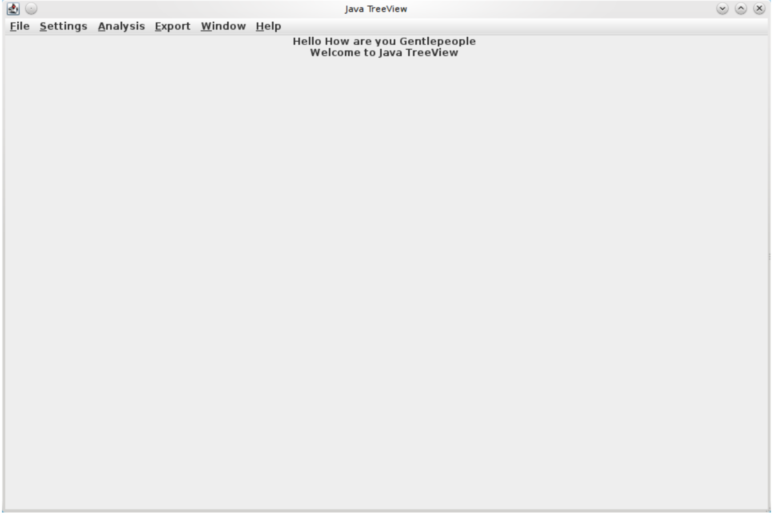
# Practical Bioinformatics

Mark Voorhies

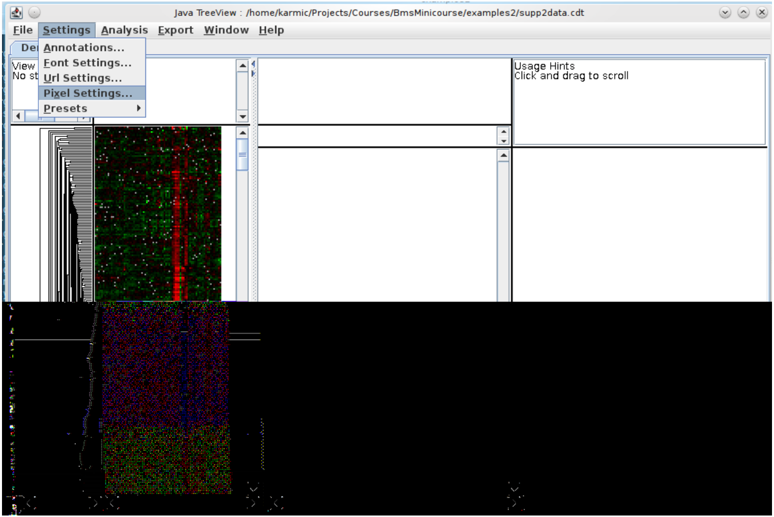
4/27/2011

# Using JavaTreeView

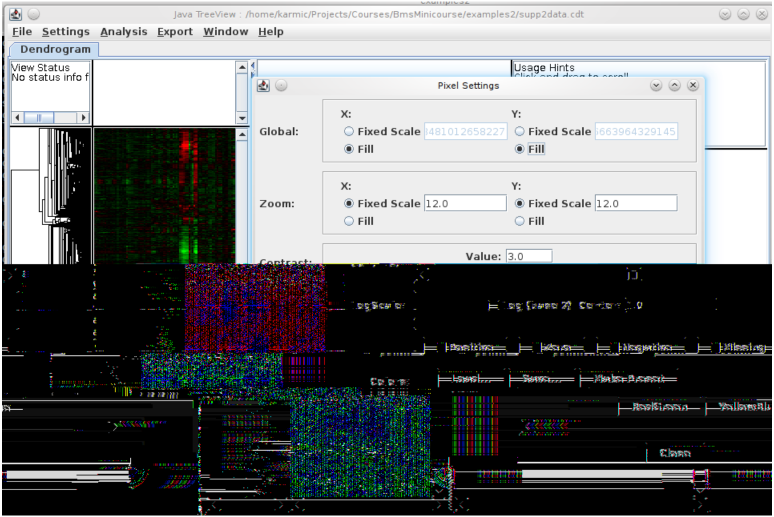


Hello How are you Gentlepeople
Welcome to Java TreeView

# Adjust pixel settings for global view



# Adjust pixel settings for global view

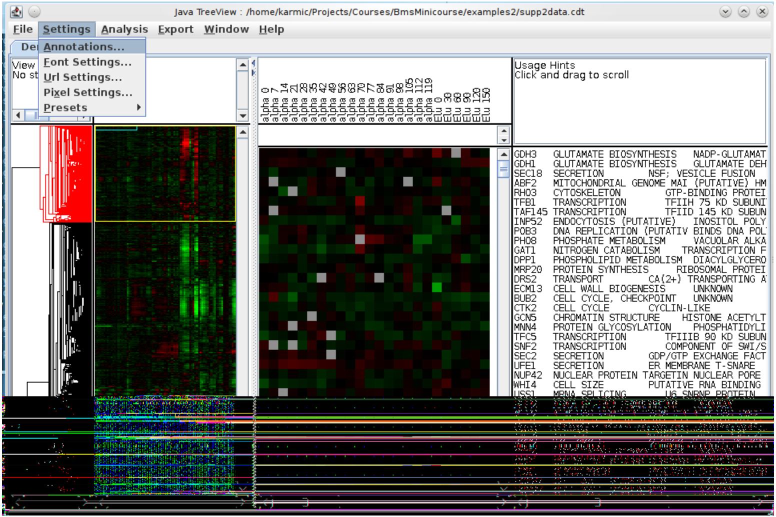


**Global:**
X: ○ Fixed Scale 481012658227 ● Fill
Y: ○ Fixed Scale 663964329145 ● Fill

**Zoom:**
X: ● Fixed Scale 12.0 ○ Fill
Y: ● Fixed Scale 12.0 ○ Fill

**Contrast:** Value: 3.0

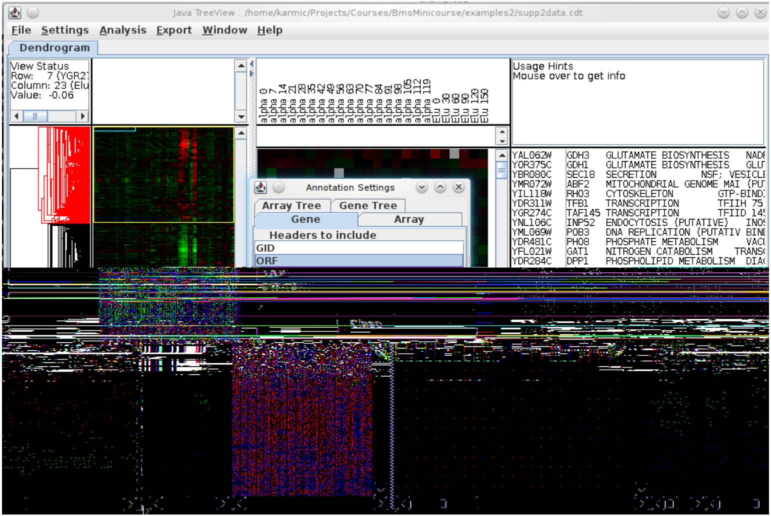
# Select annotation columns



Usage Hints
Click and drag to scroll

```
GDH3   GLUTAMATE BIOSYNTHESIS     NADP-GLUTAMAT
GDH1   GLUTAMATE BIOSYNTHESIS     GLUTAMATE DEH
SEC18  SECRETION              NSF; VESICLE FUSION
ABF2   MITOCHONDRIAL GENOME MAI (PUTATIVE) HM
RHO3   CYTOSKELETON             GTP-BINDING PROTEI
TFB1   TRANSCRIPTION            TFIIH 75 KD SUBUNI
