PCA and clustering

STEPHANIE J. SPIELMAN, PHD

BIO5312, FALL 2017

Exploratory methods for high-dimensional data

Principal components analysis (PCA)

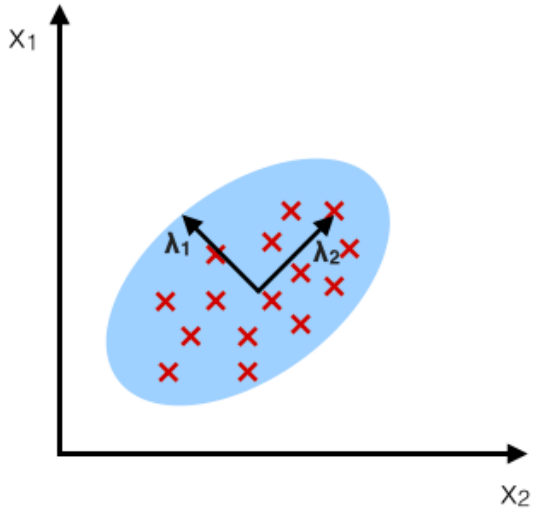
- Note there are many similar methods, e.g. linear discriminant analysis

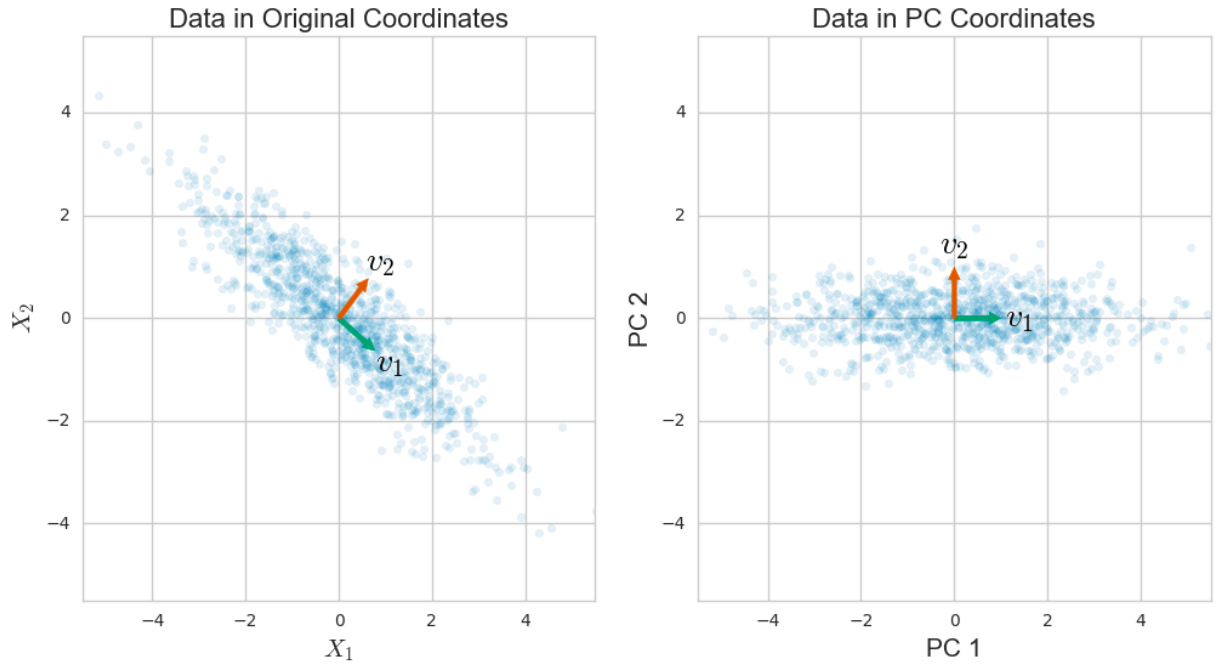
Clustering

- K-means
- Hierarchical
- Again, **many** more

Principal components analysis

Linear algebra technique to emphasize *axes of variation in the data*





PCA offers new coordinate system to emphasize variation in the data

Do it yourself!

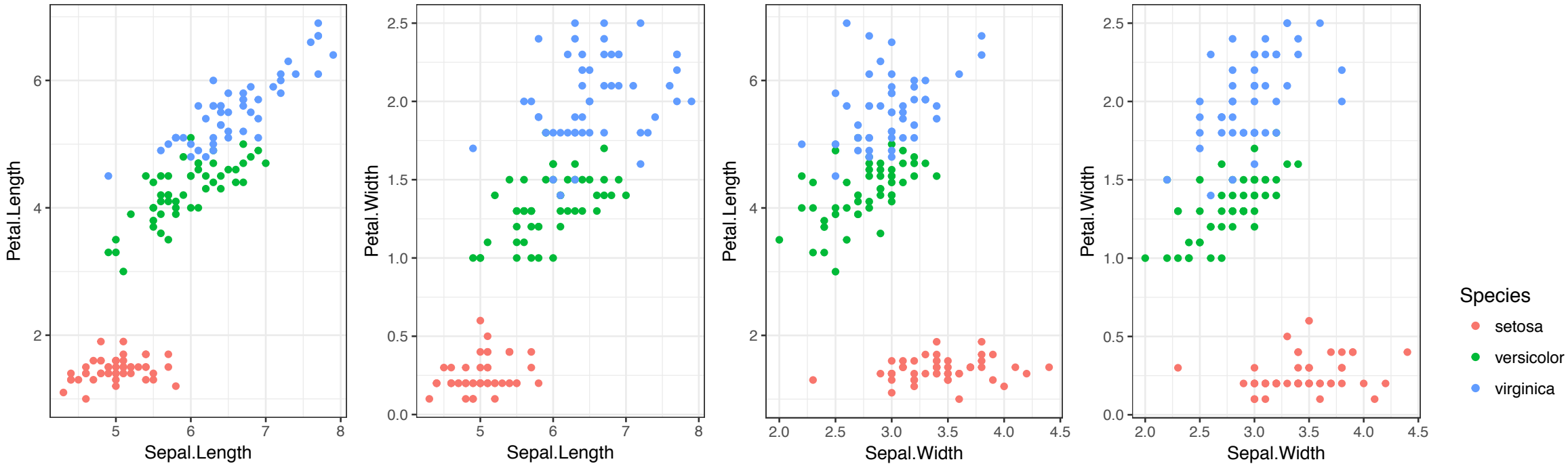
There are as many PCs are there are variables

- NUMERIC ONLY

http://setosa.io/ev/principal-component-analysis/

Example: Iris

How well we can tell species apart depends on plotting strategy



PCA on iris

```
> iris %>%
    select(-Species) %>%      ### Remove any non-numeric columns
    scale() %>%               ### Scale the data (columns in same units)
    prcomp() -> iris.pca      ### Run the PCA with prcomp()
```

PCA output

```
## Rotation matrix: Loadings are the percent of variance explained by the variable
> iris.pca$rotation
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Sepal.Length, Petal.Length, and Petal.Width load positively on PC1.
Sepal.Width shows a weaker negative loading on PC1.

PC2 is dominated by Sepal.Width, which loads strongly and negatively.

PCA output

```
#### The actual principal components
> head(iris.pca$x)
            PC1        PC2         PC3           PC4
[1,] -2.257141 -0.4784238  0.12727962  0.024087508
[2,] -2.074013  0.6718827  0.23382552  0.102662845
[3,] -2.356335  0.3407664 -0.04405390  0.028282305
[4,] -2.291707  0.5953999 -0.09098530 -0.065735340
[5,] -2.381863 -0.6446757 -0.01568565 -0.035802870
[6,] -2.068701 -1.4842053 -0.02687825  0.006586116
```

PCA output

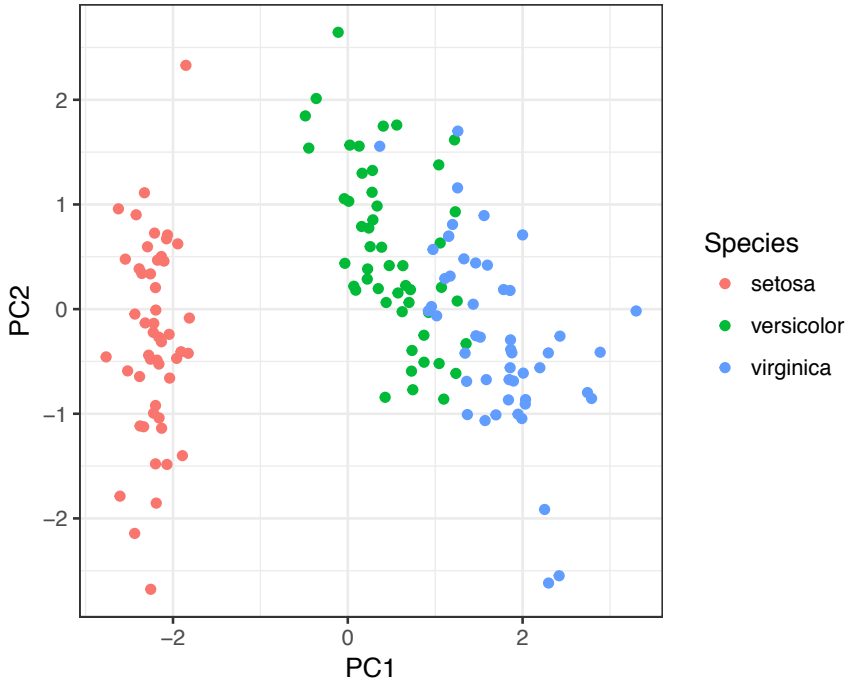
```
#### Standard deviation of components is represents the percent of variation each component explains, ish
> iris.pca$sdev
[1] 1.7083611 0.9560494 0.3830886 0.1439265


### Compute variance explained:
> (iris.pca$sdev)^2 / (sum(iris.pca$sdev^2))
[1] 0.729624454 0.228507618 0.036689219 0.005178709
```

PC1 explains ~73% of variance in the data. **By definition, PC1 explains the most variation** (and so on)
PC2 explains ~ 23% of variance in the data
etc.

Visualizing the PCA: PC vs PC

```
#### Bring back the original data for plotting
> plot.pca <- cbind(iris, iris.pca$x)
> ggplot(plot.pca, aes(x = PC1, y = PC2, color = Species)) + geom_point()
```

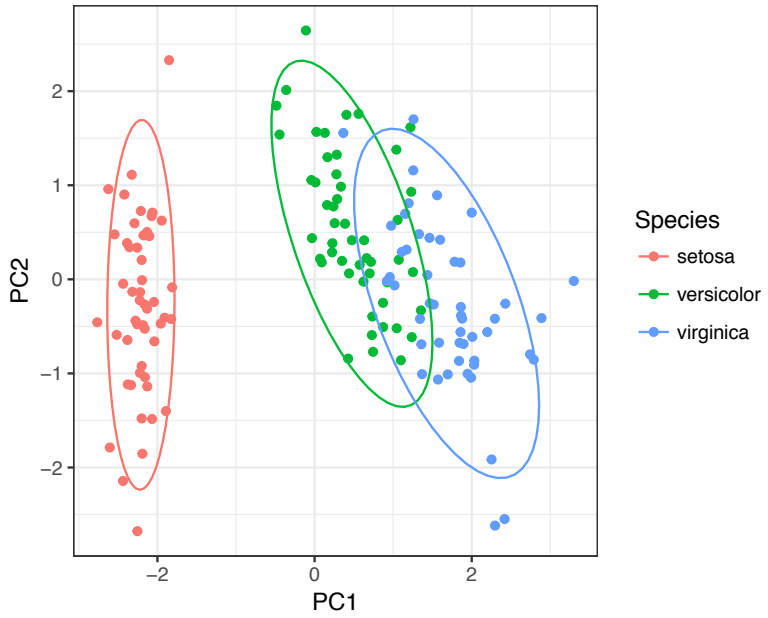


Visualizing the PCA: PC vs PC

