Exploring recent relatedness IBD and biparental ancestry

Peter Ralph and Graham Coop

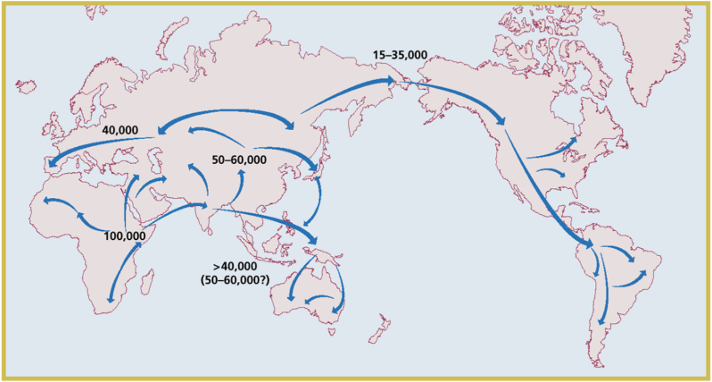
Department of Evolution and Ecology
UC Davis

June 11th, 2012, CIRM
Probability, Population Genetics and Evolution

Outline

- Introduction
- Probability: inheritance through the pedigree
- Data: common genetic ancestors
- When did these common ancestors live?
- Next steps

Population history and spatial dynamics

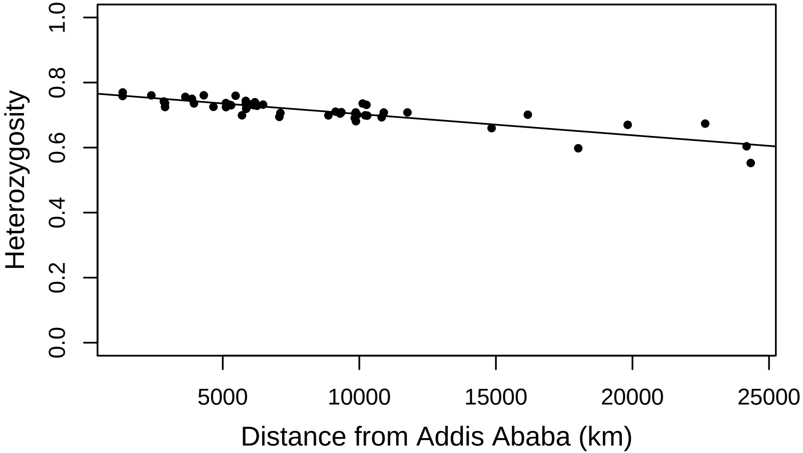


(from Feldman & Cavalli-Svorza)

Humans:

- range expansion(s)
- admixture
- adaptation locally and to local conditions

GENOMIC SIGNALS OF MIGRATION



(Ramachandran et al 2005)

History from genomes

Goal: infer recent migrations and population structure.

Method:

- Infer rates of shared ancestry
- by identifying close relatives (10th–100th cousins)
- How can we hope to do this?
  - Unlikely that any given pair are 10th cousins, but
  - many ways to be related, and
  - between thousands of samples there are millions of possibly related pairs.

History from genomes

Goal: infer recent migrations and population structure.

Method:

- Infer rates of shared ancestry
- by identifying close relatives (10^{th}–100^{th} cousins)
- How can we hope to do this?
  - Unlikely that any given pair are 10^{th} cousins, but
  - many ways to be related, and
  - between thousands of samples there are millions of possibly related pairs.

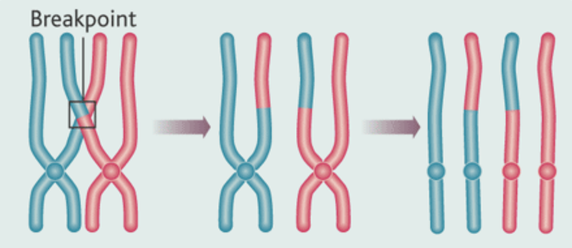
History from genomes

Goal: infer recent migrations and population structure.

Method:

- Infer rates of shared ancestry
- by identifying close relatives (10th–100th cousins)
- How can we hope to do this?
  - Unlikely that any given pair are 10th cousins, but
  - many ways to be related, and
  - between thousands of samples there are millions of possibly related pairs.

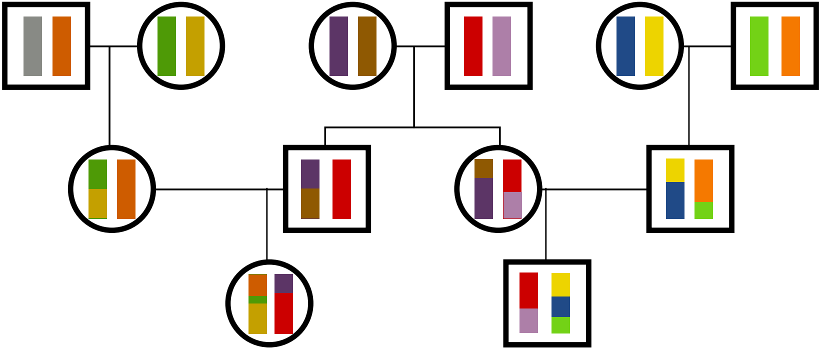
MEIOSIS & RECOMBINATION (SEX)



- You have two copies of each chromosome, one from each parent.
- When you make a gamete, the copies recombine.
- genetic distance: such that recombination rate is unity
- units of centiMorgans (cM) $\approx 10^6$ bp in humans

IBD: “Identity by Descent”

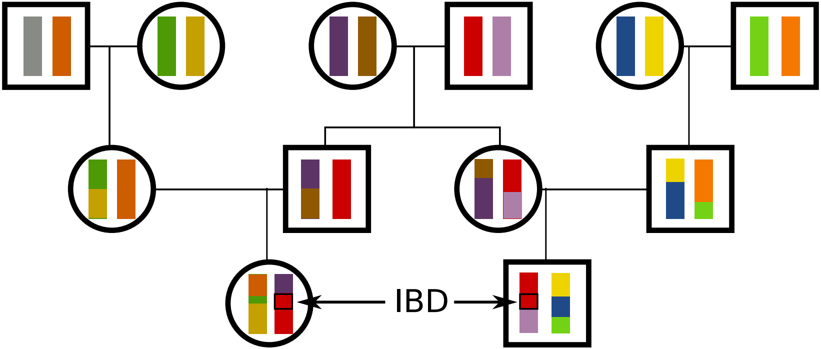
Definition: A block is IBD between two chromosomes if inherited from the same ancestor, without intervening recombinations.



Everyone is IBD everywhere, but the blocks are mostly short and old.

