Lecture 17

Whole-genome comparative genomics:
Genome Assembly, Genome Alignment, Genome Duplication

Slides credit:
Serafim Batzoglou
Pardis Sabeti
David Reich
<table>
<thead>
<tr>
<th>Project</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Lec</th>
<th>Topic</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS1</td>
<td>1</td>
<td>Thu, Sep 7</td>
<td>Introduction</td>
<td>L1</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Sep 8</td>
<td>R1</td>
<td>Recitation 1: Biology and Probability Review</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Tue, Sep 12</td>
<td>L2</td>
<td>Alignment I: Dynamic Programming, Global and local alignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Sep 14</td>
<td>L3</td>
<td>Feature extraction and machine learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Tue, Sep 19</td>
<td>L4</td>
<td>Alignment II: Database search, Rapid string matching, BLAST, BLOSUM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Sep 21</td>
<td>R2</td>
<td>Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Sep 22</td>
<td>L5</td>
<td>Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Sep 22</td>
<td>L6</td>
<td>Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mon, Sep 25</td>
<td>R3</td>
<td>No classes - student holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Oct 6</td>
<td>L7</td>
<td>Expressions Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Thu, Oct 6</td>
<td>R4</td>
<td>Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, Chrompute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 6</td>
<td></td>
<td>Project intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Tue, Oct 10</td>
<td>L8</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expressed Genes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 12</td>
<td>L9</td>
<td>Recitation 5: Affinity Propagation Clustering and Random Forest Classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Thu, Oct 17</td>
<td>L10</td>
<td>Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 20</td>
<td>R5</td>
<td>Recitation 6: ENCODE, Epigenome Roadmap, ChromHMM, Chrompute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 20</td>
<td>L11</td>
<td>Three-dimensional chromatin interactions: 3C, 5C, Hi-C, ChiA-Pet</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 20</td>
<td></td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507 at 4-5pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Tue, Oct 24</td>
<td>L12</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 26</td>
<td>L13</td>
<td>Deep Learning, Neural Nets, Convolutional NNS, Recurrent NNS, Autoencoder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Thu, Oct 31</td>
<td>L14</td>
<td>Networks review, Recommendation systems, EHR, PhenWAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 3</td>
<td></td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507 at 4-5pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Tue, Nov 7</td>
<td>L15</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 8</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 10</td>
<td>L16</td>
<td>Deep Learning, Neural Nets, Convolutional NNS, Recurrent NNS, Autoencoder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 10</td>
<td></td>
<td>No lectures - Columbus Day holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Tue, Nov 14</td>
<td>L17</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 16</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 17</td>
<td>L18</td>
<td>Deep Learning, Neural Nets, Convolutional NNS, Recurrent NNS, Autoencoder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 17</td>
<td></td>
<td>No lectures - Columbus Day holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Tue, Nov 21</td>
<td>L19</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 23</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Tue, Nov 29</td>
<td></td>
<td>No lectures - Columbus Day holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 5</td>
<td>L20</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 7</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Dec 8</td>
<td>L21</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Dec 8</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>Tue, Dec 12</td>
<td></td>
<td>No lectures - Columbus Day holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td>L22</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td>R11</td>
<td>Recitation 11: Presentation Tips – Intro, discussion, Slides, Presentation skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td>L23</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td>R12</td>
<td>Recitation 12: Presentation Tips – Intro, discussion, Slides, Presentation skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td>L24</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>Thu, Dec 12</td>
<td></td>
<td>No lectures - Columbus Day holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td>L25</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td>L26</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 12</td>
<td>L27</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>Thu, Dec 12</td>
<td></td>
<td>No lectures - Columbus Day holiday</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar
** Recitation topics will be adjusted to respond to lecture and student needs
**Module V: Comparative genomics and evolution**

| Foundation | L17 | Comparative genomics and evolutionary signatures | 4 |
| Foundation | L18 | Genome Scale Evolution, Genome Duplication | 5,6 |
| Frontiers | L19 | Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference | 27 |
| Frontiers | L20 | Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs | 28 |
| Frontiers | R9 | Recitation 9: Phylogenetic distance metrics, Coalescent Process | |

- **L17: Comparative genomics & evolutionary signatures**
  - Genome-wide studies → infer signatures → annotate elements
  - Protein-coding, non-coding RNA, microRNA, regulatory motifs
- **L18: Genomic rearrangements and genome duplication**
  - Synteny-based alignment, Beyond nucleotide mutation
  - Detecting structural changes: genome assembly/alignment
- **L19: Phylogenetics**
  - Evolutionary rates and models of evolution
  - Distance-based trees, DP on a tree, Bayesian models
- **L20: Phylogenomics**
  - Gene trees vs. species trees, reconciliation, coalescence
The age of comparative genomics

