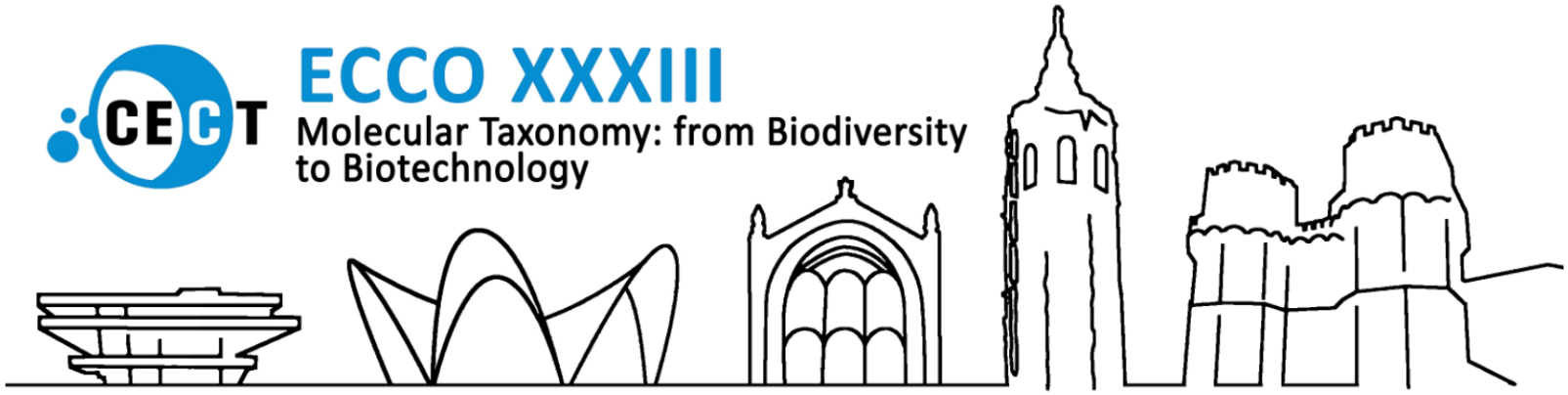


Microbial genomics and culture collections

Daniel Ramón Vidal
Biopolis SL

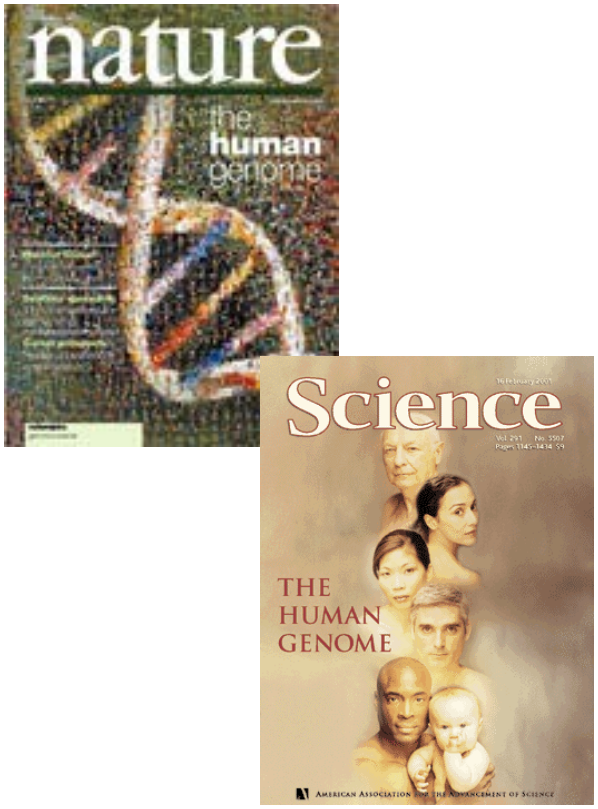
Microbial genomics and culture collections

1. Massive genome sequencing
2. Genome sequencing and its applications
3. The other “omic” strategies
4. Conclusions

Microbial genomics and culture collections

1. Massive genome sequencing
2. Genome sequencing and its applications
3. The other “omic” strategies
4. Conclusions

The first human genome draft

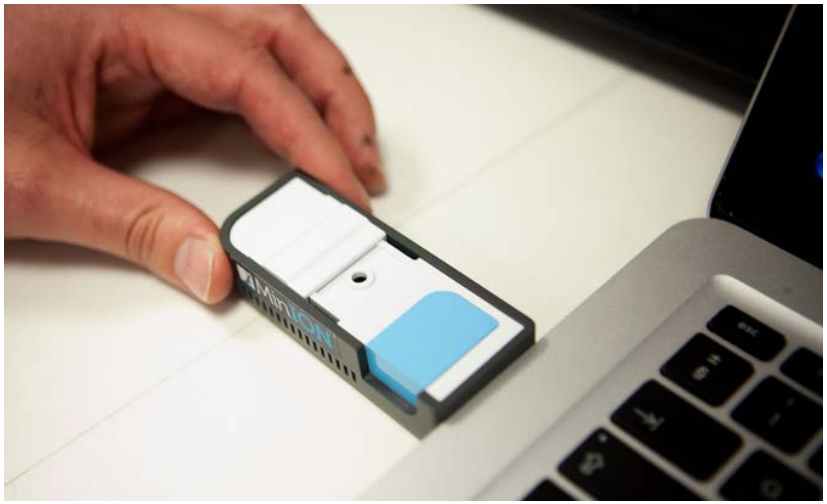


- In 2001, just a couple of days apart, two research groups published the complete sequence of the human genome
- Humans comprise about 23,000 genes interacting with the environment, but we know the functionality of only half of them
- The Human Genome Project cost of $3 billion (US) and over 3000 scientists worked on it for almost ten years
- Even so, it yielded a draft genome with over half a million errors
- These projects used Sanger sequencing tech, which yields very few sequences per run
- The future goal was to develop new more efficient platforms

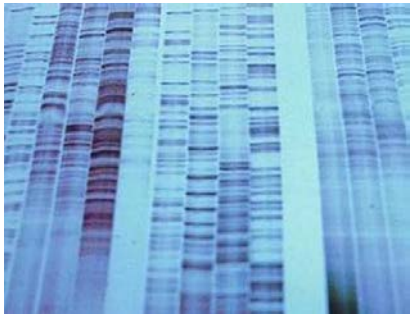
Massive genome sequencing (MGS)

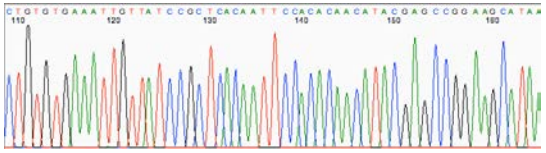
- Since 2001 new much faster and more efficient DNA sequencing methods have been developed
- These new technologies are called "massive genome sequencing" (MGS)
- There are now several different MGS platforms on the market (454 Roche pyrosequencing, Illumina, Solid 5, Ion Torrent, PacBio)
- Each is more suitable for certain tasks
- With these new platforms, genome sequencing has become increasingly cheaper and faster
- An example worth mentioning is that today it is possible to sequence a human genome for under 6000 € in just a few weeks, and the work can be done by a qualified technician

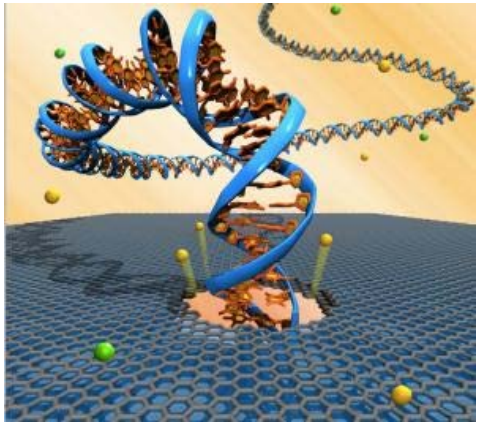
The future of MGS



9 minutes/200 $
2015?

