Bioinformatics: making large-scale science happen

Sergi Beltran

Unitat de Bioinformàtica, CCiTUB, Universitat de Barcelona. Parc Científic de Barcelona, Baldiri Reixach 10, 08028 Barcelona.

email: *beltran@ccit.ub.edu*

Abstract. Bionformatics is a rapidly evolving research field dedicated to analyzing and managing biological data with computational resources. This paper aims to overview some of the processes and applications currently implemented at CCiT-UB's Bioinformatics Unit, focusing mainly on the areas of Genomics, Transcriptomics and Proteomics.

1. Introduction

Bioinformatics is widely accepted to be the application of computational resources to the analysis and management of biological data. However, it was originally defined in 1970 by Ben Hesper and Paulien Hogeweg [1], as the "study of informatic processes in biotic systems" [2], and it was not until the rise of high throughput technologies in molecular biology that Bioinformatics, as is nowadays known, became an everyday word in science (Figure 1).

Computers and the Internet have allowed the current revolution in science, becoming essential to any field of study. On one hand, they provide communication services such as email, mailing lists, forums, papers or webinars. On the other, they are used to store and analyze data and provide online tools like databases, genome browsers, image banks or applications. Needless to say, the former services and tools have been possible through the development of programs, algorithms, statistical methods and even new programming languages, therefore generating knowledge by themselves.

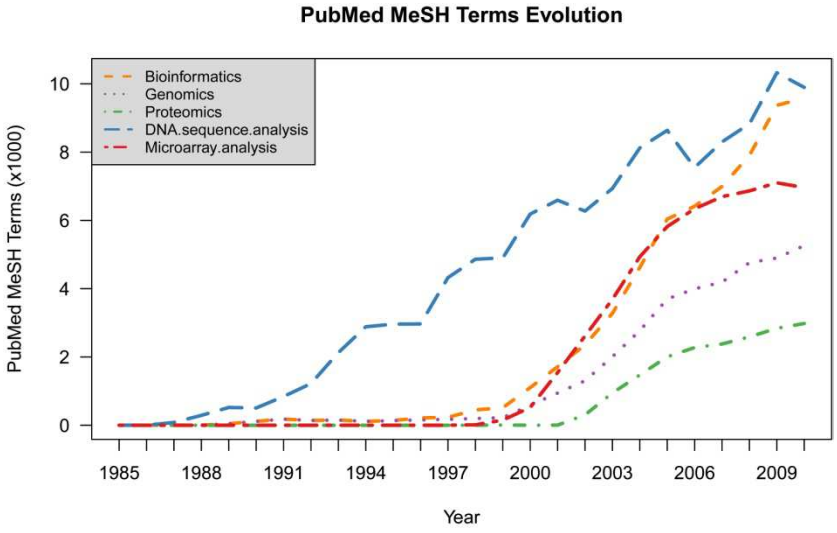

Figure 1. Number of articles by year classified by the following Medical Subject Headings (MeSH) terms from the Natio-nal Library of Medicine's con-trolled vocabulary: Bioinformatics, Genomics, Proteomics, DNA sequence analysis and Microarray analysis. Only data since 1985 are shown. Plot has been generated with R [3] and MEDSUM [4].

From a narrower point of view, Bioinformatics is associated with the analysis and management of massive amounts of data, be it DNA sequences, microarrays, in-situ hybridizations or microscope images. In that sense, Bioinformatics has been essential to the birth of Omics, fields that study the totality of a certain feature, and Systems Biology. To put it simple, the assembly and annotation of the Human genome [5, 6] or its proteome structure prediction [7] would have never been possible without Bioinformatics.

At CCiT-UB, the Bioinformatics Unit is mainly devoted to the areas of Genomics, Transcriptomics and Proteomics. It provides customized analyses, management, storage and visualization solutions to data from different sources such as next-generation *de novo* sequencing and re-sequencing, ChIP-on-chip, ChIPseq, DNA and protein microarrays, real time PCR or databases, amongst others. State of the art analyses are performed by using the latest software releases as well as developing and adapting programs to the researcher's needs.

