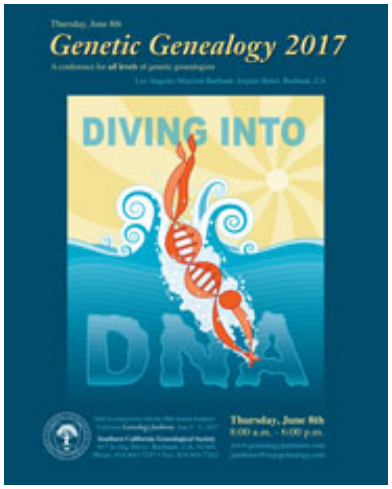
Don't Discard the Y-STRs

Genetic Genealogy
Testing Strategies in 2017



Brad Larkin

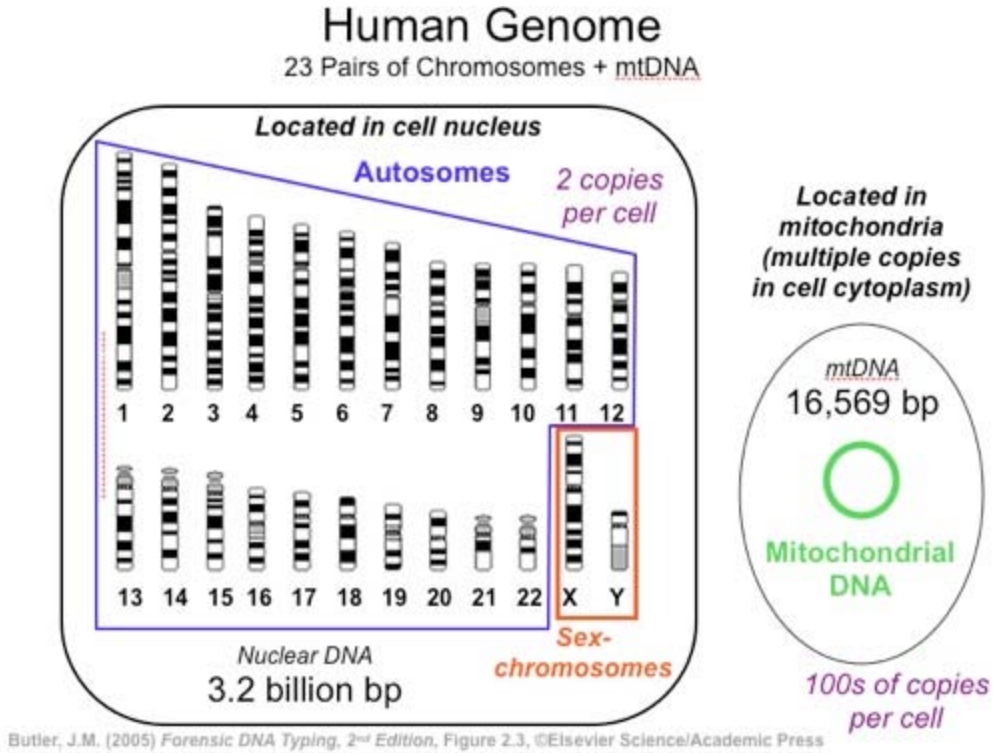
Prepared for the
Southern California Genealogical Society
Jamboree 2017

Topics

- Biology of DNA for Genealogy
- Y-DNA Testing Technologies
- Price – Value Comparison

Two Types of DNA Structures

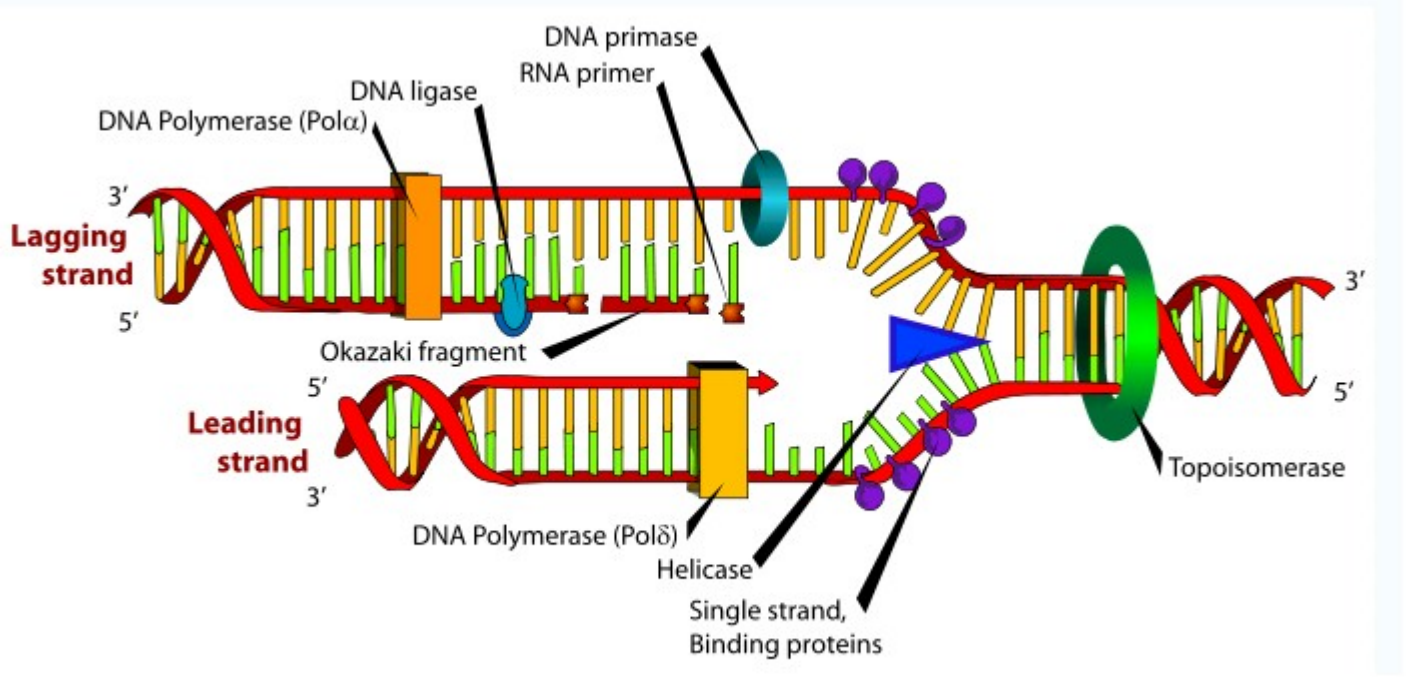
- Nuclear DNA
  - 23 chromosome pairs.
  - One set per normal cell.
- Mitochondria DNA (MtDNA)
  - 1 ring of pairs
  - Many spread throughout cell



Butler JM (2005) Forensics DNA Typing 2nd Edition
Figure 2.3, Elsevier Science Academic Press

DNA Mutations

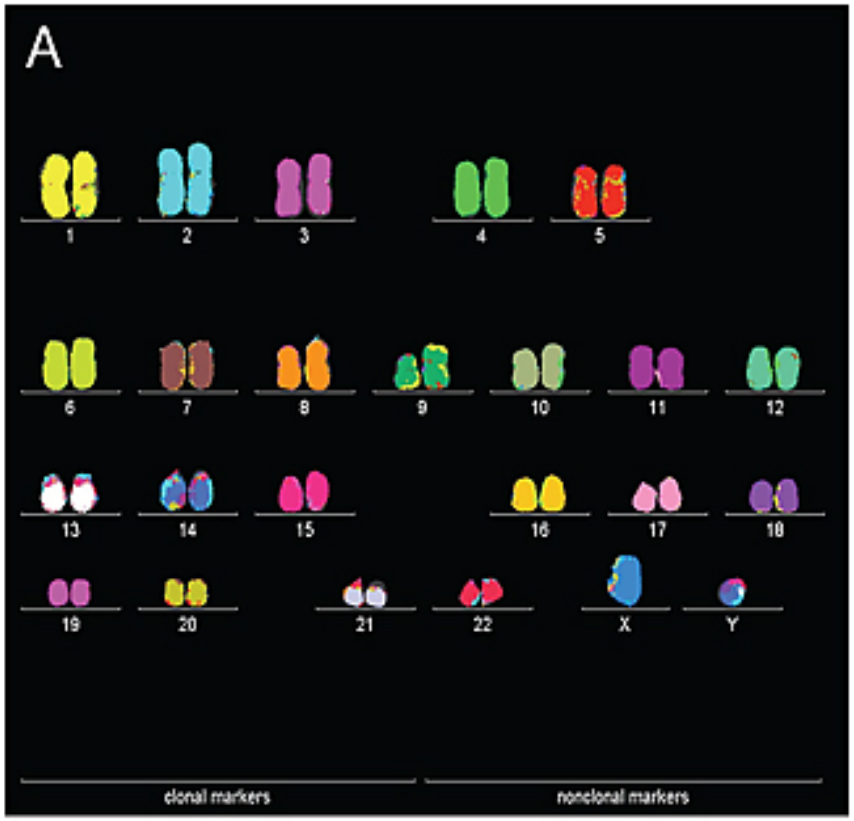
- Genetic Genealogy made possible by mutations and mixing that occurs during reproduction.
  - Individual mutations, insertions, deletions
  - Combination of maternal and paternal strands.
  - Makes it possible to identify descendants of different individuals based on presence or absence of specific mutations and combinations of mutations.