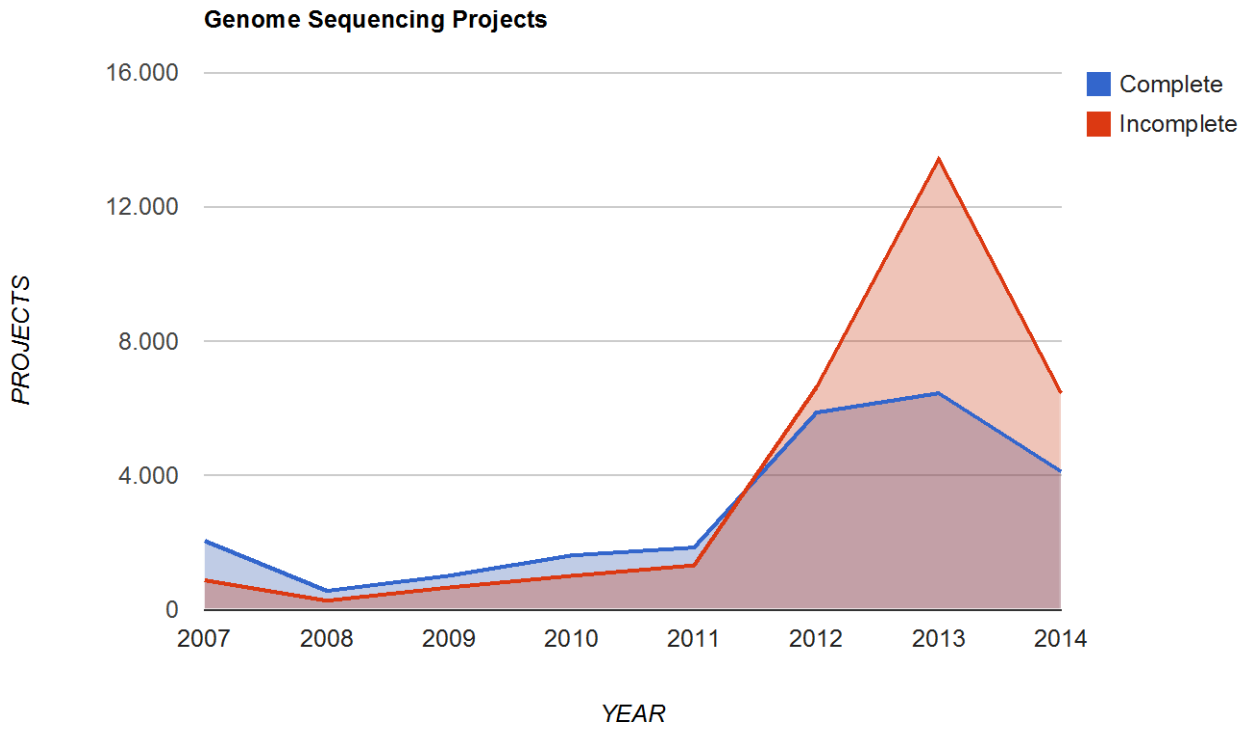






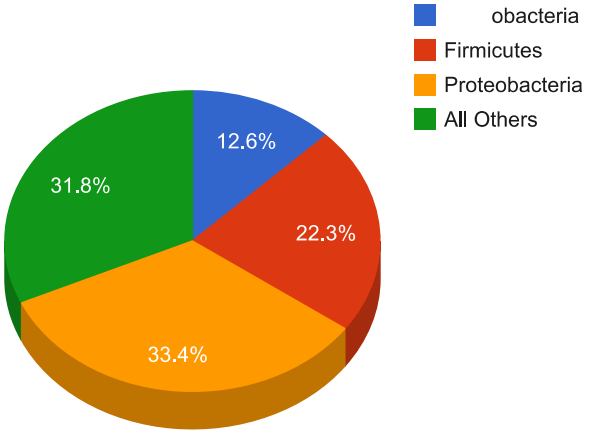
1970 1990 2015

What have we sequenced?



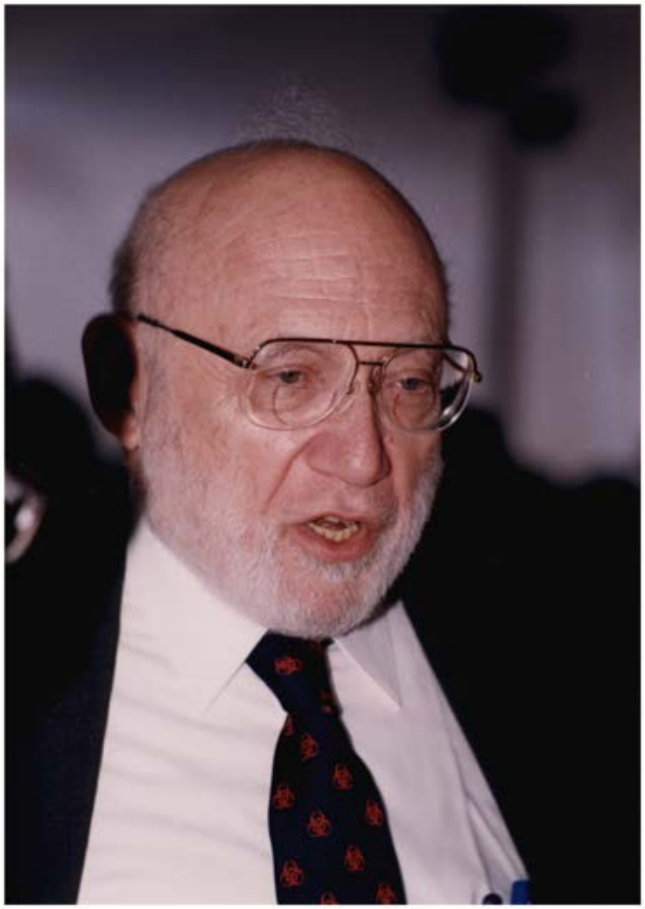
- There are hundreds of public and private laboratories sequencing the genomes of different animals, plants, microorganisms and virus
- These data are updated in the GOLD website (Genomes On Line Database; http://genomesonline.org/cgi-bin/GOLD/index.cgi)

Microbial genomes



- According to the GOLD website there are 53,660 genome sequencing projects with a total of 22,744 genomes that have been fully sequenced
- A high percentage are microbial genomes, mainly bacterial genomes
- These results pave the way to carrying out massive genomic screening and molecular identification of culture collections

Other MGS applications: microbiomes



- A microbiome is the whole community of microorganisms - including commensal, symbiotic and pathogenic bacteria - which occupy a particular ecological niche
- That niche can be a soil sample, an aqueous medium or biological remains…
- We can talk about the soil microbiome of Antarctica, or we can talk about the microbiome from the faeces of a child or a panda bear
- Until recently little was known about microbiomes and knowledge was biased because a very small percentage of the microorganisms in a sample can be grown in the lab
- MGS has helped to solve these problems

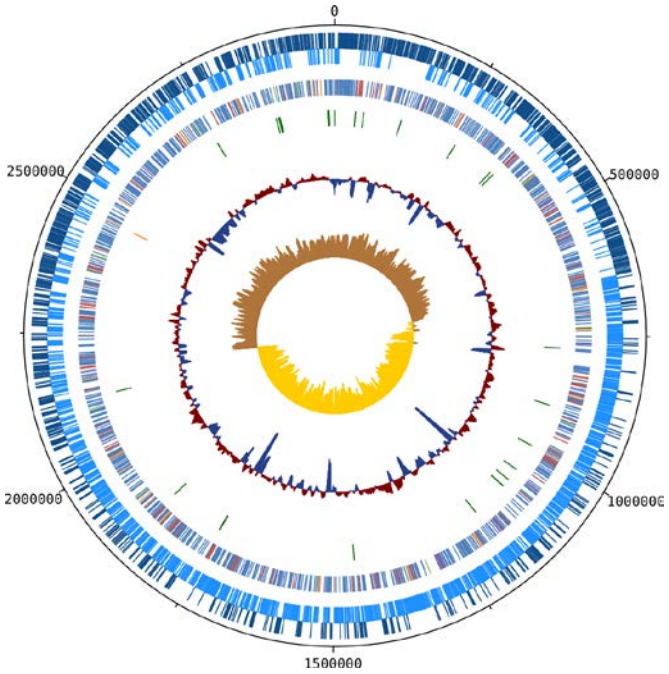
The other “omic” technologies

Microbial genomics and culture collections

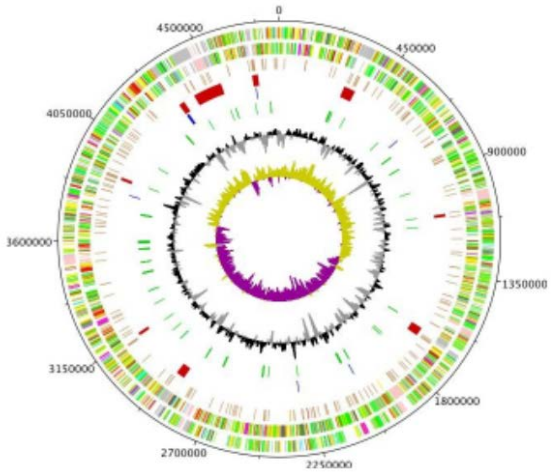
1. Massive genome sequencing
2. Genome sequencing and its applications
3. The other “omic” strategies
4. Conclusions

How to generate a bacterial genome?

