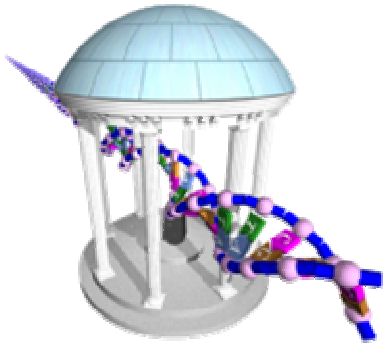
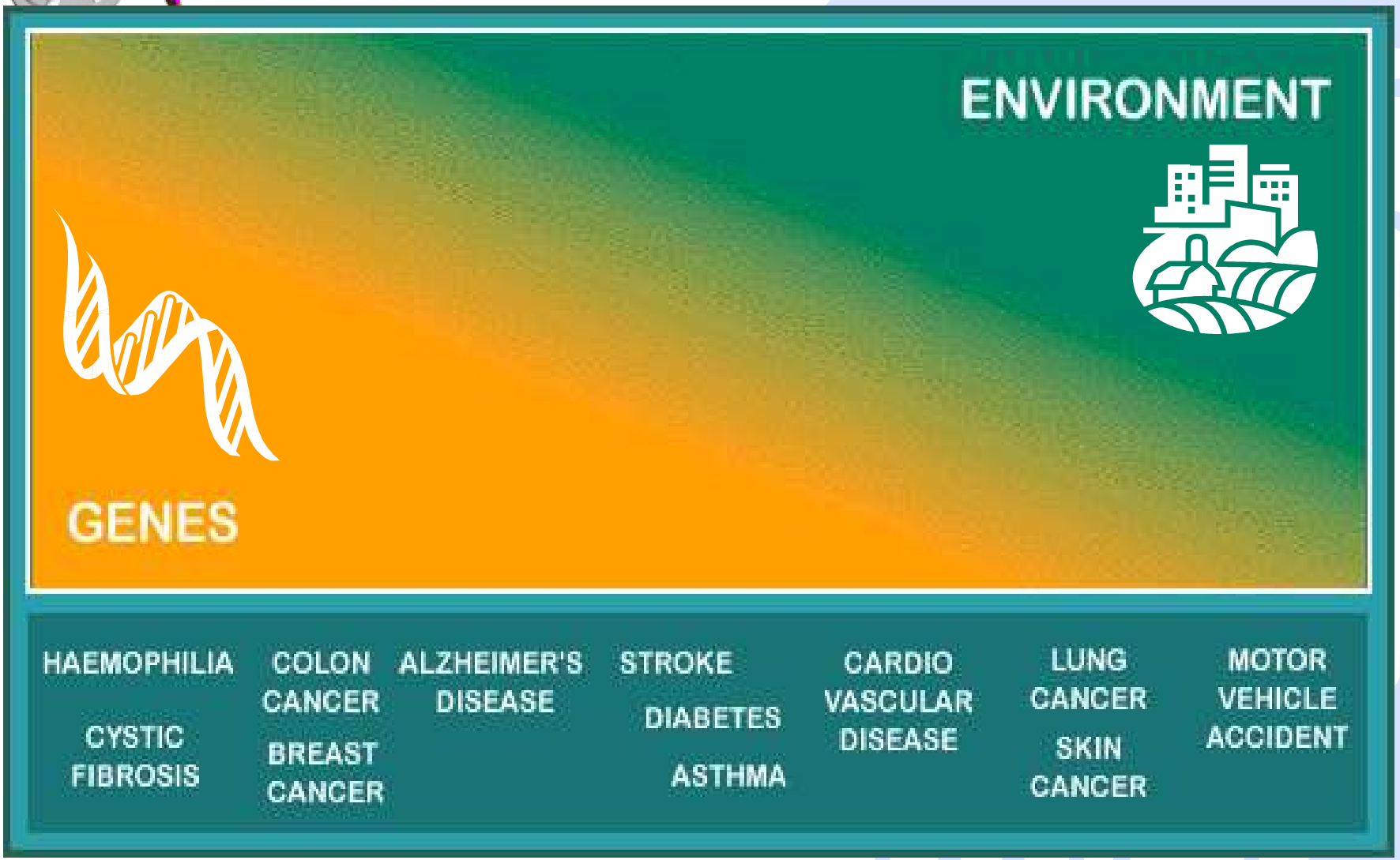


Efficient Data Mining Methods for Enabling Genome-wide Computing

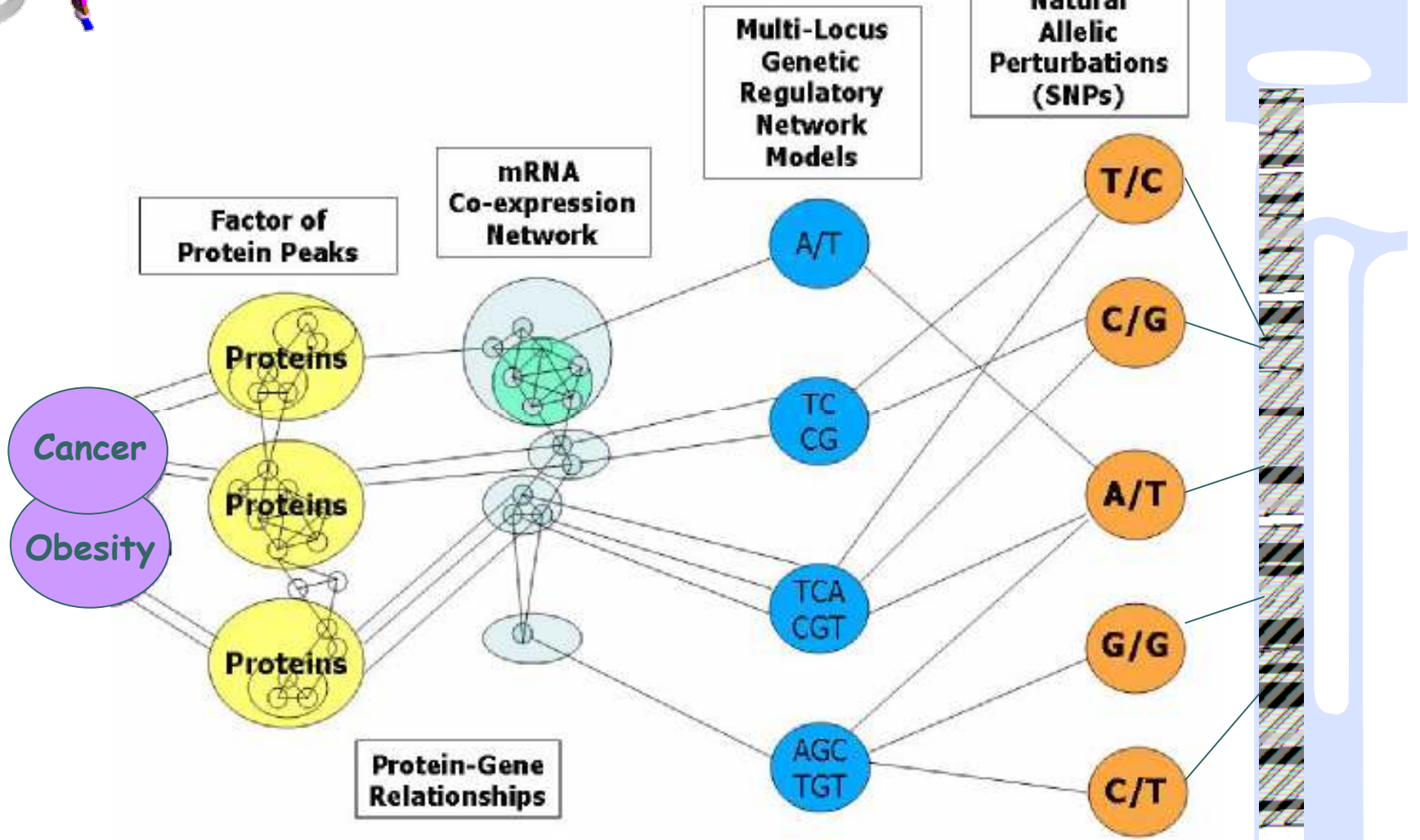
Wei Wang

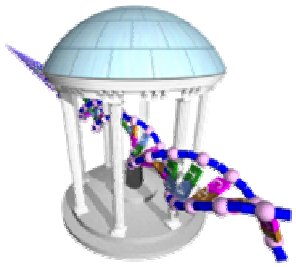
University of North Carolina at Chapel Hill

