



Published in final edited form as:

*Neurobiol Dis.* 2021 January ; 148: 105182. doi:10.1016/j.nbd.2020.105182.

## A population scale analysis of rare *SNCA* variation in the UK Biobank

Cornelis Blauwendraat<sup>a,\*</sup>,<sup>1</sup>, Mary B. Makarios<sup>a,b,c,1</sup>, Hampton L. Leonard<sup>a,d</sup>, Sara Bandres-Ciga<sup>a</sup>, Hirotaka Iwaki<sup>a,d</sup>, Mike A. Nalls<sup>a,d</sup>, Alastair J. Noyce<sup>e,f</sup>, Andrew B. Singleton<sup>a</sup>

<sup>a</sup>Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

<sup>b</sup>Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, London, UK

<sup>c</sup>UCL Movement Disorders Centre, University College London, London, UK

<sup>d</sup>Data Tecnica International, Glen Echo, MD, USA

<sup>e</sup>Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Barts and the London School of Medicine and Dentistry, Queen Mary University, London, UK

<sup>f</sup>Department of Clinical and Movement Neurosciences, UCL Institute of Neurology, London, UK

### Abstract

Parkinson's disease (PD) is a complex neurodegenerative disease with a variety of genetic and environmental factors contributing to disease. The *SNCA* gene encodes for the alpha-synuclein protein which plays a central role in PD, where aggregates of this protein are one of the pathological hallmarks of disease. Rare point mutations and copy number gains of the *SNCA* gene have been shown to cause autosomal dominant PD, and common DNA variants identified using Genome-Wide Association Studies (GWAS) are a moderate risk factor for PD. The UK Biobank is a large-scale population prospective study including ~500,000 individuals that has revolutionized human genetics. Here we assessed the frequency of *SNCA* variation in this cohort and identified 30 subjects carrying variants of interest including duplications ( $n=6$ ), deletions ( $n=6$ ) and large complex likely mosaic events ( $n=18$ ). No known pathogenic missense variants were identified. None of these subjects were reported to be a PD case, although it is possible that these individuals may develop PD at a later age, and whilst three had known prodromal features, these did not meet

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

\*Corresponding author at: Laboratory of Neurogenetics, NIA, NIH, Building 35, 35 Convent Drive, Bethesda, MD 20892, USA. [cornelis.blauwendraat@nih.gov](mailto:cornelis.blauwendraat@nih.gov) (C. Blauwendraat).

<sup>1</sup>These authors contributed equally to this work.

CRedit authorship contribution statement

**Cornelis Blauwendraat:** Conceptualization, Data curation, Formal analysis, Supervision, Methodology, Writing - original draft, Writing - review & editing. **Mary B. Makarios:** Conceptualization, Methodology, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Hampton L. Leonard:** Data curation, Writing - review & editing. **Sara Bandres-Ciga:** Data curation, Writing - original draft, Writing - review & editing. **Hirotaka Iwaki:** Data curation, Writing - review & editing. **Mike A. Nalls:** Data curation, Writing - review & editing. **Alastair J. Noyce:** Data curation, Writing - review & editing. **Andrew B. Singleton:** Conceptualization, Supervision, Data curation, Writing - review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.nbd.2020.105182>.

defined clinical criteria for being considered ‘prodromal’ cases. Seven of the 18 large complex carriers showed a history of blood based cancer. Overall, we identified copy number variants in the *SNCA* region in a large population based cohort without reported PD phenotype and symptoms. Putative mosaicism of the *SNCA* gene was identified, however, it is unclear whether it is associated with PD. These individuals are potential candidates for further investigation by performing *SNCA* RNA and protein expression studies, as well as promising clinical trial candidates to understand how duplication carriers potentially escape PD.

## Keywords

*SNCA*; Parkinson’s disease; Copy number variants; Alpha-synuclein; Mosaicism

---

## 1. Introduction

Parkinson’s disease (PD) often has a genetic component, either via rare damaging variants, more common risk variants, or a complex combination of these. Rare missense variants and copy number gains of the *SNCA* gene have been shown to cause autosomal dominant PD (Singleton et al., 2003; Polymeropoulos et al., 1997; Chartier-Harlin et al., 2004). Additionally, several non-coding common variants have been identified to moderately increase risk for PD (Nalls et al., 2019) and modify age at onset (Blauwendraat et al., 2019) with a disease mechanism that likely implies an increased expression of *SNCA* levels (Pihlström et al., 2018; Soldner et al., 2016). Importantly, *SNCA* encodes for the alpha-synuclein protein, and aggregates of this protein have been identified in Lewy bodies, which is one of the major histopathological hallmarks of PD. Overall, this puts *SNCA* at a central position in PD pathogenesis, and currently there are potential disease-modifying therapies under development aimed at lowering *SNCA* levels.

