**Glossary of essential terms and abbreviations**

**Terms:**

**Amplicon Sequence Variant** (**ASV**, aka **OTU** for Operational Taxonomic Unit). This technical term points to a cluster of closely related microbial species (strains) that are collectively identified (mapped) by 16S rDNA amplicon sequencing, which is one of the common methods of taxonomic profiling of microbial communities. The actual heterogeneity of any ASV (how many distinct species/strains are clustered within a single ASV) is *a priori* unknown. It depends on the resolution of a method (eg which region of 16S rDNA was sequenced).

**Binary metabolic phenotypes:** the foundation of our methodology lies in the ability to predict (reconstruct and infer) the presence or absence of functional metabolic pathways (eg uptake and biotransformation of numerous nutrients, such as specific carbohydrates, major sources of carbon and energy for microbial cells) via comparative genomics-based *in silico* **metabolic reconstruction**. As a result of a long-term expert curation by our team and numerous collaborators, we have reconstructed >100 pathways driving metabolism of major dietary nutrients (carbohydrates, amino acids, B-vitamins) across the entire collection of ~3,000 microbial genomes representing the Human Gut Microbiome (HGM)). For the purpose of large-scale computations and comparative analyses, we express this knowledge in a format of **binary metabolic phenotypes** (“1” or “0”) reflecting the presence or absence of respective functional pathways in each genome, and, thus, endowing the analyzed organisms with the ability or inability to benefit from the supplementation of specific nutrients in the ecological niche.

**Binary Phenotype Matrix (BPM):** a tableaggregating inferred binary metabolic phenotypes for a set of N nutrients (in our case, N~100) over a range of microbial species with completely sequenced genomes representing a microbial community (in our case, ~3,000 microbial genomes representing human gut microbiome, HGM)

**Human Gut Microbiome** (**HGM):**. Typically refers to the entirety of microbial species colonizing human guts. Needless to emphasize that HGM taxonomic composition is a subject of substantial variations between people and even within the same person over time.

**Metabolic reconstruction:** a predictive bioinformatics methodologybased on comparative genomics applying homology, genomic and functional context analyses to infer presence/absence of functional metabolic pathways in completely sequenced microbial genomes. To this end, we are using a subsystems-based approach implemented in the SEED genomic platform.

**Nutrient Impact (NI):** a predicted propensity of a given ASV to benefit (increase fitness) from dietary supplementation of a specific nutrient. NI value is derived from BPM as a weighted average of binary phenotypes of species comprising a given ASV. NI values can be non-integers on a probability-like scale from 0 to 1.

**Nutrient Impact Matrix (NIM):** a table aggregating computed NI values for N nutrients (~100) over the entire set of detected ASVs

**Nutrient Impact Score (NIS):** a cumulative (averaged) impact of several nutrients on the fitness (and, potentially, on relative abundance) of an ASV.

**Relative abundance (RA):** fractional representation (%) of an ASV in the sample computed from the number of clustered sequencing reads

**Taxonomic Profile:** same as RA table of a microbiome sample

**Abbreviations:**

**BPM:** Binary Phenotype Matrix

**HGM:** Human Gut Microbiome

**NI:** Nutrient Impact (0 to 1)

**NIM:** Nutrient Impact Matrix

**NIS:** Nutrient Impact Score(averaged; range 0 to 1)

**RA:** relative abundance (%)

**RSC:** Reference Sample Collection

**TMS**: Test Microbiome Sample

**Nutriomics challenge**

1. **Medical objective**:

Rebalance human gut microbiome (**HGM**) to enhance the diversity of its bacterial composition (referred to as *taxonomic diversity*) via supplementary nutrients rationally selected to promote the preferential growth (representation) of species “underrepresented” while providing minimal additional support to “overrepresented” species in the patient’s sample.