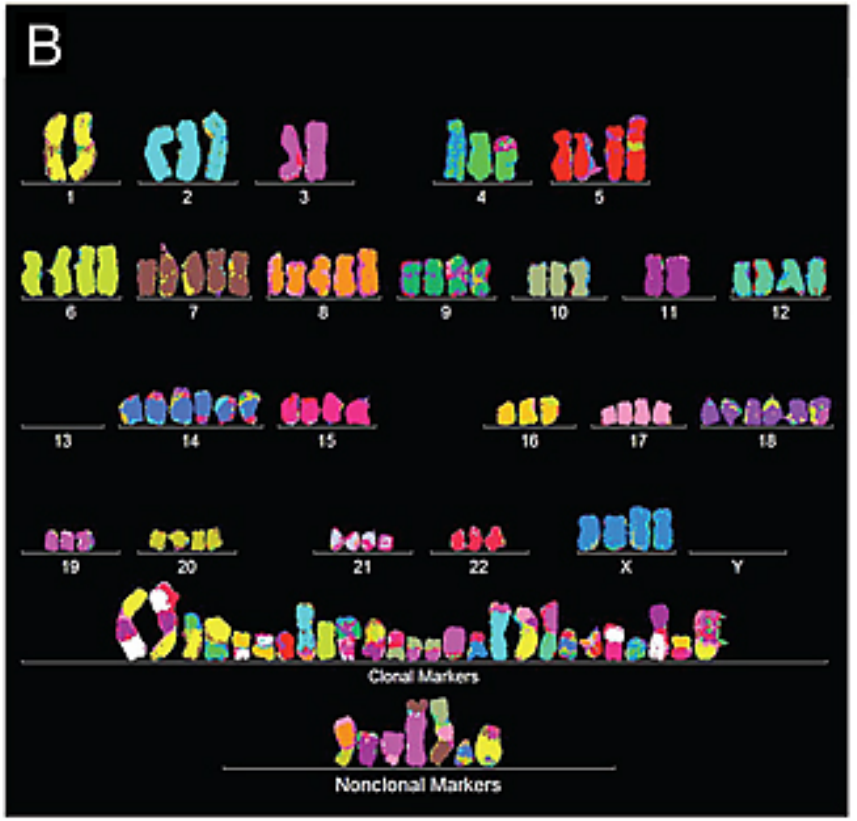


DNA Replication by Mariana Ruiz

Side Note – DNA of Cancer Cells

- Karotypes - Stained Chromosome Images
  - Illustrating chromosomal disruption called *aneuploidy in Cancer Cells*[1]

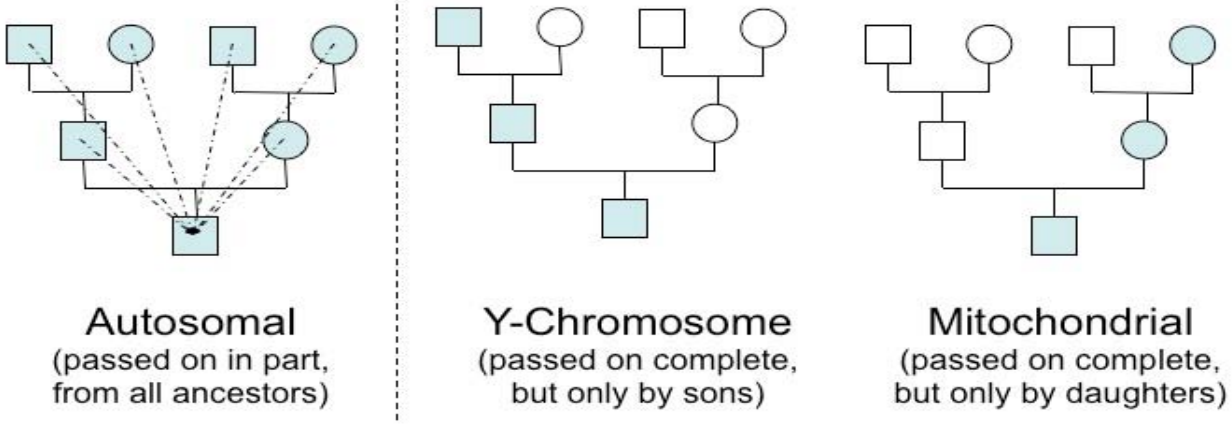




[1] Robert Sanders (2011) Are cancers newly evolved species? , UC Berkeley, Berkeley News

Chromosome Fit for Genealogy

	Autosomal (Microarray)	Y-Chromosome (Y-37 STRs)	Mitochondrial (HVR1+HVR2)
Recombination - Mixing	Yes	No	No
# Coding Genes	~ 30,000	86	37
# Markers Initial Test	708,093	37	1,120
Mutation Rate	0.5 bp/gen = 354,047 per generation	μ = 0.0041 markers/generation 1 change per 165 years	0.48 bp/MY = 1 change per 1,860 years



Autosomal
(passed on in part, from all ancestors)

Y-Chromosome
(passed on complete, but only by sons)

Mitochondrial
(passed on complete, but only by daughters)

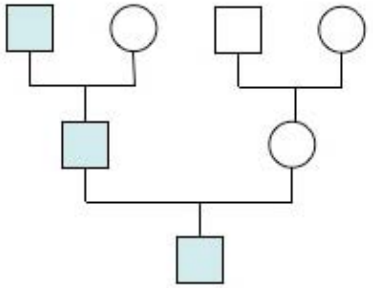
Surname-Related Research Questions

- Y-DNA focused, surname-related research questions can include:
  - Classifying worldwide linkages in diaspora populations
    - All Larkin's living today in Shannon River Valley in Ireland
    - Ashkenazi descendants of a particular 18th century Eastern European Rabbi
  - Connecting American families with common surnames to colonial roots
    - Relationship of all Reynolds families living in Texas in 1860.
  - Connecting genealogical lineages for surnames which have highly-variable spelling
    - All Robinson / Robertson / Roberson families living in Charleston South Carolina area today.
      - Y-DNA can go to much higher resolution than surname spelling.

Recap: Biology – Types of DNA Used for Genealogy

Chromosome Type	Strengths	Weaknesses
Y-DNA	Deep ancestry with measurable variation in surname era.	Only patrilineal ancestry. Only Male samples.
MtDNA	Sample any Gender Deep ancestry, strong signal	Only matrilineal ancestry. Slow mutation – no differentiation in surname era.
Autosomal	Sample any Gender Traces all lineages Adoptee research	Fuzzy signal lost after a few generations. Recombination makes interpretation challenging

Y-DNA



Y-Chromosome
(passed on complete, but only by sons)

- Biology
  - Non-recombining portion of the Y-chromosome
    - Holds signal over multiple generations
    - Mutates at a rate that we can see differences occur within families over the course of decades and centuries

SNP vs. STR Measurement

- SNP = Single Nucleotide Polymorphism
- Mutation in a single base pair at a specific position
- Expressed a 'positive' when different from all other human beings.
  - e.g. *position rs1019875*
  - Person1 T ATCC T = -
  - Person2 T ACCC T = +
- Analogous to '*Trunk and Branches of the Tree*'

- STR = Single Tandem Repeat
- Repeating patterns of multiple base pairs
- Allele Count = number of repetitions of particular pattern
  - e.g. *DYS389*

  Person1 T AACC T = 1

  Person2 T AACC AACC T = 2
- Analogous to '*Leaves on the Tree*'

DNA Markers for Y-DNA Genealogy

- SNPs
  - Used for breaking mankind into major groups, called Haplogroups. Change infrequently and thus serve as major branches in the tree of man.
- STRs
  - Used for clustering Y-DNA results and surname studies.
  - Change fast, thus providing good dividing points in the past 1000 years when surnames have been in use.
  - Inexpensive to test a lot of markers.

SNP and STR Parallel Mutations by Generations

- Integrating Patterns of SNPs and STRs is the correct way to think about mutations and matches.
- Root, Limbs, Branches, Twigs
- Homoplasy
  - STRs the same by coincidence as they are faster at mutating
  - Illustrated at right with two lineages part of R-L21 haplogroup

DYS385b	Lineage A	Lineage B
Generation 0	L21+ M222- DYS390=25	L21+ M222- DYS390=25
Generation 1	L21+ M222+ DYS390=25	L21+ M222- DYS390=25
Generation 10	L21+ M222+ DYS390=25	L21+ M222- DYS390=25 ZZ29+
Generation 20	L21+ M222+ DYS390=26	L21+ M222- DYS390=25 ZZ29+
Generation 30	L21+ M222+ DYS390=26	L21+ M222- DYS390=26 ZZ29+

Trees Need SNP and STR

	Lineage A	Lineage A1	Lineage B
Generation 30	L21+ M222+ DYS390=26	-	L21+ M222- DYS390=26 ZZ29+
Generation 31 (DNA Testing begins)			
STR Only	DYS390=26	DYS390=27	DYS390=26
SNP Only	L21+ M222+	L21+ M222+	L21+ ZZ29+
SNP + STR	L21+ M222+ ZZ29- DYS390=26	L21+ M222+ ZZ29- DYS390=27	L21+ M222- ZZ29+ DYS390=26

Topics

- Biology of DNA for Genealogy
- Y-DNA Testing Technologies
- Price – Value Comparison

Y-STR Testing

- Y-STR testing was the first Y-DNA direct-to-consumer genetic genealogy product.
  - Economically feasible in about the year 2000
- Testing of individual SNPs was expensive until later technologies
- With growth in genetic genealogy additional STR test panels were developed over time
  - e.g. 12 markers -> 25 markers -> 37 markers
- STR Result sets generally homogenous – every participant has a score at each marker
  - easy to compare

Scoring Microsatellites/STRs

Y-Chromosome Locus X[1]:

CGAATGCTTCTTATGCATGCATGCATGCATGCATGCAGGAC

ATGC repeated 6 times

Total length of PCR Product: "# of Base Pairs"

Sample	Locus	Motif	bp	Allele Value
#1	X	ATGC		6

Your Y-STR results are a listing of this Allele Value (the count of the pattern) at each of the Markers that the laboratory tested.

