Cross-disorder patient and mechanistic stratification using combinatorial analyses

S. HOGG, **K. CHOCIAN**, S. DAS, A. MALINOWSKI, K. TAYLOR, S. BEAULAH
PrecisionLife, Long Hanborough, United Kingdom

PSTR178.30

Introduction

PrecisionLife is a computational biology company focusing on precision medicine analytics in complex chronic diseases. Our mechanistic patient stratification identifies subgroups of patients who share causal drivers of disease and treatment response, generating biomarkers that inform and de-risk drug discovery and development.

Central Nervous System (CNS) diseases are often characterized by a high degree of heterogeneity across the patient populations, reflected in a wide range of disease presentations and therapy responses. In many of these indications Genome Wide Association Studies (GWAS) have identified a number of disease-associated genes, but these findings have not translated into progress in clinical trials (1). This likely reflects the limitations of GWAS in only identifying single variants, while the key to understanding complex diseases that are influenced by multiple genetic loci is to find combinations of variants that distinguish one patient subgroup from another.

Further, identifying targets underlaying multiple indications would allow us to effectively target patient subgroups across CNS indications.

Methods

QUALITY CONTROL

After defining the criteria for cases and controls for a given indication, each dataset used goes through and extensive and stringent quality control which involves filtering of SNPs based on various criteria (MAF, HWE etc.), generation of QQ plot and GWAS results.

COMBINATORIAL ANALYSIS

The datasets were analysed in the PrecisionLife platform to identify combinations of SNP genotypes that, when observed together in a patient, are strongly associated with specific CNS disorders.

SNP combinations that have high odds ratios, low *p*-values and high prevalence in cases are prioritized. This process undergoes 1,000 cycles of fully randomized permutations and combinations must meet a specified FDR threshold. SNPs are scored using a Random Forest algorithm in a 5-fold cross-validation framework and prioritized based on their ability to differentiate cases and controls. The highest scoring SNPs are then mapped to genes and clustered by the patients they co-occur in to generate a disease architecture.

Results

Figure 1. Combinatorial analysis enables PL to detect additional signal in datasets that yield limited results using GWAS

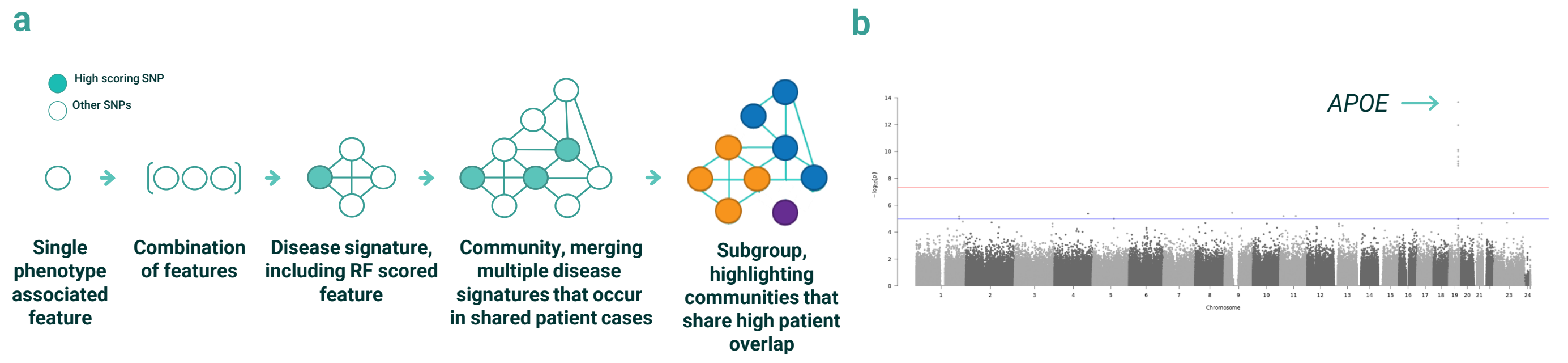
a
High scoring SNP
Other SNPs
Single phenotype associated feature
Combination of features
Disease signature, including RF scored feature
Community, merging multiple disease signatures that occur in shared patient cases
Subgroup, highlighting communities that share high patient overlap
b
APOE
Chromosome

Figure 1. (a) Conceptual representation of features, combinations, disease signatures and communities used to build up the disease architecture in the PrecisionLife combinatorial methodology. (b) Manhattan plot of genome-wide *p*-values of association for the AD UK Biobank cohort. The dashed line represents the genome-wide significance threshold at p<5e-08.

