DIGEST, a workflow to extend the incomplete genes catalogue based upon capture sequencing technology applied to human gut microbiome

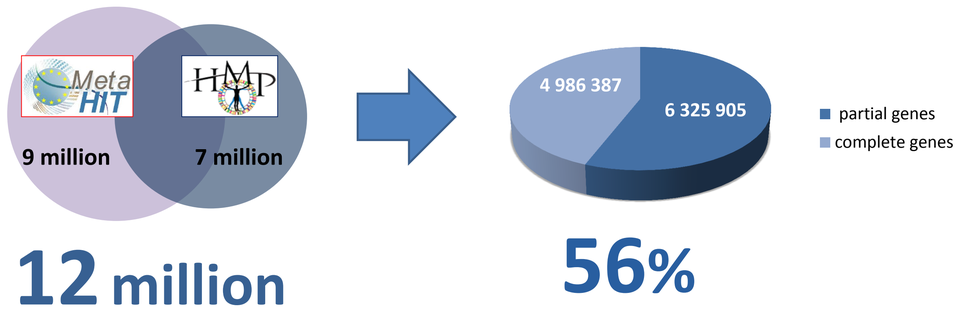
A. Felten[1], F. Le Fèvre[1], G. Gyapay[1], E. Pelletier[1], D. Muselet[1], A. Lemaçon[1], T. Richmond[2], J. Doré[1], J. Weissenbach[3], C. Médigue[1], D. Le Paslier[1], S. Cruveiller[1]

[1] CEA / Genoscope, Evry, France - [2] Roche NimbleGen, USA - [3] *INRA\UMR1319\MICALIS, France*

http://www.genoscope.cns.fr /projects/metacapture/

Background

Exploring the human gut microbiota diversity remains a challenge in understanding the relationships between human health and diseases. Two international projects, MetaHit [1] and HMP [2], produced gene catalogues containing about 8 and 5 M of genes respectively that were merged into a 11.8 million non-redundant genes of which 56% were incomplete. The challenge is to complete these genes to have a better comprehension of human microbiome.

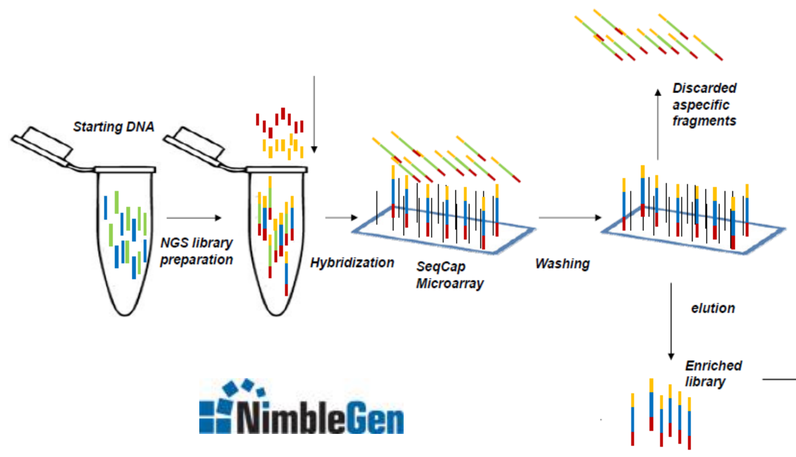


12 million
Non-redundant genes

56%
genes are partial
(START or/and STOP are missing)

A Sequencing strategy based on capture technology

A targeted re-sequencing strategy using the Roche-NimbleGen's sequence capture technology has been designed and applied on 50 individuals' fecal microbial DNA samples from the MetaHit cohort to generate read data specifically focus on gene of interest.