2. Bioinformatics applied to Genomics, Transcriptomics and Proteomics

Next-generation sequencing (NGS, see the Genomics Unit paper in this issue) is a key achievement in molecular biology that is changing the way researchers work. While before it took years and endless resources to sequence a single genome, now it is possible to *de novo* sequence or re-sequence a small genome in a few hours and a complex one in just a few days. This new era has just begun, and in the near future it will be faster and cheaper to sequence a human genome, which might imply a big leap in personalized medicine [8]. Noteworthy, NGS has multiple applications

in fields such as metagenomics, where it can be used to decipher the microorganism's composition of samples like water, soil or feces [9, 10, 11].

However, obtaining the data is one thing, and making sense out of it is another and Bioinformatics must live up to this challenge; for example, algorithms and pieces of software that were originally designed to analyze small amounts of DNA must now be optimized and scaled up in order to deal with whole genomes, or even thousands of them [12, 13]. Moreover, all this data has to be stored in an accessible way, if possible with intuitive visualization tools, and usually has to be integrated or compared with other high-throughput data.

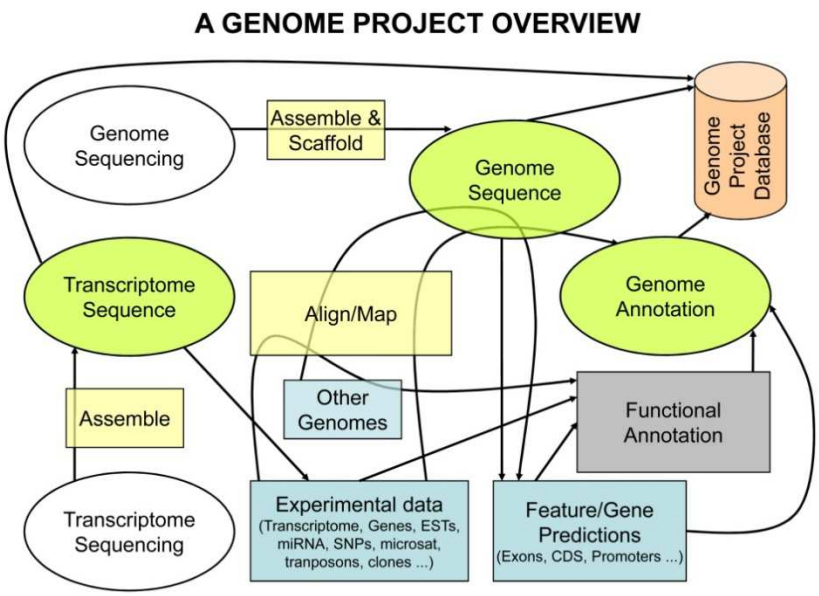

Figure 2. Non-exhaustive author's view of a genome project's workflow. Usually, a project will start by sequencing a genome or a transcriptome, which must be assembled into a consensus sequence. Annotation of this sequence can be done with *ab initio* predictions and/or using already available experimen-tal data from the same or other organisms. The relevant output from the project must be stored in the genome project's database. See main text for further discussion.

2.1. Genome-wide biology

A genome project (Figure 2) usually starts with *de novo* sequencing of the organism's DNA (genome sequencing). From the Bioinformatics point of view, it is a non-trivial step that starts by converting the experimental raw data (usually images) into usable sequences of DNA (reads). Quality control is an essential part since it must be ensured that only good quality reads are kept and primers, adapters and contaminants have been thoroughly detected and removed (Figure 3). With the aim of reconstructing the original sequence, those reads will be assembled (aligned and merged) into contigs which, if extra information like paired-end reads is available, will be sorted into scaffolds (Figures 2 and 4).

Once the genomic sequence is established, one or several annotation steps are usually performed (Figure 2). Annotation is the process of detecting and assigning features to the sequence. Amongst others, these features might be genes (Figure 4), open reading frames (ORFs), start and stop codons, transcriptions start sites (TSSs), splice donor and acceptor sites, regulatory regions, repeats, single-nucleotide polymorphisms (SNPs), miRNAs or protein motifs. Annotation can be achieved *in silico* using algorithms developed and trained to perform *ab initio* prediction of the features of interest. While some of these methods use the actual sequence as the only input, others compare it to other genomes by working under the assumption that biologically relevant strings of sequence are conserved throughout evolution. However, no matter how good the programs are, it is always desirable to have some experimental data to assess the predictions accuracy and adjust the algorithm's settings.

