Best Practice and Quality Issues On Large NGS Dataset Analysis

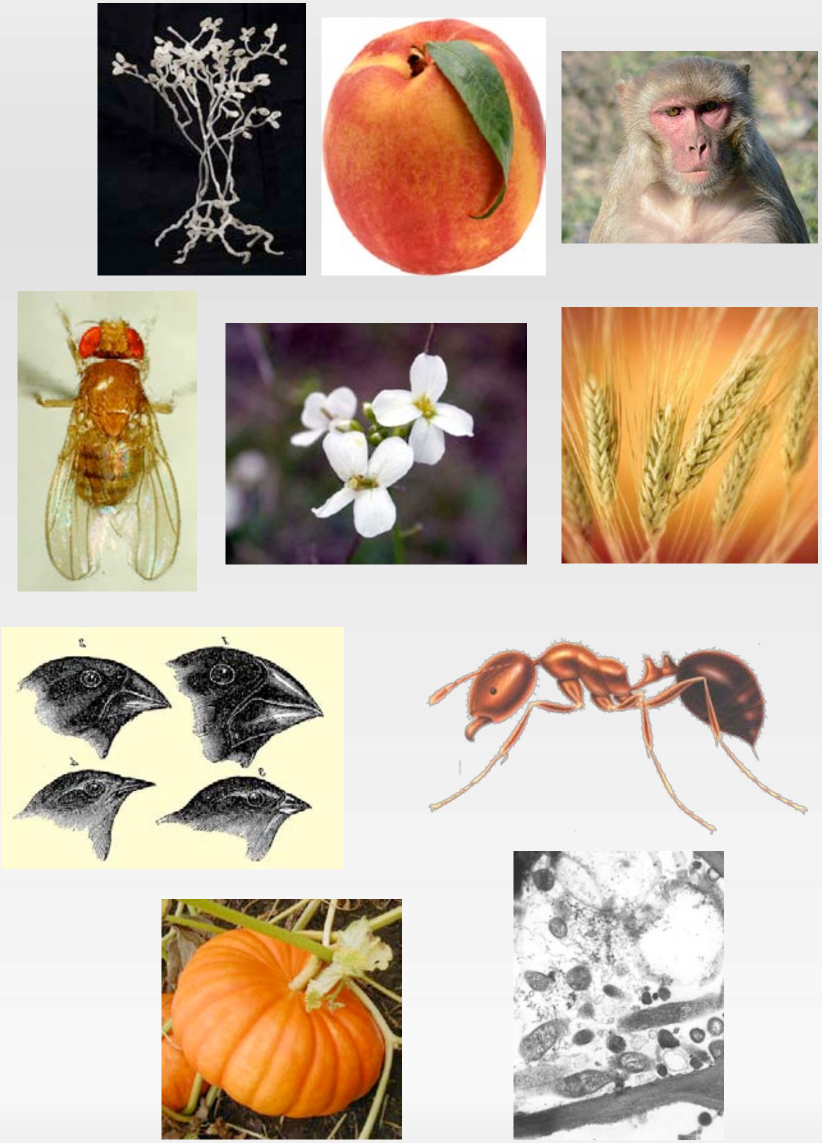
Dawei Lin, Ph.D.

Director of Bioinformatics Core
Genome Center

July 11, 2010, Boston, ISMB

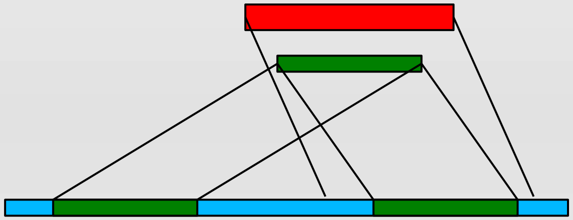
New Ideas, New People

- *De Novo* Sequencing
  - Focused sequencing
  - Large number of small genomes
- SNP Discovery
  - No reference genome sequences
- Transcriptome Profiling
  - Unknown transcriptomes
- Metagenomics
  - Special interests
- Novel Use of New Capabilities
  - High-throughput mapping



