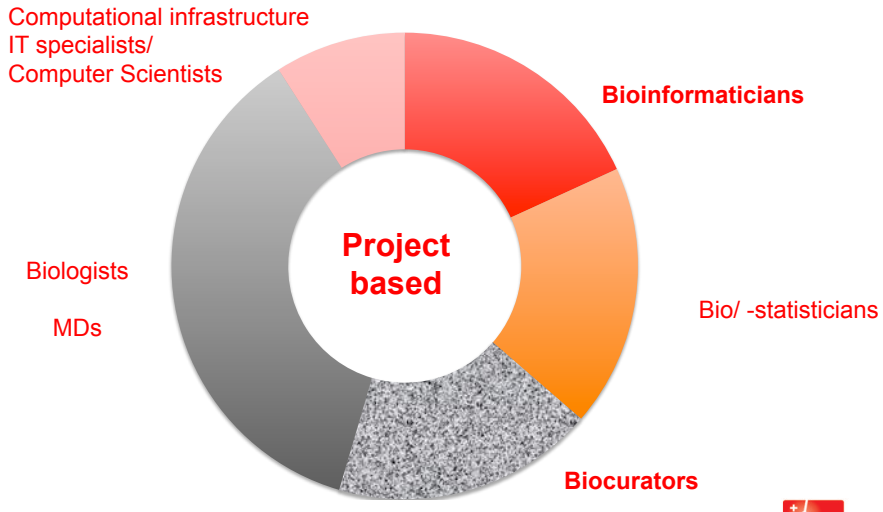
Genome, Genomics and Metagenomics

Sequencing the Universe

Swiss Institute of Bioinformatics

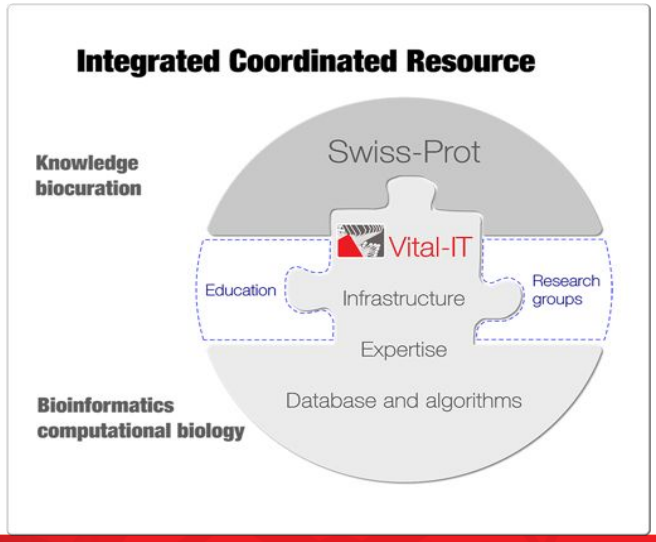
Dr. Nicolas Guex on behalf of Prof. Ioannis Xenarios
CERN, 28 May 2013 from 11:40 - 12:00

Most projects require a wide range of expertise



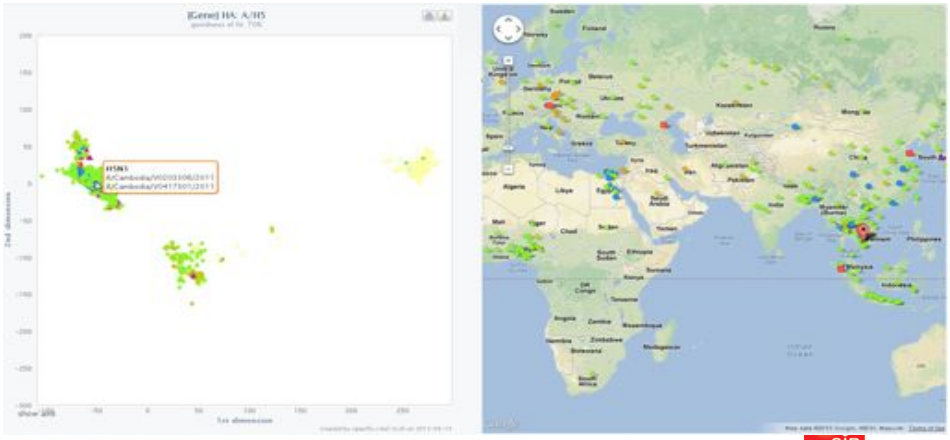
Swiss Institute of Bioinformatics

Prof. Ioannis Xenarios Ph.D.
Director Swiss-Prot/Vital-IT
Swiss Institute of Bioinformatics
Geneve-Lausanne, Switzerland



OpenFlu: large scale collection and analysis of *Influenza* genetic data

- Genetic landscape of H5Nx and cross-infection to humans



Avalanche of data

Every second a sequence informations (nucleotide) is deposited in one of the three worldwide nucleotide database (INSDC:Genbank-DDBJ-EMBL)

Every five minute a gene expression data (microarray) chip is deposited in one of the two worldwide databases (GEO-ArrayExpress)

Every day a new genome sequence is deposited (from bacteria mostly but more and more from higher eukaryotes) (Genbank-DDBJ-EMBL, ENA)

