

Clustering

Aug 7, 2025

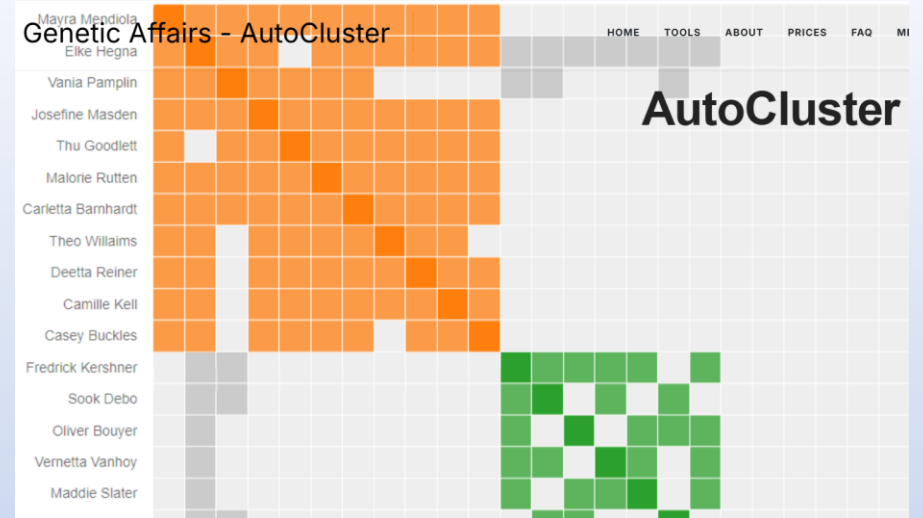
Phoenix East Valley DNA Special Interest Group



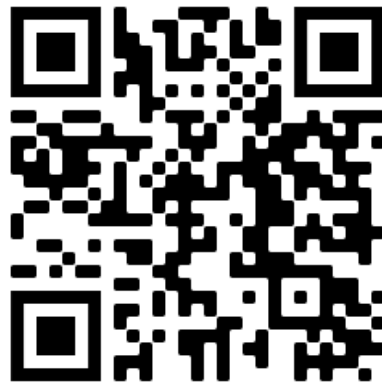
Ken waters

<http://familytreeaz.com/presentations>

Satwatcher.gen@gmail.com



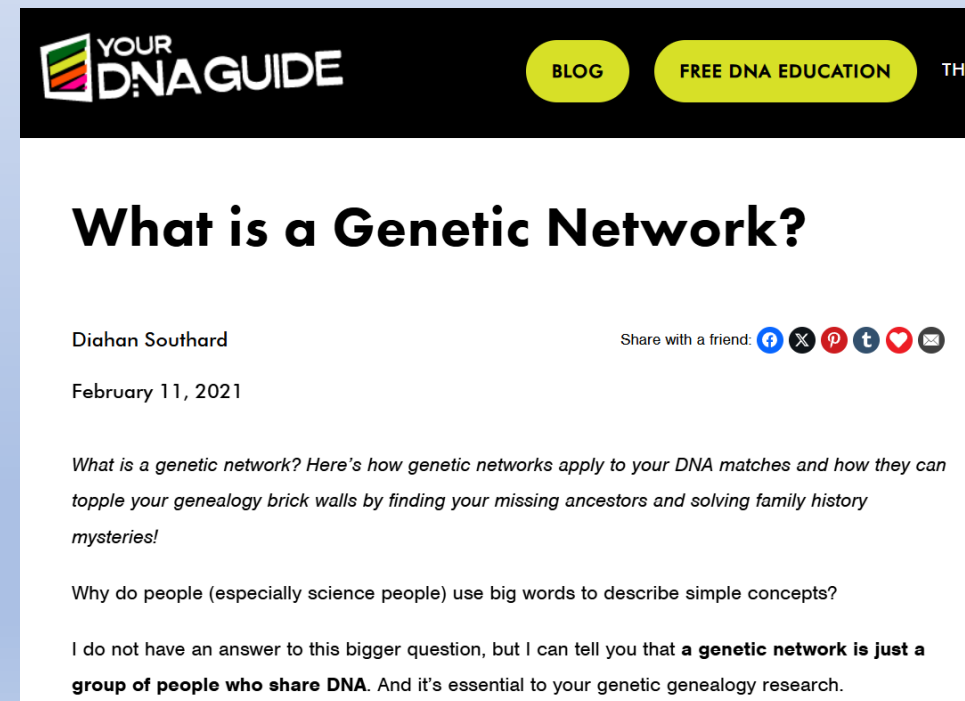
All slides and handouts can be found at:
<http://www.familytreeaz.com/Presentations/>



QR Code: take photo to
open to presentations

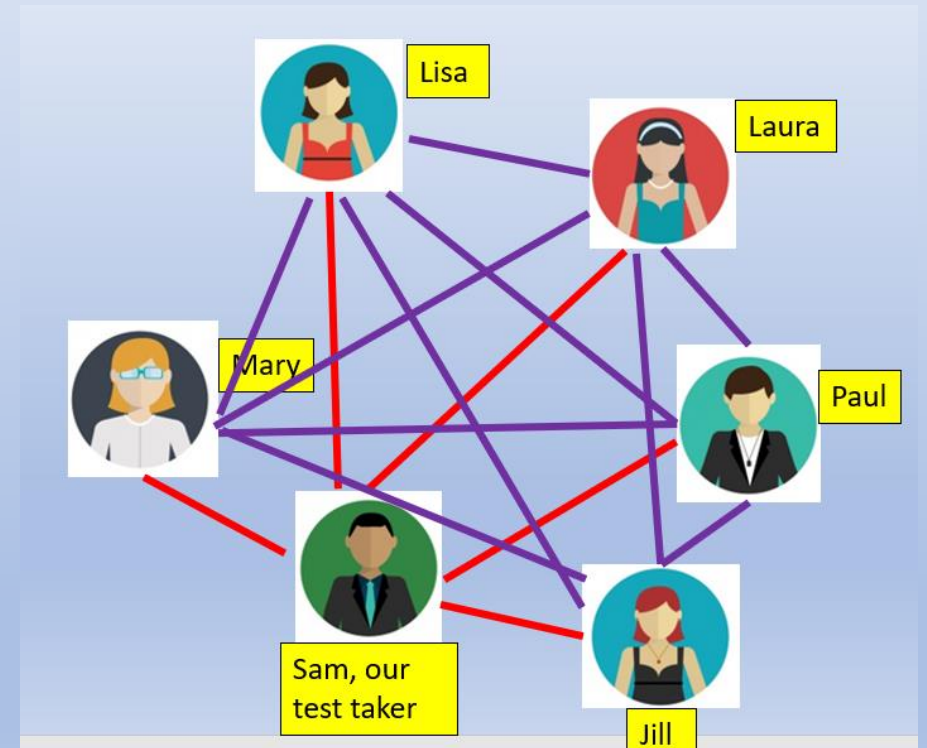
Starting Off: What is Clustering?