TAF145 TRANSCRIPTION            TFIID 145 KD SUBUN
INP52  ENDOCYTOSIS (PUTATIVE)     INOSITOL POLY
POB3   DNA REPLICATION (PUTATIV BINDS DNA POL
PHO8   PHOSPHATE METABOLISM      VACUOLAR ALKA
GAT1   NITROGEN CATABOLISM      TRANSCRIPTION F
DPP1   PHOSPHOLIPID METABOLISM  DIACYLGLYCERO
MRP20  PROTEIN SYNTHESIS       RIBOSOMAL PROTEI
DRS2   TRANSPORT              CA(2+) TRANSPORTING A
ECM13  CELL WALL BIOGENESIS      UNKNOWN
BUB2   CELL CYCLE, CHECKPOINT    UNKNOWN
CTK2   CELL CYCLE             CYCLIN-LIKE
GCN5   CHROMATIN STRUCTURE      HISTONE ACETYLT
MNN4   PROTEIN GLYCOSYLATION     PHOSPHATIDYLI
TFC5   TRANSCRIPTION            TFIIIB 90 KD SUBUN
SNF2   TRANSCRIPTION            COMPONENT OF SWI/S
SEC2   SECRETION              GDP/GTP EXCHANGE FACT
UFE1   SECRETION              ER MEMBRANE T-SNARE
NUP42  NUCLEAR PROTEIN TARGETIN NUCLEAR PORE
WHI4   CELL SIZE              PUTATIVE RNA BINDING
USS1   MRNA SPLICING            U6 SNRNP PROTEIN
```

