DNA for Genealogists (Intro)

By Robert Casey

August 11, 2012

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 up to speed
- Forums, DNA projects, etc.
- Costs of testing
- Future Trends in DNA testing

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)
  - 500 to 50,000+ years (Y-SNPs)
- Autosomal – 50 % change every generation
  - Tests all ancestral lines
  - Limited to 25 to 150 years
- Mitochondrial (mtDNA)
  - Limited to all female lines only
  - 1,000 to 50,000+ years
  - No fast mutating STRs to complement

How mtDNA works

- mtDNA is the only DNA that is not part of the nucleus (each cell has 100 -1,000 mtDNA strands available)
- 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 mutation due to the 16,000 base pair limitation
- Not recommended for testing

How Autosomal works

- Covers all ancestral lines but is limited to 100 to 150 years in accuracy (reliable for 4 or 5 generations)
- Each generation has 50 % change resulting in shorter and fewer common segments
- Requires 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 years where most brick walls exist
- A few random ancestral matches can be found at six and seven generations

How Autosomal works (Recombination at work)

- Parents – 50 % - 1950 (all)
- Grandparents – 25 % 1925 (all)
- Great grandparents – 12.5 % - 1900 (all)
- 2G grandparents – 6.3 % - 1870 (90 %)
- 3G grandparents – 3.1 % - 1850 (80 %)
- 4G grandparents – 1.4 % - 1825 (20 %)
- 5G grandparents – 0.8 % - 1800 (5 %)

GEDMATCH Database - atDNA

- Attempts to merge two major companies massive atDNA databases
- Gives a combined view database of both FTDNA and 23andme submissions
- Ancestry.com does not allow access to raw data – do not order their product
- Unfortunately, it is a volunteer download from each company, so coverage will be inconsistent
- Also has leading edge analysis tools as well
- Similar to Y-Search as public repository for atDNA tests
- Many test twice to get full access to both databases

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 does not show how lines are connected
- Overlapping haplotyes sometimes makes it impossible to assign to only one genetic cluster
- Y-SNPs greatly complement some of the shortfalls of the Y-STRs by themselves

How Y-SNPs work

- Defines genetic branches between 500 and 5,000+ years
- With 58,000,000 base pairs for possible Y-SNP testing, less than one percent of Y-SNPs have been discovered / analyzed to date
- Several new Y-SNPs being discovered every week
- Y-SNPs create father / son relationships that reveal exact genealogical relationships between Y-SNPs
- Mutations between Y-SNP and genealogical cluster define fingerprints which show common mutations
- Many unrelated Y-STRs have such common marker values that close matches are not even related (Y-SNPs help break up these clusters of unrelated Y-STR submissions)
- Y-SNPs will grow from 500 to 50,000 where future Y-SNPs will create thousands of branches within genealogical clusters

Four primary testing companies

- Family Tree DNA provides best overall value with most offerings, largest database and leading edge testing
- 23andme has strong autosomal test and useful Y-SNP test but lacks critical Y-STR testing and advanced Y-SNP testing
- Ancestry.com offers reasonable entry level Y-STR tests but has no Y-SNP testing or high resolution Y-STR testing
- Ancestry.com is beta testing autosomal but will not release actual raw data – do not order this predatory offering
- 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 NatGeo 2.0 test (orders being taken – results due in the fall) includes static test of massive Y-SNPs, extensive mtDNA and limited ethnic autosomal.

How do Y-STRs work

- Only found on Y-DNA chromosome
- Y-STRs are where patterns 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
- Most are one time mutations
- Discovers branches from 500 to 50,000 years
- Unlike Y-STRs, have a very hierarchical relationships (father / son relationships)
- Create true genealogical like descendant tree
- 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 10,000 Y-SNPs, many more to be discovered in the future

How Autosomal tests works

- Covers all ancestral lines – but limited to post-1850
- Comes from most chromosomes but not Y-DNA
- Recombines 50 % from each parent every generation
- Each recombination results in long 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 line
- 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
- 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, supports connection at 300 years
- 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 & complex
- 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 amount of excellent DNA forums & projects
- Look at Surname projects – many are excellent
- Join Surname project – 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
- 98 % of the analysis is done by amateur researchers and some are extremely skilled (as much as testing companies)
- Higher skills in Y-SNP projects (but most are biased towards anthological research but trend is changing)
- Be respectful of volunteers who help – sugar works better than vinegar – you should also test as recommended