1. **Bioinformatics objective**:

***Given***:

1. **Test Microbiome Sample(s**) (**TMS**) with a taxonomic profile provided as **R**elative **A**bundance (**RA**) table for all **A**mplicon **S**equence **V**ariants (**ASV**s) confidently mapped and quantified in a TMS. *Data format is described in Section IV, files are provided.* [Typically ~ 100-300 distinct ASVs at RA≥0.1% per TMS]..
2. **Reference Sample Collection** (**RSC**) of such profiles for a large number (S) of reference human gut microbiome (**HGM)** samples representing a generally healthy population. *Data format is described in Section IV, files are provided.* [In our case, S~1,000 preselected samples from American Gut Project or another published dataset], and
3. **Nutrient Impact Matrix** (**NIM**), which captures a predicted propensity of each ASV to benefit (increase fitness and, ultimately, RA) from dietary supplementation of any specific nutrient from the list of “N” nutrients. All numeric values of NIM (each on a probability-like scale from 0 to 1) for all ASVs (from both, TMS and RSC) are provided in a format *described in Section IV (files are provided)*. *See next section on how these values are derived from genomics-based in silico metabolic reconstruction* [In our case, N~100 dietary metabolites].

***Select:***

An optimal sub-set of *n* nutrients (n ≤10, out of N~ 100) that collectively would:

* have the largest beneficial impact promoting the growth of the **maximal number** **of** **ASVs** deemed **“underrepresented”** (*or poorly represented*) based on the RA distribution of this ASV across multiple samples in RSC,
* while having the smallest detrimental impact by promoting the **minimal number of “in-range”** and **“overrepresented”** (*or significantly represented*) **ASVs** as compared to RSC

*(note that a rational selection of metrics for this optimization task (including the assessment of ASV representation vs RSC) is a part of the assignment and a subject of discussion at an early stage.)*

1. **Premise, definitions, and assumptions:**
2. **Biology:**

* We assume that bacterial species comprising any HGM sample (whether TMS or from RSC) are in a homeostasis (steady state) defined by several individual factors including a current diet of a person (simplification).
* Therefore, our nutritional intervention would be in addition (not instead) of the current diet. Selected nutrients will be provided in excessive quantities as compared to the normal diet.
* We also assume that any added nutrient, in principle, may promote only those “underrepresented” species that are capable to utilize it (having a respective functional pathway and predicted phenotype, see below.
* That said, there is no way to be certain that a particular nutrient would indeed promote such species (*for many reasons that could be discussed*). Therefore, to increase the likelihood of a positive impact, it is beneficial to use as many promoting nutrients as possible (with a caveat about collateral promotion of “overrepresented” species, which would be detrimental and should be minimized!).