Initial Data Processing

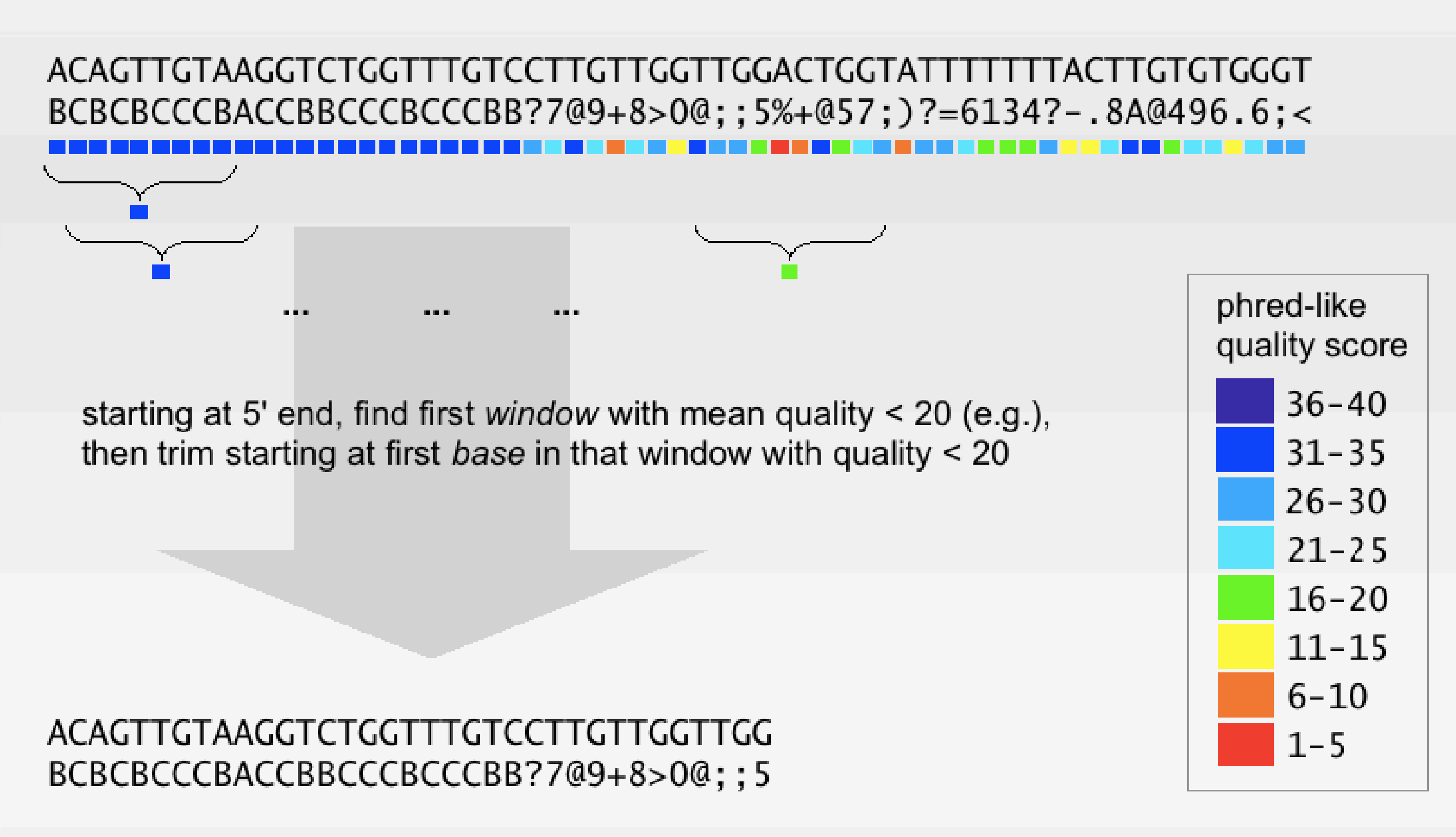
- Raw Data Examination
  - Data might not be what is specified
- Reference Source Verification
  - Version and sites
- Quick Preliminary Analysis
  - 50%-80% mapping to reference



Common Tools

- Velvet/Oases
  - Run several parameter sets and integrate the outputs
- BWA/samtools
  - Short and long reads alignment
  - .....

