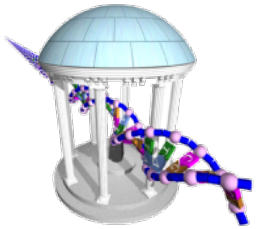
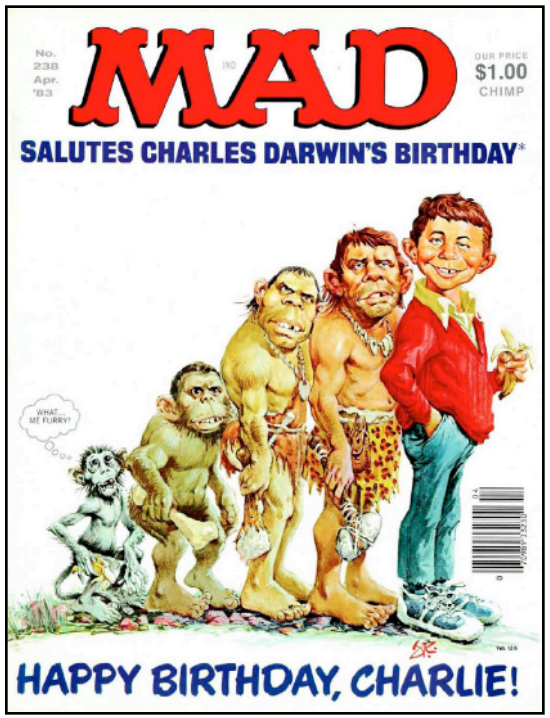
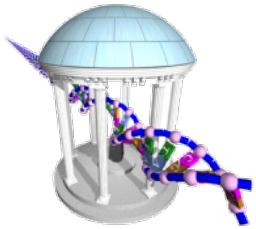


Comp 790-087

Computational Genetics

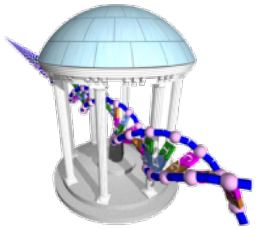
- Administrative Details
- Course Overview
- Simple Genetics





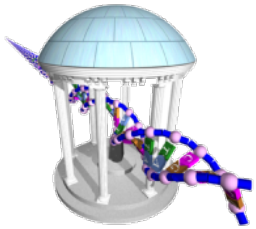
Course overview

- **Synopsis**
  - Graduate-level project course
  - Guided readings, discussions,
    in-class exercises, final projects and write-up
  - We will probably move to SN325 (will meet here next week)
- **Website**
  - To appear at:
    http://www.csbio.unc.edu/mcmillan/index.py?run=Courses.Comp790S14
- **Course Grading**
  - Class Participation 10%
  - Research Paper Presentation 20%
  - Project Proposal 20%
  - Final Project & Write-up 50%



