What to Expect When You're Clustering

Walter Steets

Houston Genealogical Forum

DNA Interest Group

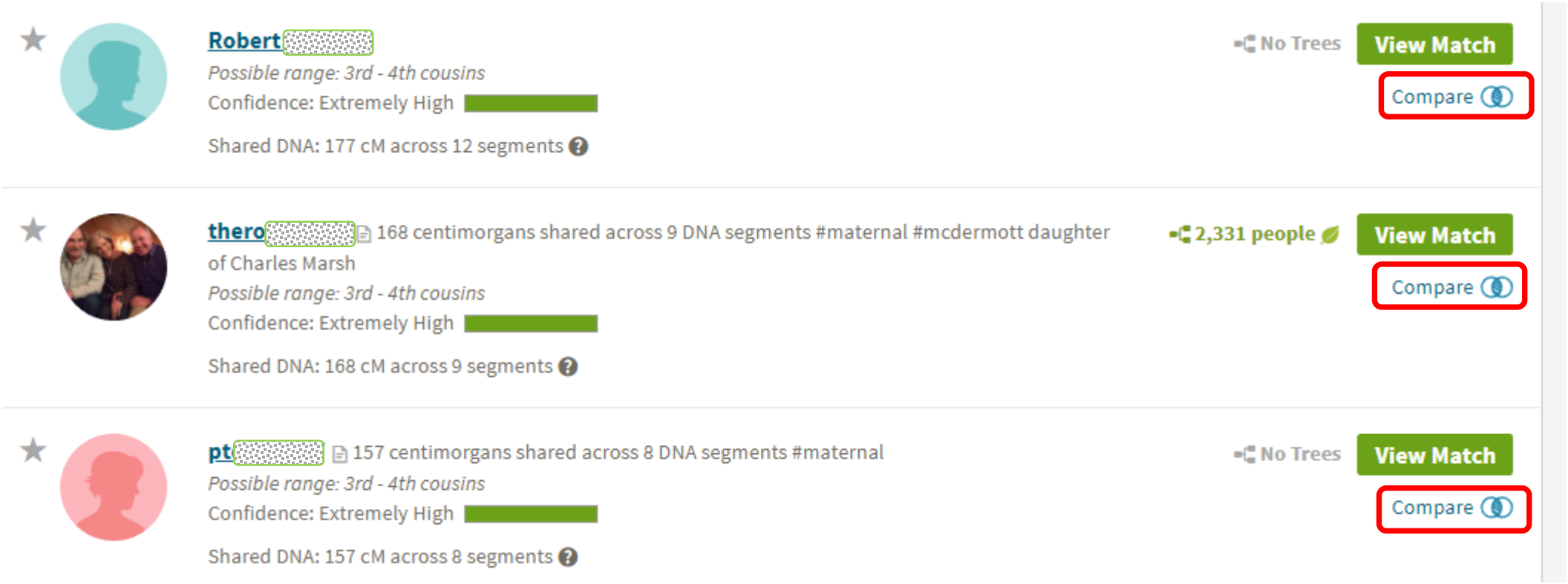
January 5, 2018

Today's agenda

- New Ancestry Match Comparison Report
- Clustering for DNA Matches
  - Describe clustering of DNA matches
  - Compare clustering tools from Genetic Affairs and DNAGedcom
  - Review example of clustering using DNAGedcom
  - DIG members experience with DNA clustering

New Ancestry Match Compare Report

- Compare Report combines information about shared DNA, Ethnicity, and Shared Marches
- Click on **Compare** to open report



AncestryDNA Compare Report
Shared DNA and Ethnicity

Shared DNA With Linda

Predicted Relationship: 3rd Cousin

Amount of shared DNA is **107** centimorgans across **6** DNA segments

Ethnicity Estimates

Walter Steets		Linda
48%	England, Wales & Northwestern Europe	77%
35%	Ireland and Scotland	21%
8%	Sweden	2%
8%	Germanic Europe	0%
1%	Finland	0%

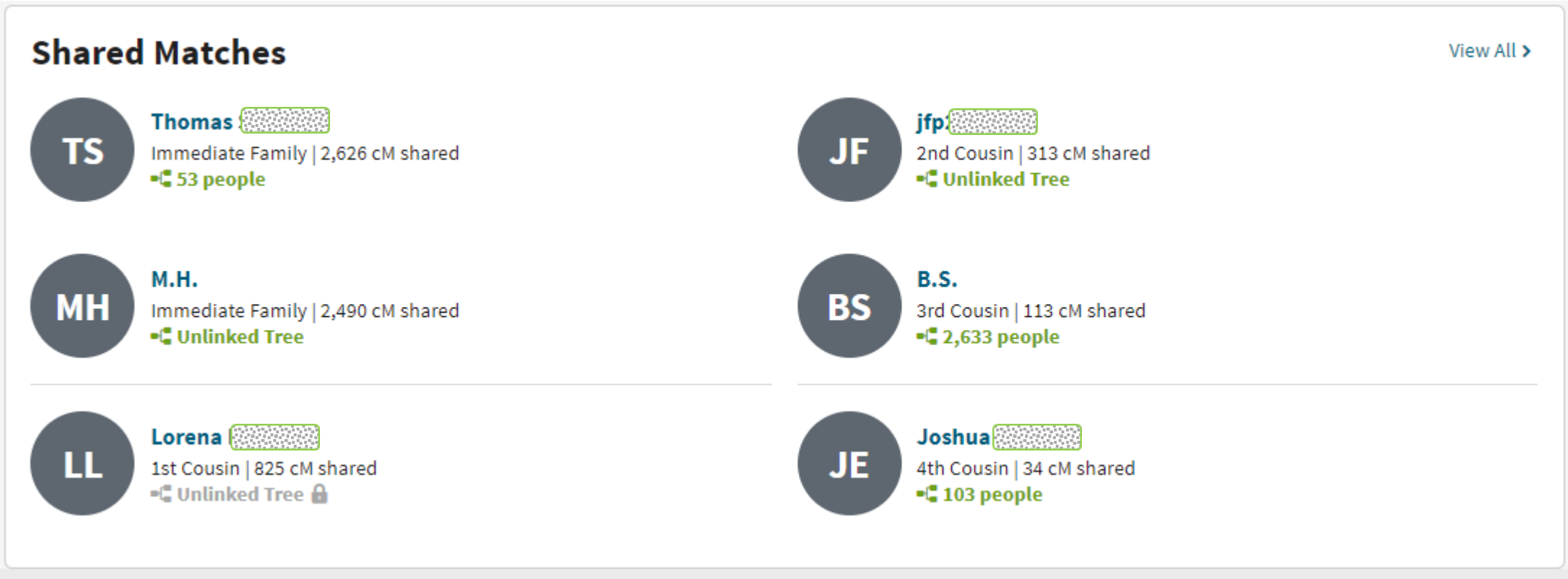
Shared Migrations

Germany & the Midwestern United States

German-speaking immigrants first came to America for familiar reasons: land, economic opportunity,...