```
#### Bring back the original data for plotting
> plot.pca <- cbind(iris, iris.pca$x)
> ggplot(plot.pca, aes(x = PC1, y = PC2, color = Species)) + geom_point() +
      stat_ellipse()
```

Species *separate* along PC1
PC1 *discriminates* species.

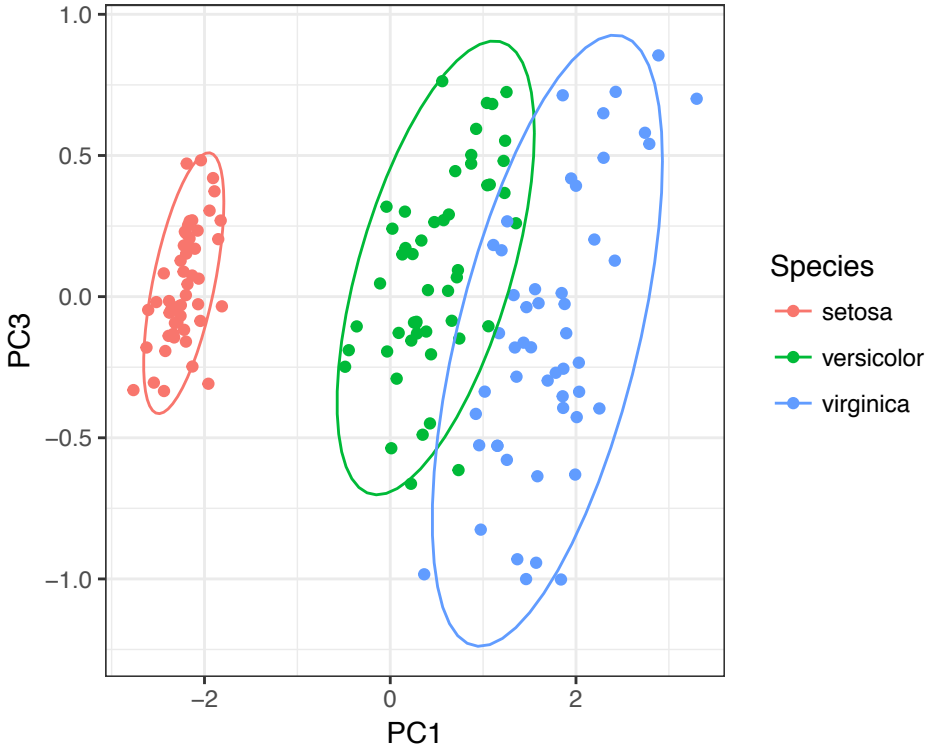
Species spread evenly across PC2.



PC1 vs PC3?

Species *separate* along PC1
PC1 *discriminates* species.

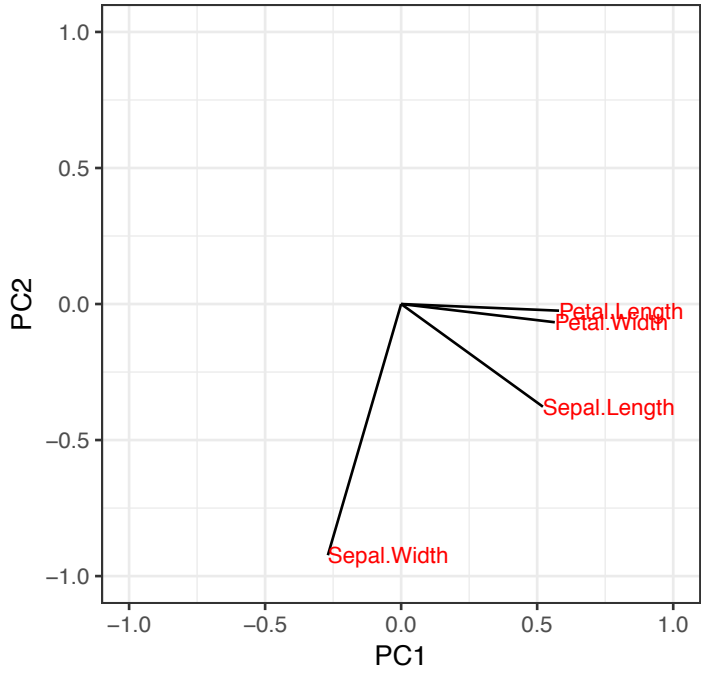
Setosa is more compact along PC3, whereas there is more spread for versicolor/virginica along PC3.



Visualizing the PCA: Loadings

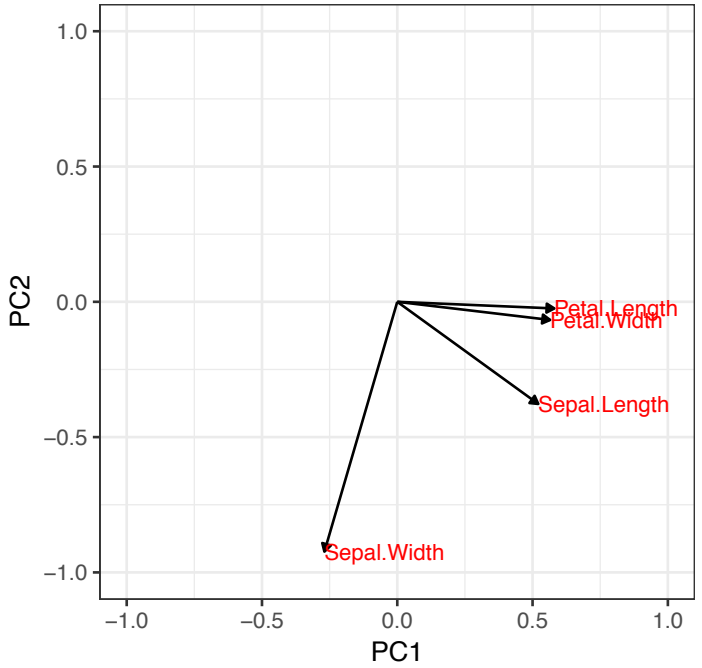
```
> as.data.frame(iris.pca$rotation) %>% rownames_to_column() -> loadings
       rowname         PC1         PC2        PC3        PC4
1 Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
2  Sepal.Width -0.2693474 -0.92329566 -0.2443818 -0.1235096
3 Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
4  Petal.Width  0.5648565 -0.06694199 -0.6342727  0.5235971
```

```
> ggplot(loadings) +
   geom_segment(x=0, y=0, aes(xend=PC1, yend=PC2)) +
   geom_text(aes(x=PC1, y=PC2, label=rowname), size=3, color='red') +
   xlim(-1.,1) +
   ylim(-1.,1.) +
   coord_fixed()
```



Loadings with arrows

```
> arrow_style <- arrow(length = unit(0.05, "inches"), type = "closed")
> ggplot(loadings) +
      geom_segment(x=0, y=0, aes(xend=PC1, yend=PC2), arrow = arrow_style) +
      geom_text(aes(x=PC1, y=PC2, label=rowname), size=3, color='red') +
      xlim(-1.,1) +
      ylim(-1.,1.) +
      coord_fixed()
```



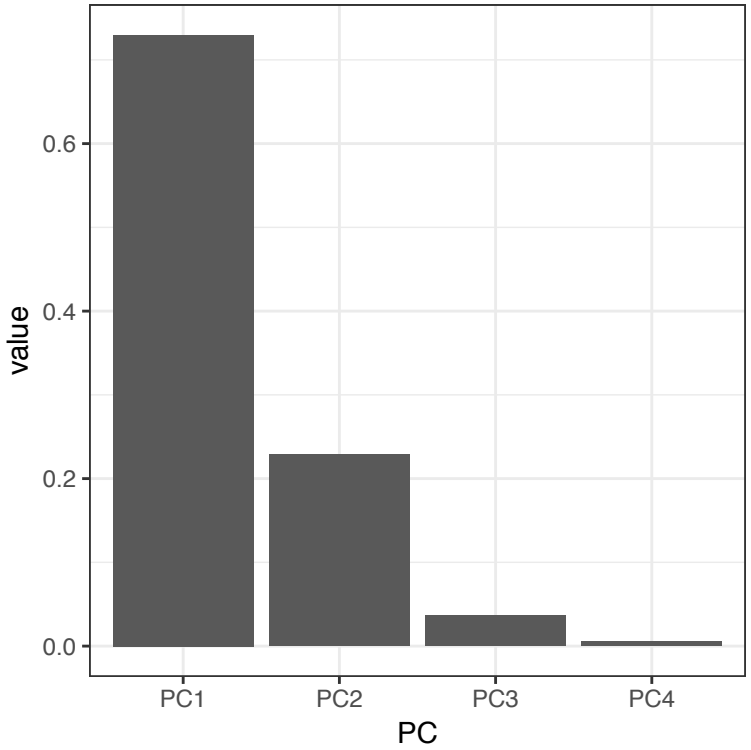
Petal.Length and Petal.Width load positively on PC1, but not at all on PC2.

Sepal.Width is *orthogonal* to petals, meaning it captures uncorrelated variation

Variation explained

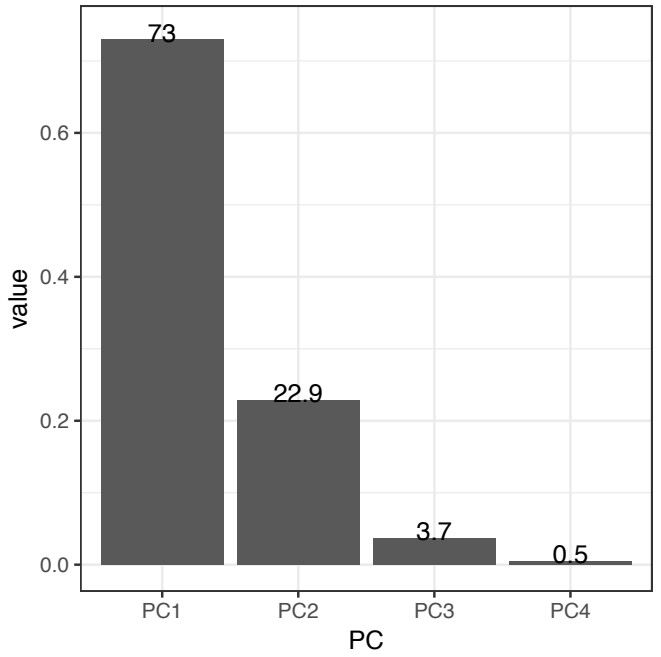
```
> as.tibble((iris.pca$sdev)^2 / (sum(iris.pca$sdev^2))) -> variance
# A tibble: 4 x 1
        value
        <dbl>
1 0.729624454
2 0.228507618
3 0.036689219
4 0.005178709

