Genome Informatics

genome.gov
National Human Genome Research Institute
National Institutes of Health

Vivien Bonazzi

Genomic data volumes

ABI 3730 30 Megabytes

Genomic data volumes

Roche 454	10 - 30 Gigabytes
Illumina -Solexa	30 - 200 Gigabytes
ABI - SOLiD	50 – 1000 Gigabytes
Heliscope	?? Terabyte range
Pacific Biosciences SMRT	?? Terabyte range

It really bytes……

Data Storage Costs

1 Base = 20 Bytes (if intensity files are kept)*

1 Base = 10 Bytes*

* ~35 bp reads, longer reads are stored more efficiently

Computational challenges

Infrastructure

- – Data Storage
- – Computing (CPU) Capacity
- - New Hardware & Software Architecture
- – Data Transfer Rates
- – Data Security

Data analysis challenges

Data analysis challenges

Developing new analysis tools

Refactoring “old” analysis tools

Optimizing analysis tools to work on new computing platforms

Visualization methods

Data analysis challenges

New and improved visualization methods

More robust analysis tools

- *non informatics specialist*

Data integration: current and new data

- *proteomics, imaging data, metadata*

Data standards

Solutions?

Data reduction of raw sequence data

- keep derived data: assemblies, SNPs etc

Actively engage the biological and computing scientific communities

- *Informatics analysis & planning workshop*
- *Cloud computing workshop*

Education

Solutions?

I THINK WE SHOULD BUILD AN SQL DATABASE.
UH-OH
S. Adams
E-mail: SCOTTADAMS@AOL.COM
DOES HE UNDERSTAND WHAT HE SAID OR IS IT SOMETHING HE SAW IN A TRADE MAGAZINE AD?
11/17
© 1995 United Feature Syndicate, Inc.(NYC)
WHAT COLOR DO YOU WANT THAT DATABASE?
I THINK MAUVE HAS THE MOST RAM.