A MULTIPLEX PRIMER DESIGN ALGORITHM FOR TARGET AMPLIFICATION OF CONTINUOUS GENOMIC REGIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AHMET RAŞİT ÖZTÜRK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF HEALTH INFORMATICS

AUGUST 2016

A MULTIPLEX PRIMER DESIGN ALGORITHM FOR TARGET AMPLIFICATION OF CONTINUOUS GENOMIC REGIONS

Submitted by AHMET RAŞİT ÖZTÜRK in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in The Department of Health Informatics Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın-Son _____
Head of Department, **Health Informatics**

Assoc. Prof. Dr. Tolga Can _____
Supervisor, **Computer Engineering**

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın-Son _____
Health Informatics, Middle East Technical University

Assoc. Prof. Dr. Tolga Can _____
Computer Engineering, Middle East Technical University

Assist. Prof. Dr. Aybar Can Acar _____
Health Informatics, Middle East Technical University

Assoc. Prof. Dr. Özlen Konu _____
Mol. Biol. and Genetics, İ.D. Bilkent University

Assist. Prof. Dr. Mehmet Tan _____
Computer Engineering, TOBB-ETÜ

Date: 01.08.2016

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this wok.

Name, Last name : Ahmet Raşit Öztürk

Signature :

ABSTRACT

A MULTIPLEX PRIMER DESIGN ALGORITHM FOR TARGET AMPLIFICATION OF CONTINUOUS GENOMIC REGIONS

Öztürk, Ahmet Raşit
Ph.D, Department of Medical Informatics
Supervisor: Assoc. Prof. Dr. Tolga Can

August 2016, 89 pages

Targeted Next Generation Sequencing (NGS) assays are cost-efficient and reliable alternatives to Sanger sequencing. For sequencing of very large set of genes, the target enrichment approach is suitable. However, for smaller genomic regions, the target amplification method is more efficient than both the target enrichment method and Sanger sequencing. The major difficulty of the target amplification method is the preparation of amplicons, regarding required time, equipment, and labor. Multiplex PCR (MPCR) is a good solution for the mentioned problems. However, finding compatible multiplex pairs is an example of a clique decision problem in graph theory and it's NP-complete by nature.

We propose a novel method to design MPCR primers for a continuous genomic region, following the best practices of clinically reliable PCR design processes. On an experimental setup with 48 different combinations of factors, we have shown that multiple parameters might effect finding the first feasible solution. Increasing the length of the initial primer candidate selection sequence gives better results, whereas waiting for a longer time to find the first feasible solution does not have a significant impact.

We generated MPCR primer design for the MEFV whole gene; and, our benchmarking experiments show that the proposed MPCR approach is able to produce reliable NGS assay primers for a given sequence in a reasonable amount of time.

Keywords: Next Generation Sequencing, target amplification, Multiplex PCR, primer design

ÖZ

KESİKSİZ GENOMİK BÖLGELERİN HEDEF ÇOĞALTMASI İÇİN MULTİPLEKS PRİMER TASARIM ALGORİTMASI

Öztürk, Ahmet Raşit
Doktora, Tıp Bilişimi Bölümü
Tez Yöneticisi: Doç. Dr. Tolga Can

Ağustos 2016, 89 sayfa

Hedeflenmiş Yeni Nesil Sekanslama (YNS) testleri Sanger sekanslamaya göre maliyet-etkin ve güvenilir alternatiflerdir. Hedef zenginleştirme yaklaşımı çok sayıda gen kümeleri için uygundur. Ancak, daha küçük bölgeler için hedef çoğaltma yöntemleri hedef zenginleştirme ve Sanger sekanslamaya göre daha etkindir. Hedef çoğaltma yönteminin en büyük zorluğu ise gerekli zaman, ekipman ve işgücü göz önüne alındığında amplikonların çoğaltılmaya hazır hale getirilmesidir. Multipleks PCR (MPCR) bahsi geçen problem için iyi bir çözüm oluşturmaktadır. Ancak uygun multipleks çiftlerin bulunması bir klik kararı problemidir ve doğası gereği NP-tam'dır.

Bu çalışmada kesiksiz genomik bölgeler için MPCR primerleri tasarlayan yeni bir yöntem öne sürülmektedir ve klinik açıdan güvenilir PCR tasarımı iyi uygulamaları benimsenmiştir. Birçok faktörün 48 farklı kombinasyonuyla oluşturulan deneysel organizasyon ile farklı parametrelerin ilk makul çözümün bulunmasını etkilediği gösterilmiştir. İlk primer adayı seçme bölgesinin uzunluğunun arttırılmasının daha iyi sonuçlar verdiği, ancak aynı şartlarda daha uzun süre beklemenin ilk makul çözümü bulma açısından anlamlı bir etkisinin olmadığı görülmüştür.

MEFV tüm geni için MPCR primer tasarımı gerçekleştirilmiş ve değerlendirme deneylerimize dayanarak, öne sürülen MPCR yaklaşımının makul bir zaman dilimi içerisinde verilen bir sekans için güvenilir YNS test primerleri üretebildiği gösterilmiştir.

Anahtar Sözcükler: Yeni Nesil Sekanslama, hedef çoğaltma, Multipleks PCR, primer tasarımı

To my wife, **Melike**,

my son, **Ömer Selim**,

my father, **Mehmet**,

and my mother, **Nermin**.

ACKNOWLEDGMENTS

I'ld like to express my appreciation to my supervisor, Assoc. Prof. Dr. Tolga Can, who made the PhD period intellectually joyful for me. I believe that he is the definition of the perfect mentor with his support, encouragement, and endless positive motivation.

I'ld like to thank Assoc. Prof. Dr. Yeşim Aydın-Son for her endless support and making the Health Informatics Department a very entertaining and fruitful place.

I'ld like to thank Assoc. Prof. Dr. Özlen Konu, my MSc supervisor, for being supportive and providing insightful advices.

I'ld like to thank Assist. Prof. Dr. Aybar Can Acar for being in my thesis committee and for his lectures. I'm grateful to him to help me understand the OOP concept comprehensively.

I'ld like to thank Assist. Prof. Dr. Mehmet Tan for reviewing my PhD thesis and being in the committee.

I'ld also like to thank Informatics Institute and Health Informatics Department faculty and staff.

Lastly, I'ld like to thank my wife Melike and my son Ömer Selim for providing great motivation to finish PhD and especially for the times they are not home during the thesis writing :)

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

PCR	Polymerase Chain Reaction
MPCR	Multiplex Polymerase Chain Reaction
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphism
VCF	Variant Calling File
SAM	Sequence Alignment/Map
BAM	Binary Alignment/Map

CHAPTER 1

Introduction

1.1 Aim of the Study

Main purpose of the study is to develop and validate an algorithm that designs Multiplex PCR (MPCR) primers for continuous genomic regions regarding the known pitfalls of primer design in order to be used with the Next Generation Sequencing (NGS) instruments. From a computational perspective, finding multiplex pairs is the problem of finding cliques among the network of all possible primer pair space covering the targeted genomic regions and it is NP-complete.

1.2 General Introduction and Rationale

Advances in Next Generation Sequencing technologies decreased the cost-per-base below Sanger sequencing (Katsanis and Katsanis 2013), leading to an increase for the demand of high-throughput and low cost NGS approaches (Metzker 2010). Despite the overall high cost of Whole Genome Sequencing (WGS), targeted sequencing assays amplifying only selected regions of the genome are developed such as target amplification, target enrichment, and molecular inversion probes (Mamanova et al. 2010; Teer et al. 2010).

Among the targeted sequencing approaches, targeted amplification method is more suitable for smaller genomic regions in order to get a uniform coverage and reliable read quality (Mamanova et al. 2010). In this method, selected genomic regions are first amplified using PCR, then, PCR products are filtered and isolated, and sequenced with a NGS instrument (Gray, Dunlop, and Elliott 2015). A major drawback of the approach is the allele dropout, caused by a SNP in the 3' end of a primer, resulting in low or no amount of expected PCR product. However, this problem can be overcome at the design level by including a primer-binding region in another PCR product (Chong et al. 2014).

In order to automate the process of amplification of a selected genomic region, special instruments, such as RainDance® are required (Orkunoglu-Suer et al. 2015). A good alternative to achieve multiple amplification using conventional PCR is the MPCR method. For example, the consensus transcript of the MEFV gene (ENST00000219596) has 10 exons and 8 of them can be easily sequenced by popular

desktop sequencers like Illumina MiSeq or Ion PGM instruments since the maximum length of the those exons is 357 bps. However the remaining 2nd and 10th exons are 633 and 554 bps, respectively. Since those lengths cannot be read in the desktop sequencers at once, they should be either amplified as shorter fragments or the whole exons should be fragmented using an experimental method, which results in additional experimental steps and more PCR experiments for those regions. However, a multiplex approach does not require additional experimental steps. In addition, costly PCR consumables like the polymerase enzyme are only used in a few tubes regardless of the number of fragments to be amplified. Therefore, sequencing cost of a small gene like the HBB gene and a larger one like the MEFV genes becomes almost the same.

The main limitation of the MPCR approach is the content of the gene itself. For a successful MPCR experiment, there should be as few secondary structures and dimers as possible whereas a feasible solution should be found among a very limited number of possible primer candidate sites (Burgart et al. 1992). To our knowledge, a method for describing the design of MPCR primers for a continuous genomic region following best practices of reliable PCR design to be used in NGS does not exist. In this study, a novel primer design method to amplify targeted genomic regions using a multiplex approach that is suitable to be used in NGS is proposed.

1.3 Computational Challenges

From a computational point of view, finding compatible multiplex pairs is an example of a clique decision problem in graph theory (Eblen et al. 2012). Nodes are primer pairs, and edges are drawn to represent compatible primer pairs. An feasible multiplex solution is a clique (complete subgraph) covering the targeted genomic region.

Current clique enumeration methods can be divided into two categories: iterative enumeration (Kose et al. 2001) and backtracing (Bron and Kerbosch 1973). Iterative enumeration is a breadth-first approach and requires extensive memory usage since all cliques regardless of its size should be stored to find the next clique, which makes longer to find a clique with a given size using a high amount of memory (Kose et al. 2001). This algorithm is suitable for smaller networks whereas it becomes unpractical for networks with a large number of nodes.

Backtracing is a depth-first traversal algorithm (Bron and Kerbosch 1973) and suitable for finding the first clique with any given size sooner than an iterative approach on average, with less memory requirements.

Among clique problems in computer size, multiplex pair finding can be defined as finding a clique with varying size, which is a non-deterministic polynomial (NP) problem. The size of the clique is driven by the possible maximum and minimum number of amplicons covering the targeted regions, given the PCR product and primer length parameters. Therefore, the problem is an NP-complete problem (Downey and Fellows 1995).

1.4 Genetic Testing

There are three major types of genetic testing to investigate genetic abnormalities that cause a genetic disease (Genetic Alliance and District of Columbia Department of Health 2010), each having various advantages and limitations ("Limitations of Cytogenetic Testing" 2016):

Cytogenetic Testing

Karyotyping, Fluorescence in-situ hybridization (FISH), Microarray and MLPA are used to detect structural abnormalities in chromosome structure. Some of the methods are limited by the capabilities of the light microscope whereas others cannot target regions that are unknown to be relevant with the disease. In addition, small abnormalities like Single Nucleotide Polymorphisms (SNP) or micro scale insertions or deletions cannot be detected by this testing methodology ("Cytogenetic Testing Methods | University of Florida Health Pathology Laboratories" 2010).

Biochemical Testing

The purpose of biochemical testing is to detect direct and indirect changes in protein activity, size, or amount. It may not be possible to undercover the underlying genetic abnormality causing the change in protein activity using biochemical tests. In addition, the overlap of biochemical marker levels in normal and carrier populations makes it difficult to identify carriers for some diseases (Pastores and Hughes 1993).

Molecular Testing

Small or single alterations in DNA can be detected using molecular testing methods like PCR, Real Time PCR (RT-PCR), or microarrays. Although it is the most reliable approach to diagnose genetic diseases (Katsanis and Katsanis 2013; Netto, Saad, and Dysert 2003), it is relatively expensive and time consuming with the classical molecular testing methods compared to the others discussed above. In addition, attempts to lower costs through specific mutation panels might result in undetected relevant rare mutations ("Testing Methods" 2016). However, using the NGS technology, a cost-effective and easy-to-implement MPCR assays can be developed, as suggested by this thesis study.

1.4.1 Potential Applications

There are several potential applications of the proposed algorithm to be used for genetic testing. Some of them are outlined below:

Newborn Screening

Newborn screening is utilized in mostly developed countries for early detection of hereditary diseases and therefore aims for either managing the disease to prevent symptoms or increase the quality and duration of life (Pourfarzam and Zadhoush 2013). Almost all of the newborn screening tests are conducted using a tandem mass spectrometry instrument with a dried blood spot from the neonatal (Yoon 2015).

Commonly, a two stage algorithm is applied for testing newborns: at the first stage, a test with a low-cost and low-specificity is made. Then, babies with positive test results are subjected to a high-cost and high-specificity test. Introduction of an MPCR-based NGS testing would decrease the steps of testing to one and increase the quality of screening to a diagnosis-level in both sensitivity and specificity (Bell et al. 2011). Cost-effectiveness of the proposed method is mentioned in the next section.

Carrier Screening

Carrier screening is mostly utilized for genetic disease risk assessment of parents planning to have a baby (Yao and Goetzinger 2016). Methods similar to newborn screening are performed for carrier screening and parents are advised in case of a risk of having a baby with a certain genetic disease. However, the scope of carrier screening is not as broad as newborn screening. Again, using the proposed method, the utility and scope of carrier testing could be extended cost-effectively.

Diagnosis of Genetic Diseases

Since the proposed method is capable of producing diagnosis-quality results, it can be used for diagnostic purposes (Muzzey, Evans, and Lieber 2015). As shown in the literature, the reliability of NGS with certain quality measures is compatible with Sanger sequencing. Several rare genetic diseases can be diagnosed and new genetic tests can be developed with little cost.

Predictive Testing

Although still controversial (Janssens et al. 2006), predictive testing for complex diseases like Alzheimer's or several cancer types for certain ethnic groups can be performed using the methods explained in this study. However, from a cost point of view, there are limitations on the size of gene panels that can be generated with an MPCR-based method. For example, cost of a probe hybridization technique is almost similar to the MPCR-based method for approximately 80 genes. In addition, sample DNA amount required for the amplification of 80 or more genes using the MPCR-based method is very high compared to probe hybridization approach.

1.4.2 Brief economical and clinical utility analysis

According to a report in the US ("Universal Newborn Screening for Cystic Fibrosis in Connecticut" 2006), cost of diagnosing Cystic Fibrosis, a common genetic disease worldwide, in a three stage approach is around $3.5. With the utilization of the proposed approach, the overall cost of screening and diagnosing babies with Cystic Fibrosis in single-stage would be around $3.0 in large scale with a custom NGS setting. Diagnosis of any certain genetic disease would also cost similar in large scale, and would not exceed $30 per test for a low-scale diagnosis volume. Compared to the cost of recently introduced gene panels for hereditary diseases (Fecteau et al. 2014), it would be still feasible to test up to 80 genes at the same time using the proposed method.

1.4.3 Current limitations of the translation of the method

Since the proposed MPCR based method is applicable through an NGS instrument, there are detection limitations of certain genomic abnormalities. Currently, the maximum length of a read using a desktop sequencer is 400 to 600 nucleotides for different platforms. Sharing similar limitations to Sanger sequencing, it is not possible to diagnose diseases like Fragile-X Syndrome since the total length of repeats for a mutated gene (Saul and Tarleton 1993) exceeds the read length of a desktop sequencer. In addition, to confirm large deletions and insertions, cytogenetic testing methods should be performed (Mahdieh and Rabbani 2013). However, as the read-lengths with reliable quality increases, the method can be easily adapted to the new instruments such as the NanoPore technology (Feng et al. 2015).

1.5 Multiplex Polymerase Chain Reaction

The Multiplex Polymerase Chain Reaction (MPCR) method performs more than one Polymerase Chain Reaction (PCR) in a single tube at the same time (Burgart et al. 1992). Thus, multiple genomic regions can be amplified in vitro in a single protocol. As the number of PCR reactions increase, the complexity of interaction would result in inhibition of some or all PCRs. Therefore; an MPCR requires special calculations and design (Thornhill and Snow 2002).

