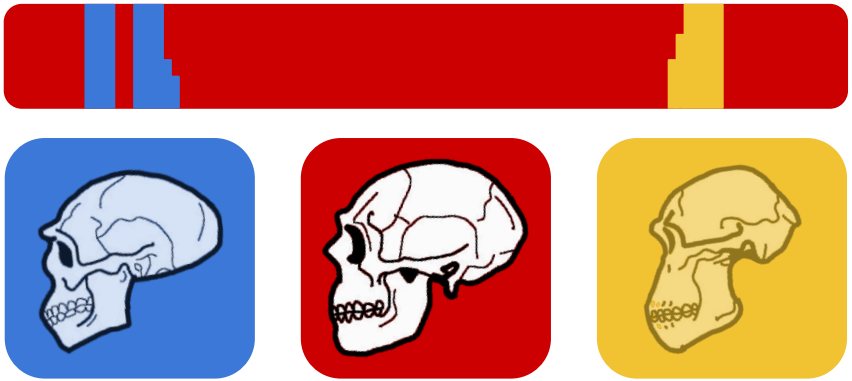
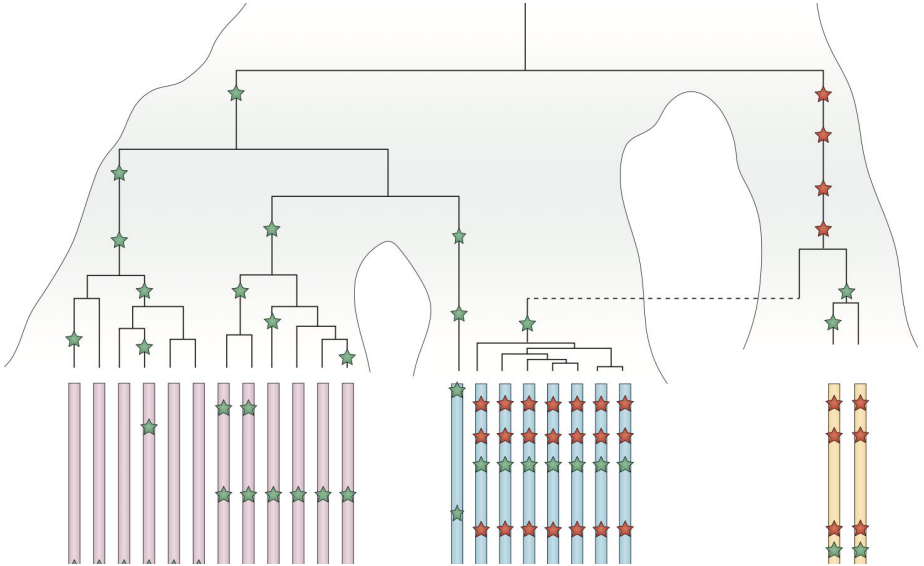
A Local Ancestry Inference Hidden Markov Model Optimized with Baum-Welch Expectation-Maximization

Motivation:

Inferring Local Ancestry with Hidden Markov Models to understand archaic introgression





Existing Literature

Prüfer (2013) *The complete genome sequence of a Neanderthal from the Altai Mountains*	• Early analysis of high-quality archaic hominin • Evidence of introgression among hominins • Simple HMM with fixed parameters
Racimo (2015) *Evidence for archaic adaptive introgression in humans*	• Overviews statistical methods for LAI • Focuses on adaptive introgression • Compares HMMs to other methods
Rabiner (1998) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*	• Influential early paper in a separate field • Allowed many HMM-specific methods to become accessible • Describes Expectation-Maximization Algorithm

