



High Throughput Scientific Computing in Vaccine Research: Challenges and Opportunities

Riccardo Beltrami
Head, Data Science Unit
Novartis Vaccines

Univ. di Siena, Dec. 10th 2014
Rev. 1.2.3

PDC# SIE-14-0923



Outline



The Economist, Feb. 25th 2010

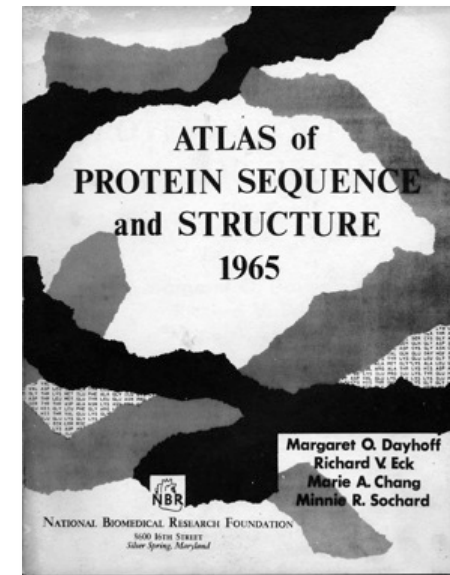
- Consequences of the Data Deluge
- Challenges for HPC / HTC Infrastructures and Organizations
- Some Use Cases and Examples
- Ideas and Opportunities Ahead

When Biology Met Information Technologies



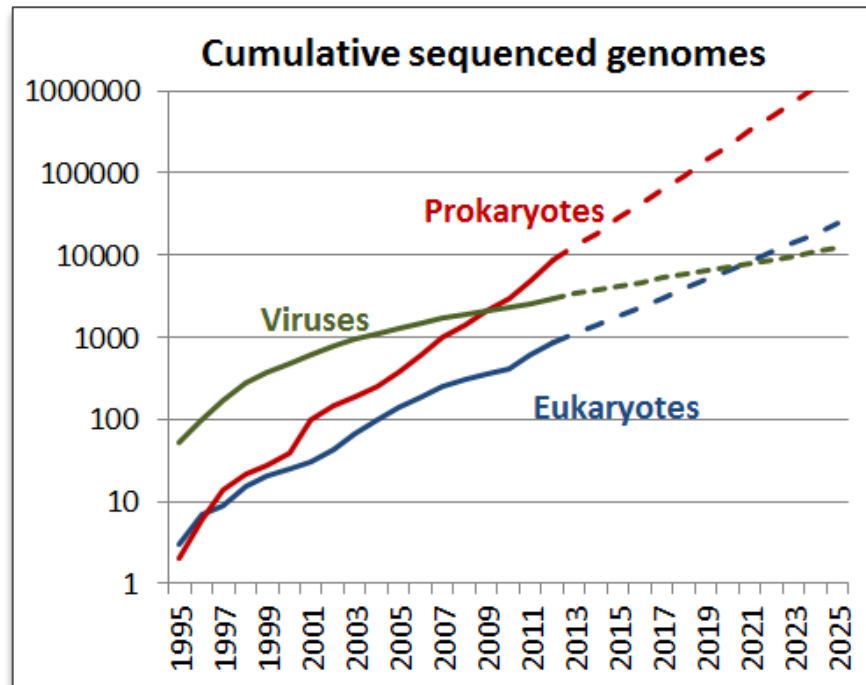
http://en.wikipedia.org/wiki/File:Margaret_Oakley_Dayhoff_cropped.jpg

Since the pioneering work of M. Dayhoff biology and medicine have become computing-intensive and data-driven sciences

