"Genomes, Genetics, and Diversity"

New sequencing technologies are emerging driven by human biomedical applications. This talk will introduce the characteristics of these new datasets, and outline the computational hurdles in both converting the raw data into useful genetic information, and in answering questions

related to population structure, dynamics, and evolution.

Dan Rokhsar

DOE Joint Genome Institute

UC Berkeley Center for Integrative Genomics

The big picture

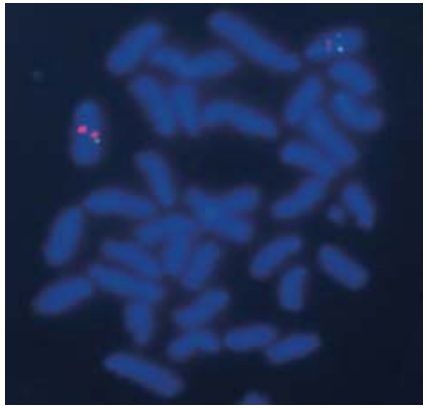
- How does genetic variation, in concert with the environment, lead to “phenotypic” variation?
  - Phenotype = behavior of the biological system (morphology, disease, efficiency, etc.)

- What genetic and phenotypic variation is present, and how is it structured in populations of organisms? Species?

Three computational/ algorithmic challenges

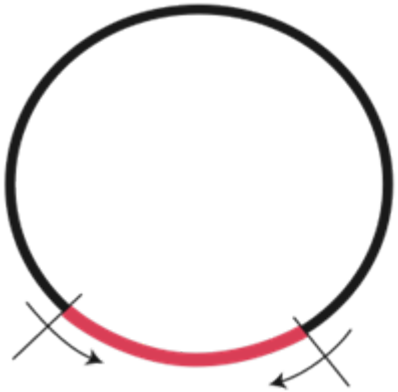
- 1. High throughput genome sequencing and “tagging” technologies
- 2. Statistical genetics: genomic variation pedigrees, and populations
- 3. Systems biology and the quantitative description of complexity

whole genome shotgun sequencing



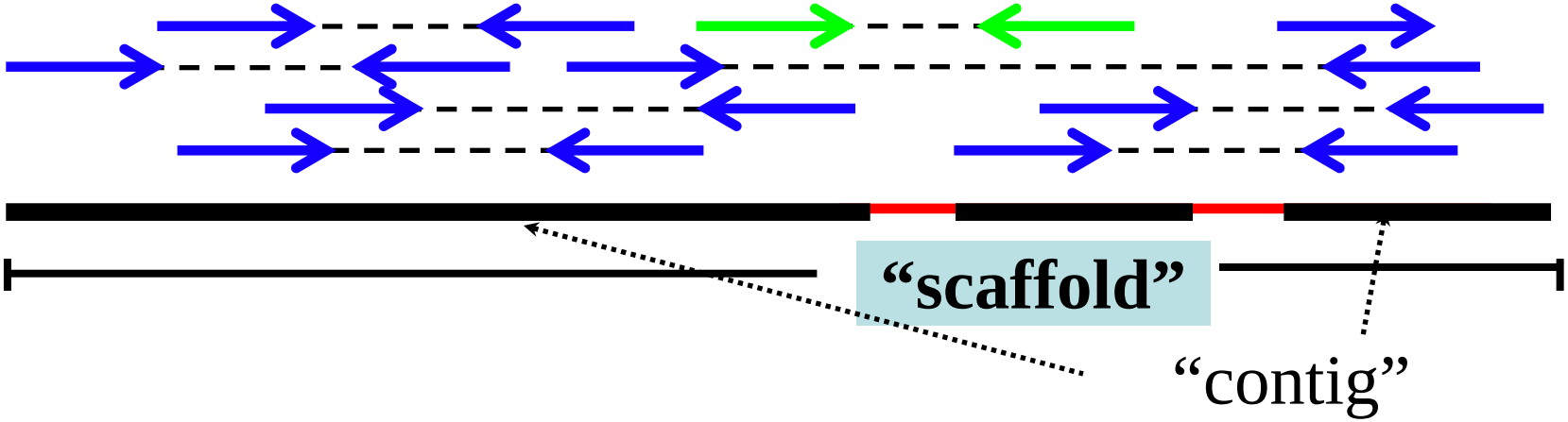
1. Isolate genomic DNA from one inbred individual, if possible.

2. Shear

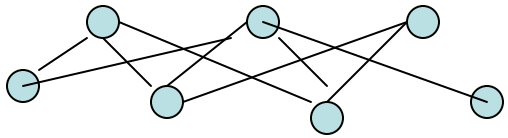


3,4. Clone & Sequence

5. Assemble



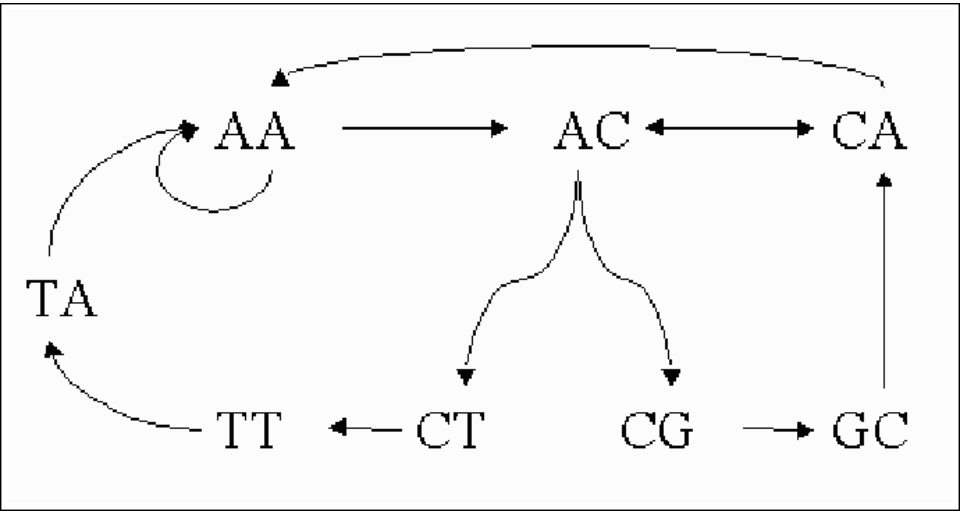
Two graphical formulations



Node = “read”

Edge = read-read alignment

Seek consistent path



Decompose into k-mers

Node = k-mer

Edge = maximally overlapping k-mers in dataset

Assembly = traversal

Scales

- Genome sizes
  - Fungi ~ 40 Mb (500K reads)
  - Human, mammals ~ 3 Gb (30M reads)
  - Wheat, pine ~ 15-20 Gb (200M reads)
- To find reads that overlap on genome, must seek all pairwise alignments
  - Use hashing schemes to avoid N^2 work
  - Can be readily parallelized
  - Repeats in genome are problematic (reads w/many alignments)
- Problem of determining optimal reconstruction of genome is hard …
  - Heuristics needed to handle various additional constraints. Big assemblies take weeks/months

