Microbiome biomarker for disease diagnosis:

Towards novel biomarker development with artificial intelligence algorithms

Krzysztof Odrzywołek[1,2], Raul Cano[3], Paweł Biernat[1], Błażej Szczerba[1], Jan Majta[1,4], Martha Carlin[3], Kaja Milanowska-Zabel[1]

1 Ardigen, Krakow, Poland, 2 Department of Computer Science, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, Krakow, Poland, 3 The BioCollective,Denver,CO, 4 Department of Computational Physics and Bioinformatics, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland

BACKGROUND

- The causes of Parkinson's disease (PD) are not fully defined, but at least some of them are genetic factors.
- 90% of cases are idiopathic - the causes are unknown (i.e. only a small part of it is genetic)
- One of the first symptoms is constipation, which is why attention was paid to the gut-brain axis.
- 75% of people with PD have gastrointestinal abnormalities, primarily constipation.
- The incidence of Parkinson's is higher in people with Crohn's disease, inflammatory bowel disease and ulcerative colitis.
- Alpha-synuclein proteins observed in the brain, can be also detected in the intestines

CHALLENGE

- "Combining TBC microbiome expertise with exciting AI tools (...) developed by Ardigen will translate exponentially to combat diseases."
- "The BioCollective extensive sample/data collection enabled well-founded discoveries"
- "Functional signatures can be translated into candidates for biomarkers, and further Tx"
- "To facilitate earlier diagnosis of Parkinson's Disease"
- "Better disease management for Parkinson's patients and their clinicians"
- "We need to understand more than species associacions - mechanism of action"

DATASETS

The BioCollective Cohort

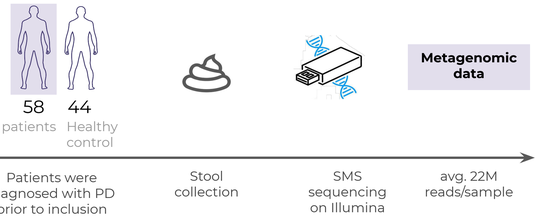

Fig. 1. Scheme of The BioCollective's process of microbiome data generation.

ANALYSIS OUTLINE

- Data analysis and interpretation were carried out using **Ardigen Microbiome Translational Platform** (MTP) developed by Ardigen.
- Shotgun sequencing data of described cohorts were processed with **Microbiome Scout** tool that identified **Metagenomic Features** associated with the analyzed phenotype.
- Selection performance was measured by ROC AUC of the used classifier and the classifier build with a selection of predictive features - Signature.
- We applied a number of microbiome processing approach to analyze their ability to expose the predictive signal.
- In terms of AI-Based profiling, we used all SMS samples for building a study-specific feature space. Ardigen's Microbiome Translational platform builds metagenomic **Feature Space** with up to 30% of previously unknown metagenomic sequences (Fig. 3a).
- All the metagenomic features sequences were encoded into a functionally-valid space (Fig. 3b).
- For classification problems, Random Forest and Logistic Regression algorithms were used as implemented in Ardigen's Microbiome Scout (Fig. 3c).
- Each model was validated in 10x repeated 5-fold cross-validation

CONFOUNDING FACTOR: DRUGS

- As gut microbiome is a fragile system, it's crucial to take into account confounding factors - in case of patients it usually a drug regiment
- In analyzed cohorts majority of patients took PD drugs (Fig. 2a) but no systematic impact of those on gut microbiome was detected (Fig 2b)

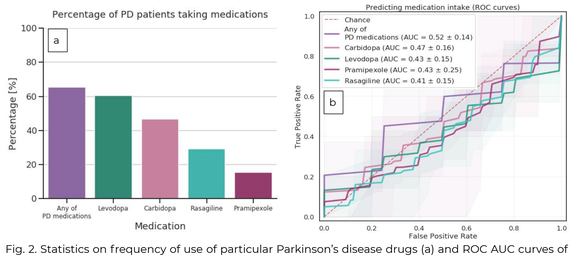

Fig. 2. Statistics on frequency of use of particular Parkinson's disease drugs (a) and ROC AUC curves of models trained to distinguish participants who uses particular drugs.