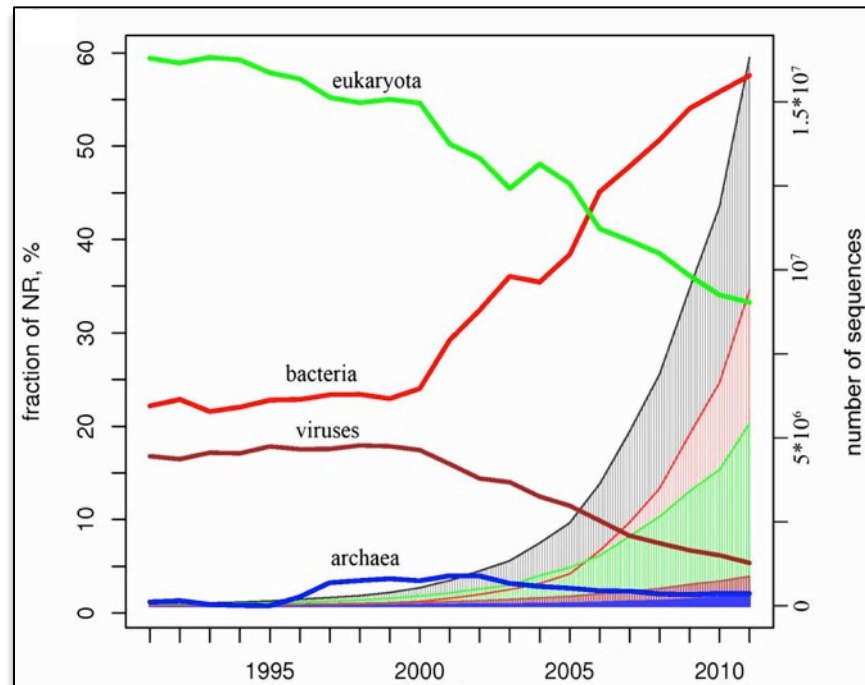


Data, data and... data

Growth of genomic data across thirty years



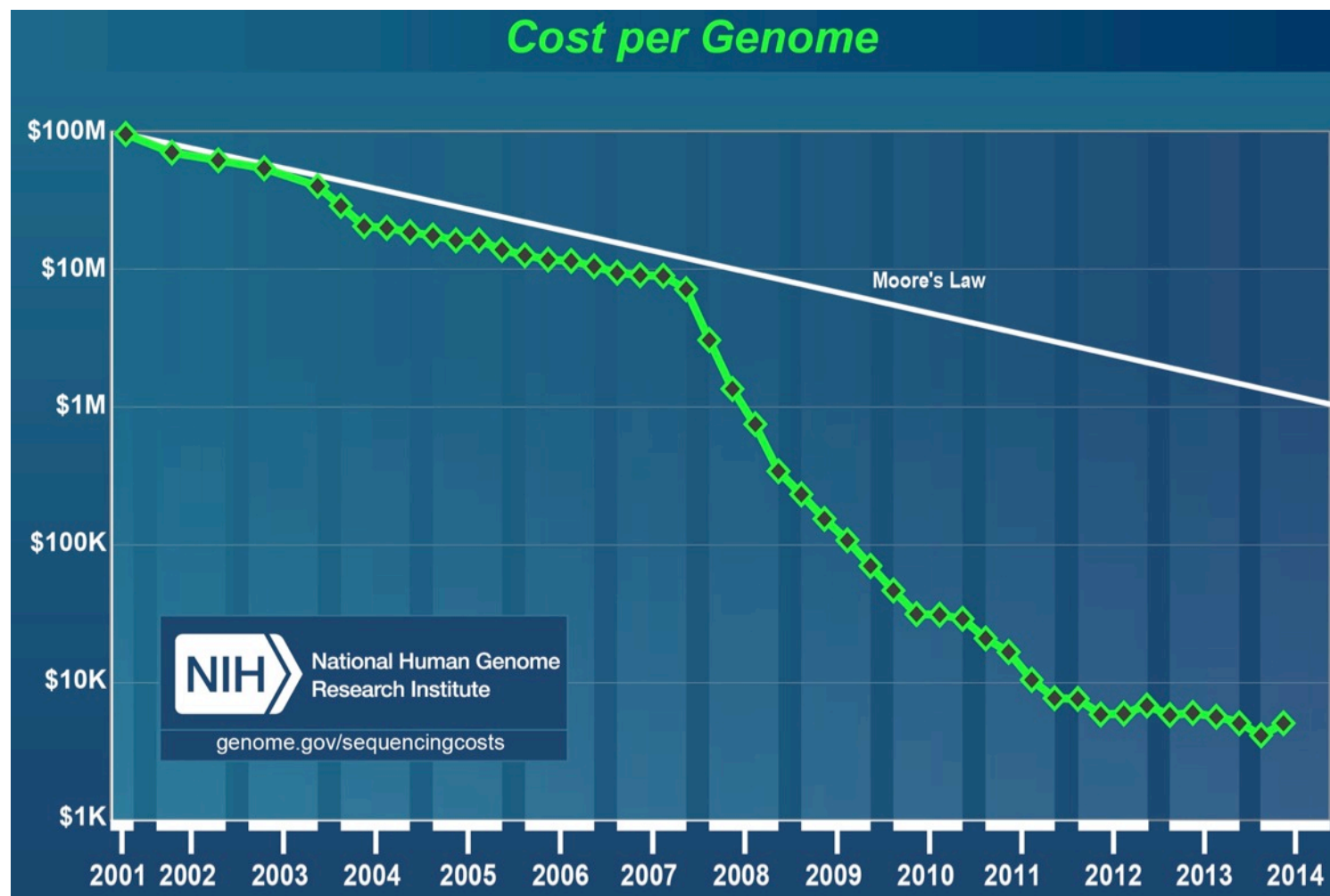
<http://sulab.org/2013/06/sequenced-genomes-per-year>



<http://www.pnas.org/cgi/doi/10.1073/pnas.1321614111>

Data, data and... data

Cost of sequencing compared to cost of computation



<http://www.genome.gov/sequencingcosts/>

HPC/HTC Infrastructures Objectives

"... nanos gigantis humeris insidentes."

■ The five 'S'

- Speed-up
- Store
- Secure
- Study
- Share



The Problems Space

- Thousands of genomes to be processed and analyzed along with an exponential growth of reference databases
- Hundreds of in-silico experiments to be performed
- Diverse, changing, growing datasets
- Evolving algorithms, analysis workflows and handling of hundreds of different data formats
- Complex, interconnected computing infrastructures
- Efficient archival and sharing of data and visualization of results

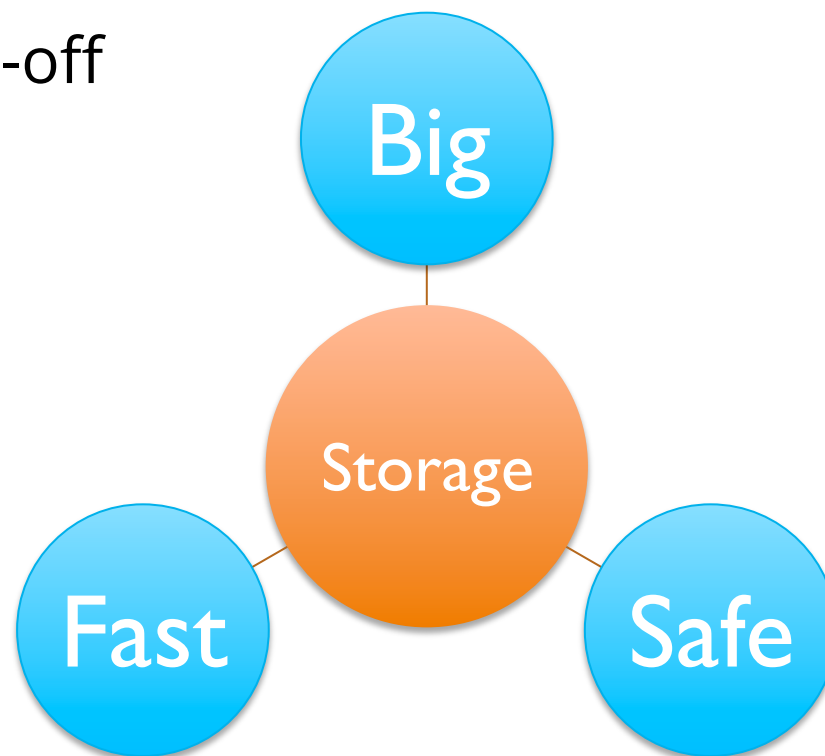
Today HPC/HTC Infrastructure Challenges

- Fast innovation of wet-lab technologies
- Data handing is the key point
- Computation and networks are not big issues, usually
- Open (stable) standards and interoperability
- Distributed and complex architectures
- Intertwined *wet-lab* and computational experiments

HPC/HTC Storage Systems

An example of current challenges on computing hardware

- Adapt to find the right trade-off
- Multiple storage tiers
- Storage is getting smarter
- Data life-cycle management
- Ease of management