New sequencing technologies: cheaper, faster (& crappier)

Driven by market for human resequencing
-- shorter reads, lower quality

Enables sequence as a "hammer" to solve other problems in new ways
-- genetic markers
-- resequencing desirable lab-created variants
-- sampling natural variation in populations

? bp/run

1-2 Billion bp/run

100-300 Million bp/run

ABI 3730: 70 kilo-bp/run

New Technology: quality v. cost

- Solexa, ABI-SOLID, …
  - Dramatically lower cost per raw base (1000X)
  - Dramatically reduced read length (~30-50 bp) & quality
- Development driven by market for human resequencing
  - need to recognize 1 SNP/1000 bp; “$1000 genome”
- Why so cheap?
  - Highly multiplexed (e.g., 10^7 fragments/experiment)
  - Optical readouts -- watch dNTP -> DNA + Pi reaction!
- Cost for another maize genome
  - Sanger: $15M (~6.5X, ~$1/Kb) high quality draft
  - Solexa: $360K (~50X, ~$3/Mb)!!! But what quality?
- On the way: single molecule methods (<five years). Terabases/day or more.

Challenges/rewards

- One run ~few days, 30 million “reads” per instrument.
- Aligning reads to reference on ~20 cpu can take ~few days
- Aligning reads to each other for de novo assembly (w/o reference to guide) takes a bit longer. Parallelizable.
- Resolving graph into sensible assembly is where algorithmic development is needed.
- >Thousandfold reduction in cost of genomes.

Can use short reads to identify variation vs. a reference sequence

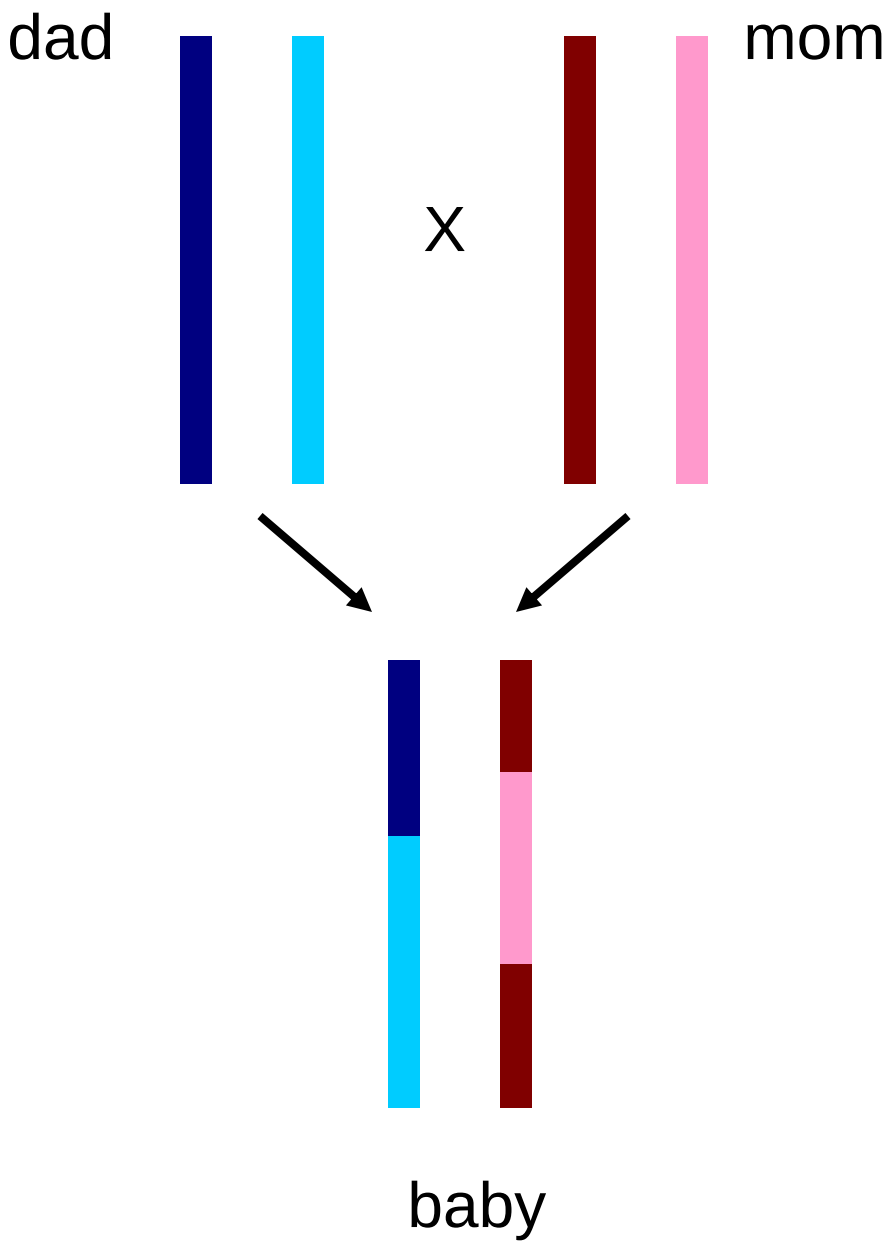
```
MELD              atcgcgtgtagctagtgctagctgctagctagctgatgca
ROM1_read12667    ............t.........................
AWA1_read00345    ..............................
SBF1_read06734    ..............................
TEX1_read00234    ............t.........................
ROM1_read10385    ............t.........................
TEX1_read39890    ..............................
```

