BIS180L
Rice Diversity Lab

Professor MALOOF

Outline

- Part I:
  - Lecture
    - Intro to rice
    - Review of population structure
    - MDS and PCA plots
  - Lab
    - Analyze rice phenotypic and genetic diversity
    - New R Skills
      - Reformatting data tables from wide to long format (and back)
      - Joining data frames with different columns
      - Plotting with ggplot
- Part II (Following lab period)
  - Lecture and Lab:
    - Genome Wide Association Mapping

First: Sensitivity and Specificity

- *Sensitivity:* How likely you are to find a sequence.
  - Increased sensitivity = more sequences detected.
  - But there may be more noise (things that you don't want).
  - Like a microphone: if you have a very sensitive microphone you will detect more sounds but this will include more background.
  - In BLAST, sensitivity comes at the expense of speed
- *Specificity:* How specific are the results to your query?
  - Increased specificity = fewer distant or unrelated sequences
- Word size
  - Small word size increases sensitivity.
    - Why? BLAST searches start when a word from the query matches the subject. Smaller words are more likely to match
  - Small word size slows down the search (more initial hits)
  - Large word size decreases sensitivity
    - A large word is less likely to find a match.

Assignment Dos and Don'ts

- DO
  - Look at your formatted html to make sure it is nicely formatted and tables are correct
  - Include code so that we can troubleshoot and give partial credit
  - Make sure code is in code blocks
  - Comment out R code that downloads files or installs packages (or use eval=FALSE in the chunk option)
  - Load libraries at beginning of file
  - Questions and Answers should be on separate line

- DON'T (you will lose points)
  - Rename the template files
  - Turn in poorly formatted work.
  - Display extremely long tables

Bad (top) vs Good (bottom)

Exercise Three Modify the above loop to produce the following output

mkdir for_example myfiles=$(ls) cd for_example echo "this" > file1.txt echo "is" > file2.txt echo "silly"> file3.txt

#echo ${myfiles}

for file in $myfiles do cat $file done

for file in $(ls) do echo "file" ${file} "contains: $(cat $file)" done

Exercise Three Modify the above loop to produce the following output

```
for file in $myfiles
  do
    echo "file $file contains: $(cat $file)"
  done
```

Bad (top) vs Good (bottom)

Word size	real time	Unique hits
9	1m45.2s	2506
11	25.34s	2506
13	14.99s	2503
15	7.466s	2485
17	5.999s	2419
19	5.100	1790
28	3.100	2506
11 (megablast)	23.83s	2506
11 (dc-megablast)	24.01	2506
11 (blastn)	23.79	2506
blastpp	17m26s	0
Exercise 11: Use a for loop that builds on the command above to count the number of unique hits in each file. Add these results to the table above. Did word size affect time and sensitivity in the direction you predicted from Exercise 9? (FYI a word size of 7 takes 16 minutes to run, but I decide to spare you that pain). If you didn't already do this above, explain why word size is affecting the results in the way that it does.		

Search Method	Search Time	# Unique Hits
Megablast	24s	2506
dc-Megablast	1m5s	2501
blastn	47s	2509
blastp	17m55s	2695

Bad

Exercise One:Write a for loop to pluralize peach, tomato, potato (remember that these end in "es" when plural) veggies="peach tomato potato" for veg in ${veggies} do echo "${veg}es" done **Exercise Two** In your own words provide a "translation" of the above loop. For each file stored in the variable myfiles, display the contents of the file. **Exercise Three** Modify the above loop to produce the following output for file in $(ls) do echo "file ${file} contains: $(cat $file)" done

Very Bad (top) vs Good (bottom)

Exercise 1: What is the sequence format for these sequences? Do the files contain RNA or DNA sequences? They contain DNA sequences in a FASTA format. **Exercise 2**: For the refseq file, the header line contains a number of fields, separated by "|". For the first entry in the refseq file, try to figure out what each header is. List each field and what you think it is.
Accession number, sequence title, sequence completeness, viral species, country of isolation, host species **Exercise 3**: How many of the viruses come from a domesticated cat (*Felis catus*) host? How many come from a human host? How many were isolated in the United States? Show the commands used to answer these questions. Number of viruses from a domesticated cat: 276 Command used: `zgrep -c ">*|*|*|*|*|Felis catus" ncbi_virus_110119_2.txt.gz` Number of viruses isolated from human hosts: 29,911 Command used: `zgrep -c ">*|*|*|*|*|Homo sapiens" ncbi_virus_110119_2.txt.gz` Number of viruses that originated in the United States: 19,893 Command used: z `grep -c ">*|*|*|*|USA" ncbi_virus_110119_2.txt.gz` **Exercise 4**: Create a simple (markdown formatted) table with the following information for the refseq and patientseq files: To find file size: `ls -lh ncbi_virus_110119_2.txt` To find number of sequences: `grep -c ">" ncbi_virus_110119_2.txt` To find number of base pairs: Two steps