> variance %>% mutate(PC = colnames(iris.pca$x)) %>%
     ggplot(aes(x = PC, y = value)) +
     geom_bar(stat = "identity")
```



Variation explained

```
> variance %>% mutate(PC = colnames(iris.pca$x)) %>%
    ggplot(aes(x = PC, y = value)) +
    geom_bar(stat = "identity") +
    geom_text(aes(x = PC, y = value+0.01, label=100*round(value,3)))
```



Breathe break

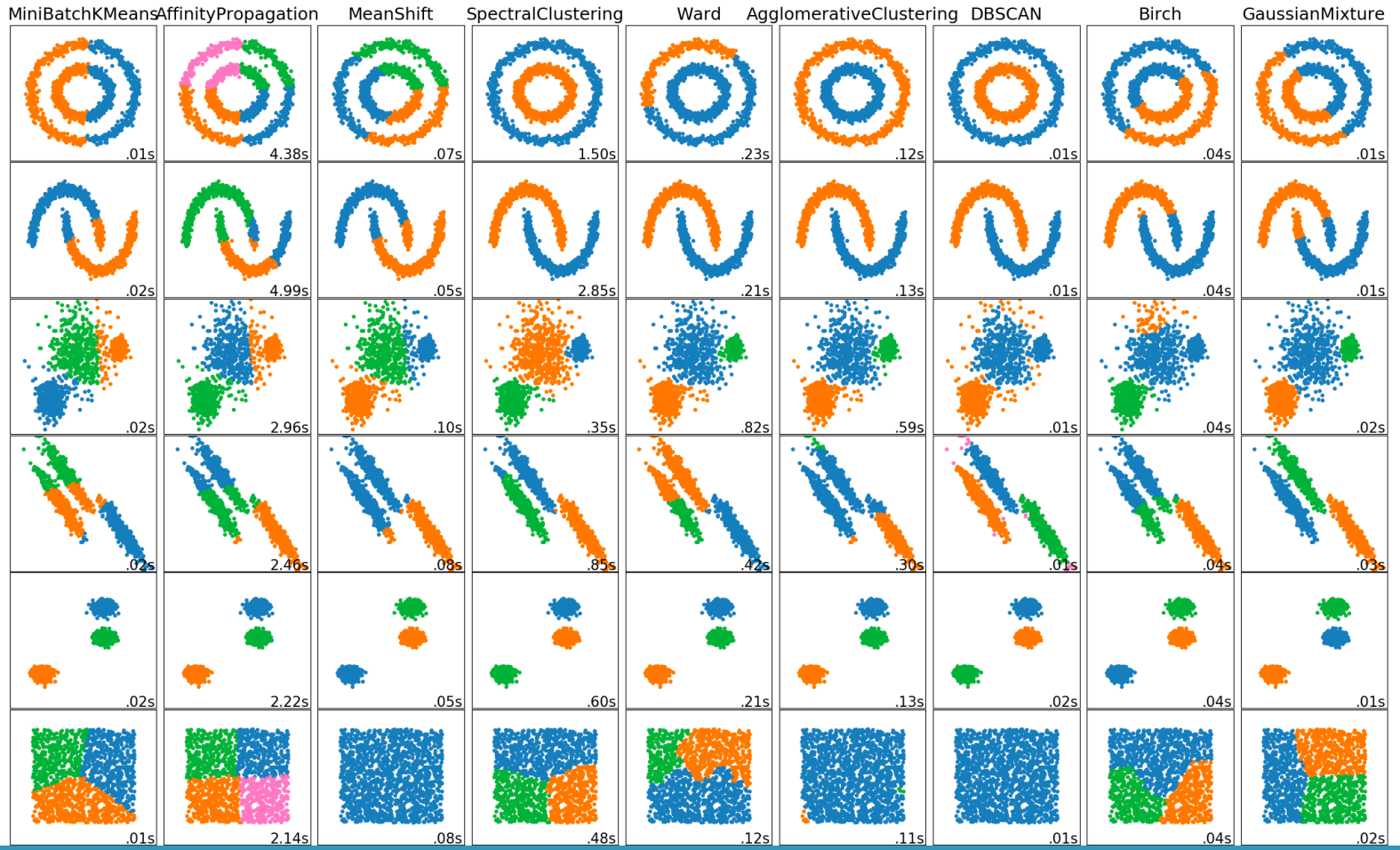
Clustering

A family of approaches to identify previously unknown or undetected groupings in data

Requires:

- A measure of distance and/or similarity among data points
- A clustering algorithm to create the groupings

There are too many algorithms and no real answers

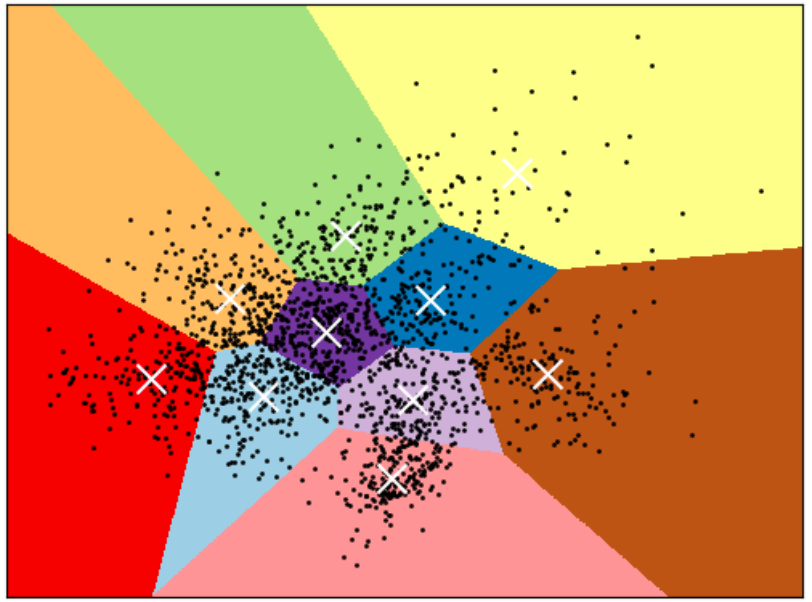


K-means clustering

Clusters data into *k* groups of equal variance by minimizing the *within-cluster sum of squares*

Divide *n* data points into ***k*** disjoint clusters, each described by its **mean** (ish)

K is specified *in advance*



K-means algorithm

1. Place k "centroids" in the data
2. Assign point to cluster k based on *Euclidian distance*
3. Re-compute each *k* centroid based on means of associated points
4. Re-assign centroids
5. Repeat until convergence

https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif

Do it yourself here:

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-means caveats

Clustering depends on initial conditions

Algorithm guaranteed to converge, but possibly on *local optima*

No real way to know if clusters have meaning beyond the math

- **This is true for all clustering!**

Example: iris with K=5

```
> iris %>%
      select(-Species) %>% ### We can only cluster numbers!
      kmeans(5)
K-means clustering with 5 clusters of sizes 50, 12, 25, 24, 39

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.006000    3.428000     1.462000    0.246000
2     7.475000    3.125000     6.300000    2.050000
3     5.508000    2.600000     3.908000    1.204000
4     6.529167    3.058333     5.508333    2.162500
5     6.207692    2.853846     4.746154    1.564103

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 5 5 5 3 5 5 5 3 5 3 3 5 3 5 3 5 5 3 5 3 5 3 5 5
 [75] 5 5 5 5 5 3 3 3 3 5 3 5 5 5 3 3 3 5 3 3 3 3 5 3 3 4 5 2 4 4 2 3 2 4 2 4
[112] 4 4 5 4 4 4 2 2 5 4 5 2 5 4 2 5 5 4 2 2 2 4 5 5 2 4 4 5 4 4 4 5 4 4 4 5 4
[149] 4 5

Within cluster sum of squares by cluster:
[1] 15.15100  4.65500  8.36640  5.46250 12.81128
 (between_SS / total_SS =  93.2 %)

Available components:


[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Example: iris with K=5… and broom!

```
> iris %>%
    select(-Species) %>% ### We can only cluster numbers!
    kmeans(5) %>%
    augment(iris) %>% ### Add clusters back into to original data frame
    head()
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species .cluster
1          5.1         3.5          1.4         0.2  setosa        4
2          4.9         3.0          1.4         0.2  setosa        4
3          4.7         3.2          1.3         0.2  setosa        4
4          4.6         3.1          1.5         0.2  setosa        4
5          5.0         3.6          1.4         0.2  setosa        4
6          5.4         3.9          1.7         0.4  setosa        4
```

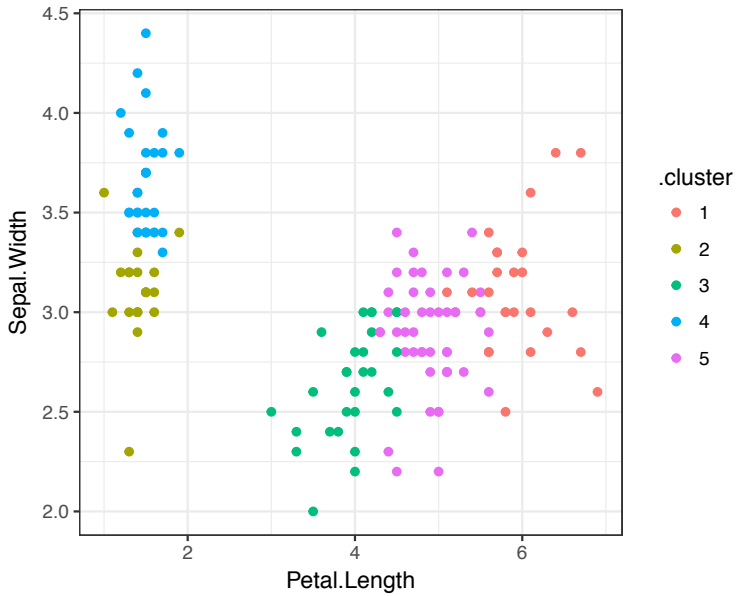
tidy() shows per-cluster information

```
> iris %>%
      select(-Species) %>% ### We can only cluster numbers!
      kmeans(5) %>%
      tidy()
         x1       x2       x3        x4 size  withinss cluster
1 5.532143 2.635714 3.960714 1.2285714   28  9.749286       1
2 6.264444 2.884444 4.886667 1.6666667   45 17.014222       2
3 4.704545 3.122727 1.413636 0.2000000   22  3.114091       3
4 7.014815 3.096296 5.918519 2.1555556   27 15.351111       4
5 5.242857 3.667857 1.500000 0.2821429   28  4.630714       5
```

Visualize the clustering

```
> iris %>%
      select(-Species) %>%
      kmeans(5) %>%
      augment(iris) %>%
      ggplot(aes(x = Petal.Length, y=Sepal.Width)) + geom_point(aes(color = .cluster))
```

No clear way to know "best" X and Y axes besides exhaustive plotting



Was K=5 reasonable?