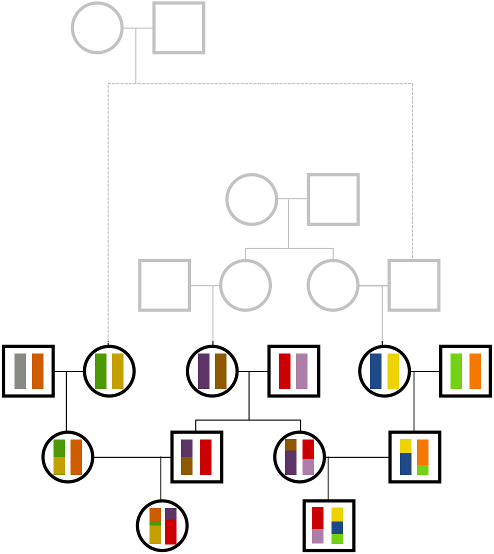
IBD: “Identity by Descent”

Definition: A block is IBD between two chromosomes if inherited from the same ancestor, without intervening recombinations.



Everyone is IBD everywhere, but the blocks are mostly short and old.

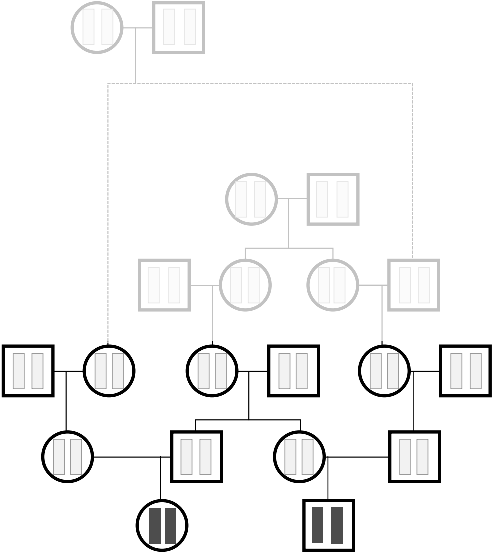
THE PEDIGREE AND IBD



Fragmentation-coalescence in the pedigree

- number of genealogical ancestors from n generations ago is 2^n
- number of genetic ancestors grows linearly
- since n meioses divides a 1M chromosome into $\sim n$ blocks

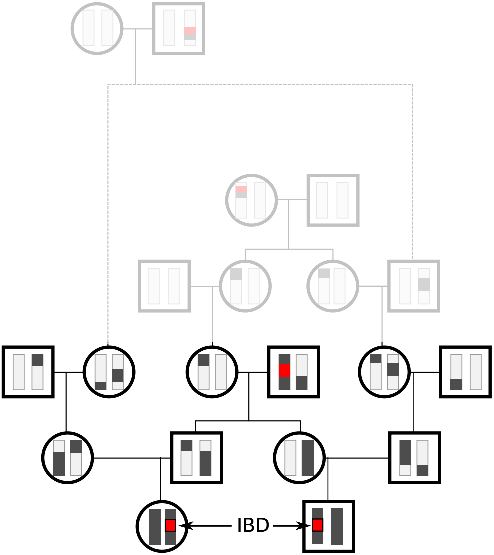
The pedigree and IBD



Fragmentation-coalescence in the pedigree

- number of genealogical ancestors from n generations ago is 2^n
- number of genetic ancestors grows linearly
- since n meioses divides a 1M chromosome into $\sim n$ blocks

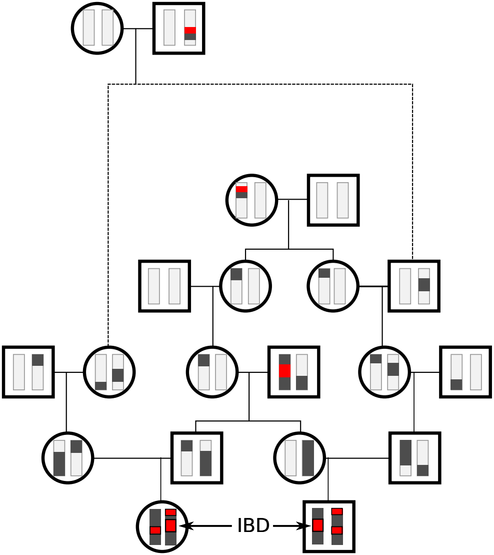
The pedigree and IBD



Fragmentation-coalescence in the pedigree

- number of genealogical ancestors from n generations ago is 2^n
- number of genetic ancestors grows linearly
- since n meioses divides a 1M chromosome into $\sim n$ blocks

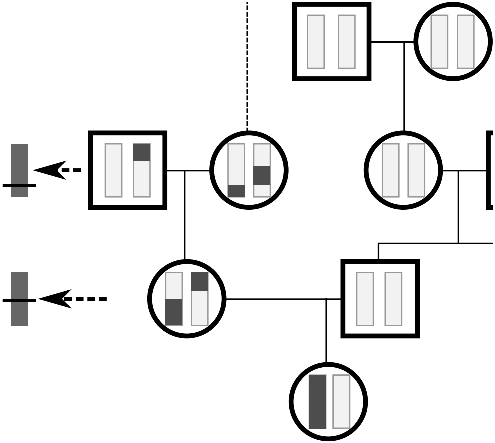
The pedigree and IBD



Fragmentation-coalescence in the pedigree

- number of genealogical ancestors from n generations ago is 2^n
- number of genetic ancestors grows linearly
- since n meioses divides a 1M chromosome into $\sim n$ blocks

The pedigree and IBD

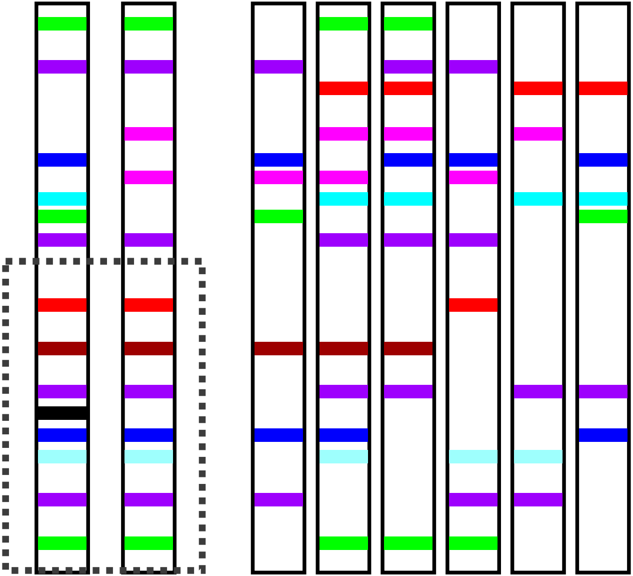


Fragmentation-coalescence in the pedigree

- number of genealogical ancestors from n generations ago is 2^n
- number of genetic ancestors grows linearly
- since n meioses divides a 1M chromosome into $\sim n$ blocks

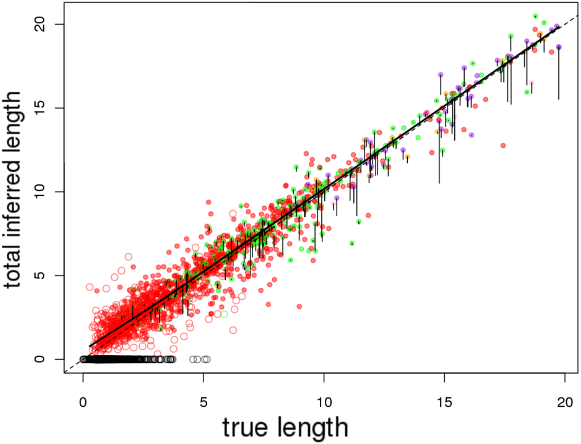
What we actually observe: mutations

Infer IBD from long regions of agreement (relative to everyone else).