Every week a protein structure is deposited either originating from a structural biology project or from medium throughput crystallography project (PDB-RCSB)

Every month a new human genome is sequenced
In 2011, 1024 genomes completed…

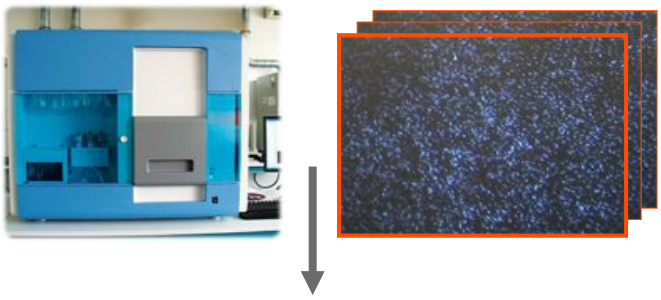
In 2012…1 genome / week is produced; over 1000 exomes on Vital-IT

In 2013…8 genomes/ week.
over 3000 exomes and 150 complete genomes
on Vital-IT/Lausanne

SIB Swiss Institute of Bioinformatics

Next Generation Sequencing redefine Genomics (again)

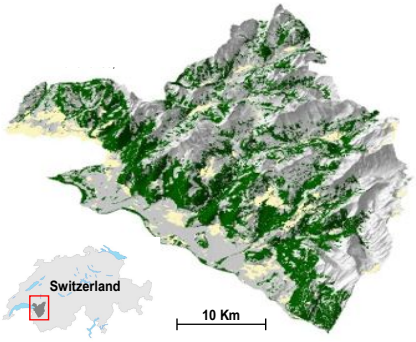
Ultra High Throughput Sequencing (UHTS)
Solexa or other…



2007 – **1 Terabytes** of raw and processed data
2008 – **2.5 Terabytes** of raw and processed data
2009 – **7 Terabytes** of raw and processed data
2012 – **14 Terabytes** of raw data

SIB Swiss Institute of Bioinformatics

Characterization of the fungal biodiversity in the grassland soils of the western Swiss Alps

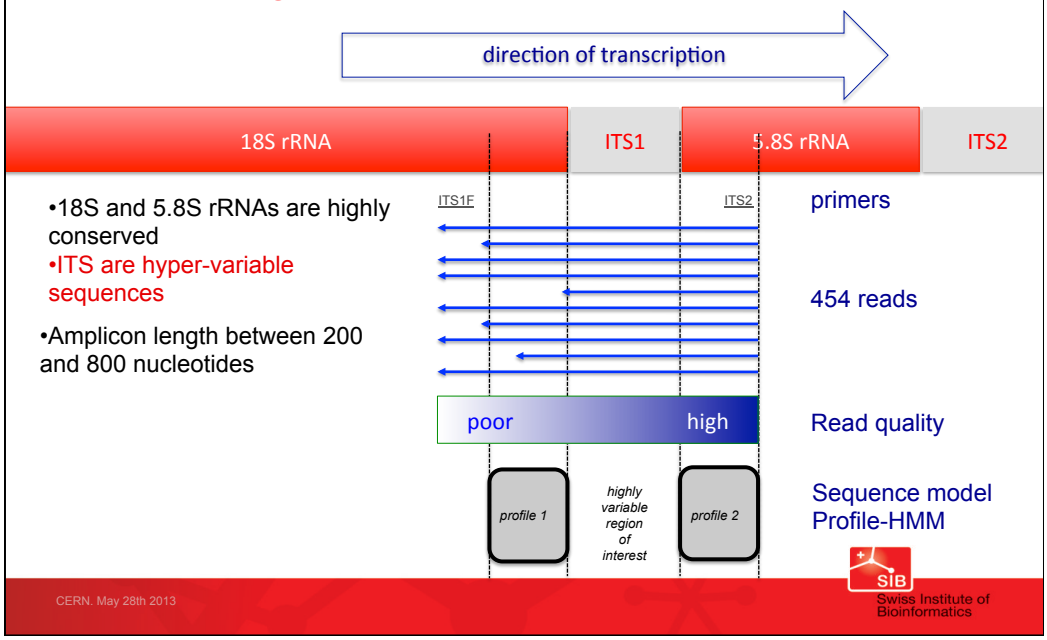


- 205 sampling sites, 4m^2 each
- DNA barcoding using Roche 454 pyro-sequencing
- Using pan-fungal primers to get a global picture
- How biotic and abiotic factors shape the distribution of the fungal biodiversity?