# Select annotation columns



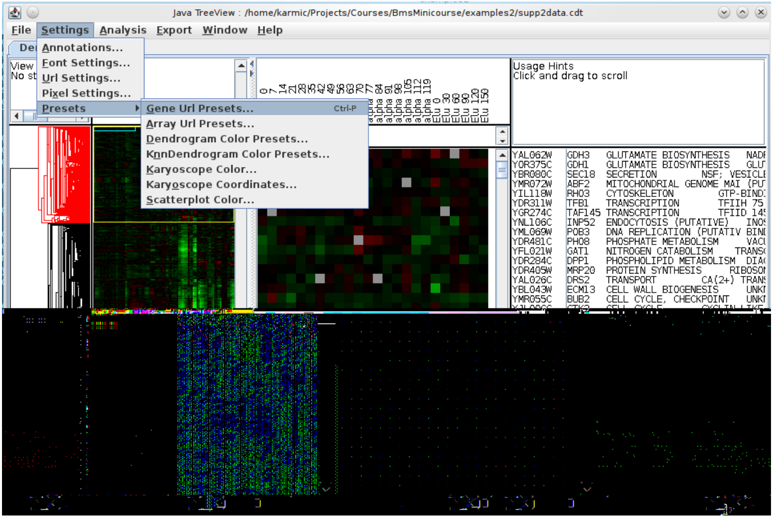
View Status
Row: 7 (YGR2
Column: 23 (Elu
Value: -0.06

Usage Hints
Mouse over to get info

Headers to include
GID
ORF

```
YAL062W  GDH3    GLUTAMATE BIOSYNTHESIS     NADP
YOR375C  GDH1    GLUTAMATE BIOSYNTHESIS     GLUT
YBR080C  SEC18   SECRETION            NSF; VESICLE
YMR072W  ABF2    MITOCHONDRIAL GENOME MAI (PUT
YIL118W  RHO3    CYTOSKELETON          GTP-BIND
YDR311W  TFB1    TRANSCRIPTION          TFIIH 75
YGR274C  TAF145  TRANSCRIPTION          TFIID 14
YNL106C  INP52   ENDOCYTOSIS (PUTATIVE)    INOS
YML069W  POB3    DNA REPLICATION (PUTATIV BIND
YDR481C  PHO8    PHOSPHATE METABOLISM       VACU
YFL021W  GAT1    NITROGEN CATABOLISM      TRANS
YDR284C  DPP1    PHOSPHOLIPID METABOLISM   DIA
```

