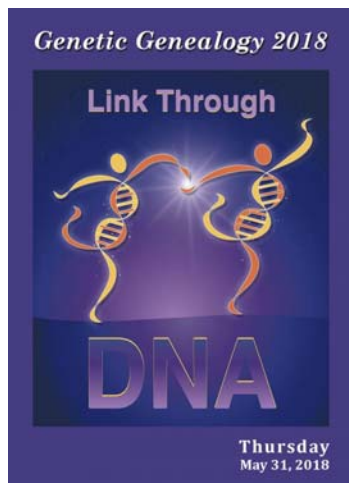


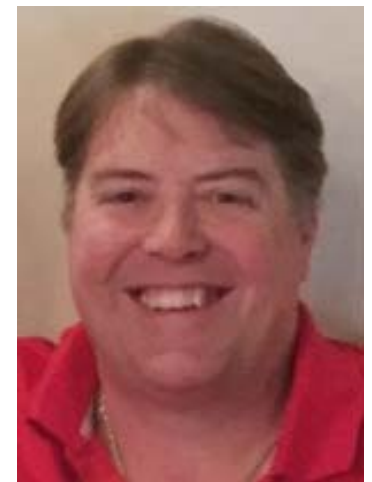
Insights on DNA Ancestral Percentages

How are those ethnicity and ancestral percentages calculated?



Brad Larkin

Prepared for the
Southern California Genealogical Society
Jamboree 2018



Topics

- Overview
- Simple DIY Example
- Methods of Major Labs
- Case Studies: Paper versus Labs
- Unexpected Result Possibilities
- Potential Future Improvements

Overview

- Many DNA testing laboratories are advertising ethnicity or ancestral population percentages for participants.
 - How are those figures calculated?
 - Why are there such differences in the results from one company to another?
- This presentation will provide insights on how these percentages are derived.

Timeline of Ancestral Estimates

- 2000
 - Family Tree DNA (FTDNA) first direct-to-consumer genetic genealogy lab formed, offering Y-chromosome testing
- 2005
 - First National Geographic *Genographic Project* recruiting participants worldwide, focused on Y and MtDNA. FTDNA lab performs the testing.
- 2007
 - 23andMe V1 begins consumer testing but focused on health-related markers.
- 2008
 - 23andMe V2 microarray chip introduced
- 2010
 - 23andMe V3 chip introduced along with *Ancestry Painting* ethnicity estimate using only three (3) reference populations.
 - FTDNA releases *Family Finder* autosomal product using HGDP as core reference population data source.
- 2012
 - Ancestry.com releases autosomal product and includes ethnicity estimates (U.S. only)
 - Launch of *Geno 2.0* product focused on SNP but also including an ethnicity estimate as well as Neanderthal & Denisovan estimate.
 - 23andMe updates to V2 *Ancestry Composition*
 - FTDNA upgrade to V2 of their *Population Finder* estimates.
- 2013
 - 23andMe V4 chip but sales inhibited by FDA. *Ancestral Composition* estimates based on 31 populations.
- 2014
 - FTDNA stops using Doug McDonald's licensed ethnicity algorithms and creates new, in-house process called *myOrigins* (V 1.0).
- 2015
 - Ancestry DNA starts sale of kits in UK, Ireland, Australia, New Zealand and Canada.
- 2016
 - Ancestry DNA kits now for sale in 29 additional countries. Begins using V2 SNP Microarray chip.
 - FTDNA adds *Ancient Origins* estimation for autosomal contribution of three prehistoric European population groups.
- 2017
 - 23andMe V5 microarray chip
 - FTDNA *myOrigins* V 2.0 adds some populations
- 2018
 - 23andMe *Populations Collaboration Program* seeking samples from underrepresented countries

Ancestral Percentage

- Defined as an estimation of one's biological, ethnic and/or geographical origins based on DNA analysis.
 - More formally called Biogeographical Ancestry (BGA)
 - aka Ethnicity Estimates, Admixture Analysis, etc.
- Mass marketed on television and other media
 - “Holy Cow, I found out I was 35% Martian and 25% Imaginerian”*

Ancestral Percentages Calculation

1. Reference Population Sampling
2. Test Participant
3. Mathematical Algorithms to attribute individual markers to Reference Population(s)
4. Aggregate individual markers to overall percentage estimate for the Participant
5. Geographic plotting for maps

1. Reference Population Sampling

- Obtain DNA test results of *single nucleotide polymorphism* (SNP) markers for groups of people thought to represent the native population of various geographical and ethnic groups around the globe.

2. Test Participant

- Test Participant on the same DNA markers used in the Reference Population datasets.

3. Algorithms Attributing Individual Markers to Populations

- Use mathematical algorithms in software to assess the probability that each of the Participant's DNA markers originate in one or more of the Reference Populations.

4. Aggregate Marker Probabilities for Individual Participant

- Aggregate the probabilities of individual markers using additional algorithms into an overall percentage estimate for the Participant's DNA sourced from Reference Populations.

5. Geographic Map Plot

- Software to produce the maps highlighting the location of the Participant's ancestral Reference Populations.
 - Typically geo-surface plots.

Topics

- Overview
- Simple DIY Example
- Methods of Major Labs
- Case Studies: Paper versus Labs
- Unexpected Result Possibilities
- Potential Future Improvements

Simple DIY Example

- DIY = *Do It Yourself*
- Illustration using published SNP markers studied in the *1000 Genomes Project*
 - Underlies all three major genetic genealogy labs reference populations