Microorganism | DNA isolation | DNA amplification | MGS | Bioinformatics



Applications

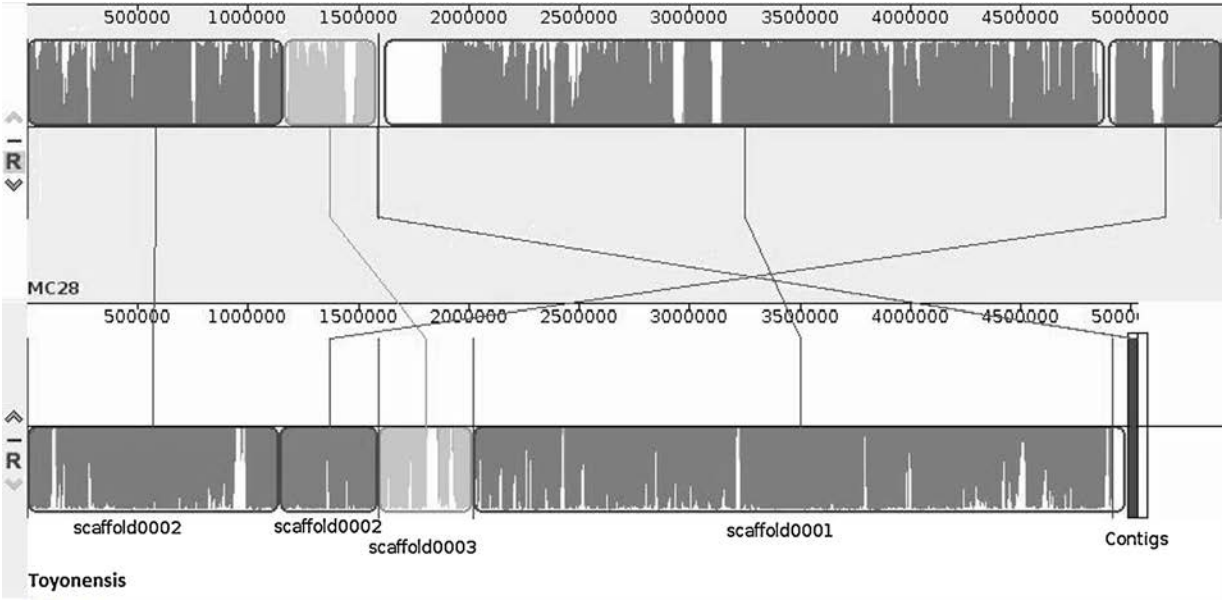


- Taxonomic identification
- Phenotype description
- Safety evaluation
- Massive genomic screening

Taxonomic identification (I)

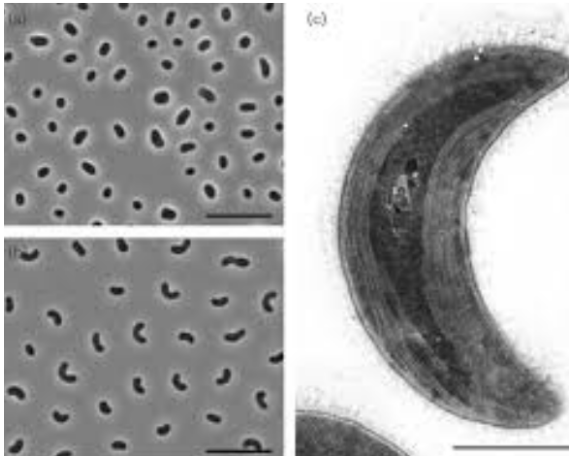
- Strain BCT-7112T was isolated in 1966 in Japan; it was primarily identified as *Bacillus cereus* var *toyoi* and has been used as a probiotic in animal nutrition for more than 30 years
- Genome sequencing showed significant genomic differences from the type strains of the *B. cereus* group that were large enough (ANI values below 92%) to allow it to be considered as a different species within the group
- It is now recognized as *Bacillus toyonensis* sp. nov with BCT-7112T (=CECT 876T; =NCIMB 14858T) being designated as the type strain
- Moreover, a pairwise comparison between the available genomes of the whole *B. cereus* group indicated that besides the eight classified species (including *B. toyonensis*), additional genomospecies could be detected

Taxonomic identification (II)



G. Jiménez, M. Urdiain, A. Cifuentes, A. López-López, A.R. Blanch, J. Tamames, P. Kämpfer, A.B. Kolstø, D. Ramón, J.F. Martínez, F.M. Codoñer, M. Castillo, R. Rosselló-Móra. (2013). Description of *Bacillus toyonensis* sp. nov, a novel species of the *Bacillus cereus* group, and pairwise genome comparisons of the species of the group by means of ANI calculations. Systematic and Applied Microbiology 36: 383-391.

Phenotype description (I)



- *Methylocystis parvus* OBBP is an obligate methylotroph considered the type species of the genus *Methylocystis;* its genome has been fully sequenced by our company
- Compared with some other *Methylocistis* strains, two *pmo*CAB particulate methane monooxygenase operons and one additional singleton *pmo*C paralog were identified in the OBBP strain sequence
- Other genes encoding proteins involved in methane oxidation were identified (i.e. a methanol dehydrogenase or a set of enzymes involved in pyrroloquinoline quinone synthesis, among others))
- Genes encoding enzymes for the serine cycle were identified, confirming that OBBP possesses a methane assimilation pathway linked to the central metabolite pathways

Phenotype description (II)

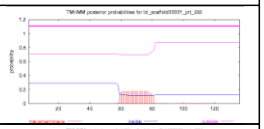
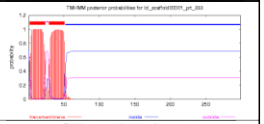
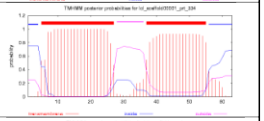
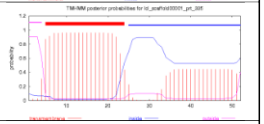
- Also genes involved in the ethylmalonyl-coenzyme A (ethylmalonyl-CoA) pathway for glyoxylate regeneration were detected in *M. parvus* OBBP
- Other genes encoding enzymes responsible for PHB metabolism were identified in the OBBP genome (two poly-β-hydroxybutyrate polymerases named PhbCI and PhbCII), two depolymerases (DepA and DepB), an one acetyl-CoA acetyltransferase (PhbA), one acetoacetyl-CoA reductase (PhbB), one polyhydroxyalkonate synthesis repressor (PhbR), and a phasin

C. del Cerro, J.M. García, A. Rojas, M. Tortajada, D. Ramón, B. Galán, M.A. Prieto, J.L. García. (2012). Genome sequence of the methanotrophic poly-b-hydroxybutyrate producer *Methylocystis parvus* OBBP. Journal of Bacteriology 194: 5709.

Phenotype description (III)

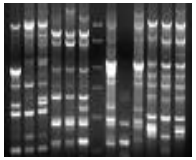
- ES1 is a *Bifidobacterium longum* probiotic for celiac patients
- A randomized, double-blind placebo-controlled trial has been conducted on 33 new diagnosed CD children aged 2-14 years, with this probiotic
- ES1 administration significantly decreased CD3+ T-lymphocytes and TNF-α concentration in peripheral blood and stool IgA
- Children ingesting ES1 underwent a statistically significant increase in height percentile and also an increase in weight (albeit non-statistically significant)
- Fecal microbiota analysis indicated a reduction in *Bacteroides fragilis* group members in the ES1 group and an increase in bifidobacteria and *Lactobacillus* counts

Phenotype description (IV)

ID	GO	NAME	TM (TMHMM)
lcl|scaffold00001_cds_330	GO:0043565	sequence-specific DNA binding	no
lcl|scaffold00001_cds_331	GO:0043565	sequence-specific DNA binding	no
lcl|scaffold00001_cds_332			
lcl|scaffold00001_cds_333	GO:0008233	peptidase activity	
	GO:0016020	membrane	
	GO:0006508	proteolysis	
lcl|scaffold00001_cds_334			
lcl|scaffold00001_cds_335			

- The ES1 genome was sequenced with high-throughput pyrosequencing technology
- This genome is around 3.5 Mb; a total of 2,078 elements were detected, where 2,020 were ORFs (1,624 canonical and 396 non-canonical) and 58 were RNAs (3 rRNA and 55 tRNA).
- Comparison with some other *B. longum* genomes revealed six differential genomic regions in ES1 affecting 52 genes
- Some of these genes encode enzymes that may explain the functionality of this strain

Food safety assessment (I)




Selection → Identification → Evaluation → Safety → Scale-up → Production → **Clinical trials**

- Bowl inflammation : *Bifidobacterium longum* ES1
- Rotavirus: *Bifidobacterium longum* subsp. *infantis* CECT 7210
- Immune booster: *Bifidobacterium breve* I-4035, *Lactobacillus paracasei* I-4034, *Lactobacillus rhamnosus* I-4036

- *Helicobacter pylori: Bifidobacterium bifidum* CECT 7366
- Metabolic syndrome: *Bifidobacterium animalis* subsp. *lactis* CECT 8145
- Vaginosis: *Lactobacillus rhamnosus*

Food safety assessment (II)

- Blast search of ES1 strain genome sequencing showed that none of the ORFs had an identity higher than 50% of that described for antibiotic resistance or virulence genes
- This indicates that only domains in the ORF are similar to some of the putative virulence and antibiotic resistance genes
- No significant or highly similar genes to virulence and antibiotic resistance were detected
- Only a DOC (death-on-curing) family protein with an identity of 87% and a significant e-value was detected when searching for virulence factors; this protein is also present in *B. longum* JDM301and ATCC 55813 strains

S. Muñoz-Quesada, E. Chenoll, J.M. Vieites, S. Genovés, J. Maldonado, M. Bermúdez-Brito, C. Gómez-Llorente, E. Matencio, M.J. Bernal, F. Romero, D. Ramón, A. Gil. (2013). Isolation, identification and characterization of three novel probiotic strains (*Lactobacillus paracasei* CNCM I-4034, *Bifidobacterium breve* CNMC I-4035 and *Lactobacillus rhamnosus* CNMC I-4036) from the faeces of exclusively breast-fed infants. British Journal of Nutrition 109: S51-S62.