1.5.1 Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) enables in vitro amplification of a specific genomic region using a special DNA polymerase directed by primers (Saiki et al. 1988; Valones et al. 2009; Garibyan and Avashia 2013). Primer is a short single strand oligonucleotide that hybridizes to a complementary DNA region and it creates an attraction site for DNA polymerase at a certain temperature. Then, DNA polymerase starts adding bases to the primer until it is detached from DNA due to the thermodynamic events. Since the hybridization of the primer to its target is the key for targeting a genomic region and its binding to its target triggers polymerization, design of the primer has a crucial role. Primers are named according to their relative position; the one on the plus DNA strand is called a forward primer, and the other one on the minus DNA strand is called a reverse primer.

DNA is a double stranded molecule and requires primers in both strands. This leads to an exponential amplification and strict limitation to the targeted region. After the first few cycle of reactions, amplified DNA fragments become the main template. Primers are consumed in each cycle as they are incorporated into the amplified DNA fragments.

PCR is a thermodynamically driven reaction and there are three temperatures and associated time periods that need attention:

Denaturation

Template double stranded DNA should be denatured into single stranded DNA in order to allow binding of primers at a later step. Although 95 °C is suggested for separating two strands, initial time periods need to be between approximately 30 to 150 seconds depending on the complexity of template DNA. Increased initial denaturation period is also used to activate engineered polymerases that are otherwise inactive, which is a beneficial feature to prevent unnecessary polymerase activity during experimental setup. In each additional cycle, DNA needs to be denatured although as low as 10 seconds is enough as suggested by polymerase manufacturers.

Annealing

Despite having a perfect complementary base order to the template DNA, primers may not bind efficiently to their targets in non-optimal temperatures. Therefore, special thermodynamic calculations are necessary to ensure the maximal hybridization efficiency of primers. Salt concentration in the tube might severely effect the optimum annealing temperature, or temperature of melting (Tm) in other words. There are several studies focused on annealing temperature calculation regarding the salt concentrations and base orders (Schildkraut and Lifson 1965; SantaLucia 1998; Owczarzy et al. 2004). An above-optimum temperature leads to less hybridization efficiency of a primer whereas a below-optimum temperature results in non-specific binding.

Extension

Following the hybridization of primers to the template DNA, temperature is increased to the optimum working temperature of the DNA polymerase used. At this step, primers are elongated by adding complementary bases using the template DNA as the reference. Depending on the properties of the polymerase, duration and temperature might differ. After the completion of extension, the same sequence of temperature cycles follow until the final cycle, which is suggested usually between 20-40 cycles, depending on the expectations of the researcher. The number of cycles might be lower in order to visualize the results of the PCR experiment in agarose gel in a shorter time, or cycles might be more to amplify the targeted regions as much as possible.

1.5.2 PCR consumables

Consumables and their concentrations also have an important role in the PCR process. Consumables and their concentration can be determined depending on the aim of the study, quality expectations and cost constraints, etc.

DNA polymerase

DNA Polymerase is an enzyme responsible for recognizing the double to single stranded DNA border created by a perfectly hybridized primer, adding complementary bases to the primer, maintaining stability during the polymerization, and proofreading, if applicable.

There are several DNA polymerases in the market. The ones that are of high interest for the study should have proofreading capability to ensure very low error rate during the amplification and high stability during the reactions. Additionally, engineered polymerases with heat-inactivation property are useful especially for the initial denaturation period; since the polymerases also do not loose their activity under long periods of high temperature, complex DNA templates like human genomic DNA denatures very well and allows efficient binding of primers in the later cycles. DNA polymerase concentrations are suggested by manufacturers; whereas, slight changes do not have severe impacts on reaction efficiency.

Primers

Since primers are responsible for directing DNA polymerase to the borders of intended amplification sites, their design has a crucial role for the success of the PCR. If a primer binds to more than one intended target site, the concentration of the primer is split into the each hybridization site, resulting in less attraction and unnecessary relocation of DNA polymerase to multiple sites. This can be due to the poor design of primers not regarding other possible binding regions, or miscalculated annealing temperature. The first problem can be overcome by performing a BLAST query, and the latter one can be overcome by incorporating salt concentrations into the calculation of annealing temperature.

Interactions between primers also have an impact on PCR efficiency. Since primers potentially target other single stranded DNA fragments, they cannot not distinguish the targeted fragment from another primer as a template. If a primer hybridizes to itself, it is called a homodimer. Hybridization of a primer to another primer is called a heterodimer. In addition, one end of a primer may bind to the other end of itself. In this case, it is called a hairpin. In either case, the concentration of the primer decreases on its targeted site that results in a weak reaction or a loss of reaction at the end.

Primer concentration also has an interesting effect: as the concentration of primers increase, shorter products are favored. In the opposite case, amplification of longer oligomers is expected, if any (Amiram et al. 2011). Therefore, careful optimization of primer concentration is a must.

Lastly, the length of a primer is also closely related to the specificity of the amplification reaction. Since there are errors each time a base is added in the manufacturing process, which is called coupling efficiency, long primers might contain truncated oligomers, which binds to undesirable locations on the template and decreases reaction efficiency (“DNA Oligo FAQ” 2016). Therefore, additional purifications might be necessary for certain applications.

Mg^{++} concentration

Mg^{++} is the cofactor of DNA polymerase and higher concentrations lead to increased DNA polymerase activity. On the other hand, this gain has a side effect: specificity is decreased (Roux 2009). Free dNTPs also bind to Mg^{++} and change its concentration in the test tube. Another effect of Mg^{++} concentration is on the annealing temperature: it increases the hybridization temperature. Having multiple effects on reaction dynamics,

Mg^{++} has a central role among PCR consumables. Careful consideration of its concentration might increase the PCR efficiency whereas incautious additions might result in a partial or complete loss of reaction.

dNTP

Free dNTP is consumed each time a DNA polymerase extends the primer into an oligomer. Therefore, there should be sufficient dNTPs in the reaction to ensure adequate amplification. On the other hand, as mentioned above, excess concentrations of dNTP might result in the occupation of Mg^{++} ions, which results in decreased DNA polymerase enzyme efficiency.

Buffer

Various chemicals might be added to the reaction to ensure better or selective DNA amplification. A few examples of these chemicals are DMSO, Glycerol, KSO_4, and BSA (Reaction buffer composition for nucleic acid replication with packed DNA polymerases 2016). PCR Kit manufacturers usually add one or more of these to get optimized reaction efficiency.

1.5.3 Considerations on Multiplex Polymerase Chain Reaction Design

Major difference of MPCR from PCR is the amount of targeted sites to be amplified. As a result, more than one pair of primers is added. The side effects are:

Primer cross reactions

As the number primers in the same tube increases, the chance of cross-reaction between different primers increases. Heterodimers between primers have two effects on the reaction efficiency:

1) If hybridization occurs on 3' end of any primer, DNA polymerase also binds and amplifies the heterodimer complex, thus reducing the overall reaction efficiency.

2) Regardless of the hybridization position, single stranded primer concentration decreases, thus reducing the overall reaction efficiency.

To prevent such unwanted interactions, each pair of primers should be tested with the others. Since each primer should be tested with itself, total number of calculations required is proportional to the square of the number of primers in the reaction.

DNA polymerase concentration

Since the DNA polymerases will be allocated to multiple sites among the template DNA, their concentration will greatly reduce depending on the number of multiplexing. Therefore, sufficient amount of DNA polymerase should be supplied in the reaction.

dNTP concentration

In case of high number of multiplexing, dNTP will be depleted very quickly. Therefore, sufficient dNTP should exist in the reaction. On the other hand, since free dNTPs bind to Mg^{++} ions, the final concentrations should be carefully determined.

Mg^{++} concentration

As the dNTP concentration increases, there should be sufficient Mg^{++} ions in the solution. However, increasing the Mg^{++} concentration too much might result in unspecific amplification and increased annealing temperature, resulting in loss of reaction efficiency.

Amplification sites with close proximity

Since the borders of an amplification site is determined by forward and reverse primers, there will be unwanted amplification sites in a multiplex PCR. The forward primer of the first amplification site and the reverse primer of second amplification site might identify the borders of another undesired amplification region. In this case, DNA polymerase will also amplify that region. If the amplification sites are close enough, the extension time period is long enough, and the binding stability of DNA polymerase is high enough, undesired DNA fragments will be amplified in the test tube. This will decrease PCR efficiency by depleting and/or occupying consumables. Two ways to prevent such a formation are either to make the extension period shorter or adjust the primer concentration.

1.5.4 Principles of primer design for MPCR

Primer design is a well-studied subject and researchers suggest several principles for a successful primer design. Although not all of might be necessary for a PCR reaction, considering all the principles below is important for a multiplex PCR design due to the increased complexity of the reaction. In this thesis, we consider the following factors in the proposed MPCR design algorithm.

Primer GC content

GC content negatively (Buck et al. 1999; Frey et al. 2008) effects the hybridization efficiency especially when it is abundant on high and low ends. Therefore, GC content is suggested to be kept near 50%.

Primer length

Considering the human genome, if 19-mer oligonucleotides were spread on the genome randomly, there will be 1.09% chance to find two similar 19-mer oligonucleotides in the same genome. Therefore, the suggested minimum primer length would be 19. For the maximum number, one should consider the coupling efficiency, which is the rate of success each time after adding a base in the manufacturing process. If a primer is 35 bases long, the coupling efficiency will be 70%. In other words, 30% of the delivered primer solution will contain truncated

oligomers, which would result in undesired amplification or reduced reaction efficiency. Primer manufacturers also provide purification options to increase purity. However, this decision should be made regarding the initial primer cost of ownership.

Primer specificity

In order to direct DNA polymerase to desired amplification sites, it should be carefully checked whether the primers potentially bind to other regions on the genome (Kennedy 2011). The BLAST sequence search tool can be used to check undesired hybridization of primers, which would increase the specificity and reaction efficiency.

Product length

Main limitation on the product length interval is the instrument of use. If the sequencer has a read limit of 500 nucleotides, the length of amplification sites including the primers should not be kept more than that number. Although there is no minimum number for an amplicon length, in order to reduce the number of primer pairs required to sequence a region, it should be better to keep the minimum amplicon size 250 bp or higher.

Mismatches in primer binding sites

Mismatches at 3' end of primer binding sites result in kinks in those regions and greatly reduce the stability of DNA polymerase initial binding (Lefever et al. 2013). The major source of unpredicted mismatches is the Single Nucleotide Polymorphisms (SNPs). Therefore, avoiding primer design containing a SNP inside the 3' end of the primer would increase the success rate of PCR.

Homopolymers in primers

Consecutive repeats of the same bases are called homopolymers. Homopolymers of 4 bases or more negatively effect reaction efficiency and therefore should be avoided if possible (Buck et al. 1999).

Annealing temperature

Above and below optimum primer annealing temperature results in either no amplification or non-specific amplification, respectively. Therefore, primers in the same tube should have very similar annealing temperatures for an efficient amplification of targeted DNA fragments (Kennedy 2011).

Hairpins

Hairpin formation has an inhibitory effect since it reduces annealing efficiency of primers to the template DNA (Kennedy 2011). Therefore, stable hairpins should be avoided during primer design.

Homodimers

Partially or fully hybridization of two primers having the same sequence is called a homodimer. Homodimers result in reduced primer concentration and unnecessary relocation of DNA polymerase if the hybridization occurs at 3' end (Kennedy 2011). Therefore, they should be avoided as well.

Allele dropout

Despite considering known SNPs for primer design, there might be unexpected mutations near the 3' end of a primer that might block primer binding to its target. As a result of a loss of a primer binding, no amplification occurs (Thornhill and Snow 2002). This phenomenon is called allele dropout and extra effort (like sequencing the inclusive region again) is required to reveal the mutation that caused mismatch. A tiling amplification design helps to avoid this: according to the design principle, every amplicon includes at least one reverse and forward primers of another amplicon, except the very first and last primers. Thus, such a mutation would be easily identified and the remaining area would be sequenced separately by skipping the mutation.

1.5.5 Current MPCR methodologies and comparison

Although the proposed method in this dissertation is novel with its continuous multiplexing capabilities, there are several studies in the literature that developed MPCR-based tools and assays in order to benefit from the multiplexing approach. They can be grouped into three categories:

Profiling

In this category, MPCR assays are utilized to identify individual categories from others. The groups are mostly different species (Rodríguez et al. 2015; Romero-Pastrana and Romero-Pastrana 2012; M. Li et al. 2016; Shi et al. 2016), different ethnic groups (Eduardoff et al. 2016) or patients with different prognosis (Raj et al. 2016). In those studies, genomic regions that are specific for each group are selected and primers are designed using tools like Primer3 (Untergasser et al. 2012, 3) and then experimentally validated to check their performance. Singleplex PCR is performed first, and multiplex experiments are conducted after that.

Sequencing

In this group of studies, the aim of the study is to sequence different regions of the genome simultaneously for mutation screening (López et al. 2007; Wang et al. 2015; X. Li et al. 2016). True multiplexing is achieved in one study for the sequencing of DMD in 10-plex PCR (Stockley et al. 2006) whereas their approach is specific to DMD since all of its exons are smaller than 500 bp and are distantly located. In another study, all exons of a gene are amplified in 5-plex PCR and then sequenced in an NGS instrument (Poon, Tan, and Koay 2016). However, again, their approach is not systematic and is not applicable to other genomic regions.

Automated Design Tools

PrimerStation, a web-based multiplex PCR primer design tool, is capable of calculating computationally expensive cross hybridization thermodynamics whereas it is only intended for the multiplex amplification of discrete genomic regions (Yamada, Soma, and Morishita 2006).

Another tool called MPrimer designs multiplex PCR primers using a graph-expanding algorithm and shows its utility by 5-plex assay design for DMD gene (Shen et al. 2010). However, as stated above, multiplexing DMD does not require multiplex PCR primer design for a continuous region and this approach does not have a practical utility for most of the genes, which have at least one exon with a length of 500 bases.

Recently, an epigenetic research focused tool for multiplex PCR primer design tool with an experimental success rate of 71% is introduced in the literature (Pandey et al. 2016). The tool takes many parameters into account like SNP locations and CpG islands; whereas, its major goal is enrichment of large genomic regions at the same time. Therefore, extra experimental steps are necessary to shear DNA to make it ready for a sequencer or microarray instrument.

Some of the commercial primer design tools also have multiplex primer design capability. However, considering their interfaces and explanations, none of them has the ability to generate multiplex PCR primer for continuous regions ("Primo" 2016; "PrimerDigital: Biotechnology Solutions" 2016).

There is one algorithm worths mentioning for its multiplexing capacity for continuous genomic regions, hence allowing multiplex PCR primer design for whole gene or whole exome sequencing assays. Belgian biotechnology company Multiplicom has developed an algorithm called Multiplexer, which is used for the generation of their multiplex assays for whole exome sequencing ("Gimv, VIB and University of Antwerp Invest EUR 2.0 Million in Molecular Diagnostics Startup Multiplicom" 2016; "Multiplicom's MASTR Technology" 2016). However, their method is very similar to AmpliSeq and a two-step amplification is used for multiplexing.

There are other primer design tools in the literature which not capable of performing multiplex primer design (You et al. 2008; Ye et al. 2012; "UCSC In-Silico PCR" 2016).

In vitro assays

A method called Universal Multiplex PCR is introduced that avoids cross-inhibition of multiple primers in the same tube (Wen and Zhang 2012; Xu et al. 2012; Universal primer sequence for multiplex DNA amplification 2016). A similar approach called AmpliSeq is utilized by ThermoFisher through an acquired patent (Methods and systems of nucleic acid sequencing 2016). In addition, a microarray-based MPCR is introduced in a study (Y. Li et al. 2011). However, major drawback of the assay is the initial cost of microarray manufacturing. Universal Multiplex PCR assay has the potential for sequencing long genomic regions; whereas, the primer design algorithm does not exist for that approach. Despite its ease of use, AmpliSeq is designed for

single tube multiplexing and therefore lacks the ability to cover a continuous genomic region without any gap or allele dropout risk.

