Challenges in precision medicine: From sequencing to big data processing

Mikel Hernaez

CARL R. WOESE **INSTITUTE FOR GENOMIC BIOLOGY**
Where Science Meets Society

Torino, July 19th 2017

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

"Tonight I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes"

President Obama, 2015

Precision Medicine

- This growing field will revolutionize how we treat disease, taking into account the individual’s unique genetic makeup, environment, and lifestyle.

- Hospitals around the world are working hard to embrace the personalized medicine paradigm.

Precision Medicine: Technology-enabled medicine

Today

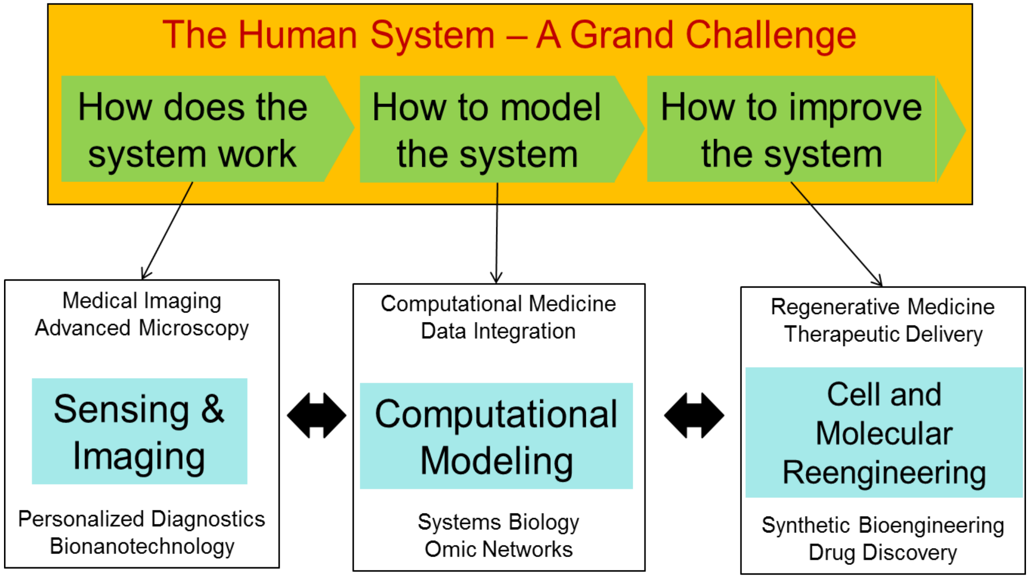
Disease Progression → EXAMINE, DIAGNOSE, TREAT

Future

Health & Wellness, Disease Etiology

↕

SYSTEMS OR PRECISION MEDICINE



The Human System – A Grand Challenge

How does the system work → How to model the system → How to improve the system

Medical Imaging
Advanced Microscopy

Sensing & Imaging

Personalized Diagnostics
Bionanotechnology

Computational Medicine
Data Integration

Computational Modeling

Systems Biology
Omic Networks

Regenerative Medicine
Therapeutic Delivery

Cell and Molecular Reengineering

Synthetic Bioengineering
Drug Discovery

The Mayo-Illinois Alliance

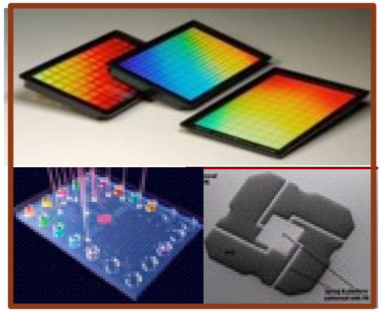
MAYO CLINIC

Top-Ranked Medical Center, World Renowned Tradition in Quality Health Care Delivery

Top-Ranked Programs in Engineering, Computation, Bioinformatics, Genomics and Nanotechnology

INFORMATION-BASED MEDICINE

GENOMICS



POINT-OF-CARE DIAGNOSTICS

The Mayo-Illinois Alliance

- The goal is to advance clinical and translational research and to harness the power of **big data** to transform precision medicine for the 21st century.