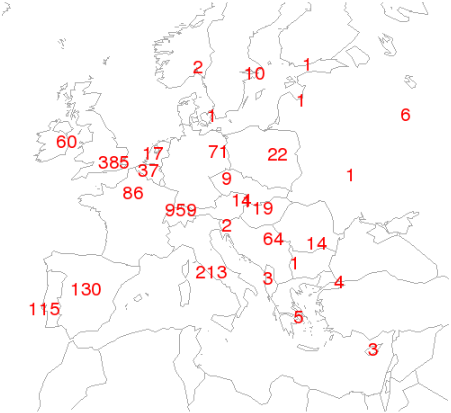


Black box IBD finding

- `fastIBD` in `BEAGLE` (Browning & Browning)
- Fits a variable length Markov chain to phase data and infer IBD blocks.
- Power analysis
- False positive rate



ON TO SOME DATA



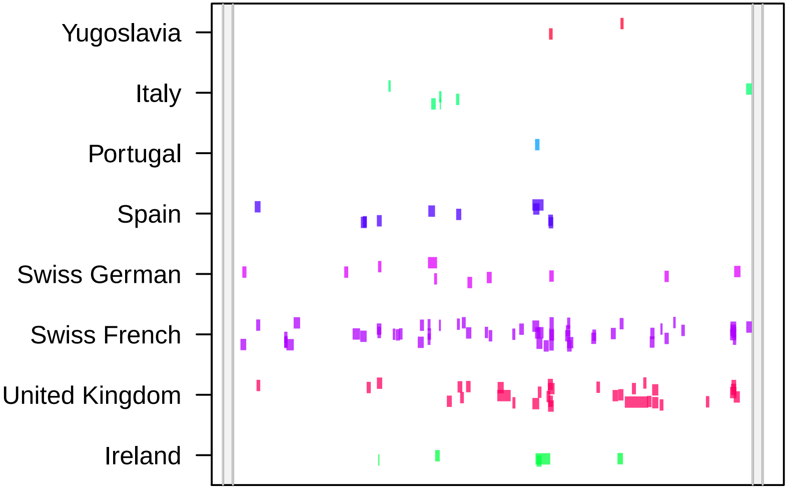
Data from POPRES:

(Nelson et al 2008)

- 2257 Europeans after removing outliers and close relatives
- have country and language data: 40 populations
- ∼500,000 SNPs

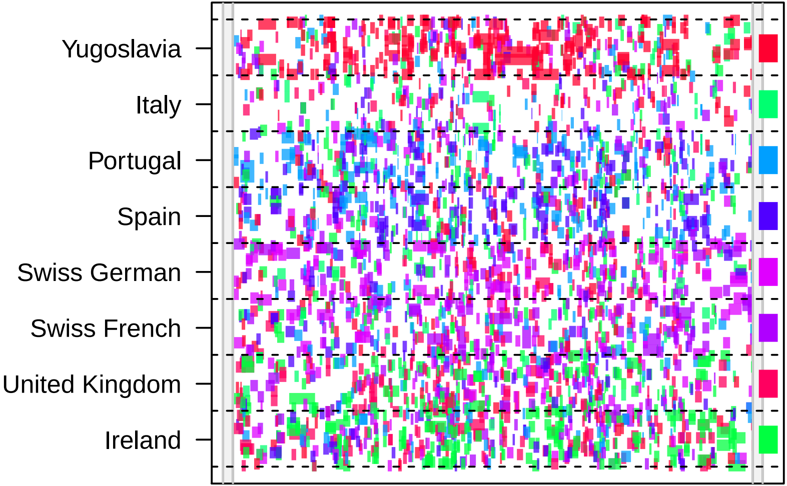
Plenty of IBD blocks

- 1877114 blocks
- 831 blocks per indiv, 0.737 per pair
- covering 30–250% of each individual

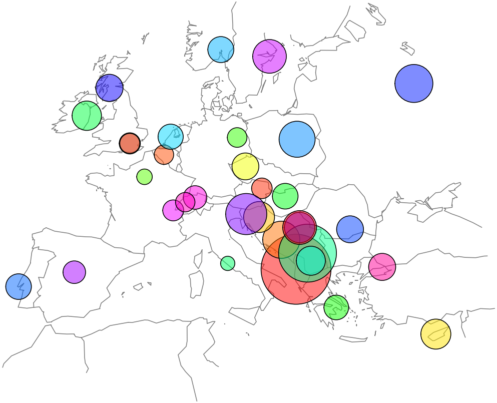


Plenty of IBD blocks

- 1877114 blocks
- 831 blocks per indiv, 0.737 per pair
- covering 30–250% of each individual