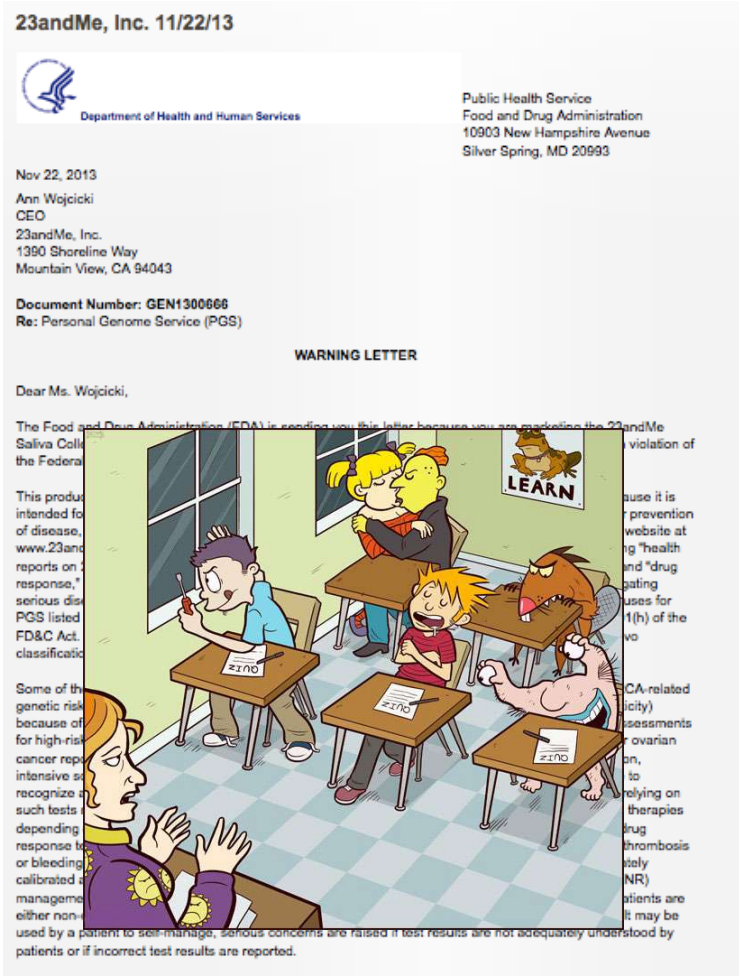
Syllabus

- Part 1. Background Topics (1/3) – each student will contribute to a shared s/w library (bring your laptops, preloaded with Python 2.7)
  - Genotyping and Sequencing
  - Recombination, phasing, genome mapping
  - Population structure and coalescent theory
  - Selection, evolution, and phylogenetic trees
  - Epigenetics, imprinting, chromatin structure, and X inactivation
- Part 2. Paper Presentations (1/3) – each student will be assigned two papers– One as a presenter, the second as a discussant
  - Haplotype assembly from sequence data
  - De novo genome and transcriptome assembly
  - Sequence archival, comparative genomic analysis, genomic compression, and query
  - Detecting structural variation using sequence and genotype data
  - Inferring chromatin structure using sequence and methylation data
- Part 3. Project Proposals (1/3)



Do you want to be genotyped?

- I have 10 donated kits and will get more if needed
- Analyze yourself using the tools we develop
- Contribute your genotypes if you want others to use
- You must be signed up (not auditing) to be genotyped
- Who in this class is your closest relative?



23andMe, Inc. 11/22/13

Department of Health and Human Services

Public Health Service
Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993

Nov 22, 2013

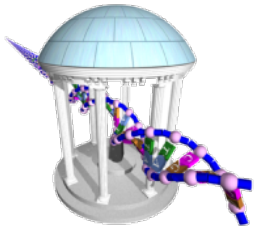
Ann Wojcicki
CEO
23andMe, Inc.
1390 Shoreline Way
Mountain View, CA 94043

Document Number: GEN1300666
Re: Personal Genome Service (PGS)

WARNING LETTER

Dear Ms. Wojcicki,

patients or if incorrect test results are reported.

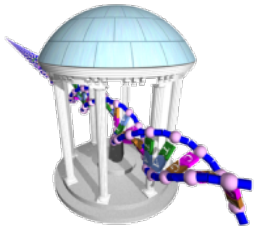


It's about genes

- Genetics is most clearly understood by considering its subject to be genes rather than organisms
- Organisms are merely vessels for assuring the survival of genes
- The central objective of a gene is to propagate itself
- Successful genes live on long after their host organism

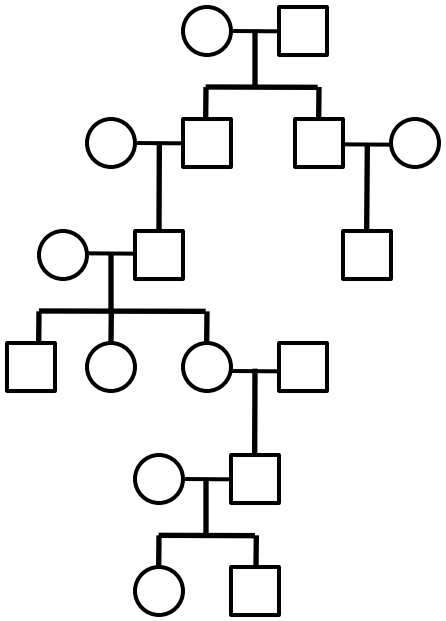
"[Genes] that survived were the ones that built survival machines for themselves to live in. But making a living got steadily harder as new rivals arose with better and more effective survivial machines. Survival machines got bigger and more eloborate, and the process was cumulative and progressive…"