- Clustering (or AutoClustering) is a semi-automated process to attempt to group your DNA matches by ancestral lines
- Your DNA matches often can be sorted into Genetic Networks, groups that share a common ancestor(s)
 - For example, descendants of a common 2nd great-grandparent couple



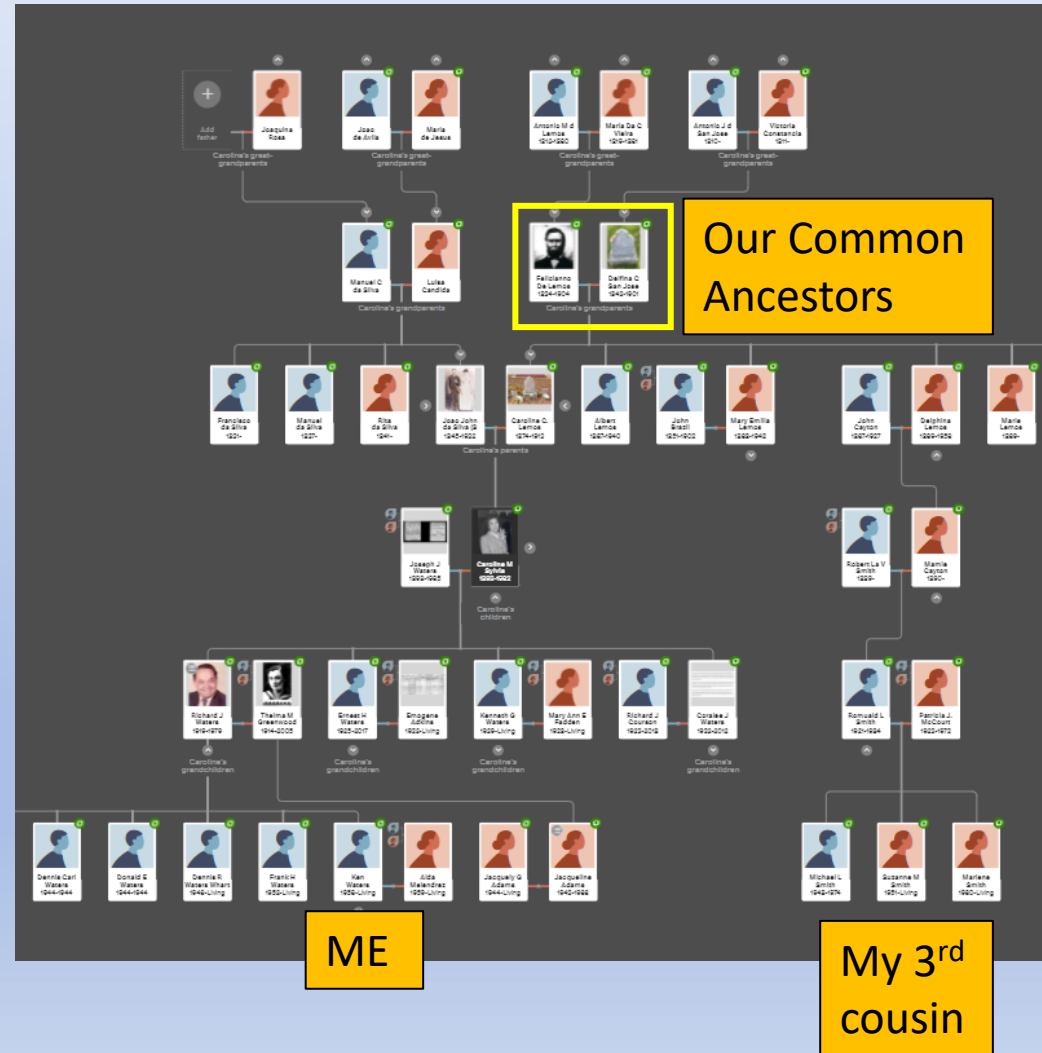
Why would we want to try AutoClustering?

- By grouping a number of your DNA matches together you can possibly identify how you are related to each of the different matches
- Sorting matches into Genetic Networks can allow you to focus on one particular network of interest, sometimes by “sorting out” networks that are not your primary focus
- Each of the members in a Genetic Network share DNA with others in the group and provide clues to a common ancestor



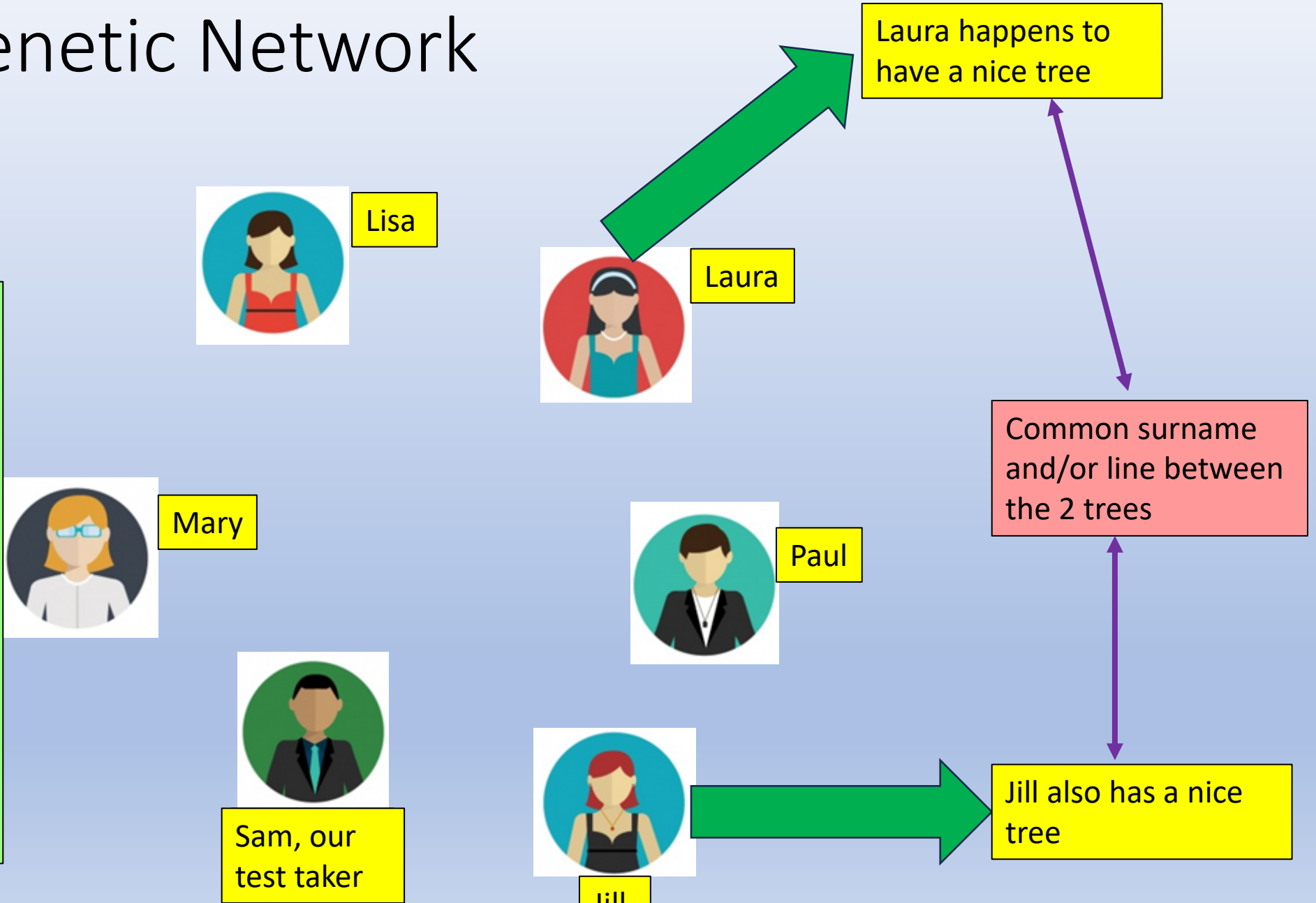
Goal: Find the Most Recent Common Ancestor (MRCA)