Goals and Questions

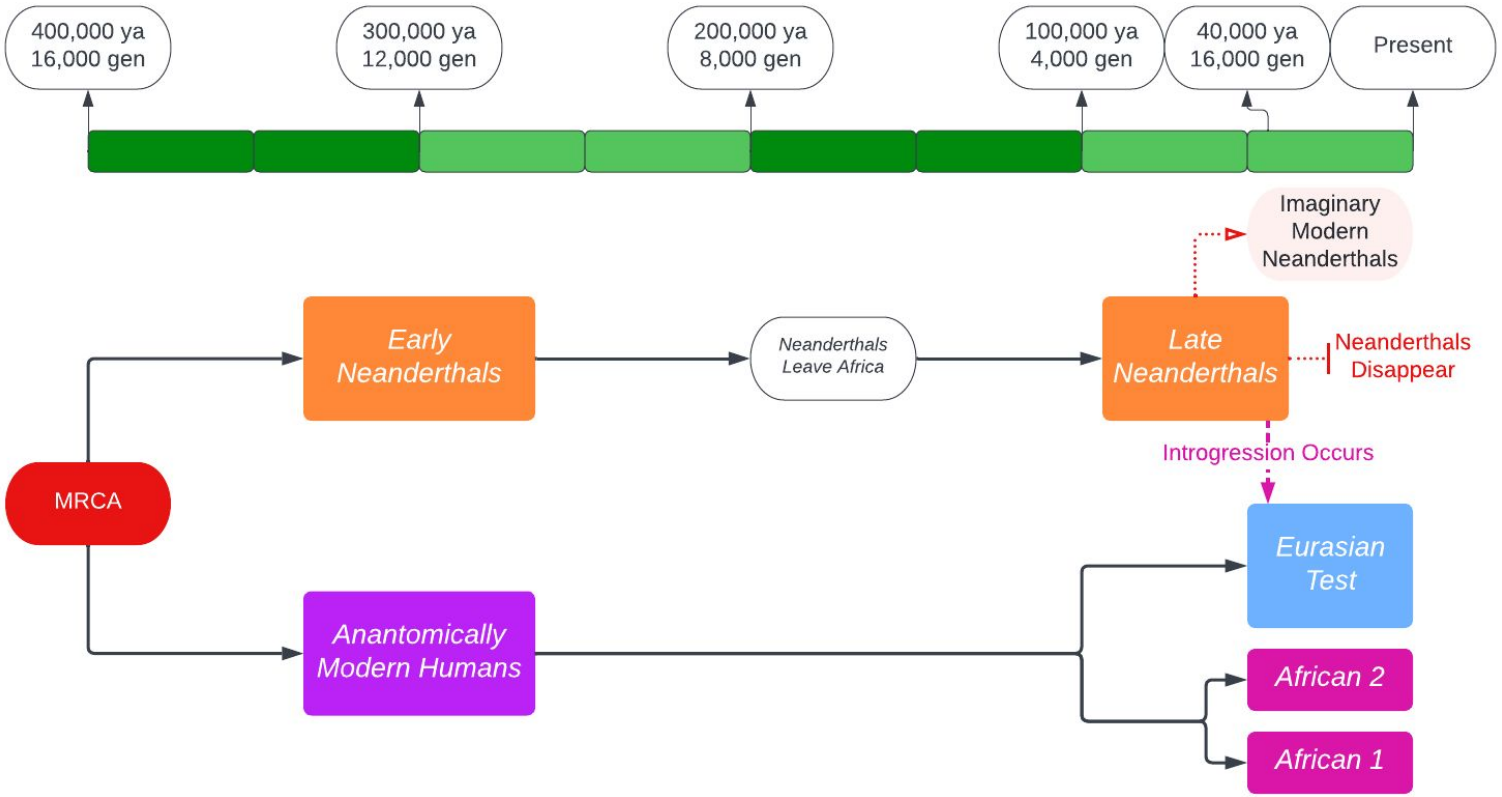
Goals:

- *Estimate the likelihood of Neanderthal introgression at each locus in a modern human genome*
- Re-implement Prüfer's HMM
- Evaluate it on simulated data
- Evaluate HMM LAI performance when using Baum-Welch Expectation-Maximization

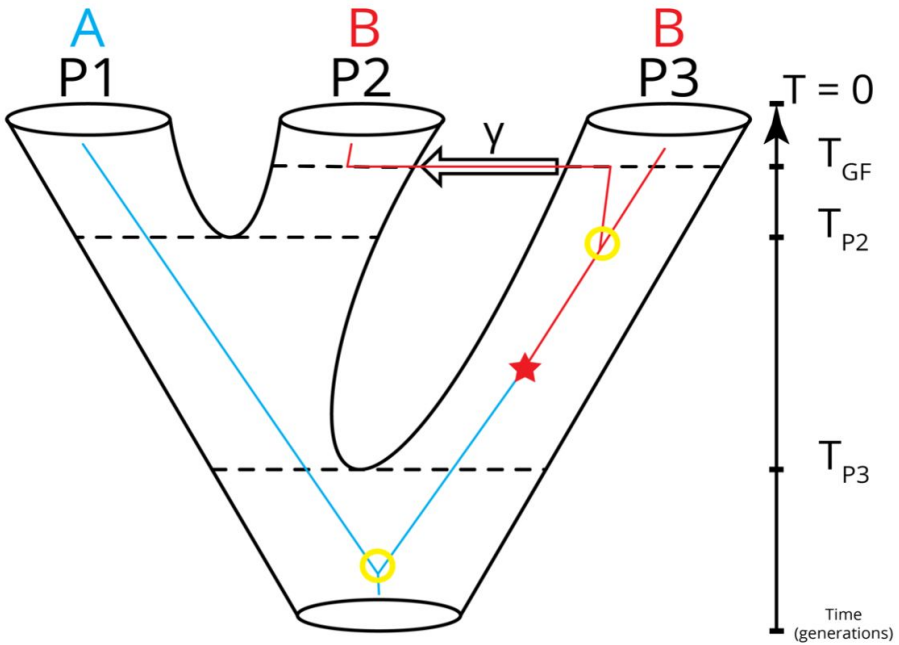
Questions:

- *How accurately does Prüfer's model perform on simulated data?*
- *How does Baum-Welch change Prüfer's assumed parameters?*
- *Can incorporating an expanded definition of introgressed site patterns improve LAI performance?*

Methods: Simulating Demographic Model



Methods: Data Generation



Topology resulting from an introgression event

P1	P1	P2	P3
AFR1	AFR2	TEST	NEAN
G	G	A	G
C	C	C	T
T	T	A	A

AFR1	AFR2	TEST	NEAN
0	0	1	0
1	1	1	0
0	**0**	**1**	**1**

Polarizing a Genotype Matrix

Methods: Site Pattern Types

PATTERN	AFR1	AFR2	EUR	NEAN
A	0	0	1	1
B	1	1	0	0
C (either)	0	0	1	1
	1	1	0	0

Methods: Genetic Distances (dxy)