GWAS	Combinatorial Analysis
Single SNP associations must be significant across large groups of patients	Specific combinations of variants associated with each patient subgroup serve as a genetic stratification biomarker
Limited insights unless disease is likely to be caused by a small number of rare variants with large effect sizes (often in gene coding regions affecting protein 3D structure)	Patient subgroups with different causes of disease or even incorrect diagnoses can be distinguished (stratified) by different mechanistic etiology
Does not account for the effects of interactions between SNPs, genes and metabolic networks	Captures epistatic and non-linear additive effects of all interactions between SNPs, genes, environmental factors and metabolic networks

Figure 2. Combining results from multiple analyses can uncover underlying patterns across CNS indications

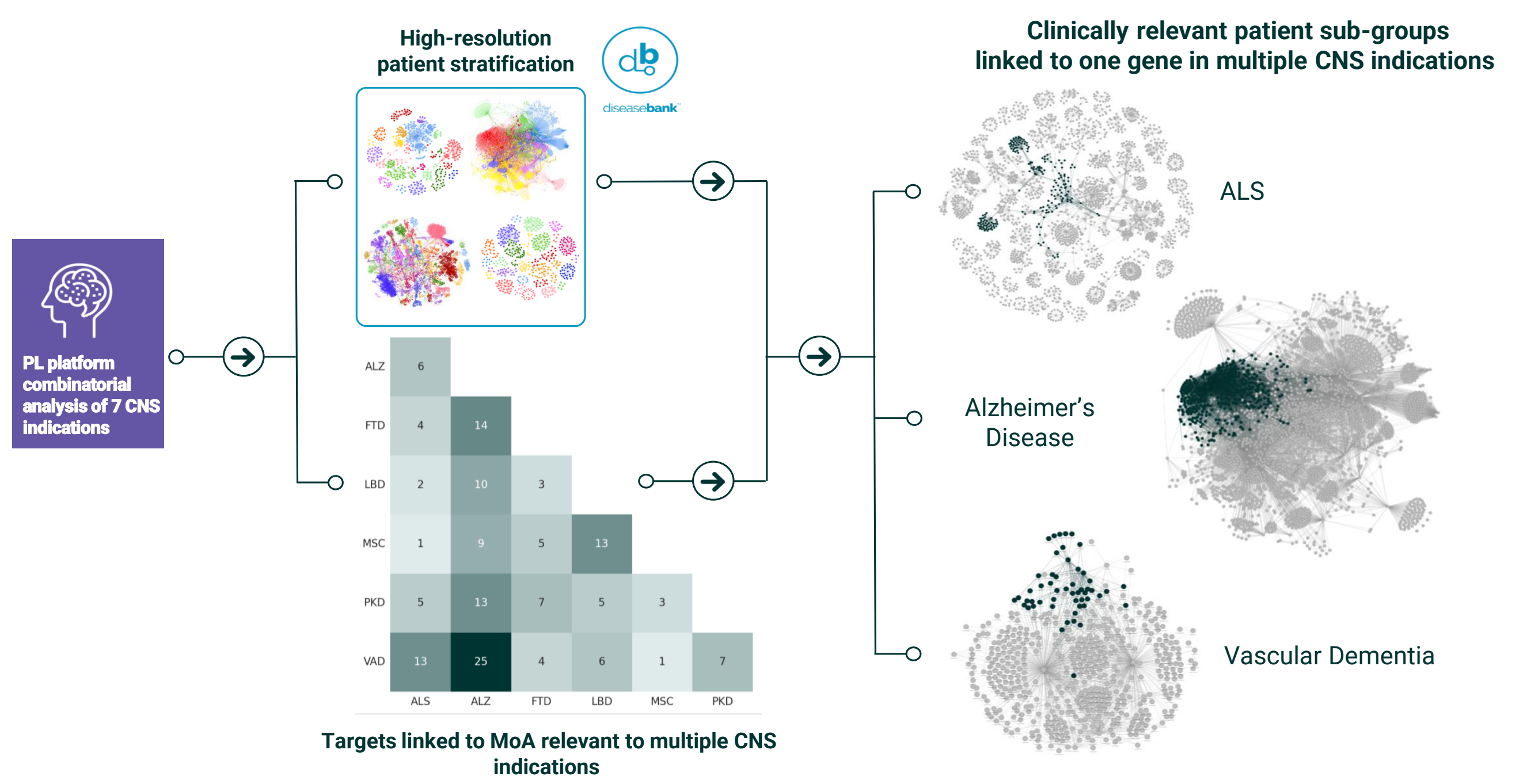
PL platform combinatorial analysis of 7 CNS indications
High-resolution patient stratification
diseasebank
Targets linked to MoA relevant to multiple CNS indications
ALZ
FTD
LBD
MSC
PKD
VAD
ALS
ALZ
FTD
LBD
MSC
PKD
6
4
14
2
10
3
1
9
5
13
5
13
7
5
3
13
25
4
6
1
7
Clinically relevant patient sub-groups linked to one gene in multiple CNS indications
ALS
Alzheimer's Disease
Vascular Dementia