Challenges for Scientific Computing Teams

Processes, policies and organizations

- Open Hardware | Source | Standard | Data
- Collaborative Efforts / Governance Schemas
- Bring Life and Computer Scientists Closer
- Ease of Data Sharing and Information Exchange
- Quality and Reproducibility of Results

Extracting Information

The end goal of our endeavor



Some examples of what we did with our HTC scientific infrastructure

REAL USE CASES

DSU Scientific Computing Infrastructure

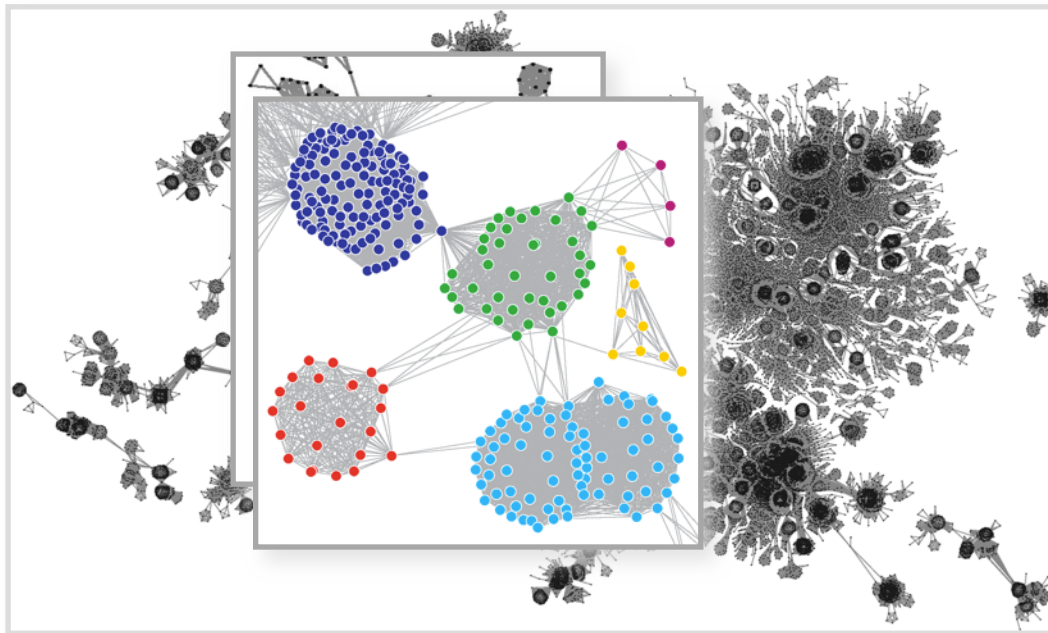
By the numbers

- Built from scratch in 1999
- Two HPC Clusters (~1K cores) + dedicated servers
- 10/40Gb Ethernet and 40Gbp IB Networks
- ~1PByte of Storage (HT scale-out, DAS & DRS)
- (Mostly) Open and standard operating systems and software
- Up to 100 users, 15% advanced users
- Max 90KVA power consumption

1. PHN FAMILIES

Protein Homology Network Families

- More than 4 millions of bacterial proteins have been identified
- “Similar” proteins could be grouped together
- Function of unknown genes products could be inferred from the most similar protein family



A. Covacci et al.
(doi:10.1371/journal.pcbi.0020173)

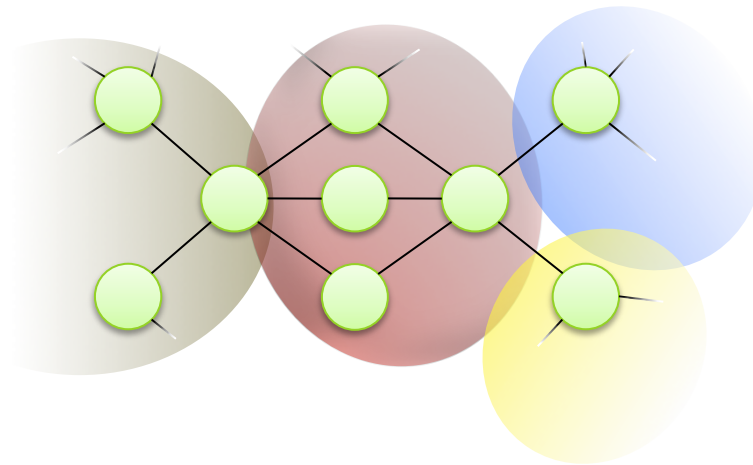
Protein Homology Network Protein Families

High level process description

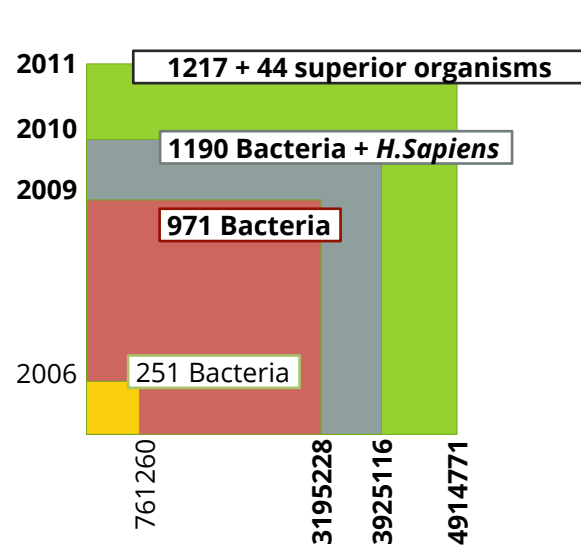
Measure distance between network elements (proteins)

