

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 Matches
- Click on **Compare** to open report

★ 	Robert  <i>Possible range: 3rd - 4th cousins</i> Confidence: Extremely High  Shared DNA: 177 cM across 12 segments 	- No Trees	View Match Compare 
★ 	thero  168 centimorgans shared across 9 DNA segments #maternal #mcdermott daughter of Charles Marsh <i>Possible range: 3rd - 4th cousins</i> Confidence: Extremely High  Shared DNA: 168 cM across 9 segments 	- 2,331 people 	View Match Compare 
★ 	pt  157 centimorgans shared across 8 DNA segments #maternal <i>Possible range: 3rd - 4th cousins</i> Confidence: Extremely High  Shared DNA: 157 cM across 8 segments 	- No Trees	View Match Compare 



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



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

[View All >](#)

Shared Matches



Thomas

Immediate Family | 2,626 cM shared

53 people



M.H.

Immediate Family | 2,490 cM shared

Unlinked Tree



Lorena

1st Cousin | 825 cM shared

Unlinked Tree



jfp

2nd Cousin | 313 cM shared

Unlinked Tree



B.S.

3rd Cousin | 113 cM shared

2,633 people



Joshua

4th Cousin | 34 cM shared

103 people



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 IC – blue-grey labels
 - ii. Maternal IC – dark-pink labels
4. Use first cousins to identify paternal and maternal matches
 - i. Paternal – light-grey labels
 - ii. Maternal – light-pink labels

		915	851	825	512	338	325	313	224	168	158	130	114	113	109	108	107	97.8	95.4	94	72	69.8	68.5	63.4	62.2	61.1
Shared cM	Matchname	Patric	Willia	Loren	Carol	charle	sillyno	jfp212	carol	theroe	ptenc	richar	Sebas	B.S.	Willia	James	Linda	Sara S	olivia	Chant	Thom	meyer	J.H.	carole	Stacia	L.C.
914.7	Patric	Grey	Blue						Blue			Blue	Blue		Blue	Blue		Blue					Blue	Blue		Blue
850.7	Willia	Blue	Grey						Blue						Blue	Blue		Blue					Blue	Blue		Blue
824.8	Loren			Grey	Blue	Blue	Blue	Blue		Blue	Blue			Blue						Blue		Blue				
512.4	Carol				Grey	Blue	Blue	Blue		Blue	Blue									Blue		Blue				
338.4	charle					Grey	Blue	Blue		Blue	Blue									Blue		Blue				
324.8	sillyno						Grey	Blue		Blue	Blue										Blue					
312.5	jfp212							Grey		Blue	Blue						Blue									
223.8	carol	Blue	Blue						Grey									Blue			Blue			Blue		
167.7	theroe			Blue	Blue	Blue	Blue	Blue		Grey	Blue															
157.5	ptenc			Blue	Blue	Blue	Blue	Blue			Grey									Blue						
130.4	richar	Blue	Blue									Grey											Blue			
113.9	Sebas	Blue	Blue										Grey		Blue	Blue		Blue								Blue
113.2	B.S.			Blue										Grey			Blue									
108.6	Willia	Blue	Blue												Grey	Blue		Blue								Blue
108.2	James	Blue	Blue													Grey		Blue								Blue
107	Linda			Blue				Blue									Grey									
97.8	Sara S	Blue	Blue															Grey								Blue
95.4	olivia								Blue										Grey		Blue					
94	Chant			Blue	Blue	Blue	Blue	Blue												Grey						
72	Thom		Blue						Blue												Grey			Blue		
69.8	meyer	Blue	Blue																			Grey				
68.5	J.H.	Blue	Blue																				Grey			
63.4	carole	Blue	Blue																					Grey		
62.2	Stacia			Blue	Blue	Blue	Blue	Blue																	Grey	
61.1	L.C.	Blue	Blue																							Grey



Clustering Example – Group Paternal and Maternal Matches

1. Separate paternal and maternal matches by sorting colors in following order:

- i. Blue-grey for paternal IC's.
- ii. Light-grey for remaining paternal matches
- iii. Dark-pink for maternal IC'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**