Learn More

AncestryDNA Compare Report
Shared Matches

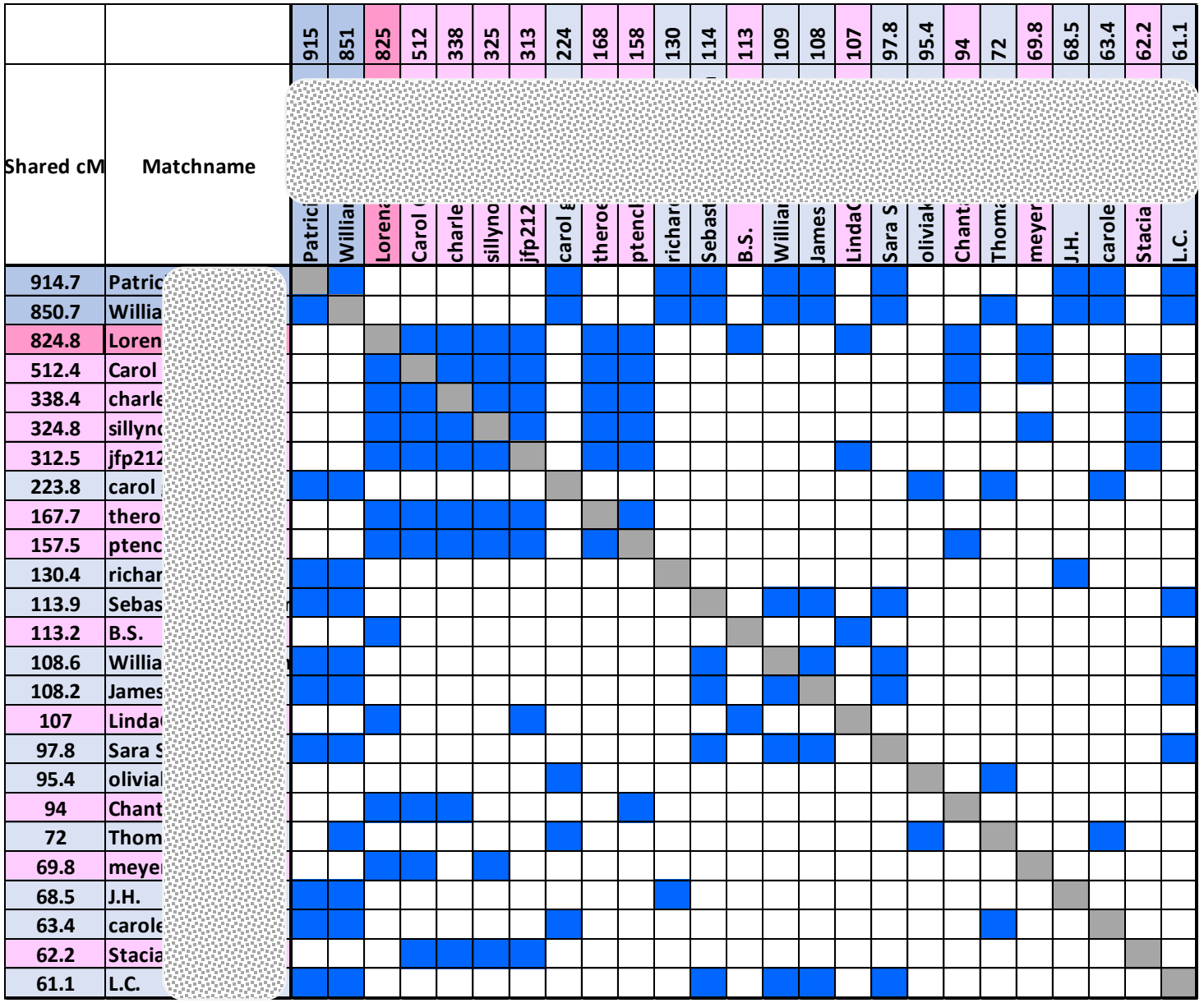


DNA Clustering

- As in the Ancestry DNA Compare Report, each of the DNA testing companies report:
  - List of people with whom we share DNA – our Matches
  - For each Match
    - Amount of DNA we share with Match and estimates of our relationship with the Match based on the amount of shared DNA
    - List of people who share DNA with both us and the Match – our Shared Matches
    - Some companies also provide list of shared DNA segments with Matches
- Problem: The location of the match in our family tree frequently cannot be identified based on this data alone
- DNA clustering can help
  - Discovers groups of Matches called clusters within our Shared Matches
    - DNA is shared with most of the other Matches in the same cluster
    - Not shared with most of the other Matches in other clusters
- By finding and displaying patterns in the relationships of Shared Matches, clustering can provide additional insight into the identities of our DNA matches.

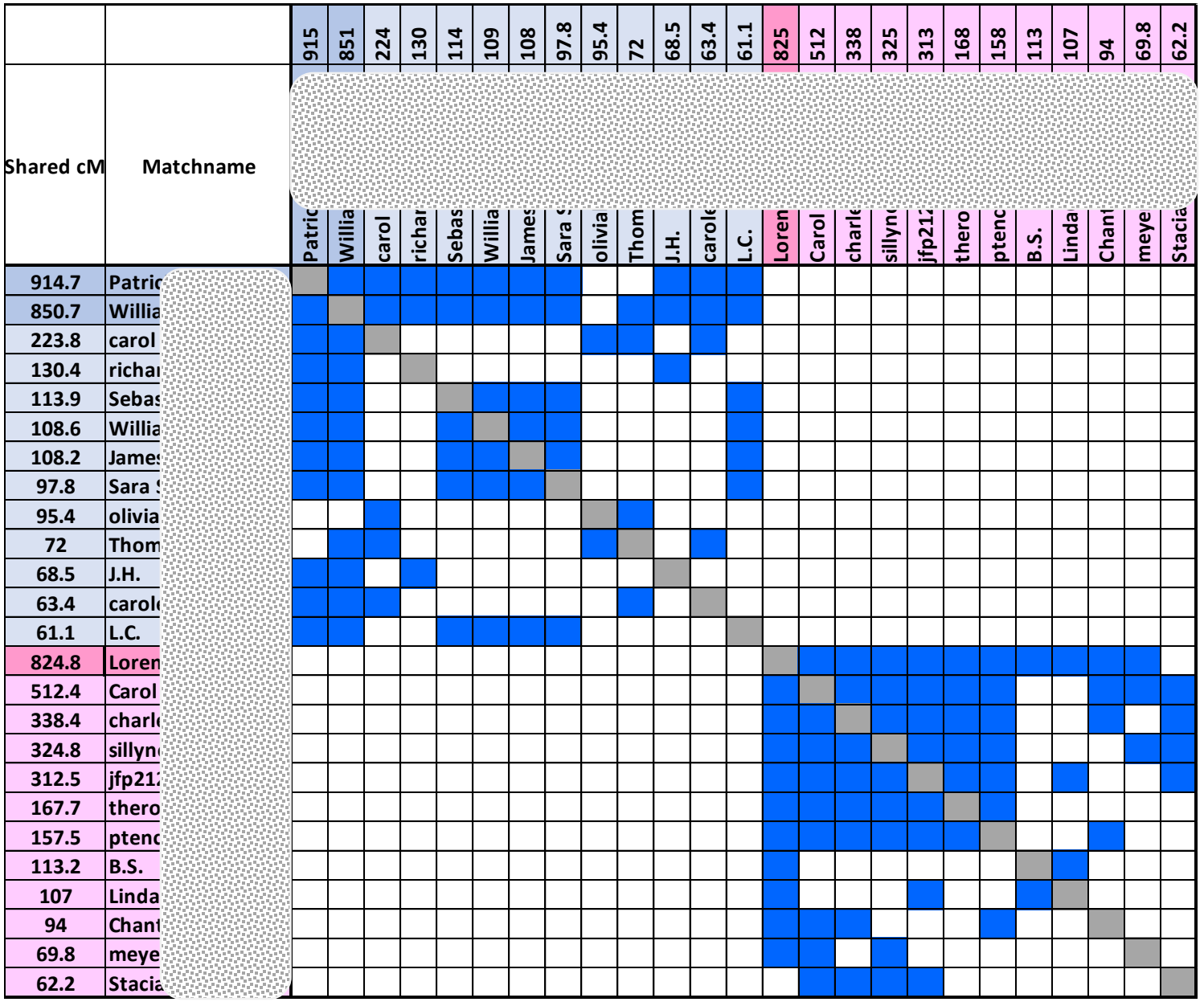
Clustering Example – Tabulate and Label Shared Matches

1. List closest matches in spreadsheet
2. Show matches
   i. Matches – blue square
   ii. Diagonal – dark-grey square
3. Identify closest relatives – in this case first cousins
   i. Paternal 1C – blue-grey labels
   ii. Maternal 1C – dark-pink labels
4. Use first cousins to identify paternal and maternal matches
   i. Paternal – light-grey labels
   ii. Maternal – light-pink labels