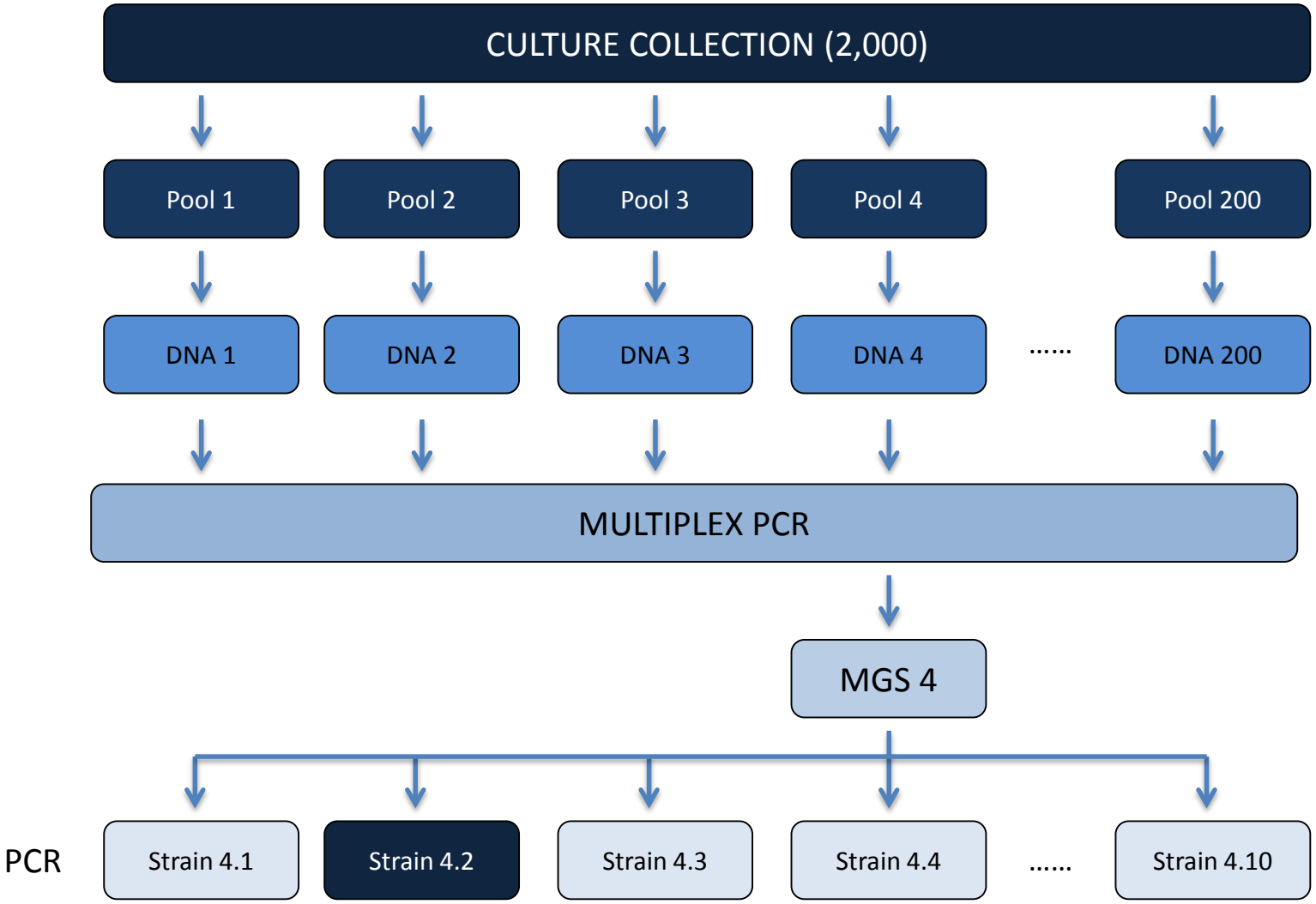
E. Chenoll, F.M. Codoñer, A. Silva, A. Ibañez, J.F. Martínez-Blanch, M. Bollati-Fogolin, Y. Sanz, D. Ramón, S. Genovés. (2013). Genomic sequence and pre-clinical safety assessment of *Bifidobacterium longum* CECT7347, a probiotic able to reduce the toxicity and inflammatory potential of gliadin-derived peptides. Journal of Probiotics and Health 1: 106 doi:10.4172/jph.1000106

E. Chenoll, F.M. Codoñer, A. Silva, J. F. Martinez-Blanch, P. Martorell, D. Ramón, S. Genovés. (2014). Draft genome sequence of *Bifidobacterium animalis* subsp. *lactis* strain CECT 8145, able to improve metabolic syndrome *in vivo*. Genome Anouncements 2: e00183-14

Massive genome screening (I)

- Project between CECT and Biopolis SL
- 2,000 different strains (bacteria, yeast and filamentous fungi)
- Five enzymatic activities of industrial relevance (agarase, cyclodextrin glucan transferase, dextran sucrase, lipoxygenase and PHA synthase)

Massive genome screening (II)



Microbial genomics and culture collections

1. Massive genome sequencing
2. Genome sequencing and its applications
3. The other “omic” strategies
4. Conclusions

Omic technologies plus model organisms

Streptococcus
(9 strains)

Lactobacillus
(75 strains)

Bifidobacterium
(15 strains)

Years/millons € → Months/thousands €

DANONE RESEARCH

Two steps screening methodology

1

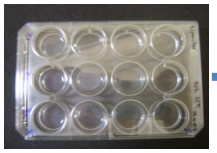
Synchronization | Feeding | Assay

C. elegans
BA17 fem-1

4 days

NG Plates
20° C

Eggs
(1mL Buffer M9)