- Finding common ancestors allows determination of exact relationship
- Once this is found then we can compare to other matches to help see how they are related as well



Our Genetic Network

If one member of the Genetic Network has a tree then it can facilitate finding the common ancestor(s)



Non-automated ways to cluster matches



DANA LEEDS

CREATOR OF THE LEEDS METHOD

<https://www.danaleeds.com/the-leeds-method/>

You have Relatives in Common
Each Relative in Common shares at least one small segment of DNA with Michael.

Relative in common	You	Michael	DNA Overlap
MH	Matthew Thomas 1st Cousin, Once Removed (5.64)	Distant Cousin (6.08)	No
CR	CR	Sister (40.67)	Request sent
MT	MT	3rd Cousin (1.55)	Yes
CW	2nd Cousin, Once Removed (2.25)	3rd Cousin (1.14)	Request sent
MT	2nd Cousin, Once Removed (2.19)	Daughter (56.06)	Yes
BV	3rd Cousin (1.61)	3rd Cousin (6.85)	Yes
DH	3rd Cousin (6.96)	5th Cousin (6.11)	Yes
JM	3rd Cousin (6.76)	3rd Cousin (6.62)	Yes
LV	3rd Cousin (6.67)	4th Cousin (6.45)	Yes
LA	3rd Cousin (6.66)	4th Cousin (6.41)	Yes

1 2 3 4 5 6 7 8 9 10 Next >

Testing Company Identified Matches in a group

Create a new group
25/64 groups created

Group name

Assign a color

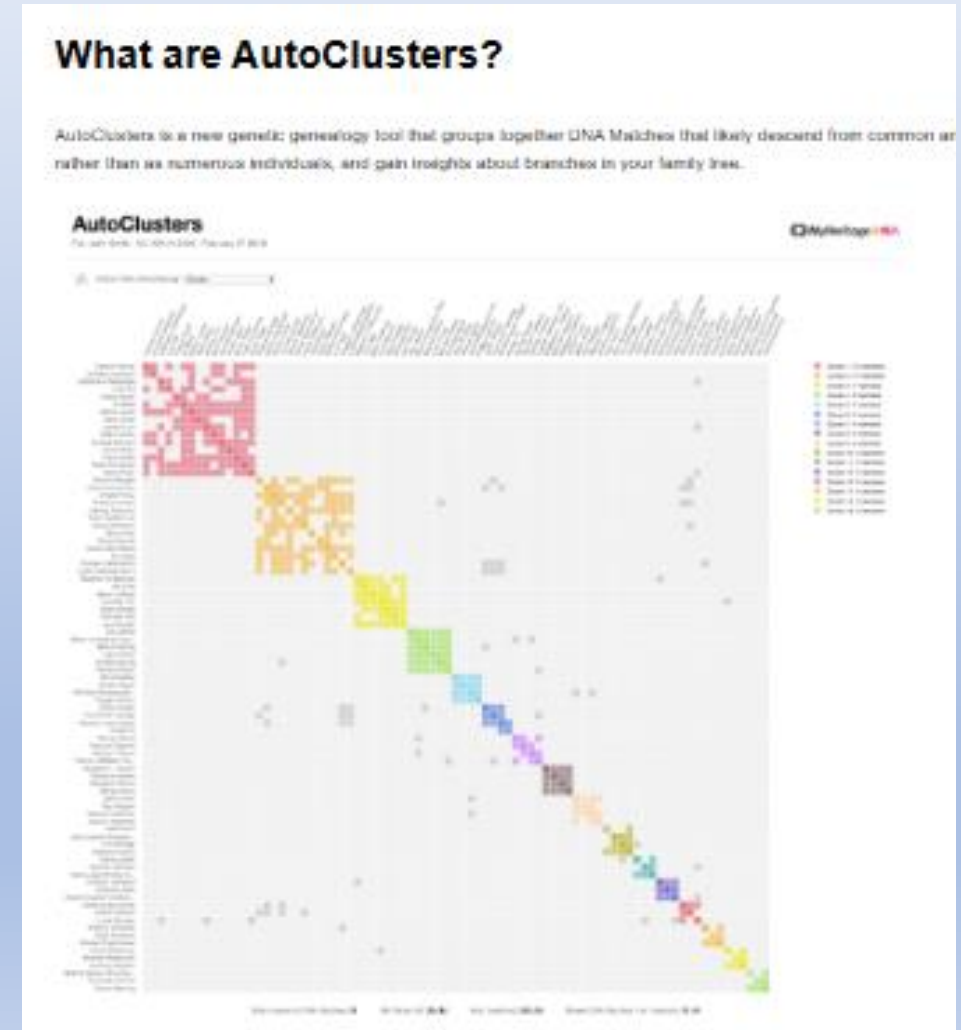
Groups

- ☐ Starred matches 33
- ☐ 0 - Distant Saved Matches 16
- ☐ 1 - Waters/Unknown 46
- ☐ 2 - Berkley/Wheeler 41
- ☐ 3 - Sylvia/Correia 95
- ☐ 4 - Lemos/San Jose 163
- ☐ 5 - Craddock/Spivey 201
- ☐ 6 - Wood/Thurman 272
- ☐ 7 - Unknown/Unknown MYSTERY 9
- ☐ 8 - Spencer/Cummings 225
- ☐ Berkley/Berkeley line from DC 49
- ☐ Dad's non-Portuguese side (Waters or Berkley or Wheeler) 75
- ☐ D -- Mom's side 15
- ☐ F Ferry/Elder on Spencer/Cummings line (maybe Tyler/Robbins) 3
- ☐ G GM Father Mystery Network #6 20
- ☐ I Inner Family 17
- ☐ L Low matches (6-8)
- ☐ M Mystery Network #7 -- Mom's side; Grandma 13