One (of many) approaches to choosing the best K is the "elbow method"

- Plot within-sum-of-squares across different K choices
- "Best" k is where you see an elbow/kink in the plot
- **Highly subjective**

Choosing K with broom

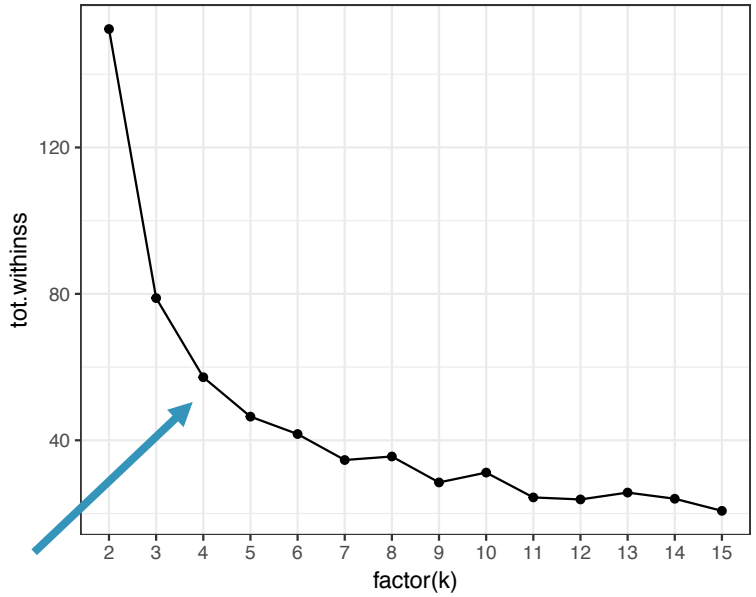
```
>iris %>%
      select(-Species) %>%
      kmeans(5) %>%
      glance()
      totss tot.withinss betweenss iter
1 681.3706     51.08942  630.2812    2
```

```
> numeric.iris <- iris %>% select(-Species)
> tibble(k = 2:15) %>%
   group_by(k) %>%
   do(kclust=kmeans(numeric.iris, .$k)) %>%
   glance(kclust)
```

	k	totss	tot.withinss	betweenss	iter
1	2	681.3706	152.34795	529.0226	1
2	3	681.3706	78.85144	602.5192	2
3	4	681.3706	57.26562	624.1050	2
4	5	681.3706	49.82228	631.5483	2
5	6	681.3706	42.42155	638.9491	4
6	7	681.3706	36.83714	644.5335	3
7	8	681.3706	40.84578	640.5248	3
...					

Choosing K with broom

```
> numeric.iris <- iris %>% select(-Species)
> tibble(k = 2:15) %>%
    group_by(k) %>%
    do(kclust=kmeans(numeric.iris, .$k)) %>%
    glance(kclust) %>%
    mutate(g = 1) %>%    ### ggplot gets angsty with geom_line without this specification
    ggplot(aes(x = factor(k), y = tot.withinss, group=g)) + geom_point() + geom_line()
```

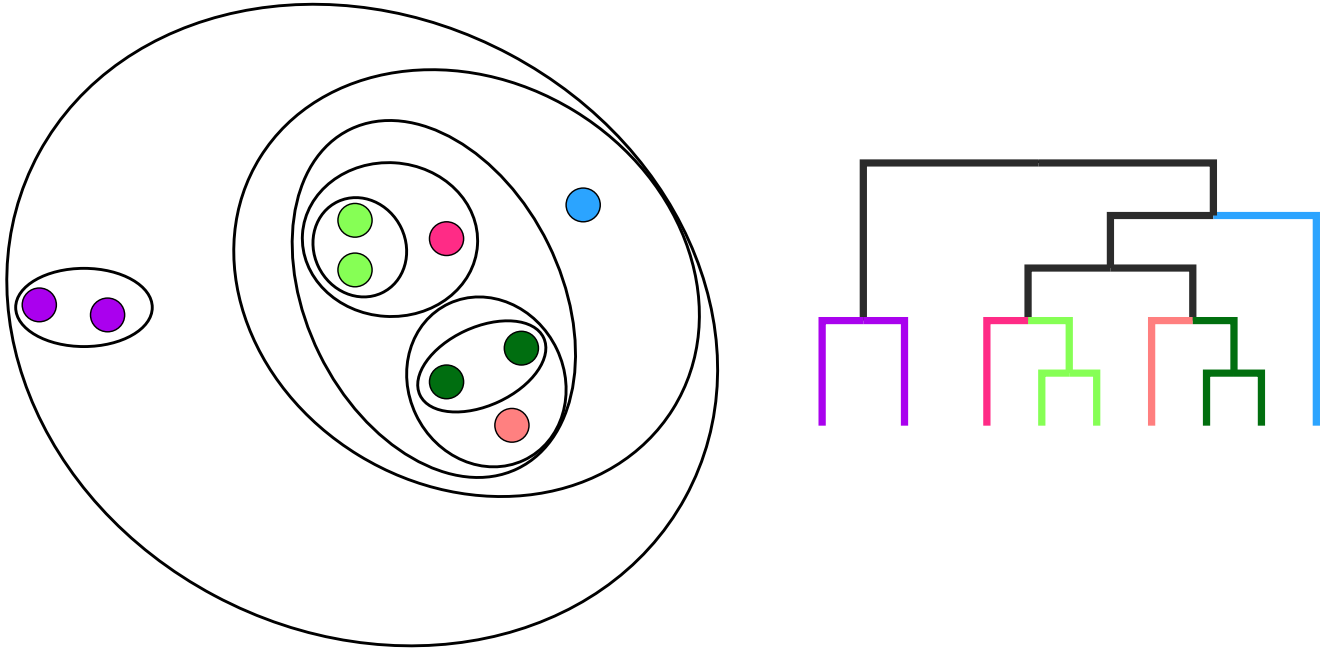


"Elbow" = shift in slope happens around K=4

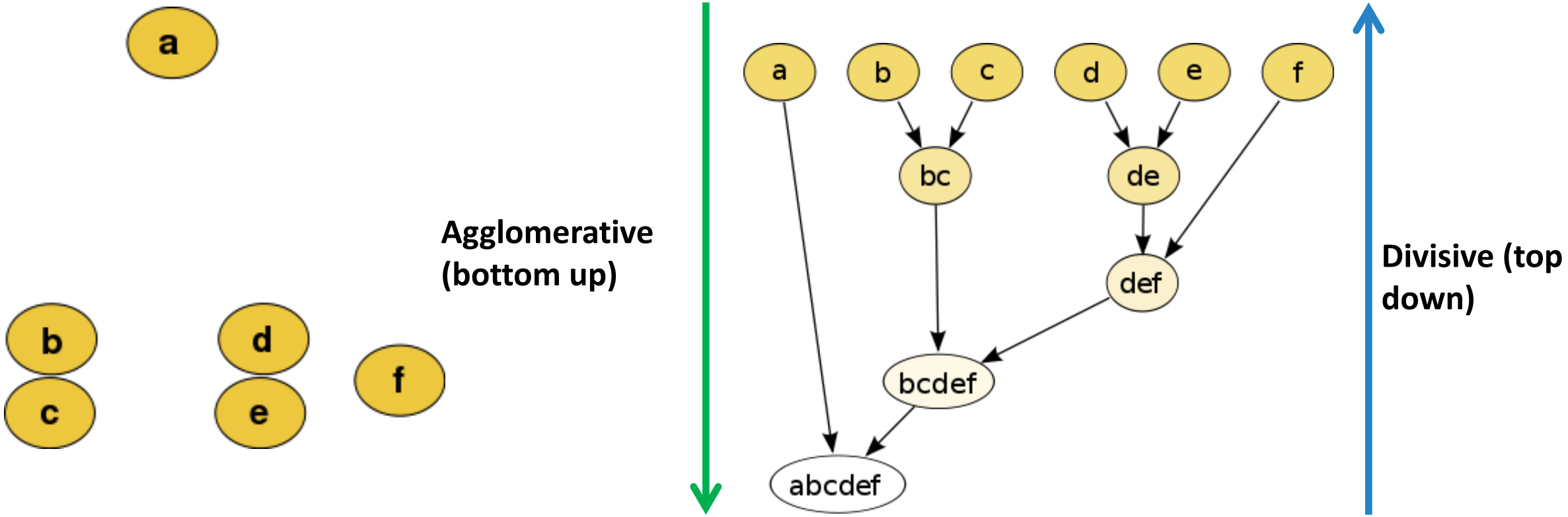
Hierarchical clustering

Extremely common in gene expression and/or systems biology studies

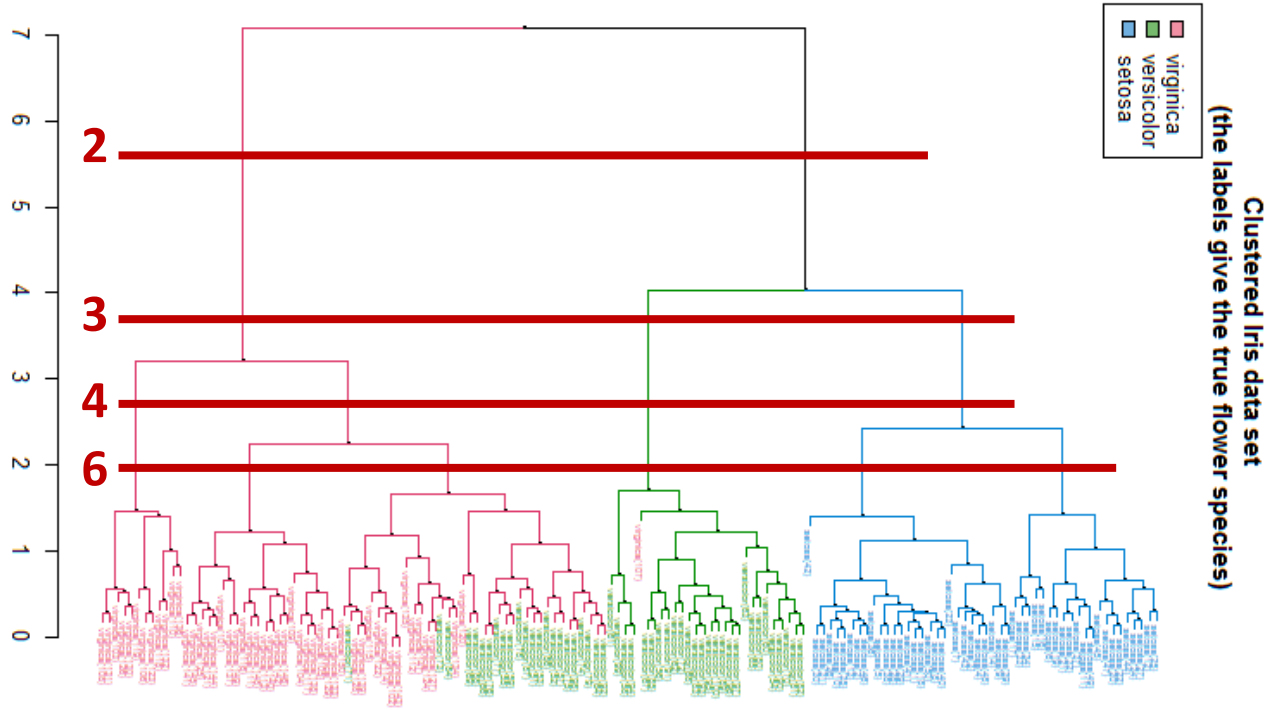
Useful when data have a hierarchical structure:



Approach

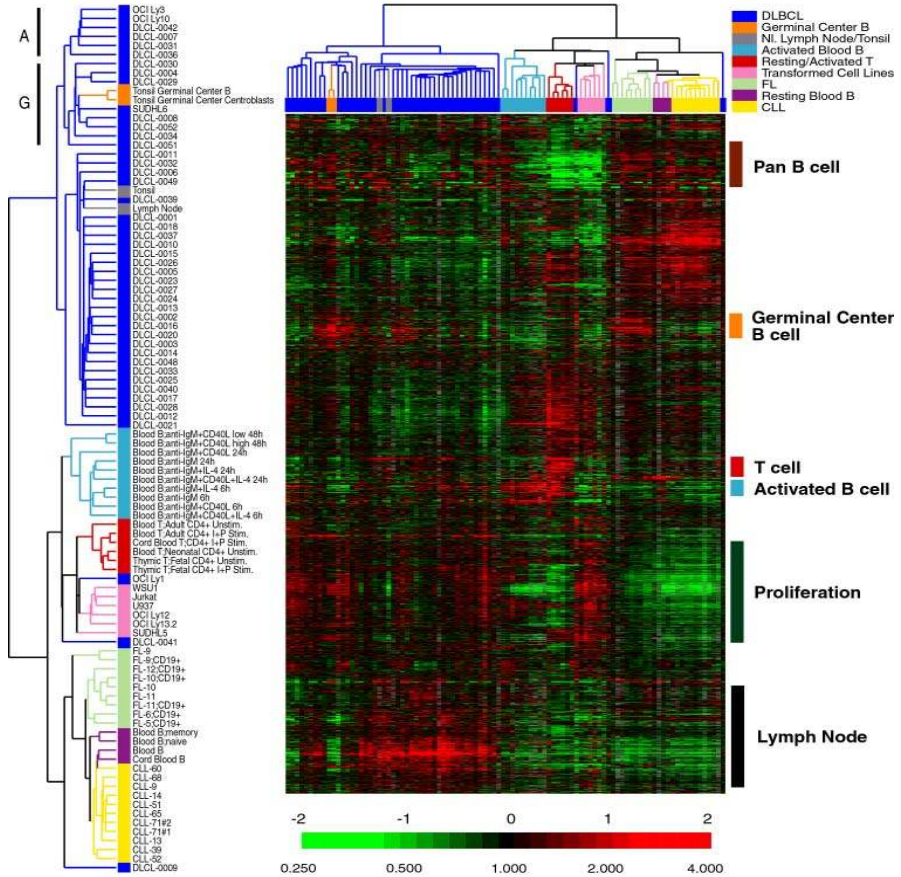

a
b
c
d
e
f
Agglomerative (bottom up)
a
b
c
d
e
f
bc
de
def
bcdef
abcdef
Divisive (top down)

Example output



https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html

You will see this figure in every –omics paper you read



Alizadeh et al. Nature 2000

What does the real world have to say?

BMC Bioinformatics

BioMed Central

Research article

Open Access

Clustering cancer gene expression data: a comparative study

Marcilio CP de Souto*[1,2], Ivan G Costa[1,3], Daniel SA de Araujo[1,2], Teresa B Ludermir[3] and Alexander Schliep[1]

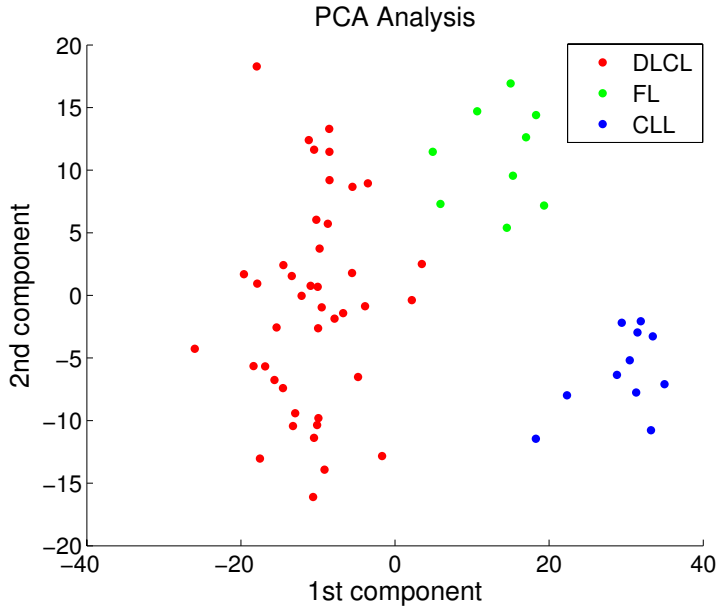
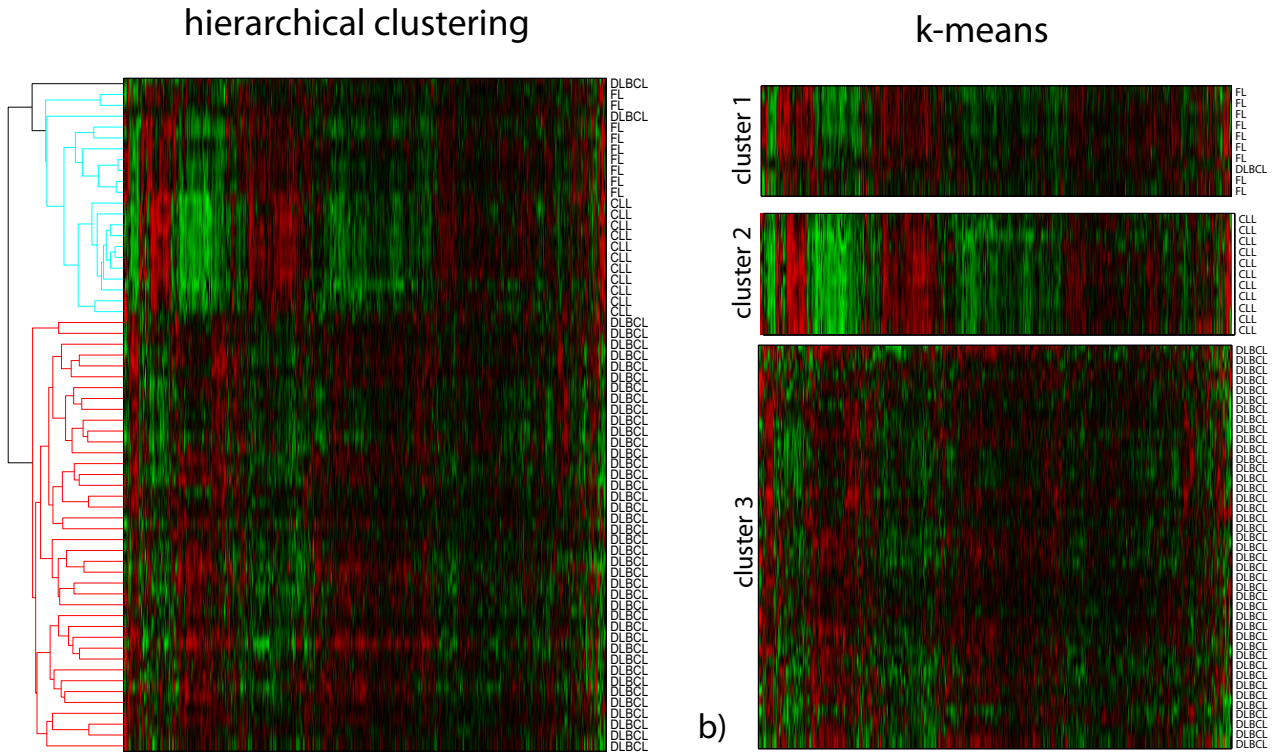


Figure 6
PCA plot for Alizadeh-V2. We display a scatter plot with the two first largest components of a PCA for Alizadeh-V2. Colors indicate the three classes in the data: diffuse large B-cell lymphoma in red (DLBCL), follicular lymphoma in green (FL) and chronic lymphocytic leukemia in blue(CLL).



ENCODE battles

PNAS

Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin[a,b,1], Yiing Lin[c,1], Joseph R. Nery[d], Mark A. Urich[d], Alessandra Breschi[e,f], Carrie A. Davis[g], Alexander Dobin[g], Christopher Zaleski[g], Michael A. Beer[h], William C. Chapman[c], Thomas R. Gingeras[g,i], Joseph R. Ecker[d,j,2], and Michael P. Snyder[a,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305; [b]Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305; [c]Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110; [d]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; [e]Centre for Genomic Regulation and UPF, Catalonia, 08003 Barcelona, Spain; [f]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [g]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742; [h]McKusick-Nathans Institute of Genetic Medicine and the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [i]Affymetrix, Inc., Santa Clara, CA 95051; and [j]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037

Contributed by Joseph R. Ecker, July 23, 2014 (sent for review May 23, 2014)

F1000Research

F1000Research 2015, 4:121 Last updated: 08 NOV 2017

Check for updates

RESEARCH ARTICLE

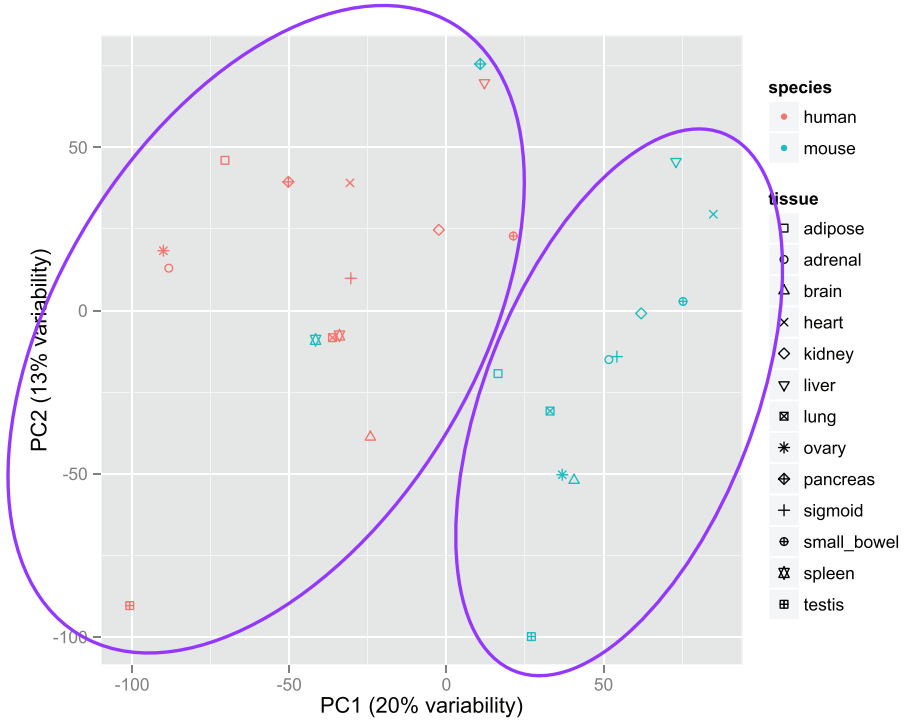
A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]