The contributions of this study can be summarized as follows:

1. A novel method for MPCR is proposed. The proposed methods is better than the state of the art methods for MPCR in the following aspects: it is capable of producing primer designed for a whole continuous genomic region of more than 500 bp, it considers several design parameters simultaneously, and takes allele dropout into consideration.

2. We tested and validated the proposed method on several genes and exons, as well as experimentally demonstrated its application on the MEFV gene.

3. Our proposed method considers multiple tubes and multiple flanking regions that have significant effect on the elapsed time to get the first feasible solution.

CHAPTER 2

Algorithm for Continuous Primer Design

2.1 Problem definition

DNA is a double stranded polar molecule. In order to determine the order of bases in a fragment of DNA, the amount of the targeted region should be amplified using a special technique called Polymerase Chain Reaction (PCR) (Saiki et al. 1988). In PCR experiments, short oligonucleotides (primers) that have the complementary sequence to the start and end positions of the targeted DNA fragment determines the boundary of the amplified region called a PCR product. The forward primer determines the start position of the region to-be-amplified. The reverse primer determines the end position of the targeted fragment and binds to the complementary strand of the template DNA. A pair of forward and reverse primers is called a primer pair for a specific PCR product. Theoretically, multiple targeted DNA regions can be amplified at the same time and this technique is called Multiplex PCR (MPCR) (Chamberlain et al. 1988). However, primer-primer interactions, primer-PCR product interactions, formation of inhibitory secondary structures, or thermodynamically favored side products prevent efficient amplification of multiple targeted DNA regions in the same tube. With careful consideration of possible interactions and their thermodynamic properties, it is possible to avoid these issues and conduct a successful MPCR experiment.

2.2 Constraints in MPCR primer design

Below are the constraints of the problem:

1) The length of a PCR product in current sequencing technologies acceptable for diagnostic use is usually limited to 500 bases. Also, for practical purposes, it should not be less than 300 bases.

2) Primers should be long enough for a specific hybridization to the targeted genomic region, but it should not be very long in order to reduce the cost of production and secondary structure formation tendency. The interval of primer length should be limited to 23 to 30 bps for optimum length.

3) To avoid non-specific PCR products, designed primers should only bind to the target region and nowhere else. Thus, each designed primer should be checked

for alternative binding regions through a BLAST search against the targeted genome.

4) Variations in the DNA sequence of individuals are heterogeneous in terms of type and genomic location. An unexpected variation in the last 3 bases of a primer results in a weakened binding of the primer to its target region in the DNA template, resulting in the formation of low PCR product concentration. Therefore, there should not be a known variation in the last three bases of a designed primer. Thus, there should not be a known SNP in the last three bases

5) Total number of G and C bases divided by the total length of an oligonucleotide gives the GC rate of given sequence. Optimum GC rate of a primer is 50%, and it should not be more than 70% or less than 30%.

6) Secondary structure formation inhibits PCR and decreases the yield of PCR products. Thus, it should be avoided when possible. Interactions between primers (either homo or heterodimers) and hairpins (self-hybridization of an oligonucleotide forming a loop structure) should not be thermodynamically favored, and their ΔG value should be more than -3 Kcal.

7) Annealing temperature is defined as the ideal temperature for the formation of a stable primer-DNA template complex. Annealing temperature of each designed primer should be very close to each other, within a difference of 0.5 °C, and annealing temperature of each primer should be within 0.5 °C of the specified optimum temperature.

8) There should not be 4bp-long or longer homopolymers in the primer.

9) Due to a phenomenon called allele-dropout resulted by variations in the 3' end of a primer, each primer region should be included in another PCR product except the first and last primers for the targeted whole DNA fragment. Therefore, MPCR primer pairs should be split into at least two test tubes so that there should be no overlapping and undesirable primer products in the same test tube.

2.3 Formulation of the MPCR design problem as a graph problem

The MPCR primer design problem can be formulated as a graph problem, with primer pairs meeting the primer design criteria as nodes in the graph and with edges between two primer pairs if they meet the interaction constraints. Among a set of feasible candidate primer pairs, a subset meeting the requirements of a complete graph can be placed in the same test tube. For a successful design, 1) there should be at least two or more cliques where their PCR products meet the constraints and 2) primer pairs in those cliques should cover the targeted DNA region. This problem corresponds to finding a clique in the graph with a varying size and is an NP-Complete problem described by Downey (Downey and Fellows 1995). The solution time to find the best primer pair design is exponential with respect to the target region length, and there are no known efficient solutions for this problem. Network enumeration to discover cliques can be classified as iterative enumeration (breadth-first traversal) and backtracing

(depth-first traversal) whereas backtracing is more practical in time and memory requirements to find the first feasible clique (Eblen et al. 2012). Therefore, a depth-first heuristic approach (Bron and Kerbosch 1973) is implemented to find the first solution that meets the given constraints since all optimum solutions meeting the criteria are experimentally acceptable.

2.4 The Proposed Method

Regarding the problem definition and constraints, finding suitable primer pairs is a tree search problem in the space of optimum primer pairs. Due to the exponential complexity of the problem, a depth-first approach is favored to find an acceptable solution within reasonable amount of time with less memory usage. The rules for designing primer pairs are given as follows:

- Leftmost forward primer should be in the first n bases of the given sequence.
- Position of the rightmost reverse primer should be in the last n bases of the given sequence.
- Next PCR product should be in a different test tube.
- Pos(Forward tube n mod m, k) < Pos(Reverse tube n-1 mod m, k)
- Pos(Forward tube n mod m, k) > Pos(Reverse tube n-2 mod m, k)
- Pos(Reverse tube n mod m, k) > Pos(Reverse tube n-1 mod m, k)
- Pos(Forward tube n mod m, k) > Pos(Reverse tube n mod m, k-1)

where

- Pos(Forward tube n mod m, k) denotes the position of the first base of the k-th forward primer in the test tube n regarding a total of m test tubes and
- Pos(Reverse tube n mod m, k) denotes the position of the last base of the k-th reverse primer in the test tube n regarding a total of m test tubes

As a result, designed primers should be in the following order:

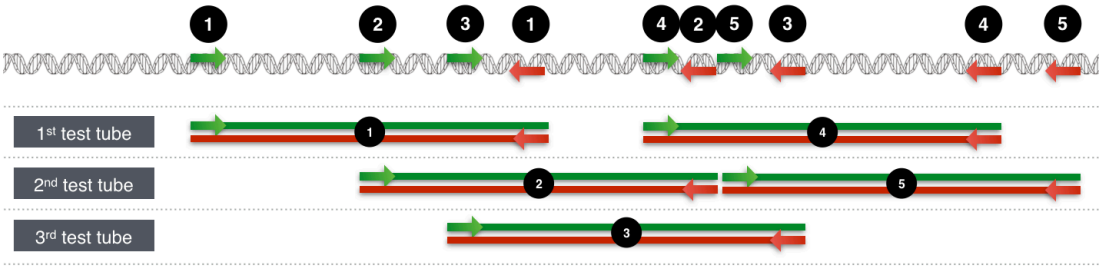

Figure 1. Example of primer and amplicon positions regarding the rules of proposed algorithm. Rightward arrows denote forward primers, leftward arrows denote reverse primers, numbers show the order of amplicons

where, rightward arrows denote forward primers, leftward arrows denote reverse primers, numbers show the order of amplicons. In addition to these rules, each primer pair in the same tube should also satisfy the design constraints stated in Section 2.2.

2.5 Implementation

The algorithm is divided into three parts for optimization purposes. The first part is responsible for primate candidate filtering, the second part is responsible for finding compatible forward and reverse primer pairs, and the third part aims for finding compatible multiplex primer pairs.

2.5.1 Primer candidate identification

In this step, there are 7 out of 9 constraints to follow in order to find feasible primer candidates: primer size interval, specificity using BLAST, SNP position control, GC content, self and cross interactions, annealing temperature difference, and the length of homopolymers. All of the possible primer candidates are created using a sliding window with given primer sizes. The number all possible primers can be calculated as the following:

$$\sum_{exon\ 1}^{exon\ n} length(exon\ i) + 2 \times length(flanking\ region) - length(primer)$$

Then, each sequence is tested for the constraints. After that, a map of feasible primer positions is visualized to check if the genomic region is suitable. This visualization serves two purposes: debugging and genomic region characterization. As will be discussed in the Discussion Chapter, not all genomic regions are equal.

Constraints for the current implementation is as follows:

Table 1. Constraints for primer candidate identification.

Parameter	Inclusive interval or value
Primer size interval	23 – 30 bases
Specificity check using BLAST	< 3 hits (up to one mismatch or one gap)
SNP position control	no SNPs in the last 3 bases (3' end)
GC rate	30% - 70%
Hairpin ΔG	> -3 Kcal
Homodimer ΔG	> -3 Kcal
Heterodimer ΔG	> -3 Kcal
3' end hybridization ΔG	> -3 Kcal
Optimum annealing temperature	60 °C
Annealing temperature difference	0.5 °C
Maximum allowed length of a homopolymer	3 bases

A sample primer candidate map of last 4 exons of the MEFV (Figure 2) gene is shown in Figure 3 and Figure 4 regarding the parameters in Table 1. Since the lengths of introns between the last four exons of MEFV gene are smaller than the flanking region length limit, they are merged into a single genomic region.

Figure 2. Exon and introns of MEFV transcript ENST00000219596.

Exon and intron lengths of the MEFV transcript is given below:

Table 2. Exon and intron lengths of MEFV transcript ENST00000219596.

Exon/Intron	Length in bases
Exon 1	317
Intron 1-2	1520
Exon 2	633
Intron 2-3	4377
Exon 3	350
Intron 3-4	426
Exon 4	96
Intron 4-5	1662
Exon 5	231
Intron 5-6	468
Exon 6	23
Intron 6-7	1936
Exon 7	116
Intron 7-8	186
Exon 8	33
Intron 8-9	361
Exon 9	33
Intron 9-10	165
Exon 10	1667

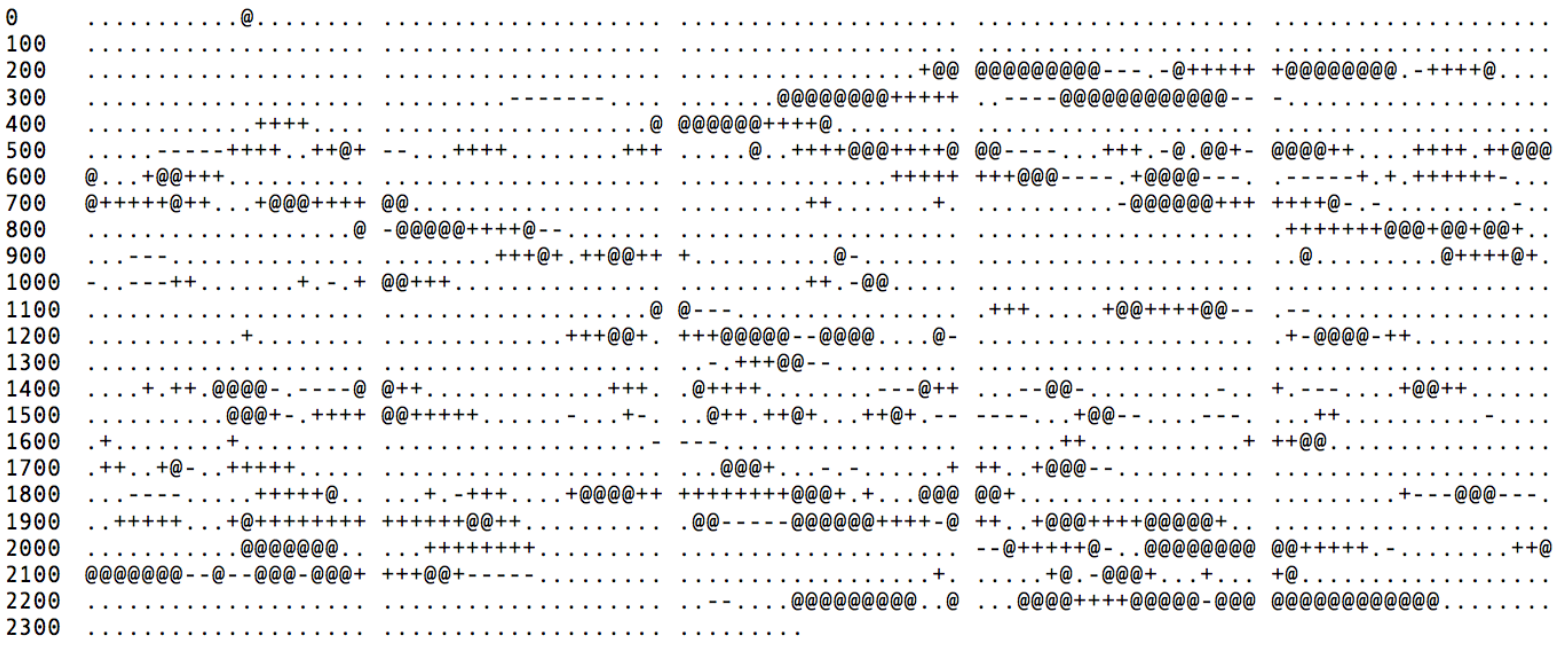

Figure 3. Primer candidate map of last four exons of MEFV gene with 450 bases flanking regions. Sequence is take from (+) strand. If a suitable primer is a forward, its leftmost position is denoted with “+” sign. If a suitable primer is a reverse, its leftmost position is denoted with “–“ symbol. If a position is the leftmost of both a suitable forward and reverse primer, it’s denoted with “@” symbol. Otherwise, “.” symbol is used for unsuitable positions for a leftmost position of a primer candidate.

The same map is generated for the (-) strand also:

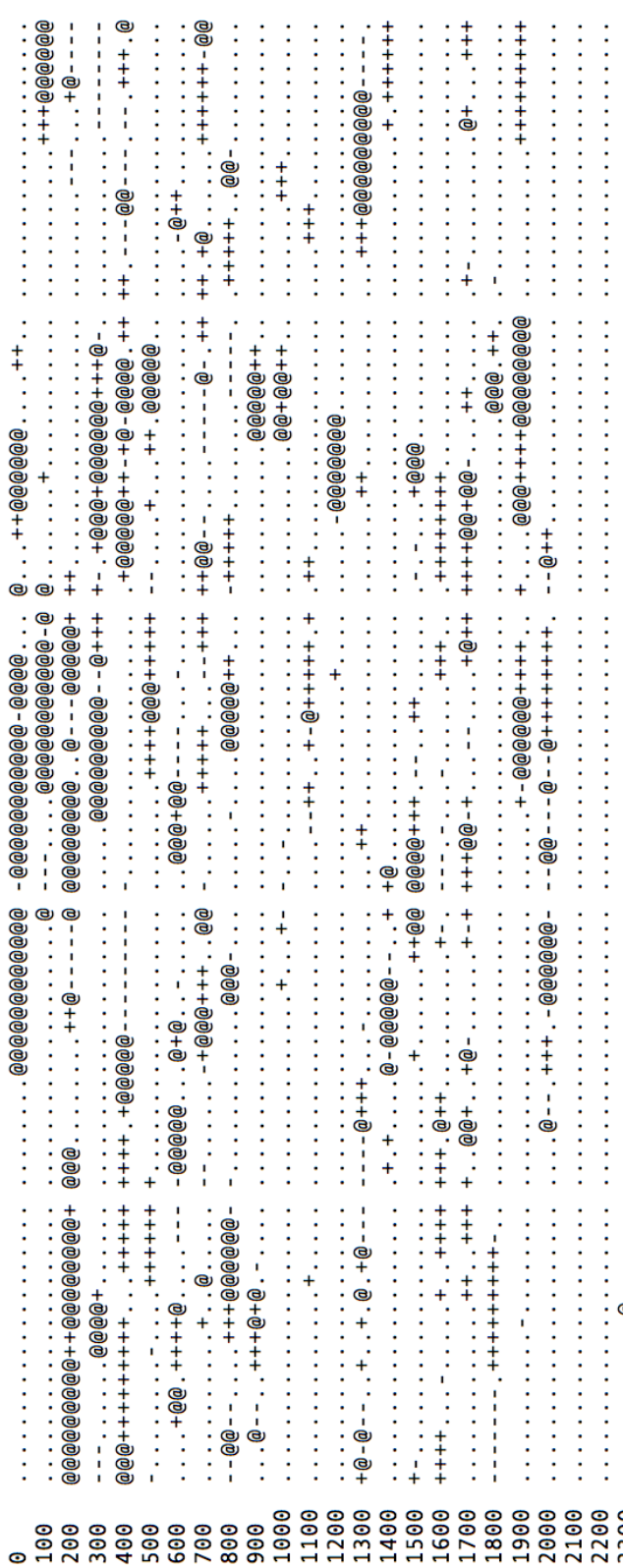

Figure 4. Primer candidate map of last four exons of MEFV gene with 450 bases flanking regions. Sequence is take from (-) strand. If a suitable primer is a forward, its leftmost position is denoted with "+" sign. If a suitable primer is a reverse, its leftmost position is denoted with "–" symbol. If a position is the leftmost of both a suitable forward and reverse primer, it's denoted with "@" symbol. Otherwise, "." symbol is used for unsuitable positions for a leftmost position of a primer candidate.