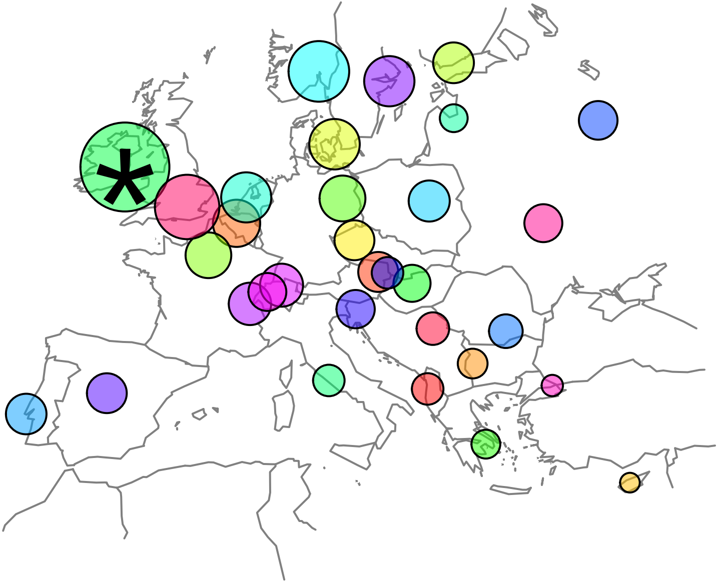


MEAN WITHIN-COUNTRY IBD RATES

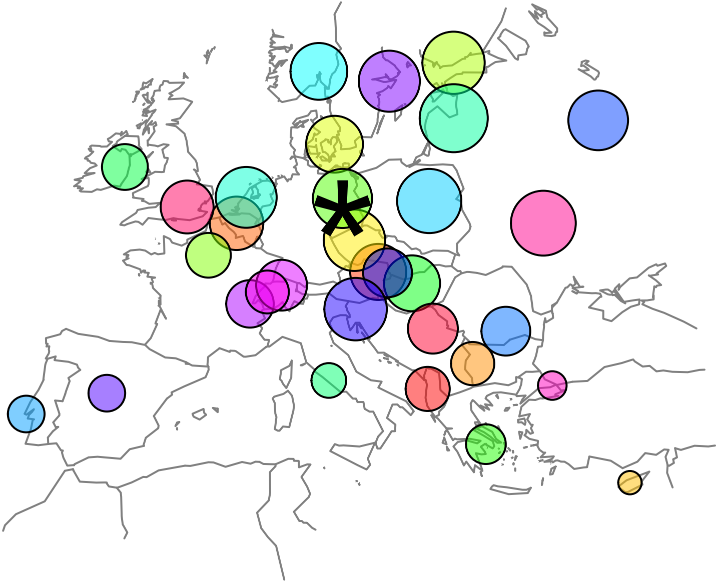


Mean # blocks $> 1\,cM$	
Italy	0.44
France	0.62
Belgium	0.97
Swiss.German	1.32
Swiss.French	1.01
Germany	1.01
Spain	1.14
Portugal	1.40
United.Kingdom	1.04
Ireland	2.15
Poland	3.40
Yugoslavia	3.59