AFR1	AFR2	EUR	NEAN
0	0	1	0
1	1	1	0
0	0	1	1
0	0	1	1
0	0	1	1

$$\frac{\sum \text{POP}(1\text{-NEAN}) + \text{NEAN}(1\text{-POP})}{\text{\# of variant sites}}$$

AFR/NEAN dxy

AFR1	AFR2	EUR	NEAN
0	0	1	0
1	1	1	0
0	0	1	1
0	0	1	1
0	0	1	1

$$\frac{\sum \text{AFR(1-NEAN) + NEAN(1-AFR)}}{\text{\# of variant sites}}$$

Equation	Distance
0(1-0)+0(1-0)	0
1(1-0)+0(1-1)	1
0(1-1)+1(1-0)	1
0(1-1)+1(1-0)	1
0(1-1)+1(1-0)	1
(0+1+1+1+1)/5	**.8**

EUR/NEAN dxy

AFR1	AFR2	EUR	NEAN
0	0	1	0
1	1	1	0
0	0	1	1
0	0	1	1
0	0	1	1

$$\frac{\sum \text{EUR(1-NEAN) + NEAN(1-EUR)}}{\text{\# of variant sites}}$$

Equation	Distance
1(1-0)+0(1-1)	1
1(1-0)+0(1-1)	1
1(1-1)+1(1-1)	0
1(1-1)+1(1-1)	0
1(1-1)+1(1-1)	0
(1+1+0+0+0)/5	**.4**

Methods: "Binning" Site Patterns into Windows

Length of simulated genotype = **20 megabases**

Window size: **0.0005 centimorgans**

Assuming 1cM = 1Mb, **window size is 500bp**

I therefore have **40,000 observations**

AFR1	AFR2	TEST	NEAN	N/C?
1	1	1	0	N
0	**0**	**1**	**1**	**C**

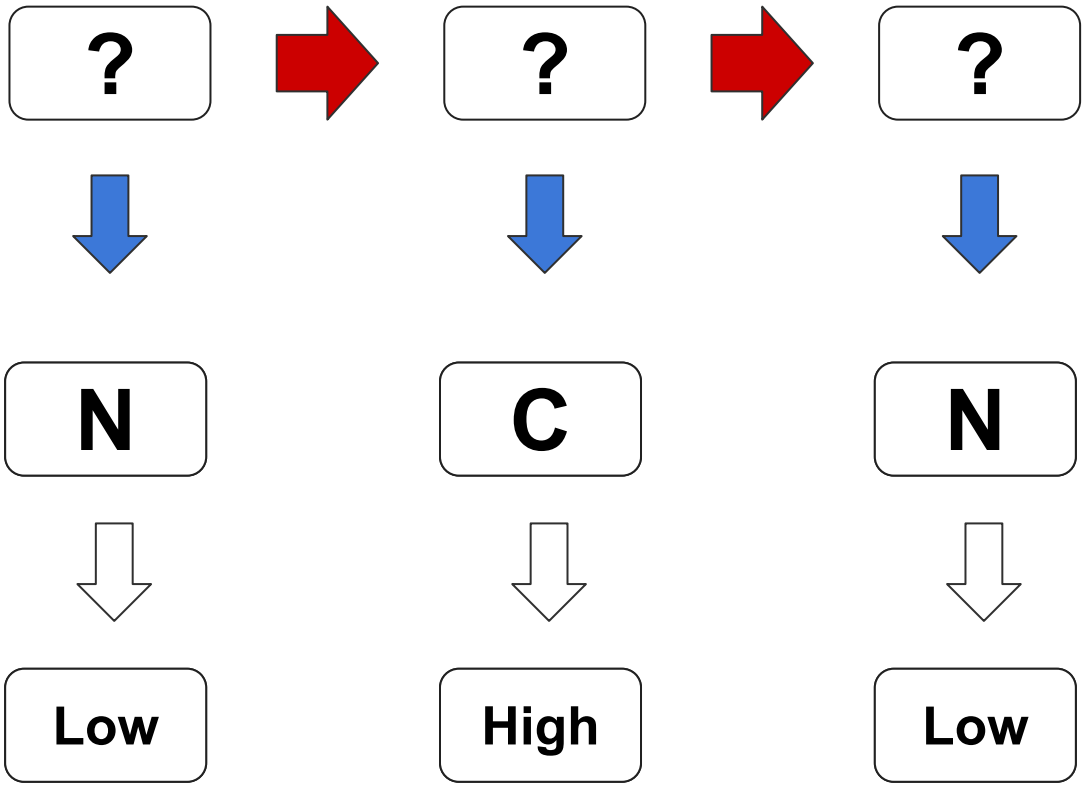
Window Range	[0-500) bp	[500, 1000) bp	[1000, 1500) bp
Variant Classes	N, N, N, … N	C, N, C, … , C	C, C, C, … C
Window Label	All N's ⇒ No Evidence, label N	One or more N's ⇒ Uninformative, label N	All C's ⇒ Strong Evidence, label C
Observed Sequence	**N**	**N**	**C**