1 day

25° C

L1 Larvae

M9 + Cholesterol
Inhibition *E. coli*
Probiotic addition

3 days

25° C

Adults

Oxidative stress

5 hours

25° C

Viability

2

C. elegans
wild –type (N2)

OP50

5 days

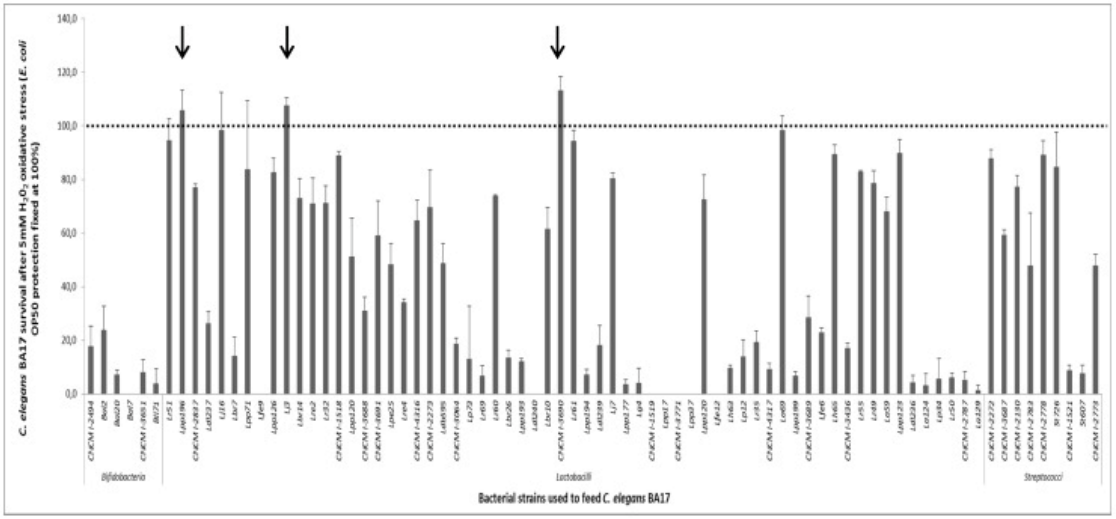
Probiotic

Oxidative stress

5 hours

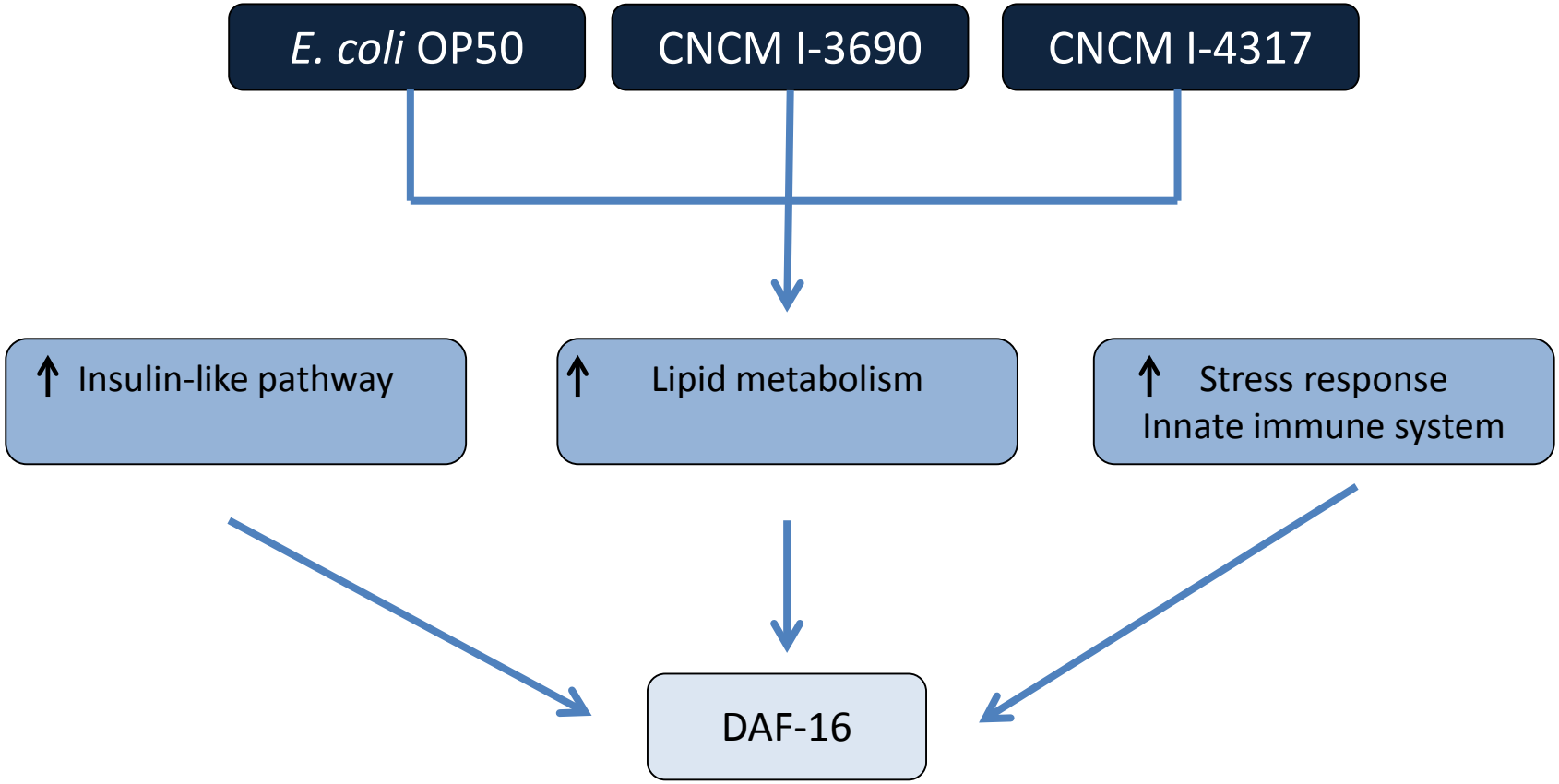
G. Grompone, M.C. Degivry, D. Ramón, P. Martorell, N. González, S. Genovés, S. Legrain-Raspaud, I. Chabaud, R, Bourdet-Siscard. (2011). Method for selecting bacteria with antioxidant activity. WO2011/083353A1

An antioxidant *Lb. rhamnosus* strain



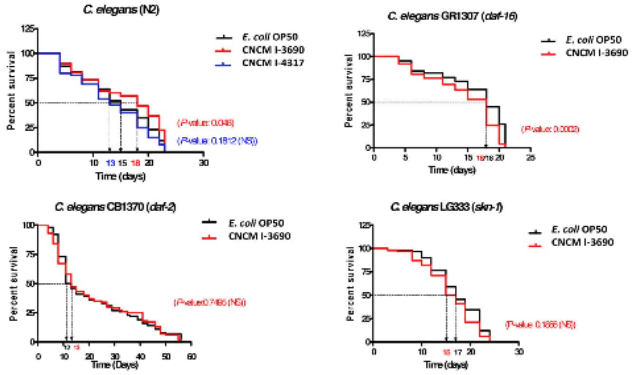
- In general, lactic acid bacteria conferred greater resistance to oxidative stress
- No effects were observed for Bifidobacteria
- An important effect was observed for eleven strains belonging to *Lactobacillus* and *Streptococcus* genera; one of these, *Lactobacillus rhamnosus* CNCM I-3690, proved highly effective

Transcriptomics as a strategic tool



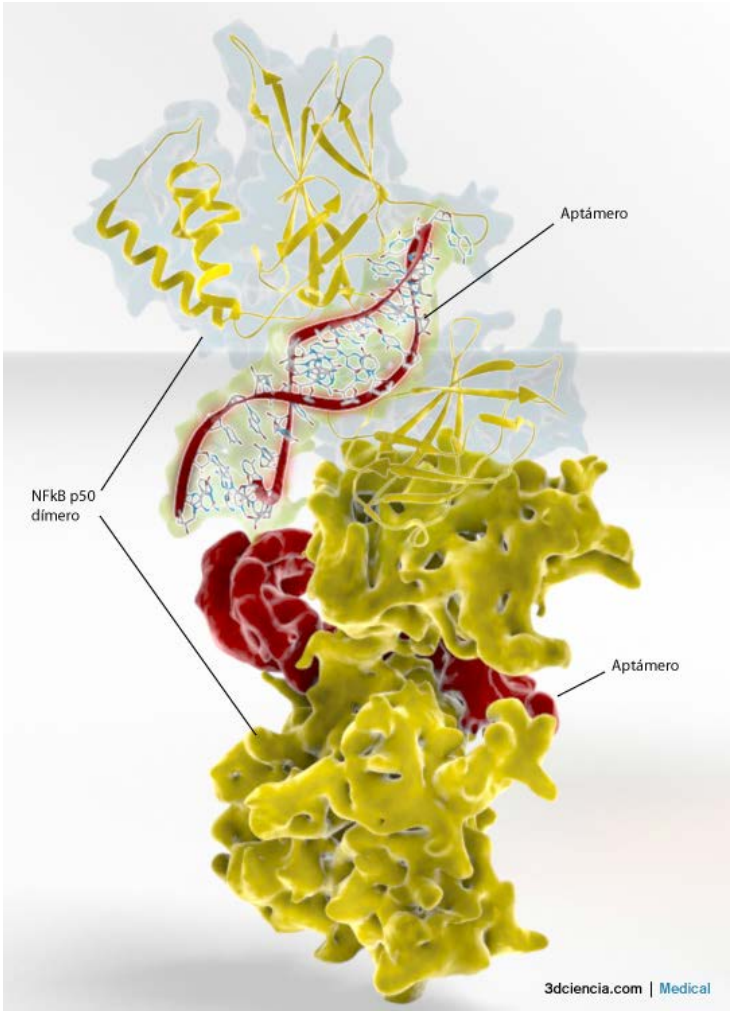
G. Grompone, D. Ramón, P. Martorell, S. Genovés, P. Ortíz, S. Llopis, N. González. (2012). *Lactobacillus rhamnosus* strain for reducing body fat accumulation. PCT/IB2012/056344

The use of mutants



- To elucidate the role of DAF-16, we studied the antioxidant activity and effects on lifespan of *Lb. rhamnosus* CNCM I-3690 *in* C. elegans daf-2 (CB1370), daf-16 (GR1307) and skn-1 (LG333) mutant backgrounds
- Data indicated that the increase in survival after 15 days observed in wild-type strain N2 was absent in GR1307, CB1370 and LG333 mutant worms
- This would indicate that the increased lifespan observed after CNCM I-3690 consumption in wild-type N2 is dependent (at least partially) on the DAF-2/DAF- 16 signaling pathway
- These phenotypes are associated with a reduction of inflammatory processes

Comparative genomics

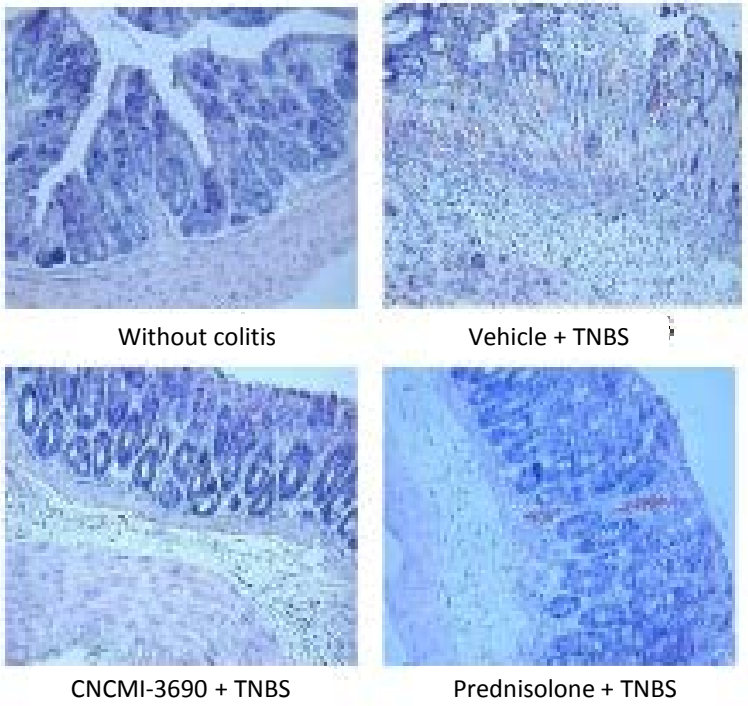


- There is evidence in mammalian systems for a direct link between signaling via longevity factors, such as FoxOs or SIRT-1, and inhibition of NF-kB signaling
- Thus, we hypoth-esized that a strain providing antioxidant protection in the *C. elegans* model through the DAF-16 transcriptional factor may also exhibit an anti-inflammatory profile in mammals
- Therefore we decided to focus further experimentation on the analysis of anti-inflammatory properties of both strains CNCM I-3690 and the control CNCM I-4317 in *in vitro* and *in vivo* models

