DNA for Genealogists (Intro)

By Robert Casey

September 14, 2013

http://www.rcasey.net/present

Intro to DNA for Genealogists

- Type of tests used
- How common tests work
- Companies that offer tests
- How does DNA get analyzed
- Autosomal – best fit for post-1850 research
- Y-DNA – best fit for pre-1850 research
- Books to get you up to speed
- Forums, DNA projects, etc.
- Costs of testing
- Future Trends in DNA testing

Four major types of tests available

- Y-DNA – uses both Y-STR tests and Y-SNP tests
  - Limited to all male lines only
  - 100 to 1,000 years (Y-STRs)
  - 1,000 to 10,000+ years (Y-SNPs) – some 500 years
- Autosomal (atDNA) – 50 % change every generation
  - Tests all ancestral lines
  - Limited to 150 to 200 years
- Mitochondrial (mtDNA)
  - Limited genealogical applications
  - Only 16K base pairs vs. 58M base pairs for Y-DNA
- Geographical origin tests
  - Not very accurate – goes back 10,000 years ?
  - Continental accuracy – expectations are too high

How mtDNA works

- mtDNA is not part of the nucleus (each cell has 100 - 1,000 mtDNA strands available – used for food)
- mtDNA passes only via all female lines
- mtDNA is a very small DNA structure with only 16,000 base pairs (vs. 58,000,000 bp for Y-DNA)
- No fast moving markers like Y-STRs available
- Only deep ancestral information available – 1,000 years or more
- Limited future growth of discovery of new mutations due to the 16,000 base pair limitation
- Minimal genealogical usage

How Autosomal (atDNA) works

- Covers all ancestral lines but limited to 150 to 200 years (reliable for only 4 or 5 generations)
- Each generation has 50 % change resulting in shorter and fewer common segments
- Most people order multiple tests to assign matching segments to various ancestral lines
- Works great for recent adoptions, breaking recent brick walls or just starting out with genealogy
- Not reliable beyond 200 to 250 years where many brick walls exist
- Random matches can be found at six and seven generations – sometimes even at eight generations

How Autosomal works (Recombination at work)

- 1 gen - Parents – 50 % - 1935 (all)
- 2 gen - Grandparents – 25 % - 1910 (all)
- 3 gen - 1G gparents – 12.5 % - 1885 (all)
- 4 gen - 2G gparents – 6.25 % - 1860 (90 %)
- 5 gen - 3G gparents – 3.12 % - 1835 (80 %)
- 6 gen - 4G gparents – 1.55 % - 1810 (20 %)
- 7 gen - 5G gparents – 0.77 % - 1785 (5 %)

GEDMATCH Database - atDNA

- Attempts to merge three major companies massive atDNA databases (Ancestry.com in early stages)
- Gives a combined view of a database of all companies but only small percentage testers upload to GEDMATCH
- Unfortunately, it is a volunteer download from each company, so coverage will be inconsistent
- Also has leading edge analysis tools as well
- Public repository for atDNA test results - similar to Y-Search as FTDNA's public repository for YSTR results
- Many take atDNA tests at both 23andme and FTDNA to get full access to both databases
- Many upload 23andme results to FTDNA database ($69)
- Most 23andme donors are only interested in medical results

How Y-STR works

- Only works with all male lines
- Relatively faster mutating DNA that matches genealogical time frame (100 to 1,000 years)
- Great for answering yes / no / maybe relationships
- Takes a lot of submissions to build a genetic cluster and determine relationships
- Generates clusters of related lines but rarely shows how lines are connected
- Overlapping haplotypes (convergence) sometimes makes it impossible to assign to one genetic cluster
- Y-SNPs greatly complement some of the shortfalls of the Y-STRs by themselves (but in early stages)

How Y-SNPs work

- Defines genetic branches between 500 and 5,000+ years
- With 58,000,000 base pairs for possible Y-SNP testing, only a few percent of Y-SNPs have been discovered / analyzed to date
- Dozens of new Y-SNPs being discovered every week
- Y-SNPs create father / son relationships that reveal exact genealogical relationships between Y-SNPs
- YSTR mutations between Y-SNP ancestor and genealogical cluster define YSTR fingerprints which show common mutations
- Some unrelated Y-STRs submissions have such common marker values that close matches are not related (Y-SNPs help break up these clusters of unrelated Y-STR matching submissions)
- Y-SNPs has grown from 2,000 to 28,000 and future new Y-SNPs should create over 100,000 branches in the next few years