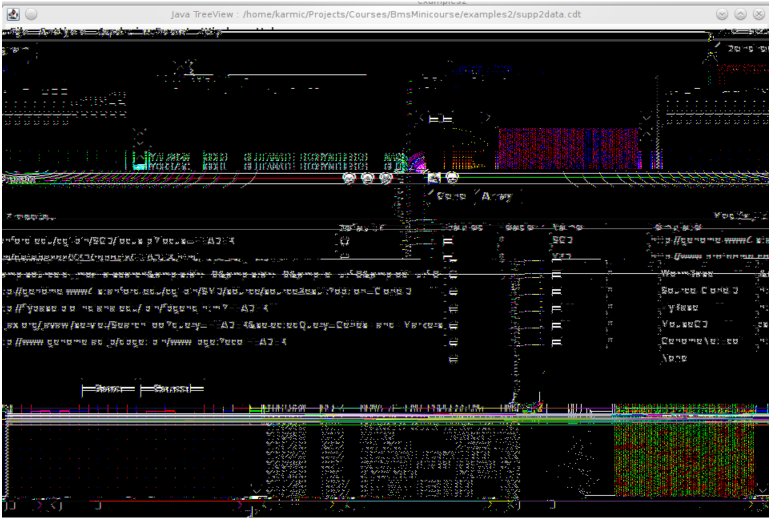
# Select URL for gene annotations



Usage Hints
Click and drag to scroll

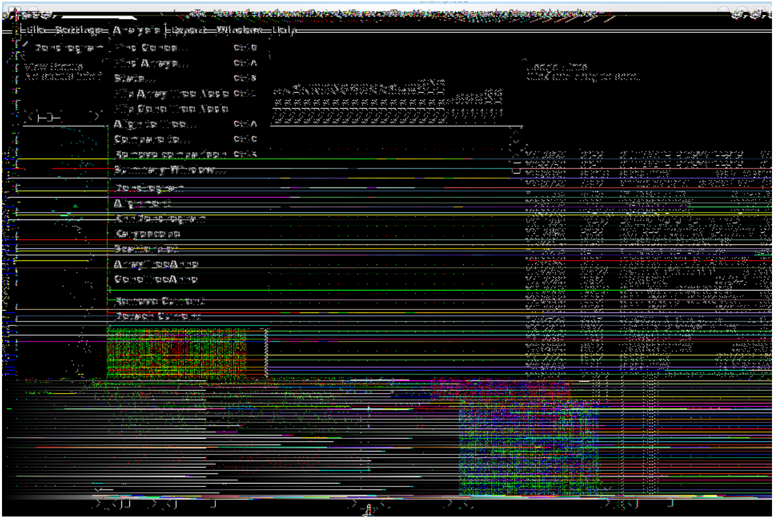
| | | | |
|---|---|---|---|
| YAL062W | GDH3 | GLUTAMATE BIOSYNTHESIS | NADF |
| YOR375C | GDH1 | GLUTAMATE BIOSYNTHESIS | GLUT |
| YBR080C | SEC18 | SECRETION | NSF; VESICLE |
| YMR072W | ABF2 | MITOCHONDRIAL GENOME MAI | (PUT |
| YIL118W | RHO3 | CYTOSKELETON | GTP-BINDI |
| YDR311W | TFB1 | TRANSCRIPTION | TFIIH 75 |
| YGR274C | TAF145 | TRANSCRIPTION | TFIID 145 |
| YNL106C | INP52 | ENDOCYTOSIS (PUTATIVE) | INOS |
| YML069W | POB3 | DNA REPLICATION (PUTATIV | BIND |
| YDR481C | PHO8 | PHOSPHATE METABOLISM | VACU |
| YFL021W | GAT1 | NITROGEN CATABOLISM | TRANSC |
| YDR284C | DPP1 | PHOSPHOLIPID METABOLISM | DIAC |
| YDR405W | MRP20 | PROTEIN SYNTHESIS | RIBOSOM |
| YAL026C | DRS2 | TRANSPORT | CA(2+) TRANS |
| YBL043W | ECM13 | CELL WALL BIOGENESIS | UNKN |
| YMR055C | BUB2 | CELL CYCLE, CHECKPOINT | UNKN |