Studies on human cells

- At Danone Research, they studied the effect of a transient co-culture of CNCM I-3690 and CNCM I- 4317 strains with HT-29 intestinal epithelial cell-line after proinflammatory stimulation with TNF-α, IL-1b and IFN-γ
- They observed a significant reduction in NF-kβ signaling
- They also studied the effect of co-culturing the two bacterial strains in a transwell-filter co-culture system with HT-29-NF-kβ-luciferase epithelial cells in the apical side and human dendritic cells
- In this case, a clear anti-inflammatory profile was induced by strain CNCM I-369; moreover, only CNCM I-3690 strain was able to reduce the pro-inflammatory cytokine ratios IL-12/IL-10, IL-6/IL-10, IL-8/IL-10 and TNFα/IL-10

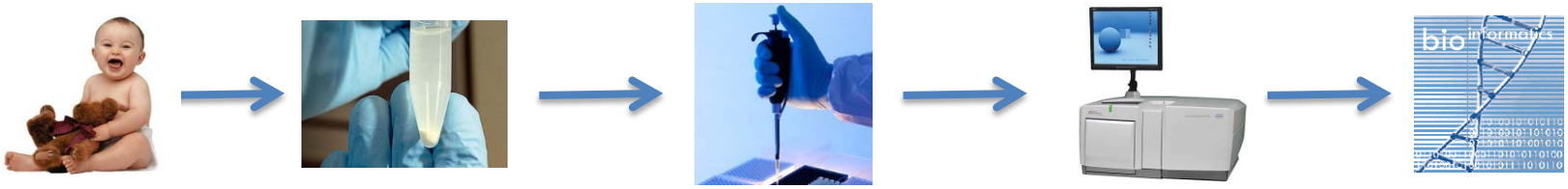
I-3690 is an anti-inflammatory probiotic



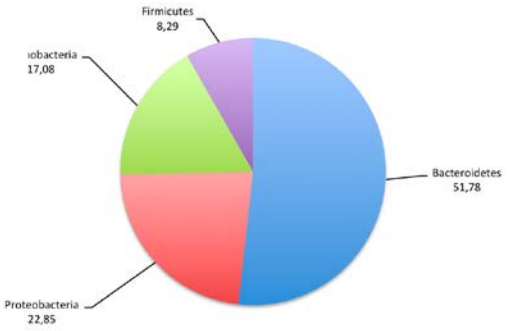
- Strain CNCM I-3690 (10^8 cfu) was administered intragastrically for 5 days in a TNBS-induced rectocolitis model of BALB/c mice
- Prednisolone was used as positive control and probiotic buffer as negative control
- Histopathological analysis was performed
- The Wallace coefficient was: 3.4 in the placebo control group; 1.5 in the prednisolone group; and 2.4 in the group treated with strain CNCM I-3690

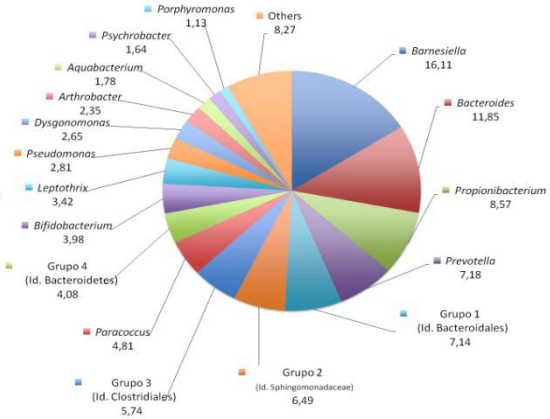
G. Grompone, P. Martorell, S. Llopis, N. González, S. Genovés, A.P. Mulet, T. Fernández-Calero, I. Tiscornia, M. Bollati-Fogolin, I. Chambaud, B. Foligné, A. Montserrat, D. Ramón. (2012). Anti-inflammatory *Lactobacillus rhamnosus* CNCM I-3690 strain protects against oxidative stress and increases lifespan in *Caenorhabditis elegans*. PLOS ONE 7: e52493

And now microbiomes



Sample | DNA isolation | DNA amplification | MGS | Bioinformatics

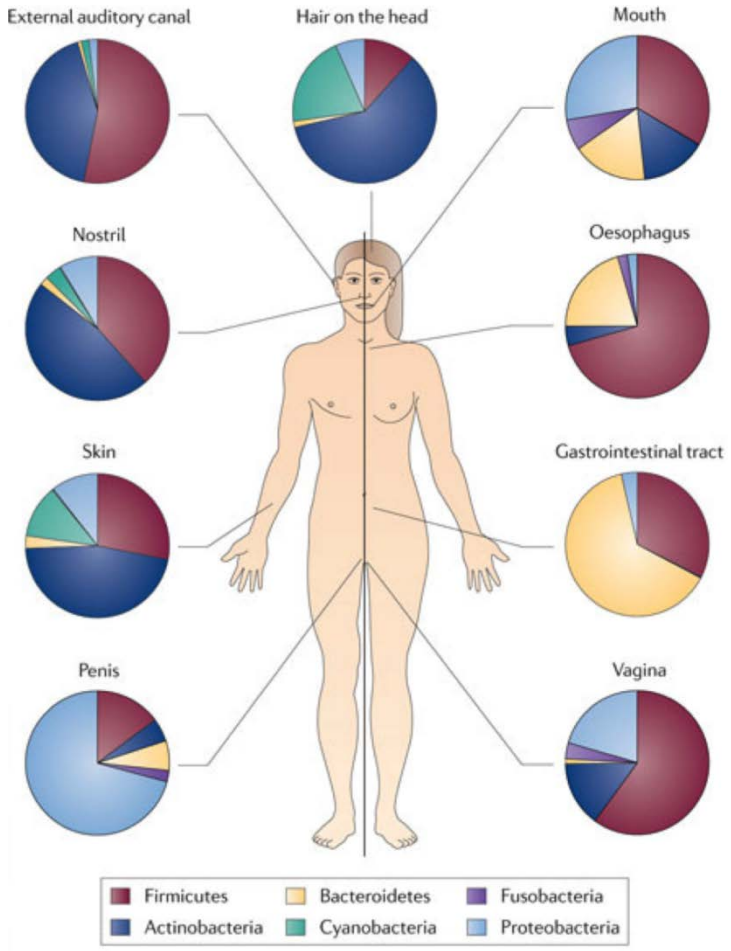




Bacterial species	Percentage
Barnesiella intestinihominis	15,12
Bacteroides acidifaciens	10,11
Propionibacterium acnes	8,47
Caulobacter leidyia	6,49
Paracoccus marcusii	4,81
Bifidobacterium breve	3,03
Prevotella buccae	2,97
Dysgonomonas mossii	2,57
Prevotella loescheii	1,98
Flavonifractor plautii	1,94
Pseudomonas migulae	1,94
Arthrobacter bergerei	1,35
Prevotella oris	1,03

Phylum | Genera | Species

The human microbiome

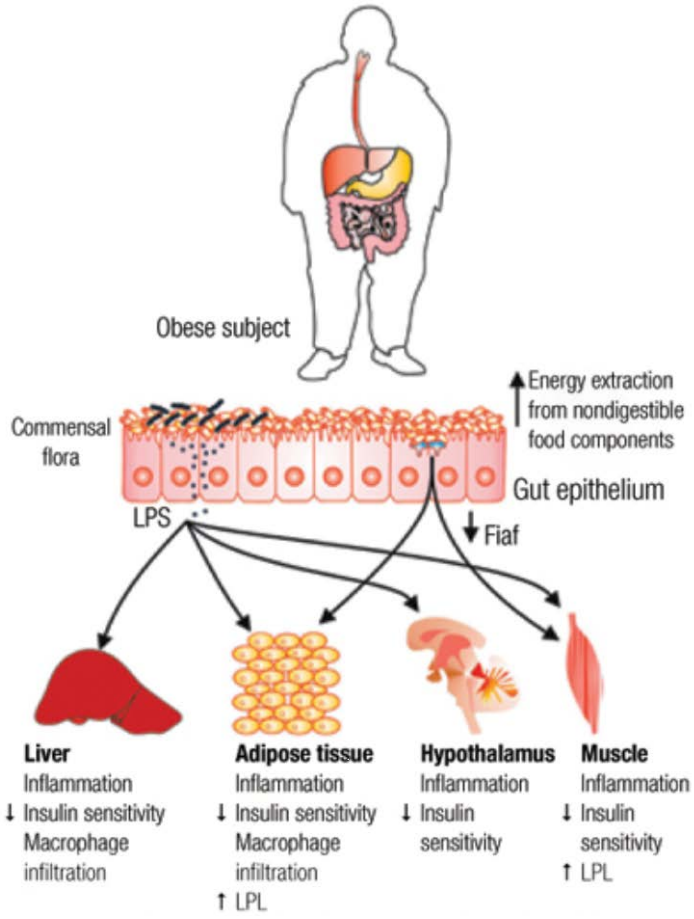


- The human body harbours 10^{14} bacteria (ten times more than all our cells together)
- The genes of these bacteria are 100 times more than all the genes of all the cells in our body
- Of over 50 known bacterial phyla, the human body harbours just 6 to 10
- There is no common pattern for microbiota and there are individual differences that may vary over time
- A large part of this microbiota is found on the skin, in the mouth, gut, and vagina
- The microbiota present in the gut is of particular relevance