		915	851	224	130	114	109	108	97.8	95.4	72	68.5	63.4	61.1	825	512	338	325	313	168	158	113	107	94	69.8	62.2
Shared cM	Matchname	[Redacted]																								
		Patric	Willia	carol	richar	Sebas	Willia	James	Sara S	olivia	Thom	J.H.	carol	L.C.	Loren	Carol	charle	sillync	jfp21	thero	ptenc	B.S.	Linda	Chan	meye	Stacia
914.7	Patric	Blue-grey																								
850.7	Willia		Blue-grey																							
223.8	carol			Blue-grey																						
130.4	richar				Blue-grey																					
113.9	Sebas					Blue-grey																				
108.6	Willia						Blue-grey																			
108.2	James							Blue-grey																		
97.8	Sara S								Blue-grey																	
95.4	olivia									Blue-grey																
72	Thom										Blue-grey															
68.5	J.H.											Blue-grey														
63.4	carol												Blue-grey													
61.1	L.C.													Blue-grey												
824.8	Loren														Dark-pink											
512.4	Carol															Dark-pink										
338.4	charle																Dark-pink									
324.8	sillync																	Dark-pink								
312.5	jfp21																		Dark-pink							
167.7	thero																			Dark-pink						
157.5	ptenc																				Dark-pink					
113.2	B.S.																					Dark-pink				
107	Linda																						Dark-pink			
94	Chan																							Dark-pink		
69.8	meye																								Dark-pink	
62.2	Stacia																									Dark-pink



Clustering Example - Define Clusters

I. 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.

		915	851	224	130	114	109	108	97.8	95.4	72	68.5	63.4	61.1	825	512	338	325	313	168	158	113	107	94	69.8	62.2	
Shared cM	Matchname																										
		Patri	Willi	carol	richa	Seba	Willi	Jame	Sara	olivi	Thor	J.H.	caro	L.C.	Lore	Caro	char	sillyr	jfp21	ther	pten	B.S.	Linda	Char	mey	Staci	
914.7	Patric																										
850.7	Willia																										
223.8	carol g																										
130.4	richar																										
113.9	Sebast																										
108.6	Willia																										
108.2	James																										
97.8	Sara S																										
95.4	olivial																										
72	Thoma																										
68.5	J.H.																										
63.4	carole																										
61.1	L.C.																										
824.8	Lorena																										
512.4	Carol C																										
338.4	charle																										
324.8	sillync																										
312.5	jfp212																										
167.7	theroc																										
157.5	ptencl																										
113.2	B.S.																										
107	Linda C																										
94	Chant																										
69.8	meyer																										
62.2	Stacia																										



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

		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
Shared cM	Matchname	Patric	Willia	carol	olivial	Thom	carole	richar	J.H.	Sebas	Willia	James	Sara S	L.C.	Loren	Carol	charle	sillync	jfp212	thero	ptenc	Chant	imeyer	Stacia	B.S.	Linda
914.7	Patric																									
850.7	Willia																									
223.8	carol																									
95.4	olivial																									
72	Thom																									
63.4	carole																									
130.4	richar																									
68.5	J.H.																									
113.9	Sebas																									
108.6	Willia																									
108.2	James																									
97.8	Sara S																									
61.1	L.C.																									
824.8	Loren																									
512.4	Carol																									
338.4	charle																									
324.8	sillync																									
312.5	jfp212																									
167.7	thero																									
157.5	ptenc																									
94	Chant																									
69.8	imeyer																									
62.2	Stacia																									
113.2	B.S.																									
107	Linda																									



Clustering Example – Interpretation of Clusters

I. Paternal Cluster

i. Blue-grey:

- a. Shared cM consistent with IC
- b. Cluster members matches almost all lower-level clusters

ii. Light-grey, purple, light-blue:

- a. Shared cM consistent with 2C, 2CIR, 3C, 3CIR
- b. Three paternal clusters which have no shared matches between them making 2CIR (father's 2C) and 3C more likely

2. Maternal Clusters

i. Pink:

- a. Shared cM consistent with IC, 2C, and 3C
- b. Two matches with smallest shared cM (Jeff Meyers and Stacia Milius, likely 3C) could have separate MRCA's