UNIL | Université de Lausanne
Prof. Jérôme Goudet
Prof. Antoine Guisan
Prof. Ian Sanders
Faculté de biologie et de médecine
Department of Ecology and Evolution

SIB Swiss Institute of Bioinformatics

Sequencing ITS1 with Roche 454



- 18S and 5.8S rRNAs are highly conserved
- ITS are hyper-variable sequences
- Amplicon length between 200 and 800 nucleotides

SIB Swiss Institute of Bioinformatics

A big dataset from Alpine soil samples

Number of reads / sequences	Alpine soils Roche454 reads	EMBL FUN + PLN + INV + ENV
Total	1'649'376	NA
After profiles extraction	1'152'122	127'577
After profiles extraction, non redundant	341'585	72'501

- We were looking for a method to recognize/define Operational Taxonomic Units (OTUs) that is
  - unsupervised i.e. independent of any pre-existing taxonomy
  - robust with respect to the sequencing errors that exists in pyro-sequences
- There are just a few algorithms that can perform the task of clustering a million of sequences in a reasonable delay

CD-hit-454

- It uses a fast sequence comparison heuristic based on k-mers
- It produces spherical clusters with a center (a "seed" sequence) and a maximal radius (user supplied)

Otupipe

- It uses an even faster sequence comparison heuristic
- It uses a different selection of the seed sequences

ESPRIT-Tree

- Fast heuristics with a metrics related to the Jaccard distance with the sequences represented as vectors of k-mers
- Agglomerative hierarchical clustering
- Clusters are more or less spherical with variable radius

Our method: DBC454

- Extremely fast and fully parallelized implementation of a density-based clustering algorithm (explained following slides)
- Input: data points must be defined in a Euclididan space with *n*-dimensions, *n*

$$dist(\vec{A}, \vec{B}) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

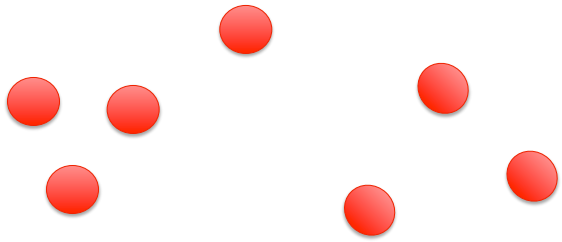
- How to represent a sequence as a point in an Euclidian space?
  - Counts of the dinucleotides XY (n = 16)
  - Counts of the dinucleotides X?Y (n =16)
  - Counts of the dinucleotides X??Y (n = 16)
- Total: 48 dimensions (16 XY + 16 X?Y + 16 X??Y)

Density-based hierarchical clustering

- Compute all pairwise distances and retain only those that are equal or smaller than a given distance threshold *T*
- A cluster is formed by single linkage and retained if it contains at least *N* points. Clusters too small are ignored
- Hierarchical clustering is obtained by repeating the clustering for increasing distance thresholds