Pathogenic missense *SNCA* variants are very rare in the general population with five disease-causing mutations (A30P, E46K, G51D, A53E and A53T). *SNCA* copy number variants, although still rare, appear to be more frequent compared to pathogenic missense variants with approximately 60 families reported to date. More rapid progression and atypical PD presentation is fairly common for all *SNCA* mutation carriers, with increased prevalence of symptoms like: dementia, rapid eye movement sleep behavior disorder, and autonomic dysfunction. Age of onset for *SNCA* missense variant and copy number variant carriers is earlier compared to idiopathic PD, with a broad range of age of onset in which late forties/early fifties tend to be the most frequent (Book et al., 2018; Konno et al., 2016).

Besides these autosomal dominant inheritance patterns, *SNCA* copy number gains also have been reported via mosaic patterns (Perez-Rodriguez et al., 2019; Mokretar et al., 2018), although larger studies are needed to confirm whether this is a general PD pathogenesis mechanism. Here, we assess the presence and frequency of pathogenic *SNCA* variants (single nucleotide variants and copy number variants) in the population scale UK Biobank cohort (Bycroft et al., 2018).

## 2. Methods

### 2.1 UK Biobank genotype data

UK Biobank B allele frequency (BAF) and Log2 Ratio (L2R) were downloaded from the UK Biobank (v2) (June 2020) containing data of 488,377 individuals. BAF and L2R values were extracted for the larger *SNCA* gene region (chr4:84992449–96412248, hg19), and these parameters were investigated considering three *SNCA* regions 1) larger *SNCA* region gene  $\pm 5$  Mb (chr4:85645250–95759466, hg19), 2) *SNCA* gene  $\pm 0.5$  Mb (chr4:90145250–91259466, hg19), 3) *SNCA* gene body (chr4:90645250–90759466, hg19). For additional wider chromosome exploration, the *SNCA* region gene  $\pm 20$  Mb (chr4:70645250–110759466, hg19) and the full chromosome 4 region were used (chr4:1–180915260, hg19). For samples with large genomic events, the remaining autosomes were further inspected. Typically, BAF values ranging between  $-0.66$  and  $-0.33$  correspond to potential duplications. As part of our approach, BAF values  $<0.85$  and  $>0.65$  +  $<0.35$  and  $>0.15$  were counted for each of the three genetic regions and if the variant count was higher than 6, the sample was plotted and visually inspected. L2R values were averaged for each of the three genetic regions, where high average values would correspond to potential duplications and low average values would correspond to a potential deletion. After selecting subjects of interest, a total of 363 individuals were manually inspected of which 30 showed an event of interest including potential duplications, deletions and more complex events. BAF and L2R values were plotted in R (<https://www.r-project.org/>) using ggplot2, and averages were calculated using the LOWESS (Locally Weighted Scatterplot Smoothing) function with variable smoother span based on size of total graph (ranging from 0.01 to 0.0001). UK Biobank genotype data (v2) was used to calculate relatedness using PLINK (v1.9) (Chang et al., 2015).

### 2.2 Exome sequencing data

UK Biobank exome sequencing data (FE dataset, field codes: 23160 and 23,161, June 2020) were downloaded from the UK Biobank containing data of 49,960 individuals. Variants were annotated using ANNOVAR (Wang et al., 2010). A comprehensive assessment of known pathogenic variants including A30P, E46K, G51D, A53E and A53T was performed. Furthermore, we screened previously reported pathogenic variants from both ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) and the Human Gene Mutation Database (HGMD) (Professional 2020.2, Qiagen), as well as potentially relevant loss-of-function variants (stop, frameshift, splicing). Exome sequencing (OQFE dataset, field codes: 23151, December 2020) was used as replication of genomic events using the allelic depth (AD) field included in the individual's VCF files. AD ratio was calculated by dividing the highest by the lowest depth value and subsequently plotted using ggplot2 in R (Wickham, 2016). A total of ten random samples were selected as negative controls. Additional population allele frequencies were obtained from the gnomAD population database (<https://gnomad.broadinstitute.org/>, v2.1.1, June 2020) (Karczewski et al., 2020).

### 2.3 UK Biobank phenotype data

UK Biobank phenotype data was obtained from ICD10 codes (field code: 41270), PD (field code: 131023), illnesses of father and mother (field codes: 20107 and 20,110), parkinsonism

(field code: 42031) or dementia (field code: 42018), genetic ethnic grouping (field code: 22006), year of birth (field code: 34) and age of recruitment (field code: 21022). The MDS prodromal criteria were used to assess a potential PD prodromal phenotype (Heinzel et al., 2019).

## 2.4 Data and code availability

All data used as part of this study is publically available upon application at the UK Biobank website (<https://www.ukbiobank.ac.uk/>) and gnomAD database (<https://gnomad.broadinstitute.org/>). Code used for both preprocessing and subsequent analyses is available on our GitHub repository [https://github.com/neurogenetics/UKbiobank\\_SNCA](https://github.com/neurogenetics/UKbiobank_SNCA).

