvent des questions à mon responsable, un généticien, et nous avions des discussions approfondies sur la génétique. Cependant, le travail que nous faisions était un service pour les laboratoires de recherche, et il était frustrant pour moi de ne pas connaître les détails des projets. J'avais besoin d'un plus grand défi.

Chaque jour, pendant mon trajet, je traversais un carrefour de l'innovation (au sens propre et figuré) de Kendall Square, près du campus du MIT, en passant devant l'un des instituts les plus connus mondialement en recherche génomique, pour son rôle majeur dans le séquençage du génome humain. Finalement, après plus d'un an, j'ai osé postuler dans un laboratoire récemment monté par un jeune chef (Dr. Yaniv Erlich). Je n'oublierai jamais mon entretien, ni le moment où mon futur patron m'a appelée pour me proposer le poste. J'avais lu ses publications clés, mais je n'avais aucune expérience en programmation ni en génomique. Cependant, comme ce domaine était en train d'émerger, il suffisait alors d'être très motivé pour apprendre, puis de suivre l'évolution rapide de la technologie et ses applications.

Le MIT est un environnement dynamique et rempli de défis, où j'ai travaillé avec des experts mondiaux en génomique, ce qui a exigé de la ténacité pour être à la pointe de ce domaine en constante transformation. À ce poste, j'ai appris les technologies de séquençage, notamment la

préparation des librairies de séquençage et j'ai élargi mes compétences en biologie moléculaire et cellulaire. Un rite de passage en biologie moléculaire est d'accepter que la plupart des expériences consistent à faire des mélanges de volumes infimes de liquides. Ce travail était certes souvent ennuyeux, mais je savais que c'était nécessaire pour répondre aux questions importantes posées, et j'ai toujours cherché à comprendre le rôle et les conséquences de l'ajout de chaque réactif dans les kits que j'ai utilisés. Cette approche m'a permis, lorsque c'était possible, de trouver des solutions aux problèmes techniques des expériences, d'être plus efficace, et d'améliorer la fiabilité des résultats. Je ne me suis jamais contentée de suivre les protocoles et de rendre compte de mes résultats ; depuis le début de ma carrière, j'ai toujours cherché à élargir mes connaissances scientifiques et mes compétences au-delà de ce qui était exigé d'un technicien.

Peu de temps après le début de ce poste en 2011, j'ai réalisé qu'il était nécessaire d'acquérir des compétences en programmation pour faire avancer ma carrière en génomique, étant donné le potentiel de la bioinformatique en tant qu'outil pour répondre à une problématique scientifique (gain de temps, de représentabilité et de fiabilité des données). La recherche en biologie, notamment l'étude multidisciplinaire de la complexité du vivant, devient de plus en plus une science quantitative et les méthodes computationnelles et technologies se développent et évoluent rapidement. La biologie computationnelle rend les concepts biologiques plus rigoureux et vérifiables, pour améliorer notre compréhension du monde naturel. J'ai donc commencé à apprendre des langages informatiques, et j'ai acquis au fur et à mesure les compétences pour pouvoir gérer un projet dans son intégralité, de la paillasse à la ligne de commande. Étant la seule dans l'équipe à faire des expériences à la paillasse, mes responsabilités n'ont cessé de s'accroître, ce qui m'a permis d'affiner ma méthodologie. Après avoir appris toutes les techniques nécessaires, j'étais chargée de la gestion de divers projets avec axés sur la génomique. L'un de mes projets consistait à trouver la mutation causale d'une maladie rare. La publication qui en résulte en 2014 est au carrefour de tous les domaines dans lesquels j'ai travaillé depuis : génomique, maladies rares, neurologie, biologie du développement, et cancer. En parallèle, j'ai collaboré avec le laboratoire du Dr. Susan Lindquist (ancienne directrice du Whitehead Institute) pour qui j'ai supervisé un projet sur les maladies neurodégénératives. Mon rôle était initialement de collecter les échantillons et de préparer les librairies de séquençage. Mais ce projet m'a rapidement motivée à accroitre mes connaissances en bioinformatique, à être autonome dans le développement et le lancement des analyses, et j'en ai présenté les résultats lors de réunions internes ainsi qu'à certaines conférences. Mes supérieurs et mes collègues m'ont poussée à me dépasser, en me chargeant non seulement de faire des expériences et des analyses statistiques, mais aussi en me donnant la responsabilité de conduire des projets de recherche collaboratifs, montrant leur confiance dans mes connaissances et mes compétences de chercheuse.

Après mûre réflexion, j'ai décidé de suivre mon responsable quand le laboratoire a déménagé du MIT au New York Genome Center en 2015. Ayant commencé ma carrière dans une petite équipe de recherche, cela m'a permis d'élargir mes compétences expérimentales et d'apprendre la bioinformatique, mais aussi de prouver mes aptitudes scientifiques : de formuler des hypothèses et concevoir ce qu'il faut pour y répondre. Cette liberté de monter en compétence et de diriger mes projets m'a permis de passer du statut de technicien à un poste de chercheuse autonome. J'ai profité de cette opportunité pour continuer à élargir mes connaissances, dans un institut axé sur la génomique. J'ai entrepris des projets informatiques pour lesquels j'ai dû acquérir de nouvelles compétences en combinant des stratégies moléculaires et informatiques. Un des sujets concernait des éléments répétés, les microsatellites. Ils sont témoins d'un dysfonctionnement cellulaire au cours de la réplication, et la longueur des répétitions peut influencer l'expression des gènes à proximité. Ils sont difficiles à étudier tant sur le plan expérimental que sur celui du calcul car le même glissement de la réplication qui se produit dans les cellules se produit également *in vitro*. Notre équipe a développé des stratégies pour les détecter à partir des données de séquençage haut débit. J'ai effectué des expériences pour valider leur présence ainsi que l'association entre la taille

des microsatellites et l'expression de gènes par des méthodes informatiques et expérimentales. À ce stade de ma carrière, mon temps était réparti à 50/50 entre la paillasse et les analyses bioinformatiques. Après m'être formée de façon autonome, je voulais approfondir et confirmer mes compétences, j'ai donc suivi un master en bioinformatique en parallèle de mon emploi. J'ai développé une meilleure compréhension des différentes méthodes d'analyse et d'extraction de l'information biologique à partir de vastes jeux de données, pour poser des questions pertinentes et y répondre.

Deux ans plus tard, je me sentais maintenant prête pour un nouveau défi de l'autre côté de l'Atlantique. J'ai été engagée à l'Institut Curie en 2017 en tant qu'ingénieure en bioinformatique, partagée entre la plateforme de bioinformatique et deux équipes de l'unité de biologie du développement. J'avais commencé ma carrière en étant à 100% dans le travail expérimental, et j'avais maintenant un rôle 100% informatique. Cette double compétence, à l'interface des deux rôles, me permet une meilleure compréhension des défis expérimentaux liés à des systèmes biologiques complexes et informatiques ; je peux ainsi traduire les questions biologiques en questions bioinformatiques. Les deux équipes auxquelles j'apportais mon expertise sont spécialisées en épigénétique, et se focalisent sur l'expression des gènes et leur association avec les modifications de la chromatine. Bien qu'étant une débutante en épigénétique, et même si les questions spécifiques étaient différentes de celles étudiées au cours de mes précédentes fonctions, les objectifs étaient très similaires : comprendre l'architecture moléculaire qui sous-tend le dysfonctionnement cellulaire dans la maladie.

Étant la seule bio informaticienne d'une des équipes, j'étais responsable d'analyser toutes les données haut débit du laboratoire. J'avais souvent plusieurs études différentes en parallèle, en plus des analyses pour la deuxième équipe dont un projet que je devais mener. L'objectif de ce dernier était de mieux comprendre les interactions entre la chromatine et la lamina nucléaire (qui joue un rôle important dans la régulation de l'expression des gènes) au niveau d'allèles particuliers, dans des lignées cellulaires des souris. J'ai dû apprendre les méthodes appropriées pour plusieurs types de données, interpréter les résultats afin de comprendre l'association de certains gènes avec la lamina, et les présenter à l'équipe. Pour les différents types de données, j'ai dû adapter des codes existants, et parfois développer entièrement un processus d'analyse dédié. Quand j'ai enfin pu interpréter les résultats, j'ai remarqué des biais dans les données, suggérant que les lignées cellulaires avaient accumulé des mutations suite à un plus grand nombre de passages en culture dans certains clones. Ce projet nécessiterait un re-séquençage, néanmoins cela n'est pas possible dans l'immédiat car le laboratoire a déménagé à l'EMBL en Allemagne. Ce nouveau double challenge (scientifique et culturel) a concrétiser l'intégration de mes acquis biologiques et informatiques, me permettant de démontrer ma capacité à comprendre le sujet, poser les bonnes questions, et de m'appuyer sur mes connaissances scientifiques, maintenant très solides. Grâce à ce projet, j'ai réalisé mon besoin de m'impliquer plus activement dans la recherche. J'ai dû faire face à la gestion de plusieurs projets complexes en parallèle, comprenant les enjeux de temps qui peuvent limiter le niveau de granularité des analyses réalisées. Finalement, j'ai approfondi mes compétences bioinformatiques et ai acquis un esprit critique sur les limites des algorithmes et sur les méthodes statistiques. Je me suis adaptée au travail avec des équipes pluridisciplinaires, à l'interface entre la biologie et la bioinformatique. Par conséquent, je suis maintenant capable de mobiliser mon expérience et mes connaissances afin de définir une démarche appropriée ou, si besoin, développer de nouvelles approches méthodologiques plus adaptées aux problématiques biologiques rencontrées.

Bien qu'ayant eu les responsabilités d'une chercheuse lorsque j'étais au MIT, ce n'est qu'à partir de mon poste à l'Institut Curie, après mon évolution de biologiste moléculaire à bioinformaticienne, que je me suis considérée comme chercheuse à part entière. En plus de mes responsabilités actuelles, j'ai été invitée à donner des conférences internationales sur le sujet sur lequel je travaillais à New York, j'ai *reviewé* des articles et j'ai continué à mener l'étude sur la maladie de Parkinson pour un laboratoire de Harvard Medical School. Pendant ce temps, j'ai aussi terminé mon Masters en

bioinformatique en 2018 à l'Université de Paris. À cette occasion, mon jury (composé de 4 chercheurs en biologie, bioinformatique, et génétique) m'a fortement suggéré de poursuivre un doctorat pour faire valoir mon niveau actuel.

En septembre 2019, j'ai été engagée en tant que chercheuse pour assurer la liaison entre une start-up médicale et une équipe de recherche Inserm de renommée mondiale dans la transplantation (Paris Transplant Group), un poste qui correspondait en tout point à mes aspirations. Mes missions comprennent le développement des stratégies d'analyse, la conception des algorithmes, le déploiement des outils de calcul pour l'exploration de très grands ensembles de données multidimensionnelles et de fournir des analyses bioinformatiques auprès des équipes et des partenaires de R&D. J'entretiens un dialogue permanent avec les membres de l'équipe de recherche, les équipes scientifiques en interne, ainsi qu'avec des partenaires du public et du privé pour conduire des projets de recherche collaboratifs. Depuis le début de ce poste, j'ai appliqué des méthodes d'apprentissage automatique (« machine learning ») pour catégoriser les rejets parmi les receveurs de greffes d'organes. Suite à l'achèvement de ce projet, le chef de l'équipe (Pr. Alexandre Loupy) m'a chargée de mettre en place et de diriger un consortium international d'experts en transplantation, et de diriger les analyses visant à valider l'application du diagnostic moléculaire sur les tissus. Ces nouvelles responsabilités sont l'aboutissement de ma montée en compétence jusqu'à un rôle de chercheuse à part entière travaillant en complète autonomie.

Mon parcours professionnel a nécessité une adaptation continue aux domaines en constante évolution, ainsi qu'à divers sujets. J'ai ainsi acquis une palette de connaissances et de compétences techniques sur des sujets interdisciplinaires. A chaque fois, j'ai accompagné les projets jusqu'à leur publication, même après mon départ, démontrant ainsi ma motivation, mon engagement, et mon professionnalisme. Mes travaux de recherche ont également contribué aux avances dans plusieurs domaines (génomique, génétique médicale, génétique des populations, théorie de l'information et codage, et épigénétique). Je me suis diversifiée depuis mes débuts au MIT, cependant toutes mes études ont eu pour point commun d'être spécialisées en génomique. La variété des sujets pourrait apparaître comme un désavantage, mais elle apparaît aujourd'hui précieuse, dans une recherche en évolution et aux besoins interdisciplinaires croissants. Plus particulièrement en génomique, il est indispensable d'être très adaptable, car la recherche progresse et évolue très rapidement. Le premier assemblage du génome humain a été achevé quand j'étudiais la biologie au lycée, il avait nécessité des milliards de dollars et plus d'une décennie. Aujourd'hui on séquence un génome en quelques heures pour quelques centaines de dollars, et seule la capacité d'analyse des données est un facteur limitant.

J'apprécie particulièrement la bioinformatique car c'est un domaine pluridisciplinaire alliant la biologie, les mathématiques, les statistiques, et l'informatique. Dans ce domaine, il faut être capable de se former de façon autonome pour progresser en raison de la constante et rapide évolution. J'ai été formée à la recherche au fur et à mesure de mes expériences diverses et enrichissantes et ai acquis la plupart de mes compétences en analysant les données que j'avais moi-même générées à la paillasse. J'ai progressivement élargi le champ de mes connaissances et compétences en biologie et en bioinformatique (notamment la maîtrise des langages de programmation, des méthodes statistiques adaptées, ainsi que l'analyse et la modélisation des données biologiques à grande échelle). J'ai démontré mes capacités à comprendre et à analyser une problématique biologique complexe, puis à définir une stratégie permettant de résoudre ces problèmes en appliquant les méthodes expérimentales et d'analyses biostatistiques les plus appropriées. Pour chaque mission, j'ai démontré mon adaptabilité en utilisant différents outils d'analyse et mon niveau de compréhension des questions biologiques sur des sujets divers pour interpréter les résultats. J'ai toujours approfondi mes connaissances théoriques pour ensuite les appliquer, et manipuler de grands ensembles de données issues de différentes conditions expérimentales. La recherche scientifique est en constante évolution et cela représente pour moi une véritable opportunité que de pouvoir évoluer en parallèle, en mettant continuellement à l'épreuve mes acquis.

II. RARE VARIANT ANALYSIS IN HUMAN NEURODEGENERATIVE DISORDERS GUIDED BY CELLULAR MODELS OF PROTEOTOXICITY

ABSTRACT

Aggregation of specific proteins within neurons and glia comprise the hallmark pathologies of neurodegenerative diseases like Alzheimer's (AD) and Parkinson's disease (PD). It remains uncertain whether different clinical diseases linked to misfolding of the same protein share common genetic risk drivers and whether different protein-aggregation pathologies are mechanistically related or distinct. In previous studies, we established amyloid-beta and alpha-synuclein yeast models that recapitulate the molecular pathologies of AD and PD, respectively. To address these questions, we performed targeted exome sequencing in 500 patients with four distinct synucleinopathies: PD, PD with dementia (PDD), dementia with Lewy bodies (LBD) and multiple system atrophy (MSA). Together with known AD and PD genes, we sequenced human homologs of genetic modulators of alpha-synuclein and beta-amyloid toxicity identified in genome-wide screens in yeast, including a broad range of orthologs to identify genetic drivers in diverse patient, cellular and disease specific contexts. Despite the small number of samples, the relatively small set of 430 genes allowed us to concentrate the statistical power on more likely targets. As a control, we compared the burden of rare variants to a cohort of more than 2500 healthy aged individuals with no reported history of cancer, cardiovascular disease, or neurodegenerative disease. Variants in both beta-amyloid (ARGHEF1, MAP2K3, PKN2, TEAD2) and alpha-synuclein (PTPRR, PTPRS, HKR1) genetic networks were enriched across different synucleinopathies compared to healthy aged controls, as were variants in both known PD and AD genes. Associations were independent of concomitant AD pathology. Gene-level tests, including a trend test developed for rare variant analysis, confirmed and extended these findings in additional MSA and PD cohorts. Our data implicate shared genetic drivers across distinct synucleinopathies and demonstrate convergent genetic networks in beta-amyloid and alpha-synuclein proteinopathies in humans. We envision that our cross-species approaches will continue to yield important insights for complex diseases with evolutionarily conserved biological mechanisms, and perhaps help fulfill therapeutic promises in the post-genomics era.

1. INTRODUCTION

We are drowning in data but starved for knowledge. John Naisbitt

1.1 ADVANCES IN GENOMICS

This year marks 20 years since the first draft of the human genome assembly was published (Jones et al., 2021). The global effort of the Human Genome Project was a model for open science, establishing strict guidelines for data sharing. However, due to privacy concerns, access to genomic data has become increasingly challenging. We have the technology to perform the incredibly large-scale population studies that are necessary to uncover genetic variation driving complex traits and diseases. It is critical to strike a balance between patient privacy and scientific progress by establishing reasonable pathways for researchers to access and share data.

Beadle and Tatum's one-gene one-enzyme hypothesis laid the foundation for molecular biology by connecting biochemistry and genetics (Strauss, 2016). This paradigm was the basis for one of the shortcomings of the Human Genome Project and the GWAS era, the missing heritability problem, that single variants do not explain much of the heritability of complex traits and diseases. Genotype-phenotype associations are far more complicated due to allelic and genetic heterogeneity, incomplete penetrance, and variable expressivity. Functional genomic strategies will become increasingly necessary to uncover causal associations between variants and disease phenotypes.

The post-genomic era has been incredibly humbling but has revealed thousands of genetic variants underlying a wide range of traits and diseases. In order to interpret the wealth of genetic data, major advances in understanding how genetic variation drives changes in cellular and physiological functions are needed. Knowledge of genetic risk factors is biased towards individuals of European ancestry, who represent a small fraction of the global population. This lack of diversity and focus

on a minority of the global population limits our understanding of disease susceptibility for the majority. Identifying the missing heritability for complex traits and diseases will require more diverse, inclusive studies across different populations.

Everyone who is born holds dual citizenship, in the kingdom of the well and in the kingdom of the sick. Although we all prefer to use the good passport, sooner or later each of us is obliged, at least for a spell, to identify ourselves as citizens of that other place.

Susan Sontag

1.1.1. TOWARDS PRECISION MEDICINE

Thousands of genome-wide association studies have identified increasing numbers of variants potentially associated with complex human traits and diseases (Visscher et al., 2017). Technological advances have led to increased throughput and decreasing costs of DNA sequencing have enabled the application of clinical genomics. For genome sequencing to become an effective clinical diagnostic tool, significant advances in the functional and mechanistic interpretation of human genetic variation are needed. Functional genomics integrates data from the genome, proteome, metabolome, transcriptome, and epigenome to understand the consequences of genomic variation on complex cellular phenotypes in human traits and disease. The post-GWAS era is marked by a call for functional studies, including initiatives like the NHLBI Trans-Omics for Precision Medicine (TOPMed) program (Taliun et al., 2021) and the iPSC Neurodegenerative Disease Initiative (iNDI).

Major breakthroughs in genomics have increased our understanding of the genetic basis of complex traits and diseases. Clinical sequencing can detect a disease long before symptoms present, enabling early diagnosis and increasing the chances of successful treatment. A major goal of medical genomics is to elucidate the genetic risk and molecular architecture of disease, ultimately providing evidence-based interpretation of pathogenic variants. On a practical level, this entails distinguishing pathogenic from benign mutations in order to establish a causal link. Individual genomes contain an average of 3 million SNPs, more than all ~500,000 health related genomic variants in the

ClinVar database, with each newly sequenced genome contributing an average of more than 8000 novel variants (Telenti et al., 2016). This data can be harnessed to close the genotype-phenotype gap by characterizing genomic regions that are intolerant to variation and validated through high-throughput functional screens.

1.2. GENETIC ASSOCIATION METHODS

Prior to the advent of high-throughput sequencing and population scale genetic association studies, disease and trait associated mutations were determined using linkage mapping in which genetic markers spanning the genome were assessed for co-segregation of genomic regions in individuals with the phenotype of interest. This approach provided evidence for a genetic component in Parkinson's disease (PD) and established a link between mutations in the SNCA gene and familial PD (Polymeropoulos et al., 1997). Family-based and population scale association studies using high-throughput sequencing have both been successfully utilized to identify genomic loci associated with complex diseases. Family-based designs are less susceptible to confounders, particularly population stratification, and can have greater power to identify significant associations as the allelic diversity is smaller. Extensive identity by descent (IBD) between individuals can distinguish true variant calls from sequencing errors. However, finding sufficiently large pedigrees, especially for complex, rare diseases can be challenging. Mendelian forms of complex diseases and family-based studies have been instrumental in understanding the genetic architecture of disease like Alzheimer's and Parkinson's disease. However, monogenic forms of both AD and PD are rare and a polygenic model is likely to explain the genetic contribution of the majority of cases.

1.2.2. RARE VERSUS COMMON VARIANTS

Population-based studies require many samples but, due to rapid human population growth, even doubling the sample size may simply enrich extremely rare variants (Coventry et al., 2010). Most genome-wide association studies have focused on low frequency or common variants (MAF>1% or 5%) and increasing sample sizes have uncovered increasing numbers of associated loci in PD,

most of which appear to have small effect (Nalls et al., 2019). More than a decade of GWASs uncovered thousands of genetic variants associated with complex human diseases and traits. However, interpretation of these single-nucleotide polymorphisms (SNPs) is challenging as the vast majority are in non-coding regions and are in linkage disequilibrium (LD) with the 'true' locus, making identification of the causal genetic variants and biological mechanisms difficult.

Common variants conferring incremental risk explain a small portion of the heritable risk for most complex diseases. Improvements in sequencing technology and statistical methods enabled investigation of the contribution of rare variants as one potential source of this missing heritability. Increasing evidence suggests that rare variants play a significant role in PD and other complex diseases (Germer et al., 2019; Jansen et al., 2017; Momozawa and Mizukami, 2021). Rare variants can still be modified by the environment or other genomic loci but they are generally expected to explain greater variance in disease susceptibility. However, even for the dominant effect model (ie. due to haploinsufficiency or gain of function), penetrance is not expected to be 100% and most individuals with complex diseases are expected to carry multiple, if not many in the case of strong heritability, rare risk variants (Manolio et al., 2009).

Recent human population growth has led to an excess of rare variation (Keinan and Clark, 2012). Most human variants are rare (MAF<1%), functional variants even more so (Keinan and Clark, 2012). Rare variant association tests (RVAS) have supported the contribution of rare variants to phenotypic variability but have also shown the need for greater sample sizes than common variant association tests (CVAS). The more we sequence, the more novel, rare variants we uncover, requiring increasing sample sizes in order to achieve sufficient statistical power to detect genetic associations. Targeted sequencing of candidate genes has been successful in detecting rare variants of large effect for complex traits, with less success for whole exome based studies, which require extremely large sample sizes (Kosmicki et al., 2016). Reducing the search space by focusing on

candidate genes or by collapsing variants in regions of interest (eg. genes) can increase statistical power in rare variant association tests (Lee et al., 2014).

The hypothesized roles of common and rare variants in complex diseases are not mutually exclusive and can even exist at the same loci. Common variants of small effect, rare variants of large effect, rare variants of small effect, and combinatorial variants are likely to all play a role in complex diseases, highlighting the need for integrative approaches to define the genetic architecture of complex diseases.

1.2.3. RARE VARIANT ASSOCIATION TESTING

Most human variants are rare and have greater independence compared to common variants, leading to greater multiple testing penalty for individual variant association tests. Further, many more samples are needed than for common variants to even observe the rare variants and identify their effects. Exome-sequencing studies are generally underpowered in uncovering deleterious mutations with a large effect for complex traits. Assessing the contribution of a single variant to disease risk involves the relatively straightforward comparison of the frequencies of alleles or genotypes at the genomic site in cases compared to controls. However, classical association tests (eg. chi-squared, Cochran-Armitage trend, Fisher's exact, and Wald tests) are not appropriate for rare variation due to their low frequencies and greater numbers than common variants, which can lead to spurious findings when observed counts are too small. Single variant tests can be useful for rare variants when only summary data are available or if the effects are very large, the sample size is sufficient, or variants are not too rare, and to assess sequencing data quality and identify population stratification. Fisher's exact test compares allele frequencies between cases and controls and is considered analogous to the allelic chi-squared test but is appropriate for all, especially small, samples sizes. However, this test can be overly conservative.

Multiple methods exist to aggregate rare variants and test the cumulative effect by gene or genomic region for disease association, but performance is largely dependent on the disease architecture

(Table 1). As this is often unknown, it can be useful to apply a combined test or multiple methods and adjust p-values across all tests/methods. The most common method to group rare variants together in population-based genetic analyses is at the gene level, usually via a collapsing test, combined multivariate and collapsing test, or sequence kernel association test (SKAT). Each method makes varying assumptions of the underlying genetic model and can be powerful in different scenarios, particularly with respect to the presence of benign variants or different directions of effect of causal variants.

Rare variant association methods

Burden tests	Collapse variants into genetic scores	CAST, CMC method, MZ test, WSS, ARIEL test
Adaptive burden tests	Data-adaptive weights or thresholds	KBAC method, VT, aSum, Step-up, EREC test, RBT
Variance-component tests	Test variance of genetic effects	SKAT, C-alpha test, SSU test
Combined tests	Combine burden and VC tests	SKAT-O, Fisher method, MiST
EC test	Exponential combination of score stats	EC test

Table 1: Methods for population-based rare variant association studies. Adapted from (Lee et al., 2014).

Burden tests collapse rare variants into specified genomic regions (eg. genes) and test the association between the weighted sum of rare alleles with disease/trait and are powerful when most variants are causal, and effects are in the same direction but lose power if there are both risk and protective variants in the same collapsed region. **Adaptive burden tests** are robust to neutral variants as well as trait-increasing and trait-decreasing variants. Weights or thresholds are applied to variants under the assumption that variants below a certain threshold or with genotype-based weights are greater in cases than controls. The data-adaptive sum test (aSum) (Han and Pan 2010) is the first to consider the difference in the direction of effect (protective or deleterious) in a specific region, using the estimated direction of effect (with w = -1 for protective variants) in the burden test. **Variance-component tests** (eg. SKAT) are powerful both when a small fraction of variants

is causal or when both trait-increasing and trait-decreasing variants are present. Unlike burden tests, VC tests lose power when most variants act in the same direction. **Combined tests** (eg. SKAT-O), which combine burden and VC tests, are more robust when tested variants have different directions of effect (trait- increasing or decreasing) or many noncausal variants. SKAT-O finds the optimal linear combination of the burden test and SKAT to maximize power, choosing between burden test and SKAT depending on how variants are distributed and additionally provides small-sample adjustment to properly control type I error, for small-sample case-control sequencing association studies. **EC tests** combine score statistics exponentially and is powerful when a very small proportion of variants are causal.

Methods for rare variant association testing have improved considerably, contributing to the prioritization of candidate variants and genes associated with complex traits and diseases. No single test is optimal and different approaches should be assessed to determine the most robust and powerful under different scenarios.

1.3. NEURODEGENERATIVE DISEASE GENETICS

1.3.1. COMPLEX ETIOLOGY OF NEURODEGENERATIVE DISEASE

Aging is a complex trait and the greatest risk factor for neurodegenerative disease (ND), with environmental and genetic risk factors contributing to physical deterioration, increasing risk of disease and death. Human life expectancy has increased by a quarter of a year each year for the past nearly 200 years and continues to increase (Oeppen and Vaupel, 2002). There are few effective treatments for progressive, irreversible neurodegenerative disease and the prevalence and associated burden of age-related disease will continue to increase as the population over the age of 65 grows. The most common neurodegenerative diseases, Alzheimer's and Parkinson's, appear predominantly in the elderly and risk increases with age. Parkinson's disease is the second most common neurodegenerative disorder and was first described more than 200 years ago (Goetz,

2011), yet no intervention exists to slow or stop the associated nigrostriatal degeneration. The global prevalence of individuals with PD has more than doubled globally in the past 20 years (GBD 2016 Parkinson's Disease Collaborators, 2018) and is expected to double within the next 20 years to more than 12 million cases (Dorsey and Bloem, 2018). One in ten individuals aged 65 years or older is diagnosed with Alzheimer's disease and Parkinson's disease has a global prevalence of 0.5% and 20% in individuals 65 or older (Hou et al., 2019). Heritable forms of PD represent 5-10% of cases and the penetrance of familial mutations varies widely and is age dependent.

Aging is the primary risk factor for most neurodegenerative diseases due in part to vulnerability to DNA damage of the postmitotic cells that comprise the affected tissues. The biological hallmarks of aging, notably epigenetic alterations, genomic instability, telomere shortening, and loss of proteostasis, are associated with neurodegenerative disease (López-Otín et al., 2013). Understanding the basic molecular and cellular mechanisms of aging and their role in neurodegenerative disease onset and progression is a major challenge in developing effective treatments.

Despite shared pathobiology across proteinopathies (NDs characterized by the aggregation of a specific protein), the exact nature of the deposited protein and the vulnerability of different brain regions is specific to each disease. Genetic, epigenetic, and environmental risk factors have been shown to be associated with PD, however, the causes of neuronal death and formation of the pathologic hallmark of alpha-synuclein aggregates (Lewy bodies) are not well-understood. Heritable forms of ND have helped to uncover associated genomic loci. Mutations in APP, PSEN1, PSEN2 are observed in rare Mendelian AD but are not risk factors for the more common form of AD, in which APOE is associated with disease susceptibility (Harold et al., 2009). This is in contrast with PD, in which common variants in Mendelian disease genes (SNCA, LRRK2, MAPT) are also risk alleles in the more common idiopathic form (Simón-Sánchez et al., 2009). Disease onset and progression are likely due to a complex interplay between genetic risk factors,

environment, as well as the aging process itself. Determining both drivers of convergent and disease-specific cellular phenotypes are critical to developing disease-targeted therapies.

1.3.2. NEUROPATHOLOGICAL SPECTRUM OF SYNUCLEINOPATHIES

Neurodegenerative diseases are defined by the specific proteins that aggregate in distinct cell types. However, different diseases can be associated with aggregation of the same protein. For example, alpha-synuclein (α-syn) is a lipid-binding protein that contributes to a class of synucleinopathies characterized by the abnormal accumulation of alpha-synuclein in the central and peripheral nervous system (Parkinson's Disease, its associated dementia PDD, Dementia with Lewy bodies, and Multiple System Atrophy). How distinct diseases arise from the misfolding of the same protein is poorly understood. Some possibilities include patient-specific cell-type vulnerabilities or different toxic protein strains, some of which may spread more easily between cells, shared biological mechanisms related to the specific protein that aggregates or glio-neuronal, neuroimmune, and neuroinflammatory processes.

A description of a shaking palsy in 6 individuals was first published in 1817 by the English physician James Parkinson and the hallmark lesion of PD (Lewy bodies) was defined in 1912 by a German-American neurologist (Engelhardt and Gomes, 2017). In 1997, the first gene (SNCA) associated with the disease was identified (Polymeropoulos et al., 1997) and the protein it encodes, α-synuclein, was identified to be the major component of the hallmark pathological cytoplasmic aggregates (Lewy bodies) in PD and Dementia with Lewy bodies (Spillantini et al., 1997). Soon after, in 1998, alpha-synuclein was identified as the main component of filamentous inclusions in multiple system atrophy (MSA) and PD, DLB, and MSA became collectively known as synucleinopathies (Spillantini et al., 1998).

Synucleinopathies are all late-onset neurodegenerative diseases, with the majority of cases showing no familial risk. Although the presence of α-syn inclusions is a common hallmark of

synucleinopathies, protein deposition is specific to each disease. Different neuroanatomical regions and cellular populations have distinct vulnerability to protein aggregation, cell dysfunction, and cell death, leading to phenotypic diversity. These diseases can be distinguished by the affected neuronal populations, cellular inclusions, and clinical presentation.

Parkinson's disease is the second most common neurodegenerative disease, after Alzheimer's, affecting approximately 2% of the population over age 60 and the most common movement disorder. It is characterized by the neuronal inclusions Lewy bodies and Lewy neurites, leading to loss of dopaminergic neurons in the substantia nigra (Fig. 1). Numerous genomic loci have been associated with the disease (Fig. 4).

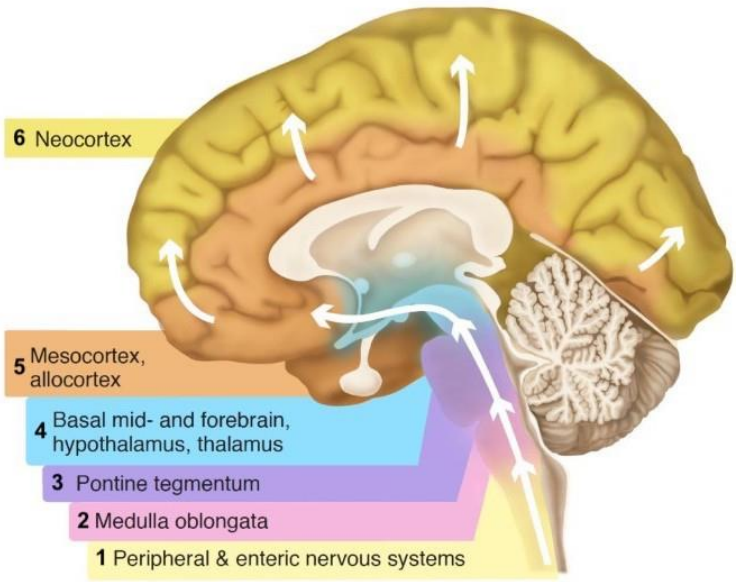


Figure 1: Parkinson's disease is characterized by temporal progression of alpha-synuclein accumulation and subsequent Lewy body pathology, divided into 6 Braak stages that extend through vulnerable regions, from the peripheral nervous system to the midbrain, leading to motor symptoms, to the neocortex at late disease stages, when cognitive symptoms appear. Image sourced from (Visanji et al., 2013).

Lewy body dementia is characterized by the same pathological hallmarks as PD (Lewy bodies and Lewy neurites) but can be clinically and pathologically similar to AD and PD and shares genetic risk factors with both neurodegenerative diseases (Geiger et al., 2016). Cognitive symptoms

typically develop more quickly than in PD after the onset of movement symptoms and appears to show decreased survival compared to AD (Price et al., 2017).

Multiple system atrophy is a rare, sporadic neurodegenerative disease with few identified genetic risks. It is characterized by autonomic, cerebellar, parkinsonian, and pyramidal symptoms with two clinical subtypes, parkinsonism (MSA-P) and cerebellar dysfunction (MSA-C), based on the predominant motor phenotype. The main histological hallmark consists of oligodendroglial cytoplasmic inclusions. Like PD and LBD, it is a late onset disease but progresses faster, with a mean survival time of 6-10 years after onset of symptoms (Fanciulli and Wenning, 2015). While MSA appears to affect males and females equally, risk of developing PD is 1.5-2 times higher for men (GBD 2016 Parkinson's Disease Collaborators, 2018) (Fig. 2).

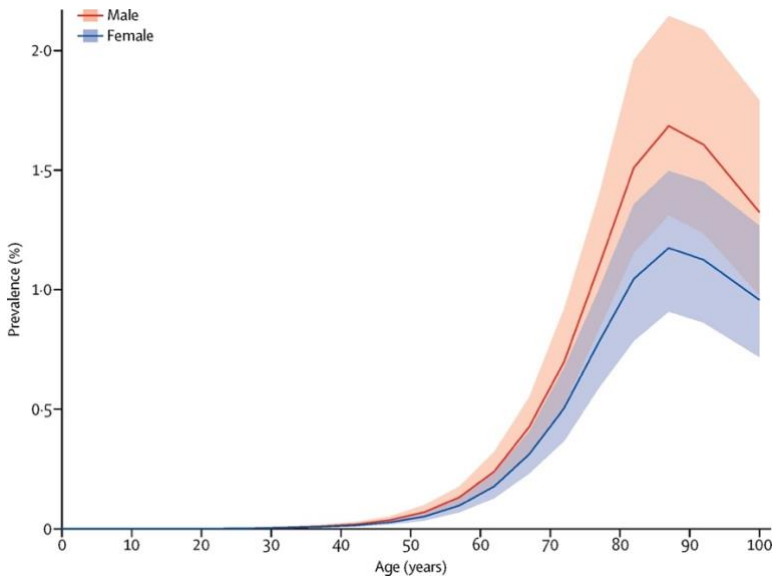


Figure 2: Global prevalence of Parkinson's disease by age and sex. Image sourced from (GBD 2016 Parkinson's Disease Collaborators, 2018).

As the most common and well-characterized synucleinopathy, PD will be the primary focus throughout.

1.3.3. GENETIC BASIS IN FAMILIAL AND SPORADIC DISEASE

The majority of synucleinopathies are sporadic, comprising 85% of Parkinson's cases and even more so in MSA, in which only a few percent of cases appear to be familial (Stefanova et al., 2009). Parkinson's disease is not a single diagnosis and includes a range of symptoms, complicating elucidation of genetic associations. Genetic risk factors identified in familial cases have been observed in the more common sporadic disease and have helped to uncover genetic risk factors. Monogenic forms of PD account for about 30% of familial and 3-5% of sporadic cases (Klein and Westenberger, 2012). Heritable synucleinopathies, even within the same family harboring the same mutation, can demonstrate different clinical diseases, from PD to PDD (Parkinson's disease dementia) to LBD (Lewy body dementia). Different clinical presentations in patients with identical primary mutations could stem from different environmental factors or genetic background modifiers that lead to either distinct alpha-synuclein conformational states or glioneuronal vulnerability patterns (or both).

Evidence of an underlying genetic component in Parkinson's disease was first published in 1990 following the identification of a large Italian family (known as the Contursi kindred) (Golbe et al., 1990). Genetic markers in a region on chromosome 4 showed linkage with the disease in the family and mapping of the first gene (SNCA) associated with the disease was reported in 1997 (Polymeropoulos et al., 1997) in both the Contursi and multiple Greek kindreds. A role for SNCA dosage in disease onset was later identified in heritable multiplications of SNCA (Singleton et al., 2003). The heterogeneity of the disorder was quickly apparent when subsequent studies identified additional genetic associations that were not in linkage.

Parkinson's disease is a genetically complex disorder for which no single gene mutation has 100% disease penetrance. Even in the case of highly penetrant rare variants, disease penetrance is likely affected by both additional genetic and non-genetic factors. Numerous genetic risk factors have been identified in familial and sporadic PD, with the largest genome-wide association (GWAS)

study to date identifying 90 independent genome-wide loci harboring rare and common variants, that latter of which explain 12-36% of heritability, depending on disease prevalence estimates (0.5-2%) (Nalls et al., 2019). Power estimates predict that increasing sample sizes to upwards of 99 000 cases will uncover additional risk variants (Fig. 3).