[1]From Taylor Edwards (2006) presentation "Laboratory Procedures, 2006" at the 2nd International Conference on Genetic Genealogy

Y-STR Result Interpretation

- For quick screening, STRs alone do a very good job for most individuals and within surname family groups.
- Compare individuals using **Correlation** of Multiple Markers
  - Especially correlation of STRs along with a shared SNP marker
  - Not necessarily any single value.
  - Remember – MUTATIONS ARE RANDOM

Interpreting STR Marker Correlation

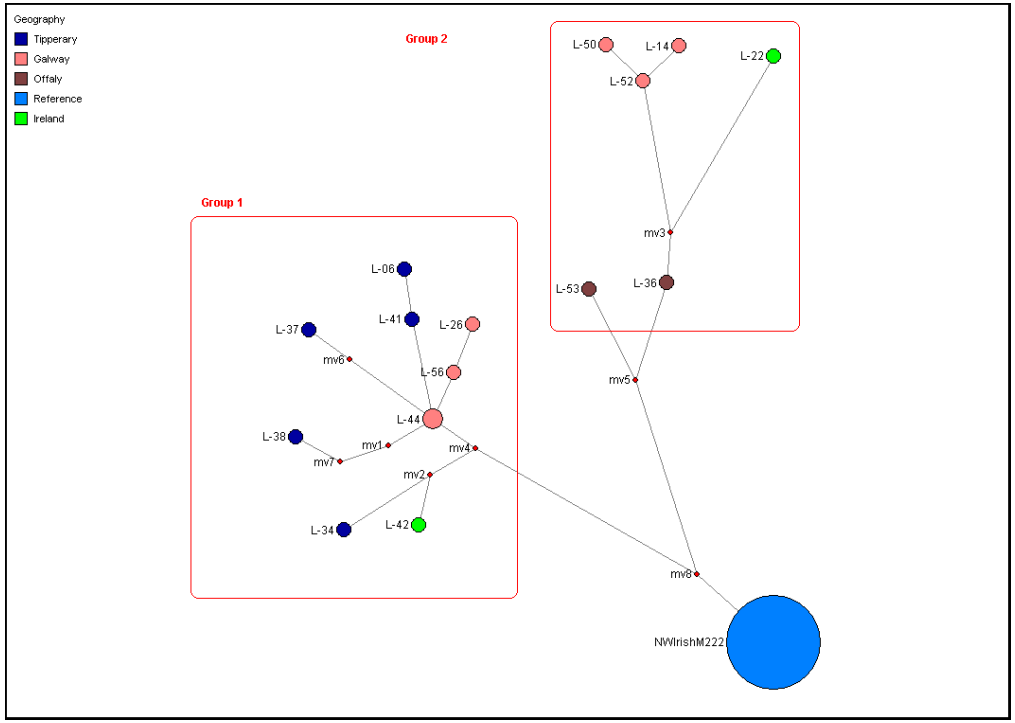
- Initial comparison between two samples with mediocre match at 12 and 25 markers.

	393	390	394	391	385a	385b	426	388	439	389-1	392	389-2
L06	13	26	14	11	11	14	12	12	13	13	14	29
L64	13	26	14	11	11	14	12	12	12	13	14	29

	Matches 12 Markers	Matches 25 Markers	Matches 37 Markers	Matches 67 Markers
L06 / L64	11/12	23/25	34/37	64/67
Percentage	92%	92%	92%	96%

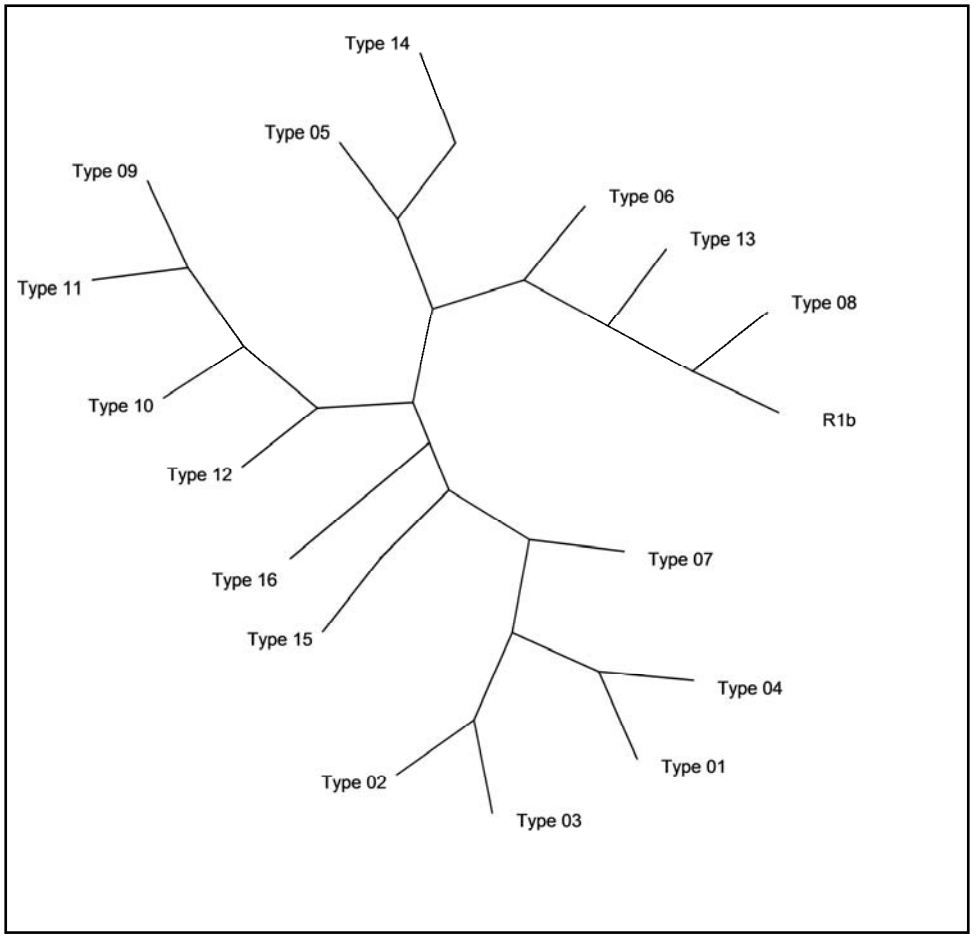
Data Analysis Identify Groups

- Even though the best genetic genealogy combines SNPs and STRs
- The economic value of Y-37 STR results are still a very important part in both identifying groups and distinguishing recent genealogy within groups.
  - Economical
  - Homogenous
  - Easier analysis
  - Freeware tools for graphing
  - Matching Databases



Phylogenetic Trees from STRs

- Larkin DNA Project Cladogram created purely with 37 marker STR-based TMRCA
  - Dec 2012
  - Using McGee's Y-DNA Comparison Utility and TreeView