Yoav Gilad, Orna Mizrahi-Man

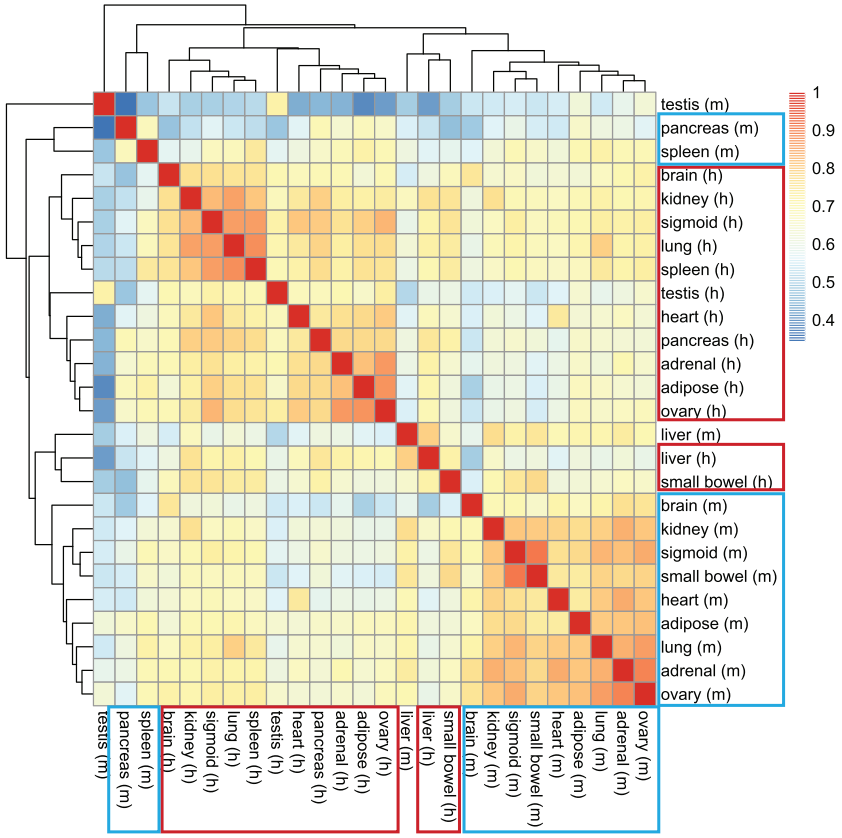
Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

Original findings: Clusters by species

a



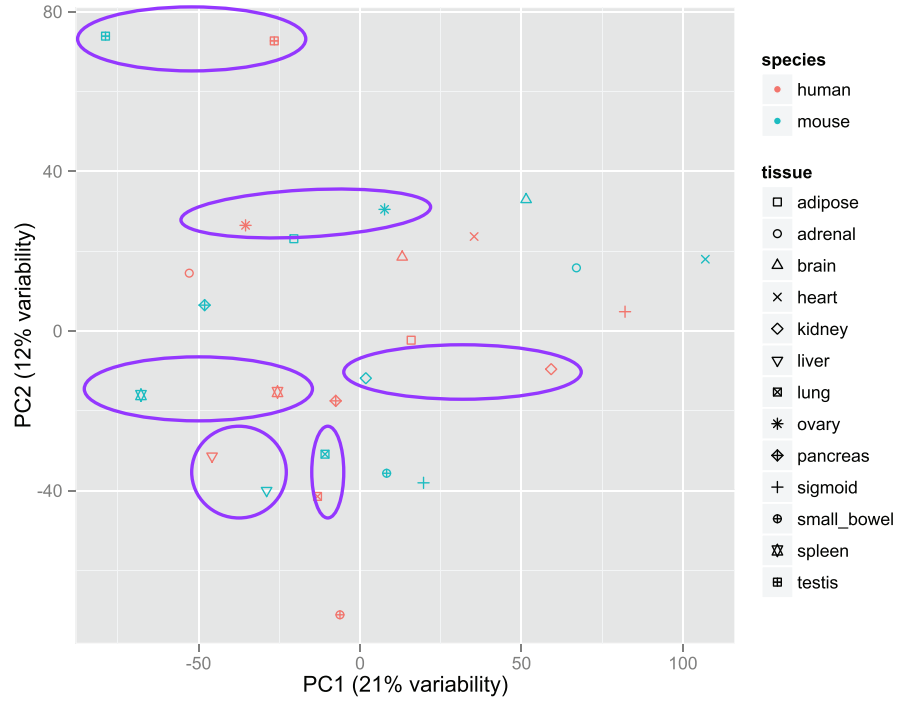
b

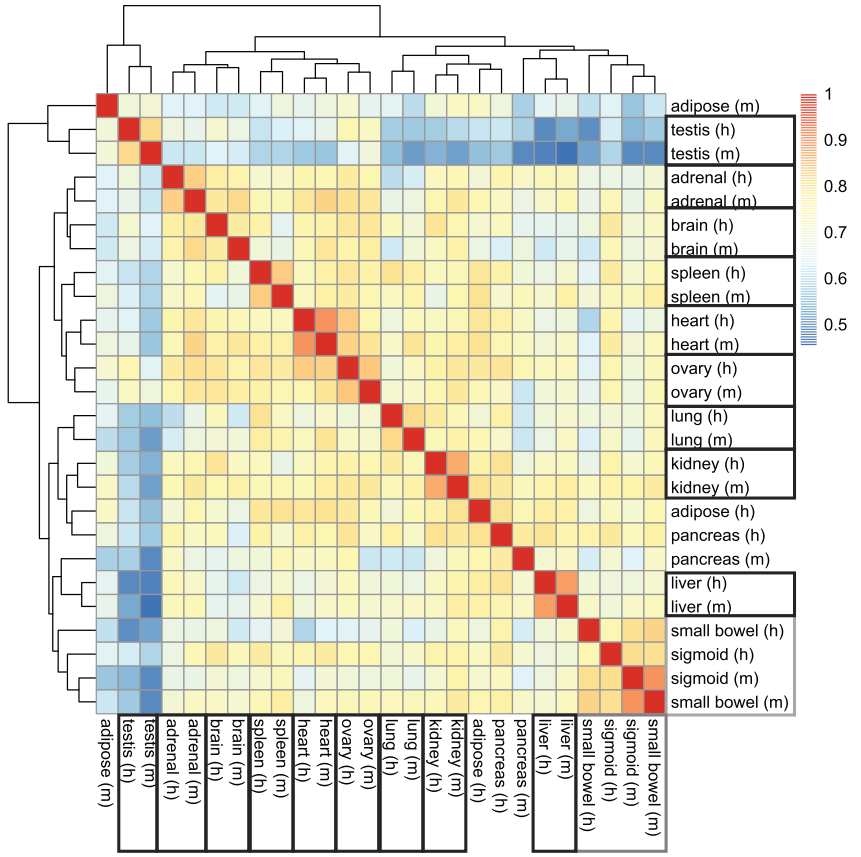


Confounding study design means results dominated by *batch effects*

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Accounting for batch effects changes the story



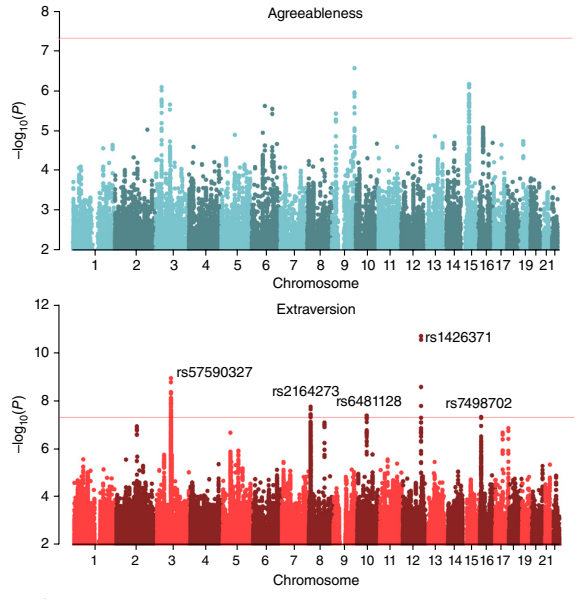


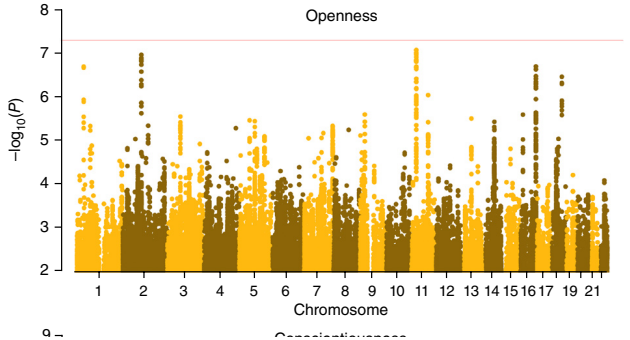
Today's very believable GWAS

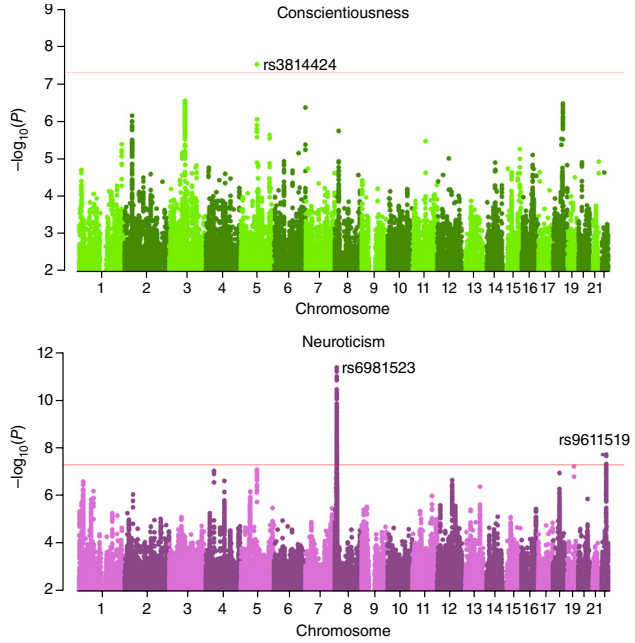
LETTERS

Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders

Min-Tzu Lo[1], David A Hinds[2], Joyce Y Tung[2], Carol Franz[3], Chun-Chieh Fan[1,4], Yunpeng Wang[5–7], Olav B Smeland[6,7], Andrew Schork[1,4], Dominic Holland[5], Karolina Kauppi[1,8], Nilotpal Sanyal[1], Valentina Escott-Price[9], Daniel J Smith[10], Michael O'Donovan[9], Hreinn Stefansson[11], Gyda Bjornsdottir[11], Thorgeir E Thorgeirsson[11], Kari Stefansson[11], Linda K McEvoy[1], Anders M Dale[1,3,5], Ole A Andreassen[6,7] & Chi-Hua Chen[1]