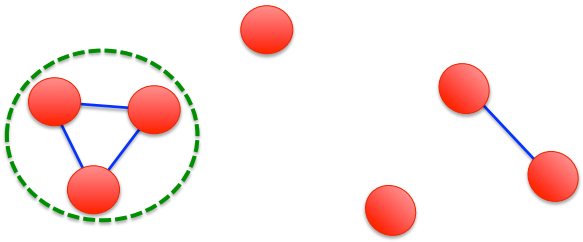
T from T_{min} to T_{max} in *s* steps

Density-based hierarchical clustering



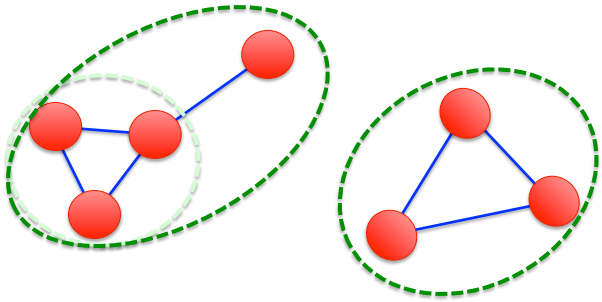
Density-based hierarchical clustering

- $T \leq 1$ and $N \geq 3$



Density-based hierarchical clustering

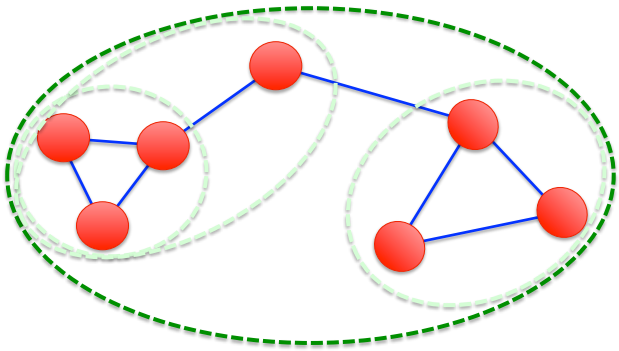
- $T \leq 2$ and $N \geq 3$



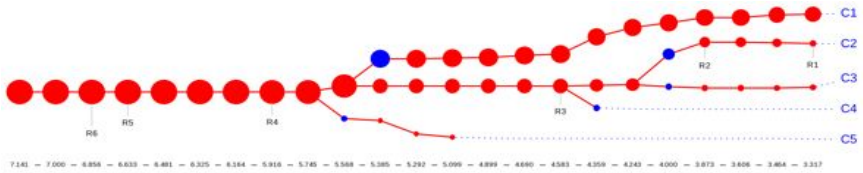
Density-based hierarchical clustering

- $T \leq 3$ and $N \geq 3$

Here, there is no such thing as a cluster center or a maximal cluster radius



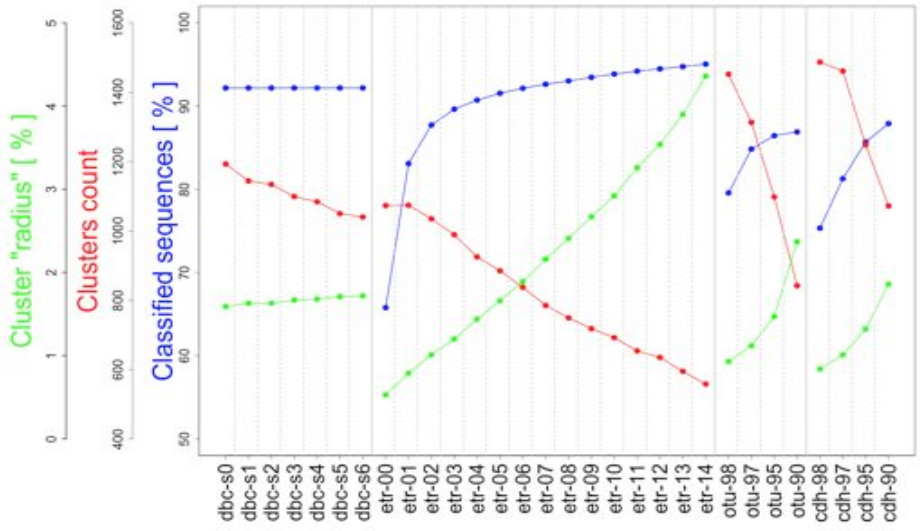
Assignment to the final partition



- At the end of the hierarchical clustering, seed clusters are determined and the points are re-attributed to the closest seed
- Optionally, reference sequences can be attributed to the closest seed lusters

SIB Swiss Institute of Bioinformatics

Algorithms comparison ($N \geq 100$)



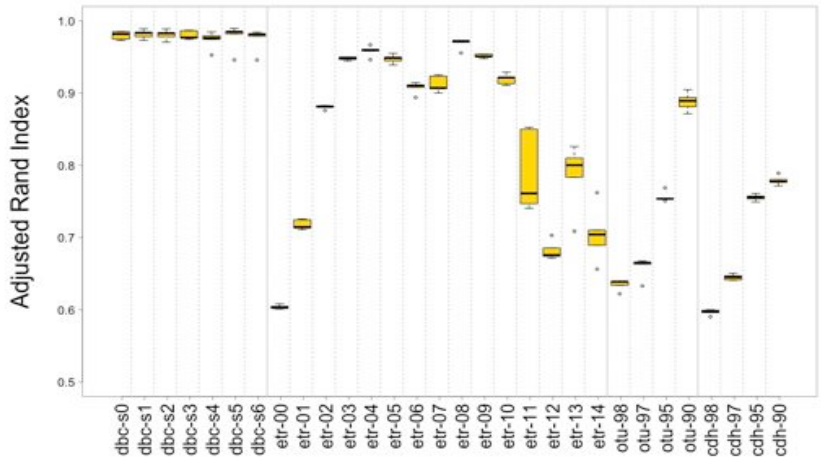
SIB Swiss Institute of Bioinformatics

Five datasets with added errors

- Five mutated dataset were created to simulate the errors that occur during the 454 sequencing:
  - 18% of the original sequences were randomly selected
    - 61% were given a single error
    - 17% two errors
    - 22% three errors.
  - The types of errors introduced were
    - 46% insertions,
    - 33 % deletions
    - 21% substitutions

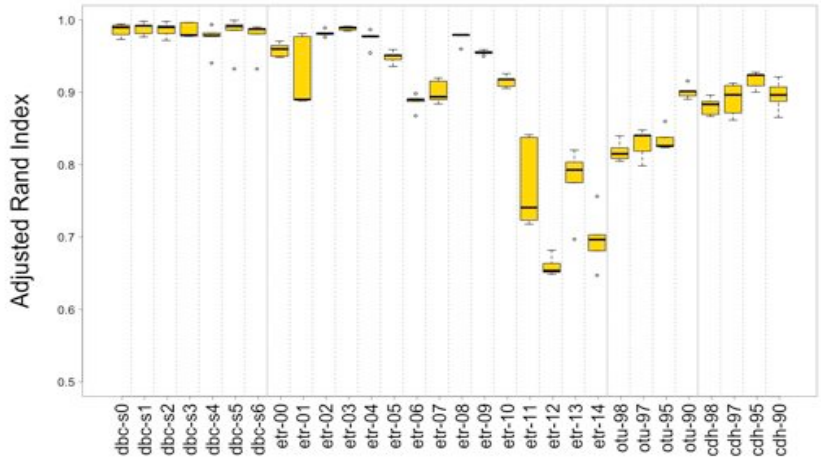
Robustness I: orginal *versus* mutated dataset