## 3. Results

Here we assessed the presence and frequency of potentially damaging *SNCA* mutations in the UK Biobank large-scale population level cohort. None of the five known *SNCA* pathogenic variants (A30P, E46K, G51D, A53E and A53T) were identified in the UK Biobank exome sequencing data. Two *SNCA* variants with conflicting or unconfirmed pathogenicity were detected (Sup Table 1) with relatively similar frequencies as in the gnomAD database. None of the variant carriers was a reported PD case, however, three showed a positive family history for PD. Thirteen subjects carried the *SNCA* P117T variant and one of these had a parent affected with PD. Additionally, a total of 28 subjects carried the *SNCA* H50Q variant of which two had a parent affected with PD. Note that none of these observations showed statistically significant results ( $p > 0.3$ ,  $2 \times 2$  Fisher exact test, Sup Table 4).

Next, we explored the presence and frequency of *SNCA* copy number variants by assessing BAF and L2R values in the UK Biobank genotyping data. A total of 30 carriers were identified and categorized into three groups, duplication carriers ( $n = 6$ ), deletion carriers ( $n = 6$ ) and large “complex” events carriers ( $n = 18$ ). Large complex events carriers were subjects where the BAF and L2R values did not meet the criteria for a “small” duplication or deletion but rather showed evidence of a large mosaicism event (Sup Table 2). Six potential *SNCA* duplications were identified (Fig. 1), of which all showed clear BAF and L2R changes implying a *SNCA* duplication. The average estimated size was  $-2$  Mb, ranging from  $-0.8$  Mb to  $-6$  Mb. Estimated breaking points are reported in Sup Table 2. Furthermore, six potential *SNCA* deletions were identified of which five were complete gene deletions and one likely partial gene deletion (covering transcription start site and first exons) but sufficient to likely result in haploinsufficiency (Fig. 2). The average estimated size was  $-0.76$  Mb and ranging from  $-0.2$  Mb to  $-1$  Mb. Besides the identification of relatively straightforward duplications and deletions, 18 complex events were further detected (Sup. Fig. 1). Inspection of the wider *SNCA* ( $\pm 20$  Mb), the full chromosome 4, as well as the remaining autosomes, suggested that most events are likely due to mosaicism mostly spanning the majority of the long-arm of chromosome 4 (4q11-4q35). Notably, one individual (#comp16) carried an additional large event on chromosome 20. After closer inspection, the majority were classified as large complex events with likely varying levels of

mosaicism, of which two were likely mosaic deletions due to a uniparental disomy event (Sup. Fig. 1).

As partial validation of the results obtained from the genotyping array data, we explored exome sequencing allelic depth data (sequencing depth of each allele) of all heterozygous variants on chromosome 4. Exome sequencing was available for 14 of the 30 subjects (3 from UK Biobank field code 23161 and 14 from field code 23151) with a potential *SNCA* alteration (Sup. Table 2). All large events (>6 Mb) showed clear differences of allelic depth in the expected areas of genomic events (Fig. 3 and Sup. Fig. 2). However, the smaller events <2 Mb did not show clear differences of allelic depth likely due to limited useable variants in these regions due to exome sequencing coverage (Sup. Fig. 2). We used 10 random subjects from the exome sequencing data as negative controls, of which none showed evidence of an allelic imbalance (Sup. Fig. 3).

When assessing the UK Biobank phenotypic data of the subjects with a potential *SNCA* alteration, none of them had been reported to be affected with PD, parkinsonism or dementia. Further inspection of ICD10 codes did not result in a potential PD diagnosis, although we cannot discard the possibility that some may be prodromal PD cases based on non-specific symptoms which are common in the older general population, including depression, anxiety, syncope, and head injury. Three participants were further evaluated to ascertain the probability of them being prodromal cases according to the recently revised MDS Prodromal Criteria. None of the participants met the clinical threshold criteria for probable prodromal PD (see Sup. Table 3). Of the 24 subjects of interest, two had a parent affected with dementia (#comp3 and #comp6) and one had a parent with PD (#comp6). Note that none of these observations showed statistically significant results ( $p > 0.1$ , Chi-square test). Of note, seven carriers (all large complex events) of the 18 (39%) potential mosaic subjects showed a history of blood-based cancer (including: lymphoma, multiple myeloma or myelodysplastic syndrome, Sup. Table 2). Average age of recruitment of *SNCA* deletion carriers was 52 (range 41–56), for *SNCA* duplication carriers average age of recruitment was 61 (range 49–69) (all European) and for *SNCA* complex event carriers average age was 61 (range 42–69) of which two were not grouped as European. Interestingly, two pairs of individuals are siblings based on their relatedness value (PIHAT –0.5) with reported birth years having –2 years difference. Pair one (subject #dup1 and #dup6) carry both a 1.5 Mb *SNCA* duplication, and pair two (subject #del2 and #del5) carry both a 0.8 Mb *SNCA* deletion.

#### 4. Discussion

*SNCA* missense and copy number gains are a clear cause of autosomal dominant PD. Here we assessed the UK Biobank cohort for potential pathogenic variants and genetic rearrangements in *SNCA*. None of the five known missense mutations were identified. Several other *SNCA* missense variants with conflicting pathogenicity reports were identified for which none of the carriers was reported to be a PD case. Although three individuals did report a positive family history for PD, the allele frequency of these variants was too low to robustly test whether this frequency is more than expected by chance. Note that we previously suggested that the *SNCA* H50Q variant is likely not pathogenic in an autosomal

dominant inheritance fashion (Blauwendraat et al., 2018) despite the presence of functional evidence (Boyer et al., 2019; Ruf et al., 2019).