Methods: Hidden States and Observations

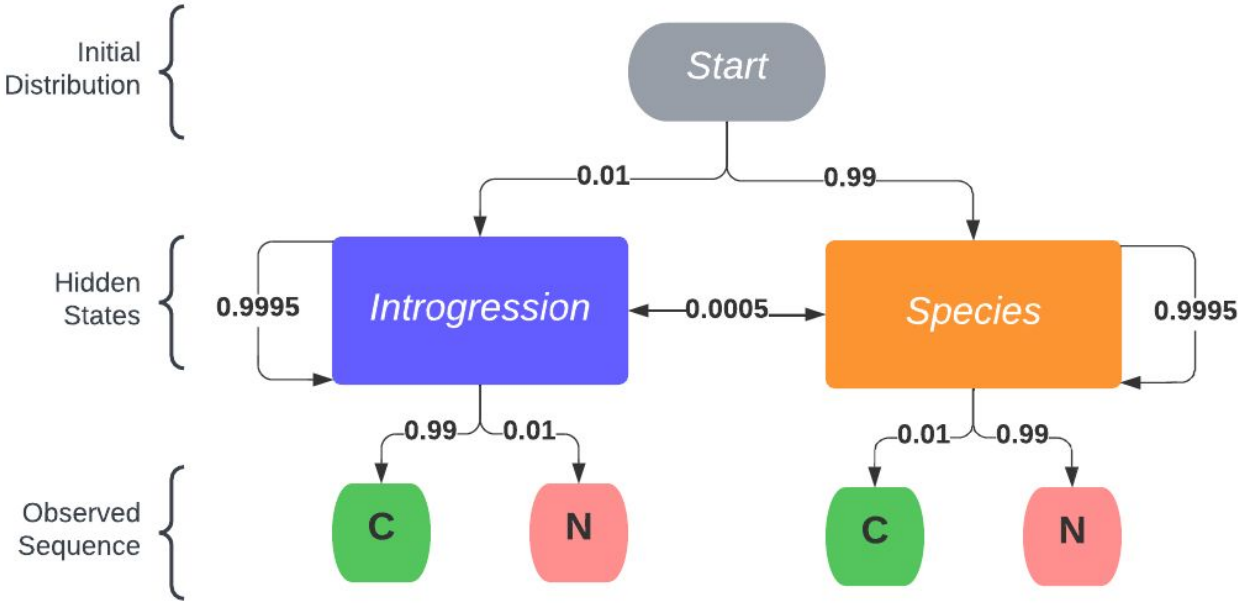
Hidden States, or unknown ancestries

Observed sequence, based on biallelic SNPs

Result: inferred likelihood of introgressed hidden state



Methods: Prüfer's Simplified HMM



Naive HMM λ

π	S	I
	.99	.01

A	S	I
S	.9995	.0005
I	.0005	.9995

B	S	I
N	.99	.01
C	.01	.99

Methods: Baum-Welch Updates Naive Assumptions

Naive HMM λ

π	S	I
	.75	.25

A	S	I
S	.75	.25
I	.25	.75

B	S	I
N	.9	.1
C	.1	.9

Naive γ matrix

Introgression chance	
N	2.1%
N	9.5%
C	85.8%
C	86.9%
N	22.2%

Converged in 6 steps

Baum-Welch λ

π	S	I
	.99	.01

A	S	I
S	.43	.57
I	.30	.69

B	S	I
N	.99	.01
C	.23	.77

New γ matrix

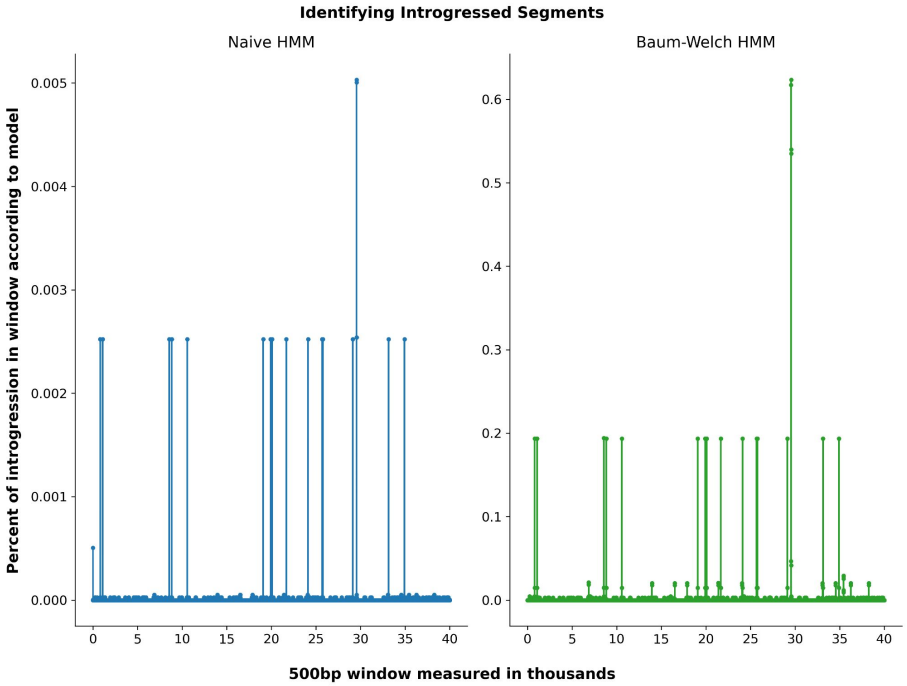
Introgression chance	
N	**.0001%**
N	18.2%
C	**98.1%**
C	**98.1%**
N	21.8%