Manually Assigning Matches to a Group using colored dots

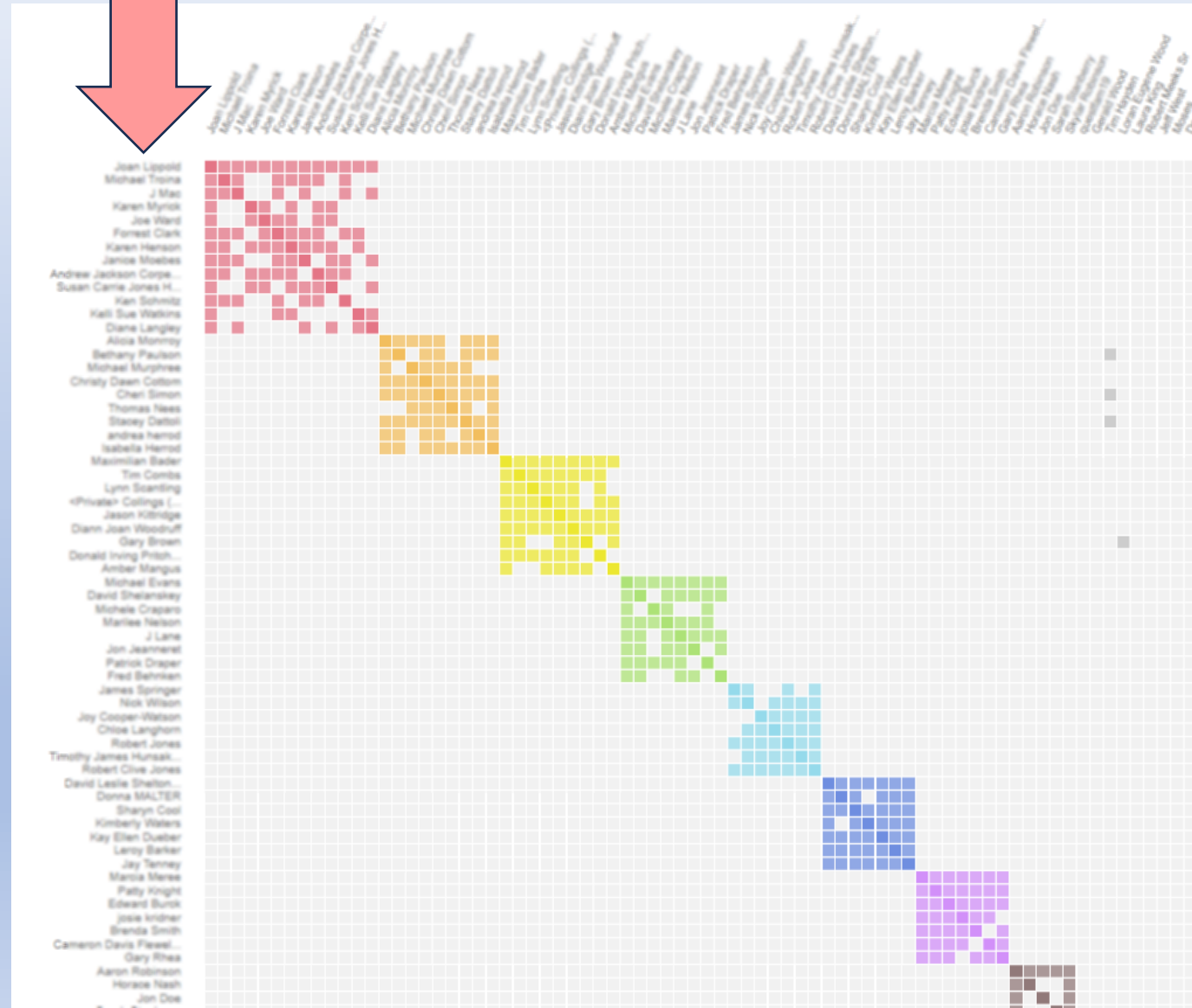
Automated Methods – On to AutoClustering

- A method to automatically take matches and sort into groups
- Result is most often a grid showing Genetic Networks



Interpreting AutoCluster Results

- Matches grouped into common groups, each a different color
- Match names across top row and left of chart
- Hovering over the different boxes provide more info
- Often there is an additional listing of the group members