- The Alliance's interests extend to numerous areas, including:
  - High-performance computing
  - Machine learning
  - Signal processing
  - Nanotechnology

The Mayo Grand Challenge:
The 10,000 Genome Project

- The aim is to develop a pipeline capable of sequencing, analyzing, visualizing, and interpreting genomes of at least 10,000 patients per year, each within 48 hours.
- Some challenges:
  - Data Volume: ~1PB a year
  - Data Processing: GWAS analysis in less than 48 h (per patient)

Problems towards personalized medicine

- Data size:
  - New sequencing machines can generate 1 TB of data per day.
  - Institutions are shipping HDs through FedEx instead of transmitting data through the internet!
- Data analysis:
  - Interoperability across computational biology methods is not fully guaranteed, causing huge headaches to researchers and data scientists from hospitals and institutions.

Evolution of Genome Sequencing

	2009	2017
Cost/Genome	$100K	$1K
Coverage	30x	> 200x
Number of reads	1 Billion	> 6 Billion
Size of raw sequencing files	0.25 TB	> 1.5 TB

Evolution of Genome Sequencing

	2009	2017
Cost/TB	$100/TB	$50/TB
Download Speed	10Mbps	100Mbps
Cost/Genome	$100K	$1K

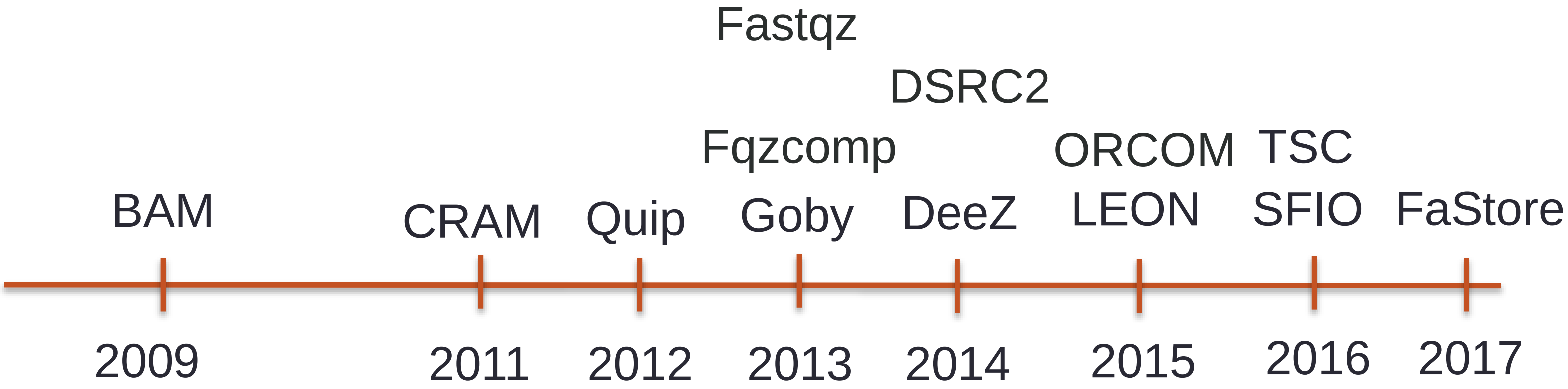
No technology is keeping with the pace of genome sequencing!

