The painting palettes of human ancestry

Daniel Falush, Dan Lawson, Garrett Hellenthal, Simon Myers.

www.paintmychromosomes.com
We want to “paint” DNA according to which ancestor it comes from

(O’Connor 2008)
Recombination ➞ “mosaic” genomes

Positions of recombination crossover events
Recombination "mosaic" genomes
Recombination \[\rightarrow\] "mosaic" genomes
Ancient ancestors

We share segments of DNA today, inherited from shared ancestors living in the distant past.

Identifying these shared segments, and insights into the genetics of people from the UK, and worldwide human migrations
Chromosome painting in practice

- (Lawson et al. 2012)
Genealogical interpretation of painting

• The painting of each segment indicates who in the painting panel has the most recent shared common ancestor with the focal chromosome for that stretch of DNA.

• Boundaries between segments correspond to the positions of ancestral recombination events.

• The time at which the ancestor lived varies from segment to segment.

• Longer segments are associated with more recent shared ancestors.
FineSTRUCTURE

1. Paint each individual using every other individual in the sample as the painting panel and record the painting palettes.

2. MCMC algorithm is used to find clustering such that:
   (a) members of the same cluster paint with the same colour
   (b) individuals within clusters have similar palettes
   (c) palettes are enriched for their own colour (mostly).

   (Algorithm takes into account the fact that individuals are not used to paint themselves).
toy FineSTRUCURE example
Notes about FineSTRUCTURE

- MCMC algorithm includes both merges and splits.
- Good convergence properties in practice.
- Number and membership of clusters is inferred based only on DNA.
- We call the collection of palettes obtained in the all-versus-all painting the “coancestry matrix”.
- If the markers are treated as unlinked, the coancestry matrix is equivalent to the covariance matrix used by SMARTPCA.
- Likelihood approximately equivalent to that of STRUCTURE for weak genetic drift
Sampled individuals had all four grandparents living within 50 miles of each other.

Genotyped using SNP chip. (500,000 markers)

54 distinct palettes inferred by fineSTRUCTURE

Build a tree: by successively joining groups with the most similar palettes.

CREDIT: Bruce Winney, Stephen Leslie, Walter Bodmer, Peter Donnelly
Clusters very strongly resemble ancient geopolitical groupings.
Traditional software (Admixture)
Continental European palettes

Hospital based sampling as controls for an association study with FineSTRUCTURE used to identify populations
British palettes
painted with a continental European panel
Mixture modelling of an English palette based on European palettes using Non-Negative Least Squares
The map shows the distribution of Anglo-Saxon people in the United Kingdom, with a focus on the regions with the highest concentration of Anglo-Saxon heritage. The map is labeled "ANGLO-SAXON (~10–20%)". The data is presented with various colors and symbols indicating different genetic contributions from various European regions.

G.Hellenthal uUGIv
Inferring history using DNA
EMBO Hum Evol 4/2/14 36 / 40
NORWAY
VIKING
(~25%)
Danish Vikings colonisations from v. similar region to previous Anglo-Saxons
World-wide dataset application (HGDP+more)

\(~475K\) SNPs on 1490 individuals from 95 pops \((5-45\) inds/pop\)

CREDIT: Christian Capelli, George Busby
169 populations inferred in HGDP data
Half-matching based on coancestry matrix
Inferring past demography

admixture events

Mixing Events two populations q"red" and "yellow" intermix r generations ago q"admixture event" r genetic pieces from each population get smaller each subsequent generation due to recombination. can we take present day samples qgeneration r and infer:

evidence of admixture?

when groups mixed?

segment sizes exponentially distributed

Information on time since mixture given by spatial structure of ancestry along the chromosome.

\[ \begin{align*}
0 & \quad \text{\textbf{Red}} & X \quad \text{\textbf{Yellow}} \\
1 & \quad \text{\textbf{Red}} & \quad \text{\textbf{Yellow}} \\
2 & \quad \text{\textbf{Red}} & \quad \text{\textbf{Yellow}} \\
3 & \quad \text{\textbf{Red}} & \quad \text{\textbf{Yellow}} \\
\vdots & \quad \text{\textbf{Red}} & \quad \text{\textbf{Yellow}} \\
\text{r} & \quad \text{\textbf{Red}} & \quad \text{\textbf{Yellow}}
\end{align*} \]
Globetrotter

Simulated admixture
30 generations ago

Mixture modelling of genomic palettes

Fitting of spatial structure of variation in palettes along the chromosome
http://admixturemap.paintmychromosomes.com
(Hellenthal et al 2014)
Can also identify complex events such as multi-way admixture
can fit curves with sum of two exponential distributions
Conclusions

Can use chromosome painting to
• Distill ancestry information
• Cluster based on genetic similarity
• Visualise genetic drift
• Identify mixtures
• Date mixture events
• Reconstruct history
A set of haplotype sequences sorted in order of reversed prefixes at position k, showing the set of values at k isolated from those before and after, and on the right hand side how the order at position \((k + 1)\) is derived from that at k as in Algorithm 1.
### CPU Time (seconds)

**88K SNPs, 100 Inds**
- ChromoPainter: 3130
- FastIBD: 2686
- PBWT: 1

**500 Inds**
- ChromoPainter: 264000
- FastIBD: 169093
- PBWT: 19

### Simulated data (100 Inds)

- ChromoPainter
- PBWT
- FastIBD

- **Better**

- (Unlinked CP)
- (unlinked)
- ChromoPainter also the EigenSTRAT normalisation
- (IBS)
- Identity by State (unlinked)

### HGDP Data (938 Inds, 800K SNPs)

<table>
<thead>
<tr>
<th></th>
<th>ChromoPainter</th>
<th>FastIBD</th>
<th>PBWT</th>
<th>CP Unlinked</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0.35</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Daniel J. Lawson** (Heilbronn Institute, University of Bristol), **Daniel Falush** (Max Planck, Leipzig)

**Simon Myers** (Oxford), **Garrett Hellenthal** (UCL), **Richard Durbin** (Wellcome Trust Sanger Institute)