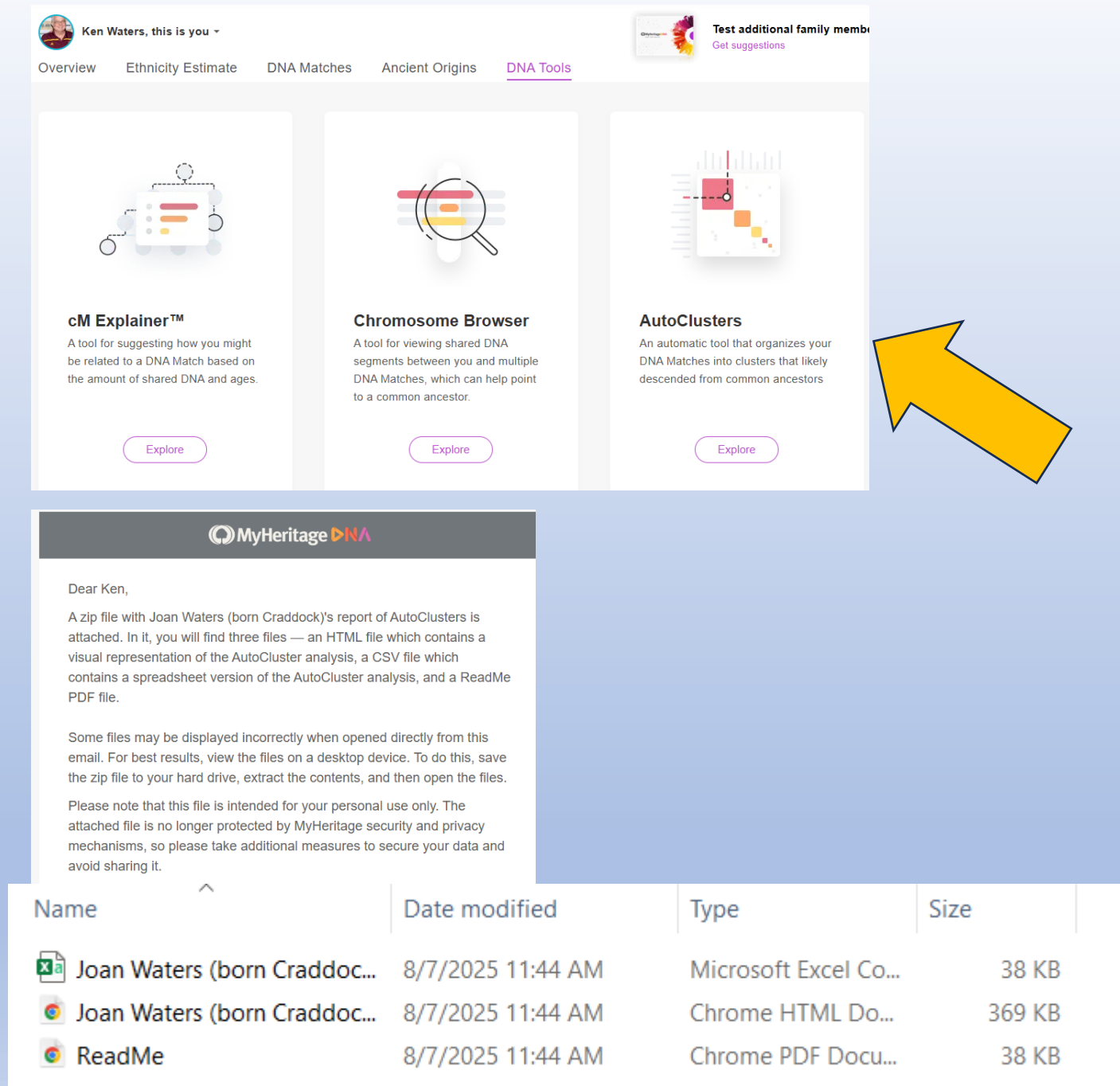


At Least 4 Ways to do AutoClustering




- MyHeritage/Genetic Affairs
 - GEDMatch
 - DNAGedcom
 - Ancestry
-
- NOTE: all of these methods do typically require additional fees

1 - MyHeritage

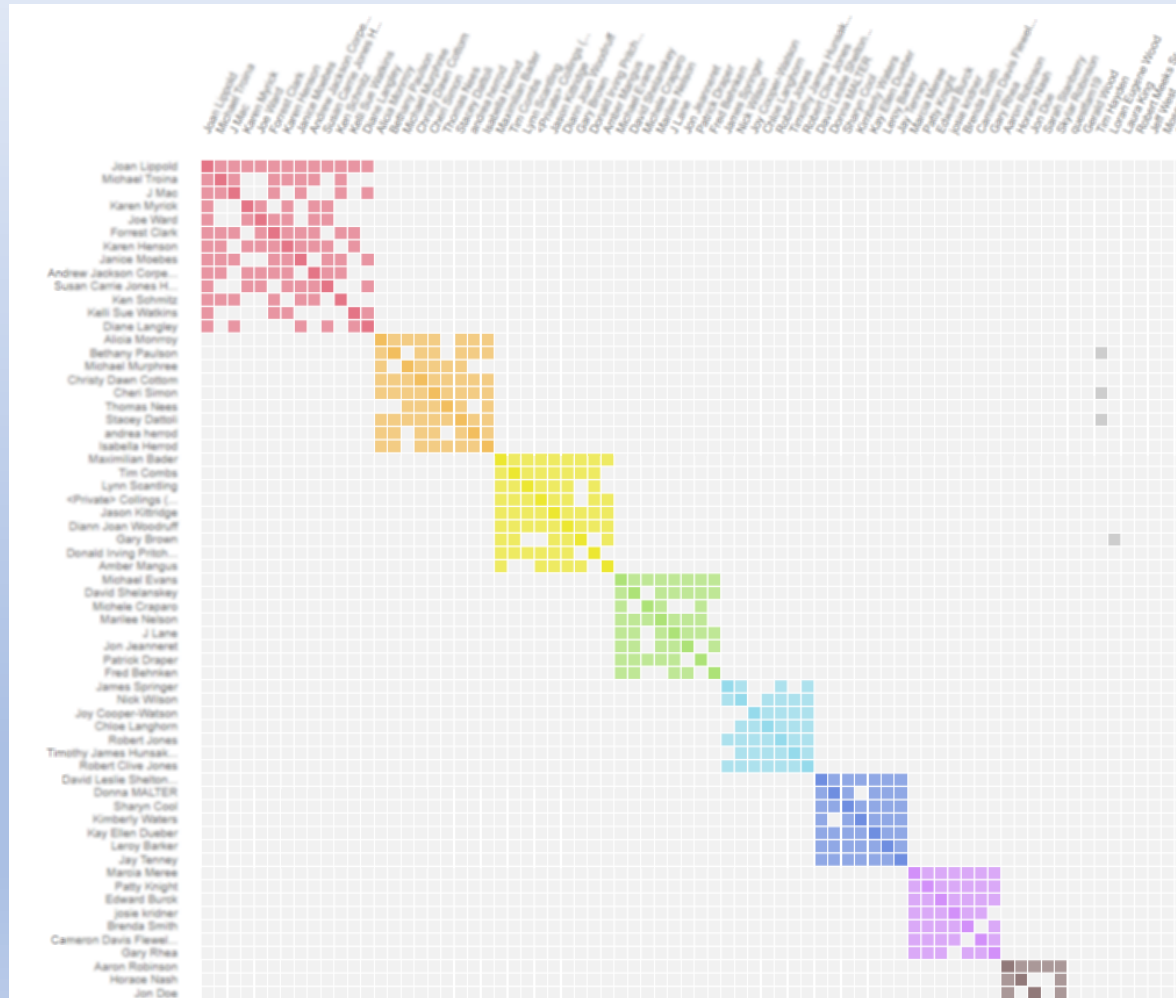
- Requires one-time upgrade fee (\$19?)
- Has set centiMorgan ranges (35-400 cMs, can not be adjusted)
- Not interactive, sends email when the cluster chart is available
- 3 files in Zip file:
 - Readme
 - HTML
 - Spreadsheet



The screenshot shows the MyHeritage DNA Tools interface. At the top, there's a navigation bar with links: Overview, Ethnicity Estimate, DNA Matches, Ancient Origins, and DNA Tools (which is highlighted). Below the navigation bar, there are three main tool cards: cM Explainer™, Chromosome Browser, and AutoClusters. Each card has an icon, a title, a brief description, and an 'Explore' button. A large yellow arrow points to the AutoClusters card. Below the tools, there's an email notification from MyHeritage DNA addressed to Ken. The email contains information about a zip file with Joan Waters (born Craddock)'s report of AutoClusters, including a list of files (HTML, CSV, and ReadMe PDF) and instructions on how to view them. At the bottom, there's a table showing the contents of the zip file.