Measure of topological similarities to rewire the network

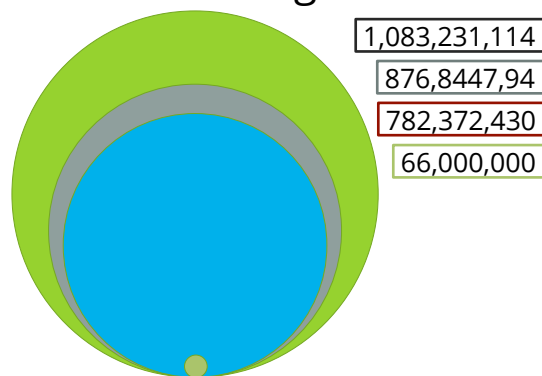
Obtain the connected components



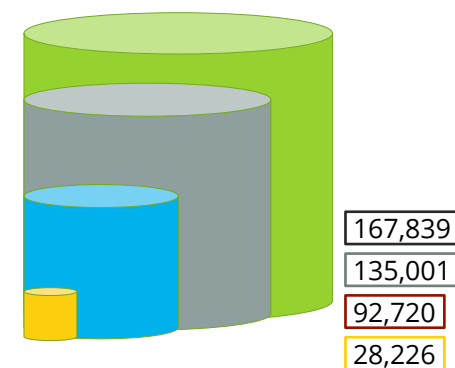
PHN Families Results



Number of connections from BLAST alignments

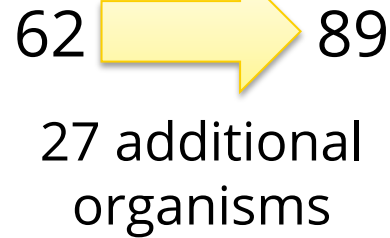


Number of families



Number of genomes	Year	Number of proteins	Number of families**	Time * (hours)
971 Bacteria	2009	3195228	92720	27
1190 Bacteria + <i>H.Sapiens</i>	2010	3925116	135001	31
1217 + 44 superior organisms	2011	4914771	167839	36

* 84 Intel Xeon E5660 processors, frequency 2.67 GHz, 6-cores; ** $E\text{-value} = 10^{-3}$, $\theta = 0.5$

20 | High Throughput Scientific Computing in Vaccine Research | R. Beltrami | Dec. 10th, 2014

PHN Protein Families

Computational challenges

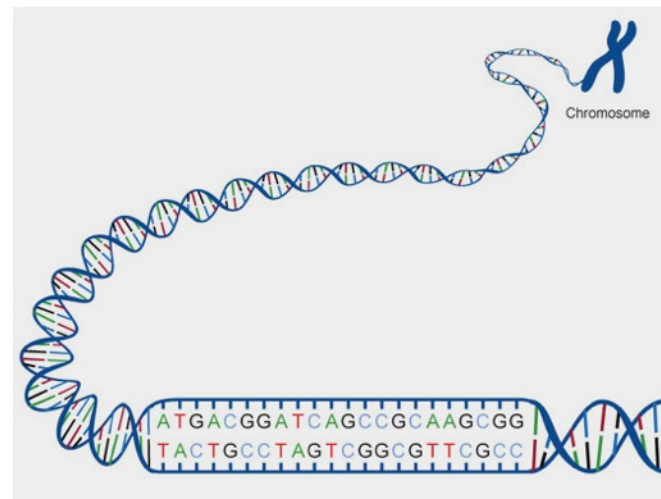
- $O(n^2)$
- Find the "optimal" parameters of the algorithm
- Large single image memory
- Large number of relatively small files generated during execution
- Highly influenced by storage performances (metadata I/O)
- Browsing and visualization of results

2. DEEP SEQUENCING

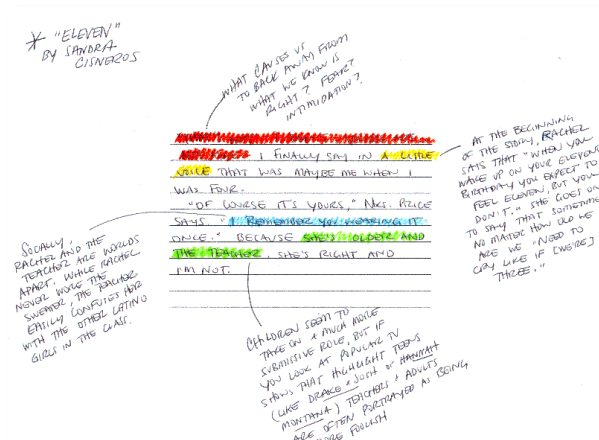
Deep Sequencing Data Analysis

Genomics pillars

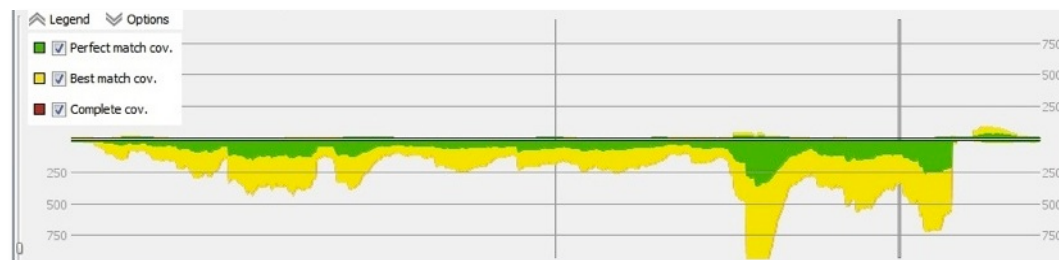
1. Sequencing & Assembly



2. Annotate

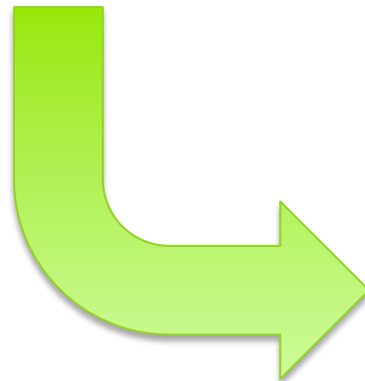


3. Evaluate expression



1. The Assembly Phase

How to get from one to the other with no instruction manual



2. The Annotation Phase

Metadata rule

Bioinformatics Advance Access published January 10, 2008

Sequence analysis **31 JAN '08**

Beyond the "Best Match": Machine Learning Sequences by Integration of Different Sources

Igor V. Tetko,^{1*} Igor V. Rodchenkov,¹ Matthias Thomas Rattei,² Hans-Werner Meier

¹Helmholtz Center Munich - German Research Center for Environmental Health, Stadt der Landstraße 1, Neuherberg, Germany
²Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

Associate Editor: Burkhard Rost

ABSTRACT
Accurate automatic assignment of protein functions remains a challenge for genome annotation. We have developed and compared the automatic annotation of four bacterial genomes employing a 5-fold cross-validation procedure and several machine learning methods. The analyzed genomes were manually annotated with FunCat categories in MIPS, providing a gold standard. Features describing a pair of sequences rather than each sequence alone were used. The descriptors were derived from sequence alignment scores, InterPro domains, synteny information, sequence length, and calculated protein properties. Following training we scored all pairs from the validation sets, selected a pair with the highest predicted score and annotated the target protein with functional categories of the prototype protein. The data integration using machine-learning methods provided significantly higher annotation accuracy compared to the use of individual descriptors alone. The neural network approach showed the best performance. The descriptors derived from the InterPro domains and sequence similarity provided the highest contribution to the method performance. The predicted annotation scores allow differentiation of reliable vs. non-reliable annotations. The developed approach was applied to annotate the protein sequences from 180 complete bacterial genomes. The FUNcat Annotation Tool (FUNAT) is available on-line as Web Services at <http://mips.gsf.de/proj/funat>.