1. **Bioinformatics (BPM and NIM):**

* *Individual values in NIM* are precomputed from the **Binary Phenotype Matrix** (**BPM**), which was built by our team of curators using genomics-based in silico *metabolic reconstruction* (as introduced during the lecture) for:
  + ~3,000 representative HGM bacterial species with completely sequenced genomes, and
  + metabolic pathways for ~100 major dietary nutrients such as carbohydrates (most of them can be utilized by bacteria but not the host), amino acids and vitamins (utilized by both, bacterial and the host). *A complete list is provided FYI along with other data tables.*
  + Note that in contrast to NIM, individual values for all nutrients in BPM are integers reflecting the presence (“1”) or absence (“0”) of a phenotype (pathway), which defines a hard-wired (genome-encoded) ability or inability of a bacteria to utilize (benefit from) respective nutrients.
  + In our BPM, each of the ~3,000 curated HGM species is assigned a string of ~100 binary phenotype values (*“*1” or “0”) representing its metabolic capabilities toward a corresponding number of potentially beneficial nutrients (those corresponding to *phenotype value=*1). These strings are metabolic fingerprints that are essentially unique for each organism and convey a crucial information about its metabolic capabilities and dietary requirements.
  + Note that while nutrient is a metabolite and phenotype is an encoded property of a bacteria, within the framework of this task they form a tightly coupled pair, and they are shown by the same notations. Thus, a phenotype: “utilization of glucose” and “vitamin B3 auxotrophy (or requirement for growth)” are denoted as Glc and B3, respectively.
  + A dietary supplementation by any one of these nutrients may, with some probability[[1]](#footnote-1), enhance competitive fitness and, thus, increase representation (RA) of all organisms endowed with the respective metabolic capability (carriers of a respective phenotype, =1).
  + A collateral is that RA of non-carriers of this phenotype (=0), with some probability, may decrease upon such supplementation due to competition (natural selection in favor of the fittest). However, this notion is not directly reflected in calculations of the nutrient impact score (NIS) as defined below.
* Two major (and interrelated) differences between **NIM** and **BPM** are:
  1. Taxonomic units in **NIM,** in general case, do not correspond to individual species with complete genomes (as in BPM) but to clusters of several related species (**ASV**s). This is due to imperfect mapping achieved by most metagenomics-based taxonomic profiling techniques (in our case, 16S rDNA amplicon sequencing);
  2. Individual **NI** values, in general case, are not strictly binary (1 or 0). They can be any number between 0 and 1, which reflects *taxonomic and phenotypic heterogeneity* within each ASV. These values are computed as weighted average of binary phenotypes of those BPM genomes that comprise (map equally well to) a given ASV. (See [PMID: 34113324] for math details).
* Thus, among multi-taxonomic descriptors of ASVs in the supporting data tables you may find those that are described as a single species (e.g. *Bacteroides uniformis*), but many correspond to a cluster of multiple “/”-separated species (e.g. *Bacteroides faecichinchillae/faecis/thetaiotaomicron*)
  + Note that although for single-species-ASVs, NI values are typically binary (0 or 1), in some cases they may be non-integers reflecting strain-level phenotype heterogeneity.
  + On the other hand, while for multi-species-ASVs NI values are often non-integers; in some cases they may as well be =0 or =1, reflecting phenotype conservation between distinct species from the same genus.

1. **Bioinformatics (Nutrient Impact Score)**

* *An additional challenging aspect* of this optimization task (beyond the obvious combinatorial complexity) is our intrinsic inability to predict *quantitatively* the outcome of any nutrient supplementation on the **RA** of any individual taxa. Therefore, in this crucial aspect, our approach is intrinsically qualitative.
* One possible (suggested) way to handle this challenge is to introduce a notion of ***Nutrient Impact Score*** (**NIS**), a cumulative (averaged) impact of several selected nutrients on each ASV. Thus:
  + For any ***ASVi***, (where “*i*” is from 1 to “*A*”, a total number of identified ASVs in RA table of a given **TMS,** typically ~ 200-300**),** the ***NIS(ASVi)*** value for a single nutrient is simply equal to its impact ***NIin* (**where “*n*” is from 1 to “N”, in our case ~100 distinct nutrients/phenotypes in **NIM**) on this ***ASVi***
  + In case of simultaneous supplementation by *“n”* distinct nutrients (up to n=10 in this project), we use a simple additive model to compute averaged ***NIS(ASVi)*** as:
    - ***NIS(ASVi) =*** *average(****NIi1 + NIi2 + … NIin***
  + Quite obviously, a range for any ***NIS(ASVi)*** is also from **0** (no impact) to **1** (maximal impact) on representation (or **RA**) of ***ASVi***.
  + *Note: feel free to suggest/rationalize an alternative way to assess the combined impact of several nutrients using the provided data and assuming that every nutrient acts independently with the same (albeit unknown) magnitude of impact.*
* Using this simplistic approach, theoretically (albeit not practically ☹), we could compute NIS values for all ASVs in a given TMS (A~200-300) and all possible combinations of up to n(=10) from a list of N(~100) nutrients.
* As defined in the beginning, the optimal combination is expected to have the **highest** cumulative NIS (cNIS) for as many as possible “underrepresented” ASVs and the **lowest** cNIS for “overrepresented” ASVs (*see below for a discussion on species classification by* ***RA*** *as compared to* ***RSC***). *Note that we assume that all nutrients can potentially affect all ASVs (minor and major) in the same way (huge simplification), which is a sole function of the respective NI (close to “0” - little impact, close to “1” – strong impact)*
* Note that (beyond combinatorial challenge) in this optimization task, there is no preconceived notion of how to strike the balance between the number of beneficially affected ASVs (from “underrepresented” category) and the magnitude of cNIS. **It is up to you to suggest and rationalize specific optimization criteria**.
  + Indeed, some nutrient combinations may have a very strong impact (close to 1) on a narrow subset of ASVs, whereas other combinations would have a milder impact but on a broader set of underrepresented ASVs. *Which is better?*
  + Other weighting criteria (such as to what extent the RA of a given ASV deviates from “norm” (see below) may be also used as a part of the optimization strategy.