MSRP ⓘ

~~$210,[illegible]50~~

$2,500

Some proposed solutions

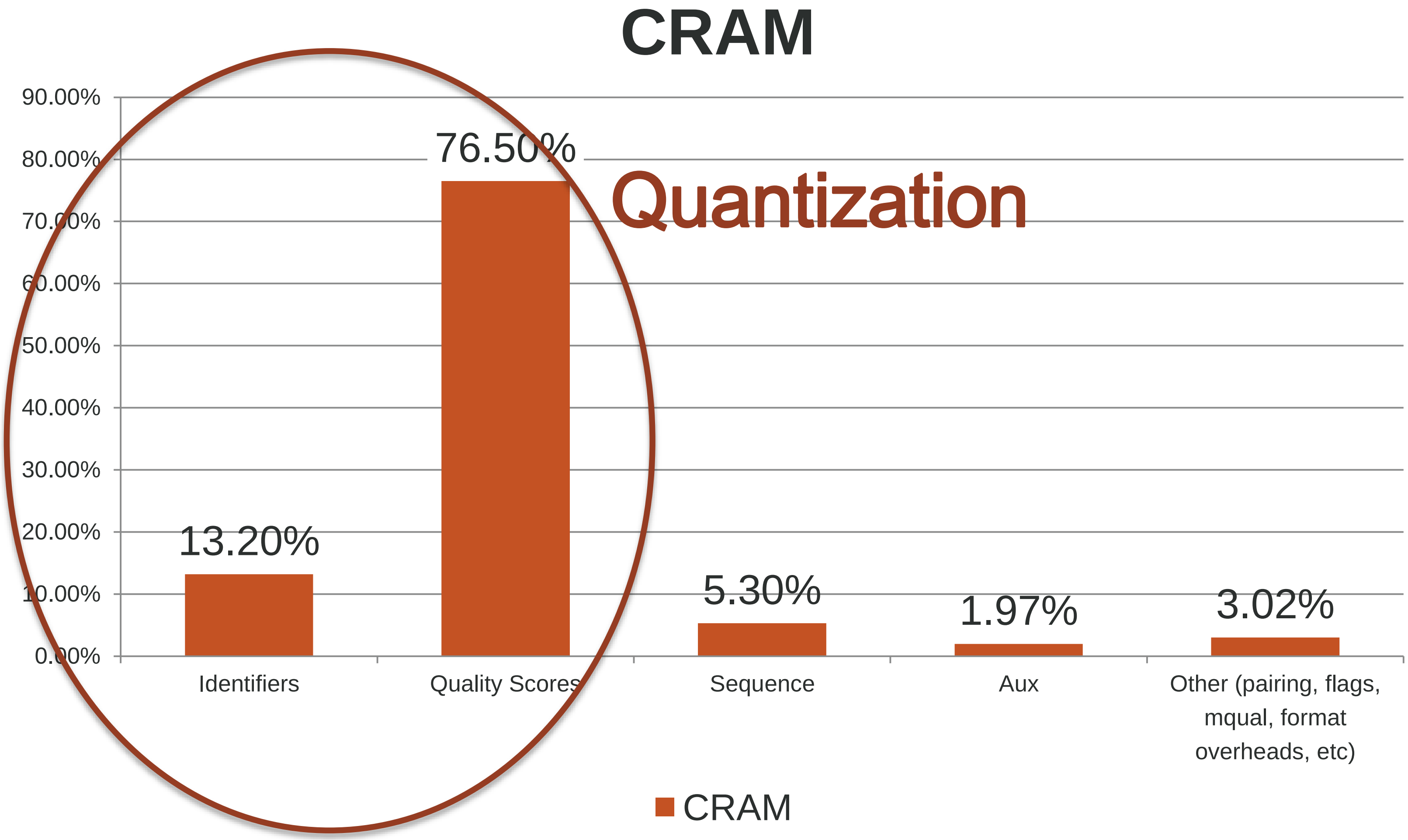


2009	2011	2012	2013	2014	2015	2016	2017
BAM	CRAM	Quip	Goby Fqzcomp Fastqz	DeeZ DSRC2	LEON ORCOM	SFIO TSC	FaStore

Contributors:

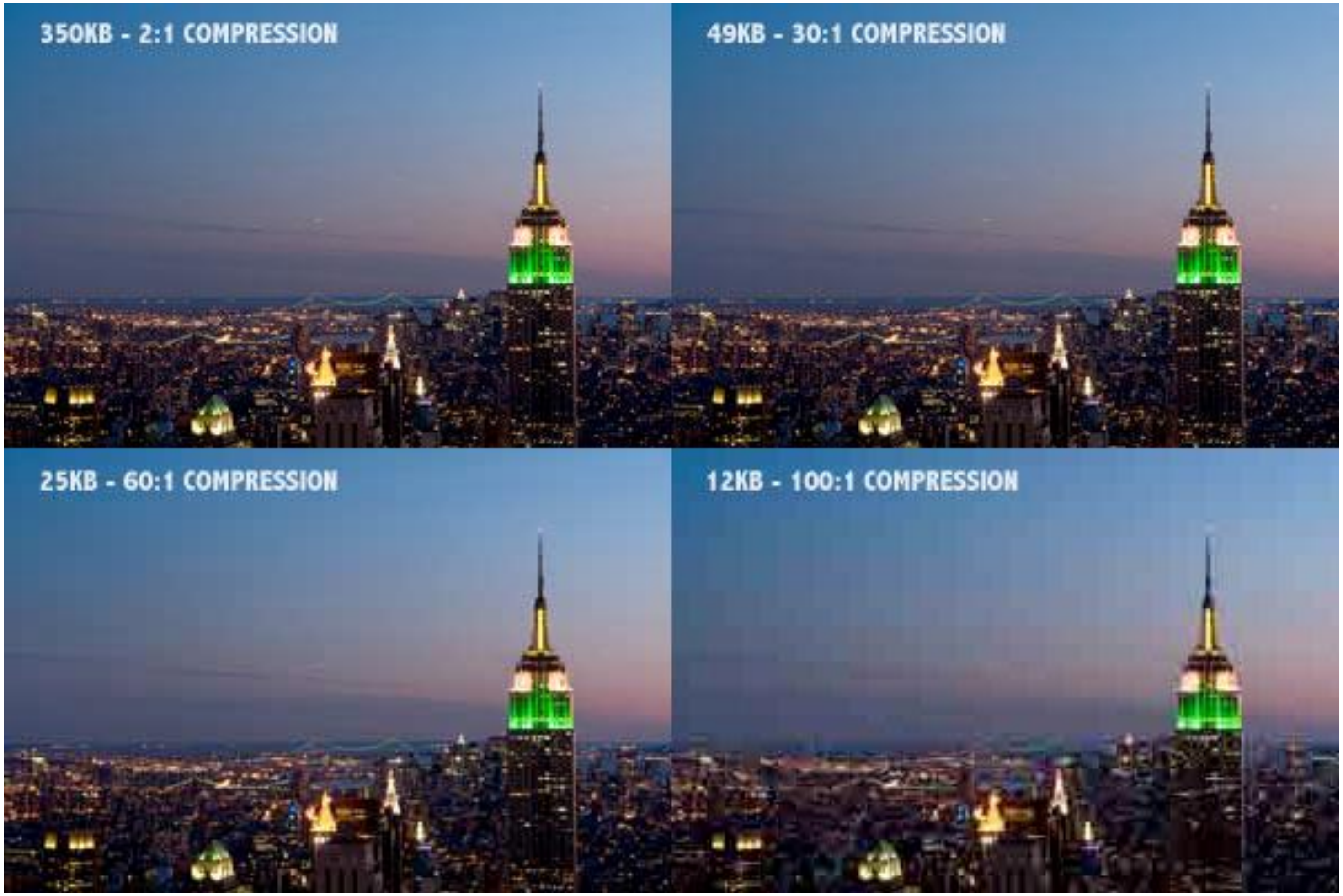
Stanford University, University of Washington, Carnegie Melon, Simon Fraser University, Cornell University, University of Hannover, European Bioinformatics Institute, Silesian University of Technology, University of Illinois, …