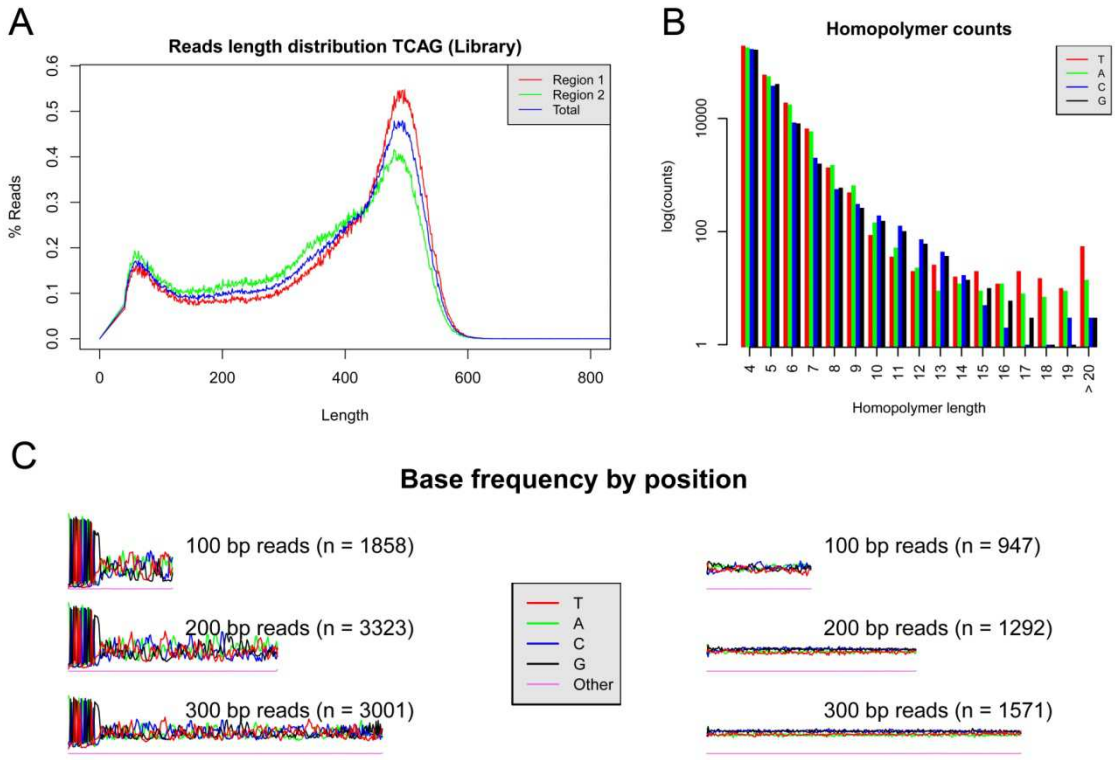

Figure 3. Some plots generated by our processing and quality control pipeline of Titanium FLX 454 sequencing runs performed at CCiT-UB's Genomics Unit. A) Percentage of reads from each length of the whole sequencing plate (blue), its first region (red) or its second region (green). B) Number of different size homopolymers from each base (see color legend). C) Percentage of each base (see legend) in each sequence position from 5' to 3' (left to right) in all reads with exactly 100bp, 200bp or 300bp. In comparison to the plot on the left, the one on the right had primers, adapters, polyA+, highly repetitive sequences and highly predominant sequences removed. All plots generated with R and Bioconductor packages [14, 15].

BT.5

High-throughput technologies enable whole genome annotation based on experimental evidence. For example, to find the binding sites of a certain protein, we can use the latter to perform chromatin immunoprecipitation, sequence the bound fragments and map them back to the genome, in a process known as ChIPseq. These results can also be used to predict putative binding site motifs (Figure 5A) that can be searched in the same or other related genomes. Or to elucidate the genes, and even their different transcripts, an organism's transcriptome can be sequenced and aligned to its corresponding genome (Figure 2).