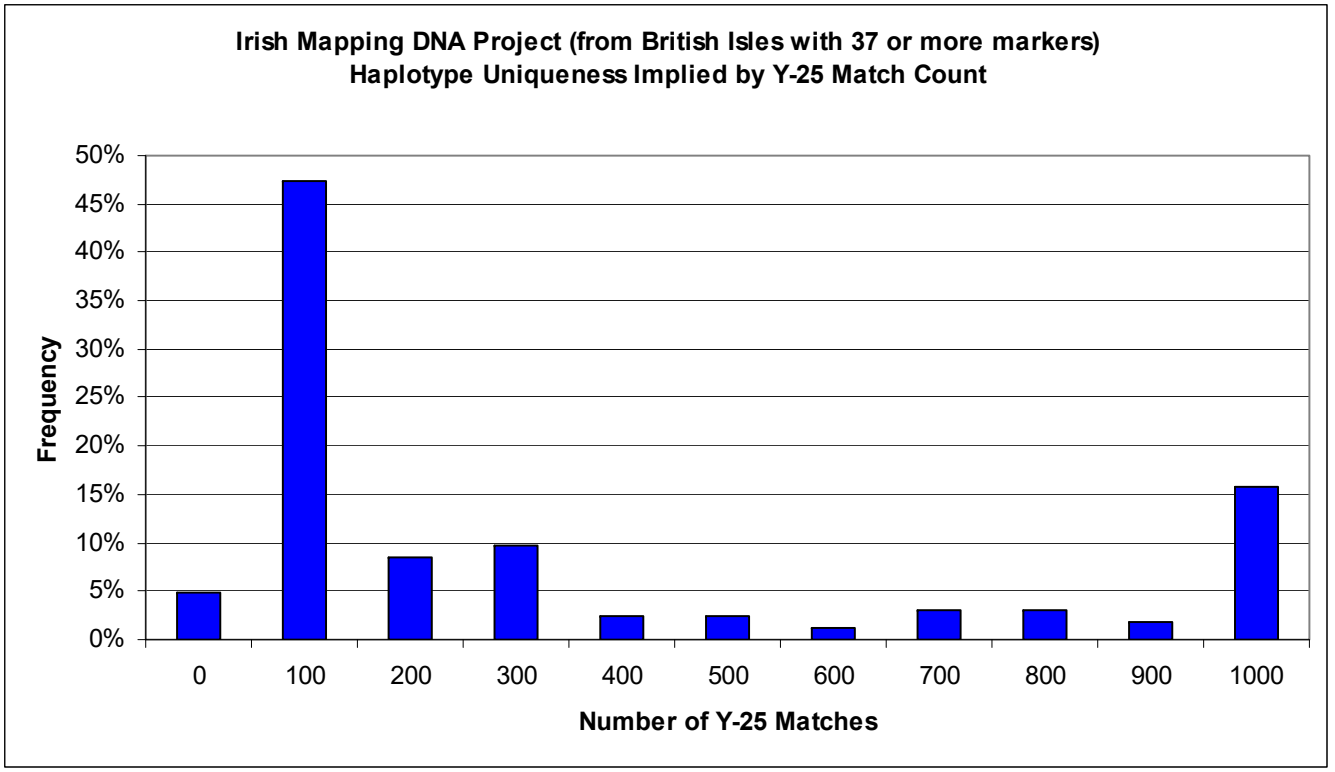
Match Count Distribution

- Histogram of Y-25 Match Count[1]
  - Bimodal Shaped Distribution

5% of samples with NO 25 marker matches

13% < 3 matches

27% > 500 matches



[1]Brad Larkin, Irish Mapping DNA Project , 2014, n=165

[2] Bennett Greenspan, Family Tree DNA, Sept 18, 2014

How Many STR Matches Is Enough?

- 12 Markers really is only enough to tell what major haplogroup (~ what continent) your deepest paternal ancestor came from.
- 37 Markers provides excellent basis for grouping Y-DNA lineages
  - Refine root and branching of groups with major SNPs
  - Costs less than $ 200 per sample
- 67 and 111 STR markers informative within family groups and where 37 markers leaves precise lineage determination ambiguous
  - But with larger SNP packages now available, an SNP package might be a better value.

Minimum 37 Y-STR Markers

- Surname era genealogical relatedness needs 37 marker matches
  - 12 marker matches – source continent
  - 25 marker matches – recent continent
  - 37 marker matches – same kin group in surname era
  - 67 marker matches – single, common ancestor in surname era

Y-Example Larkin Type 01

- STR Results for some Larkin men all in Type 01 M222, some of whom are known cousins.
  - Shows how important STRs remain for sorting out recent ancestry.
  - And combination of autosomal and STR results

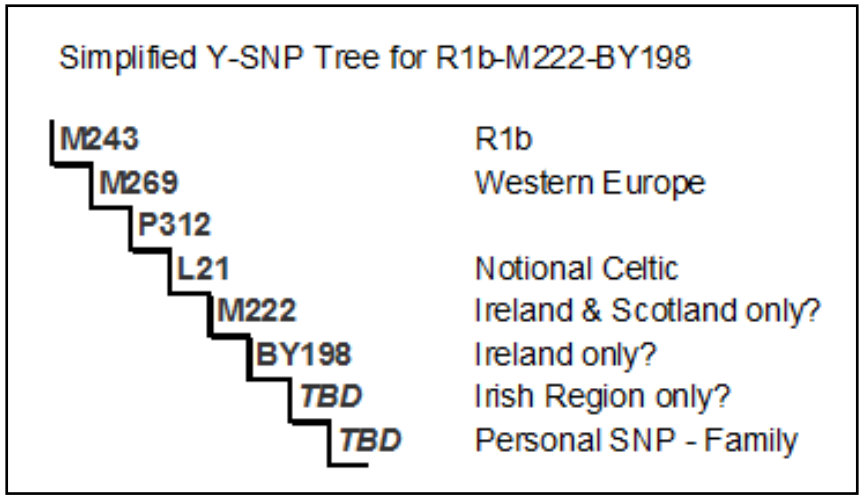
Identification	Identified SNP Differences	Autosomal Match (Shared cM)	STR Match to L-0064	Known Relationship to L-0064
L-0044 MAL	0	0	66/67	?
L-0077 LTL	0	108	37/37	3C
L-0006 BTL	0	135	110/111	3C1R
L00041 JFL	0	140	37/37	3C
L-0063 JTL	0	141	36/37	3C2R

Y-SNP Testing

- Haplogroups are ways of classifying the genetic ancestry based on distinguishing people who have an SNP mutation that is different from the rest of Humanity.
- First SNP tests for Y genetic genealogy were for major haplogroup confirmation.
  - Often Y-STR patterns for members of a haplogroup were noticed and so the Haplogroup assignment was *PREDICTED*
    - Predicted was a lot more affordable than confirmed with SNP marker value with Sanger sequencing.

SNP Tree Levels

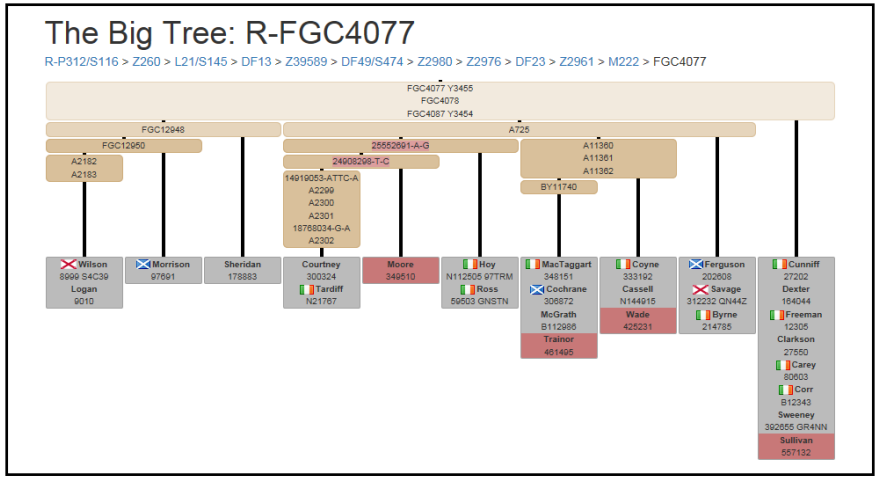
- The definition of Haplogroups has expanded as more SNP testing has been done to map the root & branches of the human Y-Chromosome Family Tree
  - Ongoing effort

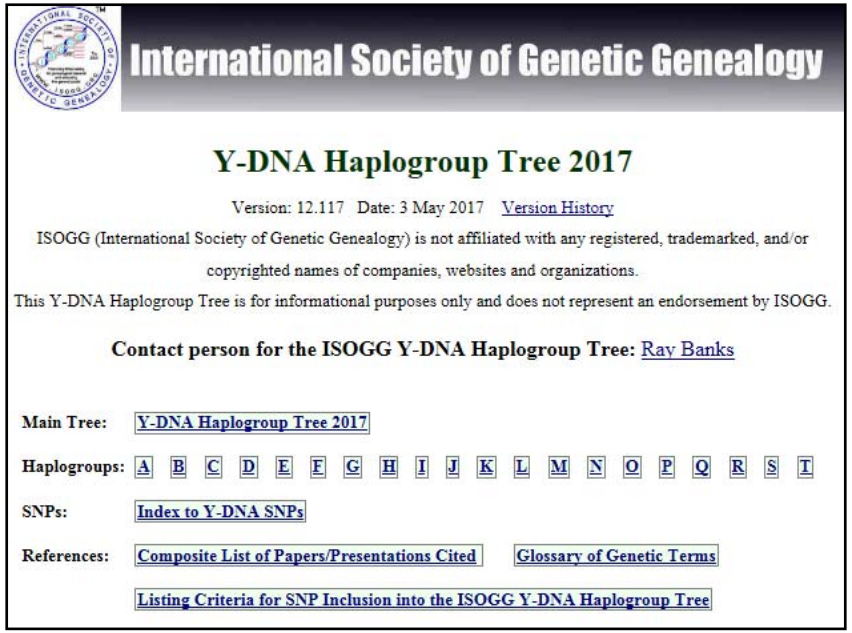


  - One can think of Y-SNP Haplogroups now as being at multiple levels
    - High Level e.g. R1b => 100 million people
    - Mid Level e.g. R-M222 => 2 million people
    - Low Level e.g. BY198 => 200,000 people
    - *Personal SNPs*, unique to a specific family in past 500 years