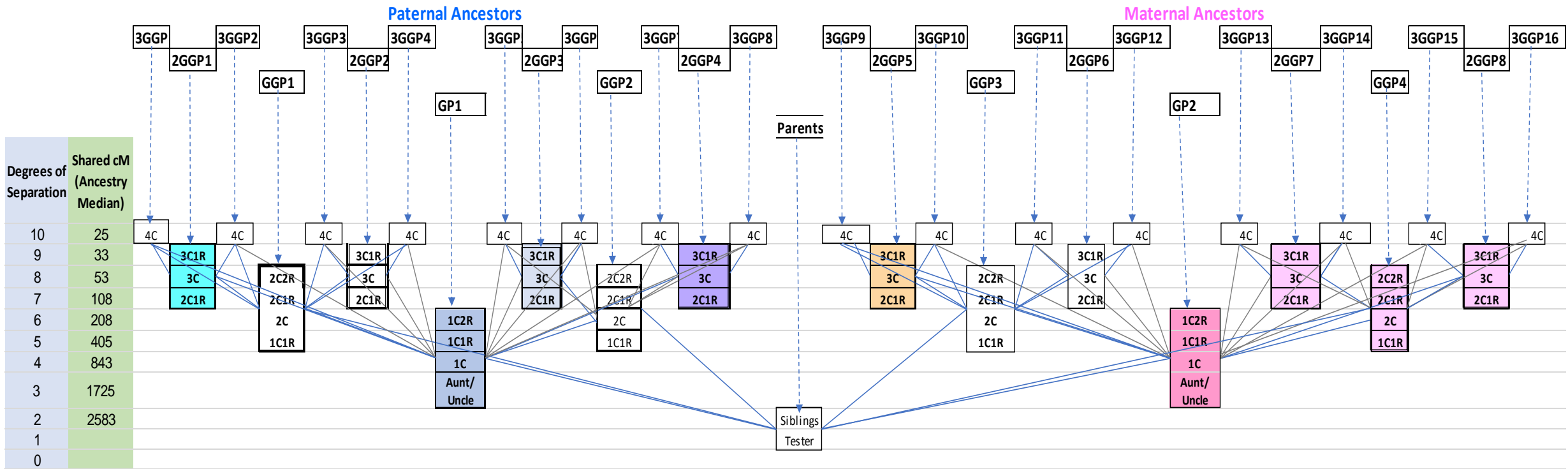
ii. Tan:

- a. Match only IC and ICIR (ICs son)
- b. Likely 3C on separate line from Pink

		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		
Shared cM	Matchname	[Redacted]																										
		Patrici	Willia	carol	olivia	Thom	carole	richar	J.H.	Sebas	Willia	James	Sara S	L.C.	Loren	Carol	charle	sillyno	jfp212	thero	ptenc	Chant	meyer	Stacia	B.S.	LindaC		
914.7	Patrici																											
850.7	Willia																											
223.8	carol																											
95.4	olivia																											
72	Thom																											
63.4	carole																											
130.4	richar																											
68.5	J.H.																											
113.9	Sebas																											
108.6	Willia																											
108.2	James																											
97.8	Sara S																											
61.1	L.C.																											
824.8	Loren																											
512.4	Carol																											
338.4	charle																											
324.8	sillyno																											
312.5	jfp212																											
167.7	thero																											
157.5	ptenc																											
94	Chant																											
69.8	meyer																											
62.2	Stacia																											
113.2	B.S.																											
107	LindaC																											



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 DNAGedcom PC client for detailed instructions. Document includes links for registration, subscription, and PC client installation
2. Register at DNAGedcom if not already registered
3. Subscribe at DNAGedcom 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.

Login Options

Username :

Password :

Login

Export folder : ...

Database Name :

Please fill in your DNAGedcom login information above.