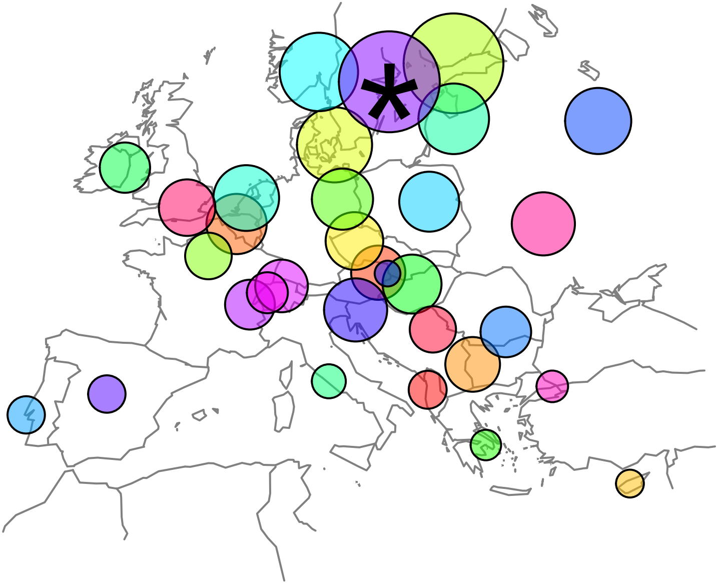
Ireland



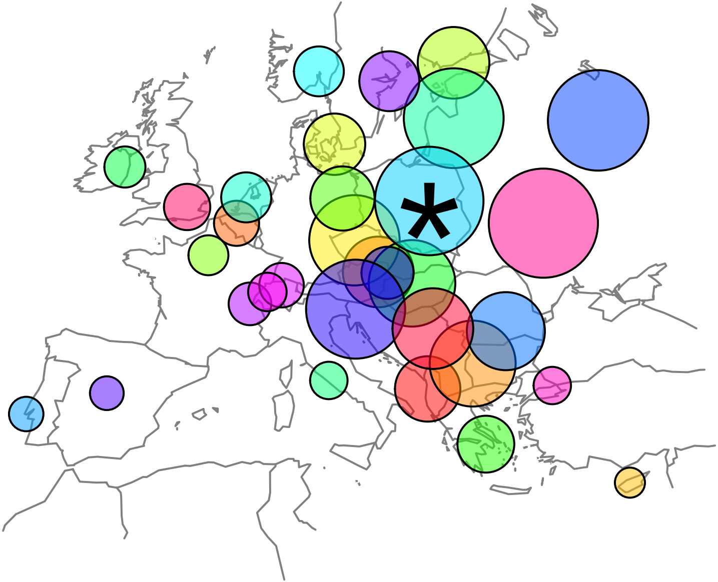
Germany



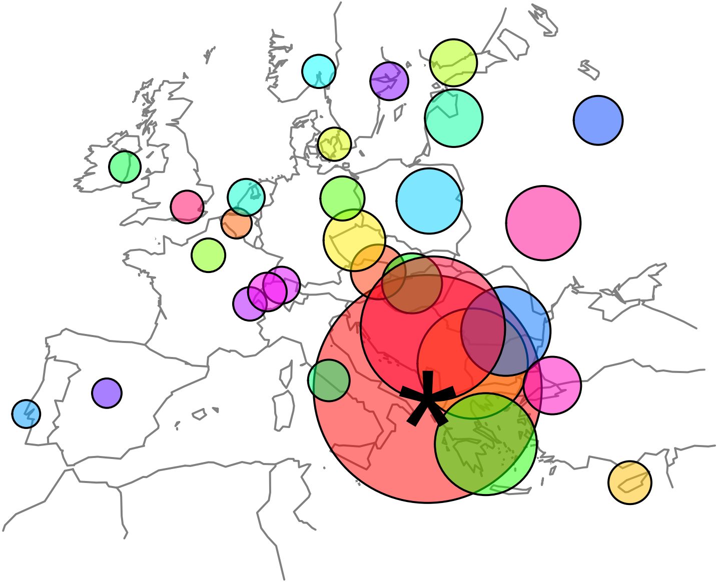
Sweden



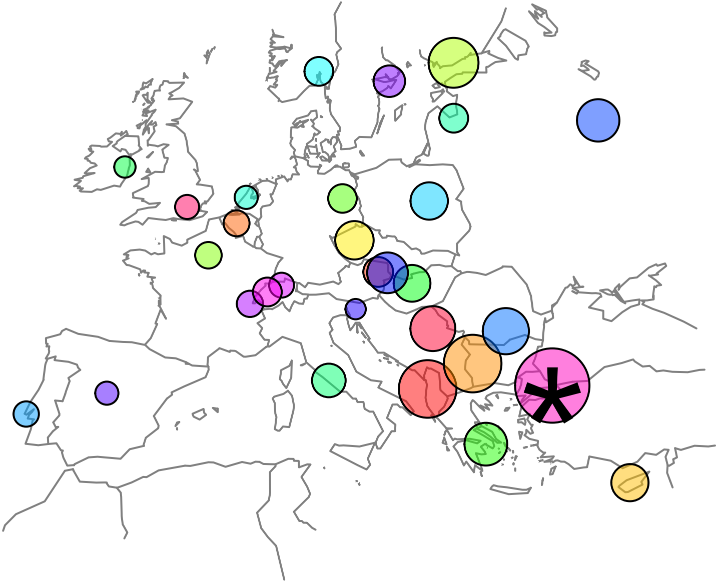
Poland



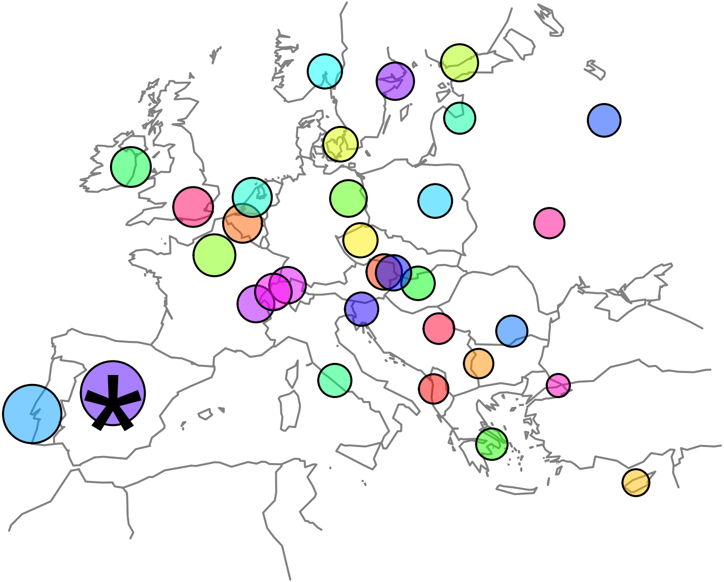
Albania



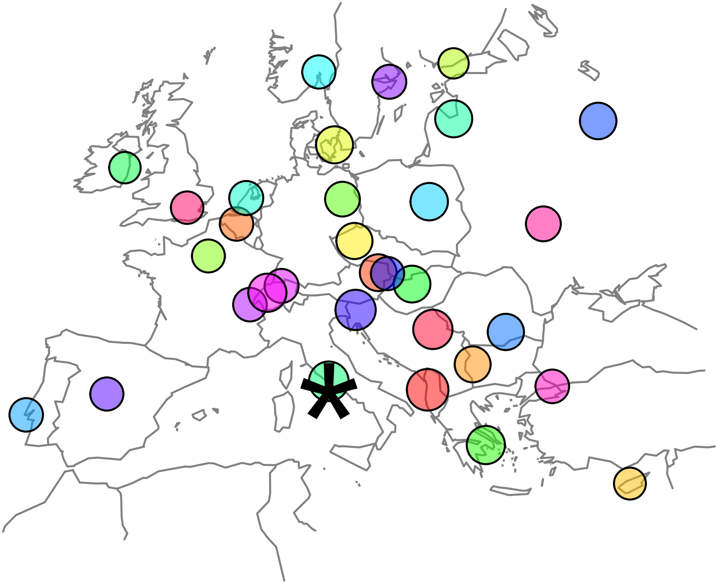
Turkey



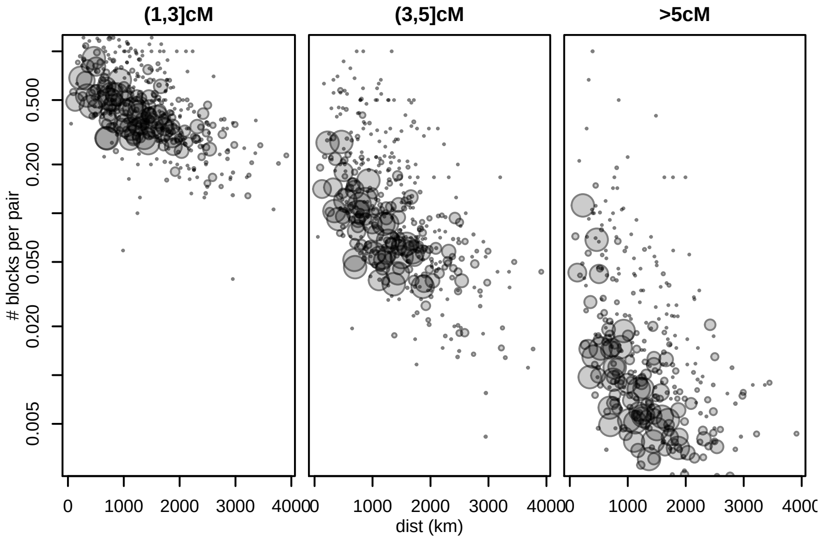
Spain



Italy

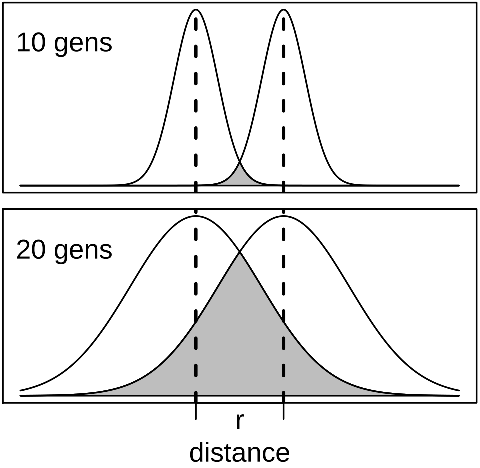


Decay of IBD with distance



Decay with distance is steeper for longer (older) blocks.
(circle size is sample size)

Common ancestors across geography



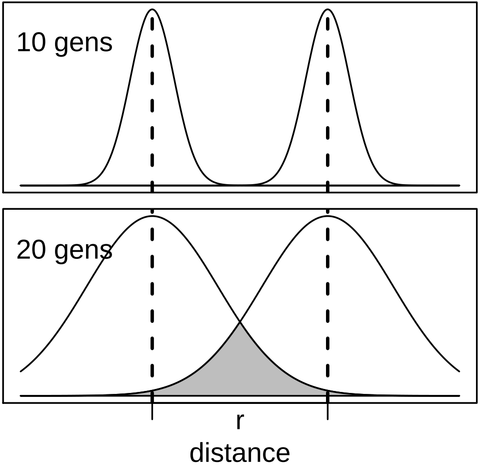
Geographic distribution of n^{th} generation co-located ancestors:

more recent
⇒ more localized

Gaussian distribution ⇒ coancestry at distance r

$$\propto \frac{C}{n} \exp\left(-\frac{r^2}{n\sigma^2}\right)$$

Common ancestors across geography



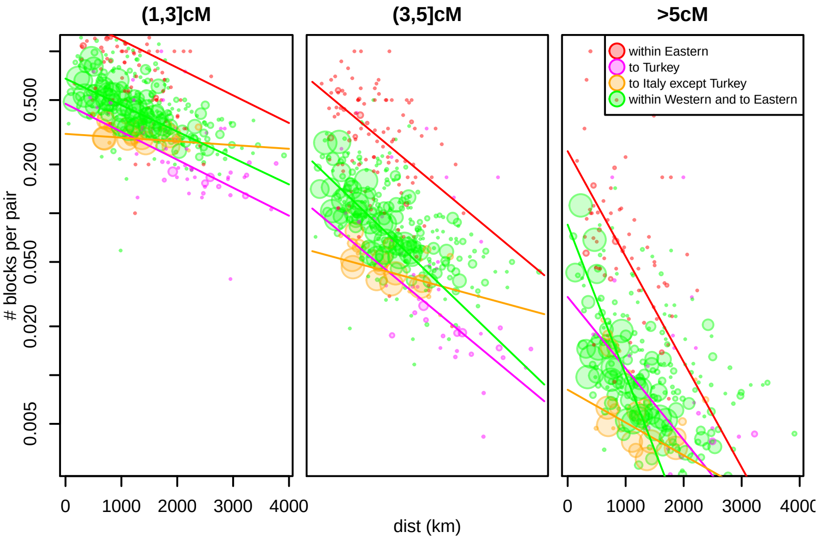
Geographic distribution of n^{th} generation co-located ancestors:

more recent
⇒ more localized

Gaussian distribution ⇒ coancestry at distance r

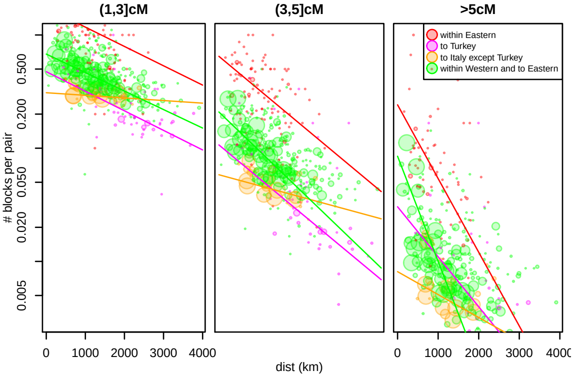
$$\propto \frac{C}{n} \exp\left(-\frac{r^2}{n\sigma^2}\right)$$

DECAY OF IBD WITH DISTANCE



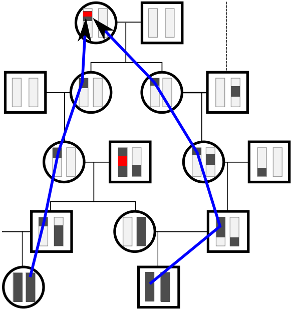
. . . and differs by location.
Italians have much slower decay (older?).

When did these common ancestors live?



Back to theory: expected IBD

- Fix pedigree (not recombinations), and two (sampled) chromosomes of length G.
- $N(x) = \#\{$ IBD blocks of length $\geq x\}$
- Decompose $N(x) = \sum_{T} N_T(x)$ by paths T through the pedigree
- $\mathbb{E}[N_T(x)] = K(|T|, x)\, 4^{-|T|}$, where
- $|T| = \#$ of meioses along T,
- $0 = R_0 \leq R_1 \leq \ldots R_k = G$ locations of recombinations
- $K(t, x) = \mathbb{E}[\#\{j : \ R_j - R_{j-1} > x\}]$.



Expected number of long IBD blocks

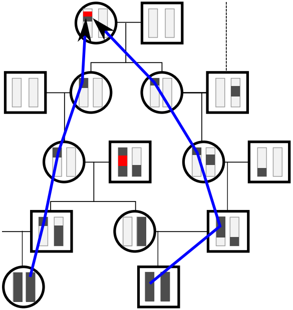
Coalescent distribution: choose path T with probability $4^{-|T|}$. Let $|T| = \tau$.

Expected block length distribution is a linear function of coalescent distribution:

$$\mathbb{E}[N(x)] = \sum_t \mathbb{P}\{\tau = t\} K(t, x)$$

If recombinations are Poisson,

$$K(t, x) = (1 + t(G - x))e^{-tx}$$



Mean block rate and coalescent distribution

Actually: mean IBD length distribution is a linear function of the coalescent distribution, so with

$$\mu(t) := \mathbb{P}\{\tau = t\},$$

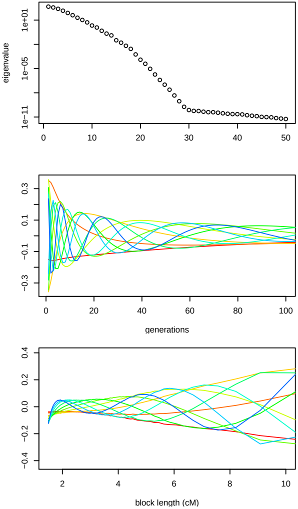
$N(x)$ is Poisson with mean density

$$\begin{aligned} -\frac{d}{dx}\mathbb{E}[N(x)] &= \sum_t \mu(t) K(t,y)\gamma(y) \int_0^G f(y,x)dy + \xi(x) \\ &= \sum_t \mu(t)\tilde{K}(t,x) + \xi(x) \end{aligned}$$

with: power γ, false positive rate ξ, and error kernel f.

. . . maximum likelihood?

Exploring the likelihood ridge



(Poisson) log likelihood function

$$\mathcal{L}(N|\mu)$$

is very flat in many directions (“ridged”).

We’re doing inference, so: need to explore it.

We do this by finding maximizers to

$$\mathcal{L}(N|\mu) - \gamma(\mu)$$

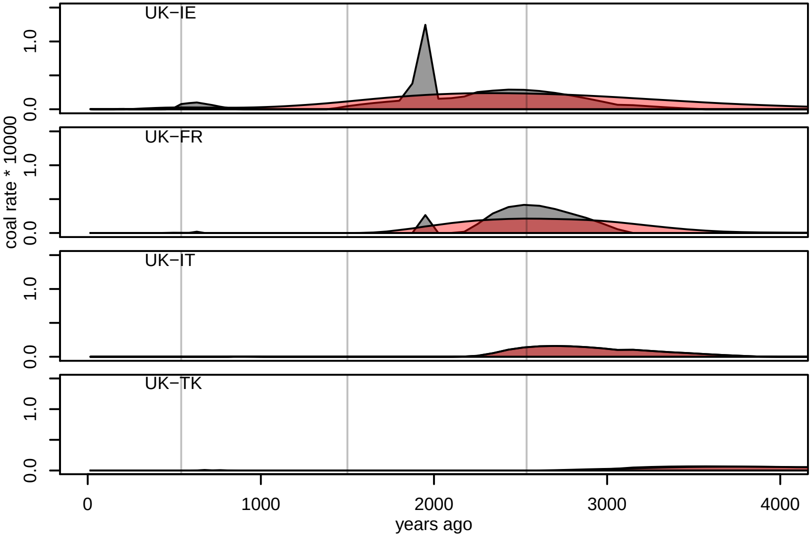
for various penalizations γ.

Call μ' *feasible* if $\mathcal{L}(N|\mu') + 2 \geq \max_\nu \mathcal{L}(N|\nu)$.