Name	Date modified	Type	Size
 Joan Waters (born Craddoc...	8/7/2025 11:44 AM	Microsoft Excel Co...	38 KB
 Joan Waters (born Craddoc...	8/7/2025 11:44 AM	Chrome HTML Do...	369 KB
 ReadMe	8/7/2025 11:44 AM	Chrome PDF Docu...	38 KB

1 - MyHeritage

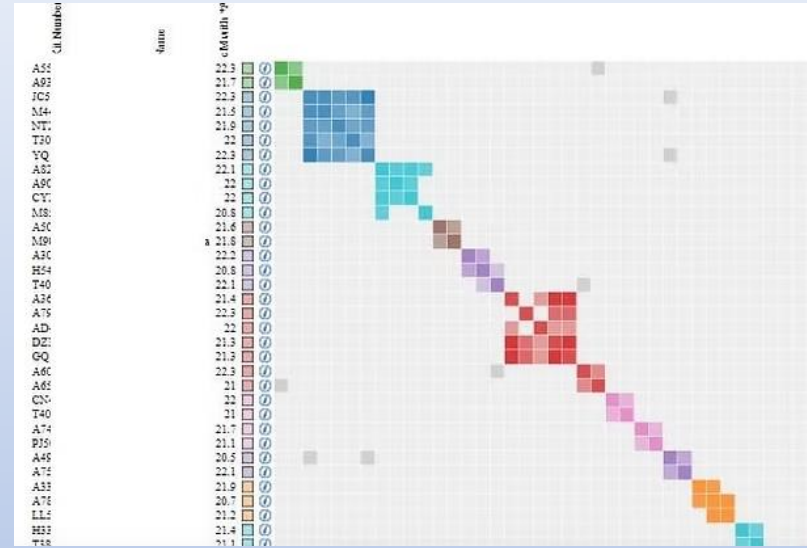


AutoClusters Information

Name	cM	Largest cM	Segments	ICW	Cluster	Tree
<input type="text" value="Search"/>	<input type="text" value="Min cM"/>	<input type="text" value="Min Largest cM"/>	<input type="text" value="Min Segmer"/>	<input type="text" value="Min #"/>	<input type="text" value="Search fc"/>	
▼ Cluster 1 (13 people)						
J [redacted]	58.2	44.1	2	14	1	
M [redacted]	45.8	45.8	1	9	1	6
J [redacted]	42.2	36.1	2	9	1	89
K [redacted]	38.3	32.3	2	5	1	13
J [redacted]	42.1	34.7	2	7	1	49
E [redacted]	41.7	35.6	2	9	1	8
K [redacted]	46.6	39.5	2	10	1	123
J [redacted]	40.4	40.4	1	9	1	7
A [redacted]	36.4	36.4	1	8	1	651
S [redacted]	38	29.4	2	7	1	178

2 - GEDMatch

- Adjustable centiMorgan range
- Requires Tier 1 subscription (\$15 one-time or \$10 monthly recurring)
- One **BIG** negative: Small database size (only ~1 million estimated)

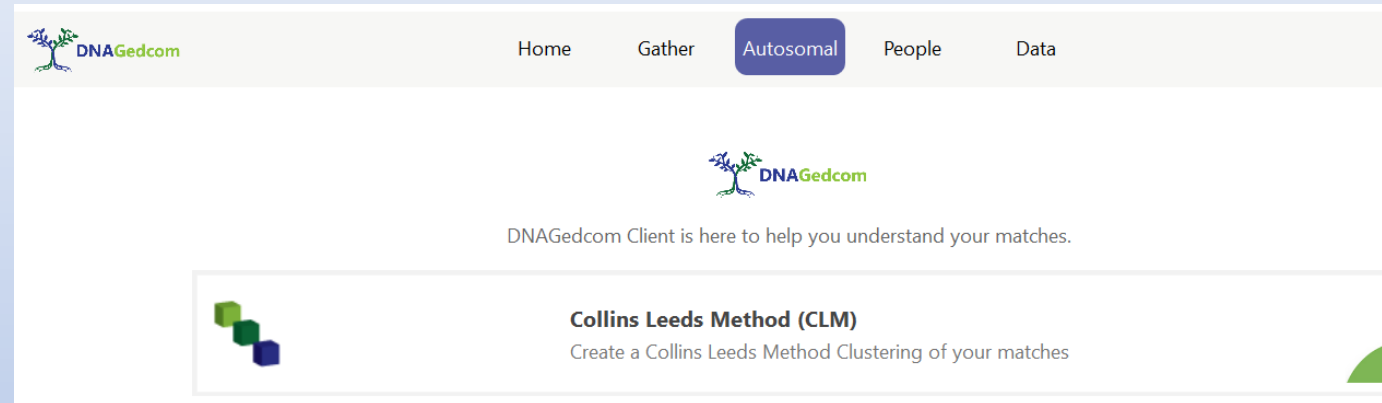


<https://www.familyhistoryfanatics.com/gedmatch-clustering>

<https://dna-explained.com/2022/02/21/autokinship-at-gedmatch-by-genetic-affairs/>

3 - DNAGedcom

- **Most flexible**
- Requires download of client and paid registration
- Requires download of match data (can take hours or even days)
- Once downloaded you have very flexible options, not just centiMorgan ranges
- Works best with Ancestry but can be affected by Ancestry's periodic backend changes; can work with other testing services but these may or may not work at any given time
- Has the CLM (C-Leeds Method) AutoClustering tool
- Cost is \$5/mo or \$50/yr