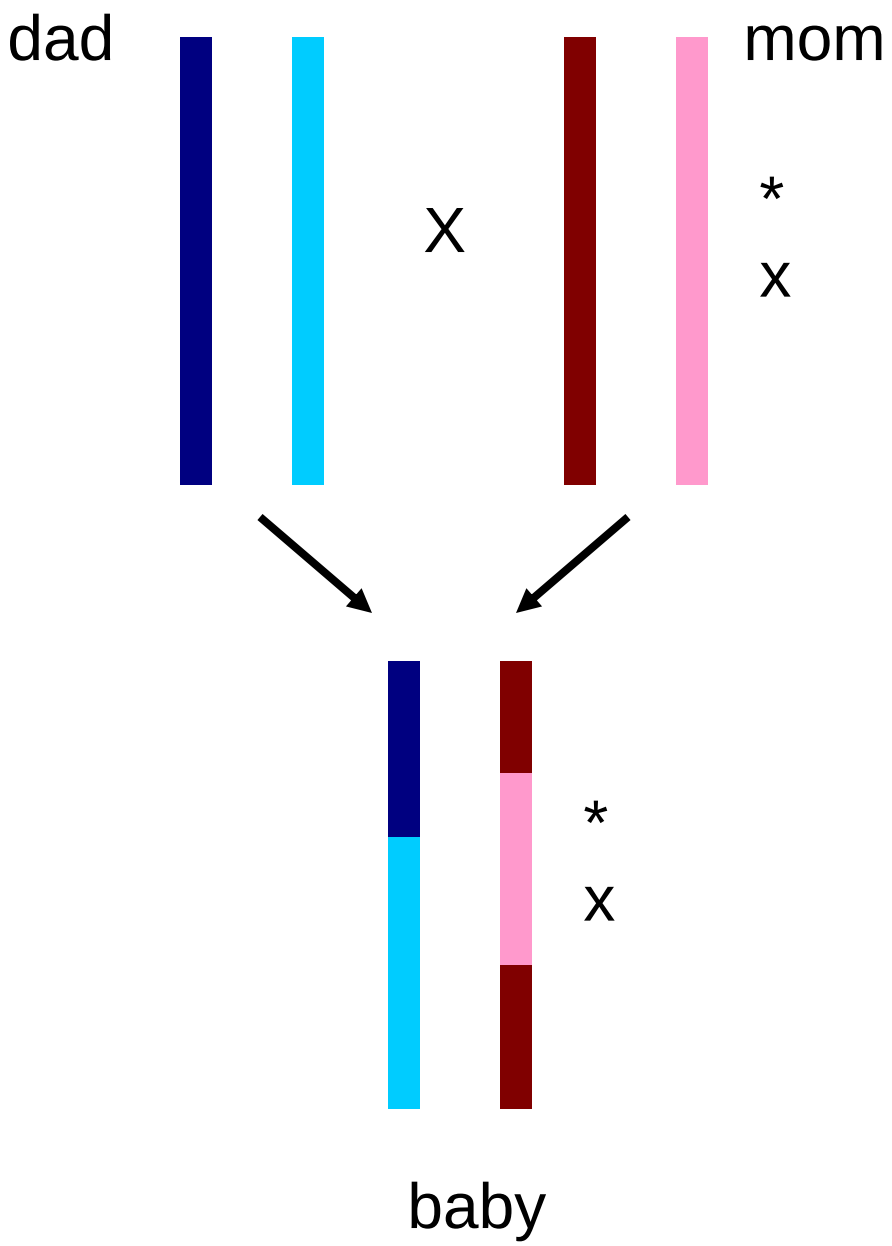
Figure 3. High quality SNP detection

- Problem: rapidly align 100 million short reads to best position in human genome.
- Identify true variations and reject false positives, using error model learned from data

crash course in genetics

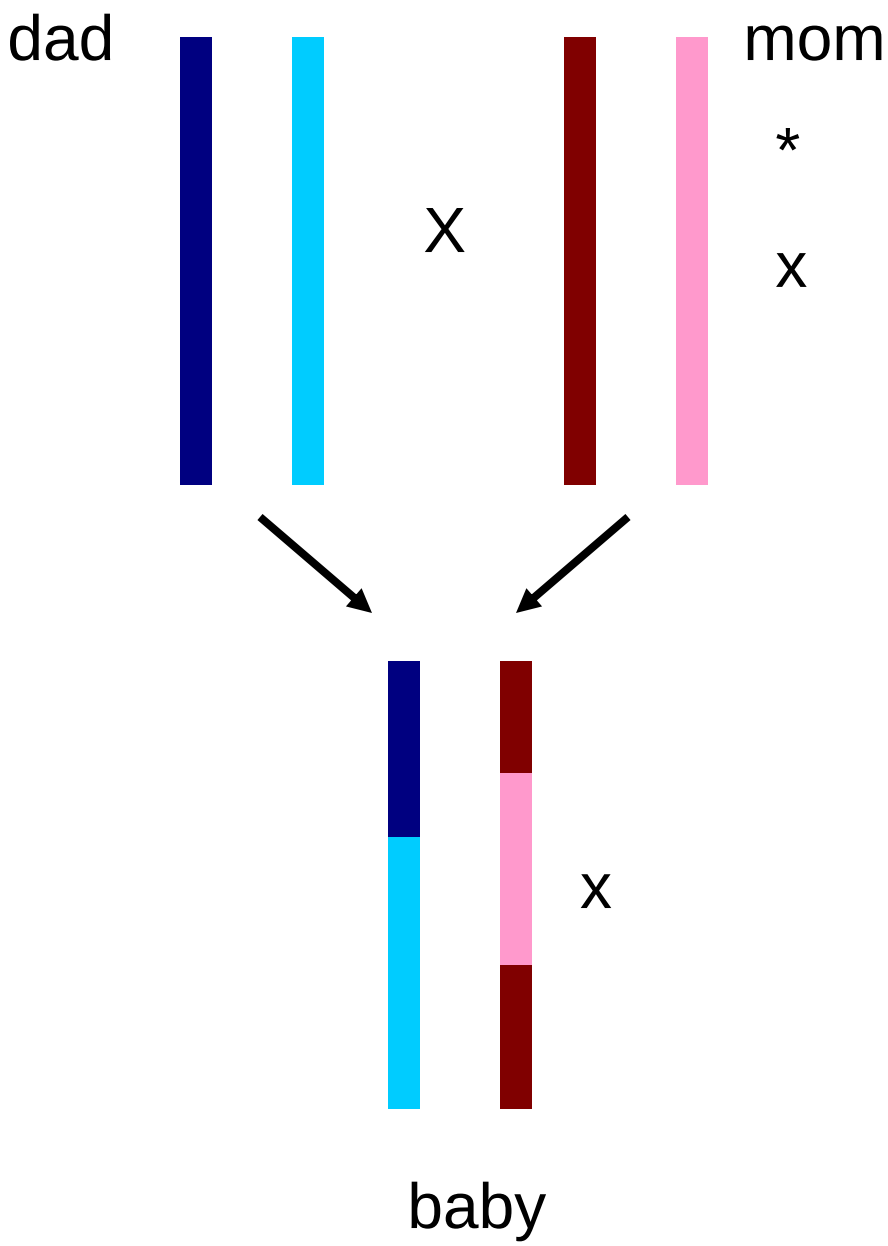
- A chromosome is a single DNA molecule, 10's or 100's of millions of "letters" long,harboring thousands of genes, associated regulatory sequences, and other stuff ("junk")
- Except for X and Y, somatic cells have two "copies" of each chromosome, one each from mom/dad
- The copies differ in detail (humans 1/1000 bp have single letter changes "SNPs"; also hundreds of small inversions, local duplications)
- When sperm/egg (or pollen/ovule, or spore) are produced, the two copies line up and exchange ~1-2 segments to make new mosaic sequences. One such mosaic is passed on to offspring from each parent.
- Chromosome pair of offspring is then one mosaic of maternal grandparental sequences, and another mosaic from paternal grandparents.


dad
mom
X
baby



If x is a causal variant and * is some other linked variant, then * and trait may co-occur in offspring

