

# Genomics Environments Characterization by means of Novel Fingerprints Features

Guillermo G. Torres<sup>1</sup>, J. H. Martinez<sup>2,3</sup>

**1** Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

**2** Universidad del Rosario, Bogotá, Colombia **3** Technical University of Madrid, Madrid, Spain

\* E-mail: Corresponding [ggtorrese@unal.edu.co](mailto:ggtorrese@unal.edu.co)

## Abstract

Bla bla bla

## Introduction

The metatranscriptomics and metagenomics are new molecular approaches developed to explore the genetic potential of the ecosystems, bringing an unprecedented understanding of the relationships between microbial communities without the need of previews knowledge of them or their environment. This approaches allowed the first large-scale insight into the ecology of the environments since both taxonomical and functional diversity outlook.

Since an ecology perspective, a key step to characterize one ecosystem is studying its biodiversity, that implies to discover, to describe and to analyze the organization of all elements involved in, classifying both by evolutionary (phylogenetic) and ecological (functional) criteria [1]. In this sense, a metagenomic and metatranscriptomic analysis is aiming to decipher the microbial community structure by characterizing some microbes residing therein and quantifying their diversity regarding some level of the biological population like species, genera, families, order or even patterns of evolutionary diversification and the richness as frequently employed diversity metric. Describing the number of distinct microbial taxa within a given unit area inhabiting a particular ecosystem, measured by the relative abundance, that refers to the quantity of rarity and commonness among taxonomical or functional individuals in the sample or community [1,2].

Currently, high-throughput sequencing (HTS) technologies have offered the opportunity to obtain genomic and transcriptomic information at increasingly high throughput and low price. Consequently, studies aiming to investigate taxonomical and functional diversity use an approach referred to as shotgun sequencing, in which genomic or transcriptomic fragments originated from organisms constituting microbiome are extracted and massively sequenced [3]. Habitually, the HTS technologies, the most commonly used Illumina and Roche454 platforms, generates millions of sequences, referred to as "reads", that could be considered to represent the compositional properties of their source genomes and transcripts. Subsequent analysis of the reads involves the process referred to as "binning". The reads derived from this mixture of different organisms are assigned to phylogenetic groups according to their taxonomic origins, analogous to machine learning process, where the reads are clustered into distinct bins using reference sequences with known taxonomic origin like a supervised learning method.

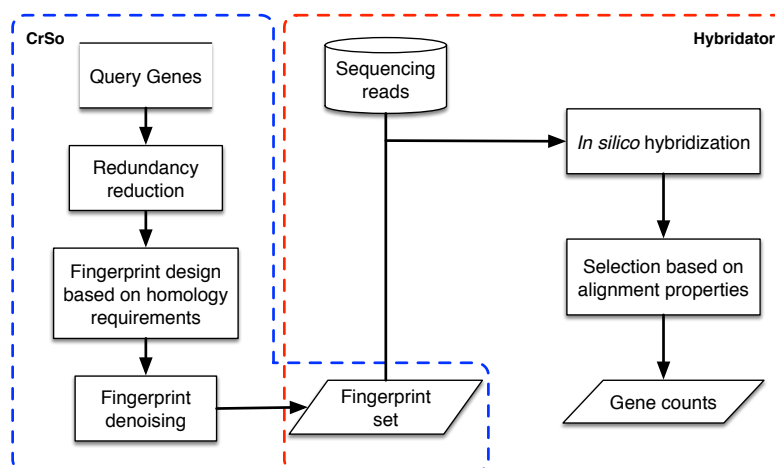
In the binning process typically we attempt to classify the reads through two strategies: composition-based or similarity-based. The first involves comparison of information related to GC content [4], codon usage [5] or K-mer frequency [6] from reads with those calculated from reference sequences. On the other hand, the similarity-based strategy uses string comparison algorithms between reads and reference sequences for data characterization. However, this last strategy can be sub-divided in two general methods: those to use Hidden Markov Models (HMM) [7] or BLAST-based homology searches [8,9].

Despite all efforts, the bioinformatic tools able to read binning are unable to make enough specific assignments for short fragments (< 400pb) [10,11]. Therefore, we present an enhancing of the HISS pipeline [12], an integrated approach that combines bioinformatic algorithms to fingerprint design and *in silico* hybridization using BLAST algorithm to characterize the ecosystems assessing critical genes involved in a functional context.