# Select URL for gene annotations

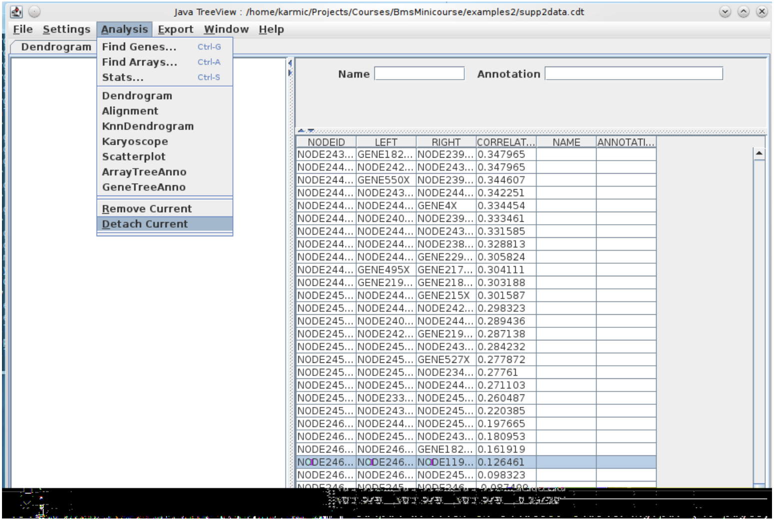


Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

# Activate and detach annotation window

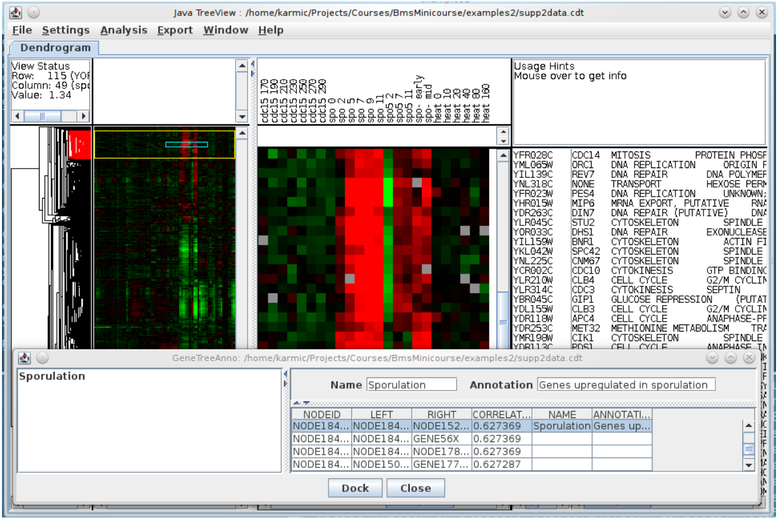


# Activate and detach annotation window



| NODEID | LEFT | RIGHT | CORRELAT... | NAME | ANNOTATI... |
|---|---|---|---|---|---|
| NODE243... | GENE182... | NODE239... | 0.347965 | | |
| NODE244... | NODE242... | NODE243... | 0.347965 | | |
| NODE244... | GENE550X | NODE239... | 0.344607 | | |
| NODE244... | NODE243... | NODE244... | 0.342251 | | |
| NODE244... | NODE244... | GENE4X | 0.334454 | | |
| NODE244... | NODE240... | NODE239... | 0.333461 | | |
| NODE244... | NODE244... | NODE243... | 0.331585 | | |
| NODE244... | NODE244... | NODE238... | 0.328813 | | |
| NODE244... | NODE244... | GENE229... | 0.305824 | | |
| NODE244... | GENE495X | GENE217... | 0.304111 | | |
| NODE244... | GENE219... | GENE218... | 0.303188 | | |
| NODE245... | NODE244... | GENE215X | 0.301587 | | |
| NODE245... | NODE244... | NODE242... | 0.298323 | | |
| NODE245... | NODE240... | NODE244... | 0.289436 | | |
| NODE245... | NODE242... | GENE219... | 0.287138 | | |
| NODE245... | NODE245... | NODE243... | 0.284232 | | |
| NODE245... | NODE245... | GENE527X | 0.277872 | | |
| NODE245... | NODE245... | NODE234... | 0.27761 | | |
| NODE245... | NODE245... | NODE244... | 0.271103 | | |
| NODE245... | NODE233... | NODE245... | 0.260487 | | |
| NODE245... | NODE243... | NODE245... | 0.220385 | | |
| NODE246... | NODE244... | NODE245... | 0.197665 | | |
| NODE246... | NODE245... | NODE243... | 0.180953 | | |
| NODE246... | NODE246... | GENE182... | 0.161919 | | |
| NODE246... | NODE246... | NODE119... | 0.126461 | | |
| NODE246... | NODE246... | NODE245... | 0.098323 | | |

# Activate and detach annotation window



View Status
Row: 115 (YOI
Column: 49 (spo
Value: 1.34

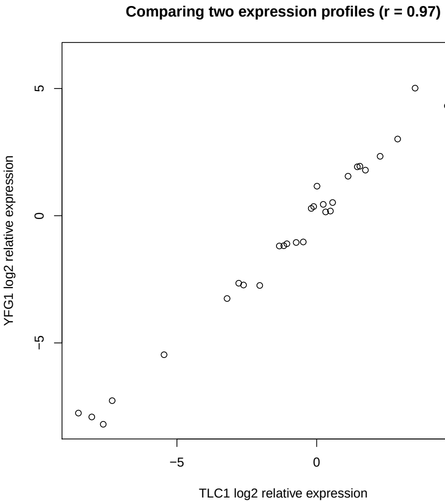
Usage Hints
Mouse over to get info

```
YFR028C  CDC14  MITOSIS                PROTEIN PHOSF
YML065W  ORC1   DNA REPLICATION        ORIGIN F
YIL139C  REV7   DNA REPAIR             DNA POLYMEF
YNL318C  NONE   TRANSPORT              HEXOSE PERM
YFR023W  PES4   DNA REPLICATION        UNKNOWN;
YHR015W  MIP6   MRNA EXPORT, PUTATIVE  RNA
YDR263C  DIN7   DNA REPAIR (PUTATIVE)  DNA
YLR045C  STU2   CYTOSKELETON           SPINDLE
YOR033C  DHS1   DNA REPAIR             EXONUCLEASE
YIL159W  BNR1   CYTOSKELETON           ACTIN FI
YKL042W  SPC42  CYTOSKELETON           SPINDLE
YNL225C  CNM67  CYTOSKELETON           SPINDLE
YCR002C  CDC10  CYTOKINESIS            GTP BINDINC
YLR210W  CLB4   CELL CYCLE             G2/M CYCLIN
YLR314C  CDC3   CYTOKINESIS            SEPTIN
YBR045C  GIP1   GLUCOSE REPRESSION     (PUTAT
YDL155W  CLB3   CELL CYCLE             G2/M CYCLIN
YDR118W  APC4   CELL CYCLE             ANAPHASE-PF
YDR253C  MET32  METHIONINE METABOLISM  TRA
YMR198W  CIK1   CYTOSKELETON           SPINDLE
YDR113C  PDS1   CELL CYCLE             ANAPHASE IN
```

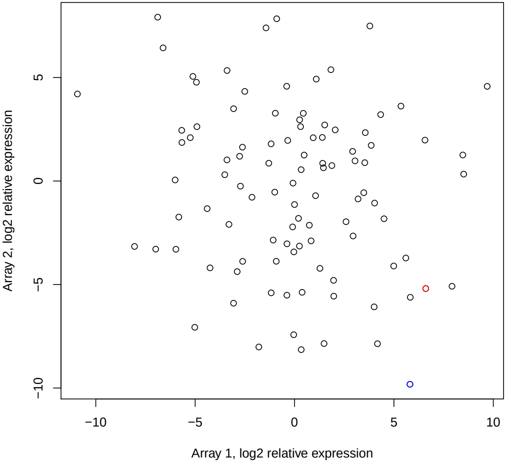
**Name** Sporulation **Annotation** Genes upregulated in sporulation

| NODEID | LEFT | RIGHT | CORRELAT... | NAME | ANNOTATI... |
|---|---|---|---|---|---|
| NODE184... | NODE184... | NODE152... | 0.627369 | Sporulation | Genes up... |
| NODE184... | NODE184... | GENE56X | 0.627369 | | |
| NODE184... | NODE184... | NODE178... | 0.627369 | | |
| NODE184... | NODE150... | GENE177... | 0.627287 | | |