Cost of Testing (Retail - FTDNA only)

- Always join project before ordering to get the project discounts & use twice a year sales
- Family Finder (autosomal) - $289
- NatGeo 2.0 (Y-SNPs, mtDNA, autosomal)- $199 – only from National Geographic extremely robust test – leading edge
- Y-STR - $169 (37), $268 (67) & $359 (111)
- Full mtDNA - $299, partial $159
- Special order Y-SNPs - $29 each
- Walk the Y - $950

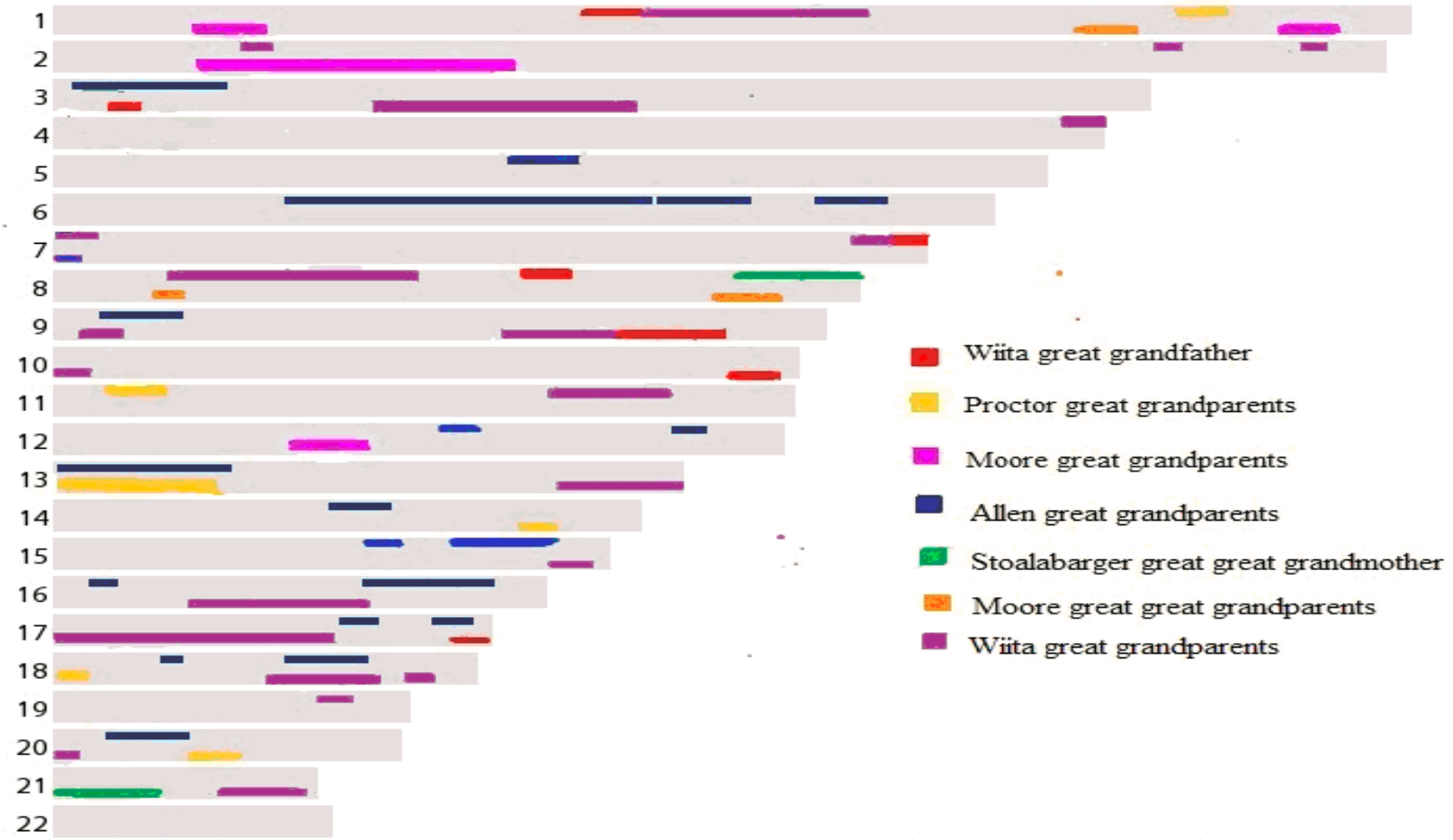
Cost of testing – other companies

- 23andme – “one size fits all” test - $299 for health, autosomal and limited Y-SNPs (no Y-STRs) – first with autosomal
- 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 raw data will be provided & scope of test not revealed do not order – very predatory offering)
- Ancestry.com – Partial mtDNA $179 (NatGeo 2.0 better offering)

Future trends

- Y-SNP tests will become the most important tests
- NatGeo 2.0 increased static Y-SNP test from 500 to 12,000 Y-SNPs (a lot to absorb)
- Potential for 10,000s of genealogical Y-SNPs & estimates show that 50,000 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) – minimal genealogical applicability

Sample atDNA comparison



War stories – atDNA for Casey

- Common ancestor believed to be born around 1700 (way too early for atDNA)
- However, common segments found between all 12 Casey related tests
- 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 South Carolina Casey genetic cluster (Y-DNA cluster that several have tested positive for 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 can 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 Pace lines claimed same ancestor
- Each had different wives, overlapping children and residences
- Both claimed same person who was proven back to Jamestown
- At least 20 books claimed the same man (including my book)
- The Jamestown line was connected to a part of London
- DNA proved both lines could not be closely related
- Two random 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 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 areas (GA & AL)
- Both lines had very similar given names
- Both lines borrowed from each other due to similar residences
- 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 MRCA states that there is less than 1 of 10,000 chance of lines being related in the last 600 years
- Genetically proved these lines are not related as once believed
- However, one line was very closely related to third different line

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)
- DNA has proven that one 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 branch and the remaining half are part of second branch
- Author of 600 page Hanvey book found out that his Hanvey line is actually a Casey line genetically and was found not to be related to their Hanvey genetically

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 is has many submissions and 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 just a few households away was one of the Bradberry NPEs that had been tested
- 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