32 mammals

14 yeasts

12 flies

human  chimp  mouse  rat  dog

opossum  armadillo  rabbit  cow  hyrax  elephant

bat  dolphin  lemur  bushbaby  pika  hedgehog  tenrec

llama  Tree shrew  pangolin

e etc...
Today: Genome evolution, alignment, evolution

- Genome assembly
  - Consensus-layout-overlap methods
  - String graph methods
- Whole-genome alignment
  - Resolving region correspondence
  - Gene-based global genome alignment
  - Mechanisms of rapid genome evolution
- Whole genome duplication (WGD)
  - Evolutionary signature of whole genome duplication
  - Post-duplication emergence of new functions
- Recent human evolution
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   - Consensus-layout-overlap methods
   - String graph methods

2. Foundations 2: Whole-genome alignment
   - Resolving region correspondence
   - Gene-based global genome alignment
   - Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   - Evolutionary signature of whole genome duplication
   - Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   - Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   - Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   - 1000Genomes population differences. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (F_st). PC correction
Fragment Assembly

Section “borrowed” from Serafim Batzoglou
Whole Genome Shotgun Sequencing

cut many times at random

plasmids (2 – 10 Kbp)
cosmids (40 Kbp)

forward-reverse paired reads

known dist

~500 bp
~500 bp

genome
Genome assembly I: Overlap-layout-consensus
Steps to Assemble a Genome

1. Find overlapping reads

2. Merge some “good” pairs of reads into longer contigs

3. Link contigs to form supercontigs

4. Derive consensus sequence
### Some Terminology

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>read</strong></td>
<td>a 500-900 long word that comes out of sequencer</td>
</tr>
<tr>
<td><strong>mate pair</strong></td>
<td>a pair of reads from two ends of the same insert fragment</td>
</tr>
<tr>
<td><strong>contig</strong></td>
<td>a contiguous sequence formed by several overlapping reads with no gaps</td>
</tr>
<tr>
<td><strong>supercontig</strong> (scaffold)</td>
<td>an ordered and oriented set of contigs, usually by mate pairs</td>
</tr>
<tr>
<td><strong>consensus sequence</strong></td>
<td>sequence derived from the multiple alignment of reads in a contig</td>
</tr>
</tbody>
</table>
1. Find Overlapping Reads

- Sort all k-mers in reads \((k = 24)\)
- Find pairs of reads sharing a k-mer
- Extend to full alignment – throw away if not >97% similar
2. Merge Reads into Contigs

Merge reads up to potential repeat boundaries

repeat region

Unique Contig

Overcollapsed Contig

Merge reads up to potential repeat boundaries
2. Merge Reads into Contigs

- **Overlap graph:**
  - Nodes: reads $r_1 \ldots r_n$
  - Edges: overlaps ($r_i$, $r_j$, shift, orientation, score)

Remove transitively inferrable overlaps
Overlap graph after forming contigs
Repeats, errors, and contig lengths

- Repeats shorter than read length are OK
- Repeats with more base pair diffs than sequencing error rate are OK

- To make the genome appear less repetitive, try to:
  - Increase read length
  - Decrease sequencing error rate

Role of error correction:
Discards ~90% of single-letter sequencing errors
  - decreases error rate
  - decreases effective repeat content
  - increases contig length
4. Derive Consensus Sequence

Derive multiple alignment from pairwise read alignments

Derive each consensus base by weighted voting

(Alternative: take maximum-quality letter)
Genome assembly II:
String graph manipulation
Popular genome assemblers

- **PHRAP (Green 2002/03/08)**
  - Early assembler, widely used, good model of read errors
  - Overlap $O(n^2)$ -- layout (no mate pairs) -- consensus

- **Celera (Myers, 2004/10)**
  - First assembler to handle large genomes (fly, human, mouse)
  - Overlap -- layout -- consensus

- **Arachne (Batzoglou, 2002)**
  - Public assembler (mouse, several fungi)
  - Overlap -- layout -- consensus

- **Euler (Pevzner, 2001/06)**
  - Indexing -- deBruijn graph -- picking paths -- consensus

- **Velvet (Birney, 2010)**
  - Short reads -- small genomes -- simplification -- error correction

- **ALLPATHS (Gnerre, 2011)**
  - Short reads -- large genomes -- jumping data -- uncertainty
Given a shotgun dataset of reads we should be able to build a graph that looks like this:

There are two possible tours:

Myers 2005
How To Build A String Graph

- Remove Transitive Overlaps
  \( O(E) \) expected-time alg.
Collapsing chains

- Collapse Chains
- Compressed Edge
- Junction
- DNA can be read in 2 directions
- Reads can be used in either direction
- Junction points are directed
- An edge can be used in both directions
Want \(a + b + c = x + y + z\)

Classify edges:

- \(= 1\): High probability that edge is unique: \(P(\text{non-unique}) < e^{-18}\)
- \(\geq 0\): No information
- \(\geq 2\): Edge is likely over-collapsed
Dealing with errors in real datasets

- Reads from multiple places in the genome (chimers)

- Some overlaps are missed due to errors and polymorphisms
Error Correction Algorithm

- Build local alignments between all read pairs
  We use a very fast $O(N+d^2)$ algorithm

- Fix parts of reads (indels, mutations) that are not supported by any read and are contradicted by at least 2