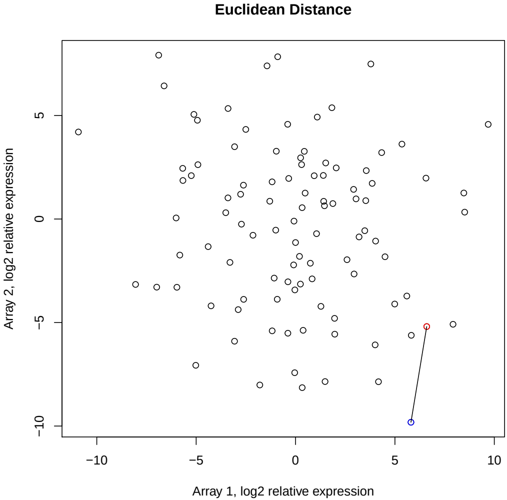
# Comparing all measurements for two genes



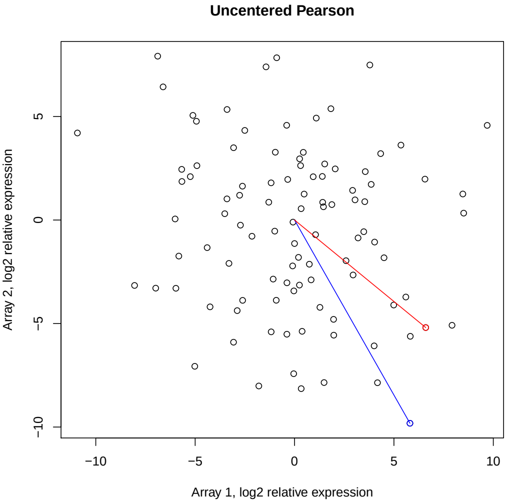
# Comparing all genes for two measurements



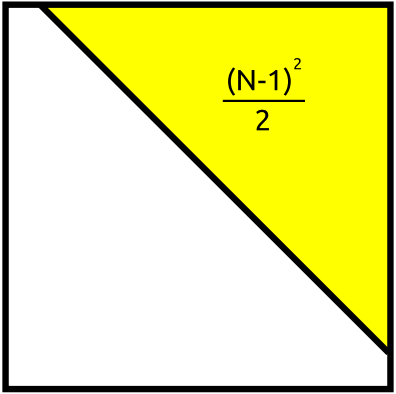
# Comparing all genes for two measurements



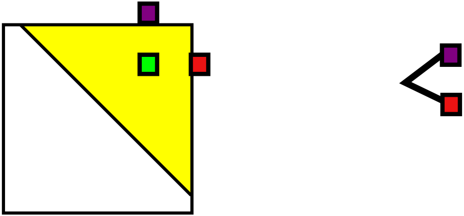
# Comparing all genes for two measurements



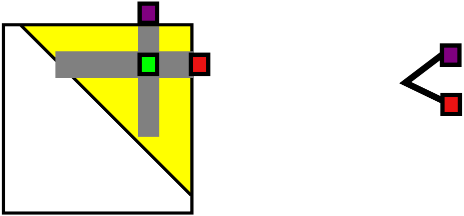
# Measure all pairwise distances under distance metric