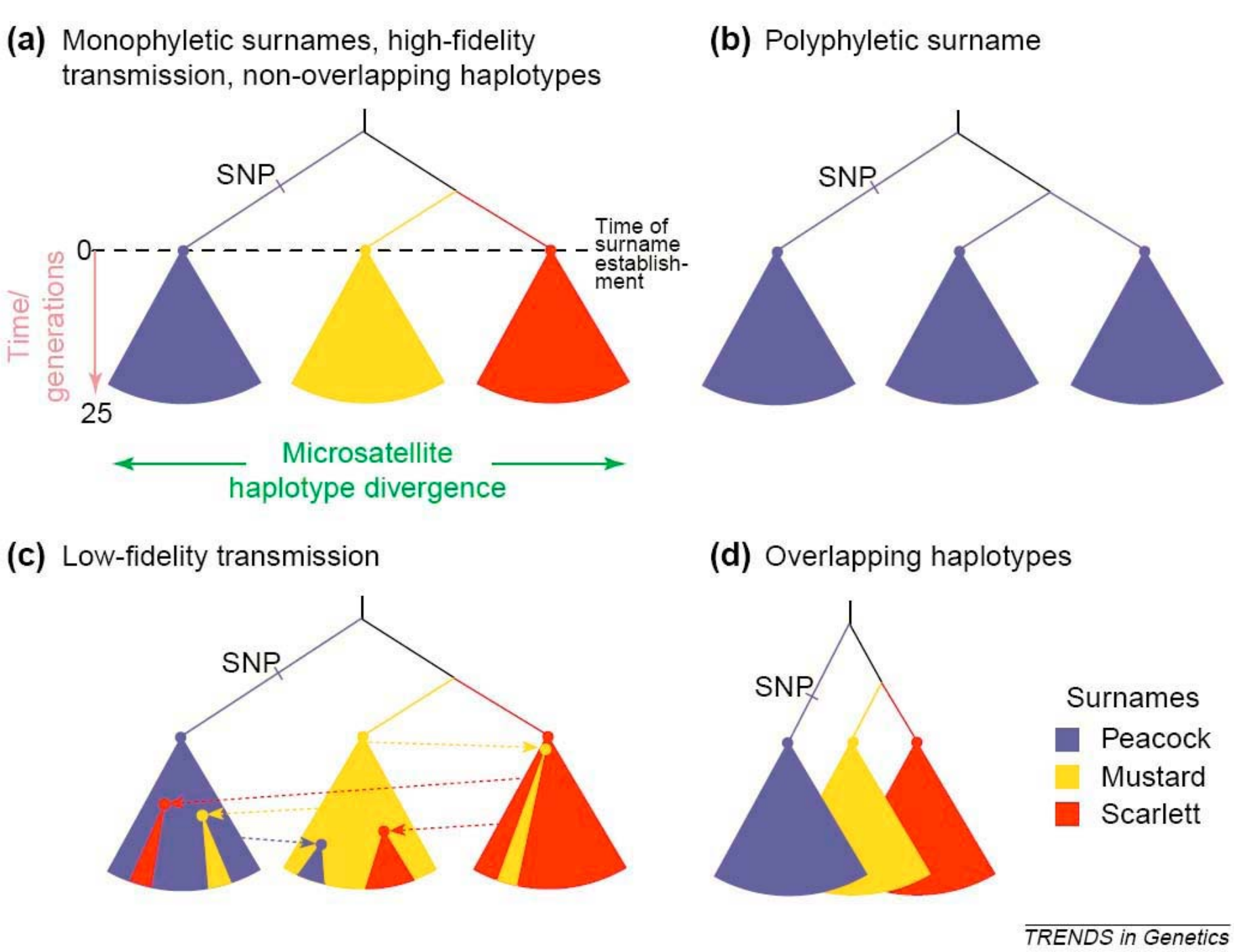
- My mother's line, Brooks, has many genealogical anomalies where the two oldest sons can not be confirmed as they were not included is extensive probate records
- However, these two unproven sons were found living in the same household via personal property tax lists and the supposed father signed the marriage bond for one son
- There were many Wade families residing in the same area and unproven family history stated the oldest ancestor married a 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
- The conclusion is that this oldest male ancestor may have married a woman who was previously married to a Wade and that the two oldest sons may have been informally adopted

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 those goals
- A lot of willing volunteers to assist – make it a two way interchange (test what they recommend)
- Before we talk about a couple of advanced topics or future trends, it is time for questions & answers