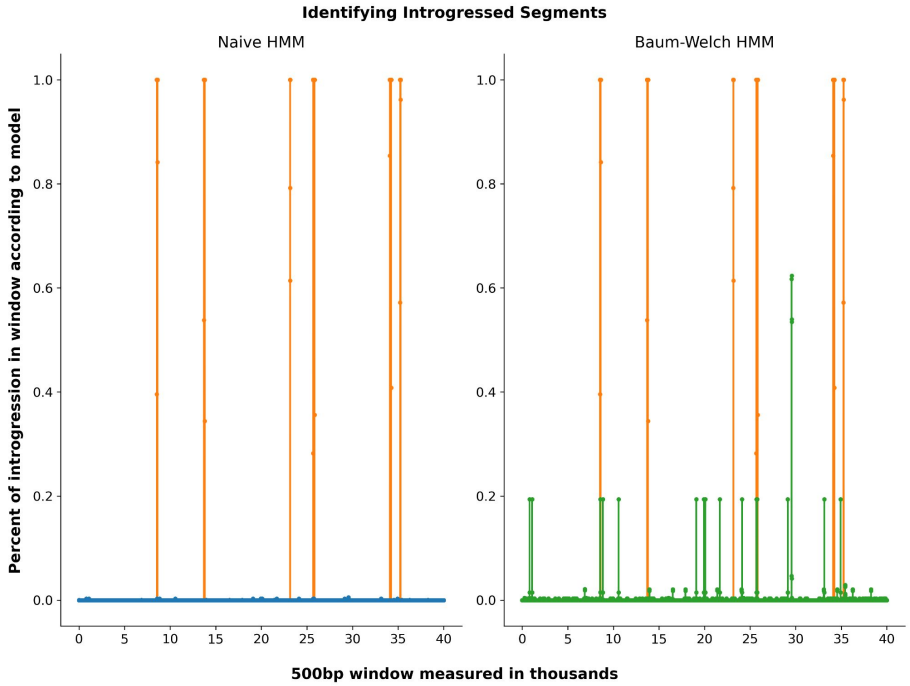
Methods: Workflow Summary

1. Demographic model with standard introgression assumptions
2. Extract observed genotypes and introgressed tracts from simulated individuals
3. Data Preprocessing: polarize the genotypes, bin into windows
4. Compute per-window introgression likelihood using HMM
5. Use Baum-Welch to optimize parameters
6. Evaluate accuracy
7. Compare model types