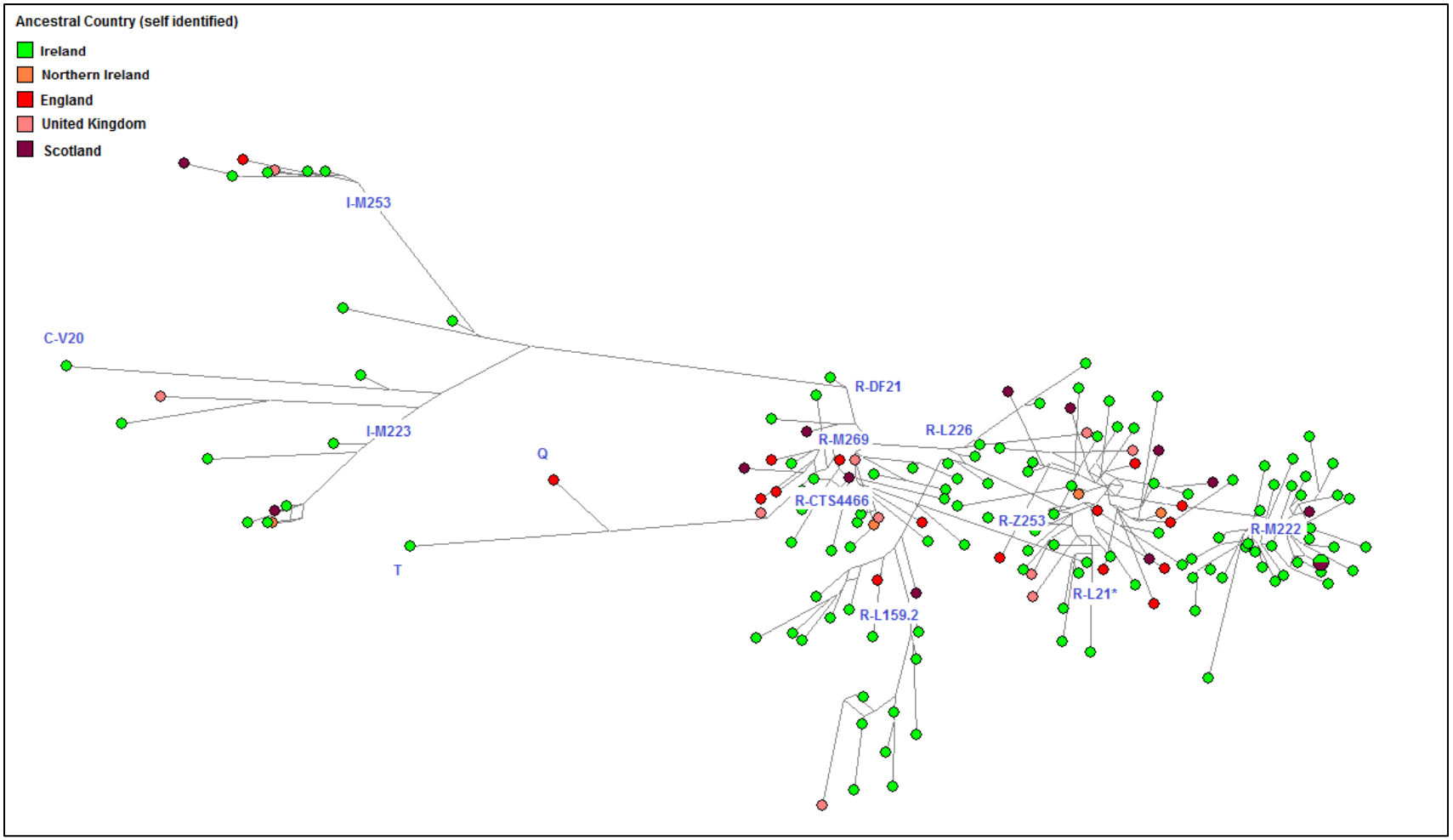
SNP Tree References

- YTree.Net and ISOGG maintain updated SNP-based Phylogenetic trees of Y-DNA.





Y-STR Clusters with Haplogroups



Brad Larkin, Irish Mapping DNA Project , 2014, samples with uniform 37 markers and ancestral county identified, n=165

SNP Chip Test Technology

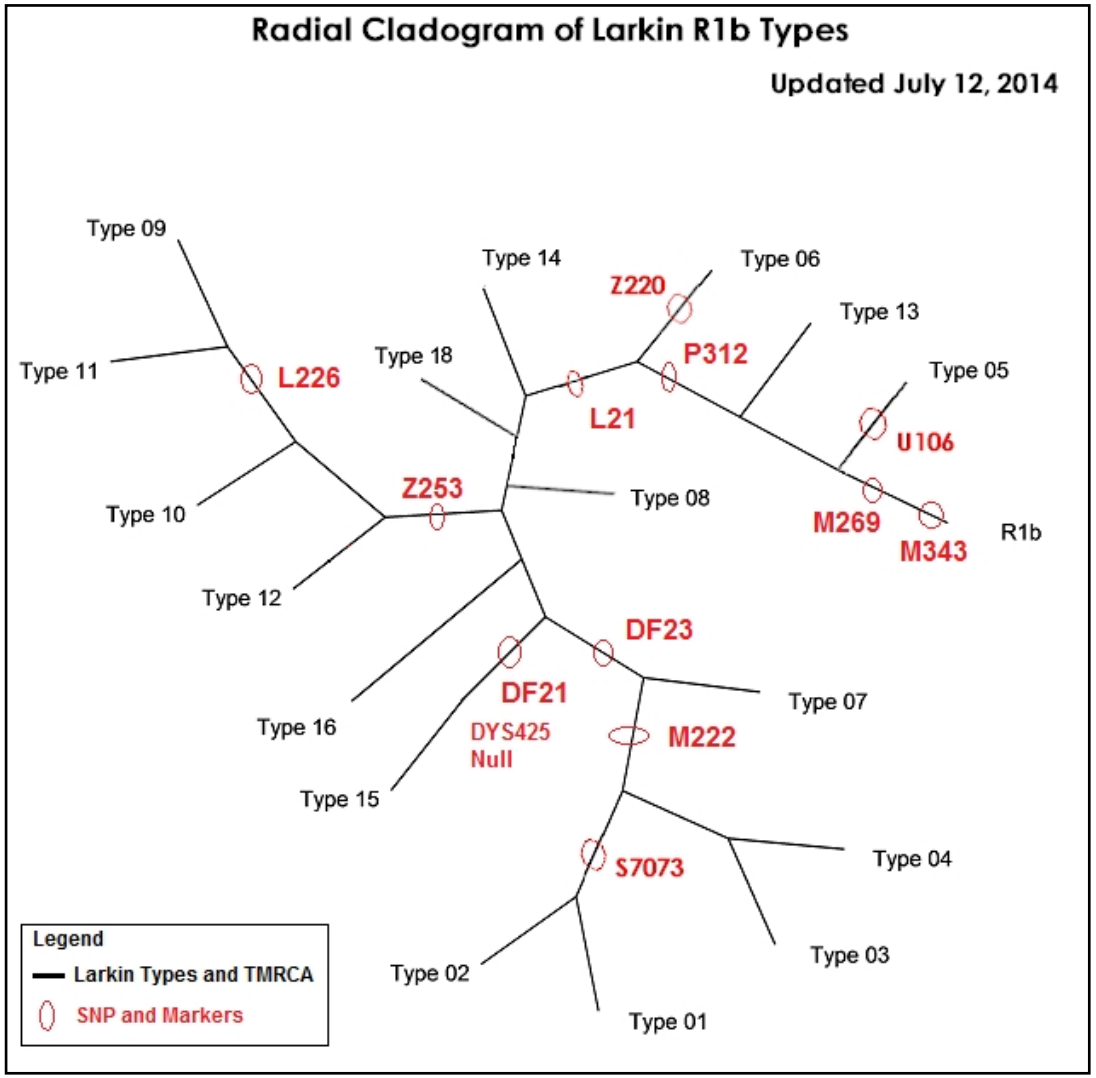
- Technically called *Microarray* testing
  - 800,000 individual SNP markers on one processing array chip (about 2 square inches in size)
- Made cost of testing a lot of SNPs at one time much more affordable
- Commercial Y-SNP Chip Products
  - Geno 2.0 Genographic Project
  - 23andMe
  - Ancestry DNA
  - Britain's DNA
  - Be aware that some companies strip the Y chromosome data out of their SNP Chip results.

SNP Chip Value for Surname Study

- Excellent for separating individuals from the root of the tree down
  - Great for population studies and paleogenetics
- Like STR testing, results sets generally homogenous
- Pricing: Very scalable for lab => decreasing prices
- Downside – no new discovery
  - The genealogically important SNP markers have to already be identified by scientists and researchers
  - Hard coded into the chip at time of design and manufacture.

Refine Branches with Major SNPs

- Major SNPs for surname groups emerge for STR-identified groups in 2014
- Branches of a few STRs had to be moved, but core groups remain.
  - Total of 11 distinguishing SNPs below M343 for Larkin DNA project members



Next-Gen DNA Testing

- aka *High Throughput*
  - Sequencing by Synthesis technique
- Breaking DNA down into many, small pieces (small read length) called *Shotgun Sequencing*
- Sequencing those pieces very quickly
- Making multiple runs so as to cover large chromosomal areas
- Assembling and interpreting those small reads with software
  - Computer technology makes this process more industrialized and scalable

Shotgun Sequencing

- Individual runs of small segments have some amount of read errors
- Individual errors overcome by making multiple runs with randomly overlapping reads for a given chromosome position.
- Testing Lab economics and policies determine the amount of coverage for a given genetic testing product.