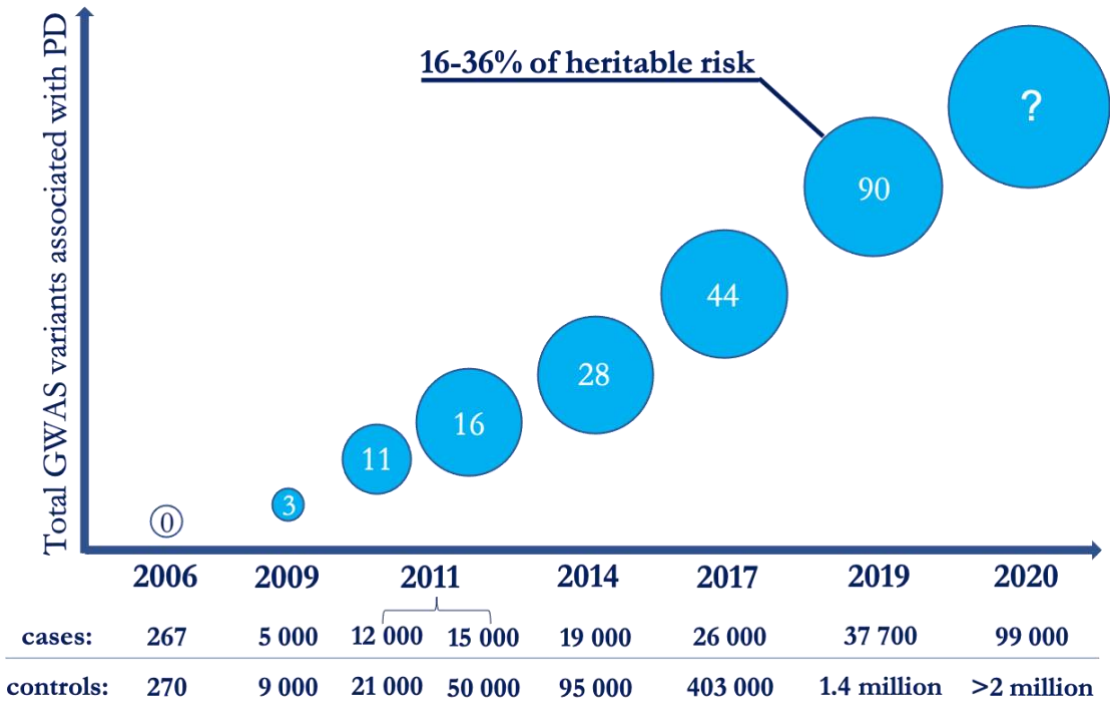


Figure 3: Since the start of the GWAS era in 2005, increasingly large case-control cohorts have identified an increasing number of genomic loci associated with PD. Common variants explain an estimated 16-36% of disease heritability. Image adapted from (Blauwendraat et al., 2020).

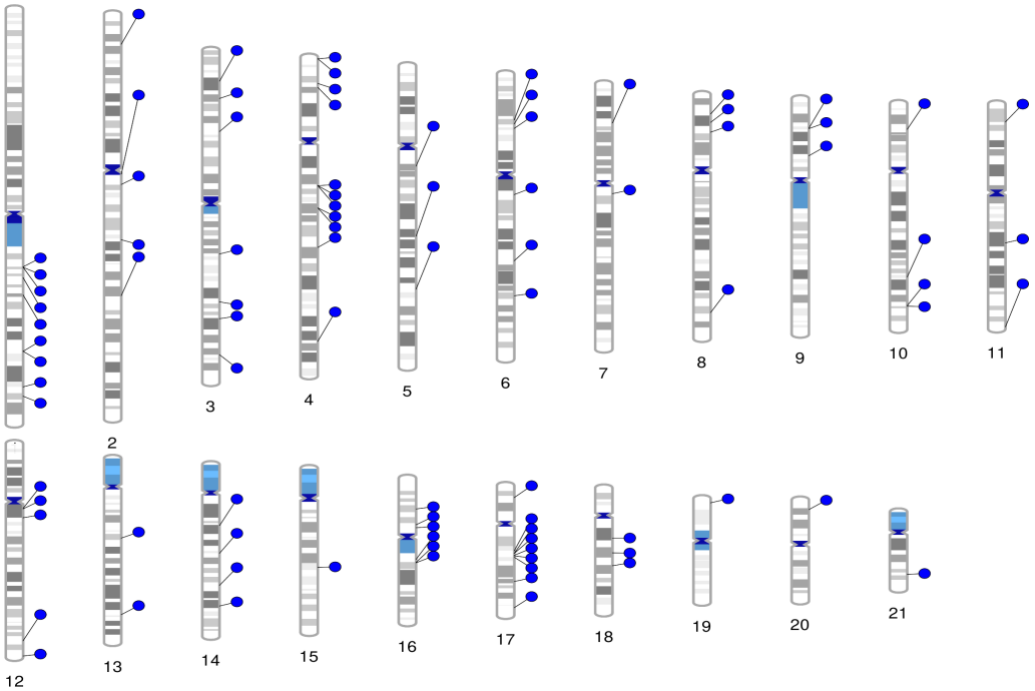


Figure 4: A phenogram plot of 90 independent GWAS loci implicated in Parkinson's disease (generated using the phenogram tool from the Ritchie lab at UPenn: visualization.ritchielab.org).

Known causal mutations (Table 2), defined as such by their occurrence in affected families or early-onset cases, appear to be rare in most PD cases. Mutations in SNCA, the gene encoding a-synuclein, are generally rare (several missense mutations as well as duplications and triplications of the entire gene have been reported) (Klein and Schlossmacher, 2006; Siddiqui et al., 2016). LRRK2 and PRKN/PARK2 are the most commonly mutated genes associated with familial PD and have been reported in 0.7% and 0.3% of individuals, respectively, with PD symptoms in epidemiological studies (Tran et al., 2020). LRRK2-PD has been shown to share clinical and pathological features of sporadic disease, with similar age of onset and disease duration (Henderson et al., 2019).

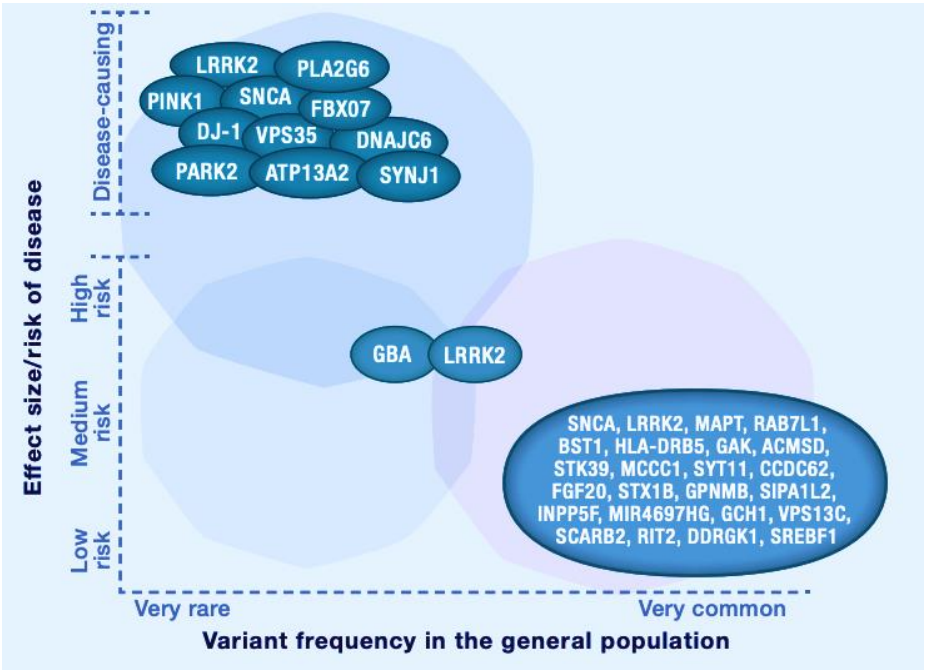


Figure 5: Genetics of PD, sourced from (Brás et al., 2015). The most common PD-associated variants tend to have the lowest disease penetrance.

A recent review found PRKN (exon rearrangements), LRRK2 (G2019S), GBA (L444P, N370S), and CHCHD2 (P2L) to be the most penetrant mutations, increasing the chances of developing PD symptoms by up to a factor 14, 10, 8 and 5, respectively. However, these variants are relatively rare and occur in ~4%, 2%, 5%, and 2% of PD patients, respectively, supporting increasing evidence that common PD mutations have lower penetrance than rare variants (Tran et al., 2020). Mutations in LRRK2 are common in early and late-onset PD and familial (4%) and sporadic disease (1%), with age-dependent risk increasing from 28% around age 60 to more than 70% around age 80 (Healy et al., 2008). The G2019S mutation is the most common but is highly variable across populations, from 0-3% in northern Europe and 2-14% in southern Europe, to 30-40% in North Africa. An international study of mutation frequency established a clear link between GBA mutations and PD (Sidransky et al., 2009). They observed variable phenotype but individuals carrying L444P and N370S mutations had earlier disease onset than non-carriers and were more likely to have a family history of PD. Mutations in PRKN cause about 15% of familial and 4% of sporadic PD cases with the onset before the age of 40 years (Konovalova et al., 2015).

gene	function	inheritance
CHCHD2	regulation of mitochondrial metabolism	AD
EIF4G1	regulation of mRNA translation	AD
GIGYF2	negative regulation of cell growth	AD
LRRK2	membrane trafficking, mitochondrial membrane maintenance	AD
SNCA	vesicle trafficking, chaperone, dopamine transmission	AD
UCHL1	ubiquitin-proteasome system and neuronal survival	AD
VPS35	membrane protein recycling	AD
GBA	lysosomal function	RF
MAPT	axonal microtubule stability	RF
ATP13A2	cation homeostasis, lysosomal function	AR
DJ-1	protection against oxidative stress	AR
DNAJC6	clathrin mediated endocytosis	AR
FBXO7	phosphorylation dependent ubiquitination	AR
HTRA2	neuroprotection	AR
PINK1	mitochondrial quality control	AR
PLA2G6	phospholipid remodelling, mitochondrial function	AR
PRKN	mitochondrial quality control	AR
SYNJ1	regulation of synaptic vesicle endocytosis	AR
VPS13C	maintenance of mitochondrial function	AR

Table 2: Functions and mode of inheritance (AD=autosomal dominant; AR=autosomal recessive; RF=risk factor) for disease genes identified in familial and early onset PD. Adapted from (Tran et al., 2020).

The identification of genes in familial and early-onset cases has increased our understanding of the likely pathogenic mechanisms resulting in disease. However causal mutations show age-dependent penetrance that is likely driven by additional genetic and environmental risk factors, as well as age-related pathobiology.

1.3.4. FROM GENETIC ASSOCIATION TO MOLECULAR MECHANISM

Genome-wide association studies have uncovered new loci associated with Parkinson's disease with increasing numbers of cases and controls, expanding our understating of the genes and pathways associated with disease risk. However, common variants tend to have small effects and polygenic risk scores based on the largest study to date would not be sufficient to predict individual risk (Nalls et al., 2019). A fundamental challenge in Parkinson's disease genetics is understanding the complex molecular mechanisms that drive the heterogeneity of disease onset and progression.

While we do not yet have a complete picture of the genetic background of PD, there is much to be done in establishing function and causality for the risk factors identified to date. Variants associated with rare, early-onset or familial forms of the disease have provided insight into disease processes, however, even for Mendelian loci, disease pathogenesis is still poorly understood.

Mutations in SNCA, the gene encoding the protein that forms the pathogenic protein aggregates, were first identified more than 20 years ago and have since been identified in both familial and sporadic PD patients. However, our understanding of their roles in PD pathogenesis is incomplete. Mutations in LRRK2, which functions as a GTPase and kinase, are the most common cause of late-onset autosomal dominant PD. Animal and cellular models have revealed a link between specific mutations and increased kinase activity, leading to degeneration of dopaminergic neurons (Nguyen et al., 2020). It is, however, unclear how mutant LRRK2 enzymatic activities contribute to neuronal toxicity. GBA mutations are the most common genetic risk factor in PD, and the functions of PD-associated mutations in the GBA gene are better established, with evidence supporting a role in autophagic and endolysosomal pathways. Pathogenic mutations are thought to diminish or reduce the activity of the encoded enzyme, glucocerebrosidase, which leads to accumulation of α-syn (Behl et al., 2021).

PD and other proteinopathies are complex, heterogenous diseases with genetic and environmental risk factors. Establishing a functional network of genes and gene regulatory elements associated with both monogenic and sporadic disease can help to identify the pathological mechanisms underlying disease-associated variants, which will in turn improve our understanding of genetic risk.

1.3.5. GENE EXPRESSION PROFILING IN NEURODEGENERATIVE DISEASE

Despite progress in uncovering genetic risk factors for Parkinson's disease, the molecular mechanisms underlying cellular degeneration in the brains of individuals with PD remain poorly

understood. Transcriptomics is increasingly utilized to interpret the functional impact of genetic associations (Cummings et al., 2017; Liu et al., 2014). This approach can be effective for splice variants, but additional methods are needed for disease-associated missense variants, which can often lead to perturbed protein-protein interactions rather than perturbed protein structure/function.

Gene expression studies in neurodegenerative disease are limited by small sample sizes, as these diseases are more common in aged individuals and disease-free brains are relatively rare. Further, results reflect perturbed expression after disease onset rather than changes that lead to PD. Despite many inherent challenges, transcriptomic analyses in PD, and other neurodegenerative diseases, have revealed genes involved in various pathways, including lysosomal and mitochondrial dysfunction, oxidative stress, neuroinflammation, and immune function.

Neurodegenerative diseases are defined in part by the affected brain region and brain tissue heterogeneity contributes to low overlap across expression studies, making validation challenging. Studies of PD brains often target the prefrontal cortex, which is not the primary brain region affected in PD, so expression changes may have a minimal effect. While the degeneration of dopaminergic neurons primarily occurs in the substantia nigra pars compacta, other brain areas, including the prefrontal cortex, do develop Lewy bodies (Dumitriu et al., 2012).

One example of the challenges of interpreting expression in diseased brains is the inconsistency of SNCA across studies. SNCA is abundant in the brain but high expression of alpha-synuclein in neurons is a risk factor for neurodegeneration, supported by the association of increased gene dosage with PD risk (Singleton et al., 2003). The susceptibility of different neuronal populations may depend on a-syn expression levels (Courte et al., 2020) and it has been suggested that low expression in pre-clinical regions, as defined by pre-symptomatic Braak stages (Braak et al., 2003) may increase vulnerability to Lewy body pathology (Kalia and Kalia, 2015). Increased copy number

and significant overexpression of SNCA has been observed in familial PD (Chiba-Falek et al., 2006; Singleton et al., 2003) but in both familial and sporadic disease, it is unclear if abnormal accumulation of alpha-synuclein is due to abnormal SNCA expression. SNCA has been found to be highly expressed in human microglia as well as peripheral monocytes of PD patients versus age-matched controls (Gardai et al., 2013; Keo et al., 2020) while other studies report decreased alpha-synuclein expression, notably in the substantia nigra, with overall lower expression in preclinical regions (brainstem) and highest in clinically involved regions (cortex and limbic system) (Dächsel et al., 2007; Keo et al., 2020).

The discrepancy in expression changes for known disease genes can likely be explained by clinical subtypes (ie. distinct underlying etiology), cell-type specific effects, and the cellular heterogeneity of brain tissues. As neurodegenerative diseases involve neuronal loss, expression studies must rely on post-mortem brain tissue, making sufficient sample sizes difficult to attain. While differential expression analysis is typically adjusted for post-mortem RNA degradation, brain tissue cellular heterogeneity has also been shown to be a confounder in expression analysis of bulk brain tissue (Nido et al., 2020). Given the evidence for cell-type specific expression, analysis of expression changes by brain region can help to elucidate pathogenic changes as the disease progresses as well as the characteristic changes in specific brain regions that define neurodegenerative disease (Courte et al., 2020; Keo et al., 2020; Taguchi et al., 2019).

1.3.6. BRIDGING THE GAP BETWEEN GENETIC MODIFIERS AND GENE EXPRESSION

Transcriptional responses appear to be distinct from genetic modifiers in alpha-synucleinopathies, a finding that is conserved from model organisms to humans (Lam et al., 2020). Genetic modifiers may be largely driving disease pathology whereas transcriptional profiling may be primarily capturing the response to proteotoxicity. In a comprehensive study in yeast, mRNA profiling and genetic screening was carried out to assess responses to genetic and chemical perturbations (eg.

DNA damage, ER stress, growth arrest). Differentially expressed genes and genetic modulators showed little overlap in response to these stressors, with genetic modifiers largely comprising regulatory proteins whereas differentially expressed genes were enriched for metabolic processes (Yeger-Lotem et al., 2009). This study applied the same approach to a yeast model of a-synuclein toxicity, confirming the distinct cellular pathways, with genetic modifiers biased towards vesicle trafficking proteins, and mRNA profiling revealed increased expression of genes with oxidoreductase activity and decreased expression of genes related to mitochondrial activity and ribosomal response to stress. To address the distinct cellular responses between genetic hits and transcriptional response, the authors harnessed the rich yeast interactome to develop a more comprehensive view of the response to cellular perturbation. They developed an algorithm (ResponseNet) to identify molecular-interaction pathways that connect hits from genetic screens and differentially expressed genes through known protein-protein interactions. When applied to a yeast model of a-syn toxicity, the major gene classes identified included vesicle-trafficking genes, kinases and phosphatases, ubiquitin-related proteins, transcriptional regulators, manganese transporters, and trehalose biosynthesis genes, providing a more systematic view of the incredibly wide range of cellular functions perturbed by a-syn expression (Yeger-Lotem et al., 2009).

Prioritizing PD associated genes ultimately requires integrative approaches to identify the molecular mechanisms and cellular responses underlying the more than 90 genomic loci associated with PD. Transcriptome-wide association studies (TWAS) measure the genetic association between gene expression and a complex phenotype by using genotypes (or GWAS summary statistics) to impute tissue-specific expression levels using a reference transcriptome panel (eg. GTEx) (Gusev et al., 2016). After estimating the component of gene expression explained by individual genotypes, expression is then correlated with the phenotype to identify etiological genes. (Li et al., 2019) integrated PD GWAS data from 13,708 cases and 95,282 controls and RNA-seq data from dorsolateral prefrontal cortex (DLPFC) and peripheral monocytes in order to identify tissues/cell

types implicated in PD. Both AD and PD-associated loci have been shown to be enriched in eQTLs from peripheral monocytes (Raj et al., 2014), demonstrating involvement of innate immune cells (peripheral monocytes and microglia). They identified associations between risk variant related gene expression and PD risk, including known AD risk genes in monocytes but not in cortex, supporting common immune-related risk factors in AD and PD. Interestingly, LRRK2 was only found to be significantly associated with PD in monocytes while there was significant enrichment of GWAS signals near genes expressed in the CNS for PD, with no such enrichment for AD, supporting the differential vulnerability of distinct cell types in neurodegenerative disease.

In addition to showing heterogeneous cell-type expression, single risk mutations typically pose low risk of developing neurodegenerative disease, necessitating approaches that model the complex disease mechanisms for multiple risk factors, especially for sporadic PD. The largest single-cell transcriptomic study in PD profiled human iPSC-derived dopaminergic neurons in response to both genetic and cytotoxic stressors (Fernandes et al., 2020). This in vitro disease model overcomes many of the aforementioned challenges of expression studies in neurodegenerative disease, namely cellular heterogeneity, and the need for post-mortem tissue. iPSC technology has the potential to revolutionize neurodegenerative disease research by recapitulating the complex interplay between multiple genetic and environmental risk factors and aging. Understanding the genetic predispositions and cellular phenotypes in disease onset and progression can contribute to the discovery and validation of targeted therapeutics.

1.3.7 ADDRESSING MISSING HERITABILITY IN PARKINSON'S DISEASE

Decades of genetic studies have been instrumental in guiding neurodegenerative research and have provided invaluable insights into pathogenic mechanisms and disease pathways. However, a large portion of genetic risk is unexplained. The heritable risk of PD due to common variants is estimated to be around 22% and power estimates suggest that increasing case numbers means that more risk variants are yet to be discovered (Keller et al., 2012; Nalls et al., 2019). These variants

typically have a modest effect size, complicating individual genetic risk prediction and our understanding of mechanistic roles in disease onset. There is increasing evidence supporting the role of rare variants in complex disease. As age-related diseases affect individuals in post-reproductive years, rare risk variants are more likely to persist in populations because they are not subject to selective pressure. A targeted approach of whole exome sequencing (WES) can identify rare coding variants not captured by GWAS and WES studies of modest sample size have successfully identified PD risk variants and genes (Jansen et al., 2017; Nuytemans et al., 2016). However, statistical tests to identify both rare and common variants associated with complex disease are generally underpowered. Novel approaches are needed in order to understand the biological basis of genetic risk and unravel the complex, polygenic mechanisms underlying disease onset and progression.

Genetic screens in model organisms or cellular models of proteotoxicity have identified homologs of known human genetic risk factors. Genome-wide screens in yeast synuclein models have recovered human orthologs of known and putative risk factors in PD, which have been validated in additional model organisms as well as induced pluripotent stem cell (iPSc)-derived neurons (Dhungel et al., 2015; Khurana et al., 2017). Human derived neurons can be generated directly from PD patients and controls and iPSC models of complex neurological disorders with limited heritability can facilitate the investigation of neuronal phenotypes *in vitro* (Burkhardt et al., 2013; Israel et al., 2012). These cross-species screens can be used to identify true modifiers in human disease, thus reducing the number of tests in genetic association studies. Cellular models with robust phenotypes can facilitate the characterization of the genetic drivers of Parkinson's and neurodegenerative disease in general.

1.3.8. CELLULAR MODELS OF PROTEOTOXICITY

Despite increasing understanding of genes and mutations implicated in neurodegenerative disease, no disease-modifying therapies exist. A treatment to prevent, slow, or stop disease progression or

reverse neuropathology is the single most important goal in neurodegenerative disease research. Although there have been several promising therapies, most merely alleviate certain symptoms. Development has been hindered by the lack of validated drug targets, largely due to uncertainty in the cause of cell death, as well as lack of an animal model that captures both disease etiology and progression.

It has become clear that human genetic analysis alone cannot fully establish causal associations between rare variants and disease phenotypes. A successful discovery platform requires integration of human and model organism-based approaches in order to identify genes, pathways, and compounds that can target α-synuclein related pathology. α-synuclein, the major constituent of PD-associated aggregates, is a large protein expressed primarily in the brain that interacts with lipid membranes and is involved in synaptic function. In the neurons of synucleinopathy patients, it adopts amyloid conformations, forming cytoplasmic protein aggregates called Lewy bodies (LB). Model systems have linked abnormal membrane interactions and alterations in vesicle trafficking with α-synuclein associated toxicity (Auluck et al., 2010).

The identification of protein misfolding and disease-causing mutations have been instrumental in creating tractable models of neurodegenerative disease. Unbiased genome-scale and candidate genetic screens have been successfully executed in model systems to identify genetic modulators of proteotoxicities. These systems, including yeast *Saccharomyces cerevisiae* (Khurana and Lindquist, 2010), the fruit fly *Drosophila melanogaster* (Shulman et al., 2003) and the nematode *Caenorhabditis elegans* (Teschendorf and Link, 2009) have uncovered disease-relevant biology, including homologs of known human disease risk factors.

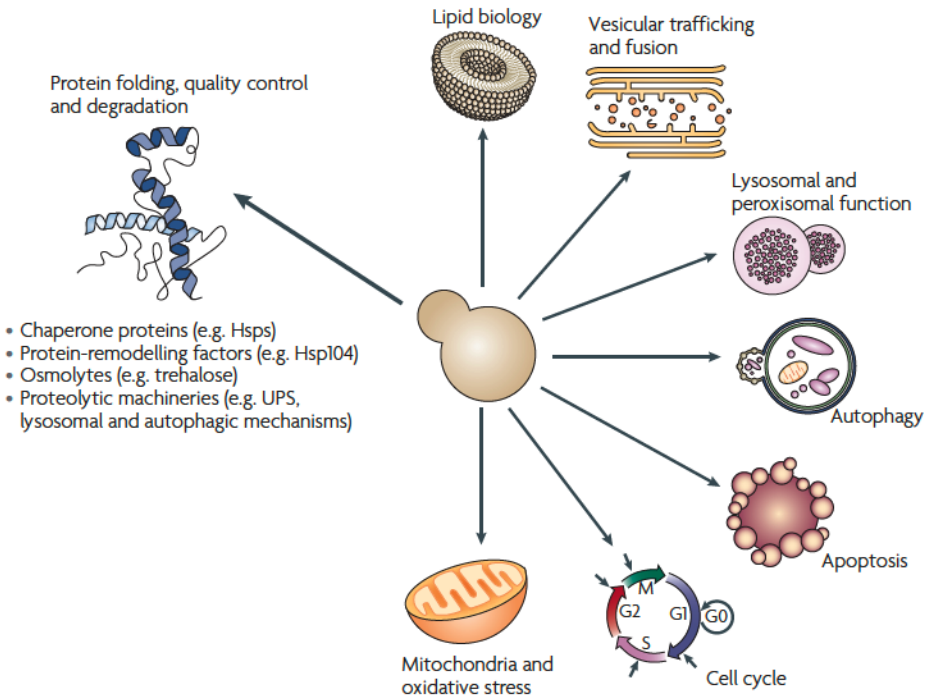
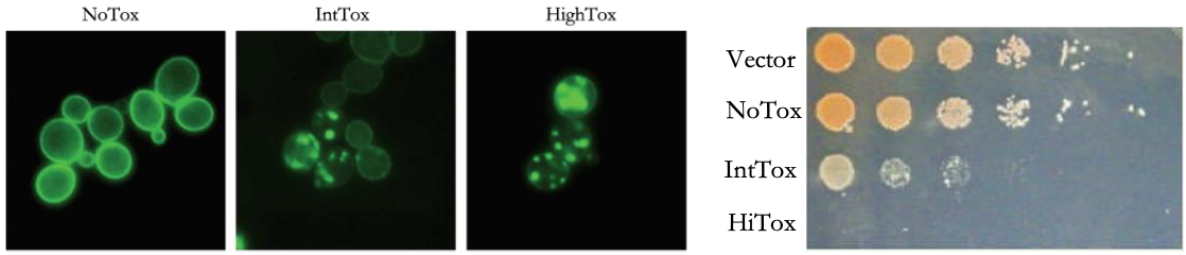


Figure 6: Cellular pathways related to neurodegeneration are conserved in yeast. Image sourced from (Khurana and Lindquist, 2010).

S. cerevisiae was the first eukaryote to have its 13Mb genome sequenced in 1996 and has since become one of the most thoroughly studied model organisms. Yeast is unicellular with a short generation time (1.5 - 3 hours) and its DNA is easily transformed, making it amenable to high-throughput genetic and small molecule screens and its reduced genetic redundancy facilitates effective gene knockout compared to more complex models (Khurana and Lindquist, 2010).

Human genetic data for neurodegenerative proteinopathies heavily implicate cellular pathways that are amongst the most highly conserved in eukaryotic evolution, including protein homeostasis and quality control, protein trafficking, RNA biology, and mitochondrial function (Brás et al., 2015; Guerreiro et al., 2015) (Fig. 6). The late onset of proteinopathies may be due to the impaired QC system in aging neurons, which cannot cope with protein misfolding and aggregation. Many of the processes associated with protein misfolding and aggregation can be modeled in yeast due to the high conservation of the cellular protein quality system. (Outeiro and Lindquist, 2003) developed a yeast model to exceed this quality control system by increasing a-syn expression. By tagging WT

and mutant (A53T or A30P) a-syn at the C terminus with a GFP reporter using a galactose inducible promoter, expression is induced when the yeast are transferred to galactose media. In cells with one copy of WT or A53T alpha-synuclein, the protein localized to the plasma membrane, whereas cells with 2 copies had cytoplasmic inclusions and reduced concentration at the membrane, as well as inhibited growth. Around the same time, (Singleton et al., 2003) reported a triplication of a-syn in early onset PD, supporting the a-syn expression model in disease related toxicity. Increasing the dosage of α-synuclein results in toxicity and the redistribution of α-syn from the surface of the yeast cell to cytoplasmic foci (Auluck et al., 2010). Serial dilution of spotting of yeast colonies shows decreased growth with higher a-syn levels (Fig. 7).



	NoTox	IntTox	HighTox
α-Syn localization	membranous	membranous/small foci	large foci
growth rate	normal	decreased	none
vesicle accumulation	mild	moderate	high
ER-to-golgi trafficking defects	absent	present	present
mitochondrial defects	none	low	high
lipid droplet accumulation	absent	rare	present

Figure 7: Localization of GFP tagged alpha synuclein in three yeast strains with low, intermediate, and high alpha-synuclein copy number (top left); adapted from (Auluck et al., 2010). A dilution spot assay (top right) shows decreased growth with higher a-syn levels (Khurana and Lindquist, 2010).

There are of course limitations to yeast models of neurodegeneration, notably, lack of specialized neuronal features. A genetic interaction within a unicellular yeast cell may play out differently in humans, depending on cell type and genetic background or disease context. However, the goal is not to model a human disease, rather the early cellular pathology caused by alpha-synuclein toxicity. Findings from high-throughput screens can be subsequently validated in neuronal model systems to account for the biological differences between yeast and complex human disease. The yeast

alpha-synuclein model recapitulates phenotypes detected in other model systems expressing human a-syn (Cooper et al., 2006; Feany and Bender, 2000; Masliah et al., 2000) and genes and small molecules identified in yeast have been validated in patient-derived neurons with the same cellular phenotypes (Chung et al., 2013).

1.3.8.1. GENOME-WIDE SCREENS FOR PROTEOTOXICITY IN YEAST

Expression of a-syn and a-beta in yeast leads to robust cytotoxicity and genetic screens in cellular models of proteotoxicity are scalable and easily scorable (Khurana and Lindquist, 2010). Each of the more than 5000 open reading frames (ORFs) in the yeast genome is overexpressed followed by assessment of growth and viability. In previous unbiased yeast genetic screens, genes encompassing ~85% of the yeast proteome were individually co-expressed with each of these toxic proteins to identify genetic modifiers of a-beta (Treusch et al., 2011) and a-syn (Yeger-Lotem et al., 2009) toxicity, respectively. These screens revealed both shared and distinct networks that were enriched in homologs of known human genetic risk factors for neurodegenerative disease and have facilitated the identification of novel human disease genes. Notably, there was no overlap between the networks of genes recovered in these two screens and many more AD and PD genetic risk factors were recovered than would be expected by chance (Khurana et al., 2017).

Given the high conservation of pathways implicated in neurodegenerative disease, human orthologs of these genes may represent additional risk factors in human genetic screens. In addition to evolutionarily conserved biological mechanisms, yeast has an incredibly rich interactome that can be exploited through the assignment of human homologs, essentially 'transposing' yeast molecular networks to augment the sparse human network. Although more than 60% of yeast genes have human homologs or at least conserved domains (Smith and Snyder, 2006), their cross-species assignment is challenging.

1.3.8.2. TRANSPOSING MOLECULAR NETWORKS ACROSS SPECIES

Consistent with human genetic analysis to date, model organism screens have found minimal overlap among genetic modifiers of proteinopathies (Khurana et al., 2017; Lam et al., 2020; Na et al., 2013). However, overlapping networks have emerged when combined with systems biology approaches that take into account protein-protein interactions (Haenig et al., 2020). Further, the mutually exclusive modifiers from genome-wide over-expression screens of a-beta (Treusch et al., 2011) and a-syn (Yeger-Lotem et al., 2009) proteotoxicity were enriched in known Mendelian genetic risk factors for AD and PD, respectively (Khurana et al., 2017).

While model organisms have been instrumental in functional genomics, assigning homology of both genes and networks can be challenging. Yeast genes frequently have many homologs in humans due to diversification in a multicellular organism. A computational approach, called TransposeNet (Khurana et al., 2017), was developed to map the yeast network of modifiers of alpha-synuclein toxicity from yeast to human based on sequence, protein structure, and known molecular interactions (Fig. 8). Yeast-human protein pairs are determined based on protein sequence similarity, evolutionary profiles and structural features to identify remote homologs, as well as protein function. Only homologs expressed in human brain tissue based on GTEx expression data are retained (GTEx Consortium, 2013). Known genetic and physical interactions between the human homologs are then used to create a network, which is augmented by the much richer yeast interactome. The final step applies the prize-collecting Steiner Forest algorithm to connect gene/protein nodes through the most confident protein-protein or genetic interactions as well as inserting predicted nodes (ie. genes not in the predicted homologs), augmenting the network by including genes with no clear yeast homologs (see (Khurana et al., 2017) for complete methods).

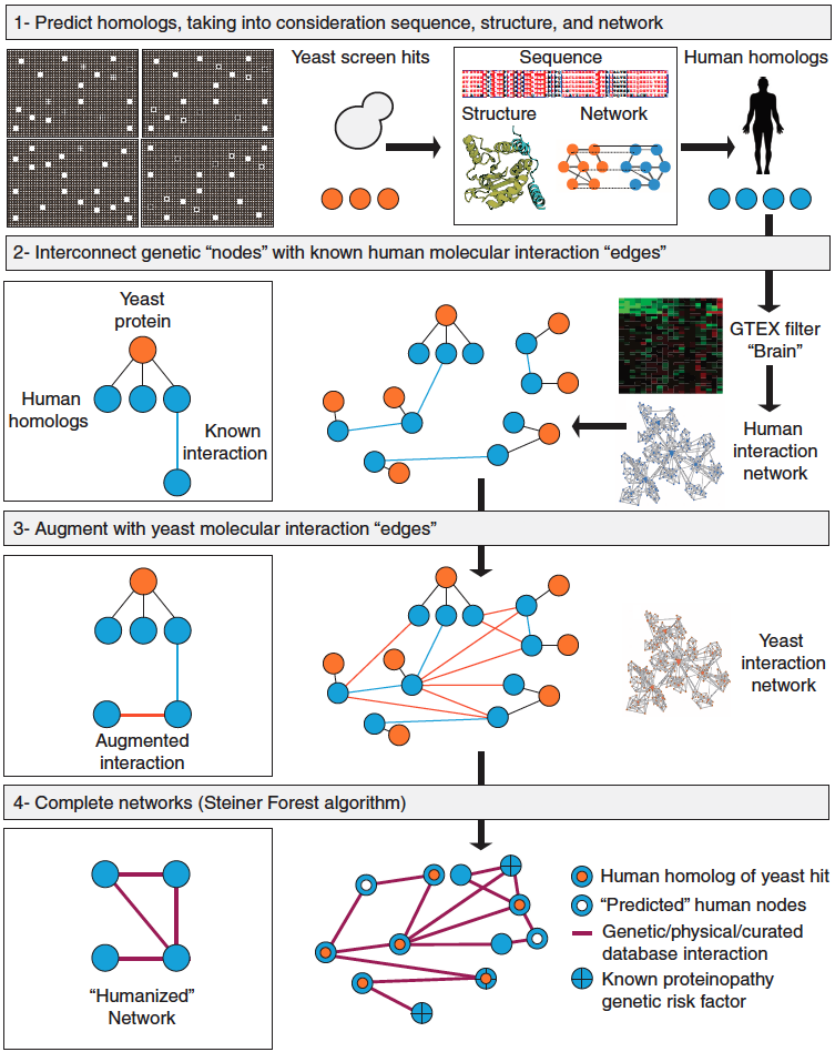


Figure 8: TransposeNet links PD-associated genes to alpha-synuclein biology through sequence, structure, and protein-protein interactions, which are then assembled into a network by exploiting the rich yeast protein interactome. Image sourced from (Lam et al., 2020).

The resulting a-syn toxicity network links parkinsonism genes and druggable targets to alpha-synuclein pathobiology. This approach identified 300 homologs of a-syn modifiers and 74 modifiers of a-beta toxicity. To directly test how these gene networks might be relevant in a human disease and patient-specific context, we targeted a wide range of orthologs of these genes in patients with synucleinopathies (PD, DLB and MSA). We performed targeted-exome sequencing of orthologs of the a-beta and a-syn modifier genes as well as known AD, PD and ataxia related disease genes. Our aim was to identify rare variants (minor allele frequency MAF <1%) in the a-beta and a-syn genetic networks to assess their relative contribution to disease risk and if there are

common genetic drivers across diverse synucleinopathies. We hypothesized that reducing the sequencing search space to a set of genes known *a priori* to be enriched in known risk factors and modifiers of alpha-synuclein toxicity would enable us to find meaningful signals even in a modestly sized cohort.

2. RARE VARIANT ANALYSIS OF SYNUCLEINOPATHIES

Nature is nowhere accustomed more openly to display her secret mysteries than in cases where she shows tracings of her workings apart from the beaten paths; nor is there any better way to advance the proper practice of medicine than to give our minds to the discovery of the usual law of nature, by careful investigation of cases of rarer forms of disease.

William Harvey

2.1. INTRODUCTION

Rare variants pose a major challenge for association studies of complex human conditions. Hypothesis free searches require an overwhelming number of individuals to reach the statistical power needed to associate these variants. Genome-wide association studies have identified more than 90 loci associated with Parkinson's disease that explain an estimated 16-36% of the heritability (Kunkle et al., 2019; Nalls et al., 2019). However, even biobank scale studies are underpowered to detect rare variants. In order to reduce the search space to detect rare variant signals, we focused on exonic variants in a targeted gene set of human homologs of proteotoxicity modifiers. Comparing our high-quality variant callset from alpha-synucleinopathy patients to a healthy, aged, cohort allowed us to prioritize these variants for subsequent functional validation, including characterization of multigenic risk and relevant perturbed pathways in patient-derived neurons.

Two distinct but complementary genetic approaches have been employed to understand the complex mechanisms driving neurodegenerative diseases: human genetic analysis and genetic analysis of tractable model systems. Using an established alpha-synuclein yeast model that recapitulates the molecular pathologies of Parkinson's disease, we narrowed the genome search

space for rare variants in neurodegenerative diseases by exploiting the underlying pathology: protein misfolding.

2.2. RESULTS

2.2.1. HUMAN GENETIC ANALYSIS

Human genetic data for neurodegenerative proteinopathies strongly implicate cellular pathways that are amongst the most highly conserved in eukaryotic evolution, including protein homeostasis and quality control, protein trafficking, RNA biology and mitochondrial function (Brás et al., 2015; Khurana and Lindquist, 2010). Genetic and small-molecule screen hits in yeast have uncovered early pathologic phenotypes in patient-derived neurons (Chung et al., 2013). Additional hits from yeast screens may represent genetic risk factors for human neurodegenerative disease.