Coalescent distribution within the UK:

Grey is “best” solution; red is “smoothest” solution

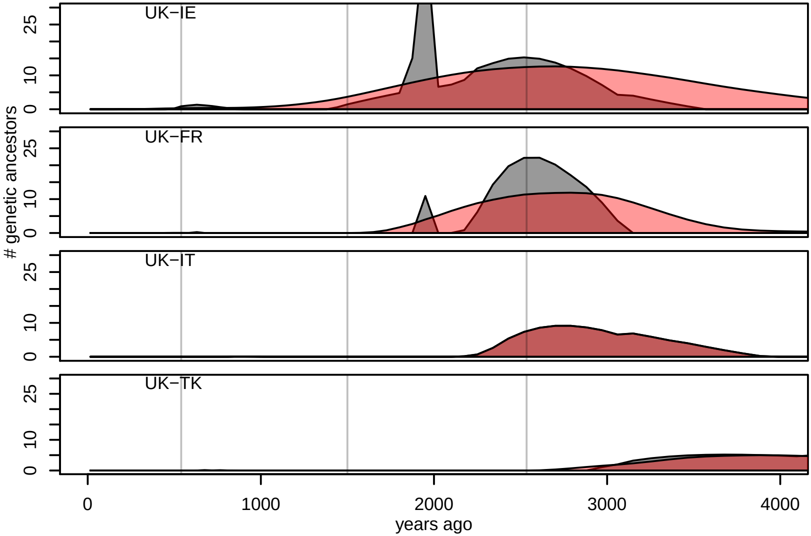
(differing by no more than 2 units of log likelihood).



Numbers of genetic common ancestors:

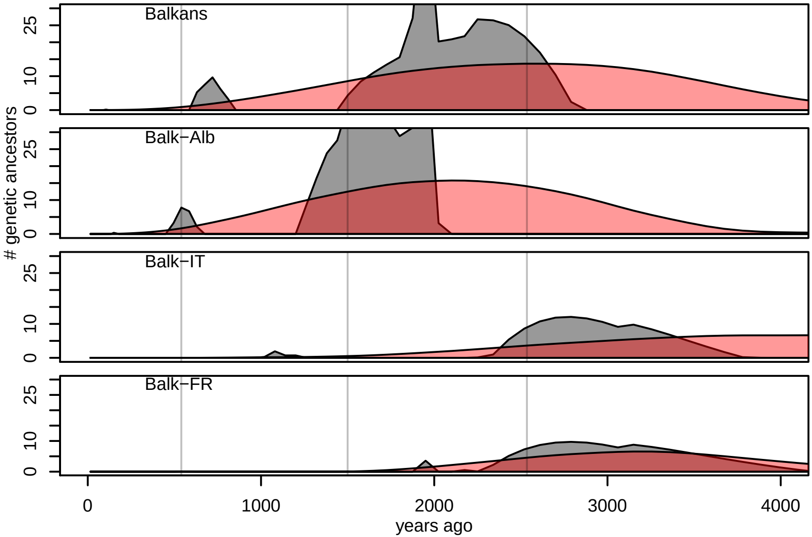
Grey is “best” solution; red is “smoothest” solution

(differing by no more than 2 units of log likelihood).



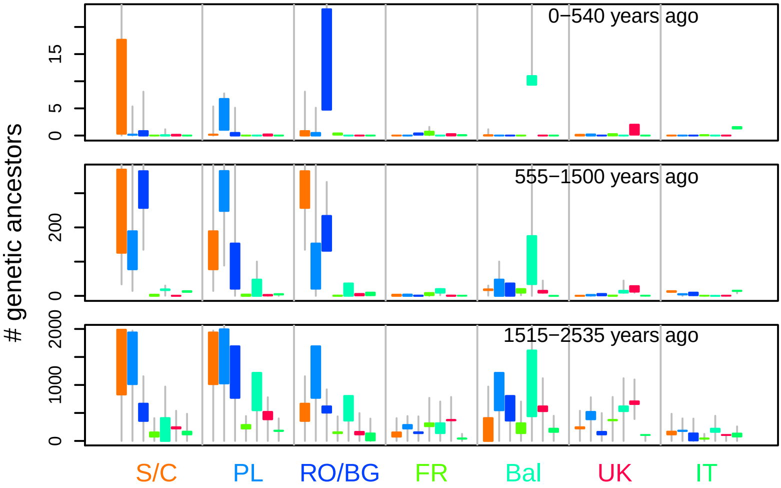
Common ancestry with the Balkans:

("Balkans" is peninsula except Albanian speakers.)



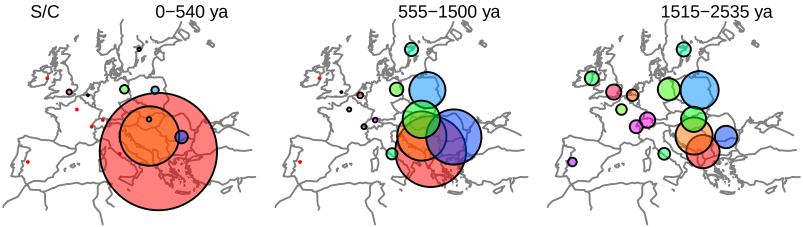
SUMMARIES OF COMMON ANCESTRY

Box: “best” & “smoothest”; whiskers: most & least

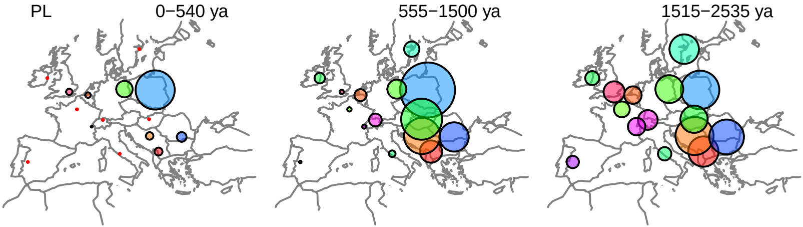


From whence eastern IBD?

