LUDC-IRC Postdoctoral Program 2020 – Project 9

The role of gut phageome in diabetes and related traits
PI: Marju Orho-Melander
Co-PI: Isabel Gonçalves

Purpose and aims

To understand the role of gut bacteriophages and prophages ("phages") in the dynamics and affiliations with gut bacteria, and in relation to diabetes and related cardiometabolic traits in large prospective population cohorts.

1. To identify and quantify all bacterial species and phages, their encoded genes and functional potential, in faecal samples of >12,000 Swedish individuals and characterize their diversity connections and interrelationships
2. To investigate how the gut microbiota composition and diversity, or specific bacteria and phages associate with cardiometabolic risk traits and predict incidence of T2D and CVD, and to investigate if the risk is mediated by specific microbiome produced circulating metabolites.
3. To identify specific potential probiotic gut bacterial strains with beneficial effects on cardiometabolic health, and how such strains may interact with phages.

State-of-the-art/background

Global increase in the prevalence of obesity and type 2 diabetes (T2D) is a major threat to cardiovascular health [1]. The poor cardiovascular prognosis already at prediabetes state is difficult to reverse by treatments available today [1]. Therefore, major efforts are needed to find effective strategies for prevention of obesity and T2D in order to decrease the risk of CVD- morbidity and mortality.

During 2018-2019, more than 10000 gut microbiota studies were published, and over the past 15 years, a large number of studies have associated obesity, T2D [2] and CVD [3] with specific changes of the gut microbiota composition, analysed in stool samples. However, the human studies have so far been relatively small case-control-, cohort- or intervention studies, often with inconclusive results [4]. Novel mechanisms of potential importance for cardiometabolic diseases have emerged from rodent- and in vitro models, and studies in germ-free mice have provided some support for causal connections. However, large prospective population based studies with high statistical power and good quality data on the numerous confounding factors are lacking, and human data favouring causality is clearly insufficient [5].

The totality of microbial genes in the human gut exceeds the number of human genes by 1000-fold, and the expression of bacterial genes essentially affects and complements the functions of the human body. Bacteriophages and prophages (from now on called "phages"), known to regulate bacteria via lytic or lysogenic infections, are likely the most abundant biological entities in the human gut, but their contribution to bacterial composition and intestinal physiology has remained under-investigated [6]. However, a recent longitudinal study of 10 healthy humans described the fecal phageome as highly individual and stable up to one year, and with clear connections to gut bacteriome [7]. Therefore, to understand the role of gut bacteria in human health and disease, it is crucial to investigate the dynamics and affiliations between gut bacteria and phages in large population cohorts. Bacteriophages are able to change the composition and function of the bacterial microflora. Therefore, investigation of a multidimensional ecosystem should not be limited on one dimension only. From the perspective of therapeutic modulation, understanding of how phages modulate the gut bacterial population is of crucial medical interest [6].

Significance and scientific novelty

Global characterization of both gut bacterial and viral (phages) composition, diversity and functional capacity, and the between-host variability, in a large population is lacking. For the first time, utilizing ultra-high phylogenic resolution metagenomics, our study will determine the interplay between the gut bacteriome

and phageome and provide prospective evidence for their role in cardiometabolic traits and diseases. Our project has the capacity to detect potential new probitic bacterial strains and to provide information on the potential of phages in future diagnostic or therapeutic applications.

Preliminary and previous results

The metagenomic sequencing of MOS and SCAPIS cohorts totaling 12,200 samples has been finalized and bioinformatic processing is ongoing and will be finalized by summer 2020. We have got preliminary processed sequence-, taxonomy- (for bacteria and phages), gene-ontology-, and KEGG-pathway pilot data for 850 MOS samples and 1500 SCAPIS-Malmö samples, to be able to set up analytical pipelines, test feasibility of approaches, and perform preliminary analyses. For this application, we have performed analyses to demonstrate feasibility and methodological advances. Preliminary data of bacteriophages is presented below in **Figure 1**.