INTRODUCTION
(Altschul, et al., 1997), FASTA (Pearson, 1996) or more sensitive methods, such as hidden Markov models (Sonnhammer, et al., 1997) or alignment-independent tools (Kocsor, et al., 2006). These methods are applied to detect a set of candidate genes, which were previously manually annotated. When such set of genes is pre-selected, several strategies are used to provide the annotation. One of the most simplest and straightforward approaches is to annotate a target gene with functions of a prototype gene. Homology-based methods use the highest alignment score ("best bidirectional match") to do such annotation. They were implemented in the first workflow-based annotation systems for sequence analysis such as GeneQuiz (Andrade, et al., 1999) and PEDANT (Mewes, et al., 1999; Riley, et al., 2007). In addition, these approaches performed supplementary analyses to add multiple evidences for the inferences made. Another strategy is to annotate proteins using clustering algorithms applied on many proteins. Orthologous families are detected and the consensus annotation of all genes in the cluster is assigned (Enright, et al., 2002; Tetko, et al., 2005). More complicated methods that make use of a combination of annotations of more than one homolog have been also proposed (Abascal and Valencia, 2003; Biswas, et al., 2002; Clare, et al., 2006; Clare and King, 2003; Krebs and Bourne, 2004; Kretschmann, et al., 2001; Levy, et al., 2005; Meinel, et al., 2005; von Mering, et al., 2007). It has been shown that the latter strategies are suited to provide an improved accuracy of annotation over the best match approach.

MLRA - KNN - ASNN
TRAINING SET → 4 GENOMI BATTERICI MANUALLY ANNOTATED DA FUN CAT E LE LUI PROPRIETA' PER I DESCRITTORI SONO FORNITE DA INTER PRO
RUNNING SET → 180 GENOMI DI CUI SONO DISP. ALLINEAMENTO E PROLISTA DA PEDANT & INTER PRO
SCORE & ANNOTAZIONE → IN ACCORDO A FUN CAT

SUGGERISCE:
LA RICERCA NA. CRATTERIZZATA COME ALLINEAMENTI DI 2 PROTEINE PULITORE CHE COME P. INGLESE RISPONDE.

DB di famiglie di proteine di proprietà e siti funzionali
nel quale feature identificabili delle proteine note possono essere applicate a sequenze proteiche non note

schema di annotazione per descrizione funzionale di proteine

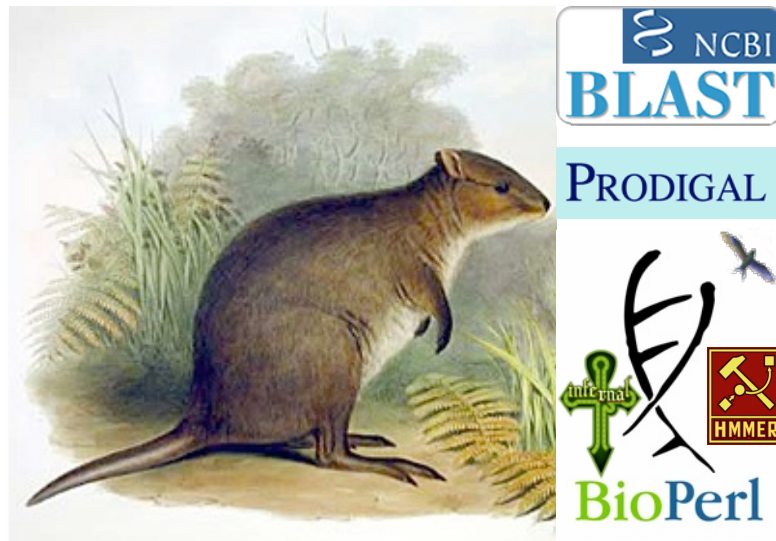
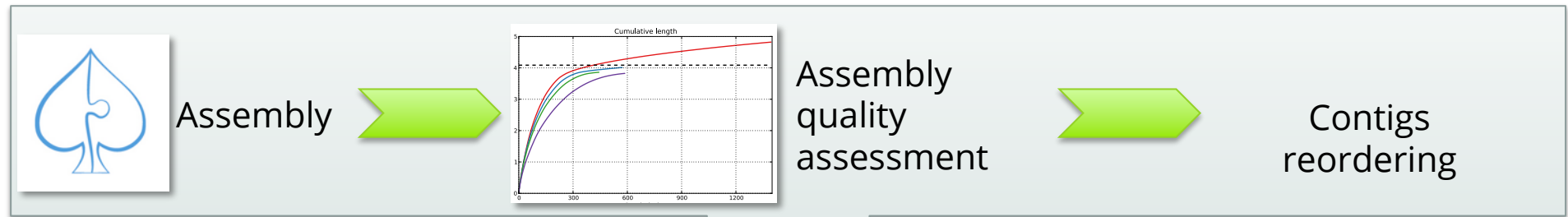
PB
accidentale
proprietà
ovvio
definiscono le feature

strategy 1
avvicinamento di sequenze/genomi note
POST-EDITION

strategy 2
clustering
che sono di sequenze generato da un solo place di un genome