For an imaginary genomic region of 30 bases, the same map would be something like that:

```
@+--++-...  -.+-.+.--.  +.-.+.....
```

Figure 5. Primer candidate map of an imaginary 30 bases long genomic region.

The leftmost position of a forward primer is denoted by “+”, leftmost position of a reverse primer is denoted by “-“, and the leftmost position of both forward and reverse primers are denoted by the “@” symbol. There are 8 forward and 9 reverse primers in this example. The same sequence can be also shown as a graph. However, at this step, there are only nodes as the primers. Edges are added in the next step.

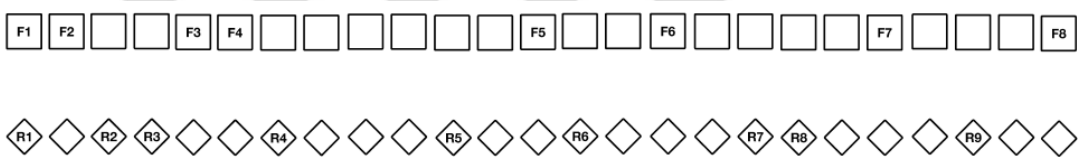

Figure 6. An example sequence is shown as a network of primers. The upper part is used to represent forward primers, and the lower part of the same sequence is used to represent reverse primers.

2.5.2 Primer pair identification

The next step is finding the compatible primer pairs among candidate primers. Each forward and reverse primer is checked for the others, and compatible pairs are kept as compatible primer pairs. The process can be visualized as an empty matrix initially; then, compatible pairs are marked on this matrix.

	F1	F2	F3	F4	F5	F6	F7	F8
R1								
R2								
R3								
R4								
R5								
R6								
R7								
R8								
R9								

Figure 7. Empty matrix of primer candidates.

For compatibility test, 2 out of 9 constraints are considered: self and cross interactions and amplicon size. For each parameter, the same matrix is created. Constraints for the current implementation are shown below:

Table 3. Constraints for primer pair identification

Parameter	Inclusive interval or value
Amplicon size	300 – 500 bases
Heterodimer ΔG	> -3 Kcal
3’ end hybridization ΔG	> -3 Kcal

The same matrix after cross interaction check is filled as above:

	F1	F2	F3	F4	F5	F6	F7	F8
R1	✗	✗	✗	✗	✗	✗	✗	✗
R2	✗	✗	✗	✗	✗	✗	✗	✗
R3	✗	✗	✗	✗	✗	✗	✗	✗
R4	✓	✓	✓	✗	✓	✓	✗	✗
R5	✗	✓	✓	✓	✗	✗	✗	✗
R6	✗	✓	✓	✓	✓	✓	✗	✗
R7	✗	✓	✓	✗	✓	✓	✗	✗
R8	✗	✗	✗	✗	✓	✗	✗	✗
R9	✗	✓	✓	✗	✓	✓	✗	✗

Figure 8. Compatible primers after cross interaction check.

Main limitation after thermodynamic calculations is now the length of amplicons. The example sequence after amplicon size filtering becomes a network like the one below in Figure 9:

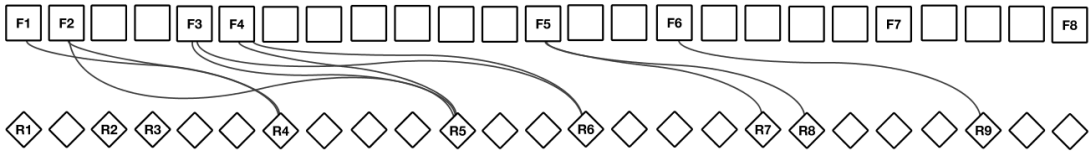

Figure 9. Compatible primers after thermodynamic and amplicon size filtering

Compatibility matrix is a sparse matrix since a lower percentage of primers can be compatible. The next step is finding the set of compatible multiplex pairs.

2.5.3 Multiplex pair identification

At this step, compatible primer pairs are again checked for each other to find the ultimate set of multiplex pairs. A depth-first based backtracing algorithm is utilized to find the feasible clique. Now, primer compatibility matrix turns into a pair compatibility matrix:

	F1-R4	F2-R4	F2-R5	F3-R5	F3-R6	F4-R5	F4-R6	F5-R7	F5-R8	F6-R9
F1-R4		✗	✗	✗	✗	✗	✗	✗	✗	✗
F2-R4			✗	✗	✓	✗	✗	✓	✗	✓
F2-R5				✗	✓	✗	✗	✗	✗	✗
F3-R5					✗	✗	✗	✗	✗	✗
F3-R6						✗	✗	✓	✗	✓
F4-R5							✗	✗	✗	✗
F4-R6								✗	✗	✗
F5-R7									✓	✓
F5-R8										✗
F6-R9										

Figure 10. Compatible pairs after cross interaction check.

Again, the matrix is a sparse matrix. At this step, the matrix can be visualized as a network. As explained in the problem definition, the optimum solution is a clique covering the targeted region..

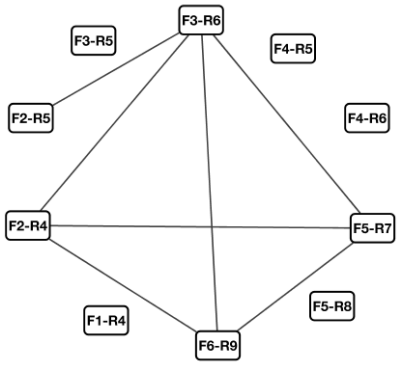

Figure 11. Network of compatible pairs after cross interaction check.

Amplification pattern of the solution is shown below:

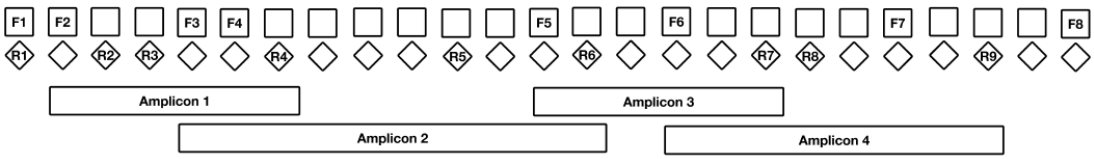

Figure 12. Primer positioning of a feasible solution.

2.5.4 Overview of the procedures

In summary, the following route is followed to find the feasible solution:

1. Firstly, all k-mer oligonucleotides in the targeted region are checked for multiple criteria explained in section 2.2 and candidate primers are identified.

2. In the next step, suitable primer pairs are assessed. Distance and cross-interaction parameters are taken into consideration for the assessment.

3. In the third step, all identified suitable primer pairs are checked against each other in terms of cross-interaction capability and multiple criteria supplied in section 2.4. Therefore, a network of all primer pairs that can work with another primer pair is created.

4. In the last step, the problem is to find the first clique covering the targeted DNA region. This is a feasible solution since all the primer pairs can work in harmony and the whole targeted region can be amplified.

Since any feasible solution can be accepted, the first feasible solution is considered as a working MPCR primer pair set.

CHAPTER 3

In silico validation

In order to test the proposed method with genomic data, the method is implemented and an in silico test scheme is created.

3.1 Implementation of the algorithm

Algorithm is implemented in Python 2.7. Major factors to choose Python are its object oriented approach, wide variety of modules including a well written thermodynamic calculation module, and a large user community.

A local BLAST server is installed and human genome database (GRCh38) is set up in order to perform BLAST queries fast. Multiprocessing ability of local BLAST server is utilized for every query.

dbSNP human build 146 is used as a reference for the sequences of known SNPs. Each chromosome sequence is indexed for a faster query.

For thermodynamic calculations, primer3 module is utilized. Thermodynamic calculations in Primer3-based tools are performed using the principles in this module (Untergasser et al. 2012).

A primer pair and a multiplex pair classes are implemented and well-optimized.

Since BLAST queries and thermodynamic calculations are CPU-intensive processes, each query or calculation is stored in a database and retrieved instead of a real-time query or calculation. Classical SQL databases are slow for the read and write queries. Therefore, noSQL databases based on memory are utilized. Redis database is utilized for data storage. However, TCP based communication become the bottleneck after many rounds of optimization. Therefore, all prior calculations are first cached into memory as a Python dictionary when the implemented script started, and the only transaction between Redis database is reduced to writing new data after that. Optimization is conducted using a Python module called cProfile, which gives diagnostic benchmarking for each function used in the script.

Functions for primer candidate selection was optimized considering the consumption of CPU time. The order of filtering functions to identify a primer candidate is as follows, regarding the CPU consumption amount and starting from the fastest:

1. Check GC rate
2. Check Homopolymers
3. Calculate annealing temperature
4. Check hairpin
5. Check homodimer
6. Check 3' end hybridization
7. Check SNP
8. Check BLAST

Multiprocessing is also heavily used when applicable. As mentioned, BLAST is utilized on multiple processes at the same time. In addition, a master script is written to distribute multiplex finding process such that each thread deals with an exon or exon group at the same time. Although this does not have a practical value for smaller genes or genes with very few exons like HBB (3 exons) or with very short introns CYP21A2 (10 exons, but should be processed as it's a single continuous genomic region), it provides a significant benefit for genes like CFTR (27 exons) or DMD (79 exons).

Another optimization approach that had a great improvement effect was randomly testing only a portion possible primer pairs for finding multiplex pairs. Due to the nature of the problem, any feasible solution can be utilized. Therefore, instead of waiting until the end of the calculations for all possible space of potential solutions, each time a subset of pairs are selected and they are checked against other pairs. This step is repeated until the first feasible solution is found.

All calculations are performed either on Apple iMac with 2.93 GHz Intel 4-Core i7 and 16GB RAM desktop computer or an Apple MacPro with 2.7 GHz 12-Core Intel Xeon E5 processors and 64 GB RAM workstation. When a time-based test is conducted, the specifications of the computer are given in the text.

3.2 Test approach

The implemented algorithm is tested for three different cases.

3.2.1 Human exon sequences

In the first case, human exon sequences with a length between 2000 to 2100 bases are selected using the Ensembl BioMart MartView interface including upstream and downstream flanking sequences, 240 bases for each. In the test, elapsed time until the first feasible solution is recorded. Three factors are evaluated: 1) the order of candidate primers in terms of base position for a given sequence interval, 2) the effect of initial primer candidate area length, since it changes the number of starting forward primer candidates, either 120 or 240 bases, and 3) the time limit required to find a feasible solution, either for 240 or 480 seconds. In total, 48 different test cases are benchmarked. This test is conducted on the Apple iMac with the specifications mentioned above.

3.2.2 A sample set of genes related with genetic diseases

In the second case, genes with a clinical utility are selected and the algorithm is tested whether it can generate multiplex primer pairs in 24 hours. This test is conducted on the Apple MacPro with the specifications mentioned above. This test is performed for the genes HBB, CFTR, SERPINA1, HEXA, BRCA1, and BRCA2.

3.2.3 The MEFV gene

In the last test case, multiplex primer pairs for MEFV gene is designed using the algorithm. In addition, found primers are experimentally validated in order to see the utility of the algorithm in a real case. Details of the experimental validation procedures are given in the next chapter. Results of in silico and in vitro tests are given in detail in Chapter 5.

CHAPTER 4

In vitro validation protocols

In addition to in silico test cases to show the theoretical utility of the proposed method, experimental validation is required to show the practical utility of the developed algorithm. In order to fulfill this goal, three different sets of experiments were conducted. Each experimental approach has its own advantages and disadvantages. In this chapter, these validation approaches will be explained in detail and their results will be given in the next chapter.

4.1 Singleplex PCR experiments

The aim of singleplex PCR experiments is to test each primer pair individually and observe their working performance qualitatively through an agarose gel.

4.1.1 Advantages and limitations

Main advantage of the classical PCR experiments is to understand the efficiency of primers using a very basic set of consumables and equipment. On the other hand, it is a very time consumable process for multiplex assays with larger number of primer pairs. To test the methodology, this is the first step of validation. However, after developing more assays, conventional PCR becomes the second step to confirm the amplification of primer pairs that work inefficiently.

4.1.2 Validation

Bio-Rad iProof High Fidelity PCR Kit 200U (Cat. No: 172-5331) is used for PCR experiments. Genomic DNA is isolated from blood using QiaGen DNeasy Blood & Tissue Kit (Cat. No: 69504) following the protocol supplied with the product.

In order to validate primer pairs for the first time, the following experiment protocol is followed, as suggested by the manufacturer:

Table 4. PCR experiment protocol.

Component	Stock concentration	Volume for 20 µl reaction
5X iProof HF Buffer	-	4 µl
dNTP Mix	10 mM	0.4 µl
Forward Primer	10 µM	1 µl
Reverse Primer	10 µM	1 µl
DNA Template	50 ng/µl	1 µl
ddH_2O	-	11.6 µl
iProof DNA Polymerase	0.4U/µl	1 µl

PCR instrument protocol for Bio-Rad T100 Thermal Cycler is as follows:

Table 5. PCR instrument protocol.

Cycle Step	Temperature	Time	Cycle
Initial denaturation	98 °C	1 min	
Denaturation	98 °C	10 s	
Annealing	63 °C	30 s	30
Extension	72 °C	30 s	
Final Extension	72 °C	10 min	

After the experiments, PCR products are run on 2% agarose gel. Gels are imaged using Bio-Rad Gel Doc EZ instrument.

4.2 Multiplex PCR experiments

Two different kinds of MPCR experiments are conducted: the first one for confirmation of the design and the second one for sequencing.

4.2.1 Advantages and limitations

There are two major advantages of MPCR experiments: consuming less DNA polymerase enzyme and using the workforce more efficiently. Also it is valuable for assay validation: success of the assay design can be visualized in a few hours with minimal amount of consumables.

Although ultimate goal of the proposed methodology is to design multiplex primers, confirmative visualization of multiplex primers is a cumbersome process. Since agarose gel does not have enough resolution to identify close bands, the multiplex primer pairs should be further separated into more tubes regarding the product size.

4.2.2 Validation

NEB Multiplex PCR 5X Master Mix (Cat. No: M0284S) is used for MPCR experiments. Genomic DNA is isolated from blood using QiaGen DNeasy Blood & Tissue Kit (Cat. No: 69504) following the protocol supplied with the product.

The following experiment protocol is followed for all MPCR experiments, as suggested by the manufacturer:

Table 6. MPCR experiment protocol.

Component	Stock concentration	Volume for 20 µl reaction
Multiplex PCR 5X Master Mix	-	4 µl
Primer Coctail	5 µM each	5 µl
DNA Template	50 ng/µl	1 µl
ddH_2O	-	10 µl

PCR instrument protocol for Bio-Rad T100 Thermal Cycler is as follows:

Table 7. PCR instrument protocol.

Cycle Step	Temperature	Time	Cycle
Initial denaturation	95 °C	1 min	
Denaturation	95 °C	20 s	
Annealing	63 °C	1 min	35
Extension	68 °C	1 min	
Final Extension	68 °C	5 min	

After the experiments, PCR products are run on 3.5% agarose. Gels are imaged using Bio-Rad Gel Doc EZ instrument.

4.3 NGS experiments

The proposed method is developed to design MPCR assays that can be sequenced using an NGS instrument. Therefore, sequencing the designed assay is the ultimate validation of the proposed algorithm.