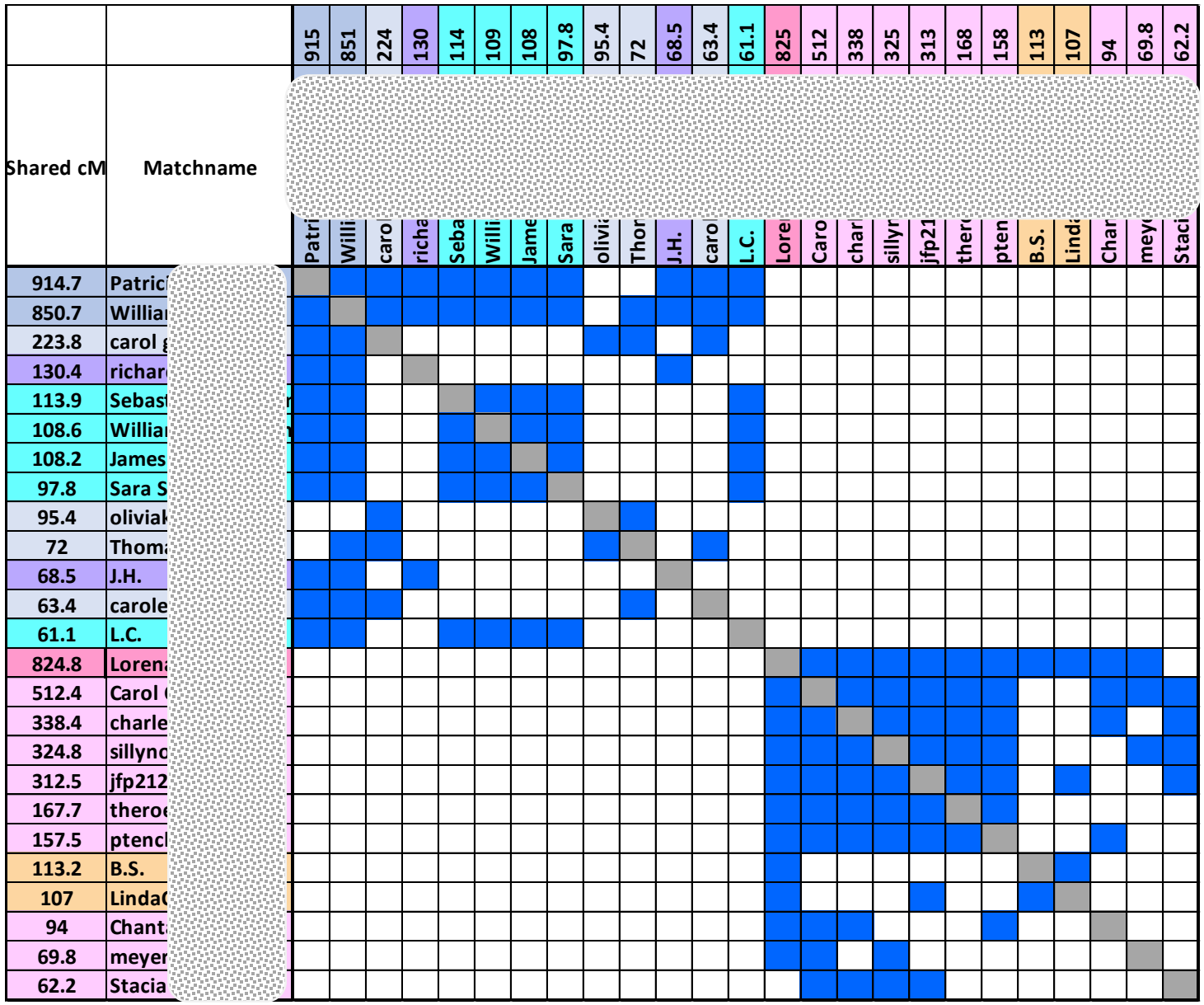
Clustering Example – Group Paternal and Maternal Matches

1. Separate paternal and maternal matches by sorting colors in following order:
   i. Blue-grey for paternal 1C's.
   ii. Light-grey for remaining paternal matches
   iii. Dark-pink for maternal 1C's
   iv. Light-pink for remaining maternal matches
2. Paternal and maternal matches are grouped together in separate blocks within the matrix – these are high-level **clusters**



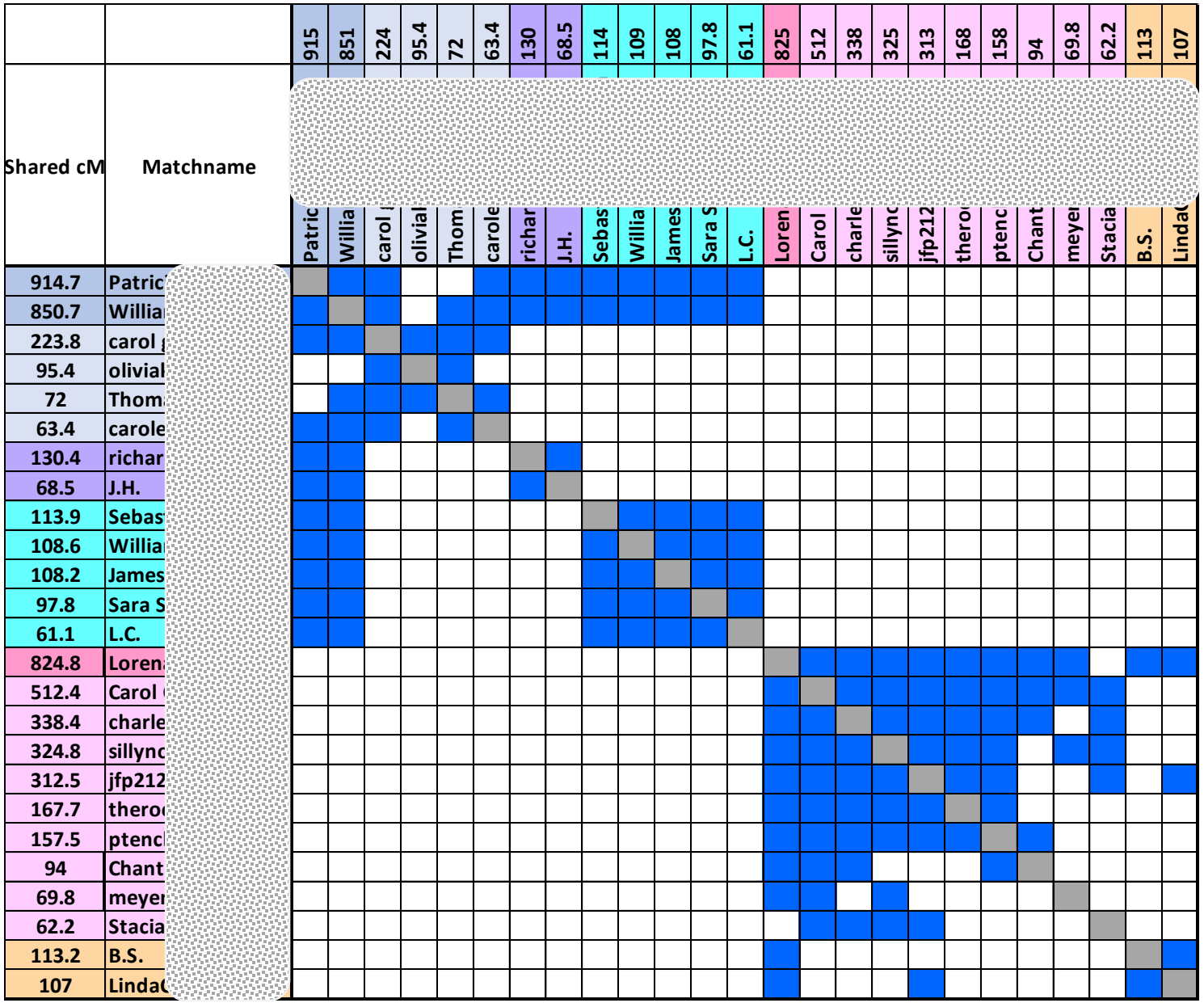
Clustering Example – Define Clusters

1. **Clusters:**
   - i. **Sets of matches who have more matches with each other than with other matches.**
   - ii. **High-level clusters likely contain matches for a range of generations e.g. 1C – 4C**
   - iii. **Larger high-level clusters may contain smaller lower-level clusters**
   - iv. **Lowest level clusters likely contain matches for only one or two distant ancestors e.g. 4C, 5C**
2. In example, matches in lower-level clusters are colored in light blue, purple, and tan.