We identified 6 copy number gains, 6 copy number loss and 18 complex *SNCA* events. Surprisingly, none of these carriers have been reported to be PD cases and although three had ICD10 codes that relate to recognised prodromal PD symptoms, none met published probability thresholds to be considered a prodromal case (Heinzel et al., 2019). Separately, one *SNCA* deletion carrier had an ICD10 code for schizophrenia which has been suggested to be a prodromal symptom in *SNCA* duplication carriers (Takamura et al., 2016). However, while *SNCA* copy number gains are a clear cause of autosomal dominant PD, *SNCA* copy number loss (deletions) variants are, to our knowledge, not reported to cause PD. One possible explanation for the lack of PD status in these subjects is that they will develop PD at a later stage. The age of recruitment for the six duplication carriers is 61 (range 49–69). A large recent meta-analysis of duplication carriers showed an average age of onset of 46.9 with a range of 30–73 (Book et al., 2018), suggesting that these subjects could develop PD at a later age.

The identification of 18 potential mosaic events is of interest, especially given the prior evidence of *SNCA* mosaicism in a PD case (Pankratz et al., 2011) and mosaicism in synucleinopathies brain tissue (Perez-Rodriguez et al., 2019; Mokretar et al., 2018). Fourteen of the identified genetic rearrangements were classified as large complex events (100 + Mb) with likely varying levels of mosaicism (Conlin et al., 2010). Three were likely mosaic deletions of the larger *SNCA* region with 20–50 Mb size, while one was likely a uniparental disomy event spanning the majority of the long arm of chromosome 4. In the UK Biobank data, clonal hematopoiesis, mosaic effects and its association with cancer have already been shown to be present on a larger scale than previously assumed (Loh et al., 2018; Tuke et al., 2020; Loh et al., 2020; Dawoud et al., 2020). Therefore, not surprisingly, seven (all large complex events) of the 18 (39%) potential mosaic subjects showed a history of blood-based cancer. It is important to note that the current data is based on blood derived DNA, and it is unclear whether these effects are also present in more PD related tissues and cells like the brain due to potential clonal hematopoiesis. Besides it is unclear what the effect of these events is on *SNCA* gene expression and a-synuclein protein expression. A deeper investigation of these genomic events is needed before robust disease associations can be made.

Although our findings are based on a population level cohort, there are several clear limitations. First, the gold standard of copy number variant detection is multiplex ligation-dependent probe amplification (MLPA). However, DNA of these subjects is not readily available making validation of these genomic events not possible. In an attempt to overcome this limitation, we partially validated a subset of the larger genomic events using the available exome sequencing data in which clear events were detected, although gene and protein expression investigations would be necessary to confirm the functional consequences. Unfortunately the smaller genomic events were not validated in the exome sequencing data likely due to limit of coverage in the smaller regions. More future extensive validation is needed to determine the exact size and downstream effects of these variants, especially of the complex mosaic events. Second, although the UK Biobank is the largest

and most comprehensive genotype/phenotype cohort to date there is the potential of missing phenotype data when using electronic health record (EHR) ICD10 codes for various reasons. Therefore, although PD or a prodromal PD phenotype could have been missed in some instances, it is unlikely that this information is missing in all carriers of potential duplication events. Third, our approach is partly based on manual interpretation of prioritized subjects and is biased towards larger events and smaller deletions (i.e. single exon) or complex rearrangements are easily missed due to the genotyping array design (spacing and number of variants in the *SNCA* gene region). Although manual interpretation is not ideal, it is more likely to be conservative than automatically generated algorithms since these often also require a manual interpretation step. Of note, multiple copy number variant events were detected outside of the *SNCA* gene body region which can have a significant impact on expression. To confirm this, RNA sequencing data would be needed.

Overall, here we identified that copy number variants and mosaicism in the *SNCA* region are present in the general population without reported PD symptoms. These subjects are outstanding candidates for more thorough investigation of *SNCA* levels, general *SNCA* biology, potential clinical trial candidates and how these duplication carriers potentially escape PD. Mosaicism of the *SNCA* gene region in blood derived genotype data is of interest, although this mechanism needs to be explored on a larger case-control scale to assess whether this is associated with disease or not.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements and funding