Exercise 2: For the refseq file, the header line contains a number of fields, separated by "|". For the first entry in the refseq file, try to figure out what each header is. List each field and what you think it is.

Fields:

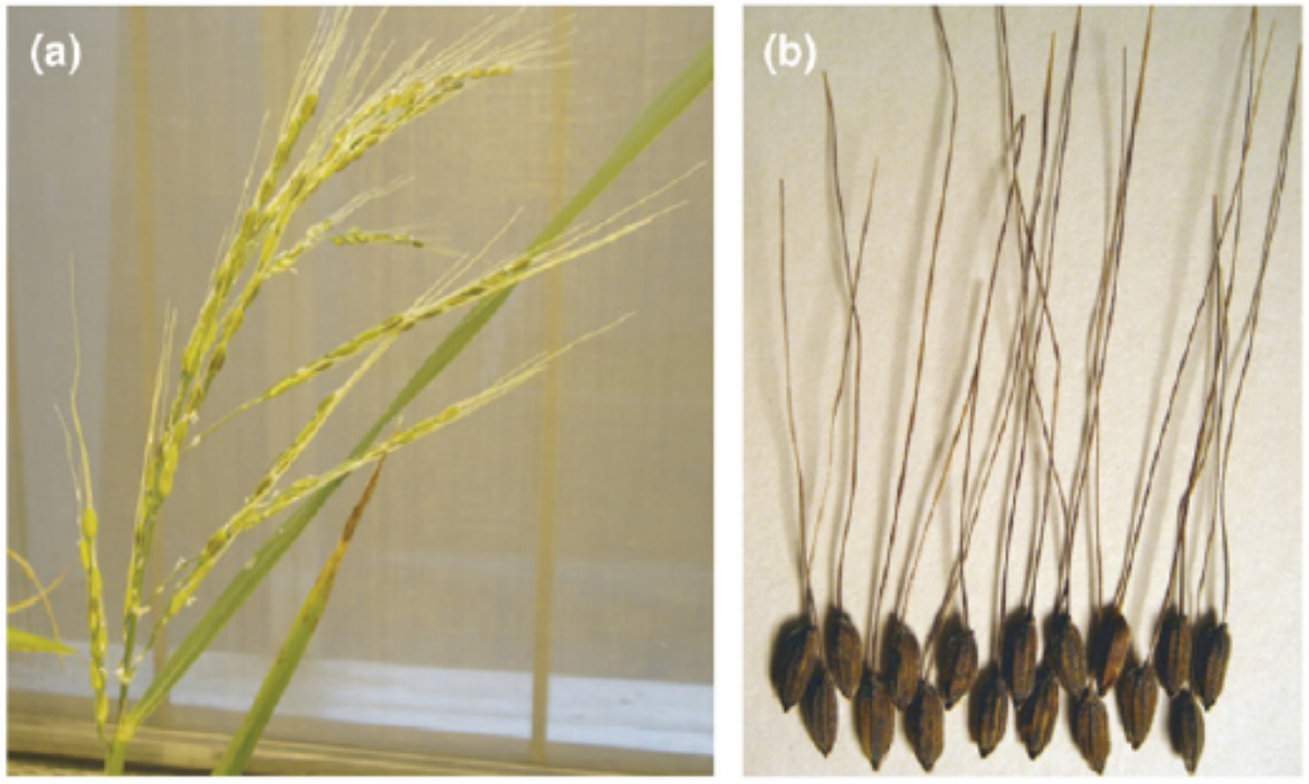
1. GenBank accession
2. Sequence title
3. Sequence completeness
4. Viral species
5. Country where virus isolated
6. Host genus and species (mystery field)