Genotype codes for phenotype

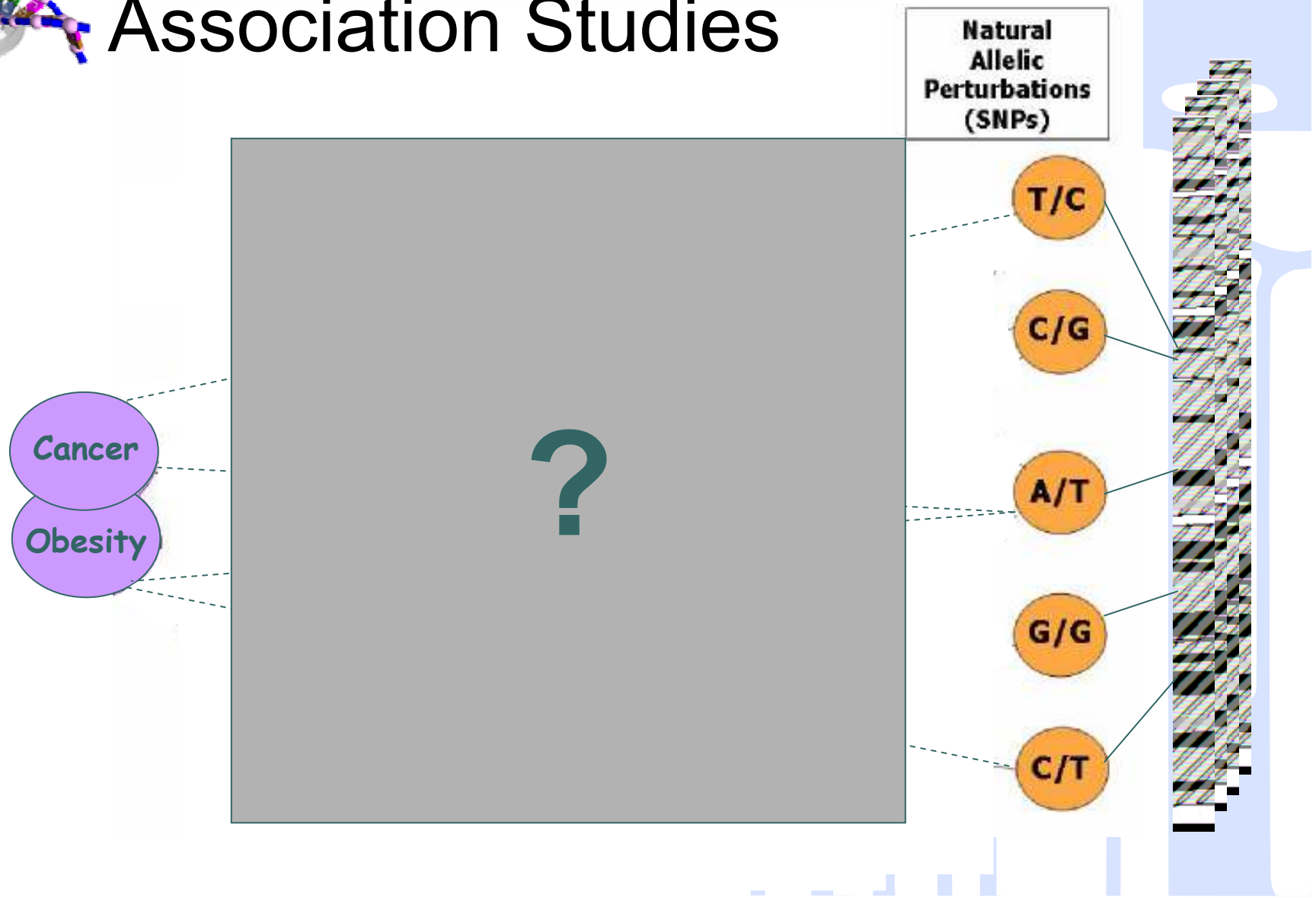


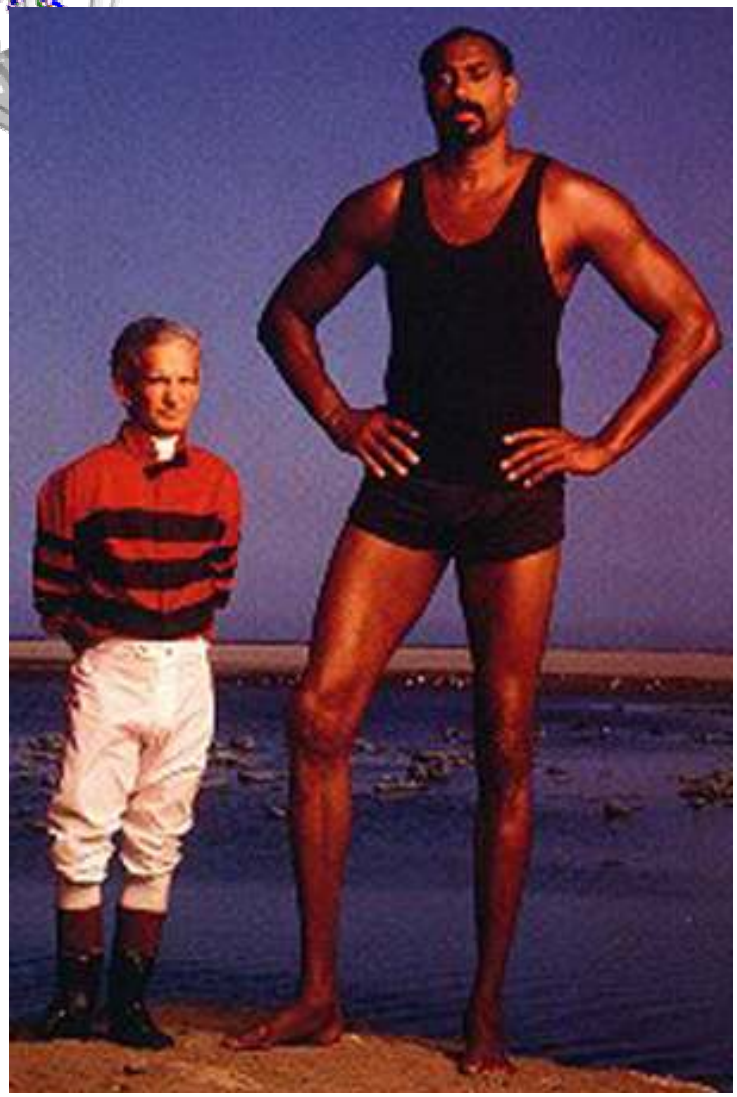
Systems Genetics View





Current View of Genome-wide Association Studies

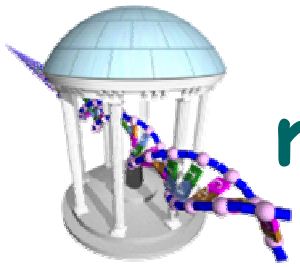




mouse populations ≅ human populations

Total mouse SNPs = ~40M
musculus, domesticus, castaneous

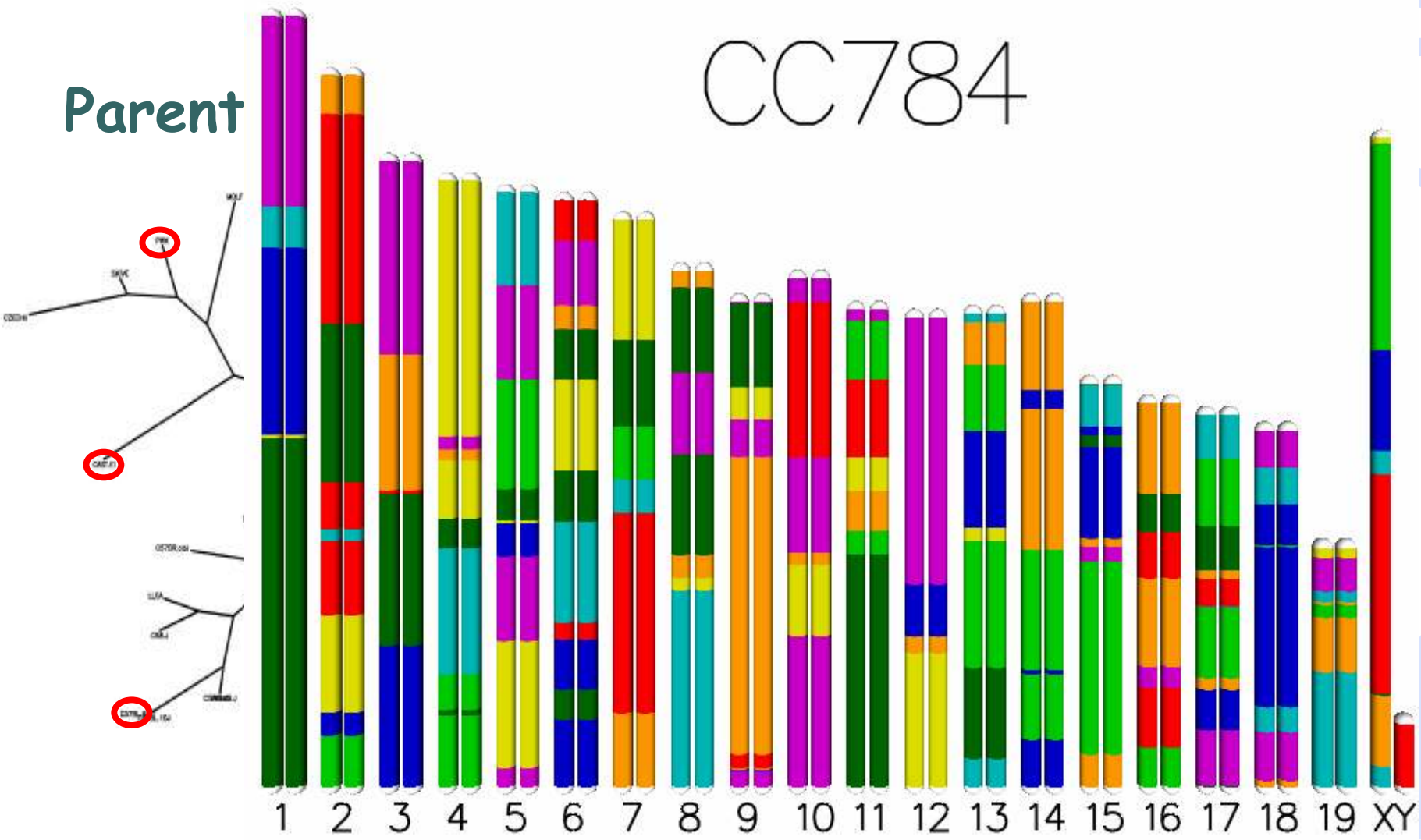
Total human SNPs = ~20M