Five primary testing companies

- Family Tree DNA provides best overall value with most offerings, largest database and leading edge testing Y-DNA testing
- 23andme has strong autosomal test and useful Y-SNP test but lacks critical Y-STR testing and advanced Y-SNP testing
- Ancestry.com has entry level Y-STR tests but has no Y-SNP testing and FTDNA is becoming the primary player for Y-STR testing
- Ancestry.com recently added atDNA but has no segment matching tool as does 23andme and FTDNA (all require a lot of manual analysis)
- Family Tree DNA offers unbelievable Y-SNP testing that will eventually become the primary tool for future genealogical research
- All three companies offer robust mtDNA test but only FTDNA offers full mtDNA test (do not recommend testing of mtDNA from any company)
- National Geographic and FTDNA recently announced the NatGeo 2.0 test includes static test of a massive 12,000 Y-SNPs, extensive mtDNA and geographical origin testing (all geographical tests are questionable)
- Full Genomes Corp is recent addition with a full Y-chromosome test and 24,000 known YSNPs, 400 YSTRs and full mtDNA plus the capability to discover 25 to 50 new YSNP mutations per test ($1,500 test)

How do Y-STRs work

- Only found on Y-DNA chromosome
- Y-STRs are patterns that repeat many times and the number of repeats vary generation to generation
- Testing companies scan the Y-DNA until they find the landmark indicating they have arrived at the Y-STR
- From that landmark, they then know how to locate the repeating patterns and count the number of repeats (Short Tandem Repeats)
- The Y-STR values (numbers of repeats) vary over time allowing genealogists to track ancestors

How do Y-SNPs work

- Only found on Y chromosome – Single Nucleotide Polymorphism (pronounced Y Snips – also called haplogroups)
- Most are one time mutations (but many mutate multiple times or later discovered to be in unstable areas – still learning mode)
- Discovers branches from 500 to 5,000+ years
- Unlike Y-STRs, Y-SNPs have a very hierarchical relationships (clearly define father / son relationships)
- Creates a true genealogical like descendant tree (haplotree)
- Once you find your most recent Y-SNP (usually 500 to 2,000 years old) Y-STRs complement Y-SNPs for more recent mutations
- Recent explosion from 2,000 to 28,000 Y-SNPs with as many as 100,000 or more to be discovered in the near future

How Autosomal tests works

- Covers all ancestral lines – but reliable to to only post 1850 (but does catch some lines back to the 1750s)
- Comes from all chromosomes except for Y-DNA
- Recombines 50 % from each parent every generation
- Each recombination results in shorter segments of common DNA that is partially passed to children
- Segments get shorter and shorter every generation until no longer reliable for identification purposes
- Multiple tests of close relatives required to sort out which segments belong to which ancestral lines
- The total amount of longer segments can estimate the degree of relationship (3rd cousin, once removed)

How mtDNA works

- Mother passes to daughter over many generations (also passes to sons but sons can not pass it on)
- Over time, mutations occur that allows us to build a all female descendant tree (haplotree)
- Can answer questions about no / maybe relationships
- If you do not share a common mutation that is 3,000 years old, you obviously are not related in the last 300 years
- If you do share a common mutation that is 3,000 years old, provides some support for connection at 300 years (but this is not very reliable)
- The more rare the mutation and the more recent the mutation, the more support there is for a connection

Books to get up to speed with

- Trace Your Roots with DNA by Megan Smolenyak & Ann Turner, 2004 solid book – not real deep
- DNA & Genealogy by Colleen Fitzpatrick & Andrew Yeiser, 2005 solid book – little more depth
- DNA & Social Networking: A Guide to Genealogy by Debbie Kennett, 2012 - updated (includes autosomal)
- Family History in the Genes by Chris Pomeroy, 2007 – by far the best on Y-DNA – more in depth & covers more complex issues
- DNA and Family History by Chris Pomeroy & Steve Jones, 2004 – both versions worth getting
- Go to Amazon.com and search DNA & genealogy

