"Celiac Disease Associated Gluten Sensitivity and the Human Gut Microbiome"

Several studies have indicated that there is a change in the gut microbiome composition and function in coeliac disease. Some of this change can even precede the onset of the disease, and contribute to the gluten intolerance observed in coeliac disease subjects. The genes HLA-DQ2 and HLA-DQ8 have been determined as causative factors for coeliac disease. In spite of this, the only effective treatment for the disease at the moment is complete avoidance of gluten in one's diet. However, latest research in coeliac gluten sensitivity has suggested that the inflammatory response triggered by wheat consumption can be inhibited by the consumption of probiotics, or by the immediate detoxification of gluten immediately after ingestion, in the gastrointestinal tract[1]. It can hence be hypothesized that an altered gut microbiome has some role to play in gluten sensitivity in coeliac disease, and an investigation into the compositional and functional differences in the gut microbiome of coeliac disease patients and healthy people can yield some interesting insights.

We hypothesize that the gut microbiome of coeliac disease patients differs significantly from the microbiome of non-coeliac people (with no gluten sensitivity) in both its composition and functional profile. The following metagenomics pipeline is proposed to test out this hypothesis:

Experimental protocol

Sample collection: An equal number of participants with coeliac disease and without coeliac disease or gluten sensitivity will be recruited. Since this is intended as an initial exploratory study, we will try to correct for any additional confounders on microbiome population by selecting participants based on the following criteria[2]:

i) Adult subjects between the ages of 18-40 years.
ii) Non-coeliac subjects must be defined as "healthy" based on the criteria described by Aagaard et al[3]. A modification of these criteria, whereby otherwise healthy subjects with a personal history of coeliacs disease, will be used to select coeliac subjects.
iii) An equal balance between the two groups (coeliacs and non-coeliacs), and between males and females within each group, must be maintained during subject recruitment.

Since the goal of this study is to identify differences in microbiome populations between coeliac and non-coeliac subjects, we must ensure we account for any biases from diet within the two groups (since the former is forcibly on a permanent gluten-free diet as a result of their condition). Thus, any recruits (from both groups) to the study must be willing to adhere to a shared, pre-determined gluten-free diet for a brief duration (5 days) in the study. The gut microbiome sampling process will entail collection of a stool specimen from each subject at multiple different timepoints, as described in Figure 1. After sample collection at t_{-1}, the subjects will all be required to switch to the study's gluten-free diet. The subsequent timeframes are chosen to span the turnover rate of the gut microbiome after significant dietary alteration (4 days), as determined by previous research[4]. The multi-timepoint samplings will help correct for any subject-specific influences on the microbiome composition, and ensure a statistically robust downstream analysis. Subjects must not make any drastic changes in their everyday diet regimen during the course of the study, unless indicated to as a part of the study design.

Data collection and processing

The study is designed as a culture-independent study. There will be two types of data collected from the stool samples – RNA-Seq (6 separate timepoints, as described in Figure 1), and 16S-rRNA (every day). Thus, with a low quantity of DNA isolate (20-50 ng) and a high sequencing depth (150x) on the Illumina MiSeq, we hope to generate high quality sequencing data from each subject's gut microbiome to be able to identify microbial communities. For the RNA-Seq data, after obtaining the RNA isolate from the stool samples, and generating short cDNA sequence reads, we will do some read quality and length filtering using previously published protocol for metagenomics processing[5]. We will use MegaBlast against the human genome to remove human contaminants. Other methods to remove human contaminants, like BLAT, are faster but can result in some off-target filtering – while this may very well account for sequencing error, since we are generating high quality data and most of the poor quality reads will hopefully be filtered out in the protocol mentioned before this step. Metagenomic assembly will be done using Celera, genome mapping will be generated using references from Silva, GREENGENES, and the Human Microbiome Project (HMP)[7]. Additionally, we will be using MetaGeneMark, an HMM based model, to identify the different genes present in each metagenomic sample[6] (Figure 2). An advantage of this model is that it can be pre-trained on a normal human gut microbiome – we will be using the 823 gut microbiota reference maps of the HMP to train a model for gene prediction.

16S ribosomal RNA (rRNA) sequences will also be obtained. The alignment and microbial community construction for each metagenomic sample will be done using Mothur[8]. In this instance, we will be using

HMMCP (mothur Community Profiling) data from the HMP for alignment against the list of known, annotated gut bacterial species. This works better than other predictive models that rely on clade specific markers (such as MetaPhlAn, AMPHORA), or a heuristic similarity search (BLAST). However, these methods are limited by the underlying databases, which tend to focus on bacteria that are easily cultivable in vitro. Currently, BLAST-based composition estimation overestimates the abundance of Proteobacteria and Firmicutes, and overlooks the potential for horizontal gene transfer within the sub communities in a metagenomic sample. Alternatively, we can also use MEGAN for binning our 16S reads, or use Phylopythia, which uses a sample specific model (that can be derived from the HMP gut reference genomes).