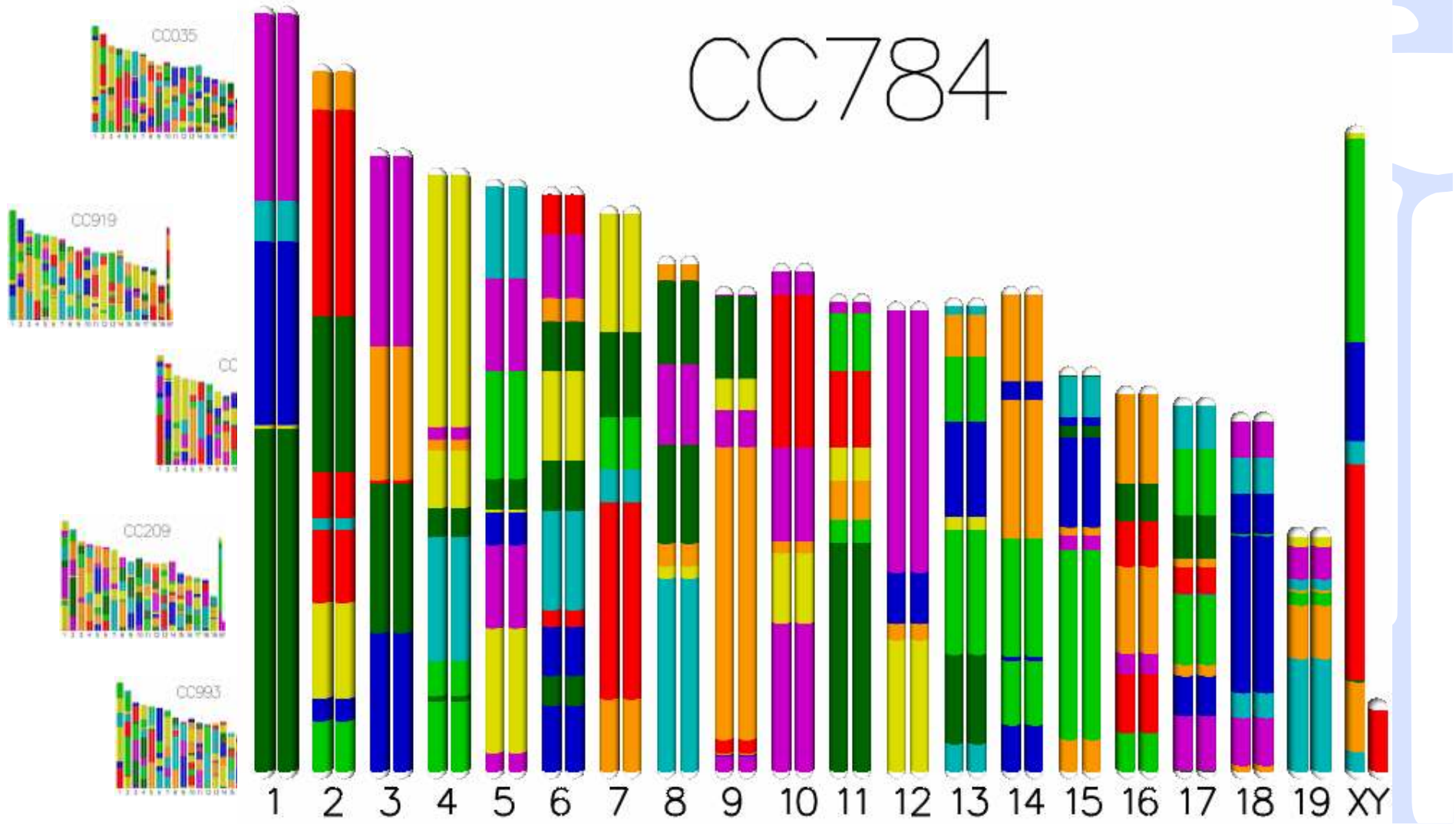
mouse populations ≅ human populations

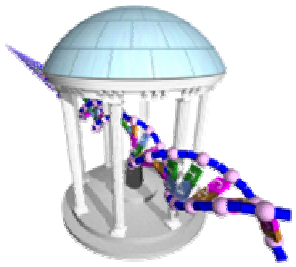
- fast generation time
- reproducibility
- gene modification

Collaborative Cross



1000 Independent Iterations

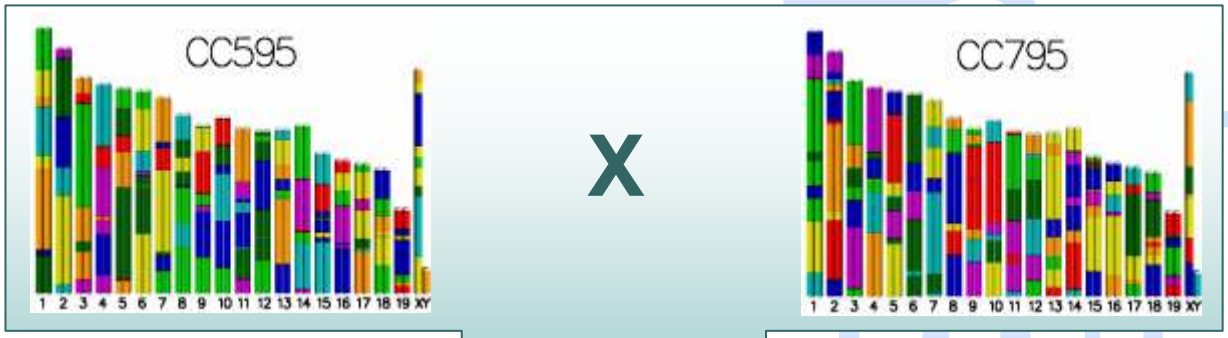


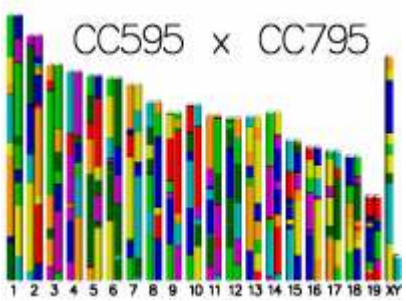


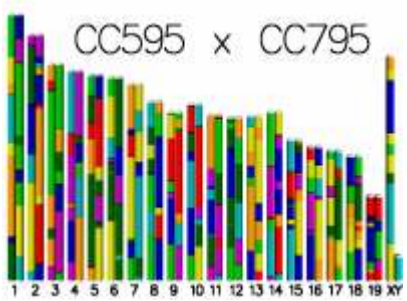
Recombinant Inbred Intercrosses (RIX)