Forums, DNA Projects

- Unbelievable good support found in DNA forums & projects
- Look at Surname projects – many are excellent
- Join Surname, Geographic, Haplogroup, clan and ethnic projects – key to having your DNA analyzed by experts
- Skill levels of forums and projects vary a lot, be prepared for minimal support from some – specially less common surnames
- Expect minimal assistance from testing companies
- 95 % of the analysis is done by amateur researchers and some are extremely skilled (as much as testing companies)
- Highest skills found in haplogroup projects (most are biased towards anthropological research but trend is changing)
- Be respectful of volunteers who help – sugar works better than vinegar (demanding) – you should also test as recommended

Cost of Testing (Retail - FTDNA only)

- Always join project before ordering to get the project discounts & catch regular sales
- Family Finder (autosomal) - $99 (reduced)
- Nat Geo 2.0 (Y-SNPs, mtDNA, geographic origins)- $199 – order from National Geographic extremely robust YSNP test
- Y-STR - $169 (37), $268 (67) & $359 (111)
- Full mtDNA - $199, partial mtDNA - $49
- Special order Y-SNPs - $39 each
- Walk The Y - $950 (recently withdrawn)

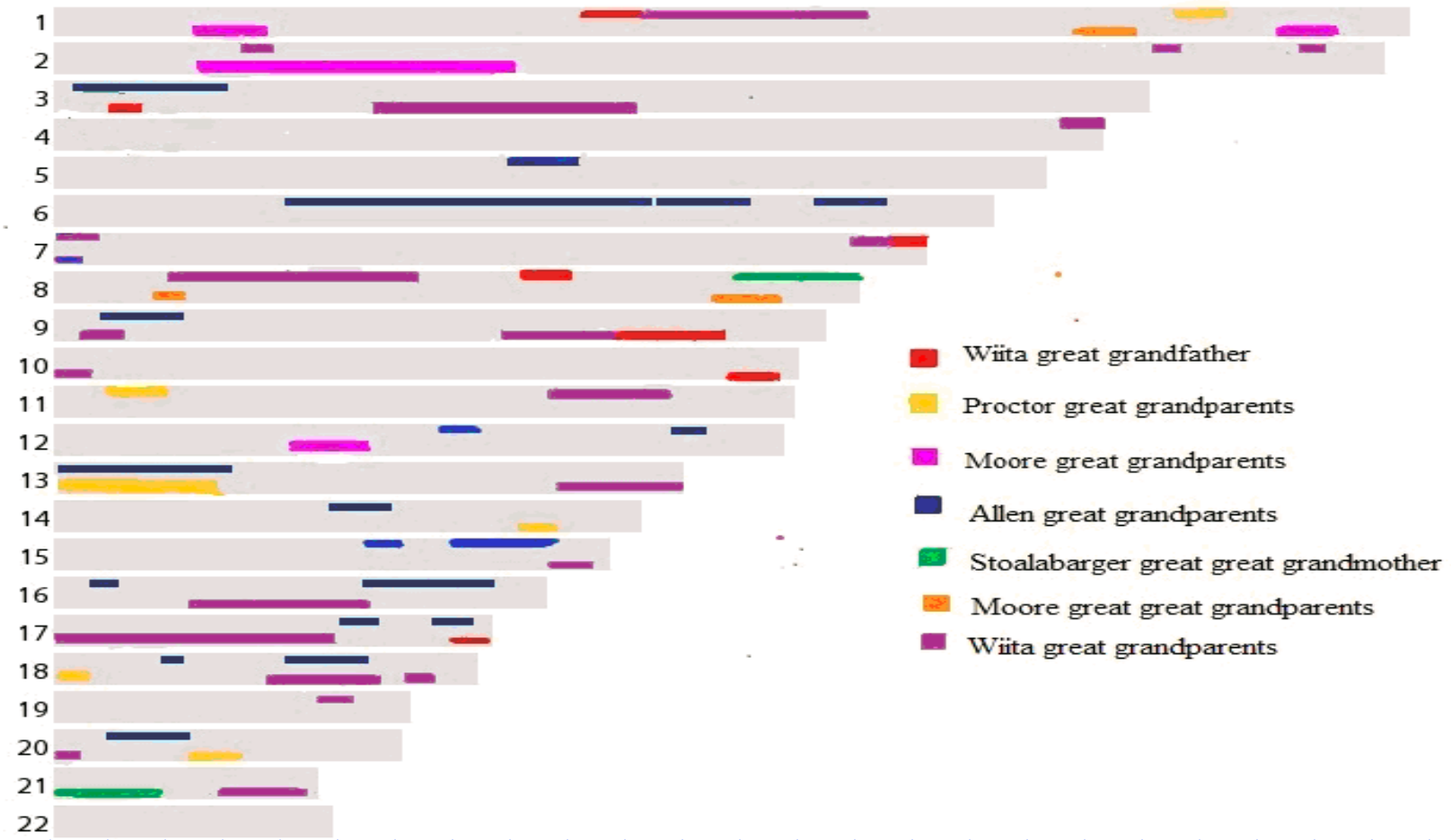
Cost of testing – other companies

- 23andme – “one size fits all” test - $99 for health, atDNA, geographical origins and limited Y-SNPs (no Y-STRs)
- 23andme - good starting point but many end up migrating to FTDNA for more Y-DNA testing – health markers are unique
- Ancestry.com – Y-STR $149 (33) & $179 (46) – No Y-SNPs
- Ancestry.com – 46 marker is good starting test but many migrate to FTDNA for more Y-STR & Y-SNP testing
- Ancestry.com – Autosomal $99 to existing customers (no segment matching tools available)
- Full Genomes Corp. - $1,499 – full Y-Chromosome scan, 28,000 Y-SNPs, 20-30 new Y-SNPs, 400 Y-STRs and mtDNA

Future trends – break for Q&A

- Y-SNP testing will become the most important test
- NatGeo 2.0 increased static Y-SNP test from 2,000 to 12,000 Y-SNPs (a lot to absorb)
- Potential for 10,000s of genealogical Y-SNPs & estimates state that 100,000 or more are possible
- Autosomal test good for post-1850 brick walls but requires a lot of tests to triangulate
- Full mtDNA is only discovery test for all female line (only available from FTDNA) – leading edge anthropological test (minimal genealogical usage)

Sample atDNA comparison



War stories – atDNA for Casey

- Common Casey ancestor believed to be born around 1700 (way too early for atDNA)