4.3.1 Advantages and limitations

Despite the singleplex sequencing property of widely used Sanger sequencing technology, NGS enables sequencing of mixture of DNA sequences at the same time. In addition, cost per base is very low compared to Sanger sequencing.

On the other hand, typical run cost of an NGS sequencing round is quite expensive and should be multiplexed enough in order to be cost-effective. In addition, maximum

read length is strict and cannot be optimized in contrast to Sanger sequencing. Therefore, maximum read length limits must be considered. This is a significant constraint in MPCR primer design.

4.3.2 Validation

Sequencing is performed on Illumina MiSeq instrument with Illumina MiSeq Reagent Nano Kit v2 500 cycle (Cat. No: MS-103-1003) following the protocols of the instrument and the kit as suggested. Genomic DNA is isolated from blood using QiaGen DNeasy Blood & Tissue Kit (Cat. No: 69504) following the protocol supplied with the product.

After the MPCR, PCR products are quantified using Invitrogen Qubit® 3.0 Fluorometer instrument and Qubit® dsDNA BR Assay Kit (Q32850). Each tube is diluted to 80ng/µl using ddH_2O.

For the library preparation step, NEBNext® Ultra™ DNA Library Prep Kit for Illumina (Cat. No: E7370S) is utilized. For multiplexing of samples, NEBNext® Multiplex Oligos for Illumina® (E7335S) is used during the library preparation. Protocols supplied with the products are followed.

Data is acquired in FASTQ format, and the data analysis up to variant calling file (VCF) is conducted in-house using SAMtools (H. Li 2011). VCF files are supplied in Appendix A.

CHAPTER 5

Results

The results part is composed of two parts: the results of in silico tests from Chapter 3 and the results of in vitro experiments from Chapter 4.

At first, in order to get an overview of the potentials of the algorithm, we investigated whether the algorithm is applicable to large genomic regions. During this in silico experiment, effects of flanking regions, elapsed time, and the selection order of initial primer sets were investigated. In the second in silico experiment, the implemented algorithm is applied to a sample set of six genes that have clinical utility. In the third in silico experiment, a multiplex primer design is conducted on the MEFV gene. Mutations in the MEFV gene cause the most prevalent genetic disease in Mediterranean region: familial Mediterranean fever ("OMIM Entry - # 249100 - FAMILIAL MEDITERRANEAN FEVER; FMF" 2016). Therefore, ability to design a multiplex assay has a crucial potential for the translation of the method to clinical studies.

After passing the in silico experiments with success, in vitro experiments are performed. In this case, there are three major experiments: singleplex PCR experiments to confirm that each primer pair works alone, multiplex PCR experiments to confirm the multiplex ability of the primer pairs and to amplify intended regions for sequencing, and the sequencing experiment to confirm the actual amplification of MPCR experiment using an NGS instrument.

5.1 In silico validation results

As described in Chapter 3, the proposed algorithm is subjected to three different in silico tests. Below are the results of these three test cases.

5.1.1 Human exon sequences

The effectiveness of a multiplex target amplification experiment depends on the following factors: 1) avoiding undesired secondary structure formation, 2) uniformity of annealing temperature of primers, 3) GC content of primers, 4) avoiding single nucleotide polymorphisms (SNPs) in the 3' end of primers, and 5) uniqueness of genomic regions which would reduce non-specific binding of the primers to other

regions other than the target site. The proposed method takes these factors into account and designs robust primers for given target sites.

Although all of the factors can be calculated, finding an acceptable solution depends mostly on the primers in initial primer candidate set, which are derived from the flanking region just before the targeted exon. Another factor that might effect the performance is the selection order of candidate primers for a given sequence interval. For example, using a forward primer very close to the targeted exon might result in lower number of tubes and less primer pairs whereas selecting the forward primer at the beginning of a flanking region might increase the number of pairs, which will increase the complexity of finding compatible primer pairs.

Data used in this test is human exon sequences limited to lengths between 2000 to 2100 nucleotides retrieved from Ensembl through Biomart Martview interface (Smedley et al. 2009). There are two main data categories that are named according to the length of flanking regions and the maximum time period allowances.

Short dataset

Upstream and downstream flanking regions in this dataset are limited to 120 bases for each exon.

Long dataset

Upstream and downstream flanking regions in this dataset are limited to 240 bases for each exon.

The performance of the algorithm is tested against each dataset. The performance criterion is the percent of exons that a multiplex primer design solution with given time limit can be successfully found. In order to test the effect of time limit in addition to the effect of the length of flanking region, the categories are broadened into three. These are called *Group A* categories for simplicity:

A1:Short240

In this category, short dataset is used and the time limit to find the first feasible multiplex primer pair solution for given exon is set to 240 seconds.

A2:Short480

In this category, short dataset is used again and the time limit to find the first feasible multiplex primer pair solution for given exon is set to a longer period: 480 seconds.

A3:Long240

In this category, long dataset is used to observe the effect of a longer flanking region. The time limit to find the first feasible multiplex primer pair solution for given exon is set to 240 seconds.

Another question of interest is the candidate primer selection order by base position. As explained above, it can increase or decrease the multiplex primer pair number per exon. Therefore, four additional categories are included in the test. These are called Group B categories. In each of these tests, a primer is selected from the pool of primer candidates:

B1:Both primers in leftmost selection order

It will be chosen starting from the leftmost position. Named shortly in the figures as *bothNormal*.

B2:Forward primer in rightmost and reverse primer in leftmost selection order

Forward primer will be chosen starting from the rightmost position and reverse primer will be chosen starting from the leftmost position. Named shortly in the figures as *fwdReverse*.

B3:Forward primer in leftmost and reverse primer in rightmost selection order

Forward primer will be chosen starting from the leftmost position and reverse primer will be chosen starting from the rightmost position. Named shortly in the figures as *revReverse*.

B4:Both primers in rightmost selection order

It will be chosen starting from the rightmost position. Named shortly in the figures as *bothReverse*.

One more question to be asked is the effect of the number of tubes for each multiplex primer pair design effort. In order to achieve this goal, another category (Group C) is added:

C1: Multiplex in 2 tubes

Aim of this test is to find first feasible solution for the specific dataset in given time in order to generate multiplex pairs in 2 tubes.

C2: Multiplex in 3 tubes

Aim of this test is to find first feasible solution for the specific dataset in given time in order to generate multiplex pairs in 3 tubes.

C3: Multiplex in 4 tubes

Aim of this test is to find first feasible solution for the specific dataset in given time in order to generate multiplex pairs in 4 tubes.

C4: Multiplex in 5 tubes

Aim of this test is to find first feasible solution for the specific dataset in given time in order to generate multiplex pairs in 5 tubes.

In order to test all possibilities within the scope of the questions asked, there should be 48 test cases as the combinations of three categories: 4 A Category test X 3 B Category Test X 4 C Category Test. The combinations of test cases can be displayed on a 2-dimensional table as below:

Table 8. Combination of test cases for human exon sequences.

Test Cases	A1	A2	A3
B1	A1,B1,(C1-C4)	A2,B1,(C1-C4)	A3,B1,(C1-C4)
B2	A1,B2,(C1-C4)	A2,B2,(C1-C4)	A3,B2,(C1-C4)
B3	A1,B3,(C1-C4)	A2,B3,(C1-C4)	A3,B3,(C1-C4)
B4	A1,B4,(C1-C4)	A2,B4,(C1-C4)	A3,B4,(C1-C4)

During the test, not all of the given multiplex design attempts resulted in a feasible solution within the limited time. Success criterion is defined as the occurrence of the feasible solution for the targeted genomic region in defined time period. For a given set of exon data and time constraints, the percentage of exons with feasible solutions is given as success rate for that category. However, success rates show differences in each case. Success rates for Short240, Short48, and Long240 test batches are shown in the following figures, respectively.

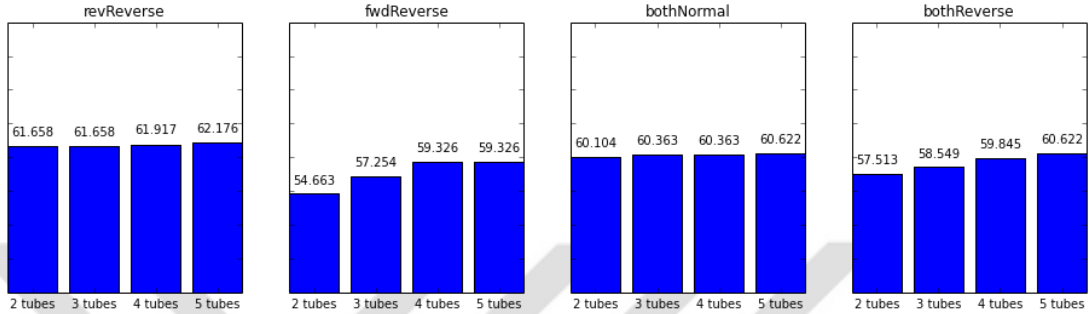

Figure 13. Group B and Group C tests for Short240 test category, success rates.

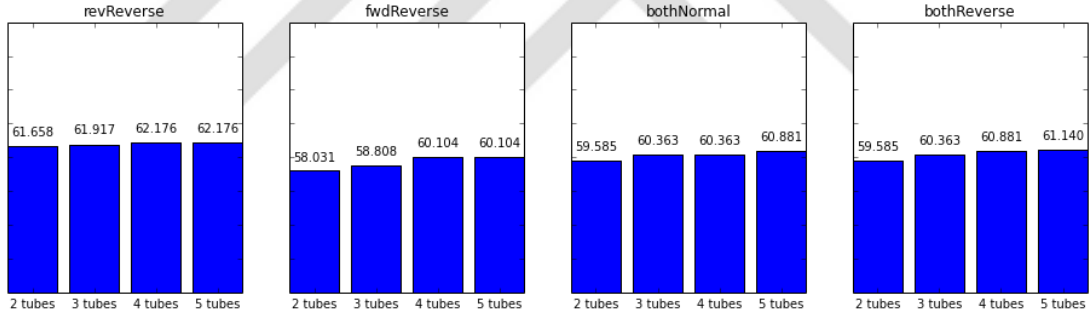

Figure 14. Group B and Group C tests for Short480 test category, success rates.

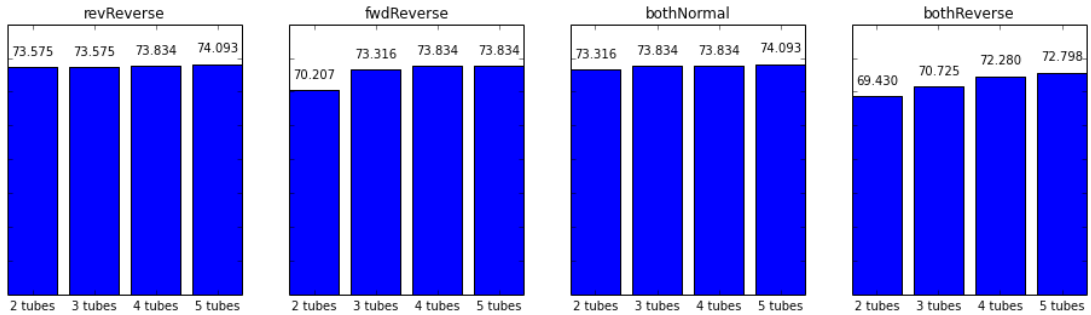

Figure 15. Group B and Group C tests for Long240 test category, success rates.

Figure 13 and Figure 14 show that increasing the time limit does not increase the success rate significantly (p-value = 1). However, Figure 15 clearly shows that increasing the initial primer candidate sequence length have a dramatic effect on success rates (p-value = 0.033) since the initial primer candidate space harshly restricts the space of overall feasible solutions.

The number of multiplex tubes used is another restriction on getting more successful solutions in limited time. In all test case groups, 2-tubes per amplification has the worst success rates (Figures 13-15). However, increasing the number of tubes from 3 to 5 does not have a significant time gain to get the first feasible solution for revReverse and bothNormal test cases (Figures 16-18) (p-value = 0.299 and p-value = 0.545, respectively).

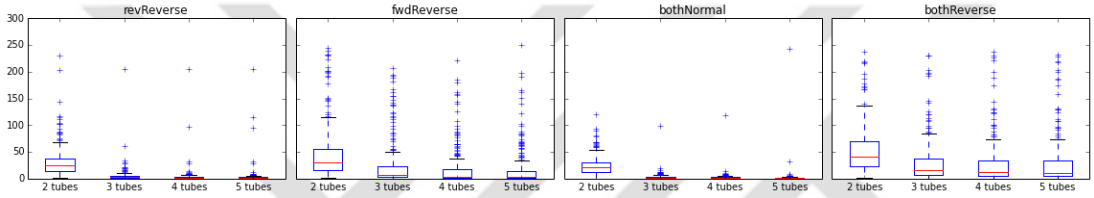

Figure 16. Elapsed time in seconds to find the first feasible solution in Group B and Group C tests for Short240 test case.

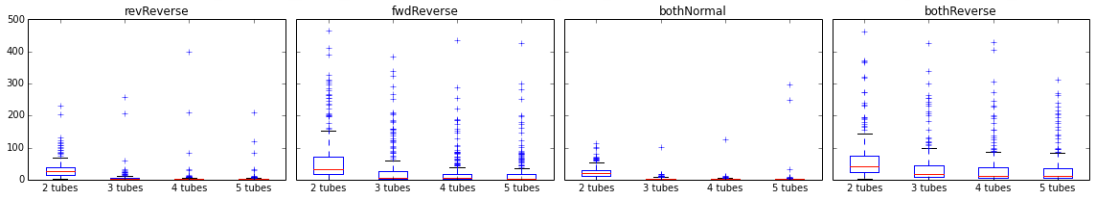

Figure 17. Elapsed time in seconds to find the first feasible solution in Group B and Group C tests for Short480 test case.

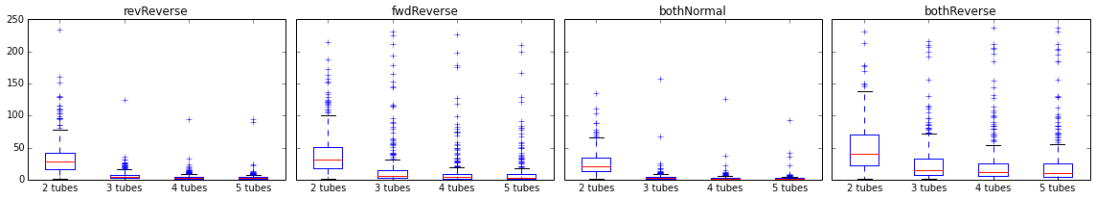

Figure 18. Elapsed time in seconds to find the first feasible solution in Group B and Group C tests for Long240 test case.

Regarding the order of primer candidate selection each time for the same candidate sequence area, there are different factors that effect the performance of the method. revReverse and bothNormal test cases provide favorable results compared to fwdReverse and bothReverse test cases in all tests (Figures 19-21).

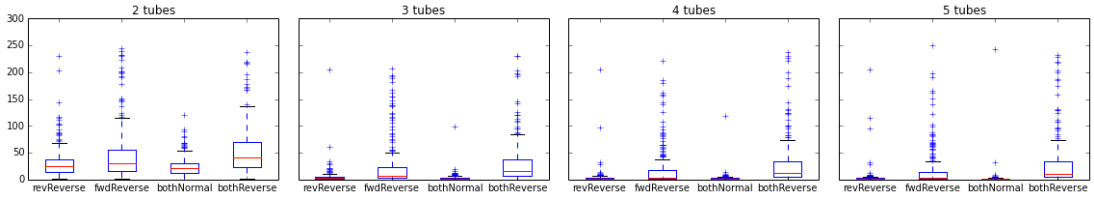

Figure 19. Elapsed time in seconds to find the first feasible solution in Group B and Group C tests for Short240 test case, grouped by tube number.