2. From Sequences to Genes

Automatic annotation pipeline

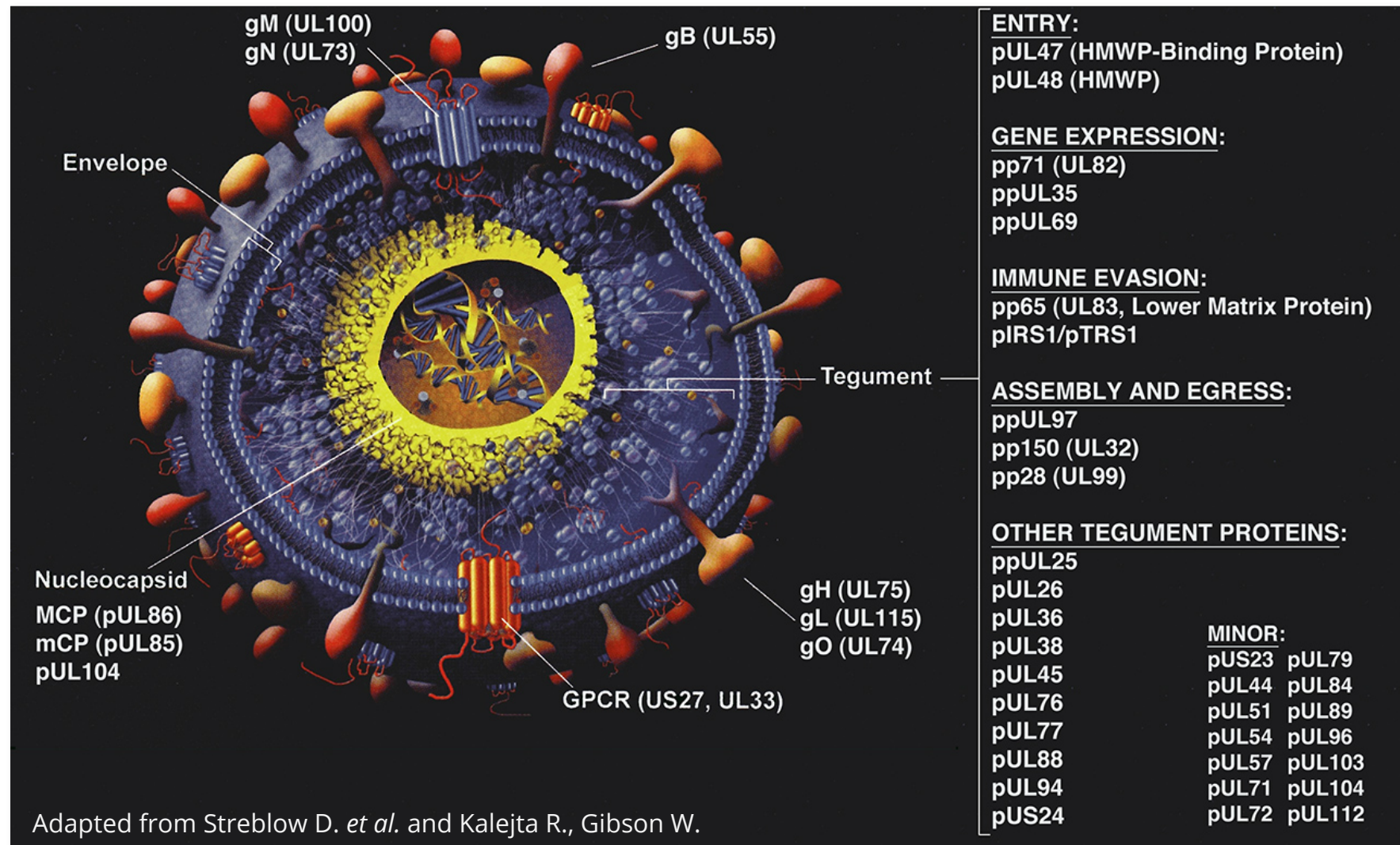


Quokka (*Setonix brachyurus*) from "Mammals of Australia" by John Gould, 1863.

Annotation process:

- Finds transfer RNAs
- Predict ribosomal RNAs
- Identify coding regions
- Spot signal peptide motifs
- Search similar proteins for functional annotation
- Explore protein families
- Predict non coding RNAs

