Metagenomics : beyond the horizon of current implementations and methods

Khachatryan, L.

Chapter 6

General discussion and possible future improvement

As one can appreciate from this thesis, metagenomics analysis can be a relevant and vital step for the improvement of many fields including human and animal health, ecology, agriculture and forensics. This research was dedicated to a better understanding of the current situation in the field of metagenomics, and extending its present application boundaries. At first, we described, classified and evaluated popular data types, sequencing platforms and algorithms aimed to collect the information provided by microbial communities. We also improved the set of metagenomics data analysis tools by developing and testing both reference-dependent and reference-free algorithms. Below, we will summarize the most important conclusions of this thesis as answers to four important questions in the field of metagenomics.

6.1 Who is inhabiting the microbiome?

So far, the only possibility to find the answer to this question is to perform so-called reference-dependent analysis of metagenomic data, comparing the reads obtained during the microbiome sequencing with a reference database. As described in Chapter 2, we created a series of benchmark bacterial mixes with a different known distribution of species. The obtained mixes were used to estimate the resolution capacity of two different metagenomic datatypes - routine 16S and costlier WGS - and to evaluate two different approaches for the taxonomic reads classification. We have shown that the use of WGS data provides a much more accurate outcome in comparison to 16S samples. This was true for expected taxa prediction, and estimations of the abundances of the observed species. This conclusion was solid across all mixes and analysis techniques. Furthermore, we demonstrated that the same microbiome, analysed using 16S sampling by different pipelines and even using different reference databases, can produce quite distinct results. Finally, it is important to note that the constructed bacterial mixes can be utilized to evaluate future algorithms for metagenomic taxonomic profiling.
The conclusions obtained during this research finalize and supplement a series of previous reports [90, 348, 185, 349, 350, 351, 186, 352] addressing the incompetence of 16S metagenomic data in accessing the true metagenome taxonomic composition, and should be considered when planning microbiome sequencing experiments. Since the cost of producing WGS metagenomic data remains rather high, it is worth considering investigating comprehensive yet cost-effective sampling techniques for taxonomic profiling. The search of new, distinct from 16S rRNA, marker genes could be one of the possible solutions.

6.2 How complex is the investigated microbiome?

Once microbiology switched from single-genome studies to the exploration of multi-organism DNA samples, the question about the complexity of the investigated sample became the most vital one. The classical routine approaches aim to answer it by mapping the metagenome sequencing reads or assembly contigs to an annotated sequence from a reference databases. The obvious weak spot of such method is the incompleteness of current databases, as well as the discrepancy between their content and the real distribution of microbial species on our planet. Another group of techniques to estimate the metagenome complexity use the sequencing of multiple samples of the same metagenome cultivated under different conditions, and analyse the reads or contigs co-occurrences. The main weakness of such methods is their technical and computational difficulty.
In Chapter 3 we proposed a reference-free method to estimate the complexity of a metagenome. Our approach was designed to classify reads within a single long read metagenomic dataset using only the sequencing information, particularly *k*-mers. This so far unique approach featured an unsupervised machine learning tSNE algorithm for non-linear dimensionality reduction, as well as a subsequent density-based clustering technique. We have shown that *k*-mer profiles can reveal relationships between reads within a single metagenome using a series of simulated long read metagenomic datasets as well as the real PacBio RSII bioreactor microbiome sequencing data.
The obtained results are highly important, as they prove the concept of substructures detection within a single metagenome operating only with the information purely found in the sequencing reads alone. The possibility of reference-free deconvolution of metagenomic data benefits the field of metagenomics greatly, as it contributes not only to the estimation of metagenome complexity, but also improves the metagenomic data assembly and enables the investigation of new bacterial species. The main limitations of the described approach - restricted number of reads that can be analysed - is caused by memory issues when calculating the dissimilarity matrix between *k*-mer profiles. We believe that in the future, this issue can be solved by calculating the distances between *k*-mer profiles "on the go", and storing only the most informative ones. The constant improvement in quality and accessibility of long-reads sequencing techniques provides a great perspective for this approach in the future.