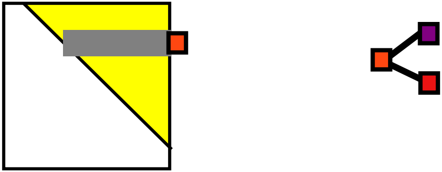
# Hierarchical Clustering



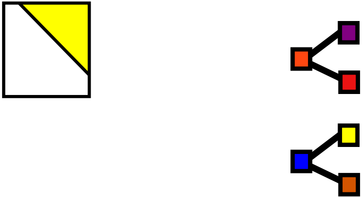
# Hierarchical Clustering



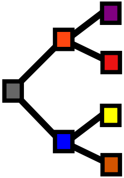
# Hierarchical Clustering



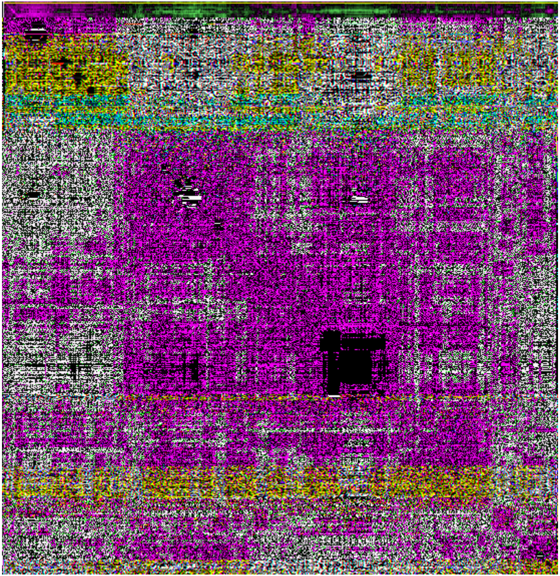
# Hierarchical Clustering



# Hierarchical Clustering



# Clustering exercises { Visualizing the distance matrix



# Clustering exercises { Visualizing the distance matrix

# Writing stand-alone scripts

```
#!/usr/bin/env python
"""Example command-line program."""
import sys
```

# Writing stand-alone scripts

```
#!/usr/bin/env python
"""Example command-line program."""
import sys
if(__name__ == "__main__"):
    print "Hello, world"
    for arg in sys.argv:
        sys.stderr.write(arg+"\n")
    for line in sys.stdin:
        sys.stdout.write(line)
```

Windows

- script.py arg1 arg2

# Writing stand-alone scripts

```
#!/usr/bin/env python
"""Example command-line program."""
import sys
if(__name__ == "__main__"):
    print "Hello, world"
    for arg in sys.argv:
        sys.stderr.write(arg+"\n")
    for line in sys.stdin:
        sys.stdout.write(line)
```

Windows

- script.py arg1 arg2

OS X

- chmod “a+x” script.py
- ./script.py arg1 arg2

# Writing stand-alone scripts

```python
#!/usr/bin/env python
"""Example command-line program."""
import sys
if(__name__ == "__main__"):
    print "Hello, world"
    for arg in sys.argv:
        sys.stderr.write(arg+"\n")
    for line in sys.stdin:
        sys.stdout.write(line)
```

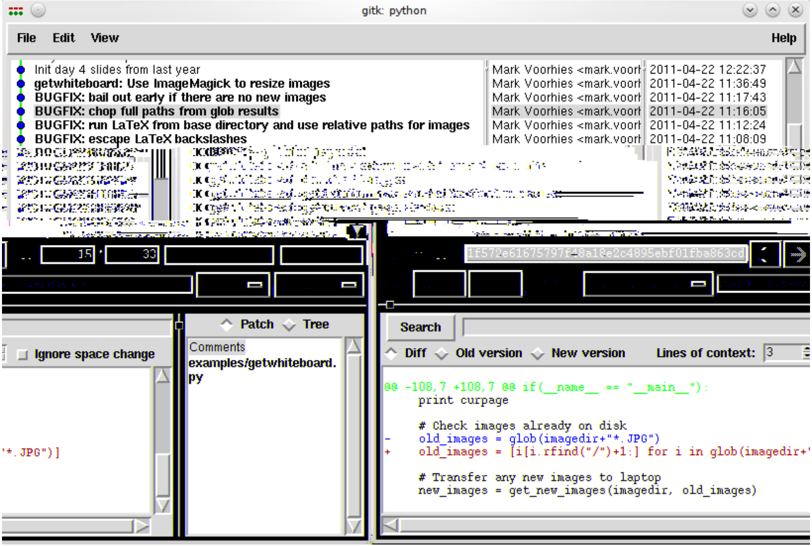
Windows

- script.py arg1 arg2

OS X

- chmod "a+x" script.py
- ./script.py arg1 arg2

# Version Control



```
Init day 4 slides from last year                                                  Mark Voorhies <mark.voorh  2011-04-22 12:22:37
getwhiteboard: Use ImageMagick to resize images                                   Mark Voorhies <mark.voorh  2011-04-22 11:36:49
BUGFIX: bail out early if there are no new images                                 Mark Voorhies <mark.voorh  2011-04-22 11:17:43
BUGFIX: chop full paths from glob results                                         Mark Voorhies <mark.voorh  2011-04-22 11:16:05
BUGFIX: run LaTeX from base directory and use relative paths for images          Mark Voorhies <mark.voorh  2011-04-22 11:12:24
BUGFIX: escape LaTeX backslashes                                                  Mark Voorhies <mark.voorh  2011-04-22 11:08:09
```

```
1f572e61675797f48a18e2c4895ebf01fba863cd
```

Comments
examples/getwhiteboard.py

```
@@ -108,7 +108,7 @@ if(__name__ == "__main__"):
     print curpage

     # Check images already on disk
-    old_images = glob(imagedir+"*.JPG")
+    old_images = [i[i.rfind("/")+1:] for i in glob(imagedir+'
                                                                   '*.JPG")]
     # Transfer any new images to laptop
     new_images = get_new_images(imagedir, old_images)
```

# Scripting Cluster

- Running Cluster3 from the command line
    - /Applications/Cluster.app/Contents/MacOS/Cluster
    - /Program Files/Stanford University/Cluster3/Cluster.exe
- Command-line programs are like functions
- “man program” is like “help(function)”
- Use the `subprocess` module to run command-line programs from within Python.