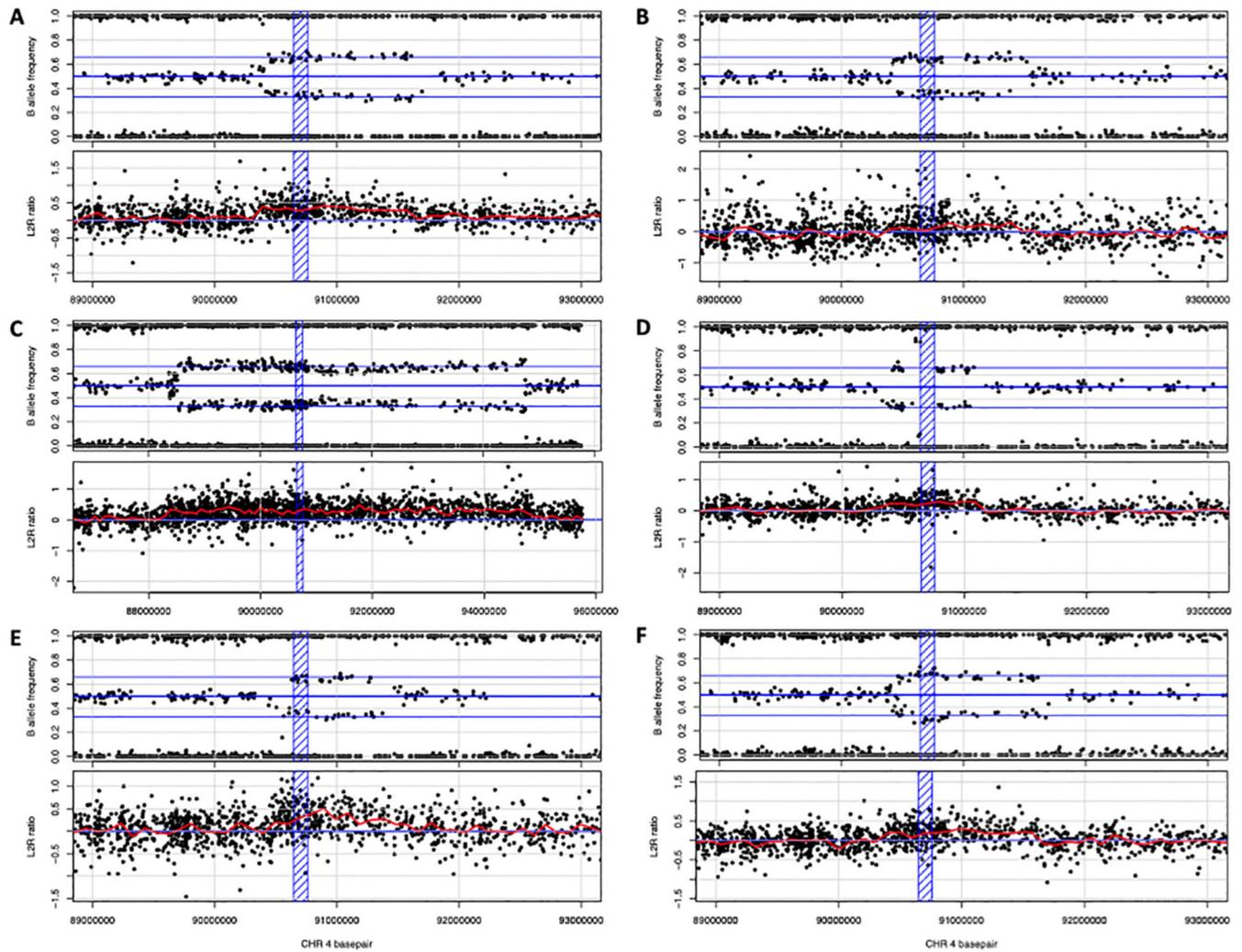
We would like to thank all of the subjects who donated their time and biological samples to be part of this study. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. This research has been conducted using the UK Biobank Resource under Application Number 33601. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (<http://hpc.nih.gov>). AJN works at the Preventive Neurology Unit which is funded by the Barts Charity.