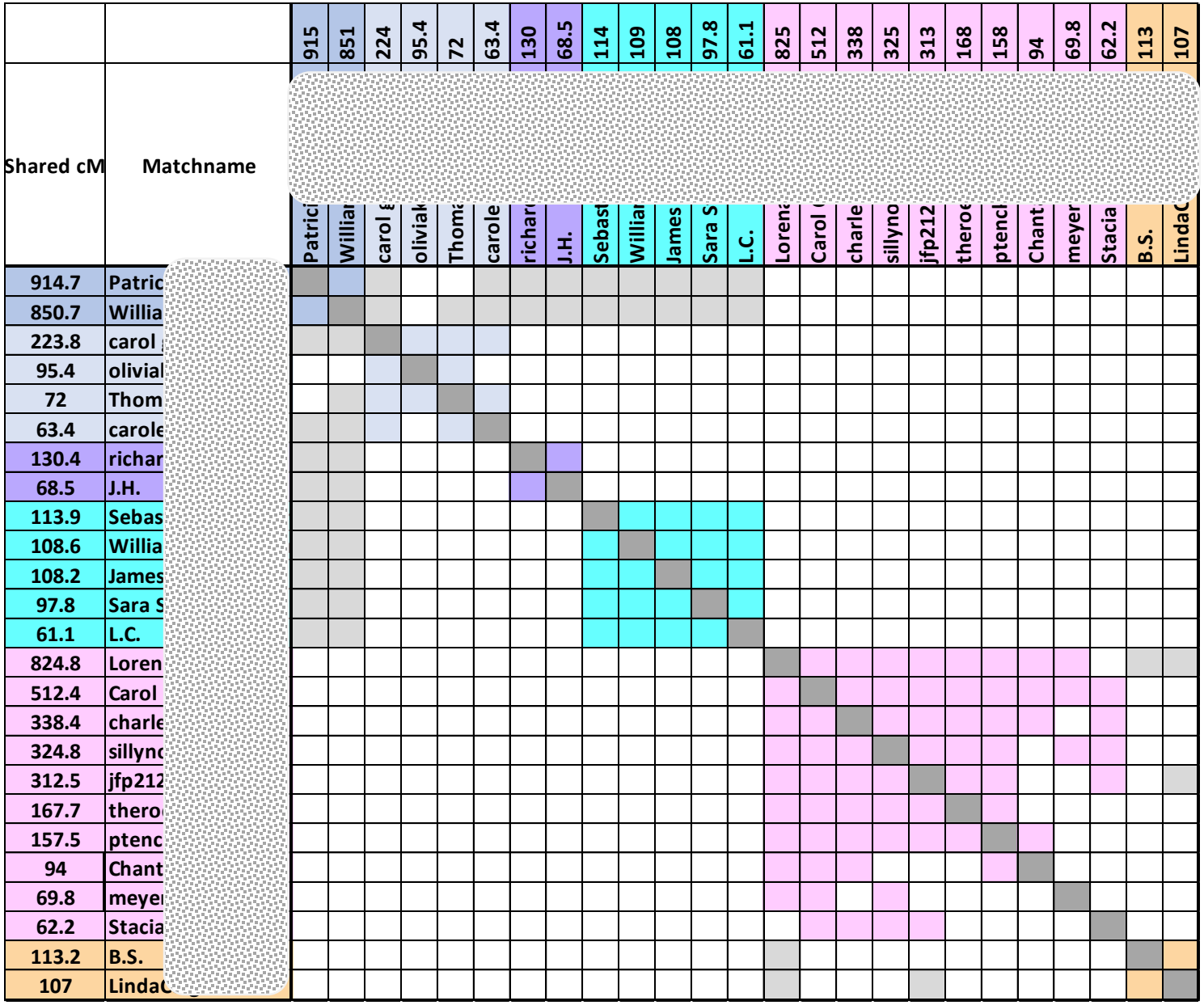
Clustering Example – Levels of Clusters

1. Sort matches by color to group multiple levels of clusters
2. Order colors for sorting by amount of shared cM in match with the largest amount of shared cM in the high-level cluster
3. Sorting has made the sets of related matches much more clear
4. Final color coding (on next slide);
   i. Color the match squares with the color of the cluster
   ii. Color the squares between clusters light grey

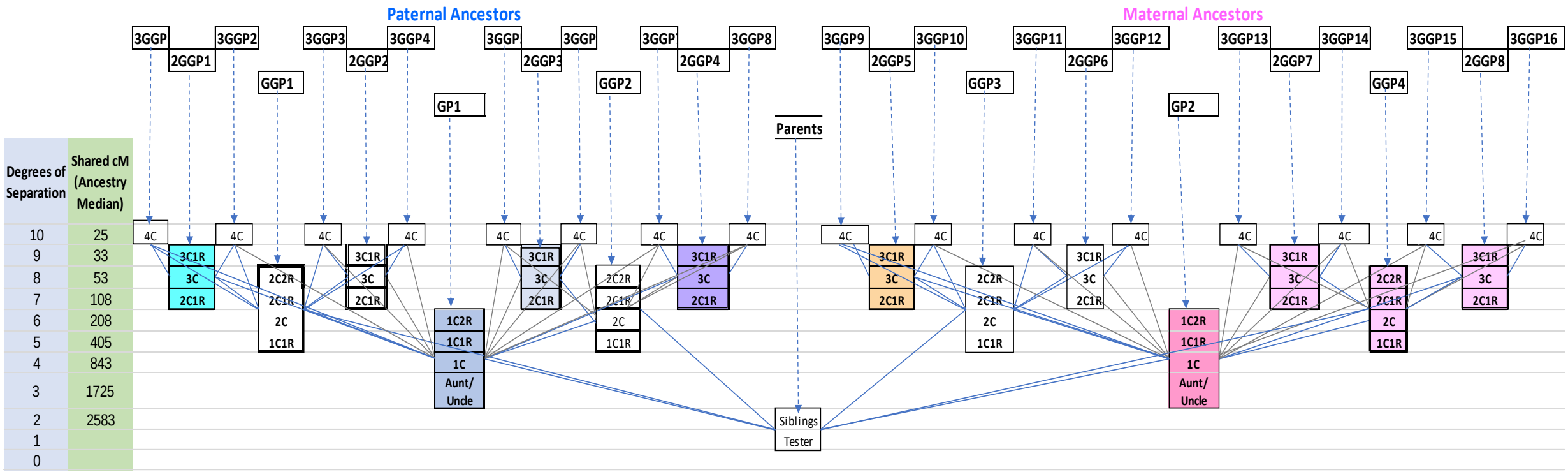


Clustering Example – Interpretation of Clusters

1. Paternal Cluster
   - i. Blue-grey:
     - a. Shared cM consistent with 1C
     - b. Cluster members matches almost all lower-level clusters
   - ii. Light-grey, purple, light-blue:
     - a. Shared cM consistent with 2C, 2C1R, 3C, 3C1R
     - b. Three paternal clusters which have no shared matches between them making 2C1R (father's 2C) and 3C more likely
2. Maternal Clusters
   - i. Pink:
     - a. Shared cM consistent with 1C, 2C, and 3C
     - b. Two matches with smallest shared cM (Jeff Meyers and Stacia Milius, likely 3C) could have separate MRCA's
   - ii. Tan:
     - a. Match only 1C and 1C1R (1Cs son)
     - b. Likely 3C on separate line from Pink