6.3 How to compare different metagenomes?

As was mentioned in the introduction to this thesis, comparative metagenomics strictly speaking does not necessarily require reference-based metagenome profiling. However, most of the scientific research uses reference-based methods to address the difference between two distinct metagenomes. In Chapter 4 we demonstrated that the comparison of metagenomic data performed using a reference-free approach provides much better resolution and allows to fetch the patterns lost during the standard reference-dependent techniques. In this thesis we presented kPal - a *k*-mer based method, that was used to resolve the level of relatedness between microbiomes. We tested kPal on a series of simulated metagenomes with different copy number of closely related bacterial genomes. Our method was sensitive to temporal changes in microbiome composition. To check whether our reference-free approach could distinguish between different human metagenomes, we tested it on a set of gut and palm 16S metagenomes, collected from different people in a period of 6 months. kPal could distinguish the datasets not only by the metagenome origin (gut or skin), but also by person! This result was better than the one demonstrated by the homology-based approach, which failed to cluster metagenomes per person in case of skin samples. The obtained results are highly significant as they allow to look at the comparative metagenomics under a different angle.
While the existing tools are following the "first annotate, then compare" model, we proposed a contrasting "first compare, then annotate" algorithm, when the comparison of the annotation-free profiles (in our case *k*-mer profiles) is followed by the investigation of the *k*-mers that contribute the most to the observed dissimilarities. The further investigation of the most informative *k*-mers and reads from which these *k*-mers belong, could allow to fetch the DNA sequences that might possibly be lost during the routine reference-based techniques. This idea can be developed further as a base for many different projects, for example metagenomics-based disease diagnostics. Another possible application is the search for species specific to a particular environment, body habitat, diet, or a person. This opens a set of new possibilities for fields like forensics, where the resolution of reference-dependent techniques was not enough to use metagenomic data in routine experiments.

6.4 What is the possible pathogenic impact of the metagenome?

Many different strategies can be implemented to find the functional profile of a metagenome. Among them are using a mapping to existing reference databases, and predicting possible functional genes with supervised machine learning techniques. Recently separated branch of metagenomics - meta-transcriptomics - provides researchers with community-wide gene expression (RNA-seq) data, which can be further utilized for metagenome functionality annotation. However, standard approaches for functional profiling fail to annotate the metagenomic data on the "sub-gene" level, when the information about allele of the particular gene is desired. In the meantime, it is known that different alleles are often responsible for distinct types of virulence. Therefore, it is important to rapidly detect not only the gene of interest, but also the relevant allele. Consequently, an approach that allows a "super-zoom" to a gene sequence, as well as a database providing the user with sequences of different alleles of the same gene, were required. Current methods are limited to mapping reads to each of the known allele reference, which is a time-consuming procedure. The other strategy is the assembly of sequencing reads with the subsequent mapping of the obtained contigs to the known allele references. The last algorithm provides fast and accurate results, but cannot be extended to metagenomic samples, since the assembly dismantles the possible variations in case of two different alleles of the same gene in the sample.

We developed BacTag (see Chapter 5), a distributed bioinformatics pipeline for fast and accurate bacterial gene and allele typing using clinical WGS sequencing data. The major advantage of this approach is a preprocessing procedure in which signatures of candidate alleles are identified and stored in a database. The subsequent identification of alleles in clinical samples is done using these signatures instead of using a traditional exhaustive search. This tool can be successfully used for diagnostic purposes. Also, and because this particular approach can be applied to uncultured samples, we expect to implement this method for cases in which time is of the essence. BacTag currently is not designed to work with samples where more than one allele of the same gene is present. However, unlike in case with other similar tools, this issue can be fixed in the future by detailed evaluation of the coverage depth, as well as the additional analysis of heterozygous variants sites. The development of reference databases, containing the allelic sequences of virulent genes, is another direction that still can be improved. Some progress in this direction is done for antibiotic-resistance genes, however, the great number of possibly virulent genes and their alleles is still not included in such databases. In the era of rising antimicrobial resistance and the existence of so-called "super-bacteria", a fast and accurate bioinformatic analysis providing the possible pathogenic impact of a microbial sample can be crucial for human health.