When this study was conceived, targeted sequencing not only allowed us to refine the search space for genes of interest, but to cost-effectively sequence additional patients. Finding sufficiently well-matched control samples is a common challenge in disease association studies. We initially relied on summary data from the large-scale sequencing Genome Aggregation Database (gnomAD) (Karczewski et al., 2020). However, individual level data is necessary to perform adequate quality control, to detect population outliers and control for differences in sequencing and data analysis protocols. Burden testing using individual exome sequencing of cases combined with publicly available data can be achieved by pairwise comparison of sequencing quality scores (eg. Quality by Depth and Variant Quality Score Log-Odds) and using benign synonymous variants to improve variant filtering and control type I error (Guo et al., 2018). However, for individuals carrying multiple qualifying variants in the same gene, using the sum of allele counts for controls compared to the total number of individual cases leads to a more conservative test.

This QC based filtering did not improve the inflation (λ=2.1) seen in our analysis using summary count data from gnomAD (v2) non-Finnish European exomes, even after downsampling our high coverage callset (Methods). This approach nonetheless allowed us to refine our variant filtering and analysis until we were able to obtain sufficient control data from the Medical Genome Reference Bank (Pinese et al., 2020), a deeply phenotyped cohort of healthy individuals 70 years of age and older with no evidence of cancer, cardiovascular, or neurological disease. The MGRB is generally depleted of both common and rare disease-associated mutations. While more than 1% of these individuals carry clinical pathogenic mutations (including pathogenic GBA and LRRK2 mutations), a subset of these individuals was thoroughly reviewed and showed no subclinical manifestation of the associated disease.

2.2.2. GENE TARGETING RATIONALE

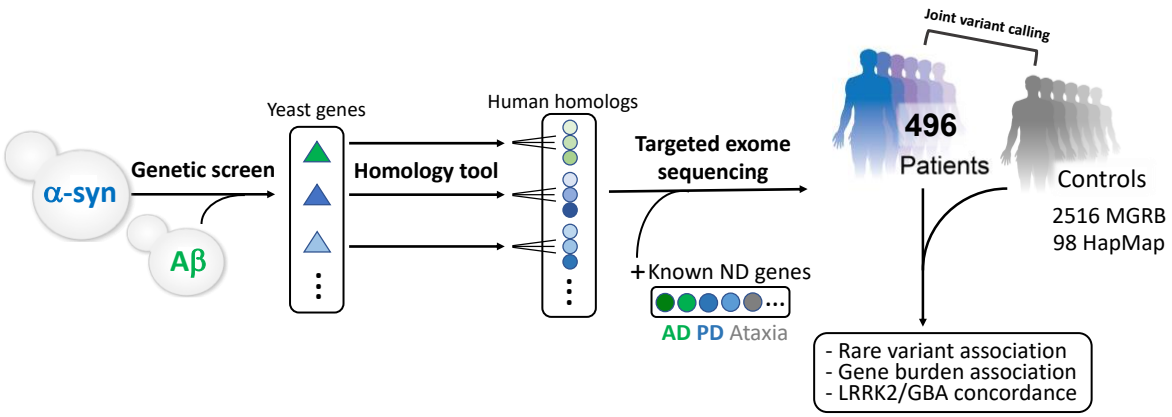
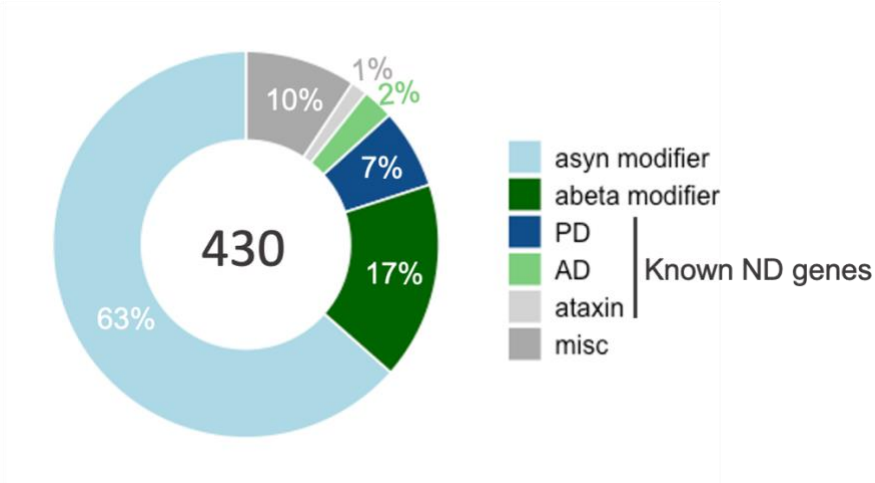


Figure 9: Study design. Genes uncovered in cellular proteotoxicity models were transposed to a human network and these human homologs were sequenced in a synucleinopathy patient cohort.

Beyond sequencing known genetic risk factors for AD and PD, the majority of targeted genes were chosen based on their implication in modifying cellular toxicity in yeast models of alpha-synuclein and beta-amyloid proteinopathies. The genes selected for targeted exome sequencing were compiled from three sources: predicted human homologs of yeast genetic modifiers of alpha-synuclein toxicity from our previous deletion and overexpression screens, human homologs of yeast genetic modifiers of amyloid-beta toxicity, and genes from previous genome-wide association

studies of neurodegenerative diseases. AD-associated genes and human homologs of yeast modifiers of a-beta toxicity were included in the screen due to evidence of concomitant a-beta pathology in DLB/PDD (Irwin et al., 2018; Siderowf et al., 2010). As the targeted genes were modifiers of proteotoxicities, we sequenced patients with a common proteinopathy, alpha-synucleinopathy, regardless of clinical diagnosis.



	a-syn	a-beta	total
enhancer	59	31	91
suppressor	164	32	197

Figure 10: Human homologs of yeast toxicity modifiers (131 yeast homologs) of a-synuclein or a-beta overexpression as well as known neurodegenerative disease (ND) genes were targeted in the exome sequencing.

2.2.3. DISCOVERY COHORT

We recruited a total of 506 patients with four distinct synucleinopathies: familial Parkinson's disease (GenePD cohort, Boston University); sporadic Parkinson's disease (Harvard Biomarker Study); pathologically confirmed synucleinopathies PD, LBD (Lewy body dementia), MSA (Multiple System Atrophy) (University of Pennsylvania Perelman School of Medicine) (Fig. 11). For cases with available age of onset (89%), the average was 64 years. The majority of samples (65%) were diagnosed clinically (ie. based on symptoms), however 35% of samples had definitive diagnoses based on histopathological assessment We also included 98 HapMap CEU samples that

were sequenced as part of the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) for the purpose of quality control of the targeted exome sequencing.

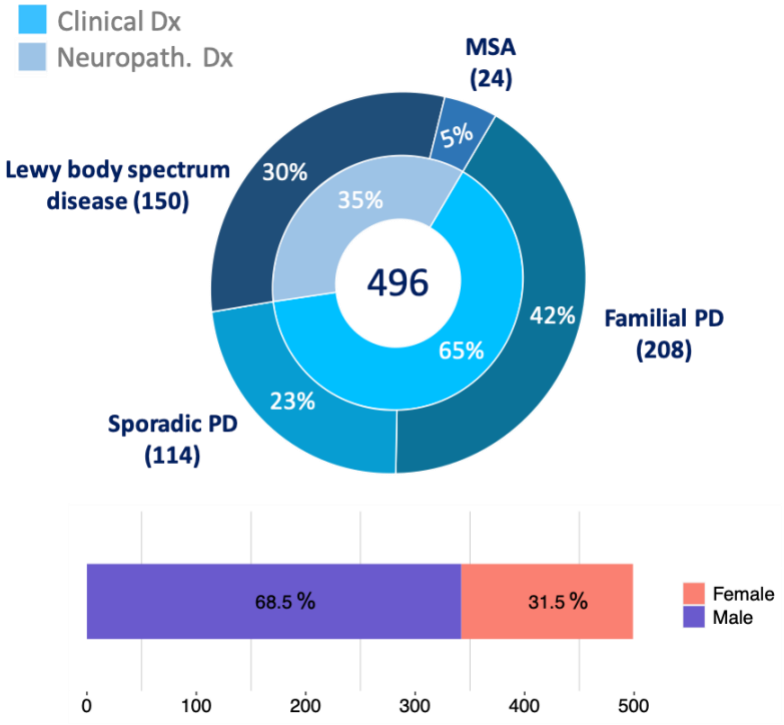


Figure 11: Breakdown of synucleinopathy cases by primary diagnosis in the derivation cohort.

Including appropriate control samples is critical in terms of both inclusion/exclusion criteria and sequencing platform, especially for complex age-related diseases. Including individuals who are at risk but do not yet display disease pathology as controls can reduce power in case-control association studies. Oversampling cases with early onset and aged controls is an efficient design, focusing on the most informative individuals. Including healthy aged controls has been shown to increase statistical power compared to age-matched cases and controls (Beecham et al., 2017). We obtained control genomes from 2570 healthy aged individuals from the Garvan Institute of Medical Research (Sydney, Australia) Medical Genome Reference Bank (MGRB). The MGRB control set consists of whole-genome sequencing data of elderly Australians who lived to at least 70 years of age with no history of cancer, cardiovascular disease, or dementia (1319 female, 1251 male). Age at the time of sample collection ranged from 64 to 95 years old (mean=78 years).

2.2.4. JOINT VARIANT CALLING WITH HEALTHY, AGED CONTROLS

A major challenge in genetic studies of diseases of aging is finding appropriate control subjects. Although the probability of a healthy control sample being affected by neurodegenerative disease is incredibly small, rare variant studies are already under-powered. We obtained individual data from 2570 healthy controls of European ancestry over the age of 70 with no history of major disease from the Medical Genome Reference Bank (Pinese et al., 2020). To reduce the false positive rate due to different sequencing approaches, genotypes were jointly called with our cases, internal HapMap controls, and MGRB samples (Methods). Joint analysis incorporates information across all samples, improving error detection and can increase sensitivity in low-coverage regions as genotype likelihoods across all samples can be used to inform genotype inference.

All cases and controls were reported to be of European ancestry, however, while it is challenging to adjust for population structure given our small (~1Mb) target region, we applied principal component analysis to the publicly available sequencing data for 1397 HapMap samples (1000 Genomes Project Consortium et al., 2015) and our cohort (cases, HapMap CEU, and MGRB controls). PCA performance depends highly on population structure and the underlying risk distribution and its utility in correcting for population stratification for rare-variant association tests is unclear. In such cases, PCA can be used to guide the matching of cases and controls. Less than 1000 SNPs remained after filtering rare variants (MAF<1%) and LD pruning, however all samples in our cohort clustered closely with the CEU (Utah residents with Northern and Western European Ancestry) and TSI (Toscani in Italia) HapMap populations and with the MGRB control samples. Principal component analysis based on WGS of the MGRB samples revealed 53 samples with potential non-European ancestry. Given the sensitivity of rare variant analysis to population structure, these samples were excluded from the analysis.

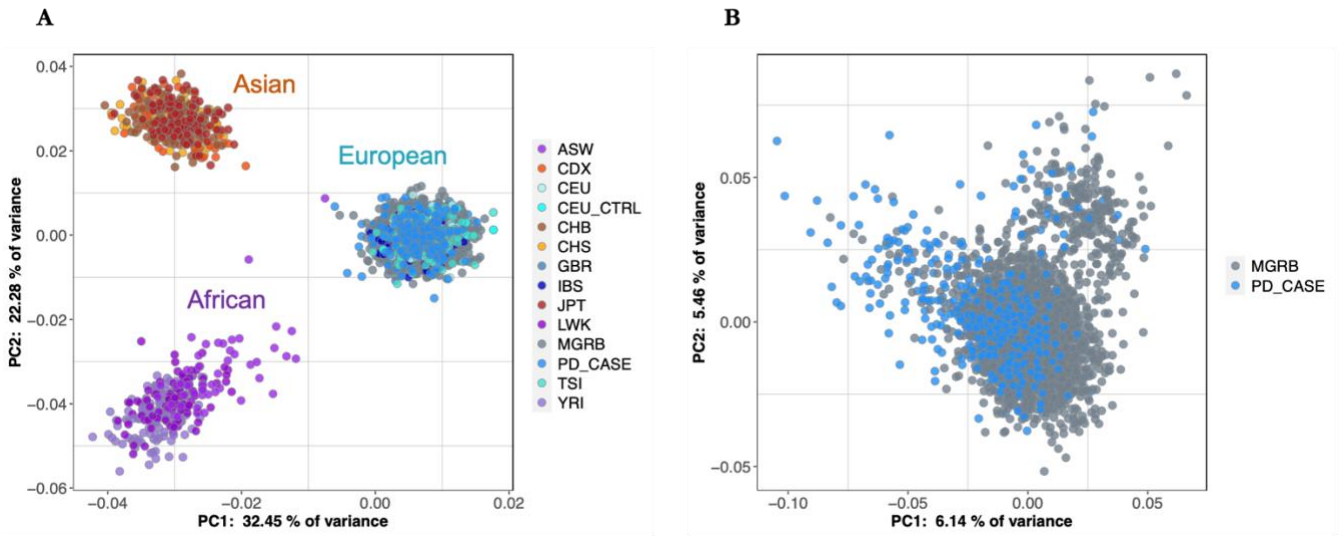


Figure 12: (A) Principal component analysis of cases, MGRB controls, internal CEU samples, and WGS data from the 1000 Genomes Project. (B) PCA of cases and MGRB controls.

2.2.5. VARIANT DETECTION AND FILTERING

The joint callset (496 cases, 98 CEU, and 2516 MGRB) included a total of 15070 variants after QC filtering of variants and samples (Methods), 75% of which were in gnomAD (v2) NFE non-neuro exomes and 81% in dbSNP build 150. The majority of variants not in gnomAD (~94%) were singletons.

Common and low frequency variants make up only ~10% and ~14%, respectively, of sites in the human genome while rare variants (MAF<0.5%) comprise ~76% (1000 Genomes Project Consortium et al., 2015). Improvements in sequencing technologies and computational methods have made rare variant analysis possible. Many of these methods focus on weighting, prioritizing or collapsing markers in order to overcome the lack of power due to limited sample sizes (Lee et al., 2014; Moutsianas et al., 2015). However, most standard quality control procedures for genome-based association studies were developed for whole genome sequencing, with a focus on common variants and have limited application for rare variants.

Joint calling of samples can overcome many of the technical biases when combining data from different studies or by testing for rare variant association by modelling sequencing reads without

genotype calling (Hu et al., 2016). (Chen et al., 2020) assessed the power to detect rare variants in deep coverage exome (~82X) and low coverage genome (~5X) sequence data and found that more than 97% of SNVs were detected by both single-study calling and the gold standard joint calling after applying similar quality control. Despite the extreme difference in depth of coverage in their analysis, they did not observe a loss of sensitivity and genotype accuracy in the individual compared to joint calling.

2.2.6. RARE VARIANT ASSOCIATION ANALYSIS

Single variant analysis using rare variant association tests are underpowered but can be useful when only summary statistics are available, as was the case in the early stages of this study. Further, many collapsing methods are not robust to the inclusion of neutral or protective variants, which can result in substantial power loss. We performed a one-sided Fisher's Exact test using allele counts in cases and MGRB healthy aged controls. The Cochran-Armitage (CA) trend test is commonly used in genome-wide association studies to compare differences in allele frequencies between cases and controls. However, Fisher's exact (FE) test can also be used to compare binomial proportions. While conservative, this test is robust to small sample sizes and guarantees type I error control. We performed both methods and the CA test resulted in 10 SNVs passing the Bonferroni threshold ($p<0.05$) and the FE test in 5 variants (Fig. 13), however, genomic inflation was well controlled for in the latter ($\lambda=1.12$ versus 1.97).

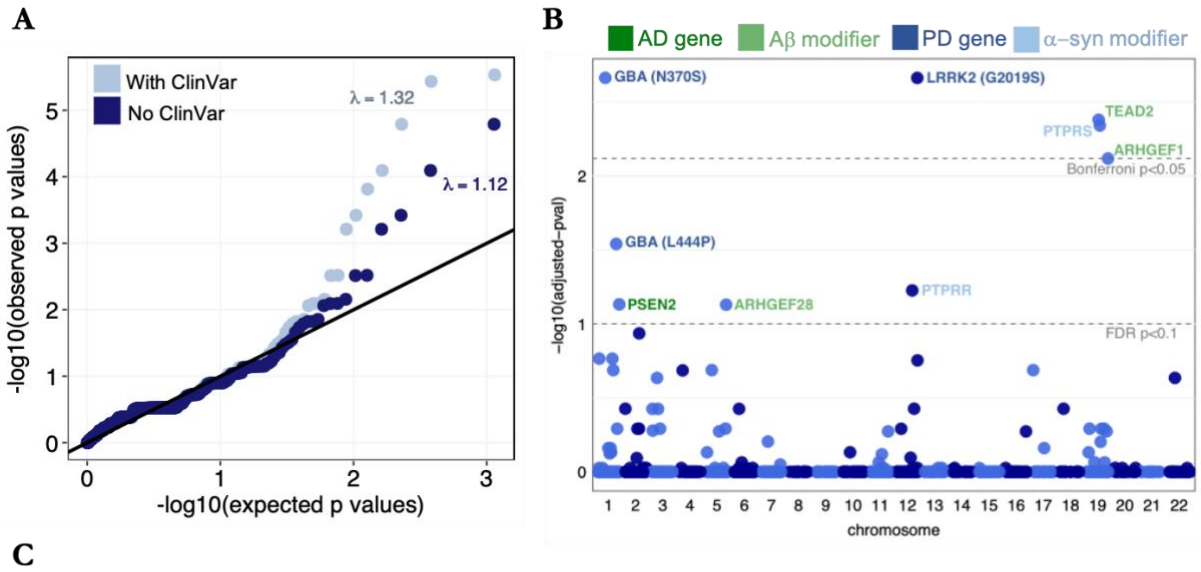
Inclusion of insertions and deletions (indels) generally leads to increased sequencing and variant-calling artefacts compared to SNVs and were excluded from the analysis. Autosomal, biallelic deleterious SNVs (missense, splice, stop gained, stop lost, start lost) with a combined MAF < 1% in cases, gnomAD non-Finnish non-neuro exomes, and MGRB controls were retained. When choosing a MAF threshold, it is reasonable to assume that a completely penetrant genetic variant should not be more common in the population than the disease that it causes. However, there is no completely penetrant causal variant in PD and lifetime risk increases with age. Establishing an

AF beyond which a variant could not credibly cause PD is challenging, especially since common variants have been demonstrated to play a substantial role in disease risk, explaining ~22% of the heritability (Nalls et al., 2019). Given the differences in sequencing technology between cases and controls, we could not include ultra-rare variants (eg. singletons). We therefore applied the standard 'rare' MAF threshold of 1% in cases, controls, and gnomAD non-Finnish non-neuro samples to assess population specific disease risk.

While rare coding variants are more likely to be deleterious, rarity is not sufficient to distinguish pathogenic from benign variants. The application of frequency-based filters should take into account disease prevalence, disease penetrance, heterogeneity, and inheritance (Whiffin et al., 2017). This approach is useful for Mendelian disorders but is challenging for complex, incompletely penetrant diseases. However, the maximum frequency of a known pathogenic variant in a large cohort, such as the Genome Aggregation Database (gnomAD), can help to estimate the maximum tolerated frequency. gnomAD v2 includes the majority of 1000 Genomes Project samples with exome sequencing data (including 80 of our 98 HapMap CEU internal controls), however it could not be confirmed if these CEU are included in non-neuro samples, defined as "samples from individuals who were not ascertained for having a neurological condition in a neurological case/control study", which were used for MAF filtering. Further, although all HapMap donors were healthy adults at the time of sample collection, this does not rule out the possibility that these individuals developed neurodegenerative disease, as is the case with most publicly available control data. However, their inclusion has the advantage of reducing technical artefacts for the tradeoff of a small probability of even 1 individual developing the disease and would have a modest effect on MAF filtering as they only comprise ~2% of our cohort.

Private singletons were excluded from all analyses as these variants are more likely to be artefacts. Pathogenic variants with no conflicting interpretations in the ClinVar database were exceptionally retained in order to assess genomic inflation. We performed a Fisher's exact test of allele counts

between cases and controls for predicted deleterious SNVs (Polyphen score >= 0.5 or CADD >= 10 for variants without a Polyphen score) and without this filter, as well as after excluding pathogenic ClinVar variants. The latter filter showed the greatest reduction in genomic inflation (Fig. 13A).



C

CHROM:POS:REF:ALT	rsID	Gene	OR	pval	Bonferroni
12:40734202:G:A	rs34637584	LRRK2	16.45	3.30E-06	0.00337
1:155205634:T:C	rs76763715	GBA	7.10	4.27E-06	0.00435
19:49852030:C:G	rs75100029	TEAD2	15.25	1.23E-05	0.01254
19:5240271:G:A	rs145504993	PTPRS	20.33	1.79E-05	0.01824
19:42392819:C:G	rs41303849	ARHGEF1	11.28	3.74E-05	0.03813

Figure 13: (A) Quantile-quantile plot of observed versus expected p-values from the variant-level Fisher's exact test for association between cases and MGRB controls and their respective genomic inflation. (B) Manhattan plot of FDR-adjusted (permutation-based case-control resampling) p-values from the variant-level association. Dashed lines indicate Bonferroni and FDR-adjusted significance thresholds. AD and PD associated significant genes are highlighted in green and blue, respectively. (C) Significant variants (Bonferroni p<0.05), that emerged from the cases v. MGRB association test.

Including all variants MAF<1% (except singletons) resulted in a highly conservative test with a p-value distribution peak at 1 of variants only found in MGRB. While it is reasonable to exclude these variants when testing each variant individually, this cannot be done when collapsing variants for gene-based tests. We thus retained only variants with AC>0 in both cases and MGRB in all analyses. Genomic inflation improved considerably (λ=1.12 compared to λ=1.89) and results for

top variants/genes were highly similar to variant and gene level tests that included all variants MAF<1%. A total of 1020 qualifying variants were tested, with 5 significant associations passing the stringent Bonferroni adjusted p-value and an additional 4 passing a modified FDR threshold.

As 3 of the top variants are all on chromosome 19, linkage between these variants was assessed, as this could affect the gene-based collapsing analysis (Methods: Linkage patterns). Although all appear to be in linkage equilibrium, this does not guarantee independence due to long-range LD.

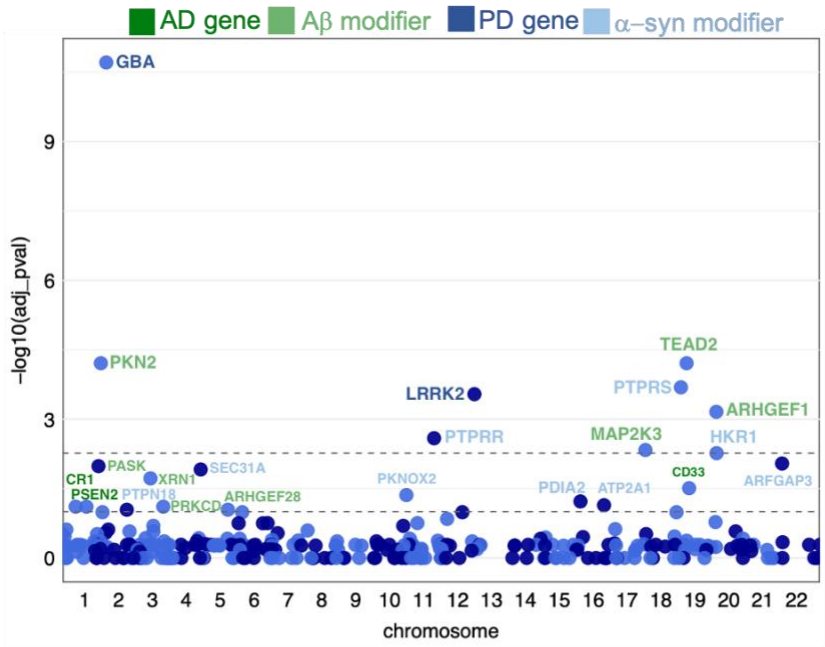
To control type I error, we performed permutation-based resampling to estimate the empirical null distribution (Methods: non-uniform p-value correction). Due to the discrete nature of rare variants and the non-uniformity of unadjusted p-values, we performed a modified FDR adjustment for multiple testing correction. For each variant or gene, the observed test statistic and p-values are calculated, with the adjusted p-value determined by calculating the number of times the observed statistic is smaller than the permuted p-value, divided by the total number of permutations.

2.2.7. GENE-BASED BURDEN ANALYSIS

Variant-level analyses are generally underpowered for rare variant analysis. Collapsing-based methods combine variants in a genomic region to increase statistical power when assessing their cumulative disease or trait association. Burden tests are more powerful when most variants in a region are causal and effects are in the same direction. SKAT is more powerful when a large fraction of the variants in a region are noncausal or the effects of causal variants are in different directions (Lee et al., 2012). Genes with at least one qualifying variant were compared between cases and MGRB controls using the gene-based sequence kernel association test - optimized (SKAT-O) with the linear weighted kernel and fixed imputation. We applied the SKAT-O method, a combined collapsing and variance component test, which is statistically efficient regardless of the direction and effect of the tested variants (Lee et al., 2012). This analysis was performed on variant

groups including all deleterious variants (missense, nonsense, and splice), moderate or high impact according to VEP, loss of function (nonsense and splice only), and missense only.

A



B

Gene	Description	Rationale for sequencing
GBA	Lysosomal hydrolase	PD risk gene
PKN2	Protein kinase N2	Homolog of Ab modifier PKC1 (OE=Enh)
TEAD2	Transcription factor, Hippo signaling	Homolog of Ab modifier Tec1 (OE=Supp)
PTPRS	Receptor tyrosine phosphatase, MAPK signaling	Homolog of a-syn modifier Ptp2 (OE=Supp)
LRRK2	Vesicle trafficking, actin cytoskeleton	PD risk gene
ARHGEF1	RhoA/cytoskeleton signaling	Homolog of Ab modifier Rom 1 (OE=Enh)
PTPRR	Receptor tyrosine phosphatase	Homolog of a-syn modifier PTP2 (OE=Supp)
MAP2K3	Protein kinase, MAPK signaling	Homolog of Ab modifier PBS2 (OE=Enh)
HKR1	Zinc finger protein	Homolog of a-syn modifier Tda9 (OE=Supp)

Figure 14: (A) Manhattan plot of SKAT-O deleterious variants (missense, nonsense, and splice) collapsed by gene (1020 variants collapsed to 293 genes). Dashed lines represent Bonferroni (upper) and modified FDR (lower) adjusted p-value thresholds. (B) Descriptions of significant genes (Bonferroni $p<0.05$). OE: Overexpression of amyloid-β or α-synuclein in yeast; Supp=toxicity suppressor; Enh=toxicity enhancer.

We applied additional methods, including a burden test (CAST) and adaptive burden tests (VT-test and KBAC). All procedures identified similar significant genes. Adaptive (rather than fixed) threshold tests can improve statistical power and reduce the influence of neutral and protective variants by using a permutation-based approach to determine the MAF below which variants are more likely to be causal (Price et al., 2010). This method was applied to validation in additional cohorts, as detailed in the following section.

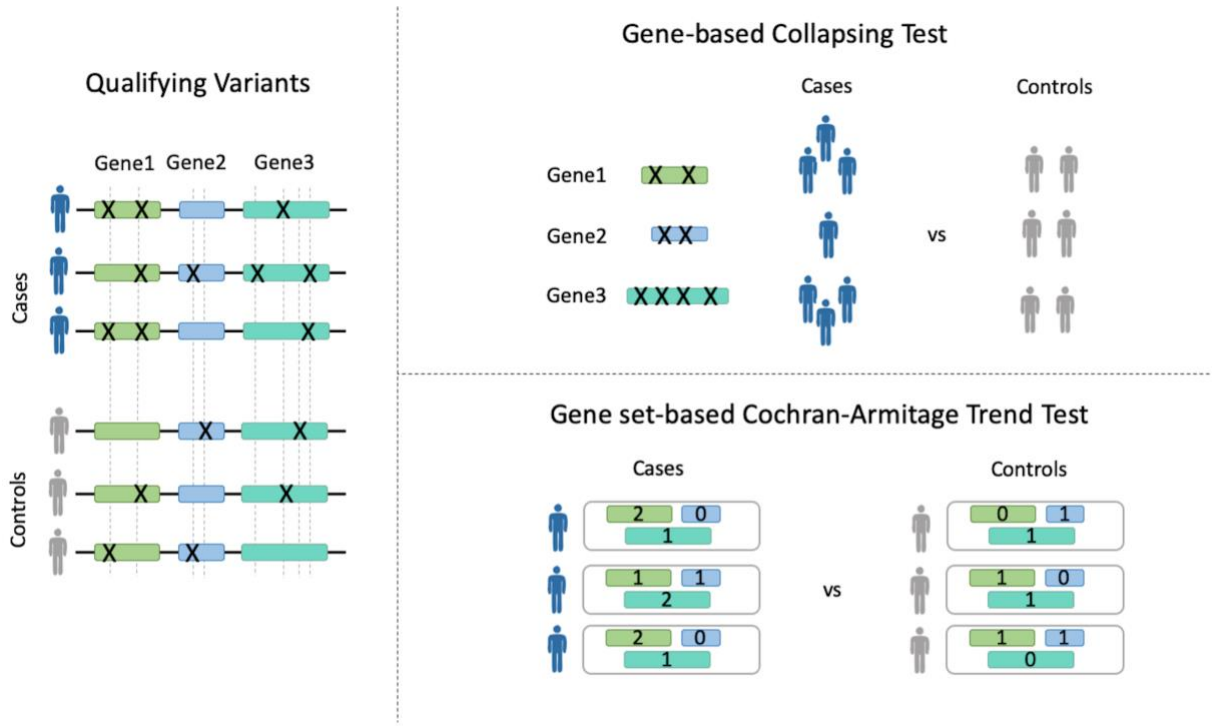


Figure 15: Rare variant association analysis. *Qualifying Variants:* Rare variants in each gene are identified in cases and controls. For variant-level association, a one-sided Fisher exact test was performed for each variant to determine if the variant is enriched in cases compared to controls. *Gene-based Collapsing Test:* The number of cases and controls harboring at least one variant in the gene of interest were counted (CMC-model) and compared using the SKAT-O test. *Gene set-based Cochran-Armitage Trend Test:* In our 'bag-of-genes' approach, we consider all the genes of interest as one single unit, and count the number of occurrences of variants in that unit for each individual. We then calculate the Cochran-Armitage statistic using the count data from two groups and determine the significance of increasing trend using a permutation-based approach.

We also applied a trend test for rare variant burden using a fixed threshold. For this method, qualifying rare variants (those with FDR < 0.1 across multiple tests) were chosen and variant counts were collapsed into a 2 X 1 contingency table for each mutation category (ie. all deleterious, missense, loss-of-function, synonymous). The observed Cochran-Armitage statistic (z-score) was then calculated. This test is often used in case-control association studies as it has more power than the chi-squared test when the suspected trend is correct. Row and column sums of the contingency table were fixed, case-control labels were permuted, and the z-score calculated M (10000) times. The permutation p-value is computed by counting the number of times (x) permutation z-scores >= observed z-score: (x+1)/(M+1).

Variants driving significant gene-based associations

Apart from MAP2K3, SKAT-O gene-based associations were driven by multiple variants. Of the top genes (Fig. 14), only 2 of the collapsed qualifying variants (TEAD2:19:49852030:C:G and PTPRS:19:5240271:G:A) passed the stringent Bonferroni threshold of $p < 0.05$ for the Cohort Allelic Sums Test when including or excluding ClinVar pathogenic variants (35 total variants). This demonstrates the increased power of the collapsing-based approach but also supports the utility of variant-level analysis.

2.2.8. ANALYSIS OF ADDITIONAL COHORTS

Lack of replication of putative variants and/or genes associated with complex genetic diseases is a challenge for both common and rare variants (Ioannidis et al., 2001). For rare variants, substantial recent genetic divergence of human populations means that validation studies within or between different populations are often not feasible and statistical power to identify rare variants decreases as the minor allele frequency decreases. However, replication of genotype–phenotype associations is a critical step in genetic discovery and in establishing causal links with the outcome of interest. To overcome some of the challenges inherent to rare variants, we used a variable threshold approach (Price et al., 2010) in which variants are collapsed using a Combined and Multivariate Collapsing (CMC) method for each gene by applying different MAF cutoffs and selecting the MAF that gives the smallest p-value as the optimal threshold for the gene in question. Current gene-based tests are not readily applicable to gene sets because they only test for the presence or absence of a variant, whereas most individuals are likely to carry at least one mutation in a set of genes simply by chance. Therefore, rather than considering the presence or absence of variants, we leveraged the frequency of occurrence of rare qualifying variants in our gene set of interest to assess association with disease. Our test, named RVTT (Rare Variant Trend Test), uses the Cochran-Armitage statistic (Armitage, 1955; Cochran, 1954) to assess the relationship between the increasing burden of rare qualifying variants with disease status. Variants were selected using a variable-

threshold approach (Price et al., 2010) and collapsed by pathway or network into combined gene sets. The significance of observing an increasing rare-variant burden is determined by a permutation-based approach.

	AMP-PD	MSA-jMorp
cases	1289	663
controls	694	8380

Table 3: Case-control cohorts used to validate our top signals from the gene-based burden analysis: AMP-PD (Accelerating Medicines Partnership: Parkinson's Disease) European cohort and Multiple System Atrophy in a Japanese cohort.

2.2.8.1. MULTIPLE SYSTEM ATROPHY COHORT

We tested for enrichment of rare variants in our targeted genes in exome sequencing data from a cohort of more than 600 MSA cases and healthy controls. Exome-sequencing of 663 MSA cases and 1226 control subjects was performed using Agilent's All Exon (v4+UTR or v5+UTR) capture followed by sequencing on an Illumina HiSeq2000. Paired-end reads were aligned to the human reference genome (hg19) with the Burrows-Wheeler Aligner software and variant calling was performed with SAMtools. Variants were annotated with SeattleSeq138, VEP (v95) (McLaren et al., 2016), and SnpEff (4.3t) (Cingolani et al., 2012) and filtered to include autosomal, predicted deleterious SNVs (missense, splice, stop gained, stop lost, start lost) with phred-scaled quality score > 20, and HWE p-value > 0.05 in controls, resulting in 6471 variants in 416 genes overlapping our target list.

We performed a one-sided Fisher's exact test for variants with AC>0 in cases as well as burden analysis of variants with AC>0 in cases or controls collapsed by gene between MSA cases and internal controls. No variants or genes survived multiple test correction. We then combined the filtered variant list with quality filtered autosomal SNVs from 8380 healthy controls from the jMorp Japanese Multi Omics Reference Panel (ToMMo 8.3KJPN Allele Frequency Panel v20200831)

(Tadaka et al., 2018). To select qualifying variants, we used the variable threshold approach (Price et al., 2010), collapsing rare non-synonymous and synonymous variants using a variable threshold for each gene by different in-cohort MAF cutoffs: 0.0005, 0.001, 0.005, and 0.01 and selected the MAF that gave the smallest p-value as the optimal threshold for each given gene. Inflation was well-controlled at the variant (λ=0.932) and gene level (λ=1.055). The SKAT-O test could not be performed as we did not have access to individual-level data for the jMorp controls. As cases and jMorp controls were not jointly called, the expectation of seeing a mutation in cases or in controls is not the same for our subset of genes and a gene-based burden analysis of summary counts is likely to amplify these differences. However, we used individual level data for cases and summary allele counts in controls as a proxy for the number of individuals, resulting in a more conservative test, as the sum of control allele counts is likely inflated due to individuals carrying multiple variants in the same genes. Qualifying variants were collapsed by gene for each significant gene based on our synucleinopathy cases versus MGRB controls. We performed a one-sided Fisher's exact test to compare the number of cases with at least one qualifying variant in each gene with the allele counts in jMorp controls. PTPRR and MAP2K3 were significant at an in-cohort MAF threshold of 1% and TEAD2 and PTPRS at a threshold of 0.1% (Table 4).

gene	number cases	total cases	number controls	total controls	threshold	pval
PTPRR	13	663	9	3294	0.01	8.33E-06
MAP2K3	19	663	32	3294	0.01	0.000328
TEAD2	5	663	1	3294	0.001	0.00067383
PTPRS	25	663	61	3294	0.001	0.00282778

Table 4: Top genes from the variable threshold test of significant genes from the synucleinopathy derivation cohort in MSA cases versus jMorp controls.

Qualifying rare variants in TEAD2 and PTPRS were found in 2 and 3, respectively, of the 24 MSA cases included in the derivation cohort. Qualifying variants in PTPRR and MAP2K3 were only observed in PD and LBD cases. Interestingly, 3 MSA cases were found to harbor rare GBA variants (2 with rs76763715 (N370S) and 1 with rs75548401). The association between GBA and MSA is

not clear with as many studies observing no association as those that have. More than 20% of autopsy-confirmed MSA cases were found to harbor GBA variants (Sklerov et al., 2016, 2017) and one of the largest studies to date (Mitsui et al., 2015) observed a strong association with MSA, although weaker than the association with PD.

2.2.8.2. INDEPENDENT PD CASE-CONTROL COHORTS

We validated the top hits from both the variant- and gene-level analysis performed on the targeted PD exomes vs. MGRB in two independent datasets: i) PPMI (Parkinson's Progression Markers Initiative) WGS, and ii) AMP-PD (Accelerating Medicines Partnership: Parkinson's Disease) WGS. These cohorts are part of large-scale collaborative initiatives to facilitate the identification of biomarkers and therapeutic targets, and include whole genome sequencing, transcriptomic, and harmonized clinical data. For both datasets we excluded any individual belonging to the genetic registry, genetic cohort, SWEDD, and prodromal categories. We included only Caucasian individuals in both cases. Additionally, we included only biallelic variants with a maximum missingness cutoff of 0.10. After filtering, the PPMI and AMP-PD datasets had 608 (PD: 433, control: 175) and 1983 (PD: 1289, control: 694) individuals respectively.

All variants with invalid reference and alternate alleles were excluded and coordinates were converted to genome build hg19/GRCh37 using picard LiftoverVcf (v2.18.17). Sites were filtered to include those overlapping the target region and annotated using the same tools as for our synucleinopathy cases and controls.

As indicated by the results of our initial variant- and gene-level analyses, there are only a few rare variants with moderate effect size implicated in PD, and those rare variants often do not collapse into the same few genes. With current available sample sizes, variant- and gene-level association tests are likely to be underpowered. Unsurprisingly, we did not observe any significant hits in the AMP-PD cohort at either the variant- or gene-level. This motivated us to develop a new statistical

test that aggregates variants at the network/pathway-level instead of a single gene or genomic region to increase power (Fig. 15).