  => Coverage Ratio

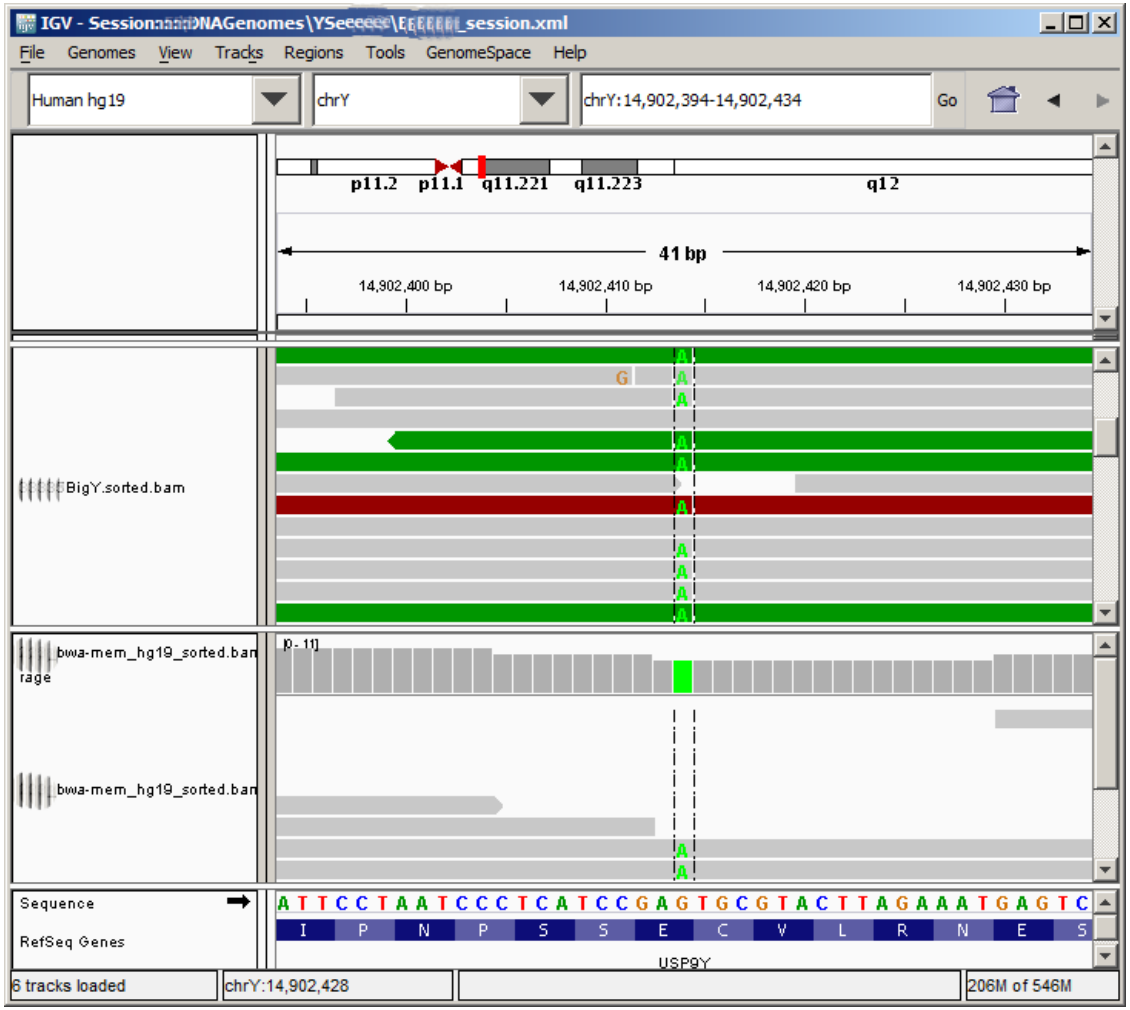
Shotgun Coverage Illustration

	Reads
Y-Chromosome Position:	0 10 20 30 12345678901234567890123456789 0
Shotgun Sequence 01	AGCATGCTGC----------
Shotgun Sequence 02	AGCATGCTGCAGTCAT----------
Shotgun Sequence 03	-------TGCAGTCATGCT----------
Shotgun Sequence 04	----------GTCATGCTTCTATGCAGTC
Shotgun Sequence 05	----------TCTATGCAGTC
Shotgun Sequence 06	----------TGCAGTC
Assembled Sequence	AGCATGCTGCAGTCATGCTTCTATGCAGTC

Summary of simplified illustration: 6 reads; 2X coverage

Shotgun Coverage - BAM

- The individual reads are stored in a binary format called a 'BAM' file.
  - Example from real BAM file from Next-Gen sequencing.
    - G->A mutation at Y-14902414 = M222



Next-Gen Products and Coverage

- Coverage Ratio Defines minimum number of times number of reads that align with each base.
  - e.g. Each base pair is observed with a value
  - 1000 Genomes Project was done with 4-5X coverage
- Next-Gen Coverage Ratios at current retail products

Current Next-Gen Y-DNA Genetic Genealogy Product Coverage and Pricing

Company & Product	# Chromosomes	Base Pair Targets (mbp)	NGS Coverage Ratio	Price (as of 4/28/2017)
Family Tree DNA (FTDNA) Big Y	1	20	55X with quality algorithm[2]	$ 575
Full Genomes Corp (FGS) Y Elite 2.1	1	16	50x at 250bp	$ 795
YSeq Whole Genome Test	25		15X + 10 unique SNP Sanger sequencing	$ 899
Full Genomes Corp (FGS) Whole Genome	25		30X 15X 10X	$ 1,250 $ 700 $ 610

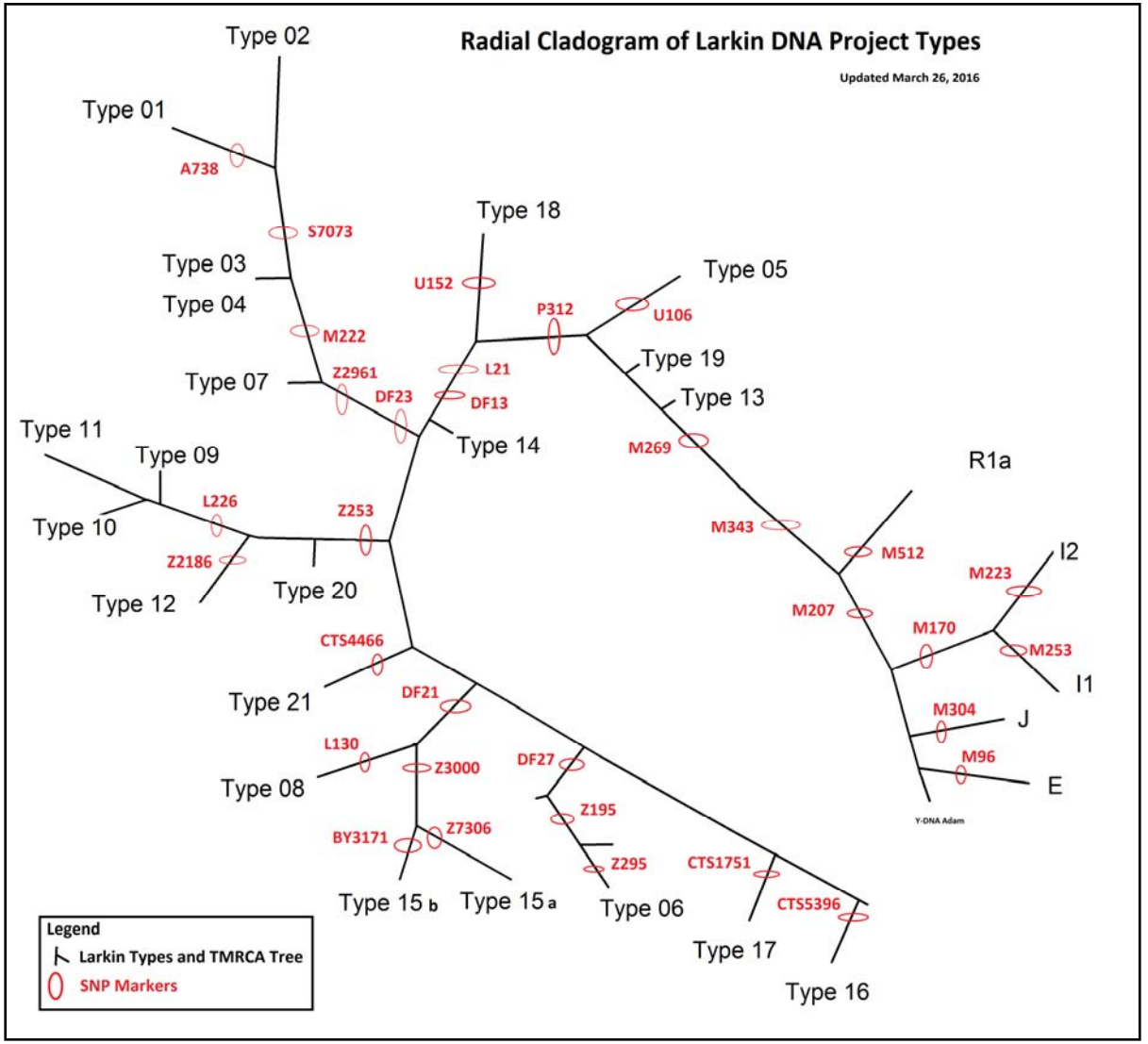
[1]Yang et al (2014) Application of Next-generation Sequencing Technology in Forensic Science

[2]FTDNA (2014) Introduction to the BigY White Paper, plus FAQ, please data file analysis. FTDNA Big Y requires preliminary order of Y-STR test as well.