16.10⁶ selected probes

6To data

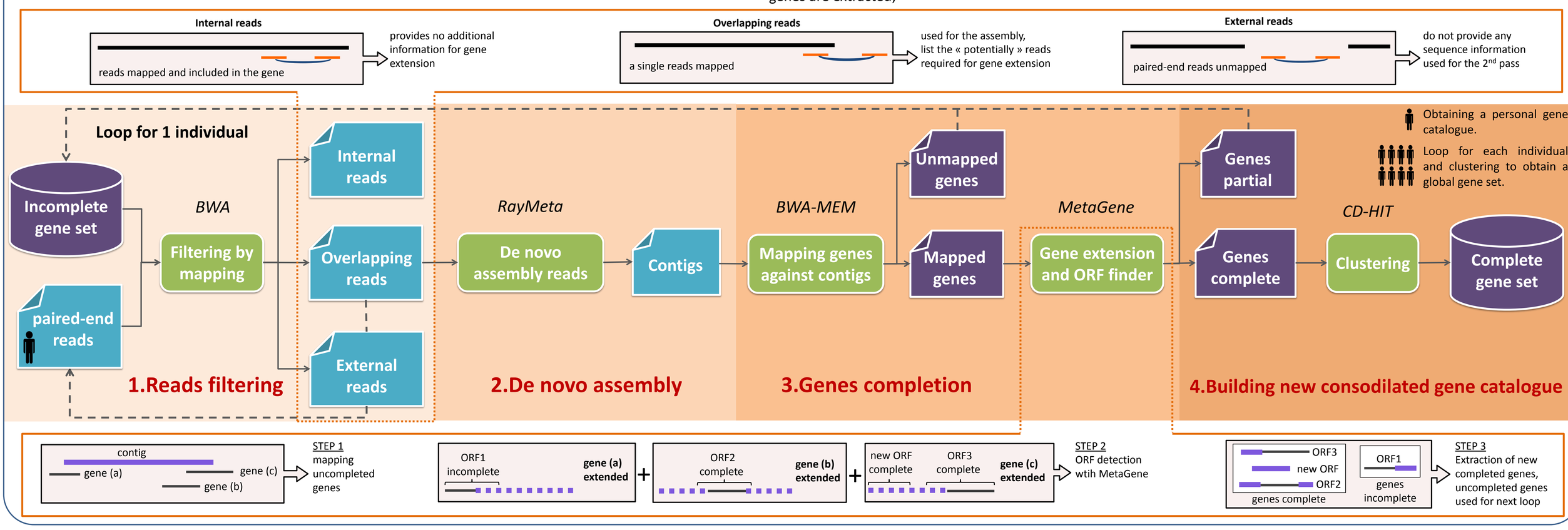
50 Individuals sequenced
4 x ¼ run per individual
paired-end (2x100bp)
Illumina Hiseq

2.10¹⁰ reads

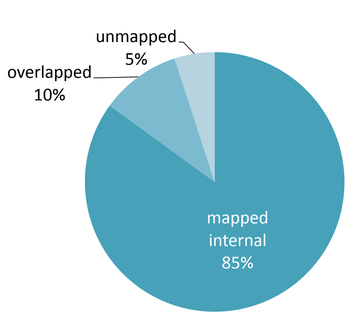
The Directed Iterative Gene Extension by Sequencing capture Technology (DIGEST) workflow

Opposite to a brute approach solution based on global assembly per individuals, a dedicated global iterative assembly of co-localized extremity paired end reads approach has been specifically designed and implemented in the DIGEST workflow composed of 4 main steps:

1. selection of informative reads for extension by read mapping on gene catalogue [3], "overlapping reads" (*i.e.* pairs of reads for which one end matches one extremity of a gene to be extended) and discard of "internal reads" and "external reads";
2. de novo assembly of local extremities using overlapping reads with RAYMeta [4];
3. merging the resulting contigs with initial genes [5]. The extended contigs, together with the set of yet unextended genes, are used as the input dataset for the next iteration. Process stops whenever the set of "external reads" is empty or remains unchanged. From the resulting contigs, genes are identified using an ORF finder software [6] and completed genes are extracted;
4. completed genes are clustered [7]; this leads to a new gene catalogue per individual. Finally all gene catalogues across individuals are merged into a non-redundant gene catalogue.