Quality Trimming



```
ACAGTTGTAAGGTCTGGTTTGTCCTTGTTGGTTGGACTGGTATTTTTTTACTTGTGTGGGT
BCBCBCCCBACCBBCCCBCCCBB?7@9+8>0@;;5%+@57;)?=6134?-.8A@496.6;<
```

starting at 5' end, find first *window* with mean quality < 20 (e.g.), then trim starting at first *base* in that window with quality < 20

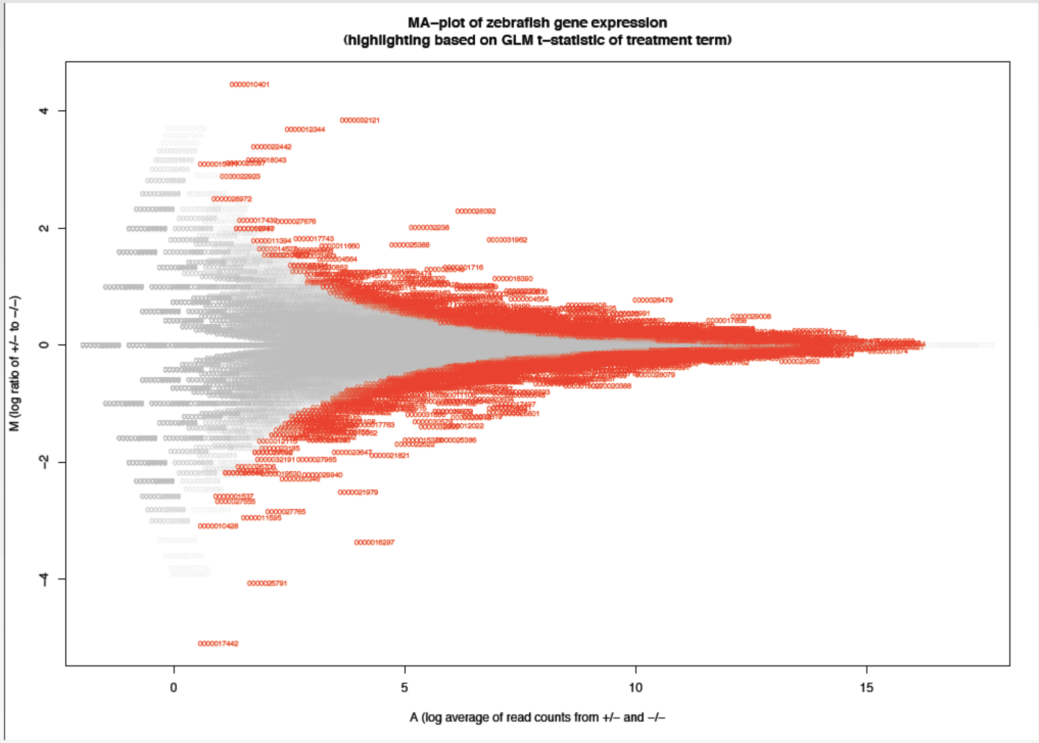
```
ACAGTTGTAAGGTCTGGTTTGTCCTTGTTGGTTGG
BCBCBCCCBACCBBCCCBCCCBB?7@9+8>0@;;5
```

phred-like quality score

🟪	36-40
🟦	31-35
🟦	26-30
🟦	21-25
🟩	16-20
🟨	11-15
🟧	6-10
🟥	1-5