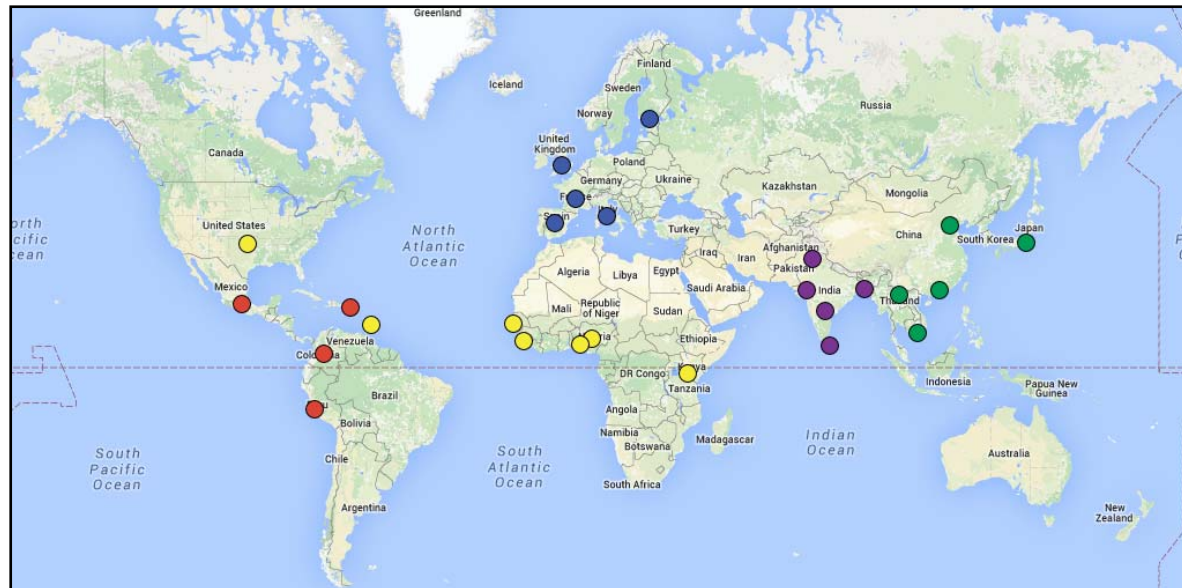


Illustration from Raw Results File

- Ancestry DNA raw data file
 - rs12562034
 - Chromosome 1, Position 768448 (hg19)
 - Participant Allele Values AA

```
AncestryDNA.txt
#AncestryDNA raw data download
#This file was generated by AncestryDNA at: 05/19/2014 08:59:31 MDT
#Data was collected using AncestryDNA array version: V1.0
#Data is formatted using AncestryDNA converter version: V1.0
rsid      chromosome  position  allele1  allele2
rs4477212      1          82154    T        T
rs3131972      1          752721   G        G
rs12562034     1          768448   A        A
```

SNP DB Illustration

- Lookup **rs12562034** on dbSNP*
 - ancestral allele: **G**

Reference SNP (refSNP) Cluster Report: **rs12562034**

dbSNP Short Genetic Variations

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez: dbSNP for Go

Reference SNP (refSNP) Cluster Report: **rs12562034**

RefSNP	Allele	HGVS Names
Organism: human (<i>Homo sapiens</i>)	Variation Class: SNV: single nucleotide variation	NC_000001.10:g.768448G>A
Molecule Type: Genomic	RefSNP Alleles: A/G (FWD)	NC_000001.11:g.833068G>A
Created/Updated in build: 120/150	Allele Origin	NR_015368.2:n.287+3964G>A
Map to Genome Build: 108/Weight 1	Ancestral Allele: G	NR_047519.1:n.287+3964G>A
Validation Status:	Variation Viewer:	NR_047520.1:n.287+3964G>A
Citation: PubMed LitVar	Clinical Significance: NA	NR_047521.1:n.287+3964G>A
	MAF/MinorAlleleCount: A=0.1919/961 (1000 Genomes)	NR_047522.1:n.287+3964G>A
		NR_047523.1:n.287+3964G>A
		NR_047524.1:n.287+3964G>A
		NR_047525.1:n.154+3964G>A
		NR_047526.1:n.287+3964G>A

SNP Details are organized in the following sections:

[GeneView](#) [Map](#) [Submission](#) [Fasta](#) [Resource](#) [Diversity](#) [Validation](#)

Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh38.p7	108	1	833068	NT_032977.10	247080	Fwd	G	Fwd	view	mapup

*dbSNP – National Institute of Health, Reference for Short Genetic Variations <https://www.ncbi.nlm.nih.gov/SNP>

Population Diversity in dbSNP

- dbSNP shows some Reference Population allele variations and percentages for various academic datasets in its *Population Diversity* section.
 - EAS = east asia: 1000 Genomes super population
 - 39.6% have A alleles at rs12562034
 - EUR = europe: 1000 Genomes super population
 - Only 9.24% have A alleles at rs12562034
 - AFR = africa: 1000 Genomes super population
 - Only 8.55% have A alleles at rs12562034
- So if we wanted to stop here, and make an ethnicity estimate for Participant using a single SNP allele rs12562034 of A based on three reference populations:
 - 39% chance of Being East Asian
 - 9% change of being European
 - 8% change of being African

rsid	chromosome	position	alt
rs4477212	1	82154	T
rs3131972	1	752721	G
rs12562034	1	768448	A