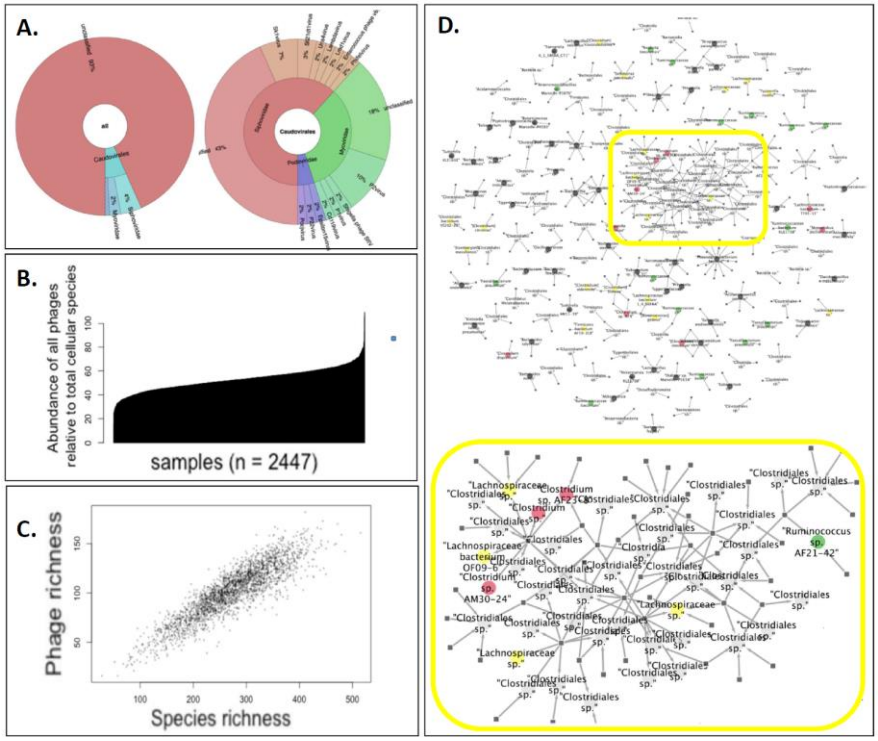
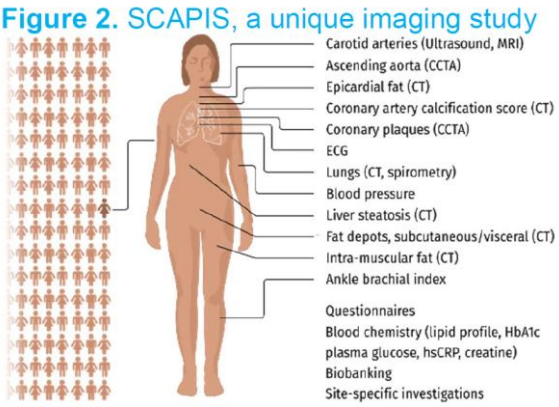


Figure 1. Profiling of bacterio-phages across 2447 deep sequenced MOS and SCAPIS-Malmö stool samples. **A.** Phages taxonomy with all profiled phages (n=972, left) and phages with sufficient similarity to existing phage reference genomes for taxonomical assignment (n= 65, right). **B.** Abundance of phages relative to the total cellular species (archaea, bacteria, fungi and protist) with an average 53% of the abundance signal from phages. **C.** Comparison of richness (total number) of cellular species and phages reveals high correlation (r2=0.8). **D.** Network of phage to bacterial host dependency associations. Phages are represented by black dots and bacterial species by colored circles, and edges represent statistically significant (FDR < 0.001) dependency associations. Most subnetworks are centered around a few species, but the encircled subnetwork is more interconnected. These results support our hypothesis of strong connections between phages and bacteria, and demonstrate the feasibility of the analytical pipeline.