- However, common segments found between all 12 Casey related tests – but not the same segments
- It is very frustrating that each pair has different segments (not as expected when they descend from the same ancestor)
- However, does imply that all are probably part of the Casey South Carolina genetic cluster (half of atDNA testers already found to be related via Y-DNA testing)
- Half of the atDNA submissions did not know any male Casey ancestor – only the maiden name of Casey female ancestor
- With no Casey males (or Casey descendant males to test), no further Y-DNA research could be conducted
- Most distant Casey lines showed only modest interest in Casey line since it was not one of their primary lines

War stories - YDNA provides answers

- Two different John Pace lines claimed the same father, Richard
- Each had different wives, children born in the same time period and resided in different areas (proven to be two different men)
- Both claimed Richard Pace who was proven back to Jamestown
- At least 20 books claimed the same man (including my book)
- The Jamestown line originated from a known part of London

- DNA proved both lines could not be closely related
- Two random Pace submissions solved the mystery
- One submission that still lived in London where the Jamestown line resided - traced back for five generations within one mile
- One submission from Canada traced back to rural England with supporting parish records one John Pace and matched the second Pace line

War stories - YDNA provides answers

- Two men named Jordan Brooks resided in common counties and neighboring counties - at least five different counties (GA & AL)
- There were only two Jordan Brooks individuals in the entire south prior to 1830 – very rare combination of given name and surname
- Both lines had very similar given names (both Methodists)
- Researchers from both lines borrowed from each other due to similar residences, similar given names and common surname
- Many publications actually turned speculation into firm connections on the Internet databases
- Male descendants were located from each line and both submitted DNA for comparison
- FTDNA relatedness tool revealed there is less than 1 of 10,000 chance of lines being related in the last 600 years (proved to be not related)
- Genetically proved these lines are not related as once believed
- However, the line that was not my line closely matched a third different line which was not researched previously by anyone