- Some errors are impossible to fix
Achieve a Feasible Flow

- Remove fewest number of reads: add back-edges
  Penalty for back-edge equal to number of reads

- Edge + back edge form a cycle: edge eliminated
Iterating Flow Solving

- On larger genomes there may not be a unique min cost flow
- We can iterate flow solving:
  - Add $\varepsilon$ penalty to all edges in solution
  - Solve flow again – if there is an alternate min cost flow it will now be smaller
  - Repeat until no new edges

- Edges are labeled
  - Required In all solutions
  - Unreliable In some solutions
  - Unneeded In no solutions
Example of a resulting string graph (C. jejuni)

- 1.7 Mb; 24,000 reads
- Initial graph: 129 nodes, 174 edges
- After Flow solving (< 3 minutes total run time):
  - 22 nodes 35 edges
  - 4 edges (5 reads) rejected
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   - Consensus-layout-overlap methods
   - String graph methods

2. Foundations 2: Whole-genome alignment
   - Resolving region correspondence
   - Gene-based global genome alignment
   - Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   - Evolutionary signature of whole genome duplication
   - Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   - Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   - Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   - 1000Genomes population differences. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction
Whole-genome alignment
Multiple alignment is $O(n^{\#\text{species}})$

$F(i,j) = \text{Score of best alignment ending at } i,j$

<table>
<thead>
<tr>
<th>F(i-1,j)</th>
<th>F(i,j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(i,j-1)</td>
<td>F(i,j-1)</td>
</tr>
</tbody>
</table>

Time: $O(n^2)$ for two seqs,  
$\Omega(n^k)$ for $k$ seqs
LAGAN: 1. FIND Local Alignments

1. Find Local Alignments
2. Chain Local Alignments
3. Restricted DP

Brudno, Do 2003
LAGAN: 2. CHAIN Local Alignments

1. Find Local Alignments
2. Chain Local Alignments
3. Restricted DP
LAGAN: 3. Restricted DP

1. Find Local Alignments

2. Chain Local Alignments

3. Restricted DP
MLAGAN: 1. Progressive Alignment

Given N sequences, phylogenetic tree

Align pairwise, in order of the tree (LAGAN)
Glocal alignment
Local & Global Alignment

Local

Global
Glocal Alignment Problem

Find least cost transformation of one sequence into another using new operations

- Sequence edits
- Inversions
- Translocations
- Duplications
- Combinations of above

AGTGCCCTGGAACCCTGACGGTGGGTCACAAAACTTCTGGA
AGTGACCTGGGAAGACCCTGAACCCTGGGTCACAAAACTC
S-LAGAN Results (CFTR)
Our goals for today

• Genome assembly
  – Consensus-layout-overlap methods
  – String graph methods

• Whole-genome alignment
  – Resolving region correspondence
  – Gene-based global genome alignment
  – Mechanisms of rapid genome evolution

• Whole genome duplication
  – Evolutionary signature of WGD
  – Emergence of new functions
Gene-based global genome alignment
Comparative Genomics

Lecture 17 (Today):
Using evolution to study genomes

Evolution

Genomics

Lectures 18-20 (module V):
Using genomics to study evolution
**Goal:** Find corresponding regions

- **Aligning four genomes: methodology**
  - Anchor genomic segments by the genes they contain
  - Resolve the correspondence of genes for each pair of species
  - Construct nucleotide-level alignment

- **What information is available**
  - Amino-acid similarity of every gene pair across genomes
  - Locations of genes in each genome

- **What makes it hard?**
  - Not all regions have one-to-one correspondence
  - Gene divergence, duplication and loss.
  - Genome rearrangements.
Gene-based global genome alignment

- **BUS algorithm (Best Unambiguous Subgroups)**
  - Resolve correspondence of genes and regions
  - Uses complete bipartite graph connectivity
  - Integrates protein similarity and gene order information

- **Correctly resolved gene correspondence**
  - More than 90% of genes have 1-to-1 correspondence
  - Identified regions and protein families of rapid change

**Computational tools give us the ability to automatically align complete genomes**
Framework: graph of gene correspondence

- **Weighted bipartite graph**
  - Graph represents gene correspondence
  - Nodes: genes (w/ coordinates)
  - Edges: sequence similarity (w/ weights)

- **Two types of evolutionary relationships**
  - Orthologs (1-to-1 matches)
  - Paralogs (1-to-many / many-to-many)

- **Method**
  - Eliminate spurious edges (simplify graph)
  - Select edges based on available information
    - Blocks of conserved gene order
    - Protein sequence similarity
Definition: Best Unambiguous Subgroups (BUS)

Extend concept of best-bidirectional hits
Implementation: Iterative refinement

Iterate

Full bipartite graph $G=(X+Y,E)$

Separate connected components of $M$

Directed Graph $D$

Maximal out-edges $M$, relative threshold $T$

Iterative refinement with increasing relative threshold
Conservation of gene order (synteny)

S. cerevisiae

S. paradoxus

 Preferentially select edges in synteny blocks
We are comparing orthologous regions
Conservation of local gene order and spacing