Weight of the lossless compressed data

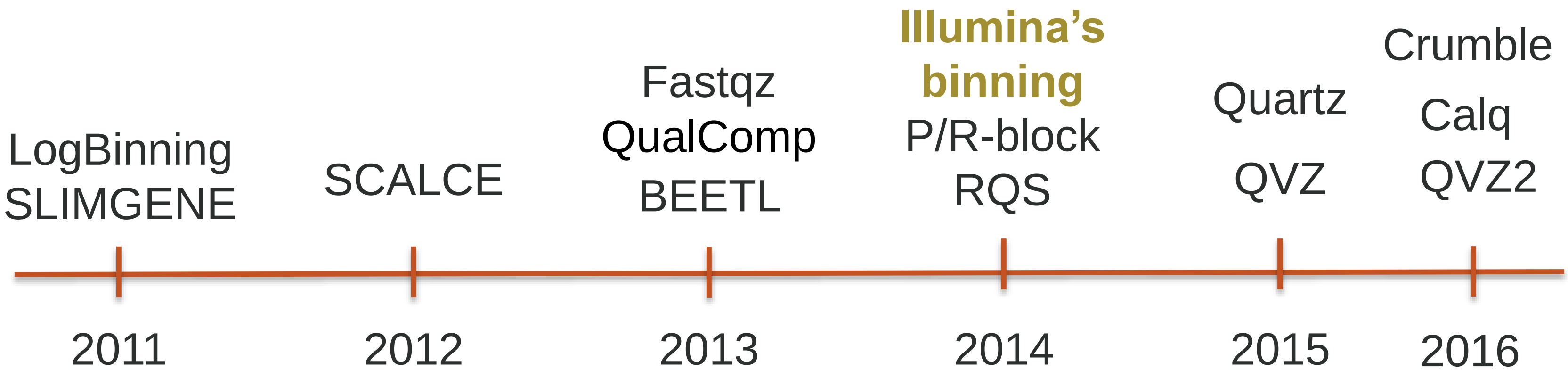


Quantization

Quantization



Quantization of quality scores



2011: LogBinning, SLIMGENE

2012: SCALCE

2013: Fastqz, QualComp, BEETL

2014: **Illumina's binning**, P/R-block, RQS

2015: Quartz, QVZ

2016: Crumble, Calq, QVZ2

Contributors:

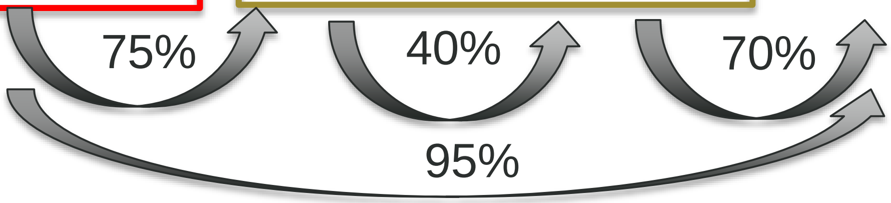
Stanford University, MIT, Simon Fraser University, University of Hannover, EBI, University of Illinois, Illumina Inc., University of Melbourne, …

Quantization of quality scores

- Order of magnitude improvements in compression.
- Several extensive analyses of its effect on variant calling and RNA-Seq gene expression:
  - Only negligible variation of results
  - Consistent improvements in variant calling also shown!