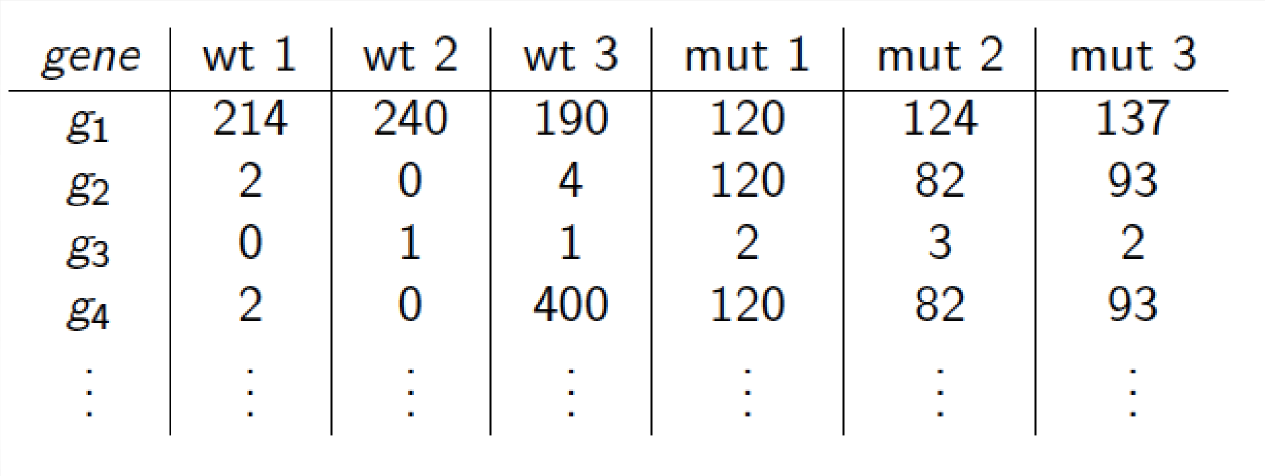
Quality Checking

- Independent or Orthogonal Validation
- Cherry Pick Cases
- Biological Significance



RNAseq Experimental Designs

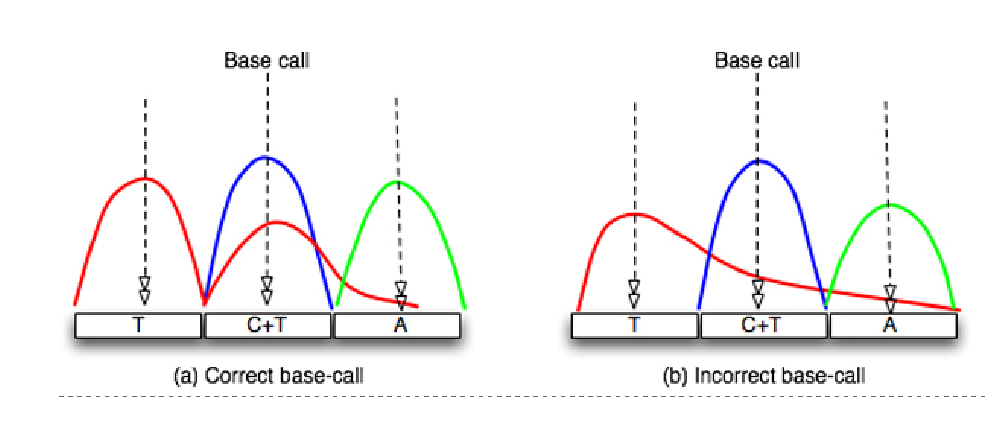
- Unreplicated Data
- Technical Replicates
- Biological Replicates
- Pooling (most popular choice)



gene	wt 1	wt 2	wt 3	mut 1	mut 2	mut 3
g_1	214	240	190	120	124	137
g_2	2	0	4	120	82	93
g_3	0	1	1	2	3	2
g_4	2	0	400	120	82	93
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Experimental Validation