2. Atlas of Gene Products



2. Comparative Genome Analysis



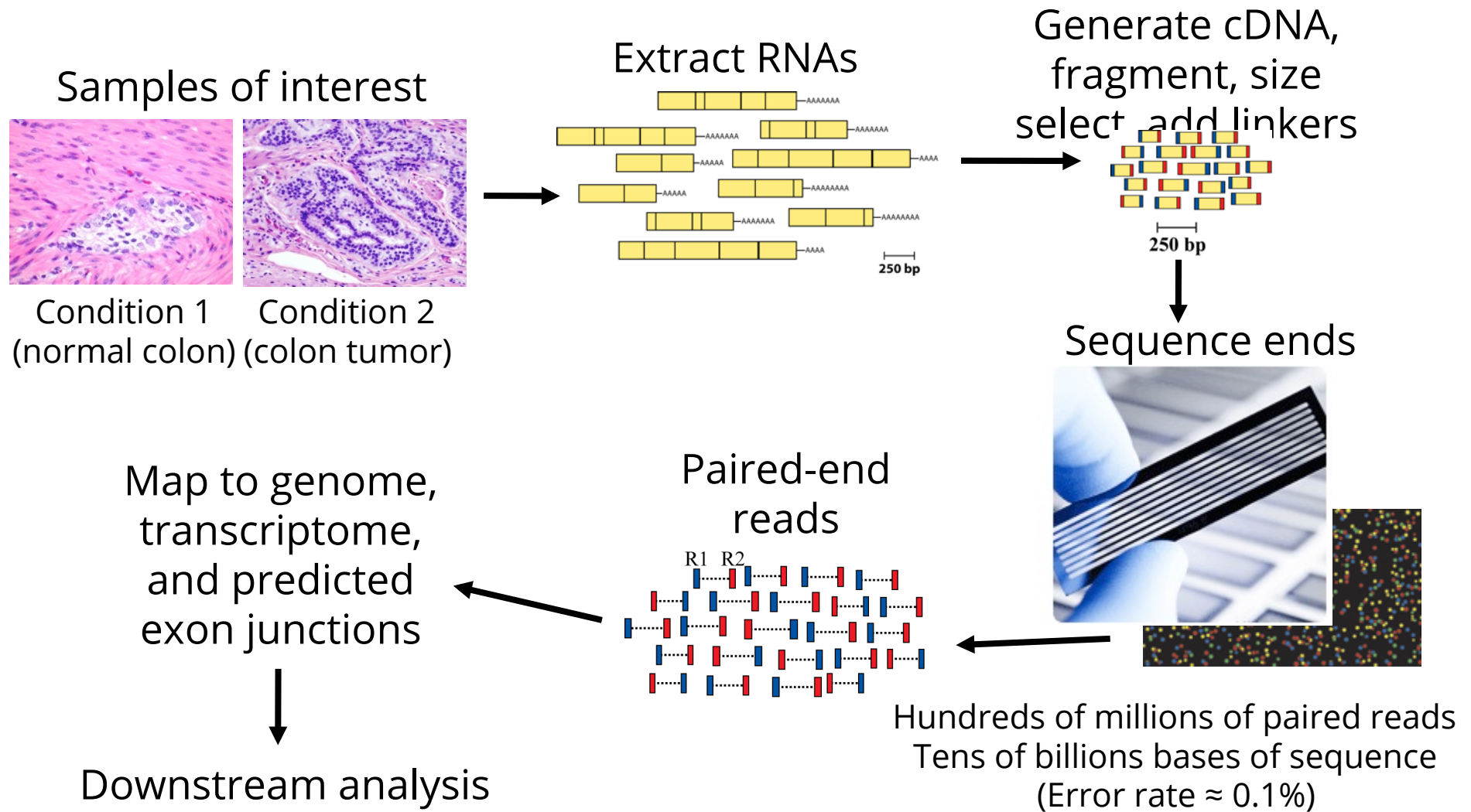
3. RNA-Seq

A genome-wide transcriptional activity snapshot



- Expression level quantification by normalized counts of reads in specific genomic regions
- Differential gene expression
- Annotation-independent detection of transcription
(large and small RNAs, novel transcripts, novel CDS, ncRNA, operons, untranslated regions)

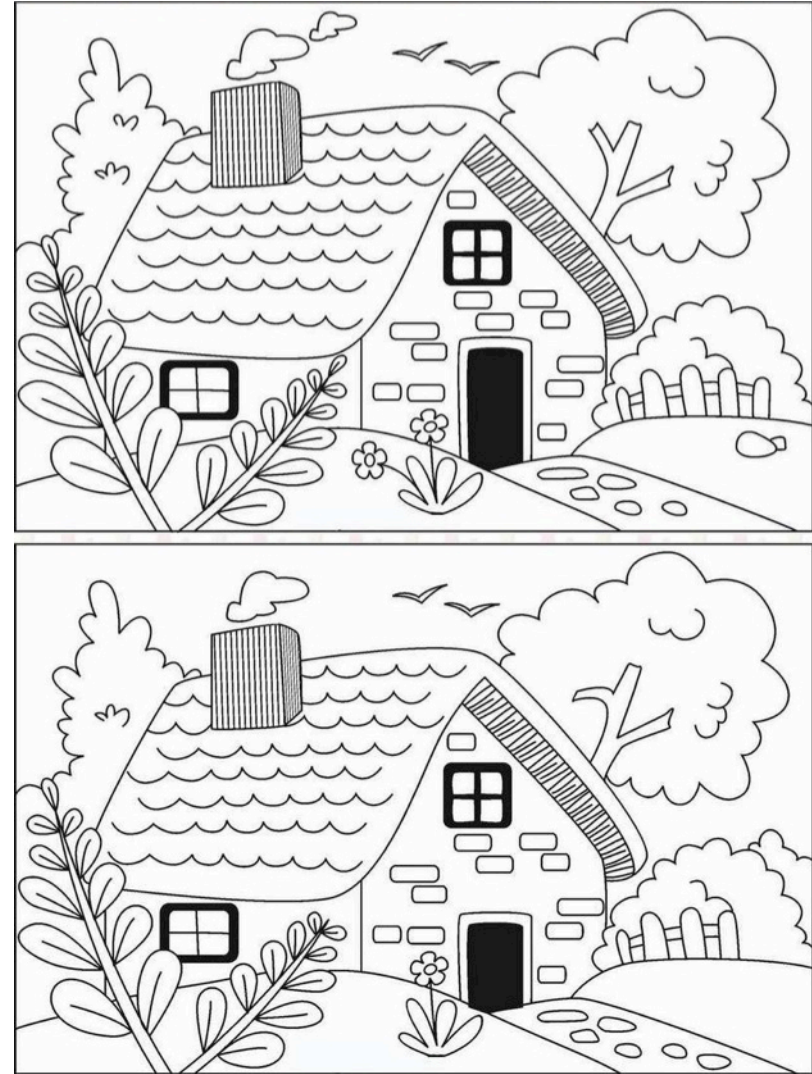
3. RNA-Seq Process Overview



3. Differential Gene Expression

A hard "spot the differences" game

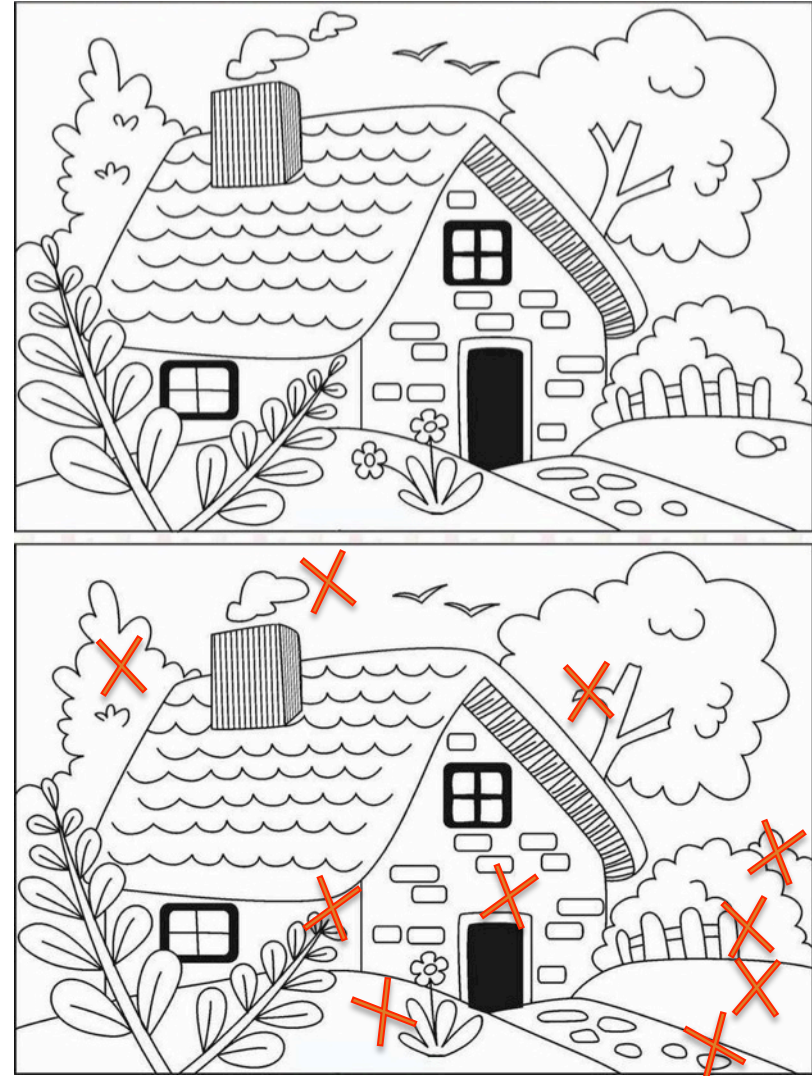
- The question we are trying to answer is:
Which genes are being expressed at different levels in different conditions ?
- Conditions can be environment, time, genetic background, etc.