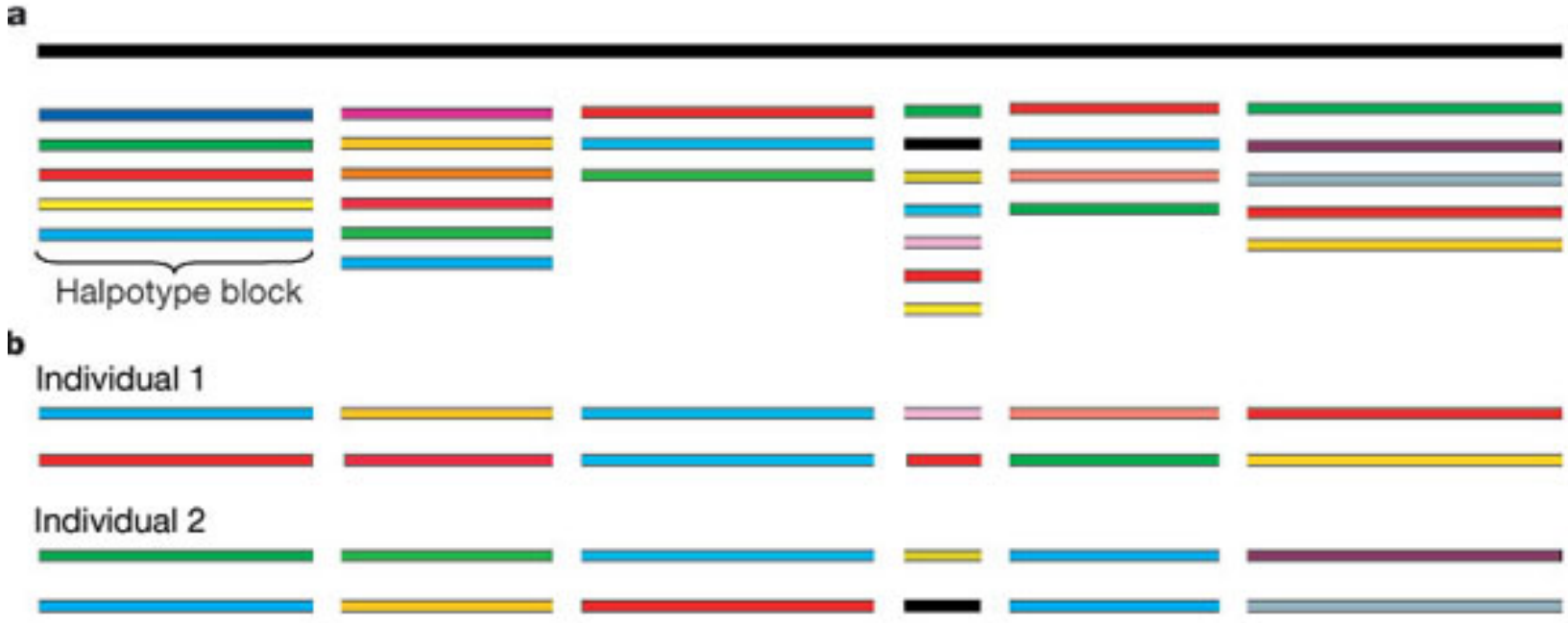
Over time, recombination will eventually separate * and x.



Variations (“markers”) that are farther apart will not be co-inherited

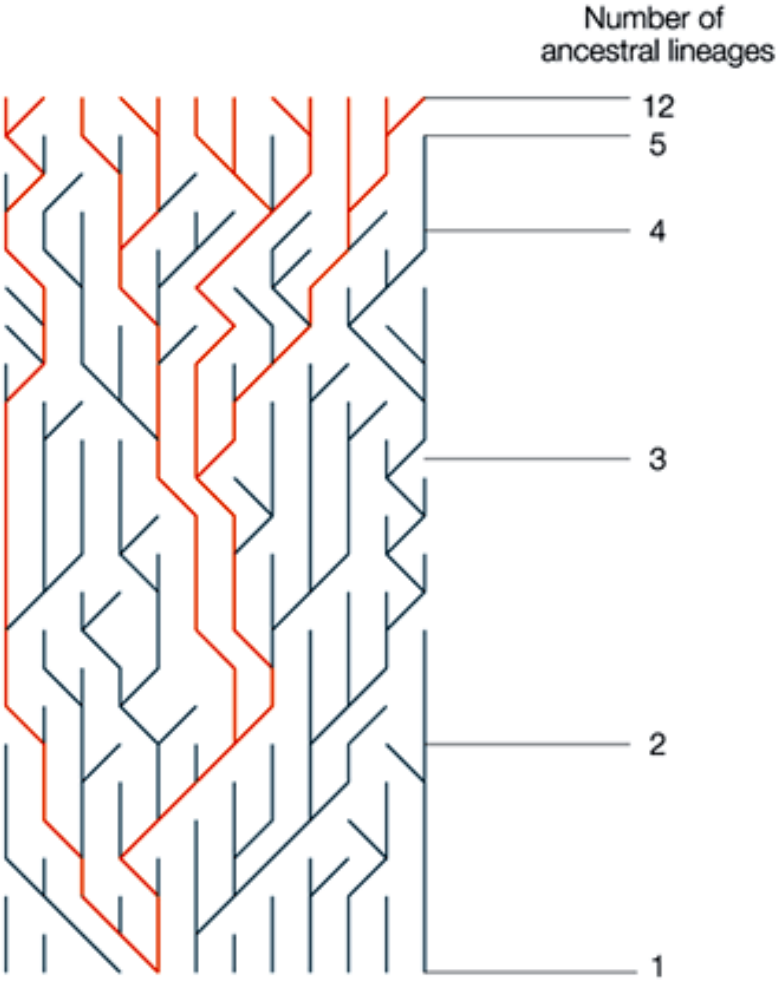
.
History of a population determines the range of this correlation