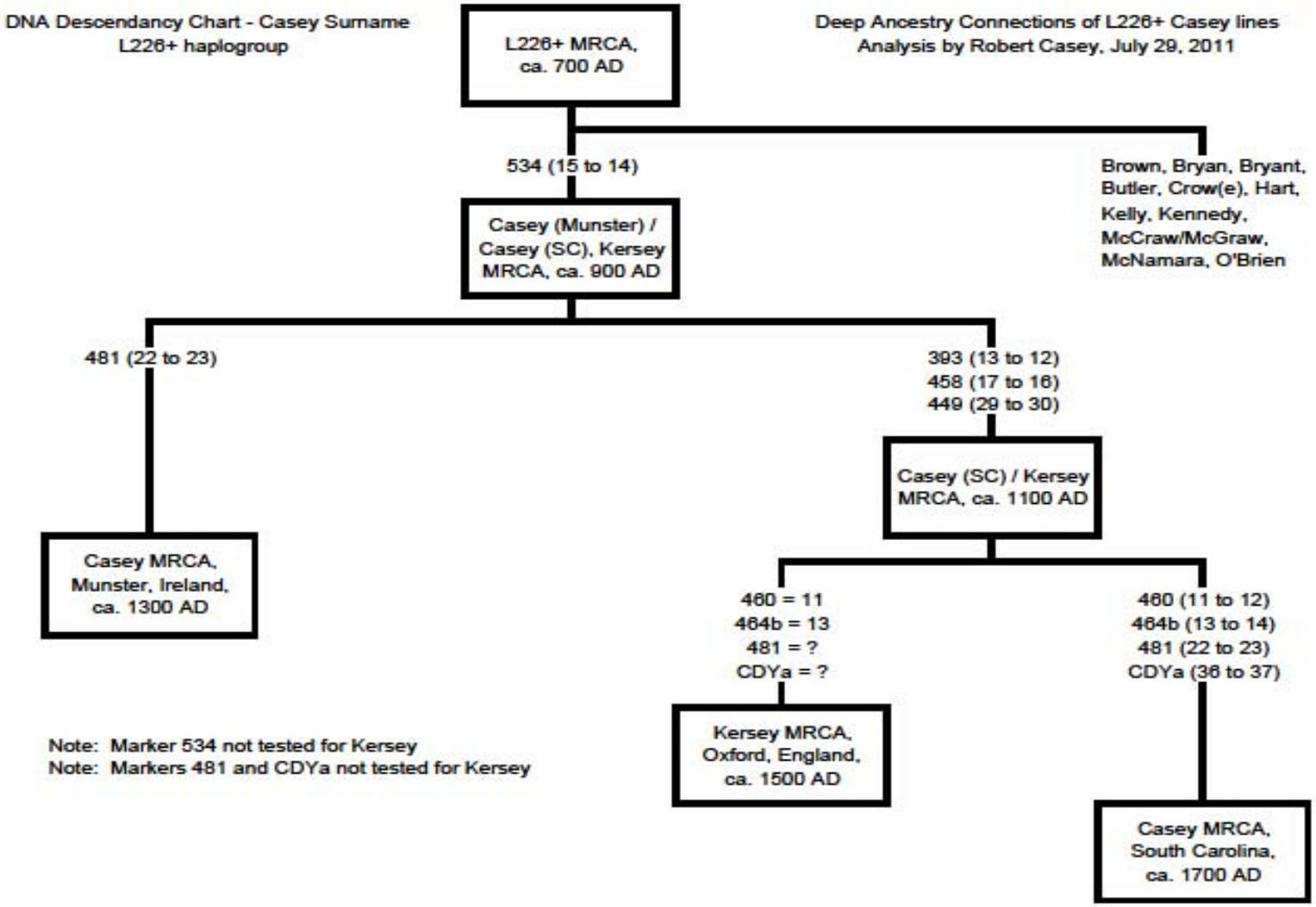
Overlapping Haplotypes

- This issue is rarely covered by books & web sites
- Some haplotypes contain such common DNA marker values that even very close matches may not related
- As much of 10 to 20 percent of all submissions fall into the "overlapping haplotype" scenario
- Genetic distance (the number of mutations that are different) is not always reliable by itself
- 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
- http://www.rcasey.net/DNA/Casey/Analysis/Analysis_South_Carolina.html


(a) Monophyletic surnames, high-fidelity transmission, non-overlapping haplotypes
SNP
0
Time/ generations
25
Time of surname establish-ment
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 are 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
- 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 500 useful Y-SNPs in the haplotree (if you ignore duplicate SNPs)
- NatGeo 2.0 will test 12,000 Y-SNPs – probably doubling 500 to 1,000 useful SNPs for western European research (majority are Chinese and Sardinian)
- It is believe that useful Y-SNP should exceed 50,000 when it becomes economical feasible to scan the entire Y-DNA strand

Y-SNP analysis is the future

- Every surname cluster should get several Y-SNPs that create branches within surname clusters
- The Y-SNPs have father-son relationships vs. Y-STRs which are only clusters of related submissions that overlap with each other
- Between all combinations of Y-STRs and Y-SNPs, most living individual as well as most deceased ancestors can 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 500 useful Y-SNPs, 50,000 Y-SNPs will produce thousands more
- Bennett Greenspan (president of FTDNA) stated that Y-SNPs will be the genealogical tree and the Y-STRs will become leaves on this tree
- FTDNA is the only genetic testing company doing any serious Y-SNP advancement while others take advantage of their research
- NatGeo 2.0 provides a static 12,000 Y-SNP test starting in November and is taking orders now (tests are processed by FTDNA and uploadable to FTDNA databases at no charge)

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 ten 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 as the should withdrawn only access for those that did not pay vs. withdrawing data 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