## References

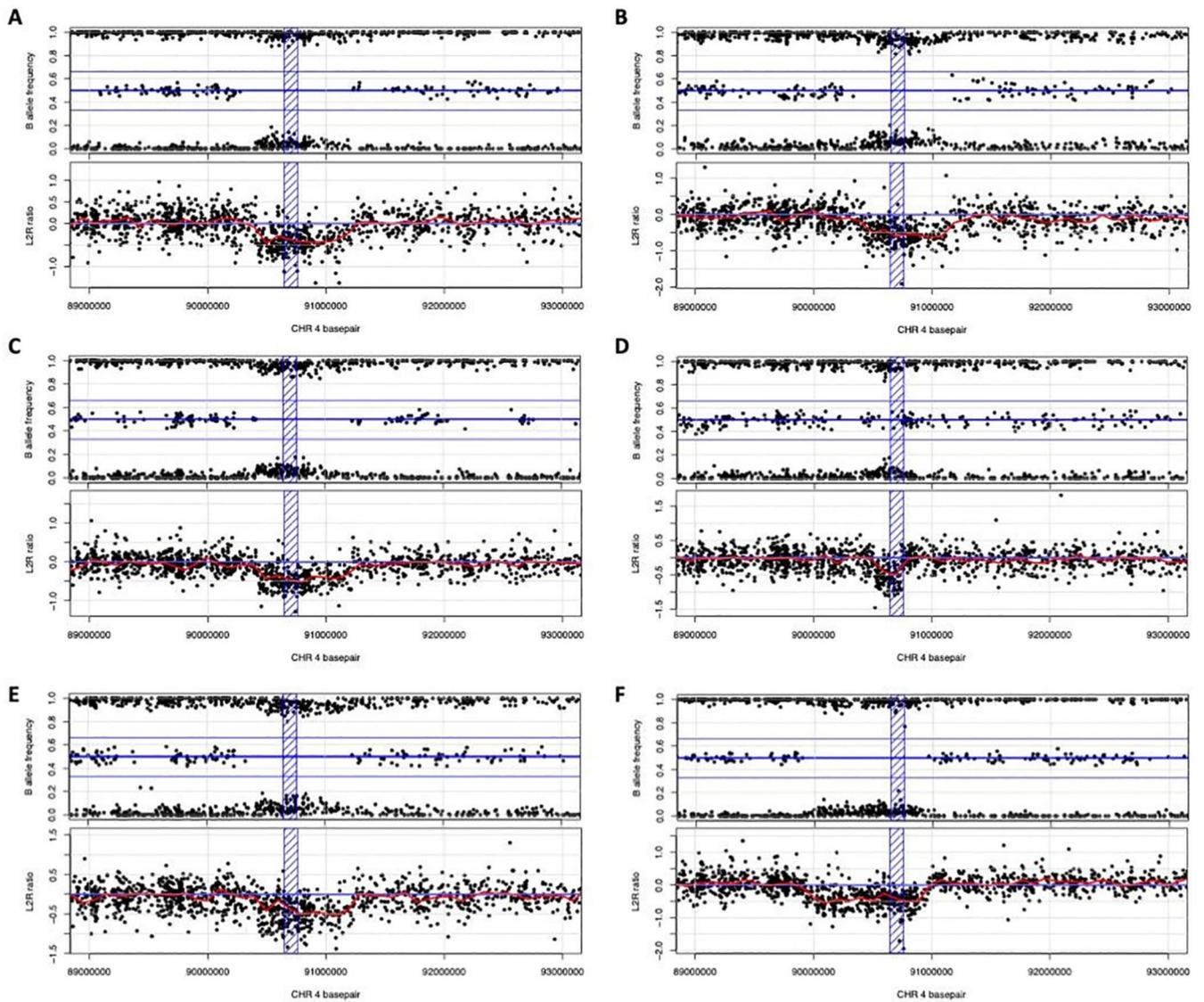
- Blauwendraat Cornelis, Kia Demis A., Pihlstrøm Lasse, Gan-Or Ziv, Lesage Suzanne, Gibbs J Raphael Ding, Jinhui, et al., 2018 Insufficient evidence for pathogenicity of SNCA His50Gln (H50Q) in Parkinson's disease. *Neurobiol. Aging* 64 (4), 159 e5–159.e8.
- Blauwendraat Cornelis, Heilbron Karl, Vallerga Costanza L., Bandres-Ciga Sara, von Coelln Rainer, Pihlstrøm Lasse, Simón-Sánchez Javier, et al., 2019 Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and  $\alpha$ -Synuclein mechanisms. *Mov. Disord* 34 (6), 866–875. [PubMed: 30957308]
- Book Adam, Guella Ilaria, Candido Tara, Brice Alexis, Hattori Nobutaka, Jeon Beomseok, Farrer Matthew J., SNCA Multiplication Investigators of the GEPD Consortium, 2018 A meta-analysis of  $\alpha$ -Synuclein multiplication in familial parkinsonism. *Front. Neurol* 9 (12), 1021. [PubMed: 30619023]
- Boyer David R., Li Binsen, Sun Chuanqi, Fan Weijia, Sawaya Michael R., Lin Jiang, Eisenberg David S., 2019 Structures of fibrils formed by  $\alpha$ -Synuclein hereditary disease mutant H50Q reveal new polymorphs. *Nat. Struct. Mol. Biol* 26 (11), 1044–1052. [PubMed: 31695184]

