UPTEC X 09 017
Examensarbete 30 hp
April 2009

Development of software tools for the design and evaluation of MLGA assays

Patrik Eriksson

UPPSALA UNIVERSITET

Bioinformatics Engineering Program

Uppsala University School of Engineering

UPTEC X 09 017	**Date of issue 2009-05**

Author

Patrik Eriksson

Title (English)

Development of software tools for the design and evaluation of MLGA assays

Title (Swedish)

Abstract

Multiplex ligation dependent genome amplification (MLGA) is a novel method for the multiplex targeted measurement and validation of genetic copy-number variations (CNVs). A design software for MLGA assays and a simple analyse software have been developed. Three design sets were created during the development of the software for evaluation purposes.

Keywords

MLGA, In silico digestion, GC content, Folding energy, polymorphism

Supervisors

Simon Fredriksson
Olink Bioscience

Scientific reviewer

Patrik Forssén
SweCrown AB

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **40**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Development of software tools for the design and evaluation of MLGA assays

Patrik Eriksson

Sammanfattning

Multiplex ligeringsberoende genomamplifiering (MLGA) är en metod utvecklad på Rudbecklaboratoriet för att under en och samma mätning kunna fastställa om flera olika DNA-sekvenser existerar eller finns i flera kopior i ett genom. Metoden använder sig av delvis dubbelsträngat DNA som binder till specifika delar av genomet, som är av intresse, för att selektivt kopiera upp delar av genomet i mängder som är mätbara med tillgängliga tekniker. Tyvärr har en del DNA fragment egenskaper som gör att de inte är lämpliga att använda tillsammans med MLGA metoden. Några av dessa egenskaper är GC-innehåll, sekundärstruktur, polymorfism och sekvenslängd.

Detta examensarbete beskriver utvecklingen av en mjukvara där egenskaper hos DNA fragment snabbt kan utvärdera för att sortera bort olämpliga fragment. I den slutgilltiga versionen av mjukvaran kan DNA fragment utvärderas med avseende på GC-innehåll, sekundärstruktur, polymorfism och sekvenslängd. Information angående utvalda fragment kan efter utvärdering sparas som PDF-filer eller projekt-filer för fortsatt analys.

Examensarbete 30 hp
Civilingenjörsprogrammet Bioinformatik
Uppsala universitet april 2009

Contents

1 Introduction

Multiplex ligation dependent genome amplification (MLGA) is a novel method for the multiplex targeted measurement and validation of genetic copy-number variations (CNVs). It is a technique that can be used to measure the differences between genomes belonging to healthy and diseased individuals, for example the detection of people with Down syndrome[1] or Turner syndrome[2]. Another application area could be to find people with higher chance of developing diseases, because of specific genetic abnormalities. MLGA can also be used to study submicroscopic variations in genomes, variations that contribute a lot to genetic diversity (Isaksson et al., 2007, p. e117).

Today, there exists software that can and has been used to design selector-probes, which are used in MLGA assays. The current programs[3] have not been constructed specifically for the development of selector probes, which can create some difficulties during the design. These programs contain a lot of functions and options that are not needed, which make it harder to learn how to use these programs to design selector probes. These software are not collected in one package, some of those you can install on your computer and other parts have to be used through the Internet. This makes design dependent on Internet connection and the providers' servers. Somehow the information also has to be transformed and transmitted between the different programs. Using the Internet or network connected programs research information is subjected to the chance of being exposed.

The goal with this project is to design one software package that contains all functions that are needed to successfully create selector probes used in an MLGA assay.

[1] Down syndrome: a person with three 21 chromosomes (aka trisomy 21).

[2] Turner: a disease where a women only has one X chromosome.

[3] PieceMaker, ProbeMaker, mfold etc.

1.1 Project goal

The primary goal within this project is the development and implementation of a software tool for rapid and simple design of selector probes to be used in MLGA assays. The software should have these features:

- Allow for the import of target DNA sequences both from text files and from public databases such as Ensemble or GenBank, specified via gene name and/or accession number.
- The software needs to perform *in silico* digestion of chosen DNA targets with a variety of restriction endonucleases and combine these fragments into MLGA sets with sufficiently large size differences.
- The user should have the option to specify important design criteria such as choice of restriction enzyme, minimal and maximal fragment length, GC content etc.
- The software needs to generate suitable selector probes for the respective targets, ensuring compatible (similar) melting temperatures of the probe ends while avoiding excessive secondary structures and homologies to other probes and non-targeted restriction fragments.

A secondary goal within this project, which may be implemented either as a separate software tool or as part of the software described above, is to automate the analysis of MLGA data. For this purpose, electropherograms generated on standard laboratory equipment from different vendors (i.e. Agilent Bioanalyzer 2100, Bio-Rad Experion, Shimadzu MultiNA, QIAGEN QIAxcel) need to be imported and analyzed according to fragment size (separation time) and signal strength (peak height; amount of fluorescence). Fragments must be matched to their respective target sequences and compared to a normal reference sample in order to calculate ratios representing the change in copy number compared to the reference.

2 Background

2.1 Polymerase Chain Reaction

Polymerase chain reaction (PCR) is an important and powerful technique that was introduced in 1987. It is used to amplify small amount of DNA to enable further analysis. Advantages with PCR are that it only needs very small amounts of DNA to start with (at least one sequence) and it does not need living cells. A limitation is, to be able to perform PCR, part of the sequence has to be known(Mathews et al., 2000, p. 922-993). An additional limit is that the success rate of the amplification procedure is dependent on the GC/AT ratio of the sequence, regions with high or lower GC can be hard to amplify(Benita et al., 2003, p. e99).

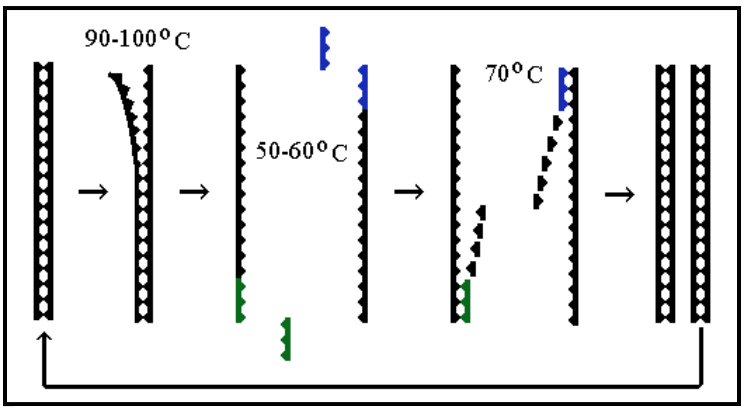


Figure 1: A PCR cycle where a double stranded DNA is used as starting material. Blue and green parts are primers and primer binding sites.

To perform PCR, a DNA sample, Taq polymerase[4], a pair of oligonucleotide primers and a PCR machine are needed. The Primers are used by Taq polymerase to start the DNA synthesis. Two primers are needed, because there are two DNA strands and each primer base-pair to one DNA strand. The PCR machine is used to change the temperature between the different steps and cycles in a PCR reaction. A PCR cycle is started by increasing the temperature of the mixture to 90-100 °C, which will break the hydrogen-bonds between the strands creating single stranded DNA. The temperature is then lowered to 50-60 °C, which make it possible for the primers to bind to the single stranded DNA. The temperature is then increased to 70 °C, which is the optimal temperature for Taq polymerase. This is repeated many times to create more products, which will act as starting material in the next cycle(Brown, 2002, p. 119-122).

[4] A thermostable DNA polymerase.

2.2 Detection systems

One of the most common techniques for detection and separation of DNA fragments (and other molecules) with different lengths, is gel electrophoresis. The technique uses an electric field in combination with a gel that consists of a network of pores. The electric field is used to migrate DNA fragments. DNA fragments will move as they are charged and therefore are drawn to the opposite charge. The agarose gel is used to separate fragments with help of pores found in the gel that will make it harder for longer DNA fragments to move (Brown, 2002, p. 37).

Today more advanced techniques for detection and separation. One such technique is the Experion automated electrophoresis system provided by Bio-Rad. The Experion system uses gel-based electrophoresis in combination with Caliper Life Science LabChip microfluidic technology to deliver size measurement and quantitation(Bio-RadLaboratories, 27 Mars 2009).

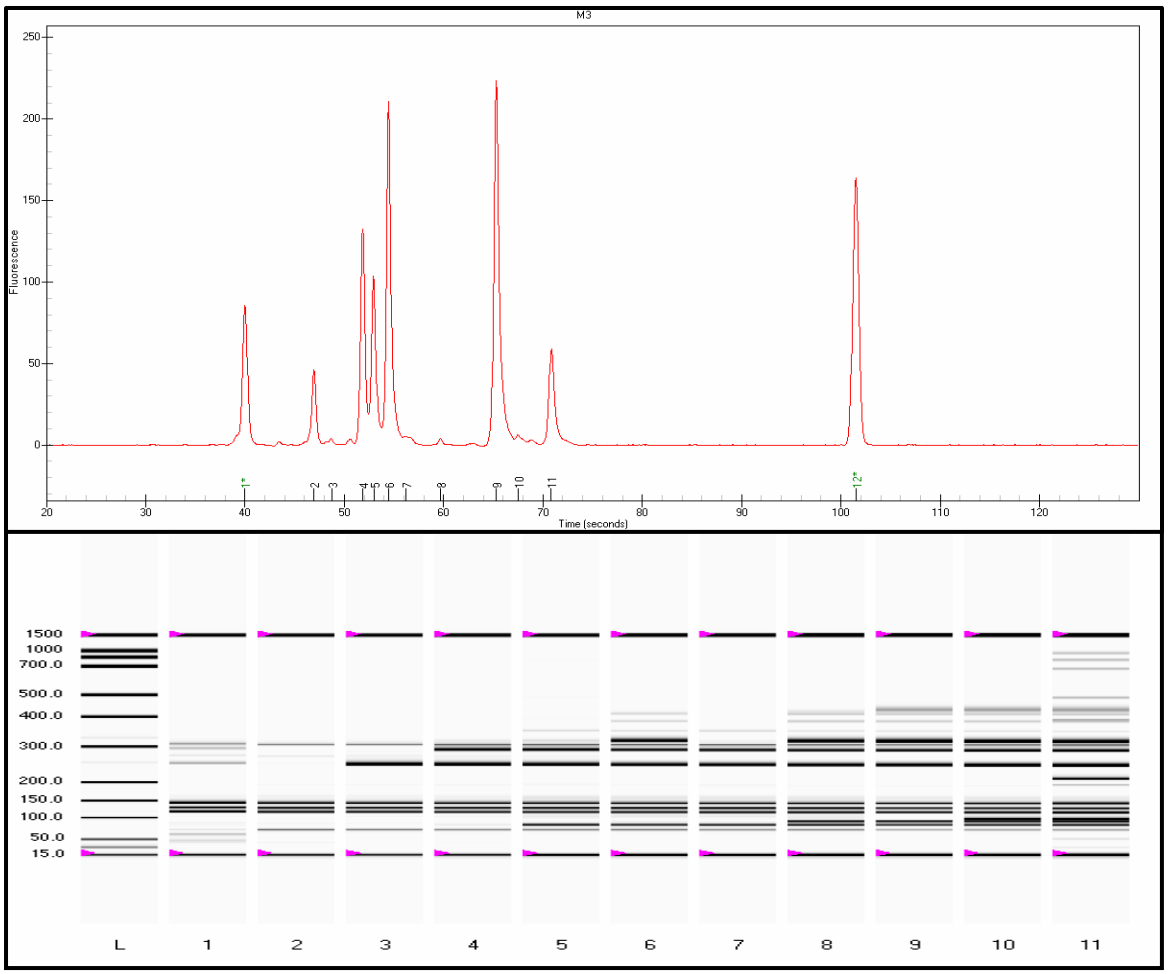


Figure 1: Two images representing output information from an ordinary electrophoresis gel (lower image) and output from an Experion system (upper image). The Experion result will be paired with a table (table 1, next page) containing information about the result.

Peak Num.	Mig. Time(secs)	Size (bp)	Correct. Area	Area ratio	Area	FWHM	Peak Height
2	46.93	72	53.63	0.4579	25.97	0.46	46.42
3	48.72	87	3.51	0.0300	1.76	0.43	4.01
4	51.87	116	128.90	1.1005	68.78	0.44	133.31
5	52.97	127	109.59	0.9356	59.69	0.48	103.85
6	54.44	142	216.59	1.8491	121.14	0.45	211.05
7	56.18	159	9.52	0.0813	5.49	1.05	5.41
8	59.69	194	3.17	0.0271	1.94	0.42	4.11
9	65.31	251	225.51	1.9253	150.65	0.53	223.99
10	67.52	273	7.50	0.0640	5.17	0.92	6.15
11	70.83	308	60.13	0.5134	43.49	0.57	58.87

Table 1: Result from the Experion system.

2.3 Selector probe

A selector probe is a partly double-stranded sequence, constructed in a way that it forms a circularized structure with one or more specific regions of the genome. It consists of two parts, target specific ends that are joined by a double stranded oligonucleotide containing a general primer-pair motif[5] and an enzyme binding site(Dahl et al., 2005, p. e71).

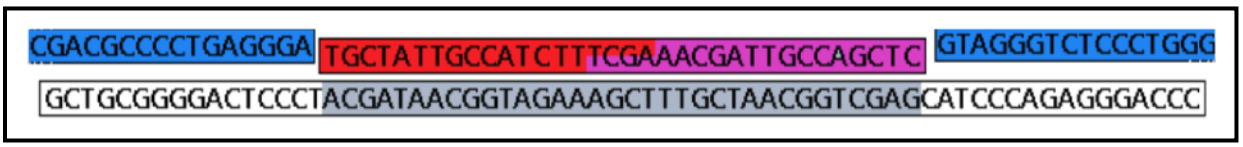


Figure 2: A representation of a selector probe, blue represent a DNA sequence, white represent the target specific ends, red represent the binding region for the reverse primer and the purple region represent the binding site belonging to the forward primer.

The target specific ends of the selector probe are constructed to be complementary sequences to the ends belonging to the DNA sequence of interest. They are joined by the double stranded oligonucleotide(Dahl et al., 2005, p. e71).

The general primer motif is incorporated in the double stranded oligonucleotide to provide binding sites for the forward and reverse primers that are used in the PCR reaction. An enzyme binding site can also be found in the oligonucleotide, the enzyme binding site is placed in a way that the enzyme does not destroy the binding site belonging to the forward and reverse primer. The enzyme is used to transform the circularized structure to a linear strand just before the amplification (Dahl et al., 2005, p. e71).

[5] A nucleotide pattern.

2.4 Multiplex ligation dependent genome amplification – MLGA

Multiplex ligation dependent genome amplification (MLGA) is a novel method that was developed by the Department of Genetics and Pathology at Rudbeck Laboratory. The method uses selector probes to amplify multiple parts of the genome. Each selector probe is designed to bind to specific DNA sequences that are created using a restriction enzyme. The selector probe and sequence forms a circularized construction, which can be selectively amplified. The amplification product can then be analyzed using an ordinary electrophoresis gel or a more advanced system like the Experion system(Isaksson et al., 2007, p. e115).

The MLGA procedure uses several selector probes each time, to investigate several loci. It consists of four different steps, restriction digestion reaction, ligation reaction, exonuclease reaction and multiplex PCR.

2.4.1 Restriction digestion reaction

A big advantage with MLGA is that specific parts of the genome can be selected for amplification, leaving the rest of the genome untouched. To be able to do this the genome has to be digested into known parts, which selector probes can be designed against. This can be done using a restriction endonuclease, which is an enzyme that cut a DNA strand at specific recognition sites[6]. There are three types of restriction endonucleases, types I, II and III. Type I and III are less reliable, as they cut at no fixed position relative to the recognition site, producing fragments with ends that are unknown. Type II, on the other hand, cut at the same place each time, either inside the recognition sequence or near to it(Brown, 2002, p. 102).

Depending on the choice of enzyme different compositions of DNA pieces will be produced, which means one cannot change the enzyme in the restriction reaction without changing the design of the selector probes.

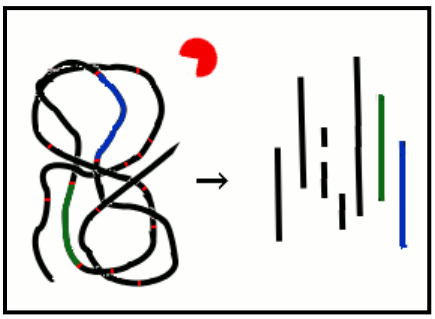


Figure 3: The left part represent unmodified DNA, where the red parts are restriction sites. The red figure is an enzyme and the green and blue parts are the part of interest. The right part represents the restriction reaction product.

[6] A specific combination of nucleotides.

2.4.2 Ligation reaction

During the ligation reaction the selector probes bind to the sequences of interest and forms circularized structures. The restriction fragments not used to design selector probes will stay unmodified.

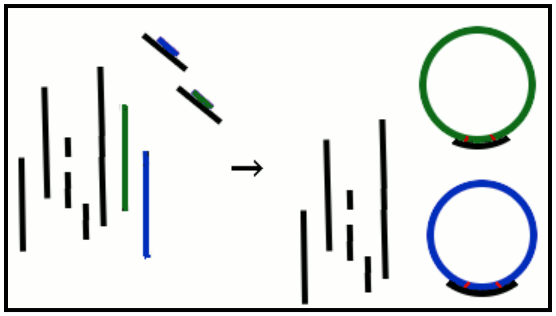


Figure 4: The left part contains a representation of restriction fragments, where blue and green lines represent the sequences of interest. The green/blue and black double lines represent selector probes.
The right part represents the ligation reaction product.

2.4.3 Exonuclease reaction

The exonuclease enzyme is introduced to digest all non-circularized pieces of DNA to much smaller pieces of DNA. The reason for this is to reduce the likelihood of other pieces of DNA to be amplified, besides the circularized sequence. If other pieces of DNA were to be amplified they would show up at the analysis step, which would make the analysis more difficult or impossible.

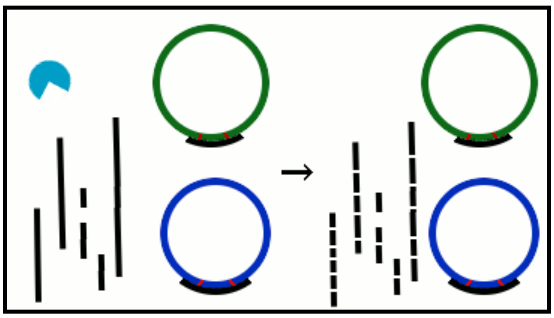


Figure 5: The left part of the image represents the unmodified restriction fragments and the ligation product. The right part represents the exonuclease product.

2.4.4 Multiplex PCR

During multiplex PCR the circularized sequences are first cut at the enzyme binding site that was incorporated into the selector probes during the design. The resulting linear sequences are then amplified many times using ordinary PCR. Producing amounts of products that are detectable, using common methods.

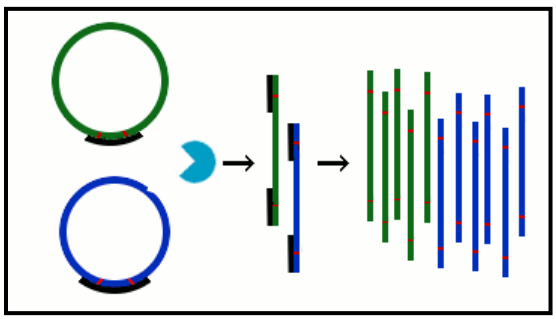


Figure 6: The right part of the image represents the circulated fragments and an enzyme. The center part of the image represents the un-circulated fragments. The right part represents the un-circulated fragments after a PCR reaction.

2.5 Secondary DNA structures

Secondary DNA structures are formations that occur when the bases belonging to a single stranded DNA (ssDNA) base-pair to each other. Some conformations that can be found in a secondary DNA structure are hairpin-loop, bulge-loop, internal-loop, multi-loops and stacking bases. They will occur depending on the properties of the surrounding environment and will theoretically always adopt the conformation with the lowest energy.

2.5.1 Bulge-loop

A bulge-loop is a structure where two parts of the ssDNA base-pair at least two times. It contains a starting and ending base-pair, where only one of the base-paring parts has one or more unpaired nucleotides between the starting and ending base-pair.

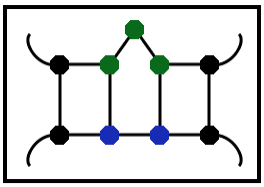


Figure 7: Bulge loop.

2.5.2 Internal-loop

An internal-loop is very similar to a bulge-loop. The difference is that the internal-loop has one or more unpaired nucleotides between the starting and ending base-pair on both base-pairing parts.

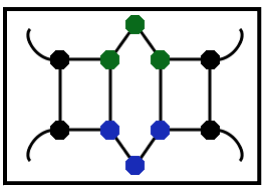


Figure 8: Internal loop.

2.5.3 Hairpin-loop

A hairpin-loop is created when two parts of the ssDNA base-pair. The hairpin-loop will be the part at the end of the structure. It contains two nucleotides that base-pair and are connected by at least three unpaired bases.

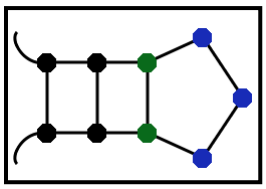


Figure 9: Hairpin-loop.

2.5.4 Stacking base-pair

A stacking base-pair Is a structure where a base-pair is directly followed by another base-pair.

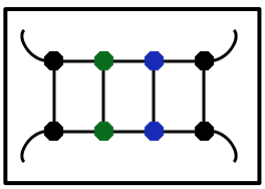


Figure 10: Stacking base-pair.

2.5.5 Stem-loop

If you combine all previous mentioned structures you will have a stem-loop. A structure started with a hairpin-loop or a base pair followed by zero or more stacking-, bulge-, or internal-loops and ended by a hairpin-loop.

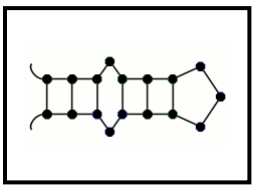


Figure 9: Stem-loop.

2.5.6 Multi-loop

A multi-loop is a loop structure that contains at least two stem-loops. It is started by a base-pair followed by a loop which will contain the starting base-pairs for two or more steam-loops and zero or more unbase-paired nucleotide.

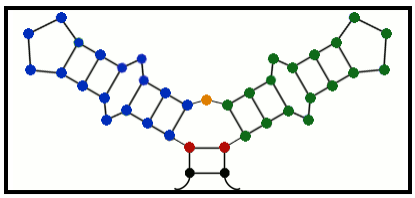


Figure 10: Multi-loop.

2.6 Polymorphism

Polymorphisms are differences between genomes within species that arise because of replication errors that are not detected by the proofreading mechanism or due to a reaction between a mutagen[7]. The change then segregates in the population. Three common forms of polymorphism are deletion, insertion and single nucleotide polymorphisms (SNPs). A deletion will remove one or more bases in a DNA sequence. The opposite of a deletion is an insertion where one or more bases are added. SNPs are the most common polymorphism and one of the biggest reasons for the difference between the genomes of different humans (GarlandScience, 27 Mars 2009). On average you can find a SNP every 2.0 kb (Brown, 2002, p. 18).

The effect of polymorphisms can vary depending on where the polymorphism has occurred (Brown, 2002, p. 428-432).

2.7 Melting temperature T_m

The melting temperature (T_m) is a value indicating at which temperature 50 percent of the DNA strands have separated. Strands with higher GC content require higher temperature to denaturate and hence they have a higher T_m. GC and AT content will not by themselves give an exact value of T_m. For example ionic concentration, pH or other factors may influence the T_m value (Lodish, 2003, p. 105). Nevertheless GC and AT content often are used to approximate T_m.

[7] Any substance that can cause a mutation.

3 Materials and methods

3.1 DNA samples

DNA sequences used to design selector probes were gathered from Ensembl (15 Sep 2008). DNA samples used to test selector probes were ordered from Coriell Institute (15 Oct 2008) and Promega Corporation (15 Oct 2008).

3.2 MLGA Protocol

Protocols[8] used to test selector probes created with the MLGA software are slightly modified versions of the protocol used in the paper *MLGA – rapid and cost-efficient assay for gene copy-number analysis* (Isaksson et al., 2007, p. e115).

3.3 Development software and language

The software has been developed using Java and NetBeans. Java is a platform independent language, which is an advantage as the software will work on most platforms, Windows, Linux, Unix, Mac and so forth (Sun Microsystems, 1 Sep 2008). NetBeans is a "free, open-source Integrated Development Environment", a program that makes it easier to develop software by providing a GUI containing file handling, API, compiler etc (Netbeans, 1 Sep 2008).

3.4 Detection

Detection of MLGA result has been done with the Experion system in combination with the 1k DNA chip. The 1k DNA chip detection range is 25-1000 nucleotides where the size has to differ, different much depending on the size of the fragments (25-100 bp has to differ 5bp, 100-700 has to differ 5 % and 700-1000 has to differ 10 %) (Bio-RadLaboratories, 27 Mars 2009).

3.5 Selector algorithm

One of the main goals with the project was to develop an algorithm that can be used to select restriction fragments that are appropriate to use during the design of selector probes.

The designed selector algorithm works as follows. The user first selects how many selector probes that should be created from each sequence and also which properties should be looked at when finding valid restriction fragments. The software then starts to evaluate restriction

[8] Appendix B.

fragments and create a set of valid restriction fragments. From this set the algorithm will select restriction fragments that will be used to create Selector probes.

3.5.1 Evaluation properties

The properties that the algorithm can use are restriction fragment length, folding value, polymorphism and GC content, and also T_m, GC content and length of the binding regions, to establish if a restriction fragment is appropriate for targeting. The user has the option of selecting which properties to use and may remove all properties except the length property. A restriction fragment that fails a test will have its valid parameter set to false and will not be selected or investigated for any remaining properties.

3.5.1.1 Length

There are at least two reasons for taking the combined length of the selector probe and restriction fragment into account; detection possibilities and circularization success rate.

To be able to detect the different selector probes and restriction fragments, they will have to vary in length. If they are of the same size they will be detected as one probe, at least with electrophoresis agarose or the Experion-system. How much they have to vary in length depend on the choice of detection system.

If the circulation is going to be successful the combined selector probe and restriction fragment length cannot be too small or too big. Successful circulations have been made in the range 100-1000 bp (Stenberg et al., 2005, p. e72).

3.5.1.2 Single stranded DNA structure

During the design of a selector probe it's important to take into account the formation of secondary structure. Because secondary structures will affect how well the selector probe are able to bind to the target DNA. Stable secondary structures may block the binding sites on the restriction fragment or make the formation between the restriction fragment and selector probe to stiff.

The algorithm used to calculate folding energy, originate from the Vienna RNA Package (Hofacker et al., 1993). It uses experimental compiled energy parameters to assign energy values to hairpin-, internal and bulge-loops and stacking base-pairs.

```
0
1     for(d = 1….n)
2          for(i = 1…d)
3               j = i+d;
4               C[ i, j ] = MIN( Hairpin( i, j ),
5                                    MIN( i < p < q < j : Interior(i , j; p,q) ),
6                                    MIN( i < k < j : FM[ i + 1, k ] + FM[ k + 1, j -  1 ] + cc))
7               F[ i, j ] = MIN(  C[ i, j],
8                                    MIN( i < k < j : FM[ i, k ] + FM[ k + 1, j ] )
9              FM[ i, j ] = MIN(  C[ i, j ] + ci,
10                                   FM[ i + 1, j ],
11                                   FM[ i + 1, j ] + cu,
12                                   FM[ i, j – 1 ] + cu,
13                                   MIN( i < k < j :  FM[ i, k] + FM[ k + 1, j ] )
14
```

Algorithm 1: Folding algorithm. C[i,j] the energy given that i and j base-pair, FM[i,j] the energy of the smallest multi-loop structure between i and j, F[i,j] the energy of the most stable secondary structure between i and j.

The Hairpin function calculates the energy of a hairpin loop started at i and ended at j. MIN(i < p < q < j : Interior(i , j; p,q)) will find the internal loop, bulge loop or stacking loop that start at the base-pair i and j and end at base-pair p and q that adopt the lowest energy. MIN(i < k < j : FM[i + 1, k] + FM[k + 1, j - 1] + cc)) calculate the energy of the multi loop starting at base-pair i and j that have the smallest energy of all multi loops that can be formed if i and j base-pair.

The structure with the lowest energy will be stored in F[i,j]. C[i,j] holds the structure with the lowest energy, that have a starting base-pair at i and j. FM[i,j] holds the energy of the structure with lowest energy, that can be created between base i and j, i and j does not have to base-pair.

3.5.1.3 GC content of the restriction fragment

The main reason for taking the GC content into account is that it will influence the PCR reaction, as it is harder to amplify fragments that have high or low GC content. This is undesirable; we need the selector probes to work well every time we use them.

The calculation of the GC content:

$$GC_{i \to j} = \frac{\#G + \#C}{j - i}$$

Formula 1: i, the position where the counting should start. j, the position where the counting should end.

3.5.1.4 Polymorphism

Polymorphisms are changes in the genome and will possibly affect the success rate of the selector probes. If a SNP is located inside the restriction fragment that a selector probe has been designed against, it will affect the selector probes success rate only if it is positioned at the binding regions of the restriction fragment. This will make the ligation between the selector probe and restriction fragment less effective. A SNP site between the binding regions will have no effect.

An insertion or a deletion is much worse, compared to a SNP. These kinds of polymorphism can make the ligation between the selector probe and restriction fragment most likely impossible if they are located at the binding regions, as they will shift the bases, producing many mismatches. An insertion or a deletion between the binding regions will probably not affect the ligation, but they will change the length. This will make the analysis harder as the peaks will move and possibly merge.

3.5.1.5 Tm, GC content and length of the binding region:

There are at least two reasons for investigating these properties, ligation effectiveness and cost. Even though GC content of the whole restriction fragment is good, the binding regions may have a GC content that is not desirable. High or low GC content at the binding regions will make the creation of secondary structures easier. This will make the ligation reaction less effective as secondary structures will block the ligation between the selector probe and restriction fragment.

The bigger the binding regions are the longer the selector probes will become. That will affect the cost, as it is more expensive to create longer selector probes.

$$T_{m,l} = 4 \times (\#A + \#T) + 2 \times (\#G + \#C)$$

Formula 2: Calculate the T_m value given the sequence length l.

The function that investigate T_m,GC content and length work as follows:

```
0
1       for each R
2               While T_m,3' < T_m
3                       Length_3'++ and recalculation of T_m,3
4               While T_m,5' < T_m
5                       Length_5'++ and recalculation of T_m,5
6               IF length_3' < minL OR length_3' > maxL OR length_5' < minL OR length_5' > maxL
9                       Valid = false;
10              ELSE if GC_3' < minGC OR GC_3' > maxGC OR GC_5' < minGC OR GC_5' > maxGC
11                      Valid = false
12              Else
13              Valide = true;
14
```

Algoritm 2: Valid bidning region algoritm. R is all restriction fragments. T_m is the minimum allowed T_m. minL is the minimum length allowed. maxL is the maximum length allowed. minGC is the minimum GC content allowed. maxGC is the maximum GC content allowed.

3.5.2 Selection

When a set of valid restriction fragments have been created, selections will be made with one requirement. That requirement is that the combined restriction fragment and selector probe length has to differ to all other already created selector probes and their restriction fragments.

The selection will be done as follows. Selector probes will be created for each sequence until all sequences have the number of selector probes that they should produce or until no more selector probes can be created without breaking the design requirements.

3.6 The MLGA design software and assays

To be able to start testing the software as early as possible a design method called evolutionary development was used, a development process where one rapidly develop an initial system from an abstract specification. The system is then gradually developed, adding new functionality and re-designing already developed components that need improvement. During all evolutionary steps the customer/user are involved to find new features that are needed and re-designing already defined features to improve the software (Sommerville, 2007, p. 68-69).

3.6.1 MLGA design 1 and software version 1

First basic functionality and design properties where implemented to make it possible to design selector probes using very few properties. Basic functions that were implemented:

- Loading sequence information from Fasta, GenBank or Ensembl files, which are files that can be downloaded from common web pages that provide DNA information.
- Using hard coded[9] vector and restriction enzyme information.
- In silico restriction digestion, finding all cut sites that the restriction enzyme will create and creating restriction fragments.
- Saving the created selector probes.
- Creating a selector probe using a selected restriction fragment and a hard coded T_m value.
- Generation of restriction fragments and automated selection of restriction fragments for selector probe creation. Using the length property to validate fragments and minimum length difference to select fragments.
- Tables containing restriction fragment information, where the user can select a specific restriction fragment to use for creating a selector probe.

Using this software a first design[10] set of 20 selector probes was designed, limiting the length to 50-500 bases and the length difference between two selector probes to at least five percent (a few selector probes that broke the design criteria were created to investigate if the design criteria could be changed).

3.6.2 MLGA design 2 and software version 2

After the first design set had been created the software went through further development to include more design parameters, GC content, polymorphism and folding energy. More functions were added to improve usability:

- A panel where the user can set design limits and which parameters to use.
- Tables containing restriction fragment were extended with functions to show more detailed restriction fragment information and the ability to calculate properties for restriction fragments.

[9] The object information is written in to the source code and cannot be changed without modifying the source code.

[10] Appendix A, Design 1: Selector Probes.

- The function to read sequence files was improved to be able to read information about polymorphism that are stored inside a sequence file.
- The option of saving a project and open it again.
- Load files containing information about vector or restriction enzyme.
- Input box for T_m value.

A second design[11] of 20 selector probes was made where the length was limited to 100-400 bp and the length difference was kept at five percent or higher, GC content was limited to the range 45-55 percent, no known SNPs, insertions or deletions were allowed (a few selector probes that broke the design criteria were created to be able to create more selector probes).

3.6.3 MLGA design 3 and software version 3

A third design and improvement of the software was made. Where a more reliable enzyme was used, to improve the restriction fragment reaction. The software usability and functionality were extended with:

- The selection algorithm was extended with the possibility to investigate the GC content and length of the target arms that met the minimum T_m value.
- A panel containing a table with information about all created selector probes, GC content of the restriction fragment, polymorphism, folding value, combined selector probe and restriction fragment length, GC content of the right target sequence, GC content of the left target sequence.
- The possibility to search for enzyme binding sites in the circulated selector probe and restriction fragment.

A third design[12] of 17 selector probes was made where the length was limited to 50-500 bp and the length difference was kept at five percent or higher, GC content was limited to the range 45-55 percent, no known SNPs, insertion or deletions were allowed (a few selector probes that broke the design criteria were created to be able to create more selector probes). End modification sequences were added to prevent ligation between selector probes.

3.7 The MLGA data analysis software

The analysis of an MLGA assay result is made in two steps. The first step is to use the MLGA data analysis software to extract all relevant information from out-put files created by the Experion

[11] Appendix A, Design 2: Selector Probes.

[12] Appendix A, Design 3: Selector Probes.

system. The extracted information is then saved to files that can be imported into Excel where the data will be processed.

The second analyze step is to normalize the data and compare it against a reference. First each sample is normalized by dividing all peak areas with the sum of all autosomal[13] chromosomes peak areas except chromosome 21 peak areas. Then all samples are compared to a reference sample by dividing each peak area in a sample with the corresponding peak area in the normalized reference sample.

[13] Any chromosome other then a sex chromosome (X and Y).

4 Result and discussion

4.1 The software

The main goal with this project was to create software[14] that will be used to design selector probes used in MLGA assays. This has been done and the software now contains these features:

- Load various sequence formats that can include information about polymorphism.
- Load end-modification-, enzyme- and vector-files, that are used to design a selector probe.
- Create selector probes by selecting restriction fragments or by using the algorithm that can select valid restriction fragments.
- Perform in silico restriction digestion on a sequence with a selected enzyme.
- Can create project file that can be used to continue work on the project, PDF file containing information about created selector probes and selector probes file that is used to order selector probes.
- Selection algorithm giving the ability of sorting out restriction fragments that does not fulfill minimum folding energy, GC content, polymorphism, length and binding region properties.
- Graphical user interface that give the user easy access to all functions. The GUI also provide graphical information using different panels:
    - Information about each sequence and its restriction fragments.
    - Information about each created selector probe.
    - A summary containing all created selector probes and their properties.
    - A panel where the user can use the algorithm to generate and select restriction fragments for all sequences at the same time.

A secondary goal was to create software used to analyze an MLGA assay result. That software now contains these features:

- Can read design XML file used to define peaks that should be found.
- Opening Experion out-put files and sorting out non-valid peaks using the design XML file.
- Creating csv file containing information about peaks that can be imported into Excel.

[14] Screen dump images, Appendix C.

4.2 Selector probe design:

4.2.1 Design 1

Selector probe success rate		
Working	Not working	Multi ligation
10 / 20	6/20	4/20

Table 2: The success rate of design selector probes. Where Working are selector probes that gave signal. Not working are selector probes that gave no signal. Multi ligation are selector probes that gave multiple signals.

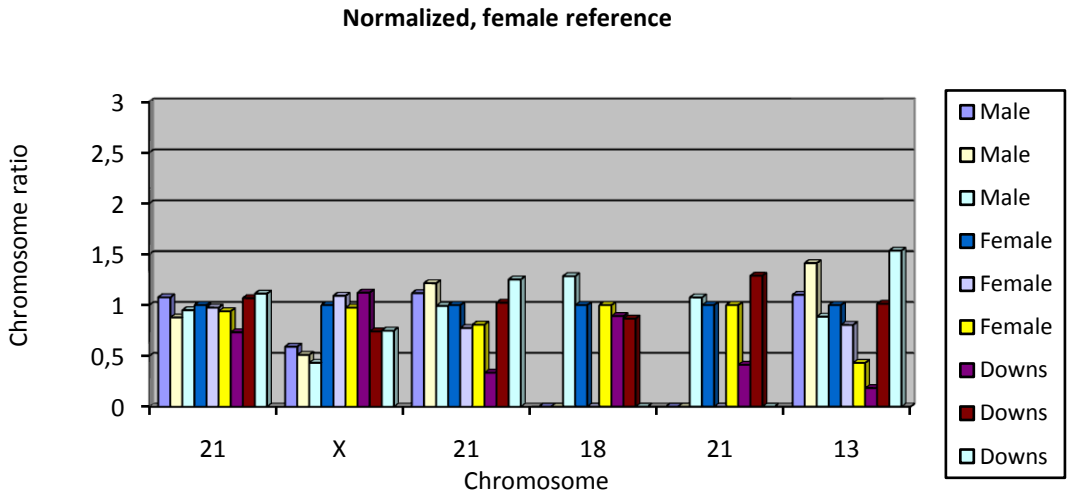


Diagram 1: The normalized result compared to a healthy female reference, which means that a chromosome ratio of 1 indicate that the sample has two set of that chromosome as a healthy female always has two set of each chromosome.

At the moment the result is not really good, it is too unreliable. We can differ between females and males, between healthy subjects and Downs subjects but just sometimes.

The result success rate for the Selector probes isn't very good. Half of the Selector probes do not work as intended, a few do not work at all other give multiple signals. The ones that work do not give a reliable result and some of them work sometimes, others are amplified uneven between different reactions. This influence the normalization negatively as we sometimes get fewer peak areas to use during the normalization and other times they work but influence the normalization too much. Chromosome 13 will for example influence the normalization a lot, as it differ so much from time to time and also differ a lot from the expected result.

A common property of the selector probes that worked sometimes was high or low GC content[15]. This could explain why they did not work very good, a high or low GC content can make it harder to amplify the sequence.

4.2.2 Design 2

Selector probe success rate		
Working	Not working	Multi ligation
13 / 20	1/20	6/20

Table 3: The success rate of design selector probes. Where Working are selector probes that gave signal. Not working are selector probes that gave no signal. Multi ligation are selector probes that gave multiple signals.

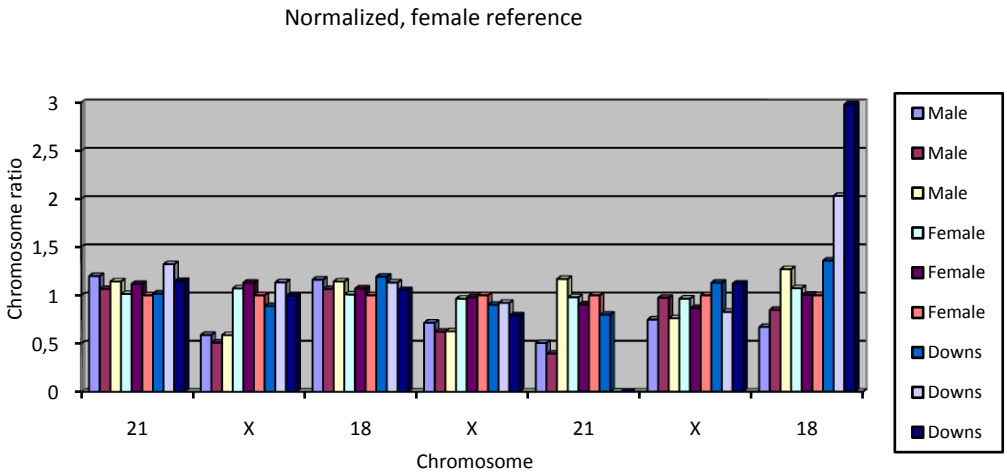


Diagram 2: The normalized result compared to a healthy female reference, which means that a chromosome ratio of 1 indicate that the sample has two set of that chromosome as a healthy female always has two set of each chromosome.

The success rate with design 2 was much higher, only 4 of the 20 probes did not work as intended. The normalized peak areas compared to a reference between different samples were more even, producing a more reliable result.

A new problem arose with design 2 when the selector probes where mixed. Some of them seemed to ligate to each other producing a big bump[16] which affected the corrected area of the peaks found there. Some were hidden and others got bigger areas.

[15] Appendix A, Design 1: Selector Probes.

4.2.3 Design 3

Selector probe success rate		
Working	Not working	Multi ligation
12 / 17	2/17	3/17

Table 4: The success rate of design selector probes. Where Working is selector probes that gives signal. Not working is selector probes that gave no signal. Multi ligation is selector probes that gave multiple signals.

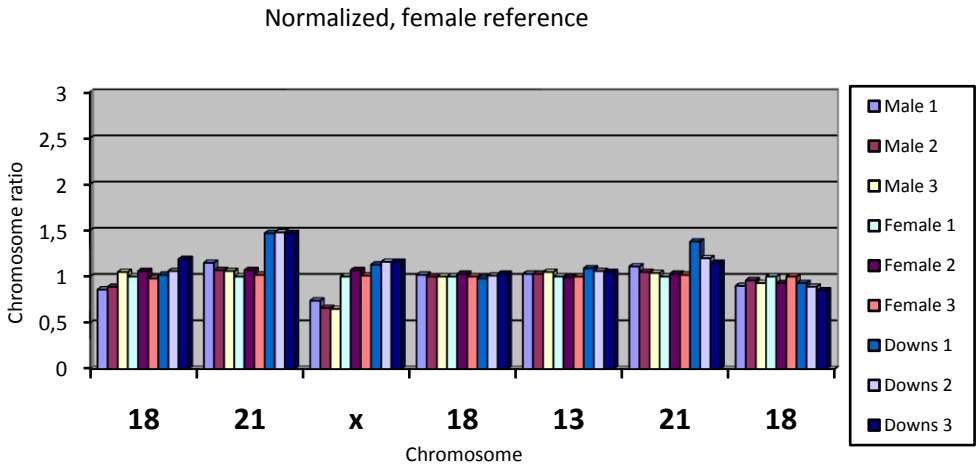


Diagram 3: The normalized result compared to a healthy female reference, which means that a chromosome ratio of 1 indicate that the sample has two set of that chromosome as a healthy female always has two set of each chromosome.

The change in Design 2 to Design 3 seems to have improved the result. Design 3 produces a normalization that can differ between males and females and between healthy samples and Down's. The normalization showed in Diagram 3 is more reliable compared to the normalization of the previous designs (Diagram 1 and Diagram 2). The peaks are more even and with the exception of the last chromosome 21 the selector probes always give the correct ratio.

The success rate of the selector probe has also improved. The success rate is at the moment 70% (table 3) compared to design 1 where the success rate was 50% and design 2 where the success rate was 65 %. The improvement from design 2 to Design 3 was not big and that is most likely because that the change from design 2 to design 3 concentrated on removing the problem with ligation between selector probes.

[16] Appendix D.

5 Conclusions

The software and the current design parameters work, but can be improved. Currently some selector probes still do not work and other give multiple peaks. Future design will be improved with each new design, as we learn new things with each new design and new unsuccessful selector probe.

5.1 The current version of the software

The current version of the design software implements the goals that were set in the project plan. The user can import new design parts like sequence, enzymes and vectors. Restriction fragments that are used to create selector probes can be sorted to remove invalid fragments by specifying design parameters like GC content, polymorphism, length and so forth. All functions are built into a GUI that provides information about created restriction fragments and selector probes.

The data analysis software is very basic at the moment. It will remove invalid peaks and produce a csv file that can be imported to excel where the normalization and comparison to a reference will be made.

5.2 Feature and improvement of the software

At the moment some selector probes will produce multiple peaks that indicate that the selector probes bind to more than one place. This problem could possibly be reduced using Blast to search for other sites that the selector probe can bind to.

The analyze software can be improved to include functions to normalize, compare to reference and to display the result. This would remove the need of Excel.

5.3 The current design

The current design has improved a lot since the first design where less than half of the selector probes worked as intended. With the new design more than half of the selector probes work. Current design take GC content, polymorphism, length and binding regions T_m value, length and GC content into account.

5.4 Design improvement

The software currently supports the calculation of secondary structure and the resulting folding energy. But we do not know how to effectively use this in the design. Future design could

possibly be improved by studying the secondary structure and folding energy of the selector probes that have failed.

6 Acknowledgements

I would like to thank my supervisor Simon Fredriksson for providing this project and helping me with any questions that have had. Thanks to Patrik Forssén for accepting to be my scientific reviewer. Also thanks to Mats Nilsson and Magnus Isaksson at Rudbeck Laboratory for giving me valuable input about possible improvements with new MLGA analysis designs and software parameters.

I final thanks to my opponents and all the people at Olink Bioscience whose help and knowledge have improved this report.

7 References

Papers:

1. BENITA, Y., OOSTING, R., LOK, M., WISE, M. & HUMPHERY-SMITH, I. 2003. Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res,* 31, e99.
2. DAHL, F., GULLBERG, M., STENBERG, J., LANDEGREN, U. & NILSSON, M. 2005. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res,* 33, e71.
3. ISAKSSON, M., STENBERG, J., DAHL, F., THURESSON, A., BONDESON, M. & NILSSON, M. 2007. MLGA--a rapid and cost-efficient assay for gene copy-number analysis. *Nucleic Acids Res,* 35, e115.
4. STENBERG, J., DAHL, F., LANDEGREN, U. & NILSSON, M. 2005. PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic Acids Res,* 33, e72.

Books:

1. BROWN, T. A. 2002. *Genomes,* Oxford, BIOS Scientific Publishers.
2. MATHEWS, C. K., VAN HOLDE, K. E. & AHERN, K. G. 2000. *Biochemistry,* San Francisco, Calif., Addison Wesley Publishing company.
3. LODISH, H. F. 2003. *Molecular cell biology,* New York, W. H. Freeman and Co.
4. SOMMERVILLE, I. 2007. *Software engineering,* Harlow, Addison-Wesley.

Webb pages:

1. Promega Corporation, http://www.promega.com, 15 Oct 2008.
2. Ensembl, http://jul2008.archive.ensembl.org/index.html, 15 Sep 2008.
3. Coriell Institute, http://www.coriell.org/, 15 Oct 2008.
4. Bio-Rad Laboratories, http://www.bio-rad.com, 27 Mars 2009.
5. Sun Microsystems, http://java.sun.com, 1 Sep 2008.
6. Netbeans, http://www.netbeans.org/features/, 1 Sep 2008.
7. Garland Science, http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hmg.section.1050, 27 Mars 2009.
8. HOFACKER, I. L., FONTANA, W., STADLER, P. F., BONHOEFFER, L. S., MANFRED & SCHUSTER, T. P. 1993, *Fast Folding and Comparison of RNA Secondary Structures* http://fontana.med.harvard.edu/www/Documents/WF/Papers/vienna.rna.pdf, 10 Dec 2008.

Appendix A

List of target loci for the MLGA selector probe set

Genome area	Chromosome
ABBC4	13
Ar	X
BRCA2	13
CYORF14	Y
DSCR6	21
L1CAM	X
MADH4	18
NFATC1	18
RPS6KA3	X
SERPINB2	18
SIM2	21
SOD1	21
SRY	Y
STCH	21
TYMS	18

Design 1: Selector Probes

Design 1	L	#SNP	#I/D	ΔG	min ΔL	#Enzy SP	#Enz sites RF	3' end GC	5' end GC	RF GC	Worked
ABCC4-P0	349	1	0	-28,18	7%	1	0	75%	33%	41%	M
ABCC4-P1	393	2	0	-32,32	7%	2	1	42%	35%	37%	N
Ar-P0	494	0	0	-22,93	7%	1	0	23%	56%	37%	M
Ar-P1	421	0	0	-25,63	7%	1	0	50%	35%	30%	N
BRCA2-P0	147	3	0	-10,79	7%	1	0	59%	65%	51%	Y
BRCA2-P1	587	2	0	-40,82	16%	1	0	35%	56%	32%	N
DSCR6-P0	274	0	0	-20,47	8%	1	0	50%	59%	41%	Y
L1CAM-P0	234	0	0	-34,53	15%	1	0	56%	75%	55%	Y
NFAT-P0	125	0	0	-21.90	6%	1	0	72%	72%	74%	N
NFAT-P1	541	0	0	-1.11	8%	1	0	55%	50%	50%	M
RPS6KA3-P0	298	0	0	-25,41	8%	1	0	35%	59%	36%	M
RPS6KA3-P1	84	0	0	-3,51	9%	1	0	65%	65%	66%	N
SERPINB2-P0	457	5	0	-34,89	7%	1	0	65%	17%	40%	Y
SOD1-P0	136	4	0	-13,74	7%	1	0	59%	80%	62%	Y
SOD1-P1	191	0	0	-10,12	7%	1	0	50%	69%	44%	Y
SRY-P0	323	1	0	-33,22	7%	1	0	65%	50%	53%	Y
SRY-P1	108	0	0	-4,16	8%	1	0	59%	59%	47%	Y
STCH-P0	117	0	0	-5,93	8%	1	0	50%	59%	40%	Y
STCH-P1	92	0	0	-3,04	9%	1	0	42%	35%	31%	N
TYMS-P0	177	0	0	-32,64	7%	1	0	69%	93%	78%	Y

Table 5: Design parameters for the first design. # = number of. I/D = Insertion/Deletion, Enz = Enzyme. 3' end GC = GC content at the 3' binding region. 5' end GC = GC content at the 5' binding region. min ΔL = minimum length difference to any other selected restriction fragment. ΔG folding energy. L = length. Y = Yes. N = No signal. M = multiple peaks. Yellow marking are probes that did not work good every time. Light-blue are selector probes that worked good every time.

Design 2: Selector Probes

Design 2	L	#SNP	#I/D	ΔG	min ΔL	#Enz SP	#Enzy sites RF	3' end GC	5' end GC	RF GC	Worked
ABCC4-P0	165	0	0	-13,03	4%	1	0	71%	53%	49%	Y
ABCC4 -P1	221	0	0	-19,39	5%	1	0	45%	61%	48%	Y
ABCC4-P2	272	0	0	-25,34	4%	1	0	61%	58%	49%	Y
Ar-P0	134	0	0	-5,46	8%	1	0	67%	45%	47%	Y
Ar-P1	289	0	0	-26,81	4%	1	0	58%	53%	49%	Y
BRCA2-P0	413	1	0	-33,25	9%	1	0	71%	38%	44%	M
CYORF14-P0	211	0	0	-15,45	5%	1	0	43%	45%	46%	Y
CYORF14-P1	260	0	0	-24,52	4%	1	0	45%	61%	51%	M
CYORF14-P2	320	0	0	-36,88	4%	1	0	43%	43%	46%	Y
DSCR6-P0	233	0	0	-24,34	5%	1	0	71%	71%	49%	M
L1CAM-P0	155	0	0	-17,74	6%	1	0	45%	50%	46%	M
MADH4-P0	192	0	0	-21,68	4%	1	0	71%	53%	52%	Y
MADH4-P1	301	0	0	-18,82	4%	1	0	67%	45%	45%	Y
RPS6KA3-P0	181	0	0	-15,84	5%	1	0	53%	67%	52%	Y
RPS6KA3-P1	246	0	0	-29,19	5%	1	0	76%	53%	50%	W
SERPINB2-P0	172	0	0	-16,95	4%	1	0	32%	53%	49%	Y
SERPINB2-P1	376	3	0	-25,34	9%	1	1	53%	38%	39%	Y
SIM2-P0	200	0	0	-15,98	4%	1	0	45%	61%	52%	M
SOD1-P0	333	0	0	-40,06	4%	1	0	36%	71%	47%	M
SRY-P0	146	0	0	-12,73	6%	1	0	61%	45%	49%	Y

Table 6: Design parameters for the first design. # = number of. I/D = Insertion/Deletion, Enz = Enzyme. 3' end GC = GC content at the 3' binding region. 5' end GC = GC content at the 5' binding region. min ΔL = minimum length difference to any other selected restriction fragment. ΔG folding energy. L = length. Y = Yes. N = No signal. M = multiple peaks.

Design 3: Selector Probes

Design 3	L	#SNP	#I/D	ΔG	min ΔL	#Enz SP	#Enz sites RF	3' end GC	5' end GC	RF GC	Worked
ABCC4-P0	84	0	0	-5,25	9%	1	0	48%	55%	50%	Y
ABCC4- P1	210	0	0	-15,12	10%	1	0	43%	50%	47%	Y
ABCC4-P2	263	0	0	-28,59	11%	1	0	58%	58%	55%	Y
ABCC4-P3	405	0	0	-32,66	11%	1	0	43%	55%	42%	Y
Ar - P0	142	0	0	-6,12	10%	1	0	48%	50%	45%	Y
BRCA2-P0	174	0	0	-12,94	8%	1	0	55%	43%	51%	M
DSCR6-P0	233	0	0	-26,16	10%	1	0	58%	50%	48%	M
DSCR6-P1	303	0	0	-35,81	11%	1	0	43%	58%	52%	Y
L1CAM-P0	114	0	0	-10,02	9%	1	0	48%	50%	55%	N
NFATC1-P0	157	0	0	-18,35	10%	1	0	58%	43%	50%	Y
NFATC1-P1	339	0	0	-35,90	6%	1	0	43%	63%	58%	Y
RPS6KA3-P1	189	0	0	-15,70	8%	1	0	58%	50%	48%	Y
RPS6KA3-P2	466	0	0	-36,22	13%	1	0	50%	43%	42%	N
SERPINB2-P0	92	0	0	-3,55	9%	1	0	55%	43%	48%	Y
SERPINB2-P2	102	0	0	-6,92	10%	1	0	43%	50%	50%	Y
SIM2-P0	125	0	0	-8,91	9%	1	0	55%	55%	49%	Y
SOD1-P0	359	0	0	-35,58	6%	1	0	55%	58%	48%	M

Table 7: Design parameters for the first design. # = number of. I/D = Insertion/Deletion, Enz = Enzyme. 3' end GC = GC content at the 3' binding region. 5' end GC = GC content at the 5' binding region. min ΔL = minimum length difference to any other selected restriction fragment. ΔG folding energy. L = length. Y = Yes. N = No signal. M = multiple peaks.

Appendix B

Protocol 1

Time restriction react.	60min at 37°C, ended with 20min 65°C

Restriction digestion (1x) 200 ng DNA				
DNA	**Volume DNA mix (ul)**	**H2O (ul)**	**Buffer & Enzyme mix (ul)**	**End volume (ul)**
Male	2,22	1,28	1,5	5
Female	3,25	0,25	1,5	5
Trisomi 21	1,41	2,09	1,5	5

Restriction digestion (12 X)				
DNA	**Volume DNA mix (ul)**	**H2O (ul)**	**Buffer & Enzyme mix (ul)**	**End volume (ul)**
Male	26,67	15,33	18	60
Female	38,96	3,04	18	60
Trisomi 21	26,67	15,33	18	60

Buffer & Enzyme Mix	
	Volume (ul)
NEB4 Buffer	0,5
BSA	0,5
Mnl I	0,5
End volume	1,5

Time ligation react.	10min at 95°C, 3x(5min at 75°C, 5min at 65°C, 5min at 60°C, 5min at 55°C, 10min at 50°C)

Ligation reaction											
Ligation mix	**conc.**		**1**	**X**	**12**	**X**	**Ligation mix conc.**		**Final conc.**	
PCR buffer	10	X	0,75	ul	9	ul	0,75	X	0,50	X
MgCl2	50	mM	2,9	ul	34,8	ul	14,5	mM	9,67	mM
NAD	10	mM	1,2	ul	14,4	ul	1,2	mM	0,80	mM
Ampligase	5	U/ul	0,6	ul	7,2	ul	0,3	U/ul	0,20	U/ul
Vector	1	uM	0,33	ul	3,96	ul	33,0	nM	0,02	nM
DNA restriction product			5	ul	60	ul				
H2O			2,72	ul	32,64	ul				
Selector probe(s)	1	nM	1,5	ul	18	ul	0,15	nM	0,10	nM
Volume			10	ul	120	ul				
Final Volume			15	ul	180	ul				

Time exonuclease react.	60min at 37°C, ended with 10min at 70°C

Exo I reaction								
Exo I Mix	**conc.**		**1**	**X**	**12**	**X**	**Final conc.**	
Buffer	10	X	3	ul	36		1	X
Exo I	10	U/ul	0,75	ul	9		0,25	U/ul
H2O			11,25	ul	135			
DNA ligation product			15	ul	180			
Final Volume			30	ul	360			

Time PCR	30min at 37°C, 5min at 95°C, 35x(15s at 95°C, 30s at 55°C, 60s at 72°C) ended with 10min at 72°C

PCR reaction									
PCR mix	**conc.**		**1**	**X**	**12**	**X**	**PCR mix**		**Final conc.**
dNTP "U"	25	mM	0,25	ul	3	ul	0,33	mM	0,25
PCR buffer	10	X	1,75	ul	21	ul	0,92	X	0,7
MgCl2	50	mM	0,25	ul	3	ul	0,66	mM	0,5
Fw primer	10	uM	1,25	ul	15	ul	0,66	uM	0,5
Rev primer	10	uM	1,25	ul	15	ul	0,66	uM	0,5
Hind III	10	U/ul	0,5	ul	6	ul	0,26	U/ul	0,2
Taq-pol	5	U/ul	0,3	ul	3,6	ul	0,08	U/ul	0,06
H2O			13,45	ul	161,4	ul			
Volume			19	ul	228	ul			
Template DNA			6	ul	72	ul			
Final Volume			25	ul	300	ul			

Protocol 2

Time restriction react.	60min at 37°C, ended with 20min 65°C

Restriction digestion (1x) 200 ng DNA				
DNA	**Volume DNA mix (ul)**	**H2O (ul)**	**Buffer & Enzyme mix (ul)**	**End volume (ul)**
Male	2,22	1,78	1	5
Female	3,25	0,75	1	5
Trisomi 21	1,41	2,59	1	5

Restriction digestion (scaled up)			12	X
DNA	**Volume DNA mix (ul)**	**H2O (ul)**	**Buffer & Enzyme mix (ul)**	**End volume (ul)**
Male	26,67	21,33	12	60
Female	38,96	9,04	12	60
Trisomi 21	26,67	21,33	12	60

Buffer & Enzyme Mix	
	Volume (ul)
NEB4 Buffer	0,5
BSA	0
CviAII	0,5
End volume	1

Time ligation react.	10min at 95°C, 3x(5min at 75°C, 5min at 65°C, 5min at 60°C, 5min at 55°C, 10min at 50°C)

Ligation reaction											
Ligation mix	**conc.**		**1**	**X**	**12**	**X**	**Ligation mix conc.**		**Final conc.**	
PCR buffer	10	X	0,75	ul	9	ul	0,75	X	0,50	X
MgCl2	50	mM	2,9	ul	34,8	ul	14,5	mM	9,67	mM
NAD	10	mM	1,2	ul	14,4	ul	1,2	mM	0,80	mM
Ampligase	5	U/ul	0,6	ul	7,2	ul	0,3	U/ul	0,20	U/ul
Vector	1	uM	0,33	ul	3,96	ul	33,0	nM	0,02	nM
DNA restriction product			5	ul	60	ul				
H2O			2,72	ul	32,64	ul				
Selector probe(s)	1	nM	1,5	ul	18	ul	0,15	nM	0,10	nM
Volume			10	ul	120	ul				
Final Volume			15	ul	180	ul				

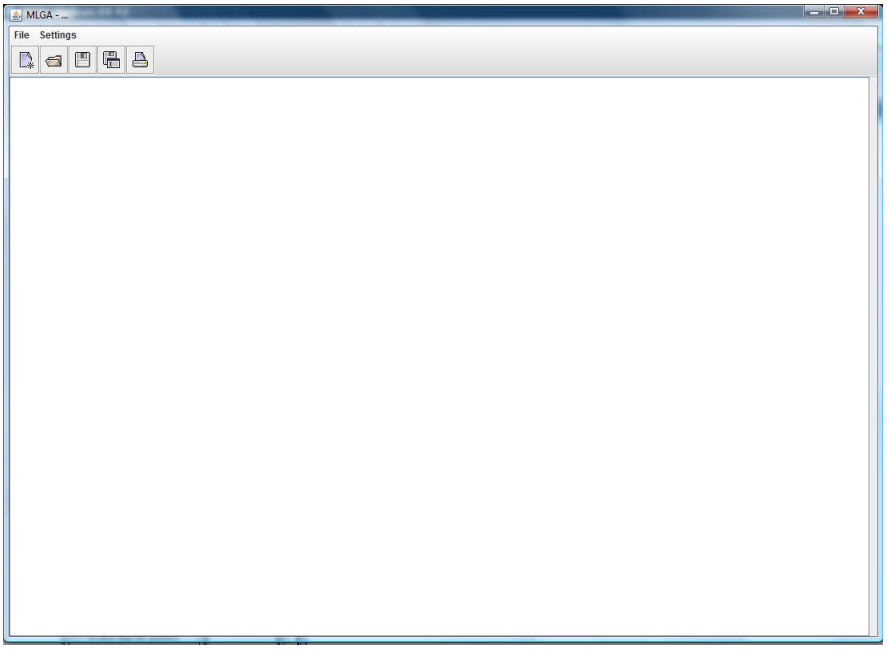
Time exonuclease react.	60min at 37°C, ended with 10min at 70°C

Exo I reaction								
Exo I Mix	**conc.**		**1**	**X**	**12**	**X**	**Final conc.**	
Buffer	10	X	3	ul	36		1	X
Exo I	10	U/ul	0,75	ul	9		0,25	U/ul
H2O			11,25	ul	135			
DNA ligation product			15	ul	180			
Final Volume			30	ul	360			

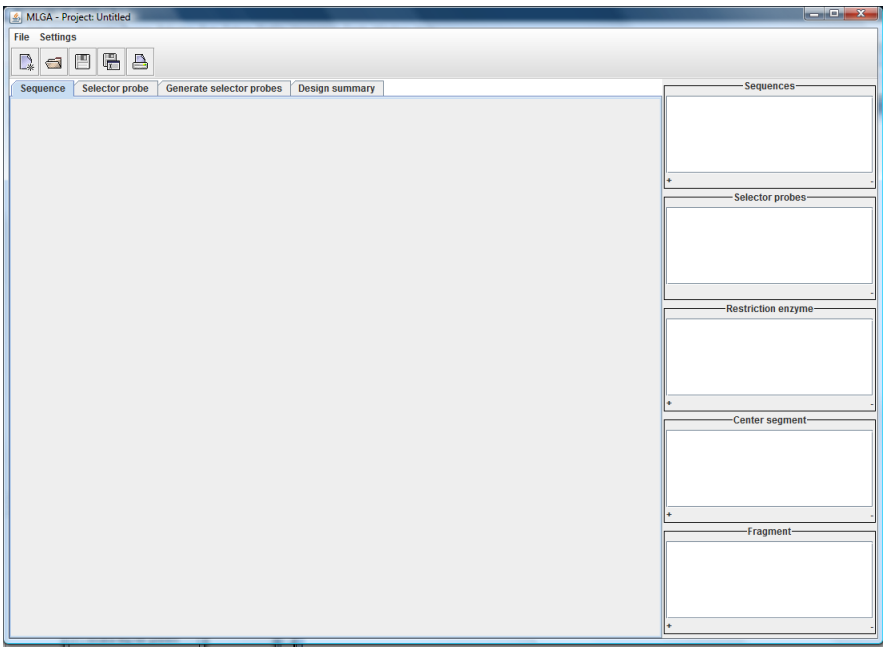
Time PCR	30min at 37°C, 5min at 95°C, 35x(15s at 95°C, 30s at 55°C, 60s at 72°C) ended with 10min at 72°C

PCR reaction									
PCR mix	**conc.**		**1**	**X**	**12**	**X**	**PCR mix**		**Final conc.**
dNTP "U"	25	mM	0,25	ul	3	ul	0,33	mM	0,25
PCR buffer	10	X	1,75	ul	21	ul	0,92	X	0,7
MgCl2	50	mM	0,25	ul	3	ul	0,66	mM	0,5
Fw primer	10	uM	1,25	ul	15	ul	0,66	uM	0,5
Rev primer	10	uM	1,25	ul	15	ul	0,66	uM	0,5
Hind III	10	U/ul	0,5	ul	6	ul	0,26	U/ul	0,2
Taq-pol	5	U/ul	0,3	ul	3,6	ul	0,08	U/ul	0,06
H2O			13,45	ul	161,4	ul			
Volume			19	ul	228	ul			
Template DNA			6	ul	72	ul			
Final Volume			25	ul	300	ul			