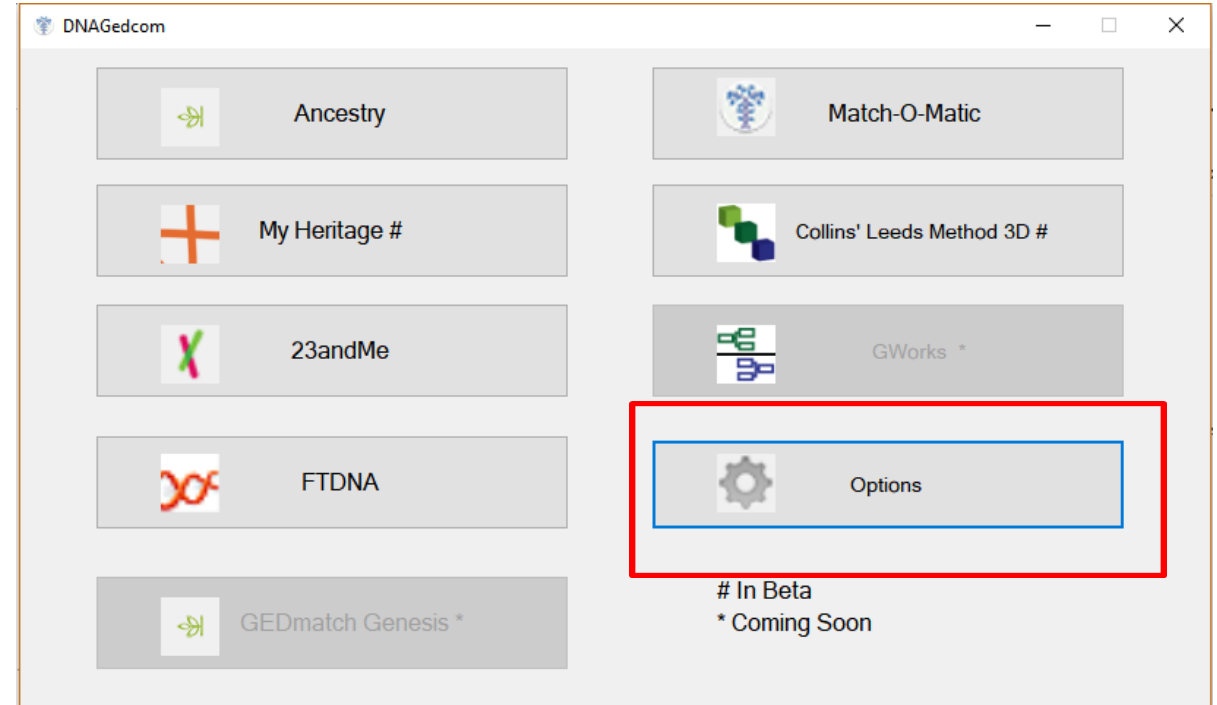
Save All Cancel

Version: 2.3.0.4



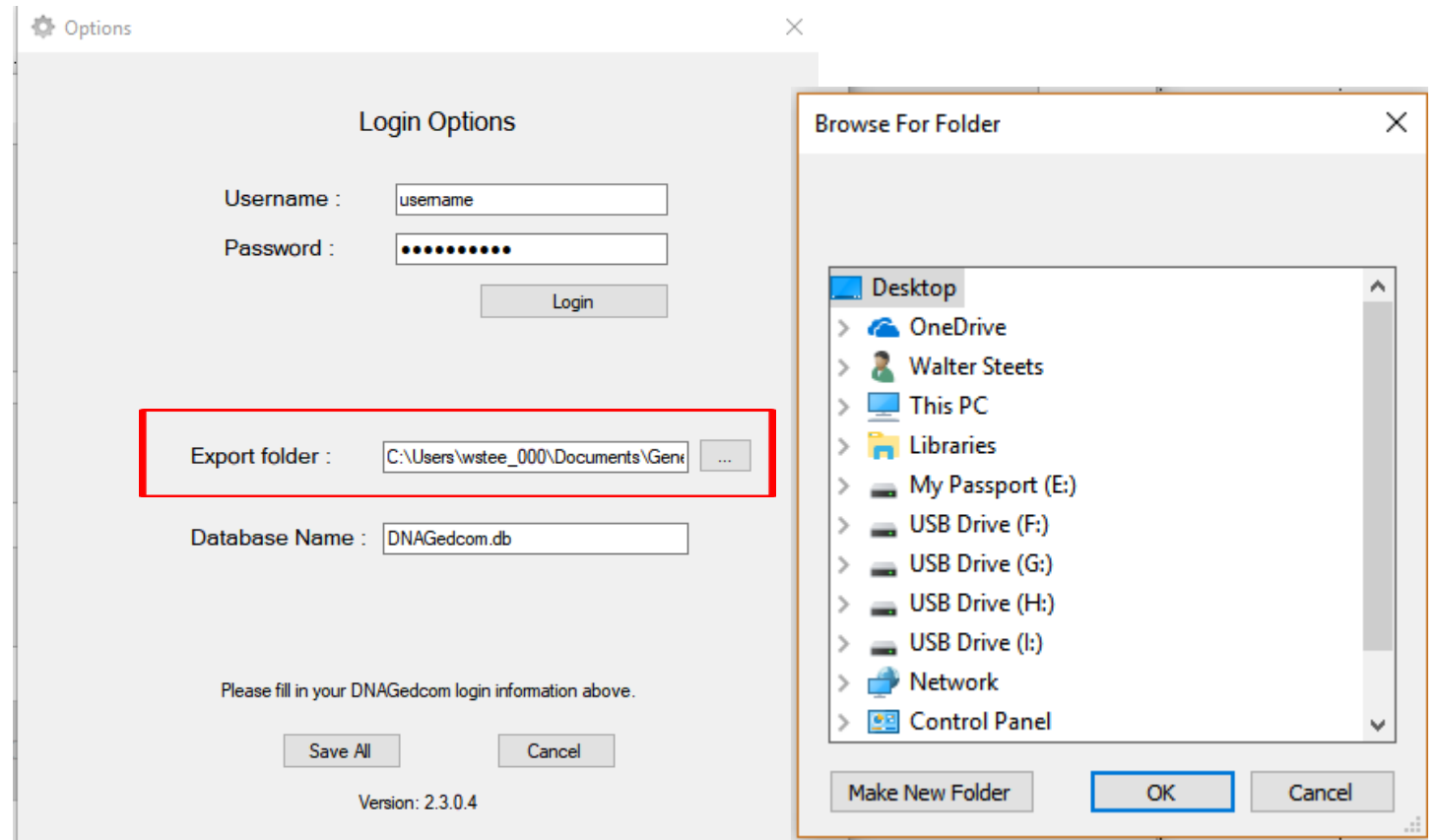
Preparing for Clustering with Ancestry

- 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.

DNAGedcom - Ancestry

Ancestry Login

Username : ancestry_user_name

Password :

Logon

Profile : [dropdown]

Gather Matches Gather Trees Gather ICW

Quicker Match Gather
(Minimum data necessary to enable Tree and ICW gathering)