Noise was pooled as a single additional cluster



Robustness II: orginal *versus* mutated dataset

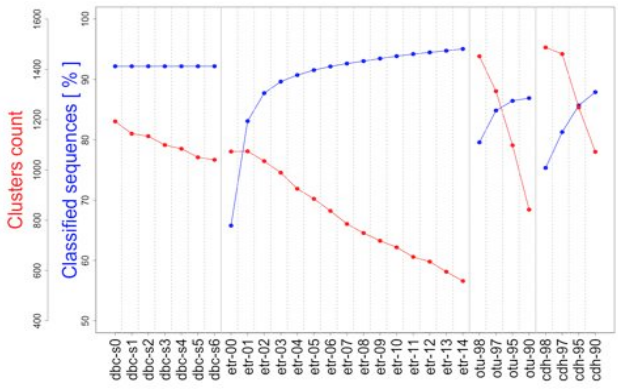
Sequences appearing in the noise were removed



SIB Swiss Institute of Bioinformatics

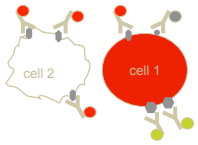
Conclusion

- Our algorithm is more robust to sequencing errors, and classifies more sequences than methods using fast heuristics

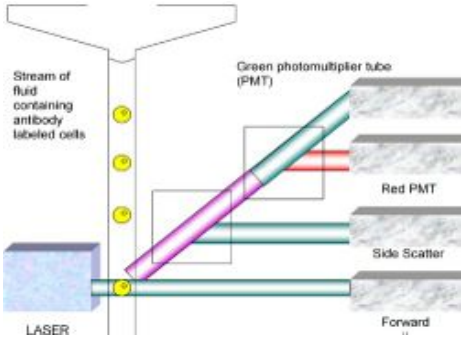


SIB Swiss Institute of Bioinformatics

Sequences are not the only ones to be sorted



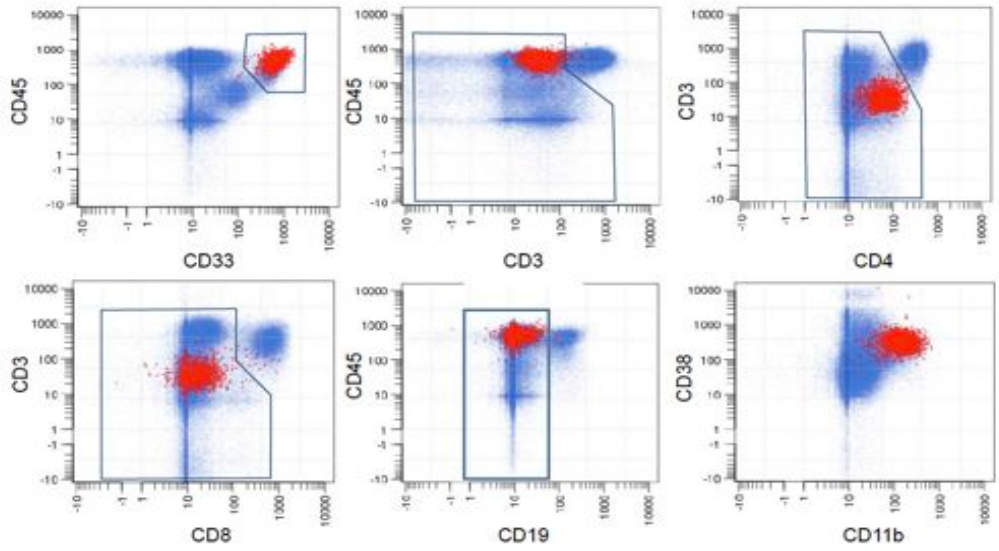
Fluorescent Activated Cell Sorting (FACS)



- Instruments can record >10 channels (colors)
- Measurements collected for 10^5 - 10^7 cells
- Analysis of cell populations is tedious and involves manual steps (biaised)

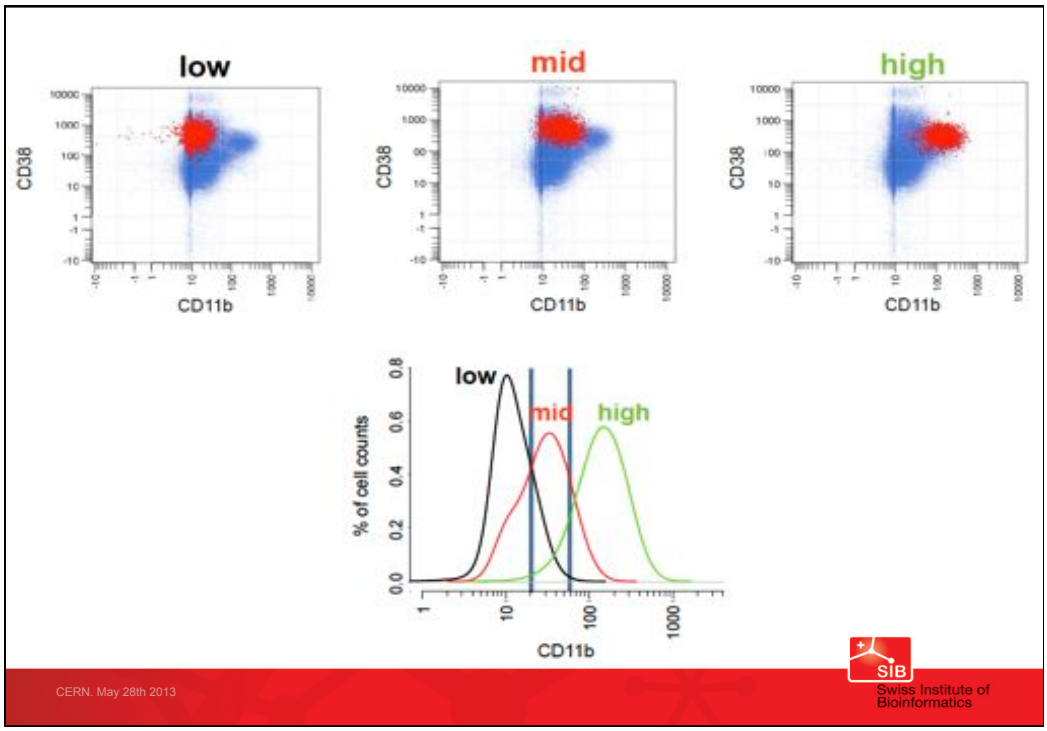
	FSC	SSC	color1	color2	color3
cell 1	100	100	100	100	200
cell 2	100	500	300	0	0