## Materials and Methods

Succeeding, we briefly describe the HISS pipeline and then present the algorithmic improvements implemented to fingerprint design and *in silico* hybridization. The HISS pipeline consist of 2 main stages (Figure 1). The first phase is the fingerprint designer, this takes interest genes sequences and designs the best non-redundant fingerprints from each of them. Later, the hybridizer performs an *in silico* hybridization between designed probes and the reads from metagenome or metatranscriptome, subsequently, the hybridization is evaluated.

**Figure 1. Overview of HISS pipeline stages.** CrSo: Fingerprint designer. Hybridator: *In silico* hybridization step.



### Fingerprints design (CrSo).

CrSo split up the fingerprint design in three general steps (Figure 1): 1) redundancy reduction 2) homology score calculation and best fingerprint compilation and 3) fingerprint denoising. At the first stage, CrSo performs a clusterization process using UCLUST [13] program, at the end of this step several sequences that covering the same biological sequence, and sequence fragments of one biological sequence that are globally alignable will be eliminated from initial input data set.

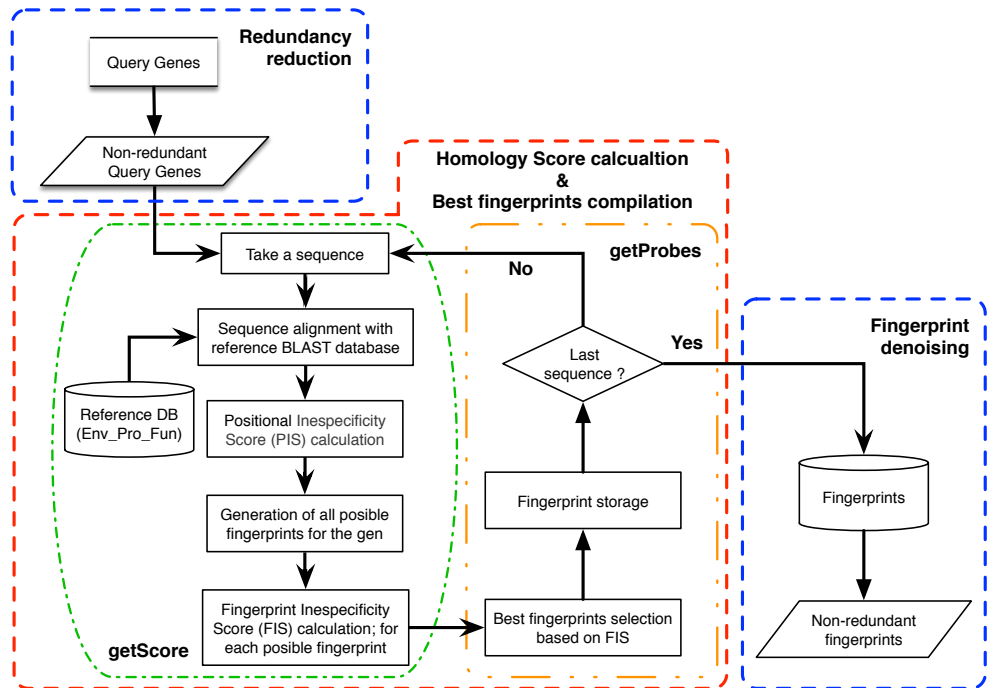
For fingerprint metrics calculation, the previous version of CrSo used the program OligoWiz 2.0 [14], however this software implement some algorithms to satisfy microarray experimental condition, such as melting temperature ( $T_m$ ) and GC content, not relevant by *in silico* hybridization, therefore here we will focus on specificity constraint of the fingerprints. A fingerprint of given gene  $g_t$  by definition will be any sub-sequence of gene  $g_t$  that is not sub-sequence of any  $g_i \in G, i \neq t$ , where  $G$  is a reference database of genes  $G = \{g_1, g_2, g_3, \dots, g_N\}$  consisting of  $N$  sequences.

In order to prevent hybridization to unintended targets (cross-hybridization), CrSo estimates this cross-hybridization evaluating the similarity of the target gene with other transcripts using BLAST. CrSo calculates a homology score for each possible oligonucleotide (FIS), based on BLAST search of the target gene against a reference database constituted by prokaryote, fungi and environmental sequences reported in EMBL-GeneBank-DDBJ [15] database. Each BLAST hit resulting is evaluated along the target sequence, where  $M$  is the number of BLAST hits regarded by  $j$  position of the target gene and  $H = \{h_{1j}, \dots, h_{Mj}\}$  be the BLAST hits identity in position  $j$ .