S. cerevisiae

S. paradoxus

S. mikatae

S. bayanus

S. cerevisiae Chromosome VI 250-300kbp
Today: Genome evolution, near and far

1. **Foundations 1: Building the genome assembly**
   - Consensus-layout-overlap methods
   - String graph methods

2. **Foundations 2: Whole-genome alignment**
   - Resolving region correspondence
   - Gene-based global genome alignment
   - Mechanisms of rapid genome evolution

3. **Evolutionary “jumps”: Whole genome duplication (WGD)**
   - Evolutionary signature of whole genome duplication
   - Post-duplication emergence of new functions

4. **Recent evolution 1: Human selection at multiple timescales**
   - Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   - Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. **Recent evolution 2: Human demographic history.**
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. **Recent evolution 3: Human relatedness and ancestry painting**
   - 1000Genomes population differences. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction
Mechanisms of genome evolution
Identify regions of rapid change

Protein family expansions in chromosome ends
Chromosomal Evolution

i Telomeric expansion

ii Chromosomal exchange

iii Transposition

iv Inversions

v Segmental duplication

vi Intein insertion, transposons
Differences in gene content

- **8-10 genes unique to each genome**
  - Metabolism, regulation/silencing, stress

- **Changes in gene dosage**
  - 10-20 tandem duplications (1-2 genes)
  - 2 segment duplications (5-6 genes)

- **Protein family expansions**
  - 211 genes (3%) with ambiguous correspondence
  - Paralog duplication and/or loss

**Different species, few novel genes**
Specific mechanisms mediate rearrangements

- 10 translocations
  - 8 across Ty elements
  - 2 across nearly identical genes

- 19 inversions
  - All flanked by tRNA genes

Evolutionary features
Rapid protein change

• Protein domain creation
  – Protein-protein interaction

• Compensatory frame-shifts
  – Explore new reading frames
  – RNA editing signals

• Stop-codon variation
  – Gain enables rapid change
  – Loss explores new diversity
  – Read-through is regulated

• Intein gain
  – Recent, present in S.cerevisiae only

Evolutionary shortcuts apparent in recent evolution
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   - Consensus-layout-overlap methods
   - String graph methods

2. Foundations 2: Whole-genome alignment
   - Resolving region correspondence
   - Gene-based global genome alignment
   - Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   - Evolutionary signature of whole genome duplication
   - Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   - Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   - Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   - 1000Genomes population differences. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction
Whole-Genome Duplication
Further back in evolutionary time

- 100 Myr: K. waltii, S. bayanus, S. mikatae, S. paradoxus, S. cerevisiae
- 20 Myr: S. cerevisiae
- 5 Myr: S. cerevisiae

Ability to ask different set of questions
Gene correspondence

S. cerevisiae chromosomes

K. waltii scaffolds
Sister regions show gene interleaving

Gene interleaving is evidence of complete duplication

Few genes remain in 2 copies
Duplicate mapping of centromeres

Recognize sister regions solely based on gene order
145 blocks cover 88% of genome

Duplicate mapping tiles \textit{S. cerevisiae}
Conclusion: Whole Genome Duplication *has* happened.
Whole Genome Duplications are everywhere!

Yeast Duplication
- Most genes 1-to-1 mapping
- Gene interleaving evidence of duplication
- Complete tiling of the genome

Vertebrate Duplication in Fish
- Fish: Gene order not conserved, only chromosomes
- Mammals: Gene order conserved, not chromosomes

Two rounds of WGD in base of vertebrate lineage
- Build clusters of related genes (use Ciona as outgroup)
- Count duplications by reconciliation
- Find regions of duplicate overlap → 4-way synteny

1R, 2R, and 3R duplications in vertebrate evolution
Genome duplication evidence in a single species
Evidence of duplication using a single genome?

- **Genomic evidence**
  - Conserved order of paralogous genes
  - Same transcriptional orientation
- **However**
  - Interspersed with single-copy genes

**Interpretation:** Genome duplication followed by gene loss
Whole genome duplication is controversial

- **Insufficient evidence**
  - Only 50% of genome in duplicate regions
  - Only 8% of genes present in two copies
  - Extensive redundancy outside duplicate regions

- **Evidence against WGD**
  - Divergence-based dating show multiple times
  - Other species have similar level of redundancy

- **Alternative evolutionary scenario proposed**
  - Independent segmental duplications
  - Also consistent with the evidence

---

- “There was a whole-genome duplication.” Wolfe, Nature ‘97
- “There was no whole-genome duplication.” Dujon, FEBS 2000
- “At least some chrom dup. occurred independently” Langkjaer, JMB, 2000
- “Dynamic equilibrium of duplications and loss” Llorente, FEBS, 2000
- “Recent evidence supports single event”. Wong, PNAS ‘02
- “Continuous block duplications and deletions” Dujon, Yeast 2003
- “Telomere-mediated duplication events” Coissac, Mol Bio Evo 1997
- “Multiple closely spaced events” Friedman, Genome Res, 2003
- “Spontaneous duplication of large chromosomal segments” Koszul, EMBO ’04
Conclusion: Whole Genome Duplication *has happened*
Our goals for today