Benefits of Quantization

Human data	# Reads	SAM	BAM	CRAM	Quantize QS
Chr20 (5X)	3.4 M	800 MB	172 MB	89 MB	27 MB
Chr20 (85X)	52 M	13 GB	3.5 GB	2.4 GB	0.7GB



Current *de facto* solutions

- Raw data: GZIP format
  - GZIP is a generic file compressor
- Aligned data: BAM format
  - BAM is a binarized and GZIP-ed raw aligned data

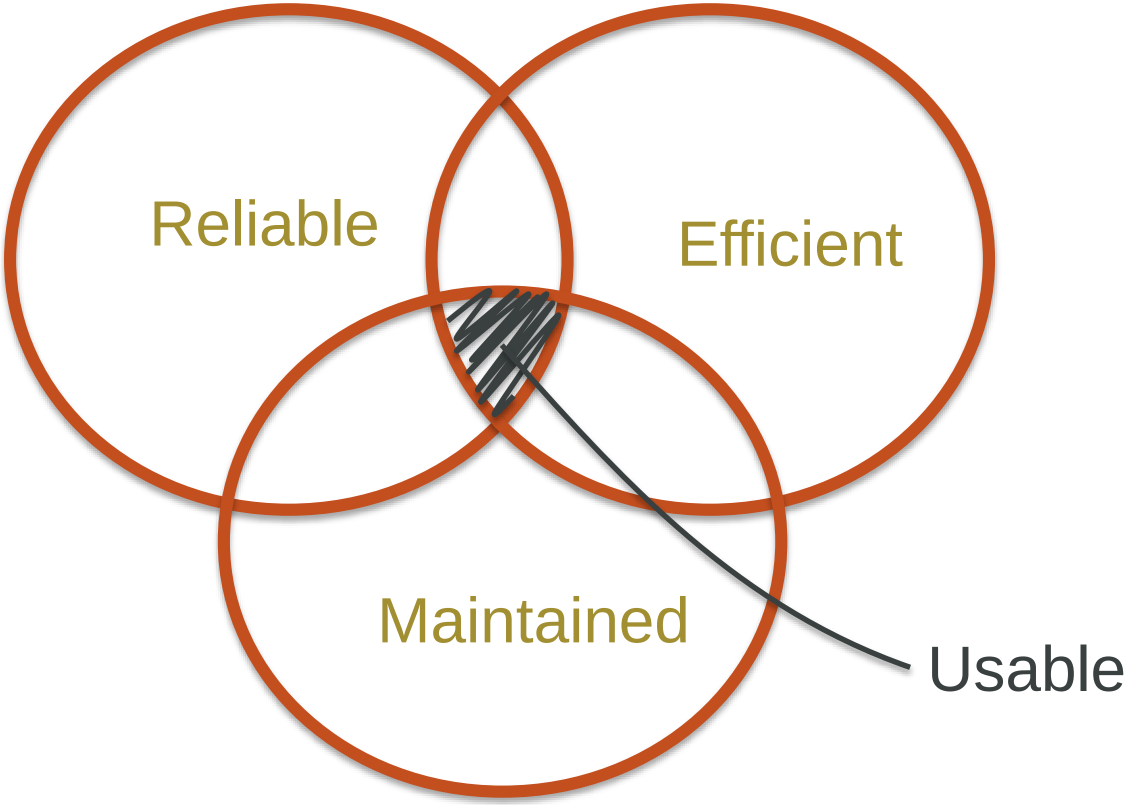
All the solutions use off-the-self general compressors!

Why are not adopted?

- Most solutions implemented by “students”
- Usually not long-term maintained
- Buggy

There is a need for a well maintained standard format!

Requirements for a good genomic data format



Requirements for a good genomic data format

- It is not only about compression:
  - Random Access over the compressed domain
  - Indexing capabilities
  - Interoperability among systems

Solutions needed by Hospitals and Institutions

- Mayo Clinic’s personalized medicine initiative is expected to generate ~ 1 PB of data per year.

- These data needs to be store, transmitted and analyzed:
  - Extremely cumbersome with current *de facto* formats !

- A well maintained data standard format is urgently needed to fully enable the personalized medicine paradigm.

Thanks!

CARL R. WOESE **INSTITUTE FOR GENOMIC BIOLOGY**

Where Science Meets Society

Torino, July 19th 2017

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN