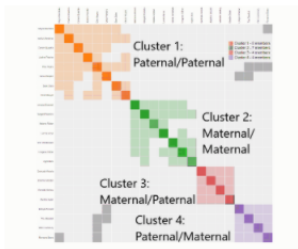




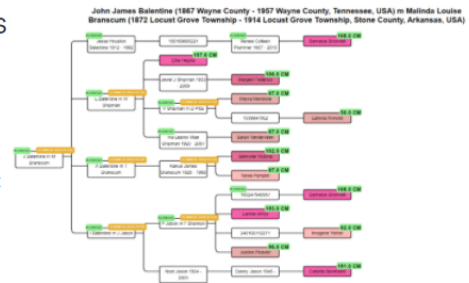
# MANUAL

## GENETIC AFFAIRS



### Unleash your DNA cousins

Genetic Affairs offers the [AutoCluster](#), [AutoFastCluster](#), [AutoSegment](#), [AutoTree](#), [AutoPedigree](#) and [hybrid AutoSegment](#) tools which groups together your DNA matches into clusters of matches that most likely descend from common ancestors. [AutoCluster](#), [AutoPedigree](#) and [AutoTree](#) can help in the identification of unknown ancestors (for instance an unknown great grandfather) or assist persons with unknown parentage to their birth families (e.g., adoptees or donor-conceived persons).



### How to start an analysis on Genetic Affairs



1. Register



2. Add website



3. Select profiles



4. Receive results in mail



5. Analyze results



DECEMBER 5, 2020

GENETIC AFFAIRS

[www.geneticaffairs.com](http://www.geneticaffairs.com)

## Table of Contents

Features of Genetic Affairs .....	4
Registration .....	6
Adding a website .....	8
Profiles – starting an analysis .....	12
AutoScan view .....	13
AutoScan - e-mail updates and notifications .....	13
AutoCluster analysis .....	16
Start an AutoCluster analysis .....	19
AutoFastCluster analysis .....	20
AutoCluster analysis using the extend cluster feature .....	22
Start an FTDNA AutoCluster analysis .....	24
DNA segment browser for 23andme and FTDNA AutoCluster analyses .....	24
Start an 23andme AutoCluster analysis .....	27
FTDNA and 23andme group-like AutoCluster analysis .....	29
Enriched Surnames and Locations for 23andme analyses .....	30
Detailed information concerning Enriched Surnames and Locations .....	31
Rule-based AutoCluster .....	32
AutoTree .....	36
AutoPedigree .....	39
AutoSegment – segment based clustering .....	44
.....	46
AutoSegment concepts .....	47
GEDmatch triangulation data .....	55
Excel cluster representation .....	56
Paternal & Maternal annotations .....	56
AutoSegment – retrieve offline data .....	58
Retrieving segment data for MyHeritage .....	58
Retrieving segment data for FamilyTreeDNA .....	60
Retrieving segment data for 23andme .....	62
Retrieving segment data for GEDmatch .....	64
Hybrid AutoSegment – combine MyHeritage, FTDNA, 23andme and GEDmatch .....	68

Additional settings..... 74  
User settings and Payments..... 75  
Blog posts and Facebook groups ..... 76  
Other AutoCluster implementations. .... 78  
Prices..... 80  
Troubleshooting ..... 82

## Introduction

This document describes the use of the website <https://www.geneticaffairs.com> and more specifically, the use of the member section which can be reached under <https://members.geneticaffairs.com>

The original purpose of Genetic Affairs was the automation of the retrieval of genetic matches from two major DNA testing companies: 23andme and FamilyTreeDNA (FTDNA). This feature, also known as **AutoScan**, allows users to register their DNA accounts and provides an interface to select which profiles updates should be provided, how often and which criteria should be applied.

Nowadays, the most important feature of Genetic Affairs is the automated clustering of shared matches, named **AutoCluster**. AutoCluster organizes your matches into shared match clusters that likely represent branches of your family.

The **rule-based AutoCluster** clustering allows for the clustering of matches obtained from different profiles by employing certain rules, for instance only using matches that do not match the matches of a known biological mother.

The **AutoTree** feature identifies common ancestors and reconstructed trees for FTDNA profiles.

In addition to the shared match clustering offered by **AutoCluster**, a new clustering that is now available that is based on overlapping segments. This clustering is available under the name **AutoSegment**. Another version of this tool, **hybrid AutoSegment** is available and allows for the clustering of MyHeritage, 23andme, FTDNA and GEDmatch data based on overlapping segments.



## Features of Genetic Affairs

The next sections will describe in more detail the various features of Genetic Affairs.

Genetic Affairs **AutoScan**<sup>™</sup> provides the following features:

- Support for two DNA testing companies (23andme and FTDNA).
- Updates concerning new matches are provided in a clear e-mail message.
- Adjustable update interval – updates can be provided daily, weekly or monthly.
- Minimal DNA match – notifications are given for specific DNA matches, e.g., minimal 3<sup>rd</sup> cousin.
- Minimal centimorgan threshold, for instance only report 4<sup>th</sup> cousins which share at least 40 cM.
- All first matches in a spreadsheet. The first analysis of a profile will include all genetic matches in a spreadsheet. This spreadsheet is added to the mail as an attachment.

Genetic Affairs **AutoCluster**<sup>™</sup> provides the following features:

- Automatic clustering of shared matches using adjustable cM ranges
- Enriched surnames/locations in the clusters of AutoCluster for 23andme profiles
- Adjustable cM limits for the largest segment for FTDNA analyses
- Adjustable cM limits for shared cM between shared matches for 23andme analyses
- Ability to perform AutoClustering using common matches from 23andme that share a segment (and which will most likely triangulate).
- Extended clustering where the shared matches from the initial analysis (given a cM range or starred matches or groups) are used as matches.
- Segment data for FTDNA and 23andme analyses as well as a chromosome browser overview.

Genetic Affairs **AutoFastClust** provides the following features:

- Instant clustering of shared matches using adjustable cM ranges
- User entered match and shared match data, saved in local storage or CSV files

Genetic Affairs **rule-based AutoCluster** provides the following features:

- Use rules to filter and/or merge their matches using matches from other profiles to focus on a particular branch of your ancestors, for instance, your paternal matches by excluding your known maternal matches.
- Three different rules allow for:
  - the exclusion (NOT rule)
  - inclusion (AND rule)
  - combination (OR rule)
- Additional visualization feature indicates which (shared) matches are added as compared to the matches of the primary profile that is used.

Genetic Affairs **AutoTree™** provides the following features for **FTDNA** profiles:

- Identification of common ancestors from trees of users by employing a three-step clustering
  - First, a surname clustering is performed
  - Second, a first name clustering is employed on the surname clusters
  - Finally, using birth and death year information, the final common ancestor clusters are determined
- Automatic reconstruction and visualization of a genealogical tree using the identified common ancestors and DNA matches
- Works for adoptees and profiles with an attached tree
- Identification of common locations using a distance clustering of birth locations of tree persons

Genetic Affairs **AutoPedigree™** provides the following features for **FTDNA** profiles:

- automates the generation and testing of hypotheses using reconstructed trees from AutoTree.
- developed to identify how a person, for instance an adoptee, fits into a reconstructed AutoTree

Genetic Affairs **AutoSegment™** provides the following features:

- perform a DNA segment-based clustering method that is available for segment data from MyHeritage, 23andme, FamilyTreeDNA and GEDmatch
- Requires segment data downloaded from testing companies
- Does not require credential information
- Allows for filtering of segments located in known pile-up regions.
- Allows for easy integration into the DNA Painter website

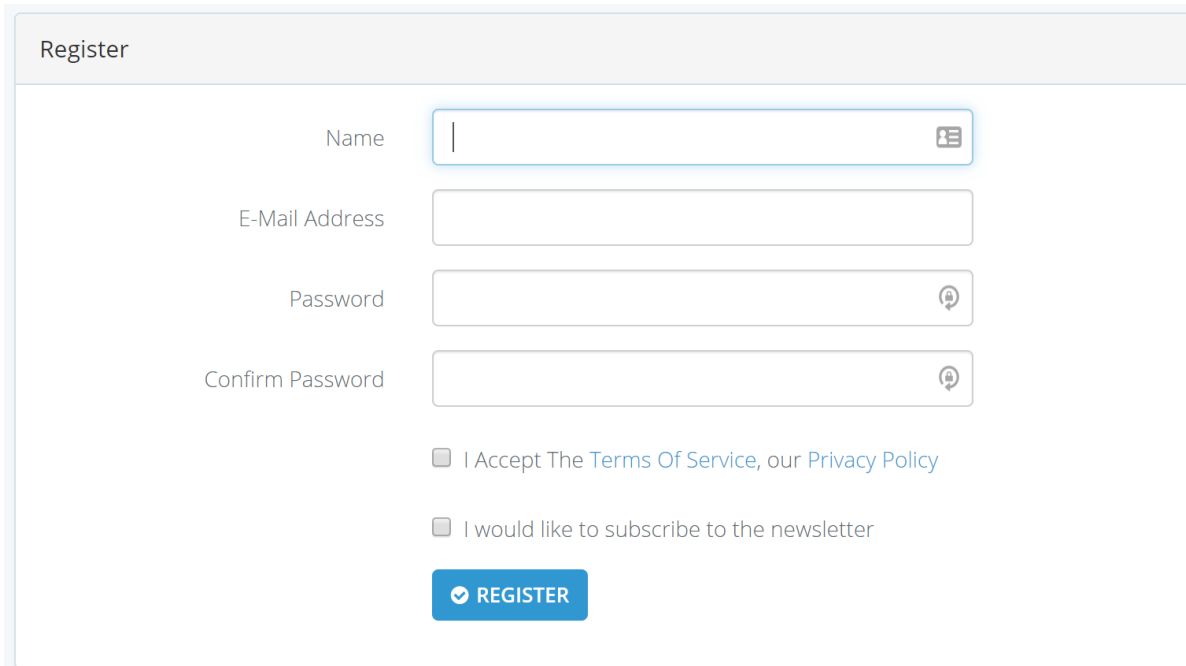
Genetic Affairs **Hybrid AutoSegment™** provides the following features:

- perform a DNA segment-based clustering method combining segment data from MyHeritage, 23andme, FamilyTreeDNA and GEDmatch into a single analysis
- Requires segment data downloaded from testing companies
- Does not require credential information
- Allows for filtering of segments located in known pile-up regions.
- Allows for easy integration into the DNA Painter website
- Allows for leftover procedure to improve segment locations for FTDNA data

Last, it is possible to recluster old AutoCluster or MyHeritage AutoCluster analyses.

## Registration

The registration page (<https://members.geneticaffairs.com/register>) shows information that is required for registration at the site (see Figure 1).



The registration page features a form with the following elements:

- Name:** A text input field with a small icon on the right.
- E-Mail Address:** A text input field.
- Password:** A text input field with a small icon on the right.
- Confirm Password:** A text input field with a small icon on the right.
- I Accept The [Terms Of Service](#), our [Privacy Policy](#)
- I would like to subscribe to the newsletter
- REGISTER:** A blue button with a white checkmark icon and the text "REGISTER".

Figure 1. Registration page of the website <https://members.geneticaffairs.com/register>

After registration, the user is redirected to the landing page (see Figure 2). The landing page offers several options. The “show websites” button shows all registered 23andme and FTDNA websites. By clicking on the “Register a new website”, users can register new FTDNA or 23andme websites, using your login credentials. This step is required for the regular AutoCluster analyses.

The “Run AutoCluster” will show all registered 23andme and FTDNA websites. The “Run AutoTree” button will display all registered FTDNA websites since the AutoTree feature is only available for FTDNA profiles.

The “Run AutoSegment” option will show another page with four options, an AutoSegment analysis for MyHeritage, 23andme, FTDNA or GEDmatch. The “Run hybrid AutoSegment” link will allow users to run a combined analysis using data from the aforementioned companies.

The “Run AutoCluster using CSV files” feature allows users to run a clustering analysis using two CSV files, one containing match data and one containing shared matches data. A similar feature is provided by the “Run online AutoFastCluster” where these datasets can be entered manually. The “Recluster MyHeritage AutoClusters” option allows you to recluster old AutoCluster or MyHeritage cluster results.

The “User settings” option will show the different options available for a user. Archived newsletters are available using the “Newsletter archive” button.

The “Show credits” button displays the amount of available credits. Credits can be acquired using the “Subscription” option.

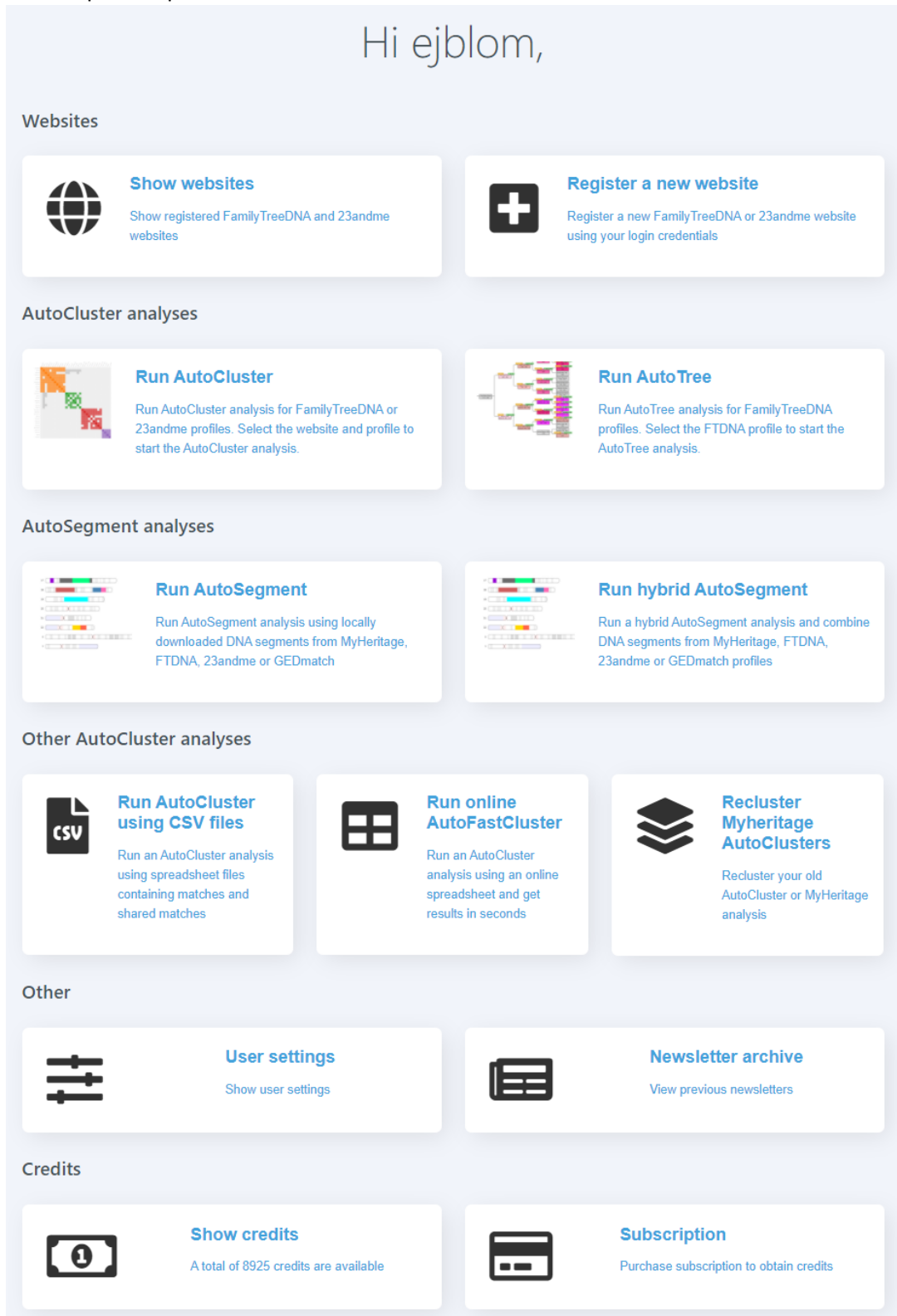


Figure 2. Landing page after registration.

## Adding a website

Upon clicking on the link: “Register a new website”, the user is redirected to a page that allows adding of website credentials (see Figure 3).

Hi ejblom,  
Add a new website account for 23andme or FamilyTreeDNA.

**23andme**  
Upon clicking on the link underneath, a login page will be displayed that requires the login information (Username or email and password) of the 23andme account. Every entered 23andme account can hold several profiles.

**FamilyTreeDNA**  
Upon clicking on the link underneath, a login page will be displayed that requires the login information (Kit Number or GAP Username and password) of the FamilyTreeDNA account. Every entered FamilyTreeDNA account is linked to a single profile.

[Add 23andme account](#)

[Add FamilyTreeDNA account](#)

Figure 3 - Landing page for adding a new website

When the “Add 23andme account or add FamilyTreeDNA” button is selected, a login page is displayed that will allow the entry of the website credentials (see Figure 4 for an example of FTDNA)

Add new website

Company	FamilyTreeDNA
Login:	login
Password:	*****
Retype password:	*****

[ADD NEW WEBSITE](#)

Figure 4. Website to add FTDNA credentials.

Upon entering the website credentials, a confirmation message will appear after pressing the “Add new website” (see Figure 5). **Note that for FTDNA profiles the KIT id is required for the login field.**

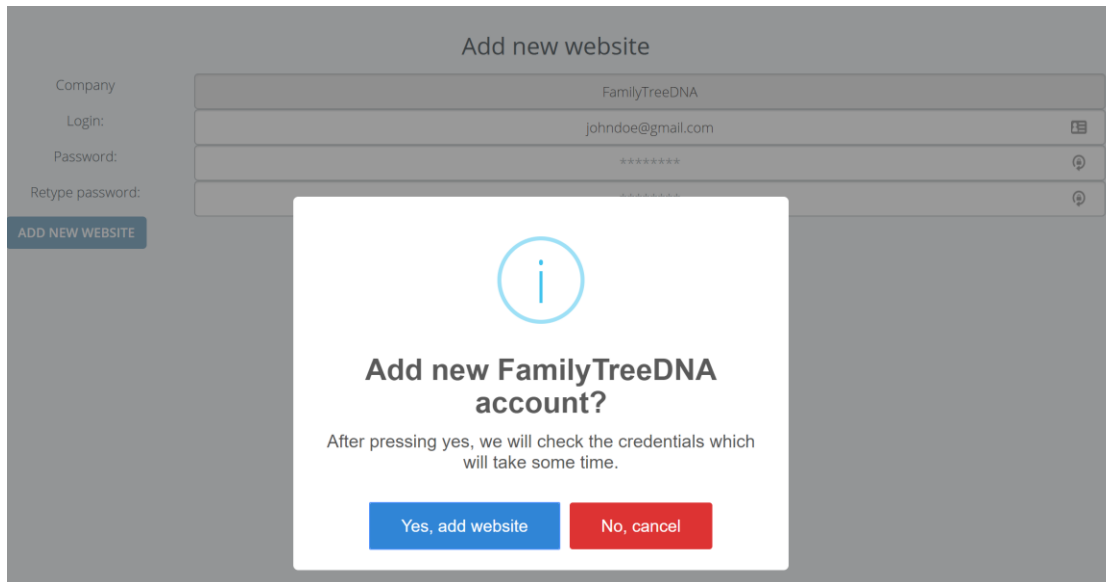


Figure 5. Information pane that will appear after pressing the "Add new website" button.

After the "Yes add website" is selected, Genetic Affairs will use the supplied user credentials to check if these are valid. During this verification, a message is displayed (see Figure 6). See our online FAQ section (<https://www.geneticaffairs.com/faq.html>) for more information concerning the privacy of your login credentials.

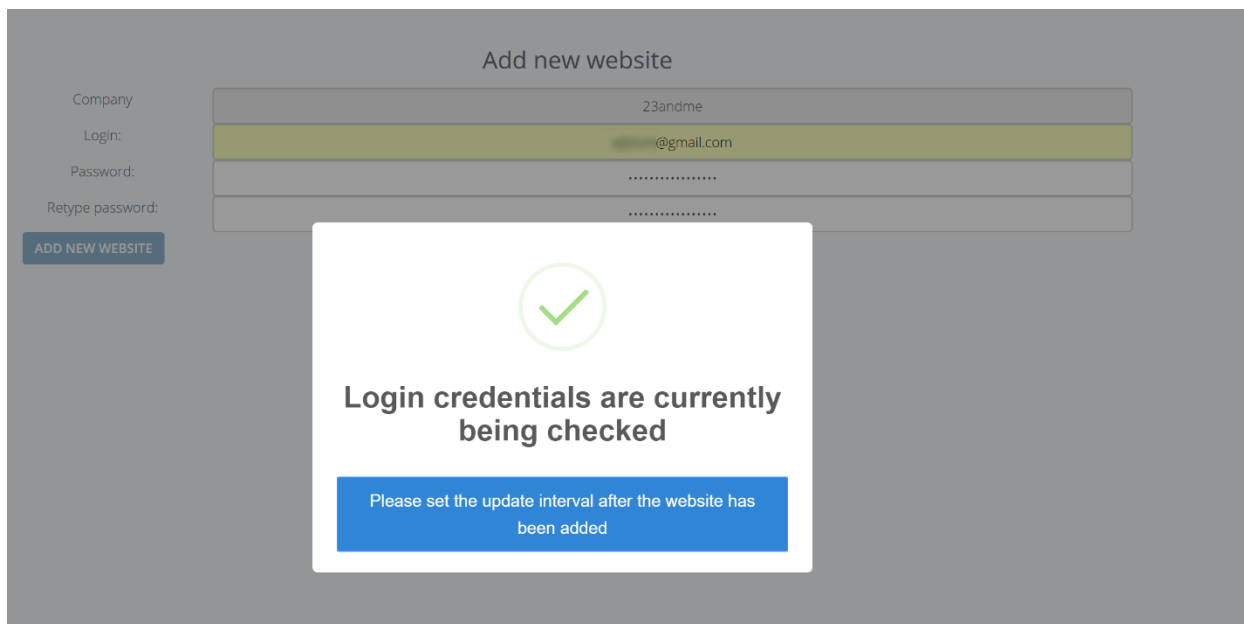


Figure 6. The message is displayed while your login credentials are verified.

**Note that browsers sometimes try to fill in credential data in the login fields. If entering credential information is not successful on multiple occasions, please try using another browser.**

The websites page will be displayed if the credentials are successfully tested or if the websites button is selected (see Figure 7).

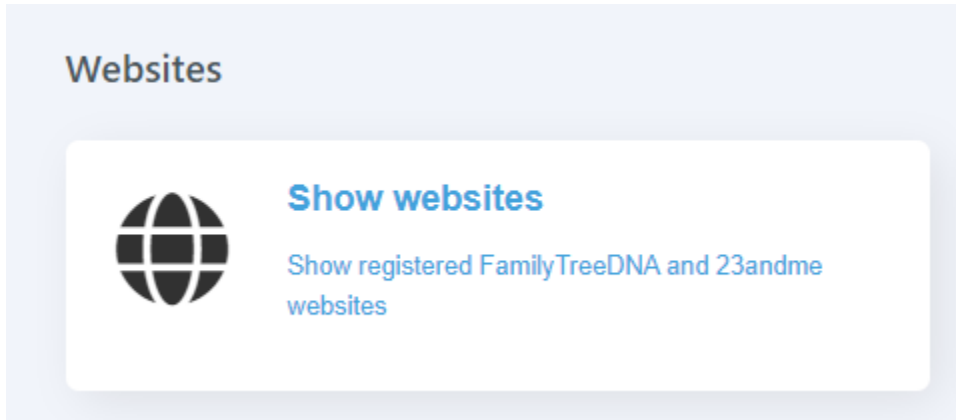


Figure 7. Website button on the main page

The websites page shows the registered FTDNA and 23andme websites (see Figure 8).

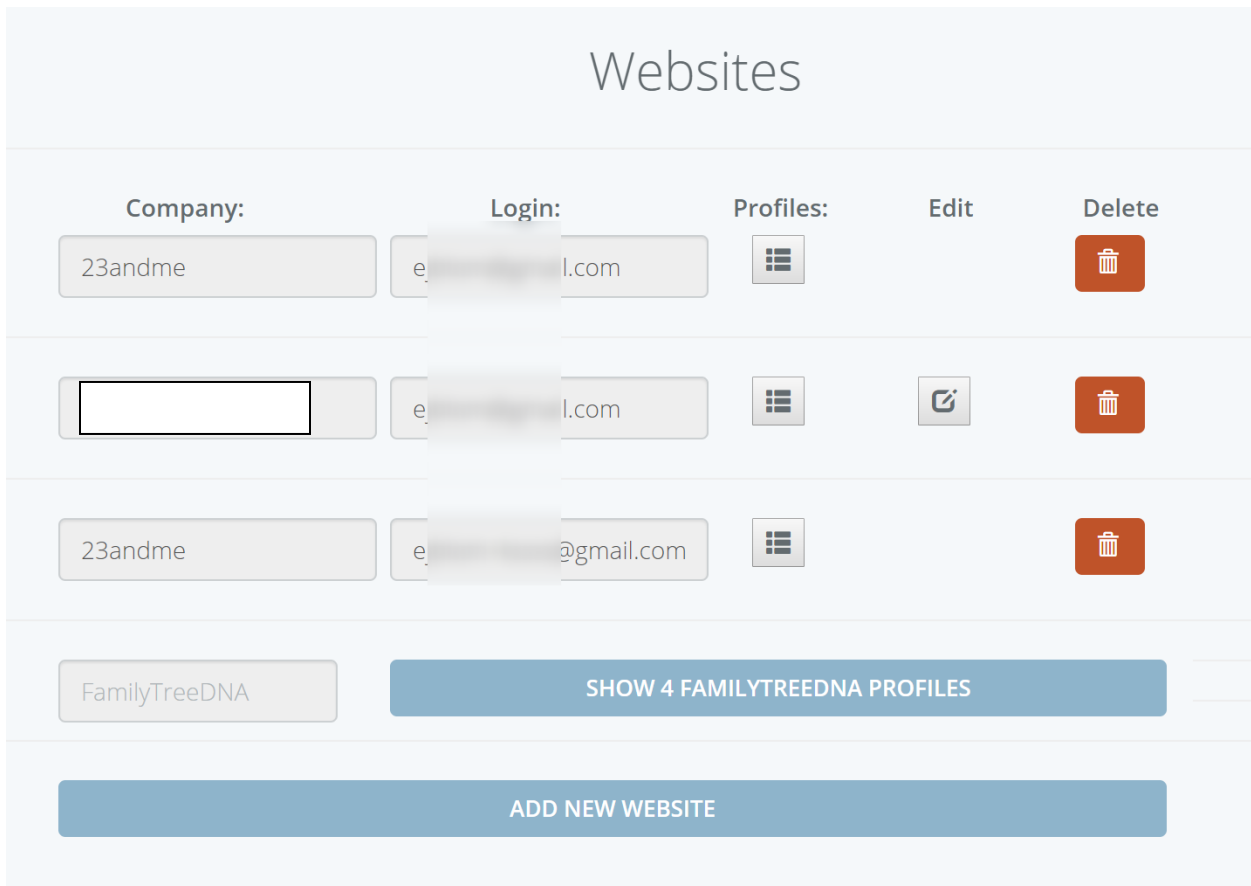


Figure 8. The main view of all the websites. FTDNA websites are grouped on a different page since they are linked to a single profile for each login.

The red button behind the 23andme websites will delete the website credentials as well as the underlying profiles and updates. The edit button can be used to change the login credentials of the website, for instance after you have changed its password.

The Profiles button will bring up the profiles for a 23andme website (for FTDNA, the website view will display a profile since each FTDNA website is linked to a single profile). The Profile view and their settings will be discussed in the next section.

Note that by default for FTDNA profiles the **KIT** identifier is shown. If multiple kits are managed, tracking the different can become difficult. It is now possible to rename the profile name to something that is easier to remember.



## Profiles – starting an analysis

After clicking on the “Profiles” button for a certain website (see Figure 8) all profiles are displayed that are linked to that website login (see Figure 9). Note that there can be several profiles linked to a single 23andme website but only one FTDNA profile linked to a registered FTDNA website.

Several analysis options are available which will be discussed below.

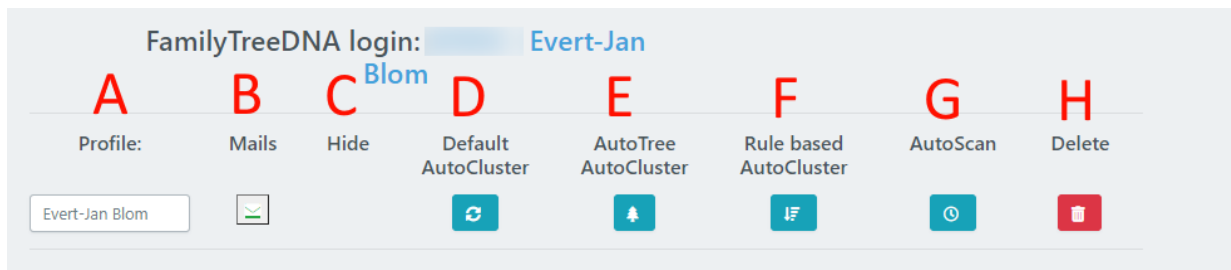


Figure 9. Profiles view for a FTDNA website that allows the selection of a specific analysis.

- A) Shows the name that is linked to the profile. This name can be edited, for instance, to replace the FTDNA kit id number to the actual profile name.
- B) When updates are available this icon turns green, when selected the latest updates are shown.
- C) Hide profile (only for 23andme profiles). When selected another unhide button will appear.
- D) AutoCluster analysis. Opens a new page to perform an AutoCluster analysis.
- E) AutoTree analysis. Opens an AutoCluster analysis page with some options already enabled to perform an AutoTree analysis (see AutoTree).
- F) Rule-based AutoCluster analysis. Opens a new page to perform a rule-based AutoCluster analysis (see Rule-based AutoCluster).
- G) AutoScan settings. Opens a new page to adjust the settings for an AutoScan analysis (see next section and Figure 10).
- H) Delete a profile. This will also delete all linked messages

## AutoScan view

By selecting the AutoScan icon (see Figure 9F), the AutoScan view is displayed. The AutoScan analysis allows users to obtain regular updates in a single email for new matches for different websites. Note that changed settings under B, C, D, and E are saved automatically.

23andme login: [redacted]

A	B	C	D	E
Profile:	Update interval:	Minimal DNA match	Minimal cM	Mails
[redacted]	weekly	4th Cousin	9 cM	<input checked="" type="checkbox"/>
[redacted]	weekly	4th Cousin	9 cM	<input checked="" type="checkbox"/>
[redacted]	monthly	4th Cousin	9 cM	<input checked="" type="checkbox"/>

All 3 profiles    leave uncha    leave uncha    leave uncha    Save all settings

Retrieve/Update new profiles for this website

Figure 10. View of all profiles linked to a website for the AutoScan analysis.

- A) Shows the name that is linked to the profile.
- B) Sets the update interval (*e.g.*; never, weekly and monthly updates). By default, all update intervals are set to never.
- C) The minimal DNA match that should be reported in the update mail. Options range from first cousins to distant cousins.
- D) A minimal cM threshold that should be employed on top of the minimal DNA match that is applied in C). This option is for instance used to only report fourth cousins that share a minimum of 40 cM.
- E) When updates are available this icon turns green, when selected the latest updates are shown.
- F) When one needs to change all settings simultaneously for all profiles, use this option. Set the option(s) that need to be changed and press the “Save all settings” button. Note that previous settings are kept if the “leave unchanged” setting for a field is selected.
- G) Retrieve new profiles for a 23andme website.

## AutoScan - e-mail updates and notifications

A first analysis to obtain the existing DNA matches will be performed once a day for a profile after an update interval has been set and saved. This analysis downloads all matches for FTDNA and 23andme profiles. After a successful analysis, an email (see Figure 11) will be sent that holds the top 20 DNA matches as well as an attachment with all matches. This comma-separated attachment can be imported in a spreadsheet application (such as Microsoft Excel or Google Sheets).

For profile [redacted] A first Ancestry search has been performed. Inbox x

Genetic Affairs <info@geneticaffairs.com>

to me ▾

Hello [redacted]

For profile [redacted] : A first Ancestry search has been performed.

The closest 20 DNA matches are listed underneath.

[redacted] total: 3457.2 cM and is a predicted parent/child with a confidence value of 100.0%.  
[redacted] shares in total: 3430.2 cM and is a predicted parent/child with a confidence value of 100.0%.  
[redacted] predicted third cousin with a confidence value of 100.0%.  
[redacted] shares in total: 157.9 cM and is a predicted third cousin with a confidence value of 100.0%.  
[redacted] is a predicted fourth cousin with a confidence value of 98.3%.  
[redacted] shares in total: 59.0 cM and is a predicted fourth cousin with a confidence value of 98.3%.  
[redacted] is a predicted fourth cousin with a confidence value of 96.7%.  
[redacted] is a predicted fourth cousin with a confidence value of 95.2%.  
[redacted] predicted fourth cousin with a confidence value of 95.2%.  
[redacted] predicted fourth cousin with a confidence value of 92.9%.  
[redacted] predicted fourth cousin with a confidence value of 91.4%.  
[redacted] shares in total: 35.6 cM and is a predicted fourth cousin with a confidence value of 89.9%.  
[redacted] total: 35.4 cM and is a predicted fourth cousin with a confidence value of 89.6%.  
[redacted] a predicted fourth cousin with a confidence value of 89.3%.  
[redacted] 4.8 cM and is a predicted fourth cousin with a confidence value of 89.1%.  
[redacted] is a predicted fourth cousin with a confidence value of 87.9%.  
[redacted] predicted fourth cousin with a confidence value of 87.4%.  
[redacted] a predicted fourth cousin with a confidence value of 87.2%.  
[redacted] shares a predicted fourth cousin with a confidence value of 84.8%.  
[redacted] predicted fourth cousin with a confidence value of 84.1%.

See the attached CSV file for an overview of the remainder of the matches. Future updates for this profile will provided on a daily basis.



Figure 11. Example of a first email. In this example, the comma separated files are displayed that contain all matches

Future weekly or monthly updates are combined into a single e-mail of which the subject contains the number of new matches that passed the minimum match (for instance 4<sup>th</sup> cousins that share at least 30 cM) criteria. The body of the e-mail contains more detailed information for each profile that was checked for updates. An example of such an e-mail is shown in Figure 12.

In addition to the mail updates, we use the notifications on the website to display the updates as well (see Figure 74. General settings from top-menu).

14 DNA matches identified for 6 profile updates Inbox x  



Genetic Affairs via amazonses.com  
to me ▾

10:55 AM (7 hours ago) ☆ ↶ ⋮

Hello ejblom,

Placed underneath are the individual update(s):

**██████████** : A weekly FamilyTreeDNA search has been performed and identified a total of 1332 DNA matches. Of the 5 new DNA matches, one match is a DNA match that is predicted as 3rd Cousin - 5th Cousin or closer (with a min cM of 15 cM) and is listed underneath.

██████████@gmail.com and shares a family tree), a predicted 1st Cousin - 3rd Cousin, is sharing in total: 363.5 cM (The largest segment is :43.2 cM.) In addition, she also shares a total of 3 segment(s) of X chromosomal DNA. This information can potentially reduce the search space of your common ancestor. Last, FTDNA has identified 61 other DNA matches in common.

**██████████** : A weekly FamilyTreeDNA search has been performed and identified a total of 1450 DNA matches. Of the 6 new DNA matches, 2 matches are DNA matches that are predicted as 3rd Cousin - 5th Cousin or closer (with a min cM of 15 cM) and are listed underneath.

██████████@gmail.com and shares a family tree), a predicted 2nd Cousin - 4th Cousin, is sharing in total: 105.8 cM (The largest segment is :21.3 cM.) Last, FTDNA has identified 37 other DNA matches in common.

██████████@gmail.com and shares a family tree), a predicted 3rd Cousin - 5th Cousin, is sharing in total: 57.3 cM (The largest segment is :12.6 cM.) Last, FTDNA has identified 20 other DNA matches in common.

**██████████** A weekly Ancestry search has been performed for ██████████. Of the 85 new DNA matches, 11 matches are DNA matches that are predicted as distant cousin or closer (with a min cM of 9 cM) and are listed underneath.

- ██████████ shares in total: 19.1 cM and is a predicted distant cousin with a confidence value of 60.3%.
- ██████████ shares in total: 15.1 cM and is a predicted distant cousin with a confidence value of 47.4%. Last, Ancestry has identified one other DNA match in common.
- ██████████ shares in total: 14.2 cM and is a predicted distant cousin with a confidence value of 44.6%.
- ██████████ shares in total: 13.6 cM and is a predicted distant cousin with a confidence value of 42.6%.
- ██████████ shares in total: 11.9 cM and is a predicted distant cousin with a confidence value of 37.0%.
- ██████████ shares in total: 10.8 cM and is a predicted distant cousin with a confidence value of 33.3%.
- ██████████ shares in total: 10.0 cM and is a predicted distant cousin with a confidence value of 30.9%.
- ██████████ shares in total: 9.8 cM and is a predicted distant cousin with a confidence value of 30.4%.
- ██████████ shares in total: 9.3 cM and is a predicted distant cousin with a confidence value of 28.9%.
- ██████████ shares in total: 9.2 cM and is a predicted distant cousin with a confidence value of 28.6%.
- ██████████ shares in total: 9.0 cM and is a predicted distant cousin with a confidence value of 28.2%.

**██████████** : A weekly Ancestry search has been performed. Of the 82 new DNA matches, no matches are found that are predicted as first cousin or closer (with a min cM of 9 cM).

**██████████** : A daily Ancestry search has been performed for ██████████. No new DNA matches were discovered.

The estimated cost for the current analysis is 42 credits. You currently have 806 credits remaining in your account.

 Reply  Forward

Figure 12. Example of an e-mail that contains regular updates

## AutoCluster analysis

AutoCluster (option D in Figure 9) organizes your DNA matches for registered 23andme or FTDNA profiles into shared match clusters that likely represent branches of your family. In the visualization of this analysis, each of the colored cells represents an intersection between two of your matches, meaning, they both match you and each other (see Figure 13). These cells, in turn, are grouped together both physically and by color to create a powerful visual chart of your shared matches clusters

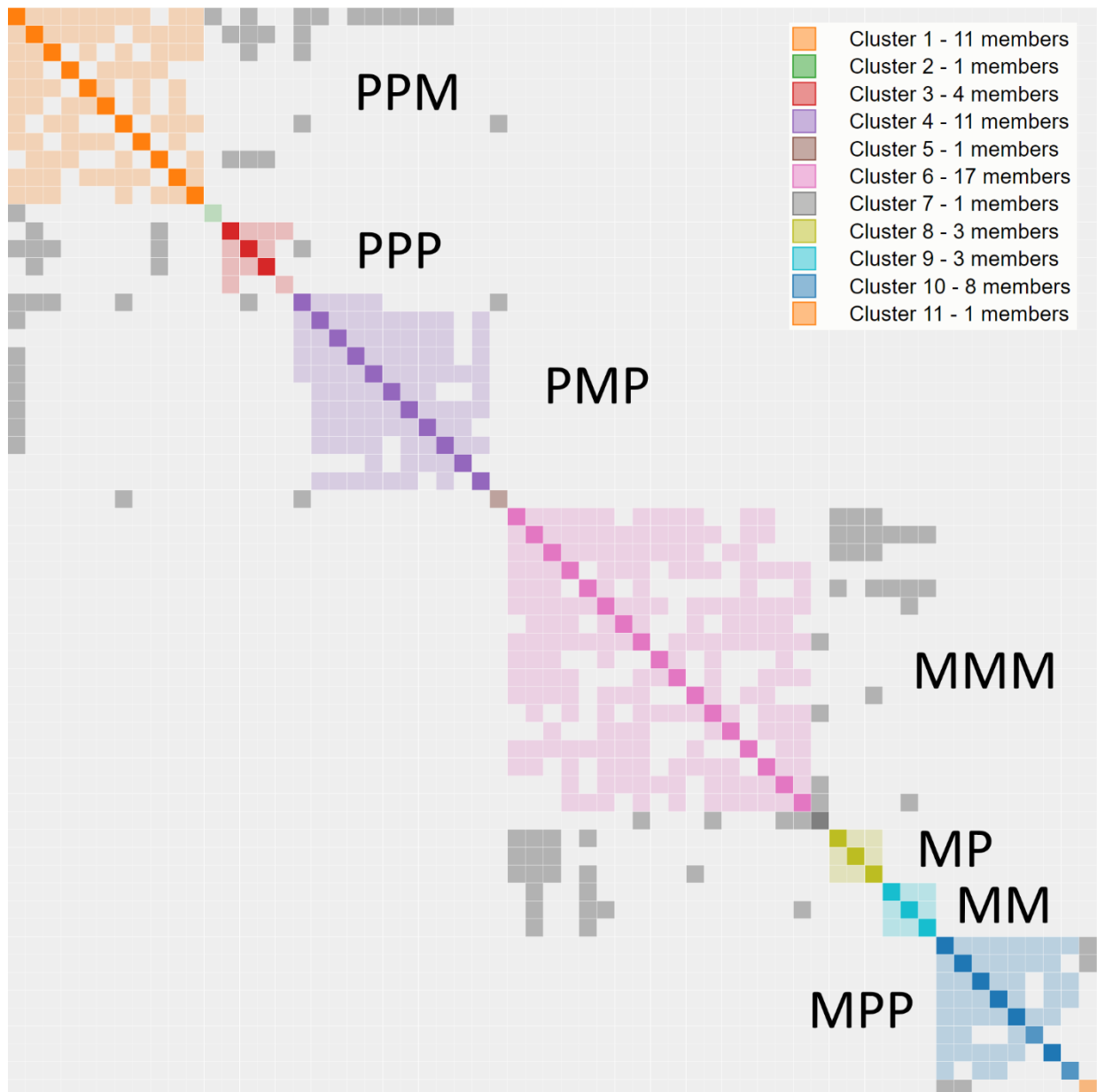


Figure 13. Clustering analysis of shared matches analysis using AutoCluster using a cM range of 600 cM – 50 cM. Various clusters have been identified which have been annotated using genealogical data (M = Maternal, P = Paternal).

Each color represents one shared match cluster. Members of a cluster match you and most or all of the other cluster members. Everyone in a cluster will likely be on the same ancestral line, although the MRCA between any of the matches and between you and any match may vary. The generational level of the clusters may vary as well. One may be your paternal grandmother's branch; another maybe your paternal grandfather's father's branch (see some genealogical annotations in the AutoCluster example in Figure 13).

You may see several gray cells that do not belong to any color-grouped cluster. They usually represent a shared match where one of the two cousins is too closely related to you to belong to just one cluster. Each of these cousins belongs to a color-grouped cluster, the gray cell indicates that one of them belongs in both clusters. In addition, the clusters are sorted based on the gray cells, these sorted clusters sometimes fall into larger cluster structures that are easy to identify in the visualization (see Figure 14).

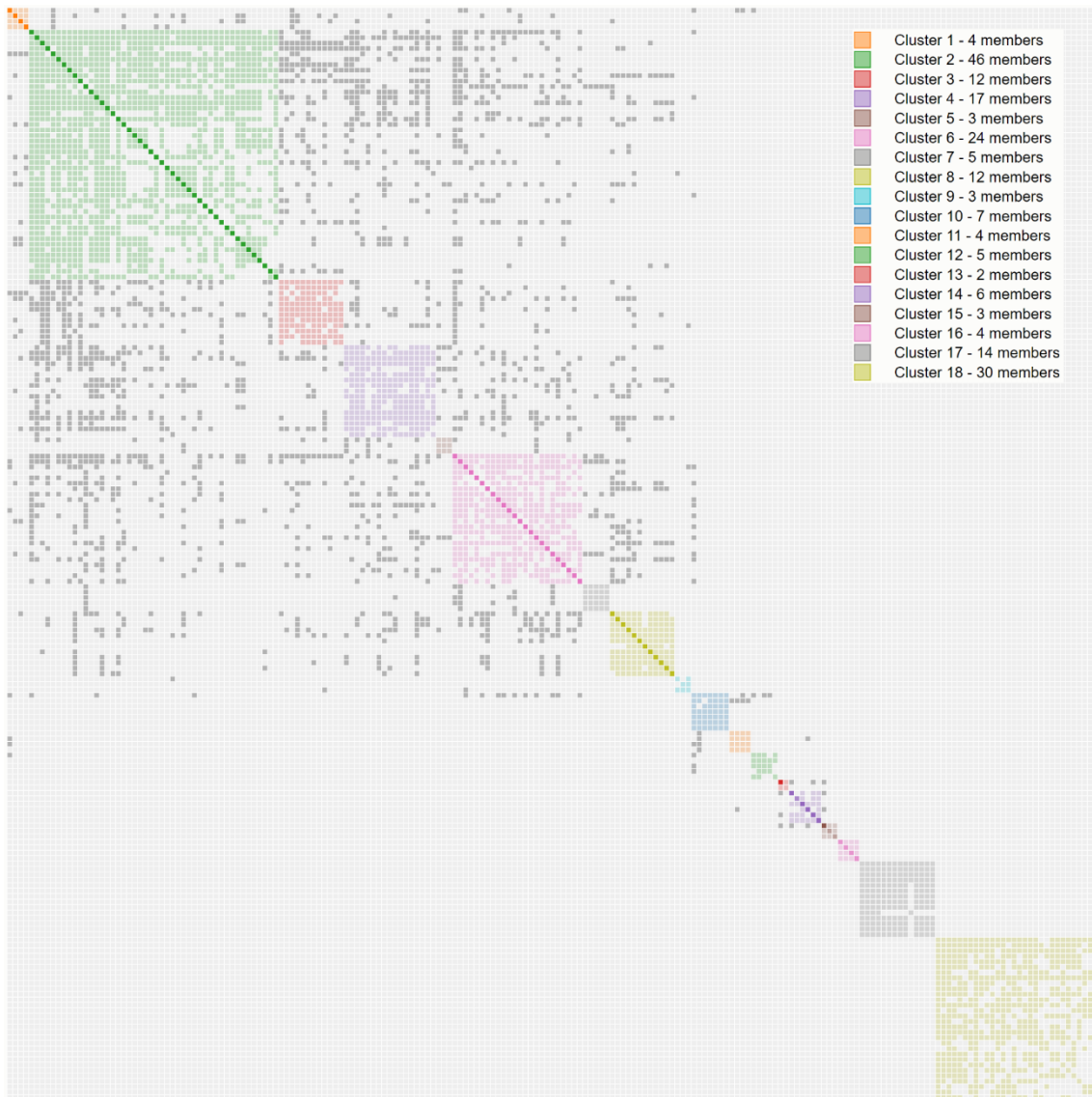


Figure 14. Example of the sorting of clusters based on the grey cells between clusters (picture provided by Robert Randolph).

Underneath the graphical representation of the clusters, some information concerning the AutoCluster is placed with respect to the employed settings as well as a searchable and sortable table (see Figure 15).

**AutoCluster match information**

Underneath, the match data are presented in a searchable, sortable table format that includes match profile and tree links.

A	B	C	D	E	F	G
Name	cM	#...	Cluster	Tree	Predicted rel...	Notes preview
<input type="text" value="Search"/>	M	Ma	<input type="text" value="Search fr"/>		<input type="text" value="Search"/>	<input type="text" value="Search"/>
<b>H</b> +	122.5	4	1		third cousin	
+	114.8	4	1		third cousin	
▼ Cluster 4 (4 items)						
+	183.3	4	4	7	third cousin	
+	165.4	5	4		third cousin	
+	94.8	4	4		third cousin	
+	92	5	4		third cousin	
▼ Cluster 3 (4 items)						
+	217.8	3	3	1003	second cousin	
+	215.4	3	3		second cousin	
+	114.1	3	3		third cousin	
+	113.6	3	3		third cousin	
▼ Cluster 2 (7 items)						
+	264.6	5	2	3	second cousin	
+	221.2	6	2		second cousin	

Figure 15 - Table view of the matches for each cluster based on an AutoCluster analysis for a profile.

- A. Name field, clickable for matches
- B. Shared cM
- C. Number of shared matches
- D. Cluster member
- E. Tree link (the number of people in the tree)
- F. Predicted relationship
- G. Preview of the notes
- H. Click this button to display the complete note

The results of the AutoCluster analysis are compressed and attached as a ZIP file. A unzipped AutoCluster report will contain:

- HTML file containing a visual representation of the AutoCluster analysis
- Excel file containing the chart visualization in a spreadsheet format, especially useful when there are a large number of matches in the HTML chart. In addition, the Excel file contains all downloaded matches (all matches for FTDNA and 23andme) and the matches per cluster.

## Start an AutoCluster analysis

From the profiles view page, click on the AutoCluster button which will display a page that allows for an AutoCluster analysis . For all three DNA testing companies, a maximum and minimum cM threshold can be selected. You can specify a range using these two cM thresholds to define criteria which DNA matches should be examined. In addition, the minimum cluster can be specified. For analyses with a low value for the minimum cM threshold a more powerful server is employed. For FTDNA analyses, this server is the default option.

Some additional options have been implemented that are only implemented for a specific DNA testing company. We will therefore discuss the AutoCluster analysis for each of the different companies.



## AutoFastCluster analysis

To generate an AutoCluster for sites that are not supported by Genetic Affairs (for instance, Living DNA) it is now possible to generate an AutoCluster analysis using user defined matches. There are two methods available. The first method allows users to run the analysis using locally generated CSV files. This feature can be reached using this link: <https://members.geneticaffairs.com/autocluster>. Patsy Coleman has blogged about this feature using LivingDNA data, the [blog post](#) contains more detail concerning this method. The second method allows users to enter matches and shared manually on the Genetic Affairs site using this link: <https://members.geneticaffairs.com/spreadsheet> (see Figure 16 for the interface). Entered matches and shared matches are saved on your local computer in the temporary storage of your browser.

The screenshot shows the AutoFastCluster interface with several panels:

- Panel A (Left):** A table for entering match data with columns: DNA Match name, cM, and Notes.
- Panel B (Right):** A table for defining shared matches with columns: DNA Match name and Shared match.
- Panel C:** Settings for Max (400 cM), Min (30 cM), and Cluster size (2).
- Panel D:** A text input field for the Name of AutoCluster analysis (John Smit).
- Panel E:** A green button labeled "PERFORM AUTOCLUSTER ANALYSIS".
- Panel F:** Buttons for exporting matches/shared matches: CSV, STATS, and EXCEL.
- Panel G:** Buttons for clearing matches/all/shared matches: MATCHES, ALL, and SHARED.
- Panel H:** Buttons for saving/clearing/loading matches/shared matches locally: SAVE, ADD ROWS, and LOAD.

Figure 16. AutoFastCluster interface

In the left panel (Figure 16A) the match data can be entered. In the right panel (Figure 16B) the shared match data can be defined (see Figure 17 for the format of this data).

The screenshot shows the AutoFastCluster interface with example data:

- Panel A (Left):** Match data table:
 

DNA Match name	cM	Notes
Jane Doe	20	from my Doe line
John Smit	600	NPE??
J.F.S	300	
- Panel B (Right):** Shared match data table:
 

DNA Match name	Shared match
Jane Doe	John Smit
Jane Doe	John Smit
John Smit	Jane Doe

Figure 17. Example match and shared match data.

The following AutoFastCluster settings can be defined, max and min cM threshold, cluster size and the name of the AutoFastCluster analysis (Figure 16C and D). The analysis is started by pressing the button “Perform AutoCluster analysis” (Figure 16E). The matches/shared matches can be exported to CSV and Excel (Figure 16F). To clear the matches, use the buttons under Figure 16G. Loading and saving the data can be performed using the buttons under Figure 16H (see Figure 18 for the message that appears after loading data).

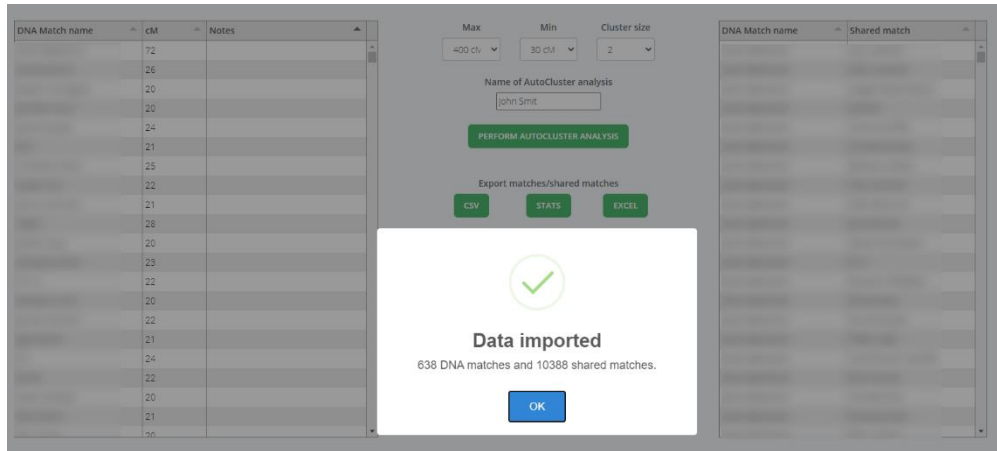


Figure 18. Message appearing after loading of previously entered data.

If an AutoFastCluster analysis is started and not enough matches are available, a popup will appear to notify the user that more matches are required (see Figure 19A). In that case, lower the min threshold and/or add more matches that are within the desired cM range. If there enough matches available, a message will appear that will inform you that the analysis is being performed (see Figure 19B).

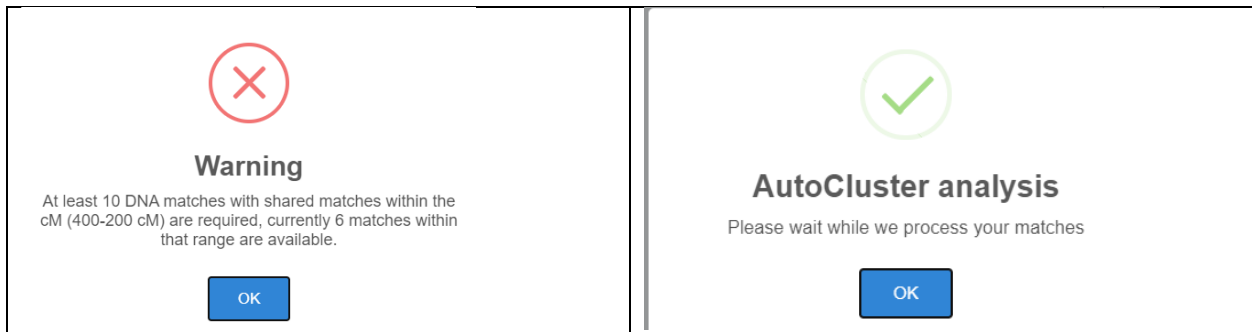


Figure 19. Message indicating more matches are required by the AutoFastCluster analysis and the message that will appear when the clustering is performed by our server.

After a couple of seconds, the AutoCluster page with the animating chart will appear (see Figure 20). The save button underneath the chart will save the analysis and allow for offline analysis.



Figure 20. Results of the AutoFastCluster analysis. In the lower pane of the chart a save button is available.

In addition to manually entering match and shared match data, it is possible to paste data from existing spreadsheets. To perform this action, first select the data (for instance, from Excel, see first figure in Figure 21) . Copy this data to the clipboard (on Windows, Ctrl-C). Next, go to the AutoFastCluster interface and select a cell. Then click on the border of that cell, a double arrow will appear. If it appears, paste the data from the clipboard.

Figure 21 illustrates the three steps to paste data from a spreadsheet into the AutoFastCluster interface. The figure is divided into three panels:

- Panel 1 (Left):** A spreadsheet with columns A, B, and C. Rows 1 through 26 contain match data. Row 1: MatchA, 25, this is a note. Row 2: MatchB, 30. Row 3: MatchC, 35. Row 4: MatchD, 40. Row 5: MatchE, 45. Row 6: MatchF, 50, this is a note. Row 7: MatchG, 55. Row 8: MatchH, 60. Row 9: MatchI, 65, note. Row 10: MatchJ, 70. Row 11: MatchK, 75. Row 12: MatchL, 80. Row 13: MatchM, 85. Row 14: MatchN, 90. Row 15: MatchO, 95. Row 16: MatchP, 100. Row 17: MatchQ, 105. Row 18: MatchR, 110. Row 19: MatchS, 115. Row 20: MatchT, 120. Row 21: MatchU, 125. Row 22: MatchV, 130. Row 23: MatchW, 135. Row 24: MatchX, 140. Row 25: MatchY, 145. Row 26: MatchZ, 150, and the last note. A dashed green border highlights the data from row 1 to row 26.
- Panel 2 (Middle):** The AutoFastCluster interface showing a table with columns: DNA Match name, cM, and Notes. A red arrow points to the border of the first cell in the cM column, and another red arrow points to the border of the first cell in the Notes column. The letters 'B' and 'A' are written in red below the arrows.
- Panel 3 (Right):** The AutoFastCluster interface showing the same table as in Panel 2, but with the data from the spreadsheet pasted into the cells. The data is: MatchA (25, this is a note), MatchB (30), MatchC (35), MatchD (40), MatchE (45), MatchF (50, this is a note), MatchG (55), MatchH (60), MatchI (65, note), MatchJ (70), MatchK (75), MatchL (80), MatchM (85), MatchN (90), MatchO (95), MatchP (100), MatchQ (105), MatchR (110), MatchS (115), MatchT (120), MatchU (125).

Below the panels are three instructions:

- Under Panel 1: Select data, copy to clipboard
- Under Panel 2: Select cell in spreadsheet, then select the border. A double arrow will appear
- Under Panel 3: Paste results.

Figure 21. Three steps to paste data.

## AutoCluster analysis using the extend cluster feature

Now one of the more complicated features, the “extend clusters” feature. The best way to illustrate this feature is by using the following example. Let’s imagine you have a high cM match, like a first or second cousin. You are interested in all shared matches with this match and see if there are clusters formed in these matches. Using the “extend cluster” feature combined with the FTDNA/23andme groups it is now possible to accomplish this. This is how it works. First, place the high cM match (or matches) in a specific group.

Next, find the FTDNA/23andme group-like identifiers. Next, enable the “extend cluster” feature. Now first the match (or matches) from the group are downloaded. Second, the shared matches are retrieved. Normally, the AutoCluster clustering would start but instead, we add the shared matches to our initial list and we download the shared matches for the shared matches. After this step, we start the clustering analysis. See Figure 22 for the result of such an analysis. The high cM match can be seen in the first

row/column, it matches all of the other matches in the chart. However, very clear clusters are identified using this approach. The “extend cluster” option can also be combined with the starred matches option.

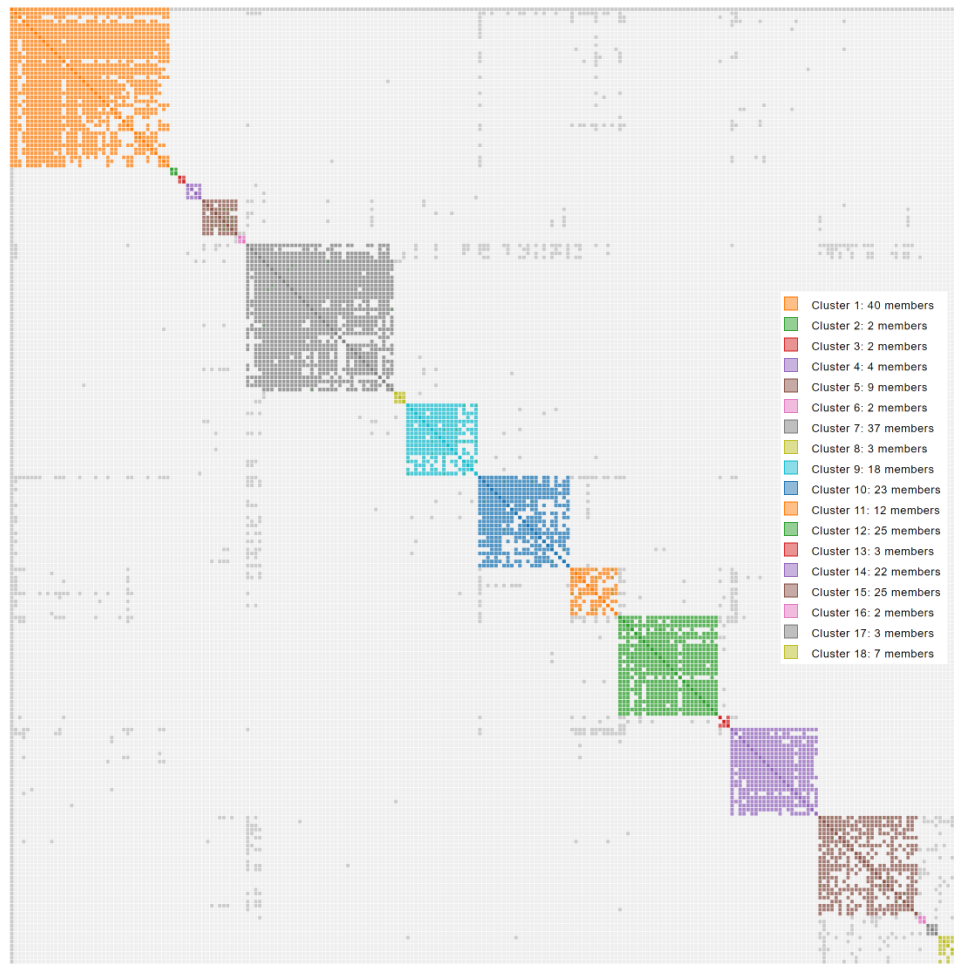


Figure 22. AutoCluster analysis performed using a **single** high cM match which was placed in a group. This group was used for the clustering analysis combined with the extend cluster feature.

However, what if you are interested in excluding this whole branch? Using the negative groups, only the high cM match would be excluded. It could be interesting to exclude the high cM match **and** its shared matches.

We could use the rule-based clustering and use a NOT rule but in some cases, you don't have access to the concerned profile. It is now possible to **discard** the high cM match as well as the shared matches you have with this match. Here is how it works. Instead of supplying the group name with a single ! you now use two ! exclamation marks. So if the group is 1001, you use !!1001

This is what happens. First, the matches of group 1001 are downloaded. Next, the shared matches of these matches are retrieved and combined with the first group. Last, each match from group 1001 and their shared matches are tagged with the negative group 1001, so getting !1001.

Now the regular AutoCluster analysis is performed, first the matches are downloaded and their shared matches. However, matches that have the !1001 group will be skipped.

## Start an FTDNA AutoCluster analysis

In addition to the minimum and maximum cM setting it is possible to specify the minimum size of the largest DNA segment (see Figure 23). In addition, the AutoTree functionality is also available for FTDNA analyses (see section below). Another new feature for FTDNA (and 23andme) profiles is the ability to integrate DNA segments into the AutoCluster analysis.

Start AutoCluster analysis with matches which share a max of	Stop AutoCluster analysis with matches which share less than	Minimum size of largest DNA segment shared with the match	AutoTree identify common ancestors from trees	Download segments for DNA Painter	Min cluster size	
250 cM	50 cM	10 cM	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	2	PERFORM ANALYSIS

Figure 23. The AutoCluster analysis page for an FTDNA profile.

## DNA segment browser for 23andme and FTDNA AutoCluster analyses

Using the segment data and a chromosome browser we can color the segments of matches from a cluster. This allows users to see how much DNA is in common with them (see Figure 24). Before we visualize the shared DNA segments we perform a clustering to group segments that are overlapping (min 5 cM). Next, these segment clusters are visualized using a certain color. In addition to the graphical representation a table is available that contains the detailed information for the segment clusters. Segments for the DNA matches for each AutoCluster cluster are available and can be accessed using the table underneath the chromosome browser. This table contains a link to the detailed chromosome browser, the number of multiple segment clusters, number of single segment clusters and number of clusters that are on the X chromosome.

It is now possible to generate a chromosome map from your DNA matches from FTDNA or 23andme clusters and import the segment data into DNA painter using the cluster auto painter tool (<https://dnapainter.com/tools/cap> also see the blog post: <https://dnapainter.com/blog/cluster-auto-painter-unravel-your-dna-test-results/>). Importing the chromosome map from your clusters of DNA matches into DNA painter allows you to:

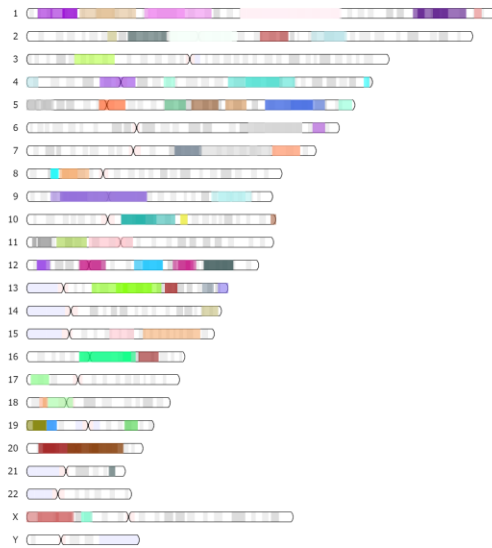
- Make notes and identify clusters as maternal or paternal
- Look at the segments behind the clusters and identify potential pile-up regions.

## Chromosome segments from DNA matches in clusters

A chromosome browser allows user to perform a graphical comparison between one or more matches to see how much DNA the user shares in common with them. Before we visualize the shared DNA segments we perform a clustering to group segments that are overlapping (min 5 cM). Next, these segment clusters are visualized using a certain color. In addition to the graphical representation a table is available that contains the detailed information for the segment clusters. Segments for the DNA matches for each AutoCluster cluster are available and can be accessed using the table underneath the chromosome browser. This table contains a link to the detailed chromosome browser, the number of multiple segment clusters, number of single segment clusters and number of clusters that are on the X chromosome.

In addition, it is now possible to generate a chromosome map from your clusters of DNA matches into [DNA painter](#) using the [cluster auto painter](#) tool. Importing the chromosome map from your clusters of DNA matches into DNA painter allows you to:

- Make notes and identify clusters as maternal or paternal
- Look at the segments behind the clusters and identify potential [pile-up](#) regions.



Chromosome browser using matches from profile Evert-Jan Blom

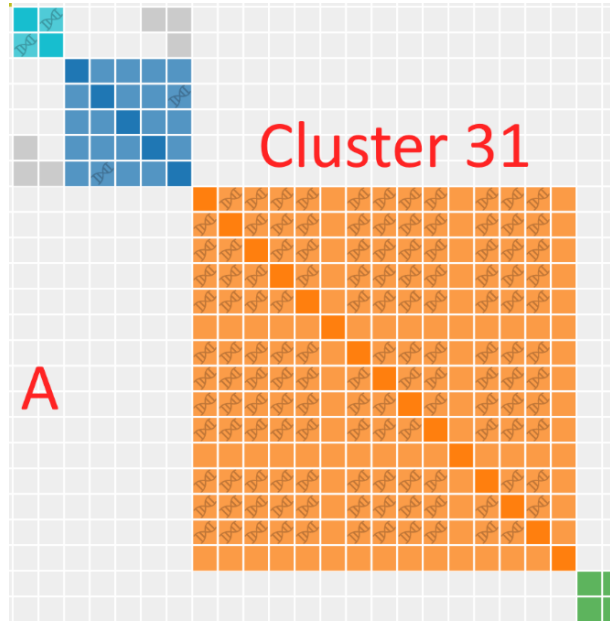


Example DNA painter overview with imported segments from AutoCluster

Figure 24. Chromosome segments from DNA matches in clusters.

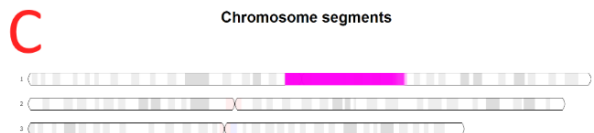
To supplement the segment analysis on DNA painter, a chromosome browser is available for the matches per cluster as well the combined matches overview (all segments from all matches from all clusters, see Figure 26).

An example for a single cluster can be seen in Figure 25. Cluster 31 from a 23andme clustering has a number of matches that probably have a triangulated segment (visualized using the helix symbol). Information concerning segment clusters, (multiple and single), x match clusters as well as paternal or maternal classification (as obtained from 23andme or FTDNA) is available in the table underneath (Figure 25B). Upon clicking on the segment link in the first column, a chromosome visualization is displayed (see Figure 25C). One segment is shared with the matches of this cluster which is also illustrated in the detailed (see Figure 25D) table that contains the start, stop positions as well as the total length of the segment.



**B** Chromosome segment statistics per AutoCluster cluster

	single_segments	multiple_segments	x_segments	Paternal	Maternal
	filter column...	filter column...	filter column...	filter column...	filter column...
▼ (23 items)					
Segments for cluster 1	2	1	0	6	0
Segments for cluster 2	1	0	0	2	0
Segments for cluster 3	2	2	0	9	0
Segments for cluster 4	2	0	0	2	0
Segments for cluster 5	2	2	2	5	0
Segments for cluster 31	0	1	0	12	0
Segments for cluster 32	1	0	0	1	0
Segments for cluster combined	34	32	6	75	0



**D** Segment Cluster Information

Cluster	C#	Start	Stop	SNP count	N...	cM	Total cM	Paternal	Maternal	AutoCluster
Segment title	Search 1	Search 2	Search 3	Search for d	Da	Misc d	Total cM	filter column	filter column	filter column...
▼ 31 (12 items)										
31	1	117759719	162351796	3244	ToL	27.1	25	F		31
31	1	117759799	161493580	2998	ML	25.9	25	F		31
31	1	117759799	161234823	2953	Ds	25.6	24	F		31
31	1	117627683	161868836	2689	act	25.9	24	F		31
31	1	117759799	160825313	2862	JE	25.3	24	F		31
31	1	117627683	160800361	2837	Eg	25.1	24	F		31
31	1	117627683	159599114	3491	Sa	25.4	21	F		31
31	1	117627683	159278251	2422	Bu	22.9	21	F		31
31	1	117627683	159278251	2422	Alk	22.9	21	F		31
31	1	117759799	158655080	2283	Nk	21.4	20	F		31
31	1	117627683	158725194	2295	Jan	21.3	20	F		31
31	1	118074219	158991822	2141	KCl	20.7	20	F		31

Figure 25. 23andme segment example. For cluster 31 quite a few triangulating segments seem to be present.



Figure 26. All segment clusters from all clusters visualized in the chromosome browser

## Start an 23andme AutoCluster analysis

Start AutoCluster analysis with matches which share a max of	Stop AutoCluster analysis with matches which share less than	Minimum shared cM between shared matches	Based on Shared matches	Based on Triangulated Groups	Surname enrichment	Download segments for DNA Painter	Min cluster size	PERFORM ANALYSIS
250 cM	50 cM	10 cM	<input type="radio"/>	<input type="radio"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	2	

For **23andme** profiles, a cM threshold between shared matches can be adjusted as well. Common matches that share a DNA segment (which typically will triangulate) with the tested person contain an additional helix symbol in the visualization (see Figure 27). It is also possible to perform the clustering analysis only for matches that share a segment (Triangulated Groups clustering). This usually results in very defined clusters and a low amount of grey cells. A [blog post of Louis Kessler](#) goes into more detail concerning this feature and discusses some of the results (and caveats) of this approach. As will be discussed on page 29, a surname enrichment analysis can be employed for the matches based on the surnames and locations of the ancestors of the matches from a specific cluster.

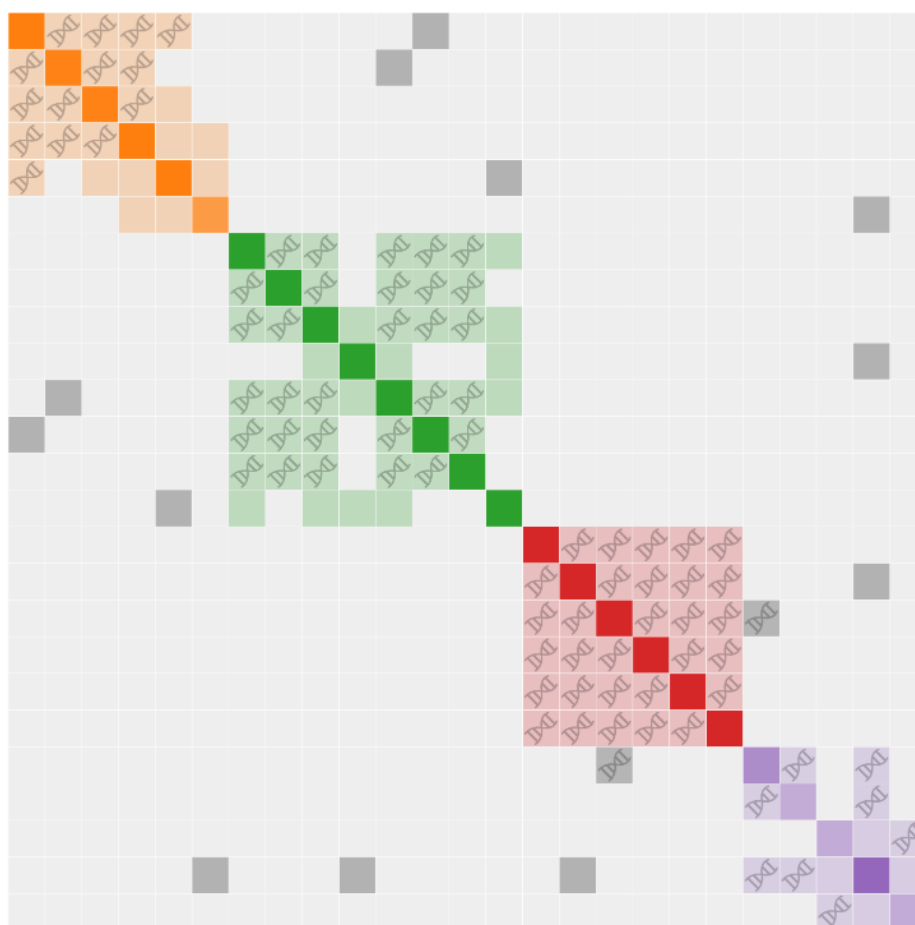


Figure 27. 23andme AutoCluster analysis, common matches that share a segment (which will most likely triangulate) with the tested person are indicated with the helix symbol.



After the AutoCluster analysis has finished, a mail will be sent to your e-mail address. If all goes well, this email will contain three zipped attachments (see the previous section for more information concerning the attachments). Note that the Excel file and HTML file will only be present if enough matches are present for the AutoCluster analysis. It is also possible that the HTML file is not present, in that case not enough matches were present for the analysis.

**Please save the attached zip files to your hard drive and unzip them. You will then be able to view the results. Opening the HTML file from the zip file will usually result in a nonfunctioning links (such as links referring to AutoTree output or chromosome browser results).**

Note that due to the computational design of our website only a certain amount of time is reserved for each AutoCluster analysis. If this time frame is exceeded, the downloaded matches and in common matches are collected after which an AutoCluster analysis is applied. If you notice that some matches are not present, restrict or modify the search settings (for instance more strict cM parameters) to obtain additional results. It is also possible to perform the AutoCluster search in the mornings or night when the servers of 23andme/FTDNA are more responsive.

***In some cases, the mail client renames the ZIP file to a file with a .dat extension. In that scenario, rename the .dat file to a file with a .zip extension and try to unzip the renamed file.***

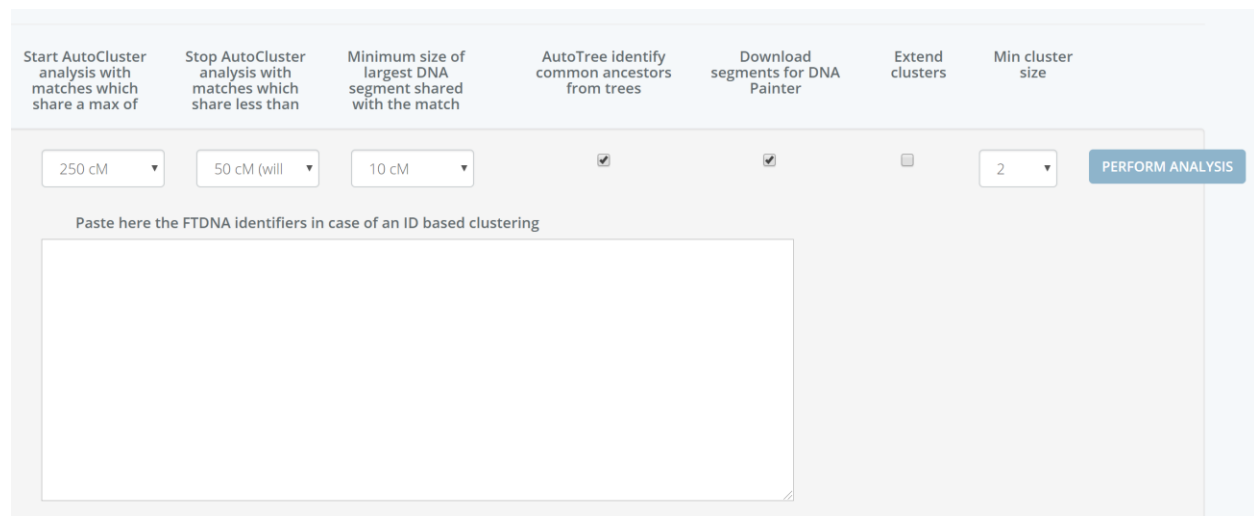
## FTDNA and 23andme group-like AutoCluster analysis

A popular feature available for profiles is the ability to use the groups. It allows you to research a particular group of matches. Since FamilyTreeDNA and 23andme are not providing a group feature, a similar system using the FTDNA and 23andme identifiers has been developed. The functionality of this feature is similar to the Ancestry version of groups.

Instead of using groups we now employ the identifiers in the Excel file that is available from each AutoCluster analysis. For FTDNA we use the "ResultID2" column, for 23andme the "ehid" column.

Why is this feature interesting for your research? Imagine the scenario where you are interested in the matches you share with a 600 cM match on FTDNA or 23andme. Perhaps some of these shared matches form clusters as well. To analyze this match (and its shared matches), find the identifier of this match in the Excel file and provide it to the AutoCluster panel (see Figure 28). Enable the "Extend cluster" feature and set the parameters. Now the shared matches of the 600 cM match are downloaded after which shared matches are downloaded for these matches (see Fig 2 for an example that employed a high cM match).

You can also remove matches by placing a single exclamation mark in front of id. And you guessed it, by using two exclamation marks, you can remove complete branches. Using the double !! feature, matches and their shared matches will be removed from the analysis.



The screenshot displays the FTDNA AutoCluster interface. At the top, there are seven settings: "Start AutoCluster analysis with matches which share a max of" (set to 250 cM), "Stop AutoCluster analysis with matches which share less than" (set to 50 cM (will)), "Minimum size of largest DNA segment shared with the match" (set to 10 cM), "AutoTree identify common ancestors from trees" (checked), "Download segments for DNA Painter" (checked), "Extend clusters" (unchecked), and "Min cluster size" (set to 2). A "PERFORM ANALYSIS" button is located to the right of these settings. Below the settings is a text box labeled "Paste here the FTDNA identifiers in case of an ID based clustering".

Figure 28. FTDNA AutoCluster interface with a text box that allows for the group-like clustering

## Enriched Surnames and Locations for 23andme analyses

For 23andme analyses, we collect the surnames and locations of recent ancestors that were entered by the different DNA matches. Surnames that are characteristic of the identified clusters can yield information concerning the historical or demographic significance of a cluster. To identify these surnames and locations, we calculate statistical evidence (i.e., p-value \*) that each surname or location is overrepresented in a given cluster compared to the background surname/location distribution over all retrieved DNA matches. Next, we rank the surnames according to the statistical evidence (i.e., smaller p-values), and select the most highly ranked surnames as the surnames that are associated to this cluster (see Figure 29).

**Enriched Surnames**

We collect the surnames and locations of recent ancestors that were entered by the different DNA matches. Surnames that are characteristic of the identified clusters can yield information concerning the historical or demographic significance of a cluster. To identify these surnames and locations, we calculate statistical evidence (i.e., p-value) that each surname or location is overrepresented in a given cluster compared to the background surname/location distribution over all retrieved DNA matches. Next, we rank the surnames according to the statistical evidence (i.e., smaller p-values), and select the most highly ranked surnames as the surnames that are associated to this cluster.

More information concerning the (calculation of the) enriched surnames is provided at the bottom of this page in the [detailed surname table](#).

Cluster	Enriched surnames
Cluster 1	Lefferts, (Hofman - Hoffman), (Corelisd - Cornelisd - Cornelis - Cornelis) (Hendricksen - Hendrickson - Hendrickse - Hendricks - Hendriks), (Teunis - Theunis) Gabes, (Huizinga - Huizenga), Sakes, (Buwens - Bouwens - Buwes), Eelses, (Heemstra - Hiemstra) (Edes - Edses), Christiaans, (Folkert - Folkerts - Folckerts), (Sibrens - Sybrens) (Tibbes - Tjibbes), (Auke - Auke), (Tjebbes - Tjibbes), (Tjebbes - Tjibbes) (Heeres - Heres - Heerkes), (Douwes, Willems, (Jounes - ) Of these 6 DNA matches, 2 have been identified in this cluster 1 (which holds 8 members of which 4 have surnames). We will now list the details for this surname: Folkert: 1 members of this cluster have this surname listed. A total of 1 DNA matches have this surname listed. Folkerts: 1 members of this cluster have this surname listed. A total of 4 DNA matches have this surname listed. Folckerts: 1 member of this cluster have this surname listed. A total of 1 DNA matches have this surname listed. Folkertsma: no members of this cluster have this surname listed. A total of 1 DNA matches have this surname listed.
Cluster 4	(Hovde - Hove), (Carlsen - Kai (Johansen - Johanssen - John (Tomasdotter - Thomasdatter - Tomsdatter), (Johannesson - Johannesen - Johansson - Johannessen - Johannisson) (Eriksdatter - Eriksdotter), (Hansdatter - Hansdotter), Svendsdotter Hansson, (Jonson - Jonsson - Jonasson), (Jonsdotter - Jansdotter - Johansdotter)
Cluster 6	Doble, Govier, Gahagan, Kulken, Lumby, Meggott, (Newbett - Newbit - Newbitt) Skins, Boomsma, Bowering, Brammer, Lantinga, Spiller, Trask, Blick Heeringa, Purscy, Marks, Terpstra, Hannan, Dowland, Lane, Dixon, Ford Miles, Martin, Wilson, Johnson, Vine, Bradley, Gainsboro, Lincs, Devonsomerset (Tjummärum - Tzummarum), Ydema, Coats, Lunn, Summerhayes, Parsons, Minnertsga Brewer. (Netherlands - NetherlandIs). Clark. Heath
Cluster 15	Hittscher, Schmidt

Figure 29. Enriched surnames table with the most overrepresented surname clusters for each cluster

Note that we first perform clustering of surnames to combine surnames that have similar spelling. Members of these surname clusters that are found to be enriched are placed between parenthesis. For example in cluster 1 in Figure 29 there is a surname cluster highlighted that contains 4 members. Three of these surnames (Folkert - Folkerts - Folckerts) are linked to members of cluster 1. If you hover your mouse over the surnames, you get this additional information in the mouse tooltip. In addition, the surname cluster also contains the surname Folkertsma which is linked to 1 DNA match which is not a member of Cluster 1.

Next, the surname table can also hold overrepresented locations that are provided by the users. This can be either a part of the surname (for instance for FTDNA users) or specifically supplied by the users in the case of 23andme matches.

\* Note that we don't employ strict [multiple testing corrections](#) which is important when testing many hypotheses. Please use these p-values to rank the results and solely as guidance for future research.

## Detailed information concerning Enriched Surnames and Locations

In some cases, certain sets of enriched surnames or locations are linked to a specific combination of matches. This more detailed information is provided in the table located at the bottom of the page (see Figure 30). Moreover, additional information is provided concerning the DNA matches with surnames and the resulting surname clusters.

### Detailed enriched surname table

As mentioned in the section above, we collect the surnames and locations of recent ancestors that were entered by the different DNA matches. Surnames that are characteristic of the identified clusters can yield information concerning the historical or demographic significance of a cluster. To identify these surnames and locations, we calculate statistical evidence (i.e., **p-value**) that each surname or location is overrepresented in a given cluster compared to the background surname/location distribution over all retrieved DNA matches. However, since the writing of some surnames can change over time, we first perform a clustering of similar surnames. This ensures that we also find enriched surnames that are highly similar, which might be missed if only unique surnames are taken into account.

Here is some information concerning the identified surname clusters: Of the initial 1347 DNA matches, 563 have surnames which were used for the surname clustering. A total of 9718 surname clusters were created, which were condensed into 3330 clusters when 6388 clusters were discarded that only contained one member.

In some cases certain enriched surnames in a cluster are often linked to the same set of DNA matches. We therefore combine these occurrences as well before placing them in the table. This means that the table is divided per cluster and the underlying clusters are divided by members. Next, we rank the clusters with enriched surnames according to the statistical evidence (i.e., smaller p-values), and sort the clusters to first show the clusters with the most enriched surnames.

Cluster	Members	Best p-value																								
<input type="text" value="Search fr"/>	<input type="text" value="Search"/>	<input type="text" value="pvalue"/>																								
▼ Cluster 6 with 5 members of which 4 have surnames (4 items)																										
6	<a href="#">View all Members</a> , <a href="#">View all Locations</a> , <a href="#">View all Surnames</a>	3.666E-7																								
<table border="1"> <thead> <tr> <th>Overrepresented surnames</th> <th>total # members wit...</th> <th># members with sur...</th> <th>p-value</th> </tr> </thead> <tbody> <tr> <td><input type="text" value="Search"/></td> <td><input type="text" value="in all clusters"/></td> <td><input type="text" value="in this cluster"/></td> <td></td> </tr> <tr> <td><b>Doble</b></td> <td>3</td> <td>3</td> <td>3.666E-7</td> </tr> <tr> <td colspan="4">                     This surname cluster contains of a single surname which is listed in the names of the ancestors of 3 DNA matches. Of these 3 DNA matches, 3 have been identified in this cluster 6 (which holds 5 members of which 4 have surnames). We will now list the details for this surname:                      Doble: 3 members ( ) of this cluster have this surname listed. A total of 3 DNA matches have this surname listed.                 </td> </tr> <tr> <td><b>Kuiken</b></td> <td>3</td> <td>3</td> <td>3.666E-7</td> </tr> <tr> <td><b>Lumby</b></td> <td>3</td> <td>3</td> <td>3.666E-7</td> </tr> </tbody> </table>			Overrepresented surnames	total # members wit...	# members with sur...	p-value	<input type="text" value="Search"/>	<input type="text" value="in all clusters"/>	<input type="text" value="in this cluster"/>		<b>Doble</b>	3	3	3.666E-7	This surname cluster contains of a single surname which is listed in the names of the ancestors of 3 DNA matches. Of these 3 DNA matches, 3 have been identified in this cluster 6 (which holds 5 members of which 4 have surnames). We will now list the details for this surname: Doble: 3 members ( ) of this cluster have this surname listed. A total of 3 DNA matches have this surname listed.				<b>Kuiken</b>	3	3	3.666E-7	<b>Lumby</b>	3	3	3.666E-7
Overrepresented surnames	total # members wit...	# members with sur...	p-value																							
<input type="text" value="Search"/>	<input type="text" value="in all clusters"/>	<input type="text" value="in this cluster"/>																								
<b>Doble</b>	3	3	3.666E-7																							
This surname cluster contains of a single surname which is listed in the names of the ancestors of 3 DNA matches. Of these 3 DNA matches, 3 have been identified in this cluster 6 (which holds 5 members of which 4 have surnames). We will now list the details for this surname: Doble: 3 members ( ) of this cluster have this surname listed. A total of 3 DNA matches have this surname listed.																										
<b>Kuiken</b>	3	3	3.666E-7																							
<b>Lumby</b>	3	3	3.666E-7																							

Figure 30. Detailed enriched surnames table

The most notable difference between the more condensed enriched surname table (Figure 29) and this table is the sorting of clusters and the surnames within clusters. In the detailed enrichment table, the clusters are first sorted based on the lowest p-value (i.e., most significant surname enrichment). Next, within each cluster, enriched surnames are combined based on the DNA matches from the cluster that one of the surnames listed. This ensures that we see all enriched surnames for a single set of matches combined in one view. Note that there are more surnames listed in this table as compared to the condensed table (Figure 29). In that table, only surnames are listed with a p-value lower or equal to 0.05 while in the detailed table (Figure 30) all surnames are listed that occur two times or more in a cluster.

## Rule-based AutoCluster

The rule-based AutoCluster allows users to employ rules to filter and/or merge their matches using matches from other profiles. Three different rules allow for the exclusion (NOT rule), inclusion (AND rule) or combination (OR rule) of matches. The resulting matches are used for an AutoCluster analysis. The usage of these rules allows for a focus on matches from a particular branch of the family, for instance, paternal or maternal matches. The usage of these rules is illustrated using the following family tree (see Figure 31) and an explanation of each of the three rules (see Figure 32). Note that it is possible to combine different rules. First, the AND and NOT rules are processed and used to filter the primary profiles and matches from OR rules.

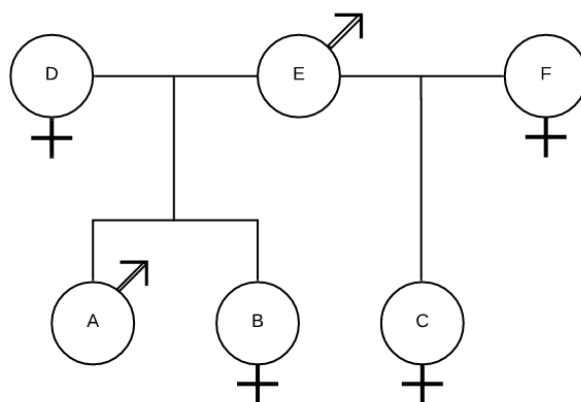


Figure 31. Example family tree to illustrate the usage of different adoptee rules

	<p><b>NOT</b> (or except) rule. In this scenario, <b>C</b> is a person with unknown parentage to her birth family (for instance adoptees or donor-conceived persons) that has matches of her biological mother <b>F</b>. By applying the <b>NOT</b> rule on the matches of her biological mother <b>F</b>, the AutoCluster analysis is performed without her maternal matches. By using this strategy she can focus on her paternal matches to identify her biological father <b>E</b>.</p>
	<p><b>AND</b> (or intersection) rule. In this scenario, <b>B</b> is a person with unknown parentage to her birth family (for instance adoptees or donor-conceived persons) that has identified a half-sister (<b>C</b>). By applying the <b>AND</b> rule, only paternal matches that are in common with her half-sister are used for the AutoCluster analysis. This allows them to focus on the identification of their shared biological father <b>E</b>.</p>
	<p><b>OR</b> (or union) rule. In this scenario, two persons (<b>A</b> and <b>B</b>) with unknown parentage to their birth families (for instance adoptees or donor-conceived persons) would like to combine their matches (and shared matches) to identify one or both biological parents. If the biological mother already is known and tested, we can add another NOT rule which will exclude the matches of the biological mother <b>D</b>. Or if another half-sister <b>C</b> and her biological mother <b>F</b> are known, we could add an OR rule to include the matches of <b>C</b> and a NOT rule to exclude the matches of <b>F</b>.</p>

Figure 32. Explanations of three different adoptee rules

The last approach using the OR rules enables donor-conceived persons that have identified a large number of half-brothers/sisters to combine all of their matches (using the OR rules) and remove the matches of their biological mothers (using the NOT rule). By employing this strategy, the majority of the paternal matches will be retrieved and used for clustering.

To select the rule-based clustering, go to the website's view (see Figure 8) and click on the profiles link. From there, select the rule-based AutoCluster icon for the profile that will be the primary profile. Upon clicking on the rule-based AutoCluster link, a new selection view is displayed (see Figure 33). The profile for which the rule-based AutoCluster icon was selected is placed on the top and will be used as the primary profile. Next, every available registered profile from the same DNA testing company as the primary profile is listed underneath. For instance, if an FTDNA profile is selected, every other FTDNA profile is shown in this view.

The screenshot shows the 'rule-based AutoCluster' interface. At the top, the primary profile is 'EJ Blom'. To its right are dropdown menus for 'Maximal cM' (set to 250 cM) and 'Minimal cM' (set to 50 cM). Below this is a text box explaining the filtering logic: 'The primary profile EJ Blom is used as the basis of this analysis. As mentioned above, the matches from this primary profile can be filtered using the NOT and AND rules. Note that a match does not get removed because of an NOT rule if shared more than 1.5 times the cM as the match from the NOT profile. For instance, if the match shares 80 cM with the primary profile and 50 cM with NOT profile, it will not be removed. If the primary profile would share 60 cM, it would get removed. This filtering has been added to avoid removing matches for which the primary or OR profile only shares a fraction of the cMs.' Below this is another text box: 'Matches can be combined with matches of the primary profiles using the OR rules. In case of overlapping matches, we will check which match shares the most cM. If the match from the OR profile shares more cM as compared to the primary profile, we will then keep the shared cM from the OR profile.'

Profile:	Do not use	OR	AND	NOT	Maximal cM	Minimal cM
EJ Blom	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	250 cM	50 cM
Frédéric Blom de Vries	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	250 cM	50 cM
Thea Orger Blom	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	250 cM	50 cM
Marcel Blom Godefrid	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	250 cM	50 cM
Wim Blom	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	250 cM	50 cM

**Pricing**

The price of the analysis is the default price of an AutoCluster analysis (**25 credits**) times the number of analyses performed. So an analysis using the primary profile, one OR rule and one NOT rule will cost 3 times 25 credits which makes **75 credits**.

**PERFORM ANALYSIS**

Figure 33. Rule-based AutoCluster page with the primary selection of different rules.

By default, each of the rules associated with the listed profiles is not used. To add a rule based on the matches of these profiles, find the concerned profile and set the desired rule. Next, adjust the cM settings for this rule. The cM settings are used to filter the matches from that rule. For instance, a NOT rule with cM settings 900 cM – 50 cM will only remove matches from the primary profile (and if available, matches from OR rules) for matches that fall between 900 and 50 cM. So select a broad cM

range if you want to use all possible matches from that profile to be removed from the primary and OR profiles. In addition, these cM limits are also applied for the download process of shared matches from profiles based on an OR rule.

The process of downloading matches employs the following logic. First, the matches from NOT and AND profiles are obtained. Next, the primary profile and available profiles based on OR rules are obtained and filtered using the matches from the previous NOT and AND rules. No filtering will occur if no NOT and AND rules are defined. The result of these filtering steps is mentioned in the HTML report, for instance:

*“A rule-based AutoClustering analysis was started with the primary profile EJ Blom (using cM settings 600cM - 20 cM).*

*The 1103 primary matches from profile EJ Blom were filtered using the provided rules. After applying the AND rule(s), the 1103 matches were condensed to 237 matches. We downloaded shared matches for 47 DNA matches. The following rule was applied:*

*1). AND rule, only using matches that overlap based on matches from profile Mommie (using cM settings 900cM - 9 cM). For this AND rule, we downloaded 1092 23andme matches.*

*After applying one rule, a total of 47 matches are used from the primary profile to perform an AutoCluster analysis.”*

In case of analysis with OR rules, it is interesting to know which (shared) matches are added to the primary profile. In the following example (see Figure 34), the matches of two unrelated persons were combined. The newly added matches are visualized using an additional plus symbol in the visualization. Note that complete new matches will get this symbol in the diagonal of the cluster. This allows users to distinguish between new shared matches for existing matches and completely new matches.

In addition to the HTML chart, an Excel formatted file is available that holds the identified clusters as well as the DNA matches (see Figure 35) and their annotations (in a separate worksheet named AllDnaMatches). New matches in the AllDnaMatches worksheet are colored grey.

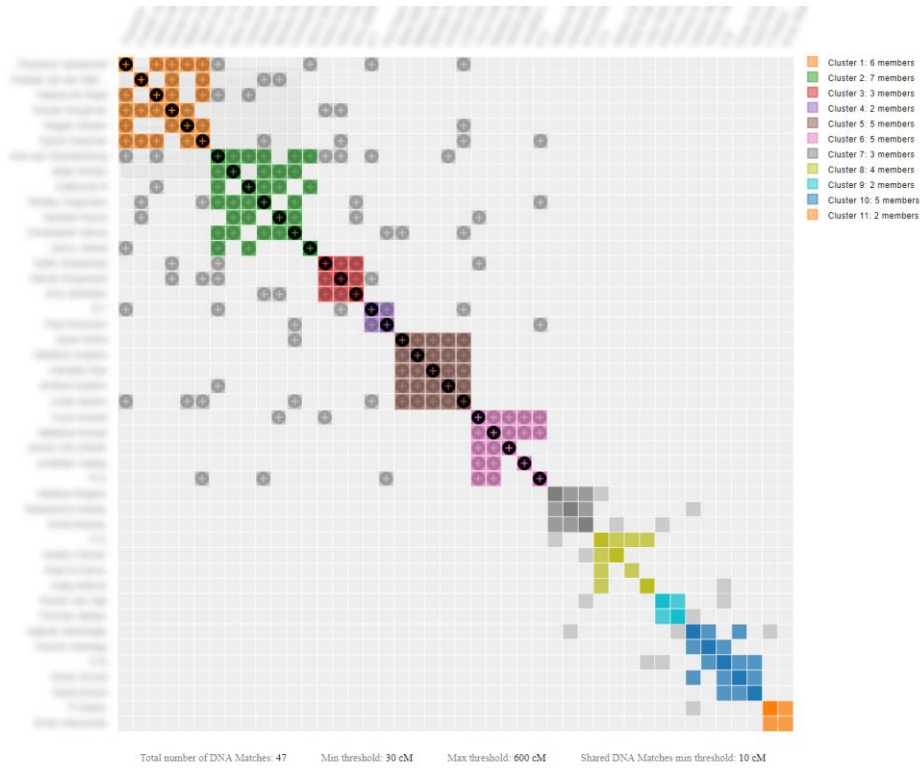


Figure 34. Rule-based AutoCluster analysis using single OR rule for two unrelated persons. Note that the (shared) matches from the first six clusters are visualized using an additional plus symbol which indicates that the (shared) match is new.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Identifi	Name	total sh	matche	cluster	notes	tree												
2			474	none	1		tree												
3			69	none	1		tree												
4			50	none	1		tree												
5			59	none	2														
6			51	none	2		tree												
7			66	none	3														
8			43	none	3														
9			58	none	4														
10			48	none	4														
11			60	none	5														
12			54	none	5		tree												
13			52	none	5														
14			212	none	6														
15			70	none	6														
16			70	none	6		tree												
17			69	none	6														
18			48	none	6														
19			47	none	6		tree												
20			56	none	7		tree												
21			48	none	7		tree												
22			57	none	8		tree												
23			57	none	8														
24			54	none	8		tree												
25			51	none	8		tree												
26			47	none	8		tree												
27			45	none	8														

Figure 35. Excel representation of rule-based AutoCluster. In this picture, the vertically dashed cell A represents a new shared match, the match itself was already available. B represents a newly added match since the diagonal is vertically dashed as well.



## AutoTree

AutoCluster first organizes your DNA matches into shared match clusters that likely represent branches of your family. Everyone in a cluster will likely be on the same ancestral line, although the MRCA between any of the matches and between you and any match may vary. The generational level of the clusters may vary as well. One may be your paternal grandmother's branch, another may be your paternal grandfather's father's branch.

By comparing the linked and unlinked trees from the members of a certain cluster, we can identify ancestors that are common amongst those trees. First, we collect the surnames that are present in the trees and create a network using the similarity between surnames. Next, we perform clustering on this network to identify clusters of similar surnames. A similar clustering is performed based on a network using the first names of members of each surname cluster. Our last clustering uses the birth and death years of members of a cluster to find similar persons. As a consequence, initially, large clusters (based on the surnames) are divided up into smaller clusters using the first name and birth/death year clustering. See also the blog post of Roberta Estes from DNAExplained that covers the AutoTree feature: <https://dna-explained.com/2019/12/02/genetic-affairs-reconstructs-trees-from-genetic-clusters-even-without-your-tree-or-common-ancestors/>

The common ancestor and location analysis is calculated using members of AutoCluster clusters as well as all using all matches from these clusters. This last step ensures that common ancestors that are present in different clusters (for instance clusters that are part of a supercluster, for instance, the first clusters from the chart of Figure 14) are identified as well.

The overview of these analyses is displayed in the main AutoCluster HTML file in a table (see Figure 36). For each AutoCluster cluster the number of common ancestors, common locations (for two distances) and common surnames are shown. The fields of the tree, common ancestor and common location are clickable and will show more detailed information.

Tree	# ancestors	# Radius 100m	# Radius 5000m	surnames
Search for surnames				
▼ (7 items)				
Tree(s) of cluster 1	334 common ancestors	96 common locations	93 common locations	Balentine (37) Sport (18) Baxter (9) Branscum (23) Crisel (11) Bumgarner (9) Harrison (4) Owings (7) Proctor (6)

Figure 36. AutoTree overview which shows the number of common ancestors, locations and surnames for each of the AutoCluster clusters as well as the combined analysis.

If there are not too many matches, an AutoTree analysis is performed on all matches. This particular analysis combines all matches from the clustering in a single cluster and perform the AutoTree. It sometimes finds common ancestors between members of separate clusters (e.g., in a supercluster).

Next, in addition to identifying the common ancestors, we combine the common ancestors and try to reconstruct the underlying genealogical tree. In most cases, only parts of the trees can be reconstructed. But, with some manual efforts, most automatically generated trees can be combined into one or several larger trees. To improve the analysis of the trees, we use a color gradient to differentiate between different DNA matches. In addition, persons in the tree are highlighted when you hover over the edges if they appear in different trees (see Figure 37).

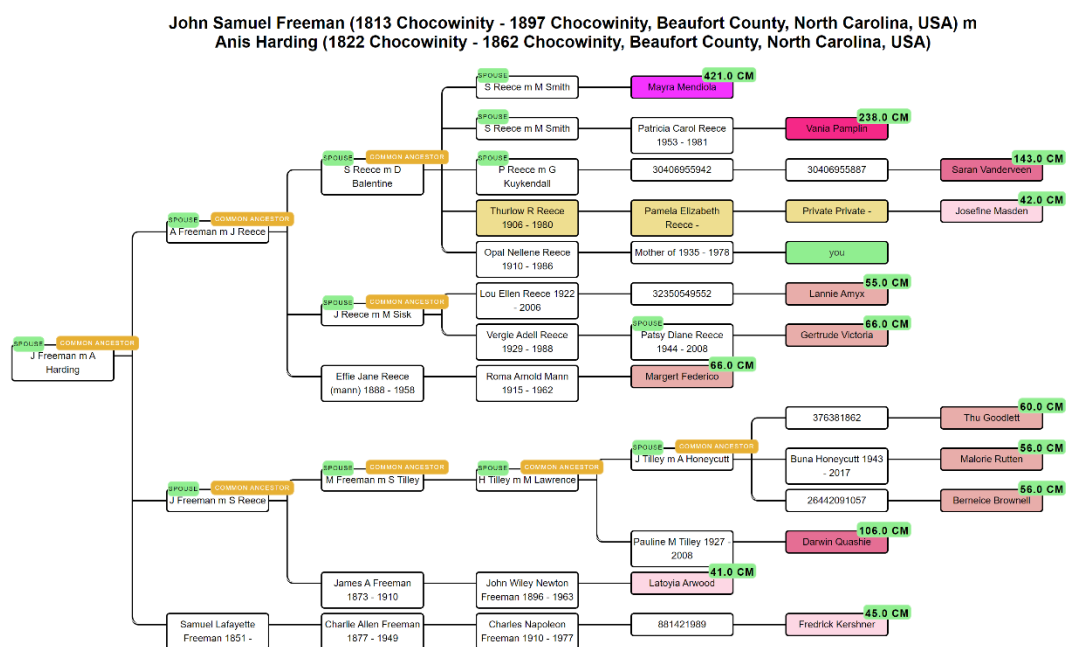


Figure 37. Reconstructed genealogical tree based on the descendants of common ancestor J Freeman and A Harding. Quite some of the descendants are a common ancestor as well. At the far right, the linked DNA matches are displayed (in green the profile of the tested person). In yellow/brown are tree persons that are retrieved from an unlinked tree.

The visualized tree persons are clickable which will show more detailed information. In addition, upon clicking on the cM value the relationship probabilities are displayed (adapted from the [Shared cM Project 3.0 tool v4 project](#), see Figure 38).

[Elke Hegna](#) with tree 8417734 shares 176.0 cM. Relationship probabilities:

- 51.03%** GGGG Aunt / Uncle Great-Great-Great-Great-Aunt / Uncle Half Great-Great-Great-Aunt / Uncle 1C3R 2C1R Half 1C2R Half 2C
- 34.51%** Great-Great-Great-Grandparent GGG Aunt / Uncle Great-Great-Great-Aunt / Uncle Half GG-Niece / Nephew Half GG-Aunt / Uncle Half Great-Great-Niece / Nephew Half Great-Great-Aunt / Uncle Half Great-Great-Niece / Nephew 1C2R 2C Half 1C1R
- 10.59%** Half GGGG-Aunt / Uncle 1C4R 2C2R 3C Half 1C3R Half 2C1R
- 2.04%** Great-Great-Grandparent Great-Great-Aunt / Uncle Great-Great-Niece / Nephew Half Great-Aunt / Uncle Half Great-Niece / Nephew 1C1R Half 1C
- 1.82%** 1C5R 2C3R 3C1R Half 1C4R Half 2C2R Half 3C

Figure 38. Relationship probabilities for a DNA match (relationship probabilities adapted from shared cM project 3.0 v4).

The complete list of common ancestor clusters with information regarding the underlying tree persons, their spouses, descendant, and linked DNA matches are available in the common ancestor's table (see Figure 39). This table is also available in de the Excel file.

### Common Ancestor clusters

Common ancestors are identified using three analyses. First, a surname clustering is performed followed by a first name clustering. Last a clustering applied based on the birth and death years. In some cases the common ancestor clusters are not visualized in the reconstituted tree. In this scenario, descendants of that common ancestor are covered by another branch of tree or via the husband. Information for each of the persons in each common ancestor cluster and their linked DNA matches is displayed in the table underneath.

Tree	Name	Birth	Death	Descendant	Name	cM	Notes preview
tree name	first name	location	location	Search	Search	cf	Search
▼ William Conrad Easterly 1865 - 1937 m Emily Jane Chism 1859 - 1920 (3 items)							
+	8417734 (27)	William Conrad Easterly 1865 -	Big Flat, ...	Maggie Eilen East...	Elke Hegna	176...	
+	157289583 (29)	William Conrad Easterly 1865 -	Big Flat, ...	William Hopkins E...	Vania Pamplin	134...	
+	den Braber/Easterly (155)	William Conrad Easterly 1865 - 1937	Fulton Co...	Jesse Pinkney Ea...	tested person	-1	
▼ Emily Jane Chism 1859 - 1920 m William Conrad Easterly 1865 - (3 items)							
+	8417734 (27)	Emily Jane Chism 1859 - 1920	Fowler, M...	Maggie Eilen East...	Elke Hegna	176...	
+	157289583 (29)	Emily Jane Chism 1859 - 1920	Fowler, M...	William Hopkins E...	Vania Pamplin	134...	
+	den Braber/Easterly (155)	Emily Jane Chism 1859 - 1920	Izard Cou...	Jesse Pinkney Ea...	tested person	-1	

Figure 39. Common ancestor table that contains all identified common ancestors, husbands, descendants and linked DNA matches.

Using the birth location information from tree persons, we calculate location clusters using a minimum location distance. These clusters are available using the common location table which contains the location, linked tree, linked DNA match and the tree persons linked to this location (see Figure 40).

### Common locations table using 5000m

We collect locations of recent ancestors that are present in the trees from different DNA matches. Locations that are characteristic of the identified clusters can yield information concerning the historical or demographic significance of a cluster. To identify these location clusters, we perform a clustering based on the distances between the entered birth locations of the persons from each of trees of the DNA matches. Birth locations that are identical or in close proximity (i.e. within a certain radius of meters) are placed into location clusters. These common locations are displayed in the table underneath.

Cluster			
Search for cluster			
▼ 0 (1 item)			
Location cluster with 3 trees and 6 people using a distance...			
Common lo...	Tree	Match	Persons
Search	Search	Search	Search
Tennessee	MARGARET METZ EA...	sharondroberts (194.505 cM)	Mary Polly Gibbons 1814 - 1919 m Zachariah Kitchens 1820 - Wiley F Hale 1817 - 1865 m Malinda King 1810 - Zachariah Kitchens 1820 - m Mary Polly Gibbons 1814 - 1919 Margaret E. Hale 1846 - 1902 m John T. Kitchen 1846 - 1898
Tennessee	Cook Family Tree	Fredrick Kershner (133.951 cM)	Nancy C Rose 1855 - m Berry A Rose 1847 - 1939
Tennessee	Charles Mckean Famil...	Mayra Mendiola (92.193 cM)	Louisa Mitchell 1827 - 1872 m James Wesley Easterly 1826 - 1875

Figure 40. Common locations table that contains the location clusters and matches of this cluster.

## AutoPedigree

AutoPedigree is a feature that employs the AutoTree predictions. It is developed to identify how a person, for instance an adoptee, fits into a reconstructed AutoTree. In short, the AutoPedigree feature automates the generation and testing of hypotheses using reconstructed trees from AutoTree.

Our approach has been inspired by the WATO tool that has been built to help solve DNA puzzles (including unknown parentage cases) by undertaking calculations as described by Leah Larkin in her series Science the heck out of your DNA. developed to identify how a person, for instance an adoptee, fits into a reconstructed AutoTree. Our approach has been inspired by the WATO tool that has been built to help solve DNA puzzles (including unknown parentage cases) by undertaking calculations as described by Leah Larkin in her series Science the heck out of your DNA.

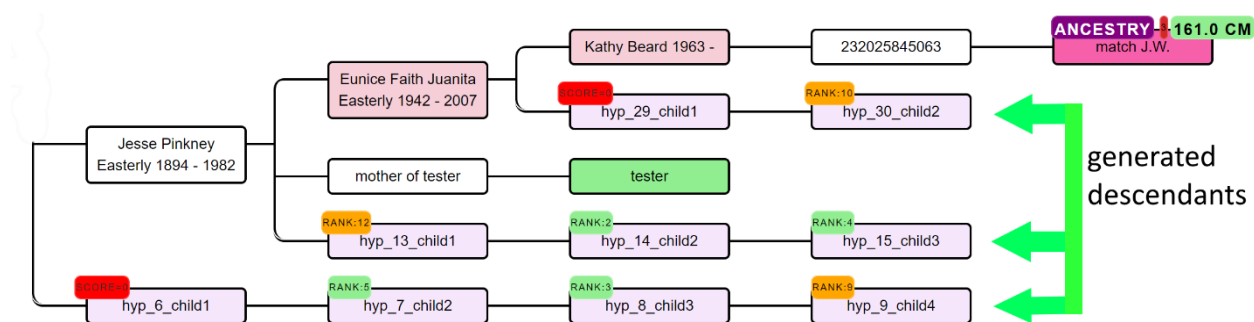


Figure 41. Partial AutoPedigree with different generated hypotheses and their ranks. The actual place in the tree is represented with the green tester rectangle

Based on the common ancestors from the reconstructed trees, we create siblings for each of the identified ancestors. Next, we generate descendants (also called hypotheses) that could serve as a hypothesis (see example in Figure 41). What that means is the following, each generated descendant could represent the actual test taker (for instance an adoptee). But given the cM values of the DNA matches in the tree, some generated descendants are more probable than others. This probability is a measure of how likely a certain relationship is to occur. For instance, a DNA match sharing 229 cM has a probability of 54% of being a 2C but a 0% probability of being a 4C (see Figure 42 for screenshot of the shared cM project that employs these probabilities as well).

By multiplying each of the probabilities for each of the DNA matches in the tree, a score can be calculated for a certain generated hypothesis. This type of analysis can

### The Shared cM Project 4.0 tool v4

**Filter**  
Enter the total number of cM for your match here:

[or enter %](#)

Then any relationships that fit will stand out below  
[Click here for a shareable link to the cM amount above](#)

**New** [Click on any relationship to view a histogram](#)

**Relationship probabilities (based on stats from The DNA Geek)**

<b>53.99%</b>	Half GG-Aunt / Uncle 2C Half 1C1R 1C2R Half GG-Niece / Nephew
<b>34.51%</b>	Half 2C 2C1R Half 1C2R 1C3R
<b>8.96%</b>	Great-Great-Aunt / Uncle Half Great-Aunt / Uncle Half 1C 1C1R Half Great-Niece / Nephew Great-Great-Niece / Nephew
<b>2.54%</b>	Half 1C3R † Half 2C1R † 3C 2C2R

† this relationship has a positive probability for 229cM in theDNAgeek's table of probabilities, but falls outside the bounds of the recorded cM range (99th percentile)

Figure 42. Example probability for a match sharing 229 cM.

also manually be performed using the online WATO tool (see Figure 43).

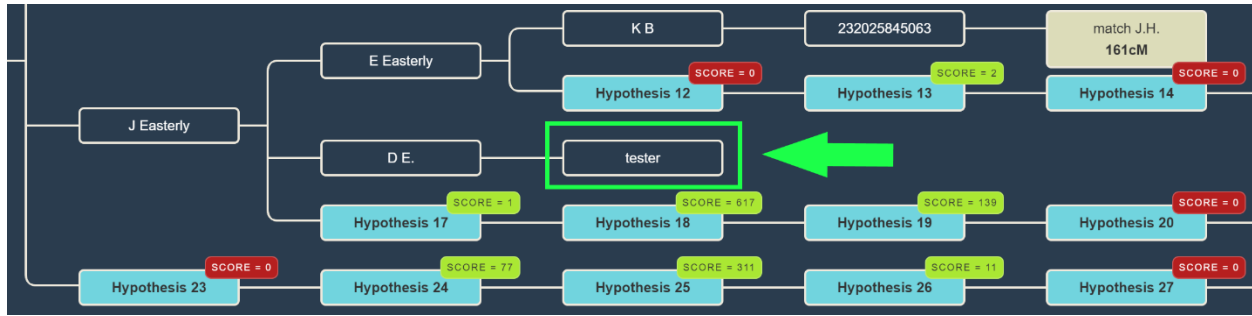


Figure 43. WATO representation of the AutoPedigree output. The different hypotheses are provided with a score.

All generated AutoPedigree trees are available in the WATO format, allowing users to import them into WATO. This allows for further tweaking of the trees, for instance if the AutoTree wrongly identified a common ancestor. In addition, matches from other companies that are known to be descendants as well can then be added.

The table underneath the visualized AutoPedigree tree summarizes the different hypotheses (see Figure 44). Each row represents a hypothesis, the MRCA and the ranked combined odds ratios. This odds ratio score is calculated based on the probability of that hypothesis divided by the smallest probability of another hypothesis. Next, we compare the score to the next, slightly smaller, score and calculated the ratio between them. For instance, if the best combined odds ratio is 200 and the second best is 50, the compared score would be 4 (200 divided by 50). The last columns show the probability of each DNA match and the generated hypothesis. By clicking on the download button above the table, the complete table becomes available as an Excel file.

[Download spreadsheet with all scores](#)

Hypotheses		Most Common Recent Ancestor		Calculations with 0 ignored		Probability of each DNA match and hypothesis			
Hypothesis	Child	MCRA	MCRA	Combined odds ratio	Compared to previous	match J.W. 161.0 cM	match M.S. 102.6 cM	match A.C 134.6 cM	match J.H. 91.2 cM
filter colour	filter c	filter column...	filter column...						
hyp_44	child2	James Wesley Easterly 1826 - 1875	Louisa Mitchell 1827 - 1877	634	1	14.16%	38.89%	50.57%	32.15%
hyp_14	child2	Jesse Pinkney Easterly 1894 - 1962	Opal Nellene Reece 1910 - 1986	617	2	28.14%	19.06%	50.57%	32.15%
hyp_8	child3	William Conrad Easterly 1865 - 1937	Emily Jane Chism 1859 - 1920	311	2.2	14.16%	19.06%	50.57%	32.15%
hyp_15	child3	Jesse Pinkney Easterly 1894 - 1962	Opal Nellene Reece 1910 - 1986	139	1.8	52.78%	5.37%	23.28%	29.83%
hyp_7	child2	William Conrad Easterly 1865 - 1937	Emily Jane Chism 1859 - 1920	77	1.2	52.78%	30.3%	16.11%	4.23%
hyp_45	child3	James Wesley Easterly 1826 - 1875	Louisa Mitchell 1827 - 1877	63	3.9	4.24%	30.3%	23.28%	29.83%
hyp_43	child1	James Wesley Easterly 1826 - 1875	Louisa Mitchell 1827 - 1877	16	1.2	52.78%	6.39%	16.11%	4.23%
hyp_24	child2	Beard Hopkins Easterly 1857 - 1925		13	1.2	4.24%	6.39%	23.28%	29.83%
hyp_9	child4	William Conrad Easterly 1865 - 1937	Emily Jane Chism 1859 - 1920	11	6.2	4.24%	5.37%	23.28%	29.83%
hyp_30	child2	Eunice Faith Juanita Easterly 1942...		2	1.6	0.68%	5.37%	23.28%	29.83%
hyp_3	child3	William Hopkins Easterly 1897 - 1965	Grace U McClung 1899 - 1982	1	1.1	4.24%	5.37%	16.11%	4.23%
hyp_13	child1	Jesse Pinkney Easterly 1894 - 1962	Opal Nellene Reece 1910 - 1986	1	1	0.68%	30.3%	16.11%	4.23%
hyp_1	child1	William Hopkins Easterly 1897 - 1965	Grace U McClung 1899 - 1982			52.78%	30.3%	0.0%	0.0%
hyp_2	child2	William Hopkins Easterly 1897 - 1965	Grace U McClung 1899 - 1982			14.16%	19.06%	0.0%	0.0%
hyp_4	child4	William Hopkins Easterly 1897 - 1965	Grace U McClung 1899 - 1982			0.0%	0.0%	50.57%	32.15%
hyp_5	child5	William Hopkins Easterly 1897 - 1965	Grace U McClung 1899 - 1982			0.0%	0.0%	23.28%	29.83%
hyp_6	child1	William Conrad Easterly 1865 - 1937	Emily Jane Chism 1859 - 1920			28.14%	38.89%	0.0%	0.0%

Figure 44. AutoPedigree scores table

Some hypotheses contain probabilities that are 0.0%, indicating that this relationship is not possible when taking into account the cM value of the DNA match and the proposed genealogical link. For instance, a DNA match that shares 200 cM cannot have a relationship of a 4C.

The generated hypotheses are visualized as descendants in the reconstructed AutoTree visualizations. For instance, hyp\_14\_child2 represents the 14th hypothesis and the second child. Instead of supplying the scores, we are providing the rank of the score in a badge. Scores that have a probability of 0 are placed in a red badge, the top 5 scores are placed in a green badge and the remainder in an orange badge. Upon clicking on the badge, a popup will appear that holds more information concerning the calculation of the score (see Figure 45).

A lot of hypotheses are tested for AutoPedigree. Therefore, in order to improve the visibility we prune the AutoTree tree by only displaying generated descendants if positive probabilities are available for that branch.

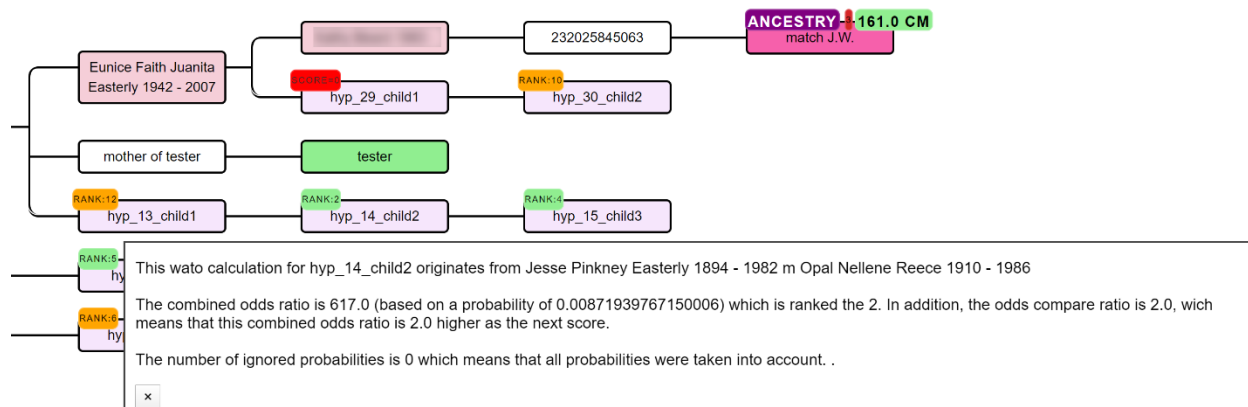


Figure 45. Pop up visualization upon clicking on the rank badge.

In some cases a DNA match has multiple links with the tested person, for instance DNA match M.M. that shares 157.9 cM and has common ancestor J Ozinga and B Ozinga (see Figure 46). The amount of cMs that is shared with the tested person is therefore inflated. To correct for this, we employ an approach that divides the amount of shared cM based on the different genealogical paths. We therefore attribute a larger fraction of cM to the DNA match if the tested hypothesis has a shorter path to the hypothesis as compared to the other path(s). Despite this measure, caution should be taken when encountering these DNA matches. The grey cells in the AutoCluster charts can be indicative of matches that are linked to more clusters and therefore linked via multiple ancestors.

Jan Fokkes Ozinga (1822 Anjum - 1890 Anjum, Oostdongeradeel, Friesland, Nederland) m Antje Jacobs Elzinga (1823 Nes - 1895 Nes, Westdongeradeel, Friesland, Netherlands)

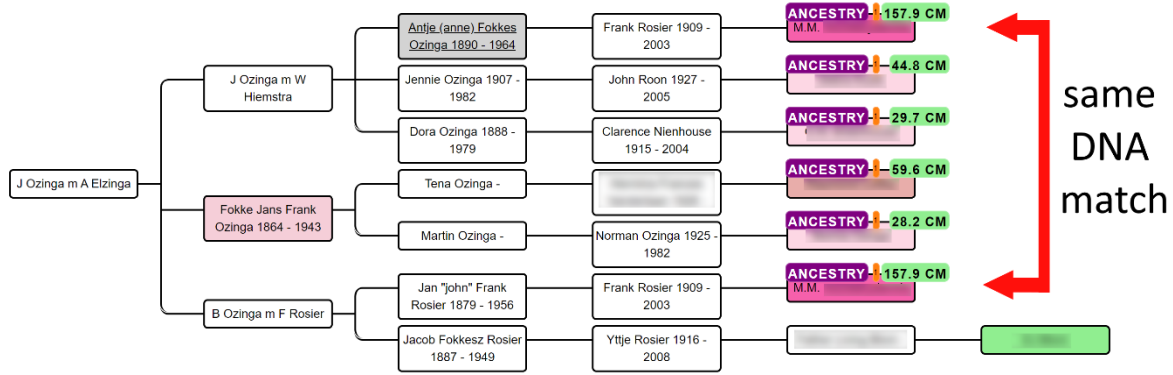


Figure 46. AutoTree reconstruction with a DNA match that has multiple links with the tested person.

A single inaccurate hypothesis in the tree can potentially nullify the overall hypothesis, making its score zero. Unfortunately, these hypotheses are sometimes inevitable, for instance because a match is related via multiple genealogical links whereas only one line is identified. In this case you might end up with a predicted 4C (based on the reconstructed tree) that shares much more DNA with the tester as is expected based on the 4C relationship. If necessary, we therefore also perform the same automated analysis while ignoring one or two of these cases. If the rank badge starts with a digit, this digit will represent the number of ignored probabilities (2\_RANK:4 indicates a hypothesis ranked 4th for which 2 probabilities were ignored). See Figure 47 for an example pop up with information concerning an ignored probability.

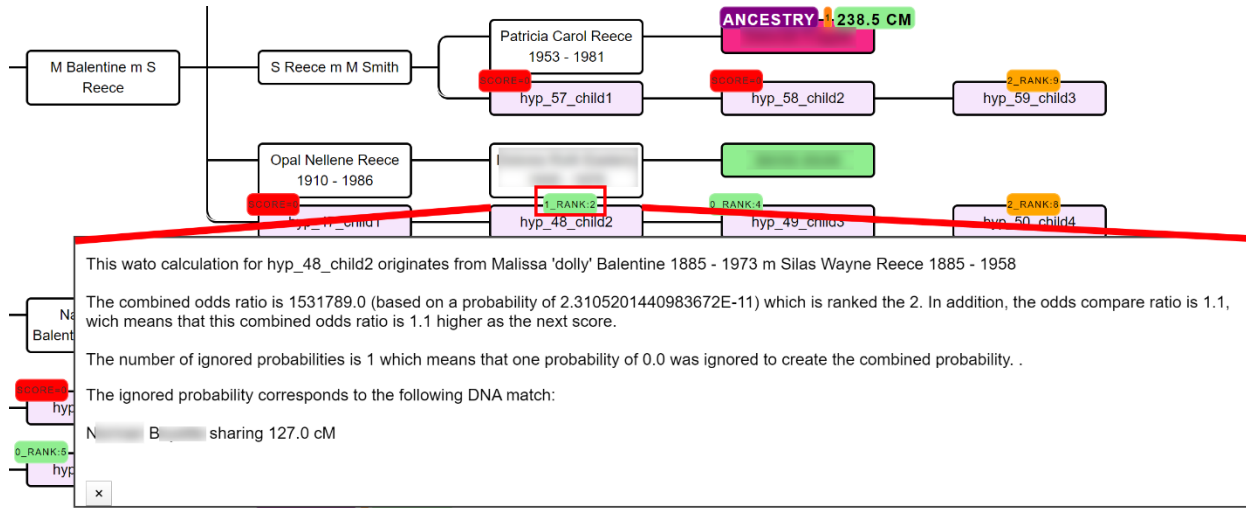


Figure 47. AutoPedigree with an ignored probability.

Invoking the AutoPedigree can be accomplished by going to the AutoCluster or AutoTree interface (see Figure 48). Select the common ancestor (AutoTree) option as well as the AutoPedigree. Next, select the min cM threshold for the analysis. This threshold indicates the minimum of shared cM that a match should share with the tested person. It is advised to use a 40 cM limit but it is possible to go down to 30 cM.

## Perform custom AutoCluster analysis with AutoTree feature for profile **EJ Blom**

Start AutoCluster analysis with matches which share a max of	Stop AutoCluster analysis with matches which share less than	Only use starred matches	Use Ancestry groups	Run on more powerful server	AutoTree identify common ancestors from trees	AutoPedigree generate hypotheses automatically	Min shared cM for AutoPedigree hypotheses	Include half relations	Extend clusters	Min cluster size
250 cM	50 cM	<input type="checkbox"/>	<input type="text"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	40	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>
<input type="button" value="PERFORM ANALYSIS"/>										

Figure 48. AutoPedigree interface, the min shared cM and half relationships can be set.



## AutoSegment – segment based clustering

AutoSegment automatically organizes your matches using shared segment clusters. It employs locally downloaded segment files, therefore no scraping and website credentials are not needed. It works for MyHeritage, 23andme, FTDNA and GEDmatch segments (see Figure 49 for MyHeritage example). An AutoSegment analysis can be started using this link: <https://members.geneticaffairs.com/autosegment>

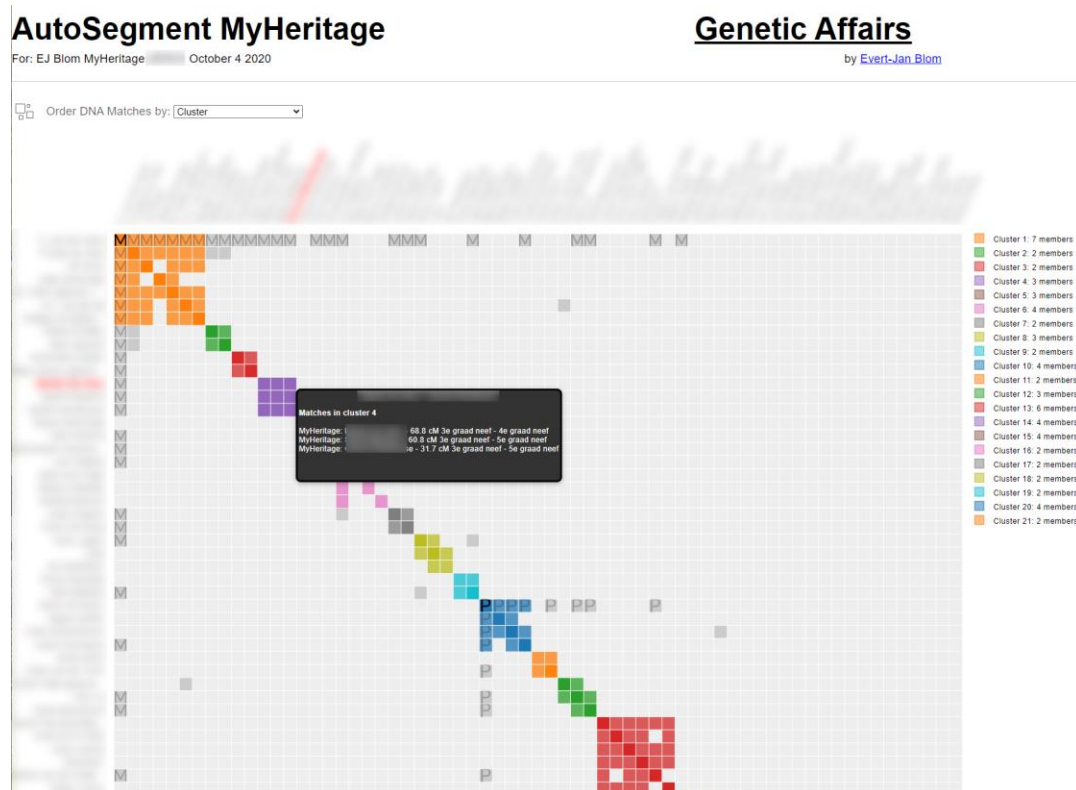


Figure 49. AutoSegment chart for MyHeritage profile with manually added maternal/paternal annotations

Let's first examine some visual examples before explaining how the AutoSegment works. Underneath the main visual chart, a table (see Figure 50) is available that contains general characteristics of the clusters. For instance, for which chromosomes segment clusters have been found, the number of paternal/maternal matches linked to a cluster, the number of segments and number of segment clusters for that cluster.

### Chromosome segment statistics per AutoSegment cluster

The following table shows the AutoSegment statistics per cluster. A link to each cluster is provided. In addition, the chromosomes linked to the segment clusters and the number of DNA matches for each cluster is shown. Last, the number of paternal/maternal DNA matches is available as well as the number of segments and number of segment clusters.

[Download spreadsheet AutoSegment statistics](#)

Cluster	chr	#mat...	Pat...	Mat...	Nr of se...	Nr of Se...
	13	<input type="text" value="filter colt"/>	<input type="text" value="filter cc"/>	<input type="text" value="filter cc"/>	<input type="text" value="filter columr"/>	<input type="text" value="filter columr"/>
▼ (8 items)						
<a href="#">AutoSegment cluster 1</a>	1,2,3,4,5,6,7,8,9,10,12,13,...	9		1	24	117
<a href="#">AutoSegment cluster 15</a>	1,8,9,10,13	5			6	26
<a href="#">AutoSegment cluster 42</a>	7,13,18	25			3	35
<a href="#">AutoSegment cluster 43</a>	13	6			1	7
<a href="#">AutoSegment cluster 57</a>	13	5			1	5
<a href="#">AutoSegment cluster 73</a>	13	8			1	8
<a href="#">AutoSegment cluster 76</a>	13,20	5			2	7
<a href="#">AutoSegment cluster 86</a>	11,13	5			2	8

Figure 50. Table that shows main characteristics of each AutoSegment cluster

After clicking on one of the links, a report displayed. This chart will display a chromosome representation (see Figure 51) of the chromosome locations that are underlying these segments.

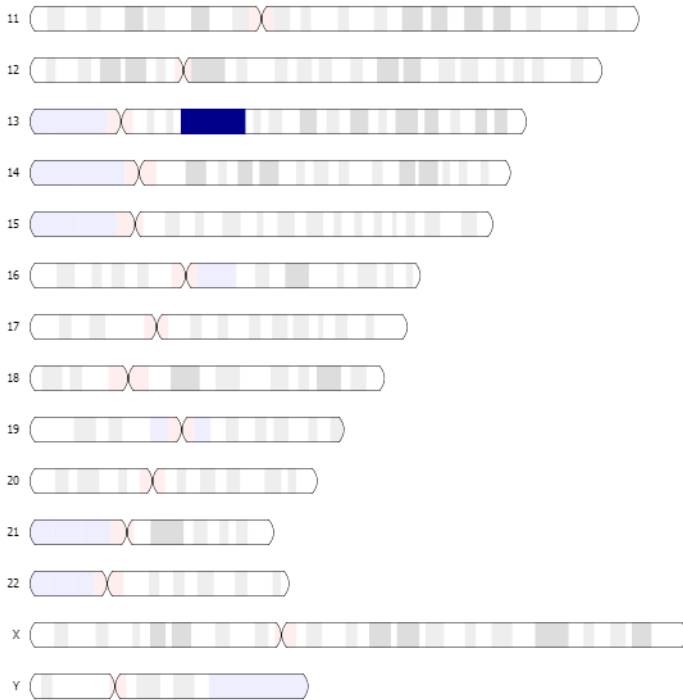


Figure 51. Chromosome visualization of the segment clusters

In this case, the segment cluster contains segments located on chromosome 13.

Underneath the chromosome browser a table is displayed that contain the segment clusters that are identified (see Figure 52). This is the core of AutoSegment, these are the overlapping segments that are identified from the analysis. This table shows the main cluster where the DNA matches are residing, the segment cluster, chromosome, start and end positions. To quickly assess the segment overlap between the segments, a visual segment representation is added. The number of SNPs is reported as well as the name of the underlying DNA matches, the DTC, the cM of the segment as well of the total cM of the match. If the match has a paternal or maternal annotation, this information will be displayed in the last two columns.

Segment Cluster Information

Clus...	Segment ...	C...	Start	Stop	Segment representation	SN...	Name	DTC	cM	To...	Pa...	M...
Segme	Segment clu	Sea	Search	Search		Search	Search for name	DTC	Max †	Total	filter †	filter †
▼ 67 (8 items)												
73	67	13	29732994	41948151		7424		MyHer...	19.1	32.5		
73	67	13	29732994	42438349		7680		MyHer...	19.6	27.2		
73	67	13	29732994	42438349		7680		MyHer...	19.6	26.2		
73	67	13	29732994	42438349		7680		MyHer...	19.6	19.6		
73	67	13	29732994	42438349		7680		MyHer...	19.6	19.6		
73	67	13	30079484	42696870		7552		MyHer...	18.6	26.2		
73	67	13	31275633	39495776		4992		MyHer...	12	33.7		
73	67	13	33593270	41948151		4864		MyHer...	11.7	19.3		

Figure 52. Segment cluster table

However, and this is also a word of caution, **AutoSegment identifies overlapping segments that do not necessarily triangulate!** The reason that we want to identify overlapping segments is because some of them will actually triangulate, and we are interested in identifying triangulated segments because they can point to a common ancestor. Conveniently, MyHeritage provides a chromosome browser (<https://www.myheritage.nl/dna/chromosome-browser>) that will show if these segments are triangulating (see Figure 53). The segments identified were checked and found to be triangulating.

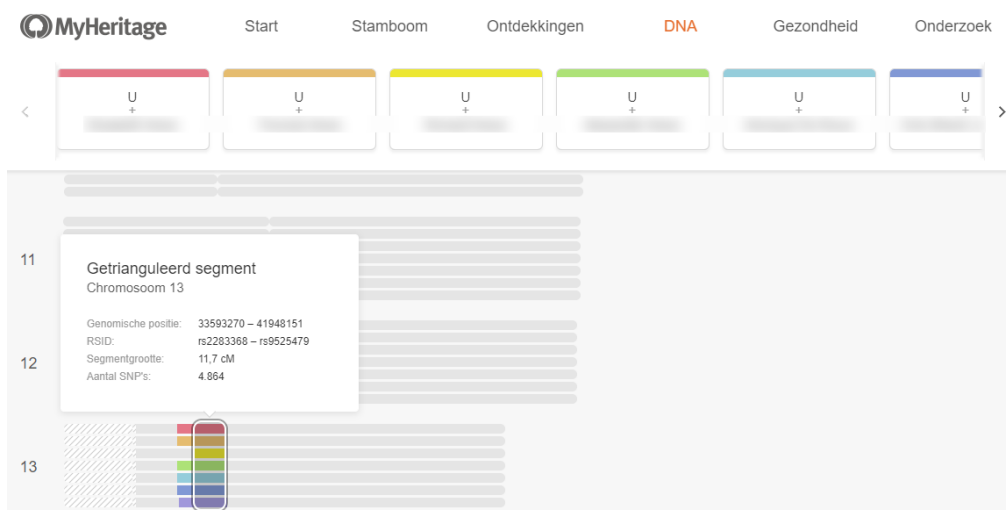


Figure 53. Chromosome browser representation of MyHeritage, indicating that the visualized segments triangulate.

## AutoSegment concepts

We will now discuss the underlying concepts of the AutoSegment clustering. The regular AutoCluster analyses are all based on shared match data, meaning that if a DNA match shares DNA with another match, they are shared matches and they might end up in the same cluster.

Let's examine one of these clusters. In this example, we will look at a cluster (see Figure 54) from a AutoCluster analysis based on a FamilyTreeDNA profile. As you can see, it's a well connected cluster, meaning that all members are sharing DNA with each other.



Figure 54. Example cluster from regular AutoCluster analysis for a FTDNA profile.

Now let's examine these matches using the chromosome browser of FamilyTreeDNA (see Figure 55). As we can immediately see in the chromosome browser, all shared DNA segments are not shared by other DNA matches. The exception seems to be on chromosome 7, but the two matches are closely related (mother/daughter). The only overlap with another match seems to be on chromosome 19.



Figure 55. Chromosome browser of all members of a certain FTDNA cluster

So in a way, by only looking at overlapping segments we are restricting ourselves. The fact that these matches in the aforementioned cluster do not share a cluster does not make them less useful. They will still probably share a common ancestor. The question now might be asked, why then look at overlapping segments? There are several reasons. First, if the overlapping segments are found to be triangulating, they might be good candidates for DNA painting procedures. Second, the overlapping segment analyses can be performed using locally downloaded files, so there is no need to supply the website credentials and scrape the website. Third, some DNA testing companies, such as MyHeritage and GEDmatch, were not covered by our website. Last, since the overlapping segments are a generic procedure, we can employ the same approach DNA segment data from several DNA testing companies, thereby combining the matches into one single clustering. We will cover this hybrid approach in a later stage.

Another example of the use of shared match data vs overlapping segments is illustrated in Figure 56. This figure shows a regular AutoClusters chart as compared to the AutoSegment output. Since the AutoSegment analysis is restricted to overlapping segments, it shows a lower number of grey cells.

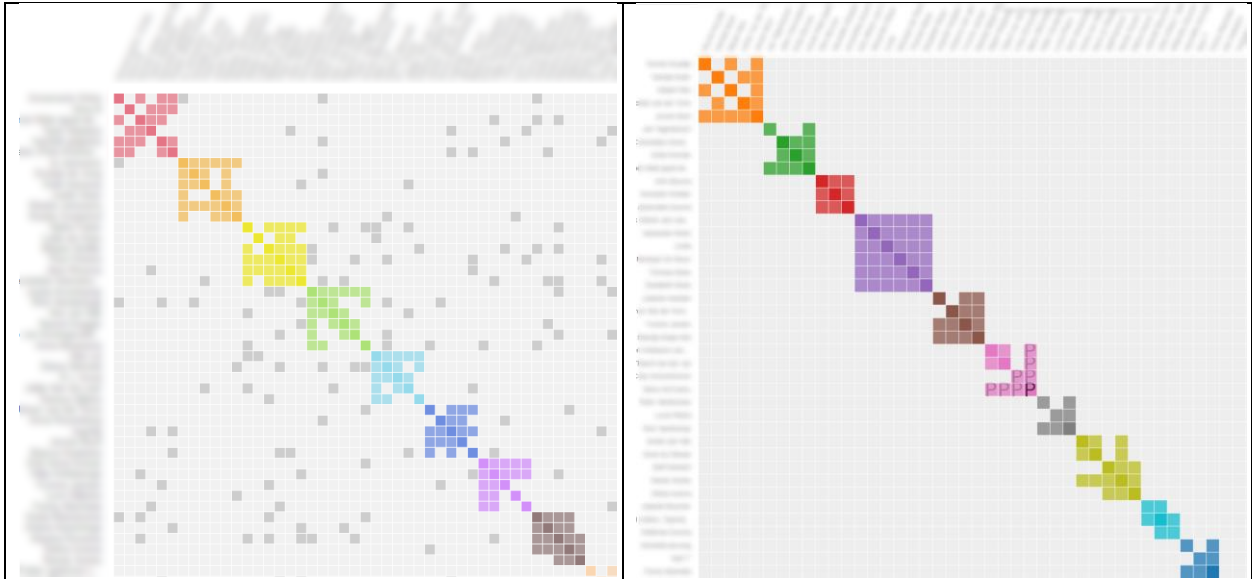


Figure 56. MyHeritage AutoClusters (based on shared matches), vs AutoSegment (based on overlapping segments).

Now for the underlying concepts of an AutoSegment analysis. Remember that DNA matches are linked via their shared matches but in this case, we don't have these available. Instead, we use segments and use these to link the matches.

The analysis starts out like a regular AutoCluster analysis. The user can define a specific max-min cM (for instance matches between 400 cM – 20 cM) but in addition a min segment overlap needs to be provided. For instance, 10 cM.

This segment overlap will be used as a minimum overlap measure, segments that share less than the provided amount, will not be linked. We calculate the segment overlap between 2 segments by looking at the overlap (see Figure 57). To provide the exact amount of overlapping cM information, we employ a human genetic map (build 37). However, before we calculate this overlap, we first check if the maternal/paternal is set, and if so, is the same for the underlying matches.



Figure 57. Segment overlap calculation

If the identified segment overlap exceeds the minimum overlap, the segments are linked. After linking all overlapping segments, we can examine these segment networks and identify segment clusters in them (see Figure 58 where we visualize 9 linked segments and are able to identify 2 segment clusters).

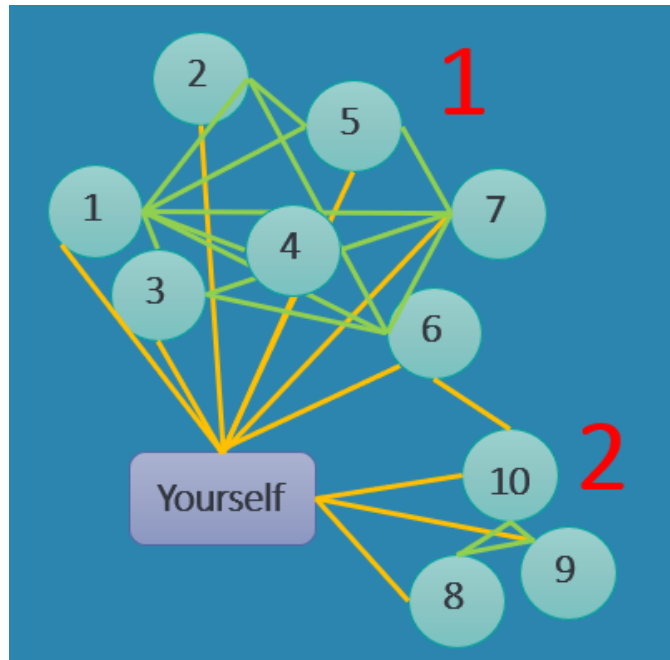


Figure 58. Linked segments. Numbers represent segments whereas a link between segments indicates an overlap.

After identifying many of these segment clusters, it's time to go back to our DNA matches. Ultimately, we want to cluster the DNA matches based on the segment clusters (see Figure 59).

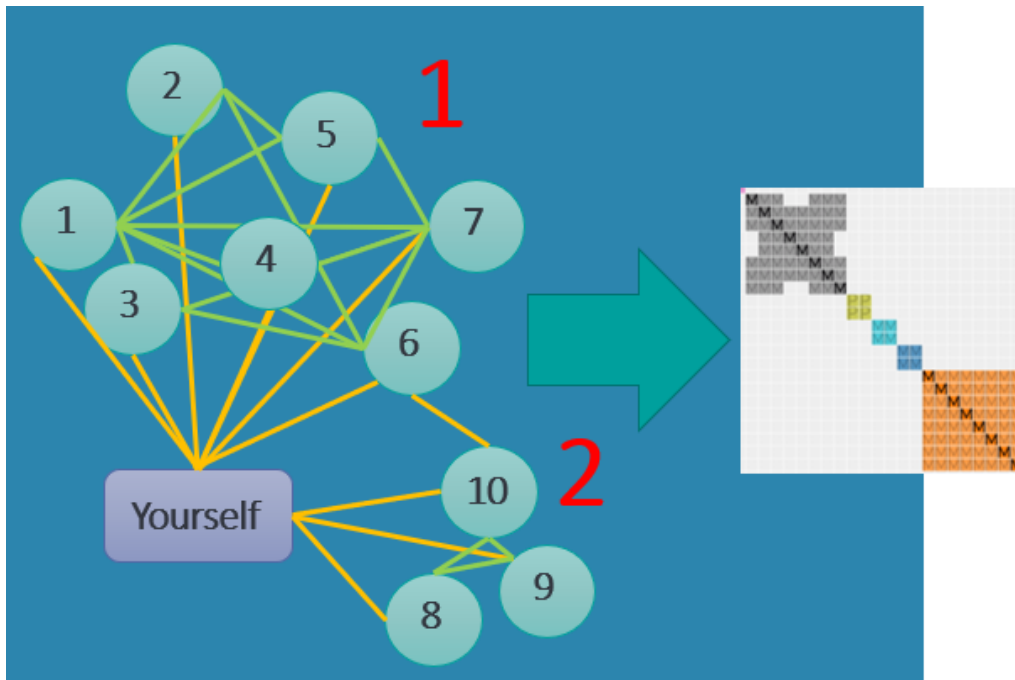


Figure 59. Convert segment clusters to DNA match clusters

By using these segment clusters we can link DNA matches. Here is how it works, we examine all of the segment clusters and check if a certain DNA matchA and DNA matchB have a segment linked to this segment cluster. If so, a link between these two matches is created. Now the table described in Figure 50 makes more sense. Some high cM matches might be linked via several segment clusters and others only via one segment cluster.

Underneath the segment cluster chart, as described before, a table is displayed (see Figure 60) that contains the general segment statistics, cluster and segment cluster id, chromosome, start, end and visual representation.

### Segment Cluster Information

Cluster	Segment ...	Chr	Start	Stop	Segment representation	SNP c...	Name	cM	To...
<input type="text" value="Segment clu"/>	<input type="text" value="Segment clu"/>	<input type="text" value="Sear"/>	<input type="text" value="Search ft"/>	<input type="text" value="Search ft"/>		<input type="text" value="Search ft"/>	<input type="text" value="Segment clu"/>	<input type="text" value="Max €"/>	<input type="text" value="Total"/>
▼ 22 (3 items)									
1	22	1	14576441	20755790		2076		13.4	0
1	22	1	14128833	20755790		2276		14.3	0
1	22	1	14576441	20755790		2076		13.4	0
▼ 25 (2 items)									
1	25	1	30872516	47629198		4065		18.2	0
	25	1	30660319	47629198		4165		18.7	0
▼ 31 (2 items)									
1	31	11	76690268	108633181		8400		28.5	0
	31	11	87132158	108633181		2231		20.3	0
▼ 39 (2 items)									
1	39	1	56987614	99291699		11597		43.5	0
1	39	1	66512322	99626055		8797		30.3	0

Figure 60. Segment cluster information with general segment statistics.



In addition to the segment cluster tables displayed on each cluster page, there is a large table on the main page that describes all identified segment clusters. This overview is useful because in some cases DNA matches are not clustered but do appear in certain segment clusters. This complete overview allows the analysis of the DNA matches that did not make it into the main clustering (see Figure 61).

### Individual segment Cluster Information

The following table shows the 376 DNA segments for each of the 96 identified segment clusters.

[Download spreadsheet with segment clusters](#)

Cluster	Segmen...	C...	Start	Stop	Segment representation	S...	Name	DTC	cM	To...	Pa...	M...
<input type="text" value="Search f"/>	<input type="text" value="Segment cl"/>	<input type="text" value="Sez"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>		<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Max ct"/>	<input type="text" value="Max"/>	<input type="text" value="Total"/>	<input type="text" value="filter"/>	<input type="text" value="filter"/>
▼ 65 (4 items)												
<a href="#">59</a>	65	10	100076444	116653136		8704		MyHerit...	15.9	30.8		
<a href="#">59</a>	65	10	97345469	116653136		10112		MyHerit...	18.8	27.3		
<a href="#">59</a>	65	10	98118257	117713221		10112		MyHerit...	18.7	26.2		
<a href="#">59</a>	65	10	99329418	112305321		6400		MyHerit...	10.2	23.5		
▼ 40 (2 items)												
<a href="#">75</a>	40	2	101598312	119751867		8320		MyHerit...	17	50.3		
<a href="#">75</a>	40	2	106811149	117686719		4480		MyHerit...	10.1	18.4		
▼ 87 (2 items)												
<a href="#">40</a>	87	7	101630597	121217473		8576		MyHerit...	16.3	24.5		
<a href="#">40</a>	87	7	105088693	121217473		7040		MyHerit...	13.8	20.2		
▼ 74 (2 items)												
<a href="#">51</a>	74	8	1022799	4446859		4352		MyHerit...	10.6	23.1		
<a href="#">51</a>	74	8	164984	5323185		6144		MyHerit...	13.6	28.8		
▼ 76 (2 items)												
<a href="#">50</a>	76	9	103675970	124109921		12544		MyHerit...	28	28		
<a href="#">50</a>	76	9	107370498	118885534		7552		MyHerit...	17.9	24.6		

Figure 61. . Individual segment cluster overview for all identified segment clusters

In some areas in the genome, there are regions where a considerable amount of people are sharing certain DNA segments. These regions are known as so-called pile-up regions. For example, the segments from the segment cluster represented in

### Segment Cluster Information

Cluster	Segment ...	C...	Start	Stop	Segment representation	SNP c...	Name	cM	To...
<input type="text" value="Segment clu"/>	<input type="text" value="Segment clu"/>	<input type="text" value="Sea"/>	<input type="text" value="Search ft"/>	<input type="text" value="Search ft"/>		<input type="text" value="Search ft"/>	<input type="text" value="Segment clu"/>	<input type="text" value="Max i"/>	<input type="text" value="Total"/>
▼ 6 (103 items)									
1	6	15	20004966	33965738		6144		43.3	138.6
1	6	15	20004966	33439811		5632		41.3	125.3
1	6	15	20004966	34079300		6272		43.7	122.7
1	6	15	20004966	33439811		5632		41.3	121.1
1	6	15	20004966	33871785		6016		42.7	104.4
1	6	15	20004966	29581108		3840		33.9	97.9
1	6	15	20004966	33439811		5632		41.3	95.4
1	6	15	20004966	33965738		6144		43.3	94.4
1	6	15	20004966	33719535		5888		42.5	93.9
1	6	15	20004966	33871785		6016		42.7	93.3
1	6	15	20004966	27770160		3072		30.8	92.2
1	6	15	20004966	28328485		3456		31.9	90.7
1	6	15	20004966	33719535		5888		42.5	90.4
1	6	15	20004966	33439811		5632		41.3	89.9
1	6	15	20004966	33871785		6016		42.7	88.7
1	6	15	20004966	33582463		5760		41.9	88.2
1	6	15	20004966	27770160		3072		30.8	87.7
1	6	15	20004966	27770160		3072		30.8	87.6
1	6	15	20004966	33965738		6144		43.3	87.2

Figure 62 all are found on chromosome 15. This region overlaps with pile-up regions that are identified in a study of Li *et al* 2014 (see Figure 63).

### Segment Cluster Information

Cluster	Segment ...	C...	Start	Stop	Segment representation	SNP c...	Name	cM	To...
Segment clu	Segment clu	Sea	Search fr	Search fr		Search fr	Segment clu	Max r	Total
▼ 6 (103 items)									
1	6	15	20004966	33965738		6144		43.3	138.6
1	6	15	20004966	33439811		5632		41.3	125.3
1	6	15	20004966	34079300		6272		43.7	122.7
1	6	15	20004966	33439811		5632		41.3	121.1
1	6	15	20004966	33871785		6016		42.7	104.4
1	6	15	20004966	29581108		3840		33.9	97.9
1	6	15	20004966	33439811		5632		41.3	95.4
1	6	15	20004966	33965738		6144		43.3	94.4
1	6	15	20004966	33719535		5888		42.5	93.9
1	6	15	20004966	33871785		6016		42.7	93.3
1	6	15	20004966	27770160		3072		30.8	92.2
1	6	15	20004966	28328485		3456		31.9	90.7
1	6	15	20004966	33719535		5888		42.5	90.4
1	6	15	20004966	33439811		5632		41.3	89.9
1	6	15	20004966	33871785		6016		42.7	88.7
1	6	15	20004966	33582463		5760		41.9	88.2
1	6	15	20004966	27770160		3072		30.8	87.7
1	6	15	20004966	27770160		3072		30.8	87.6
1	6	15	20004966	33965738		6144		43.3	87.2

Figure 62. Example segment cluster that overlaps with a known pile up region

Chromosome	Starting position	Ending position	Genetic length (in cM)
chr9	38,293,483	72,605,261	8.15
chr8	10,428,647	13,469,693	7.96
chr21	16,344,186	19,375,168	6.91
chr10	44,555,093	53,240,188	7.58
chr22	16,051,881	25,095,451	20.82
chr2	85,304,243	99,558,013	6.53
chr1	118,434,520	153,401,108	9.95
chr15	20,060,673	25,145,260	10.46
chr17	77,186,666	78,417,478	5.66
chr15	27,115,823	30,295,750	9.29
chr17	59,518,083	64,970,531	6.23
chr2	132,695,025	141,442,636	9.16
chr16	19,393,068	24,031,556	6.18
chr2	192,352,906	198,110,229	5.04
<b>Total</b>	<b>14 regions</b>		119.92

doi:10.1371/journal.pgen.1004144.t003

Figure 63. Identified pile up regions in a study of Li *et al* 2014

We have added an option to the AutoSegment interface that allows users to filter these known pile-up regions. Segments that are located in these regions are removed from the analysis.

However, there can still be some regions in the genome that are enriched with respect to segments. To provide insights into these regions, we visualize the number of segments for each chromosome (see Figure 64). The pileup report link is found in the main HTML, the file called `pileup_report.html` and created in the `chromosomes` folder.

## Pile up regions

[AutoSegment](#) identifies DNA segment clusters by clustering overlapping segment data from various DNA testing companies (MyHeritage, 23andme, FamilyTreeDNA and GEDmatch) using a minimum segment overlap of 10 cM. In some cases, certain areas on your DNA might be overrepresented with respect to the number of DNA matches that match on that area. These areas are also known as pile-up regions. By plotting the segment occurrences for each chromosome, we can obtain insights in these regions. For this particular analysis, previously identified pile up regions from the paper of [Li et al \(2014\)](#) were used to remove segments that are in these pile-up regions. As a consequence, a total of 349 segments were filtered that were present in known pile-up regions.

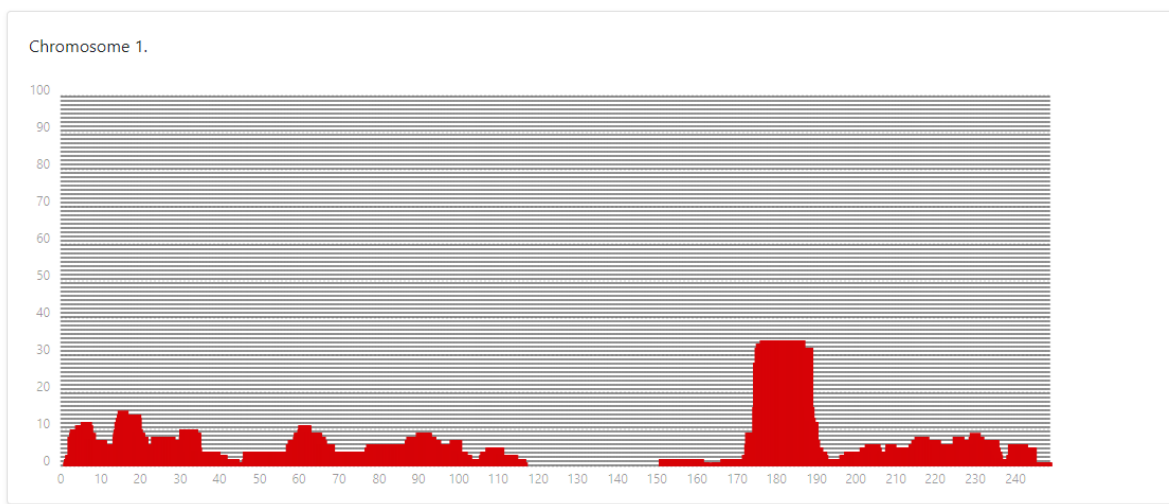


Figure 64. Personal pileup region report

GEDmatch triangulation data

The word of caution that was mentioned in the beginning needs additional explanations with respect to the GEDmatch analysis. For MyHeritage and FTDNA analyses we employ two files, one containing the DNA match data and one containing the segments. For 23andme analyses, the segments and DNA matches combined into one file. For GEDmatch we also employ 2 files, one segment file but the other file contain triangulation data (see next section how to obtain this data set). Since we have this triangulation data, we can now verify overlapping segments and only keep the overlapping segments that have a triangulation segment linked to them. This greatly improves the accuracy of the predictions.

### Excel cluster representation

In some cases, the numbers of cluster in the HTML file is too large to open using a browser. Luckily, an Excel is created that will show a spreadsheet representation of the generated clusters. Usually, this file can be opened if the HTML is unresponsive.

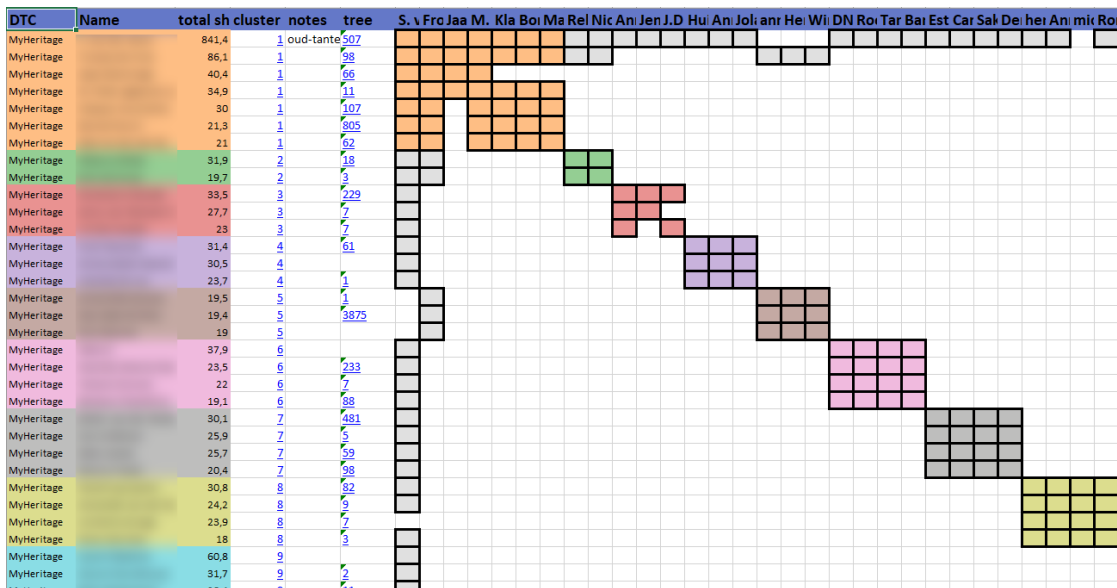


Figure 65. Excel representation of the AutoSegment clusters

In addition, to still view the cluster table and segment cluster table information is contained in the HTML, open the file that ends with “\_chart.html” which is the same HTML file without the chart.

### Paternal & Maternal annotations

Profiles from FTDNA and 23andme that have tested some close relatives can sometimes have the maternal and paternal labels (see Figure 66). This information is used to assess if certain overlapping segments are valid. It is also possible to add these labels manually. To enable this, open the DNA match file CSV using a spreadsheet edit and add the “maternal” or “paternal” annotation in the last column of this spreadsheet. Save the modified spreadsheet as a CSV file and use it for the AutoSegment clustering.



All (2004)		Paternal (209)	Maternal (297)	Both (2)	Calculating Family Matching			
Name	Match Date	Relationship Range	Shared cM	Longest Block	X-Match	Linked Relationship	Ancestral Surnames	
<input type="checkbox"/> 	07/06/2015	Father/ Son	3,384	267		Father	(van der Veen) / Abeles / Aukes / Aetzes / Adema / Alberts / Alles /	
<input type="checkbox"/> 	02/11/2016	Mother/ Daughter	3,384	267	X-Match	Mother	(Boetje) / (Deddes) / (Hoekstra) / (Jager) / (Luimstra) / (van der Harst) / Aabes / Abels	

Figure 66. Paternal and maternal annotations for FTDNA

## AutoSegment – retrieve offline data

This section describes how to obtain the DNA match and/or segment matches for MyHeritage, 23andme, FamilyTreeDNA and segment/triangulation data for GEDmatch.

### Retrieving segment data for MyHeritage

Login to MyHeritage and visit the DNA matches page: <https://www.myheritage.com/dna/matches>

Click on the three vertical dots to expand the menu. Select the first option to retrieve the DNA match list, the second option will allow the retrieval of the segment data. Both files will be mailed.

The screenshot shows the MyHeritage DNA matches interface. At the top, there's a 'DNA results' header with a profile picture and a 'Test additional family members' banner. Below this are navigation tabs: 'Overview', 'Ethnicity Estimate', 'DNA Matches' (highlighted), and 'Tools'. A red arrow points to the three vertical dots menu icon in the top right corner of the matches list. The matches list shows 'Showing 1–10 of 4,273 DNA Matches'. The menu is expanded, showing three options: 'Export entire DNA Matches list', 'Export shared DNA segment info for all DNA Matches', and 'What are DNA Matches?'. The 'Export shared DNA segment info for all DNA Matches' option is highlighted with a red box. Below the matches list, there are buttons for 'Review DNA Match' and 'View tree'.

Save the attachments from the e-mails and go to the Genetic Affairs AutoSegment page for MyHeritage: <https://members.geneticaffairs.com/autosegment/addWebsite/MyHeritage>

Hi ejblom,

Run an AutoSegment analysis for MyHeritage using the segment data and DNA match file.

The MyHeritage **DNA match file** is available on the [DNA Matches](#) page by clicking the 'three vertical dots' icon at the top right of the list of matches to expand the menu. Click on 'Export complete DNA match list' and the file will be emailed to you within a few minutes. Please unzip the file before uploading.

The **segment file** is on the same page as the DNA match file. Click on 'Export shared DNA segment info for all DNA Matches' and the file will be emailed to you within a few minutes. Please unzip the file before uploading.

For more info concerning the segment download, check [the faq section](#) of DNA Painter. Click [here](#) for a blog post from [Patsy Coleman](#) that describes her findings with AutoSegment.

Start AutoSegment analysis with matches which share a max of	Stop AutoSegment analysis with matches which share less than	Min overlapping segment size	Min cluster size	Remove known pileups	AutoSegment name	Select match file	Select segment file
250 c ▾	25 c ▾	15 c ▾	2 ▾	<input type="checkbox"/>	<input style="width: 100%;" type="text"/>	Bestand kiezen <small>Geen bestand</small>	Bestand kiezen <small>Geen bestand gekozen</small>
<input style="background-color: #4CAF50; color: white; padding: 5px 15px; border: none;" type="button" value="PERFORM AUTOSEGMENT ANALYSIS"/>							

Adjust the search parameters, fill in the name field and select the segment file and match file. After you start the analysis, results will appear within 15 min in your mailbox. If no results appear, it is possible that too low cM values were employed. Try raising the minimum cM settings and retry the analysis.



## Retrieving segment data for FamilyTreeDNA

Login into FamilyTreeDNA and visit your Family Finder page:

<https://www.familytreedna.com/my/familyfinder>. Go to the bottom of the page and click on “Download Matches: CSV”

<input type="checkbox"/>	Profile	Date	Relationship	Shared Segments	Shared cM	Actions
<input type="checkbox"/>	[Profile]	02/25/2014	2nd Cousin - 4th Cousin	53	21	[User] + [Close]
<input type="checkbox"/>	[Profile]	12/20/2018	2nd Cousin - 4th Cousin	53	17	[User] + [Close]
<input type="checkbox"/>	[Profile]	05/06/2019	2nd Cousin - 4th Cousin	44	17	[User] + [Close] Waterink (Hardenberg)
<input type="checkbox"/>	[Profile]	01/16/2017	2nd Cousin - 4th Cousin	42	17	[User] + [Close]

Download Matches: **CSV**  
Download Filtered Matches: **CSV**

1-30 of 1970 << < > >> Page 1 / 66 Go

Save the matches file to your local drive. Next, visit the chromosome browser page for the segment data: <https://www.familytreedna.com/my/family-finder/chromosome-browser>

Compare

Evert-Jan Blom YOU

With

To compare, start by selecting up to 7 from your DNA matches

DNA Matches

All Matches

Search First or Last Name

<input type="checkbox"/>	Name	Relationship Range ↑	Shared Segments	Shared cM	Longest Block	Actions
<input type="checkbox"/>	[Profile]	[Relationship]	[Shared Segments]	[Shared cM]	[Longest Block]	[More]
<input type="checkbox"/>	[Profile]	[Relationship]	[Shared Segments]	[Shared cM]	[Longest Block]	[More]
<input type="checkbox"/>	[Profile]	[Relationship]	[Shared Segments]	[Shared cM]	[Longest Block]	[More]
<input type="checkbox"/>	[Profile]	[Relationship]	[Shared Segments]	[Shared cM]	[Longest Block]	[More]

**Download All Segments**

Click on “Download all segments” to download the segment data.

Next, go to the Genetic Affairs AutoSegment page for FamilyTreeDNA:

<https://members.geneticaffairs.com/autosegment/addWebsite/FamilyTreeDNA>

Hi ejblom,

Run an AutoSegment analysis for FamilyTreeDNA using the segment data and match file.

The FTDNA **DNA match file** is available on the bottom of the [FamilyFinder](#) page. Click on 'Download Matches: CSV' link and the file will be made available to you.

The **segment file** is available via the button 'Download All Segments' in the [ftDNA chromosome](#) browser. The button is at the top right of the page above the list of matches. Please note that if you transferred to ftDNA, you'll need to unlock their chromosome browser before you can download your data.

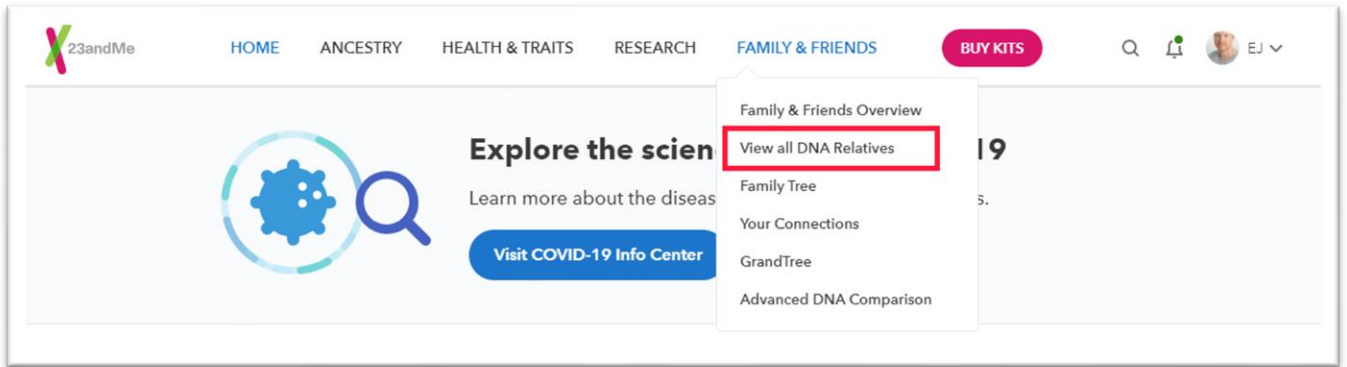
For more info concerning the segment download, check [the faq section](#) of DNA Painter. Click [here](#) for a blog post from [Patsy Coleman](#) that describes her findings with AutoSegment.

Start AutoSegment analysis with matches which share a max of	Stop AutoSegment analysis with matches which share less than	Min overlapping segment size	Min cluster size	Remove known pileups	AutoSegment name	Select match file	Select segment file
250 cM	45 cM	15 cM	2	<input type="checkbox"/>		<input type="button" value="Bestand kiezen"/> Geen bestand	<input type="button" value="Bestand kiezen"/> Geen bestand gekozen

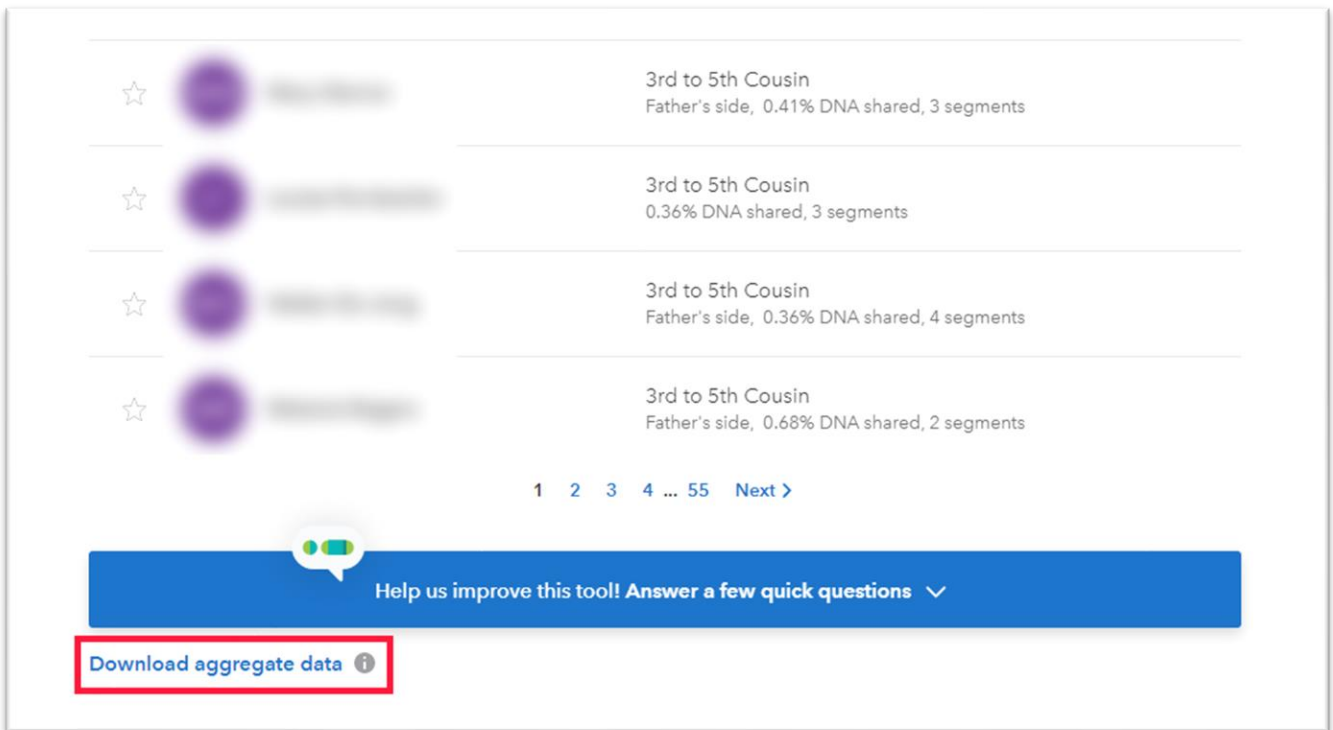
Adjust the search parameters, fill in the name field and select the segment file and match file. After you start the analysis, results will appear within 15 min in your mailbox. If no results appear, it is possible that too low cM values were employed. Try raising the minimum cM settings and retry the analysis.

Retrieving segment data for 23andme

The AutoSegment analysis for 23andme is the only analysis that requires one file. The reason for this is because 23andme has merged the DNA match data into the segment file. To obtain this file, first go to <https://www.23andme.com> and login. Next, go to the DNA relatives page (<https://you.23andme.com/tools/relatives/>):



On the DNA relatives page, scroll to the bottom and click on “Download aggregate data”



This will allow you to download all segment/match data to your local drive. You can also obtain this file by using to this direct link: <https://you.23andme.com/tools/relatives/download/>

Next, go to the Genetic Affairs AutoSegment page for 23andme:  
<https://members.geneticaffairs.com/autosegment/addWebsite/23andme>

Hi ejblom,

Run an AutoSegment analysis for 23andme using the segment data. The file is available via the link 'Download aggregate data' on the 23andme DNA Relatives page under 'Family and Friends' (you will need to scroll to the bottom of the page to see this link). If you are logged in, the following direct link will also download the file directly to your computer: [Download Segment Data from 23andme](#) (usually called [name]\_relatives\_download.csv)

For more info concerning the segment download, check [the faq section](#) of DNA Painter. Click [here](#) for a blog post from [Patsy Coleman](#) that describes her findings with AutoSegment.

Start AutoSegment analysis with matches which share a max of	Stop AutoSegment analysis with matches which share less than	Min overlapping segment size	Min cluster size	Remove known pileups	AutoSegment name	Select segment file
250 cM	25 cM	15 cM	2	<input type="checkbox"/>		Bestand kiezen   Geen bestand gekozen

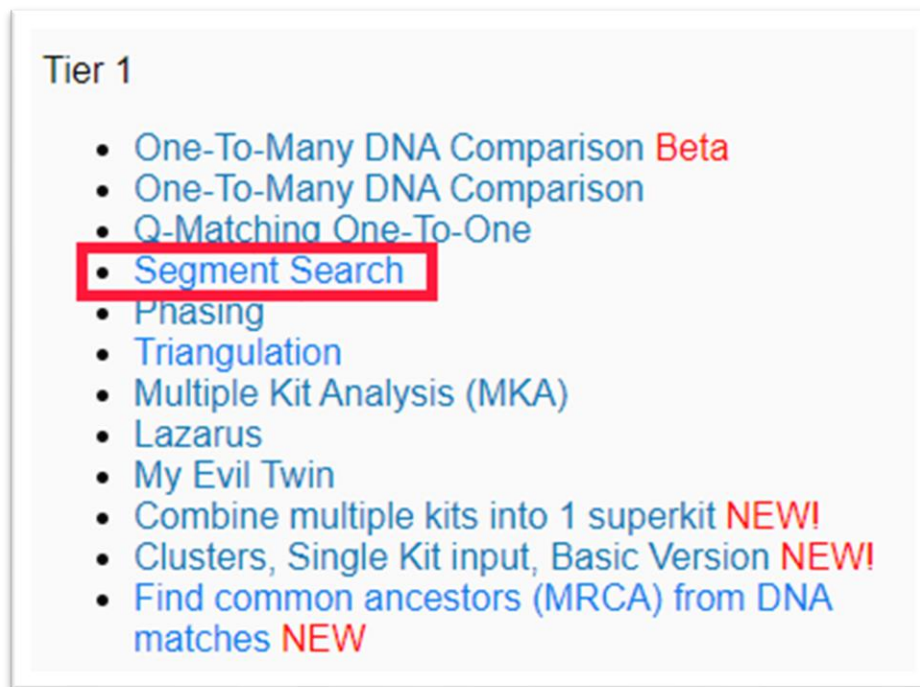
**PERFORM AUTOSEGMENT ANALYSIS**

Adjust the search parameters, fill in the name field and select the segment/match file. After you start the analysis, results will appear within 15 min in your mailbox. If no results appear, it is possible that too low cM values were employed. Try raising the minimum cM settings and retry the analysis.


Retrieving segment data for GEDmatch

GEDmatch is the only company that provides triangulated segments in addition to regular DNA segment data. AutoSegment employs both files since the regular DNA segment file contains information concerning the DNA matches that is not available in the triangulated segments file. The triangulated segment file greatly improves the quality of the AutoSegment predictions.

Log into GEDmatch and select the Tier 1 – Segment Search ([https://www.gedmatch.com/segment\\_search.php](https://www.gedmatch.com/segment_search.php)):



Fill in the concerned Kit Number, select the max number of closest matches to consider (for instance 5000) and enable the “Prevent Hard Breaks” option:



**GEDmatch** Tools for Genealogy Research

[Home](#) [Log out](#)

---

## GEDmatch<sup>®</sup> DNA Segment Search

This utility allows you to find other kits with matching chromosome segments. (Note that matches closer than 2100 cm's are skipped to save resources.)

Kit Number:

Max number of closest matches to consider:  Max Kits

Build to Display (Choose just one):  B36  B37  B38

SNP count minimum threshold to be considered a matching segment  
(Leave blank for dynamically-calculated value (200 - 400))

Minimum segment cM size to be included in total:  
(Leave blank for default value = 7)

Prevent Hard Breaks (default is to create hard breaks when distance between SNP's exceeds 500,000 base positions):

Chromosome to scan (or all)

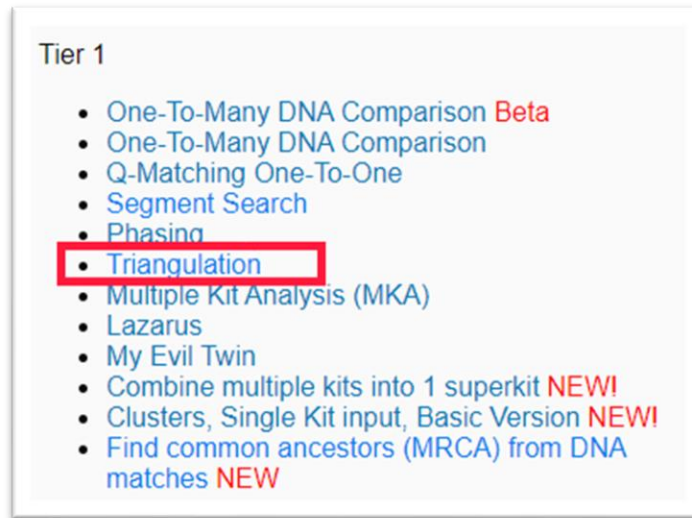
Optional segment start to match:  
(Only use if a specific chromosome is specified)

Optional segment end to match:  
(Only use if a specific chromosome is specified)

Show graphic bar for Chromosome?  Yes  No

Click on submit and let the tool analyze the data. After it is finished, click on the “here” button in the top of the screen and save the **csv** file to your local drive.

Next, we will download the triangulated data from GEDmatch. From the main page, select the triangulation option:



Fill in the concerned Kit Number, select the max number of closest matches to consider (for instance 5000) and start the analysis by clicking on the submit button:

**GEDmatch** Tools for Genealogy Research Home Log out

### GEDmatch Segment Triangulation

This utility finds people who match you with your top close matches as shown in the one-to-many results and below the upper threshold limit that you specify. It then compares those matches against each other. Results can be sorted by chromosome and position, or by kit number, chromosome and position, and then displayed in tabular and graphical format for each matching segment larger than 7 cM. Close relatives can be excluded from results by specifying an upper segment threshold limit. All kits must have completed batch processing to be included in results.

Kit Number:

Max number of closest matches to consider:  Max Kits

Upper Segment Threshold Limit:  cM

Minimum Segment length:  Minimum cM

Chromosome to triangulate (or all):

Build to Display (Choose just one):  B36  B37  B38

Display Options:

- Show results sorted by chromosome, segment start position
- Show results sorted by kit\_number, chromosome, segment start position
- Show results sorted both ways

Cross-match triangulated segments with others within chromosome.

Cross-match limit per chromosome:  Cross-Match Limit

Note: For maxKits > 500, limiting the cross-match limit to 200 is advised, as cross-matching is CPU-intensive and when a set of triangulated segments for a chromosome is large (>200), the time to cross-match segments can grow very large.

Only triangulated segments > 7 cM considered for cross-matching.

Browse to the end of the page and locate the download link:

## GEDmatch Segment Triangulation -- (V0.3)

### Triangulation with Kit

All kits shown in columns Kit1 and Kit2 are taken from the closest 3000 matches to M020545 with a total matching segment count less than 3000 cM.

Matches above 3000 cM (total) are not shown.

3-Way (Triangulated) segment matches are shown in **green**. This is an indication of common ancestry.

Segments shown are larger than 7 cM and between 200 and 400 SNPs.

Triangulated Segments : 501 of 501

Click [HERE](#) to download triangulated segment data to a comma-separated CSV file.

Triangulated results sorted by Chromosome, Start Position:

Chr	Kit 1	Kit 2	Start	End	cM	
1						
1			776546	4488979	11.7	█
1			798959	3000924	7.3	█
1			798959	3000924	7.3	█

Download the triangulated data by clicking on the 'here' link in the bottom of the screen. This will allow you to save a **CSV** file to your local drive.

Next, go to the Genetic Affairs AutoSegment page for GEDmatch:

<https://members.geneticaffairs.com/autosegment/addWebsite/GEDmatch>

Hi ejblom,

Run an AutoSegment analysis for GEDmatch using the segment data and triangulated data (please note: this is available to Gedmatch Tier 1 subscribers only).

Gedmatch provides a downloadable file of all segments via their 'Segment Search' report. Please make sure to include enough matches, for instance 5000. In addition, enable the option "Prevent Hard Breaks"

GEDmatch also provides a [triangulated segments](#) which are used to verify identified overlapping segments.

Click [here](#) for a blog post from [Patsy Coleman](#) that describes her findings with AutoSegment and GEDmatch.

Start AutoSegment analysis with matches which share a max of	Stop AutoSegment analysis with matches which share less than	Min overlapping segment size	Min cluster size	Remove known pileups	AutoSegment name	Select triangulated file	Select segment file
250 c	15 cM	9 cM	2	<input type="checkbox"/>		<input type="button" value="Bestand kiezen"/> Geen bestand	<input type="button" value="Bestand kiezen"/> Geen bestand gekozen
<input type="button" value="PERFORM AUTOSEGMENT ANALYSIS"/>							

Adjust the search parameters, fill in the name field and select the triangulated and segment file. After you start the analysis, results will appear within 15 min in your mailbox. If no results appear, it is possible that too low cM values were employed. Try raising the minimum cM settings and retry the analysis.



## Hybrid AutoSegment – combine MyHeritage, FTDNA, 23andme and GEDmatch

The logical successor of AutoSegment would be a version that would create a clustering output using the segments of all four different companies. This hybrid AutoSegment is now available using the link <https://members.geneticaffairs.com/hybridautosegment> (see Figure 67).

Hi ejblom,

Run a hybrid AutoSegment analysis that combines segment data for MyHeritage, FamilyTreeDNA, 23andme and GEDmatch segment data.

Click [here](#) for a tutorial how to obtain the DNA match and segment data.

Name of analysis: hybrid\_clustering

Min overlapping segment size: 10 cM

Remove known pileups:

Perform liftover for FTDNA:

Min cluster size: 2

Company	Max shared	Min shared	Match file	Segment file
MyHeritage	250 cM	25 cM	<input type="checkbox"/> Bestand kiezen	<input type="checkbox"/> Bestand kiezen
FamilyTreeDNA	250 cM	35 cM	<input type="checkbox"/> Bestand kiezen	<input type="checkbox"/> Bestand kiezen
GEDmatch	250 cM	15 cM	<input type="checkbox"/> Bestand kiezen	<input type="checkbox"/> Bestand kiezen
23andme	250 cM	25 cM	<input type="checkbox"/> Bestand kiezen	<input type="checkbox"/> Bestand kiezen

Triangulated file:  Bestand kiezen

Match/segments file:  Bestand kiezen

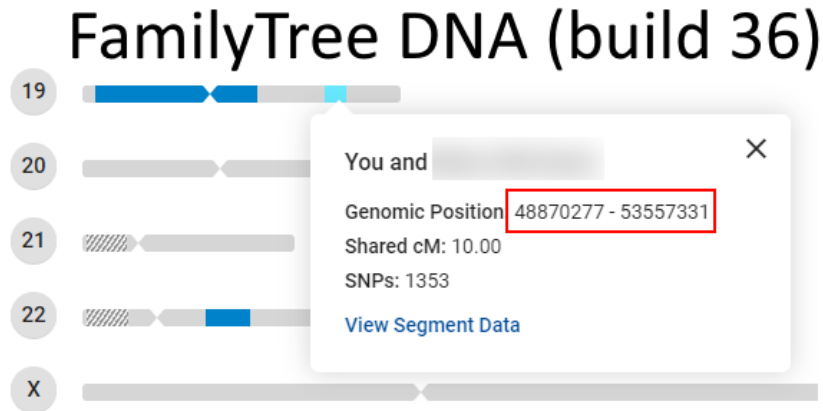
**PERFORM HYBRID AUTOSEGMENT ANALYSIS**

Figure 67. Hybrid AutoSegment interface

The interface options are quite similar to the default AutoSegment. However, the min overlapping segment size is now set for all four companies. In addition, there is an option to enable a liftover procedure for FTDNA. The reasoning for this option is as follows. Since this AutoSegment clustering relies on overlapping segments, it is crucial to obtain accurate genome coordinates. However, FTDNA employs human genome build 36 (NCBI36) to report its coordinates while the other companies (except Ancestry) report them using build 37. Using liftover tools, we can convert the genomic coordinates between different assemblies.

For my own data, I found an interesting segment on chromosome 19 (see Figure 68). In this case, the 10.0 cM segment on FTDNA is reported: start 48.870.277 - end 53.557.331

and on MyHeritage the 10.7 cM segment is reported: start 42.262.891 - end 48.995.691



### MyHeritage (build 37)

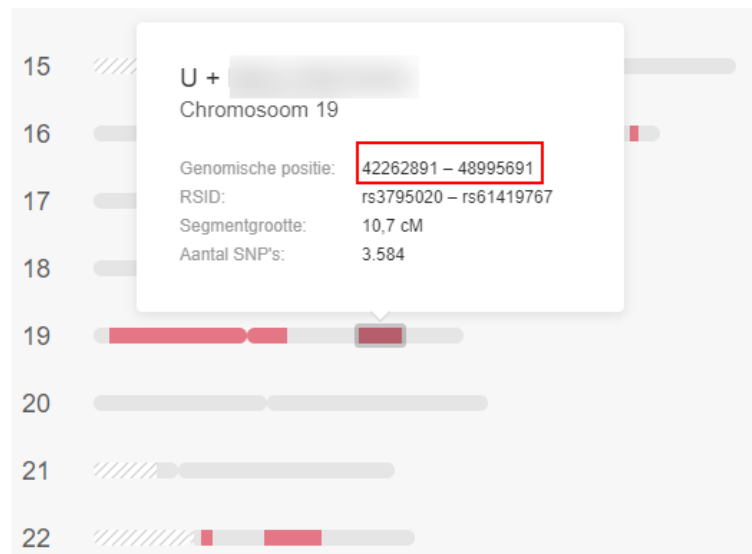


Figure 68. Difference between FTDNA build 36 and MyHeritage build 37

When comparing the segments between FTDNA and MyHeritage, it can be seen that they barely overlap. However, after applying the leftover method, the following coordinates are created:

before leftover start 48.870.277 - end 53.557.331

after leftover: start 44.178.437 - end 48.865.519

New segment length 8.9 cM

The updated coordinates overlap with the ones from MyHeritage. So with the updated coordinates, the segments will be overlapping whereas with the initial coordinates they were not.

The interface of the hybrid AutoSegment is similar to the regular AutoSegment. The overlay messages will show which matches and from which company they are (see Figure 69).

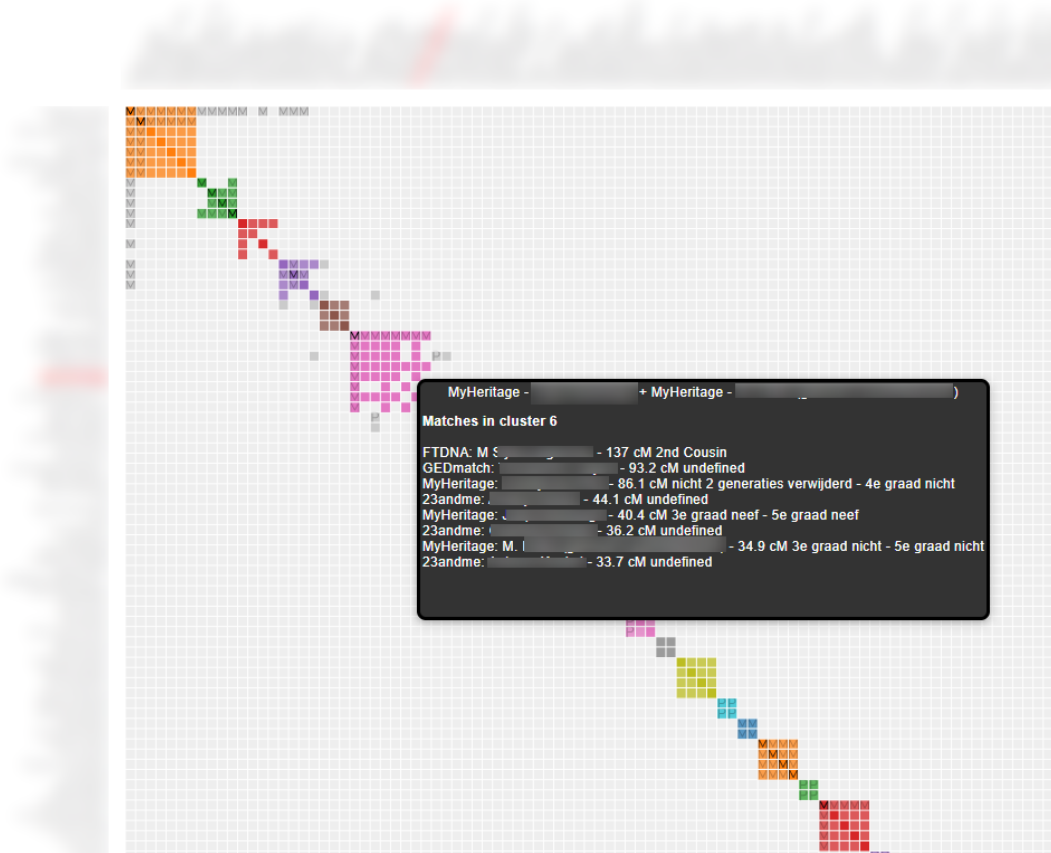


Figure 69. Hybrid AutoSegment chart with overlay

The table that contains some general statistics and links to the individual cluster reports is now also reporting the number of matches per DNA testing company (see Figure 70).

### Chromosome segment statistics per AutoSegment cluster

The following table shows the AutoSegment statistics per cluster. A link to each cluster is provided. In addition, the chromosomes linked to the segment clusters and the number of DNA matches for each cluster is shown. Last, the number of paternal/maternal DNA matches is available as well the number of segments and number of segment clusters.

[Download spreadsheet AutoSegment statistics](#)

Cluster	chr	#mat...	#F	#23	#MH	#G	Pat...	Mat...	Nr of se...	Nr of Se...
	<input type="text" value="filter colt"/>	<input type="text" value="filter colt"/>	<input type="text" value="filter"/>	<input type="text" value="filter"/>	<input type="text" value="filter"/>	<input type="text" value="filter"/>	<input type="text" value="filter cc"/>	<input type="text" value="filter cc"/>	<input type="text" value="filter column"/>	<input type="text" value="filter column"/>
▼ (44 items)										
<a href="#">AutoSegment cluster 1</a>	1,3,7,10,11	7	2	1	2	2		2	7	25
<a href="#">AutoSegment cluster 3</a>	3,12	4			4				3	7
<a href="#">AutoSegment cluster 4</a>	8,11,18	4	1	1	2			1	3	9
<a href="#">AutoSegment cluster 5</a>	11,18	3		1	2				2	6
<a href="#">AutoSegment cluster 6</a>	1,2,3,4,8,X	8	1	3	3	1		1	7	22

Figure 70. General cluster statistics

The same holds true for the table in the lower section of the main HTML page. This now contains the individual segment cluster information with information per DNA testing company.

**Individual segment Cluster Information**

The following table shows the 204 DNA segments for each of the 71 identified segment clusters.

[Download spreadsheet with segment clusters](#)

Cluster	Segmen...	C...	Start	Stop	Segment representation	S...	Name	DTC	cM	To...	Pa...	M...
<input type="text" value="Search"/>	<input type="text" value="Segment cl"/>	<input type="text" value="Sez"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>		<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Max"/>	<input type="text" value="Total"/>	<input type="text" value="filter"/>	<input type="text" value="filter"/>
▼ 11 (3 items)												
<a href="#">38</a>	11	5	1	5050830		1329		23andme	12.5	40		
<a href="#">37</a>	11	5	14782	4841505		3840		MyHeritage	12.2	91.5		
<a href="#">39</a>	11	5	81437	5176672		493		GEDmatch	15.9	26.6		
▼ 50 (2 items)												
-	50	10	100076444	116653136		8704		MyHeritage	15.9	30.8		
<a href="#">1</a>	50	10	92994805	117029683		6600		FTDNA	24	351		M
▼ 60 (2 items)												
<a href="#">1</a>	60	7	100159726	123897167		4700		FTDNA	18.3	351		M
-	60	7	97601052	129663496		1561		GEDmatch	24.3	23.3		
▼ 24 (5 items)												
<a href="#">23</a>	24	20	10752610	52308169		21504		MyHeritage	51	51		
<a href="#">23</a>	24	20	17599087	44076324		11648		MyHeritage	24.5	32.1		
<a href="#">23</a>	24	20	17770834	43374640		5400		FTDNA	23.6	53		M
<a href="#">23</a>	24	20	17771604	44076324		11520		MyHeritage	24.2	30.2		
<a href="#">23</a>	24	20	35335891	52308169		10112		MyHeritage	26.7	33.6		

Figure 71. Hybrid AutoSegment segment cluster chart

Information concerning the DNA match and segment files is provided in the main HTML file (see Figure 72). Also information is provided if there are any warnings concerning issues while importing these files. The results of the pile-up removal procedure (if selected) are also provided. The same for the FTDNA liftover analysis, with information concerning the largest change.

For GEDmatch some additional information is provided for the triangulation file. In some cases these triangulation reports generate data that are not provided in the segment file. We still would like to keep these segments since they represent high quality triangulated data.

To predict the cM values of these unknown matches, we first cluster (flatten) the triangulated segments such that we obtain a single segment cluster per segment location. The reason this is important because the triangulation file can contain many segments that are more or less on the same location. After clustering these matches we take the largest segment from the segment cluster and combine the cM values for each them. The combined cM values for these flattened cM values are then the total shared cM with this particular DNA match. It's most likely a deflated score but at least we use the available data to come to a certain value. These reconstructed matches can easily be detected by looking at the notes field of GEDmach matches. That will contain information concerning the reconstruction of this DNA match.

## Settings used for this AutoSegment analysis

A **MyHeritage** analysis was performed for "aaa". From the supplied CSV file(null), a total number of 6508 matches were retrieved. After applying the cM settings (min 30.0 cM, max 600.0 cM), we removed a total of 6391 matches and continued the clustering analysis using 117 DNA matches. A total of 14 rules (14 rules that exclude segments, excluding thereby 31 segments) were employed.

A **FTDNA** analysis was performed for "aaa". From the supplied CSV file(s), a total number of 1943 matches were retrieved. After applying the cM settings (min 45.0 cM, max 600.0 cM), we removed a total of 1536 matches and continued the clustering analysis using 407 DNA matches. After importing the segment file, 4 warnings were obtained: Warning, for match Robert George White we already imported some segments, the match might be duplicated within the segment file? Warning, for match Claus Behrendt Møller we already imported some segments, the match might be duplicated within the segment file? Warning, for match John Wright Ballard we already imported some segments, the match might be duplicated within the segment file? Warning, for match Nicholas de Groot we already imported some segments, the match might be duplicated within the segment file? A total of 14 rules (14 rules that exclude segments, excluding thereby 7 segments) were employed.

### *FamilyTreeDNA leftover procedure*

The provided segments by the DNA testing companies are all based on a certain [human reference genome](#). Since the hybrid AutoSegment tool identifies putative triangulating segments based on DNA segment positions, it is important to have the correct coordinates. FamilyTreeDNA employs the human genome build 36 whereas the other companies all support build 37. Most FamilyTreeDNA segments will have similar coordinates for both builds but in some cases, there can be a large difference (especially for chromosome 19). There are [lifter](#) tools that can convert the coordinates between different reference genomes. For 124 segments, we changed the start/end positions for total number of 208211196 basepairs. The largest segment change was found for: Daryl G. May Chr19:46394266-54350202 (15.6 cM) after lifter: 41702426:49658390 (12.8 cM)

A **23andme** analysis was performed for "aaa". From the supplied CSV file(s), a total number of 1334 matches were retrieved. After applying the cM settings (min **30.0** cM, max **600.0** cM), we removed a total of 1309 matches and continued the clustering analysis using 25 DNA matches. During importing the segments from the CSV file, we couldn't find segment data for 301 DNA matches (probably due to their DNA sharing settings). A total of 14 rules (14 rules that exclude segments, excluding thereby 5 segments) were employed.

A **GEDmatch** analysis was performed for "aaa". From the supplied flat segment CSV file(s), a total number of 3889 matches were retrieved. After applying the cM settings (min 20.0 cM, max 600.0 cM), we removed a total of 3852 matches and continued using 37 DNA matches.

A total number of 7868 triangulated segments were obtained from the GEDmatch triangulation report. Of these segments, a total number of 48 segments were linked to an existing segment that was retrieved from the previously imported segment file. A total number of 6944 triangulated segments were not linked because the minimum segment size criterium (10cM) was not met. In addition, a total of 871 triangulated segments could not be linked to a DNA match from the existing DNA match list. However, based on the kit numbers, we were able to identify 81 previously unknown DNA matches using 871 segments. The unlinked triangulated DNA segments were converted to DNA matches after which the triangulating segments will be used to assess their total shared cM.

A total of 871 unlinked segments that triangulate with another segment were retrieved for 81 unknown DNA matches. These triangulating DNA segments often overlap with other triangulation segments and should therefore be condensed. After performing this step, a total of 47 flattened segments were obtained. These flattened segments are better representatives of the actual segments shared by the kit owner and can also be used to assess the total amount of shared cM. Due to the cM settings (600.0-20.0 cM), a total of 80 reconstructed DNA matches based on triangulated matches were discarded, as were 0 triangulating segments because they did not meet the minimum segment size of 10 cM. A total of 1 DNA matches were reconstructed based on the triangulated segments and added to the existing set of 37 DNA matches. A total of 14 rules (14 rules that exclude segments, excluding thereby 4 segments) were employed.

Figure 72. Hybrid AutoSegment settings

In addition to the general statistics we also supply information concerning the DNA segments. This will include the number of segments imported, after filtering and the number of overlapping segments (see Figure 73).

#### DNA Segment statistics

Based on the 587 DNA matches we were able to retrieve 317 DNA segments (after filtering using a minimum segment size of 10cM, a total of 317 segments remained). Interestingly, we weren't able to retrieve segment data from the provided segment data file for the following 4 matches: [REDACTED], [REDACTED], [REDACTED], [REDACTED]. A total number of 7 DNA segments were imported that originated from the X chromosome (chromosome 23). A total of 646 segment overlapping segment pairs were found (using the min cM overlap of 10 cM) whereas 99192 segment combination did not overlap and were therefore not used for the clustering analysis. Based on the 646 overlapping segment combinations a total of 71 segment clusters (containing 204 segments) resulted from the segment clustering and are linked to 152 DNA matches (59 **FTDNA** matches, 65 **MyHeritage** matches, 16 **23andme** matches, 12 **GEDmatch** matches, ).

#### DNA match cluster analysis using overlapping segments.

As was mentioned in the previous sections, a clustering of overlapping segments was performed that resulted in segment clusters. This clustering links individual segments, so certain high cM DNA matches that share multiple DNA segments might be linked to multiple DNA segment clusters. To link the DNA matches, we do another clustering using their segment cluster membership. So 2 DNA matches can be linked if they are a member of the segment cluster. This clustering tries to mimic a regular clustering analyses that is performed using shared matches. Using this methodology, we were able to cluster 145 (56 **FTDNA** matches, 63 **MyHeritage** matches, 16 **23andme** matches, 10 **GEDmatch** matches, ) DNA matches. Note that you can still obtain valueable information for the missing matches by examining the segment clusters (a summary is available in a table on this page). A summary concerning the discarded cluster members is provided underneath:

Note: The following 7 matches met the inclusion parameters but were placed in an cluster that contains an amount of matches that is lower than the min cluster size of 2 matches and are therefore not included in the chart: [REDACTED]

Figure 73. Hybrid AutoSegment segment statistics.



## Additional settings

More additional settings and links are available from the top-menu (see Figure 74). The different settings and options will be discussed below.

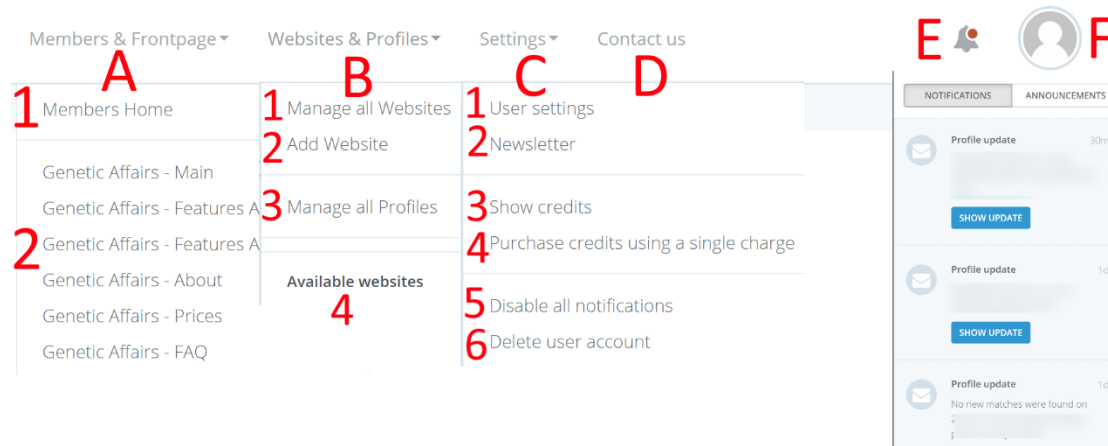


Figure 74. General settings from top-menu

### A) Members & Frontpage

- 1) Members Home
- 2) Genetic Affairs frontpage links

### B) Websites & Profiles

- 1) Shows all websites
- 2) Allows adding a new website
- 3) Manage all profiles. This view shows all profiles for all websites. It can also be used to modify the settings for all profiles.
- 4) This section will display all registered websites

### C) Settings

- 1) User settings – allows the modification of user settings with respect to user credentials, subscriptions and payment methods. The next section will go into more detail
- 2) Subscribing or unsubscribing to the newsletter can be performed using this link
- 3) This page shows all available credits
- 4) When a valid credit card has been registered, it is possible to use this page to obtain credits based on a single charge
- 5) Allows disabling all regular updates with one action
- 6) Delete user account

### D) Contact us option

### E) Notifications pane and general announcements. The notifications panel will display account info updates, DNA match updates, information about single or monthly subscriptions that will be available as notifications. The general announcements will be used to show more general messages like new functionalities or promotional information.

### F) User settings allow changing information concerning your name, e-mail address, password, subscriptions, payment information, and invoices. Our next section will discuss this in more detail.

## User settings and Payments

The user settings are available from the top-menu after selecting the user icon (most right option) and clicking on the “Your Settings” link (see Figure 74C). The different settings and options will be discussed below.

The screenshot shows a user settings page. On the left, there are two main sections: 'Settings' and 'Billing'. Under 'Settings', there are 'Profile' (marked with a red 'A') and 'Security' (marked with a red 'B'). Under 'Billing', there are 'Subscription' (marked with a red 'C'), 'Payment Method' (marked with a red 'D'), and 'Invoices' (marked with a red 'E'). On the right, the 'Contact Information' section contains input fields for 'Name' and 'E-Mail Address', and an 'UPDATE' button.

Figure 75. User settings

- A) Profile settings – allows changing the name and e-mail address
- B) Allows changing the password
- C) Shows the different monthly subscription options (see Figure 76). The monthly subscriptions allow for a monthly addition of credits to your account. In addition, monthly subscription yields 10% bonus credits on top of the acquired credits.
- D) Store your credit card settings in this section. We use the Stripe ([www.stripe.com](http://www.stripe.com)) payment processing platform to store credit card information. In addition, the Stripe platform also performs monthly or single payments. Note that the **VAT** field only is required for companies, most users can ignore this option.
- E) Invoices from the monthly subscriptions can be downloaded from this invoices page.

The screenshot shows a subscription options page. At the top, there is a 'Free Trial' section with a yellow background, stating 'You are currently within your free trial period. Your trial will expire on October 15th, 2018.' Below this is the 'Subscribe' section, which says 'You are not subscribed to a plan. Choose from the plans below to get started.' and 'All subscription plan prices are excluding applicable VAT.' The main content is a table of subscription plans, all labeled 'Pro', with prices ranging from \$5.00 to \$15.00 per month. Each row includes a 'PLAN FEATURES' button, the price, and a 'SELECT' button.

Figure 76 - Subscription options



## Blog posts and Facebook groups

After the launch of AutoCluster a new user group has been founded, please [visit us on Facebook](#).

In addition, several (often AutoCluster related) blog posts have been created in the last year. These describe the use of AutoCluster and sometimes illustrate this tool with genealogical evidence.

[Comparison of ICW AutoCluster and AutoSegment AutoCluster by Patricia Coleman](#)

[Genetic Affairs Hybrid AutoSegment Cluster by Patricia Coleman](#)

[GEDmatch AutoSegment by Patricia Coleman](#)

[Genetic Affairs AutoFastClusters by Patricia Coleman](#)

[Manual AutoClusters for LivingDNA by Patricia Coleman](#)

[DNAeXplained - Genetic Affairs: AutoPedigree Combines AutoTree with WATO to Identify Your Potential Tree Locations](#)

[Kitty Cooper - Automated tree building with Genetic Affairs](#)

[DNAeXplained – Genetic Genealogy - Genetic Affairs Reconstructs Trees from Genetic Clusters – Even Without Your Tree or Common Ancestors](#)

[Dana Leeds Blog](#)

[Kitty Cooper's Blog - Automatic Clustering from Genetic Affairs](#)

[Kitty Cooper's blog - More Clustering Tools!](#)

[DNAeXplained – AutoClustering by Genetic Affairs](#)

[Hartley DNA & genealogy - A New Look for AutoClusters](#)

[Behold Genealogy - Genetic Affairs Clustering at 23andMe](#)

[Anne's Family History - DNA: experimenting with reports from GeneticAffairs.com](#)

[DNA sleuth - Clustering Tools for DNA matches](#)

[Genea Musings - Using GeneticAffairs.com to Create DNA Match AutoClusters](#)

[HistorTree - Analyzing DNA Auto-Clusters with Pedigree Collapse](#)

[MyHeritage DNA - Introducing AutoClusters for DNA Matches](#)

[The Genealogy Guys Blog - Genetic Affairs, a New DNA Tool](#)

[Matt's Genealogy Blog - Auto-Clustering of DNA Matches](#)

Genetic Affairs: Clustering DNA Matches - Part 1 (2020)

With Founder EJ Blom

Part 1

Genealogy TV  
Watch later Share

# Genetic Affairs

## Clustering DNA Matches

Forming and Using Genetic Networks with Genetic Affairs

Goal = compare trees

Theory = all members of a cluster have a common ancestor

Family Tree LIVE

20TH & 27TH APRIL 2019  
ALEXANDRIA PALACE

Come & hear Donna Rutherford's talk 'Making the most of your autosomal DNA test'

Family Tree Live speaks

Making the most of your autosomal DNA test

Donna Rutherford  
April 2019

# Cluster Tool

A Segment of DNA

## Other AutoCluster implementations.

After the initial launch of AutoCluster in December 2018, a lot of requests were made concerning a version for MyHeritage. Fortunately, MyHeritage decided to license the tool and together with their team, we implemented the AutoCluster in their infrastructure. It was released during Rootstech 2019 in Salt Lake City. The visualization of AutoCluster from MyHeritage has been changed a bit, for instance, the usage of different cluster colors (see Figure 77). In addition, an algorithm was implemented that uses the results of various clustering analyses to find a clustering that produces a chart containing around 100 members. This ensures that the user experience will be more constant. Moreover, this clustering analysis also considers the shared cM between common matches. This implementation has been shown to improve the clustering results of people from [endogamous](#) populations (for instance, Ashkenazi or Acadian).

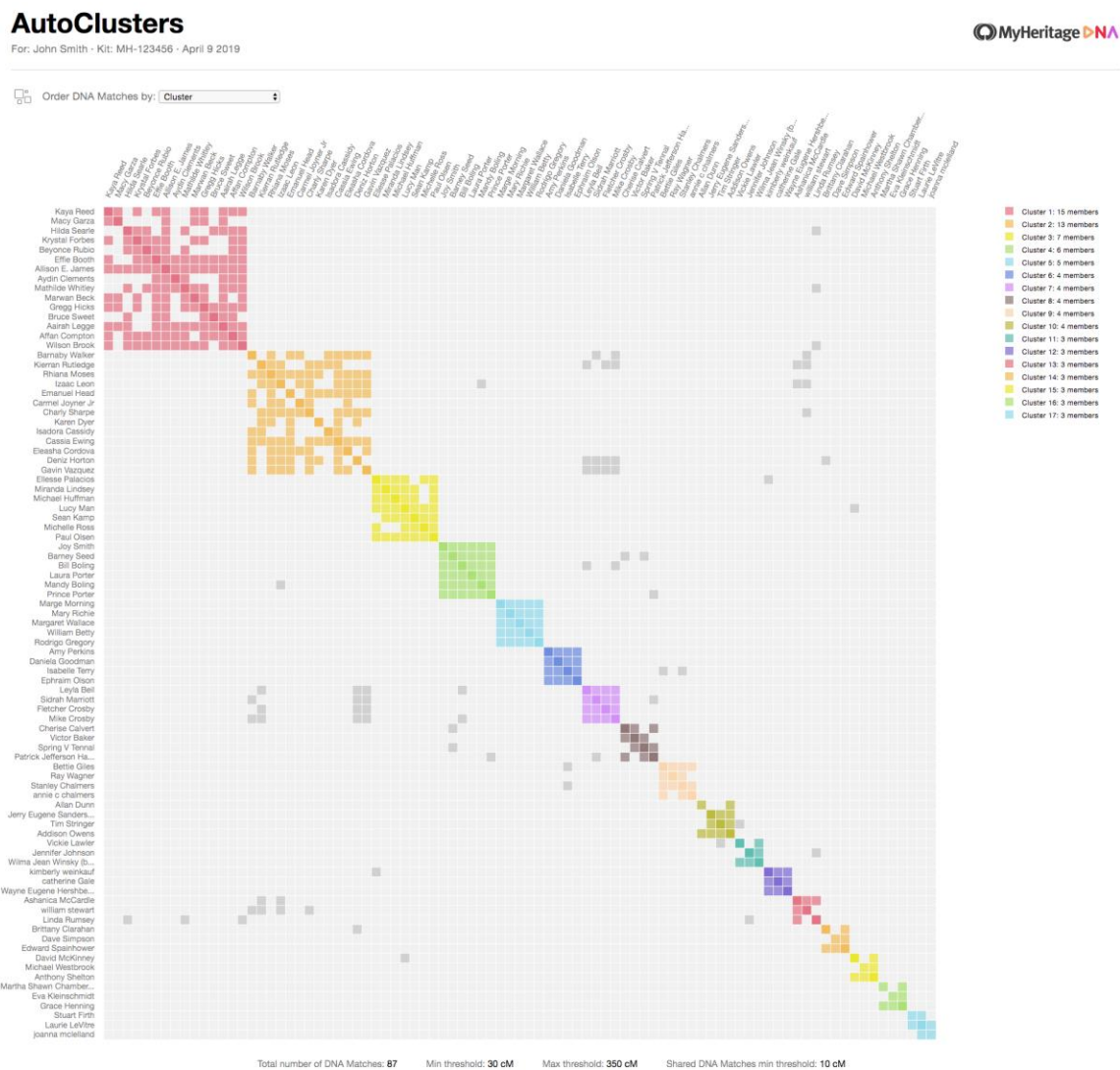


Figure 77. AutoCluster implementation from MyHeritage. Note that the current implementation does not contain the sorting of clusters based on the grey cells, instead, they are sorted based on cluster size.

Another website that now employs the AutoClustering algorithm and visualization is GEDmatch (see Figure 78). This analysis is available for Tier 1 users. GEDmatch has gone a step further by creating an interactive AutoCluster analysis by combining the Multi Kit Analysis with the AutoCluster clustering. As a result, members of a cluster can now be selected for further analysis, for instance, to see which DNA segments are shared.

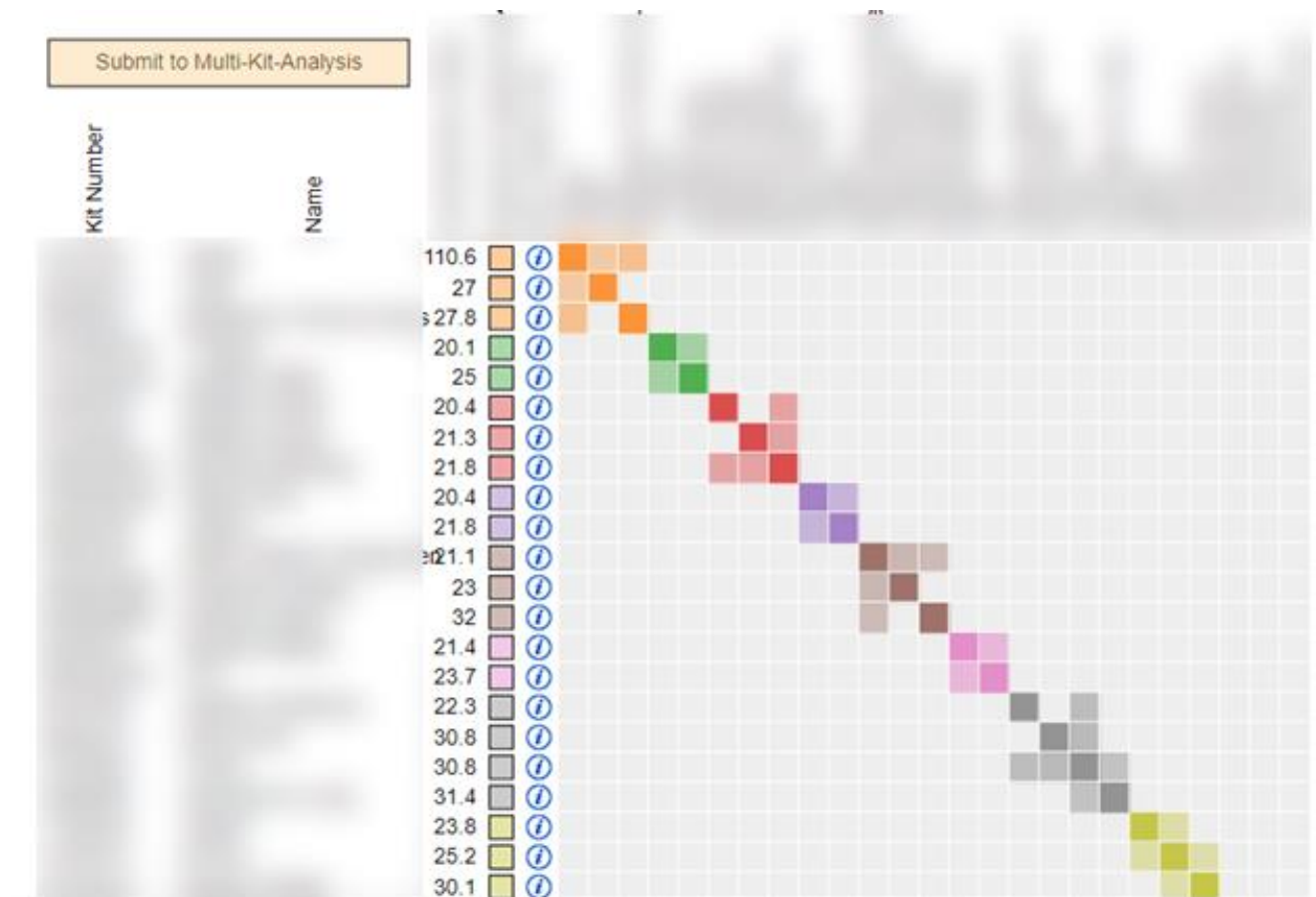


Figure 78. GEDmatch implementation of the AutoCluster algorithm with selectable clusters and members for downstream analysis by the MKA.

## Prices

The free trial provides a total of 200 credits that can be used to try several features (see Figure 79). By purchasing a subscription on our site, several premium features are unlocked as well as the removal of some restrictions.

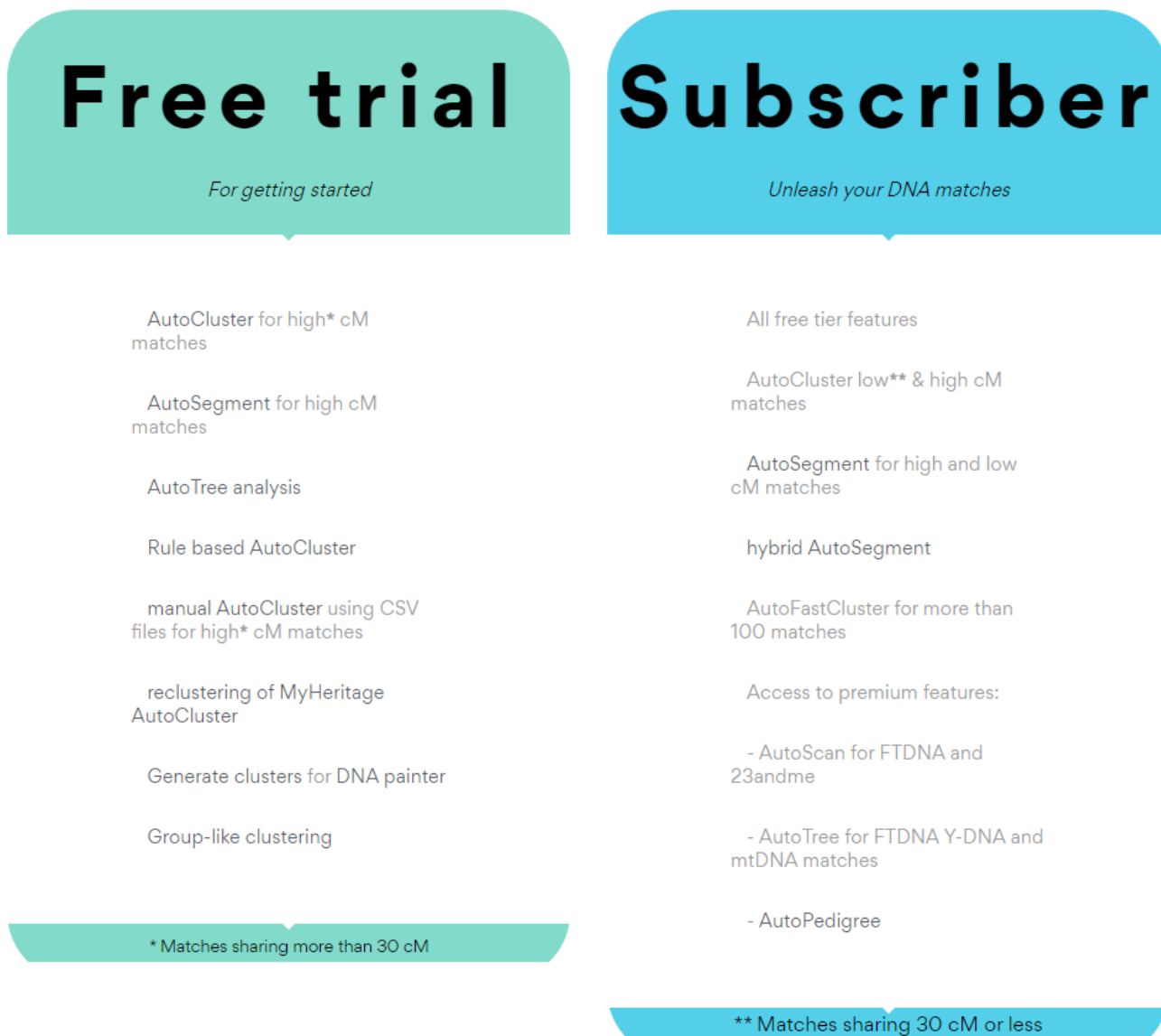


Figure 79. Free trial vs subscriber membership

The front page holds information concerning the costs of the analyses (see <https://www.geneticaffairs.com/prices.html> and Figure 80).

### What do analyses on Genetic Affairs cost?

- Reclustering, CSV or AutoFastCluster analyses are **50** credits.
- A default AutoCluster analysis using shared matches for 23andme or FTDNA is **75** credits per search.
- An AutoSegment analysis costs **75** credits per search.
- Rule based analyses\* are **50** credits per applied rule and **50** credits for the primary profile
- AutoCluster analyses with the **AutoTree** or **AutoPedigree\*** features are **100** credits.
- Hybrid AutoSegment clustering costs **100** credits for 2 datasets\*.
- Hybrid AutoSegment clustering costs **125** credits for 3 datasets\*.
- Hybrid AutoSegment clustering costs **150** credits for 4 datasets\*.

\* premium feature

*Figure 80. Cost of the analyses*

## Troubleshooting

Sometimes the websites (FTDNA or 23andme) are unreachable, for instance, if they are under maintenance. In this case, a message concerning this error will be mentioned in the regular mail. In addition, in the case of a weekly or monthly search, the delay is saved. When the upcoming search is successful, we use this delay to correct the weekly or monthly interval (which would otherwise shift by one or more days).

In the case of endogamous populations, the AutoCluster might not work properly. This is partly due to the large number of matches which a long time to download. The limited-time slot of a single AutoCluster analysis does not provide for these long downloads. In some cases, selecting the more powerful server will help. If it still doesn't work, alternative methods such as [DNAGedcom](#) might be a good option since they also perform clustering of the matches based on shared matches. The difference, however, is that this tool can run on your own computer and doesn't suffer from the time constraints that AutoCluster has.

See our [frequently asked questions](#) for more information concerning security.