(Figure 2).

the homology score for the probes the fingerprint was consider a sub-string of the genes, where they were a strigs with

**Figure 2. Overview of CrSo pipeline.**



In the second stage, HISS uses the fingerprints designed to make an *in silico* hibridization with the reads from metagenomic or metatranscriptomic sample following a general rule: each read aligned significantly with a fingerprint should be considered originated or homologous of the enzyme gene that fingerprint represents.

$$D_{coll} = \frac{D_f + \frac{[S]^2}{K_D S_T} D_S}{1 + \frac{[S]^2}{K_D S_T}}, D_{sm} = \frac{D_f + \frac{[S]}{K_D} D_S}{1 + \frac{[S]}{K_D}}, \quad (1)$$

**Figure 3.** Figure Title first bold Figure A: Lorem. B: Consectetur.

### *In Silico* Hibridization

1. react
2. diffuse free particles
3. increment time by dt and go to 1

## Results

Nulla Table 1 volutpat.

**Table 1.** Table caption title.

Heading1				Heading2			
cell1row1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
cell1row2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
cell1row3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes.

### Probe Design.

Maecenas.

### Annotation.

Nulla

**Taxonomical annotation** Nulla.

**Functional annotation** Nulla.

## Discussion

Nulla Table 1 volutpat.

### LOREM and IPSUM Nunc.

CO<sub>2</sub> Maecenas. For more information, see S1 Text.

## Supporting Information

### S1 Video

**Bold the first sentence.** Maecenas.

### S1 Text

**Lorem Ipsum.** Maecenas.

## S1 Fig

Lorem Ipsum. Maecenas.

## S1 Table

Lorem Ipsum. Maecenas.

## Acknowledgments

Cras.

## References

1. Colwell, R. K. (2009). Biodiversity: concepts, patterns, and measurement. The Princeton Guide to Ecology.
2. Mouillot, D., Mason, W. H. N., Dumay, O., Wilson, J. B. (2004). Functional regularity: a neglected aspect of functional diversity. *Oecologia*, 142(3), 353–359. doi:10.1007/s00442-004-1744-7
3. Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978), 37–43. doi:10.1038/nature02340
4. Foerstner, K. U., Mering, von, C., Bork, P. (2006). Comparative Analysis of Environmental Sequences: Potential and Challenges. *Philosophical Transactions: Biological Sciences*, 361(1467), 519–523.
5. Noguchi, H., Park, J., Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.*
6. Sandberg, R. R., Winberg, G. G., Bränden, C. I. C., Kaske, A. A., Ernberg, I. I., Cöster, J. J. (2001). Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genes & Development*, 11(8), 1404–1409. doi:10.1101/gr.186401
7. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9), 755–763. doi:10.1093/bioinformatics/14.9.755
8. Huson, D. H., Auch, A. F., Qi, J., Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3), 377–386. doi:10.1101/gr.5969107
9. Haque, M., Ghosh, T. S., Komanduri, D., Mande, S. S. (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics (Oxford, England)*, 25(14), 1722–1730. doi:10.1093/bioinformatics/btp317
10. Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6(9), 673–676. doi:10.1038/nmeth.1358

11. Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS One*, 7(2), e31386. doi:10.1371/journal.pone.0031386
12. Torres-Estupiñan, G. G., Barreto-Hernández, E. (2014). In Silico Hybridization System for Mapping Functional Genes of Soil Microorganism Using Next Generation Sequencing. In *Advances in Computational Biology* (Vol. 232, pp. 337–344). Cham: Springer International Publishing. doi:10.1007/978-3-319-01568-2\_48
13. Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* (Oxford, England), 26(19), 2460–2461. doi:10.1093/bioinformatics/btq461
14. Wernersson, R., Nielsen, H. B. (2005). OligoWiz 2.0—integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Research*, 33(Web Server issue), W611–5. doi:10.1093/nar/gki399
15. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2009). GenBank. *Nucleic Acids Research*, 38(Database), D46–D51. doi:10.1093/nar/gkp1024