Most Recent Common Ancestors and Shared DNA

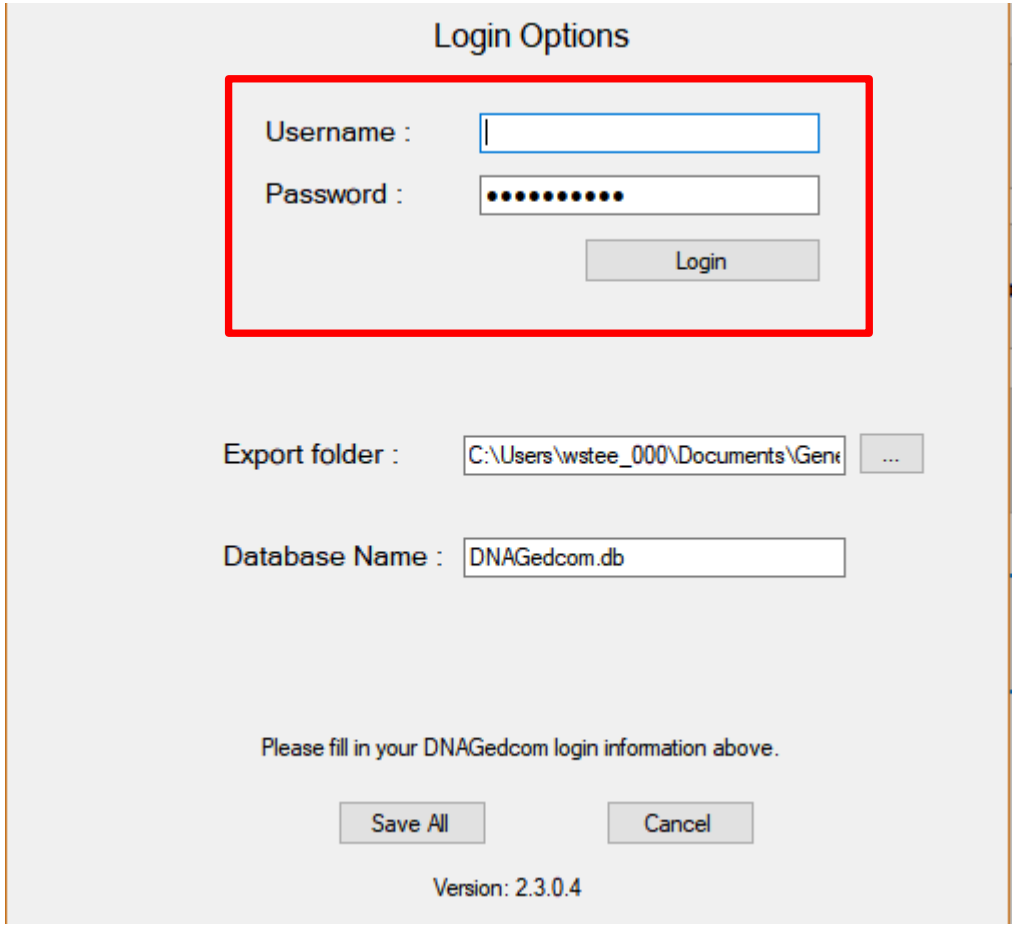


New DNA Clustering Tools

	DNAgedcom	Genetic Affairs
Name	Collins' Leeds Method – 3D (CLM3D)	AutoClustering
Release Date	14 December 2018 New releases every week or so	1 December 2018 New releases every week or so
Supported Testing Companies	AncestryDNA 23andMe FamilyTreeDNA ?	AncestryDNA 23andMe FamilyTreeDNA MyHeritage
Platforms	Website	PC and Mac
Results	Spreadsheet and html file of clusters sent by email	Spreadsheet and html file of clusters placed in local directory on PC
Cost	Small monthly charge plus charge for AutoClusters above initial credit. See website.	$ 5.00/month – Silver – DNAgedcom PC client $10.00/month – Gold – Additional support and documentation
Website	geneticaffairs.com	dnagedcom.com

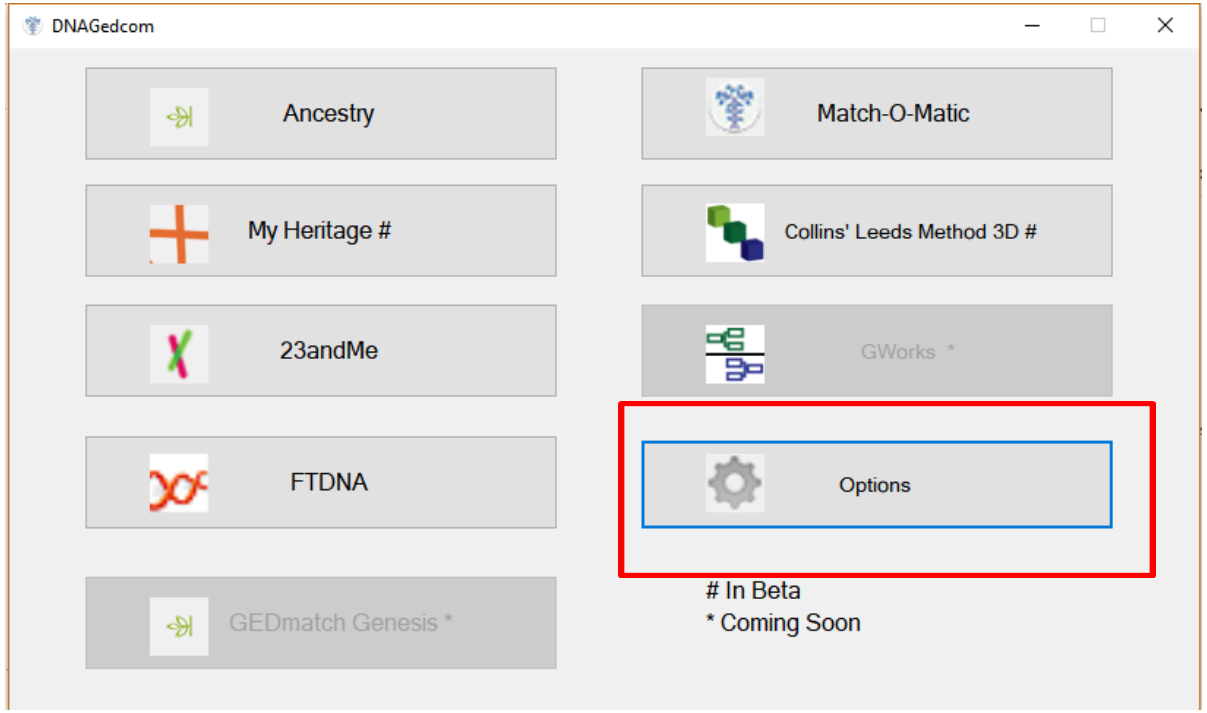
Getting Started with DNAgedcom

1. Go to DNAgedcom website for DNAgecom PC client for detailed instructions. Document includes links for registration, subscription, and PC client installation
2. Register at DNAgencom if not already registered
3. Subscribe at DNAgencom and pay fee to obtain DNAgedcom PC/Mac client software - required for clustering
4. Follow instructions and links in document to install DNAgedcom client on PC or Mac.
5. Following instructions are for PC client. Mac version has a few small differences.
6. DNAgedcom should place an shortcut icon on the PC desktop which you can click to start the application.
7. Enter your DNAgedcom Username and Password and click Login and Save All. Your username and password will be saved on your PC.



Preparing for Clustering with Ancestry

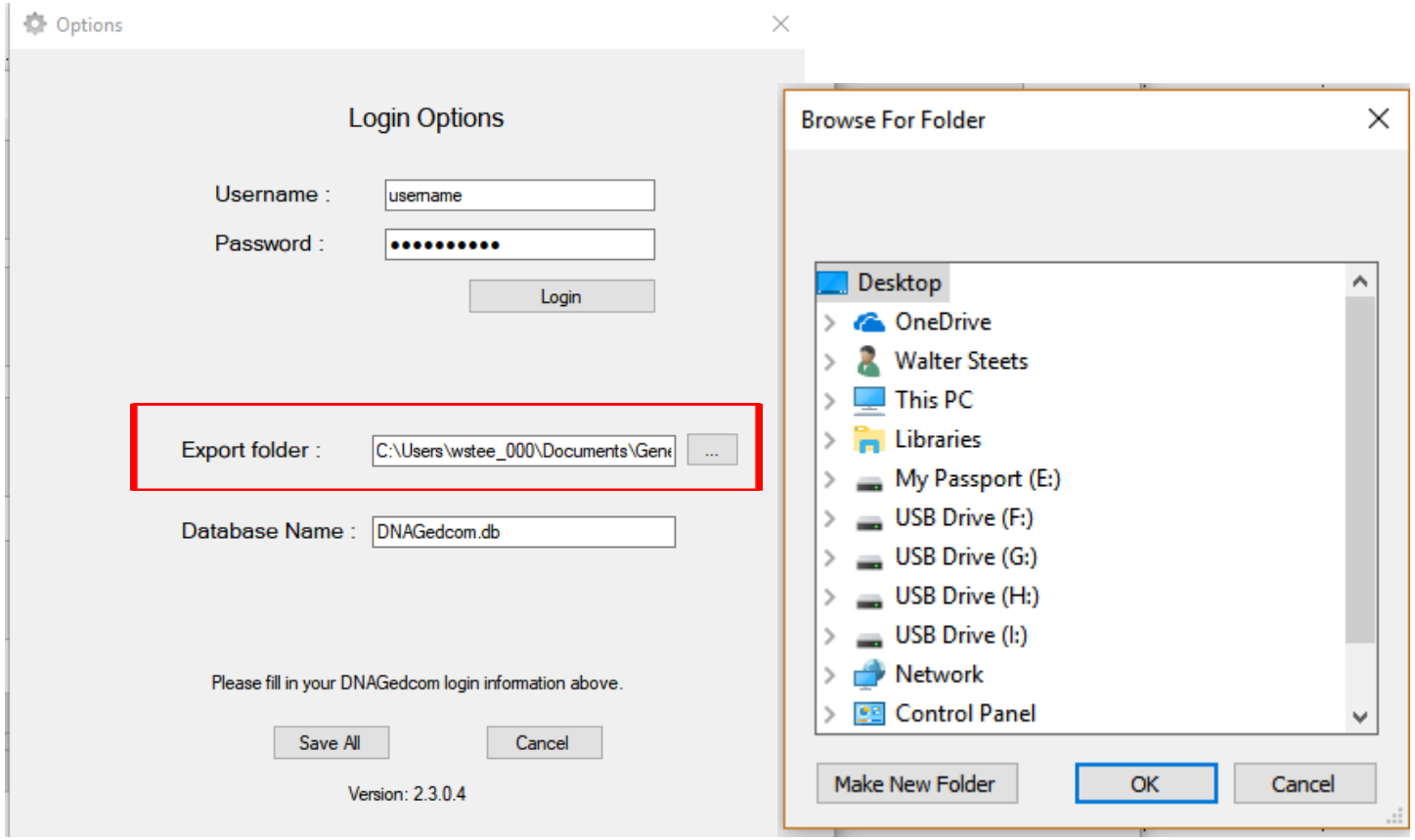
- DNAgedcom creates a number of files for each Ancestry kit which is analyzed.
- The files for each kit **must** be stored in separate Windows directories or the database needs be renamed for each new kit. Its easier to keep track of the new files generated by clustering if you create new directories.
- I find it easier to create new directories using Windows Explorer than to use the limited capability in DNAgedcom