8 markers: CD45, CD33, CD3, CD4, CD8, CD19, CD38, CD11b

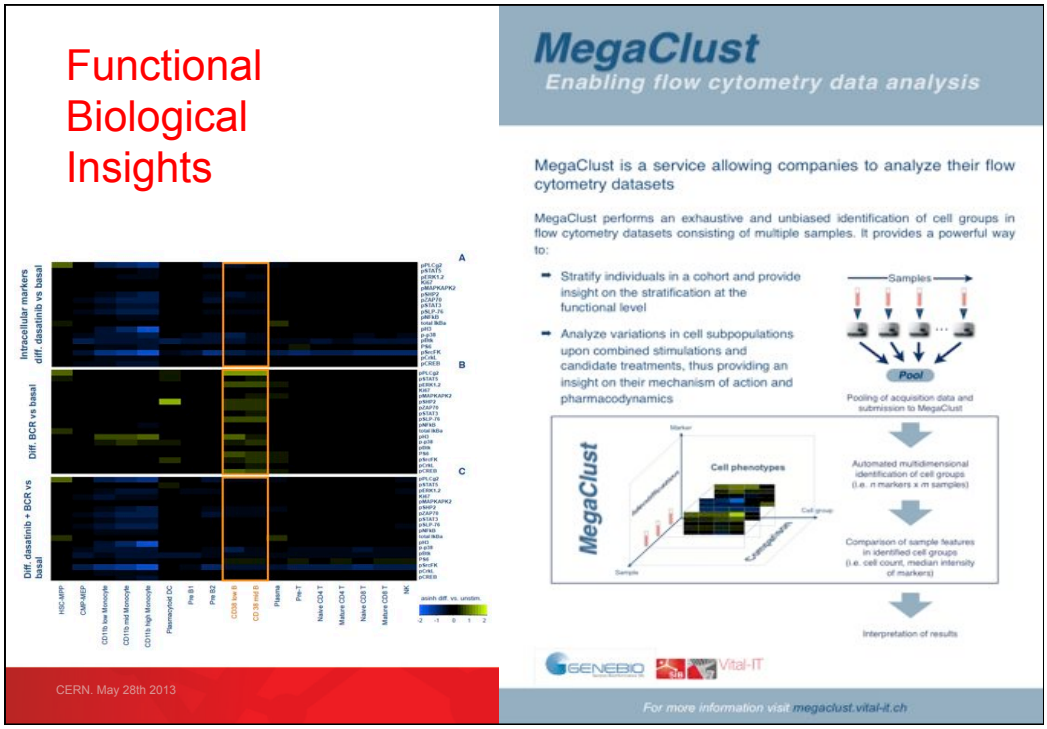


CERN. May 28th 2013

Swiss Institute of Bioinformatics

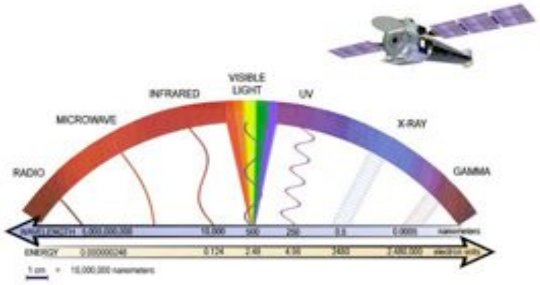


Functional Biological Insights



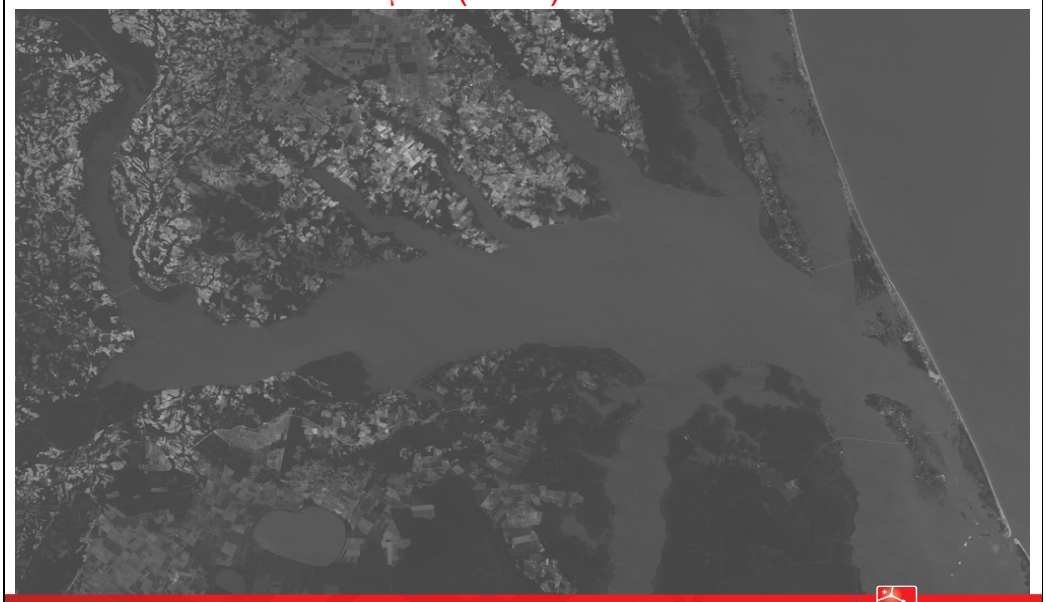
Applications to other large multidimensional datasets