Figure 2. Following the combinatorial analysis, PL creates patient stratification analysis and a gene overlap map. These approaches allow us to identify patient subpopulations that are connected to the genes or pathways overlapping across different indications.

Table 1. Datasets used for the PL combinatorial analysis in CNS indications

Amyotrophic Lateral Sclerosis (ALS)	Project MinE (WGS + UK2 and UK3 genotype datasets)
Alzheimer's disease (ALZ)	UKBB, GenADA
Frontotemporal dementia (FTD)	dbGAP DEMENTIA-SEQ
Lewy Body dementia (LBD)	dbGAP DEMENTIA-SEQ
Multiple sclerosis (MSC)	dbGAP
Parkinson's disease (PKD)	NeuroX
Vascular dementia (VAD)	UKBB

Table 1. Variety of data sources were used for PL combinatorial analysis. Because of the inherent differences in chip design, we expected to see less overlap between PKD and other CNS diseases. All datasets were processed according to PL standard prior to the analysis.

Figure 3. Cross-CNS similarities based on Gene Ontology: Biological Process enrichment and semantic similarity score

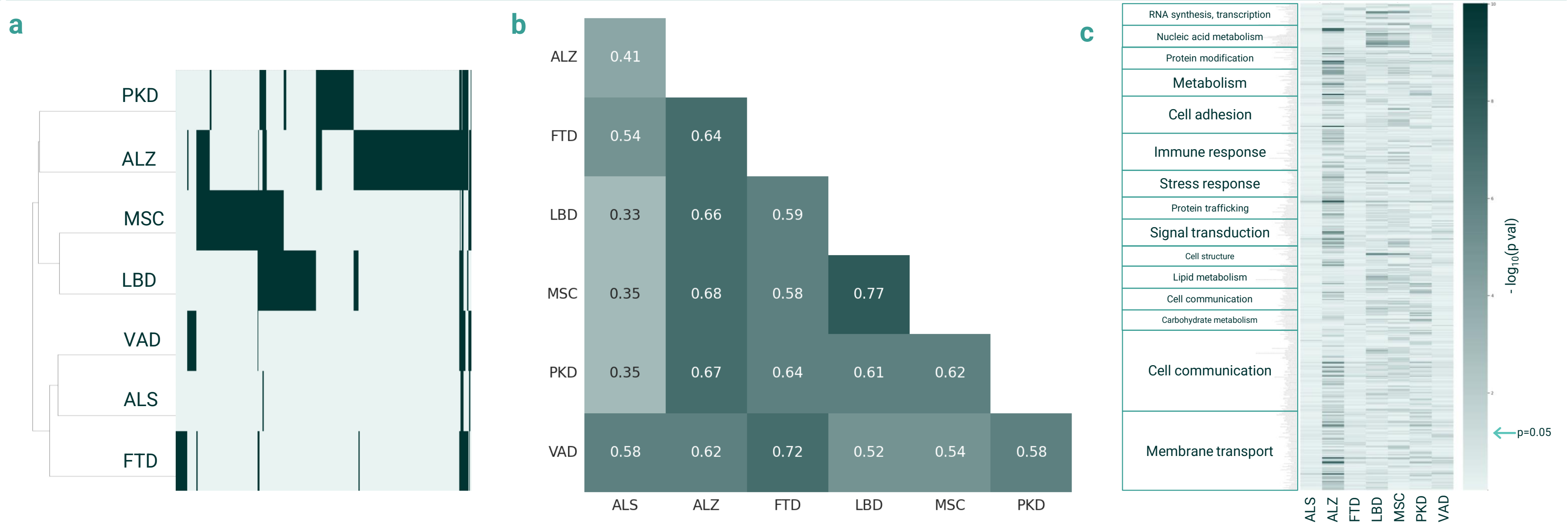

Figure 3 (a). Gene Ontology enrichment analysis of gene list from each indication was performed using g:Profiler (2). Clustering using scipy (Jaccard metric) is based on presence or absence of GO term in the list of enriched terms for a given indication (b) GOGO semantic similarity score (3) was calculated between the list of enriched GO terms for each indication The compound score across the lists of GO terms was calculated using Average Best-Matches (ABM) approach (4) (c) Heatmap of GO:Biological Process enriched terms in each of the indications (p<0.05, *p*-value correction for multiple testing using 'Benjamini-Hochberg', heatmap values correspond to - log10(p value)). GO terms were grouped using CateGOrizer (5) to visualise the main biological processes.