1. Use Windows Create a Windows directory for each kit to be analyzed
2. If you've already opened DNAgedcom once and saved your username and password, start DNAgedcom
3. Click on Options

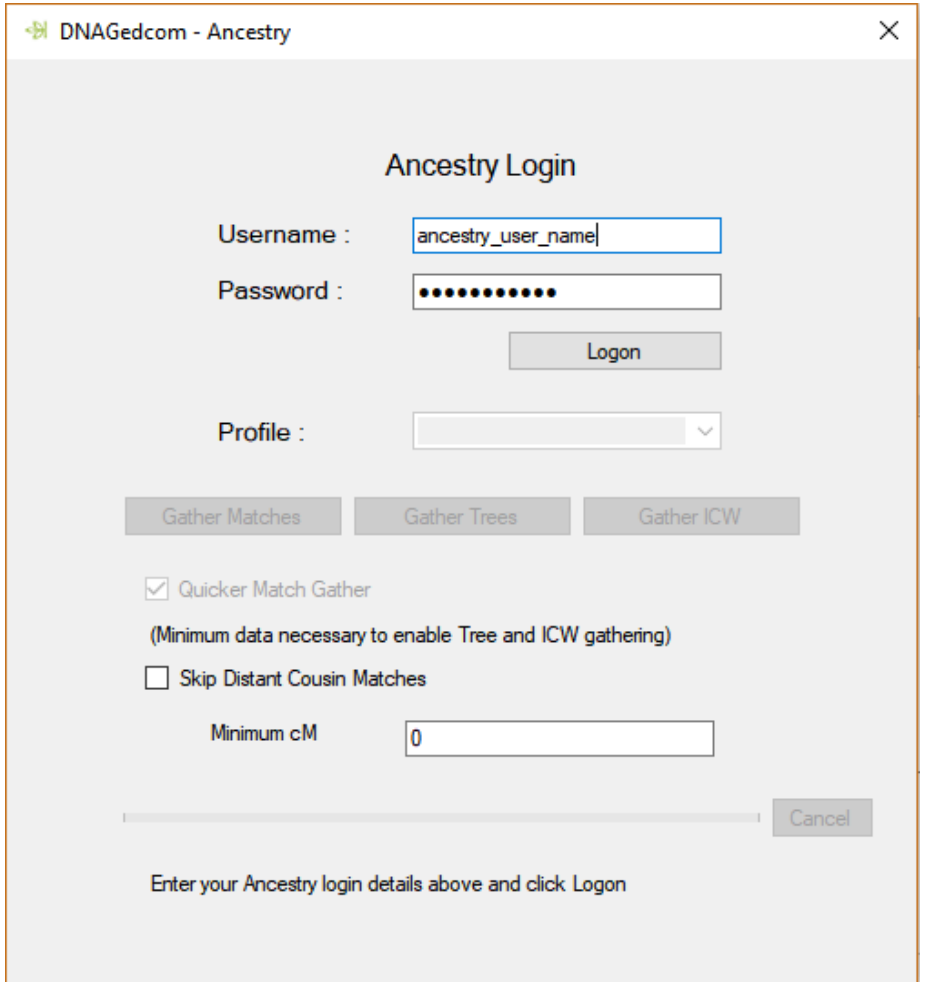
Preparing for Clustering with Ancestry (con't)

4. The Login Options should correct from initial login to DNAgedcom client.
5. In the Export folder field, click on the three dots to the right of the field and navigate to the directory for this AncestryDNA kit.
6. Don't change Database Name.
7. Click on Save All to save the name of the directory.



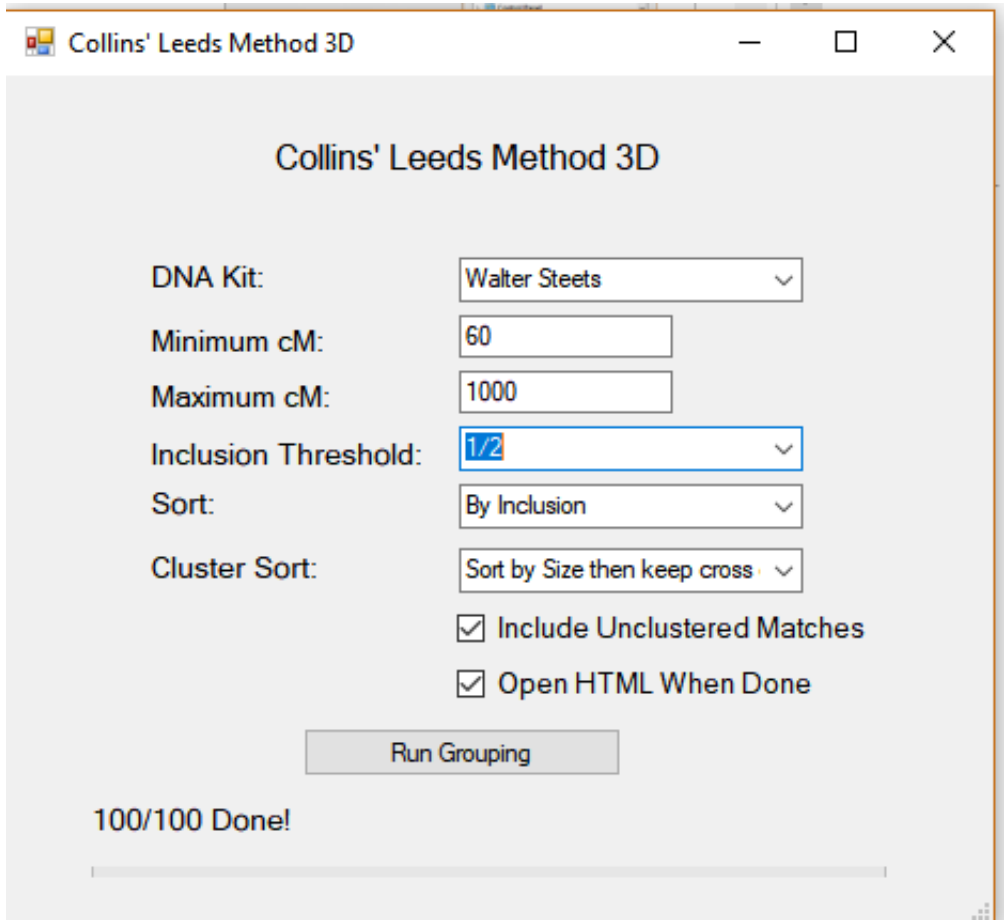
Downloading Ancestry Match and ICW Files

1. Click on Ancestry on DNAgedcom main menu
2. Enter your **Ancestry** Username and Password in the Ancestry Login fields and click Logon. This can be Ancestry username for a full or DNA-only Ancestry account.
3. When a name appears in the Profile field, you've successfully connected to Ancestry.
4. If you manage more than one Ancestry kit, select the one corresponding to the Windows directory specified in the DNAgedcom Options.
5. To gather data for clustering, the Quicker Match Gather and Skip Distant Cousin Matches can be checked. Leave Minimum cM at default of 0.
6. Click Gather Matches and wait one to two hours. A progress bar will appear. Don't turn off your PC.
7. Once Gather Matches has completed, click Gather ICW and wait another one to two hours. Don't turn off your PC.