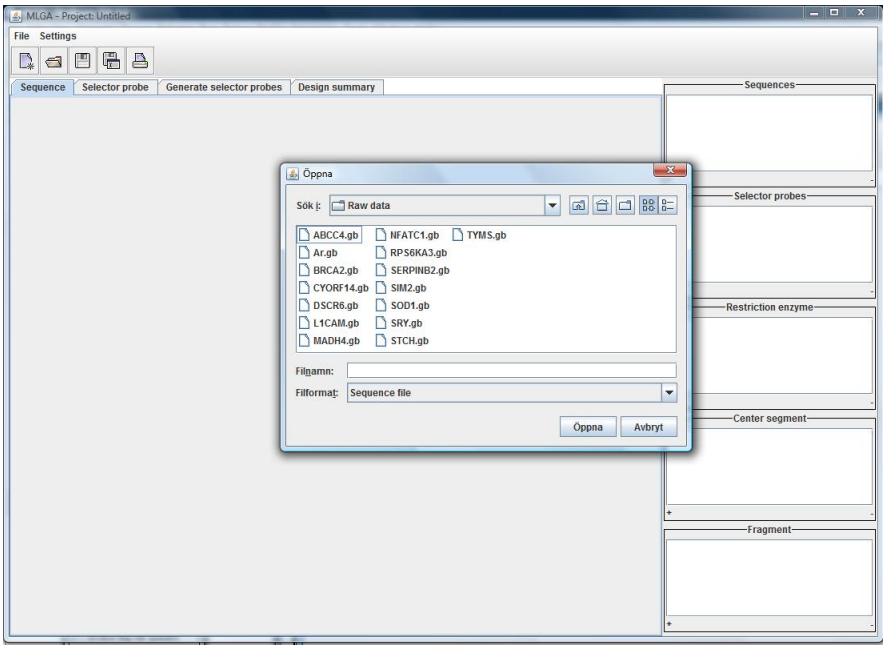
Appendix C



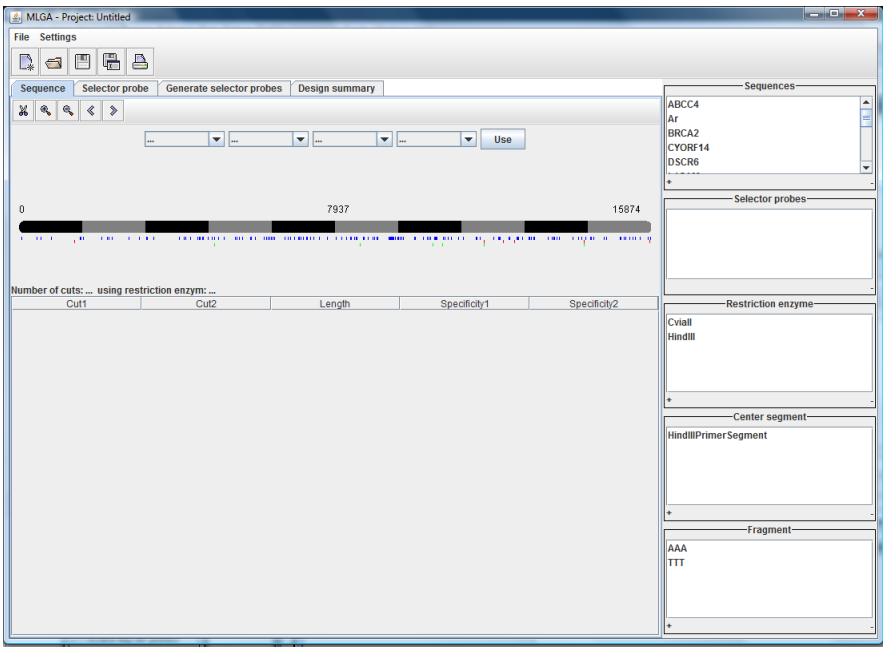
Screen image 1: A screen dump of the design software just after it has been started.



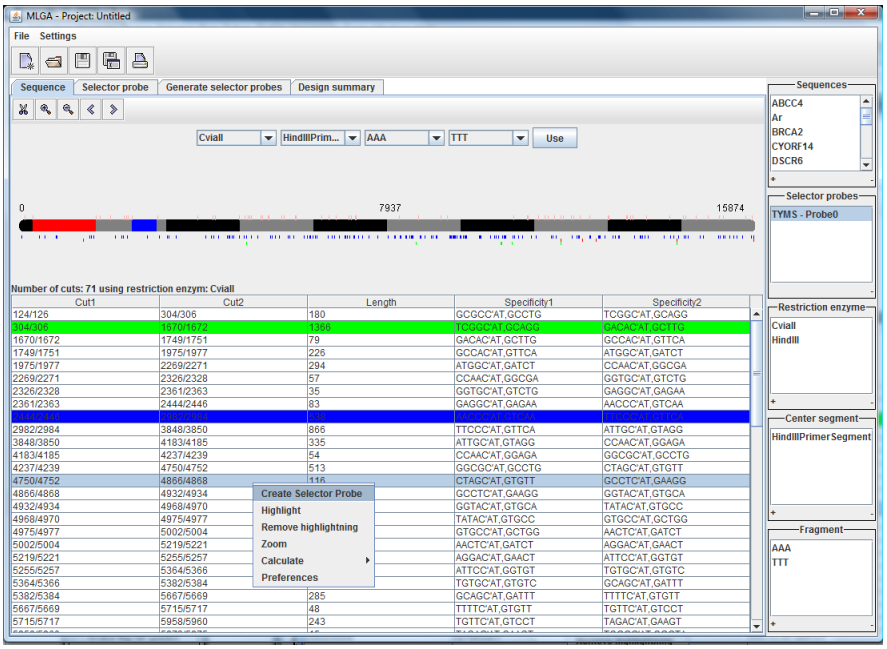
Screen image 2: Screen dump of new project.



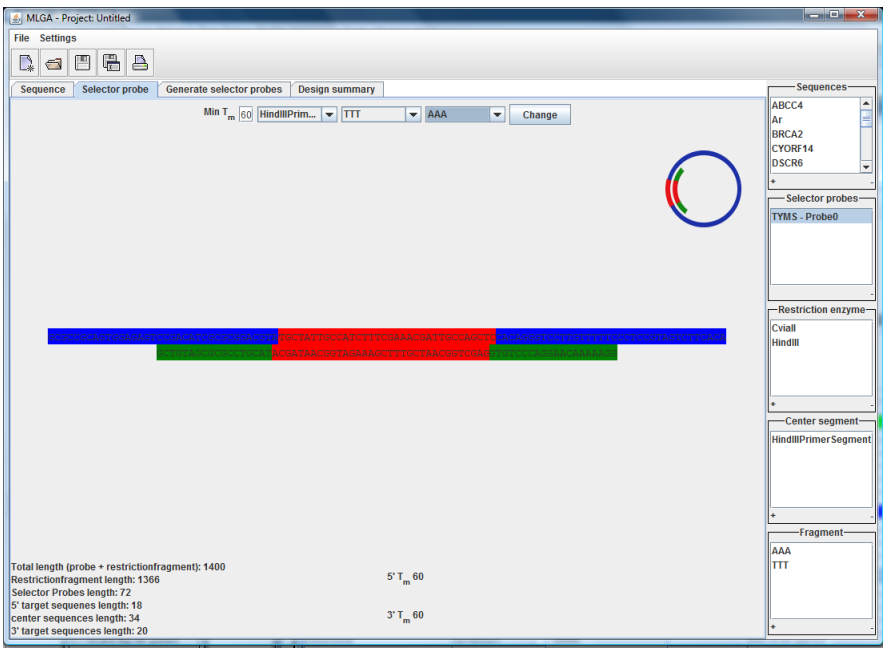
Screen image 3: A screen dump where the user are selecting sequence that should be loaded.



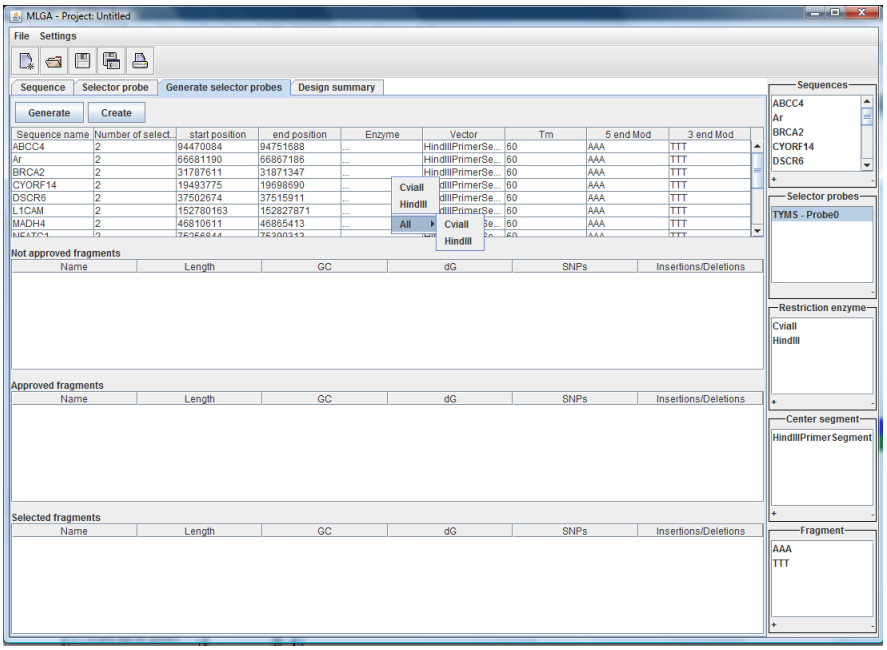
Screen image 4: Screen dump presenting information about a sequence that have been loaded. Also on the right the user have information about loaded sequences, enzymes, vectors and end modification sequences that have been loaded.



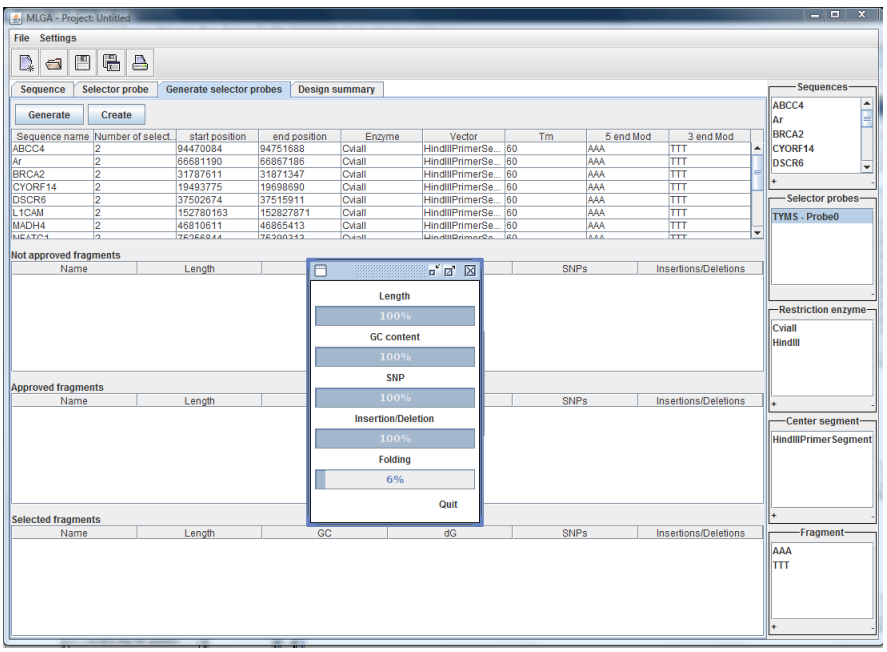
Screen image 5: Screen dump representing information about a sequence that has had enzyme, vector and end modification sequences set. The sequence has been digested using the selected enzyme, the restriction products are displayed in the table. One restriction fragment has been used to create a selector probe (table->green, sequence image-> red), on restriction fragment has been highlighted (table->blue, sequence image->blue). A third restriction fragment has been selected for processing (popup window).