We validated the top hit genes from our targeted exome analysis in the AMP-PD cohort using RVTT: i) genes containing significant rare missense variants (Fisher's exact test) excluding ClinVar and functional effect variants, namely, TEAD2, PTPRS, ARHGEF1, PTPRR; ii) genes containing significant rare missense variants (Fisher's exact test) including ClinVar variants, namely, LRRK2, GBA, TEAD2, and PTPRS; and iii) top hit genes (using non-synonymous variants) from the SKAT-O test, namely, GBA, TEAD2, PKN2, PTPRS, LRRK2, ARHGEF1, PTPRR, MAP2K3, and HKR1. In each case, we observed a significant increasing trend in rare missense and damaging missense variants (adjusted p-value < 0.1, 10000 permutations) in the gene set of interest in PD cases compared to controls. As a control experiment, we tested for enrichment of synonymous variants in the same gene sets. We observed no significant trend in synonymous variants in cases compared to controls. We also performed RVTT on 100 random gene sets of length four to ten and observed no significant trends in missense and damaging missense variants.

Rare variant trend test *(carried out by Sumaiya Nazeen, PhD)*

As we did not find any significant hits when testing individual variants and after collapsing by gene, we assessed the burden of rare qualifying variants in gene sets of interest using the rare variant trend test (RVTT) based on the Cochran-Armitage statistic (Armitage, 1955; Cochran, 1954) to assess the relationship between the increasing burden of qualifying rare variants in our gene set of interest with disease status. The burden of qualifying rare variants is treated as an ordinal variable and the disease status as a binary variable. We then count the number of occurrences of rare qualifying variants in the gene set of interest at the optimal in-cohort frequency cutoff selected using the variable threshold (VT) approach (Price et al., 2010). The null hypothesis assumes that there is no linear trend in binomial proportions of disease status across an increasing number of qualifying rare variants in the gene set of interest. The significance of observing a trend is determined by a permutation-based approach. Single gene-based collapsing tests are often under-

powered to detect the presence of a trend in the rare-variant space; a gene set or 'bag-of-genes' approach increases the possibility of seeing an increasing number of qualifying variants and thereby detecting any trend.

Validation of variant-level results

We first tested genes harboring significant non-synonymous hits excluding ClinVar and functional effect variants, namely, TEAD2, PTPRS, ARHGEF1, PTPRR. We identified the presence of a significant trend in rare missense variants (including variants with predicted damaging, neutral, and unknown functional consequences) in this gene set in both the PPMI and AMP-PD datasets. No significant trend in the synonymous variants was observed in the same gene set (Table 5).

cohort	type	threshold	z-score	pval
PPMI	missense	0.035	2.2414	0.0550
	synonymous	0.002	0.6995	0.7227
AMP-PD	missense	0.037	2.5469	0.0292
	synonymous	0.015	0.4693	0.8897

Table 5: Rare variant trend test on top hit genes from the variant-level Fisher's exact test on the targeted exome sequencing dataset of PD cases and MGRB controls (in-cohort optimal threshold, 1000 permutations).

We repeated our rare-variant trend test using the genes from top missense variants (FDR adjusted $p < 0.05$) list including the ClinVar variants (LRRK2, GBA, TEAD2, and PTPRS). In the AMP-PD dataset, we observed a significant enrichment of missense variants in cases compared to controls. Although the strength of the association was weak in the PPMI data, we did observe a similar trend in this dataset (Table 6).

cohort	type	threshold	z-score	pval
PPMI	missense	0.001	1.6560	0.2369
	synonymous	0.004	0.6254	0.7510
AMP-PD	missense	0.022	3.1557	0.0058
	synonymous	0.017	0.8988	0.5536

Table 6: Rare variant trend test on top hit genes from the variant-level Fisher's exact test on whole genome sequencing PPMI and AMP-PD datasets (10000 permutations).

Validation of gene-level results from the SKAT-O test

We performed RVTT to assess the burden in the gene set consisting of significant hits detected in the gene-based SKAT-O test of deleterious variants (Bonferroni-corrected p-value < 0.05), namely, GBA, PKN2, TEAD2, PTPRS, LRRK2, ARHGEF1, PTPRR, MAP2K3, and HKR1. Our results showed the presence of a significant trend in damaging variant (predicted as possibly/probably damaging by Polyphen2 or deleterious by SIFT) burden in the AMP-PD cases in the gene set (Table 7). We did not observe such a trend in the PPMI dataset. One possible reason could be the presence of a large number of GBA carriers in AMP-PD cases.

Cohort	Type	threshold	z-score	pval
AMP-PD	Damaging missense	0.008	2.4537	0.0390
	Missense	0.022	2.2988	0.0707
	Synonymous	0.030	1.3705	0.3614
PPMI	Damaging missense	0.032	1.1623	0.4362
	Missense	0.035	1.4235	0.3404
	Synonymous	0.002	0.9134	0.6295

Table 7: Rare variant trend test results for the PPMI and AMP-PD cohorts based on top hit genes from the gene-based SKAT-O test in the synucleinopathy derivation cohort.

We repeated the test for genes that were significant in the gene-based SKAT-O test of loss-of-function variants (Bonferroni-corrected p-value < 0.05), namely, GBA, PTPN1, and ARHGEF1 in the PPMI and AMP-PD cohorts. Our results showed the presence of a significant trend in

missense and damaging missense variant burden in this gene set in AMP-PD data (Table 8). Although not significant, a similar trend was observed in the smaller PPMI dataset.

cohort	type	threshold	z-score	pval
AMP-PD	Damaging missense	0.008	2.5022	0.0206
	Missense	0.022	3.4703	0.0017
	Synonymous	0.003	1.8545	0.1099
PPMI	Damaging missense	0.002	1.6918	0.1105
	Missense	0.005	1.6399	0.2222
	Synonymous	0.005	1.7791	0.8204

Table 8: Rare variant trend test on top hit genes from the gene-based SKAT-O test of LoF variants collapsed by gene in the derivation cohort.

2.2.8.3. EXPRESSION PROFILING OF POST-MORTEM BRAIN SAMPLES

A subset of 28 cases that were sequenced in our targeted exome analysis included RNA sequencing of prefrontal cortex samples obtained from familial PD patients (GEO accession GSE68719). Given the overlap of significant associations of genetic variants and associated genes across disease boundaries, we explored shared gene expression patterns in AD, PD, and MSA brains. These samples, along with 50 healthy control brains (not a part of our control cohort) represent the largest single cohort of PD brain RNA-seq cohort published to date (Dumitriu et al., 2016; Labadorf et al., 2017). All brain samples were obtained from males of European ancestry. Cases showing indication of AD pathology based on the assessment of neuropathology data of plaques and tangles (grading of plaques and Braak stage) were excluded.

Differential expression analysis was conducted using DESeq2 (Love et al., 2014). Briefly, counts are modeled using a Poisson-like negative binomial distribution followed by a Wald test to compare expression between cases and controls. Only genes that were also targeted in the exome sequencing analysis were retained in the analysis (after converting discordant gene aliases 418/430 genes were

detected in the RNA-seq dataset). Significant differences were observed for age at death (t-test p=0.00164) and RNA integrity (Mann-Whitney Exact test p=5.56e-5) between cases and controls. Significant differences were not observed (Mann-Whitney U test p=0.14) for post-mortem interval (PMI), defined as the total hours between death and tissue sample collection. However, PMI-related mRNA degradation can be gene or tissue-specific and genotype-dependent (Zhu et al., 2017). Further, increasing PMI has been found to affect SNCA expression levels in PD cases but not in controls (Dumitriu et al., 2012). These 3 potential confounding variables (age at death, RIN, and PMI) were included in the regression model for each gene. RIN was categorized to <=7 or >7 as 8 is considered high quality, and binning the data reflects the qualitative nature of the score. Two control samples missing RIN data were excluded from the analysis. Raw p-values were corrected for multiple testing using the Benjamini-Hochberg procedure and genes passing an FDR cutoff of 0.05 were considered significant (41/418). 35 of these DEGs from the targeted enrichment analysis were also significant (q<0.05) in the transcriptome-wide analysis.

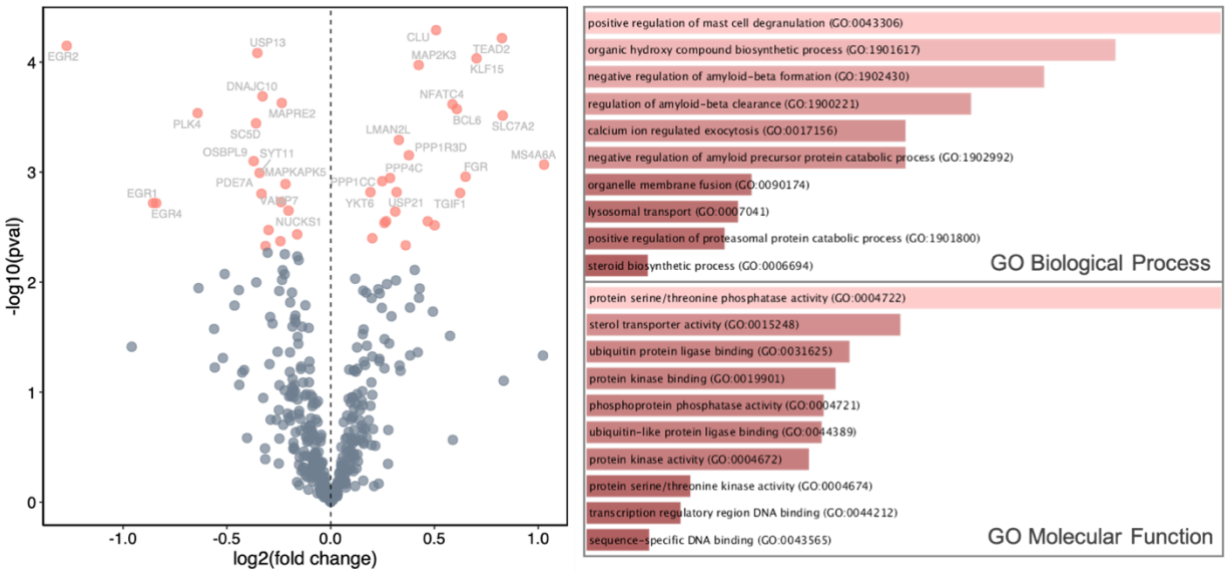


Figure 16: Left panel: Volcano plot of differential expression analysis of our gene set in 28 familial PD cases compared to 50 pathologically normal brains (top 30 significant genes q<0.05 are labeled). Right panel: Top GO terms for the 35 DEGs that were also significant transcriptome-wide.

Gene ontology (GO) enrichment analysis was performed with GOseq (Young et al., 2010) using the Wallenius method to calculate a probability weighting function to quantify the probability of a gene being detected as DE as a function of its transcript length. No significant GO enrichment terms were observed when testing the top DEGs (n=41) with our gene set (n=418) as the background. We then performed enrichment analysis for 1) the 35 DEGs significant in both our targeted differential expression analysis and the transcriptome-wide analysis (Fig. 16) and 2) all transcriptome-wide genes passing the stringent Bonferroni adjusted threshold <0.05 (n=47) using Enrichr (Kuleshov et al., 2016) (Table 9).

GO Biological Process	GO term	q-value
response to unfolded protein	GO:0006986	3.48E-07
chaperone mediated protein folding requiring cofactor	GO:0051085	0.00006923
'de novo' posttranslational protein folding	GO:0051084	0.0001178
negative regulation of transcription from RNA pol II promoter in response to stress	GO:0097201	0.0001454
negative regulation of inclusion body assembly	GO:0090084	0.0001454
regulation of inclusion body assembly	GO:0090083	0.0001664
cellular response to heat	GO:0034605	0.0003192
cellular response to unfolded protein	GO:0034620	0.001144
cellular response to topologically incorrect protein	GO:0035967	0.001723
regulation of cellular response to heat	GO:1900034	0.001723
regulation of microtubule nucleation	GO:0010968	0.003693
regulation of cellular response to stress	GO:0080135	0.004352
positive regulation of TNF-mediated signaling pathway	GO:1903265	0.004368
amino acid transport	GO:0006865	0.007032
carboxylic acid transport	GO:0046942	0.007371
positive regulation of NF-kappaB transcription factor activity	GO:0051092	0.007371
negative regulation of cellular component organization	GO:0051129	0.008223
regulation of protein modification by small protein conjugation or removal	GO:1903320	0.01331
purine ribonucleotide metabolic process	GO:0009150	0.01331

Table 9: Top 20 terms from enrichment analysis of transcriptome-wide significant differentially expressed genes.

GO Biological Processes including protein folding, heat shock/stress response, and protein aggregation were significantly enriched as well as WikiPathways (Martens et al., 2021) NRF2 pathway (WP2884), MAPK signaling (WP382), and Parkin-Ubiquitin Proteasomal System pathway (WP239). MAPK signaling and UPS dysfunction have been implicated in both AD and PD and

NRF2, which regulates antioxidant signaling, is an emerging therapeutic target in neurodegenerative disease (Cuadrado et al., 2019; Gureev and Popov, 2019). Enrichment analysis of the significant 35 DEGs overlapping our gene set recapitulate the functions of the top genes in our exome sequencing analysis, notably amyloid-beta and protein serine-threonine kinase and phosphatase activity. Results from the transcriptome-wide analysis included more general terms related to heat shock response and protein folding.

Two of the significant DEGs were also significant in the exome sequencing gene-based kernel association test. TEAD2 and MAP2K3, which passed the more stringent Bonferroni correction ($p<0.05$), were identified as human homologs of a yeast suppressor and enhancer, respectively, of amyloid beta toxicity in the a-beta overexpression screen and were both more highly expressed in PD cases. None of the 28 cases overlapping the RNA-sequencing data harbored rare variants (deleterious or synonymous) in TEAD2 or MAP2K3. Twelve of these cases harbored a total of 18 rare (MAF<1%) heterozygous mutations (16 missense and 2 splice region) in 12 of the significantly differentially expressed genes. Three cases harbored the same missense mutation in OSBPL9 (rs61739207) which showed significantly lower expression in cases ($q= 0.0208$).

These results are difficult to interpret as only 28 of 496 cases included RNA sequencing data and non-overlapping control samples. Gene-based associations were driven largely by missense variants and the majority of missense disease mutations do not appear to impair protein folding or stability (Sahni et al., 2013). However, it is interesting that, as in our exome-sequencing cohort, associations were independent of concomitant AD pathology, as cases displaying AD pathology were excluded from the RNA-seq analysis.

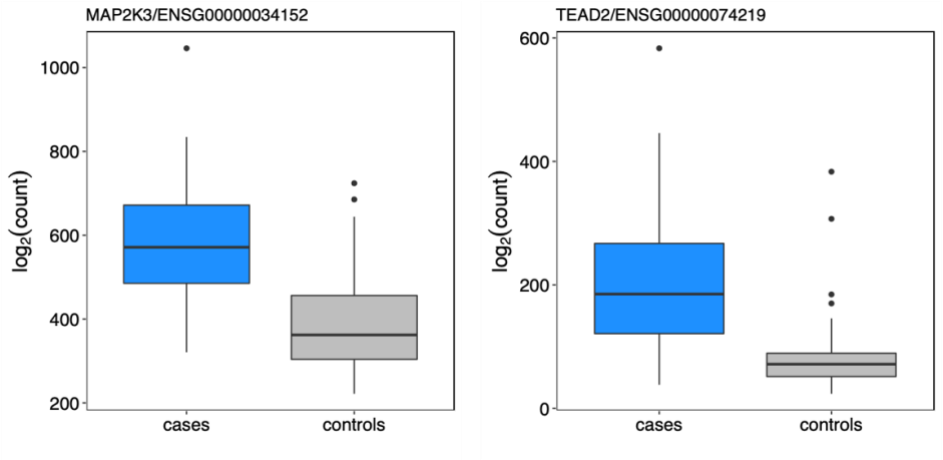


Figure 17: Boxplots of normalized counts for the 2 significant differentially expressed genes that overlapped top hits from the exome analysis, showing increased expression in cases compared to MGRB controls.

Given the limitations of small sample sizes, there have been several gene expression meta-analyses of PD cohorts. (Kelly et al., 2019) performed a meta-analysis of microarray data from 69 PD and 57 control substantia nigra samples. Their approach utilized combined gene sets from all included studies (rather than restricting to overlapping genes) and differentially expressed genes were identified by calculating the combined effect size across all samples. They observed a significant overlap in DEGs between PD and AD with expression changes in the same direction for more than 99% of the common DEGs (n=436). Neither LRRK2 nor GBA were among the significant genes, however SNCA showed significantly higher expression in both AD and PD cases (Bonferroni adjusted p-values: PD=2.06e-05 and AD=1.58e-08). Raw count data was not available, however, 3 of the 1046 significant DE genes (FDR<0.05) were also significantly differentially expressed in the same direction in RNA sequencing of our 28 prefrontal cortex samples. All 3 genes (PRKACB, BCL6, and DNAJC10) are homologs of yeast modifiers of a-syn toxicity. The authors also observed that DEGs were more likely to contain disease associated variants than non-DEG but that the association was weak (Fisher's exact p=0.04). Lack of evidence supporting an effect of genotypes on gene expression does not exclude a role for disease-associated variants, especially for missense SNVs, which rarely lead to unstable protein expression (Fragoza et al., 2019; Sahni et al., 2013). Known PD associated genes often do not show significantly altered expression

in cases compared to controls. Perturbed expression after disease onset may not capture the effect of risk variants that lead to PD but can reveal regulators of disrupted pathways.

Expression analysis of AD cases

The largest gene overlap with both RNA sequencing of 28 of our cases and our exome-based analysis was with a meta-analysis of AD expression data from six previous microarray studies comprising 450 AD and 212 healthy frontal cortex samples (Li et al., 2015). Only summary statistics of significantly DEGs were available, however, 19 of the 41 genes found to be differentially expressed in our 28 cases were also significant in the AD cohort. Five of the significant genes identified in our exome sequencing gene-based association passed a Bonferroni adjusted transcriptome-wide p-value of 0.05 (Table 10). Four of these five genes are AD-associated based on our sequencing rationale, including 3 a-beta modifiers (TEAD2, MAP2K3, and PKN2) and 1 AD risk gene (CD33).

gene	meta.pval	Bonferroni	FDR
TEAD2	4.65E-12	1.09E-07	4.05E-10
CD33	4.13E-11	9.72E-07	2.34E-09
PTPRR	1.56E-10	0.00000367	6.82E-09
MAP2K3	4.13E-07	0.00972617	4.03E-06
PKN2	1.53E-06	0.03591423	1.21E-05

Table 10: Differential expression analysis of AD prefrontal cortex samples.

TEAD2 and MAP2K3 also showed significantly increased expression in RNA-seq analysis of 263 AD (from the ROSMAP study) compared to 151 cognitively normal age matched prefrontal cortex samples (Canchi et al., 2019). A total of 4 genes were significant transcriptome-wide and in the same direction in both our 28 PD cases and 2 studies of AD, all from prefrontal cortex. It is possible that the 2 AD studies include overlapping samples, but this information was not available. However, the replication of TEAD2 and MAP2K3 across diseases, both of which were significant

in our exome-sequencing analysis, is of interest and may help to elucidate shared mechanisms of neurodegeneration.

	28 PD		Li et al.		Canchi et al.	
symbol	**log2FC**	**p.adj**	**log2FC**	**p.adj**	**log2FC**	**p.adj**
KLF15	0.677	0.00547	0.485	8.11E-11	0.431	0.000789
TEAD2	0.795	0.006047	0.367	4.05E-10	0.386	0.000148
MAP2K3	0.398	0.00686	0.0976	4.03E-06	0.214	0.0151
EGR1	-0.898	0.0274	-0.454	3.03E-09	-0.527	0.0142

Table 11: Overlapping transcriptome-wide differentially expressed genes between our 28 familial PD cases and 2 AD expression studies.

Expression analysis of MSA cases

We obtained raw count data from RNA sequencing of cerebellar white matter samples from the largest expression study of MSA, including 2 cohorts of MSA cases and healthy controls (Piras et al., 2020): 19 cases and controls and another with 47 cases and controls. Differential expression analysis was performed separately for each cohort using DESeq2 for our target genes, including PMI and age as covariates (no significant association was observed in a Chi-square test between sex and case/control status).

Two genes (BCL6 and FGR) overlapped between the 19 MSA cases and our 28 PD cases. The small number of overlapping genes is not surprising given the small sample sizes and differences in tissue analyzed as the MSA study assessed cerebellar white matter.

While the perturbed expression of individual genes can be difficult to interpret, especially in the absence of correlation between genotypes and gene expression, replication across diseases recapitulate some of our key gene level associations. Expression changes may be due to non-coding regulatory variants or perturbed protein interactions, necessitating integration of genomic and functional data, including gene expression and interaction networks.

2.2.9. RARE VARIANT BURDEN IN KNOWN PATHOGENIC MUTATION CARRIERS

For relatively rare neurodegenerative diseases that display Mendelian inheritance, mutations in a single gene are considered causal, however most cases are sporadic and disease occurrence is likely due to a combination of genetic and environmental risk factors. The occurrence of known causal PD mutations is rare in the PD population (<2%) (Tran et al., 2020). Carriers of pathogenic PD mutations conferring high risk are typically excluded from case-control analysis, however, this information is often known only after sequencing. Even for strong PD risk factors, including low frequency LRRK2 and GBA mutations, penetrance increases with age and is highly variable. Additional risk variants in patients with clinically relevant pathogenic variants may reveal genetic risk factors that influence penetrance.

Given the enrichment of AD related genes in our analysis, we considered cases with AD or PD associated clinical variants. All except 2 AD-associated pathogenic genes according to the ClinVar database (CSF1R and GRN) were included in our target genes (APOE, APP, MAPT, PSEN1, PSEN2, and VCP). However, no ClinVar pathogenic/likely pathogenic AD associated mutations were found in our filtered variant callset. The APOE-ε4 risk allele (rs429358) was detected but did not pass the 95% genotype call rate threshold in cases. This missense variant is classified as likely pathogenic in ClinVar and predicted by SIFT and Polyphen to be tolerated and benign, respectively, and has a MAF=14.9% in non-Finnish European gnomAD exomes and MAF=13.6% in our joint callset, failing inclusion filters for rare variant analysis. A total of 11 ClinVar pathogenic or likely pathogenic PD variants (out of a total of 91 SNVs in 23 genes) were detected in our cases in GBA, LRRK2, and PARK2 (Table 12).

CHR:POS:REF:ALT	rsID	GENE	ACcases	ANcases	ACmgrb	ANmgrb	ACgnomad	ANgnomad
1:155204793:C:T	rs75822236	GBA	2	972	0	5024	1	55500
1:155204987:G:A	rs80356771	GBA	1	992	1	5028	13	89546
1:155205043:A:G	rs421016	GBA	9	992	6	5020	119	89112
1:155205518:C:G	rs1064651	GBA	1	992	1	4936	19	89092
1:155205563:C:A	rs80356769	GBA	1	992	0	4950	0	89532
1:155205634:T:C	rs76763715	GBA	14	992	10	4974	176	89534
1:155207244:C:T	rs78973108	GBA	1	992	1	5026	6	89546
12:40704236:C:T	rs33939927	LRRK2	1	992	0	5016	2	89532
12:40734202:G:A	rs34637584	LRRK2	10	992	3	4856	25	89544
6:161771240:C:T	rs191486604	PARK2	0	992	1	5010	12	88302
6:162206852:G:A	rs34424986	PARK2	3	992	26	4858	285	89468
6:162394349:G:A	rs137853054	PARK2	0	992	1	4970	23	89550
6:162394435:T:A	rs137853060	PARK2	1	992	0	4980	5	89538
1:20975602:C:T	rs45539432	PINK1	0	990	1	4836	7	89454
1:20975710:C:T	rs34208370	PINK1	0	988	2	4888	7	89368

Table 12: ClinVar PD associated pathogenic variants detected in the joint callset. gnomAD counts are for non-Finnish European non-neuro exomes (AC: allele count; AN: allele number).

The two most persistent GBA mutations, N370S and L444P, were significantly enriched in our cases compared to healthy aged controls, in addition to the LRRK2 G2019S mutation. The GBA N370S and GBA L444P mutations were detected in 13/496 (2.6%) and 9/496 (1.8%) of cases compared to 10/2516 (0.39%) and 6/2516 (0.23%) of controls, respectively. Of the 21 cases harboring the N370S variant (one case harbored both the LRRK2 G2019S and GBA L444P variant), 8 were sporadic PD, 7 LBD, 6 familial PD, and 1 MSA. The LRRK2 G2019S mutation is reported in up to 1-2% of sporadic and 4-5% of familial PD cases (Healy et al., 2008). Our cohort included a total of 9/496 (1.8%) cases, all familial PD, with the G2019S mutation and 3/2516 (0.12%) controls. More common variants in LRRK2 and GBA appear to increase risk but a pathogenic mechanism has not been established. Occurrence of these mutations in both PD cases and healthy, aged individuals suggests a role for additional genetic or environmental factors in disease pathogenesis.

As cases with known pathogenic mutations may harbor additional risk variants, we did not exclude them from the primary analysis. We additionally tested whether there exist variants statistically associated with these 3 variants by performing a Cohort Allelic Sums Test (Morgenthaler and

Thilly, 2007) (1-sided Fisher's exact test) to compare the frequency of enrichment of additional rare variants (MAF<0.01) in cases harboring the LRRK2 G2019S (n=9) or either GBA N370S (n=13) or L444P (n=9) mutation compared to all other cases. 14/30 (47%) cases and 5/19 (26%) controls harboring a LRRK2 or GBA variant had at least 1 additional rare (MAF<1%) deleterious mutation (singletons were included as comparisons are within our case cohort) in a significant gene (GBA, LRRK2, TEAD2, ARHGEF1, PKN2, PTPRS, PTPRR, MAP2K3, HKR1).

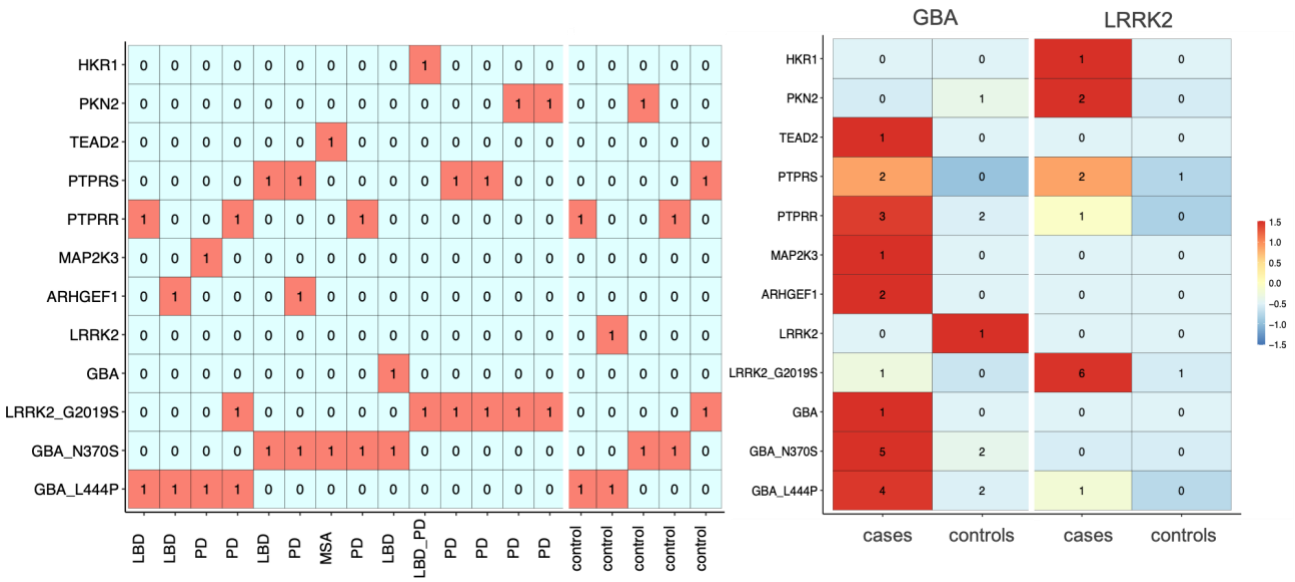


Figure 18: Left panel: Deleterious variants MAF<1% in top genes in cases (labeled by primary diagnosis) harboring known pathogenic mutations in GBA (N370S or L444P) or LRRK2 (G2019S). Right panel: Total cases or controls harboring a GBA or LRRK2 variant and a rare variant in a top gene. Heatmap colors represent scaled cohort frequency per gene.

GBA cases

GBA variants represent the most common genetic risk factor for PD, accounting for more than 10% of PD patients and both pathogenic and non-pathogenic (in Gaucher disease) variants have been associated with accelerated disease progression (Stoker et al., 2020). However, disease penetrance in carriers of these mutations is highly variable (Lee et al., 2017; Rana et al., 2013). Most carriers of GBA variants are asymptomatic, supporting a role for additional genetic modifiers and/or environmental risk. GBA mutations can increase risk for PD, however most of both heterozygous carriers and individuals with Gaucher disease do not appear to develop PD (Rana et

al., 2013). The penetrance of GBA has been studied in the context of Gaucher disease and familial PD (Balestrino et al., 2020) and the estimated penetrance in unselected PD patients was found to be higher than in Gaucher disease cohorts and lower than in familial PD cohorts. Reduced penetrance and phenocopies in apparent GBA-PD families, as demonstrated by non-symptomatic GBA carriers and PD non-carriers, respectively, may be explained by additional genetic risk factors (Aslam et al., 2021; Blauwendraat et al., 2020; Gan-Or et al., 2015).

A study of more than 500 PD probands with familial PD showed a relatively high estimated penetrance of GBA mutation carriers, increasing from 7-30% from age 50 to 80. The authors did not observe a significant difference in penetrance at 70 years between carriers of the common N370S or L444P or rarer mutations (Anheim et al., 2012). Even in Gaucher disease, in which individuals are homozygous for GBA mutations, the highly variable clinical presentation suggests the contribution of genetic modifiers.

We compared all 22 cases in the derivation cohort with either GBA mutation versus all other cases as well as each subset of carriers separately. We tested rare (MAF<1%) variants with moderate to high predicted impact according to VEP. After applying a Bonferroni correction threshold (0.05), we did not identify any significant variants in GBA (N370S or L444P) carriers versus all other cases or L444P carriers versus all other cases. We did however identify 2 missense variants significantly enriched (Bonferroni adjusted $p<0.05$) in GBA-N370S carriers compared to all other cases: PTPRB (rs202134984), and FREM2 (rs150928081). We then collapsed genotypes to count the total number of individuals with at least 1 variant in each gene. One gene (MICAL3) was significantly enriched for rare variants and 2 additional genes (PTPRB and PRKCD) had small nominal p-values but did not pass the stringent Bonferroni adjusted threshold (Table 13). The association was driven by GBA-N370S carriers, as these genes were significantly enriched in N370S cases and no genes were significant in L444P carriers versus all other cases.

gene	odds ratio	pval	Bonferroni	BH
MICAL3	12.3	0.000135	0.0379	0.0379
PRKCD	46	0.00027	0.0762	0.0381
PTPRB	5.2	0.0181	1	1

Table 13: Top genes from the cohort allelic sums test of variants with moderate to high predicted variant effect in GBA-N370S carriers versus all other cases.

MICAL3 (Microtubule Associated Monooxygenase, Calponin And LIM Domain Containing 3) and PTPRB (Protein Tyrosine Phosphatase Receptor Type B) are both human homologs of yeast genes (YML083C and PTP2, respectively) found to suppress alpha-synuclein toxicity in our overexpression screen. PRKCD (Protein Kinase C Delta) is a human homolog of a yeast enhancer of amyloid-beta toxicity (PKC1) and protein kinase C isozymes have been implicated in both AD (de Barry et al., 2010; Callender et al., 2018; Du et al., 2018) and PD (Gordon et al., 2016; Zhang et al., 2007).

LRRK2 cases

The LRRK2 G2019S mutation is a strong risk factor for PD but risk factors other than age that affect disease penetrance in carriers are not well understood. The first genome-wide association study for modifiers of LRRK2 related PD did not identify any SNP associations but this study was underpowered and included only 99 cases (Latourelle et al., 2011). A recent meta-analysis of 3 cohorts of PD cases and controls with the G2019S variant revealed an association between polygenic risk scores from 89 variants with PD, supporting a possible genetic contribution to PD penetrance among G2019S carriers (Iwaki et al., 2020). Additional genetic modifiers of age of onset have been reported in the literature (Brown et al., 2021; Trinh et al., 2016; Wang et al., 2012).

We performed a cohort allelic sums test for the 9 G2019S carriers versus all other cases. Rare (MAF<1%) deleterious SNVs (missense, splice, stop gained, stop lost, start lost) in PICALM (rs34013602), PKN2 (rs200490316) and ATP12A (rs61998252, rs61740542) were associated with LRRK2-G2019S with small nominal p-values ($p<0.01$), but did not pass the stringent Bonferroni

threshold. We then tested for associations at the gene level and observed no significant enrichment of rare variants in LRRK2-G2019S carriers, including synonymous or benign (synonymous, CADD<5, PH<0.15, REVEL<0.4) variants or collapsed by gene.

Clinical heterogeneity and age-dependent penetrance support a role for additional genetic and/or environmental risk factors in carriers of known mutations. Rare variants enriched in carriers of known GBA/LRRK2 mutations are candidate loci for combinatorial gene-gene interactions. As known pathogenic mutations represent a fraction of the genetic risk for PD, elucidating combinatorial effects is crucial to our understanding of how genetic risk variants contribute to disease pathology. Genome-wide studies comparing a sufficient number of diseased and healthy mutation carriers can facilitate the identification of genetic modifiers of disease penetrance. Their identification can have important implications for ongoing and future clinical trials of GBA and LRRK2 modifying therapies as well as understanding the risk of mutation carriers to develop PD.

2.2.10. CONCOMITANT A-BETA AND A-SYNUCLEIN PATHOLOGY

Soon after the association of SNCA and PD was published, alpha-synuclein inclusions were identified in other neurodegenerative disorders (Spillantini et al., 1997). Co-pathology is more prevalent than previously thought, with pure neurodegenerative disease making up the minority of diagnoses (Robinson et al., 2018). Despite little overlap in genetic risk factors between neurodegenerative diseases, upwards of 50% of PD patients have a secondary diagnosis of AD, based on the development of sufficient numbers of amyloid-β plaques and tau-containing neurofibrillary tangles (Irwin et al., 2013), and abnormal deposition of alpha synuclein is also seen in AD and other neurodegenerative diseases.

When we examined carriers of the top-hit variants among the 151 cases with pathological confirmation, we observed that carriers of variants in top-hit amyloid-beta genetic networks or known AD genes did not necessarily exhibit concomitant AD pathology, suggesting shared genetic

risk factors and mechanisms among different synucleinopathies. Modifiers of a-beta toxicity are associated with synucleinopathy risk, independent of concomitant AD pathology. Immunohistochemistry on brain sections from cases harboring one of the top variants (PTPRS:rs145504993) recapitulated our finding that top signals cross clinical disease boundaries (Fig. 19). For the case showing AD pathology, amyloid-beta deposits were present in the cerebellum, which only occurs in cases with advanced amyloid deposition.

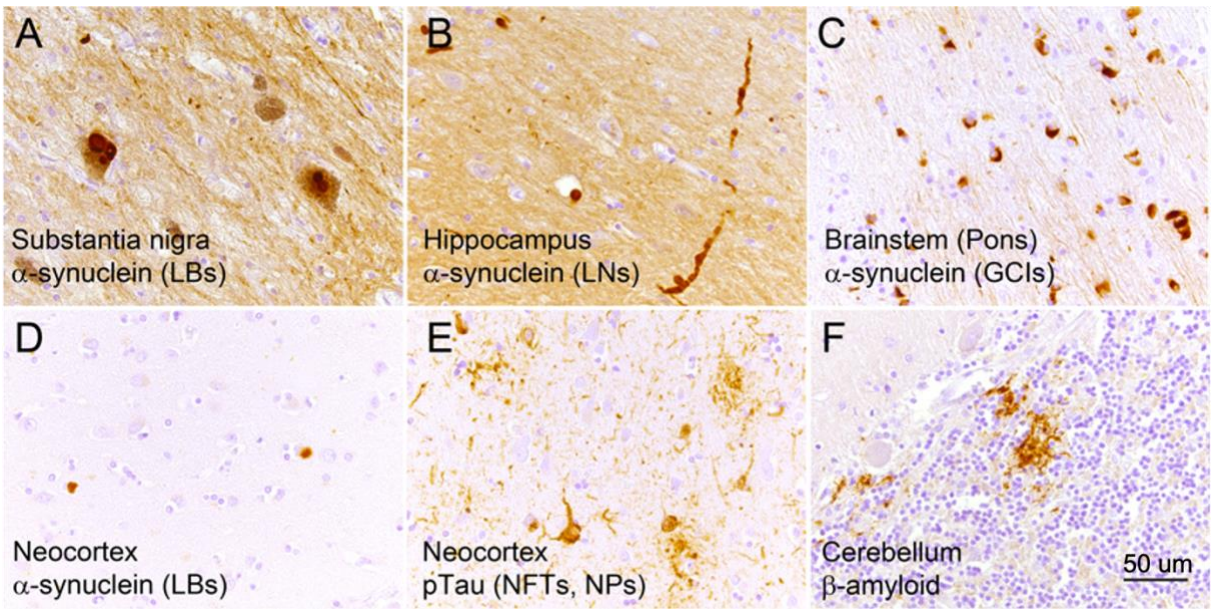


Figure 19: Histopathology for 3 cases harboring the PTPRS variant. (A) and (B) are from a clinically diagnosed PD case with LBD and no AD neuropathology. (C) A classic case of MSA showing no AD neuropathology. (D), (E), (F) A PDD case with LBD and a high level of AD neuropathologic change. (D) Neocortical Lewy bodies. (E) Phospho-tau stain showing neocortical neurofibrillary tangles and neuritic plaques. (F) Beta-amyloid deposits in the cerebellum. LB: Lewy bodies; LN: Lewy neurites; GCI: glial cytoplasmic inclusions; NFT: neurofibrillary tangles; NP: neuritic plaques. Images courtesy of Dr. Eddie Lee (University of Pennsylvania).

Co-pathology may reflect the co-existence of different disease processes in the same patient or could be due to a mechanistic link among different proteinopathies. Human genetic analysis has largely identified non-overlapping genetic risk factors in proteinopathies, and animal models have shown conflicting evidence. Combined expression of amyloid-beta and alpha-synuclein in mice appeared to worsen phenotypes, leading to increased deposition of intraneuronal fibrillar alpha-synuclein, amyloid-beta and/or tau and accelerated motor and cognitive dysfunction (Clinton et

al., 2010; Masliah et al., 2001). Conversely, knockdown of alpha-synuclein in an AD mouse model reduced degeneration of cholinergic fibers and hippocampal neurons (Spencer et al., 2016). However, increased amyloid plaque load has been observed upon knockdown of SNCA in APP mice (Kallhoff et al., 2007; Khan et al., 2018), suggesting a role for alpha-synuclein in inhibiting the formation of amyloid-beta plaques. Given the spatiotemporal progression of AD and PD according to Braak staging (Braak et al., 2003, 2006), concomitant AD/PD pathology depends on the relative stage of each disease (Visanji et al., 2013) (Fig. 20).