Doing Clustering with DNAgedcom

1. On the DNAgedcom main menu, click on the Collins' Leeds Method 3D #
2. Fields
   i. DNA Kit: Choose the DNA kit corresponding to the directory specified on the DNA Options screen. This directory should contain the data downloaded from Ancestry.
   ii. Minimum cM and Maximum cM: Limits on amount of DNA Matches share with tester. DNAgedcom recommends using a max of around 400 cM. I prefer to use a max around 1200 to include first cousins to identify and group together maternal and paternal matches.
   iii. Inclusion Threshold: For a Match to be included in a cluster, the Match must share DNA with more than the Inclusion Threshold fraction of the other members of the cluster. Choosing a higher Threshold (e.g. 2/3) produces a larger number of small dense clusters



Doing Clustering with DNAgedcom (cont'd)

I. Fields (cont')

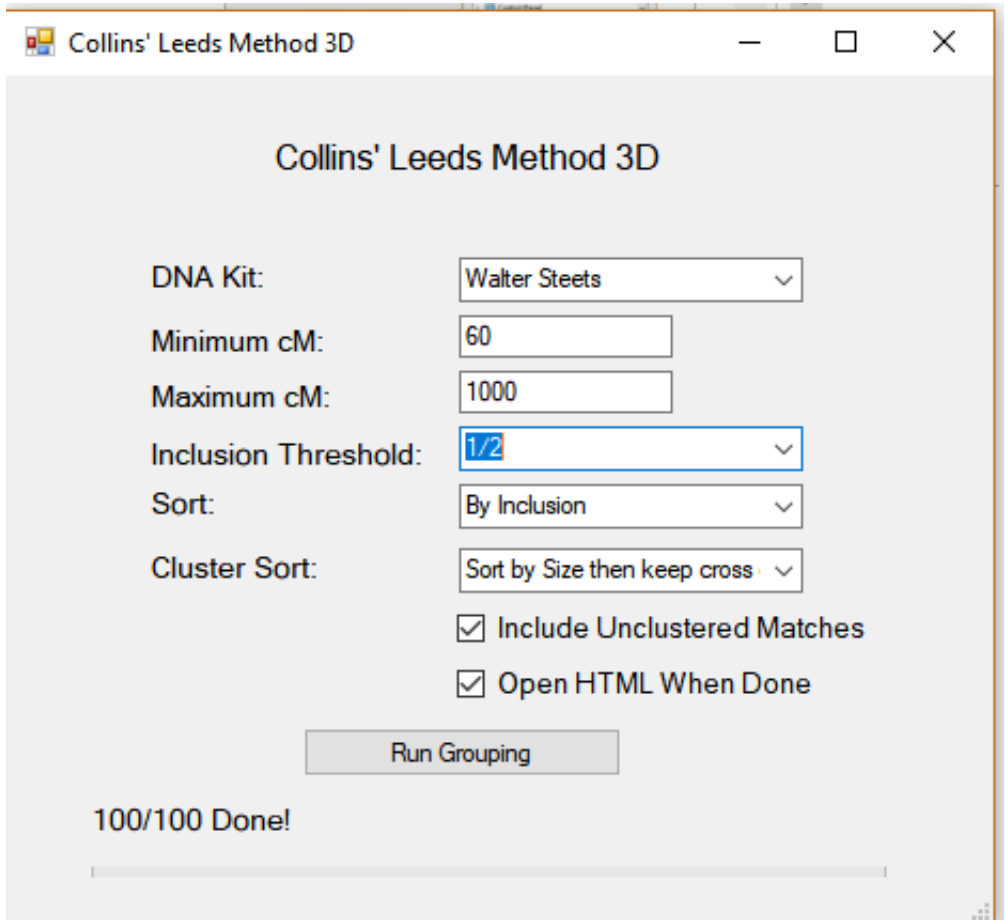
i. **Sort:** Determines how the Matches are sorts within the cluster. By Inclusion put Matches in the upper left quadrant which match the greatest number of other matches. By cM, puts the Matches with the largest amount of shared cM in the upper left of the cluster

ii. **Cluster Sort:** Four options for arranging the clusters. Choose either **Sort by Size then keep cross clusters together** or **Sort by max cM then keep cross clusters together.**

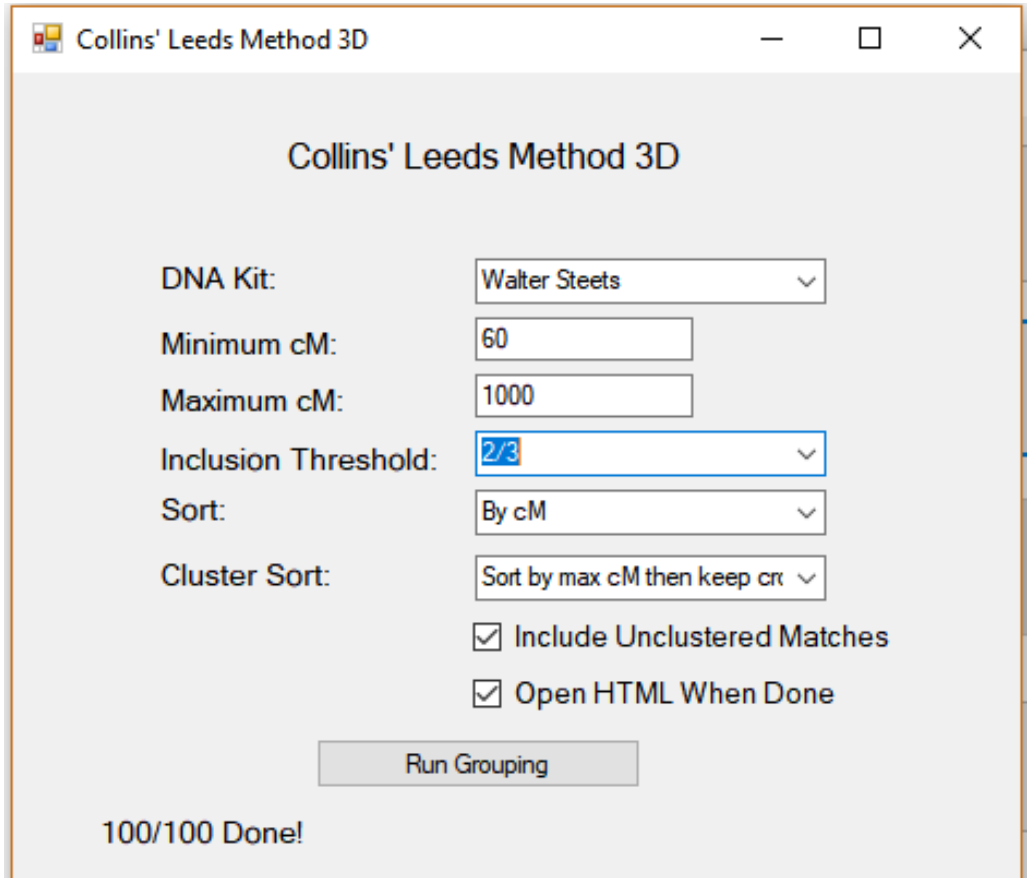
iii. **Include Unclustered Matches:** Check this box to include clusters with only a single Match. Generally probably should click on.