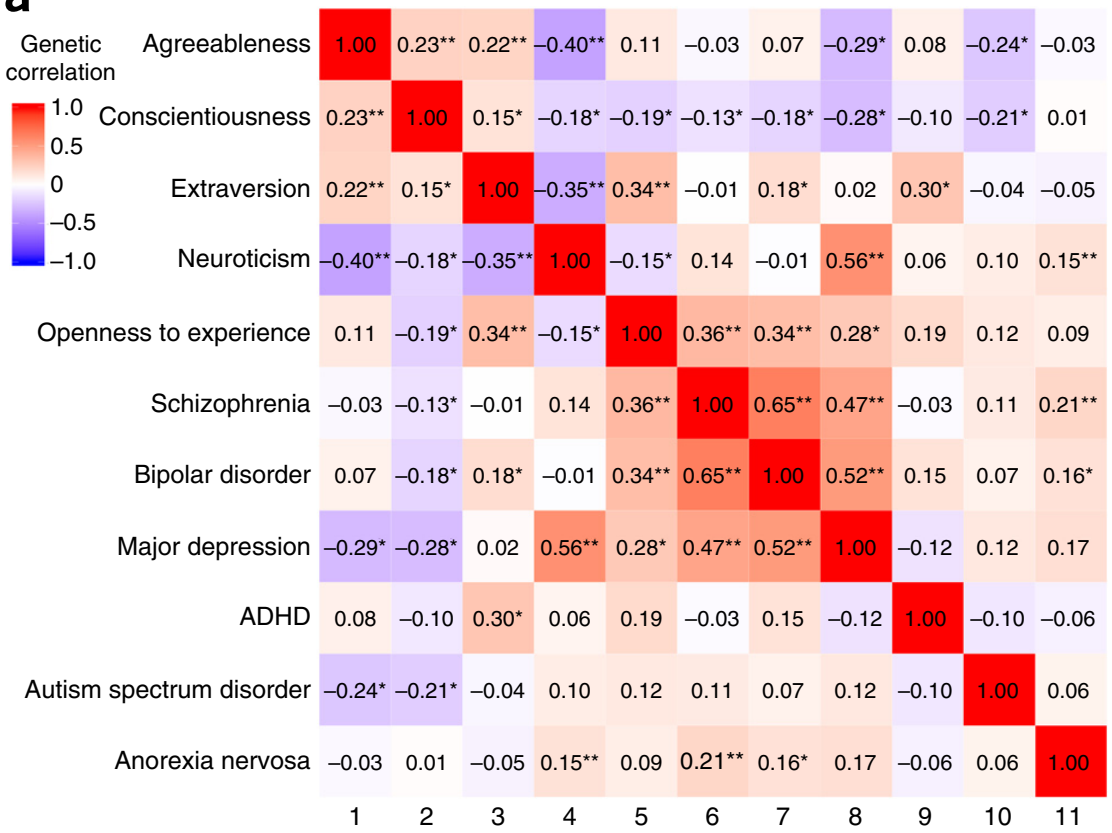




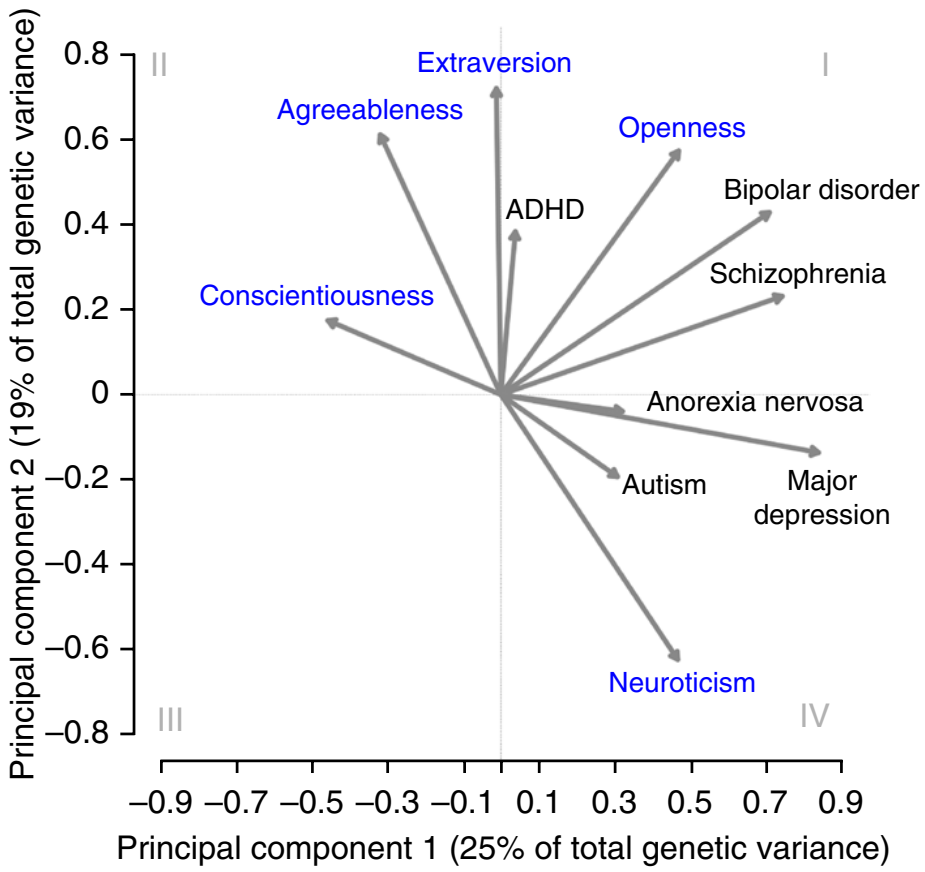


PCA loadings from GWAS

a



b



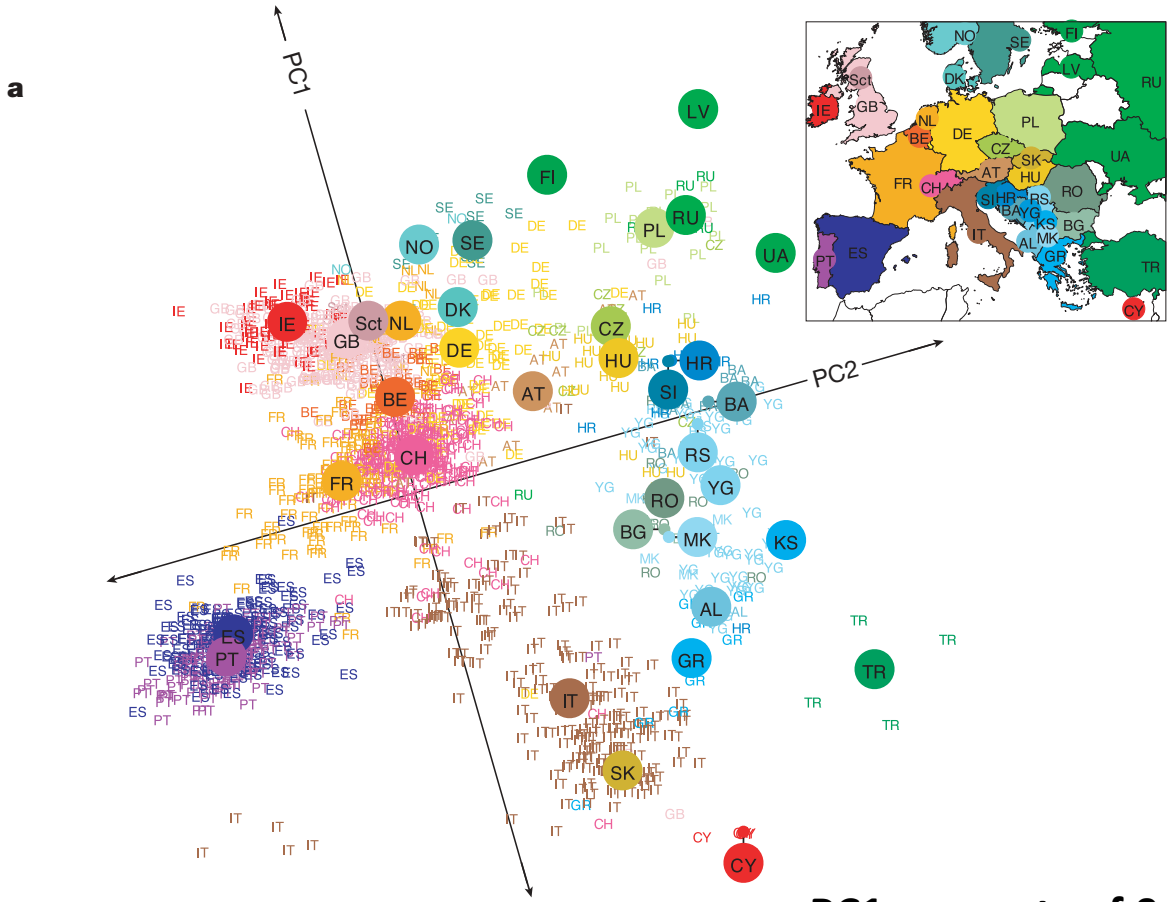
One of the coolest papers

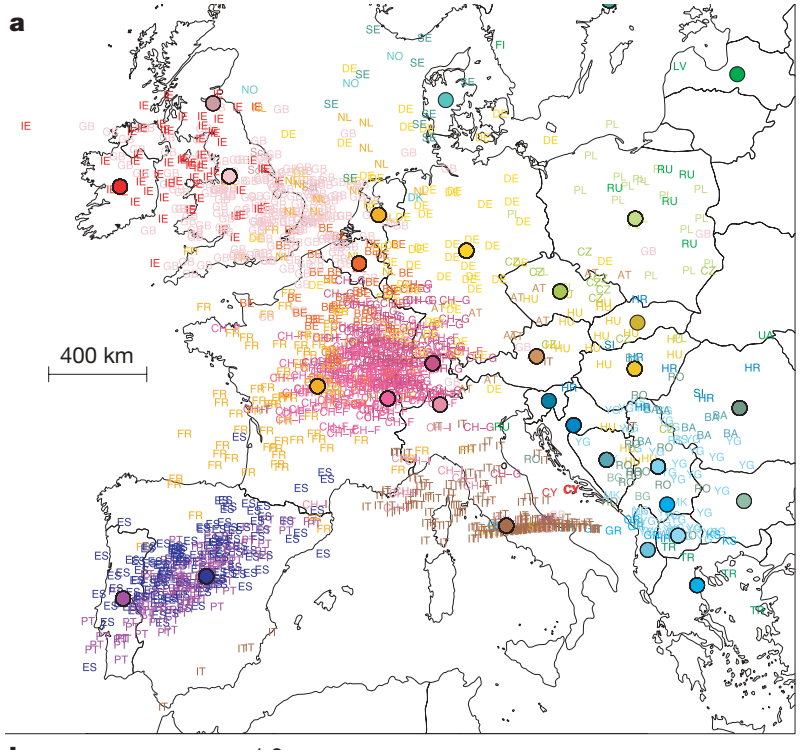
nature Vol 456|6 November 2008|doi:10.1038/nature07331