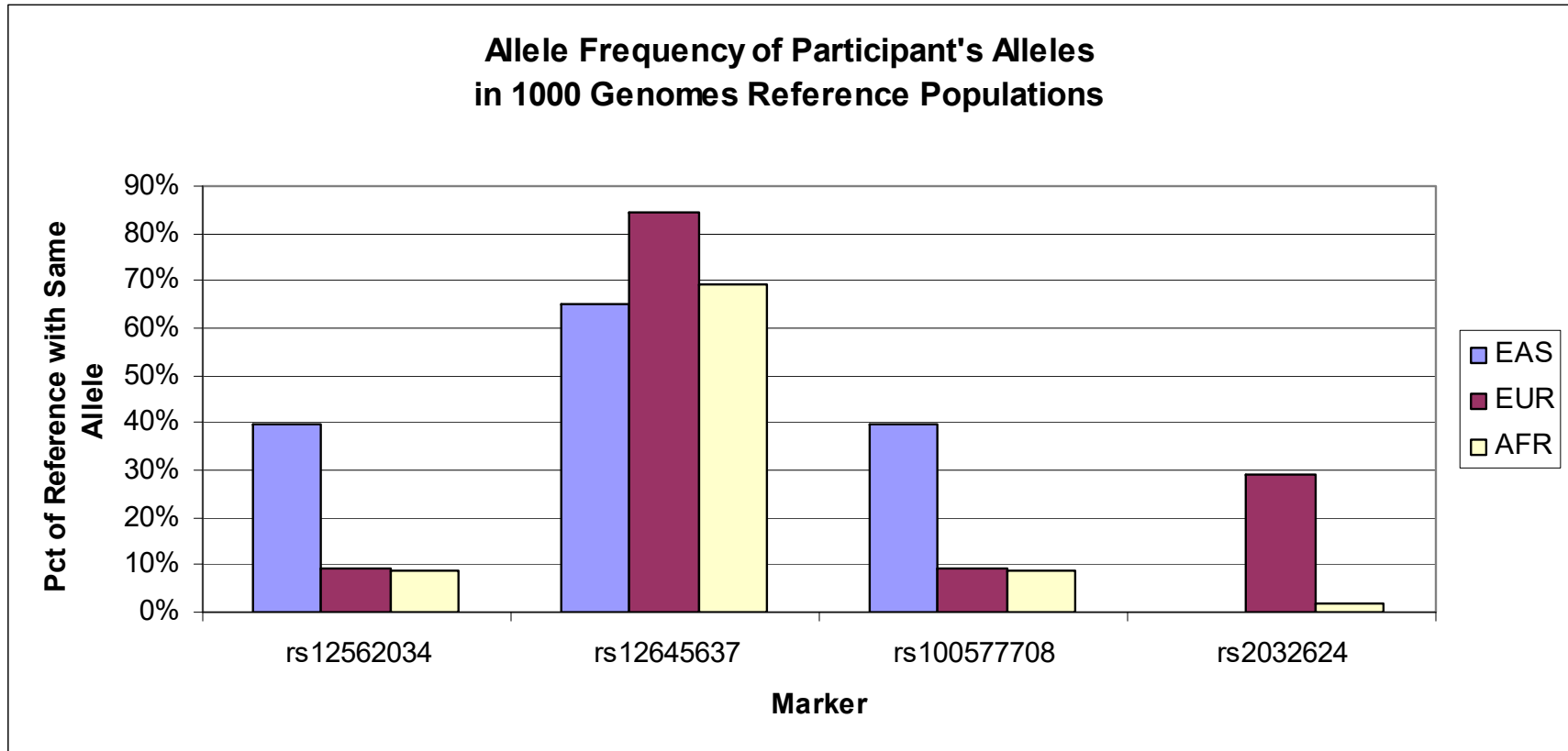
Population Diversity (Alleles in RefSNP orientation) . See additional population frequency from 1000Genome [\[here\]](#)

ss#	Sample Ascertainment			Genotype Detail				Alleles		
	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/G	G/G	HWP	A	G
ss1289339247	EAS		1008	AF					0.39579999	0.60420001
	EUR		1006	AF					0.09240000	0.90759999
	AFR		1322	AF					0.08550000	0.91450000
	AMR		694	AF					0.08930000	0.91070002
	SAS		978	AF					0.30059999	0.69940001

Improve Estimate by Adding Markers

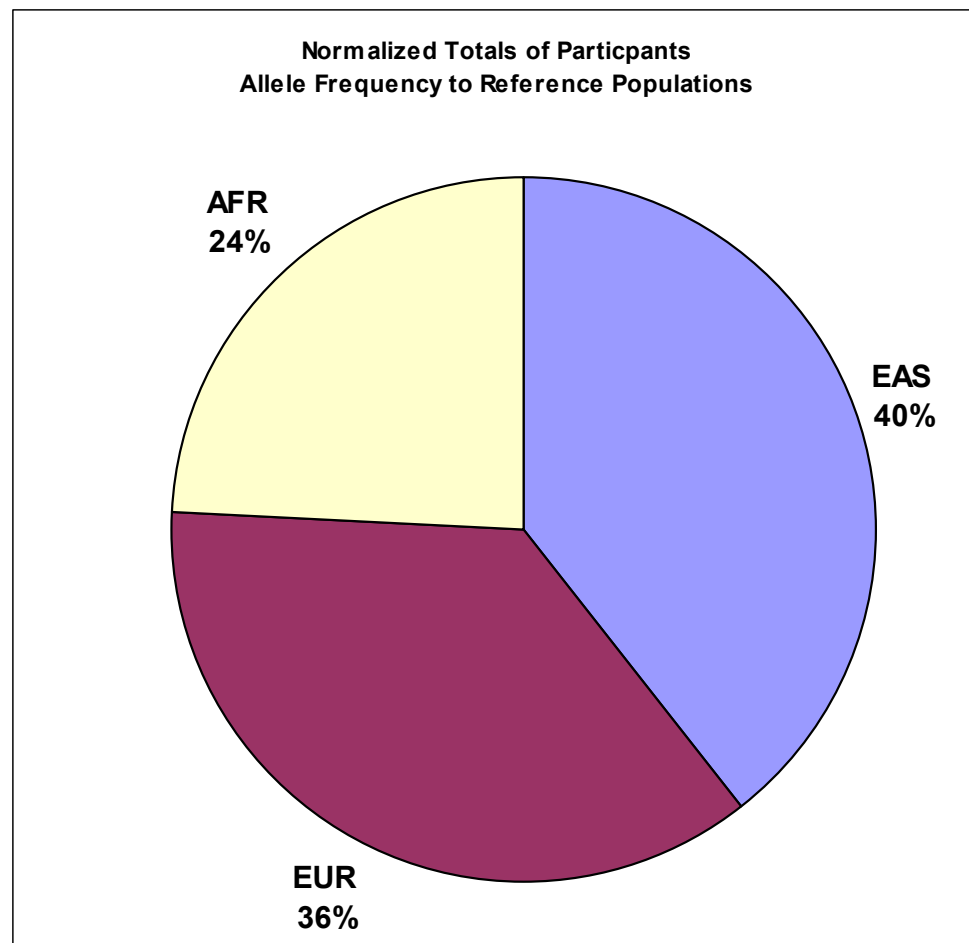
- rs12645637
 - Ancestral allele: C
 - Participant: CC
 - Population Diversity:
 - EAS = 64.9% C
 - EUR = 84.6% C
 - AFR = 69.4% C
- rs10057708
 - Ancestral: G
 - Participant: AA
 - Population Diversity:
 - EAS = 39.6% A
 - EUR = 9.2% A
 - AFR = 8.6% A
- rs2032624
 - Ancestral: T
 - Participant: CC
 - Population Diversity:
 - EAS = 0.2% C
 - EUR = 29.2% C
 - AFR = 1.9% C

Comparing Multiple Markers



Simplified Ancestral Estimate

- Our Algorithm:
 - Add up the Frequency of Participant's Allele for each Reference Population
 - Normalize totals to 100% of Participant's DNA.