Figure 4. High level Reactome pathways connected to immune and stress functions

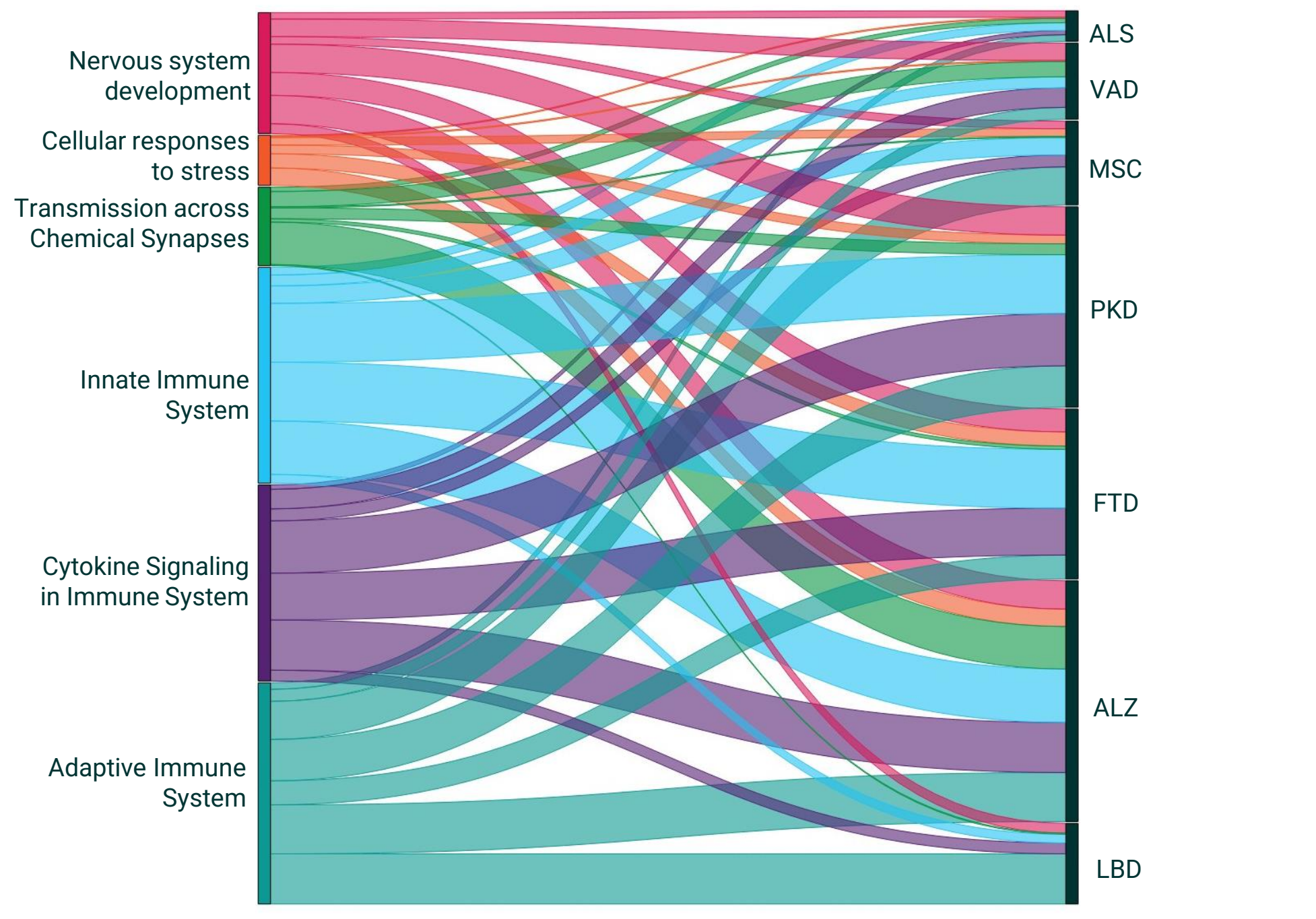

Figure 4. Sankey plot of the Reactome (6) level 2 ancestor pathways connected to neuronal, immune-response or stress-response pathways. Width of the Sankey ribbon is proportional to the number of connections between the high level Reactome pathway, and the genes connected to each indication in the PL knowledge graph.