The gut microbiome

- In a person weighing 70 kg, one kilo corresponds to their intestinal microbiota
- Bacteroidetes and Firmicutes are the dominant phyla and constitute 90 % of our microbiota
- The composition varies depending on which part of the gut is analyzed
- The metabolism of these microorganisms has a direct impact on our physiology and nutrition, and therefore on our health
- Some people have compared this microbiota inhabiting the gut to another organ, whose function was unknown

Obesity and gut microbiome



- Studies report an increase in Firmicutes and a decrease in Bacteroidetes in the digestive microbiome of murine models of obesity and in obese human volunteers
- Bacteroidetes increased in obese individuals on low-calorie diets, correlating with weight loss and reductions in BMI
- The affected metabolic pathways can be defined
- In obese rats, weight can be reduced by administering prebiotics or probiotics but results in humans vary
- All these data suggest that dietary intervention could be a strategy to manage obesity and associated metabolic disorders

Diabetes type II and gut microbiome

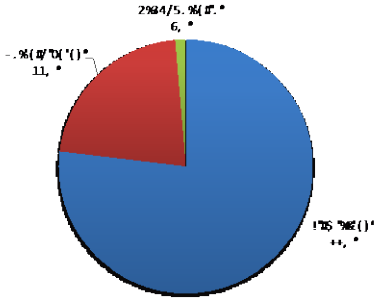
- People with type II diabetes show an increased Firmicutes /Bacteroidetes ratio and a decline in *Bifidobacterium* species and *Faecalibacterium prausnitzii*, both related to anti-inflammatory capacity
- Scientific data suggest that these species control enteroendocrine cell metabolism and the endocannabinoid system, which would explain the intestinal barrier dysfunction, metabolic endotoxemia and low-grade inflammation characteristic of the disease
- Some researchers have demonstrated in experimental murine models the positive effects of ingesting certain probiotics, mainly species of the *Bifidobacterium* genus

Brain behaviour and gut microbiome

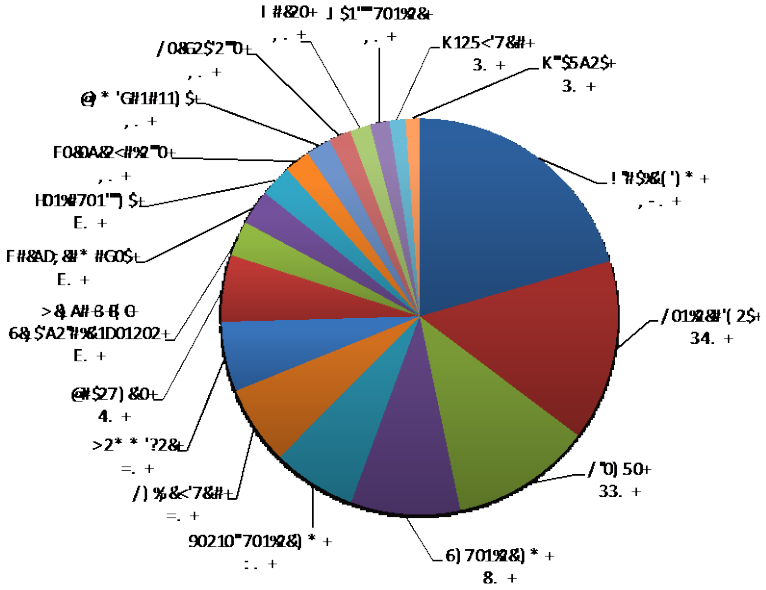
- Scientific data show there is a relationship between the gut microbiota and the central nervous system
- Studies into the regulation of anxiety, mood or pain -in animal models- suggest the brain-gut axis may be modulated through the intestinal microbiota
- An experimental mouse model of stress (induced by corticosterone) showed the intake of a *Lactobacillus rhamnosus* strain reduced symptoms of anxiety and depression by a mechanism related to the vagus nerve
- Very recently in a mouse model of autism, symptoms of the disease were reversed when the animals were fed with a strain of *Bacteroides fragilis*, a bacterium found in reduced numbers in autistic people

Colon cancer and microbiome (I)

A

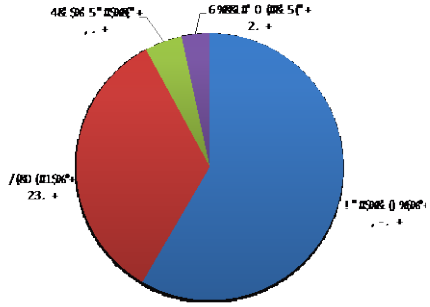


B

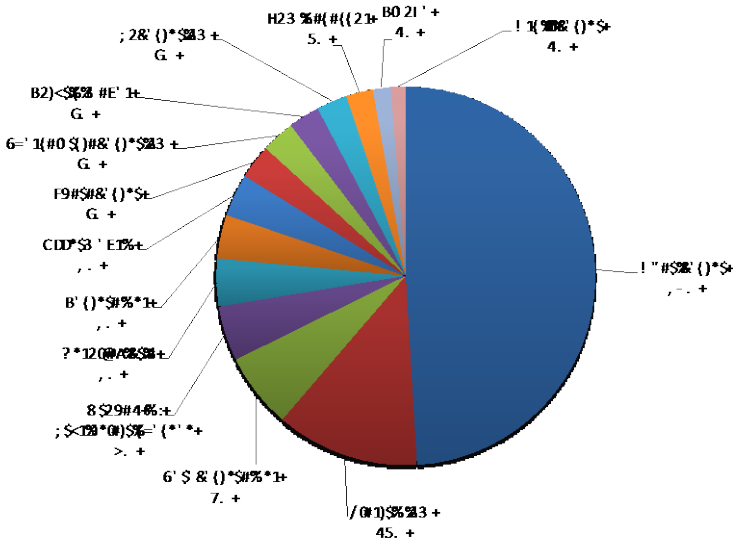


Colon cancer and microbiome (II)

A



B



Microbiomes and culture collection

Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice

Andrew L. Goodman[1], George Kallstrom, Jeremiah J. Faith, Alejandro Reyes, Aimee Moore, Gautam Dantas, and Jeffrey I. Gordon[2]

Center for Genome Science and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108

Contributed by Jeffrey I. Gordon, February 24, 2011 (sent for review January 21, 2011)

The proportion of the human gut bacterial community that is recalcitrant to culture remains poorly defined. In this report, we combine high-throughput anaerobic culturing techniques with gnotobiotic animal husbandry and metagenomics to show that the human fecal microbiota consists largely of taxa and predicted functions that are represented in its readily cultured members. When transplanted into gnotobiotic mice, complete and cultured communities exhibit similar colonization dynamics, biogeographical distribution, and responses to dietary perturbations. Moreover, gnotobiotic mice can be used to shape these personalized culture collections to enrich for taxa suited to specific diets. We also demonstrate that thousands of isolates from a single donor can be clonally archived and taxonomically mapped in multiwell format to create personalized microbiota collections. Retrieving components of a microbiota that have coexisted in single donors who have physiologic or disease phenotypes of interest and reuniting them in various combinations in gnotobiotic mice should facilitate preclinical studies designed to determine the degree to which tractable bacterial taxa are able to transmit donor traits or influence host biology.

gut bacterial diversity | nutrient–microbe interactions | translational medicine pipeline for human microbiome

Efforts to dissect the functional interactions between microbial communities and their habitats are complicated by the long-standing observation that, for many of these communities, the great majority of organisms have not been cultured in the laboratory (1). Methodological differences between culture-independent and culture-based approaches have contributed to the challenge of deriving a realistic appreciation of exactly how much discrepancy exists between the culturable components of a microbial ecosystem and total community diversity. Table S1 gives examples of these methodological differences.

The largest microbial community in the human body resides in the gut: Its microbiome contains at least two orders of magnitude more genes than are found in our *Homo sapiens* genome (2). Culture-independent metagenomic studies of the human gut microbiota are identifying microbial taxa and genes correlated with host phenotypes, but mechanistic and experimentally demonstrated links between key community members and specific aspects of host biology are difficult to establish with these methods alone. The goals of the present study were (*i*) to evaluate the representation of readily cultured phylotypes in the human gut microbiota; (*ii*) to profile the dynamics of these cultured communities in a mammalian gut ecosystem; and (*iii*) to determine whether a clonally arrayed, personalized strain collection could be constructed to serve as a foundation for reassembling varying elements of a human's gut microbiota in vitro or in vivo.

Results

To estimate the abundance of readily cultured bacterial phylotypes in the distal human gut, primers were used to amplify variable region 2 (V2) of bacterial 16S ribosomal RNA (rRNA) genes present in eight freshly discarded fecal samples obtained from two healthy, unrelated anonymous donors living in the United States (n = 1 complete sample per donor at t = 1, 2, 3, and 148 d). Amplicons were subjected to multiplex pyrosequencing, and the results were compared with those generated from DNA prepared from ~30,000 colonies cultured from each sample under strict anaerobic conditions for 7 d at 37 °C on a rich gut microbiota medium (GMM) composed of commercially available ingredients ("cultured" samples; details of the culturing technique are given in *SI Materials and Methods*, and a description of GMM is given in Table S2). The resulting 16S rRNA datasets were de-noised to minimize sequencing errors (3, 4), reads were grouped into operational taxonomic units (OTUs) of ≥97% nucleotide sequence identity (ID), and chimeric sequences were removed (*SI Materials and Methods*).