Reference Populations	rs12562034	rs12645637	rs100577708	rs2032624	Population Totals	Normalized
EAS	0.396	0.649	0.396	0.002	1.443	40%
EUR	0.092	0.846	0.092	0.292	1.322	36%
AFR	0.086	0.694	0.086	0.019	0.885	24%

Topics

- Overview
- Simple DIY Example
- Methods of Major Labs
- Case Studies: Paper versus Labs
- Unexpected Result Possibilities
- Potential Future Improvements

Comparison of Lab Methods

	FTDNA	Ancestry.com	23andMe
Product Name*	<i>MyOrigins 2.0</i>	<i>Ethnicity Estimate V2</i>	<i>Ancestry Composition</i>
Principal Authors	Rhazid Khan & Rui H	Ball, Barber et al	Eric Durand, Chuong Do, et al
Analytical Software	<i>Admixture</i> Software (Alexander 2009) Block relaxation approach applied to Bayesian inference from <i>Structure</i> (Pritchard 2000) accelerated with fast sequential quadratic programming and a quasi-Newton acceleration method.	<i>Admixture</i> Software with many layers of refinement and quality control.	Proprietary <i>Ancestry Deconvolution</i> computer data processing (pipeline) focused on ancestral origin of chromosome segments (~ 100 markers each) using 3 stage process involving machine learning, phasing; Support Vector Machines.

*For a list of references, see conference syllabus or online list at <https://www.surnamedna.com/?p=1976>

Comparison of Reference Populations

	FTDNA	Ancestry.com	23andMe
Reference Population Sources	GeneByGene (FTDNA customer database); 1000 Genomes Project; Human Genome Diversity Project (CEPH-HGDP); HapMap Project; Estonian Biocenter	Proprietary Ancestry DNA reference collection (Sorenson database): 1,500; Ancestry DNA customers: 1,800; 1000 Genomes Project; Human Genome Diversity Project (CEPH-HGDP): 800; Utah Resident with European Ancestry (CEU); HapMap Project; Chinese & Japanese (CHB+JPT) Projects; Yoruba, Ibadan; Nigeria, West Africa (YRI)	23andMe customers self-reported: 8,906; 1000 Genomes Project: 765; CEPH-HGDP: 941; HapMap3: 87
Size of Reference Population	2,943	4,245 candidates to 2,995 used in models	11,091 (self referencing)
# Reference Populations	55 populations resolved into 24 clusters	52 countries into 26 distinct populations New: 166 regions	31 genetic populations in white paper, New: 151 regions

Comparison: Strengths & Weaknesses

	FTDNA	Ancestry.com	23andMe
# of SNPs in common with Ref Populations	245,039	~ 300,000	not stated, but limited by markers in academic projects ~ 300,000
Acknowledged Limitations	Statistical model-based, number of clusters chosen drives results. If preconceived model of world is wrong, predictions will inherently have error.	<i>“Array currently performs best in European populations (as expected), and captures the least amount of variation in African populations.”</i>	<i>“In Europe the classifier is usually able to distinguish Northern from Southern from Eastern European haplotypes, but encounters difficulty at the sub-regional, let alone the national level.”</i>
Advantages	1 st genetic genealogy company, large, diverse database of global population samples.	Best family trees and integration with DNA results. Sorenson samples. Large marketing effort & customer database.	Attempts chromosome segment phasing aligned to historical genetic inheritance mechanisms.

Undocumented Filtering

- All the labs employ a number of layers and phases of filtering the Reference Population databases, and the DNA markers used for analysis
 - Trying to remove ambiguity and bias in the ancestral attributions.
 - But it introduces a type of bias as well
 - The specific markers tested are sort of hard-coded into each generation of the microarray-based lab equipment
 - Changes are possible but only infrequently
 - Backward compatibility limits number of markers used for comparison
 - Because older, smaller academic datasets are still being used as references, only about 300,000 markers per Participant are probably being used for ethnicity estimates
 - » (0.01% of a full genome test).
 - Unlike Y-DNA and MtDNA testing, labs are currently not disclosing the specific markers used in their models & processes.

Contradictions Between Test Companies

- Why are there such differences in the results from one company to another?
 - Markers Selected for Comparison
 - Reference Population databases
 - Especially where using customer results as references
 - Differences in Population Models
 - How many populations are there anyway?

Topics

- Overview
- Simple DIY Example
- Methods of Major Labs
- Case Studies: Paper versus Labs
 - British Isles Examples
 - American Examples with European Roots
- Unexpected Result Possibilities
- Potential Future Improvements

Kennett Case Study

- Debbie Kennett did a nice comparison of FTDNA *myOrigins* estimates across three family generations
 - Husband's results not consistent with his 100% British genealogy
 - Americans being 'more British' than some of her British family