Low Level SNPs Subgrouping

- For 2016, we have eleven (11) FTDNA *Big Y* Next-Gen results for men with Larkin Ancestry.

- There are at least 26 distinguishing SNPs below M343 for Larkin DNA project members.



Next-Gen Dilemmas for Genealogy

- Genealogically-meaningful SNP mutations can occur at many points on the Y-chromosome.
- 100-300 base pair read length for many Next-Gen instruments may give ambiguous results at low coverage.[1]
  - Analogy: a foot long hot dog and a 6 inch bun.
  - However, some evidence that Next-Gen at very high coverage may be more accurate allele measure than traditional capillary electrophoresis[2]

[1]Zavodna et al (2014) The Accuracy, Feasibility and Challenges of Sequencing Short Tandem Repeats Using Next-Generation Sequencing Platforms, PLOS One 9(12): e113862 DOI 10.1371/journal.pone.0113862

[2]Darby et al (2016), Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. DOI 10.1002/ece3.2221

Next-Gen Inconsistent Observations

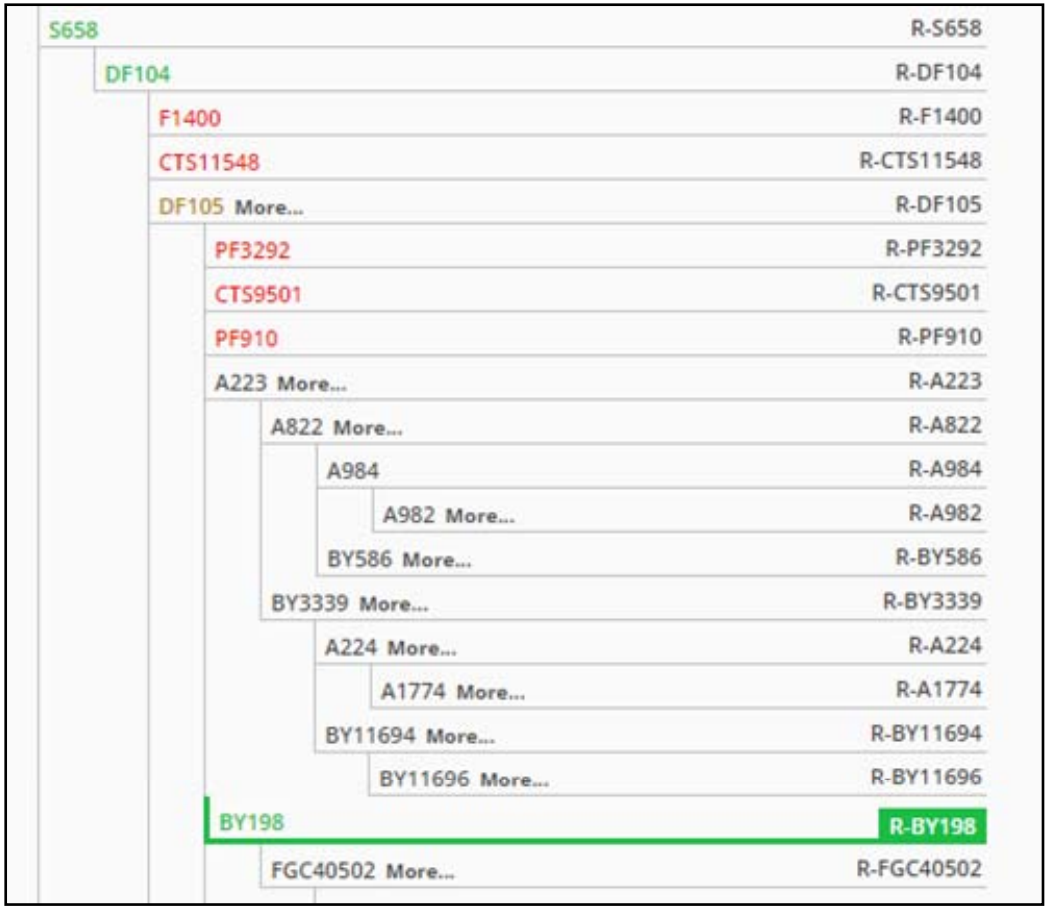
- Two Samples with same surname, from same part of Ireland, and several SNP matches below M-222.
  - Distinction between positive, negative, and NOT OBSERVED
    - Top example not observed for FGC4087
    - Lower example not observed for A738
    - 4 of 8 low level markers NOT OBSERVED (50%) in this real example
  - Cannot resolve the phylogeny due to non-homogenous datasets - inconsistent observations of Next-Gen sequencing
    - Resolution requires ad hoc Sanger sequencing for individual SNP candidates.

LastName	Ancestral Geography	Y-14902414 [10-] *M222 Page84 PAGES00084 rs20321*	Y-26078887 [10-] *S7073 FGC462*	Y-22540855 [10-] *S660 DF109 FGC4101 Y2845*	Y-14624294 [10-] *PF682 S569 rs9786370*	Y-17303280 [10-] *A738 BY198*	Y-18028717 [10-] *FGC4087 Y3454*	Y-8157356 [10-] *S7072 FGC449 Y2596*	Y-13686261 [10-]
Larkin	Ireland, Tipperary, , Nenagh	1	1	1	1	1	-	1	1
Larkin	Ireland, Galway, Srahaun	1	1	-	0	-	1	1	-

Example of Markers Missed with Next-Gen

In Haplotree at right, SNP markers in black and brown ink were not observed

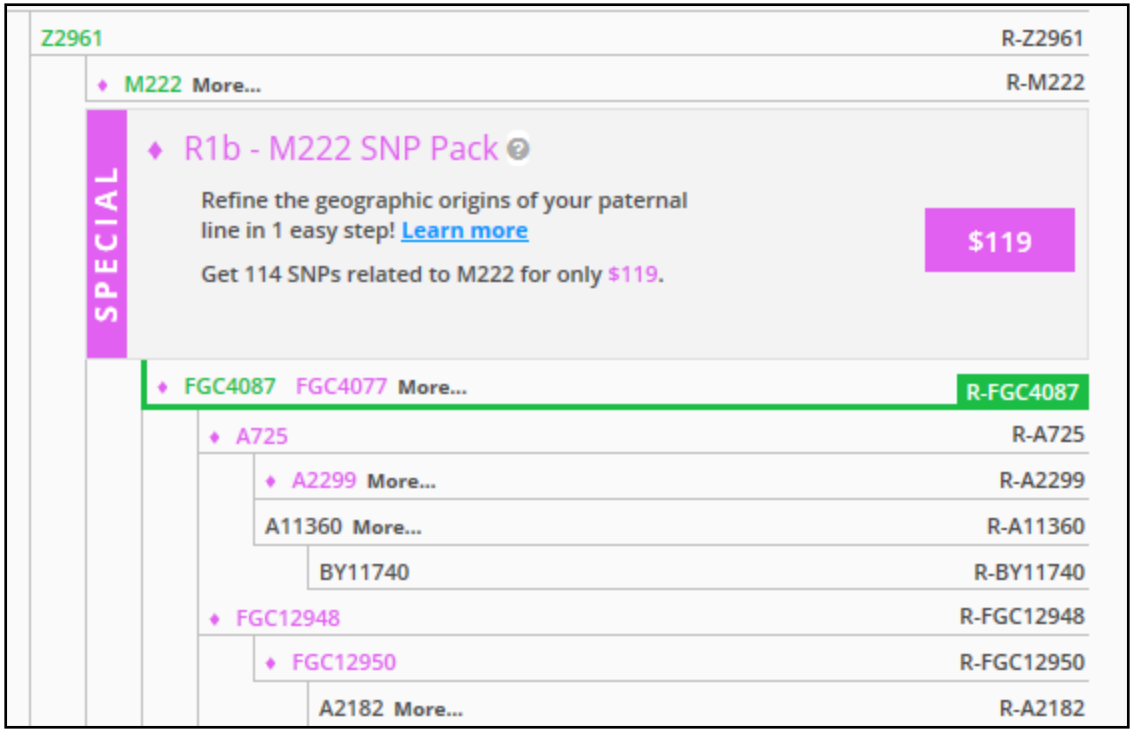
=> 11/20 Not Observed



SNP Pack Technology

- SNP Packs are groups of lab-designed Sanger sequencing probes of about 100 markers per pack.