NO LOW HANGING FRUITS

- When standard bioinformatics methodology was applied, no difference in terms of the gut microbiome was found between Parkinson's disease patients and control group

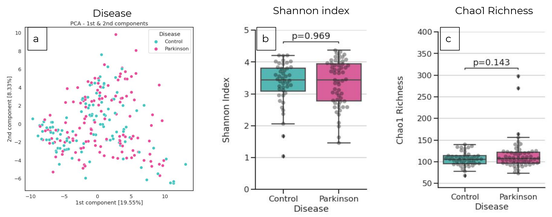

Fig. 3. PCA done on taxonomy profiles of study participants (a) and alpha diversity analysis of study participants (b,c) showing no significant differences between subgroups.

MICROBIOME-BASED CLASSIFICATIONS

- Initial approach was to use Machine Learning with taxonomy profiles as an input
- It led to significant improvement when compared to putative biomarker - abundance of Streptococcus genus (Fig 4a,b,c)
- Application of function-based features let to further improvement of results (d,e)

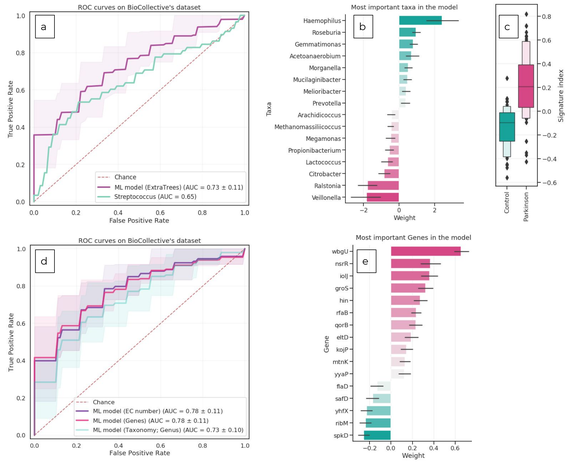

Fig. 4. ROC plots of a models based on taxonomy profiling of samples, compared to the one based on Streptococcus genus abundance (a) and the ones based on functional profiling (d), bacterial taxa (b) and genes (e) most important for PD prediction and a box plot showing PD Signature Score values for the cohort (c)

THE CHALLENGE OF THE UNKNOWN

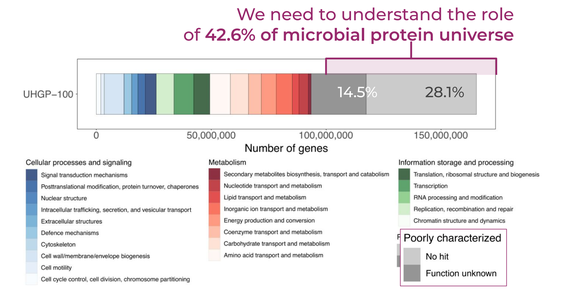

Fig. 5. Functional annotations of protein sequences stored in Unified Human Gastrointestinal Genome (UHGP [5]) - the largest gut microbiome protein collection currently available.

- Almost half of bacterial proteins are not understood in terms of its function (Fig 6.).
- Metagenomic sequencing studies detect more and more novel, unexplained proteins.
- Having this understanding gap growing, Ardigen aims at leveraging the information we already have.
- Advances in unsupervised learning methods enable the utilization of information stored in protein sequences catalogs, but requires functional discovery step to be included in the procedure.

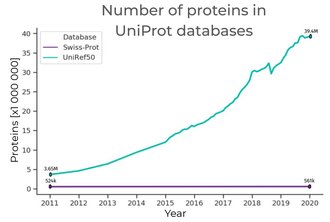

Fig. 6. Dynamic of growth of UniRef50 (unreviewed & automatically annotated) and Swiss-Prot (reviewed & manually-annotated) databases in 2010s

ARDIGEN'S FUNCTIONAL DISCOVERY ALGORITHM

- Ardigen develops a Functional Discovery algorithm for unannotated microbial proteins.
- We tested our approach on manually curated catalog of protein sequences: Swiss-Prot.
- The algorithm was trained to predict multiple functional annotations with high performance.
- Obtained results (Fig. 8) show that Functional Discovery algorithm properly assigns functions to unannotated sequences.
- Initial benchmark of Functional Discovery module's performance was done on protein sequences from UHGP catalog.
- We selected sequences that were annotated with standard tools: EggNOG mapper and InterProScan.
- Ardigen returns the functional annotation for 45-95% more proteins despite sequential dissimilarity.