Estimated Ancestral Percentage from British Isles for Kennett family of England*

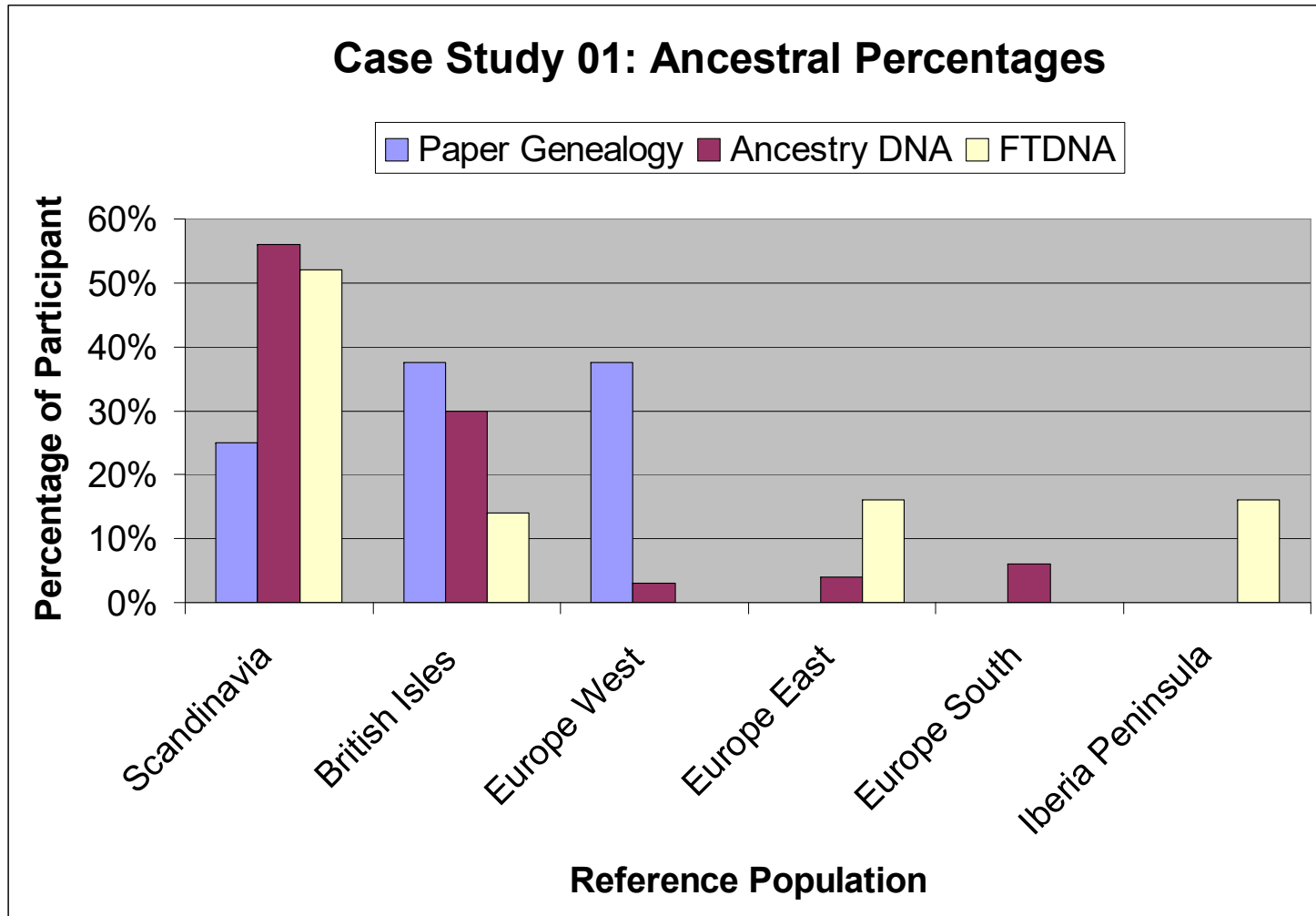
Tester	FTDNA myOrigins 1.0	FTDNA MyOrigins 2.0
Debbie's dad	40%	99%
Debbie's mum	7%	100%
Debbie	57%	100%
Debbie's husband	38%	15%
Debbie's son	75%	100%

*Debbie Kennett, Cruwys news, [Three Generations of FTDNA MyOrigins 2.0 results from Family Tree DNA](#)