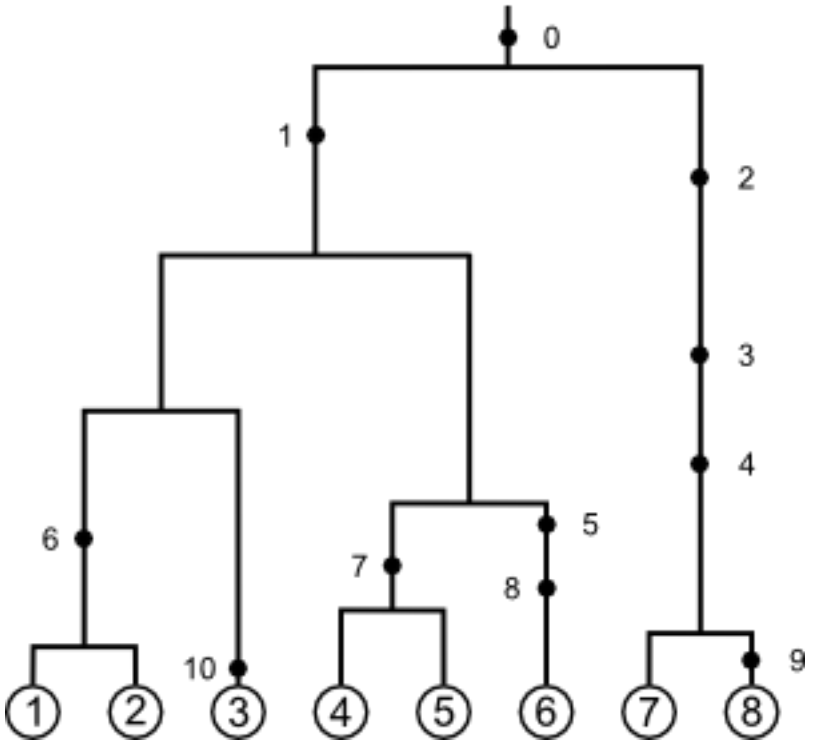
-- Dawkins, The Selfish Gene

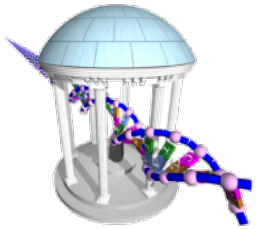


Top-down and bottom-up

- Genetics results from inheritance
- Ancestral properties can be inferred from extant populations

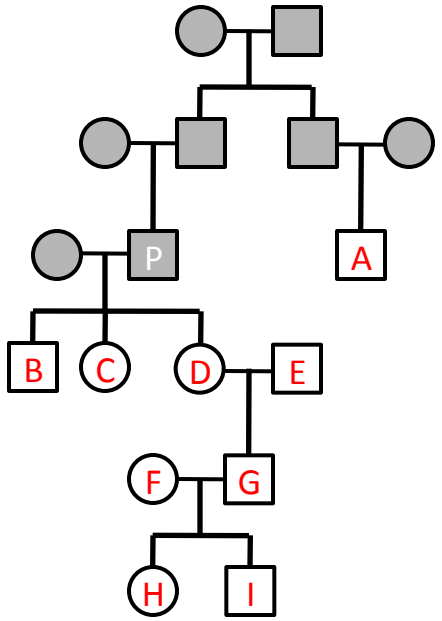


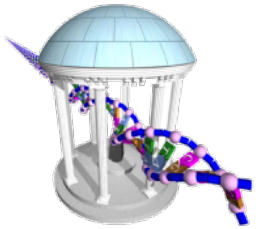




Exercise

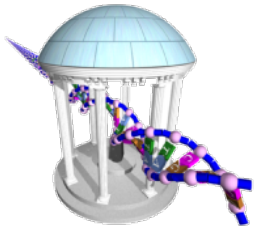
- Answer the following
  - How many distinct Y chromosomes are in this pedigree and who shares them?
  - How many distinct Mitochondrial DNAs are in this pedigree and who shares them?
  - Explain how you might reconstruct the *paternal* X chromosome of siblings B, C, and D?
  - What fraction of DNA is shared between G and H?
  - What fraction of DNA is shared by H and I?
  - Suppose that the samples, P, A, D, G, and I have a common phenotype not present in the remainder. Which two samples are most helpful for localizing the genetic component?