Study system: Rice

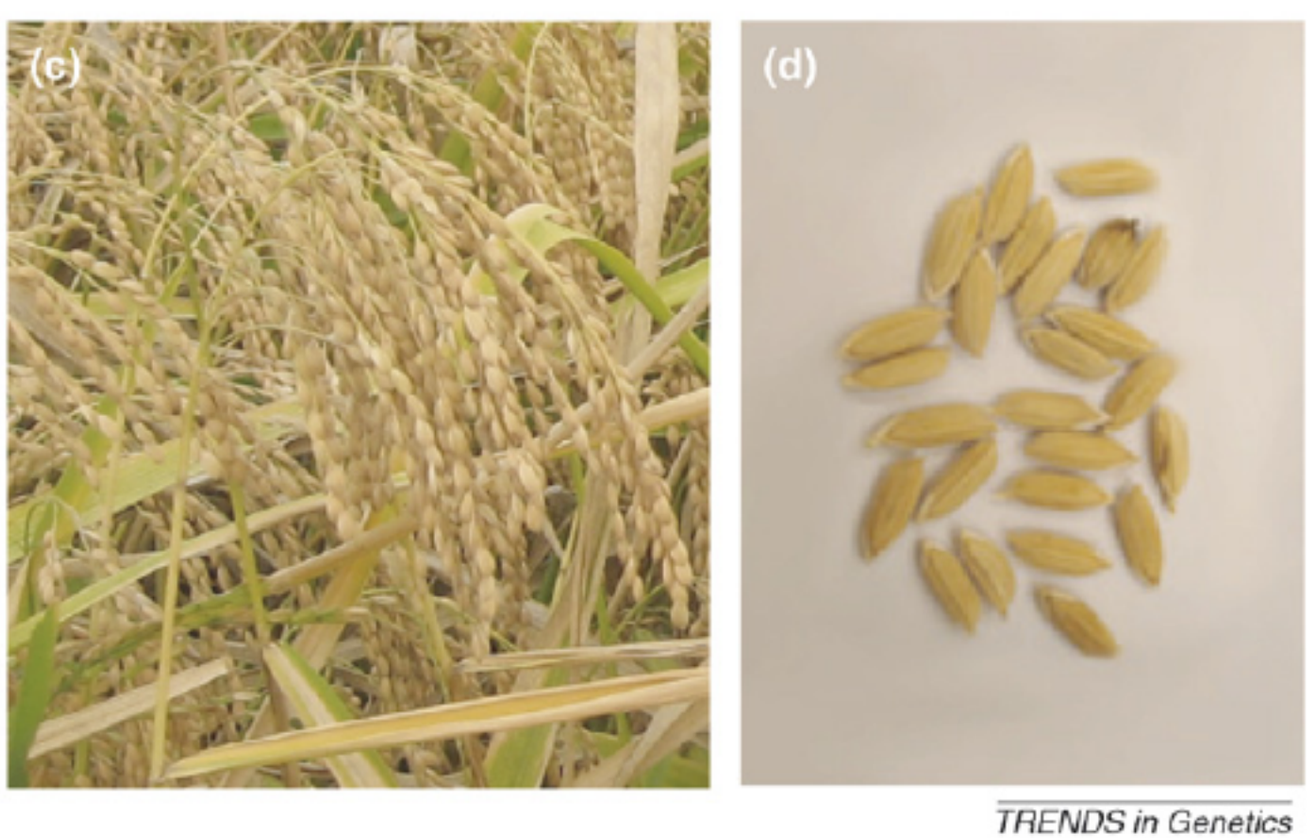
- This Week:
  - Study diversity in ~ 400 rice strains

- Why Rice?