Skip Distant Cousin Matches

Minimum cM 0

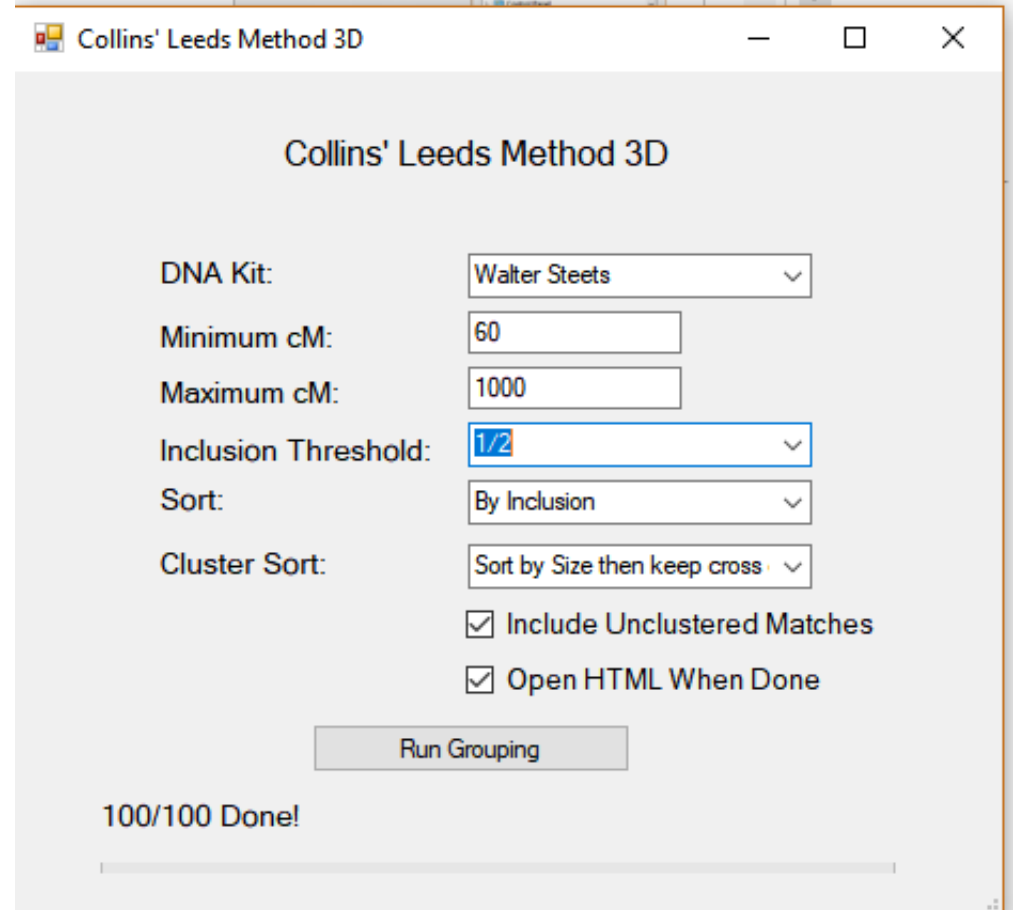
Cancel

Enter your Ancestry login details above and click Logon



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



The screenshot shows a window titled "Collins' Leeds Method 3D" with the following settings:

- DNA Kit: Walter Steets (dropdown menu)
- Minimum cM: 60 (text input)
- Maximum cM: 1000 (text input)
- Inclusion Threshold: 1/2 (dropdown menu)
- Sort: By Inclusion (dropdown menu)
- Cluster Sort: Sort by Size then keep cross (dropdown menu)
- Include Unclustered Matches
- Open HTML When Done

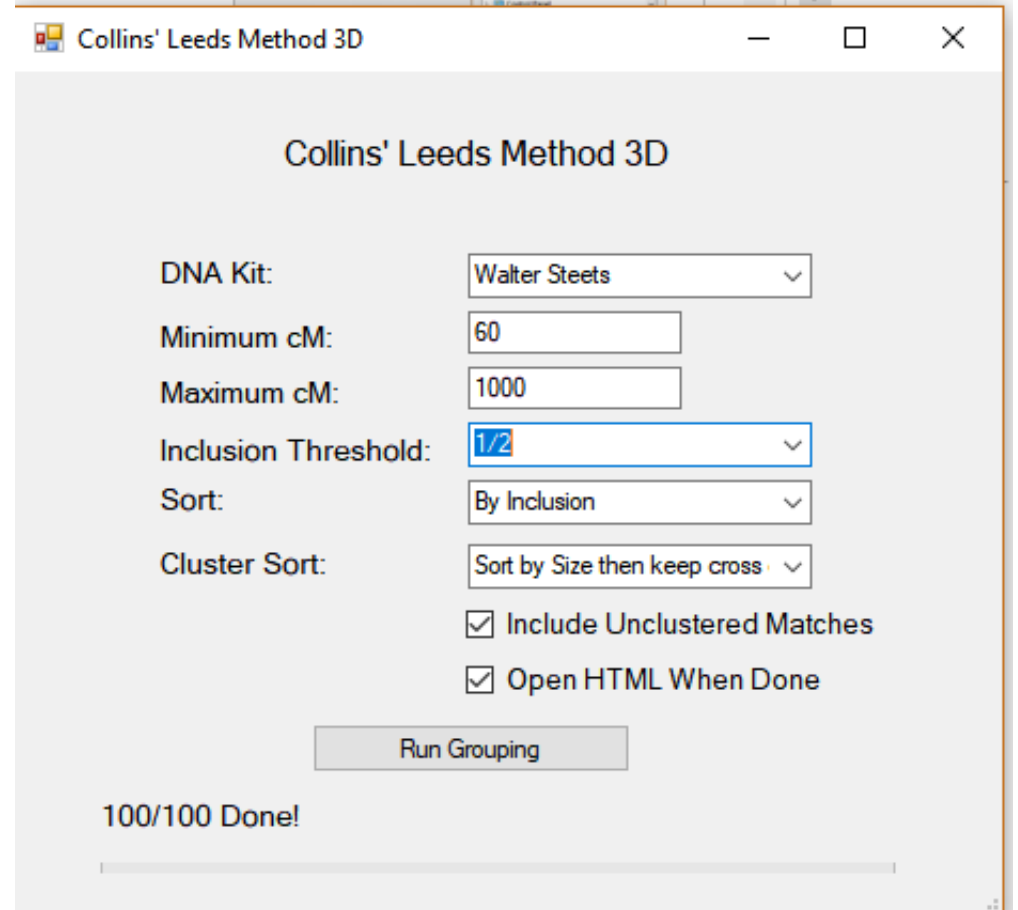
A "Run Grouping" button is located below the settings. At the bottom of the window, it displays "100/100 Done!" and a progress bar.