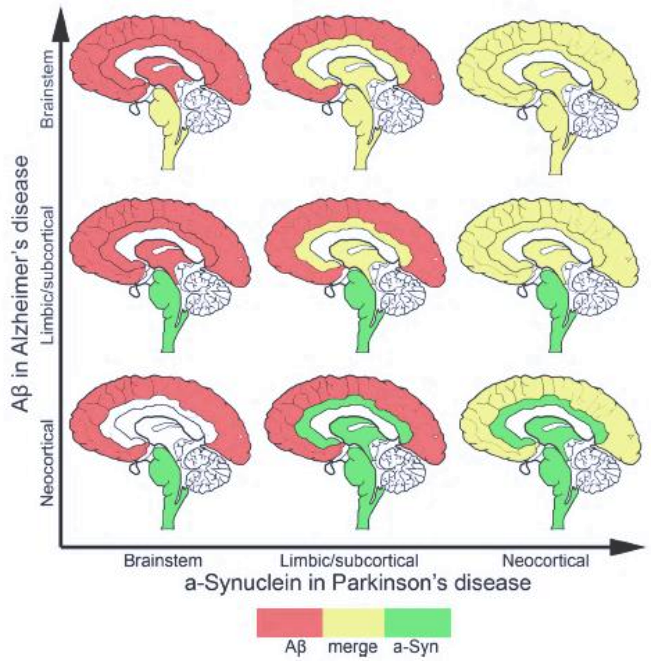


Figure 20: Predicted overlap of alpha-synuclein and amyloid-beta based on disease progression. Image sourced from (Visanji et al., 2013).

The occurrence of co-pathologies across neurodegenerative diseases can have major implications for clinical trials, which typically focus on therapies targeting a single disease-specific pathology. Understanding the interaction of different proteins in comorbid proteinopathies can help to distinguish general cellular responses to protein misfolding (ie. disruption of the proteostasis machinery) from responses related to the intrinsic function of the misfolded protein.

2.2.11. FUNCTIONAL VALIDATION

No disease-modifying or preventative therapies exist for proteinopathies. For Parkinson's disease, current strategies focus on relieving motor symptoms due to the loss of dopaminergic neurons and subsequent lack of dopamine, however therapies targeting the dopamine pathway have not been shown to alter disease duration (van der Brug et al., 2015). Model organisms, cellular models, and genomic studies have highlighted convergent pathways contributing to disease onset and progression but there are still major gaps in understanding of the contribution of individual variants and genes.

By the time symptoms appear, a significant proportion (more than 50%) of cells in the substantia nigra has typically been lost (Cheng et al., 2010). Understanding the complex interplay between genetic and environmental risk factors and aging is critical to diagnose patients before disease onset and to successfully target underlying pathologies. Drug mechanisms with direct genetic evidence have a greater chance of success in clinical development (Nelson et al., 2015). The wealth of genetic data can be harnessed to prioritize targets for the successful development of new therapeutics but requires robust models to establish the cellular impact of risk variants and potential disease-modifying mechanisms.

The complex pathophysiology of neurodegenerative disease has impeded the development of disease-modifying treatments, necessitating integrative approaches to functionally validate and prioritize drug targets with a high probability of success. Numerous family-based and population-scale genetic studies have established genetic predisposition in PD and have helped to prioritize candidate disease genes and genomic loci. However, our understanding of the biological functions of the risk alleles identified so far is limited. This is largely due to the complexity of establishing a molecular mechanism to pathogenicity, which is complicated by tissue specific expression, as well as pleiotropic and polygenic effects. Computational tools that predict pathogenicity often perform poorly for variants associated with late-onset diseases as known causal mutations have been

predicted to be benign (Foo et al., 2012). Interpreting the impact of genetic variants on gene function is essential to prioritize genes and pathways for validation. Disease associated variants are typically classified based on clinical rather than molecular effects on protein function with causality defined by familial occurrence or early disease onset. Most of these causal mutations are rare and associated with the rarer Mendelian form of the disease. Variant effect predictions reveal information about the potential impact of amino acid changes on individual protein function but not how these mutations impair protein interactions. Functional genomic strategies, using biologically meaningful assays, will become increasingly necessary to uncover causal associations between variants and disease phenotypes.

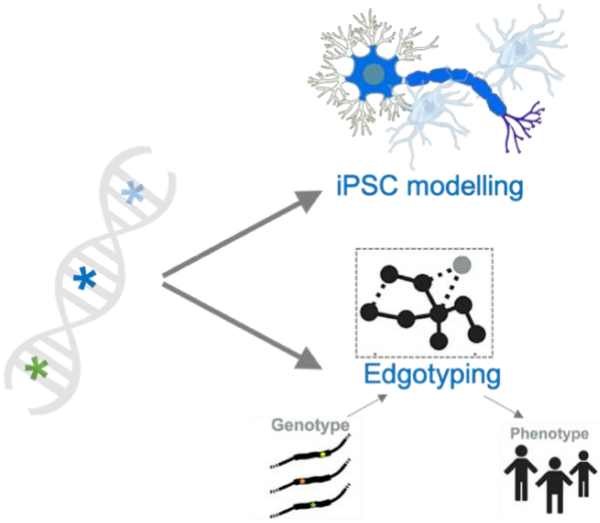


Figure 21: Distinguishing neutral from pathogenic variants and elucidating molecular mechanisms in neurodegenerative disease onset and progression are critical steps to the development of disease-modifying therapies. High-throughput assays can facilitate the functional characterization of variants, from yeast to neurons.

Distinguishing benign from pathogenic mutations and understanding the biological basis of genetic risk are necessary for accurate molecular diagnosis and to delineate their role in disease etiology and progression. We established a cross-species approach, from yeast to human neurons, to functionally validate the effect of variants identified in rare variant association testing, including protein interaction screens in yeast and phenotypic assays in human stem-cell derived neurons (Fig. 21).

2.2.11.1. FUNCTIONAL CHARACTERIZATION OF MISSENSE VARIANTS

The following experiments were significantly delayed due to the pandemic but are being carried out in the Khurana lab at Brigham and Women's Hospital/Harvard Medical School.

A major bottleneck in medical genomics is in distinguishing the small fraction of pathogenic variants from the enormous background of largely benign human genetic variation identified in association studies. The majority of associated variants are missense, which lead to amino acid sequence substitutions, which was long assumed to perturb protein function by impairing proper folding. Large-scale characterization of disease-associated missense mutations found that most do not impair protein folding or stability but do perturb macromolecular interactions (Sahni et al., 2013). A recent study observed enrichment of disease-associated variants in sequences encoding protein–protein interaction (PPI) interfaces, demonstrating the utility of prioritizing variants that disrupt PPIs (Cheng et al., 2021).

Demonstrating the functional consequences of the overwhelming number genetic variants identified from large-scale sequencing studies is crucial in order to establish a link between genotype and disease. Characterizing variants of uncertain significance (VUS) can be especially challenging for cell-type specific effects and different mutations within the same gene, which can lead to distinct disease phenotypes. A major focus is on missense mutations, which are associated with more than half of known inherited diseases (Stenson et al., 2017). The functional consequence of amino acid substitutions at the protein level can be more challenging to decipher than frameshift, splice-site, stop-gain, stop- or start-lost mutations.

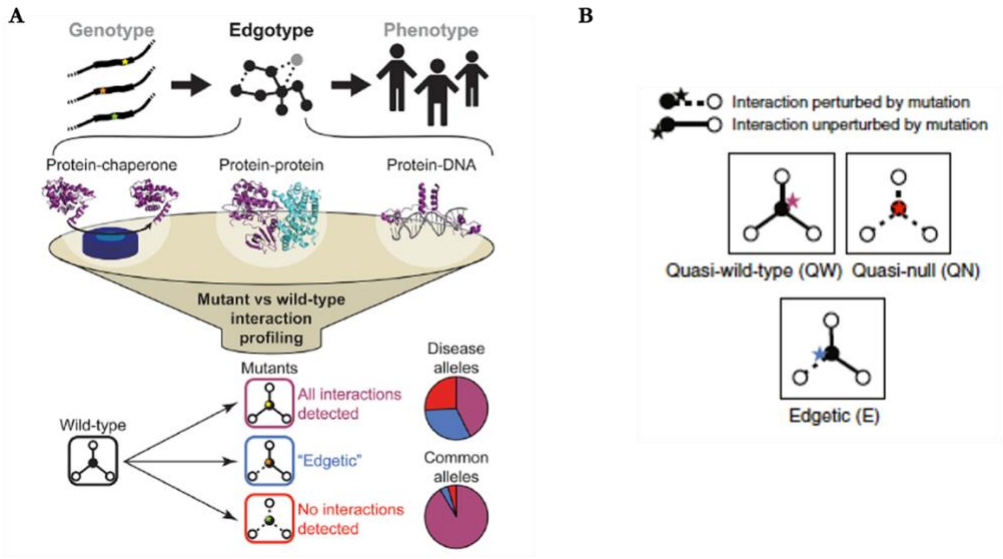


Figure 22: (A) Missense mutations can have no functional consequences, disrupt protein structure, or perturb interactions with macromolecules, including protein-chaperone, protein-protein, or protein-DNA. Edgotyping of missense variants can establish a mutational link between genotype and phenotype by comparing mutant and wild-type interaction profiles. (B) Variants can result in different interaction profiles: wild type, in which no interactions are perturbed; null, in which all interactions are perturbed; or 'edgetic', meaning a subset of interactions are perturbed. Image sourced from (Sahni et al., 2015).

Systematic analysis of disease associated missense mutations have shown that these mutations perturb protein-protein interactions (PPIs) as opposed to altering protein folding and stability. Large-scale analyses of missense mutations have highlighted a correlation between inheritance mode and protein structure. (Hijikata et al., 2017) assessed the effect of pathogenic missense mutations on protein structure and found that recessive mutations generally result in loss-of-function and are largely in the interior of the protein whereas dominant mutations are enriched in regions involved in molecular interactions. A previous study profiled several thousand missense mutations in Mendelian disorders, two-thirds of which exhibit perturbed molecular interaction profiles, compared to benign SNPs, which tend to show molecular interaction profiles similar to the wild type allele (Sahni et al., 2015). Half of these PPI disrupting mutations are considered 'edgetic', affecting only a subset of interactions. A more recent study (Gerasimavicius et al., 2020) tested the performance 13 different protein stability predictors in discriminating between pathogenic and benign missense variants and found high heterogeneity across proteins and low

performance compared to variant effect predictors in distinguishing pathogenic mutations, supporting a role for perturbed interactions.

We hypothesized that we could infer whether a variant exerts a functional effect at the protein level based on its molecular interaction profile. That is, variants exhibiting an edgetic profile (loss of interaction with some proteins, but retention of interaction with other proteins in its interaction profile), or quasi-null profile (loss of all detectable protein interactions) are likely to indicate that the variant has an effect on the protein it encodes, which may result in molecular and downstream cellular defects. Alternatively, variants exhibiting a quasi-wild-type interaction profile are more likely to have few detrimental effects and may be benign or neutral polymorphism (Fig. 22).

Briefly, protein-chaperone interactions (PCIs) are examined by quantitative LUMIER assay (Taipale et al., 2014) on the basis that unstable/misfolded proteins exhibit enhanced binding to chaperones. Protein-protein interactions (PPIs) are measured by systematic pairwise yeast two-hybrid assay with a panel of ~20,000 cloned ORFs (ORFeome Collaboration, 2016). For variants in genes encoding transcription factors (TFs), protein-DNA interactions are measured by enhanced yeast one-hybrid assay (Sahni et al., 2015) between DNA baits from a library of human enhancers and the TF as protein prey.

Recent human population growth has led to increasing numbers of rare variants. The rarest variants (singletons), which occur in a single sample, are often excluded from genetic association studies due to lower confidence in genotype accuracy. However, these variants may have a large effect on disease risk and could be incorporated into risk prediction models. As our targeted analysis included only 430 genes, it is feasible to prioritize and functionally validate singletons. Of all QC filtered, biallelic, autosomal, deleterious SNVs, 12% were singletons (appearing in 1 case and no controls). The majority (90%) are missense variants, suitable for interaction profiling, and 10% include start-lost, stop-gained, or splice variants. We selected missense variants of unknown significance (VUS)

for edgetic analysis, including those from variant level tests, those in significant genes, and private variants (Methods).

Even in Mendelian disease, genotype-phenotype correlations can be complex as a result of environmental factors, pleiotropy, and epistasis with different mutations in the same gene resulting in distinct diseases. Importantly, genes and gene products do not function in isolation, which is of particular relevance for complex diseases. Evaluating the functional consequences of rare variants will assist in revealing novel genetic risk factors, providing a more thorough understanding of the genetic architecture of PD. On a broader level, biological approaches to study the functional consequences of rare variants can be generalizable to other diseases and will facilitate effective interpretation of genetic variants in the clinic, where many variants remain of unknown significance.

2.2.11.2. GENOME EDITING TO TEST VARIANTS IN IPSC MODELS

Cell-based assays can be powerful tools to uncover the mechanisms underlying human diseases and prioritize therapeutic targets. In addition to a complex genetic architecture and environmental causes, existing animal models may not be appropriately modeling the disease, especially for idiopathic cases, which account for the majority of patients. The interplay of genetic risk, environmental stressors, and age-associated cell vulnerability on disease leads to heterogeneity of neuronal phenotypes, making the development of a single therapy challenging. Despite the heterogeneity of PD, iPSC models have revealed converging molecular pathways in familial and sporadic PD (Sandor et al., 2017).

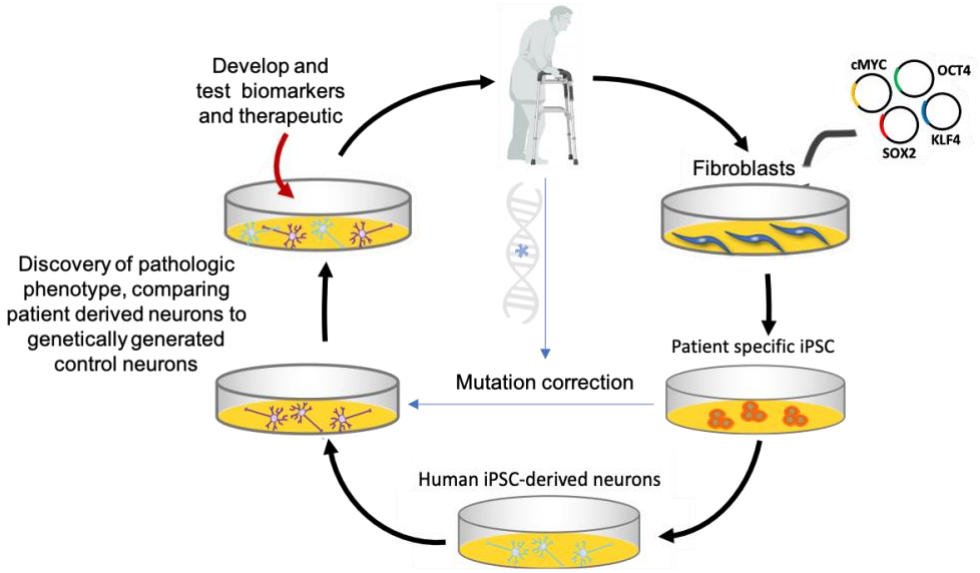


Figure 23: Genome editing of human iPSC-derived neurons can allow for systematic testing of variants in isogenic cell lines. Human iPSc alpha-synuclein A53T mutation and isogenic mutation-corrected cellular models are gene edited using CRISPR/Cas9 to introduce the base pair alteration corresponding to the gene variant. Subclones containing the correct genotype are differentiated into neuronal cultures enriched for cortical or dopaminergic neurons and subsequently evaluated for a range of pathologic phenotypes identified in these neurons.

No single mutation is 100% penetrant in Parkinson's disease, which is challenging for both the design of model-based systems as well as the development of disease modifying therapies. Multiple genetic risk factors are likely to play a role in both familial and sporadic disease, requiring novel approaches to understand the complex, polygenic mechanisms underlying disease onset and progression, especially for sporadic disease, which occurs in the vast majority (~85%) of patients. Neurons derived from PD patients through directed reprogramming to iPSCs and neuronal differentiation can establish robust disease phenotypes (Fig. 23). Combinations of variants recapitulating those found in patients can be tested to uncover those that contribute to a phenotype only in combination with additional risk variants. Differentiation into neuronal cell subtypes can allow investigation of selective vulnerability of different brain regions.

Differences in genetic background between iPSC lines from cases and controls can lead to variability in neuronal phenotypes unrelated to disease. Genetic and phenotypic variability can be addressed by generating isogenic cell lines through CRISPR/Cas9 genome editing, differing only

in the disease-associated mutation of interest. Given the low penetrance of most PD associated mutations, the genetic background of a patient is likely to affect cellular phenotypes. However, reducing variance with isogenic cell lines can lower the threshold to detect disease-relevant phenotypes. While alpha-synuclein is associated with familial PD in a dose-dependent manner, the role of elevated a-syn in the pathophysiology of idiopathic cases is not clear. Thus, a-syn overexpression may not be the appropriate model for the majority of PD cases. We have established isogenic cell lines, including a series of patient-derived SNCA triplication series, which includes the patient's reprogrammed iPSC line with 4 copies of SNCA, an engineered knockdown with 2 copies, and a complete knockout with 0 copies of SNCA, that will be used to test genes and variants of interest. As no SNCA mutations were detected in the patients in our cohort, these cell lines will allow us to test whether the gene or variant of interest displays a phenotype independently of SNCA copy number.

A limitation of in vitro models is the lack of non-genetic factors, notably, environmental stressors. Numerous cellular stressor reagents have been used to mimic inflammation, oxidative, proteostatic, synaptic, ER, and mitochondrial stress, as well as cellular aging through increased expression of progerin (a truncated form of lamin A), which is associated with premature aging (Tran et al., 2020).

Optimization of differentiation and programming protocols and establishing robust disease phenotypes are necessary to accurately model neurodegenerative disorders *in vitro*. Patient derived iPSC models have the potential to more accurately capture the complex interplay of human genetic risk factors and identify targets for effective neuroprotective therapy. A major challenge to modeling synuclein cellular pathology in human iPSc-derived neurons is that mature Lewy body-like aggregates do not form in immature iPSc-derived neurons. We developed iPSc models to capture advanced pathology and performed a proof of principle experiment using synthetic recombinant a-syn pre-formed fibrils (PFFs) visualized with an aggregation reporter (GFP). Neurons were engineered by CRIPSR targeting α-syn aggregation reporter to a novel gene locus.

The robust a-syn aggregation and inclusion toxicity phenotypes allow gene knockout and variant studies in these models (Fig. 24). This model will allow us to study the effect of top genes/variants that emerged from the targeted exome sequencing on inclusion formation or toxicity.

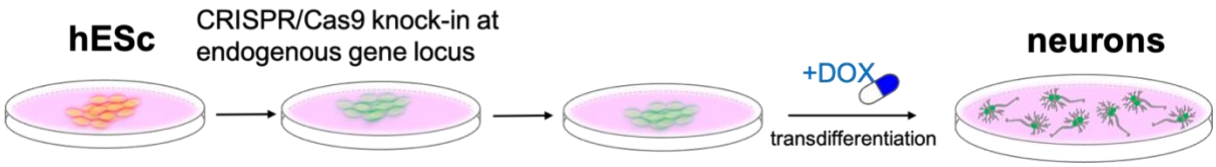


Figure 24: Proteotoxicity reporter isogenic cortical neurons. Human embryonic or induced pluripotent stem cells are knocked in with the reporter. Piggybac bicistronic constructs are then introduced into these knock-in lines, inducing simultaneously transdifferentiation (via Ngn2 forced expression) and over-expression of untagged/tagged toxic proteins (alpha-synuclein) or control fluorescent proteins. The tagged alpha-synuclein and tau constructs are used to enable visualization of seeded aggregation by exogenous pre-formed fibrils.

2.3. DISCUSSION AND PERSPECTIVES

Increasingly large cohorts in human genetic association studies have identified increasing numbers of risk loci in neurodegenerative disease. Case-control studies of both common and rare variants have uncovered more than 90 genomic loci linked to Parkinson's disease (Nalls et al., 2019). These variants explain ~22% of disease risk and increasing sample sizes will likely continue to uncover additional risk variants, most of which have a small effect size, accounting for a fraction of variance in disease risk. Rare coding variants have larger effect sizes on average, compared to common coding or non-coding variants but require incredibly large sample sizes. We hypothesized that reducing the sequencing search space to a set of genes known *a priori* to be enriched with genetic modifiers of alpha-synuclein proteotoxicity would enable us to find meaningful signals even in a modestly sized cohort.

The major challenge ahead lies in distinguishing benign from deleterious mutations and closing the genotype-phenotype gap through robust, scalable functional assays to elucidate the interplay of putative risk loci in disease onset and progression. Cross-species studies of proteinopathies have

uncovered disease-relevant biology, including homologs of known human disease risk factors. While the same genes are not necessarily recovered, findings often converge on the same cellular pathways. We uncovered new candidate risk factors for synucleinopathies, all linked to basic alpha-synuclein pathobiology through unbiased genome-wide yeast screens.

Rare variants in both amyloid-beta and alpha-synuclein genetic networks were enriched across different synucleinopathies compared to healthy aged controls, as were variants in both known PD and AD genes and associations were independent of concomitant AD pathology. Gene-level tests, including a trend test developed for rare variant analysis, confirmed and extended these findings in additional MSA and PD cohorts. Our findings implicate shared genetic drivers across distinct synucleinopathies and demonstrate convergent genetic networks in beta-amyloid and alpha-synuclein proteinopathies in humans. Despite little overlap in genetic risk factors between neurodegenerative diseases, concomitant AD and PD associated pathology are observed. Understanding the interaction of different proteins in comorbid proteinopathies can help to distinguish general cellular responses to protein misfolding from responses related to the intrinsic function of the misfolded protein and is crucial in order to properly stratify patients and develop effective disease-modifying therapies.

Assessing the functional consequences of variants is a critical step towards identifying direct contributors or risk factors, eventually enabling more accurate predictions of disease risk and the development of neuroprotective therapies. Most genome-wide screens against proteotoxicity have focused on single genetic modifiers. Genes and their products are part of a complex network of physical and biochemical interactions between DNA, RNA, and proteins. Development of effective treatments requires understanding the combinatorial effects between genetic risk factors and their relation to a-syn aggregation. Functional validation requires rigorous evaluation in appropriate cellular and/or environmental contexts in order to recapitulate the complex interplay between age-associated stressors, genetic, and environmental risk factors. Validation in simple

models (eg. yeast) can be followed by tissue and cell-type specific functional assays in more complex models. Advances in iPSC technology have established what is likely to become a robust model to validate therapeutic targets.

As the prevalence of PD and ensuing socioeconomic impact are projected to increase proportionally with population longevity, there is a critical need for therapeutic interventions. We have developed a cross-species biological platform that can be used to interrogate emerging data from large-scale sequencing in neurodegenerative disease. The combination of variant molecular interaction profiles and genetic interactions with alpha-synuclein toxicity will provide a global view of the functional consequences of rare variants identified in patients with PD as well as evidence for their biological relevance to cellular pathological events. These findings have the potential to identify novel risk factors and provide new avenues of exploration for treatments.

2.4. METHODS

2.4.1. TARGETED EXOME SEQUENCING AND JOINT CALLING

We performed targeted exome sequencing of 506 familial and sporadic PD, LBD, and MSA cases as well as 98 HapMap CEU samples that were sequenced as part of the 1000 Genomes Project for the purpose of quality control (Coriell Institute for Medical Research, Camden, NJ, USA). Exome enrichment and sequencing library preparation were done following a hybrid selection protocol (Blumenstiel et al., 2010) using Agilent SureSelect baits targeting the coding region of 430 genes (~1 Mb). Samples were sequenced in three separate runs at the Broad Institute (Cambridge, MA, USA) on the Illumina HiSeq (75 bp paired end). Raw reads were aligned to the human reference genome (GRCh37/hg19) using BWA (0.5.9). Paired reads were aligned separately with bwa aln using the following parameters: -q 5 -l 32 -k 2 –o 1. Results were combined and suffix array indexes converted to SAM format using bwa sampe. SAM files were converted to BAM format and

duplicate reads were tagged with Picard (2.18.17). All analyses and results are according to assembly GRCh37/hg19 coordinates.

Kinship estimates are important to remove cryptic relatedness but is challenging for sparse sequencing data. Relatedness statistics were calculated using VCFtools (0.1.14) to determine the unadjusted Ajk statistic based on (Yang et al., 2010). The expectation of Ajk is zero for individuals with populations and 1 for individuals with themselves. This equation is not appropriate for our small cohort and sparse targeted data as relatedness estimates are relative to the base population, but we nonetheless used 2 father/daughter duos to exclude potentially related individuals. One individual from each of 4 pairs of related (Ajk > 0.3) samples were excluded. This threshold was chosen based on the known kinships. Two cases showing excessive heterozygosity and 1 with excess genotype missingness were excluded and 3 potential population outliers based on the PCA, for a total of 496 cases. Sex imputation was not possible given the small number of sex chromosome SNPs available and was not reported for all samples.

Sequencing and alignment details for the MGRB whole genome sequencing samples can be found in the publication (Pinese et al., 2020).

The following steps were performed on both our cohort and MGRB samples. Individual variant calling was performed on each BAM file using GATK HaplotypeCaller (v3.7) to generate intermediate gVCFs. As the majority of samples were above the recommended minimum of 1E8 total base pairs per read group, base quality scores were recalibrated with GATK BaseRecalibrator using known sites of variation from dbSNP (build 138), 1000 Genomes Phase 1 indels, and Mills and 1000 Genomes gold standard indels. All variant files were obtained from the Broad Institute's GATK resource bundle (b37/hg19).

They were subsequently combined using GenomicsDBImport for joint genotyping of the 598 targeted exomes and 4032 whole genomes from the entire MGRB with GATK GenotypeGVCFs

of the resequencing target region with 100 bp padding around each interval. Variant ID annotation was done using dbSNP (build 138). Variant Quality Score Recalibration (VQSR) was not performed due to the insufficient number of sites in our targeted gene panel.

2.4.2. SEQUENCING QUALITY CONTROL

As recommended for data that cannot support VQSR, we used GATK VariantFiltration to apply the following hard filters separately for SNPs: QD > 2.0, SOR < 3.0, MQ > 40.0, FS < 60, MQRankSum > -12.5, ReadPosRankSum > -6.0 and indels: QD > 2.0, SOR < 10.0, FS < 200.0, ReadPosRankSum > -10.0) with Q > 50 and GQ > 20. These thresholds were chosen after visualization of density plots of variant annotation values, to balance sensitivity and precision. Only variants that passed all GATK filters and that overlapped our target region were retained. Indels were left-aligned and trimmed using bcftools norm. Variants were further filtered to include biallelic sites with a maximum of 5% missing genotypes, HWE p-value > 1E-6 in controls (HapMap and MGRB) and covered at >= 10X in >= 95% of samples. Sites with depth > 2x the mean were excluded, resulting in a total of 15655 variants (15237 SNPs and 418 indels). MGRB controls had nearly twice as many singletons (AC=1) as cases but roughly the same percentage (~60%) were found in gnomAD NFE.

variant effect prediction	cases	MGRB
synonymous variants	1988	4463
nonsynonymous variants	2494	6873
missense variants	2396	6532
frameshift variants	35	131
nonsense variants	33	115
splice variants	27	79

Table 14: QC filtered variants by functional effect prediction according to VEP.

We also compared SNV calls in our internal CEU samples with those from the 1000 Genomes Project and observed a concordance of greater than 99.9% for all chromosomes. Although

HapMap controls are self-declared healthy at sample collection, phenotype information is not available, however no significant deleterious variants or genes were identified in a cohort allelic sums test.

2.4.3. INDIVIDUAL LEVEL FILTERING

Only the healthy aged MGRB controls of European ancestry (n=2516) were retained in the analysis. No samples were found to have excess heterozygosity (>= 5 standard deviations from mean inbreeding coefficient = abs(F) >= 0.32). Individuals with >25% missingness (1 case and 1 MGRB control) and known related individuals were excluded. We applied principal component analysis to the publicly available sequencing data for 1397 HapMap samples and our joint callset, after excluding these samples. Plink format files were downloaded from the Centre for Applied Genomics (http://www.tcag.ca/tools/1000genomes.html). The joint callset VCF was converted to PLINK format using PLINK (v1.90b3x) and filtered to include biallelic SNPs with MAF > 1% and genotype call rate of > 95%. SNPs were pruned to be in approximate linkage equilibrium using plink --indep-pairwise (window size=50 kb, step size=5 variants, r^2 threshold=0.2) to avoid capturing too much variance in LD regions, resulting in 813 SNPs. All samples in our cohort clustered closely with the CEU and TSI HapMap populations and with the MGRB control samples. No samples were found to be further than 6 standard deviations away from the center of the first 5 PCs (Figure 25).

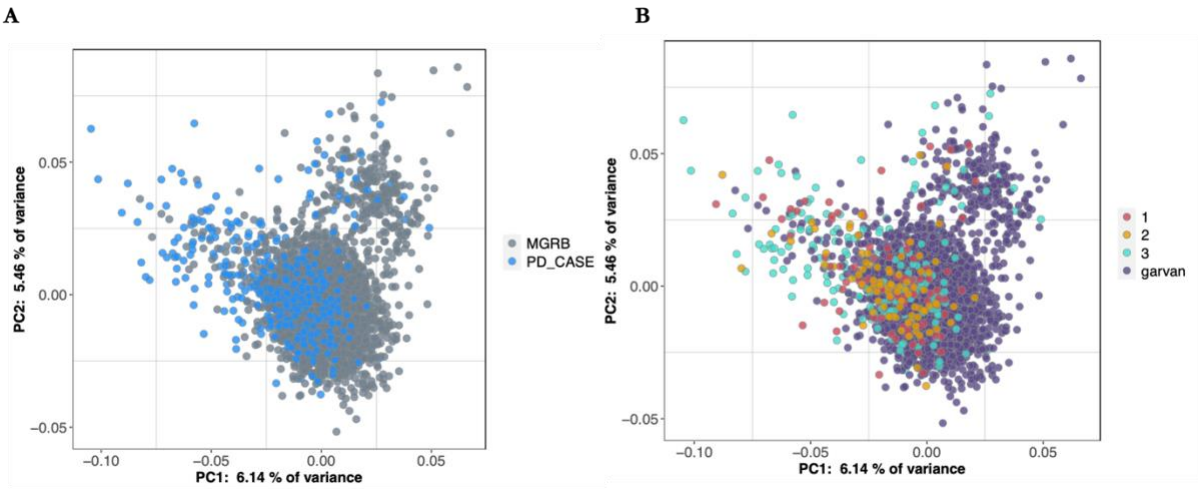


Figure 25: (A) Principal component analysis of common LD-pruned SNPs from jointly genotyped cases and MGRB controls, after removing potential population outliers. (B) PCA of jointly called cases and MGRB controls labeled by sequencing round (exome sequencing round 1=92, round 2=87, round 3=317, Garvan Institute MGRB=2516 samples).

2.4.4. VARIANT LEVEL FILTERING

Variants were further filtered to include those with >=10X coverage in >95% of all jointly called samples, genotype call rate > 95% in targeted exomes, whole genomes, and across all samples and HWE p-value > 10E-6 in internal HapMap or MGRB control samples. Sites with depth greater than 2x the mean depth across all variants were excluded as these are likely to be sequencing artefacts. Only variants passing the "AC0 filter" (GQ < 20; DP < 10; and AB < 0.2 for heterozygous calls) were retained in gnomAD Non-Finnish European non-neuro controls. Variants were annotated with SeattleSeq138, VEP (v95) (McLaren et al., 2016), and SnpEff (4.3t) (Cingolani et al., 2012), and REVEL (Ioannidis et al., 2016) and rsIDs were updated to dbSNP build 150. Finally, deleterious SNVs (missense, splice, stop gained, stop lost, start lost) with a minor allele frequency (MAF) < 1% in cases, 98 internal HapMap CEU controls, gnomAD, and MGRB controls were retained. To reduce the potential for technical artefacts, singletons, ie. variants which appeared exactly only once in both PD cases and MGRB controls were excluded because of the significant deviation observed in their distribution between the two cohorts.

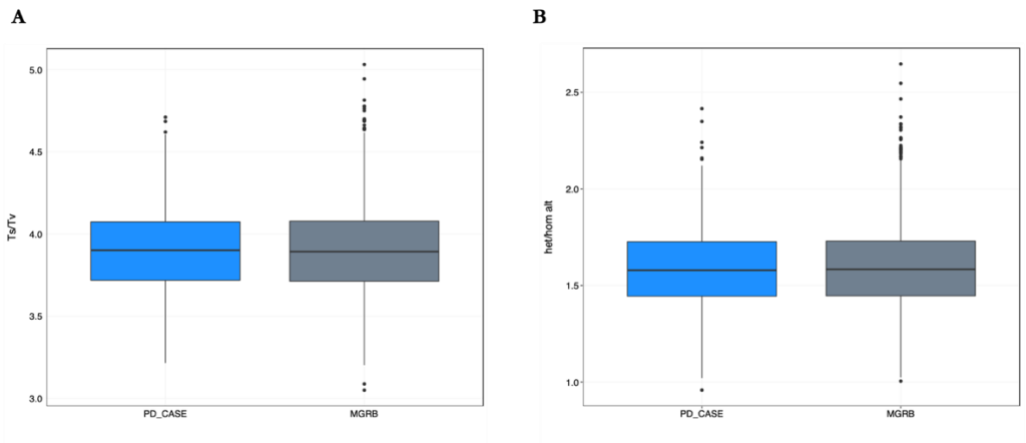


Figure 26: Boxplots of Transition/transversion (Ts/Tv) ratio (Wilcoxon p=0.87) and heterozygosity, measured as the ratio between heterozygous and homozygous alternate SNVs (Wilcoxon p=0.76) for all QC filtered variants in cases and MGRB controls.

2.4.5. RARE VARIANT DISTRIBUTION BETWEEN CASES AND CONTROLS

To reduce the potential for technical artefacts, singletons, ie. variants which appeared exactly once in PD cases or MGRB controls were excluded because of the significant deviation observed in their distribution between the two cohorts (Table 15).

In a balanced case-control study, in the absence of population structure and technical artefacts, if the target region does not harbor any alleles associated with the disease or trait, the distribution of allele counts should follow a binomial distribution, irrespective of the variant type (Neale 2011). The null hypothesis is that the outcomes, in this case the total number of variants, are equal between cases and controls. The alternative hypothesis is that there are less or greater shared total variants in cases versus. We compared the expected distribution of binomial counts with the observed counts for singleton, doubleton, and tripleton variants between cases and an equal number of controls.

In order to assess differences between cases and MGRB controls, we performed a binomial exact test for deviation from the expected distribution for singletons (variants with AC=1 in cases or controls), doubletons (AC=2). The binomial test was done using functional consequence

predictions for the canonical transcript. For the association analysis, all annotations were used when selecting qualifying variants.

The probability of success in a Bernoulli trial is as follows for each variant class:

singletons	$\frac{n_{cases}}{n_{cases} + n_{controls}}$
doubletons	$2 * \frac{n_{cases}}{n_{cases} + n_{controls}} * \frac{n_{controls}}{n_{cases} + n_{controls}}$
tripletons	$1 - \left(\frac{n_{cases}}{n_{cases} + n_{controls}}\right)^3 - \left(\frac{n_{controls}}{n_{cases} + n_{controls}}\right)^3$

		cases	MGRB	shared	total	prob. success	p-value
missense	singleton	990	4398	--	5388	0.162	**0.000017**
	doubleton	49	572	215	836	0.272	0.175
	tripleton	8	195	142	345	0.408	0.573
synonymous	singleton	744	3352	--	4096	0.162	**5.63E-04**
	doubleton	44	487	216	747	0.272	0.8606
	tripleton	6	149	113	268	0.408	0.695

Table 15: Results from the binomial exact test for missense or synonymous singleton, doubleton, and tripleton SNVs.

2.4.6. ANALYSIS OF SYNONYMOUS VARIANTS

Most sequencing studies focus on variants that alter the encoded amino acid, thereby affecting the protein sequence, structure, and function. Synonymous variants are not considered as contributors to disease, as they do not change the resulting protein sequence. They are typically discarded or used as quality control for variant filtering with the expectation that there should be no significant differences between well calibrated cases-control cohorts. However, there is increasing evidence supporting a role in cellular function and human disease (Plotkin and Kudla, 2011). Synonymous variants have been implicated in transcription factor regulation, mRNA folding, protein synthesis, microRNA binding, splicing regulation, and protein folding (Shi et al., 2019). Nonsynonymous and

synonymous SNPs can also have similar likelihood and effect size in disease association (Cheng et al., 2010).

Synonymous variants are not true negatives, as they are generally not verified to have no effect on protein function or no relationship to disease (Bromberg et al., 2013). Functional validation can be more challenging but minigene assays (Gaildrat et al., 2010) can assesses potential effects on splicing accuracy and efficiency. A study of 67 Alzheimer's disease kinships identified rare synonymous variants that segregate with the disease (Tang et al., 2020). Association was validated in independent cohorts and further supported by perturbed expression of these genes in post-mortem brains. Functional validation was performed using minigene reporter assays: constructs containing the reference and alternate alleles were transiently transfected into HEK293 cells followed by quantitation of cDNA products and comparison of size and quantity of reference and variant minigene products. No changes in splicing patterns between reference and alternate alleles were observed, however fluorescent quantitation of splicing products showed altered splicing efficiency between reference and variant allele constructs for 2 of the variants.

In addition to splicing efficiency, synonymous codon bias has been shown to play a role in disease (Fornasiero and Rizzoli, 2019; McCarthy et al., 2017). Analysis of rare synonymous variants that affect codon frequency, collapsed by gene, revealed an association with AD (Miller et al., 2018). Alternative splicing contributes to proteome diversity and plays an important role in most biological processes, with more than 90% of human protein-coding genes encoding multiple transcripts through this process (Wang et al., 2008). Global changes in alternative splicing occur during both normal aging and in age-related disease, supporting a possible role for synonymous codon changes in abnormal splicing.

We performed a cohort allelic sums test (CAST) of synonymous variants, collapsed by gene and observed enrichment in 4 genes (SORCS1, RABAC1, ATP1A3, KLF14). We also performed a

Fisher's exact test for all variants and observed enrichment of 4 synonymous variants in 3 genes (ATP1A3, OSBPL9, and RABAC1) in cases compared to MGRB controls (Bonferroni $p<0.05$) (Table 16).

variant	gene	odds ratio	pval	Bonferroni
19:42474639:A:G	ATP1A3	40.7	4.19E-06	0.00316
19:42473665:G:A	ATP1A3	39.9	4.78E-06	0.0036
19:42462977:C:A	RABAC1	39.8	4.84E-06	0.00365
1:52246956:C:T	OSBPL9	9.4	1.09E-05	0.00823

Table 16: Predicted synonymous variants enriched in cases versus controls.