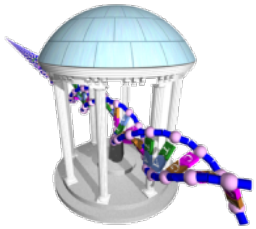
Population genetics

- Population genetics differs from classical genetics
  - Analysis rather than synthesis
  - Depends on models, which attempt to explain observations
  - Less emphasis on Darwin's natural selection
- Considers population dynamics
  - Isolation
  - Bottlenecks



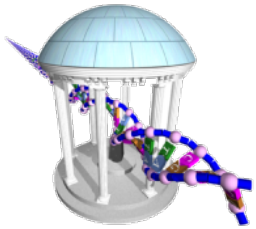
Why computer science

- Classically, genetics, both top-down (classical) and bottom-up (population) have focused on mathematical/statistical models
- Access to genetics data has recently outpaced our capability to process, interpret, and analyze it
- As model complexity increases, it becomes harder to find closed-form solutions
- Relies more and more on computational modeling to infer structure, account for noise etc.



Wright-Fisher model

- One of the first, and simplest models of population genealogies was introduced by Wright (1931) and Fisher (1930).
- Model emphasizes transmission of genes from one generation to the next
- For simplicity we'll first focus on a fixed population size, each with a distinct gene variant

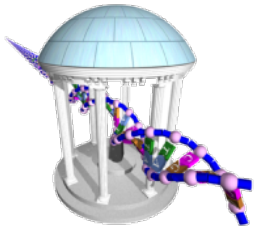


Simple haploid model

- Rules
  - Antecedent genes are chosen randomly, with replacement, from their parental generation
  - No selection
  - Fixed population size

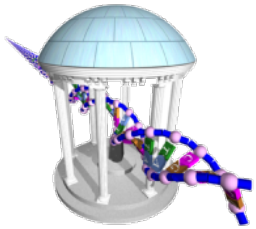
```
G0: ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J']
G1: ['J', 'A', 'H', 'B', 'I', 'E', 'D', 'G', 'A', 'B']
G2: ['A', 'J', 'E', 'G', 'D', 'E', 'B', 'I', 'A', 'A']
G3: ['A', 'A', 'E', 'J', 'I', 'A', 'I', 'A', 'J', 'B']
G4: ['E', 'A', 'B', 'B', 'A', 'E', 'A', 'A', 'A', 'A']
G5: ['A', 'A', 'B', 'A', 'A', 'E', 'A', 'A', 'A', 'B']
G6: ['A', 'A', 'A', 'A', 'A', 'A', 'A', 'B', 'A', 'A']
G7: ['B', 'A', 'A', 'B', 'A', 'A', 'A', 'A', 'A', 'A']
```

What will this population eventually look like?