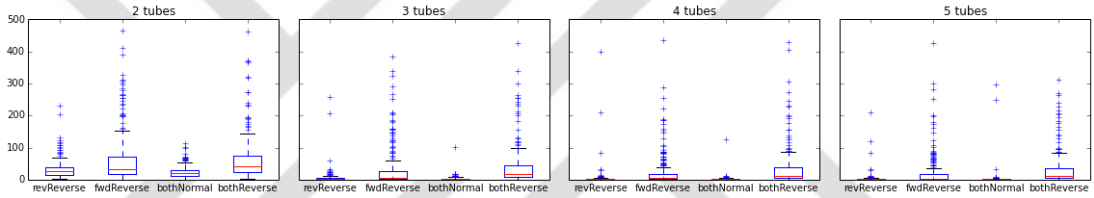

Figure 20. Elapsed time in seconds to find the first feasible solution in Group B and Group C tests for Short480 test case, grouped by tube number

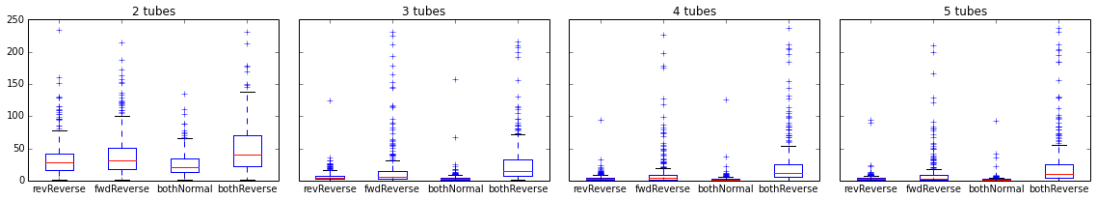

Figure 21. Elapsed time in seconds to find the first feasible solution in Group B and Group C tests for Long240 test case, grouped by tube number

Lastly, it is observed that the number of primer pairs found for each multiplex primer solution is also affected by the order of candidate primer selection. bothNormal primer candidate selection order provides the lowest number of primer pairs for each solution, regardless of the number of tubes, time limit, or initial sequence length (Figures 22-24).

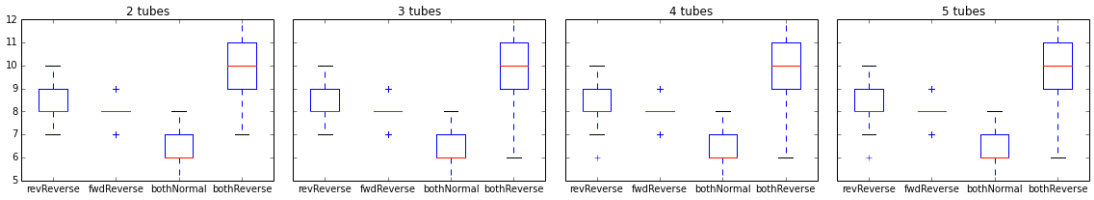

Figure 22. Number of multiplex primer pairs of the first feasible solution in Group B and Group C tests for Short240 test case.

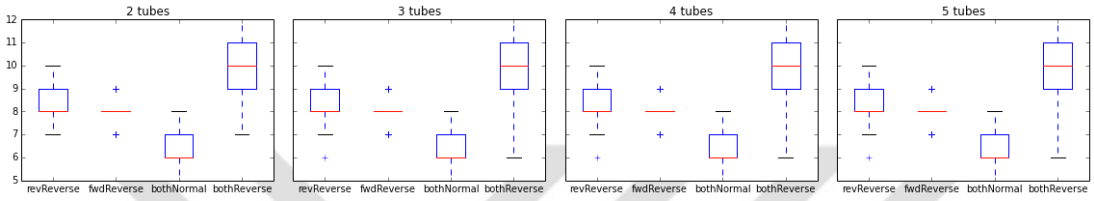

Figure 23. Number of multiplex primer pairs of the first feasible solution in Group B and Group C tests for Short480 test case.

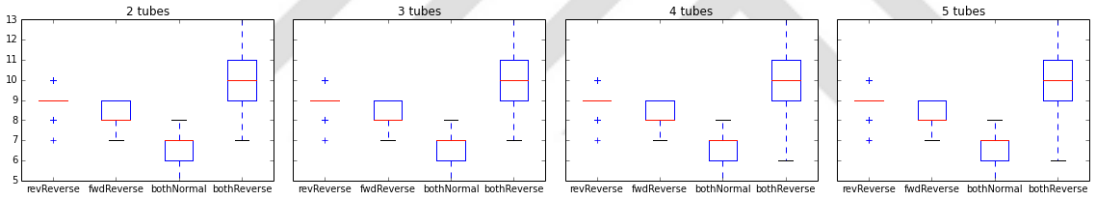

Figure 24. Number of multiplex primer pairs of the first feasible solution in Group B and Group C tests for Long240 test case

5.1.2 A sample set of genes related with genetic diseases

The purpose of thesecond set of *in silico* tests is to test the implemented algorithm with respect to its ability to design multiplex primer pairs for selected genes in a given time. Time limit is set as 24 hours per gene for practical purposes. Selected genes are HBB, CFTR, SERPINA1, HEXA, BRCA1, and BRCA2. Mutations in those genes plays a central role for the development of the following genetic diseases, respectively: Beta-thalassemia (Cao and Galanello 2010), cystic fibrosis (Moskowitz et al. 2008), alpha-1-antitrypsin deficiency (Fregonese and Stolk 2008), Tay-Sachs disease (Mcginniss et al. 2002), and hereditary breast cancer (Fackenthal and Olopade 2007).

Multiplex PCR primers are successfully designed within the given time frame for all exons of genes of interest, for two tubes (2-plex) each. Number of primer pairs for each gene is given below:

Table 9. Summary of the properties of selected genes.

Gene	Length of transcript	# of coding exons	Average GC	# of SNPs	# of found primer pairs
HBB	628 bp	3	51 %	576	7
CFTR	4443 bp	27	42 %	2644	39
SERPINA1	1257 bp	4	52 %	530	7
HEXA	1590 bp	14	52 %	800	17
BRCA1	5592 bp	22	41 %	2996	44
BRCA2	10257 bp	26	36 %	4315	63

Locations of multiplex primer pairs for each gene are mapped below on each figure. In ech figure, the bar on the top represents the targeted genomic region of the gene, the number on the left shows the start of the relative genomic position, the number on the right shows the end of the genomic position, regarding the sequence in Ensembl database. First row of primer pairs are from the first tube, and the second row of primer pairs are from the second tube. Forward and reverse primers are shown in color, and the PCR product in between are shown in light gray.

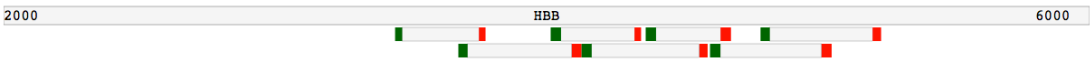

Figure 25. Multiplex primer pairs for HBB gene.

Figure 26. Multiplex primer pairs for CFTR gene.

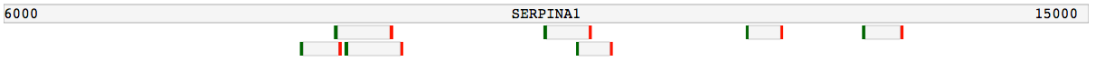

Figure 27. Multiplex primer pairs for SERPINA1 gene.

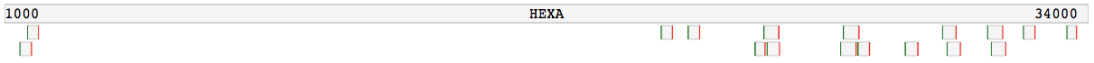

Figure 28. Multiplex primer pairs for HEXA gene.

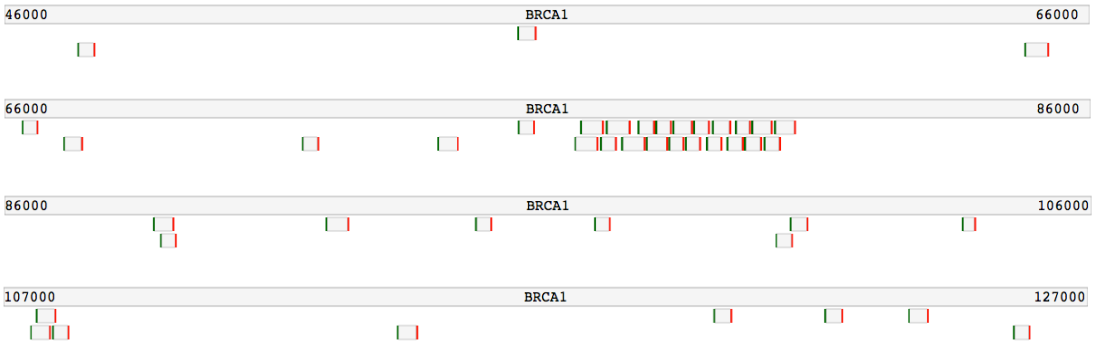

Figure 29. Multiplex primer pairs for BRCA1 gene.

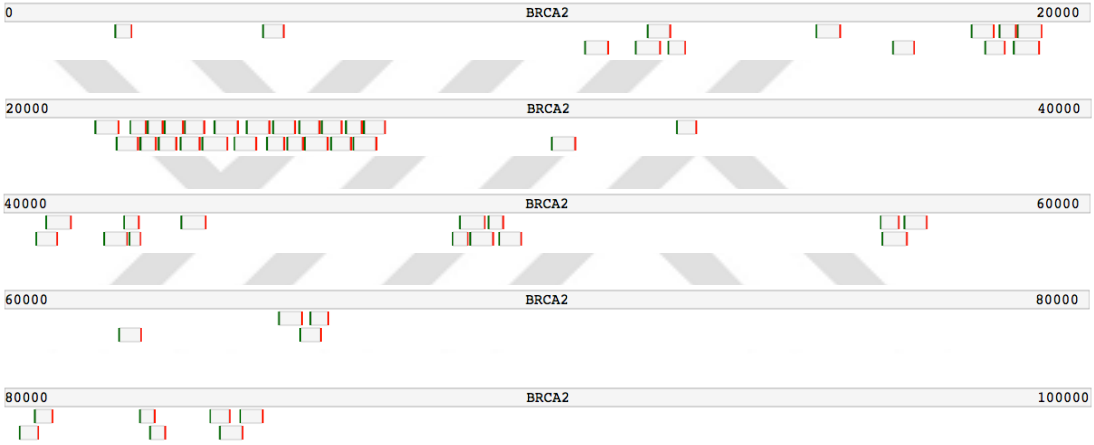

Figure 30. Multiplex primer pairs for BRCA2.

As a result, it is shown that the algorithm can design multiplex primers for a given gene in a reasonable time frame.

5.1.3 The MEFV gene

Using the same approach above, a multiplex pair is successfully designed for the exons of the MEFV gene. 2-plex primer design for the exons of MEFV resulted in 18 primer pairs.

With this last test, in-silico tests are successfully completed and the utility of the algorithm is shown theoretically. In the next step, in vitro validation of the MEFV primer design is performed.

5.2 In vitro validation results

In order to test the clinical utility potential of the algorithm, in vitro experiments are conducted using the protocols given in Chapter 4.

5.2.1 Singleplex PCR

36 primers (18 primer pairs) are ordered from LGC Biosearch Technologies Inc., Petaluma, CA, USA as lyophilized and suspended into two types of stocks: 10 µM stocks for each primer (a total of 36 stocks) and a mixture of primers (primer coctail) (a total of 2 stocks) with a final 5 µM concentration for each primer in the solution. In each primer cocktail, there are 18 primers (9 primer pairs). These primer coctails are named as A and B. Experiments are conducted following the protocols given in the previous chapter and repeated twice. Places of each primer pair on MEFV gene is shown below:

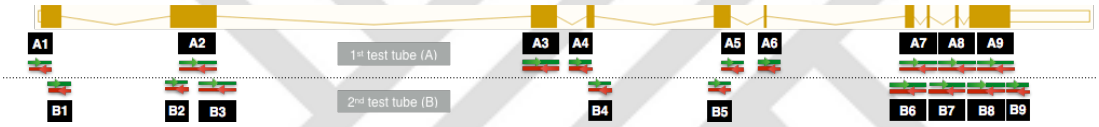

Figure 31. Positions of MEFV multiplex primer pairs.

Nucleotide length of each amplicon is shown on the following table:

Table 10. Lengths of each amplicon in MEFV multiplex experiment.

Amplicon	Length
A1	321
A2	484
A3	487
A4	300
A5	300
A6	353
A7	470
A8	426
A9	449

Amplicon	Length
B1	324
B2	301
B3	486
B4	328
B5	386
B6	486
B7	391
B8	446
B9	322

As shown in the following picture, targeted regions are successfully amplified and the sizes are as expected:

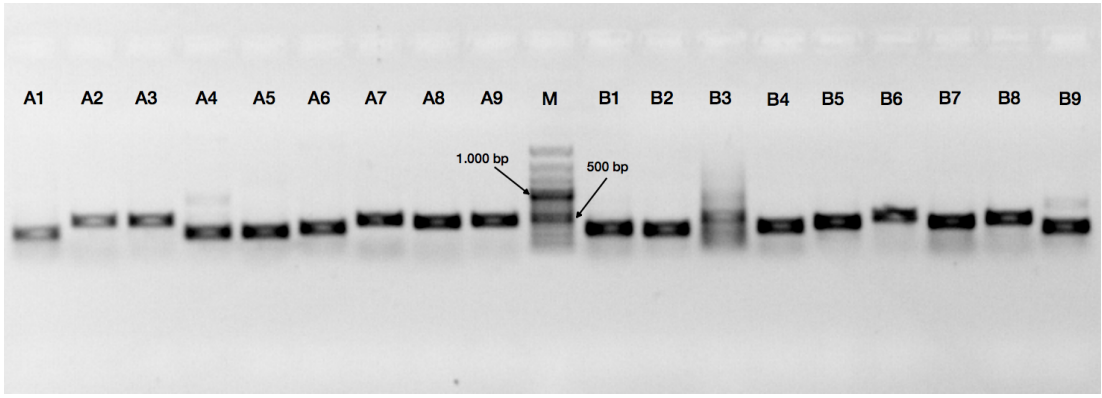

Figure 32. Singleplex PCR agarose gel image for each amplicon.

A denotes the first tube and B denotes the second tube. Marker is shown by M. Two thick bands in the marker are 500 bp and 1.000 bp signatures. Although the intensities of the bands differ, they contain enough PCR products to be sequenced, as will be shown in NGS experiment results.

B3 has non-specific amplification resulted in a smear. The main band is visible and other non-specific amplification bands are not an important case because they do not have an inhibitor effect for the NGS. Non-specific bands will be also sequenced. However, their cumulative effect is minimal.

This also reveals another potential of the proposed algorithm: there is no need to carry out any additional optimization steps, at least for the singleplex PCR validation.

In addition to the experiment shown above, two different researchers conducted experiments and validated the same results independently.

5.2.2 Multiplex PCR

Since all bands on the agarose gel from the previous section are well enough, the next step is conducting an MPCR experiment. However, as seen from the lengths of the amplicons, it is not possible for the primer pairs to work together without any inhibitory effect. For example, 4th and 5th amplicons in tube A have the same lengths: 300 nucleotides. The same is true for the 3rd and 6th amplicons in tube B have the same lengths: 486 nucleotides.

At this point, a two-step approach is employed. At the first step, both tubes are separated into two groups (4-plex) and they've been tested for their multiplex capability. In the second step, a 2-plex experiment will be conducted as planned, and the products are sequenced in an NGS instrument.

4-plex PCR

A and B groups are divided into two groups each according to their length differences as shown in the following table. Although it is a 4-plex MPCR experiment, any inhibitory affect that would result in a loss of band can be visualized in a cost-effective manner.

Table 11. MEFV 4-plex multiplex design of amplicons.

$A1_{mix}$	
Amplic	Length
A1	321
A2	484
A4	300
A9	449

$A2_{mix}$	
Amplic	Length
A3	487
A5	300
A6	353
A7	470
A8	426

$B1_{mix}$	
Amplic	Length
B1	324
B2	301
B3	486
B5	386
B8	446

$B2_{mix}$	
Amplic	Length
B4	328
B6	486
B7	391
B9	322

This time, the image below is acquired:

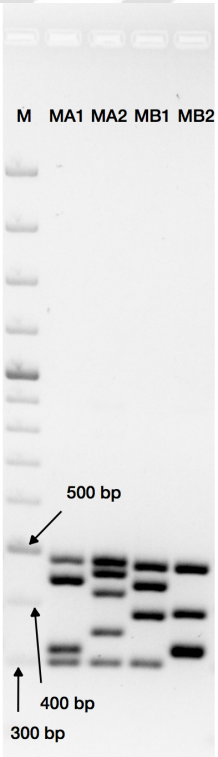

Figure 33. MEFV 4-plex MPCR agarose gel image.

Although $A1_{mix}$ and $A2_{mix}$ works as expected, there are one band missing from each $B1_{mix}$ and $B2_{mix}$.

5-plex MPCR

The lower band in $B2_{mix}$ is a thick one from the previous figure where there can be two amplicons with very similar lengths: 322 bp and 328 bp. This time, B tube is separated into three instead of two:

Table 12. B tube seperated into three groups, creating a 5-plex.

B1'$_{mix}$	
Amplic	Length
B3	486
B5	386
B9	322

B2'$_{mix}$	
Amplic	Length
B1	324
B6	486
B8	446

B3'$_{mix}$	
Amplic	Length
B2	301
B4	328
B7	391

The following gel image is acquired as a result of the 5-plex MPCR experiment:

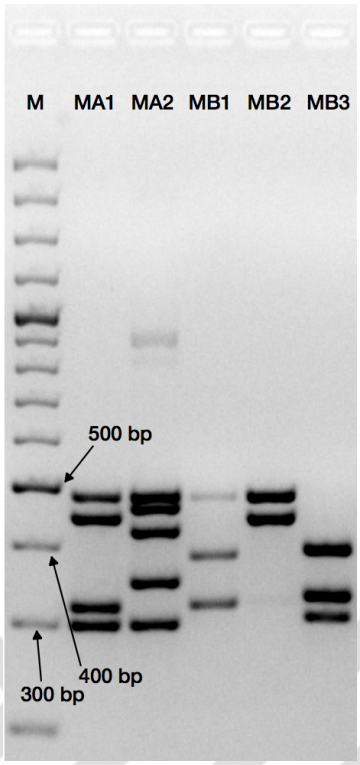

Figure 34. MEFV 5-plex MPCR agarose gel image

As shown, the lower band from B2'$_{mix}$ is almost non-existent. This band corresponds to the amplicon B1. Further experiments are conducted and it is observed that although B1 can be amplified using the singleplex PCR Kit, it cannot be amplified using multiplex PCR kit even alone. At this point, 17 out of 18 amplicons can be amplified in 4-plex. However, the picture is not complete yet, and it should be checked whether there is any loss when the amplicons are amplified in a 2-plex setting.

5.2.3 NGS results

Experiments

The ultimate answer can be answered by sequencing. Therefore, 2-plex MPCR is conducted first. Three samples are used for the experiment: 1 FMF patient and 2 healthy people. The gel image to confirm the amplification in each tube is given below in Figure 28:

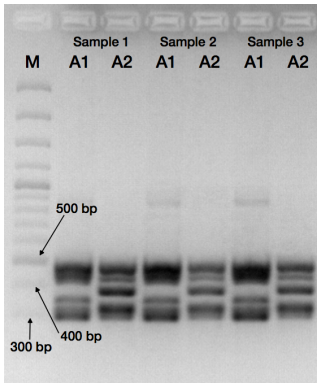

Figure 35. Gel image of 2-plex MPCR experiments for three samples.

The gel image above does not tell about exactly which bands are there, but it gives an overall view about the success of MPCR experiment in each tube. However, it is expected that amplicon B1 should not be there.

After this point, before-sequencing and sequencing experiments are performed according to the kit and instrument protocols, as mentioned in the previous chapter.

NGS run metrics

NGS run is performed on an Illumina MiSeq instrument with Illumina MiSeq Reagent Nano Kit v2 500 cycle flow cell. Cluster density is 102 K/ mm^2 which means that the flow cell is underused. The reason for such a low cluster density is that the flow cell is capable of sequencing up to 62 samples for MEFV gene in 1000x coverage. In this experiment, only 3 samples were sequenced.

QScore distribution is also good: 80.5% of the reads have the quality score of Q30 or more. The distribution of QScore is shown in Figure 29:

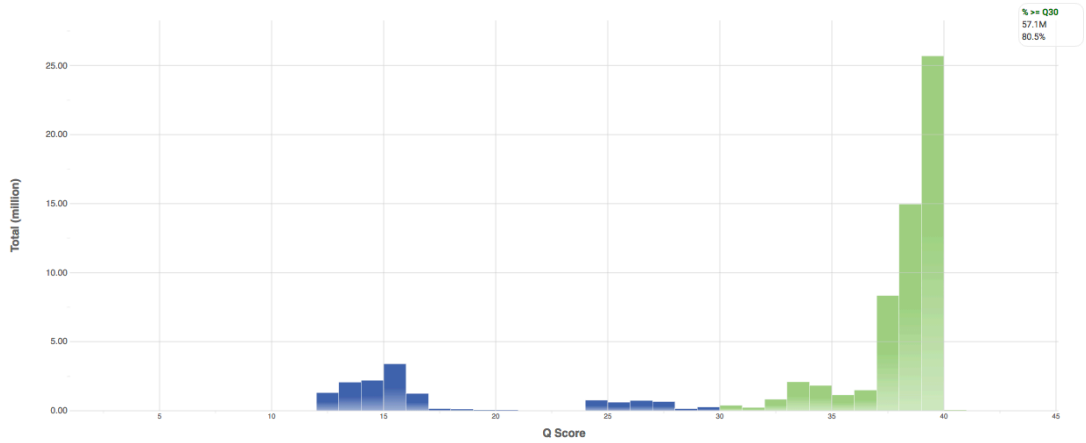

Figure 36. QScore distribution of the NGS run.

QScore by cycle number is another metric to observe anomalies during the sequencing process, if any. The next figure shows QScore by cycle number:

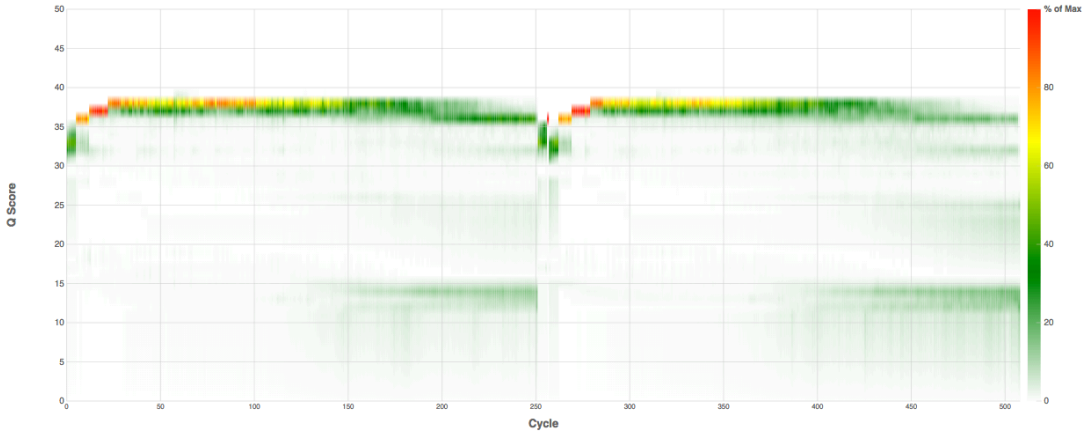

Figure 37. QScore by cycle in the NGS run.

Up to know, all quality metrics show that the experiments before NGS run was properly done, and everything went smooth in the NGS run.

NGS run data

Raw data is acquired in FASTQ format for each sample. SAM, BAM, and VCF files are created in order. Number of reads per exon is visualized using Integrative Genomics Viewer (IGV 2.3) (Robinson et al. 2011). Firstly, all read counts are displayed:

Figure 38. Visualization of read counts per region using autoscale.

The amount of amplification for each amplicon is very heterozygous. Maximum read counts per sample is 9911, 11140, and 5543, respectively. In order to visualize the same data with a 200x maximum read count displaying limit, the performance of the assay can be assessed better:

Figure 39. Visualization of read counts per region with a 200x maximum read count displaying limit.

It is clearly seen that B1 amplicon is almost never amplified. Amplification efficiency of A2 is lower on the Sample 3, and the same is valid for A3. Even for A2, minimum read count is 38 and for A3 it is 86.

Variants

VCF file is annotated using dbSNP build 126 data. Sample 1 is the disease sample whereas Sample 3 and Sample 3 are healthy samples. Variants with clinical significance according to clinVar (Landrum et al. 2014) are identified:

Table 13. clinVar annotated variations in Sample 1.

Variation	Clinical significance
c.1764A>G (p.Pro588=)	Benign
c.442G>C (p.Glu148Gln)	Conflicting interpretations of pathogenicity

Table 14. clinVar annotated variations in Sample 2.

Variation	Clinical significance
c.1530T>C (p.Asp510=)	Benign
c.1422G>A (p.Glu474=)	Benign
c.942C>T (p.Arg314=)	Benign
c.414A>G (p.Gly138=)	Benign
c.306T>C (p.Asp102=)	Benign

Table 15. clinVar annotated variations in Sample 3.

Variation	Clinical significance
c.605G>A (p.Arg202Gln)	Benign
c.414A>G (p.Gly138=)	Benign
c.306T>C (p.Asp102=)	Benign

CHAPTER 6

Discussion and Future Perspectives

Although primer design is a trivial process nowadays with many online tools, designing multiple primers that can work in harmony in a single tube is a tedious work. Multiplex PCR is a convenient method for targeted NGS studies in terms of consumable cost, labor cost, and labor time compared to conventional PCR when amplifying multiple DNA fragments at the same time.

As explained in Chapter 2, the main difficulty of the problem can be better observed by redefining the problem as a network problem. If primers are denoted as nodes, and an edge is put between two primes that can work in a harmony, then the solution set should be complete network, or a clique in technical terms. The problem is exponential in the numbers of primers and finding the optimum solution, i.e., the clique covering the targeted region, becomes intractable. Therefore, finding the first solution using a greedy depth-first search is sufficient to find a suitable set of primers that can function without inhibiting each other.

In this study, an algorithm to find the first feasible solution is developed. In addition to its firm algorithm design, it is tested both in silico and in vitro.

6.1 Discussion on in silico results

6.1.1 Human exon sequences

Due to practical reasons, benchmarking is limited with sequences between 2000 to 2100 bps long and with two different flanking sequence alternatives of either 120 or 240 bps. In addition, time to wait for the first feasible solution is limited to either 240 or 480 seconds.

On an experimental setup with 48 different combinations of factors, we have shown that multiple parameters might effect finding the first feasible solution. Increasing the length of the initial primer candidate selection sequence gives better results whereas waiting for a longer time to find the first feasible solution does not have a significant impact. Designing multiplex primers for 2 tubes is a more time-consuming problem than 3 tubes, but it does not increase dramatically when the number of tubes is increased from 3 to 5. Lastly, the selection order of candidate primers for a given

sequence interval effects the duration of finding the first feasible solution as well as the number of primer pairs in a multiplex design solution. Selecting the candidate primers in normal order with regards to the increasing base location gives the best results in terms of both getting the lowest number of primer pairs and shortest duration for the first feasible solution.

Although these settings clearly show the effect of changing the flanking sequence length and waiting time, a different setting with longer flanking sequence alternatives would increase the first set of primer candidates which in fact is the major factor of filtering out further primer candidates that are not thermodynamically compatible with the previous ones. Lastly, although selected sequences are human exons, the method can be applied to other organism to show the potential of the approach to be used for comparative genome studies.

6.1.2 A sample set of genes related with genetic diseases

6 genes are selected due to their significance in clinical utility in terms of genetic testing. Since more primers means more complexity, the implemented algorithm took several tries to find a feasible set of primers that can work in harmony for BRCA2. However, this challenge is also an indicator of the potential of the algorithm: it can be applied to many more genes in the genome.

The next step would be to design MPCR primers for each gene in the human genome. However, it would require more than 50 years to accomplish that goal using a 12-core workstation. GPU programming is promising whereas it is not suitable for all kinds of calculations using the direct interface, and more generic programming interfaces to utilize the potential of the technology are not mature yet.

On the other hand, introduction of Intel® Xeon Phi™ Coprocessors with reasonable prices enables the MPCR design in a much more shorter time. A typical cost per processor is less than $10 (*Intel BC31S1P Xeon Phi 31S1P Coprocessor*, n.d.). It is very cost-efficient to utilize those coprocessors compared to cloud computing: with the same cost, one can purchase at most 20.8 days of a similar AWS EC2 (Amazon Web Services Elastic Compute Cloud) c4.4xlarge service (“EC2 Instance Pricing – Amazon Web Services (AWS)” 2016).

6.1.3 The MEFV gene

18 primer pairs are designed for the MEFV gene with the implemented algorithm. By comparison, the only MPCR-based whole exon MEFV test in the diagnosis market (Multiplicom 2015) is accomplished 2-plex in 21 amplicons. Therefore, the algorithm is able to design MPCR assays with less primer pairs which results in lower MPCR assay development costs and shorter assay validation time.

Familial Mediterranean Fever is the most common genetic disease for people having a Mediterranean-related ancestry. In Turkey, more than 26.000 FMF tests were

purchased through public tender by public hospitals in 2015 ("EKAP (Electronic Public Procurement Platform)" 2016). The next genetic disease on the list was Cystic Fibrosis, with only 2.225 tests. Therefore, the ability to design MPCR primers for the exons of the MEFV gene has a strategic importance.

6.2 Discussion on in vitro validation results

6.2.1 Singleplex PCR

Although all of the primer pairs resulted qualitatively accepted amounts of amplifications, there were differences in the amount of intensities. When sequenced, there will be a great variation among the read counts. In order to generate clinically acceptable reads, all of the intensities should be increased to keep the minimum read depth above a certain threshold. As a result, surface of the flow cell will be consumed by those unnecessary clusters of the same kind.

One practical approach would be to measure the signal intensities quantitatively by performing qPCRs using SYBR Green as a stain. After that, further optimizations of primer concentrations would be conducted for a more uniform distribution of reads. However, since number of indexes for sample multiplexing in the same run is still quite low (("NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1) | NEB" 2016)), there is more room on the flow cell. Therefore, not performing an optimization experiment is more cost-effective considering the overall MPCR assay development process.

The smear effect seen in B3 lane of agarose gel might be due to the unintended amplification of other PCR products, which can be checked by performing a BLAST query for the forward and the reverse primers of that amplicon. However, there was no unusual property for BLAST and other primer parameters. The maximum and minimum lengths of smear also does not suggest any kind of binding due to truncated primers, which is a source of error during primer synthesis. However, the loss of smear in multiplex experiment shows that the unintended PCR products are thermodynamically not favored compared to the intended PCR product.

6.2.2 Multiplex PCR

Although all of the targeted sequences were amplified in the singleplex PCR experiments, it is understood that one of the amplicons (B2) was not amplified. It's understood that it's not because of a cross-dimer, but because of the formula of the multiplex PCR kit itself. Additional cross interaction changes are performed and no significant cross-reaction was found. Primer length or GC content of the primer pairs also do not exhibit extraordinary properties. BLAST and SNP results are also normal and any of the properties of those primer pairs are in extreme ends. Although not officially printed on user manuals of the utilized MPCR kit, it is shown in a patent that addition of special proteins to multiplex proteins increase the binding efficiency and prevent formation of weaker hybridization during the annealing step (Reaction buffer

composition for nucleic acid replication with packed DNA polymerases 2016). Although the intensity of B2 on agarose gel do not exhibit any weakness, there might be an interaction in the multiplex reaction that inhibits the binding of B2 to its target. In addition, the loss of smear for band B3 suggests that an ingredient in the MPCR kit weakens or strengthens reactions in the PCR tube.

One solution for the missing band is to create a 3-plex MPCR assay instead of 2-plex: two of the tubes will contain the working 2-plex primer pairs, and the last one will contain the primer pairs for B2 amplicon. As a result, assay is still usable, and the extra tube does not create a significant additional cost in terms of consumables. The only limitation would come from a labor perspective: this time, instead of 48, only 32 samples will be able to amplified in a 96-well thermal cycler.