Research plan

SCAPIS-Malmö and -Uppsala are part of the Swedish Cardio-Pulmonary bioImage Study (SCAPIS) [8], a population-based study of 30,000 participants in the age range of 50-64 years, including detailed heart, vessel- and pulmonary imaging (**Figure 2**) and with main funding by the Swedish Heart and Lung Foundation. The study has been conducted at five university hospitals including Malmö (SCAPIS-Mö) and Uppsala (SCAPIS-Up). Each participant was guided through a core 2-day program with blood sampling, questionnaires (including diet), extensive clinical examinations and detailed computed tomography imaging. Routine blood analyses as fasting glucose and blood lipids were performed at baseline. *Stool collection and microbiome sequencing in SCAPIS-Mö and SCAPIS-Up are add on investigations at Malmö and Uppsala nodes, with the applicant Orho-Melander as the main responsible for SCAPIS-Mö and Professor Tove Fall for SCAPIS-Up*. Main studies were completed 2018 and in total 5007 (Mö) and 5036 (Up) participants provided stool samples that have been metagenomic sequenced (**Table 1**).



Malmö Offspring Study (MOS) is an ongoing inter-generational population-based study where all adult offspring, i.e. children and grandchildren (age >18y, N~5 000) to participants of the Malmö Diet and Cancer

Study-Cardiovascular Cohort (MDC-CC) are invited to participate. At baseline examination fasting blood, saliva and stool samples are collected; anthropometry, blood lipids and vascular traits (blood pressure, pulse wave velocity and ultrasound of the carotid arteries) are measured; and dietary data is collected through a 4-day web-based dietary record. Further, a self-administrated web-based questionnaire including questions on environmental factors, family history, educational achievement, socioeconomic factors, bowel symptoms, self-perceived stress, physical activity, drug-usage and medical history are filled in. To date, >4900 individuals have participated in MOS and data cleaning has been done for 2600 (**Table 1**) and fecal metagenomic sequencing has been performed for 2200.

Table 1. Selected clinical characteristics and dietary intakes of participants in MOS, SCAPIS-Malmö and SCAPIS-Uppsala

	MOS N=2600	SCAPIS-Mö N=5007	SCAPIS-Up N=5036
Age (years)	38.9 ± 7.3	57.4 ± 4.3	57.6 ± 4.4
Women (%)	52	54	51
BMI (kg/m^2)	25.8 ± 4.7	27.3 ± 4.6	27.0 ± 4.4
Overweight [1] (%)	33.3	43.1	43.7
Obese [2] (%)	15.6	23.8	21.2
Diabetes (%)	5.1	9.1	9.0
Prediabetes [3](%)	10.5	19.2	22.4
Carotid plaque [4] (%)	15.8/13.8	15.7/12.8	11.2/10.4
Current smoker (%)	6.8	15.0	9.9
Fat (E%)	37 ± 7	36 ± 7	TBA
Carbohydrates (E%)	45 ± 8	47 ± 8	TBA
Protein (E%)	18 ± 4	17 ± 4	TBA
Fiber (g/1000kcal)	10 ± 3	12 ± 4	TBA
Fruit/veget (g/day)	262 ± 165	382 ± 275	TBA
Whole grain (g/day)	34 ± 40	43 ± 44	TBA

Data is mean± SD or %; 1) BMI >25 but <30 kg/m^2; 2) BMI>30 kg/m^2; 3) Fasting plasma glucose 6.1-6.9 mmol/l and/or HbA1c 42-47; 4) Two or more plaques on left/right carotid artery. TBA, to be analyzed (we have not yet analyzed Uppsala diet).Dietary data available in 1791 MOS and 4345 SCAPIS-Mö samples.

Dietary data collection and analysis: In MOS, dietary intakes are assessed using "Riksmaten 2010", a web-based 4-day record tool developed by the Swedish National Food Agency ("Livsmedels-verket"), and a food frequency questionnaire (FFQ) to assess habitual intake of foods that may not be captured within the 4 days. In SCAPIS, diet is assessed by a web-based FFQ. These dietary assessment tools have been validated. Associations with gut microbiota composition (bacterial species and phages), diversity and functional capacity will be examined across quantiles of intakes with the general linear model adjusting for age, sex, BMI, physical activity level, smoking and alcohol consumption, and other identified confounders.