After few thousand generations, “block” structure remains



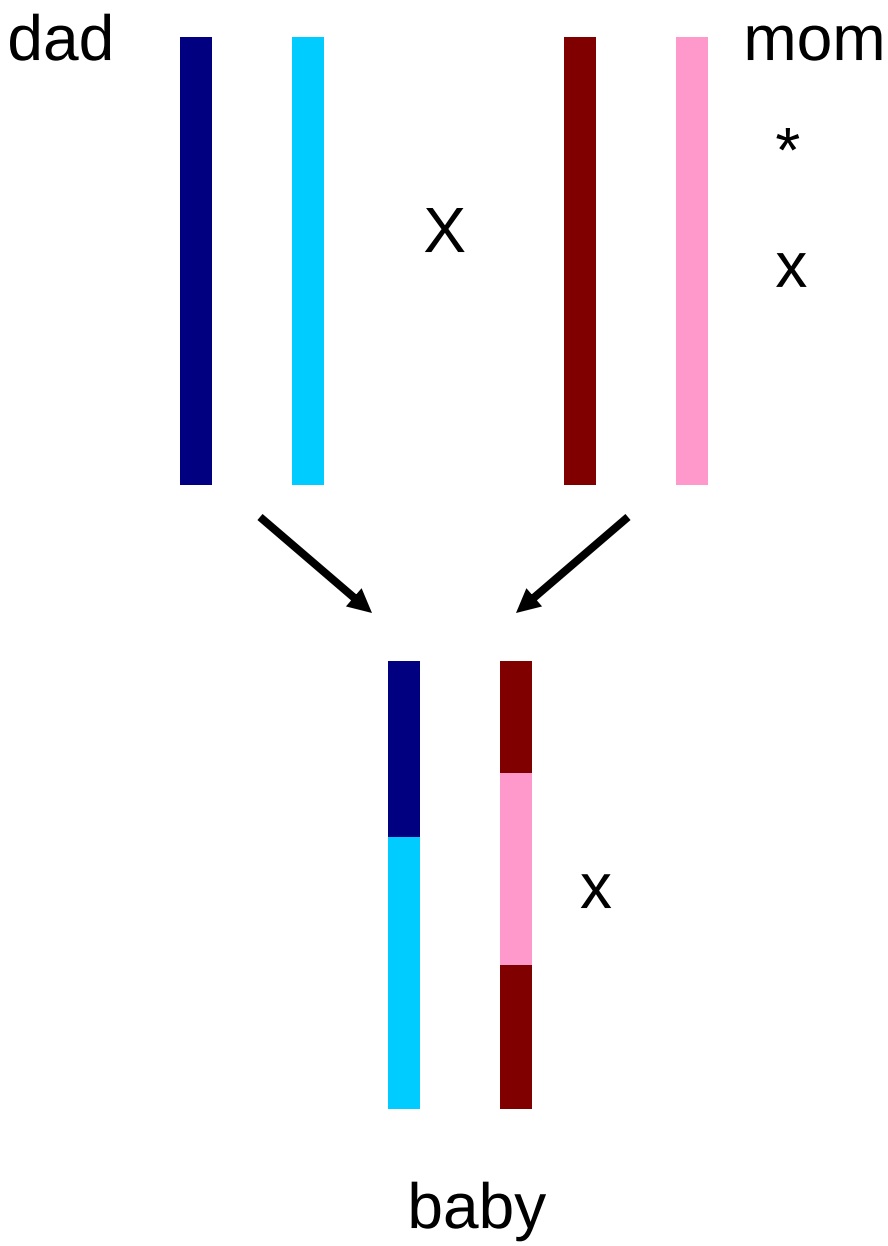
from Pääbo Nature 421, 409-412(23 January 2003)

Common ancestry



Nature Reviews | Genetics

- M generations ago, 2^M (great)$^{M-2}$grandparents contributed to your genome
- But the human population gets smaller as we go back in time
- So there must be common ancestry
- Analytical calculations for simple cases; simulations required for more complex modeling

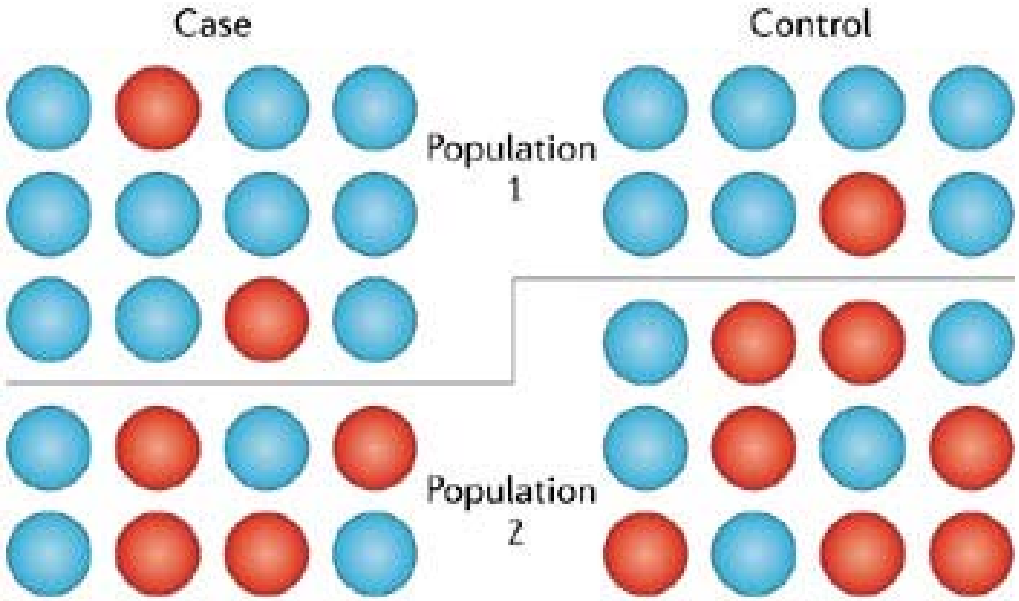


Variations (“markers”) that are farther apart will not be co-inherited

Given a set of known variants across the genome, we can ask whether or not each is co-inherited (“associated”) with a trait of interest.

Can have a severe “multiple testing” problem in assessing significance of a correlation.

Population structure can produce spurious associations

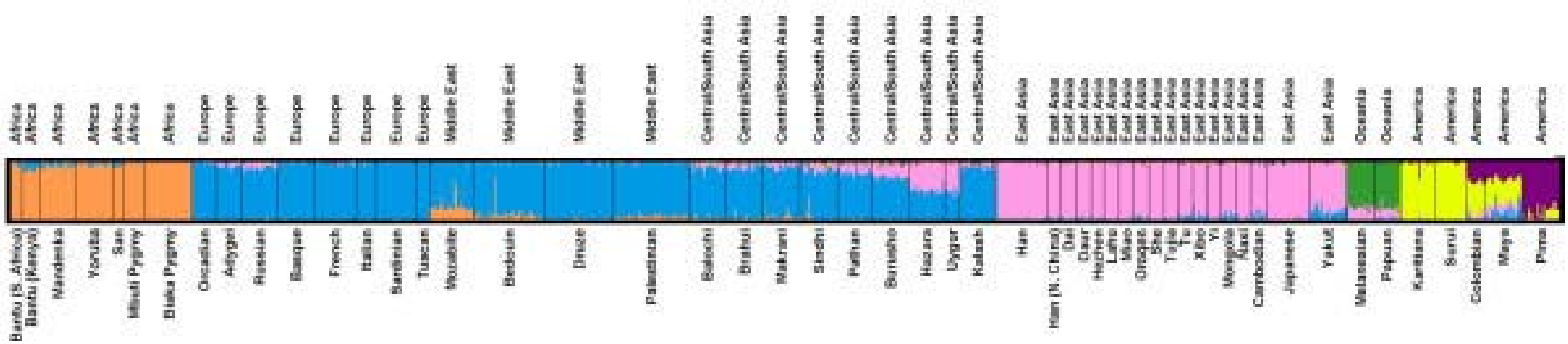


Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

Caution: the blue allele is overrepresented among cases -- but only because it is more frequent in population 1