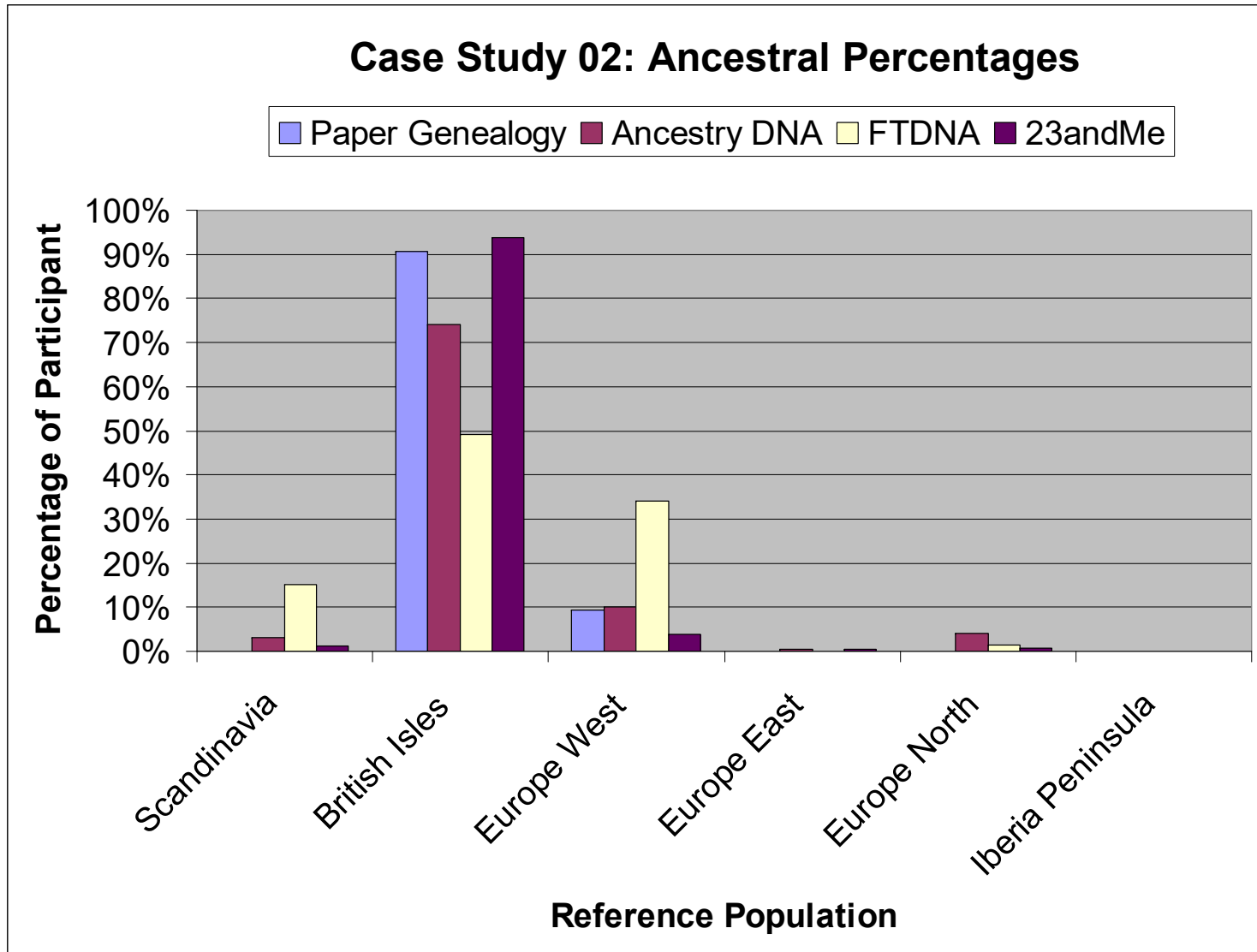
Case Study 01: Regional Estimates

Region	Paper Genealogy	Ancestry DNA	FTDNA
Scandinavia	25%	56%	52%
British Isles	38%	30%	14%
Europe West	38%	3%	0%
Europe East	0%	4%	16%
Europe South	0%	6%	0%
Iberia	0%	0%	16%

Case Study 01 - Graph



Case Study 02: Graph



Case Study 01: Continent Estimates

Region	Paper Genealogy	Ancestry DNA	FTDNA
Europe	100%	>99%	98%
Africa	0%	0%	0%
Asia	0%	0%	0%
Native American	0%	0%	0%

=> Very consistent at the continental level, even if not so consistent on regional or country-level.

Topics

- Overview
- Simple DIY Example
- Methods of Major Labs
- Case Studies: Paper versus Labs
- Unexpected Result Possibilities
- Potential Future Improvements

Paradoxes

- For all that we '*know*' about genetic genealogy, there is more that we do not know.
- There are many paradoxes in our current understanding as the molecules of DNA, the test equipment, the labs, and biology all have a lot of complexity.
 - e.g. 2,948,611,470 base pairs sequenced in the Human Reference Genome (hg38p12) yet there are an estimated 139,658,362 base pairs un-sequenced; plus 10,972,074 groups (Scaffolds) of sequenced pairs whose position is unclear*

*Analysis by author using SAMTOOLS, SQL Server based on hg38p12 Human Reference Genome
<https://www.ncbi.nlm.nih.gov/grc/human/data>

Population Paradoxes

- Original American colonists were about 100,000 from 17th century British Isles
 - A subset of all English lineages
 - There is more genetic diversity in the British Isles than in their (more-numerous) American descendants
- Genetic Diversity within Africa is higher than other continents
 - Most African-Americans descend from only ~ 300,000 Africans

Contradictions to Oral History and Self Image

- My grandmother told me she was part [*ethnicity X*] which is not showing up in my results?

Possibility 1 for Unexpected Result

- There is little or none of that particular ancestor's DNA left in your DNA.
 - The amount of DNA you inherit from any single ancestor halves with every generation
 - Averages less than 2% total beyond five (5) generations.
 - Beyond eight (8) generations, you will likely have ancestors from whom you have inherited No (0) autosomal DNA
 - They are still your ancestor, but may not be in your genes.
 - But some of your known cousins could have inherited DNA from that ancestor so that's why it is good to test as many known family members as possible.

Possibility 2 for Unexpected Result

- Some source populations are not well distinguished
 - Genetic Similarity
 - There may have been a lot of mixture and movement between geographical areas.
 - France vs. Germany
 - Scotland vs. Ireland
 - Incomplete Sampling
 - Subpopulations or ethnic groups simply missed in sampling to-date in parts of the world
 - 26 global populations vs. more than 6,500 human languages*

*Steven R Anderson, Linguistic Society of America, [How many languages are there in the world?](#)

Possibility 3 for Unexpected Result

- Your ancestor's genetics may not be adequately represented in the Reference Populations used in the current generation of Ancestry Estimates.
 - The family story could be confirmed with expanded marker panels and future Reference Population sampling.
 - Unfortunately, Test Labs do not provide marker-specific attribution by allele value and population

Possibility 4 for Unexpected Result

- Adoption & Fosterage
 - Your ancestor in question may have been raised in a particular culture, place, or population and self-identified with that group
 - But he or she was not genetically descended from that population's ancestors

Possibility 5 for Unexpected Result

- It could be a myth.
 - Myths are not uncommon in genealogy and history.
 - [Why Do So Many Americans Think They Have Cherokee Blood?](#) in Slate Magazine

Topics

- Overview
- Simple DIY Example
- Methods of Major Labs
- Case Studies: Paper versus Labs
- Unexpected Result Possibilities
- Potential Future Improvements

Forecasting the Future

- Potential Improvements in Ancestral DNA Estimation
 - Phasing and Imputation
 - Inheritance Trees
 - Full Genomic Sequencing
 - Ancient DNA Sampling

Future – Phasing and Imputation

- Improved phasing of autosomal results matching specific DNA markers to specific ancestors in our tree
 - Figuring out DNA markers from current generation going backwards in time
 - Imputation of immediate ancestor DNA markers
 - leading to better ethnicity estimates for our immigrant ancestors

Future – Inheritance Trees

- Construction of authoritative inheritance trees of DNA mutations (SNPs)
 - Figuring out SNP mutations from ancient DNA and tracing the mutations forward in time.
 - SNPs spread via population movements and mixing
 - Inheritance trees now getting strong with other types of DNA tests
 - Y-chromosome
 - MtDNA