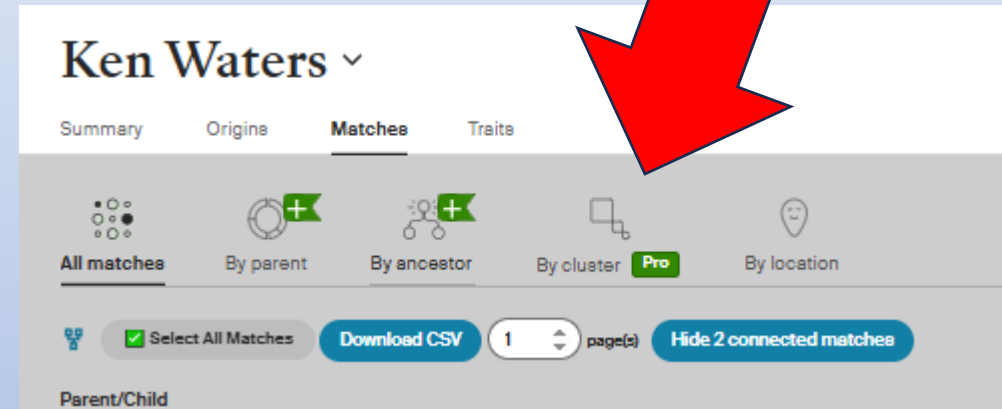


Most Flexible

Allows filtering by cM range as well as grouping (your colored dots) --- that's a BIG deal as it allows working on just a single ancestral group of interest

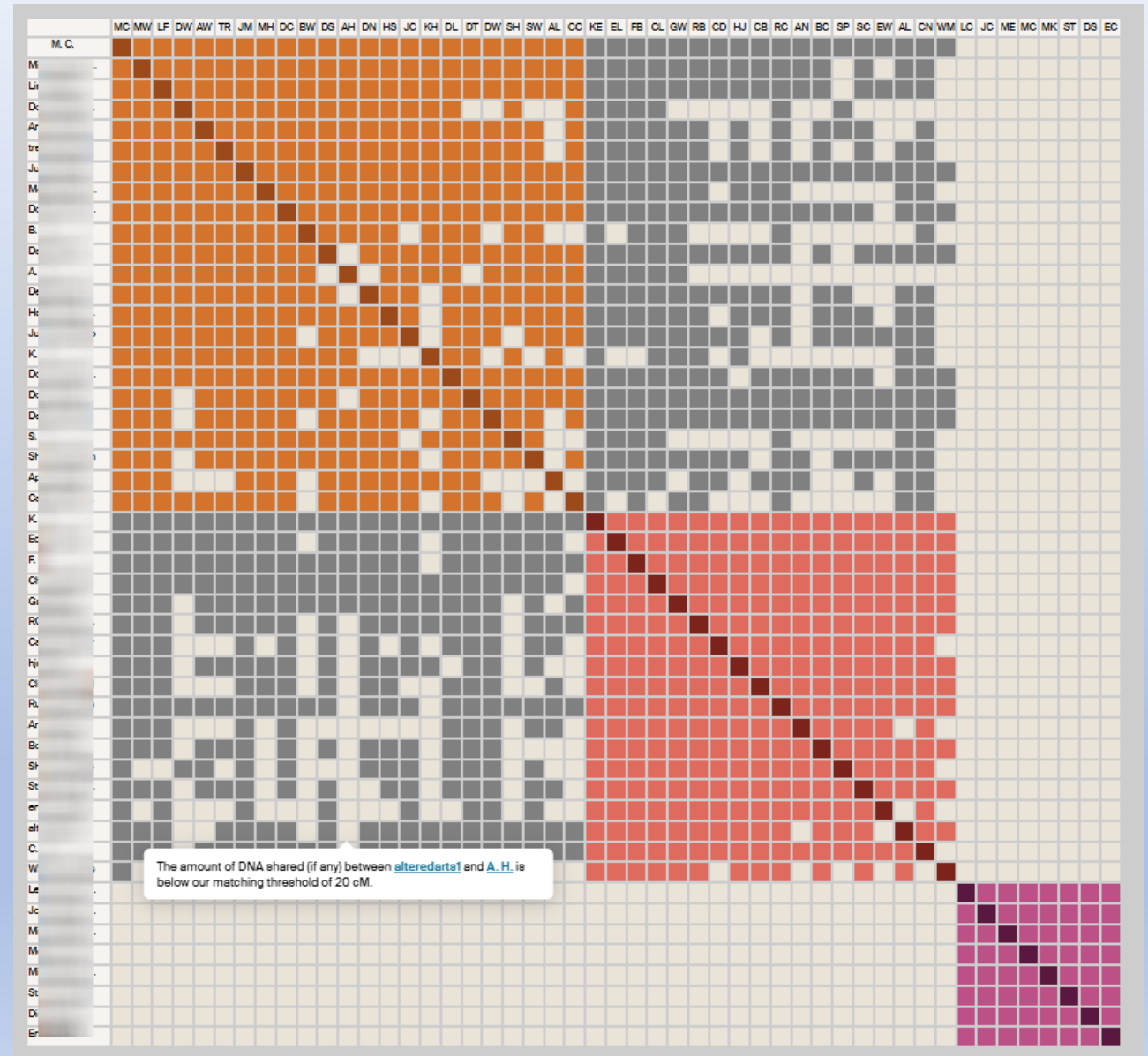
4 - Ancestry

- Just released in 2025
- Takes advantage of the largest database of DNA kits
- Currently the centiMorgan range is fixed at 65 to 1300 cMs but the ability to modify the ranges has been announced to come out soon
- Does require ProTools subscription (~\$45/quarter)



4 - Ancestry

- My kit only produced three clusters, two of them related to each other (see the grey boxes); the large one is paternal and the small one is maternal
 - Not very useful without ability to adjust cM ranges