- Bycroft Clare, Freeman Colin, Petkova Desislava, Band Gavin, Elliott Lloyd T., Sharp Kevin, Motyer Allan, et al., 2018 The UK biobank resource with deep phenotyping and genomic data. *Nature* 562 (7726), 203–209. [PubMed: 30305743]
- Chang Christopher C., Chow Carson C., Tellier Laurent Cam, Vattikuti Shashaank, Purcell Shaun M., Lee James J., 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4 (2), 7. [PubMed: 25722852]
- Chartier-Harlin Marie-Christine, Kachergus Jennifer, Roumier Christophe, Mouroux Vincent, Douay Xavier, Lincoln Sarah, Levecque Clotilde, et al., 2004 Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* 364 (9440), 1167–1169. [PubMed: 15451224]
- Conlin Laura K., Thiel Brian D., Bonnemann Carsten G., Medne Livija, Ernst Linda M., Zackai Elaine H., Deardorff Matthew A., Krantz Ian D., Hakonarson Hakon, Spinner Nancy B., 2010 Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum. Mol. Genet* 10.1093/hmg/ddq003.
- Dawoud Ahmed A.Z., Tapper William J., Cross Nicholas C.P., 2020 Clonal myelopoiesis in the UK biobank cohort: ASXL1 mutations are strongly associated with smoking. *Leukemia* 34 (10), 2660–2672. [PubMed: 32518416]
- Heinzel Sebastian, Berg Daniela, Gasser Thomas, Chen Honglei, Yao Chun, Postuma Ronald B., MDS Task Force on the Definition of Parkinson's Disease, 2019 Update of the MDS research criteria for prodromal Parkinson's disease. *Mov. Disord* 34 (10), 1464–1470. [PubMed: 31412427]
- Karczewski Konrad J., Francioli Laurent C., Tiao Grace, Cummings Beryl B., Alföldi Jessica, Wang Qingbo, Collins Ryan L., et al., 2020 The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581 (7809), 434–443. [PubMed: 32461654]
- Konno Takuya, Ross Owen A., Puschmann Andreas, Dickson Dennis W., Wszolek Zbigniew K., 2016 Autosomal dominant Parkinson's disease caused by SNCA duplications. *Parkinsonism Relat. Disord* 22 (Suppl 1 (1)), S1–S6. [PubMed: 26350119]
- Loh Po-Ru, Genovese Giulio, Handsaker Robert E., Finucane Hilary K., Reshef Yakir A., Palamara Pier Francesco, Birman Brenda M., et al., 2018 Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* 559 (7714), 350–355. [PubMed: 29995854]
- Loh Po-Ru, Genovese Giulio, McCarroll Steven A., 2020 Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* 584 (7819), 136–141. [PubMed: 32581363]
- Mokretar Katya, Pease Daniel, Taanman Jan-Willem, Soenmez Aynur, Ejaz Ayesha, Lashley Tammarny, Ling Helen, et al., 2018 Somatic copy number gains of  $\alpha$ -Synuclein (SNCA) in Parkinson's disease and multiple system atrophy brains. *Brain J. Neurol* 141 (8), 2419–2431.
- Nalls Mike A., Blauwendraat Cornelis, Vallerga Costanza L., Heilbron Karl, Ciga Bandres, Sara Chang, Diana Tan, Manuela, 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 (12), 1091–1102. [PubMed: 31701892]
- Pankratz Nathan, Dumitriu Alexandra, Hetrick Kurt N., Sun Mei, Latourelle Jeanne C., Wilk Jemma B., Halter Cheryl, et al., 2011 Copy number variation in familial Parkinson disease. *PLoS One* 6 (8), e20988.
- Perez-Rodriguez Diego, Kalyva Maria, Leija-Salazar Melissa, Lashley Tammarny, Tarabichi Maxime, Chelban Viorica, Gentleman Steve, et al., 2019 Investigation of somatic CNVs in brains of synucleinopathy cases using targeted SNCA analysis and single cell sequencing. *Acta Neuropathol. Commun* 7 (1), 219. [PubMed: 31870437]
- Pihlstrøm Lasse, Blauwendraat Cornelis, Cappelletti Chiara, Berge-Seidl Victoria, Langmyhr Margrete, Henriksen Sandra Pilar, van de Berg, Wilma DJ, et al., 2018 A comprehensive analysis of SNCA -related genetic risk in sporadic Parkinson disease. *Ann. Neurol* 10.1002/ana.25274.
- Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, et al., 1997 Mutation in the alpha-Synuclein gene identified in families with Parkinson's disease. *Science* 276 (5321), 2045–2047. [PubMed: 9197268]
- Ruf Viktoria C., Nübling Georg S., Willikens Sophia, Shi Song, Schmidt Felix, Levin Johannes, Botzel Kai, Kamp Frits, Giese Armin, 2019 Different effects of  $\alpha$ -Synuclein mutants on lipid binding and aggregation detected by single molecule fluorescence spectroscopy and ThT fluorescence-based measurements. *ACS Chem. Neurosci* 10 (3), 1649–1659. [PubMed: 30605594]