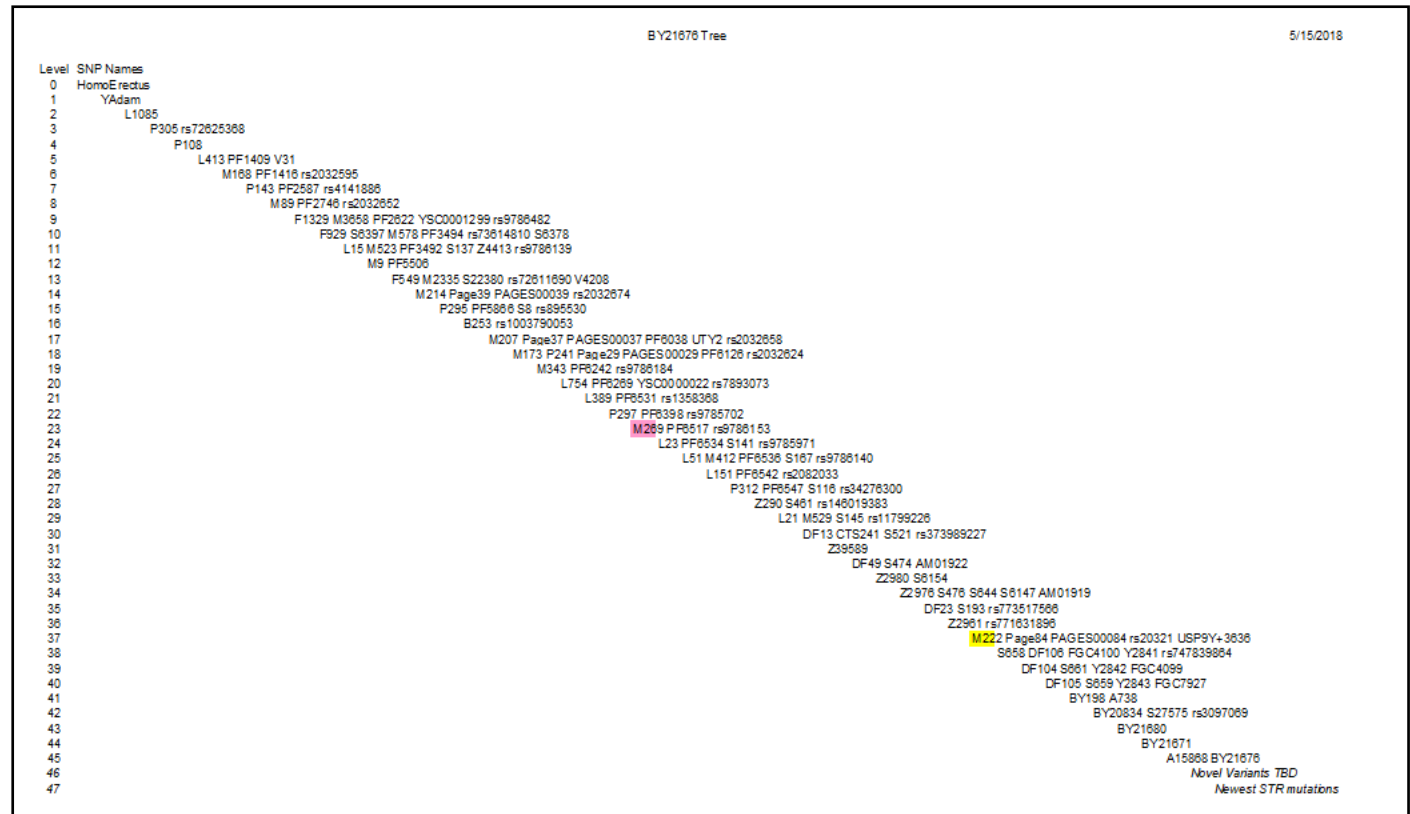
Y-DNA Phylogenetic Inheritance Tree

- Example of actual inheritance tree branch of Y chromosome marker SNP *BY21676*

•44 branches from Y-Adam

•22 branches below M269

•8 branches below *M222*



Future – Full Genomic Sequence

- Autosomal Matching on Full Genomic Sequences of ~ 2.8 billion markers is now possible although computational tools are still in their early days.
 - *“Your match results will be ready in approximately 3.7 years”*