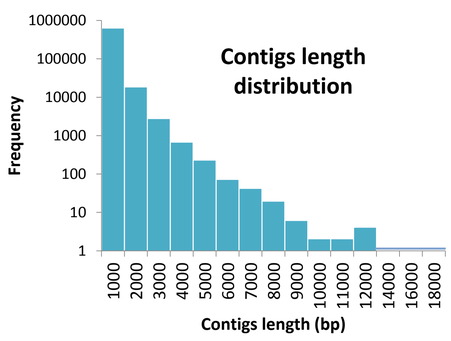


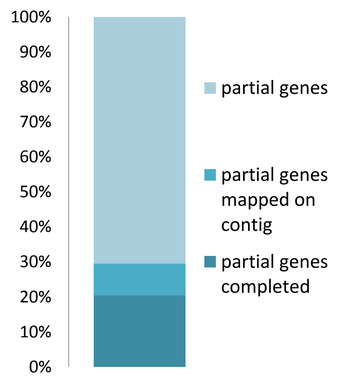
Applying DIGEST on the human gut microbiome gene set, results for 1 individual



The DIGEST's prototype workflow has been tested on the sequencing data of one individual. On the 200 million paired end reads :
- 85% are internal,
- 5% external,
- 10% overlapping.

From the 11.8 million initial genes, 3 millions have been found in this individual.





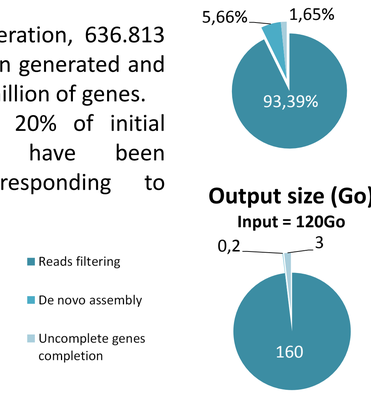
20% i.e.

1,3 million
genes partial have been completed for 1 individual

From more than 6 million of partial genes, 1.9 million (30%) have been mapped on the 636.813 contigs.

150k potential new genes

After a single iteration, 636.813 contigs have been generated and mapped by 1.9 million of genes.
After extension, 20% of initial partial genes have been completed corresponding to 1.209.092 genes.



Deployment and availability for the community

The procedure is being applied to the 49 other individuals to build a consolidated gene catalogue which could be used for accurate profiling and functional annotation in integrated platform such as MicroScope [http://www.genoscope.cns.fr/agc/microscope, 8].
By specifically targeting the sequencing on interesting part of any microbiome or environmental metagenomics dataset, the capture sequencing approach, coupled with the DIGEST workflow, can be of great value for large scale projects by easing the access to the rare fraction of a complex sample.

DIGEST will be both deployed at France Génomique and IFB clusters.
https://www.france-genomique.org/

330 core
400 To storage

HPC MIGRATION
- Raw data migration
- Software deployment
- Job scheduler adaptation

References

[1] Qin et al., **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature, 464:59-65*, 2010.
[2] Turnbaugh et al**., The Human Microbiome Project**. *Nature*, page 804-810, volume 449, 2007.
[3] Li and Durbin, **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics*, page 1754-1760, volume 25, 2009.
[4] Boisvert et al., **Ray Meta: scalable de novo metagenome assembly and profiling**. *Genome Biology*, page R122, volume 13, 2012.
[5] Heng Li, **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. 2013.
[6] Noguchi et al**., MetaGene: prokaryotic gene finding from environmental genome shotgun sequences**. *Nucleic Acids Research*, page 5623-5630, volume 34, No. 19, 2006.
[7] Limin Fu et al., **CD-HIT: accelerated for clustering the next generation sequencing data**. *Bioinformatics*,, 28 (23): 3150-3152, 2012.
[8] Vallenet et al., **Microscope – an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data**. *Nucleic Acids Research*, volume 41, 2012.

Funding

The Genoscope hight-throughput sequencing facility is a member of the "France Génomique" consortium (ANR-10-INBS-0009). The work was supported by the "France Génomique" project (ANR-10-INBS-0009).