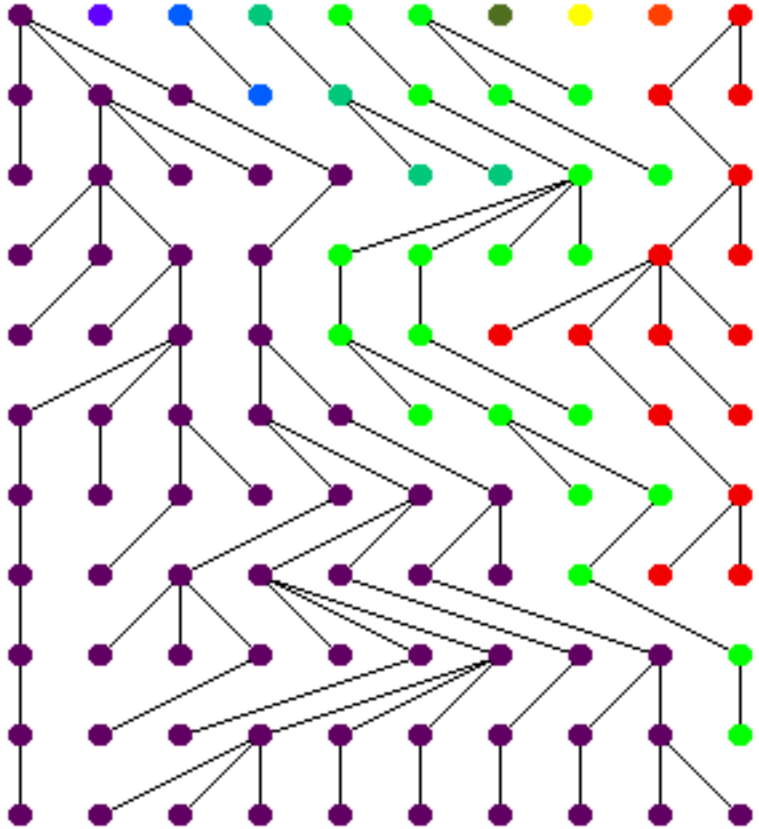
Assumptions of Wright/Fisher

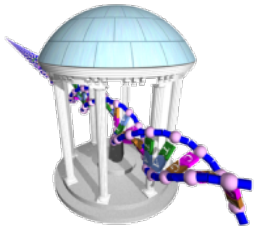
- Discrete and non-overlapping generations
- Haploid individuals
- Populations size is constant
- All individuals are equally fit
- No population of social structure
- Genes segregate independently



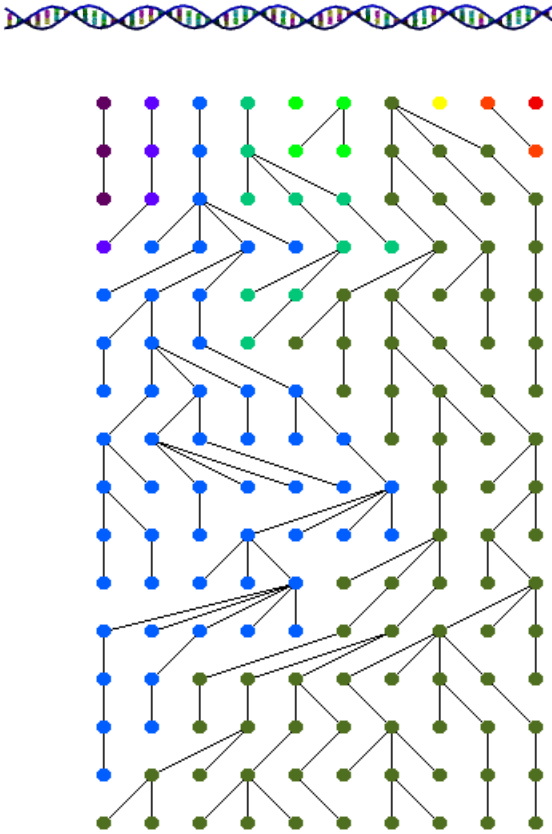
Some Graphical Abstractions

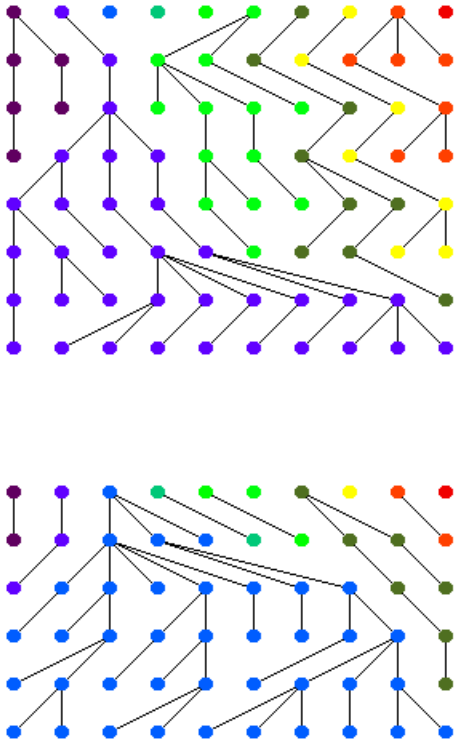
- Replace letters with colors
- Draw lineages
- Sort topologically