Nevertheless, once a set of genes is described, functional annotation will provide them with a meaning by assigning names and functions (Figure 2). One of the most popular functional annotation methods is based on similarity comparison using BLAST [18] (Figure 4), in which the unknown nucleic and proteic sequences are queried against well annotated databases. The accuracy of the process increases as evolutionary distance between species decreases. Those results can then be used to group the genes/proteins into GeneOntology [19] categories or KEGG pathways [20] (Figure 6).

Finally, a decision has to be made on which data to store. Even with the actual continuous development of storage hardware, the rate at which new data are generated is posing a strain on storage facilities and, in many cases, only processed data can be kept. These data have to be organized in databases easily searchable by researchers and, if possible, accompanied with visualization tools like genome browsers (Figure 4).

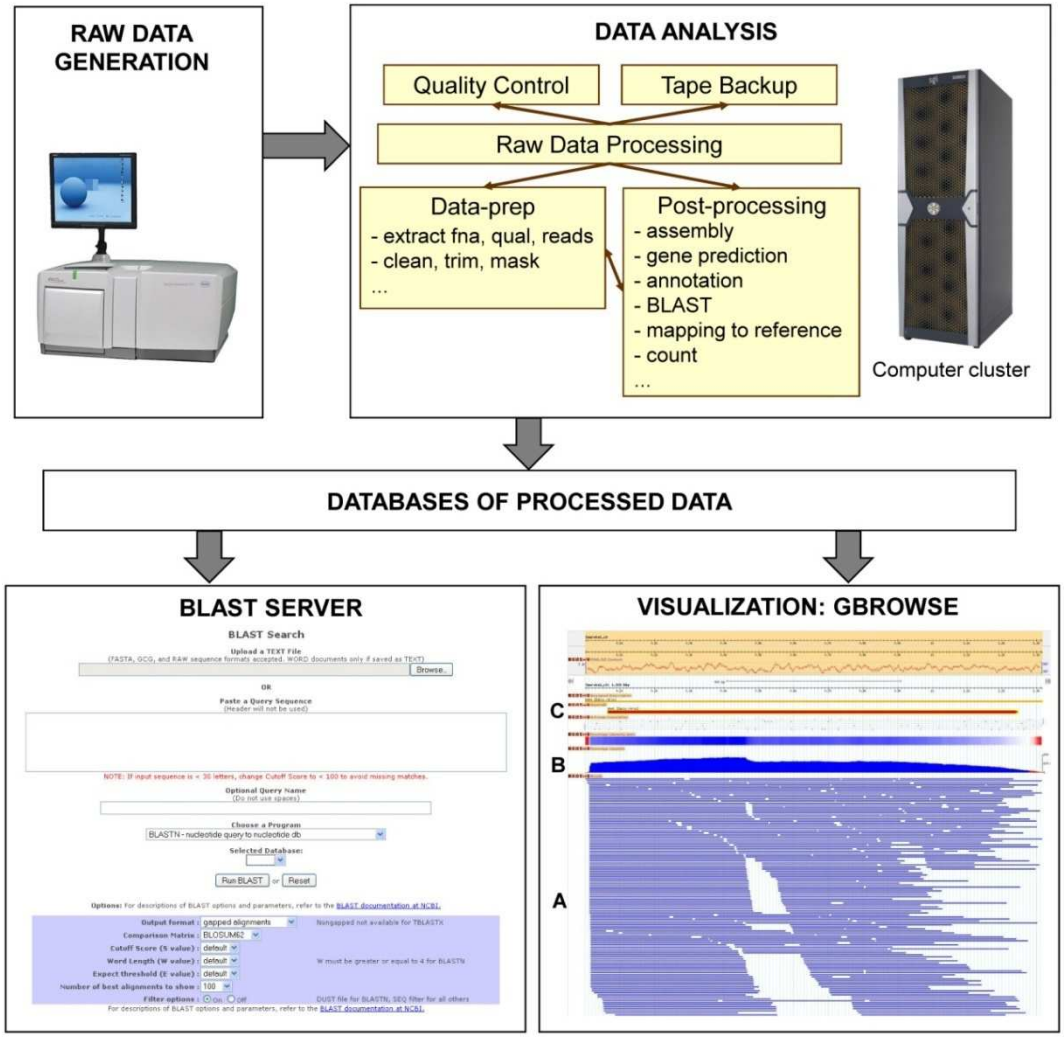

Figure 4. Bioinformatics Unit typical data workflow for a 454 sequencing project. Raw data generated by the Genomics Unit is transferred to our analysis cluster to conduct processing, quality control and backup. The nature of the experiment will determine the kind of data preparation and post-processing pipelines that must be applied. Processed data is then stored in databases that can be queried online to retrieve information in text format (not shown). In addition, results like assembled contigs can be blasted online (left) or visualized with GBrowse [16] (right). In this GBrowse example, some tracks from a contig are shown. In A), each horizontal blue line is one of the 454 reads that has been assembled to generate the contig. In B), the coverage of the contig is shown: y axis shows how many times each consensus base has been sequenced (this is, how many reads cover that base). In C), a putative gene (red box) inferred from similarity to another species. Preliminary data used in this example has been described in Martínez-Rodríguez et al. [17].