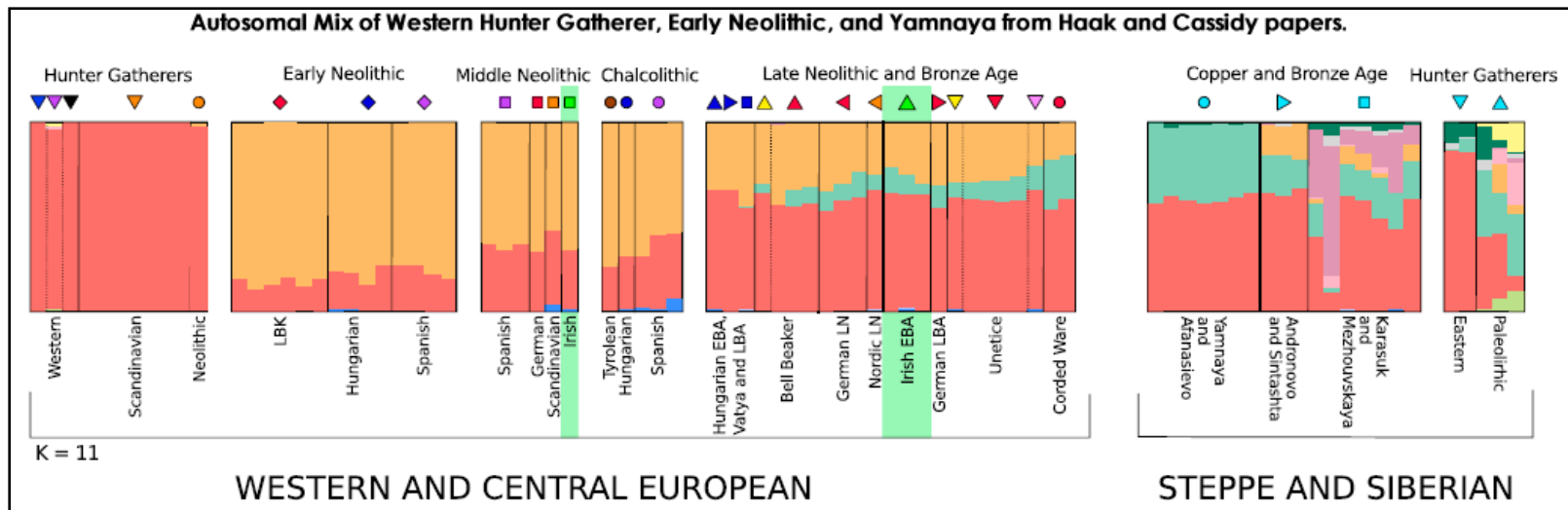
Future – Ancient DNA

- Incorporation of ancient DNA samples from grave sites of Reference Populations.
 - FTDNA already does this a bit with their *Ancient Origins* analysis
 - Estimates your autosomal DNA composition from three (3) prehistoric source populations of Europe.



Ancient DNA is Hot!

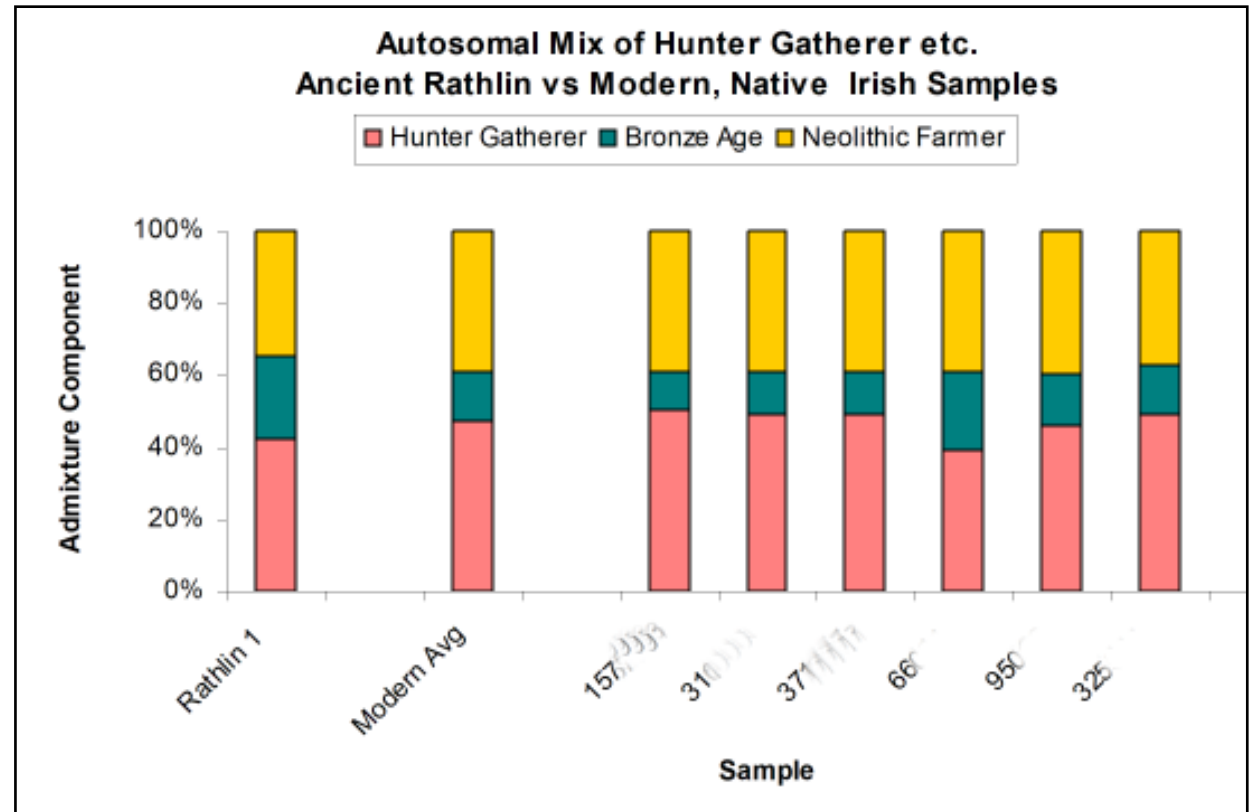
- Recent breakthroughs on ancient DNA extraction has dramatically increased the number of samples and markers available for ancient population studies.
- FTDNA Ancient Origins percentages directly comparable to ancient European population research!



¹Haak et al (2015), Massive migration from the steppe is a source for Indo-European languages in Europe, [Nature](#)

Ancient vs. Modern Samples

- Comparison of Admixture values of Rathlin 1 Bronze Age aDNA¹ to modern Irish²
 - Fairly similar mixture of three source populations


















¹Cassidy et al (2015), Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome, [PNAS](#) ² [Irish Mapping](#) and Larkin DNA Projects 2017 author compilation

Ask For Improvement

- Lab could disclose which DNA markers they are using in their Ancestral Percentage analysis
 - What the ancestral characterization is used for each marker
 - As was done with the 1000 Genomes Project

Population Diversity (Alleles in RefSNP orientation) . See additional population frequency from 1000Genome [\[here\]](#)

ss#	Sample Ascertainment				Genotype Detail				Alleles	
	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/G	G/G	HWP	A	G
ss1289339247	EAS		1008	AF					0.39579999	0.60420001
	EUR		1006	AF					0.09240000	0.90759999
	AFR		1322	AF					0.08550000	0.91450000
	AMR		694	AF					0.08930000	0.91070002
	SAS		978	AF					0.30059999	0.69940001

Conclusion

- Ancestral Percentage Calculation is successful in a broad sense
 - Between very distant reference populations
 - European versus Sub-Saharan Africa
 - Native American versus European
- Be cautious with High Resolution distinctions between countries on the same continent
- Biggest Difference between labs is probably the SNP markers used for analysis.