- Small number of parameters measured
  - multi spectral data (various wavelengths, such as IR, visible, X-ray, ...)



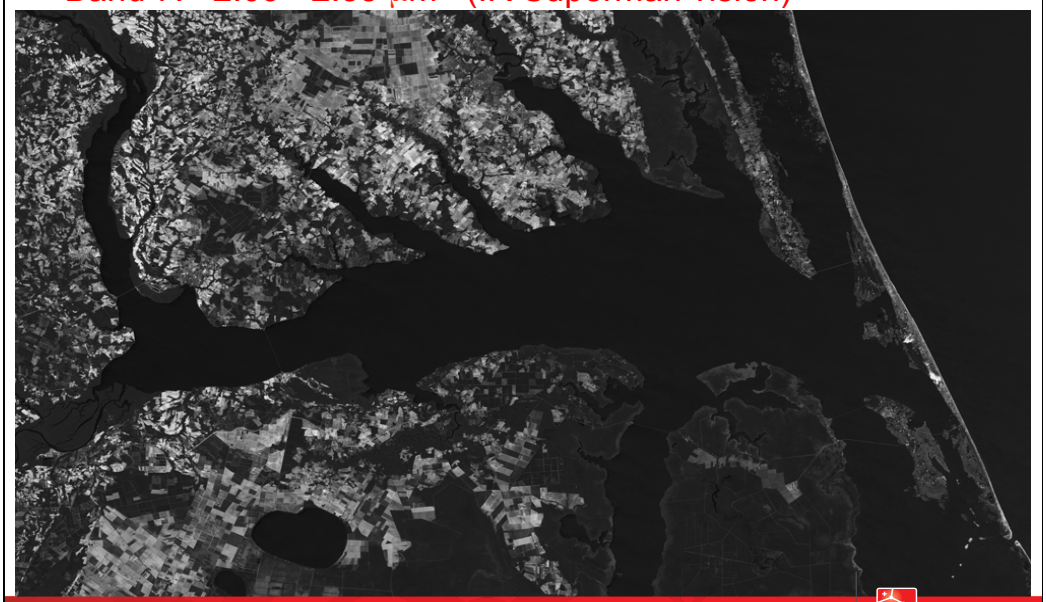
- Large number of measurements
  - pixels !

Band 1: 0.45 - 0.52 μm (violet)

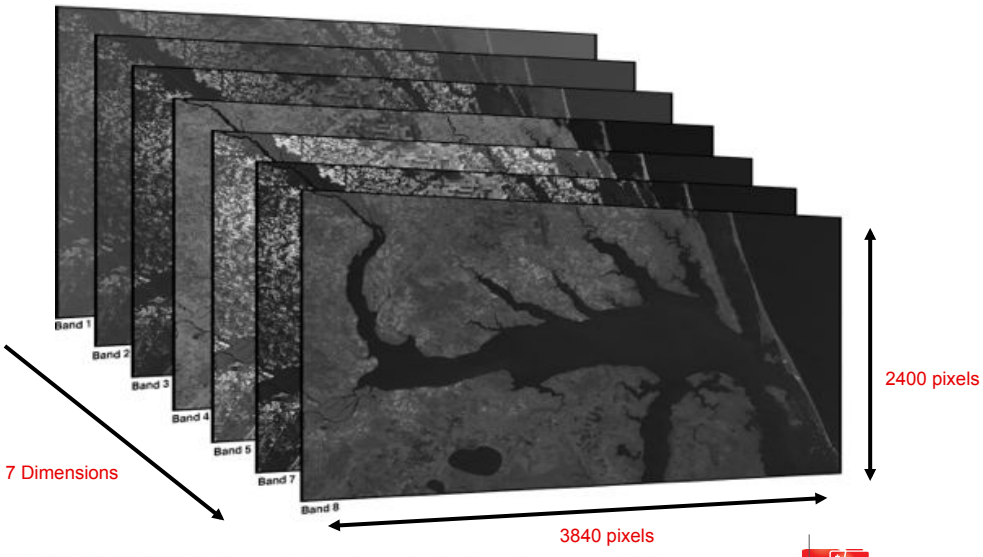


Band 8: 0.53 - 0.90 μm (green to red)