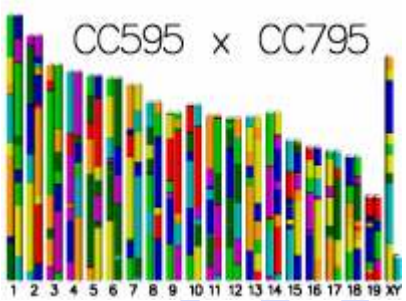
Reproducible Outbred Population

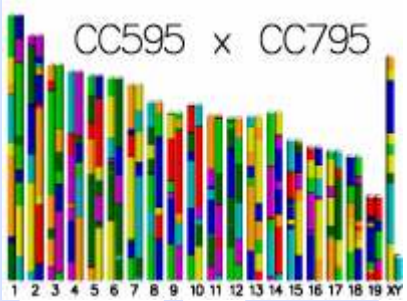
~1,000,000 possible genomes





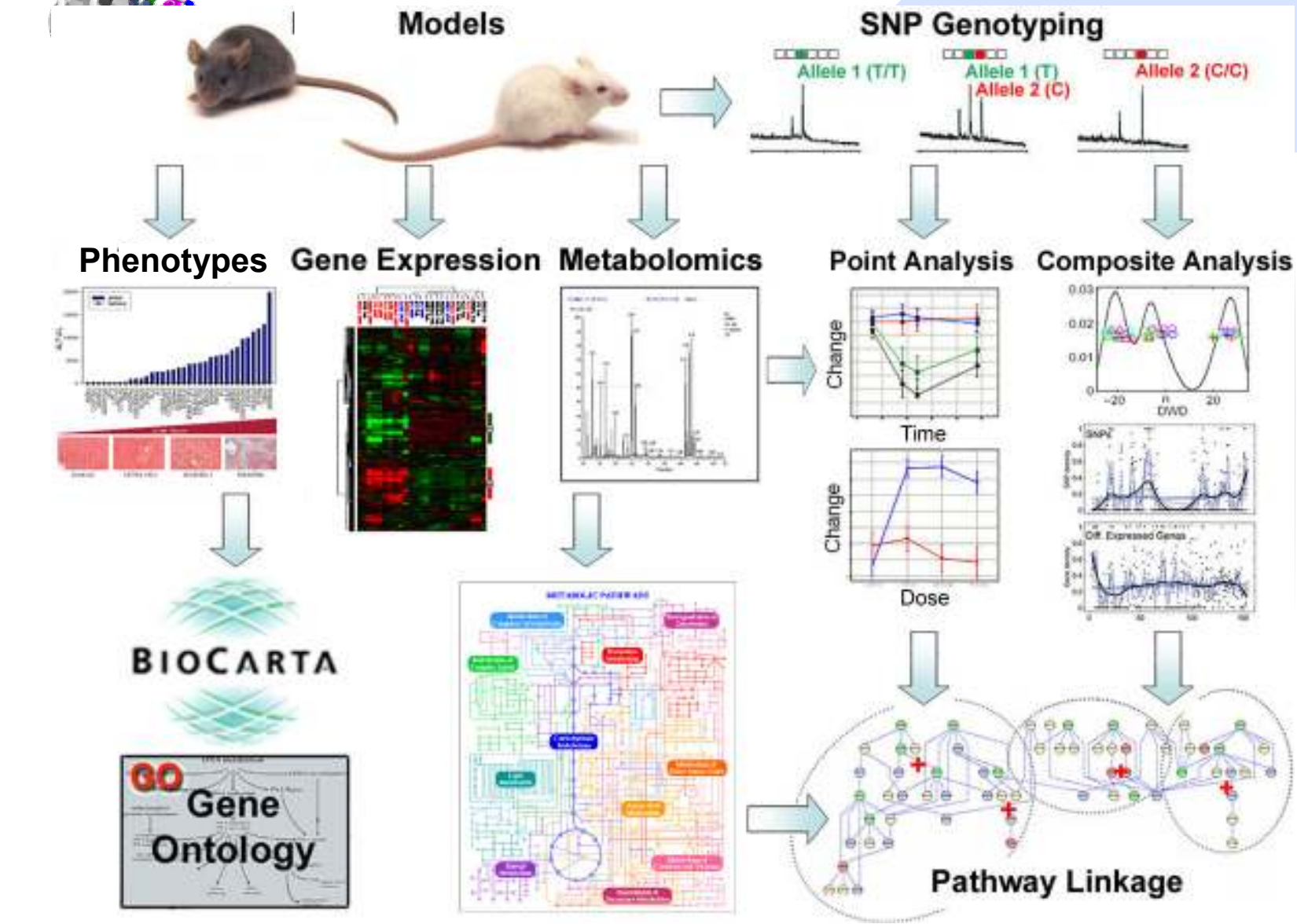


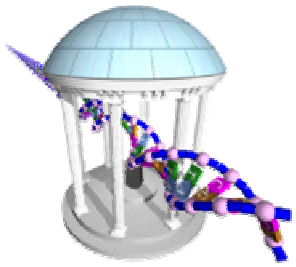




Genetics 170:1299, 2005

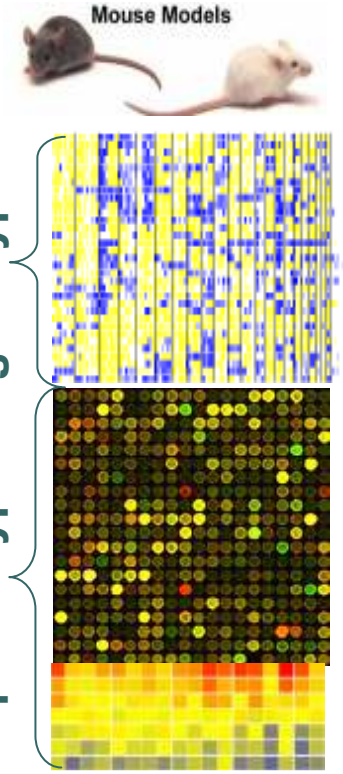
Data and Knowledge Integration over TIME and SPACE





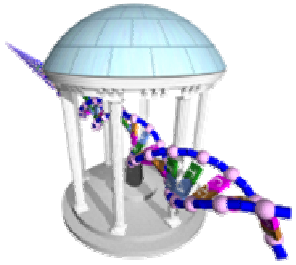
What we are facing …

DATA DATA …… and more **DATA**

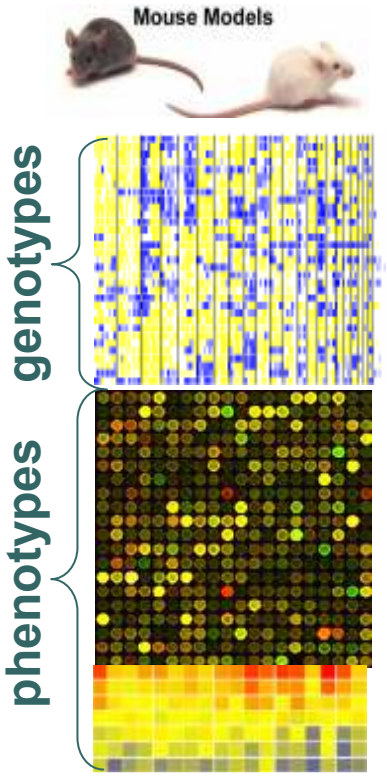


In the near future, we will have

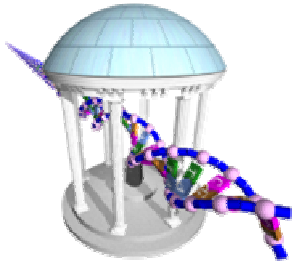
- a thousand RILs → a million RIXs
  - How to select lines to design crosses having desired features?
- tens of millions of SNPs
  - Can we infer phylogenetic structures?
  - Can we estimate historical recombination events?
- millions of phenotypic measurements (molecular and physiological) and other derived variables.
  - How to dissect complex correlations and causal relationships between variables?
  - How to efficiently assess the statistical significance of the results?



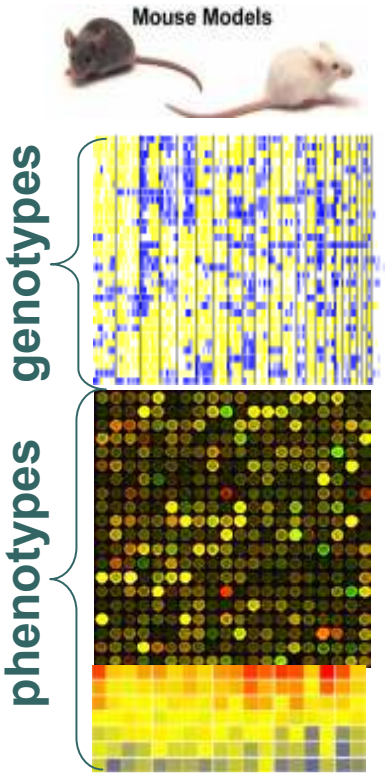
A Data Miner's View



- The dimensionality is extremely high
  - How do we cope with the curse of dimensionality?
  - Is it just a dimensionality reduction problem?
- The data matrix is comprised of disparate measurements including both continuous and discrete variables, which may not be directly comparable to each other.
  - How do we normalize data?
- The data matrix is not static, but growing both in terms of adding new samples and measurements.
  - How do we make the algorithms incremental and adaptive?