• **Genome assembly**
  – Consensus-layout-overlap methods
  – String graph methods

• **Whole-genome alignment**
  – Resolving region correspondence
  – Mechanisms of rapid genome evolution

• **Whole genome duplication**
  – Evolutionary signature of WGD
  – Emergence of new functions
Post-duplication evolution
Whole-genome duplication resolved

Number of genes

<table>
<thead>
<tr>
<th>Time</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Myrs</td>
<td>5,000</td>
</tr>
<tr>
<td>Today</td>
<td>5,500 (~500 gained)</td>
</tr>
</tbody>
</table>

WGD

Gene Loss
Fate of duplicated genes

- 457 genes kept in two copies, result of selection
  - Involved in sugar metabolism and fermentation

Evidence of accelerated protein divergence?
Scenarios for rapid gene evolution

One copy faster

- Scer - copy1
- Scer - copy2
- Kwal

Both copies faster

- Scer - copy1
- Scer - copy2
- Kwal

Ohno, 1970
Lynch, 2000

20% of duplicated genes show acceleration
95% of cases: Only one copy faster
Emerging gene functions after duplication

- **Origin of replication → silencing**
  - 4-fold acceleration
    - Scer - Sir3 (silencing)
    - Scer - Orc1 (origin of replication)
    - Kwal - Orc1

- **Translation initiation → anti-viral defense**
  - 3-fold acceleration
    - Scer - Ski7 (anti-viral defense)
    - Scer - Hbs1 (translation initiation)
    - Kwal - Hbs1

Asymmetric divergence → recognize ancestral / derived
## Distinct functional properties

<table>
<thead>
<tr>
<th>Gene deletion</th>
<th>Ancestral function</th>
<th>Derived function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lethal (20%)</td>
<td>Never lethal</td>
</tr>
</tbody>
</table>

Gain new function and lose ancestral function
## Distinct functional properties

<table>
<thead>
<tr>
<th></th>
<th>Ancestral function</th>
<th>Derived function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene deletion</td>
<td>Lethal (20%)</td>
<td>Never lethal</td>
</tr>
<tr>
<td>Expression</td>
<td>Abundant</td>
<td>Specific</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(stress, starvation)</td>
</tr>
<tr>
<td>Localization</td>
<td>General</td>
<td>Specific</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(mitochondrion, spores)</td>
</tr>
</tbody>
</table>

Gain new function and lose ancestral function
Evolutionary genomics in yeast

- **Genome ancestry resolved**
  - Whole-genome duplication
  - Massive gene loss

- **Emergence of new functions**
  - Asymmetric acceleration
  - Ancestral and derived functions
  - Repository for buffering mutations
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   – Consensus-layout-overlap methods
   – String graph methods

2. Foundations 2: Whole-genome alignment
   – Resolving region correspondence
   – Gene-based global genome alignment
   – Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   – Evolutionary signature of whole genome duplication
   – Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   – Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   – Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   – Population size, dynamics, bottlenecks, expansion, effective pop size.
   – Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   – 1000Genomes population differences. SNP sharing. IBD/heritability.
   – Admixture, ancestry painting. PCA/MDS clustering (\(F_{ST}\)). PC correction
Molecular evolution

- Fundamental question: how does allele frequency change over time?
- Under strong assumptions (no mutation, no migration, no selection, large population size, random mating), allele/genotype frequencies remain in Hardy-Weinberg equilibrium

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.6</td>
</tr>
<tr>
<td>G</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.6 \times 0.6 = 0.36</td>
</tr>
<tr>
<td>AG/GA</td>
<td>2 \times 0.6 \times 0.4 = 0.48</td>
</tr>
<tr>
<td>GG</td>
<td>0.4 \times 0.4 = 0.16</td>
</tr>
</tbody>
</table>
• Under strong assumptions (no selection, **constant** population size, random mating, **no overlapping generations**), change in allele frequency of a new mutation over time is described by the **Wright-Fisher process**

• Without selection, main force on allele frequency is **genetic drift** (random chance)

• Natural selection changes the course of the process (either towards 0 or 1)
• Fundamental question: how have allele frequencies changed over time?
• Different methods give insight into change over different time scales
Test for positive selection

- Mutations which were beneficial to *H. sapiens* specifically will not be conserved with other primates.
- The **McDonald-Kreitman test** identifies sites which are under selection within a species.
- Idea: compare the amount of variation within the species (polymorphism) with the amount between species (substitutions).

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>$D_s$</td>
<td>$P_s$</td>
</tr>
<tr>
<td>Non-synonymous</td>
<td>$D_n$</td>
<td>$P_n$</td>
</tr>
</tbody>
</table>

\[
\frac{D_n}{D_s} < \frac{P_n}{P_s} \text{ implies purifying selection} \\
\frac{D_n}{D_s} > \frac{P_n}{P_s} \text{ implies positive selection}
\]
Molecular evolution

• Fundamental question: how have allele frequencies changed over time?
• Different methods give insight into change over different time scales
Measuring Derived Allele Frequency

Non-ancestral (derived) allele is T, it is at 20% in the population
Molecular evolution

- Fundamental question: how have allele frequencies changed over time?
- Different methods give insight into change over different time scales
We can measure the differences in allele frequency in populations by the statistic $F_{st}$.