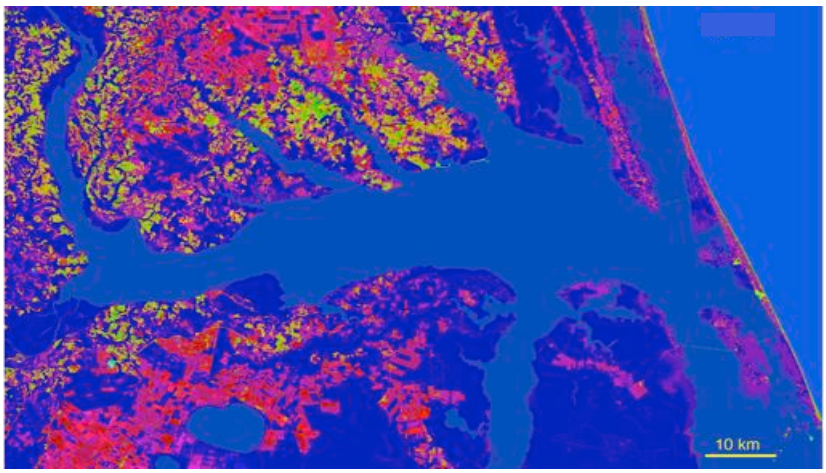
Band 7: 2.09 - 2.35 μm (IR Superman vision)



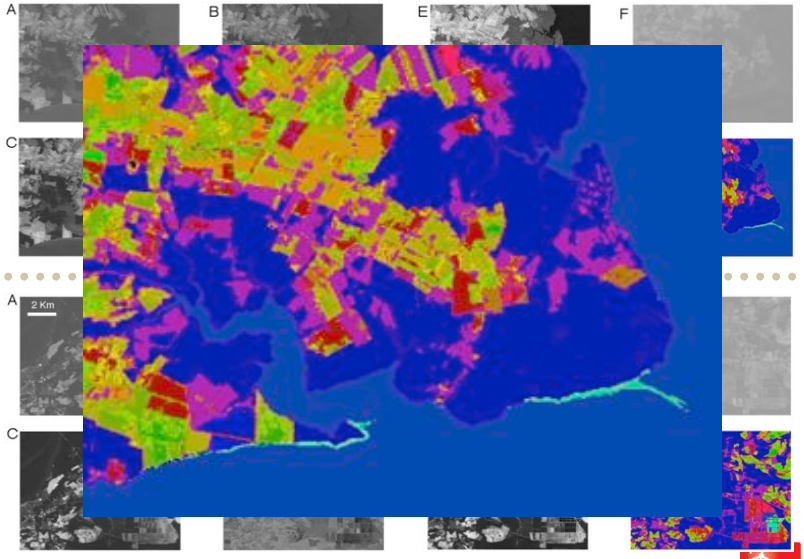
Multispectral image



Multispectral image Clustering Results



Enlargements



Funding sources

SIB

NATIONAL INSTITUTES OF HEALTH

SIXTH FRAMEWORK PROGRAMME

SEVENTH FRAMEWORK PROGRAMME

UNIVERSITÄT BERN

UNIVERSITAS FRIBURGENSIS

UNIVERSITÉ DE GENÈVE

UNIL | Université de Lausanne

SystemsX.ch
The Swiss Initiative in Systems Biology

FNSNF
Swiss National Science Foundation

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

imi
Innovative Medicines Initiative

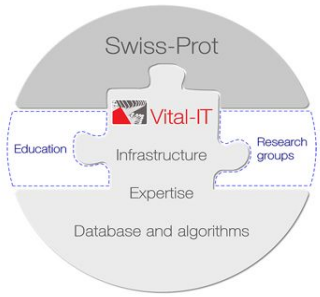
imidia
European combined excellence in diabetes research

oncosuisse

SIB
Swiss Institute of Bioinformatics

Acknowledgments

- Ioannis Xenarios
- Marco Pagni
- Dimitri Kuztenov
- Robin Liechti
- Nicolas Budin
- Roberto Fabbretti
- Volker Vlegel
- Thierry Schuepbach
- the Vital-IT & SWISSPROT teams
- Marie-Agnes Doucey
- Hélène Niculita-Hirzel
- Loïc Pellissier
- Anne Dubuis
- Antoine Guisan
- Ian R. Sanders
- Jérôme Goudet



FNSNF
FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE

UNIL | Université de Lausanne
Faculté de biologie et de médecine
Department of Ecology and Evolution

SIB Swiss Institute of Bioinformatics