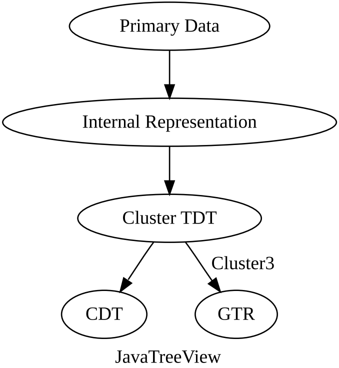
# Programs as functions

USAGE: cluster [options]

| | |
|---|---|
| -f filename | File loading |
| -u jobname | Allows you to specify a different name for the output files<br>(default is derived from the input file name) |
| -g [0..8] | Specifies the distance measure for gene clustering<br>0: No gene clustering<br>1: Uncentered correlation<br>2: Pearson correlation<br>3: Uncentered correlation, absolute value<br>4: Pearson correlation, absolute value<br>5: Spearman's rank correlation<br>6: Kendall's tau<br>7: Euclidean distance<br>8: City-block distance<br>(default: 0) |
| -m [msca] | Specifies which hierarchical clustering method to use<br>m: Pairwise complete-linkage<br>s: Pairwise single-linkage<br>c: Pairwise centroid-linkage<br>a: Pairwise average-linkage<br>(default: m) |

# Scripting the Protocol

```
from subprocess import check_call
check_call(
    # Which program to run
    ("cluster",
     # Input file
     "-f","supp2data.tdt",
     # Output prefix
     "-u","supp2data.Uncentered.Complete",
     # Clustering method: complete linkage
     "-m","m",
     # Distance function: uncentered Pearson
     "-g","1"))
```



# Clustering exercises { Scripting Cluster

Modify the clustering protocol script to run Cluster3 multiple times on the same input, varying distance metric and/or clustering method. Be sure to give the output files distinct names.

# Clustering exercises { Scripting Cluster

Modify the clustering protocol script to run Cluster3 multiple times on the same input, varying distance metric and/or clustering method. Be sure to give the output files distinct names.

```
metrics = ("None",
           "Uncentered",
           "Pearson",
           "UncenteredAbs",
           "PearsonAbs",
           "Spearman",
           "Kendall",
           "Euclidean",
           "City")
linkage = (("Complete","m"),
           ("Single","s"),
           ("Centroid","c"),
           ("Average","a"))

# Loop over all 32 possible methods
print "Starting hierarchical clustering runs..."
from subprocess import check_call
for metric in xrange(1,len(metrics)):
    print "    ",metrics[metric],"..."
    for (linkname, link) in linkage:
        print "       ",linkname
        check_call(("cluster","—f","shuffled.txt",
                    "—u",".".join(("shuffled",
                                   metrics[metric],
                                   linkname)),
                    "—m",link,"—g",str(metric)))
```

# Dictionaries

```
dictionary = {"A":"T","T":"A","G":"C","C":"G"}
dictionary["G"]
dictionary["N"] = "N"
dictionary.has_key("C")
```

# Dictionaries

```
geneticCode = {"TTT":"F","TTC":"F","TTA":"L","TTG":"L",
               "CTT":"L","CTC":"L","CTA":"L","CTG":"L",
               "ATT":"I","ATC":"I","ATA":"I","ATG":"M",
               "GTT":"V","GTC":"V","GTA":"V","GTG":"V",

               "TCT":"S","TCC":"S","TCA":"S","TCG":"S",
               "CCT":"P","CCC":"P","CCA":"P","CCG":"P",
               "ACT":"T","ACC":"T","ACA":"T","ACG":"T",
               "GCT":"A","GCC":"A","GCA":"A","GCG":"A",

               "TAT":"Y","TAC":"Y","TAA":"*","TAG":"*",
               "CAT":"H","CAC":"H","CAA":"Q","CAG":"Q",
               "AAT":"N","AAC":"N","AAA":"K","AAG":"K",
               "GAT":"D","GAC":"D","GAA":"E","GAG":"E",

               "TGT":"C","TGC":"C","TGA":"*","TGG":"W",
               "CGT":"R","CGC":"R","CGA":"R","CGG":"R",
               "AGT":"S","AGC":"S","AGA":"R","AGG":"R",
               "GGT":"G","GGC":"G","GGA":"G","GGG":"G"}
```

# Exercise: Transforming sequences

1. Write a function to return the antisense strand of a DNA sequence in 3'→5' orientation.
2. Write a function to return the compliment of a DNA sequence in 5'→3' orientation.
3. Write a function to translate a DNA sequence