The top 3 variants are all in LD with each other ($R2 > 0.1$), based on haplotype frequencies in 1000 Genomes CEU, TSI, GBR, and IBS populations. While these variants may play a functional role, they may simply be tagging a deleterious variant. However, LD analysis is challenging for rare variants, many of which are not detected in population scale sequencing databases. Nonetheless, differences in synonymous variants in case-control cohorts should be routinely assessed for association with disease.

2.4.7. GENOMIC CONTROL

Unequal distribution of cases and controls among different ethnic groups can lead to spurious genetic associations. Genomic control is an important quality control step in case-control association studies. In the absence of population stratification, technical artefacts, and cryptic relatedness, the test statistics for the majority of variants should follow the distribution under the null hypothesis of no association with the trait (Bacanu et al., 2000). We expect moderate inflation given the small subset of genes, which is enriched for known disease genes. We assessed the genomic inflation factor, defined as the median of the observed test statistic divided by the expected median of the null distribution. This value should be close to 1 in the absence of spurious association. We observed inflation in cases compared to the MGRB controls ($\lambda \sim 1.3$). Test statistics can be corrected for genomic inflation, but this can reduce power, especially for rare

variant studies. Genomic control is generally more useful in assessing rather than correcting stratification (Kiezun et al., 2012).

As we did not observe outliers from the PCA with cases and controls and our cohort clustered closely with known HapMap European samples, we decided to assess the effect of the difference in coverage between cases and MGRB controls on genomic inflation. Reads were downsampled in the cases to obtain similar coverage as the MGRB controls. The case BAM files were downsampled to median library size (4,200,740 total reads) and to 1 million total reads per sample using Picard (2.18.17) DownsampleSam (FRAC = median/total_reads; STRATEGY = HighAccuracy; ACCURACY = 0.0000001). Duplicates were again flagged with Picard MarkDuplicates, as the downsampling procedure can alter a sequencing read's duplicate status, and QC and variant calling were repeated as described for the complete callset. A total of 6785 SNPs and indels passed all GATK filters with GQ>20 and Q>50 and overlapped the target region in the full callset with 6712 after downsampling to the median library size and 6711 variants detected after downsampling to 1 million reads.

In the absence of population structure and technical artefacts, increased genomic inflation is expected in the presence of polygenic inheritance, as with complex traits and diseases (Yang et al., 2011). Principal components could be included as covariates in certain association tests (eg. kernel-based association test) but can further reduce power, especially if causal rare variants are spatially (ie. physically) clustered (Mathieson and McVean, 2013). Inferring population structure typically requires both a large number of individuals and a large number of markers. The regions in targeted sequencing are short and cover a small fraction of the genome and do not contain sufficient variation to infer global ancestry. It is unclear how reliable the results of the PCA are given the small number of SNPs, which may be insufficient to accurately capture the variance between groups. Further, stratification based solely on targeted disease-susceptibility loci, which are likely to contain trait-associated variants, could mask true association signals (Wang et al., 2014). Off-

target reads (those that map outside of the targeted exons) are typically discarded from analysis as they tend to be low quality and randomly/sparsely distributed across the genome. (Wang et al., 2015) applied a PCA and Procrustes analysis to project samples with sparse sequencing data (ie. low coverage and/or targeted sequencing) into a PCA space using reference data from the HGDP (Foster, 2008). Without the projection from high-dimensional reference PCs, ancestry estimates based on ~4000 SNPs in 385 samples of European ancestry (ancestry estimation is more challenging, requiring more data to detect fine-scale population structure) did not reflect population structure patterns and showed a low similarity score with genome-wide SNP data. (Zhang et al., 2013) showed that including classical PCs based on variants with low MAF can lead to power loss in association analysis. They also showed that classical PCA, when applied to variants with MAF < 1% is sensitive to outliers and does not effectively separate subpopulations. Given the sparsity of our data, we did not include PCs as covariates.

Based on the rationale of our targeted sequencing approach, we anticipated higher inflation than would be observed for genome- or exome-wide analysis, as we have enriched for disease-associated genes. With this in mind, we measured the genomic inflation before and after excluding pathogenic SNVs with no conflicting interpretations associated with Parkinson's disease according to ClinVar (91 SNVs in 22 genes). The genomic inflation factors in the jointly genotyped callset, the combined separately genotyped full and downsampled callsets were computed after the variant association analysis. The λ values were 1.32, 1.33, 1.34, 1.33 for the joint callset, the full combined callset (ie. separate variant calling for cases and controls), the median-depth downsampled combined callset and the 1-million-read downsampled combined callset, respectively. After excluding 15 ClinVar pathogenic variants, the corresponding λ values became 1.12, 1.14, 1.160 1.14 respectively. Given that our targeted analysis enriches for disease genes, we expect a modest amount of inflation. The decrease in genomic inflation indicates that it was primarily driven by the enrichment of known pathogenic mutations.

We observed a similar effect on inflation when including only variants with a REVEL score >0.6. Variants with scores >0.5 are considered "likely disease causing" and only 10.9% of neutral variants and 75.4% of disease mutations were found to have a score above 0.5 (Ioannidis et al., 2016). Inflation was 3.2 when including only such variants and decreased to 2.2 after excluding pathogenic ClinVar mutations.

2.4.8. NON-UNIFORM P-VALUE CORRECTION

For largescale multiple testing, the use of an empirical null distribution rather than the theoretical null distribution can be critical for correct inference from the observed data. Since observed counts of rare variants in target genes are not continuous, the raw p-values are not uniformly distributed between 0 and 1. We took this into account when correcting for multiple comparisons. We first applied the stringent Bonferroni correction to give the most conservative list of significant hits without any possible false positives. Next, we leveraged a modified false discovery rate (FDR) correction approach (Weghorn and Sunyaev, 2017), which simulates the null p-value distribution by generating the counts of rare variants in genes in 10000 independent permutations of the case-control labels. The adjusted p-values are calculated in a manner analogous to the Benjamini-Hochberg procedure for the uniform case:

$$q_g = \left(p_g^{obs} + \sum_{i=1}^{M} \{ p_i^{null} \, | p_i^{null} < p_g^{obs} \} \right) * \frac{M}{g}$$

$$p_g^{adj} = \min\left(1, \min_{i \geq g}(q_i)\right)$$

Here, *M* denotes the total number of tests, p_x^{obs} denotes the raw p-value with rank *x*, and p^{null} denotes the simulated null p-value distribution (Lee et al., 2014).

The lambda inflation factor was estimated using estlambda2 from the QQperm package in R. This method does not assume the null distribution to be uniform and instead uses the permutation-based expected null distribution, which is more representative of the true null distribution of

Fisher's exact p-values for the given case-control scenario. Empirical and permutation based empirical null p-values are used to generate quantile distributions based on the non-central chi-squared distribution with 1 degree of freedom, followed by regression of observed onto expected distribution.

2.4.9. COVARIATE ADJUSTMENT

It is common in case-control genetic association studies is to adjust for sex and age, or any covariate with a moderate to strong association with the outcome of interest, which can result in spurious correlation between genotype and phenotype due to correlation with both the tested genotype and phenotype. Covariate adjustment is done in order to discover genetic variants associated with disease after accounting for other factors that are known to affect disease risk, especially those that may be correlated with genetic variation. When the variants have no effect on the covariate or if the correlation between that covariate and outcome is due to a direct effect on the outcome, the adjusted and unadjusted estimates would correspond to direct and total (direct + indirect) genetic effect of the variant on the outcome, respectively. If the correlation between the covariate and the outcome is due to shared genetic and/or environmental risk factors, the adjusted estimate can be biased (Aschard et al., 2015).

As it is rarely possible to measure the effect of all covariates on the outcome of interest, non-confounding covariates are often included as they can increase power by explaining phenotypic variation that might instead appear as noise (ie. reducing residual variance). Including covariates can increase power in case-control studies of common disease but can significantly reduce power when disease prevalence is lower than a few percent (Pirinen et al., 2012).

Advanced age is the greatest risk factor in both familial and sporadic PD and risk can be 1.5 to 2 times higher in males compared to females. Adjusting for age should account for the increase in disease prevalence with increasing age, independent of genetic risk. Study designs typically include

age of onset for cases and age of sample collection or last known age with no evidence of disease-related symptoms for controls. For neurodegenerative disease, this typically results in a lower mean age in cases compared to controls, as in our case-control cohorts, which differ in mean age by approximately 10 years (age of onset was not available for 52/496 cases). In terms of covariate adjustment in a case-control logistic regression, the model would infer a negative effect for age on neurodegenerative disease risk, that younger individuals are more likely to develop disease. Differences in age between cases and controls and incorrectly modeling disease risk by age can substantially reduce power (Le Guen et al., 2021). Age was ultimately not included as a covariate in the primary analysis due to missingness and differences in available data between patient cohorts (age at disease onset, study enrollment, or death).

Individual sex was not available for all cases and controls when the primary analysis was conducted, however, once we were able to obtain this data, we performed SKAT-O with sex as a covariate and two genes (MAP2K3 and HKR1) dropped out from our significant results. The MAP2K3 association was driven by a single variant in 4 male patients, found in 1 female control and HKR1 by 5 variants found in 13 male/4 female cases and 11 male/17 female controls. Although variation in these genes appears to correlate with sex, it is not possible to ascertain the true association as it may represent sexual dimorphism or lack of power to detect association in females due to the limited sampling in our case cohort. As MAP2K3 was replicated at the gene level in the MSA cohort, we retained all significant genes in the unadjusted analysis while taking into consideration the potential bias of sex specific effects.

2.4.10. LINKAGE PATTERNS

Linkage disequilibrium (the non-random correlation between neighboring alleles) can be useful for fine mapping of common diseases and common variants as tag SNPs can capture correlation with SNPs that are not directly genotyped but can also affect the statistical power in association tests. SNPs with different MAFs have different LD properties (Pritchard and Przeworski, 2001). While

common methods to detect LD tend to perform poorly when rare variants are included (Turkmen and Lin, 2017), rare variants tend to have weaker correlation with other SNPs and are generally assumed to be independent. However, (Moutsianas et al., 2015) found power to be inversely related to the degree of LD between causal variants for gene-based association tests.

It has been shown that LD decreases with increasing physical distance (1000 Genomes Project Consortium et al., 2015), even within haplotype blocks (Shifman et al., 2003) but can extend over much larger regions than expected in the human genome. Haplotype blocks (loci in strong LD) have been reported between a few kb to more than 100 kb. (Reich et al., 2001) observed LD extending to 60 kb in a European sub-population (much less so in a Nigerian population). While LD tends to decrease as a result of population growth, strong, recurrent bottlenecks can result in long-range linkage disequilibrium (LRLD) between distant variants on the same chromosome. LD patterns can be challenging to assess even with comprehensive population data, particularly for LRLD, which can arise due to chance as a result of random sampling within a population or if the MAFs of a variant are very small compared to the sample size, biasing the haplotype frequency at equilibrium (Park, 2019).

We used the LDpair tool (Machiela and Chanock, 2015) to assess LD patterns of qualifying variants in publicly available haplotype data from 1000 Genomes populations (restricted to counts from the European populations: CEU, TSI, GBR, and IBS). Alleles are considered correlated if $r^2 > 0.1$.

For allele A and allele B with frequencies p_A and p_B, we expect to observe the AB haplotype at frequency $p_{AB} = p_A p_B$ if the 2 loci are independent, ie. in linkage equilibrium. If p_{AB} is higher or lower than the expected frequency, the alleles tend to be observed together and are in LD.

LD is calculated by constructing a 2x2 contingency table of haplotype counts and allele frequencies for the 2 variants being queried. The disequilibrium coefficient is thus $D_{AB} = p_{AB} - p_A p_B$. As D can

be difficult to interpret and varies with allele frequency, it is typically normalized to D' = D/D_{min} where $D_{min} = \min(p_A p_B, p_a p_b)$.

$$r^2 = \frac{D_{AB}{}^2}{pA(1-pA) * pB(1-pB)}$$

r^2=1 implies that the markers are in LD. A Chi-square test can be used to test whether the observed haplotype counts deviate from the expected values (H_0: no linkage disequilibrium): $X^2 = r^2N$, where N is the total number of chromosomes (ie. 2 x total individuals).

SNP 1	SNP 2	distance (Mb)	D'	r^2	pval
rs145504993 (PTPRS)	rs41303849 (ARHGEF1)	37.2	1.0	0.0	0.9602
rs145504993 (PTPRS)	rs75100029 (TEAD2)	44.6	1.0	0.0	0.9513
rs75100029 (TEAD2)	rs41303849 (ARHGEF1)	7.5	1.0	0.0	0.9311

Table 17: Population based linkage for top significant variants. All variants are in linkage equilibrium based on haplotype data from European individuals.

Based on the Chi-square p-values, the haplotype counts do not appear to deviate from the expected values, thus the null hypothesis cannot be rejected, suggesting that the variants are in linkage equilibrium. D'=1 implies that at least 1 expected haplotype combination is not observed. If both loci have very rare alleles, and they do not occur together in a haplotype block, it is possible that D'=1 because one of the haplotypes does not occur in the population and for r^2 to be small if the alleles for the three remaining haplotypes are not correlated. Rare SNPs tend to have higher pairwise D' values and lower pairwise r^2 values than common SNPs, especially for small sample sizes (Pe'er et al., 2006). This is because the minor allele may only appear on one chromosome and consequently be in perfect LD with that haplotype. When the population MAFs of one variant are small, this bias in haplotype frequency could generate LD by chance. For recently admixed populations, different allele frequencies in the subpopulations can results in LD between unlinked sites, but generally breaks down quickly (Pritchard and Przeworski, 2001). A nearby 'proxy' SNP (rs61757816) for PTPRS, within 5kb of rs145504993, appears to be in LE (r^2<0.1) with TEAD2: rs75100029 (D'=1, r^2=0.0001) and ARHGEF1: rs41303849 (D'=0.4821, r^2=0.0161). The top

significant variants do not appear to be in LD but may indeed not be independent due to long-range LD (LRLD). However, LD generally decays rapidly beyond a physical distance of a several hundred kilobases (kb) to a few megabases (Mb) in outbred human populations (Pritchard and Przeworski, 2001).

2.4.11. EDGETIC ANALYSIS

Edgotyping experiments are being carried out by Dr. Isabel Lam in collaboration with the Vidal lab at Harvard Medical School.

Cloning of human variants and edgotyping analysis

We selected missense variants of unknown significance (VUS) for edgetic analysis, including those from variant level tests, those in significant genes, and private variants. Amino acid changes for missense variants were obtained from SnpEff (v4.3t). Control variants for genes of interest were obtained from ClinVar and gnomAD (v2.1) databases. Of the 150 targeted genes, 108 were found in ClinVar, 42 of which harbored a germline SNV classified as "pathogenic" or "likely pathogenic". For genes with more than 50 control variants, strictly missense pathogenic were retained. Negative controls were defined as high frequency missense variants which are likely not to be disease causing and thus would not perturb disease specific PPIs. Qualifying common control variants were defined as nonsynonymous SNVs that passed all filters with MAF >= 5% in gnomAD exomes (v2.1), after excluding ClinVar variants. The human VUS were then mapped to the human ORFeome (protein-encoding open reading frames representing human genes). A variant is successfully matched when the reference nucleotide, reference amino acid, and alternate amino acid match at least one ORF (several ORFs may exist for the same gene).

Variants were cloned using a high-throughput site-directed mutagenesis pipeline, as described in (Sahni et al., 2015). Wild-type ORFs used as PCR templates were from the human ORFeome library and were in Gateway cloning-compatible donor vectors pDONR223. Site-directed mutations were incorporated by two-step PCR, in which the first round of PCRs generate two gene

fragments: one fragment encompassing the 5' attB site, ORF ATG start site, and the desired single nucleotide change incorporated during PCR reaction through the reverse primer, and the second fragment encompassing the desired single nucleotide change (incorporated through the forward primer during PCR), ORF stop codon, and 3' attB site. The second PCR step ("stitch" or "fusion" PCR) used the two primary PCR fragments as template and amplified using universal primers encompassing the attB sites on both ends of the ORF to generate the full-length ORF with the desired mutation. PCR products were subsequently cloned into the Gateway donor vector pDONR223 by BP reaction followed by bacterial transformation into DH5α cells and selected for spectinomycin resistance. Two clones per allele were selected and presence of the desired mutations in the plasmids was confirmed by PacBio sequencing.

Pairwise testing of protein-protein interactions by Y2H (edgotyping)

Sequence-confirmed variants will be subjected to edgotyping analysis with previously detected yeast two-hybrid (Y2H) interactors. First, wild-type ORFs will be re-tested by yeast two-hybrid (Y2H) for previously identified binary protein-protein interactions (PPIs) from the most recent human interactome. The wild-type interactions will be scored, and only interactions meeting specific criteria will be tested further in a wild-type and variants Y2H pairwise test against interactors. Criteria for wild-type interactions are Y2H growth score ≥2 with at least one interactor. In other words, the wild-type and variants pairwise test will be performed to compare variant and wild type PPI profiles for only ORFs that result in detectable PPI in the wild-type re-test. This pairwise test only screens for variants with no change or loss of known interactions and does not inform on variant-specific gain of interactions, since the screen is not proteome-wide. PPI profiles are compared between variants found in our PD patient cohort and the respective wild-type alleles.

For the variants pairwise test, wild-type and variant ORF entry clones will be subcloned into pDEST-DB (CEN-based) vector by Gateway LR reaction. The resulting plasmids express yeast Gal4 DNA-binding domain fusion proteins (DB-ORF). The LR products are then subsequently

transformed into bacterial competent cells DH5alpha and selected for ampicillin resistance. DB-ORF plasmids will be purified and transformed into haploid yeast Y8930 and selected on synthetic complete (SC) without leucine (SC-Leu) solid media to generate Y2H bait strains. The Y2H prey strains consisted of interactor ORF fusion proteins with Gal4 activation domain (AD-ORFs) in yeast strain Y8800. Yeast strains with AD-ORFs will be selected on SC media without tryptophan (SC-Trp). For Y2H pairwise test, 5ul of glycerol stocks of prey (DB-WT and variants) and bait (interactors) strains will be inoculated in 200ul selective media (SC-Leu or SC-Trp) in 96-well format and grown for 3 days at 30C. Prey and bait strains will be re-arrayed into the appropriate layout for the pairwise test by inoculating into fresh selection media, and grown for 2 days at 30C. Prey and bait strains are mated by combining 5ul of each strain in 180ul YPD, and grown overnight at 30C. The next day, diploids are enriched by inoculating 10ul of overnight YPD cultures in 180ul SC-Leu-Trp double selection media. Cultures will be grown overnight and robotically spotted onto SC-Leu-Trp-His + 1mM 3AT (modifications for alleles pairwise test: also spot on 10mM 3AT, and dilute overnight cultures 1:3 in water before spotting) and control SC-Leu-Trp-His agar plates. Yeast growth will be monitored at 3, 4, and 5 days after spotting. To test for DB-ORF autoactivators that could activate *GAL1::HIS3* reporter gene expression in the absence of an AD-ORF plasmid, all DB-ORF yeast strains will be mated with yeast strain harboring empty pDEST-AD vectors (pDEST-AD-CEN, pDEST-AD-2micron N-ter, or pDEST-AD-2micron C-ter). DB-ORF strains that grow on SC-Leu-Trp-His + 3AT media when mated with empty AD strains are identified as auto-activators and scored accordingly in the Y2H analysis.

RESUME COMPLET EN FRANÇAIS

ABSTRACT

L'agrégation de protéines spécifiques dans les neurones et la glie constitue la pathologie caractéristique des maladies neurodégénératives comme la maladie d'Alzheimer (MA) et la maladie de Parkinson (MP). Cependant, on ne sait toujours pas si les différentes maladies cliniques liées au mauvais repliement d'une même protéine ont des facteurs de risque génétiques communs et si les différentes pathologies liées à l'agrégation des protéines sont mécaniquement liées ou distinctes. Dans des études précédentes, nous avons établi des modèles de levure amyloïde-bêta et alpha-synucléine qui récapitulent les pathologies moléculaires de la MA et de la MP, respectivement. Pour répondre à ces questions, nous avons effectué un séquençage ciblé de l'exome chez 500 patients atteints de quatre synucléinopathies distinctes : MP, MP avec démence (MPD), maladie à corps de Lewy (MCL) et atrophie de systèmes multiples (ASM). En plus des gènes connus de la MA et de la MP, nous avons séquencé les homologues humains des modulateurs génétiques de la toxicité de l'alpha-synucléine et de la bêta-amyloïde identifiés dans des cribles à l'échelle du génome dans la levure, y compris un large éventail d'orthologues afin d'identifier les moteurs génétiques dans divers contextes spécifiques aux patients, aux cellules et aux maladies. L'ensemble relativement restreint de 430 gènes nous a permis de concentrer la puissance statistique sur des cibles plus probables. À titre de contrôle, nous avons comparé le poids des variantes rares à une cohorte de plus de 2 500 personnes âgées en bonne santé, sans antécédents déclarés de cancer, de maladie cardiovasculaire ou de maladie neurodégénérative. Les variants des réseaux génétiques de la bêta-amyloïde et de l'alpha-synucléine étaient enrichis dans les différentes synucléinopathies par rapport aux témoins âgés sains, tout comme les variants des gènes connus de la MP et de la MA. Les associations étaient indépendantes de la pathologie concomitante de la MA. Des tests au niveau des gènes, y compris un test de tendance développé pour l'analyse des variants rares, ont confirmé et étendu ces résultats dans des cohortes supplémentaires d'ASM et de la MP. Nos données impliquent des facteurs génétiques communs à des synucléinopathies distinctes et démontrent des réseaux génétiques convergents dans les protéinopathies bêta-amyloïde et alpha-synucléine chez l'homme. Nous pensons que nos approches inter-espèces continueront à fournir des informations importantes sur des maladies complexes dont les mécanismes biologiques sont conservés au cours de l'évolution, et qu'elles permettront peut-être de répondre aux promesses thérapeutiques de l'ère post-génomique.

INTRODUCTION ET MOTIVATION

La maladie de Parkinson est la deuxième maladie neurodégénérative la plus fréquente après la maladie d'Alzheimer et a été décrite pour la première fois il y a plus de 200 ans. Cependant, aucune intervention n'existe pour ralentir ou arrêter la dégénérescence progressive et irréversible associée. L'agrégation de protéines spécifiques dans les neurones et la glie constitue la pathologie caractéristique des maladies neurodégénératives comme la maladie d'Alzheimer (MA) et la maladie de Parkinson (MP). Cependant, on ne sait toujours pas si les différentes maladies cliniques liées au mauvais repliement d'une même protéine ont des facteurs de risque génétiques communs et si les différentes pathologies liées à l'agrégation des protéines sont mécaniquement liées ou distinctes. Il est essentiel de comprendre les déterminants des phénotypes cellulaires convergents et spécifiques à la maladie pour développer des thérapies ciblées.

La prévalence mondiale des personnes atteintes de la MP a plus que doublé au niveau mondial au cours des 20 dernières années (GBD 2016 Parkinson's Disease Collaborators, 2018) et devrait doubler au cours des 20 prochaines années pour atteindre plus de 12 millions de cas (Dorsey et Bloem, 2018). Malgré une pathobiologie commune aux protéinopathies (maladies neurodégénératives caractérisées par l'agrégation d'une protéine spécifique), la nature exacte de la

protéine déposée et la vulnérabilité des différentes régions du cerveau sont spécifiques à chaque maladie. La compréhension des mécanismes moléculaires et cellulaires fondamentaux du vieillissement et de leur rôle dans l'apparition et la progression des maladies neurodégénératives est un défi majeur pour le développement de traitements efficaces.

En 1997, le premier gène (SNCA) associé à la maladie a été identifié (Polymeropoulos et al., 1997) et la protéine qu'il produit, l'α-synucléine, a été identifiée comme le principal composant des agrégats cytoplasmiques pathologiques caractéristiques (corps de Lewy) dans la MP et la maladie à corps de Lewy (MCL) (Spillantini et al., 1997). Peu après, en 1998, l'alpha-synucléine a été identifiée comme le principal composant des inclusions filamenteuses dans l'atrophie systémique multiple (ASM) et la MP, la MCL, et l'ASM sont devenues collectivement connues sous le nom de synucléinopathies (Spillantini et al., 1998). Les synucléinopathies sont toutes des maladies neurodégénératives à apparition tardive, la majorité des cas ne présentant aucun risque familial. Bien que la présence d'inclusions α-syn soit une caractéristique commune des synucléinopathies, le dépôt de protéines est spécifique à chaque maladie. Différentes régions neuroanatomiques et populations cellulaires présentent une vulnérabilité distincte à l'agrégation des protéines, au dysfonctionnement et à la mort cellulaire, ce qui entraîne une diversité phénotypique. Ces maladies peuvent être distinguées par les populations neuronales affectées, les inclusions cellulaires, et la présentation clinique.

La majorité des synucléinopathies sont sporadiques, comprenant 85 % des cas de maladie de Parkinson et encore plus dans le cas de l'ASM, où seuls quelques pour cent des cas semblent être familiaux (Stefanova et al., 2009). La maladie de Parkinson n'est pas un diagnostic unique et comprend un éventail de symptômes, ce qui complique l'élucidation des associations génétiques. Les facteurs de risque génétiques identifiés dans les cas familiaux ont été observés dans la maladie sporadique plus courante et ont permis de découvrir des facteurs de risque génétiques. Les formes monogéniques de la MP représentent environ 30 % des cas familiaux et 3 à 5 % des cas sporadiques (Klein et Westenberger, 2012).

La maladie de Parkinson est une maladie génétiquement complexe pour laquelle aucune mutation génétique unique ne présente une pénétrance génétique de la maladie de 100 %. Même dans le cas de variants rares à forte pénétrance, la pénétrance de la maladie est probablement affectée par d'autres facteurs génétiques et non génétiques. De nombreux facteurs de risque génétiques ont été identifiés dans la MP familiale et sporadique, la plus grande étude d'association pangénomique (GWAS) réalisée à ce jour ayant identifié 90 loci pangénomiques indépendants comprenant des variants rares et communs, ces derniers expliquant 12 à 36 % de l'héritabilité, selon les estimations de la prévalence de la maladie (0,5 à 2 %) (Nalls et al., 2019). Les estimations de puissance prédisent que l'augmentation de la taille des échantillons à plus de 99 000 cas permettra de découvrir des variants de risque supplémentaires. Les mutations causales connues, définies comme telles par leur occurrence dans les familles touchées ou les cas à début précoce, semblent être rares dans la plupart des cas de MP. L'identification de gènes dans les cas familiaux et de début de maladie a permis de mieux comprendre les mécanismes pathogéniques probables de la maladie. Cependant, les mutations causales présentent une pénétrance qui dépend de l'âge et qui est probablement due à d'autres facteurs de risque génétiques et environnementaux, ainsi qu'à la pathobiologie liée à l'âge. L'un des défis fondamentaux de la génétique de la maladie de Parkinson est de comprendre les mécanismes moléculaires complexes qui déterminent l'hétérogénéité de l'apparition et de la progression de la maladie. Malgré les progrès réalisés dans la découverte des facteurs de risque génétiques de la maladie de Parkinson, les mécanismes moléculaires qui sous-tendent la dégénérescence cellulaire dans le cerveau des personnes atteintes de la maladie restent mal compris. L'établissement d'un réseau fonctionnel de gènes et d'éléments régulateurs de gènes associés à la maladie monogénique et sporadique peut aider à identifier les mécanismes pathologiques sous-jacents aux variants associées à la maladie, ce qui améliorera notre compréhension du risque génétique.

MODELES CELLULAIRES DE PROTÉOTOXICITÉ

Malgré une meilleure compréhension des gènes et des mutations impliqués dans les maladies neurodégénératives, il n'existe aucun traitement modificateur de la maladie. Un traitement permettant de prévenir, de ralentir ou d'arrêter la progression de la maladie ou d'inverser la neuropathologie est l'objectif le plus important de la recherche sur les maladies neurodégénératives. Il est devenu évident que l'analyse génétique humaine ne peut à elle seule établir pleinement les associations causales entre les variants rares et les phénotypes des maladies. Une plateforme de découverte réussie nécessite l'intégration d'approches basées sur l'homme et sur des organismes modèles afin d'identifier les gènes, les voies et les composés qui peuvent cibler la pathologie liée à l'α-synucléine. L'α-synucléine, le principal constituant des agrégats associés à la maladie de Parkinson, est une grande protéine exprimée principalement dans le cerveau qui interagit avec les membranes lipidiques et est impliquée dans la fonction synaptique. Dans les neurones des patients atteints de synucléinopathie, elle adopte des conformations amyloïdes, formant des agrégats protéiques cytoplasmiques appelés corps de Lewy (LB). Des systèmes modèles ont établi un lien entre des interactions membranaires anormales et des altérations du trafic des vésicules et la toxicité associée à l'α-synucléine (Auluck et al., 2010). Des cribles génétiques non biaisés à l'échelle du génome et des cribles génétiques candidats ont été exécutés avec succès dans des systèmes modèles pour identifier les modulateurs génétiques des protéotoxicités. Ces systèmes, y compris la levure *Saccharomyces cerevisiae* (Khurana et Lindquist, 2010), la mouche à fruits *Drosophila melanogaster* (Shulman et al., 2003) et le nématode *Caenorhabditis elegans* (Teschendorf et Link, 2009), ont permis de découvrir des éléments biologiques pertinents pour la maladie, y compris des homologues de facteurs de risque de maladie humaine connus.

L' espèce de levure *S. cerevisiae* a été le premier eucaryote dont le génome de 13 Mb a été séquencé en 1996 et est devenu depuis l'un des organismes modèles les plus étudiés. La levure est unicellulaire, son temps de génération est court (1,5 à 3 heures) et son ADN est facilement transformé, ce qui la rend propice aux cribles génétiques à haut débit et aux cribles de petites molécules. Sa redondance génétique réduite facilite l'élimination efficace des gènes par rapport à des modèles plus complexes (Khurana et Lindquist, 2010). Les données génétiques humaines relatives aux protéinopathies neurodégénératives impliquent fortement des voies cellulaires qui sont parmi les plus conservées au cours de l'évolution eucaryote, notamment l'homéostasie et le contrôle de la qualité des protéines, le trafic des protéines, la biologie de l'ARN et la fonction mitochondriale (Brás et al., 2015 ; Guerreiro et al., 2015). Il y a bien sûr des limites aux modèles de neurodégénérescence chez la levure, notamment l'absence de caractéristiques neuronales spécialisées. Une interaction génétique au sein d'une cellule de levure unicellulaire peut se dérouler différemment chez l'homme, en fonction du type de cellule et du bagage génétique ou du contexte de la maladie. Cependant, l'objectif n'est pas de reproduire une maladie humaine, mais plutôt la pathologie cellulaire précoce causée par la toxicité de l'alpha-synucléine.

Les résultats des cribles à haut débit peuvent ensuite être validés dans des systèmes modèles neuronaux afin de tenir compte des différences biologiques entre la levure et les maladies humaines complexes. Le modèle d'alpha-synucléine de la levure récapitule les phénotypes détectés dans d'autres systèmes modèles exprimant l'a-syn humaine (Cooper et al., 2006 ; Feany et Bender, 2000 ; Masliah et al., 2000) et les gènes et petites molécules identifiés dans la levure ont été validés dans des neurones dérivés de patients présentant les mêmes phénotypes cellulaires (Chung et al., 2013). L'expression d'a-syn et d'a-beta dans la levure entraîne une cytotoxicité robuste et les cribles génétiques dans les modèles cellulaires de protéotoxicité sont évolutifs et faciles à mettre en œuvre (Khurana et Lindquist, 2010). Chacun des plus de 5000 cadres de lecture ouverts (ORF) du génome de la levure est surexprimé, puis la croissance et la viabilité sont évaluées. Lors de précédents cribles génétiques de levure non biaisés, des gènes englobant ~85 % du protéome de la levure ont été co-exprimés individuellement avec chacune de ces protéines toxiques afin d'identifier les modificateurs génétiques de la toxicité de l'a-beta (Treusch et al., 2011) et de l'a-syn (Yeger-Lotem et al., 2009),

respectivement. Bien que plus de 60 % des gènes de levure aient des homologues humains ou au moins des domaines conservés (Smith et Snyder, 2006), leur attribution inter-espèces est difficile.

TRANSPOSITION DES RÉSEAUX MOLÉCULAIRES ENTRE LES ESPÈCES

Les gènes de levure ont fréquemment de nombreux homologues chez l'homme en raison de la diversification dans un organisme multicellulaire. Une approche computationnelle, appelée TransposeNet (Khurana et al., 2017), a été développée pour cartographier le réseau des modificateurs de la toxicité de l'alpha-synucléine de la levure à l'homme en se basant sur la séquence, la structure des protéines et les interactions moléculaires connues. Les paires de protéines levure-homme sont déterminées en fonction de la similarité des séquences protéiques, des profils évolutifs et des caractéristiques structurelles afin d'identifier les homologues éloignés, ainsi que la fonction des protéines. Seuls les homologues exprimés dans le tissu cérébral humain sur la base des données d'expression GTEx sont retenus (Consortium GTEx, 2013). Les interactions génétiques et physiques connues entre les homologues humains sont ensuite utilisées pour créer un réseau, qui est augmenté par l'interactome beaucoup plus riche de la levure. La dernière étape consiste à appliquer l'algorithme de la forêt de Steiner pour connecter les nœuds gènes/protéines par le biais des interactions protéines-protéines ou génétiques les plus fiables, ainsi qu'à insérer des nœuds prédits (c'est-à-dire des gènes ne figurant pas dans les homologues prédits), augmentant ainsi le réseau en incluant des gènes sans homologues connus dans la levure.

ANALYSES DES VARIANTS RARES DANS LES SYNUCLÉINOPATHIES

Pour tester directement la pertinence de ces réseaux de gènes dans le contexte d'une maladie humaine et d'un patient spécifique, nous avons ciblé un large éventail d'orthologues de ces gènes chez des patients atteints de synucléinopathies : MP, MCL, et AMS (**Figure 1**). Nous avons réalisé un séquençage ciblé de l'exome des orthologues des gènes modificateurs a-beta et a-syn ainsi que des gènes connus de la maladie d'Alzheimer, de la maladie de Parkinson et des maladies liées à l'ataxie. Notre objectif était d'identifier des variants rares (fréquence d'allèle mineur (MAF) <1%) dans les réseaux génétiques a-beta et a-syn afin d'évaluer leur contribution relative au risque de maladie et de déterminer s'il existe des facteurs génétiques communs à diverses synucléinopathies. Nous avons émis l'hypothèse que la réduction de l'espace de recherche par séquençage à un ensemble de gènes connus *a priori* pour être enrichis en facteurs de risque connus et en modificateurs de la toxicité de l'alpha-synucléine nous permettrait de trouver des signaux significatifs même dans une cohorte de taille modeste.

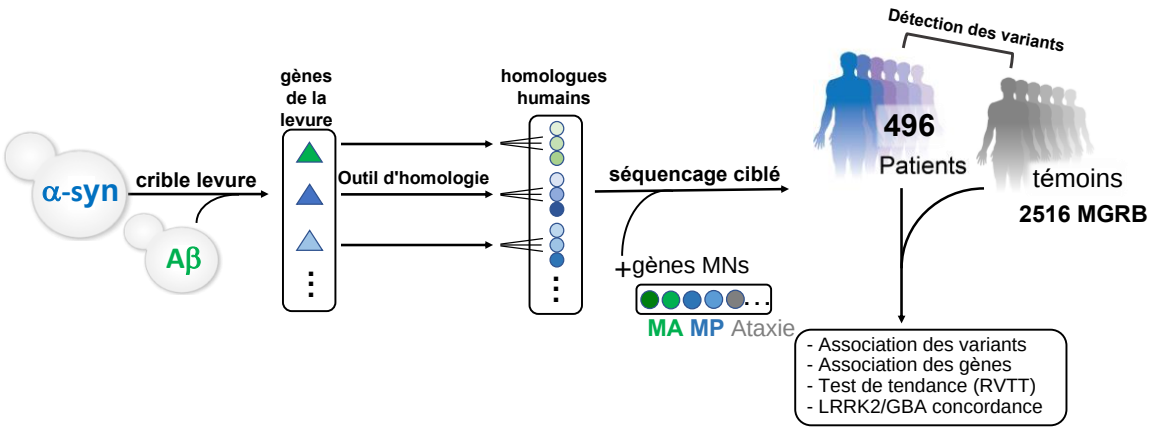


Figure 1 : Schéma de l'étude. Les gènes découverts dans les modèles de protéotoxicité cellulaire ont été transposés dans un réseau humain et ces homologues humains ainsi que gènes connus de maladies neurodégénératives (MNs) ont été séquencés dans une cohorte de patients atteints de synucléinopathie.

Les variants rares représentent un défi majeur pour les études d'association de pathologies humaines complexes. Les recherches sans hypothèse nécessitent un nombre écrasant d'individus pour atteindre la puissance statistique nécessaire à l'association de ces variants. Les études

d'association à l'échelle du génome ont identifié plus de 90 loci associés à la maladie de Parkinson (**Figure 2**) qui expliquent environ 16 à 36 % de l'héritabilité (Kunkle et al., 2019 ; Nalls et al., 2019). Cependant, même les études à l'échelle de la biobanque sont sous-puissantes pour détecter les variants rares. Afin de réduire l'espace de recherche pour détecter les signaux de variants rares, nous nous sommes concentrés sur les variants exoniques dans un ensemble de gènes ciblés des homologues humains des modificateurs de la protéotoxicité.

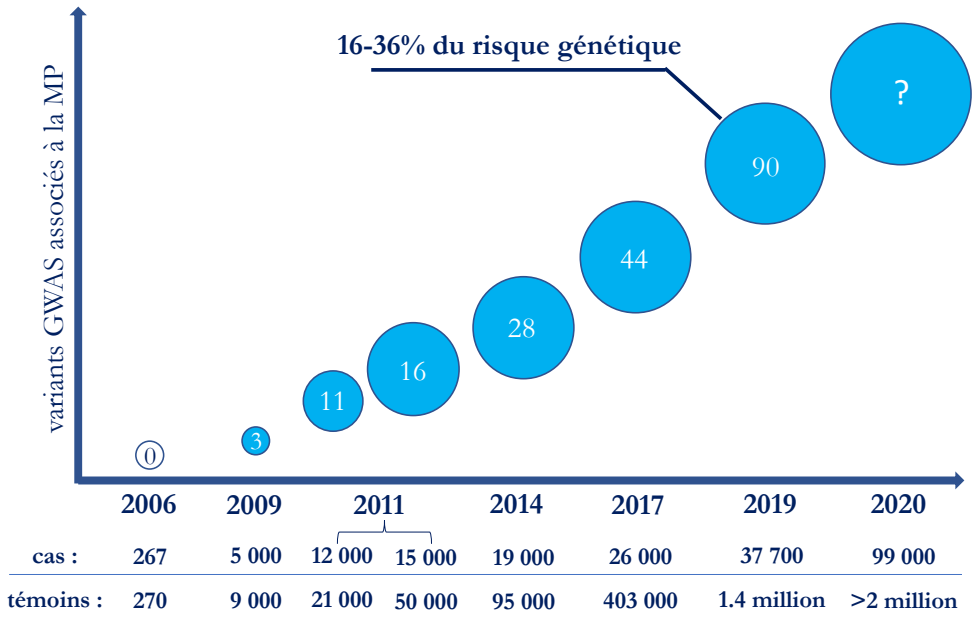


Figure 2 : Depuis le début de l'ère des GWAS en 2005, des cohortes cas-témoins de plus en plus importantes ont identifié un nombre croissant de loci génomiques associés à la MP. Les variants communs expliquent environ 16 à 36 % de l'héritabilité de la maladie. Image adaptée de (Blauwendraat et al., 2020).