Relative Community Abundance Analysis

16S rRNA sequencing data will be used to track the changes in microbiome composition over the 15 days spanning this study. We will be following the Human Microbiome Project's protocol, Mothur, in order to interpret community composition. Mothur allows us to calculate the α and β diversity of each sample between the different dietary groups and timepoints. We quantify the microbial diversity within each subject at a given timepoint using α diversity as the index of measure, and gauge the changes in microbial diversity with the diet switch (at t=0, t=5) with the β diversity metric[4]. The significance of the change, in each case, is measured after correction for multiple-testing with the Bonferroni-corrected Mann-Whitney *U* test. This way, we test the hypothesis of whether differences in the composition of the gut microbiome have an influence on gluten sensitivity in coeliac patients and non-coeliac patients. It is expected that the α diversity will not change significantly within each time range (baseline, gluten-free, washout), but the β diversity is expected to change drastically in the non-coeliacs (switching to a gluten-free diet should not make any difference in the coeliac group). We will also do between-group comparisons of β diversity during the gluten-free diet stage, to assess whether the compositional complexity of one group is higher than the other. An alternate approach entails using Limma for differential gene expression analysis, on the RNA-Seq data, however, such a metric seems heavily confounded by the different microbial communities in the sample, and since we do not have a prior understanding of the community structure for each sample, it would be better to either use β diversity for between-group abundance analysis, or a tool developed specifically for metagenomics (ex. metagenomeSeq) that identifies features like Operational Taxonomic Units (OTUs) that are differentially abundant between groups.

Pathway Abundance Analysis

We will work with the gene annotated, filtered RNA-Seq data in this analysis. We will use the gene annotation and read mapping information to generate a functional profile for our metagenomic sample. In this section, we aim to identify the specifically enriched pathways, and score the coverage of different pathways with a pathway abundance graph. The procedures for this approach are described in Figure 2. This approach looks at overall metabolic and functional capabilities of the microbiome, instead of its specific communal architecture. As has been noted elsewhere with metagenomic analysis in humans, it may be possible that it is not specific bacterial subpopulations, but core pathways that are critical for a gluten-sensitive and a gluten-insensitive gut microenvironment[10]. Thus, here we test the hypothesis of whether there is a selection on function, rather than phylogenetic identity, which influences gluten sensitivity in coeliac patients[10].

Comparison with other methods

As a way of comparison with existing methodologies, we can compare the results of our analyses pipelines with the output from MG-RAST for all our metagenomes. MG-RAST generates phylogenetic classification of sequences, and functionally classifies and compares multiple samples[11]. MG-RAST identifies protein encoding genes using BLASTX, GREENGENES, and Silva. It does a cross-sample comparison by looking at sequence similarity to any given protein in a subsystem. Since it does not add additional information from pathway specific databases (unlike our approach), we would expect to get more significant hits from our pipeline than with MG-RAST, but it would be worth a comparison!

Concluding remarks

The novelty of this approach relies on the time-based distinction between the gluten-free and normal diets for the subjects, and in the incorporation of both community profiling (16S rRNA) and functional profiling (RNA-Seq) for each subject in each group. Using the time-based analysis, we are able to account for any differences between subjects and groups that may exist because of individual food choices and environmental influences. This makes the metagenomics analysis much more powerful in terms of the interpretations that can be drawn from the results.

Bioinformatics pipeline overview: Attached as pdf

Figures

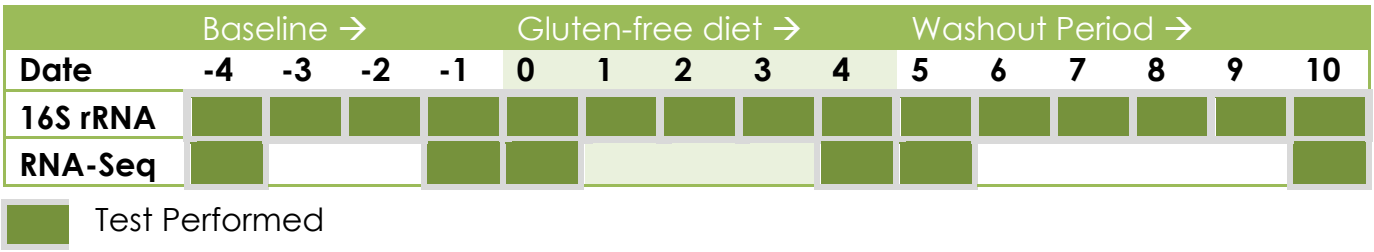

Figure 1. Sampling protocol for all study subjects. This study design is adapted from **4**. All dates are defined relative to the start of the gluten-free diet regimen. The subjects are studied 4 days preceding the diet switch in order to establish the baseline metagenome composition, then tracked for 5 days during the gluten-free diet, and lastly for 6 days after being allowed to revert to their normal, pre-study diet (washout period). Thus, we have 3 time ranges in this approach: (t_{-4}-t_{-1}, t_0-t_4, and t_5-t_{10}). Comparative analysis was done with four samples per person per diet, with two biological replicates in each diet term.