2.2. Biomedical applications

Bioinformatics are essential for the analysis of high-throughput experiments aimed to novel discovery, but can also be applied to the development and analysis of biomedical applications. For example, an annotated genome can be useful to design a DNA microarray aimed at classifying tumor types based on their gene expression profile, which might help to decide upon a treatment. In the same direction, *in silico* protein structural modeling can aid to find novel drug targets.

Whole genomes and transcriptomes from individuals sharing the same disease can be sequenced and compared in order to discover common variations amongst patients. If proofed, those variants become new treatment targets or, at least, candidates for developing a new molecular diagnostic kit based, for example, in nucleic acids hybridization or whole-gene sequencing. For that purpose, technologies like targeted sequence capture and molecular barcoding can help to bring costs down

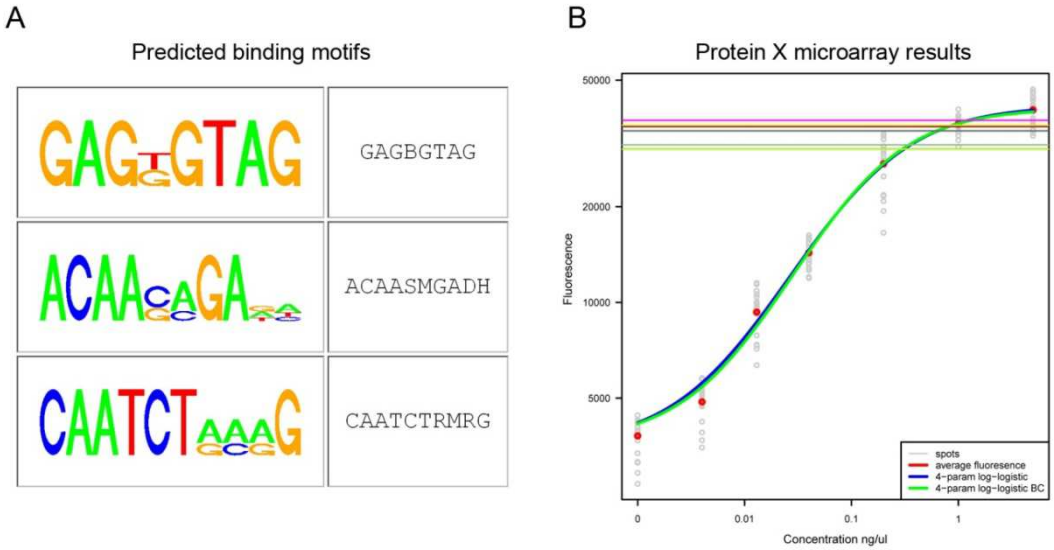

Figure 5. A) SeqLogos (left) and consensus sequence (right) of protein N putative binding sites predicted with BCRANK [21, 22]. B) Preliminary results from a protein microarray prototype produced at the PCB-UB Transcriptomics Platform by A. Sierra and L. Muixí from IDIBELL and L. Sevilla and J.I. Pons from PCB. Red circles show average normalized fluorescence at different concentrations from protein X replicates (in gray). Thick lines display the fitted 4 parameters log-logistic model with (green) or without (blue) Box-Cox transformation. Horizontal lines display observed fluorescence from different samples for which concentration of protein X wants to be assessed. Statistics and plots generated with R packages using scripts developed at CCiT-UB's Bioinformatics Unit.