Nous avons effectué un séquençage ciblé de l'exome chez 500 patients atteints de quatre synucléinopathies distinctes : MP, MP avec démence (MPD), la maladie à corps de Lewy (MCL) et atrophie de systèmes multiples (AMS). Dans des études précédentes, nous avons établi des modèles de levure amyloïde-bêta et alpha-synucléine qui récapitulent les pathologies moléculaires de la MA et de la MP, respectivement. En plus des gènes connus de la MA et de la MP, nous avons séquencé les homologues humains des modulateurs génétiques de la toxicité de l'alpha-synucléine et de la bêta-amyloïde identifiés dans des cribles à l'échelle du génome dans la levure, y compris un large éventail d'orthologues afin d'identifier les moteurs génétiques dans divers contextes spécifiques aux patients, aux cellules et aux maladies. L'ensemble relativement restreint de 430 gènes nous a permis de concentrer la puissance statistique sur des cibles plus probables. À titre de contrôle, nous avons comparé la charge des variants rares à une cohorte de plus de 2 500 personnes âgées en bonne santé, sans antécédents déclarés de cancer, de maladie cardiovasculaire ou de maladie neurodégénérative qui font parties du Medical Genome Reference Bank (MGRB). La sélection d'échantillons témoins appropriés est essentielle, tant en termes de critères d'inclusion/exclusion que de plateforme de séquençage, en particulier pour les maladies complexes liées à l'âge. L'inclusion d'individus à risque mais ne présentant pas encore de pathologie de la maladie en tant que témoins peut réduire la puissance des études d'association cas-témoins.

DÉSCRIPTIONS DES COHORTES ET CONTROLE DE QUALITÉ

La croissance récente de la population humaine a conduit à un excès de variations rares (Keinan et Clark, 2012) et la majorité des variants du génome humain sont rares (MAF < 0,05 %), soit 76 % (1000 Genomes Project Consortium et al., 2015). Les améliorations des technologies de séquençage et des méthodes de calcul ont rendu possible l'analyse de ces variants. Cependant, la plupart des

procédures de contrôle de qualité standard pour les études d'association basées sur le génome ont été développées pour le séquençage du génome entier, avec un accent sur les variants communs et ont une application limitée pour les variants rares. La détection conjointe d'échantillons (« joint calling ») peut surmonter de nombreux des biais techniques lors de la combinaison de données provenant de différentes études ou en testant l'association de variants rares en modélisant les lectures de séquençage sans appel de génotype (Hu et al., 2016). Le contrôle de la qualité est une étape essentielle dans les études d'association génétique. Nous avons effectué un filtrage au niveau des individus et des variants en adaptant ces méthodes à notre ensemble de données, notamment en supprimant les valeurs aberrantes de la population, les échantillons présentant une mauvaise qualité de séquençage et en estimant la parenté pour exclure les individus potentiellement apparentés.

Tous les cas et les contrôles ont été signalés comme étant d'ascendance européenne. Cependant, bien qu'il soit difficile d'ajuster la structure de la population étant donné notre petite région cible (~1Mb), nous avons appliqué l'analyse en composantes principales (ACP) aux données de séquençage publiquement disponibles pour les échantillons HapMap (1000 Genomes Project Consortium et al., 2015) et notre cohorte (cas, contrôles HapMap CEU et MGRB). La performance de l'ACP dépend fortement de la structure de la population et de la distribution du risque sous-jacente, et son utilité pour corriger la stratification de la population pour les tests d'association à variant rare n'est pas claire. Dans de tels cas, l'ACP peut être utilisée pour guider l'appariement des cas et des témoins. Moins de 1000 SNP sont restés après le filtrage, cependant tous les échantillons de notre cohorte se sont étroitement regroupés avec les populations européennes HapMap et avec les échantillons de contrôle MGRB. Une analyse en composantes principales basée sur le séquençage du génome entier des échantillons du MGRB a révélé 53 échantillons ayant une ascendance non-européenne potentielle. Étant donné la sensibilité de l'analyse des variants rares à la structure de la population, ces échantillons ont été exclus de l'analyse. Les individus présentant un excès d'hétérozygotie et une forte absence de génotype ont également été exclus.

Les variants ont été filtrés pour inclure ceux qui avaient une couverture >=10X dans >95% de tous les échantillons appelés conjointement, un taux d'appel de génotype > 95% dans les exomes ciblés, les génomes entiers et dans tous les échantillons, une valeur p HWE > 10E-6 dans les échantillons de contrôle internes HapMap ou MGRB et les sites dont la profondeur est supérieure à 2x la profondeur moyenne de tous les variants ont été exclus car ils sont susceptibles d'être des artefacts de séquençage.

Lors du choix d'un seuil de fréquence des allèles mineurs (MAF), il est raisonnable de supposer qu'une mutation génétique à pénétration complète ne devrait pas être plus fréquente dans la population que la maladie qu'elle provoque. Cependant, il n'existe pas de variant causale complètement pénétrante dans la MP et le risque à vie augmente avec l'âge. L'établissement d'une FA au-delà de laquelle un variant ne pourrait pas causer de manière crédible la MP est un défi, d'autant plus qu'il a été démontré que les variants communs jouent un rôle substantiel dans le risque de maladie, expliquant ~22 % de l'héritabilité (Nalls et al., 2019). Compte tenu des différences de technologie de séquençage entre les cas et les témoins, nous ne pouvions pas inclure les variants ultra-rare (ceux qui apparaissent chez un seul individu). Nous avons donc appliqué le seuil standard de MAF « rare » de 1 % dans les cas, les contrôles et les échantillons gnomAD non finlandais non neuro pour évaluer le risque de maladie spécifique à la population.

CHARGE DE VARIANTS RARES CHEZ LES PORTEURS CONNUS

L'occurrence de mutations causales connues de la MP est rare dans la population de la MP (<2%) (Tran et al., 2020). Les porteurs de mutations pathogènes de la MP conférant un risque élevé sont généralement exclus des analyses cas-témoins, mais cette information n'est souvent connue qu'après le séquençage. Même pour les facteurs de risque forts de la MP, notamment les mutations LRRK2 et GBA de faible fréquence, la pénétrance augmente avec l'âge et est très variable. Des

variants de risque supplémentaires chez des patients présentant des variants pathogènes cliniquement pertinentes peuvent révéler des facteurs de risque génétiques qui influencent la pénétrance. Comme les cas présentant des mutations pathogènes connues peuvent abriter des variants de risque supplémentaires, nous ne les avons pas exclus de l'analyse primaire. Nous avons en outre vérifié s'il existait des variants statistiquement associées à ces trois variants en effectuant un test de somme allélique de cohorte (Morgenthaler et Thilly, 2007) (test exact unilatéral de Fisher) pour comparer la fréquence d'enrichissement en variants rares supplémentaires (MAF<0,01) chez les cas porteurs de la mutation LRRK2 G2019S (n=9) ou GBA N370S (n=13) par rapport à tous les autres cas. Étant donné l'enrichissement des gènes liés à la maladie d'Alzheimer dans notre analyse, nous avons considéré les cas présentant des variants cliniques associées à la maladie d'Alzheimer ou à la maladie de Parkinson.

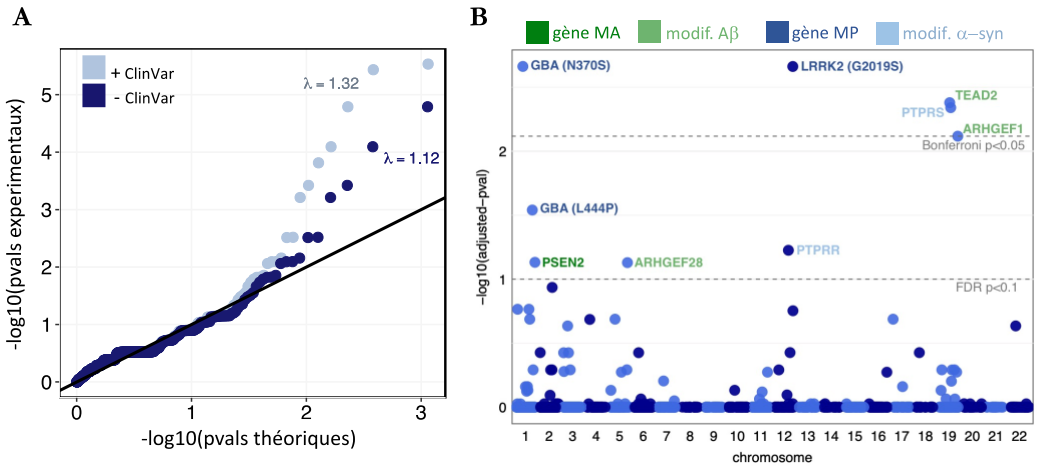
Tous les gènes pathogènes associés à la maladie d'Alzheimer selon la base de données ClinVar, à l'exception de deux gènes (CSF1R et GRN), ont été inclus dans nos gènes cibles (APOE, APP, MAPT, PSEN1, PSEN2 et VCP). Cependant, aucune mutation ClinVar pathogène/probablement pathogène associée à la MA n'a été trouvée dans notre jeu d'appels filtré. Les allèles APOE (rs429358 et rs7412) ont été détectés mais n'ont pas passé le seuil du taux d'appel de génotype de 95%. De plus, les deux variants ont un MAF>1% dans la population européenne ainsi que dans notre cohorte, ce qui ne permet pas d'utiliser les filtres d'inclusion pour l'analyse des variants rares. Au total, 11 variants pathogènes ou probablement pathogènes de la maladie de Parkinson ont été détectés dans nos cas dans GBA, LRRK2 et PARK2. Nous avons profité des données sur les variants au niveau individuel pour tester s'il existe des variants statistiquement associés à deux variants pathogènes connus : GBA rs76763715 (N370S) et LRRK2 rs34637584 (G2019S). La pénétrance génétique du phénotype de la MP chez les porteurs de ces mutations est très variable (10-50%). La pénétrance réduite et les phénocopies dans les familles avec une association apparente entre GBA et Parkinson, telles que démontrées par les porteurs de mutations GBA non symptomatiques et les non porteurs avec MP, respectivement, peuvent être expliquées par des facteurs de risque génétiques supplémentaires qui influencent la pénétrance de la maladie (Aslam et al., 2021 ; Blauwendraat et al., 2020 ; Gan-Or et al., 2015). Nous avons effectué un test de sommes alléliques de cohorte (CAST), entre les 9 porteurs de G2019S par rapport à tous les autres cas, mais nous n'avons pas observé de facteurs de risque génétiques supplémentaires. La même analyse chez les porteurs de GBA-N370S par rapport à tous les autres cas a identifié 2 variants faux-sens significativement enrichis (p<0,05 ajusté par Bonferroni) dans : PTPRB (rs202134984), et FREM2 (rs150928081). Nous avons ensuite regroupé les génotypes pour compter le nombre total d'individus présentant au moins un variant dans chaque gène. Un gène (MICAL3) était significativement enrichi en variants rares) et 2 autres gènes (PTPRB et PRKCD) présentaient de petites valeurs p nominales mais n'ont pas passé le seuil strict ajusté par Bonferroni. Comme les mutations pathogènes connues ne représentent qu'une fraction du risque génétique de la MP, l'élucidation des effets combinatoires est cruciale pour notre compréhension de la façon dont les variants de risque génétique contribuent à la pathologie de la maladie. Des études à l'échelle du génome comparant un nombre suffisant de porteurs de mutations malades et sains peuvent faciliter l'identification des modificateurs génétiques de la pénétration de la maladie. Leur identification peut avoir des implications importantes pour les essais cliniques en cours et à venir des thérapies modifiant la GBA et LRRK2, ainsi que pour la compréhension du risque de développer une MP chez les porteurs de mutations.

ANALYSE D'ASSOCIATION DE VARIANTS RARES

L'évaluation de la contribution d'un seul variant au risque de maladie implique la comparaison relativement simple des fréquences des allèles ou des génotypes sur le site génomique chez les cas par rapport aux témoins. Cependant, les tests d'association classiques (par exemple, le test du chi carré, le test de tendance de Cochran-Armitage, le test exact de Fisher et le test de Wald) ne sont pas appropriés pour les variants rares en raison de leurs faibles fréquences et de leur nombre plus

élevé que celui des variants communs, ce qui peut conduire à des résultats fallacieux lorsque les nombres observés sont trop faibles. L'analyse d'un seul variant à l'aide de tests d'association de variants rares est insuffisamment puissante mais peut être utile lorsque seules des statistiques sommaires sont disponibles, comme c'était le cas dans les premières étapes de cette étude. Nous avons effectué un test exact de Fisher unilatéral en utilisant le nombre d'allèles chez les cas et les témoins âgés sains du MGRB. Le test de tendance de Cochran-Armitage (CA) est couramment utilisé dans les études d'association pangénomique pour comparer les différences de fréquences alléliques entre les cas et les témoins. Bien que conservateur, ce test est robuste aux petites tailles d'échantillon et garantit le contrôle de l'erreur de type I. Les variants pathogènes sans interprétations contradictoires dans la base de données ClinVar ont été exceptionnellement retenues afin d'évaluer l'inflation génomique. Nous avons effectué un test exact de Fisher du nombre d'allèles entre les cas et les témoins pour les SNV délétères prédits avec et sans variants cliniquement pathogènes (selon la base de données ClinVar). Ce filtre a entraîné la plus grande réduction de l'inflation génomique.

En raison de la nature discrète des variants rares et de la non-uniformité des valeurs p non ajustées, nous avons effectué un ajustement FDR modifié pour la correction des tests multiples via un rééchantillonnage basé sur la permutation pour estimer la distribution nulle empirique. Au total, 1020 variants admissibles ont été testés, avec 5 variants significatifs dans 5 gènes (LRRK2, GBA, TEAD2, PTPRS et ARHGEF1) passant la stricte valeur p ajustée de Bonferroni et 4 autres passant un seuil FDR modifié (**Figure 3**).



C

CHROM:POS:REF:ALT	rsID	gene	OR	pval	Bonferroni
12:40734202:G:A	rs34637584	LRRK2	16.45	3.30E-06	0.00337
1:155205634:T:C	rs76763715	GBA	7.10	4.27E-06	0.00435
19:49852030:C:G	rs75100029	TEAD2	15.25	1.23E-05	0.01254
19:5240271:G:A	rs145504993	PTPRS	20.33	1.79E-05	0.01824
19:42392819:C:G	rs41303849	ARHGEF1	11.28	3.74E-05	0.03813

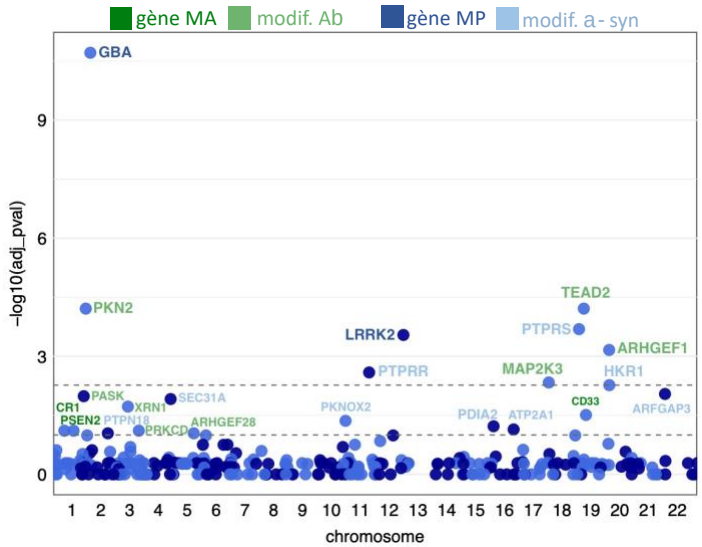
Figure 3 : (A) Graphique quantile-quantile des valeurs p observées par rapport aux valeurs p attendues du test exact de Fisher au niveau des variants pour l'association entre les cas et les témoins MGRB et leur inflation génomique respective. (B) Graphique de Manhattan des valeurs p ajustées par le FDR (rééchantillonnage cas-témoins basé sur la permutation) de l'association au niveau des variants. Les lignes pointillées indiquent les seuils de signification ajustés par Bonferroni et FDR. Les gènes significatifs associés à la MA et à la MP sont mis en évidence en vert et en bleu, respectivement. (C) Variants significatifs (Bonferroni p<0,05), qui ont émergé du test d'association cas vs MGRB.

ANALYSE D'ASSOCIATION AU NIVEAU DES GÈNES

Les analyses au niveau des variants sont généralement insuffisamment puissantes pour l'analyse des variants rares. Il existe de multiples méthodes pour agréger les variants rares et tester l'effet

cumulatif par gène ou région génomique pour l'association avec une maladie, mais les performances dépendent largement de l'architecture de la maladie. Comme celle-ci est souvent inconnue, il peut être utile d'appliquer un test combiné ou plusieurs méthodes et d'ajuster les valeurs de p sur l'ensemble des tests/méthodes. La méthode la plus courante pour regrouper les variants rares dans les analyses génétiques de population est au niveau du gène, généralement par le biais d'un test de regroupement, d'un test combiné multivarié et de regroupement, ou d'un « sequence kernel association test » (SKAT). Chaque méthode suppose des hypothèses différentes sur le modèle génétique sous-jacent et peut être puissante dans différents scénarios, notamment en ce qui concerne la présence de variants bénins ou de directions différentes de l'effet des variants causaux. De nombreuses méthodes de regroupement de variants ne sont pas robustes à l'inclusion de variants neutres ou protecteurs, ce qui peut entraîner une perte de puissance substantielle. Nous avons appliqué la méthode SKAT-O, un test combiné de regroupement et de composante de variance, qui est statistiquement efficace quels que soient la direction et l'effet des variants testés (Lee et al., 2012). Cette analyse a été réalisée en regroupant les 1020 variants délétères filtrés dans leurs 293 gènes respectifs. Au total, 9 gènes ont étés enrichis en variants rares chez les patients atteints de synucléinopathie par rapport à la cohorte de contrôle âgée et saine (MGRB) : GBA, PKN2, TEAD2, PTPRS, LRRK2, ARHGEF1, PTPRR, MAP2K3 et HKR1 (**Figure 4**). Ces gènes sont associés aux réseaux génétiques de l'alpha-synucléine et de l'amyloïde-bêta. Nous avons également observé que les porteurs de variants dans les réseaux génétiques de l'amyloïde-bêta les plus importants ou dans les gènes connus de la MA ne présentaient pas nécessairement une pathologie concomitante de la MA.

A



B

Gène	Description	Rationale
GBA	Lysosomal hydrolase	gène MP
PKN2	Protein kinase N2	Homologue Ab modifier PKC1 (OE=Enh)
TEAD2	Transcription factor, Hippo signaling	Homologue Ab modifier Tec1 (OE=Supp)
PTPRS	Receptor tyrosine phosphatase, MAPK signaling	Homologue a-syn modifier Ptp2 (OE=Supp)
LRRK2	Vesicle trafficking, actin cytoskeleton	gène MP
ARHGEF1	RhoA/cytoskeleton signaling	Homologue Ab modifier Rom 1 (OE=Enh)
PTPRR	Receptor tyrosine phosphatase	Homologue a-syn modifier PTP2 (OE=Supp)
MAP2K3	Protein kinase, MAPK signaling	Homologue Ab modifier PBS2 (OE=Enh)
HKR1	Zinc finger protein	Homologue a-syn modifier Tda9 (OE=Supp)

Figure 4 : (A) Graphique de Manhattan des variants délétères (faux sens, non-sens et épissage) regroupés par gène. Les lignes pointillées représentent les seuils de valeur de p ajustés par Bonferroni (en haut) et FDR modifié (en bas). (B) Descriptions des gènes significatifs (Bonferroni p<0,05). OE : surexpression d'amyloïde-β ou de α-synucléine dans la levure ; Supp=suppresseur de toxicité ; Enh=amplificateur de toxicité.

PATHOLOGIE CONCOMITANT DE B-AMYLOID ET DE L'A-SYNUCLEIN

Un avantage de notre étude est que les patients ont été séquencés sur la base de leur protéinopathie et non sur la base de leur maladie. De plus, pour 151 d'entre eux, nous disposions de données neuropathologiques, ce qui nous a permis de nous demander si les variants les plus importants traversaient les délimitations de la maladie clinique. L'immunohistochimie sur des coupes de cerveau de cas hébergeant l'un des variants signficatifs (PTPRS:rs145504993) a récapitulé notre conclusion selon laquelle les signaux supérieurs traversent les frontières de la maladie clinique, ce qui suggère des facteurs de risque et des mécanismes génétiques partagés entre les différentes synucléopathies. Pour le cas présentant une pathologie MA, des dépôts de bêta-amyloïde étaient présents dans le cervelet, ce qui ne se produit que dans les cas de dépôt amyloïde avancé. Les modificateurs de la toxicité a-bêta semblent être associés au risque de synucléinopathie, indépendamment de la pathologie concomitante de la MA. La copathologie peut témoigner de la coexistence de différents processus pathologiques chez un même patient ou être due à un lien mécanistique entre différentes protéinopathies. L'analyse génétique humaine a largement identifié des facteurs de risque génétiques non chevauchants dans les protéinopathies, et les modèles animaux ont montré des preuves contradictoires. L'apparition de comorbidités dans les maladies neurodégénératives peut avoir des implications majeures pour les essais cliniques, qui se concentrent généralement sur des thérapies ciblant une seule pathologie spécifique. Comprendre l'interaction des différentes protéines dans les protéinopathies comorbides peut aider à distinguer les réponses cellulaires générales au mauvais repliement des protéines (c'est-à-dire la perturbation de la machinerie de la protéostase) des réponses liées à la fonction intrinsèque de la protéine mal repliée.

VALIDATION DANS DES COHORTES EXTERNES

La réplication des associations génotype-phénotype est une étape critique dans la découverte génétique et dans l'établissement de liens de causalité avec le résultat d'intérêt. Le manque de réplication des variants et/ou des gènes putatifs associés à des maladies génétiques complexes est un défi tant pour les variants communs que pour les variants rares (Ioannidis et al., 2001). En ce qui concerne les variants rares, l'importante divergence génétique récente des populations humaines signifie que les études de validation au sein ou entre différentes populations sont souvent irréalisables et que la puissance statistique pour identifier les variants rares diminue à mesure que la fréquence de l'allèle mineur diminue.

Pour surmonter certains des défis inhérents aux variants rares, nous avons utilisé une approche à seuil variable (Price et al., 2010) dans laquelle les variants sont regroupés à l'aide d'une méthode de regroupement combinée et multivariée (CMC: Combined and Multivariate Collapsing). Les tests à seuil adaptatif (plutôt que fixe) peuvent améliorer la puissance statistique et réduire l'influence des variants neutres et protecteurs en utilisant une approche basée sur la permutation pour déterminer le fréquence de l'allèle mineur (MAF) en dessous duquel les variants sont plus susceptibles d'être causals (Price et al., 2010). Pour chaque gène, différents seuils de MAF sont appliqués et le MAF qui donne la plus petite valeur de p est sélectionné comme le seuil optimal pour le gène en question. Les tests actuels basés sur les gènes ne sont pas facilement applicables aux ensembles de gènes car ils ne testent que la présence ou l'absence d'un variant, alors que la plupart des individus sont susceptibles de porter au moins une mutation dans un ensemble de gènes, simplement par hasard. Par conséquent, plutôt que de considérer la présence ou l'absence de variants, nous avons tiré parti de la fréquence d'apparition de variants qualifiants rares dans notre ensemble de gènes d'intérêt pour évaluer l'association avec la maladie. Notre test, appelé RVTT (« Rare Variant Trend Test »), regroupe tous les gènes d'intérêt en une seule unité, et considère le nombre d'occurrences de variants dans chaque unité pour chaque individu (**Figure 5**). Nous appliquons ensuite la statistique de Cochran-Armitage (Armitage, 1955 ; Cochran, 1954) pour évaluer la relation entre la charge croissante de variants qualificatifs rares et le statut de la maladie. Les variants ont été sélectionnés

à l'aide d'une approche à seuil variable (Price et al., 2010) et regroupés par voie ou réseau en ensembles de gènes combinés.

Nous avons validé les résultats les plus significatifs de l'analyse des variants et des gènes effectuée sur les exomes ciblés de la maladie de Parkinson par rapport à la cohorte de témoins de la MGRB dans deux ensembles de données indépendants : i) PPMI (Parkinson's Progression Markers Initiative) et ii) AMP-PD (Accelerating Medicines Partnership : Parkinson's Disease). Ces cohortes font partie d'initiatives de collaboration à grande échelle visant à faciliter l'identification de biomarqueurs et de cibles thérapeutiques, et comprennent le séquençage du génome entier, la transcriptomique et des données cliniques harmonisées. Après avoir filtré les cohortes afin d'inclure des individus appariés sur le plan ethnique et de s'assurer qu'il n'y a pas de chevauchement avec notre cohorte de cas, les ensembles de données PPMI et AMP-PD comptaient respectivement 608 (PD : 433, contrôle : 175) et 1983 (PD : 1289, contrôle : 694) individus.

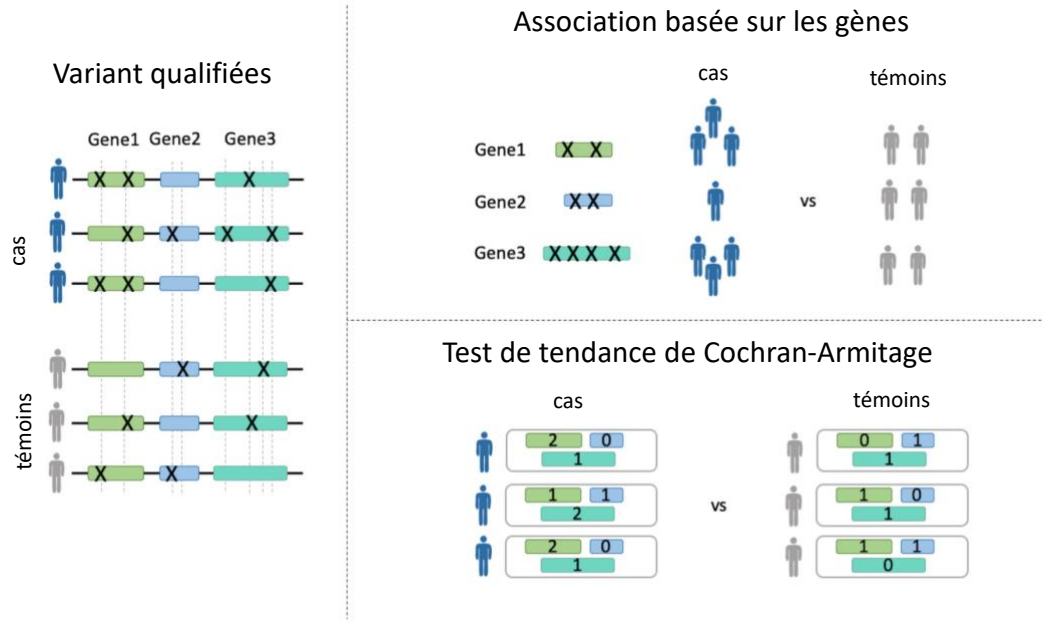


Figure 5 : Analyse d'association de variants rares. Variants qualifiés : Les variants rares dans chaque gène sont identifiés dans les cas et les contrôles. Pour l'association au niveau des variants, un test exact de Fisher unilatéral a été effectué pour chaque variant afin de déterminer si le variant est enrichie dans les cas par rapport aux contrôles. Test d'association basé sur les gènes : Le nombre de cas et de témoins hébergeant au moins un variant dans le gène d'intérêt a été compté (modèle CMC) et comparé à l'aide du test SKAT-O. Test de tendance de Cochran-Armitage basé sur un ensemble de gènes : dans notre approche RVTT, nous considérons tous les gènes d'intérêt comme une seule unité, et nous comptons le nombre d'occurrences de variants dans cette unité pour chaque individu. Nous calculons ensuite la statistique de Cochran-Armitage en utilisant les données de comptage de deux groupes et déterminons la signification de la tendance à la hausse en utilisant une approche basée sur la permutation.

Comme l'indiquent les résultats de nos premières analyses au niveau des variants et des gènes, il n'y a que quelques variants rares avec une taille d'effet modérée impliquées dans la MP, et ces variants rares ne sont souvent pas regroupées dans les mêmes quelques gènes. Avec les tailles d'échantillon actuellement disponibles, les tests d'association au niveau des variants et des gènes sont susceptibles d'être sous-puissants. Sans surprise, nous n'avons pas observé de résultats significatifs dans la cohorte AMP-PD, que ce soit au niveau du variant ou du gène. Cela nous a incités à développer un nouveau test statistique qui regroupe les variants au niveau du réseau/de la voie plutôt qu'au niveau d'un seul gène ou d'une seule région génomique afin d'augmenter la puissance. Nous avons validé les gènes les plus performants de notre analyse ciblée de l'exome dans la cohorte AMP-PD

en utilisant le RVTT. Dans chaque cas, nous avons observé une tendance significative à l'augmentation des variants faux-sens rares et faux-sens délétères dans l'ensemble de gènes d'intérêt chez les cas de MP par rapport aux témoins. Nous avons validé les gènes les plus significatifs de notre analyse ciblée de l'exome dans la cohorte AMP-PD en utilisant le RVTT, en observant une tendance significative à l'augmentation des variants faux-sens rares et faux-sens délétères dans l'ensemble de gènes d'intérêt dans les cas de PD par rapport aux contrôles. À titre d'expérience de contrôle nous avons testé l'enrichissement des variants synonymes dans les mêmes ensembles de gènes. Nous n'avons pas observé de tendance significative dans les variants synonymes chez les cas par rapport aux contrôles. Nous avons également effectué un RVTT sur 100 ensembles de gènes aléatoires d'une longueur de quatre à dix et n'avons observé aucune tendance significative dans les variants faux sens et faux sens dommageables. Bien que non significative, une tendance similaire a été observée dans l'ensemble de données PPMI. Ceci est probablement dû au petit nombre de cas et de contrôles dans cette cohorte.

PROFILS D'EXPRESSION DES ÉCHANTILLONS DE CÉRÉBRES

En plus de la validation externe de nos associations génétiques significatives, nous avons évalué les changements d'expression dans plusieurs cohortes de séquençage d'ARN de maladies neurodégénératives. Un sous-ensemble de 28 cas qui ont été séquencés dans notre analyse ciblée des exomes comprenait le séquençage de l'ARN d'échantillons de cortex préfrontal obtenus auprès de patients atteints de la MP familiale (accession GEO GSE68719). Étant donné le chevauchement des associations significatives de variants génétiques et de gènes associés au-delà des frontières de la maladie, nous avons exploré les modèles d'expression génique partagés dans les cerveaux de la MA, de la MP et de l'ASM. Ces échantillons, ainsi que 50 cerveaux de contrôle sains (qui ne font pas partie de notre cohorte de contrôle) représentent la plus grande cohorte unique d'ARN-seq de cerveaux de PD publiée à ce jour (Dumitriu et al., 2016 ; Labadorf et al., 2017). Tous les échantillons de cerveau ont été obtenus auprès d'hommes d'ascendance européenne. Les cas présentant une indication de pathologie de la MA sur la base de l'évaluation des données neuropathologiques des plaques et des enchevêtrements (grading des plaques et stade de Braak) ont été exclus. Seuls les gènes qui étaient également ciblés dans l'analyse de séquençage de l'exome ont été retenus dans l'analyse. Les valeurs p brutes ont été corrigées pour les tests multiples en utilisant la procédure de Benjamini-Hochberg et les gènes passant un seuil FDR de 0,05 ont été considérés comme significatifs (41/418). 35 de ces gènes différentiellement exprimés (DEGs) issus de l'analyse d'enrichissement ciblée étaient également significatifs ($q<0,05$) dans l'analyse transcriptomique.

Deux des DEG significatifs étaient également significatifs dans le test d'association « SKAT-O » basé sur le séquençage de l'exome. TEAD2 et MAP2K3, qui ont passé la correction stricte de Bonferroni ($p<0,05$), ont été identifiés comme des homologues humains d'un suppresseur et d'un amplificateur de levure, respectivement, de la toxicité du bêta-amyloïde dans le crible de surexpression a-bêta et étaient tous deux plus fortement exprimés dans les cas de MP. Aucun des 28 cas chevauchant les données de séquençage de l'ARN ne contenait de variants rares (délétères ou synonymes) dans TEAD2 ou MAP2K3. Douze de ces cas présentaient des mutations hétérozygotes rares dans 12 des gènes significativement exprimés de façon différentielle. Ces résultats sont difficiles à interpréter car seuls 28 des 496 cas comprenaient des données de séquençage de l'ARN et des échantillons de contrôle non chevauchants. Les associations basées sur les gènes étaient largement dues à des variants faux-sens et la majorité des mutations faux-sens des maladies ne semblent pas altérer le repliement ou la stabilité des protéines (Sahni et al., 2013). Cependant, il est intéressant de noter que, comme dans notre cohorte de séquençage d'exome, les associations étaient indépendantes de la pathologie concomitante de la MA, car les cas présentant une pathologie de la MA ont été exclus de l'analyse RNA-seq.

Le plus grand chevauchement de gènes à la fois avec le séquençage de l'ARN de 28 de nos cas et notre analyse basée sur l'exome était avec une méta-analyse des données d'expression de la MA provenant de six études antérieures sur les puces à ADN comprenant 450 échantillons de cortex

frontal de la MA et 212 échantillons de cortex frontal sain (Li et al., 2015). Cinq des gènes significatifs identifiés dans notre association basée sur le séquençage de l'exome ont passé une valeur p ajustée de Bonferroni de 0,05 à l'échelle du transcriptome. Quatre de ces cinq gènes sont associés à la MA d'après notre raisonnement de séquençage, dont 3 modificateurs a-beta (TEAD2, MAP2K3 et PKN2) et 1 gène de risque de MA (CD33). TEAD2 et MAP2K3 ont également montré une expression significativement accrue dans l'analyse RNA-seq de 263 MA (de l'étude ROSMAP) par rapport à 151 échantillons de cortex préfrontal appariés à l'âge cognitivement normal (Canchi et al., 2019).

Les gènes connus associés à la MP ne présentent souvent pas d'expression significativement modifiée (par exemple, GBA, LRRK2, SNCA) chez les cas par rapport aux témoins. L'expression perturbée après l'apparition de la maladie peut ne pas refléter l'effet des variants à risque qui mènent à la MP, mais peut révéler les régulateurs des voies perturbées. L'absence de preuves soutenant un effet des génotypes sur l'expression des gènes n'exclut pas un rôle pour les variants associés à la maladie, en particulier pour les SNV faux-sens, qui conduisent rarement à une expression instable des protéines (Fragoza et al., 2019 ; Sahni et al., 2013). En outre, la réplication de TEAD2 et MAP2K3 à travers les maladies, qui étaient toutes deux significatives dans notre analyse de séquençage d'exome, est intéressante et pourrait aider à élucider les mécanismes partagés de la neurodégénérescence.

DISCUSSION ET CONCLUSION

Nous avons effectué un séquençage ciblé de l'exome chez 500 patients atteints de quatre synucléinopathies distinctes : MP, MP avec démence (MPD), la maladie à corps de Lewy (MCL) et atrophie de systèmes multiples (ASM). En plus des gènes connus de la MA et de la MP, nous avons séquencé les homologues humains des modulateurs génétiques de la toxicité de l'alpha-synucléine et de la bêta-amyloïde identifiés dans des cribles à l'échelle du génome dans la levure, y compris un large éventail d'orthologues. Les variants dans les réseaux de gènes de la bêta-amyloïde et de l'alpha-synucléine étaient enrichies dans les différentes synucléinopathies par rapport aux témoins âgés sains, tout comme les variants dans les gènes connus de la MP et de la MA. Des tests au niveau des gènes, y compris un test de tendance développé pour l'analyse des variants rares, ont confirmé et étendu ces résultats dans des cohortes supplémentaires d'AMS et de MP. Nos données impliquent des facteurs génétiques communs à des synucléinopathies distinctes et démontrent des réseaux génétiques convergents dans les protéinopathies bêta-amyloïde et alpha-synucléine humaines.

Des cohortes de plus en plus grandes dans les études d'association génétique humaine ont permis d'identifier un nombre croissant de loci à risque dans les maladies neurodégénératives. Des études portant sur des variants communs et rares ont mis en évidence plus de 90 loci génomiques liés à la maladie de Parkinson (Nalls et al., 2019). Ces variants expliquent ~22% du risque de maladie et l'augmentation de la taille des échantillons continuera probablement à découvrir des variants de risque supplémentaires, dont la plupart ont une petite taille d'effet, représentant une fraction de la variance du risque de maladie. Les variants codants rares ont une taille d'effet plus importante en moyenne, par rapport aux variants codants ou non codants communs, mais nécessitent des tailles d'échantillon incroyablement grandes. Nous avons émis l'hypothèse que la réduction de l'espace de recherche par séquençage à un ensemble de gènes connus *a priori* pour être enrichis en modificateurs génétiques de la protéotoxicité de l'alpha-synucléine nous permettrait de trouver des signaux significatifs même dans une cohorte de taille modeste.

Le principal défi à venir consiste à distinguer les mutations bénignes des mutations délétères et à combler l'écart entre les génotypes et les phénotypes grâce à des essais fonctionnels robustes et évolutifs permettant d'élucider l'interaction entre les loci de risque putatifs dans l'apparition et la progression de la maladie. Les études inter-espèces des protéinopathies ont permis de découvrir des éléments biologiques pertinents pour la maladie, y compris des homologues de facteurs de

risque de maladie connus chez l'homme. Bien que les mêmes gènes ne soient pas nécessairement retrouvés, les résultats convergent souvent vers les mêmes voies cellulaires.