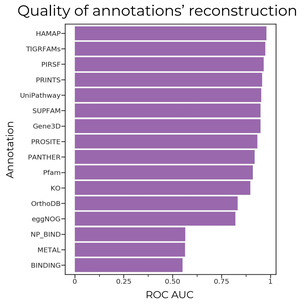

Fig. 7. Proportion of proteins from Swiss-Prot database accurately annotated with Ardigens algorithm

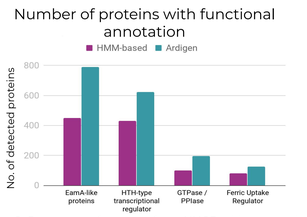

Fig. 8. Proportion of proteins from UHGP with assigned functional annotation with HMM-based, and Ardigen's method

AI-ENABLED PREDICTIONS

- Application of AI-based Functional Groups led to significant increase in ROC AUC value from 0.78±0.11 to 0.85±0.08 (Fig 9a)
- In the next step we rarefied the best performing and obtained a signature composed of ten AI-encoded Functional Groups (Fig 9b,c)

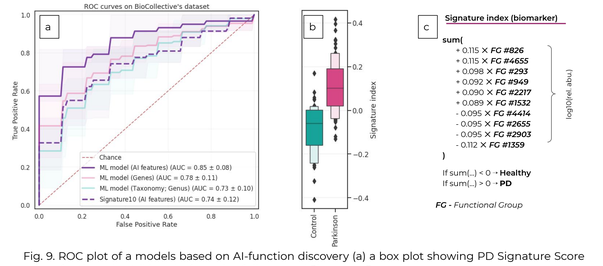

Fig. 9. ROC plot of a models based on AI-function discovery (a) a box plot showing PD Signature Score values for the cohort (b) and a formula of discovered Signsture Score (c).

POTENTIAL DIAGNOSTIC ASSAY

- Our AI-Function Discovery module allows for understanding of microbial functions behind the top predictive features (Fig 7, Fig 8)
- Glycosylation-related genes (e.g., wbgU) can provide markers of the altered intracellular biochemical processes and as such a potential diagnostic target (Fig 10a)
- In our study, the gene panel of wbgU, nsrR, iolJ, and groS/qorB is highly predictive in the diagnosis of Parkinson's Disease. (Fig 10a,b,c)

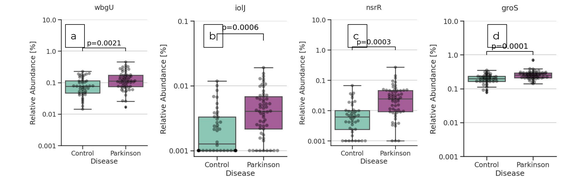

Fig. 10. Abundances of Genomic Features annotated with four bacterial genes that were identified as crucial for PD prediction

- We identified a set genomic features co-occuring in bacteria originating from the same *Bacteroides* genus suggesting that bacterial strains of this genus may be crucial for understanding of gut-PD relation

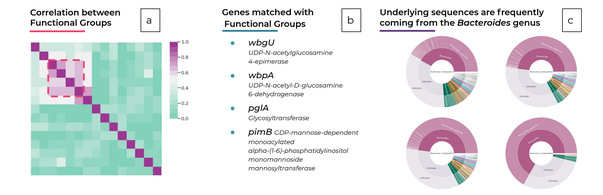

Fig. 11. A correlation matrix of top predictive Functional Groups' abundances (a), genes that Functional Groups were annotated to (b) and taxonomy profiles of Functional Groups' sequences (c)

OUTLOOK AND DEVELOPMENT DIRECTIONS

- There exists several molecular, enzymatic and chemical assays to assess presence/absence and quantification of **biomarkers for PD** used for
  - **diagnosis**
  - **stratification**
- A panel of **4 fecal microbiome expressed genes** (In *wbgU*, *nsrR*, *iolJ*, and *groS/qorB*) is highly predictive in the diagnosis of Parkinson's Disease.
- **Enzyme-based assays** of the above genes:
  - *wbgU* for GALE (galactose epimerase) or downstream enzymes (GalNAC-T3) by ELISA or immunohistochemistry
- **HPLC-based** blood/plasma as sample
  - Iron, ferritin, SOD, catalase,NOSx (*nsrR* gene)
  - Myo-inositol (*iolJ* gene)
  - Vitamin K/phylliquinone (*qorB* gene

In this work we used a technology developed in the project "BioForte Technology for in Silico Identification of Candidates for a New Microbiome-based Therapeutics and Diagnostics" cofunded by European Regional Development Fund (ERDF) as part of Smart Growth Operational Programme 2014-2020 (POIR.01.01.01-00-0347/17) awarded to Ardigen S.A.

contact: jan.majta@ardigen.com