BT.5

by restricting the area to sequence and allowing sample pooling, respectively. Alternatively, a protein microarray could be design to monitor the levels from relevant proteins in samples (Figure 5B).

What all these applications have in common is the need of Bioinformatics, first to analyze preliminary data, then to design it and, finally, to generate a method that ensures the standardized and correct analysis of the output.

2.3. Metagenomics

Metagenomics is a rapidly growing science that applies genomics to whole, usually microbial, communities, avoiding isolation and culture steps. Therefore, instead of trying to physically isolate and individually sequence all the organisms from a sample, its whole nucleotidic material, or certain common regions, can be sequenced at once and split afterwards by computational means.

There are endless potential applications derived from metagenomics analysis, ranging from health issues to earth matters and the bioinformatic pipelines must adapt accordingly. For example, without aiming to know what organisms they belong to, the sequences from a sample can be classified into evolutionary trees and quantity or proportions of the microbes, as well as diversity, can be computed. One application is to classify samples according to their microbial community composition, which might be useful to assess if a drug is working well in treating a gut disorder or if the soil is recovering from a certain contamination.

In addition, sequences can be compared to those from databases to elucidate which organisms, and in which quantity/proportion, are found in the community. In addition, the former can be

functionally classified. For example, certain organisms are only found in clean waters or are known to cause a disease, so quality control or preventive measures can be taken according to metagenomics results.

Last but not least, metagenomics can be applied to the discovery of new microbes and functions that might help develop novel biotechnological products directed towards pharmacy, food production, plague control, biofuel production or decontamination, just to mention a few. In fact, one of the most ambitious current science projects, the Global Ocean Sampling Expedition, is sailing around the globe taking samples to be sequenced and analyzed [23] in a worldwide enterprise benefitting from high-throughput sequencing and Bioinformatics.

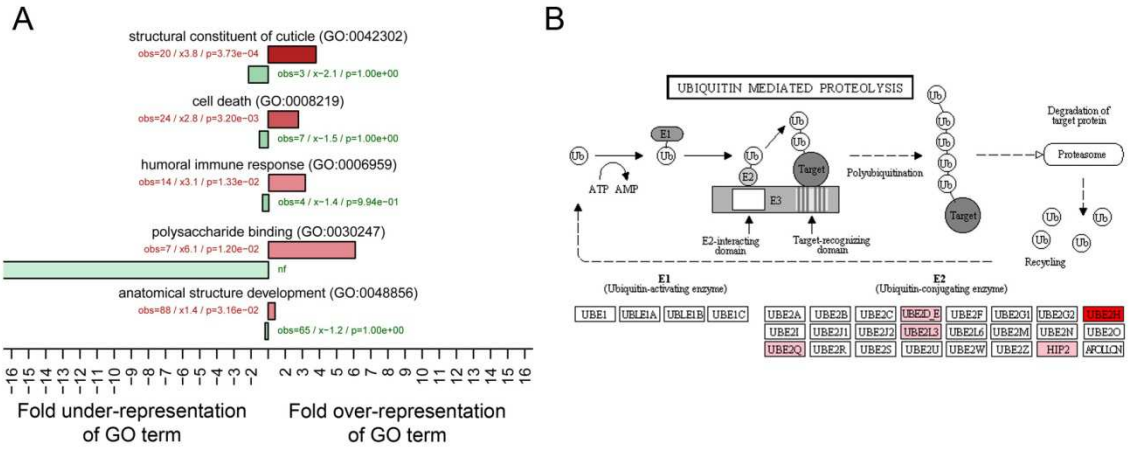

Figure 6. Examples of GeneOntology profiles and KEGG pathways visualization. A) GeneOntology barplots generated by Bioinformatics Unit on data processed with scripts adataped from [24] and [25]. Each bar's length shows that GO term fold under- (left) or over-representation (right) amongst down- (green) or up-regulated (red) genes from cut *D. melanogaster* wing discs at 24 hours compared to 0 hours [26]. obs = observed, x = fold enrichment or depletion (-), p = FDR adjusted p value, nf = not found. B) Upregulated genes (over 1.5x in pink and over 2.0x in red) from *D. melanogaster* dSAP18 mutants [27] shown in Ubiquitin mediated proteolysis KEGG pathways diagram. Only E1 and E2 are shown.