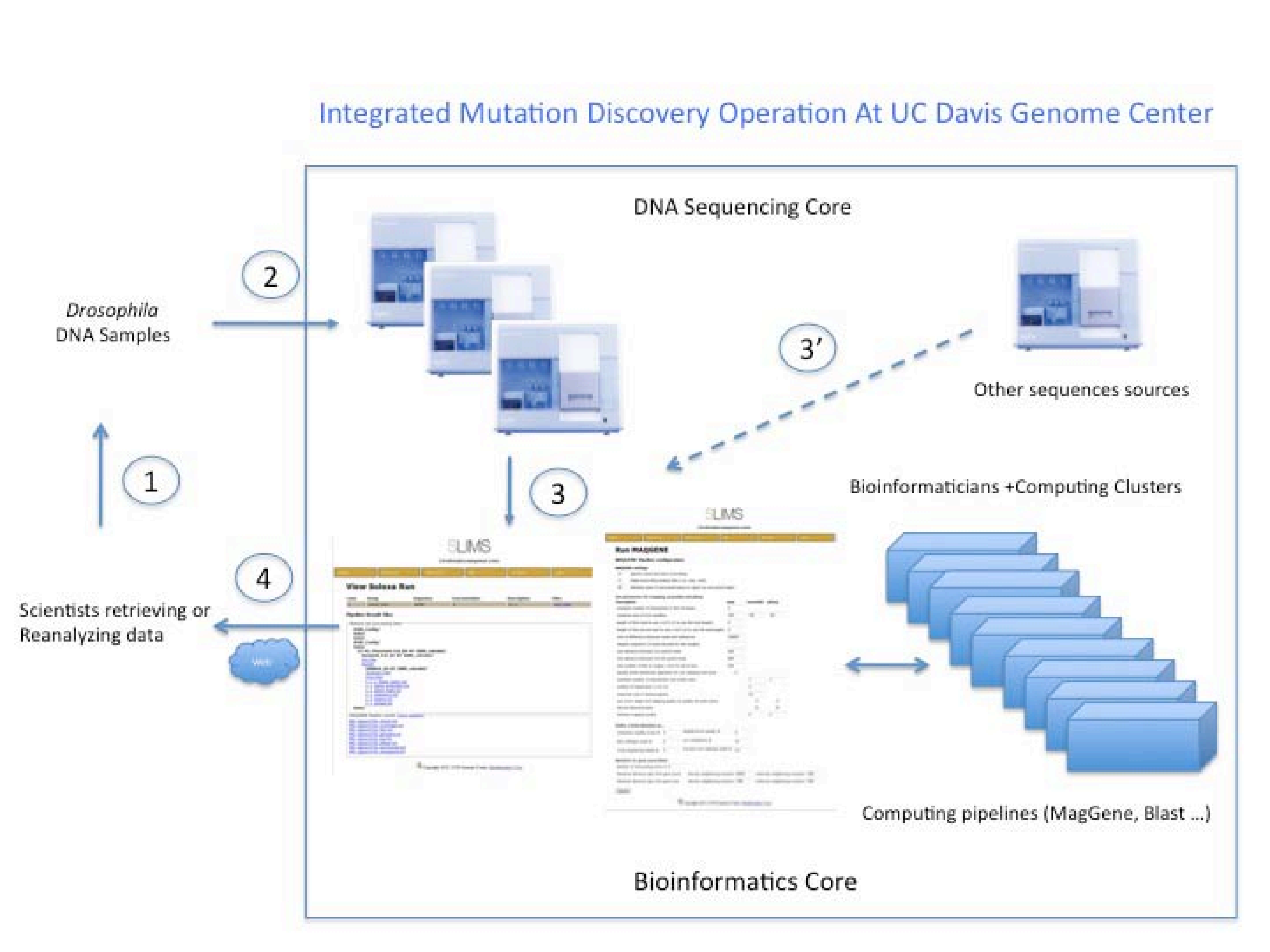
- Small Scale Test, Large scale gain
- Experimenters are not always right
  - Samples have different genetic background
  - Rare events (Dr. Joe Fass, Poster A57)



(a) Correct base-call

(b) Incorrect base-call

Mutation Discovery Pipeline


Integrated Mutation Discovery Operation At UC Davis Genome Center
DNA Sequencing Core
2
Drosophila DNA Samples
3'
Other sequences sources
1
3
Bioinformaticians +Computing Clusters
SLIMS
SLIMS
View Solexa Run
Run MAQGENE
4
Scientists retrieving or Reanalyzing data
Web
Computing pipelines (MagGene, Blast ...)
Bioinformatics Core

Worm Breeder’s Gazette, June 2010, http://bit.ly/cgnyhD

Cloud and Portable Computing

- Amazon Cloud (http://aws.amazon.com/)
- Portable Ubuntu (run Ubuntu on Windows without rebooting)