Metagenomic analysis of gut (stool) microbiome: DNA-extraction, metagenomic sequencing (6 Gb Illumina 150 bp Paired-End) and bioinformatics processing is performed at Clinical Microbiomics (Copenhagen) by shotgun metagenomic analysis utilizing the so called *metagenomic species (MGS) concept* [described in detail in ref 9, Nature Biotech. 2014, with Chief Scientific Officer of Clinical Microbiomics, Henrik Bjørn Nielsen, PhD as the first author]. The approach is described in **Figure 3**. The MGS pipeline allows us to get to the strain level i.e. sub-species resolution based on identification of single nucleotide variants (SNV) within each MGS. The SNV profile then allows identification of sub-species populations and tracking of e.g. probiotic strains (**Figure 3B**).

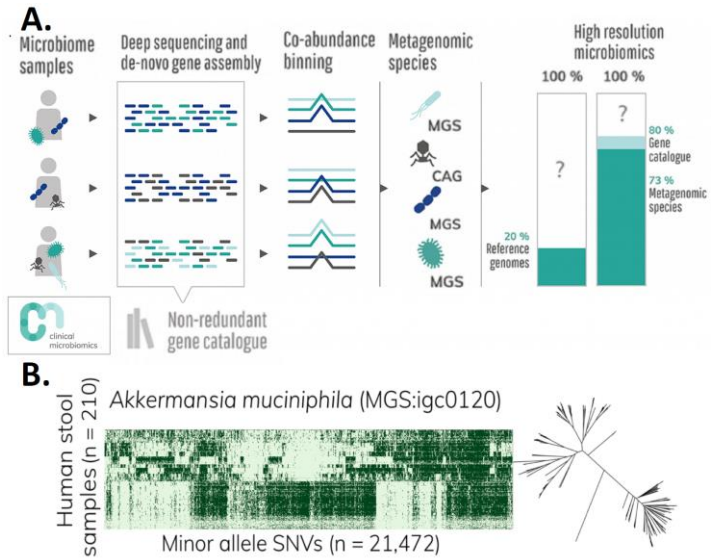


Figure 3. Overview of the metagenomic pipeline [9]. **A.** High quality microbial DNA is shotgun sequenced, genes are de-novo assembled, identified and integrated to form a cross-sample, non-redundant gene catalog. The bioinformatic approach is based on clustering of co-abundant genes into metagenomic species (MGS) that enables identification of both known and unknown species through de novo discovery of cohort-specific microorganisms thus covering the great majority of the diversity independent of reference sequences. The use of sample-specific sequence reads in the assemblies helps discriminate between closely related strains. The approach has also been adapted for cohort specific de novo discovery and profiling of bacteriophages (called co-abundancy gene groups, CAG). **B.** Example of the single nucleotide variation (SNV) based sub-species resolution. The data illustrates the SNV profile identified for the Akkermansia muciniphila MGS in human stool samples (left) at Clinical Microbiomics. The SNV-based sub-species populations is visualized with a phylogenetic tree (right).

Bioinformatic and statistical analysis of the metagenomic data: Microbiota data will be analyzed both at the community level and as specific bacteria and phages. For community level studies we use ecological analyses to assess diversity: i.e. *alpha-diversity* is calculated to assess richness by Shannon diversity index and *beta-diversity* is calculated to assess differences in the microbiome composition between individuals (or specific outcomes) by Bray-Curtis dissimilarity index. Significance of differences are tested by permutation tests. The UniProt Reference Clusters (UniRef50, www.uniprot.org) are further mapped to pathways from KEGG: Kyoto

Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg) and grouped into clusters of orthologous groups from the EggNOG data (www.egnogdb.embl.de, www.geneontology.org). The gene family abundances are further grouped into broader functional categories based on annotations for the UniProt proteins and gene ontology (GO). KEGG functional profiling of microbial communities is done using the integrated gene catalogue of the human gut microbiome (IGC).

Phage definition and taxonomy annotation: The phage data is profiled de novo similarly but with specific criteria related to the phage gene content and number of genes (**Figure 3**). A Co-abundant Gene group (CAG) qualifies as "phage-like" if 4 or more of 16 key phage protein families (Phage Tail, Terminase, Phage Capsid, Phage Portal, Phage DNA packaging, Phage Baseplate, Holine etc) are found in the CAG and that it is statistically enriched for genes that encode these key phage proteins, or the CAG has a one to one sequence similarity to a known phage [9]. When the phages have been identified they are assigned taxonomy and the reads are mapped to phage definitions. The relative abundance of the phages is transformed to be relative to the bacterial reads.

Dependency associations to affiliate phages to bacterial species: We will identify significant depend¬ency associations by comparing the absence-presence profiles throughout all samples for all pairs of phages and bacterial species using Fisher's exact test [9]. The relationships where a potential dependent phage is observed independently of the potential hosting bacteria are excluded and directional networks are created from the dependency-associated phages to the hosting organisms.

Metabolomic analysis of fasting plasma: In **SCAPIS** cohorts, a dense metabolic phenotyping is performed at Metabolon, US. The analysis includes four separate runs with different settings to cover essentially all metabolic pathways. Metabolic features are compared with an in-house standard library of >4,000 compounds to provide the chemical entity at highest level of confidence, and measurements of ~1000 named metabolites and ~200 non-named metabolites are expected. In **MOS**, profiling of plasma metabolites has been performed using a UPLC-QTOF-MS System (Agilent Technologies 1290 LC, 6550 MS, Santa Clara, CA, USA) and we have published details of the method in [10]. Non-targeted metabolite feature extraction was performed using Agilent Mass Hunter Profinder B.06.00 in pooled plasma samples and metabolite features that were present in at least 80% of the pooled samples were extracted from analytical samples and quality control samples.

Association between gut microbiota, metabolites and cardiometabolic traits, prediabetes and incidence of T2D and CVD: For the cross-sectional and prospective analyses of binary variables we will use logistic regression to calculate the odds ratios between obesity, prediabetes and the incidence of T2D, CAD and stroke (from register-based follow-ups) and microbiota characteristics, adjusting for potential confounding factors (e.g. sex, age, obesity, diet, smoking habits, physical activity, drug use, education and ethnicity). Linear regression will be used for continuous cardiometabolic risk variables (such as BMI, fasting glucose, blood pressure, blood lipids, carotic plaque, pulse wave velocity etc). Several factors such as medication (in particular protein pump inhibitors, metformin, statins), diet and smoking have been shown to affect gut microbiota composition, which will be investigated further in our large cohorts utilizing both self-reported data and data from Swedish drug registers. Due to the skewed distribution of the microbiome we use spearman correlation and negative binominal regression when investigating association between the microbiome, metabolites or clinical parameters. We use orthogonal partial least square (OPLS) regression, which aims at identifying the relationship between X and Y to find the metabolites (X) that explain as much as possible of the variance of the clinical parameter of interest (Y). Given the high dimensionality of the microbiota and other omics data we will use different types of machine learning techniques to pick up microbiome features (taxa, pathways, metabolites) that are most strongly associated with investigated phenotypes and endpoints. **Statistical power and replication:** By summer of 2020, we will have > 12,000 samples sequenced (MOS N=2200; SCAPIS-Mö N= 5007, SCAPIS-Up N=5036 (PI for Uppsala Prof Tove Fall), which sample size is far much larger compared to any so far published population cohorts. After the collection of MOS is completed in 2021, the 2nd half of MOS samples will be sequenced, increasing the total sample size to ~15,000. We use one cohort as the discovery, and the others as replication cohorts, adjusting for multiple testing.

Identification of potential probiotic strains with beneficial effects on cardiometabolic health and interventional strategies: This is a collaborative project with Probi (Ulrika Axling, PhD, Kasper Krogh Andersen, PhD and Titti Niskanen, PhD). Two approaches will be employed to develop novel probiotic strains with the potential to be used for cardiometabolic disease prevention (**Figure 4**).