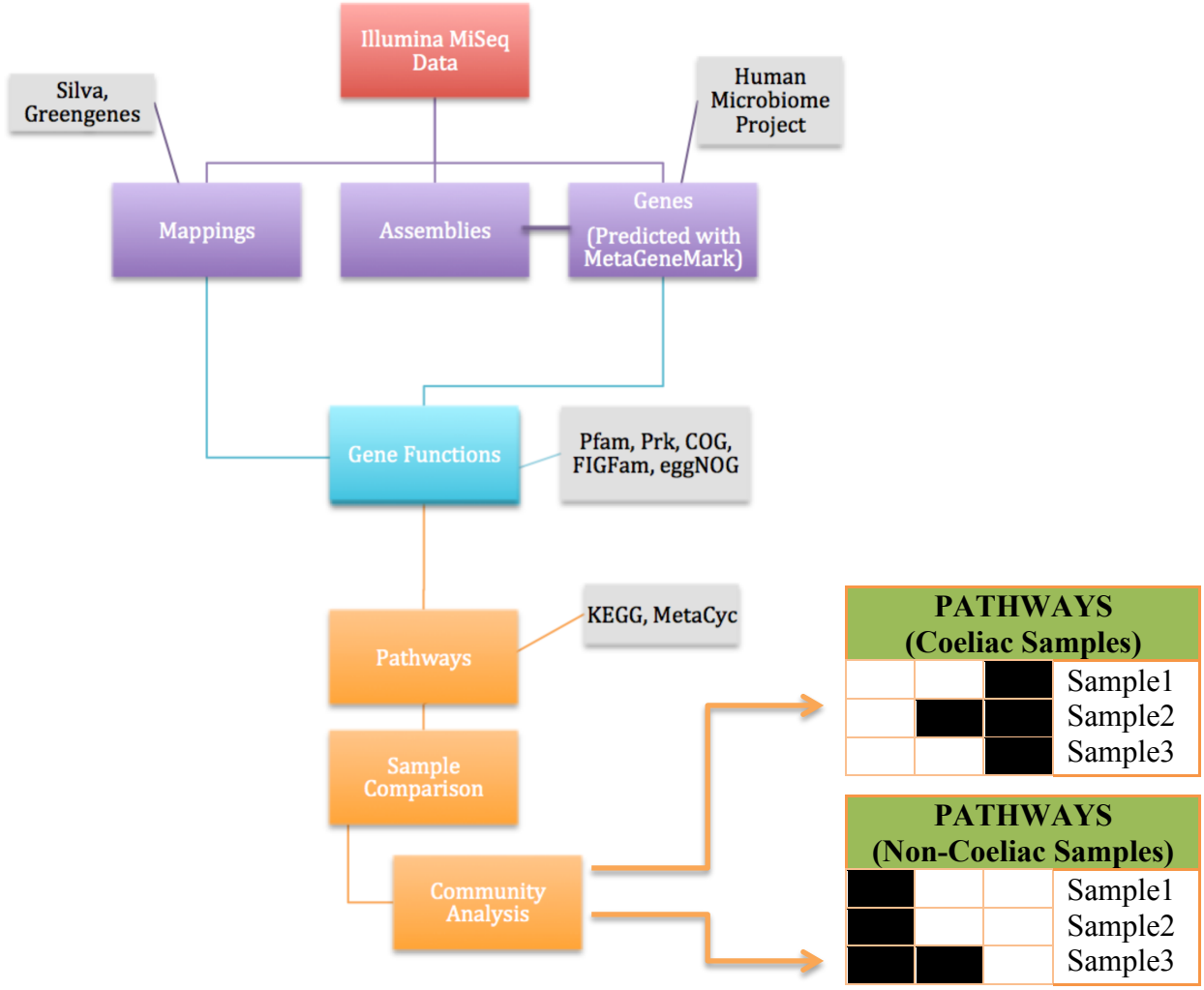

Figure 2. Workflow of the Pathway Abundance Analysis. The intent of this section is to highlight which pathways specifically enriched across the two groups (coeliac versus non-coeliac). This is

expected to provide insight into population-wide systems that may be influencing differences in coeliac and non-coeliac gut microbiome behaviours. This pipeline is inspired Cavalleri et al[9]. The expected outcome of this workflow is summarized in the two separate 'heatmap blocks', whereby we see some pathways that are on/off arbitrarily between the two groups, some that are specifically enriched only in the coeliac samples' metagenomes, and some that are specifically enriched only in the non-coeliac samples' metagenomes.