$F_{st}$ estimates the reduction in heterozygosity expected when 2 different populations are erroneously grouped.
2 isolated populations

Subpopulation 1
P(A) = 1
P(AA) = 1

Subpopulation 2
P(T) = 1
P(TT) = 1
**Fst**

<table>
<thead>
<tr>
<th>Subpopulation 1</th>
<th>Subpopulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(A) = 1</td>
<td>P(A) = 0</td>
</tr>
<tr>
<td>P(T) = 0</td>
<td>P(T) = 1</td>
</tr>
</tbody>
</table>

Total

P(A) = 0.5
P(T) = 0.5

Heterozygosity (total) = 2pq = 0

\[
Fst = \frac{\text{Heterozygosity (total)} - \text{Heterozygosity (subpopulations)}}{\text{Heterozygosity (total)}}
\]

\[
Fst = \frac{0.5 - 0}{0.5} = 1
\]
Population 1: AA (4) and TT (0), AT (1)
Population 2: AA (2) and TT (0), AT (3)
Fst

Subpopulation 1
P(A) = 0.9
P(T) = 0.1

Subpopulation 2
P(A) = 0.7
P(T) = 0.3

Total
P(A) = 0.8
P(T) = 0.2

Heterozygosity (total) – Heterozygosity (subpopulations)

Heterozygosity (total)

H = 0.30

H = 0.18

H = 0.42

H = 0.32

Fst = \frac{\text{Heterozygosity (total)} - \text{Heterozygosity (subpopulations)}}{\text{Heterozygosity (total)}}

Fst = \frac{0.32 - 0.30}{0.32} = 0.0625
Molecular evolution

- Fundamental question: how have allele frequencies changed over time?
- Different methods give insight into change over different time scales

Diagram showing evolutionary timelines and relationships between Africa, Asia, and Europe over 6, 250, 75, and 25 kya, with nodes indicating proportion of functional changes, heterozygosity/rare alleles, high frequency derived alleles, population differences, and length of haplotypes.
Decay of LD a function of the recombination rate \( r \) and time.
• Signature of recent positive selection: common variants on long haplotypes (why?)
Positive selection on lactase persistence

The gene LCT encodes the enzyme lactase, required to digest milk.

Evidence of positive selection on European LCT: the persistence haplotype is present in 77% of Europeans, extends 1MB.

Sabeti et al. Science 2006
Genomics Signals of Natural Selection

Tests:
1) Long-range correlations
   • iHS, XP-EHH
2) High frequency derived
3) High differentiation
   • $F_{ST}$

Differentiated
Combining scores can localize regions

1. Long Haplotype

2. Derived allele

3. Differentiated

![Diagram showing haplotype and allele frequency distribution across different positions.](image-url)
Combining scores can localize regions

**Single Test**

- Causative SNP
- Neutral Neighbors
- Neutral

**Composite**

- Frequency range: 0.00 to 0.10
- Frequency range: 0.00 to 0.11

**Position relative to causal (cM)**

- ~50k
- ~50k b
~300 regions detected in genome surveys
Forces Shaping Human Evolution

1950s: HBB
Resistance to Malaria

1990s: LCT
Lactose tolerance

2000s: Scans of the Human genome
Positive selection on height

- Difference in height between northern and southern Europeans driven by selection
- Analysis of ancient Eurasians reveals selection for reduced height in Iberia, increased height in steppe populations

Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   - Consensus-layout-overlap methods
   - String graph methods

2. Foundations 2: Whole-genome alignment
   - Resolving region correspondence
   - Gene-based global genome alignment
   - Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   - Evolutionary signature of whole genome duplication
   - Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   - Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   - Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   - 1000Genomes population differences. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction
Purifying selection on coding sequences

• Selection against new mutations drives allele frequency to zero

• Exome Aggregation Consortium (60K samples!) has shed new insight into purifying selection on protein-coding sequences

• Idea: predict how many new mutations should fall in a gene given its properties, then compare to the observed number

• Discover excess of de novo mutations in rare disease cases (ASD)

Samocha et al. *Nat Genet* 2014
Purifying selection on non-coding sequences

• Selection against common mutations must be weaker (why?)
• Leads to difference in heterozygosity at SNPs in predicted regulatory regions

Ward and Kellis Science 2012
Purifying selection also operates on transcription factor binding sites
Purifying selection on non-coding sequences

• Purifying selection on putative functional elements also leads to reduction in derived allele frequency
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   – Consensus-layout-overlap methods
   – String graph methods

2. Foundations 2: Whole-genome alignment
   – Resolving region correspondence
   – Gene-based global genome alignment
   – Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   – Evolutionary signature of whole genome duplication
   – Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   – Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   – Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   – Population size, dynamics, bottlenecks, expansion, effective pop size.
   – Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   – 1000Genomes population differences. SNP sharing. IBD/heritability.
   – Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction
Many models in population genetics assume population sizes are constant (coalescent process)

The pairwise sequentially Markov coalescent is used to model population growth
• Mitochondrial DNA is inherited only from the mother, doesn’t recombine appreciably
• Accumulation of mutations in mtDNA corresponds to population divergence
Evidence for out of Africa migration

Analysis of aboriginals in Australia showed that:
• they share a bottleneck (~50k years ago) with Eurasians
• show divergence from Eurasians shortly thereafter (50-70k years ago)
⇒ same out-of-Africa migration event as Eurasians!
North Indians show evidence of admixture with Eurasians, unlike South Indians.