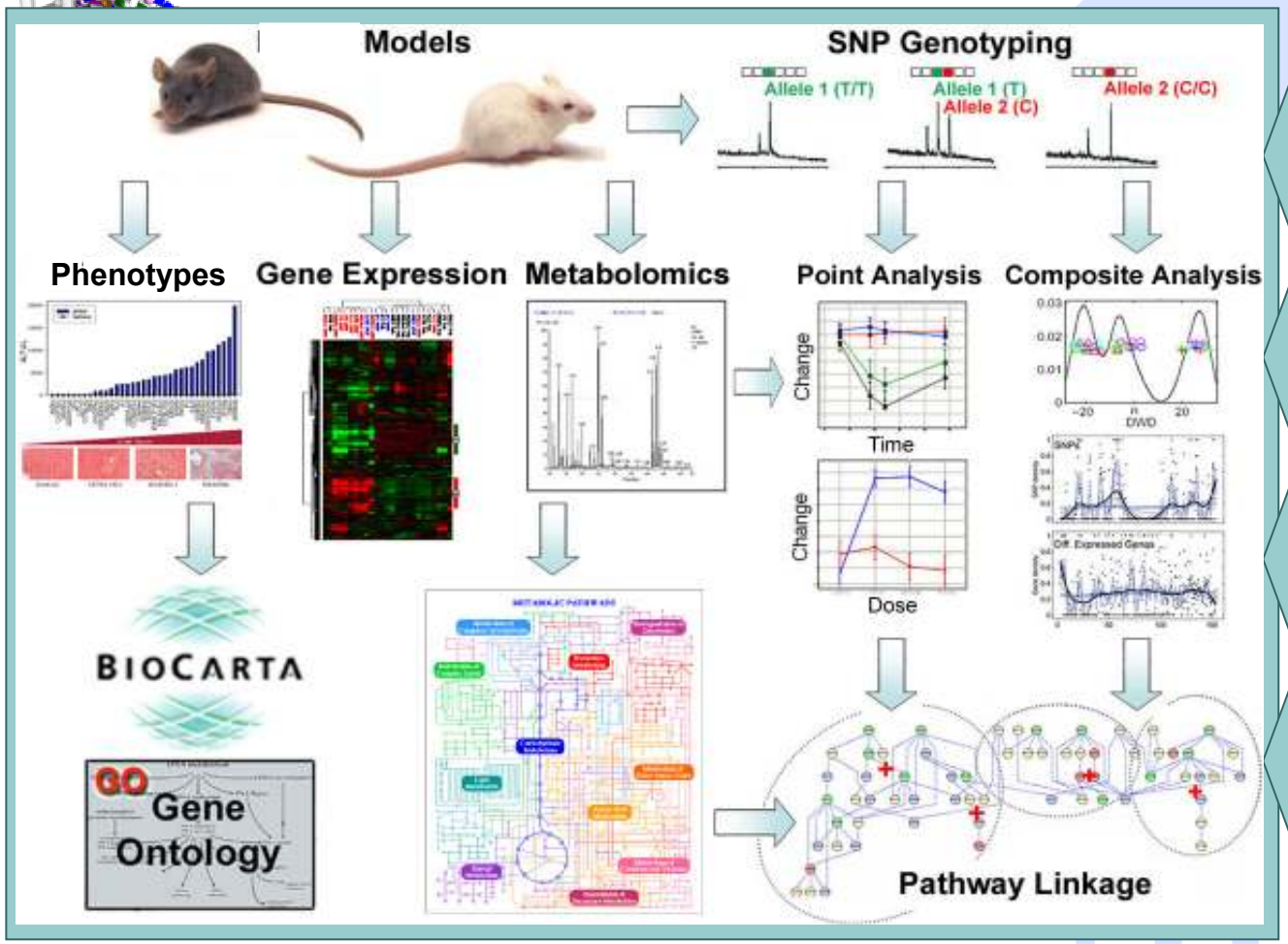


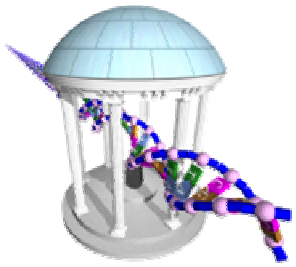
A Data Miner's View



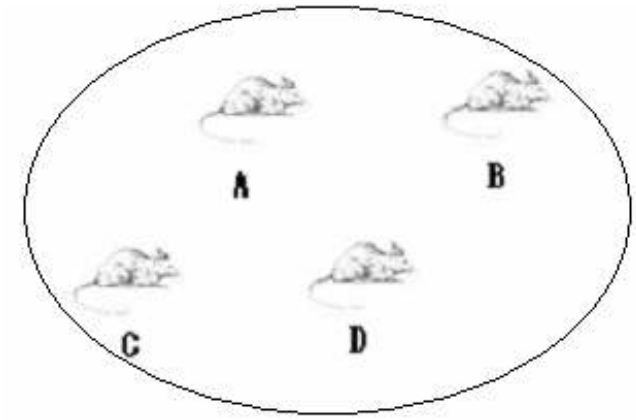
- Individual items may be contaminated, noisy or simply missing, which makes detectable relationships hard to "see", and thus hard to interpret.
  - How do we model noise?
  - How to make the algorithms robust to noise?
  - How to infer the missing value?
  - Can we formulate it as a classification or regression problem?
- The number of unknowns far exceeds the number of knowns
  - How to incorporate knowns in the methods?
- A large number of permutation tests are often needed to establish statistical significance
  - How to speed up this repeated (but necessary) computation?

Human Interaction

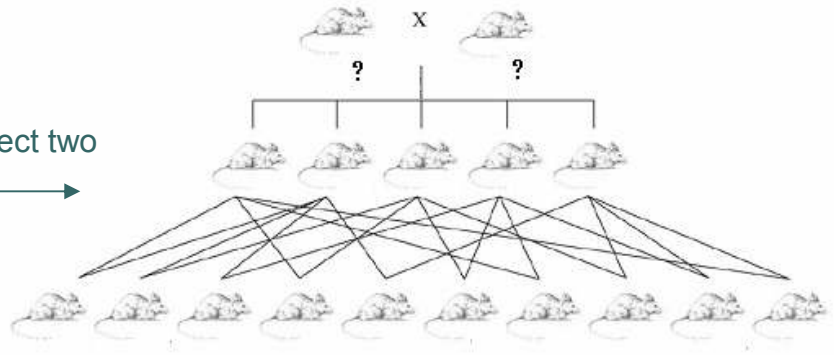




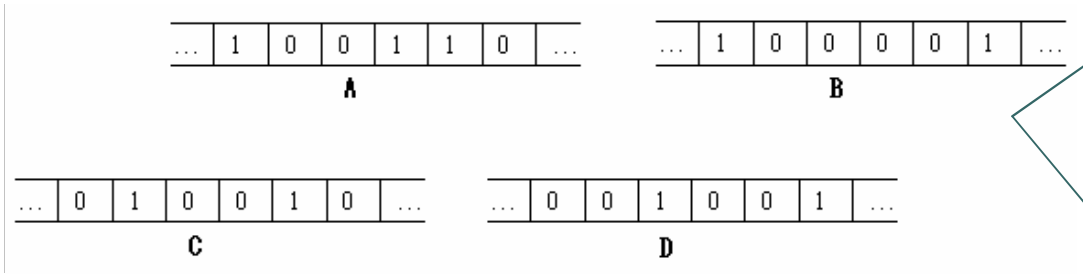
Sample Selection Maximizing Genetic Diversity

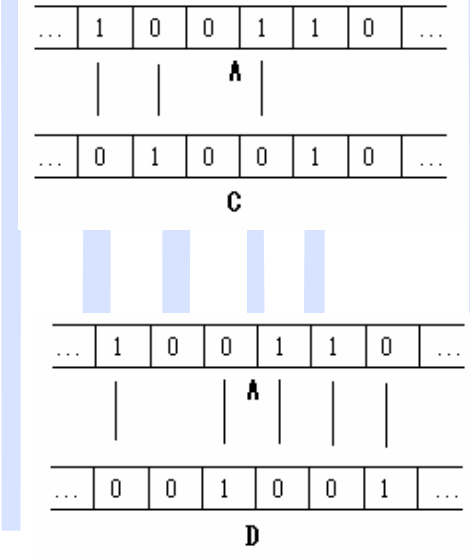


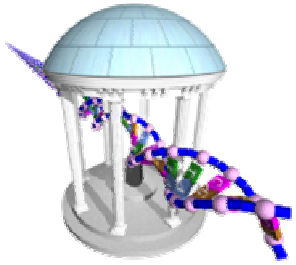
select two



genotypes, at biomolecular level,
Single Nucleotide Polymorphisms
(SNP)