Doing Clustering with DNAGEDCOM (cont'd)

I. Fields (cont')

- i. **Sort:** Determines how the Matches are sorted 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 clustering.



Collins' Leeds Method 3D

DNA Kit:

Minimum cM:

Maximum cM:

Inclusion Threshold:

Sort:

Cluster Sort:

Include Unclustered Matches

Open HTML When Done

100/100 Done!



Cluster Results

The Collins' Leeds Method 3D for Walter Steets

Collins' Leeds Method 3D

DNA Kit:

Minimum cM:

Maximum cM:

Inclusion Threshold:

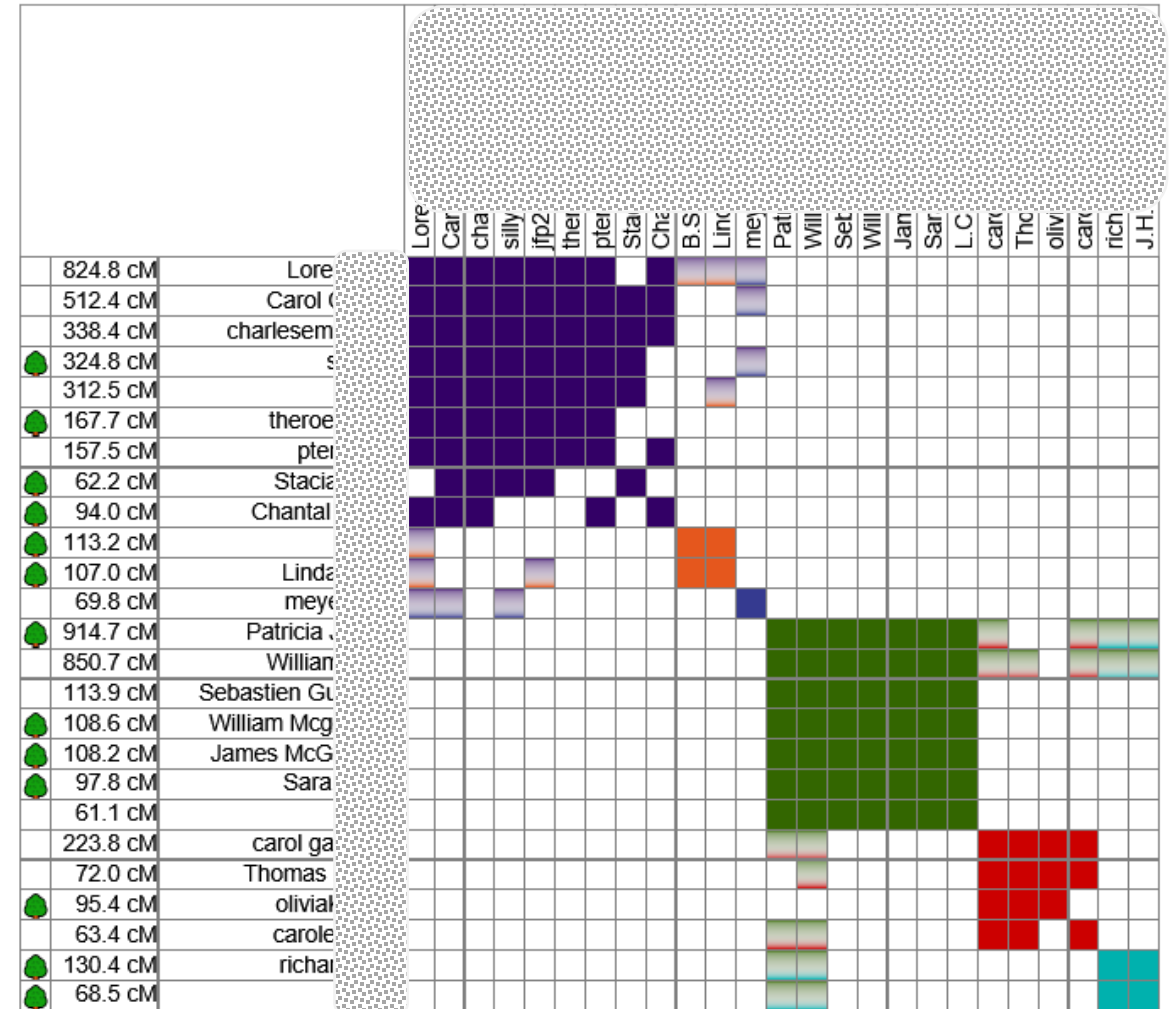
Sort:

Cluster Sort:

Include Unclustered Matches

Open HTML When Done

100/100 Done!



