Title

Subtitle

Author Name

More text

more text

Department Name

University Name

Date

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur porta enim a ipsum hendrerit aliquam. Phasellus at enim gravida, lobortis ante id, vestibulum velit. Aliquam iaculis maximus ligula quis ornare. Nunc ut orci id ante malesuada posuere id sit amet dolor. Nulla facilisi. In vitae est eu lacus finibus tincidunt vitae in neque. Aliquam tincidunt ipsum sed ligula rhoncus vestibulum gravida at erat. Quisque ligula elit, laoreet ut aliquet vel, finibus ac orci.

Introduction

Methods

A total of 31 HQ genomes (29 SGBs and 2 reference genome) were used in this study.

The species' pangenome was obtained using Roary (citation) based on Prokka annotation files (.gff).

Results and Discussion

Description of the set of bins

Information about the sequences used for this project were retrieved from the *metadata* and the *bin data* files that were provided with the data. *Metadata* provides an insight about the experiments from which our MAGs were obtained. This includes data about the hosts from whom the samples were taken: their health conditions, age, gender, and country, as well as information about the sequences: number and length of the reads, number of bases and other information concerning the experiments. The two reference genomes are curated by the NCBI and are not reported in this file, and for the other datasets not all the features are always provided. The *bin data* file reports completeness and redundancy of the MAGs, including the reference genomes.

The set of MAGs come from 9 different datasets, all of them working on stool samples. For what concerns the patients involved, three main study conditions can be identified: colorectal cancer (CRC), adenoma and control. The control group is the most numerous group, composed by 19 samples. The CRC and the adenoma groups present respectively 7 and 3 samples. Patients in the control group are cancer-free but not all of them are healthy: 7 of them suffer from fatty liver, hypertension, type two diabetes or a combination of the three. Moreover, all cancer patients are reported to be associated with one or more of these diseases (figure). All hosts are aged 35 or older, in particular 9 of them are considered senior (older than 65 years old). The samples are almost equally distributed between female and male patients (12 females and 13 males). The hosts come from 8 different countries, most of them from Europe with Austria being the most represented nation (18 samples), except for one patient from Argentina and one from Guinea. Only one sample was obtained from a non-westernized host and its country of origin is Argentina (figure). All the analyzed MAGs were obtained from sequencing experiments carried out with the IlluminaHiSeq technology and the minimum read length indicated for each dataset shows that quality control was performed and only high-quality reads longer than 30 nucleotides were retained.

All MAGs analyzed have high completeness that spans from 90.33% to 100%. In addition to the reference genomes, there are 2 other datasets presenting 100% completeness.

Genome Annotation

Pangenome Analysis

Pangenome analysis found 4670 total genes (figure **??**), among which 1017 were attributed to the *core* (above 90% prevalence in MAGs), 2074 to the *shell* (from 15% to 89% prevalence), and 1579 were classified as *cloud* (from 0% to 15% prevalence). The results were robust with respect to many rounds of computation.

The number of conserved genes appears to reach a plateau (figure **??**) when the number of MAGs increases, suggesting that this species has a closed pangenome. This is further confermed by the trend of unique genes plotted against the number of genomes (figure **??**).

(a) (b)

Figure 1

(a) (b)

(c)

Figure 2: Three simple graphs

Phylogenetic Structure and association with host data

Conclusion

(a) Number of new genes

(b) Number of conserved genes

(c) Number of genes in the pan-genome

(d) Number of unique genes

(e) Number of blastp hits with different percentage identity

Figure 3: Three simple graphs

Supplementary data