Figure 5. Cross-indication look at gene expression in neuronal cell types

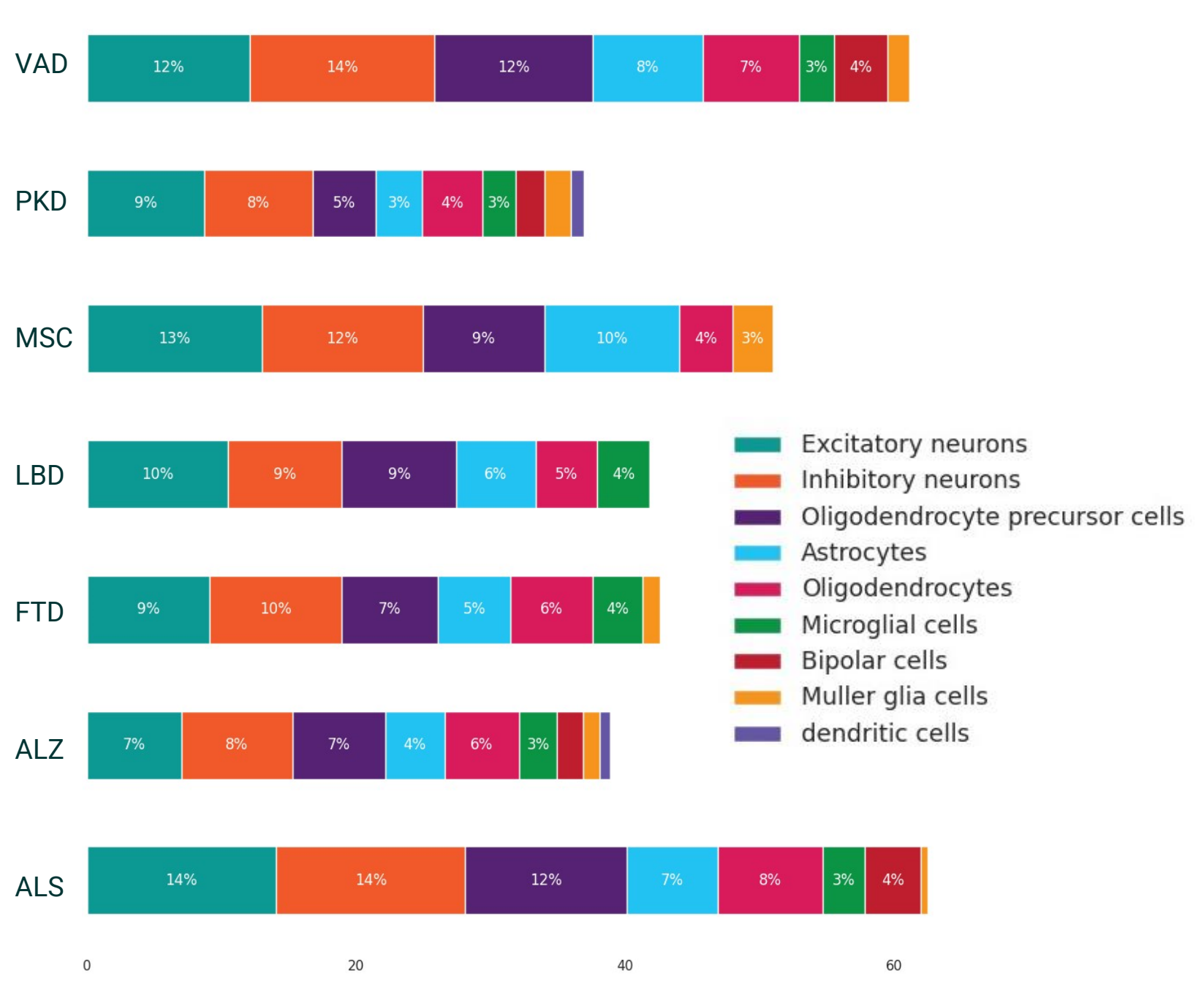

Figure 5. The composition bar plot showing the percentage of genes found in each indication that are expressed (and enriched) in a given neuronal cell type in Human Protein Atlas (7). For each of the indications, between 45-65% of the genes could be connected to single cell expression profile.