Numbers of common ancestors shared with:
Serbo-Croatian speakers

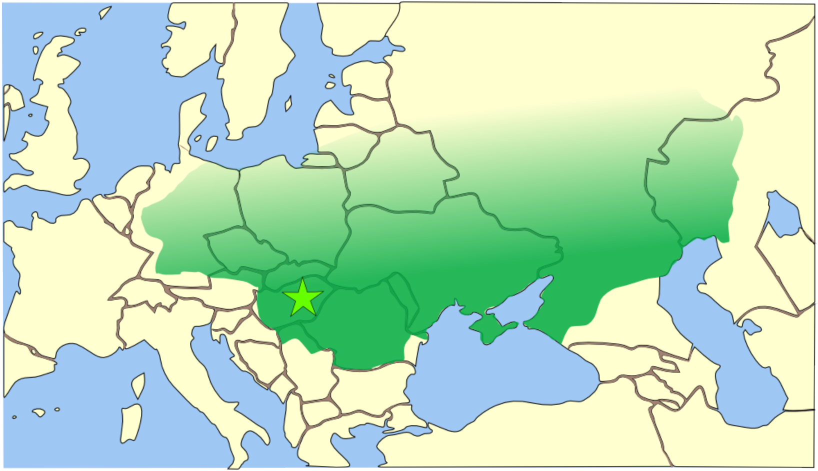


Poland



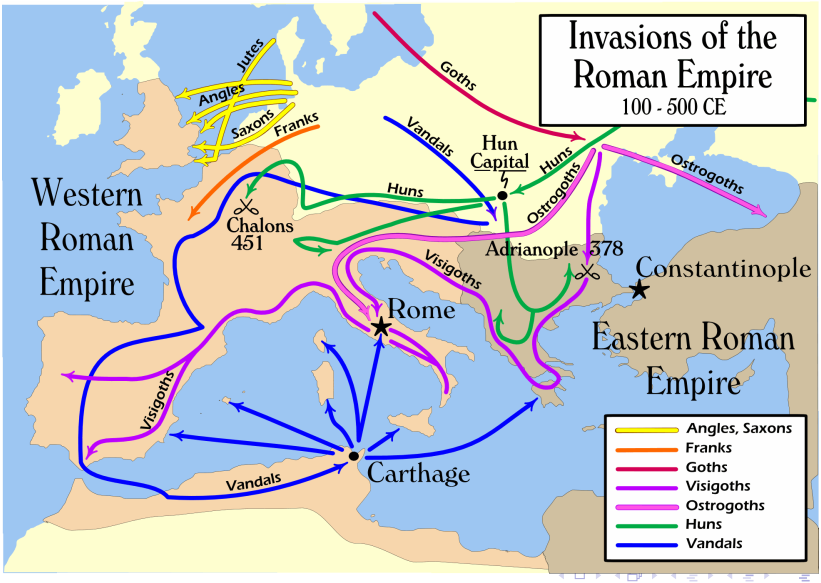
Signs of past invasions?

Hunnic empire, 1550ya:

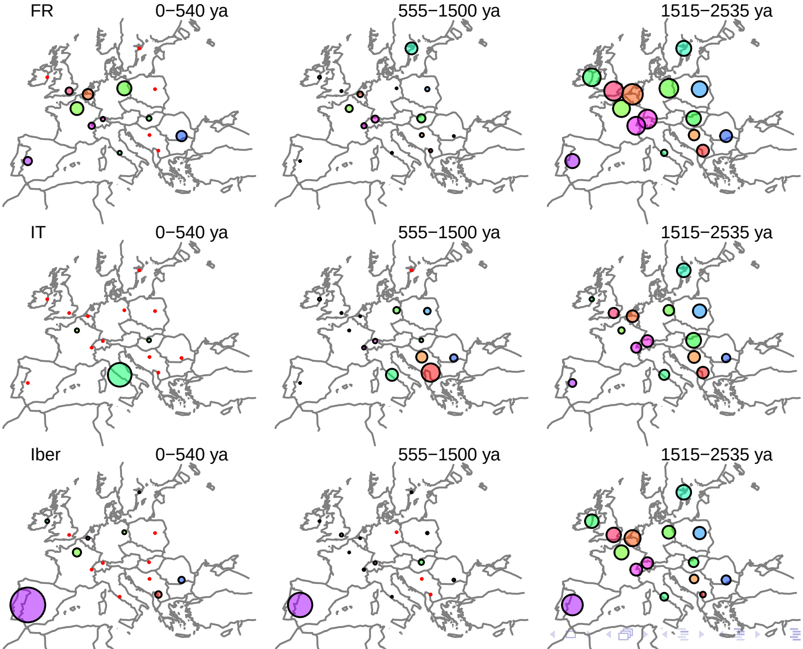


credit: wikipedia

What about the Germanic movements?



What about the Germanic movements?

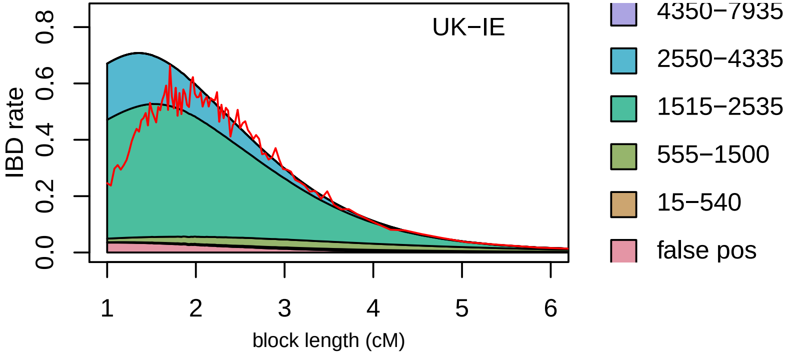


Further directions

- Further work on fragmentation-coalescence (in a pedigree?)
- Look at the process along the genome.
- Lack of fit at short lengths: improve the model.
- Geographic method of coalescent distribution inference – more than pairwise?
- What does the coalescent distribution tell us, anyhow?

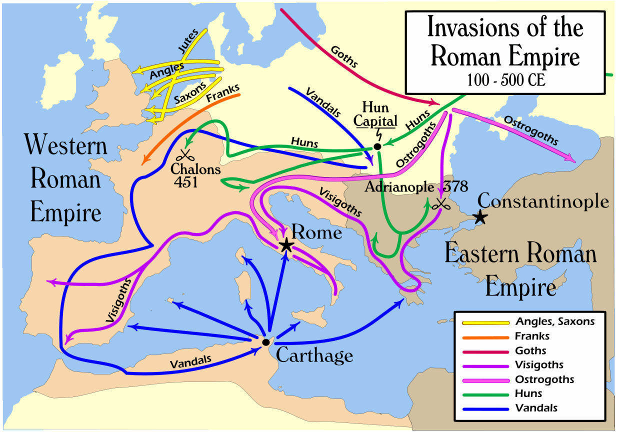
A problem with small blocks

Some assumptions of the model break down at short lengths:



(we used blocks $> 2cM$ for timing inference)

GEOGRAPHY: MORE THAN PAIRWISE



We find coalescent distribution: pairwise, nonparametric.
Could fit more than pairwise in parametric model. Other ideas?

COALESCENT TIME DISTRIBUTIONS?

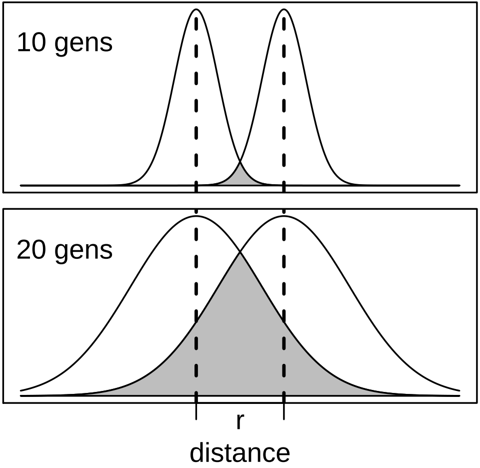
We can* infer distribution of coalescent time in the pedigree, across samples.

What does this tell us about shared history?

Intuition is mostly from:

- n-deme migration (usually, $n = 2$; rigorous)
- cartoons (not so much)

* maybe



Summary

Patterns of recent relatedness between Europeans are shaped primarily by:

- continuous, local gene flow – isolation by distance
- large population expansions – Slavs, Huns?
- other historical factors – language, stability, . . .

Summaries of long shared tracts of genome from recent common ancestors:

- have lots of signal about recent history in modern datasets
- can be used to infer statistical properties of the recent pedigree
- but may have fundamental drawbacks

Thanks

Graham Coop

Steve Evans, Charles Langley, Yaniv Brandvain, Torsten Günther

Funding: NIH