1. **Bioinformatics (ASV classification by RA)**

* The biggest problem (in general and for this project) and the best kept secret is that despite many ongoing efforts on balancing a person’s microbiome or shifting it from an “abnormal” (meaning pathological or dysbiotic) to “normal” (meaning healthy or well-balanced) state (homeostasis), there is no good (or even decent) operational definition of this very “normal” state. The lack of such operational definition makes our objective ill-defined and elusive.
* One reasonable way is to define a normal state of microbiome (at least for a certain population) is by analyzing numerous samples and, thus, defining species distributions(explicitly as “ranges” or implicitly by considering **RSC** as a training set for ML methods).
* In this **comparative distribution approach**, a taxonomic composition (RA table) of a **TMS** is compared to **RSC** (which is presumed to represent a “normal” core with some deviations/outliers). Within this approach, the objective of *Nutriomics* would be to shift the microbiome composition of a sample as close as possible to this normal core.
* As mentioned above, **feel free to choose, rationalize and implement an approach for classifying ASVs** by RA as “underrepresented”, “normal (in-range)” or “overrepresented. You may consider standard “percentile”-based cutoffs, weights (based on distance from median), etc,
* *Note: some ASVs (from the full list in RSC and NIM) will have RA below threshold (typically ≤0.02%). They should be considered “absent”, and, thus, not a subject of rebalancing via nutrient supplementation.*
* The main problem with this approach is in humongous variations within the “norm”, far exceeding those in human population genetics. This creates an obvious challenge that we have to deal with.

1. **Input data and specific tasks:**

* All needed data for the assignment have the following format *(Files will be provided)*:
  1. A set of “deviant” (potentially pathologic) TMSs (up to 100) described by a two-column tables ([ASV\_taxonomic ID], [RA%].
  2. **RSC** of S (~ 1,000) microbiome samples (described by the same type of RA tables as above) that are presumed to be “mostly normal” to be used as a definition of a normal core (with outliers).
  3. **NIM** providing NI values for N nutrients for all relevant ASVs. Columns: [ASV\_taxonomic ID], [**NI*n*** values], where “*n*” is from 1 to “N” (in our case ~100 distinct nutrients/phenotypes). (Plus a table explaining phenotype/nutrients abbreviations – just FYI)*.*
* **Specific tasks:**
  1. Select metrics for microbiome balancing/optimization.
     + You may use a version of a **comparative species distribution** approach outlined above. (*Note: for simplicity, we consider RAs of all ASVs as independent variables, ignoring possible dependencies and interactions.)*
     + Alternatively, if you prefer, you may use standard **metrics of diversity** (eg Shannon entropy) that are popular in the microbiome field and set the goal to maximize diversity of your TMS to make it closer to what is realistically observed in RSC. (*a subject of separate discussion if requested*)
  2. Based on a selected metric(s), suggest and implement the computational approach that would allow you to analyze the RSC as a foundation for further categorization of ASVs in TMSs.
  3. Based on the above, design, implement and test an algorithm for personalized (sample-specific) selection of up to n (=10) nutrients aimed to balance/optimize microbiome composition observed in “deviant” TMSs.

Please note that this is indeed an open-ended research task. It means that we do not have any preconceived notion of a right or wrong approach to define metrics of success that would enable optimization (selecting between various combinations of nutrients). In fact, **the main challenge of this task is to propose, rationalize and test such approach.**

***Good luck!***

1. In a real-life microbiome, it is defined by a variety of biological and ecological factors that can be discussed separately. In the current tasks, they are not explicitly considered, except that the impact of each nutrient is considered as a propensity, which is quantitatively reflected in NIM. [↑](#footnote-ref-1)