Figure 6. Overlap in behavioural mouse phenotypes across CNS indications

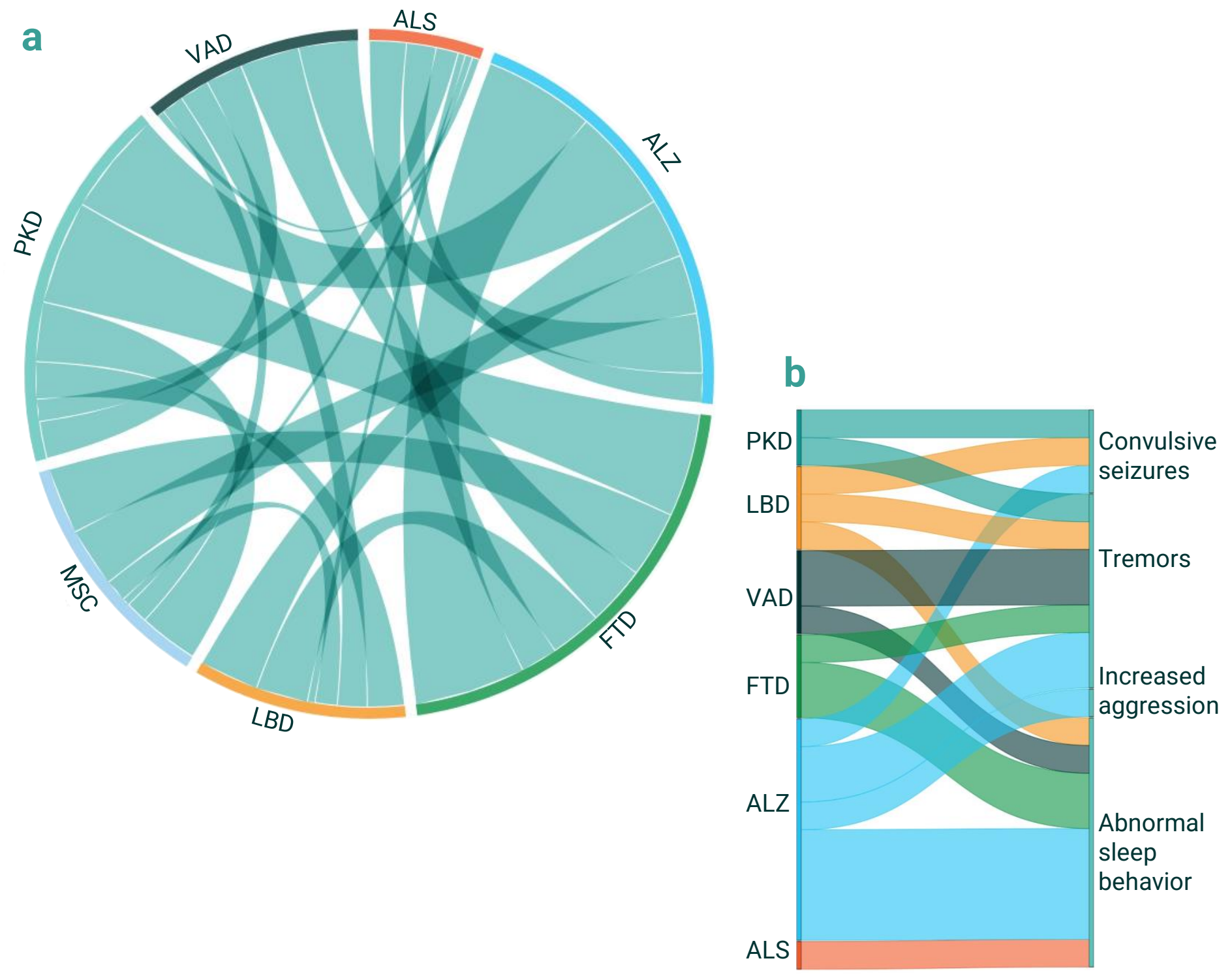

Figure 6 (a). Chord plot showing the behavioural mouse phenotypes shared between the CNS indications (from Mouse Phenotype Ontology (8)), the thickness of the ribbon is proportional to the number of phenotypes shared between the diseases. (b) Sankey plot showing a subset of behavioral phenotypes of interest and CNS diseases.

PATIENT STRATIFICATION

Clustering SNP genotypes combinations, based upon the patients in which they were found, generates distinct disease subgroups that can be defined by their genetic markers and specific biological functions, e.g., neuroinflammation, autophagy, serotonin receptor signaling, metal ion homeostasis, and adipose tissue differentiation/ fatty acid synthesis.

CROSS-DISEASE ANALYSIS

Results of each of the analysis are combined with the Patient Stratification results, and the PL Knowledge Graph database, to identify overlap in genetic markers, affected pathways and tissues, as well as protein function and model organisms' phenotypes.

Conclusion

Utilising a variety of techniques, from enrichment analysis, through semantic clustering and data mining, allows PL to identify genetic targets that can contribute to the underlying cause of multiple CNS indications or can be relevant for multiple patient subgroups across diseases. Further exploration of pathways of interest can be advantageous when investigating a specific MoA.

PL can use these insights to identify more effective therapeutic strategies and accompanying biomarker sets which match to patient subgroups across multiple CNS indications and can highlight opportunities for drug repurposing.