War stories - YDNA provides answers

- There are believed to be around 40 different Casey lines residing in four neighboring SC counties from 1760s to 1820s
- Many years of research has been unable to make much progress in tying these Casey lines together
- DNA has proven that about 12 Casey lines are very closely related (and DNA contains extremely unique marker values – genetically isolated)
- DNA has proven that one SC Casey line is not closely related
- DNA has allowed the most probable DNA Descendancy Chart (connections of these lines based on DNA information)
- Around one half of the submissions are part one early branch and the remaining half are part of second branch (pretty rare scenario)
- Author of 600 page Hanvey book found out that his Hanvey line is actually a Casey line genetically and did not match Hanvey lines as he should have (Casey boy was probably informally adopted by Hanveys)

War stories - YDNA provides answers

- Was contacted by Butler submission that had a known NPE event in the 1850s
- This Butler line was an out of wedlock birth of an unknown male which would normally be very difficult to make any progress
- DNA showed a match with a Brooks cluster (I am co-admin for the Brooks surname as well)
- This Brooks cluster has many submissions and is very actively researched and included possible NPE lines with similar DNA
- The Butler fingerprint closely aligned with a Bradberry NPE line
- A Butler sister married a Bradberry and a Bradberry in 1860 census was found just a few households away was one of the Bradberry NPEs that had been tested in the Brooks cluster
- The conclusion is that the father of Butler boy was very likely a Bradberry – but it could be one of many Bradberry males

War stories - YDNA provides answers

- My mother's line, Brooks, has many genealogical anomalies where the two oldest sons can not be confirmed as they were not included is extensive probate records of their speculated father
- However, these two unproven sons were found living in the same household via personal property tax lists and the speculated father signed the marriage bond for one son plus major common migrations
- There were many Wade families residing in the same area and unproven family history stated our oldest proven Brooks ancestor married Brambly Wade and this couple also named a son named Wade
- DNA then showed that the very unique DNA for Brooks submissions was a common DNA pattern for many Wade submissions in the same area in the mid 1700s but did not match the DNA of a Brooks uncle
- The conclusion is that this speculated father may have married a woman who was previously married to a Wade and that the two oldest Wade sons were probably informally adopted – so family tradition appears to be correct and solid primary documentation is misleading