Early Results: Prüfer vs. Baum-Welch

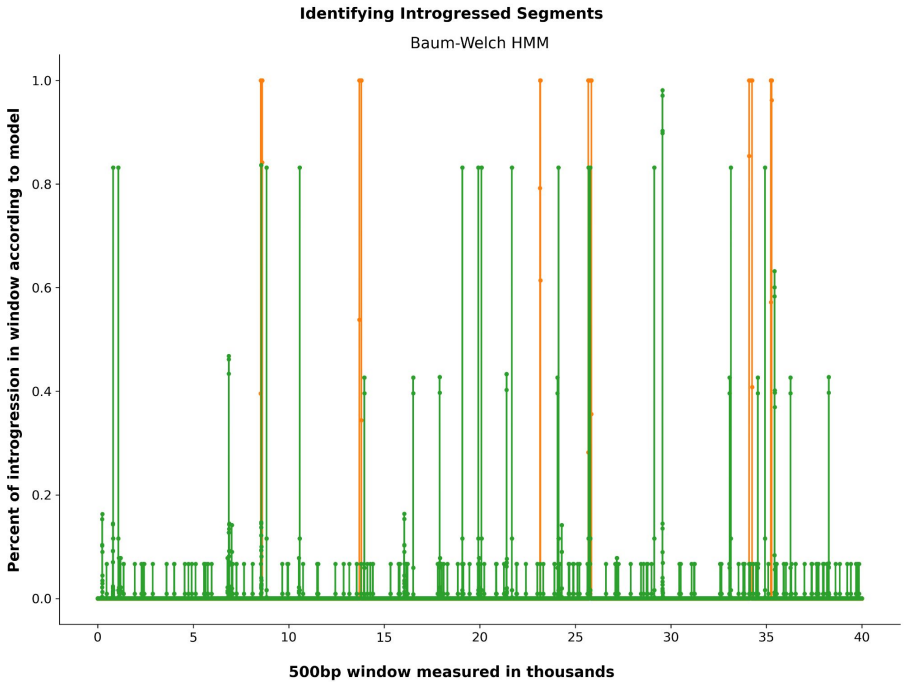
Baum-Welch HMM visualized after 5 steps

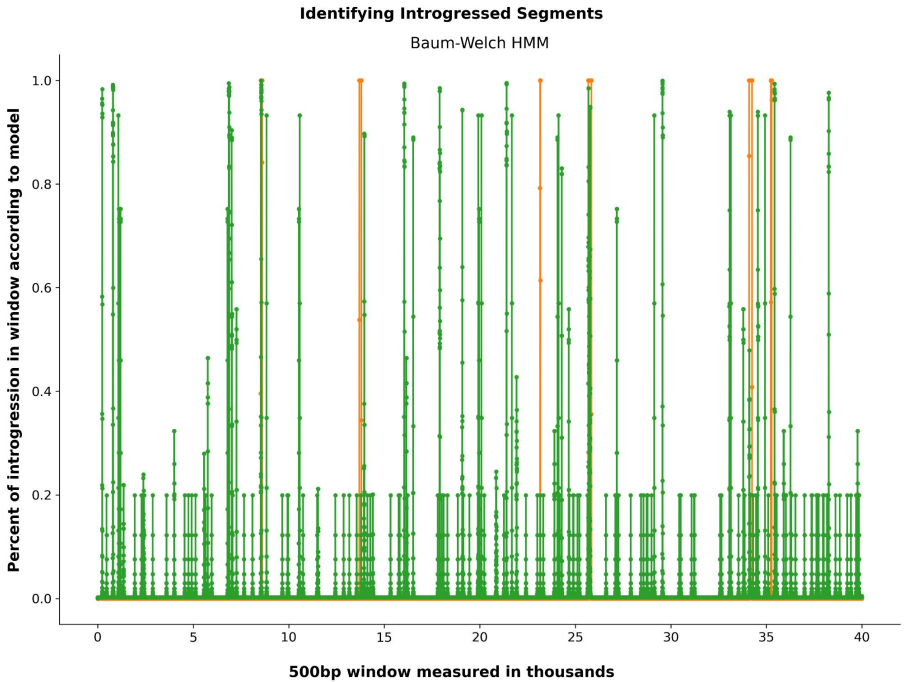




Early Results: Prüfer vs. Baum-Welch

Baum-Welch HMM visualized after 10 and 20 steps





Hidden States != Observed States

Observed sequence around introgressed segment 1, covering 500bp windows 8534 to 8614
NN**C**NNNNNNNNNNN **C**NN**C**NNNNN**CC**NNNN**C**NNN**C**NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Observed sequence around introgressed segment 2, covering 500bp windows 13696 to 13802
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN **C**NNNNNN
NNNNNNNNN

Observed sequence around introgressed segment 3, covering 500bp windows 23152 to 23170
NNNNNNNNNN **C**NNNNNNN

Observed sequence around introgressed segment 4, covering 500bp windows 25678 to 25842
N**C**NNNNNNN**CC**NNNNNNNN**C**NNNNNNNNNNNN **C**NNNNNN**C**NNNNNNNNNNNNNNNNNNNN **C**NNNN**C**NNNNNNNNNNNNNNNNNNNNNN **C**NNNNNNNN
NN**CC**NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Observed sequence around introgressed segment 5, covering 500bp windows 34097 to 34249
N**C**NNNNNNNN**C**NNNNNNNN**C**NNNNNNNNNNNNNNNNNNN **C**NNNNNNNNNNNNNNNNNNNNN **C**NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Observed sequence around introgressed segment 6, covering 500bp windows 35230 to 35276
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Loosening Window Assumptions...

Pattern A (0011) - NOT CONSISTENT

AFR1	AFR2	EUR	NEAN
1	1	0	0
0	**0**	**1**	**1**

Loosening Window Assumptions...

Pattern B (1100) - NOT CONSISTENT

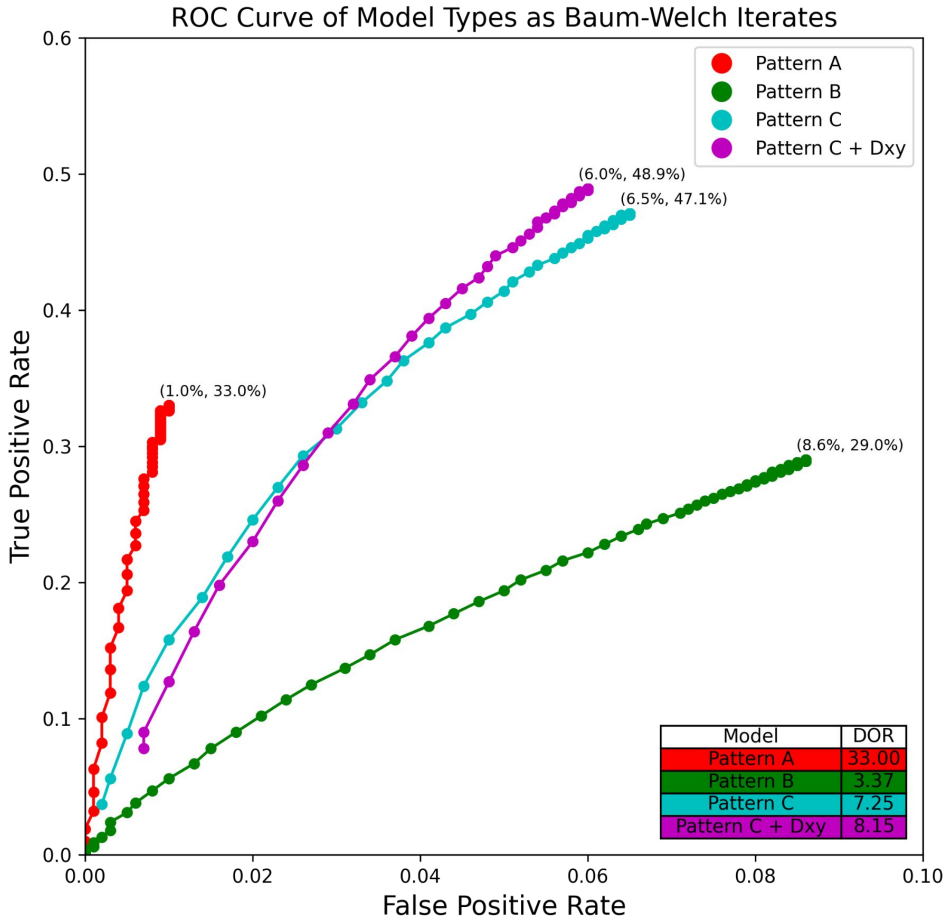
AFR1	AFR2	EUR	NEAN
1	**1**	**0**	**0**
0	0	1	1