Evidence suggests migration 1-4KYA, concordant with history/mythology.
Recent migration into India

- Evidence of a cline (gradient) of admixture going from Europe, to North India, to South India.

Moorjani et al. *Am J Hum Genet* 2013
Conquest of South America

• Present day South Americans are a mixture of Native American, European, and African ancestry
• Sex bias in inferred ancestry: increased Native American ancestry on the X chromosome (why?)

Homburger et al. PLoS Genet 2015
Conquest of South America

• Admixture of European and Native American populations obvious from PCA, ancestry painting
• Cline of Native American admixture following geography
• Cline of European admixture in Argentina recapitulates history (arrived through the Andes, spread to the Atlantic)

Recent admixture in African Americans

Present day African Americans are an admixture of European, African, Native American ancestries.

Regional differences in proportion of European ancestry.

Possibly explained by one migration event (1802) or two events (1714, 1854).

Analysis of ancient human DNA

- 2% of all non-African genomes derives from the Neanderthal genome
- The Ust’ Ishim man (45 KYA) also has 2% Neanderthal DNA!
- Interbreeding between *H. sapiens* and Neanderthal 50-60 KYA
South Asians have approx. 5% ancestry from Denisovans, another hominid subspecies.

They have larger Denisovan haplotypes than Neanderthal fragments, implying ordering of interbreeding events (why?).

More Denisovan, Neanderthal ancestry on the X chromosome.
ARGs: Ancestral recombination graphs

- Idea: history of a locus across individuals described by a **coalescent tree** (like a phylogenetic tree)
- Loci on one haplotype are explained by the same tree
- Recombinations correspond to changing the tree between positions
- Model as pairwise sequentially Markov coalescent process
- Build an HMM where hidden states are coalescent trees (**ancestral recombination graph**)
- Integrate over all paths to estimate divergence times, population sizes

Summary: re-writing human population history

- Bottlenecks/founder effects: rare alleles suddenly rise in frequency due to small population size
- Selective sweeps: rare alleles suddenly rise in frequency due to positive selection
- Loss of heterozygosity in non-African populations
- Admixture between previously isolated populations
- Interbreeding between ancient humans and other hominids
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   – Consensus-layout-overlap methods
   – String graph methods

2. Foundations 2: Whole-genome alignment
   – Resolving region correspondence
   – Gene-based global genome alignment
   – Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   – Evolutionary signature of whole genome duplication
   – Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   – Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   – Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   – Population size, dynamics, bottlenecks, expansion, effective pop size.
   – Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   – 1000Genomes population differences. SNP sharing. IBD/heritability.
   – Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction
Number of variants varies greatly by population

- Over 100 million observed variants: 4-5M positions differ between each of us and the human reference
- Each of us carries 2-3K structural variants affecting 20mb of sequence
- Each of us carries hundreds of protein truncating variants, 10Ks of non-synonymous mutations
- African individuals have more variation in their genomes (why?)

Population size, bottlenecks and expansion

- **Effective population size**: number of individuals needed in idealized model to recapitulate population properties
- Here, recapitulate the **coalescent time** (next section, L20): time to most recent common ancestor
- **Pairwise Markov sequential coalescent model** with population splits/growth enables comparison within vs. between populations
- 1KG suggests shared history beyond 150 kya
- Non-African populations sustained a **bottleneck** 15-20 kya (migration out of Africa)
- After migration, rapid population expansion (with interesting exceptions: Finland, Peru, Mexico)

Ancestry painting (e.g. admixed individual)

Ancestry Composition tells you what percent of your DNA comes from each of 31 populations worldwide. This analysis includes DNA you received from all of your recent ancestors, on both sides of your family. The results reflect where your ancestors lived before the widespread migrations of the past few hundred years.