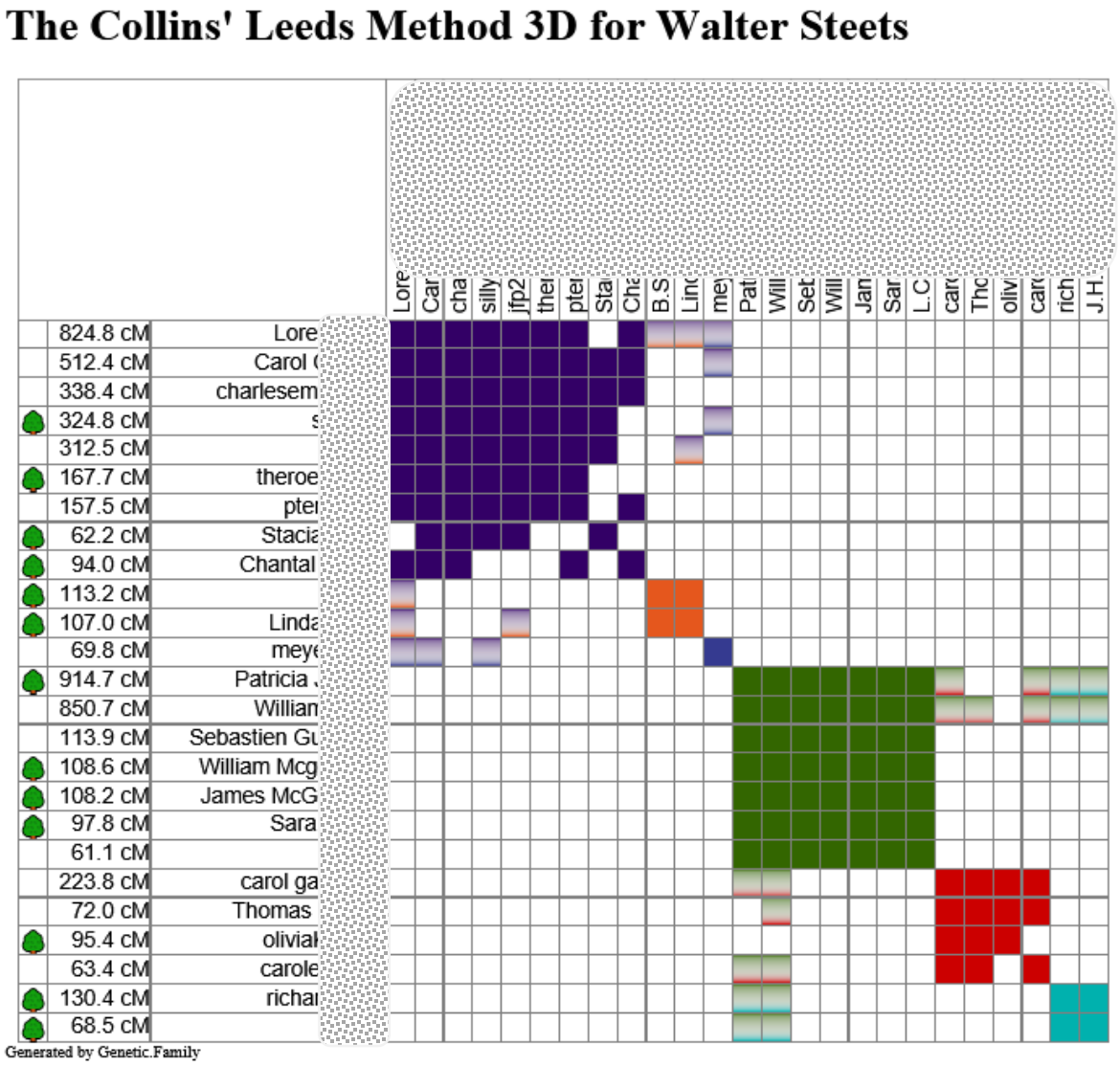
iv. **Open HTML When Done:** Displays Cluster Matrix in browser screen.

v. Run Grouping to do clusting.



Cluster Results



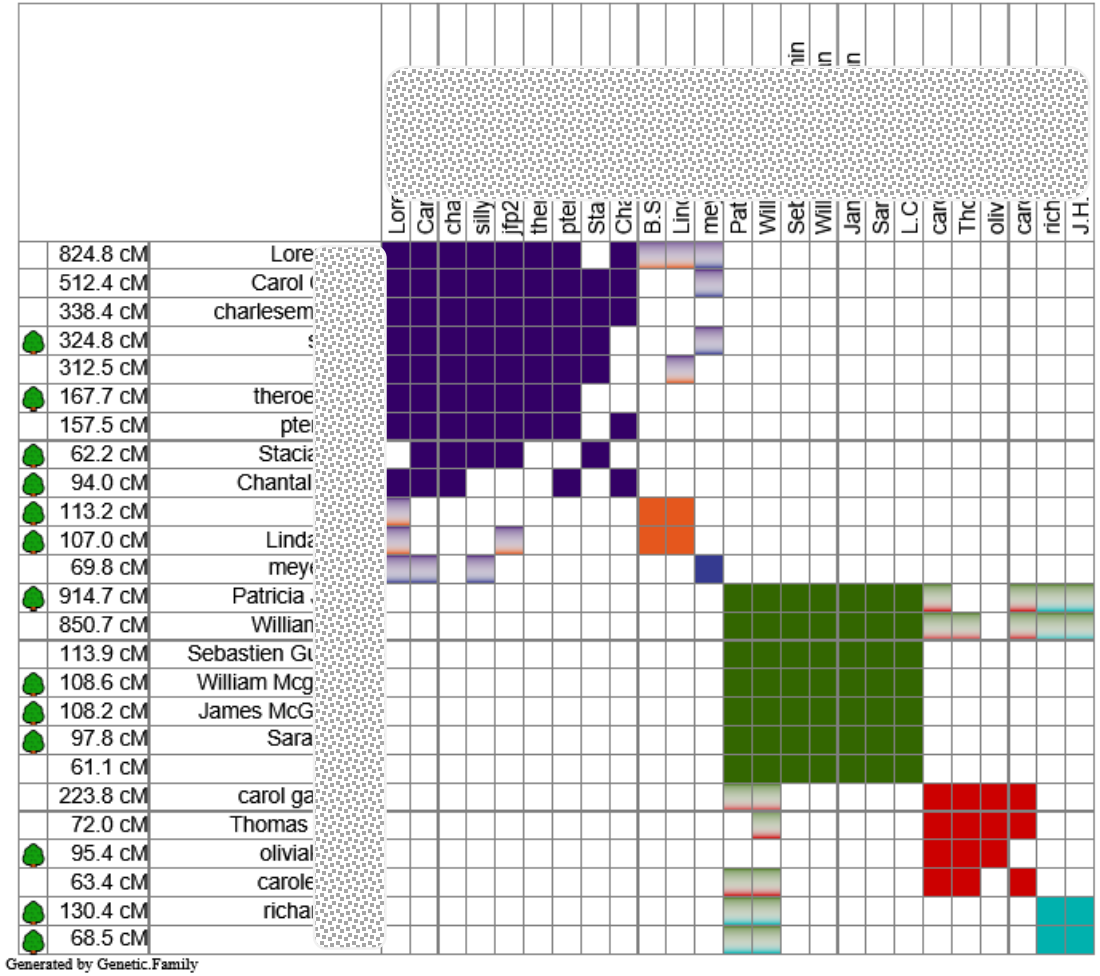


Comparison of DNAgedcom Matrix with Manual Cluster

- Results of Two Methods are nearly identical
- Interpretation of DNAgedcom results are same as for Manual method

Shared cM	Matchname	915	851	224	95.4	72	63.4	130	68.5	114	109	108	97.8	61.1	825	512	338	325	313	168	158	94	69.8	62.2	113	107
		Patri	Willi	caro	olivi	Thor	caro	richa	J.H.	Seba	Willi	Jame	Sara	L.C.	Lore	Caro	char	sillyr	jfp21	ther	pten	Char	mey	Staci	B.S.	Lind
914.7	Pat																									
850.7	Wil																									
223.8	car																									
95.4	oliv																									
72	Thc																									
63.4	car																									
130.4	rich																									
68.5	J.H.																									
113.9	Seb																									
108.6	Wil																									
108.2	Jan																									
97.8	Sar																									
61.1	L.C.																									
824.8	Lor																									
512.4	Car																									
338.4	cha																									
324.8	silly																									
312.5	jfp2																									
167.7	the																									
157.5	pte																									
94	Cha																									
69.8	me																									
62.2	Sta																									
113.2	B.S																									
107	Linc																									

The Collins' Leeds Method 3D for Walter Steets



Summary

- Benefits
  - Discovery process uses algorithms to find sets of Matches who have more matches with each other than with other Matches and so are likely to shared one or more MRCA's.
  - The cluster discovery process is only weakly dependent on amount of shared cM which vary widely for a given relationship. Clusters can find Shared Matches who have a common MRCA which would have been difficult to determine using only shared DNA.
  - Software-assisted clustering with parameters which allow the clustering process to be fine tuned (e.g. the Inclusion Threshold in DNAgedcom) can quickly find a helpful set of clusters
- Opportunities for Additional Development
  - Clustering should give good results for Matches covering a wide range of DNA – 20 to 2000 cM.
  - Clustering should be augmented with other genealogical input to provide more precise and identifiable clusters.
  - The effect of "half" relations (e.g. half 1C) need to be explored

Questions ?

Most Recent Common Ancestor and Shared DNA