Acknowledgements

V. Gordo, P. Altube, C. Ligero and A. Godoy have implemented GBrowse, mySQL and BLAST server functionality, have developed scripts for automatic database updates and have scaled-up and improved our customized GeneOntology scripts. Thanks to B. Fusté and A. Arasanz (CCiTUB's Genomics Unit) for insightful discussions and help setting up the data processing and quality control pipeline for 454 sequencing data. Thanks also to Suport Informàtic for working towards adapting CCiTUB's resources to deal with big data loads.

References

1. Ben PH. Bioinformatica: een werkconcept. *Kameleon*. 1970;1(6):28-29.
2. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*. Mar 2011;7(3):e1002021.
3. Core. *R: A Language and Environment for Statistical Computing*. Vienna, Austria 2010.
4. Galsworthy M. MEDSUM: an online MEDLINE summary tool
5. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. Feb 2001;409(6822):860-921.
6. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*. Feb

2001;291(5507):1304-1351.

7. Bonneau R. The Human Proteome Folding Project
8. Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med*. 2010;2(11):84.
9. Shi Y, Tyson GW, DeLong EF. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature*. May 2009;459(7244):266-269.
10. Bell TH, Yergeau E, Martineau C, Juck D, Whyte LG, Greer CW. Identification of Nitrogen-Incorporating Bacteria in Petroleum-Contaminated Arctic Soils Using 15N DNA-SIP and Pyrosequencing. *Appl Environ Microbiol*. Apr 2011.
11. Reyes A, Haynes M, Hanson N, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. Jul 2010;466(7304):334-338.
12. Project 1G, Durbin RM, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. *Nature*. Oct 2010;467(7319):1061-1073.
13. Genome IC, Hudson TJ, Anderson W, et al. International network of cancer genome projects. *Nature*. Apr 2010;464(7291):993-998.
14. Gentleman RC, Carey VJ, Bates DM,. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*. 2004;5:R80.
15. Pages H, Aboyoun P, Gentleman R, DebRoy S. *Biostrings: String objects representing biological sequences, and matching algorithms*.
16. Stein LD, Mungall C, Shu S, et al. The generic genome browser: a building block for a model organism system database. *Genome Res*. Oct 2002;12(10):1599-1610.
17. Martínez-Rodríguez G, Halm S, Planas J, Beltran S, Yúfera M. Transcriptomic analysis of gilthead seabream larvae by 454 pyrosequencing. Preliminary annotation results., 2010.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. Oct 1990;215(3):403-410.
19. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. May 2000;25(1):25-29.
20. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. Jan 2000;28(1):27-30.
21. Ameur A. BCRANK: Predicting binding site consensus from ranked DNA sequences. 2010.
22. Bembom O. *seqLogo: Sequence logos for DNA sequence alignments*.º
23. Yooseph S, Sutton G, Rusch DB, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol*. Mar 2007;5(3):e16.
24. Beltran S, Blanco E, Serras F, et al. Transcriptional network controlled by the trithorax-group gene ash2 in Drosophila melanogaster. *Proc Natl Acad Sci U S A*. Mar 2003;100(6):3293-3298.
25. Beltran S, Angulo M, Pignatelli M, Serras F, Corominas M. Functional dissection of the ash2 and ash1 transcriptomes provides insights into the transcriptional basis of wing phenotypes and reveals conserved protein interactions. *Genome Biol*. 2007;8(4):R67.
26. Blanco E, Ruiz-Romero M, Beltran S, et al. Gene expression following induction of regeneration in Drosophila wing imaginal discs. Expression profile of regenerating wing discs. *BMC Dev Biol*. 2010;10:94.
27. Costa E, Beltran S, Espinàs ML. Drosophila melanogaster SAP18 protein is required for environmental stress responses. *FEBS Lett*. Jan 2011;585(2):275-280.