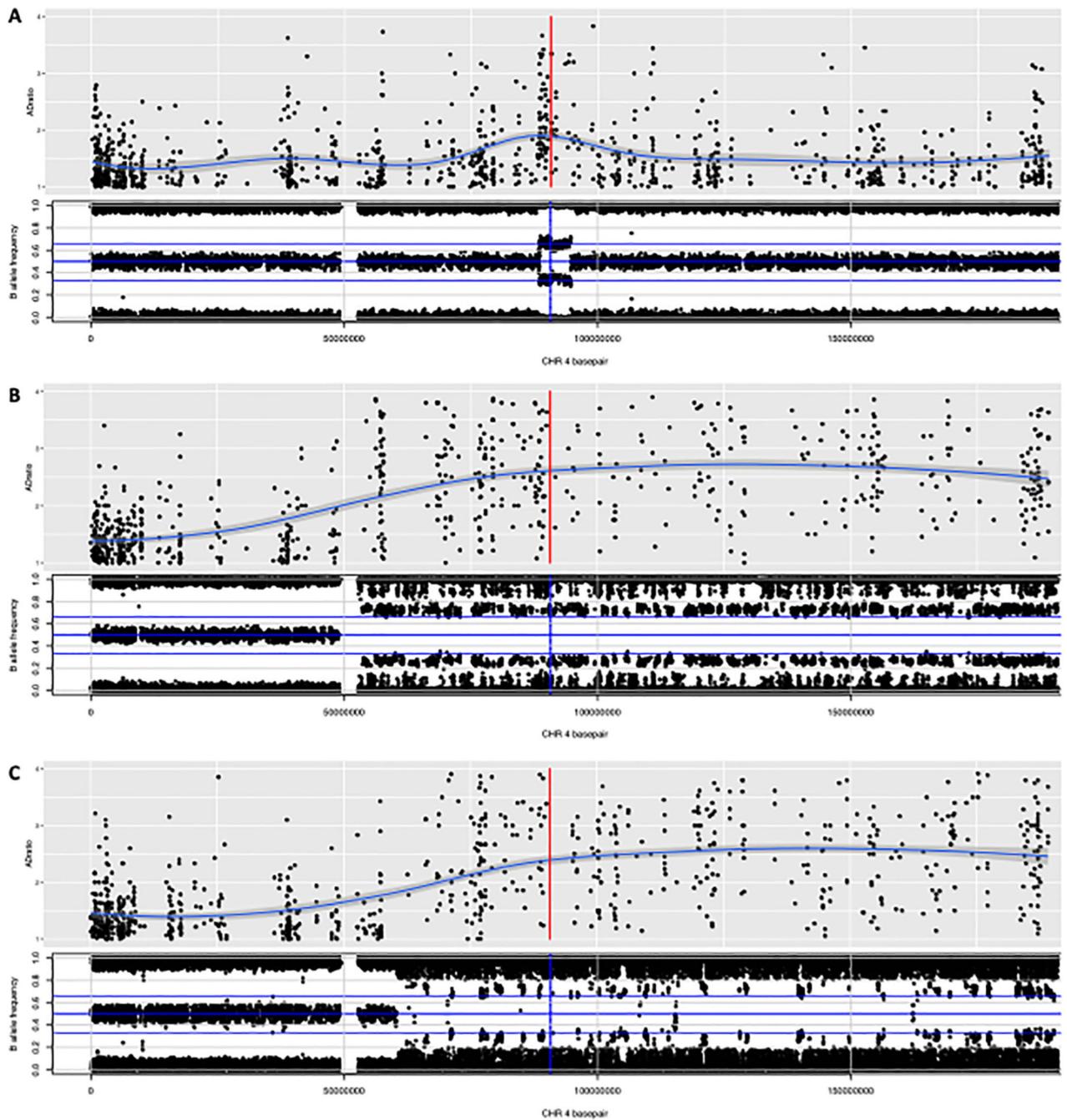
- Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, Hulihan M, et al., 2003 Alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 302 (5646), 841. [PubMed: 14593171]
- Soldner Frank, Stelzer Yonatan, Shivalila Chikdu S., Abraham Brian J., Latourelle Jeanne C., Barrasa Inmaculada M., Goldmann Johanna, Myers Richard H., Young Richard A., Jaenisch Rudolf, 2016 Parkinson-associated risk variant in distal enhancer of  $\alpha$ -Synuclein modulates target gene expression. *Nature* 533 (7601), 95–99. [PubMed: 27096366]
- Takamura Shogo, Ikeda Aya, Nishioka Kenya, Furuya Hirokazu, Tashiro Mari, Matsushima Takashi, Li Yuanzhe, et al., 2016 Schizophrenia as a prodromal symptom in a patient harboring SNCA duplication. *Parkinsonism Relat. Disord* 25 (4), 108–109. [PubMed: 26880146]
- Tuke Marcus, Tyrrell Jessica, Ruth Katherine S., Beaumont Robin N., Wood Andrew R., Murray Anna, Frayling Timothy M., Weedon Michael N., Wright Caroline F., 2020 Large copy-number variants in UK biobank caused by clonal hematopoiesis may confound penetrance estimates. *Am. J. Hum. Genet* 107 (2), 325–329. [PubMed: 32574563]
- Wang K, Li M, Hakonarson H, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 10.1093/nar/gkq603.
- Wickham Hadley, 2016 ggplot2: Elegant Graphics for Data Analysis. Springer.



**Fig. 1.**  
*SNCA* whole gene duplications identified in the UK Biobank cohort.



**Fig. 2.** *SNCA* whole gene deletions and one (likely) partial *SNCA* deletion (D) identified in the UK Biobank cohort.



**Fig. 3.** Partial validation of genomic events of the genotyping array data using exome sequencing data. A) Duplication of the *SNCA* gene region (subject #dup3), B) Partial duplication of the long arm of chromosome 4 (subject #comp8), C) Partial duplication of the long arm of chromosome 4 (subject #comp17) Red and blue vertical lines represents the *SNCA* gene body.