3. Differential Gene Expression

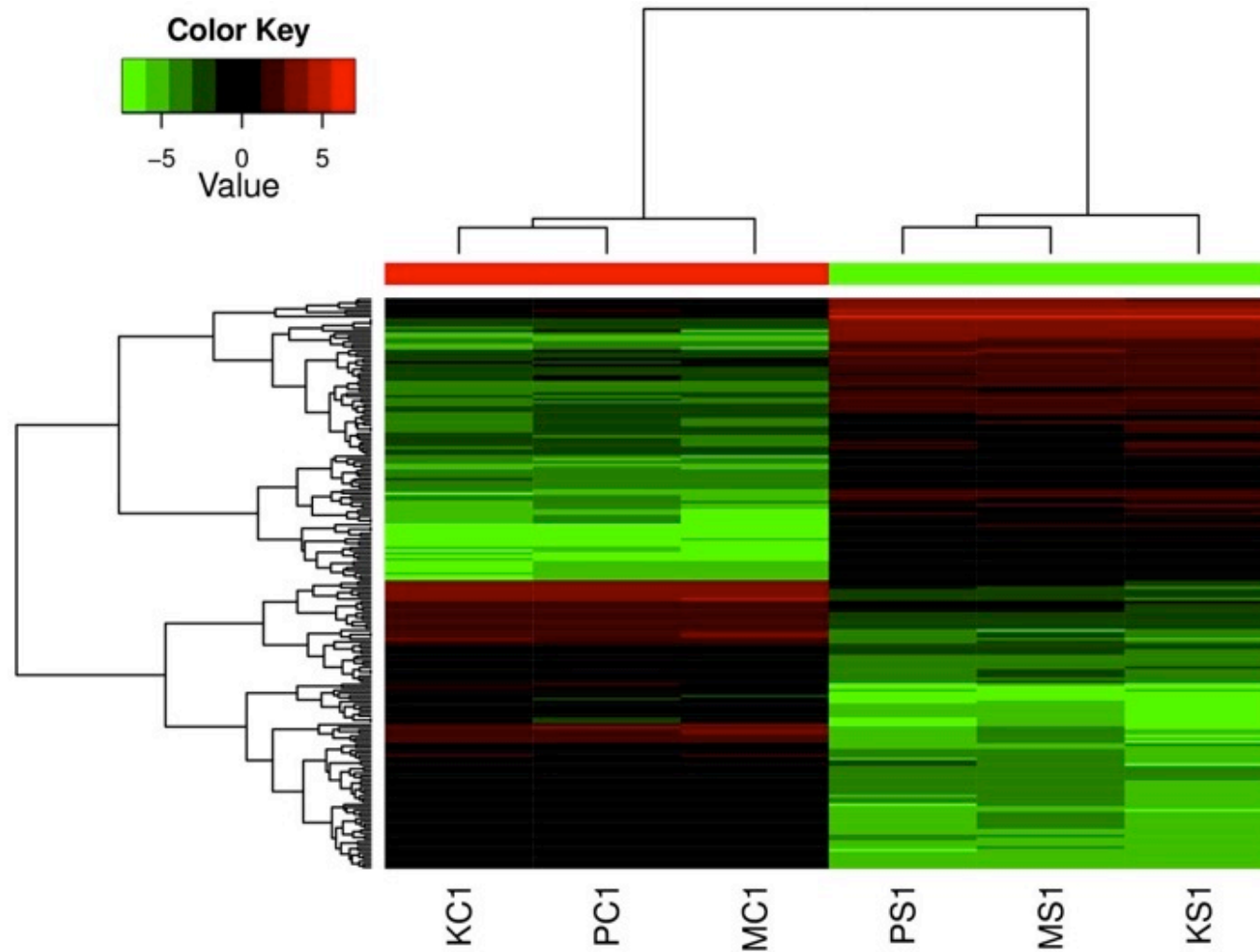
A hard "spot the differences" game

- The question we are trying to answer is:
Which genes are being expressed at different levels in different conditions ?
- Conditions can be environment, time, genetic background, etc.



3. An Example of Actual Results

Adapted from: Thumma *et al.*, BMC Genomics 2012, 13:364



Deep Sequencing Data Analysis

Computational challenges

- Very large amount of data
- Relatively new algorithms, quickly changing and evolving
- Conflicting software stacks required by analysis pipeline
- Highly influenced by storage performances (I/O throughput)
- Biological samples, data and experiment management
- Browsing and visualization of results

Ideas and Opportunities

- **New Discoveries on How to Prevent and Cure Diseases**
- Foster Interdisciplinary Collaborations
- New Innovative Computing Technologies
- HPC / HTC as an Everybody's Tool
- Applied Computer Science Research
- Environment-friendly Computing
- Help a Sustainable Economy



Summary



- Biotech innovation is shaping scientific computing infrastructures and organizations
- Computational Biology and Life Sciences in general are data-intensive research areas undergoing a “data deluge”
- Openness, scalability, flexibility and collaborative efforts are the keys to sustain innovation
- Learn from experience and look for creative research to the ultimate goal of extracting information and building knowledge



Data Science Unit

*Alessandra Anemona, Lorenzo Argante, Riccardo Beltrami,
Sonia Budroni, Andrea Cavallone, Anna Dari, Filippo Fiorini,
Nicola Pacchiani, Roberto Palmas*

THANKS !

riccardo.beltrami @ novartis.com