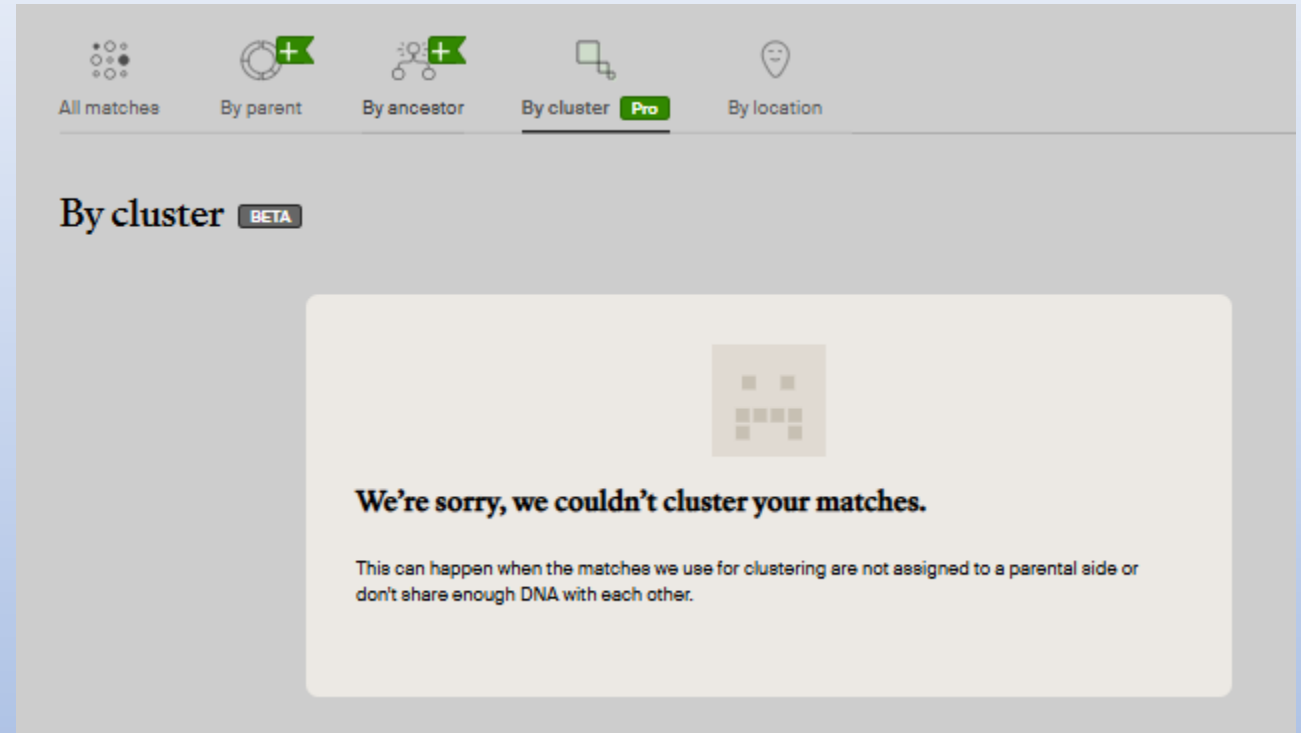


Why is it useful to be able to adjust ranges?

- Allows fine tuning of the clusters to get the desired grouping
- Focus on a particular cluster of interest that may not be shown with set ranges currently on MyHeritage and Ancestry (supposedly soon to change)
- At this time, only 2 services allow setting custom ranges:
 - GEDMatch
 - DNAGedcom/CLM

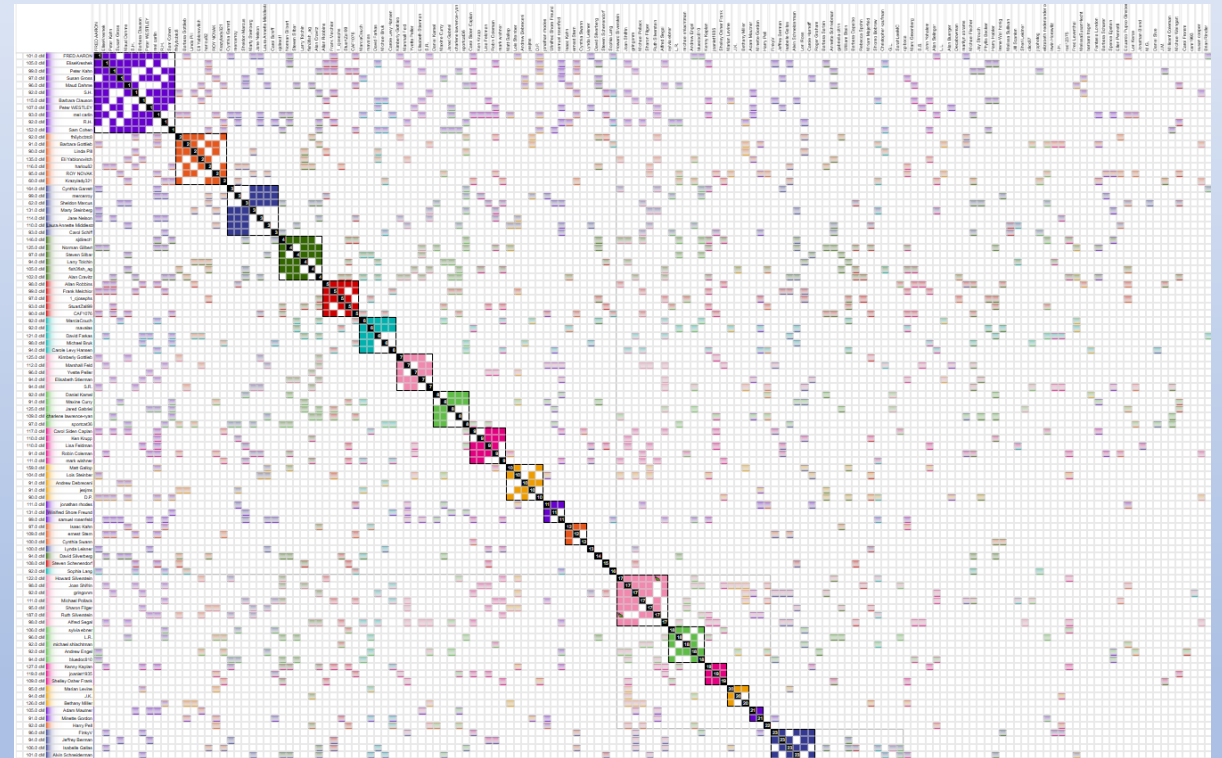
One Note on Endogamy

- AutoClustering of endogamous kits can be particularly problematic due to too many interrelated kits
- Ancestry won't even try to build a cluster (this is for an identified 100% Ashkenazi Jewish kit)



One Note on Endogamy

- AutoClustering of endogamous kits can be particularly problematic due to too many interrelated kits
- This is what the DNAGedcom CLM (AutoCluster) tool shows for a 100% Ashkenazi kit
 - Notice all the grey boxes and the small cluster sizes
 - Multiple relationships---makes interpreting very difficult



My Recommendation...?

- For now, best option is DNAGedcom (most flexible and powerful) but does require lengthy download
- Second preference goes to Ancestry due to the large database; maybe after they implement the ability to adjust cM ranges it might be my top pick but for now it's limited due to fixed ranges
- Are these useful...?
 - My opinion: can be helpful to visualize grouping of matches --- but: mostly require additional analysis/grouping using spreadsheets or other methods



Presentations:

<http://familytreeaz.com/Presentations>



Contact:

Ken Waters

E-Mail: **satwatcher.gen@gmail.com**

Google Voice Phone: (480) 331-5889