Loosening Window Assumptions...

Pattern C (0011 or 1100) - CONSISTENT!

AFR1	AFR2	EUR	NEAN
1	**1**	**0**	**0**
0	**0**	**1**	**1**

Results:


ROC Curve of Model Types as Baum-Welch Iterates
Pattern A
Pattern B
Pattern C
Pattern C + Dxy
(6.0%, 48.9%)
(6.5%, 47.1%)
(1.0%, 33.0%)
(8.6%, 29.0%)
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.00
0.02
0.04
0.06
0.08
0.10
Model
DOR
Pattern A
33.00
Pattern B
3.37
Pattern C
7.25
Pattern C + Dxy
8.15

Further Analysis

A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy

Russell Corbett-Detig[1,2]*, Rasmus Nielsen[2,3]

Detecting archaic introgression using an unadmixed outgroup

Laurits Skov[1]*, Ruoyun Hui[2], Vladimir Shchur[3], Asger Hobolth[1], Aylwyn Scally[2], Mikkel Heide Schierup[1], Richard Durbin[2,3]*

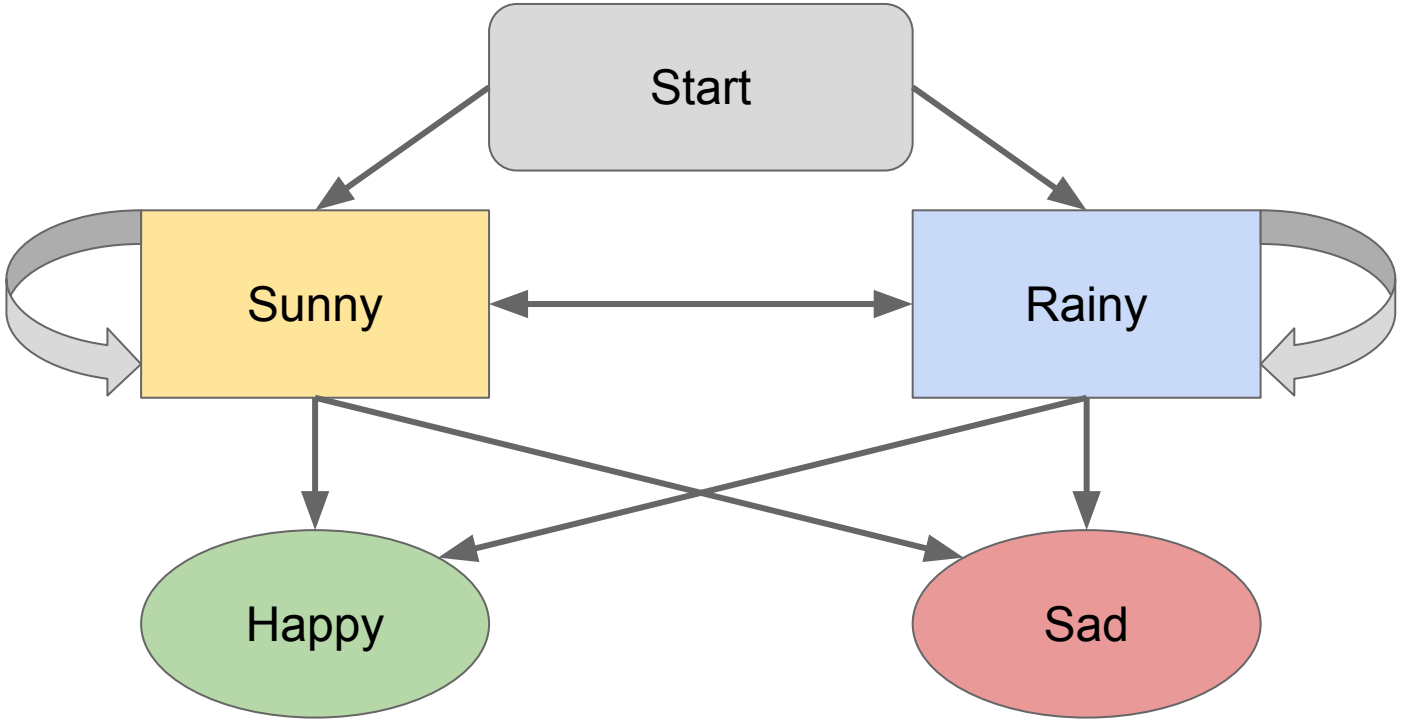
100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains

Benjamin M Peter[1]


Questions?