- **79.0%** Sub-Saharan African
- **72.3%** West African
- **2.9%** Central & South African
- **3.6%** Broadly Sub-Saharan African
- **18.4%** European
  - **2.5%** Northern European
  - **0.2%** Scandinavian
  - **11.4%** Broadly Northern European
  - **0.6%** Ashkenazi
  - **0.5%** Southern European
  - **3.3%** Broadly Southern European
- **1.9%** East Asian & Native American
  - **0.8%** Native American
  - **0.8%** Southeast Asian
  - **0.2%** Broadly East Asian & Native American
- **0.7%** Unassigned

100% TL Dixon

Which segments of a genome are shared with what populations
Ancestry painting using Latent Dirichlet Allocation (LDA)

- Originally proposed for document classification
- Words are associated with multiple topics
- Documents are associated with multiple topics, depending on which words they contain
- Here, individuals are documents, variants are words, and populations are topics
Latent Dirichlet Allocation (LDA)

- Assume each variant $l$ is independent
- The observed genotype $G$ arises from two unobserved haplotypes
- Each individual’s genome has probability (mixture proportion) $Q$ of coming from each ancestral population
- That means each haplotype $Z$ (per individual) has the same mixture proportion
- The allele frequency $P$ of each variant could differ in each ancestral population

The observed genome of an individual depends on:
(a) their ancestry mixture
(b) the genotype frequency in the corresponding populations

Challenging model to fit:
- ADMIXTURE (maximum likelihood)
- fastSTRUCTURE (variational)
- Terastructure (stochastic variational)
RFMix

- Idea: use **random forest classifiers** to predict the source population of each local segment
- Input: Observed allele frequencies in reference populations
- Output: which population did the observed segment come from
- Use a **conditional random field** to model the correlations between neighboring segments
- Intuition: switching ancestry every segment is unlikely (like HMM)

Maples et al. *Am J Hum Genet* 2013
Ancestry painting: population-level

Goal: infer ancestry of segments of the genome, population structure (patterns of relatedness between ancestry groups)

Sharing of genetic variants enables ancestry painting of individual genomes

The history of migration, settlement, conquest is written on our genomes

Thousand Genomes Consortium Nature 2016
Genetic relatedness and geography

- Can we decompose genetic variation into the major forces shaping it?
  - PCA/SVD decomposition
  - First components correspond to population structure.
  - Population structure is shaped by geography! (people near each other are more likely to mate)
  - In Europe, First two components correspond to N-S and E-W migration axes
  - Country neighbors & borders visible at the genetic level

Principal components analysis

- Idea: find a new basis for high dimensional data
- First principal component (axis) must explain most variation
- Second PC is orthogonal and explains maximum amount of the remaining variation
- PCs computed using singular value decomposition

Importance of PCA for GWAS

- Geography/demography causes genome-wide differences between individuals
- Confounds specific differences between individuals with different phenotypes
- Two main strategies to correct for population structure:
  - Include principal components as fixed covariates in linear model
  - Incorporate polygenic effect of population structure using linear mixed models

Measuring divergence between populations

- The **fixation index** $F$ measures the loss in heterozygosity in a population.
- Originally developed by Wright for studying livestock inbreeding.
- The special case $F_{ST}$ compares pairs within a subpopulation against pairs between subpopulations.

<table>
<thead>
<tr>
<th>Population</th>
<th>AA freq</th>
<th>AG freq</th>
<th>GG freq</th>
<th>Eff pop size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR</td>
<td>0.16</td>
<td>0.48</td>
<td>0.16</td>
<td>10000</td>
</tr>
<tr>
<td>EUR</td>
<td>0.01</td>
<td>0.18</td>
<td>0.81</td>
<td>4000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0.117</td>
<td>0.394</td>
<td>0.489</td>
<td>14000</td>
</tr>
</tbody>
</table>

$$F_{ST} = \frac{H_{total} - \bar{H}}{H_{total}} = \frac{0.394 - 1/2(0.48 + 0.18)}{0.394} = 0.16$$

Average individual $H$  
Joint $H = H_{total}$
Measuring divergence between populations

• \( F_{ST} \) equivalently compares the average coalescent time (time to most recent common ancestor) of pairs within a population to pairs between populations

• The underlying model also allows us to estimate effective population sizes from observed levels of heterozygosity

McVean *PLoS Genet* 2009
Evolutionary interpretation of PCA

- Under suitable assumptions, Euclidean distance in the first principal components corresponds to the mean coalescent time
- Predict exactly where individuals will appear in principal component space
- Illustrates a deep connection between population genetics and evolution

McVean PLoS Genet 2009
Today: Genome evolution, near and far

1. Foundations 1: Building the genome assembly
   – Consensus-layout-overlap methods
   – String graph methods

2. Foundations 2: Whole-genome alignment
   – Resolving region correspondence
   – Gene-based global genome alignment
   – Mechanisms of rapid genome evolution

3. Evolutionary “jumps”: Whole genome duplication (WGD)
   – Evolutionary signature of whole genome duplication
   – Post-duplication emergence of new functions

4. Recent evolution 1: Human selection at multiple timescales
   – Neutral evo, Hardy-Weinberg, Evidence of positive/directional selection
   – Negative/purifying selection: SNP density, Derived allele freq, Heterozy.

5. Recent evolution 2: Human demographic history.
   – Population size, dynamics, bottlenecks, expansion, effective pop size.
   – Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

6. Recent evolution 3: Human relatedness and ancestry painting
   – 1000Genomes population differences. SNP sharing. IBD/heritability.
   – Admixture, ancestry painting. PCA/MDS clustering ($F_{ST}$). PC correction