War stories - YSNPs provide new answers

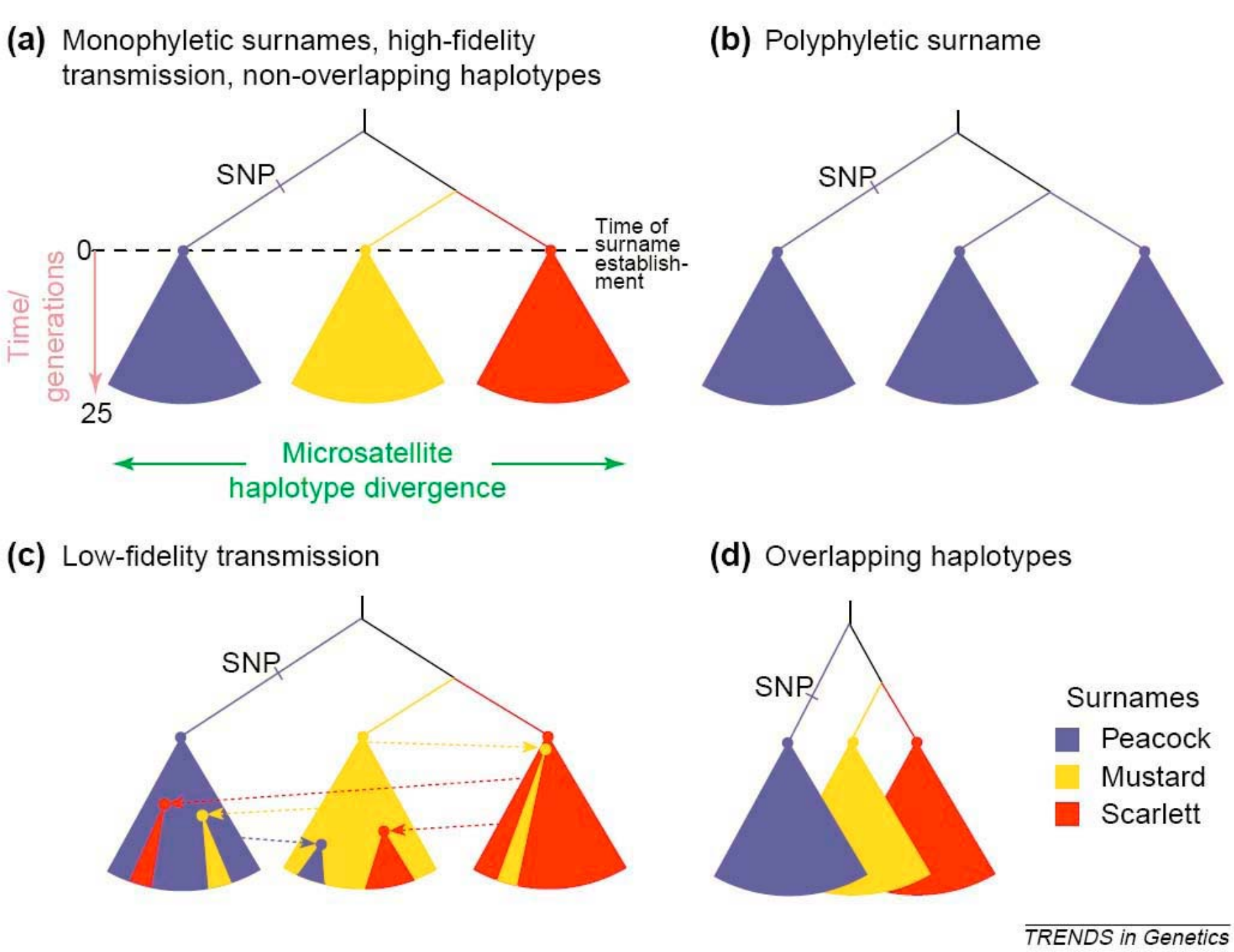
- A recent “Walk The Y” test discovered three new Y-SNPs under DF5. Unlike many Y-SNPs, submissions have mixture of positive and negative test results for these three less broad Y-SNPs
- L627 is the oldest Y-SNP and includes both Cooper and Reynolds surnames. L626 is the middle Y-SNP and is only has Reynolds surname submissions. L625 is the youngest Y-SNP and splits the Reynolds surname cluster into two branches (appears to be genealogical SNP)
- Conclusion - L625 happened after the first usage of the Reynolds surname and currently is a true genealogical only Y-SNP
- Only four submissions have been tested positive to date for these newly discovered Y-SNPs, so the scope of these Y-SNPs could change with more testing
- Very few testing candidates have been identified to date, therefore the scope of these Y-SNPs appear to be quite limited compared to most Y-SNPs

Questions & Answers

- It takes a while to get up to speed. Genetic DNA takes as many skills as traditional genealogy
- Too high expectations by many
- Just get started – be sure to have well defined goals so you can later assess if you met your goals
- A lot of willing volunteers will assist – make it a two way interchange (test what they recommend)
- Future testing will provide an unbelievable amount of new information (specially Y-STR and Y-SNP testing)
- Before we talk about a couple of advanced topics or future trends, it is time for questions & answers