Rice progenitor:
Oryza rufipogon



Domesticated rice
Oryza sativa



Kovak et al, Trends in Genetics, 2007

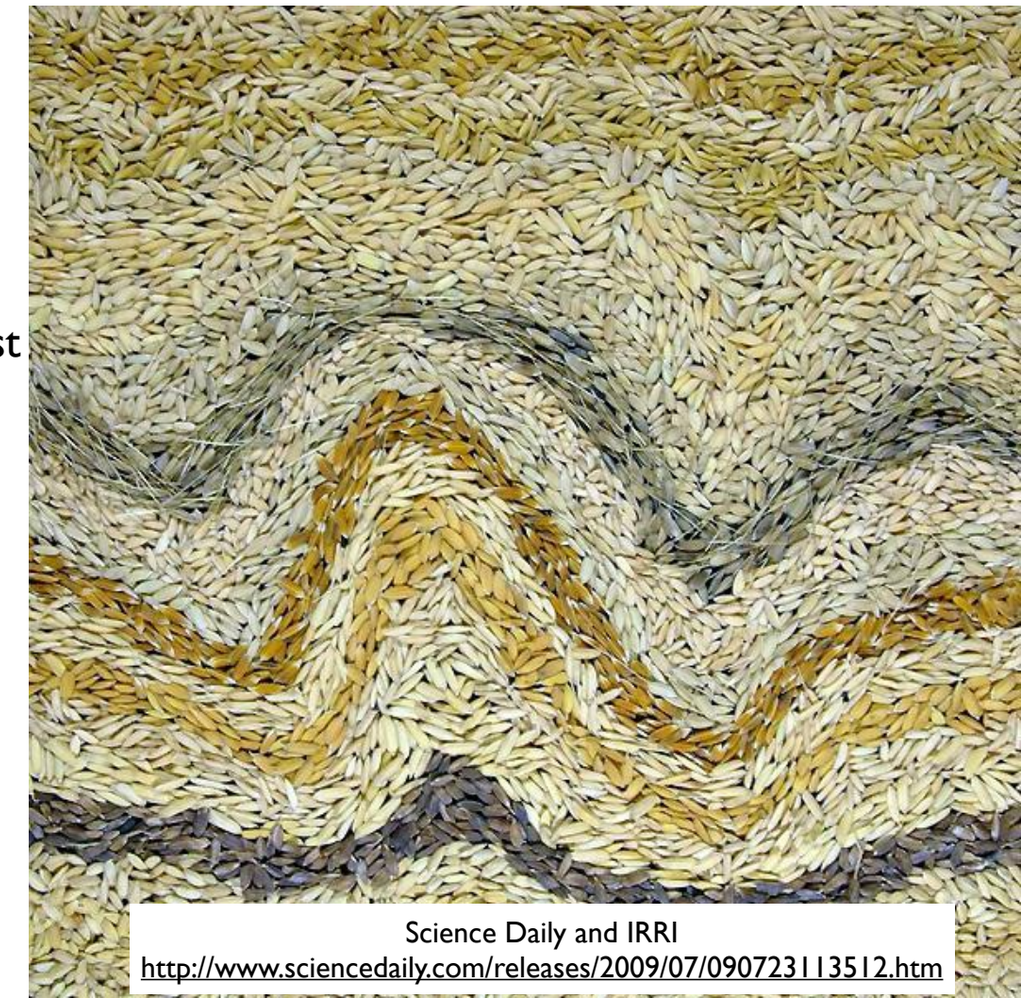
Rice Diversity

- Two main sub populations: *Indica* and *Japonica*
  - Thought to have been domesticated independently from *rufipogon*
  - Estimated divergence time: 100,000 years ago or more.
- *Indica*
  - Grown in lowland tropical areas
  - S. and S.E. Asia and China
  - Further subdivided into *indica* and *aus*
- *Japonica*
  - Grown in lowland and high-elevation upland areas of tropical SE Asia
  - Also grown in colder temperate climates including NE Asia, Europe, Western US. Chile, Australia
  - Subdivided into *tropical japonica*, *temperate japonica*, and *aromatic*

Using natural variation for rice improvement

- Urgent need for crop improvement:
  - World population expected to grow to 9,000,000,000 by 2050
  - Climate change and increasing unpredictability will reduce yield
  - Increasing demand for meat and biofuels put further strains on agriculture

- One path forward is to use the natural genetic diversity ("natural variation") already present in rice
  - 120,000 different rice strains have been deposited in seed banks
  - These harbor different genetic variants
    - Many are just random
    - some provide adaptation to specific environments/stresses
      - drought, flooding, heat, pathogens, etc
    - some determine specific grain characteristics of consumer interest
      - stickiness
      - grain length
      - aromatic (basmati, jasmine)
      - color
- Identify genes or genomic locations of variants
  - Why?



Science Daily and IRRI
http://www.sciencedaily.com/releases/2009/07/090723113512.htm

We will focus on root traits

- Why Roots?