Fit population structure to mixtures of “k” different subpopulations

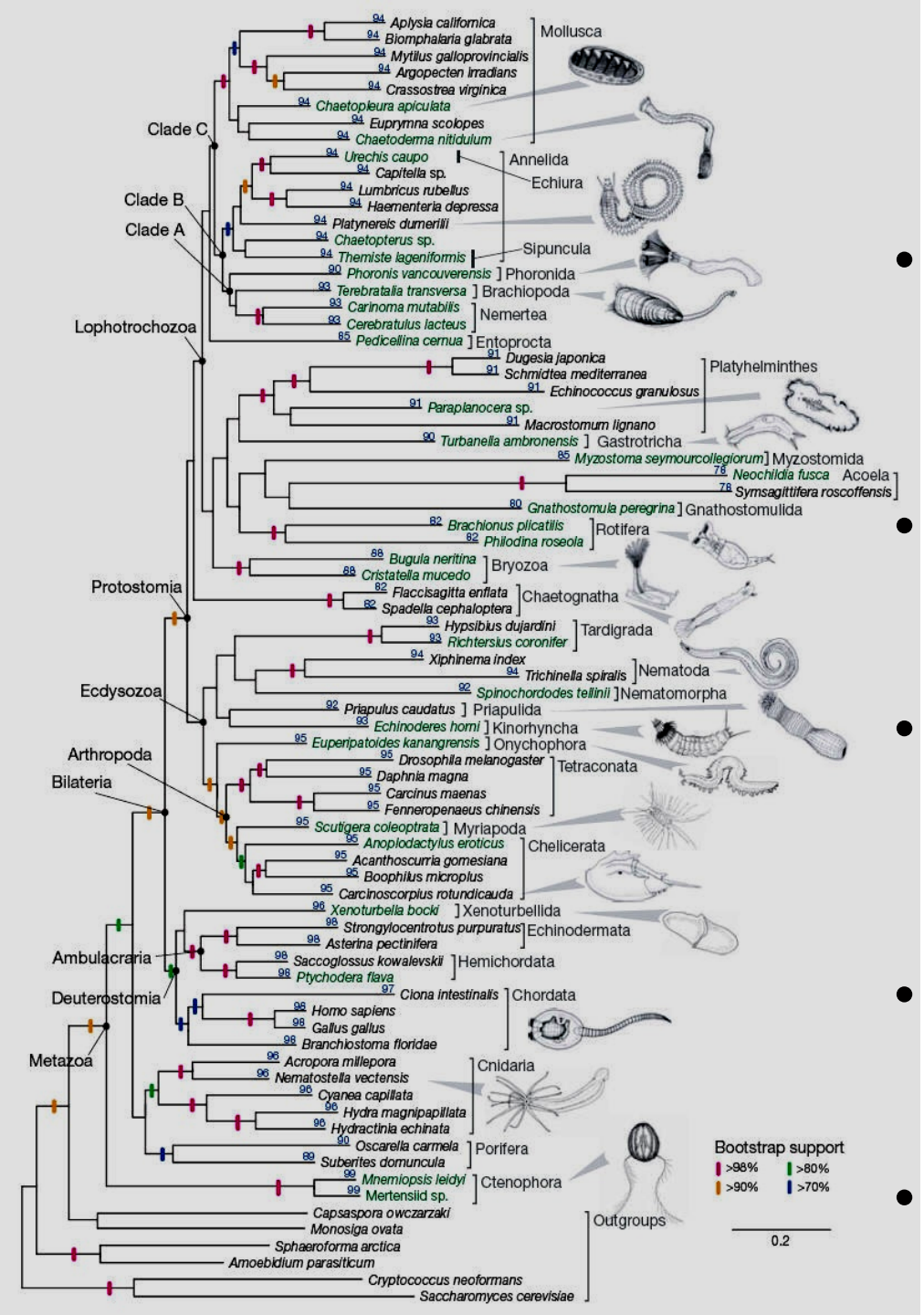
- Efficient algorithms exist for inferring population substructure (e.g., similarities within geographic subgroups) for dozens-hundreds of genomes. But 1000 Genomes Project underway, and hundreds of thousands will be genotyped.