Nous avons découvert de nouveaux facteurs de risque candidats pour les synucléinopathies, tous liés à la pathobiologie de base de l'alpha-synucléine grâce à des cribles à l'échelle du génome de la levure. Les variants rares des réseaux génétiques de la b-amyloïde et de l'a-synucléine étaient enrichis dans les différentes synucléinopathies par rapport aux témoins âgés sains et les associations étaient indépendantes de la pathologie concomitante de la MA. Des tests au niveau des gènes, y compris un test de tendance (RVTT) développé pour l'analyse des variants rares, ont confirmé et étendu ces résultats dans des cohortes supplémentaires de synucléinopathies. Nos résultats impliquent des facteurs génétiques communs à des synucléinopathies distinctes et démontrent des réseaux génétiques convergents dans les protéinopathies bêta-amyloïde et alpha-synucléine chez l'homme. Malgré le peu de chevauchement des facteurs de risque génétiques entre les maladies neurodégénératives, on observe une pathologie concomitante de la MA et de la MP. La compréhension de l'interaction de différentes protéines dans les protéinopathies comorbides peut aider à distinguer les réponses cellulaires générales au mauvais repliement des protéines des réponses liées à la fonction intrinsèque de la protéine mal repliée et est cruciale pour stratifier correctement les patients et développer des thérapies efficaces de modification de la maladie.

L'évaluation des conséquences fonctionnelles des variants est une étape cruciale vers l'identification des contributeurs directs ou des facteurs de risque, permettant éventuellement des prédictions plus précises du risque de maladie et le développement de thérapies. La plupart des cribles pangénomiques contre la protéotoxicité se sont concentrés sur des modificateurs génétiques uniques. Les gènes et leurs produits font partie d'un réseau complexe d'interactions physiques et biochimiques entre l'ADN, l'ARN et les protéines. Le développement de traitements efficaces nécessite de comprendre les effets combinatoires entre les facteurs de risque génétiques et leur relation avec l'agrégation de l'a-syn.

Comme la prévalence de la MP et son impact socio-économique devraient augmenter proportionnellement à la longévité de la population, il existe un besoin crucial d'interventions thérapeutiques. Nous avons développé une plateforme biologique inter-espèces qui peut être utilisée pour interroger les données émergentes du séquençage à grande échelle dans les maladies neurodégénératives. La combinaison des profils d'interaction moléculaire des variants et des interactions génétiques avec la toxicité de l'alpha-synucléine fournira une vue globale des conséquences fonctionnelles des variants rares identifiés chez les patients atteints de la MP ainsi que des preuves de leur pertinence biologique pour les événements pathologiques cellulaires. Nos approches ont le potentiel de fournir des informations importantes sur des maladies complexes dont les mécanismes biologiques sont conservés au cours de l'évolution et ont le potentiel d'identifier de nouveaux facteurs de risque et de fournir de nouvelles voies d'exploration pour les traitements.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Anheim, M., Elbaz, A., Lesage, S., Durr, A., Condroyer, C., Viallet, F., Pollak, P., Bonaïti, B., Bonaïti-Pellié, C., Brice, A., et al. (2012). Penetrance of Parkinson disease in glucocerebrosidase gene mutation carriers. Neurology *78*, 417–420.

Armitage, P. (1955). Tests for Linear Trends in Proportions and Frequencies. Biometrics *11*, 375–386.

Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am. J. Hum. Genet. *96*, 329–339.

Aslam, M., Kandasamy, N., Ullah, A., Paramasivam, N., Öztürk, M.A., Naureen, S., Arshad, A., Badshah, M., Khan, K., Wajid, M., et al. (2021). Putative second hit rare genetic variants in families with seemingly GBA-associated Parkinson's disease. NPJ Genom Med *6*, 2.

Auluck, P.K., Caraveo, G., and Lindquist, S. (2010). α-Synuclein: membrane interactions and toxicity in Parkinson's disease. Annu. Rev. Cell Dev. Biol. *26*, 211–233.

Bacanu, S.A., Devlin, B., and Roeder, K. (2000). The power of genomic control. Am. J. Hum. Genet. *66*, 1933–1944.

Balestrino, R., Tunesi, S., Tesei, S., Lopiano, L., Zecchinelli, A.L., and Goldwurm, S. (2020). Penetrance of Glucocerebrosidase (GBA) Mutations in Parkinson's Disease: A Kin Cohort Study. Mov. Disord. *35*, 2111–2114.

de Barry, J., Liégeois, C.M., and Janoshazi, A. (2010). Protein kinase C as a peripheral biomarker for Alzheimer's disease. Exp. Gerontol. *45*, 64–69.

Beecham, G.W., Bis, J.C., Martin, E.R., Choi, S.-H., DeStefano, A.L., van Duijn, C.M., Fornage, M., Gabriel, S.B., Koboldt, D.C., Larson, D.E., et al. (2017). The Alzheimer's Disease Sequencing Project: Study design and sample selection. Neurol Genet *3*, e194.

Behl, T., Kaur, G., Fratila, O., Buhas, C., Judea-Pusta, C.T., Negrut, N., Bustea, C., and Bungau, S. (2021). Cross-talks among GBA mutations, glucocerebrosidase, and α-synuclein in GBA-associated Parkinson's disease and their targeted therapeutic approaches: a comprehensive review. Transl. Neurodegener. *10*, 4.

Blauwendraat, C., Nalls, M.A., and Singleton, A.B. (2020). The genetic architecture of Parkinson's disease. Lancet Neurol. *19*, 170–178.

Blumenstiel, B., Cibulskis, K., Fisher, S., DeFelice, M., Barry, A., Fennell, T., Abreu, J., Minie, B., Costello, M., Young, G., et al. (2010). Targeted exon sequencing by in-solution hybrid selection. Curr. Protoc. Hum. Genet. *Chapter 18*, Unit 18.4.

Braak, H., Del Tredici, K., Rüb, U., de Vos, R.A.I., Jansen Steur, E.N.H., and Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol. Aging *24*, 197–211.

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., and Del Tredici, K. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol. *112*, 389–404.

Brás, J., Guerreiro, R., and Hardy, J. (2015). SnapShot: Genetics of Parkinson's disease. Cell *160*, 570-570.e1.

Bromberg, Y., Kahn, P.C., and Rost, B. (2013). Neutral and weakly nonneutral sequence variants may define individuality. Proc. Natl. Acad. Sci. U. S. A. *110*, 14255–14260.

Brown, E.E., Blauwendraat, C., Trinh, J., Rizig, M., Nalls, M.A., Leveille, E., Ruskey, J.A., Jonvik, H., Tan, M.M.X., Bandres-Ciga, S., et al. (2021). Analysis of DNM3 and VAMP4 as genetic modifiers of LRRK2 Parkinson's disease. Neurobiol. Aging *97*, 148.e17-148.e24.

van der Brug, M.P., Singleton, A., Gasser, T., and Lewis, P.A. (2015). Parkinson's disease: From human genetics to clinical trials. Sci. Transl. Med. *7*, 205ps20.

Burkhardt, M.F., Martinez, F.J., Wright, S., Ramos, C., Volfson, D., Mason, M., Garnes, J., Dang, V., Lievers, J., Shoukat-Mumtaz, U., et al. (2013). A cellular model for sporadic ALS using patient-derived induced pluripotent stem cells. Mol. Cell. Neurosci. *56*, 355–364.

Callender, J.A., Yang, Y., Lordén, G., Stephenson, N.L., Jones, A.C., Brognard, J., and Newton, A.C. (2018). Protein kinase Cα gain-of-function variant in Alzheimer's disease displays enhanced catalysis by a mechanism that evades down-regulation. Proc. Natl. Acad. Sci. U. S. A. *115*, E5497–E5505.

Canchi, S., Raao, B., Masliah, D., Rosenthal, S.B., Sasik, R., Fisch, K.M., De Jager, P.L., Bennett, D.A., and Rissman, R.A. (2019). Integrating Gene and Protein Expression Reveals Perturbed Functional Networks in Alzheimer's Disease. Cell Rep. *28*, 1103-1116.e4.

Chen, Z., Boehnke, M., and Fuchsberger, C. (2020). Combining sequence data from multiple studies: Impact of analysis strategies on rare variant calling and association results. Genet. Epidemiol. *44*, 41–51.

Cheng, F., Zhao, J., Wang, Y., Lu, W., Liu, Z., Zhou, Y., Martin, W.R., Wang, R., Huang, J., Hao, T., et al. (2021). Comprehensive characterization of protein-protein interactions perturbed by disease mutations. Nat. Genet. *53*, 342–353.

Cheng, H.-C., Ulane, C.M., and Burke, R.E. (2010). Clinical progression in Parkinson disease and the neurobiology of axons. Ann. Neurol. *67*, 715–725.

Chiba-Falek, O., Lopez, G.J., and Nussbaum, R.L. (2006). Levels of alpha-synuclein mRNA in sporadic Parkinson disease patients. Mov. Disord. *21*, 1703–1708.

Chung, C.Y., Khurana, V., Auluck, P.K., Tardiff, D.F., Mazzulli, J.R., Soldner, F., Baru, V., Lou, Y., Freyzon, Y., Cho, S., et al. (2013). Identification and rescue of α-synuclein toxicity in Parkinson patient-derived neurons. Science *342*, 983–987.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). 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 *6*, 80–92.

Clinton, L.K., Blurton-Jones, M., Myczek, K., Trojanowski, J.Q., and LaFerla, F.M. (2010). Synergistic Interactions between Abeta, tau, and alpha-synuclein: acceleration of neuropathology and cognitive decline. J. Neurosci. *30*, 7281–7289.

Cochran, W.G. (1954). Some Methods for Strengthening the Common χ^2 Tests. Biometrics *10*, 417–451.

Cooper, A.A., Gitler, A.D., Cashikar, A., Haynes, C.M., Hill, K.J., Bhullar, B., Liu, K., Xu, K., Strathearn, K.E., Liu, F., et al. (2006). Alpha-synuclein blocks ER-Golgi traffic and Rab1 rescues neuron loss in Parkinson's models. Science *313*, 324–328.

Courte, J., Bousset, L., Boxberg, Y.V., Villard, C., Melki, R., and Peyrin, J.-M. (2020). The expression level of alpha-synuclein in different neuronal populations is the primary determinant of its prion-like seeding. Sci. Rep. *10*, 4895.

Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. *1*, 131.

Cuadrado, A., Rojo, A.I., Wells, G., Hayes, J.D., Cousin, S.P., Rumsey, W.L., Attucks, O.C., Franklin, S., Levonen, A.-L., Kensler, T.W., et al. (2019). Therapeutic targeting of the NRF2 and KEAP1 partnership in chronic diseases. Nat. Rev. Drug Discov. *18*, 295–317.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*.

Dächsel, J.C., Lincoln, S.J., Gonzalez, J., Ross, O.A., Dickson, D.W., and Farrer, M.J. (2007). The ups and downs of alpha-synuclein mRNA expression. Mov. Disord. *22*, 293–295.

Dhungel, N., Eleuteri, S., Li, L.-B., Kramer, N.J., Chartron, J.W., Spencer, B., Kosberg, K., Fields, J.A., Stafa, K., Adame, A., et al. (2015). Parkinson's disease genes VPS35 and EIF4G1 interact genetically and converge on α-synuclein. Neuron *85*, 76–87.

Dorsey, E.R., and Bloem, B.R. (2018). The Parkinson Pandemic-A Call to Action. JAMA Neurol. *75*, 9–10.

Du, Y., Zhao, Y., Li, C., Zheng, Q., Tian, J., Li, Z., Huang, T.Y., Zhang, W., and Xu, H. (2018). Inhibition of PKCδ reduces amyloid-β levels and reverses Alzheimer disease phenotypes. J. Exp. Med. *215*, 1665–1677.

Dumitriu, A., Moser, C., Hadzi, T.C., Williamson, S.L., Pacheco, C.D., Hendricks, A.E., Latourelle, J.C., Wilk, J.B., Destefano, A.L., and Myers, R.H. (2012). Postmortem Interval Influences α-Synuclein Expression in Parkinson Disease Brain. Parkinsons Dis. *2012*, 614212.

Dumitriu, A., Golji, J., Labadorf, A.T., Gao, B., Beach, T.G., Myers, R.H., Longo, K.A., and Latourelle, J.C. (2016). Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. BMC Med. Genomics *9*, 5.

Engelhardt, E., and Gomes, M. da M. (2017). Lewy and his inclusion bodies: Discovery and rejection. Dement Neuropsychol *11*, 198–201.

Fanciulli, A., and Wenning, G.K. (2015). Multiple-system atrophy. N. Engl. J. Med. *372*, 249–263.

Feany, M.B., and Bender, W.W. (2000). A Drosophila model of Parkinson's disease. Nature *404*, 394–398.

Fernandes, H.J.R., Patikas, N., Foskolou, S., Field, S.F., Park, J.-E., Byrne, M.L., Bassett, A.R., and Metzakopian, E. (2020). Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. Cell Rep. *33*, 108263.

Foo, J.-N., Liu, J.-J., and Tan, E.-K. (2012). Whole-genome and whole-exome sequencing in neurological diseases. Nat. Rev. Neurol. *8*, 508–517.

Fornasiero, E.F., and Rizzoli, S.O. (2019). Pathological changes are associated with shifts in the employment of synonymous codons at the transcriptome level. BMC Genomics *20*, 566.

Foster, M.W. (2008). Human Genome Diversity Project (HGDP) (Chichester, UK: John Wiley & Sons, Ltd).

Fragoza, R., Das, J., Wierbowski, S.D., Liang, J., Tran, T.N., Liang, S., Beltran, J.F., Rivera-Erick, C.A., Ye, K., Wang, T.-Y., et al. (2019). Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat. Commun. *10*, 4141.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. *653*, 249–257.

Gan-Or, Z., Amshalom, I., Bar-Shira, A., Gana-Weisz, M., Mirelman, A., Marder, K., Bressman, S., Giladi, N., and Orr-Urtreger, A. (2015). The Alzheimer disease BIN1 locus as a modifier of GBA-associated Parkinson disease. J. Neurol. *262*, 2443–2447.

Gardai, S.J., Mao, W., Schüle, B., Babcock, M., Schoebel, S., Lorenzana, C., Alexander, J., Kim, S., Glick, H., Hilton, K., et al. (2013). Elevated alpha-synuclein impairs innate immune cell function and provides a potential peripheral biomarker for Parkinson's disease. PLoS One *8*, e71634.

GBD 2016 Parkinson's Disease Collaborators (2018). Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. *17*, 939–953.

Geiger, J.T., Ding, J., Crain, B., Pletnikova, O., Letson, C., Dawson, T.M., Rosenthal, L.S., Pantelyat, A., Gibbs, J.R., Albert, M.S., et al. (2016). Next-generation sequencing reveals substantial genetic contribution to dementia with Lewy bodies. Neurobiol. Dis. *94*, 55–62.

Gerasimavicius, L., Liu, X., and Marsh, J.A. (2020). Identification of pathogenic missense mutations using protein stability predictors. Sci. Rep. *10*, 15387.

Germer, E.L., Imhoff, S., Vilariño-Güell, C., Kasten, M., Seibler, P., Brüggemann, N., International Parkinson's Disease Genomics Consortium, Klein, C., and Trinh, J. (2019). The Role of Rare Coding Variants in Parkinson's Disease GWAS Loci. Front. Neurol. *10*, 1284.

Goetz, C.G. (2011). The history of Parkinson's disease: early clinical descriptions and neurological therapies. Cold Spring Harb. Perspect. Med. *1*, a008862.

Golbe, L.I., Di Iorio, G., Bonavita, V., Miller, D.C., and Duvoisin, R.C. (1990). A large kindred with autosomal dominant Parkinson's disease. Ann. Neurol. *27*, 276–282.

Gordon, R., Singh, N., Lawana, V., Ghosh, A., Harischandra, D.S., Jin, H., Hogan, C., Sarkar, S., Rokad, D., Panicker, N., et al. (2016). Protein kinase Cδ upregulation in microglia drives neuroinflammatory responses and dopaminergic neurodegeneration in experimental models of Parkinson's disease. Neurobiol. Dis. *93*, 96–114.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Guerreiro, R., Brás, J., and Hardy, J. (2015). SnapShot: Genetics of ALS and FTD. Cell *160*, 798-798.e1.

Guo, M.H., Plummer, L., Chan, Y.-M., Hirschhorn, J.N., and Lippincott, M.F. (2018). Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. Am. J. Hum. Genet. *103*, 522–534.

Gureev, A.P., and Popov, V.N. (2019). Nrf2/ARE Pathway as a Therapeutic Target for the Treatment of Parkinson Diseases. Neurochem. Res. *44*, 2273–2279.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

Haenig, C., Atias, N., Taylor, A.K., Mazza, A., Schaefer, M.H., Russ, J., Riechers, S.-P., Jain, S., Coughlin, M., Fontaine, J.-F., et al. (2020). Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains. Cell Rep. *32*, 108050.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A., et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat. Genet. *41*, 1088–1093.

Healy, D.G., Falchi, M., O'Sullivan, S.S., Bonifati, V., Durr, A., Bressman, S., Brice, A., Aasly, J., Zabetian, C.P., Goldwurm, S., et al. (2008). Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol. *7*, 583–590.

Henderson, M.X., Sengupta, M., Trojanowski, J.Q., and Lee, V.M.Y. (2019). Alzheimer's disease tau is a prominent pathology in LRRK2 Parkinson's disease. Acta Neuropathol Commun *7*, 183.

Hijikata, A., Tsuji, T., Shionyu, M., and Shirai, T. (2017). Decoding disease-causing mechanisms of missense mutations from supramolecular structures. Sci. Rep. *7*, 8541.

Hou, Y., Dan, X., Babbar, M., Wei, Y., Hasselbalch, S.G., Croteau, D.L., and Bohr, V.A. (2019). Ageing as a risk factor for neurodegenerative disease. Nat. Rev. Neurol. *15*, 565–581.

Hu, Y.-J., Liao, P., Johnston, H.R., Allen, A.S., and Satten, G.A. (2016). Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. PLoS Genet. *12*, e1006040.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. Nat. Genet. *29*, 306–309.

Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

Irwin, D.J., Lee, V.M.-Y., and Trojanowski, J.Q. (2013). Parkinson's disease dementia: convergence of α-synuclein, tau and amyloid-β pathologies. Nat. Rev. Neurosci. *14*, 626–636.

Irwin, D.J., Xie, S.X., Coughlin, D., Nevler, N., Akhtar, R.S., McMillan, C.T., Lee, E.B., Wolk, D.A., Weintraub, D., Chen-Plotkin, A., et al. (2018). CSF tau and β-amyloid predict cerebral synucleinopathy in autopsied Lewy body disorders. Neurology *90*, e1038–e1046.

Israel, M.A., Yuan, S.H., Bardy, C., Reyna, S.M., Mu, Y., Herrera, C., Hefferan, M.P., Van Gorp, S., Nazor, K.L., Boscolo, F.S., et al. (2012). Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. Nature *482*, 216–220.

Iwaki, H., Blauwendraat, C., Makarious, M.B., Bandrés-Ciga, S., Leonard, H.L., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Faghri, F., International Parkinson's Disease Genomics Consortium (IPDGC), et al. (2020). Penetrance of Parkinson's Disease in LRRK2 p.G2019S Carriers Is Modified by a Polygenic Risk Score. Mov. Disord. *35*, 774–780.

Jansen, I.E., Ye, H., Heetveld, S., Lechler, M.C., Michels, H., Seinstra, R.I., Lubbe, S.J., Drouet, V., Lesage, S., Majounie, E., et al. (2017). Discovery and functional prioritization of Parkinson's disease candidate genes from large-scale whole exome sequencing. Genome Biol. *18*, 22.

Jones, K.M., Cook-Deegan, R., Rotimi, C.N., Callier, S.L., Bentley, A.R., Stevens, H., Phillips, K.A., Jansen, J.P., Weyant, C.F., Roberts, D.E., et al. (2021). Complicated legacies: The human genome at 20. Science *371*, 564–569.

Kalia, L.V., and Kalia, S.K. (2015). α-Synuclein and Lewy pathology in Parkinson's disease. Curr. Opin. Neurol. *28*, 375–381.

Kallhoff, V., Peethumnongsin, E., and Zheng, H. (2007). Lack of alpha-synuclein increases amyloid plaque accumulation in a transgenic mouse model of Alzheimer's disease. Mol. Neurodegener. *2*, 6.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.

Keller, M.F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simón-Sánchez, J., Mittag, F., Büchel, F., Sharma, M., Gibbs, J.R., et al. (2012). Using genome-wide complex trait analysis to quantify "missing heritability" in Parkinson's disease. Hum. Mol. Genet. *21*, 4996–5009.

Kelly, J., Moyeed, R., Carroll, C., Albani, D., and Li, X. (2019). Gene expression meta-analysis of Parkinson's disease and its relationship with Alzheimer's disease. Mol. Brain *12*, 16.

Keo, A., Mahfouz, A., Ingrassia, A.M.T., Meneboo, J.-P., Villenet, C., Mutez, E., Comptdaer, T., Lelieveldt, B.P.F., Figeac, M., Chartier-Harlin, M.-C., et al. (2020). Transcriptomic signatures of brain regional vulnerability to Parkinson's disease. Commun Biol *3*, 101.

Khan, S.S., LaCroix, M., Boyle, G., Sherman, M.A., Brown, J.L., Amar, F., Aldaco, J., Lee, M.K., Bloom, G.S., and Lesné, S.E. (2018). Bidirectional modulation of Alzheimer phenotype by alpha-synuclein in mice and primary neurons. Acta Neuropathol. *136*, 589–605.

Khurana, V., and Lindquist, S. (2010). Modelling neurodegeneration in Saccharomyces cerevisiae: why cook with baker's yeast? Nat. Rev. Neurosci. *11*, 436–449.

Khurana, V., Peng, J., Chung, C.Y., Auluck, P.K., Fanning, S., Tardiff, D.F., Bartels, T., Koeva, M., Eichhorn, S.W., Benyamini, H., et al. (2017). Genome-Scale Networks Link Neurodegenerative Disease Genes to α-Synuclein through Specific Molecular Pathways. Cell Syst *4*, 157-170.e14.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.

Klein, C., and Schlossmacher, M.G. (2006). The genetics of Parkinson disease: Implications for neurological care. Nat. Clin. Pract. Neurol. *2*, 136–146.

Klein, C., and Westenberger, A. (2012). Genetics of Parkinson's disease. Cold Spring Harb. Perspect. Med. *2*, a008888.

Konovalova, E.V., Lopacheva, O.M., Grivennikov, I.A., Lebedeva, O.S., Dashinimaev, E.B., Khaspekov, L.G., Fedotova, E.Y., and Illarioshkin, S.N. (2015). Mutations in the Parkinson's Disease-Associated PARK2 Gene Are Accompanied by Imbalance in Programmed Cell Death Systems. Acta Naturae *7*, 146–149.

Kosmicki, J.A., Churchhouse, C.L., Rivas, M.A., and Neale, B.M. (2016). Discovery of rare variants for complex phenotypes. Hum. Genet. *135*, 625–634.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90-7.

Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.

Labadorf, A., Choi, S.H., and Myers, R.H. (2017). Evidence for a Pan-Neurodegenerative Disease Response in Huntington's and Parkinson's Disease Expression Profiles. Front. Mol. Neurosci. *10*, 430.

Lam, I., Hallacli, E., and Khurana, V. (2020). Proteome-Scale Mapping of Perturbed Proteostasis in Living Cells. Cold Spring Harb. Perspect. Biol. *12*.

Latourelle, J.C., Hendricks, A.E., Pankratz, N., Wilk, J.B., Halter, C., Nichols, W.C., Gusella, J.F., Destefano, A.L., Myers, R.H., Foroud, T., et al. (2011). Genomewide linkage study of modifiers of LRRK2-related Parkinson's disease. Mov. Disord. *26*, 2039–2044.

Le Guen, Y., Belloy, M.E., Napolioni, V., Eger, S.J., Kennedy, G., Tao, R., He, Z., Greicius, M.D., and Alzheimer's Disease Neuroimaging Initiative (2021). A novel age-informed approach for genetic association analysis in Alzheimer's disease. Alzheimers. Res. Ther. *13*, 72.

Lee, A.J., Wang, Y., Alcalay, R.N., Mejia-Santana, H., Saunders-Pullman, R., Bressman, S., Corvol, J.-C., Brice, A., Lesage, S., Mangone, G., et al. (2017). Penetrance estimate of LRRK2 p.G2019S mutation in individuals of non-Ashkenazi Jewish ancestry. Mov. Disord. *32*, 1432–1438.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23.

Li, X., Long, J., He, T., Belshaw, R., and Scott, J. (2015). Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease. Sci. Rep. *5*, 12393.

Li, Y.I., Wong, G., Humphrey, J., and Raj, T. (2019). Prioritizing Parkinson's disease genes using population-scale transcriptomic data. Nat. Commun. *10*, 994.

Liu, J., McCleland, M., Stawiski, E.W., Gnad, F., Mayba, O., Haverty, P.M., Durinck, S., Chen, Y.-J., Klijn, C., Jhunjhunwala, S., et al. (2014). Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. Nat. Commun. *5*, 3830.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194–1217.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics *31*, 3555–3557.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., et al. (2021). WikiPathways: connecting communities. Nucleic Acids Res. *49*, D613–D621.

Masliah, E., Rockenstein, E., Veinbergs, I., Mallory, M., Hashimoto, M., Takeda, A., Sagara, Y., Sisk, A., and Mucke, L. (2000). Dopaminergic loss and inclusion body formation in alpha-synuclein mice: implications for neurodegenerative disorders. Science *287*, 1265–1269.

Masliah, E., Rockenstein, E., Veinbergs, I., Sagara, Y., Mallory, M., Hashimoto, M., and Mucke, L. (2001). beta-amyloid peptides enhance alpha-synuclein accumulation and neuronal deficits in a transgenic mouse model linking Alzheimer's disease and Parkinson's disease. Proc. Natl. Acad. Sci. U. S. A. *98*, 12245–12250.

Mathieson, I., and McVean, G. (2013). Reply to: "FaST-LMM-Select for addressing confounding from spatial structure and rare variants." Nat. Genet. *45*, 471.

McCarthy, C., Carrea, A., and Diambra, L. (2017). Bicodon bias can determine the role of synonymous SNPs in human diseases. BMC Genomics *18*, 227.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Miller, J.E., Shivakumar, M.K., Risacher, S.L., Saykin, A.J., Lee, S., Nho, K., and Kim, D. (2018). Codon bias among synonymous rare variants is associated with Alzheimer's disease imaging biomarker. Pac. Symp. Biocomput. *23*, 365–376.

Mitsui, J., Matsukawa, T., Sasaki, H., Yabe, I., Matsushima, M., Dürr, A., Brice, A., Takashima, H., Kikuchi, A., Aoki, M., et al. (2015). Variants associated with Gaucher disease in multiple system atrophy. Ann Clin Transl Neurol *2*, 417–426.

Momozawa, Y., and Mizukami, K. (2021). Unique roles of rare variants in the genetics of complex diseases in humans. J. Hum. Genet. *66*, 11–23.

Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M.A., Gaulton, K.J., Albers, P.K., GoT2D Consortium, McVean, G., Boehnke, M., et al. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. *11*, e1005165.

Na, D., Rouf, M., O'Kane, C.J., Rubinsztein, D.C., and Gsponer, J. (2013). NeuroGeM, a knowledgebase of genetic modifiers in neurodegenerative diseases. BMC Med. Genomics *6*, 52.

Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. *18*, 1091–1102.

Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860.

Nguyen, A.P.T., Tsika, E., Kelly, K., Levine, N., Chen, X., West, A.B., Boularand, S., Barneoud, P., and Moore, D.J. (2020). Dopaminergic neurodegeneration induced by Parkinson's disease-linked G2019S LRRK2 is dependent on kinase and GTPase activity. Proc. Natl. Acad. Sci. U. S. A. *117*, 17296–17307.

Nido, G.S., Dick, F., Toker, L., Petersen, K., Alves, G., Tysnes, O.-B., Jonassen, I., Haugarvoll, K., and Tzoulis, C. (2020). Common gene expression signatures in Parkinson's disease are driven by changes in cell composition. Acta Neuropathol Commun *8*, 55.

Nuytemans, K., Maldonado, L., Ali, A., John-Williams, K., Beecham, G.W., Martin, E., Scott, W.K., and Vance, J.M. (2016). Overlap between Parkinson disease and Alzheimer disease in ABCA7 functional variants. Neurol Genet *2*, e44.

Oeppen, J., and Vaupel, J.W. (2002). Demography. Broken limits to life expectancy. Science *296*, 1029–1031.

ORFeome Collaboration (2016). The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nat. Methods *13*, 191–192.

Outeiro, T.F., and Lindquist, S. (2003). Yeast cells provide insight into alpha-synuclein biology and pathobiology. Science *302*, 1772–1775.

Park, L. (2019). Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. Sci. Rep. *9*, 11380.

Pe'er, I., Chretien, Y.R., de Bakker, P.I.W., Barrett, J.C., Daly, M.J., and Altshuler, D.M. (2006). Biases and reconciliation in estimates of linkage disequilibrium in the human genome. Am. J. Hum. Genet. *78*, 588–603.

Pinese, M., Lacaze, P., Rath, E.M., Stone, A., Brion, M.-J., Ameur, A., Nagpal, S., Puttick, C., Husson, S., Degrave, D., et al. (2020). The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly. Nat. Commun. *11*, 435.

Piras, I.S., Bleul, C., Schrauwen, I., Talboom, J., Llaci, L., De Both, M.D., Naymik, M.A., Halliday, G., Bettencourt, C., Holton, J.L., et al. (2020). Transcriptional profiling of multiple system atrophy cerebellar tissue highlights differences between the parkinsonian and cerebellar sub-types of the disease. Acta Neuropathol Commun *8*, 76.

Pirinen, M., Donnelly, P., and Spencer, C.C.A. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. Nat. Genet. *44*, 848–851.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. *12*, 32–42.

Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., et al. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science *276*, 2045–2047.

Price, A., Farooq, R., Yuan, J.-M., Menon, V.B., Cardinal, R.N., and O'Brien, J.T. (2017). Mortality in dementia with Lewy bodies compared with Alzheimer's dementia: a retrospective naturalistic cohort study. BMJ Open *7*, e017504.

Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.-J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.

Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. *69*, 1–14.

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science *344*, 519–523.

Rana, H.Q., Balwani, M., Bier, L., and Alcalay, R.N. (2013). Age-specific Parkinson disease risk in GBA mutation carriers: information for genetic counseling. Genet. Med. *15*, 146–149.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.

Robinson, J.L., Lee, E.B., Xie, S.X., Rennert, L., Suh, E., Bredenberg, C., Caswell, C., Van Deerlin, V.M., Yan, N., Yousef, A., et al. (2018). Neurodegenerative disease concomitant proteinopathies are prevalent, age-related and APOE4-associated. Brain *141*, 2181–2193.

Sahni, N., Yi, S., Zhong, Q., Jailkhani, N., Charloteaux, B., Cusick, M.E., and Vidal, M. (2013). Edgotype: a fundamental link between genotype and phenotype. Curr. Opin. Genet. Dev. *23*, 649–657.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647–660.

Sandor, C., Robertson, P., Lang, C., Heger, A., Booth, H., Vowles, J., Witty, L., Bowden, R., Hu, M., Cowley, S.A., et al. (2017). Transcriptomic profiling of purified patient-derived dopamine neurons identifies convergent perturbations and therapeutics for Parkinson's disease. Human Molecular Genetics ddw412.

Shi, F., Yao, Y., Bin, Y., Zheng, C.-H., and Xia, J. (2019). Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. BMC Med. Genomics *12*, 12.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. *12*, 771–776.

Shulman, J.M., Shulman, L.M., Weiner, W.J., and Feany, M.B. (2003). From fruit fly to bedside: translating lessons from Drosophila models of neurodegenerative disease. Curr. Opin. Neurol. *16*, 443–449.

Siddiqui, I.J., Pervaiz, N., and Abbasi, A.A. (2016). The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication. Sci. Rep. *6*, 24475.

Siderowf, A., Xie, S.X., Hurtig, H., Weintraub, D., Duda, J., Chen-Plotkin, A., Shaw, L.M., Van Deerlin, V., Trojanowski, J.Q., and Clark, C. (2010). CSF amyloid {beta} 1-42 predicts cognitive decline in Parkinson disease. Neurology *75*, 1055–1061.

Sidransky, E., Nalls, M.A., Aasly, J.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N. Engl. J. Med. *361*, 1651–1661.

Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., et al. (2003). alpha-Synuclein locus triplication causes Parkinson's disease. Science *302*, 841.

Sklerov, M., Cortes, E., Kang, U., Vonsattel, J.P., Nichols, W., Pauciulo, M., and Alcalay, R. (2016). Increased Rate of GBA Variants in MSA Brains at a Brain Bank in New York City (P1.006). Neurology *86*.

Sklerov, M., Kang, U.J., Liong, C., Clark, L., Marder, K., Pauciulo, M., Nichols, W.C., Chung, W.K., Honig, L.S., Cortes, E., et al. (2017). Frequency of GBA variants in autopsy-proven multiple system atrophy. Mov Disord Clin Pract *4*, 574–581.

Smith, M.G., and Snyder, M. (2006). Yeast as a model for human disease. Curr. Protoc. Hum. Genet. *Chapter 15*, Unit 15.6.

Spencer, B., Desplats, P.A., Overk, C.R., Valera-Martin, E., Rissman, R.A., Wu, C., Mante, M., Adame, A., Florio, J., Rockenstein, E., et al. (2016). Reducing Endogenous α-Synuclein Mitigates the Degeneration of Selective Neuronal Populations in an Alzheimer's Disease Transgenic Mouse Model. J. Neurosci. *36*, 7971–7984.

Spillantini, M.G., Schmidt, M.L., Lee, V.M., Trojanowski, J.Q., Jakes, R., and Goedert, M. (1997). Alpha-synuclein in Lewy bodies. Nature *388*, 839–840.

Spillantini, M.G., Crowther, R.A., Jakes, R., Cairns, N.J., Lantos, P.L., and Goedert, M. (1998). Filamentous alpha-synuclein inclusions link multiple system atrophy with Parkinson's disease and dementia with Lewy bodies. Neurosci. Lett. *251*, 205–208.

Stefanova, N., Bücke, P., Duerr, S., and Wenning, G.K. (2009). Multiple system atrophy: an update. Lancet Neurol. *8*, 1172–1178.

Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum. Genet. *136*, 665–677.

Stoker, T.B., Camacho, M., Winder-Rhodes, S., Liu, G., Scherzer, C.R., Foltynie, T., Evans, J., Breen, D.P., Barker, R.A., and Williams-Gray, C.H. (2020). Impact of GBA1 variants on long-term clinical progression and mortality in incident Parkinson's disease. J. Neurol. Neurosurg. Psychiatry *91*, 695–702.

Strauss, B.S. (2016). Biochemical Genetics and Molecular Biology: The Contributions of George Beadle and Edward Tatum. Genetics *203*, 13–20.

Tadaka, S., Saigusa, D., Motoike, I.N., Inoue, J., Aoki, Y., Shirota, M., Koshiba, S., Yamamoto, M., and Kinoshita, K. (2018). jMorp: Japanese Multi Omics Reference Panel. Nucleic Acids Res. *46*, D551–D557.

Taguchi, K., Watanabe, Y., Tsujimura, A., and Tanaka, M. (2019). Expression of α-synuclein is regulated in a neuronal cell type-dependent manner. Anat. Sci. Int. *94*, 11–22.

Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., Lin, Z.-Y., Larsen, B., Choi, H., Berger, B., Gingras, A.-C., and Lindquist, S. (2014). A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. Cell *158*, 434–448.

Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.

Tang, M., Alaniz, M.E., Felsky, D., Vardarajan, B., Reyes-Dumeyer, D., Lantigua, R., Medrano, M., Bennett, D.A., de Jager, P.L., Mayeux, R., et al. (2020). Synonymous variants associated with Alzheimer disease in multiplex families. Neurol Genet *6*, e450.

Telenti, A., Pierce, L.C.T., Biggs, W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., et al. (2016). Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. U. S. A. *113*, 11901–11906.

Teschendorf, D., and Link, C.D. (2009). What have worm models told us about the mechanisms of neuronal dysfunction in human neurodegenerative diseases? Mol. Neurodegener. *4*, 38.

Tran, J., Anastacio, H., and Bardy, C. (2020). Genetic predispositions of Parkinson's disease revealed in patient-derived brain cells. Npj Parkinson's Disease *6*, 1–18.

Treusch, S., Hamamichi, S., Goodman, J.L., Matlack, K.E.S., Chung, C.Y., Baru, V., Shulman, J.M., Parrado, A., Bevis, B.J., Valastyan, J.S., et al. (2011). Functional links between Aβ toxicity, endocytic trafficking, and Alzheimer's disease risk factors in yeast. Science *334*, 1241–1245.

Trinh, J., Gustavsson, E.K., Vilariño-Güell, C., Bortnick, S., Latourelle, J., McKenzie, M.B., Tu, C.S., Nosova, E., Khinda, J., Milnerwood, A., et al. (2016). DNM3 and genetic modifiers of age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and association study. Lancet Neurol. *15*, 1248–1256.

Turkmen, A., and Lin, S. (2017). Are rare variants really independent? Genet. Epidemiol. *41*, 363–371.

Visanji, N.P., Brooks, P.L., Hazrati, L.-N., and Lang, A.E. (2013). The prion hypothesis in Parkinson's disease: Braak to the future. Acta Neuropathol Commun *1*, 2.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Wang, C., Cai, Y., Zheng, Z., Tang, B.-S., Xu, Y., Wang, T., Ma, J., Chen, S.-D., Langston, J.W., Tanner, C.M., et al. (2012). Penetrance of LRRK2 G2385R and R1628P is modified by common PD-associated genetic variants. Parkinsonism Relat. Disord. *18*, 958–963.

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., FUSION Study, Fulton, R., et al. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. Nat. Genet. *46*, 409–415.

Wang, C., Zhan, X., Liang, L., Abecasis, G.R., and Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am. J. Hum. Genet. *96*, 926–937.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Weghorn, D., and Sunyaev, S. (2017). Bayesian inference of negative and positive selection in human cancers. Nat. Genet. *49*, 1785–1788.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., MacArthur, D., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genet. Med. *19*, 1151–1158.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M., et al. (2011). Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. *19*, 807–812.

Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat. Genet. *41*, 316–323.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.

Zhang, D., Anantharam, V., Kanthasamy, A., and Kanthasamy, A.G. (2007). Neuroprotective effect of protein kinase C delta inhibitor rottlerin in cell culture and animal models of Parkinson's disease. J. Pharmacol. Exp. Ther. *322*, 913–922.

Zhang, Y., Guan, W., and Pan, W. (2013). Adjustment for population stratification via principal components in association analysis of rare variants. Genet. Epidemiol. *37*, 99–109.

Zhu, Y., Wang, L., Yin, Y., and Yang, E. (2017). Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. Sci. Rep. *7*, 5435.