EU Root/Adam Price. http://www.euroot.eu/gallery/album1/

Root Aluminum Tolerance

- Aluminum: Most abundant metal in earth's crust.
- Solubilized under highly acidic soil conditions
- Toxic to plant roots
- Can we find variation in resistance to Aluminum among rice varieties?

Famosa et al, Plant Phys., 2011

Study Design

- 413 Rice varieties from 82 countries
- Phenotypes:
  - Roots in control and Al+ conditions
  - Grain length
  - Amylose content
  - Flowering time
  - ...
- Genotypes:
  - Sequence 20 strains, find common SNPs
  - Design Affymetrix "SNP chip"
  - Assay 44,000 Single Nucleotide Polymorphisms on each of the 413 varieties
  - Yields ~ 1 SNP per 10kb

Overview of Questions about Rice Data:

- Population Structure:
  - Are these samples from a single randomly mating population or are there sub-populations?
    - Multi-dimensional scaling plot of genotype data to examine diversity and relationships
    - fastStructure to assign varieties to ancestral populations
- Can we find SNPs associated with Aluminum tolerance or other traits?
  - What are the genes underlying these SNPs?

Principal Components Analysis (PCA) and Multi-Dimensional Scaling (MDS)

- SNP Data is multi-dimensional; each SNP site can be considered an axis
- PCA and MDS techniques are ways to visualize this multi-dimensional data in 2D
- PCA:
  - Find the vector through the data that explains the most variance. This is the first principal component
  - Find the vector that explains the most remaining variance. This is the second principal component.
  - repeat...
- MDS
  - Project from multiple dimensions onto 1 or 2 dimensions in a way that preserves distances present in multi-dimensional space
-

population structure

- Project SNPs in reduced dimensional space.
- Clumps of individuals represent population structure