Acknowledgements

Research described in this study has been conducted using data from the UK Biobank Resource (application number 44288).

References

1. Tam V, Patel N, Turcotte M. et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet 20, 467–484 (2019).
2. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H: g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update) Nucleic Acids Research, May 2023
3. Zhao, C., Wang, Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. Sci Rep 8, 15107 (2018).
4. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274–1281 (2007).
5. Hu Z-L, Bao J and Reecy JM (2008) "CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories". Online Journal of Bioinformatics. 9 (2):108-112.
6. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C, Matthews L, Gong C, Deng C, Varusai T, Ragueneau E, Haider Y, May B, Shamovsky V, Weiser J, Brunson T, Sanati N, Beckman L, Shao X, Fabregat A, Sidiropoulos K, Murillo J, Viteri G, Cook J, Shorser S, Bader G, Demir E, Sander C, Haw R, Wu G, Stein L, Hermjakob H, D'Eustachio P, The reactome pathway knowledgebase 2022, Nucleic Acids Research, 2021
7. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Tissue-based map of the human proteome. Science 2015 347(6220):1260419.
8. Groza T, Lopez Gomez F, Haseli Mashhadi H, Muñoz-Fuentes V, Gunes O, Wilson R, Cacheiro P, Frost A, Keskivali-Bond P, Vardal B, McCoy A, Kwan Cheng T, Santos L, Wells S, Smedley D, Mallon A, Parkinson H, The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023

precisionlife
improving health | for everyone

For more information:
www.precisionlife.com