- Designed around specific phylogenetic subclade
- No new discovery.
  - Depend on the SNPs already being identified.
- Conflicting nomenclature across labs
  - Genetic Homeland DNA Marker Index



Topics

- Biology of DNA for Genealogy
- Y-DNA Testing Technologies
- Price – Value Comparison

Best Tree: SNP + STR

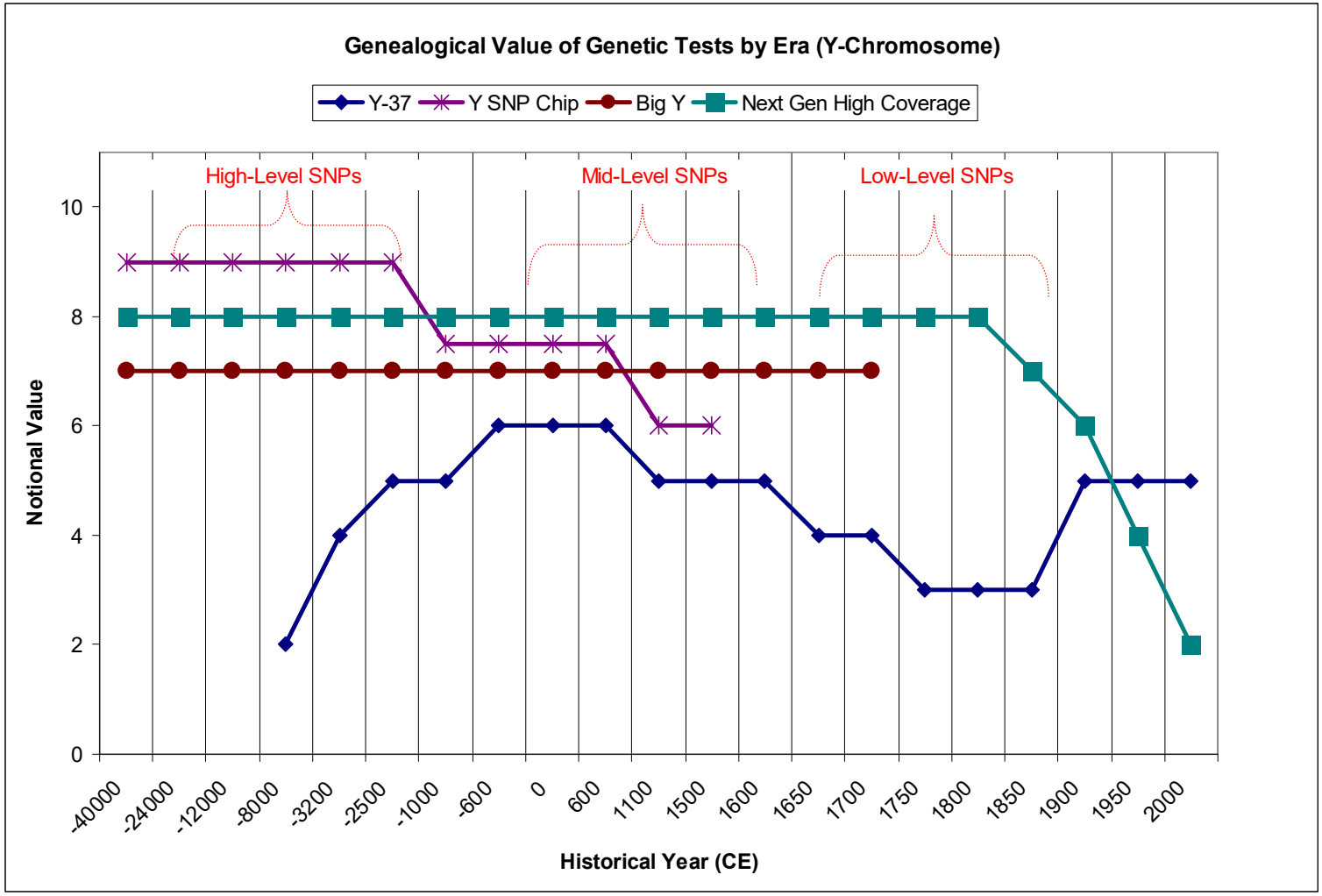
- Joint application of SNPs and STRs can provide biological insight not available from investigation of either marker type in isolation.[1]
  - Because of population bottlenecks and explosion in last 2500 years, best measure of relatedness is to confirm same branch with mid-level SNP and then compare STR allele correlation.

[1]*Payseur and Cutter (2006) Integrating patterns of polymorphisms at SNPs and STRs*

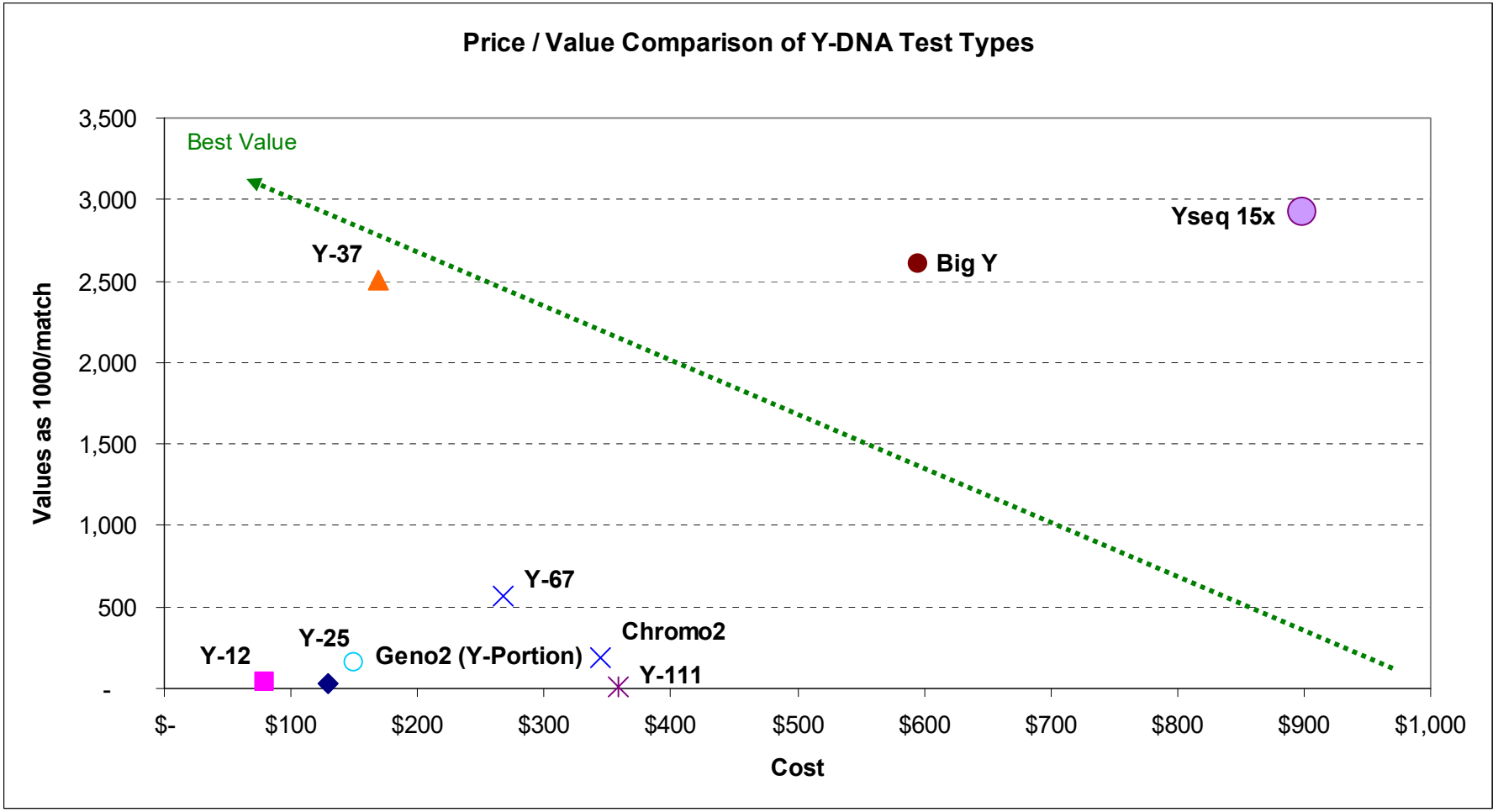
FGS BAM vs STRs

- full genomic sequencing (FGS) of all 26 chromosomes is now commercially available for less than $ 900.
- In theory, the ‘BAM’ file of your FGS results could be decoded to show your STR values.
  - Therefore, why would you order a traditional Y-STR test?

Value of Technologies by SNP Era



Price-Value Comparison



Conclusion

- Best genetic genealogy entry point depends on budget
- Under $ 200
  - a Y-37 STR or equivalent test is recommended.
    - Keeps number of markers and interpretation more straight forward
    - From there, you are in a good position to evaluate matches, applicable SNPs, and get advice from those with more experience in your haplogroup or surname.

Testing Strategy

- Define Objective
  - e.g. to see if two persons with same surname share a common ancestor with that surname
    - Your Line
    - Other line
- Start with Y-STR 37
  - If your line matches the other line
    - Consider upgrading to more STR markers
  - See how close your matches are to the modal for your lowest level SNP

When Will STR Testing be Obsolete?

- Next-Gen (2nd generation) high coverage and read length evolve at lower cost
- 3rd generation sequencing, SMRT becomes cost effective
  - Single-molecule real-time sequencing
  - Much longer read length: should measure STRs directly
  - Better de novo SNP detection
- Phylogenetic Trees and Database become more complete
  - Majority of pedigree and DNA databases remain STR-based.

Where to get Y-STR Testing

- Family Tree DNA
  - aka FTDNA
  - DNA Projects and Administrators
  - Large Database
  - De Facto Standard
- YSeq
  - Can match FTDNA 37 markers with clever ordering
    - YSeq Alpha Panel + Beta Panel + DYS442
- What about?
  - Not Ancestry.com (cancelled STR program 2014)
  - Not 23andMe (never had STR program)