In total, 632 distinct 97%ID OTUs were observed in the complete samples, and 316 were identified in the cultured samples. The average abundance of cultured OTUs in the complete samples was 0.4%, but the average abundance of uncultured OTUs (i.e., those observed in the complete but not the cultured samples) was significantly lower (0.06%; $P < 10^{-6}$ by an unpaired, two-tailed Student's t test, not assuming equal variances) (5).

To evaluate the representation of readily cultured taxa in the human gut microbiota at varying phylogenetic levels, we assigned taxonomic designations to each 97%ID OTU (*SI Materials and Methods*). Each 16S rRNA read from the complete fecal sample was scored as "cultured" if it had a taxonomic assignment that also was identified in the corresponding cultured population. If a 97%ID OTU in the complete sample could not be placed in any known taxonomic group, it was scored as "cultured" only if the same 97%ID OTU was observed in the cultured sample. This analysis indicated that 99% of the 16S rRNA reads derived from the complete fecal samples from either donor belong to phylum-, class- and order-level taxa that are also present in the corresponding cultured sample; 89 ± 4% of the reads are derived from readily cultured family-level taxa, and 70 ± 5% and 56 ± 4% belong to readily cultured genus- and species-level taxa, respectively (Fig. 1*A Upper*). Two alternate taxonomic binning

Author contributions: A.L.G., G.K., J.J.F., G.D., and J.I.G. designed research; A.L.G., G.K., J.J.F., and A.M. performed research; A.R. contributed new reagents/analytic tools; A.L.G., J.J.F., A.R., and J.I.G. analyzed data; and A.L.G. and J.I.G. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession nos. SRA026269, 026270, and 026271).

Freely available online through the PNAS open access option.

[1]Present address: Section of Microbial Pathogenesis and Microbial Diversity Institute, Yale University, New Haven, CT 06536.

[2]To whom correspondence should be addressed. E-mail: jgordon@wustl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1102938108/-/DCSupplemental.

RESEARCH ARTICLE

A Catalog of Reference Genomes from the Human Microbiome

The Human Microbiome Jumpstart Reference Strains Consortium*†

The human microbiome refers to the community of microorganisms, including prokaryotes, viruses, and microbial eukaryotes, that populate the human body. The National Institutes of Health launched an initiative that focuses on describing the diversity of microbial species that are associated with health and disease. The first phase of this initiative includes the sequencing of hundreds of microbial reference genomes, coupled to metagenomic sequencing from multiple body sites. Here we present results from an initial reference genome sequencing of 178 microbial genomes. From 547,968 predicted polypeptides that correspond to the gene complement of these strains, previously unidentified ("novel") polypeptides that had both unmasked sequence length greater than 100 amino acids and no BLASTP match to any nonreference entry in the nonredundant subset were defined. This analysis resulted in a set of 30,867 polypeptides, of which 29,987 (~97%) were unique. In addition, this set of microbial genomes allows for ~40% of random sequences from the microbiome of the gastrointestinal tract to be associated with organisms based on the match criteria used. Insights into pan-genome analysis suggest that we are still far from saturating microbial species genetic data sets. In addition, the associated metrics and standards used by our group for quality assurance are presented.

The human microbiome is the enormous community of microorganisms occupying the habitats of the human body. Different microbial communities are found in each of the varied environments of human anatomy. The aggregate microbial gene tally surpasses that of the human genome by orders of magnitude. Understanding the relationship of the microbial content to human health and disease is one of the primary goals of human microbiome studies. Determining the structure and function of any microbial community requires a detailed definition of the genomes that it encompasses and the prediction and annotation of their genes.

In 2007, the National Institutes of Health (NIH) initiated the Human Microbiome Project (HMP) as one of its Roadmap initiatives (*1*) to provide resources and build the research infrastructure. One component of the HMP is the production of reference genome sequences for at least 900 bacteria from the human microbiome, which will catalog the microbial genome sequences from the human body and aid researchers conducting human metagenomic sequencing in assigning species to sequences in their metagenomic data sets.

The HMP catalog of reference sequences is being produced by the NIH HMP Jumpstart Consortium of four genome centers: the Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, the J. Craig Venter Institute, and the Genome Center at Washington University. The challenges for the Jumpstart Consortium include selecting strains to sequence and identifying sources, creating standards for sequencing and annotation to ensure consistency and quality, and the rapid release of information to the community.

Reference genome progress. To date, 356 genomes, including 117 genomes at various stages of upgrading, have been produced by the Jumpstart Consortium and released into public databases. At the time of manuscript preparation, 178 had been completely annotated and are presented in the analysis here. The process for the selection of these strains is described in (*2*). The strains sequenced to date are distributed among body sites as follows: gastrointestinal tract (151), oral cavity (28), urogenital/vaginal tract (33), skin (18), and respiratory tract (8). They also include one isolate from blood (*3*). These are the five major body sites targeted by the HMP.

The broad phylogenetic distribution of the sequenced strains is presented in Fig. 1, which represents a 16S ribosomal RNA (rRNA) overlay of HMP-sequenced genomes on 16S rRNA sequences from cultured organisms with sequenced genomes (*4*). HMP-sequenced genomes represent two kingdoms (Bacteria and Archaea), nine phyla, 18 classes, and 24 orders. Additional rRNA overlay figures broken down by individual body sites are available in (*5*).

To obtain high-quality draft genomes and a meaningful gene list, minimum standards were defined for the assembly and annotation of draft genomes. Three reference bacterial genome assemblies were evaluated for efficacy of gene predictions and genome completeness. Based on the analysis, metrics for assembly characteristics and annotation characteristics were defined [for more details, see (*2*)]. The quality of HMP genome assemblies is summarized in Table 1 and exceeds the Jumpstart Consortium standards described in (*2*), with the exception of some genomes produced before the standards were in place.

Genome improvement. As described in (*2*), there are justifications for upgrading these high-quality draft assemblies. The Jumpstart Consortium has completed initial improvement work on 26 bacterial genomes that differed significantly with respect to GC content and assembly metrics to explore the effort required and resulting benefits (Fig. 2). The average contig N50 increased 3.63-fold, from 109 kb at draft to 396 kb after improvement. *Bacteroides pectinophilus* displays substantial improvement in N50, from 163 kb in the draft sequence to 862 kb after improvement. *Lactobacillus reuteri* illustrates the opposite extreme, with improvement leading to a smaller contig N50 change, 56 kb to 72 kb. As more genomes improve and some graduate to higher levels of improvement, the assembly state or group of states most useful to the HMP scientific goals will be evaluated.

Pan-genome analysis. A bacterial species' pan-genome can be described as the sum of the core genes shared among all sequenced members of the species and the dispensable genes, or those genes unique to one or more strains studied. To start addressing questions about pan-genomes, we identified all species within our sequenced reference genome catalog for which there was more than one sequenced and annotated genome. Of the nine species identified, four of them have five or more annotated genomes that were generated either by the HMP or by external projects publicly available at the National Center for Biotechnology Information (NCBI); five genomes is the minimum number for which a curve can reliably be fit to pan-genome data. These are *L. reuteri*, *Bifidobacterium longum*, *Enterococcus faecalis*, and *Staphylococcus aureus*. The genomic data used for the analysis consisted of both complete and draft genomes, the only requirement being that >90% of the genome be represented in the available annotated contigs or scaffolds.

Pan-genome curves (*6*) of the gastrointestinal tract isolates *L. reuteri*, *B. longum*, and *E. faecalis* (figs. S3 to S5) are consistent with an open pan-genome model, suggesting that more genome sequencing needs to be undertaken to characterize the actual makeup of the species as a whole. Preliminary results suggest core genome sizes of approximately 1430 genes, 1800 genes, and 1600 genes for *B. longum*, *E. faecalis*, and *L. reuteri*, respectively. Based on the current core gene plots, *L. reuteri* (fig. S3) appears to be approaching a closed pan-genome model, with newly sequenced strains contributing very small numbers of new genes to the pan-genome; however, we see an interesting community substructure within this species. Our current *L. reuteri* pan-genome analysis of seven isolates suggests that four of the

*All authors with their affiliations and contributions are listed at the end of this paper.

†To whom correspondence should be addressed. E-mail: kenelson@jcvi.org

Molecular taxonomy: from biodiversity to biotechnology

Microbial genomics and culture collections

1. Massive genome sequencing
2. Genome sequencing and its applications
3. The other “omic” strategies
4. Conclusions

Genomics and culture collections

- Microbial culture collections hold an enormous treasure of biodiversity of social and industrial relevance
- Classical screening methods based on plate culture are time-consuming, expensive and moderately efficient
- During the last ten years, omic technologies and bioinformatic techniques have improved greatly
- We are technically ready to enter in a new era of massive genome screening of culture collections
- This is only the beginning; systems biology will be the future

The attitude

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less"

(Marie Curie, 1867-1934)

Contact

daniel.ramon@biopolis.es
Phone: (+34) 963160299
Biópolis SL
Parc Cientific Universitat de València
C/ Catedrático Agustín Escardino 9
Building 2
46980 Paterna (Valencia)
Spain