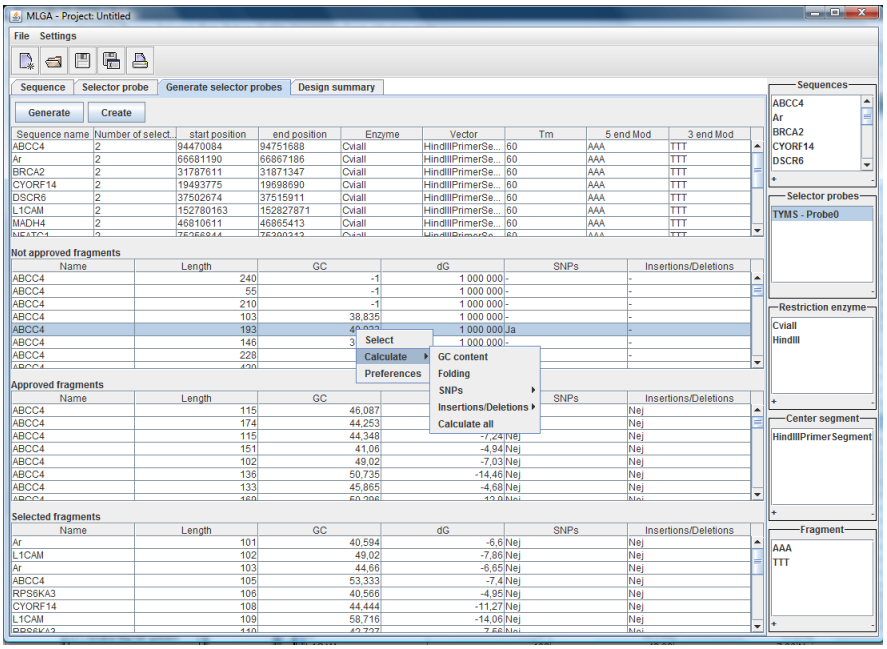
Screen image 6: Screen dump presenting information about a created selector probe, length, T_m and bases that build up the selector probe. Red->Vector, Green-> binding segment, blue->restriction fragment.



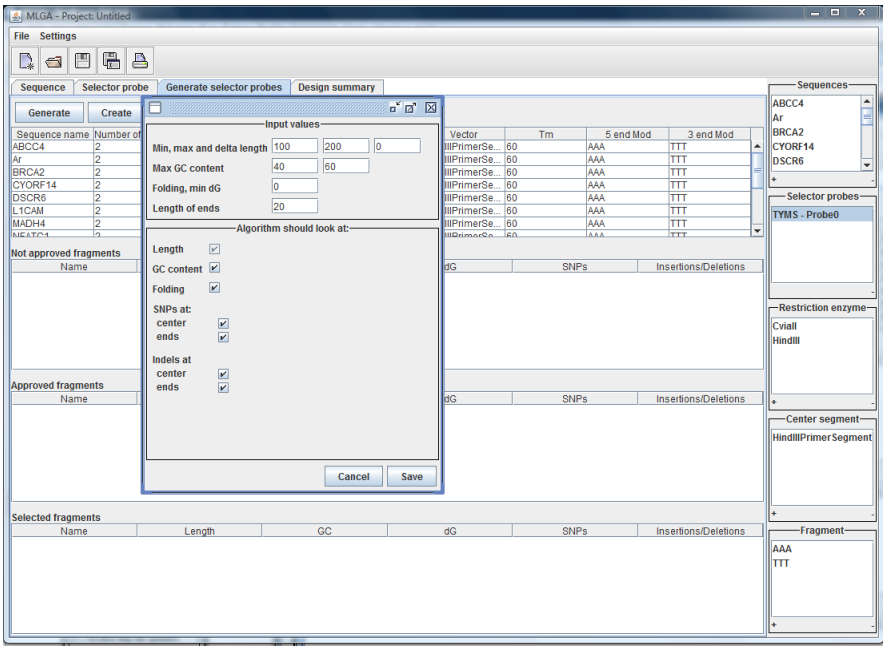
Screen image 7: Screen dump presenting the part of the software where the user can use the selection algorithm. The first table holds all loaded sequences; here the user can set what enzyme, vector, end modification sequences that should be used. The user can also set number of selector probes that should be created and with what T_m.



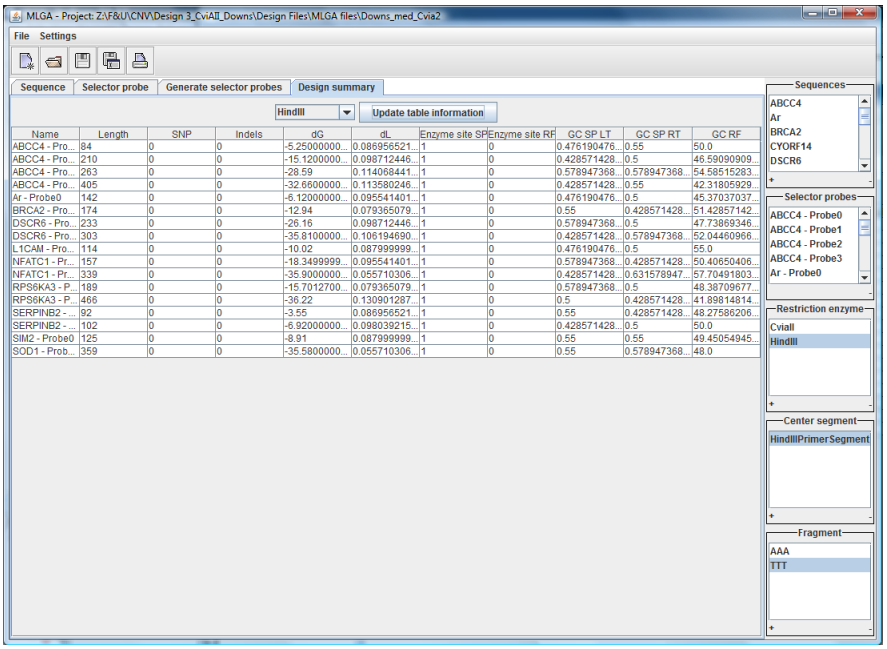
Screen image 8: Screen dump showing the progress of the selection algorithm.



Screen image 9: Screen dump showing the result of the selection algorithm. The second table holds restriction fragments that failed a test, the third restriction fragments that passed all test and the fourth table restriction fragments that have been selected.

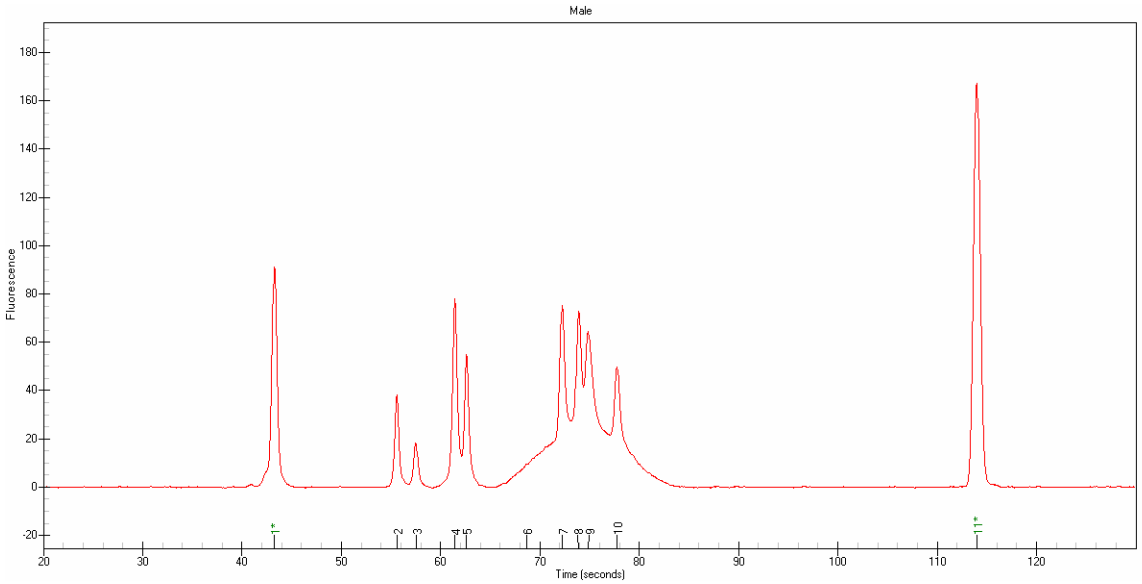


Screen image 10: Screen dump showing the panel used to set what parameters that the algorithm should use and the values used to approve restriction fragments.

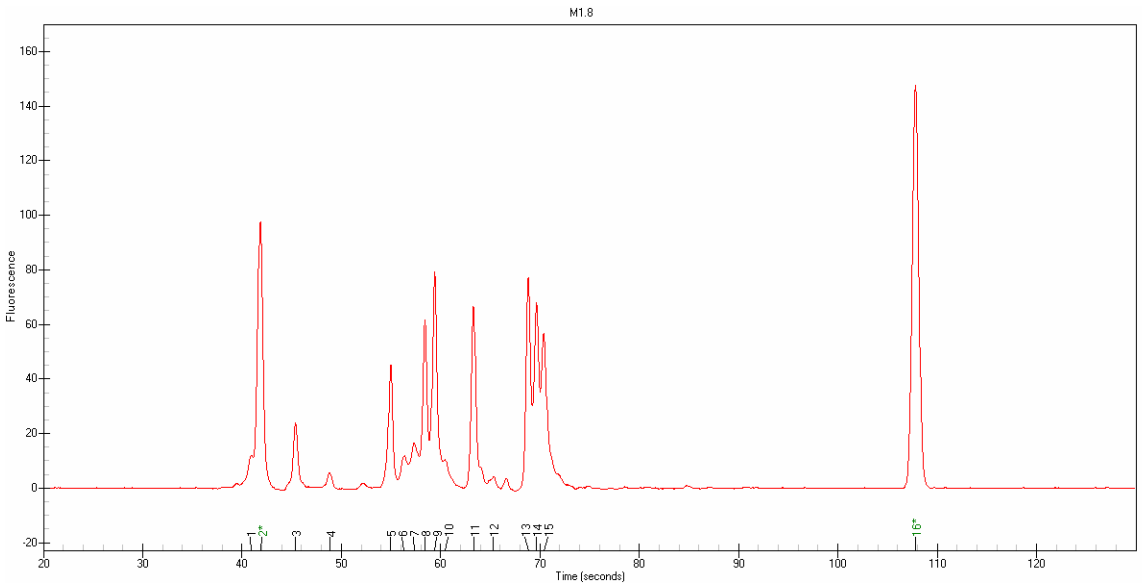


Screen image 11: Screen dump showing the design summary panel containing information about all created selector probes.

Appendix D



Experion result 1: An example of the bump found in a run. The bump starts after the fifth peak.



Experion result 2: An example of a more normal run where the bump is not present.