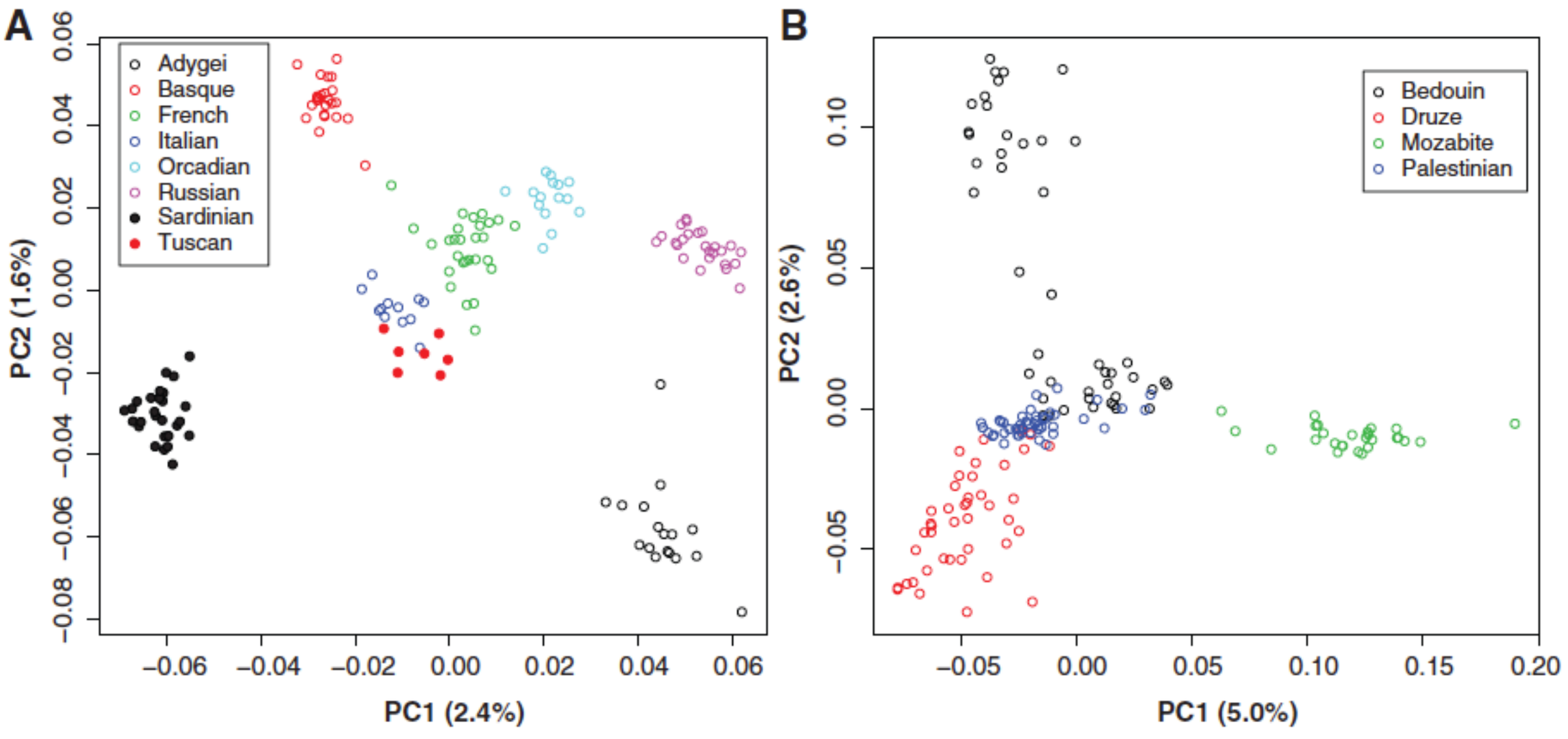
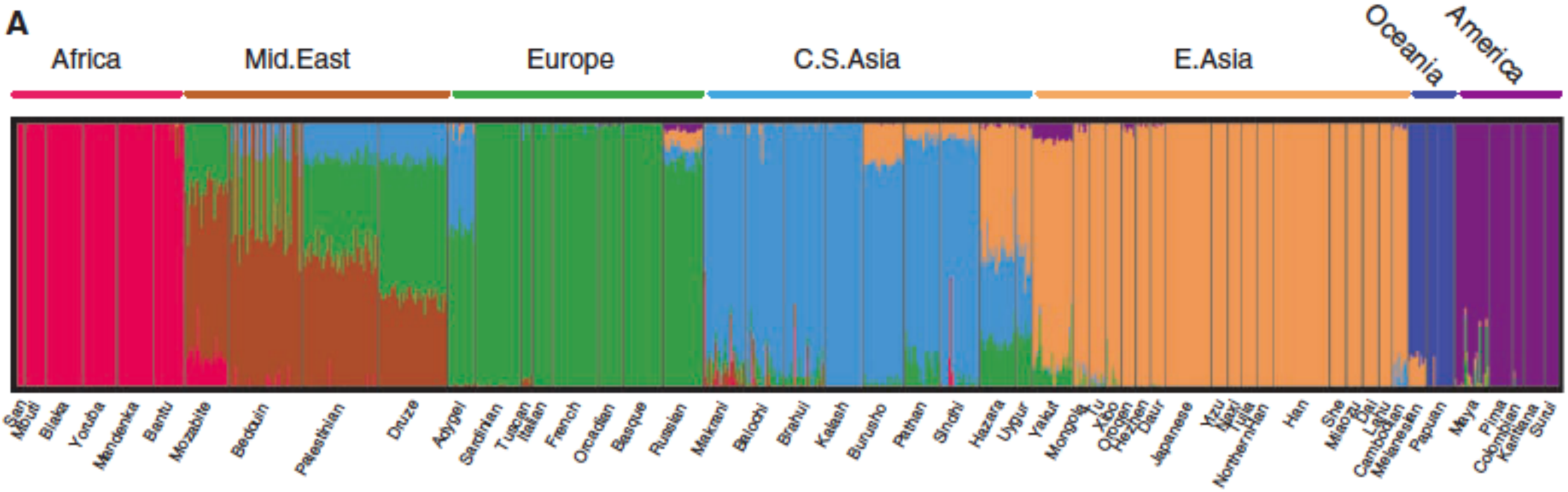


Fig. 2. Fine-scale population structure principal component analyses in two geographic regions, using all autosomal SNPs. (**A**) Europe. (**B**) The Middle East.

Li et al (2008) Science

Model based assignment of individuals to populations



- a priori decide on the number of likely ancestral populations
- use an evolutionary genetics model to assign the most likely ancestry to each individual

Questions to be considered:

- Is there population structure?
  - How does structure relate to region of origin?
  - How does structure relate to amylose content?
  - How does structure relate to Aluminum tolerance
- Are there GWAS hits for the trait you are studying (TBD)?
- What are the candidate genes for your GWAS hit?