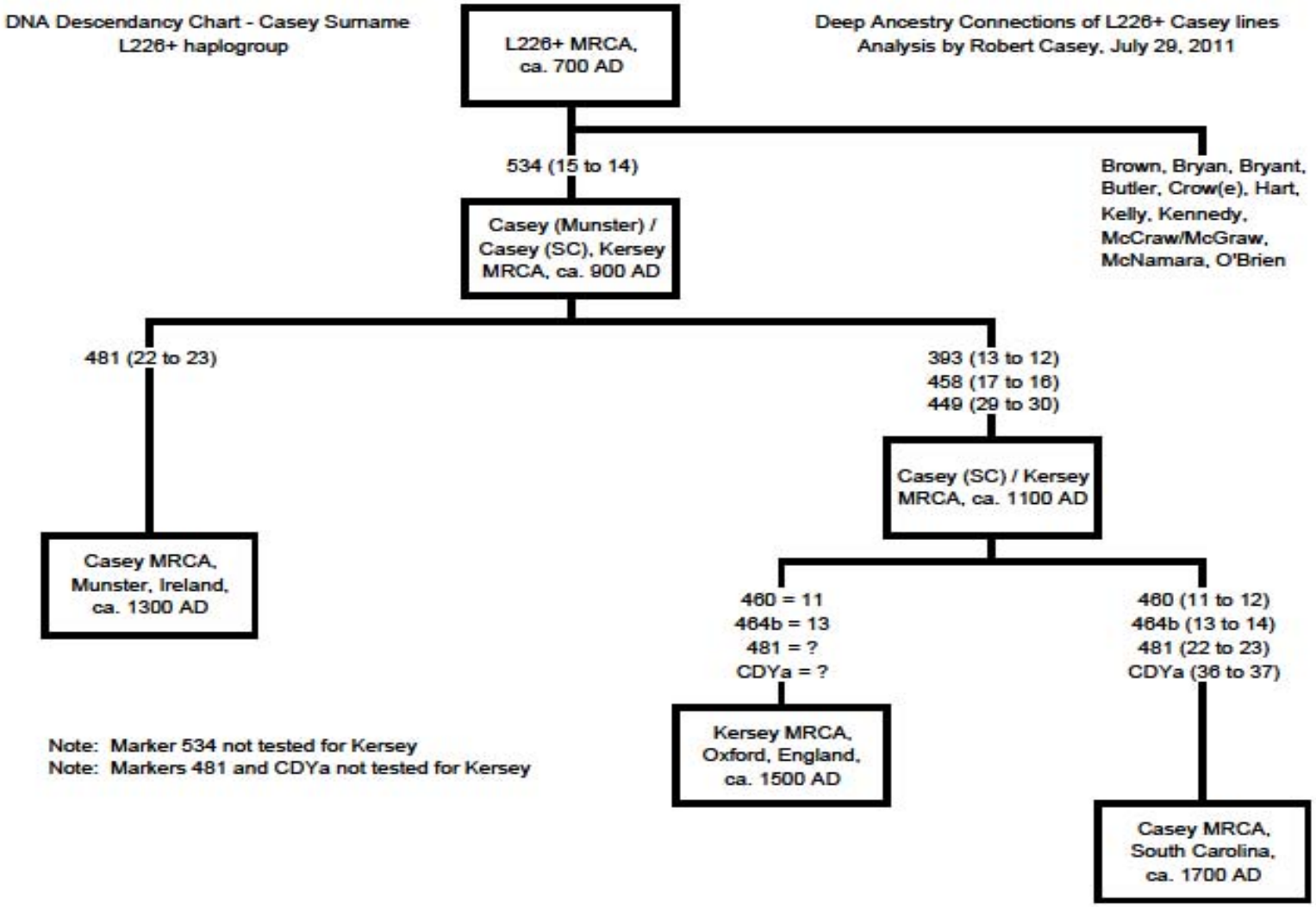
Overlapping Haplotypes (Convergence)

- This issue is not covered by books or many web sites
- Some haplotypes contain such common DNA marker values that even very close matches may not related
- As much of 5 to 10 percent of all submissions fall into the “overlapping haplotype” scenario (also called convergence)
- Genetic distance (the number of mutations that are different) is not always reliable by itself as testing companies assume
- You want to categorize non-surname matches into two categories: “overlapping haplotypes” or “possible NPEs”
- Overlapping haplotypes need to be filtered out by Y-SNP tests
- NPEs can be a new gold mine of genealogical treasures
- There are methodologies for determining the category (too advanced for this session)
- http://www.rcasey.net/DNA/Casey/Sources/Overlapping_Haplotypes_Jobling_2000.pdf


(a) Monophyletic surnames, high-fidelity transmission, non-overlapping haplotypes
SNP
0
Time of surname establish-ment
Time/ generations
25
Microsatellite haplotype divergence
(b) Polyphyletic surname
SNP
(c) Low-fidelity transmission
SNP
(d) Overlapping haplotypes
SNP
Surnames
Peacock
Mustard
Scarlett
TRENDS in Genetics

Fingerprints are key

- Most analysis of Y-STRs depends too much only on genetic distance (number mutations that differ)
- The common mutations are a much better indicator of relatedness
- Determine the haplotype of your Y-SNP and determine the fingerprint of your genetic cluster (the mutations between the Y-SNP and your cluster)
- Combination of Y-SNP, fingerprint matches, genetic distance and surname is a power combination of information that must be used in any analysis