References

1. Lindfors, K.; Blomqvist, T.; Juuti-Uusitalo, K.; Stenman, S.; Venäläinen, J.; Mäki, M.; Kaukinen, K. (2008-06-01). "Live probiotic Bifidobacterium lactis bacteria inhibit the toxic effects induced by wheat gliadin in epithelial cell culture". *Clinical & Experimental Immunology* **152** (3): 552–558. doi:10.1111/j.1365-2249.2008.03635.x. ISSN 1365-2249. PMC 2453197. PMID 18422736.
2. Consortium, The Human Microbiome Project (2012-06-14). "A framework for human microbiome research". *Nature* **486** (7402): 215–221. doi:10.1038/nature11209. ISSN 0028-0836. PMC 3377744. PMID 22699610.
3. Aagaard, Kjersti; Petrosino, Joseph; Keitel, Wendy; Watson, Mark; Katancik, James; Garcia, Nathalia; Patel, Shital; Cutting, Mary; Madden, Tessa (2013-03-01). "The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters". *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **27** (3): 1012–1022. doi:10.1096/fj.12-220806. ISSN 1530-6860. PMC 3574278. PMID 23165986.
4. David, Lawrence A.; Maurice, Corinne F.; Carmody, Rachel N.; Gootenberg, David B.; Button, Julie E.; Wolfe, Benjamin E.; Ling, Alisha V.; Devlin, A. Sloan; Varma, Yug (2014-01-23). "Diet rapidly and reproducibly alters the human gut microbiome". *Nature* **505** (7484): 559–563. doi:10.1038/nature12820. ISSN 0028-0836. PMC 3957428. PMID 24336217.
5. Abubucker, Sahar; Segata, Nicola; Goll, Johannes; Schubert, Alyxandria M.; Izard, Jacques; Cantarel, Brandi L.; Rodriguez-Mueller, Beltran; Zucker, Jeremy; Thiagarajan, Mathangi (2012-01-01). "Metabolic reconstruction for metagenomic data and its application to the human microbiome". *PLoS computational biology* **8** (6): e1002358. doi:10.1371/journal.pcbi.1002358. ISSN 1553-7358. PMC 3374609. PMID 22719234.
6. Noguchi, Hideki; Park, Jungho; Takagi, Toshihisa (2006-11-01). "MetaGene: prokaryotic gene finding from environmental genome shotgun sequences". *Nucleic Acids Research* **34** (19): 5623–5630. doi:10.1093/nar/gkl723. ISSN 0305-1048. PMC 1636498. PMID 17028096.
7. Abubucker, Sahar; Segata, Nicola; Goll, Johannes; Schubert, Alyxandria M.; Izard, Jacques; Cantarel, Brandi L.; Rodriguez-Mueller, Beltran; Zucker, Jeremy; Thiagarajan, Mathangi (2012-06-01). "Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome". *PLoS Computational Biology* **8** (6). doi:10.1371/journal.pcbi.1002358. ISSN 1553-734X. PMC 3374609. PMID 22719234.
8. Schloss, Patrick D.; Westcott, Sarah L.; Ryabin, Thomas; Hall, Justine R.; Hartmann, Martin; Hollister, Emily B.; Lesniewski, Ryan A.; Oakley, Brian B.; Parks, Donovan H. (2009-12-01). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities". *Applied and Environmental Microbiology* **75** (23): 7537–7541. doi:10.1128/AEM.01541-09. ISSN 1098-5336. PMC 2786419. PMID 19801464.
9. Filippo, Carlotta De; Ramazzotti, Matteo; Fontana, Paolo; Cavalieri, Duccio (2012-11-01). "Bioinformatic approaches for functional annotation and pathway inference in metagenomics data". *Briefings in Bioinformatics* **13** (6): 696–710. doi:10.1093/bib/bbs070. ISSN 1467-5463. PMC 3505041. PMID 23175748.
10. Shafquat, Afrah; Joice, Regina; Simmons, Sheri; Huttenhower, Curtis (2014-05-01). "Functional and phylogenetic assembly of microbial communities in the human microbiome". *Trends in microbiology* **22** (5): 261–266. doi:10.1016/j.tim.2014.01.011. ISSN 0966-842X. PMC 4008634. PMID 24618403.

11. "Analyzing Metagenomic Data: Inferring Microbial Community Function with MG-RAST | Argonne National Laboratory". *www.mcs.anl.gov*. Retrieved 2016-04-12.