LETTERS

Genes mirror geography within Europe

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]





PC1 accounts of 0.3% of the variation

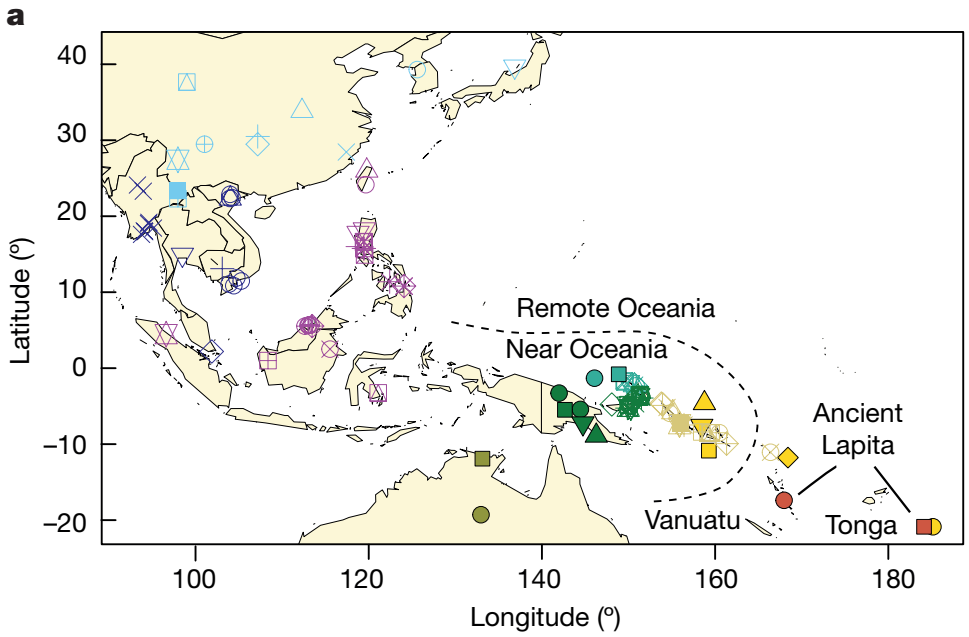
Ancient DNA

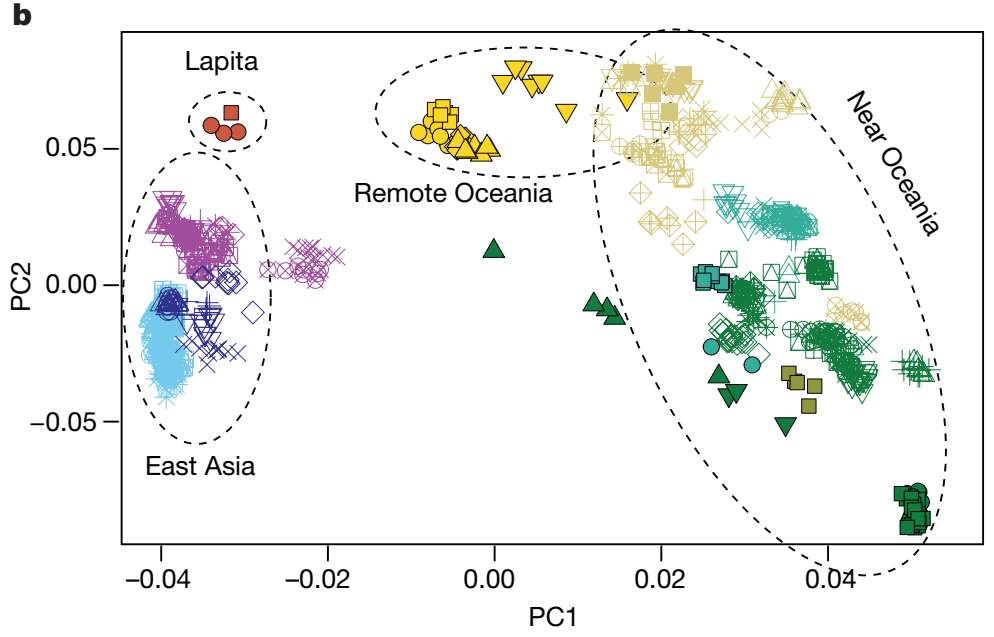
LETTER

doi:10.1038/nature19844

Genomic insights into the peopling of the Southwest Pacific

Pontus Skoglund[1,2,3], Cosimo Posth[4,5], Kendra Sirak[6,7], Matthew Spriggs[8,9], Frederique Valentin[10], Stuart Bedford[9,11], Geoffrey R. Clark[11], Christian Reepmeyer[12], Fiona Petchey[13], Daniel Fernandes[6,14], Qiaomei Fu[1,15,16], Eadaoin Harney[1,2], Mark Lipson[1], Swapan Mallick[1,2], Mario Novak[6,17], Nadin Rohland[1], Kristin Stewardson[1,2,18], Syafiq Abdullah[19], Murray P. Cox[20], Françoise R. Friedlaender[21], Jonathan S. Friedlaender[22], Toomas Kivisild[23,24], George Koki[25], Pradiptajati Kusuma[26], D. Andrew Merriwether[27], Francois-X. Ricaut[28], Joseph T. S. Wee[29], Nick Patterson[2], Johannes Krause[5], Ron Pinhasi[6]§ & David Reich[1,2,18]§





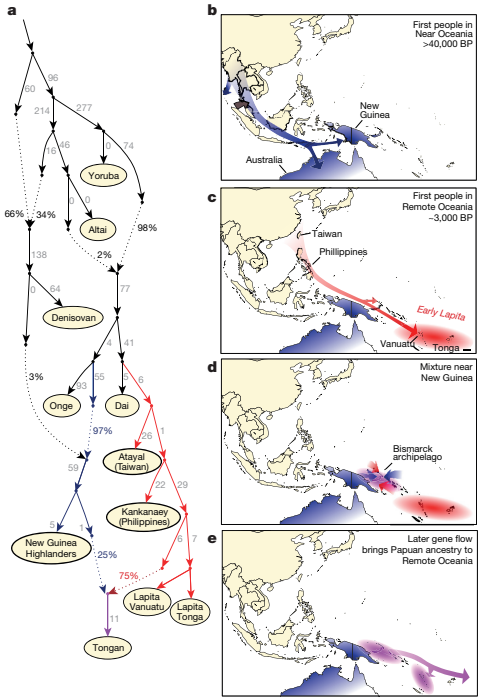


Figure 3 | A model of population history. a, A model of population

Honeybee gene expression

Honeybees show *division of labor* (common in social insects)

- Young worker bees care for broods ("nursing") and transition to foraging at 2-3 weeks
- Change is hormonally determined

Studied 72 bees with 108 microarrays = **high dimensional data**

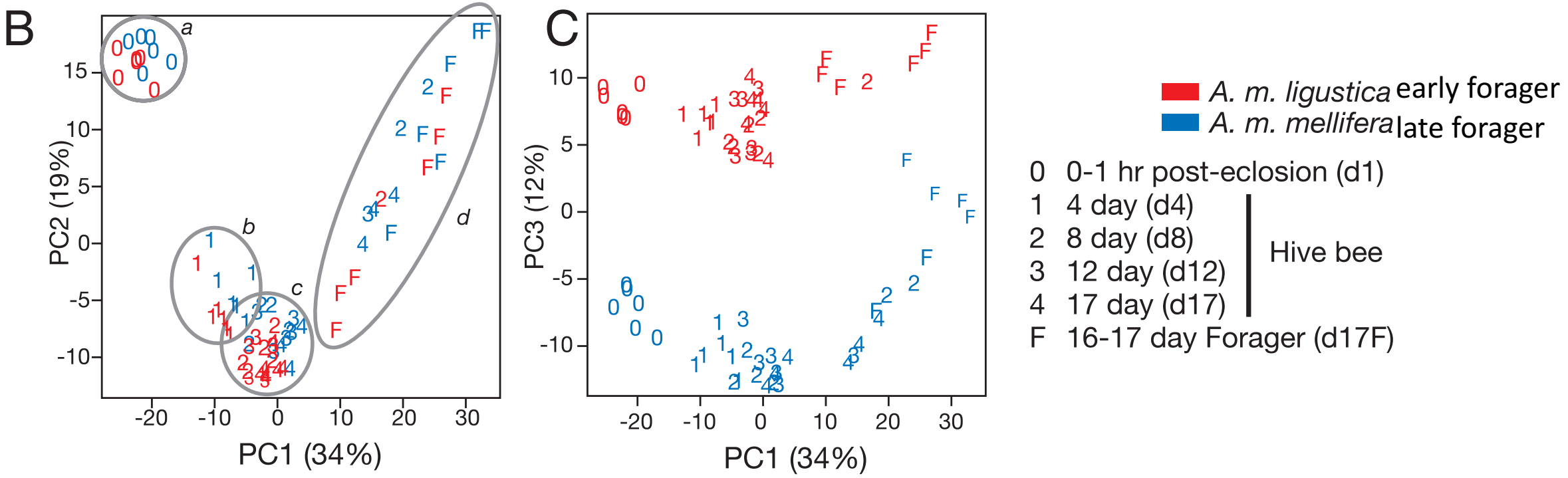
Genomic dissection of behavioral maturation in the honey bee

Charles W. Whitfield*†‡, Yehuda Ben-Shahar§¶, Charles Brillet‖, Isabelle Leoncini‖, Didier Crauser‖, Yves LeConte‖, Sandra Rodriguez-Zas*†‡, and Gene E. Robinson*†‡††**

Departments of *Entomology and **Animal Science, †Neuroscience Program, and ‡Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; §Howard Hughes Medical Institute, ¶University of Iowa College of Medicine, Iowa City, IA 52242; and ‖Laboratoire Biologie et Protection de l'Abeille, Ecologie des Invertébrés, Unité Mixte de Recherche, Institut National de la Recherche Agronomique/Université d'Avignon et des Pays de Vaucluse, Site Agroparc, Domaine Saint-Paul, 84914 Avignon Cedex 9, France

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on May 3, 2005.

PCA on gene expression



People love social insects

Wasp Gene Expression Supports an Evolutionary Link Between Maternal Behavior and Eusociality

Amy L. Toth,[1*] Kranthi Varala,[2] Thomas C. Newman,[1] Fernando E. Miguez,[2] Stephen K. Hutchison,[3] David A. Willoughby,[3] Jan Fredrik Simons,[3] Michael Egholm,[3] James H. Hunt,[4] Matthew E. Hudson,[2] Gene E. Robinson[1,5]