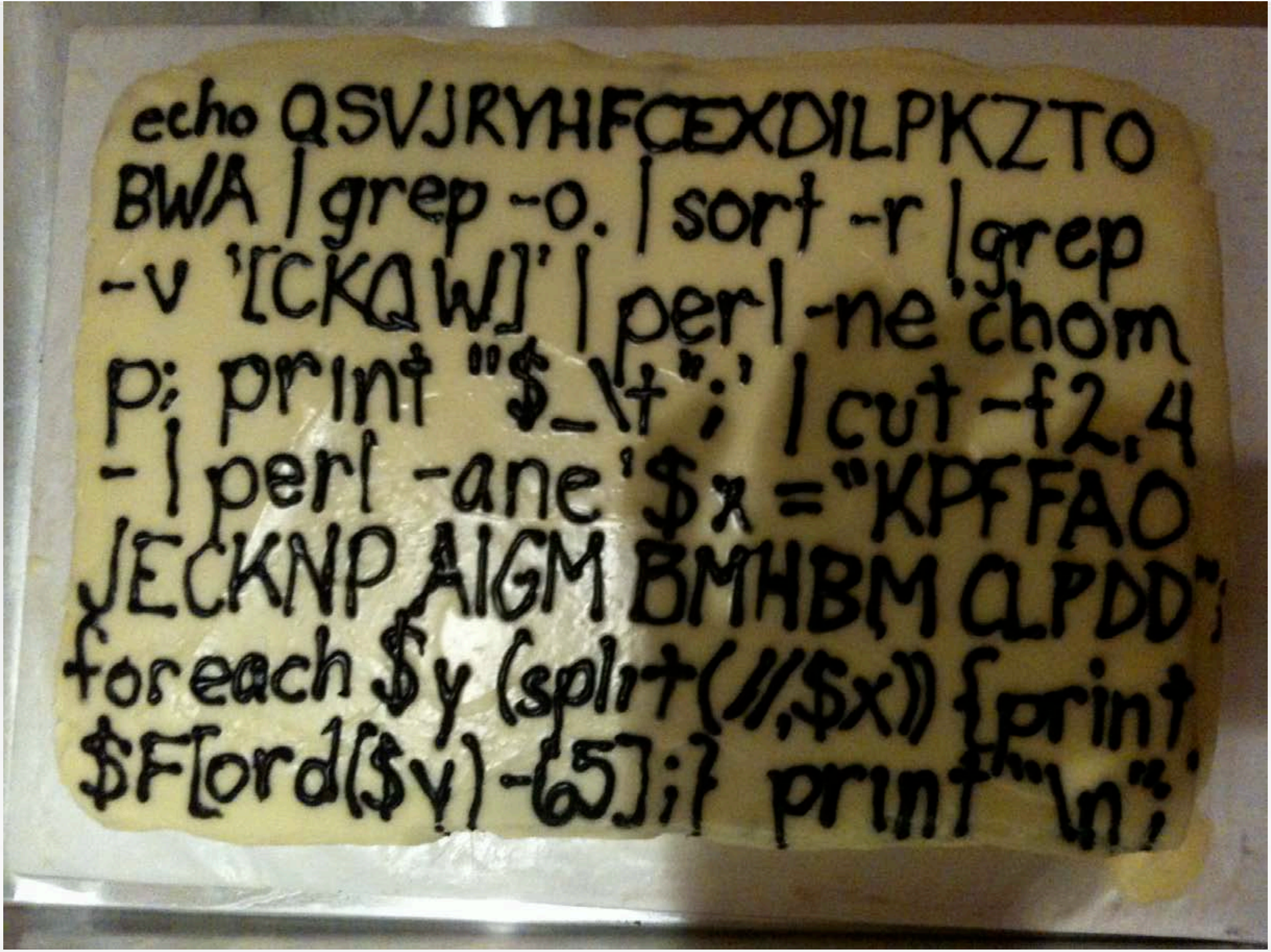
Workshops at UC Davis Genome Center

3rd Intensive Next Generation Data Analysis & Cloud Computing Workshop

Sept. 13th – Sept 17th and Sept. 21st – Sept. 22, 2010

http://bioinformatics.ucdavis.edu or http://bit.ly/9FCBGH

Bioinformatics Geeks' Cake



```
echo QSVJRYHFCEXDILPKZTO
BWA |grep -o. |sort -r |grep
-v '[CKQW]' |perl -ne 'chom
p; print "$_\t";' |cut -f2,4
-|perl -ane '$x="KPFFAO
JECKNP AIGM BMHBM CLPDD";
foreach $y (split(//,$x)) {print
$F[ord($y)-65];} print "\n";'
```

crypticmessgefromnikHAPPYBIRTHDAYJOEVELVETFASS