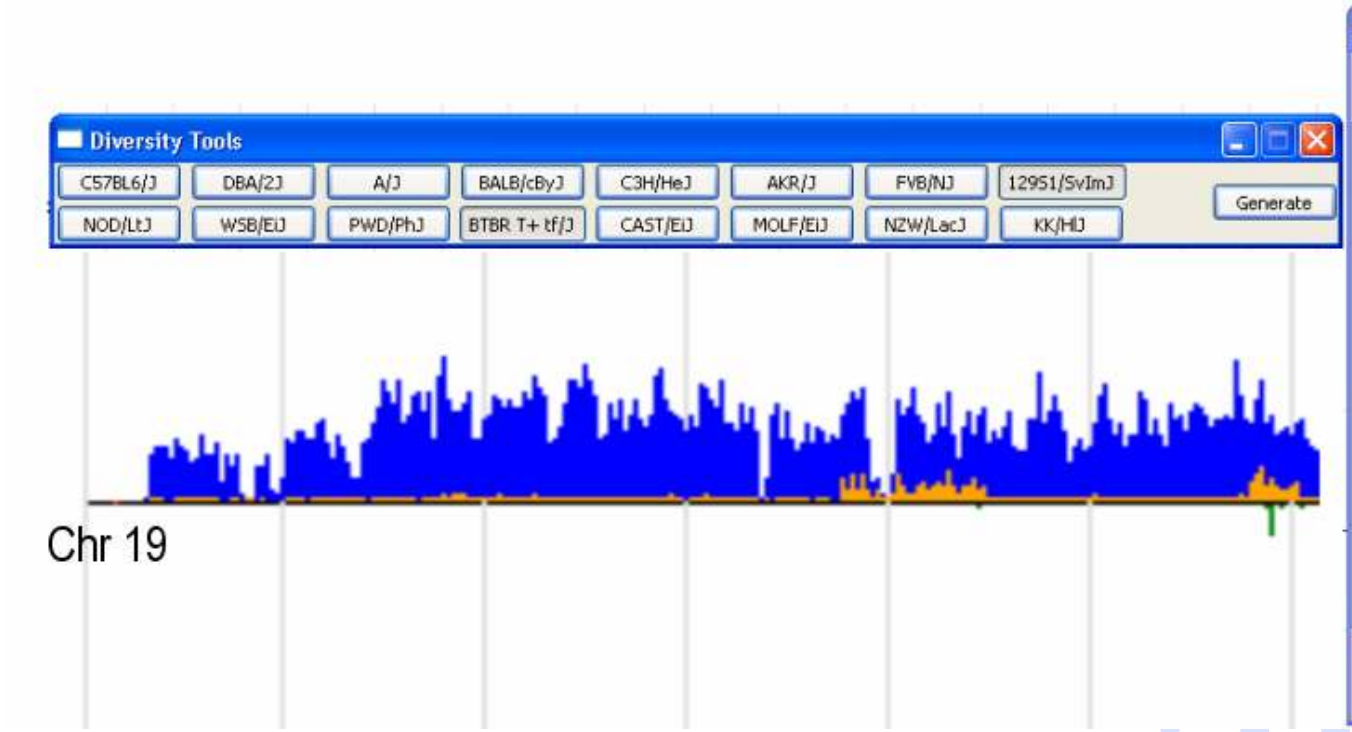


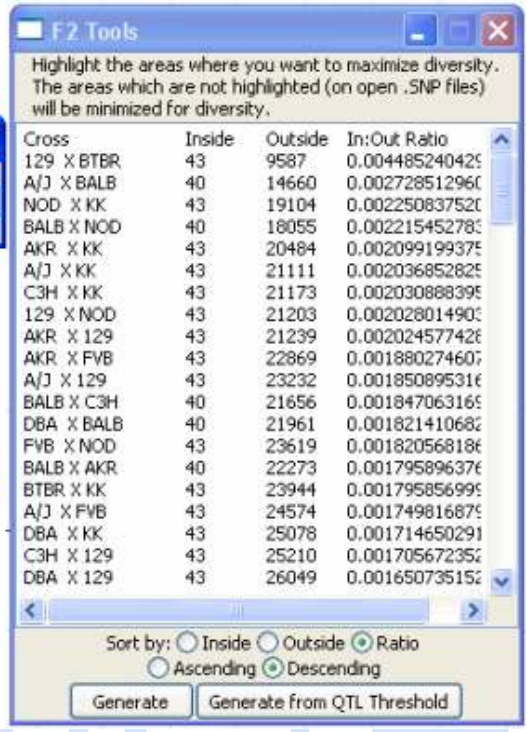


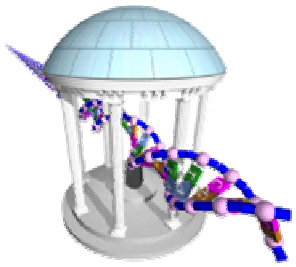


NP-complete

- maximizing the diversity within targeted regions
- minimizing the diversity outside the regions