Rosenberg et al. Plos Genetics 2005

Variation across species

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

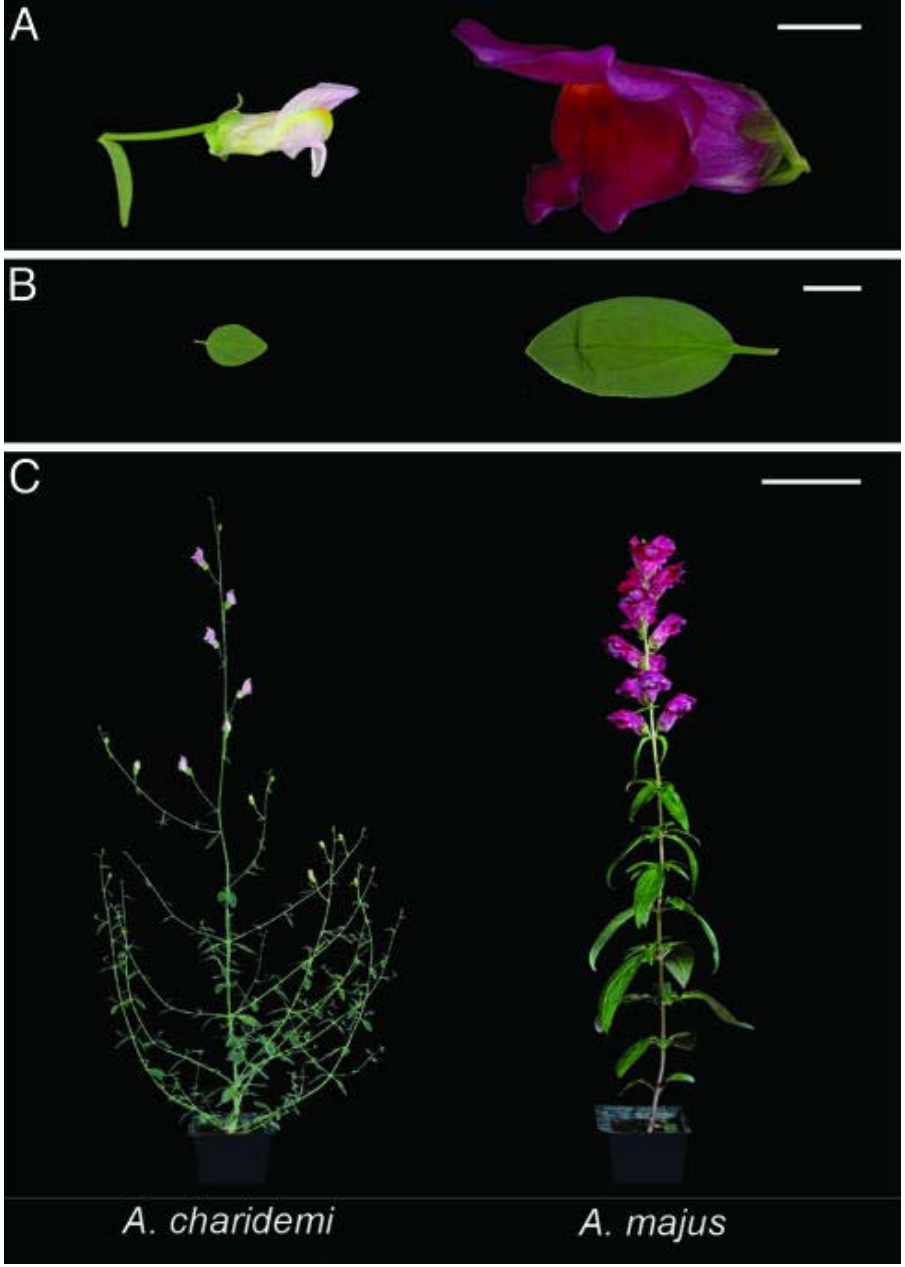


Determining relationships

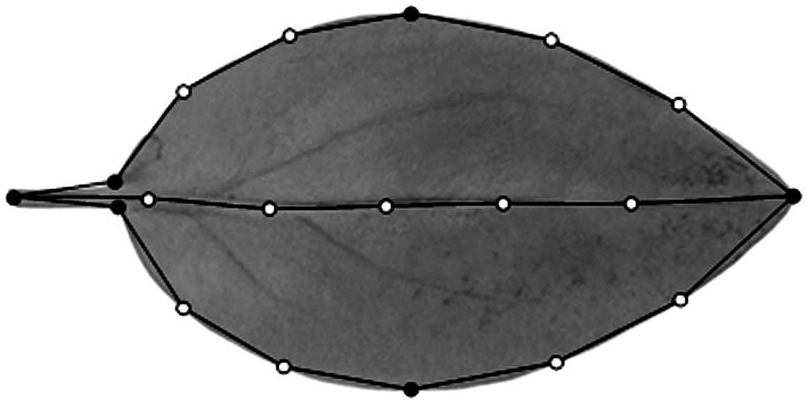
- Identify shared/different characteristics across species (hair/no hair; A/T/C/G, etc.)
- Evaluate likelihood of a particular topology and evolutionary model
- Find tree or trees with maximum likelihood (or evaluate probabilities in ensemble of trees)
- This is do-able with dozens of species, problematic with hundreds/thousands
- Not guaranteed to converge with more data

Variation

- Typical plant genome has 25-35,000 genes
- Related species have (mostly) “the same” genes
- Protein-coding and gene regulatory sequences differ at few % level between populations/related species

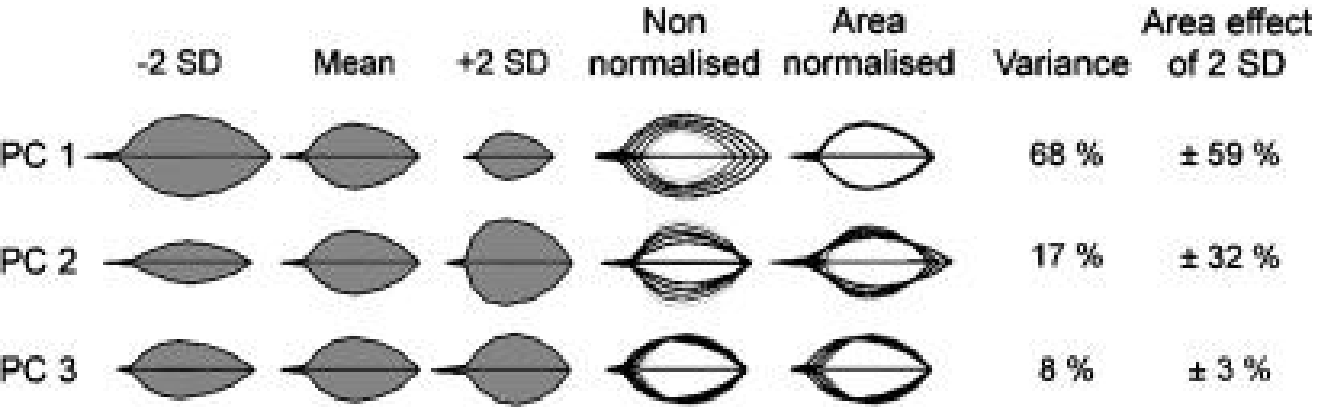


Langlade et al. PNAS. 2005 July 19; 102(29):10221–6.



Leaf shape variation

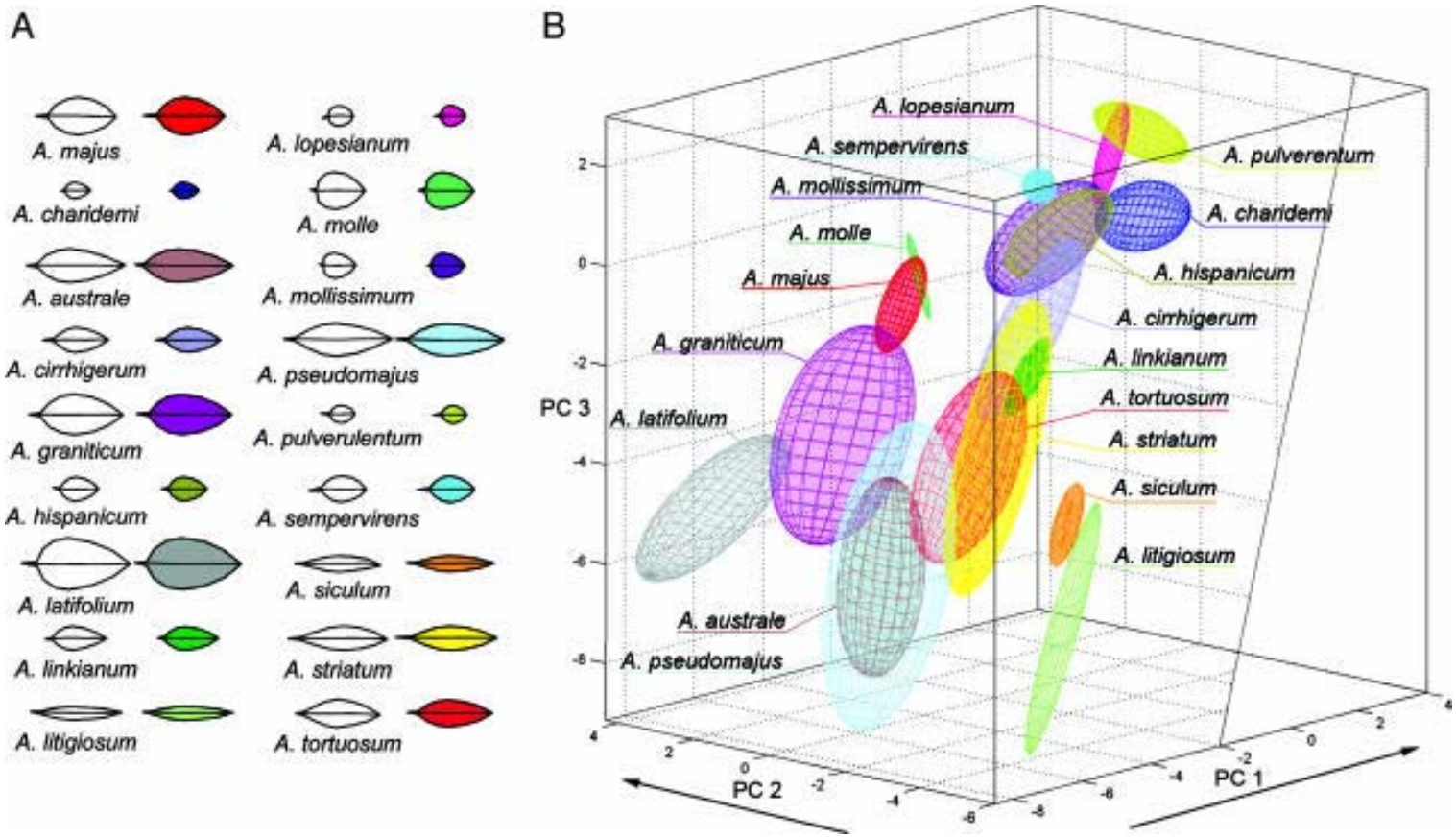
Points used to capture leaf shape.