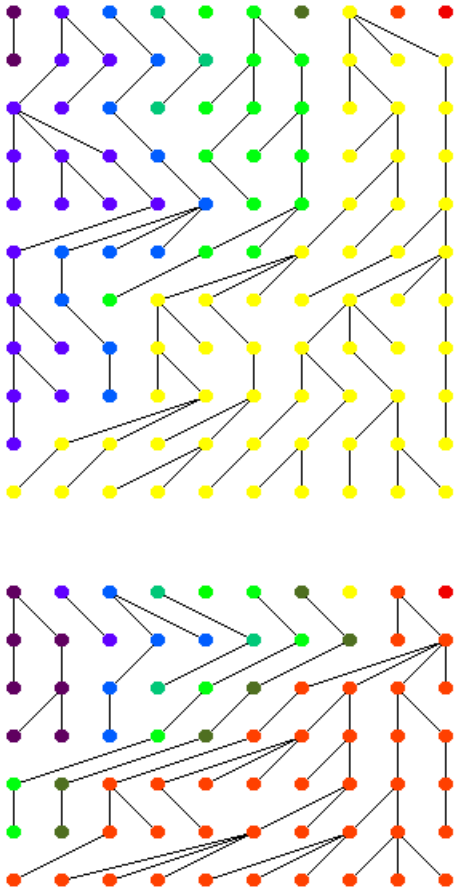




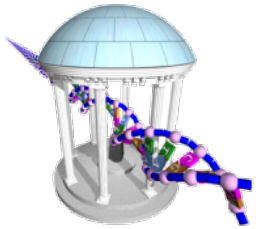
Repeats





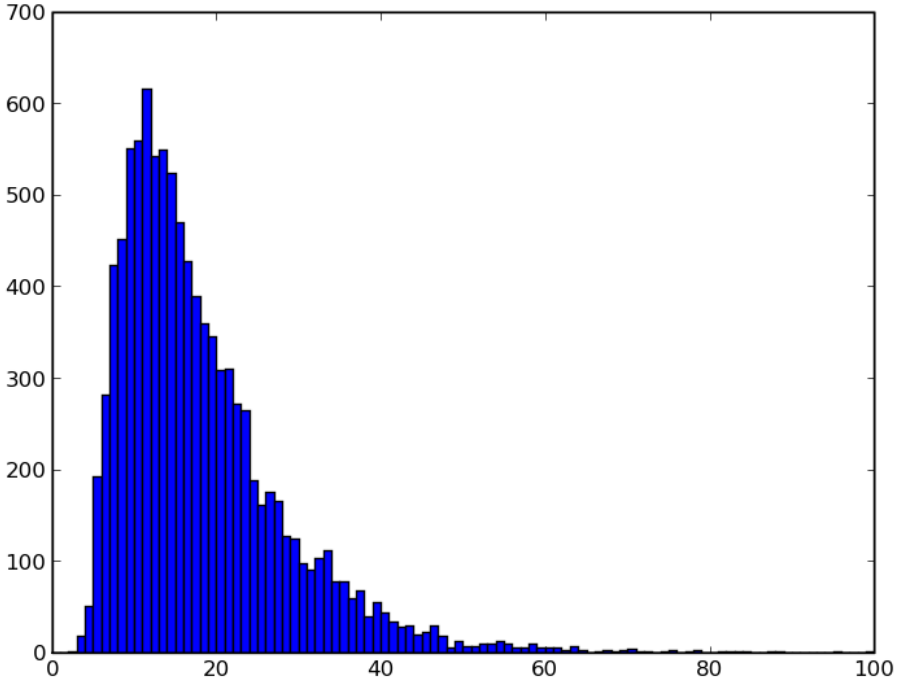


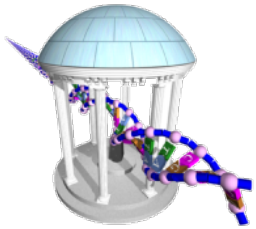
Every population results in just one gene



Onset of uniformity

- 10000 trials
- Mode = 11 (616)
- Mean = 17.5





Next Time

- Genetics Background
  - Inbreeding
  - Coefficient of relatedness
  - Diploidy
  - Recombination
  - Phasing
- Bring your laptops loaded with Python 2.7
  - We will analyze real genotypes
- A list of recent papers to choose from