DNA Descendancy Chart - Casey Surname
L226+ haplogroup
L226+ MRCA, ca. 700 AD
Deep Ancestry Connections of L226+ Casey lines
Analysis by Robert Casey, July 29, 2011
534 (15 to 14)
Brown, Bryan, Bryant, Butler, Crow(e), Hart, Kelly, Kennedy, McCraw/McGraw, McNamara, O'Brien
Casey (Munster) / Casey (SC), Kersey MRCA, ca. 900 AD
481 (22 to 23)
393 (13 to 12)
458 (17 to 16)
449 (29 to 30)
Casey (SC) / Kersey MRCA, ca. 1100 AD
Casey MRCA, Munster, Ireland, ca. 1300 AD
460 = 11
464b = 13
481 = ?
CDYa = ?
460 (11 to 12)
464b (13 to 14)
481 (22 to 23)
CDYa (36 to 37)
Kersey MRCA, Oxford, England, ca. 1500 AD
Casey MRCA, South Carolina, ca. 1700 AD
Note: Marker 534 not tested for Kersey
Note: Markers 481 and CDYa not tested for Kersey

The future

- The costs of testing of the full genome will be under $1,000 in the next few years – so you never have test any donor again just analyze
- The amount of useful data will increase by 1,000,000 fold !
- atDNA is currently under 1,000,000 base pairs – it could be extended to 10M or 100M base pairs – but the usefulness of the information exponentially decreases
- mtDNA is only 16,000 base pairs – already being analyze since it is such a small DNA strand (no major increase – just analysis of same base pairs)
- Y-STRs are estimated to between 400 and 500 useful Y-STRs
- You have to double the number to have an impact or use faster mutating markers which require more submissions to analyze
- Y-SNPs – FTNDA has only around 2,000 useful Y-SNPs in the haplotree (if you ignore duplicate SNPs) – now testing for 28,000 has begun recently
- New NatGeo 2.0 tests 12,000 Y-SNPs – probably doubling useful Y-SNPs for western European research (more being discovered daily)
- It is believe that useful Y-SNP should exceed 100,000 when it becomes economical feasible to scan the entire Y-DNA strand
- Full Genome recently offered 28,000 Y-SNPs for analysis

Y-SNP analysis is the future

- Every surname cluster should get many Y-SNPs that will create many branches within surname clusters
- The Y-SNPs have father-son relationships vs. Y-STRs which are only clusters of related submissions – plus overlap of clusters is common
- Between all combinations of Y-STRs and Y-SNPs, most living individuals as well as most deceased ancestors will have unique haplotypes assigned
- It will take several years to establish the connections between the thousands of Y-SNPs and most research is done by fellow researchers
- Genealogical Y-SNPs are already being discovered with only around 2,000 useful Y-SNPs, 100,000 Y-SNPs will produce thousands more
- Bennett Greenspan (CEO of FTDNA) stated that Y-SNPs will be the branches of our descendant tree of mankind and the Y-STRs will be leaves on the branches
- FTDNA recently announced withdrawal of “Walk the Y” test which will be soon replaced with a “much better” test for discovering new YSNPs
- NatGeo 2.0 provides a static 12,000 Y-SNP test starting in November, 2011 and Full Genomes recently offered 28,000 Y-SNP test in September, 2013 (with discovery of 20 or 30 new Y-SNPs per test)

IT costs will drive testing costs

- Almost all people really hate this chart and this chart will provide an extreme challenge to the genetic testing community
- Testing costs will continue to decline between 80 % and 90 % per year – most companies will just offer more data vs. lower costs
- Eventually, you will be able to order a full genome test from China for a fraction of cost of existing testing companies – but you only get raw data
- With amount of data required for analysis increasing by ten times each year and the testing costs declining by several times each year is an issue
- Eventually, you will have to pay for online analysis separate from testing costs as existing companies start losing business to Chinese companies
- 23andme had the correct idea – but the wrong strategy to go with it. They should withdrawn only access for those that did not pay vs. removing data with no contract which was a huge marketing disaster
- Really advanced analysis tools are possible, but nobody wants to make major investments and not get paid for it – so we deserve the tools that we get which are marginal at best
- As the data gets so huge in the future, today's manual methods will not scale to databases that large and tools will be required for analysis