- Area
- W/L ratio
- Axial asymmetry (dart/lance)

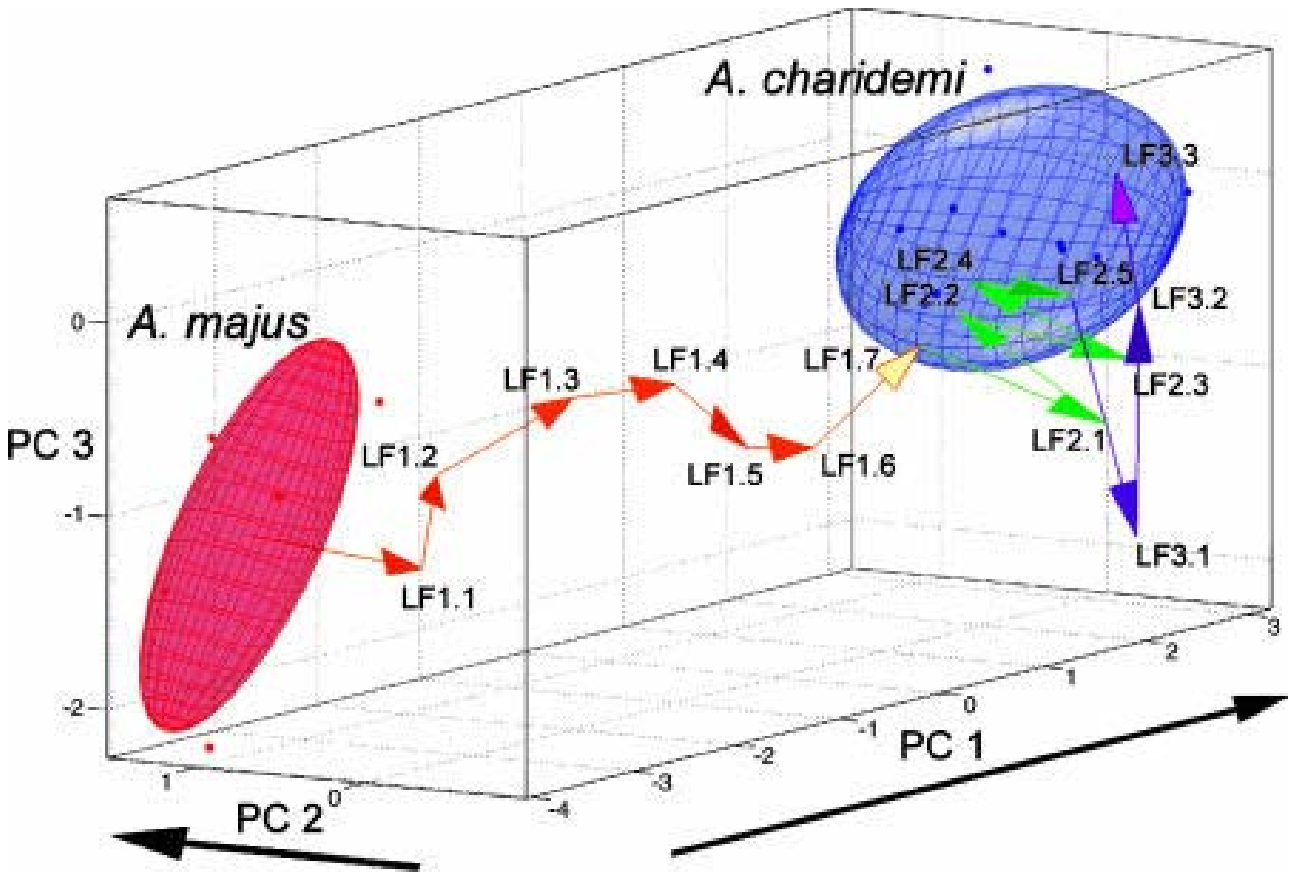
Leaf shapes and sizes described by variation along the first three Principal Components of the leaf allometry model

3 PCs describe range of related species of snapdragon



Size and shape of leaves from 18 snapdragon species captured by the allometric model.

Variation at ~15 genetic loci can describe "path" from one species to another



A. majus and *A. charidemi* species clouds displayed in allometric space with all QTL vectors laid end to end, starting from the center of *A. majus*.

Future genomic data analysis

- Genotyping and availability of genome sequences across populations will be commonplace. $1000 genome for humans is only a few years away, will also apply to tumors; agricultural species; microbial communities.
- Rich phenotypic information in space and time; gene expression, ‘omics will be associated with genotypes.
- What are the significant genotype-phenotype associations? What are artifacts of population structure, statistical noise? Need predictive power for diagnostics, personalized treatment, targeted improvement of crops, conservation biology.

- What is the high level science motivation for your work?
- What are some of the different science questions addressed in your field?
- What specific science questions do *you* expect to address with your work?
- What are the specific challenges you see in the analysis of your data?

- Does your work involve simulations? Experiments? Both? If both, how do simulation and experiment influence each other?
- Can you speak do where you expect to be, in terms of the satisfaction of your data analysis needs, in the next 5, 10, 15 years?
- Are there any real-time requirements for your analysis? Or time vs accuracy trade-offs?
- How much of the data currently being collected or generated is analyzed? How much of it needs to be?
- * How big is your data now? How big will it be in 5, 10 15 years?
- What are the kinds of analysis you would like to do, but cannot, due to lack of tools or techniques?
- Given what you foresee in 5, 10, 15 years, what analysis tools and techniques are working now but will soon breakdown?