Generated by Genetic.Family

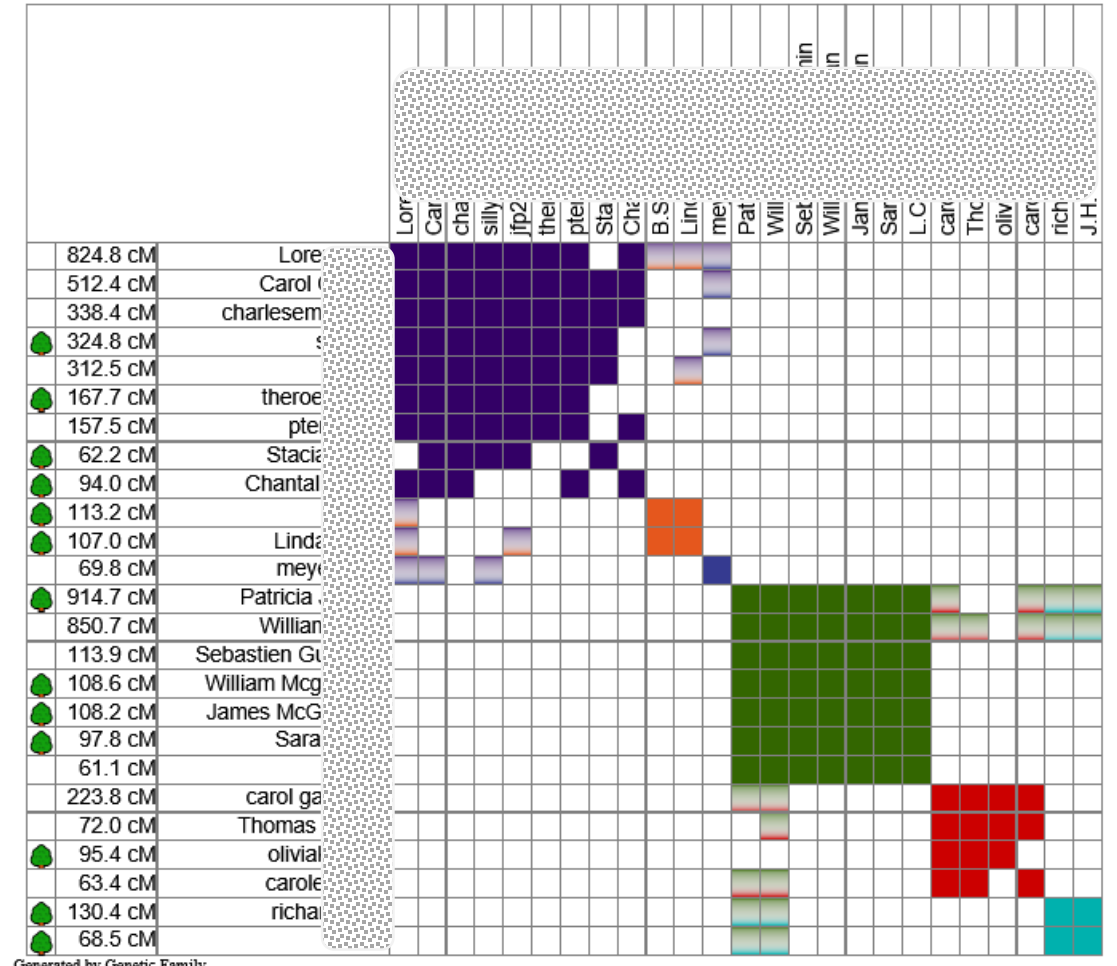


Comparison of DNAGEDCOM Matrix with Manual Cluster

- 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
		Patr	Willi	caro	olivi	Thor	caro	richa	J.H.	Seba	Willi	Jame	Sara	L.C.	Lore	Caro	char	sillyr	jfp21	theri	pten	Char	meyi	Staci	B.S.	Lindt.
914.7	Pat																									
850.7	Wil																									
223.8	car																									
95.4	olivi																									
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	Lindt																									

The Collins' Leeds Method 3D for Walter Steets



Generated by Genetic.Family



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 IC) need to be explored



Questions ?



Most Recent Common Ancestor and Shared DNA