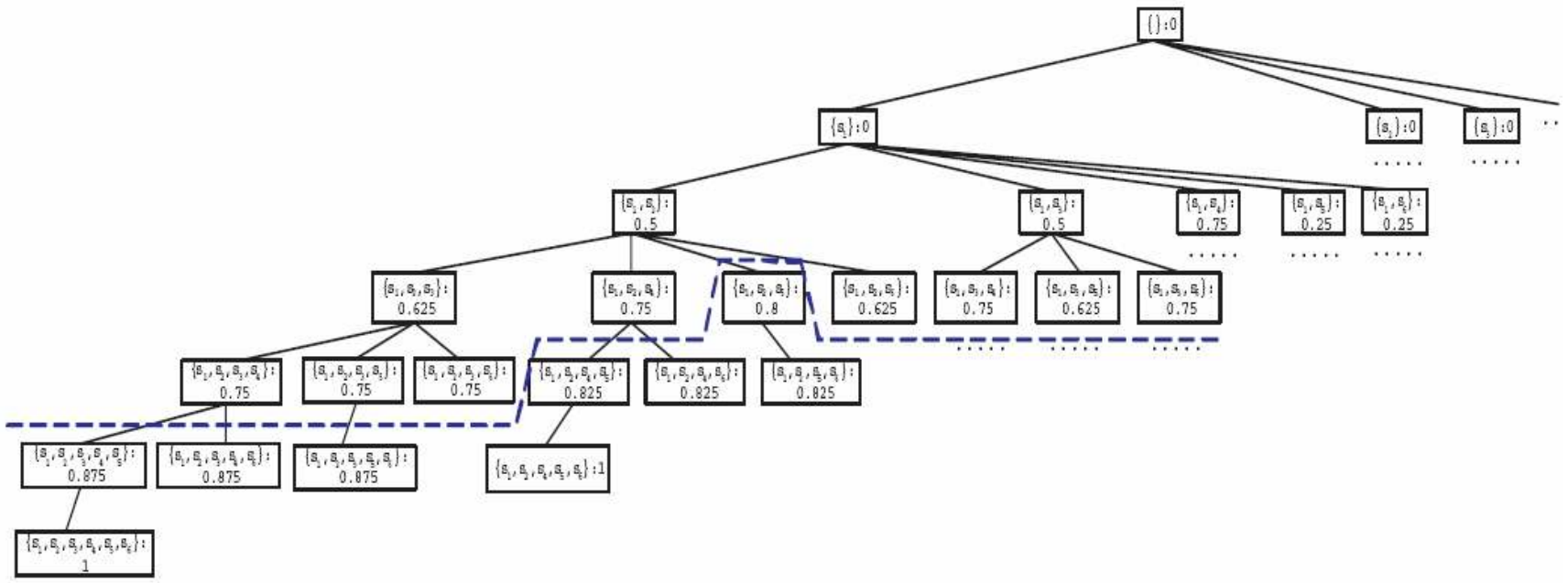


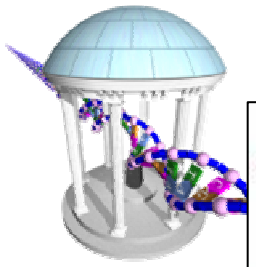


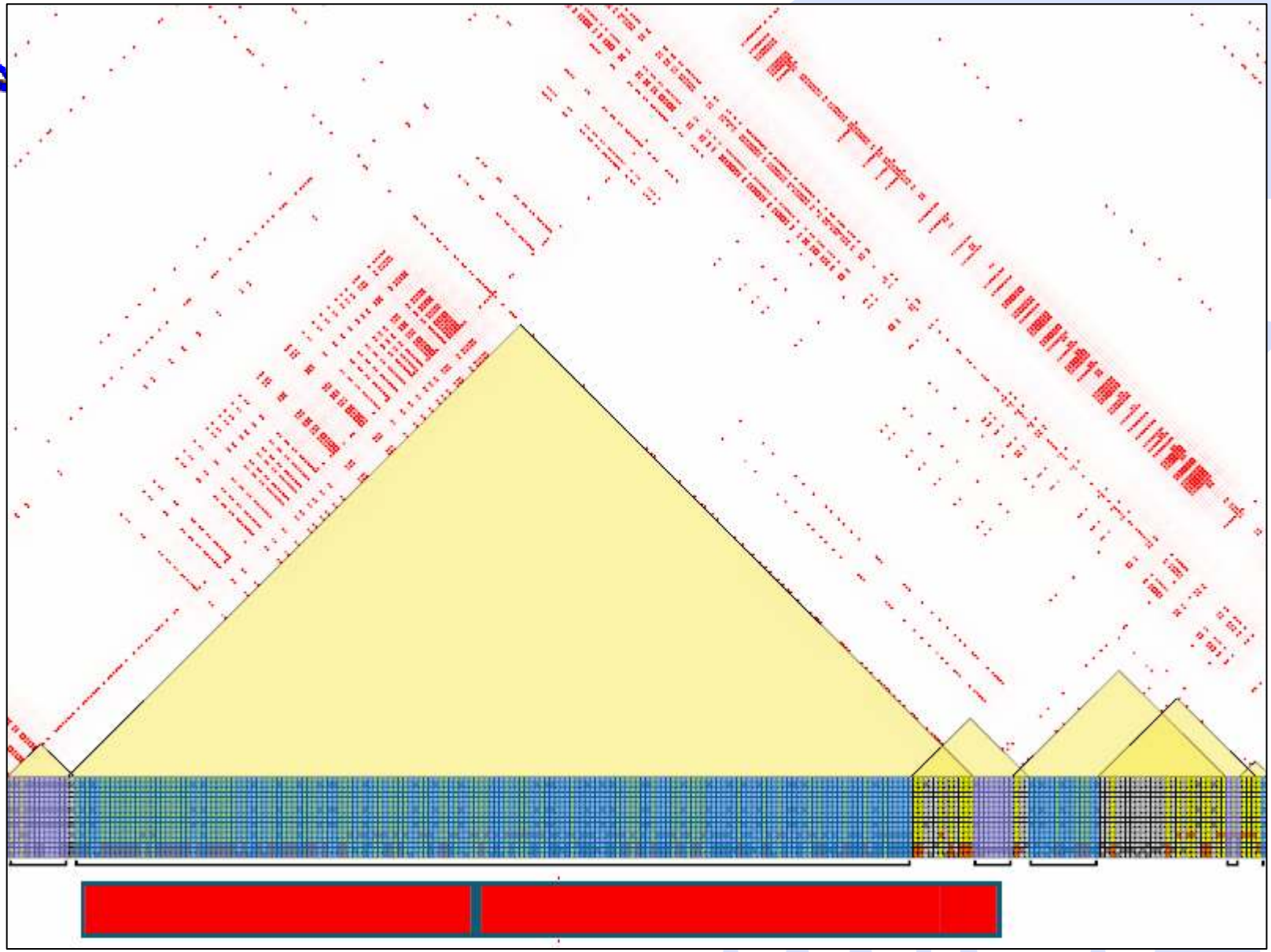


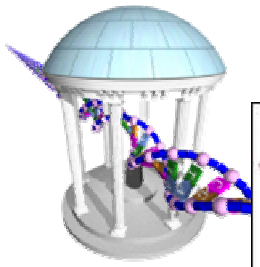
Searching Algorithms

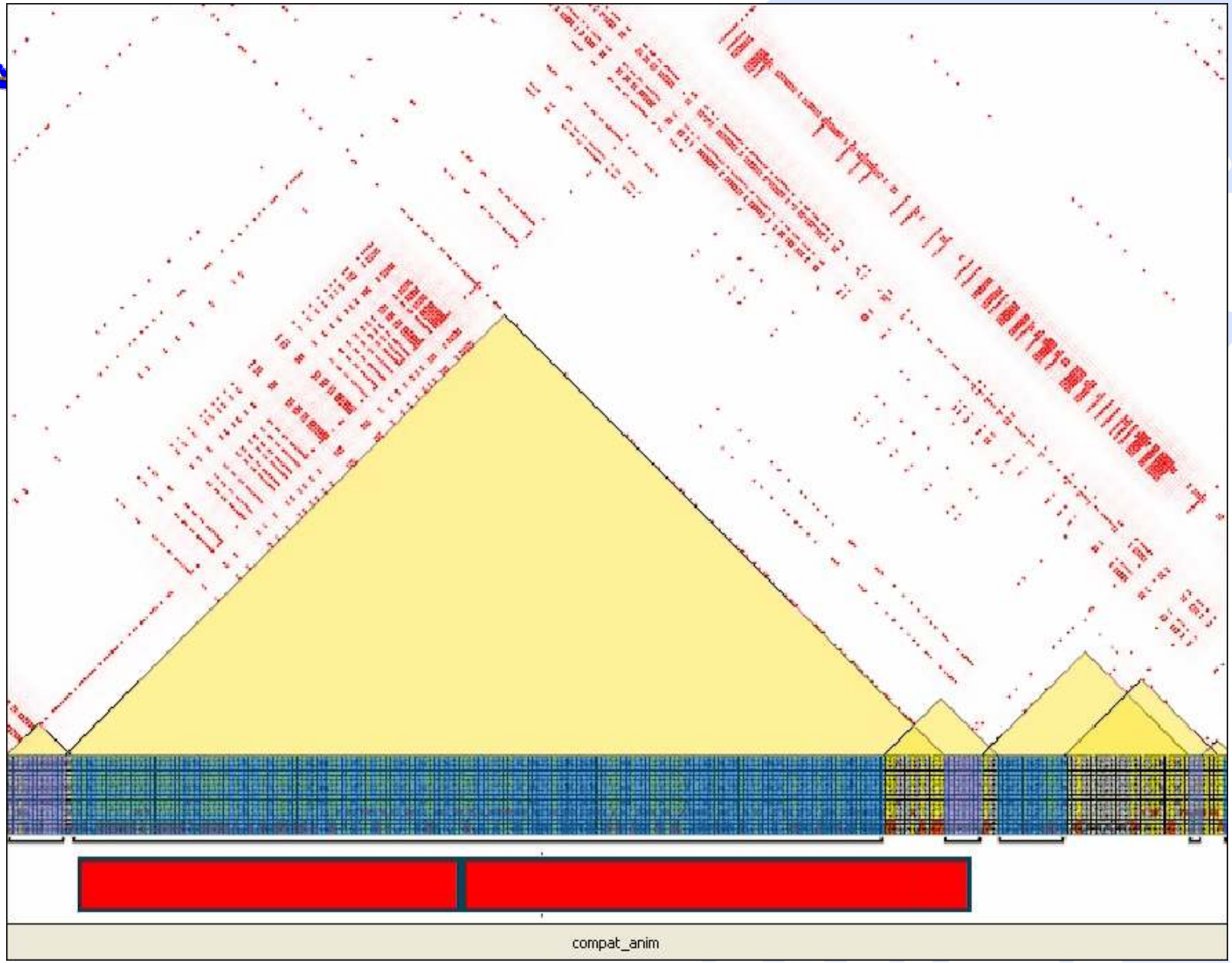
- Systematically enumerates all possible combinations of samples from smaller subsets to larger ones with effective pruning strategies
  - based on pair-wise diversity