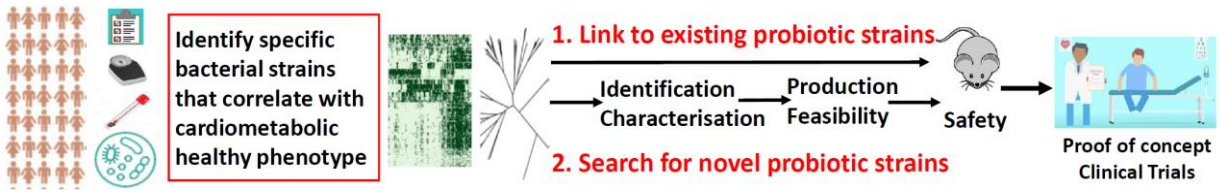


Figure 4. Two approaches to identify and develop novel probiotic strains

In the first approach the presence of bacterial strains with similarities to already characterized probiotic strains will be identified from the fecal samples from MOS and SCAPIS. We will explore whether these strains are differentially distributed between healthy subjects compared to individuals with cardiometabolic risk phenotypes. For strains with significant human evidence, animal studies will be performed to confirm the beneficial effect of the specific probiotic strains, followed by a human proof-of-concept study. This part involves probiotic strains that are already well-characterized and safe and are thus easily used for both animal and human clinical studies. The second approach aims at identifying new potential probiotic strains (*"Next Generation Probiotics"*) based on associations between bacterial strains and cardiometabolic health profiles in MOS and SCAPIS-Mö. For strain identification in human samples, see **Figure 3B**. Probi's complete strain library will be made available for mining of the proposed strains, and the therapeutic potential of such strains will initially be characterized in vitro followed by animal studies to further explore the therapeutic potential (**Figure 4**). Candidate strains will then undergo complete safety assessment and in-depth characterization prior to the initiation of the clinical phase with the overall aim of a randomized placebo-controlled proof-of-concept study to confirm the effects on cardiometabolic risk factors These novel probiotic strains ultimately have the potential to prevent or delay cardiometabolic disorders in subjects at risk.

Participating researchers: In addition to Isabel Goncalves, co-applicant with important role in cardiovascular- and atherosclerosis phenotypes, the most important collaborators are 1) The microbiome sequencing and bioinformatics experts at the Clinical Microbiomics (Copenhagen, Dr Henrk Björn Nielsen and Dr Jacob Bak Holm, microbiology, bioinformatics, systems biology); 2) Collaboration with Professor Tove Fall, Uppsala (SCAPIS-Uppsala, with focus in studies of mechanisms of atherosclerosis, role of gut microbiota and metabolomics, Mendelian Randomization), 3) Collaboration with PROBI (Ulrika Axling, PhD, Kasper Krogh Andersen, PhD and Titti Niskanen), 5) Prof Olle Melander (GWAS, plasma and stool metabolomics) and 9) Prof Peter Nilsson (PI of MOS). The bioinformatics unit of LUDC is crucial for the project.

Timeline: All three projects will be run parallel. All metagenomic and metabolomic data will be available from the project start.

References

1. Cosentino et al. ESC Scientific Document Group: 2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. *Eur Heart J*. 41:255-323, 2020
2. Brunkwall L, **Orho-Melander M**: The gut microbiome as a target for prevention and treatment of hyperglycaemia in type 2 diabetes: from current human evidence to future possibilities. *Diabetologia* 60:943-951, 2017
3. Tang et al: Intestinal Microbiota in cardiovascular Health and disease. *J Am Coll Cardiol*. 73:2089-2105, 2019
4. Schmidt et al. The Human Gut Microbiome: From Association to Modulation. *Cell* 172: 1198-2015, 2018
5. Cani PD: Microbiota and metabolites in metabolic diseases. *Nature Reviews Endocrinology*, 2019
6. Shkoporov et al: Bacteriophages of the human gut: The "known unknown" of the microbiome. *Cell Host Microbe* 25: 195-208, 2019
7. Shkoporov et al. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe*. 26:527-541, 2019
8. Bergström et al: The Swedish CArdioPulmonary BioImage Study: objectives. *J Int Med* 278: 645-659, 2015
9. Nielsen et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnol*. 32:822-8, 2014
10. Ottosson F*, Brunkwall L*, Ericson U, Nilsson PM, Almgren P, Fernandez C, Melander O and **Orho-Melander M**: Connection between BMI related plasma metabolite profile and gut microbiota. *J Clin Endocrin Metab*. 103: 1491-1501, 2018