Thank you!

HMM Example



Parameters λ

π	S	R
	.75	.25

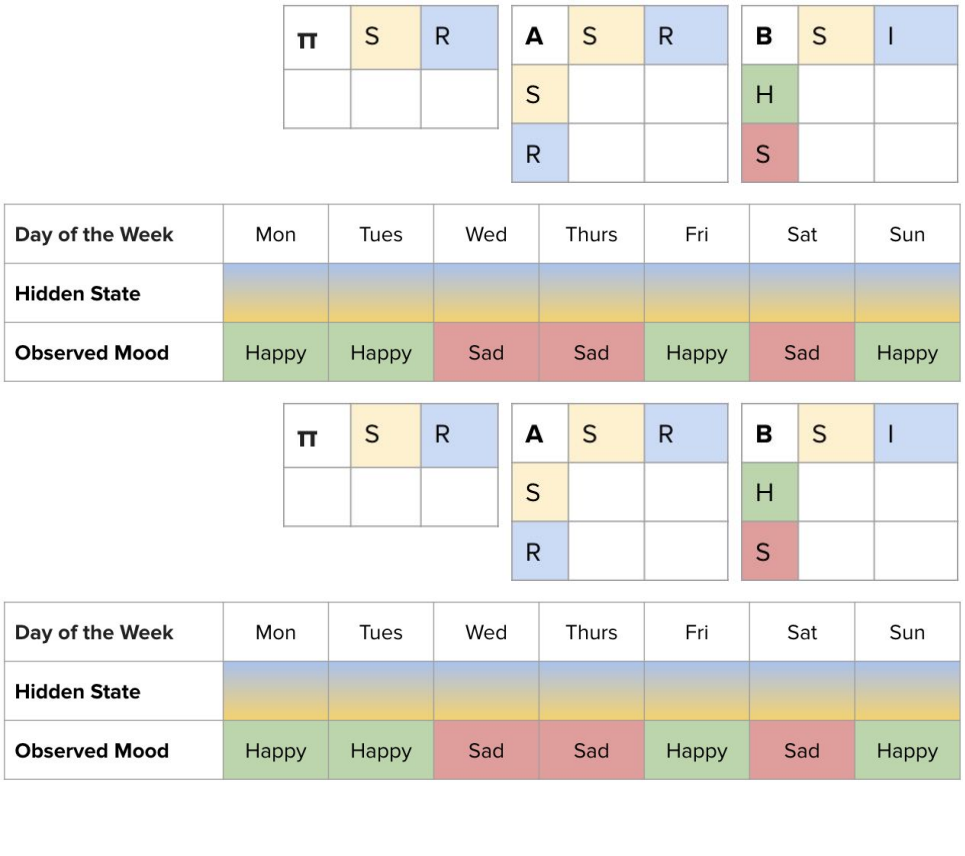
A	S	R
S	.75	.25
R	.25	.75

B	S	I
H	.9	.1
S	.1	.9

Hidden State/Observed State Example

Day of the Week	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
Hidden State							
Observed Mood	Happy	Happy	Sad	Sad	Happy	Sad	Happy

How Baum-Welch Works



Use parameters λ to calculate state probabilities γ

Use **distribution** of state probabilities γ to **refine** parameters λ

Use **updated** parameters λ to **recalculate** state probabilities γ

Recur until convergence

Works Cited:

1. Durand, E. (2011). *Testing for Ancient Admixture between Closely Related Populations*

2. Green, R. (2010). *A Draft Sequence of the Neanderthal Genome*

3. Higham T. (2014). *The Timing and Spatiotemporal Patterning of Neanderthal Disappearance*

4. Kakumani, R. (2012). *Identification of CpG islands in DNA sequences using statistically optimal null filters*

5. Lewis, B. (2012). *Understanding Humans: An Introduction to Physical Anthropology and Archaeology*

6. Peede, D. (in prep). *Leveraging Patterns of Ancestral & Derived Allele Sharing to Infer the Admixture Proportion*

7. Prüfer, K. (2014). *The complete genome sequence of a Neanderthal from the Altai Mountains*

8. Racimo, F. (2015). *Evidence for archaic adaptive introgression in humans*

9. Racimo, F. (2017). *Signatures of Archaic Adaptive Introgression in Present-Day Human Populations*

10. Rabiner, L. (1989). *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*

11. Sankararaman, S. (2014). *The genomic landscape of Neanderthal Ancestry in Present-Day Humans*

12. Seguin-Orlando, A. (2014). *Genomic structure in Europeans dating back at least 36,200 years*

13. Soares, P. (2010). *The Archaeogenetics of Europe*

14. Won, K. (2006). *Evolving the Structure of Hidden Markov Models*