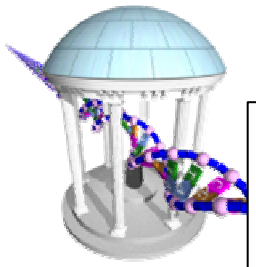


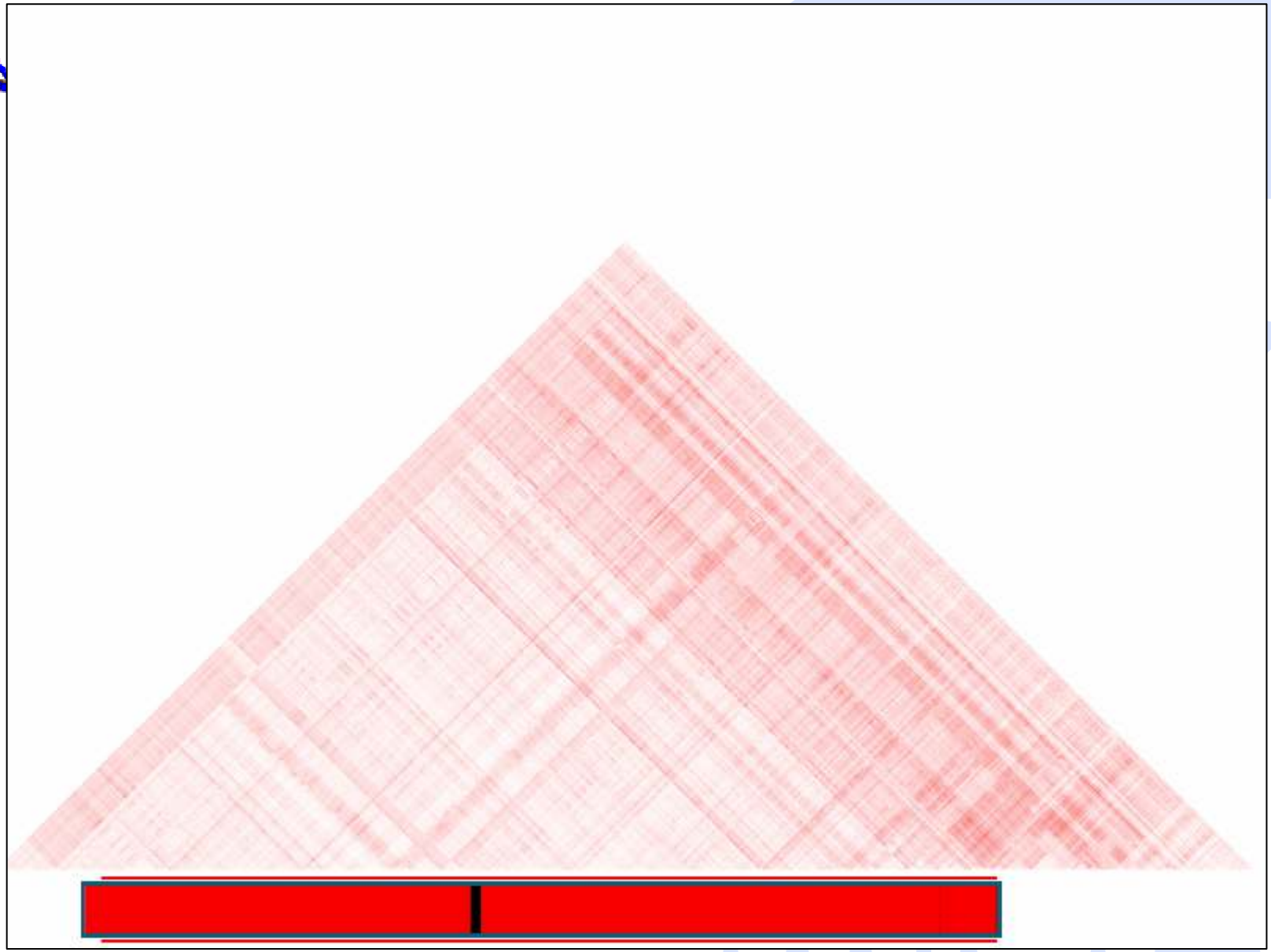




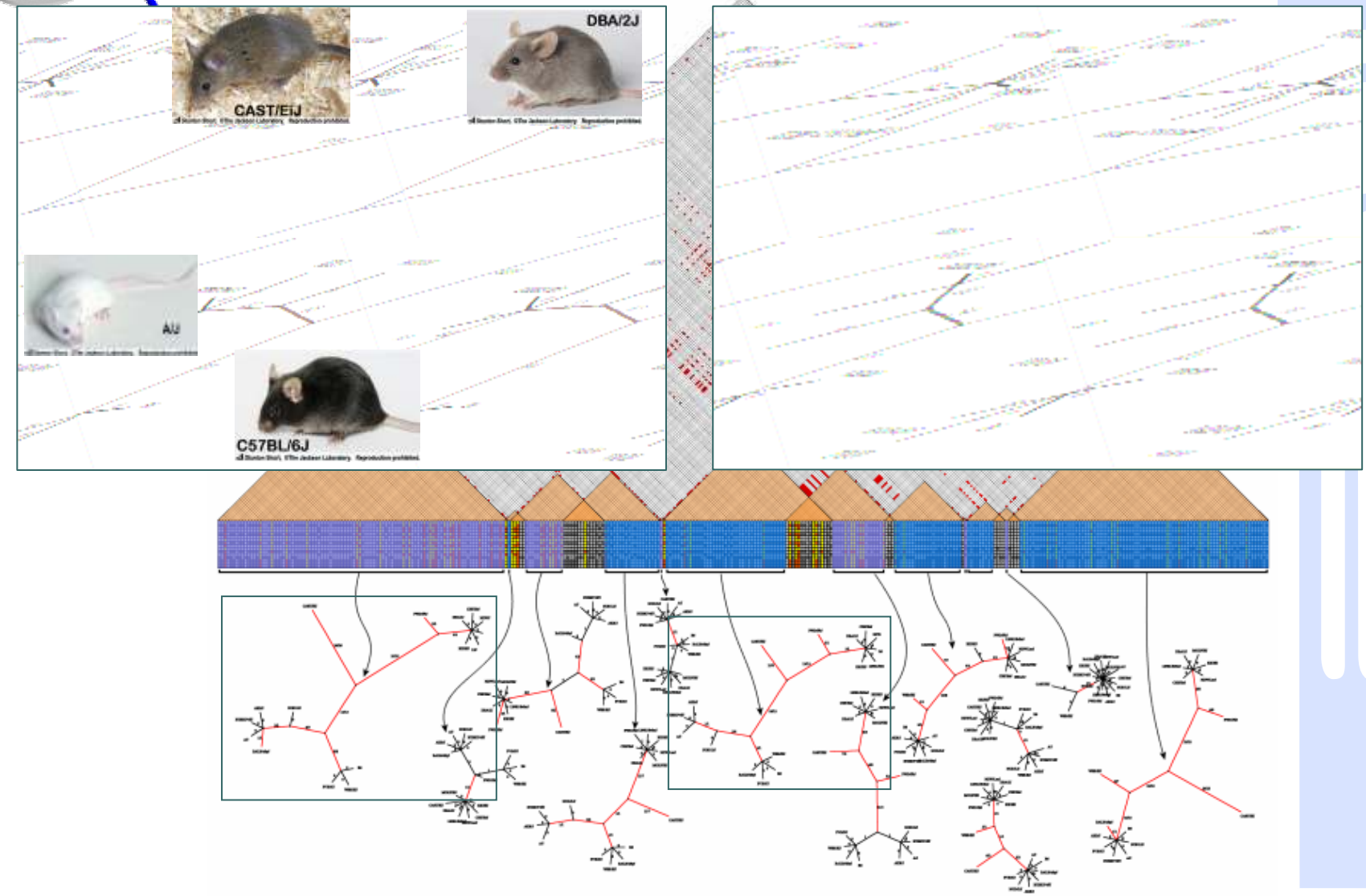


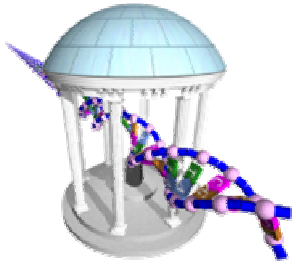

compat_anim





Local Perfect Phylogeny Trees

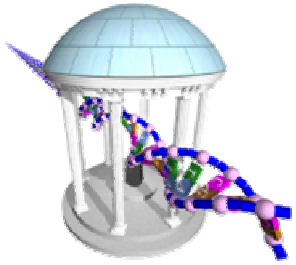




Local Perfect Phylogeny Trees

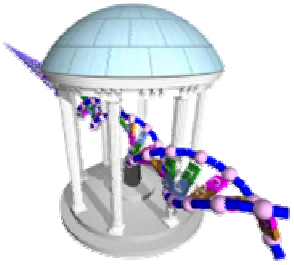
- When quadratic time/space is too much,
  - what is the minimal number of trees needed to describe an entire genome?
  - how to compute all local perfect phylogeny trees efficiently?
  - what are the common trees/subtrees?
  - how to perform phylogeny tree-based association studies efficiently?

Linear Complexity



Conclusion Remarks

- The ability to gather, organize, analyze, model, and visualize large, multi-scale, heterogeneous data sets rapidly is crucial.
- The massive scale and dynamic nature of data dictate that data mining technologies be fast, flexible, and capable of operating at multiple levels of abstraction.
- Novel data mining techniques are required to extract information, expose knowledge, and understand complex data.



Acknowledgements

- This is a joint project with






http://compgen.unc.edu/

- **NSF IIS 0534580:** “Visualizing and Exploring High-dimensional Data”
- **EPA STAR RD832720:** “Environmental Bioinformatics Research Center to Support Computational Toxicology Applications”
- **NSF IIS 0448392:** “CAREER: Mining Salient Localized Patterns in Complex Data”
- **NIH U01 CA105417:** “Integrative Genetics of Cancer Susceptibility”