The problem of variant band intensities is still observed as expected in MPCR experiment gel images. However, there is one interesting phenomenon there: although the band intensity of B3 in 4-plex gel image was quite strong, it got weaker in the 5-plex experiments. This might be due to the introduction of B9 to the group of B3, which would result in an unknown cross-dimer interaction. The same outcome is observed in NGS results, too.

6.2.3 NGS

The NGS experiment is performed after the amplification of MEFV exons in 2-plex. At this point, since all of the primer pairs in each tube are in close contact, any kind of cross-interaction may be observed. Unfortunately, due to the resolution of the agarose gel and the very similar amplicon sizes, they cannot be observed before the sequencing. Looking at the 2-plex gel images prior to sequencing, it is confirmative that all of the samples have the same band motif.

Visualization of the amplicon read counts through Integrative Genomics Viewer suggests that the read counts are in concordance with the band intensities of the 5-plex PCR. B1 amplicon is missing, and the remaining amplicons have at least 38x read depth, which is larger than the critical threshold (Meynert et al. 2013).

B1 can be incorporated into the assay as a third test tube, which would result in a 3-plex MPCR assay. Primer concentration adjustments according to the read counts can also be performed for better uniformity of the read counts among amplicons.

6.3 Conclusion

According to the literature and patent searches, the proposed algorithm is a novel method for multiplex primer design for continuous genomic regions. It is shown in the in silico validation tests that it can be applied to a majority of genes. In addition, designed primers are experimentally validated. As a result, it can be claimed that the developed algorithm can be used to develop cost-effective and reliable genetic testing assays.

REFERENCES

Amiram, Miriam, Felipe Garcia Quiroz, Daniel J. Callahan, and Ashutosh Chilkoti. 2011. "Highly Parallel Method for Synthesis of DNA Repeats Enables Discovery of 'Smart' Protein Polymers." *Nature Materials* 10 (2): 141–48. doi:10.1038/nmat2942.

Bell, Callum J., Darrell L. Dinwiddie, Neil A. Miller, Shannon L. Hateley, Elena E. Ganusova, Joann Mudge, Ray J. Langley, et al. 2011. "Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing." *Science Translational Medicine* 3 (65): 65ra4–65ra4. doi:10.1126/scitranslmed.3001756.

Bron, Coen, and Joep Kerbosch. 1973. "Algorithm 457: Finding All Cliques of an Undirected Graph." *Commun. ACM* 16 (9): 575–77. doi:10.1145/362342.362367.

Buck, G. A., J. W. Fox, M. Gunthorpe, K. M. Hager, C. W. Naeve, R. T. Pon, P. S. Adams, and J. Rush. 1999. "Design Strategies and Performance of Custom DNA Sequencing Primers." *BioTechniques* 27 (3): 528–36.

Burgart, L. J., R. A. Robinson, M. J. Heller, W. W. Wilke, O. K. Iakoubova, and J. C. Cheville. 1992. "Multiplex Polymerase Chain Reaction." *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* 5 (3): 320–23.

Cao, Antonio, and Renzo Galanello. 2010. "Beta-Thalassemia." *Genetics in Medicine* 12 (2): 61–76. doi:10.1097/GIM.0b013e3181cd68ed.

Chamberlain, J. S., R. A. Gibbs, J. E. Ranier, P. N. Nguyen, and C. T. Caskey. 1988. "Deletion Screening of the Duchenne Muscular Dystrophy Locus via Multiplex DNA Amplification." *Nucleic Acids Research* 16 (23): 11141–56.

Chong, Hansook Kim, Tao Wang, Hsiao-Mei Lu, Sara Seidler, Hong Lu, Steven Keiles, Elizabeth C. Chao, A. J. Stuenkel, Xiang Li, and Aaron M. Elliott. 2014. "The Validation and Clinical Implementation of BRCAplus: A Comprehensive High-Risk Breast Cancer Diagnostic Assay." *PloS One* 9 (5): e97408. doi:10.1371/journal.pone.0097408.

"Cytogenetic Testing Methods | University of Florida Health Pathology Laboratories." 2010. August 20. http://pathlabs.ufl.edu/services/cytogenetics/cytogenetic-testing-methods.

"DNA Oligo FAQ." 2016. *ThermoFisher Scientific DNA Oligo FAQ*. Accessed July 9. https://www.thermofisher.com/tr/en/home/products-and-services/product-

types/primers-oligos-nucleotides/invitrogen-custom-dna-oligos/technical-resources-for-oligonucleotides/dna-oligo-faq.html.

Downey, Rod G., and Michael R. Fellows. 1995. “Fixed-Parameter Tractability and Completeness II: On Completeness for W[1].” *Theoretical Computer Science* 141 (1): 109–31. doi:10.1016/0304-3975(94)00097-3.

Eblen, John D, Charles A Phillips, Gary L Rogers, and Michael A Langston. 2012. “The Maximum Clique Enumeration Problem: Algorithms, Applications, and Implementations.” *BMC Bioinformatics* 13 (Suppl 10): S5. doi:10.1186/1471-2105-13-S10-S5.

“EC2 Instance Pricing – Amazon Web Services (AWS).” 2016. *Amazon Web Services, Inc.* Accessed July 14. //aws.amazon.com/ec2/pricing/.

Eduardoff, M., T. E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Børsting, et al. 2016. “Inter-Laboratory Evaluation of the EUROFORGEN Global Ancestry-Informative SNP Panel by Massively Parallel Sequencing Using the Ion PGM™.” *Forensic Science International. Genetics* 23 (July): 178–89. doi:10.1016/j.fsigen.2016.04.008.

“EKAP (Electronic Public Procurement Platform).” 2016. Accessed July 14. https://ekap.kik.gov.tr/.

Fackenthal, James D., and Olufunmilayo I. Olopade. 2007. “Breast Cancer Risk Associated with BRCA1 and BRCA2 in Diverse Populations.” *Nature Reviews. Cancer* 7 (12): 937–48. doi:10.1038/nrc2054.

Fecteau, Heather, Kristen J. Vogel, Kristen Hanson, and Shannon Morrill-Cornelius. 2014. “The Evolution of Cancer Risk Assessment in the Era of Next Generation Sequencing.” *Journal of Genetic Counseling* 23 (4): 633–39. doi:10.1007/s10897-014-9714-7.

Feng, Yanxiao, Yuechuan Zhang, Cuifeng Ying, Deqiang Wang, and Chunlei Du. 2015. “Nanopore-Based Fourth-Generation DNA Sequencing Technology.” *Genomics, Proteomics & Bioinformatics* 13 (1): 4–16. doi:10.1016/j.gpb.2015.01.009.

Fregonese, Laura, and Jan Stolk. 2008. “Hereditary Alpha-1-Antitrypsin Deficiency and Its Clinical Consequences.” *Orphanet Journal of Rare Diseases* 3: 16. doi:10.1186/1750-1172-3-16.

Frey, Ulrich H., Hagen S. Bachmann, Jürgen Peters, and Winfried Siffert. 2008. “PCR-Amplification of GC-Rich Regions: ‘Slowdown PCR.’” *Nature Protocols* 3 (8): 1312–17. doi:10.1038/nprot.2008.112.

Garibyan, Lilit, and Nidhi Avashia. 2013. “Research Techniques Made Simple: Polymerase Chain Reaction (PCR).” *The Journal of Investigative Dermatology* 133 (3): e6. doi:10.1038/jid.2013.1.

Genetic Alliance, and District of Columbia Department of Health. 2010. *Understanding Genetics: A District of Columbia Guide for Patients and Health Professionals*. Genetic Alliance Monographs and Guides. Washington (DC): Genetic Alliance. http://www.ncbi.nlm.nih.gov/books/NBK132149/.

"Gimv, VIB and University of Antwerp Invest EUR 2.0 Million in Molecular Diagnostics Startup Multiplicom." 2016. Accessed July 9. http://www.vib.be/en/news/Pages/Gimv,-VIB-and-University-of-Antwerp-invest-EUR-2-0-million-in-molecular-diagnostics-startup-Multiplicom.aspx.

Gray, Phillip N., Charles L. M. Dunlop, and Aaron M. Elliott. 2015. "Not All Next Generation Sequencing Diagnostics Are Created Equal: Understanding the Nuances of Solid Tumor Assay Design for Somatic Mutation Detection." *Cancers* 7 (3): 1313–32. doi:10.3390/cancers7030837.

Intel BC31S1P Xeon Phi 31S1P Coprocessor. n.d.

Janssens, A. Cecile J. W., Yurii S. Aulchenko, Stefano Elefante, Gerard J. J. M. Borsboom, Ewout W. Steyerberg, and Cornelia M. van Duijn. 2006. "Predictive Testing for Complex Diseases Using Multiple Genes: Fact or Fiction?" *Genetics in Medicine* 8 (7): 395–400. doi:10.1097/01.gim.0000229689.18263.f4.

Katsanis, Sara Huston, and Nicholas Katsanis. 2013. "Molecular Genetic Testing and the Future of Clinical Genomics." *Nature Reviews Genetics* 14 (6): 415–26. doi:10.1038/nrg3493.

Kennedy, Suzanne. 2011. *PCR Troubleshooting and Optimization: The Essential Guide*. Horizon Scientific Press.

Kose, F., W. Weckwerth, T. Linke, and O. Fiehn. 2001. "Visualizing Plant Metabolomic Correlation Networks Using Clique-Metabolite Matrices." *Bioinformatics (Oxford, England)* 17 (12): 1198–1208.

Landrum, Melissa J., Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. 2014. "ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype." *Nucleic Acids Research* 42 (Database issue): D980–85. doi:10.1093/nar/gkt1113.

Lefever, Steve, Filip Pattyn, Jan Hellemans, and Jo Vandesompele. 2013. "Single-Nucleotide Polymorphisms and Other Mismatches Reduce Performance of Quantitative PCR Assays." *Clinical Chemistry*, September, clinchem.2013.203653. doi:10.1373/clinchem.2013.203653.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics (Oxford, England)* 27 (21): 2987–93. doi:10.1093/bioinformatics/btr509.

Li, Meng, Zhixun Xie, Zhiqin Xie, Jiabo Liu, Liji Xie, Xianwen Deng, Sisi Luo, et al. 2016. “Simultaneous Detection of Four Different Neuraminidase Types of Avian Influenza A H5 Viruses by Multiplex Reverse Transcription PCR Using a GeXP Analyser.” *Influenza and Other Respiratory Viruses* 10 (2): 141–49. doi:10.1111/irv.12370.

“Limitations of Cytogenetic Testing.” 2016. *Limitations of Cytogenetic Testing*. Accessed July 9. http://www.labs.gosh.nhs.uk/media/384333/Limitations%20of%20cytogenetic%20testing.pdf.

Li, Xiaodong, Mona Anand, Josh D. Haimes, Namitha Manoj, Aaron M. Berlin, Brian A. Kudlow, Marisa R. Nucci, Tony L. Ng, Colin J. R. Stewart, and Cheng-Han Lee. 2016. “The Application of next Generation-Sequencing-Based Molecular Diagnostics in Endometrial Stromal Sarcoma.” *Histopathology*, March. doi:10.1111/his.12966.

Li, Yang, Shu-Juan Guo, Ning Shao, Shun Tu, Miao Xu, Zhao-Rui Ren, Xing Ling, Guo-Qing Wang, Zhi-Xin Lin, and Sheng-Ce Tao. 2011. “A Universal Multiplex PCR Strategy for 100-Plex Amplification Using a Hydrophobically Patterned Microarray.” *Lab on a Chip* 11 (21): 3609–18. doi:10.1039/C1LC20526A.

López, Mónica, Pilar Giraldo, Patricia Alvarez, R. Cornudella, Miguel Pocoví, Antonio Martínez, Jordi Fontcuberta, and José Manuel Soria. 2007. “Multiplex Assay for Genetic Testing of Thrombophilia: A Method for Routine Clinical Care.” *Journal of Clinical Laboratory Analysis* 21 (6): 349–55. doi:10.1002/jcla.20183.

Mahdieh, Nejat, and Bahareh Rabbani. 2013. “An Overview of Mutation Detection Methods in Genetic Disorders.” *Iranian Journal of Pediatrics* 23 (4): 375–88.

Mamanova, Lira, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner. 2010. “Target-Enrichment Strategies for next-Generation Sequencing.” *Nature Methods* 7 (2): 111–18. doi:10.1038/nmeth.1419.

Mcginniss, Matthew J., David H. Brown, Andrea Fulwiler, Molly Marten, Joyce S. T. Lim-Steele, and Michael M. Kaback. 2002. “Eight Novel Mutations in the HEXA Gene.” *Genetics in Medicine* 4 (3): 158–61. doi:10.1097/00125817-200205000-00010.

Methods and systems of nucleic acid sequencing. 2016. Accessed July 9. http://www.google.com/patents/EP1472335A2.

Metzker, Michael L. 2010. “Sequencing Technologies - the next Generation.” *Nature Reviews. Genetics* 11 (1): 31–46. doi:10.1038/nrg2626.

Meynert, Alison M, Louise S Bicknell, Matthew E Hurles, Andrew P Jackson, and Martin S Taylor. 2013. “Quantifying Single Nucleotide Variant Detection Sensitivity in Exome Sequencing.” *BMC Bioinformatics* 14 (June): 195. doi:10.1186/1471-2105-14-195.

Moskowitz, Samuel M., James F. Chmiel, Darci L. Sternen, Edith Cheng, Ronald L. Gibson, Susan G. Marshall, and Garry R. Cutting. 2008. “Clinical Practice and Genetic Counseling for Cystic Fibrosis and CFTR-Related Disorders.” *Genetics in Medicine* 10 (12): 851–68. doi:10.1097/GIM.0b013e31818e55a2.

Multiplicom. 2015. “FMF MASTR Dx.” *Multiplicom*. December 11. http://www.multiplicom.com/product/fmf-mastr-dx.

“Multiplicom’s MASTR Technology.” 2016. *Multiplicom’s MASTR Technology - A Powerful Multiplex PCR Based Approach Enabling MPS Based Diag*. Accessed July 9. http://www.tcag.ca/events/120627MASTR.html.

Muzzey, Dale, Eric A. Evans, and Caroline Lieber. 2015. “Understanding the Basics of NGS: From Mechanism to Variant Calling.” *Current Genetic Medicine Reports* 3 (4): 158–65. doi:10.1007/s40142-015-0076-8.

“NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1) | NEB.” 2016. Accessed July 14. https://www.neb.com/products/e7600-nebnext-multiplex-oligos-for-illumina-dual-index-primers-set-1.

Netto, George J., Rana D. Saad, and Peter A. Dysert. 2003. “Diagnostic Molecular Pathology: Current Techniques and Clinical Applications, Part I.” *Proceedings (Baylor University. Medical Center)* 16 (4): 379–83.

“OMIM Entry - # 249100 - FAMILIAL MEDITERRANEAN FEVER; FMF.” 2016. Accessed July 11. http://www.omim.org/entry/249100.

Orkunoglu-Suer, Funda, Arthur F. Harralson, David Frankfurter, Paul Gindoff, and Travis J. O’Brien. 2015. “Targeted Single Molecule Sequencing Methodology for Ovarian Hyperstimulation Syndrome.” *BMC Genomics* 16: 264. doi:10.1186/s12864-015-1451-2.

Owczarzy, Richard, Yong You, Bernardo G. Moreira, Jeffrey A. Manthey, Lingyan Huang, Mark A. Behlke, and Joseph A. Walder. 2004. “Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures.” *Biochemistry* 43 (12): 3537–54. doi:10.1021/bi034621r.

Pandey, Ram Vinay, Pulverer Walter, Rainer Kallmeyer, Gabriel Beikircher, Stephan Pabinger, Albert Kriegner, and Andreas Weinhäusel. 2016. “MSRE-HTPrimer: A High-Throughput and Genome-Wide Primer Design Pipeline Optimized for Epigenetic Research.” *Clinical Epigenetics* 8: 26. doi:10.1186/s13148-016-0190-9.

Pastores, Gregory M., and Derralynn A. Hughes. 1993. “Gaucher Disease.” In *GeneReviews(®)*, edited by Roberta A. Pagon, Margaret P. Adam, Holly H. Ardinger, Stephanie E. Wallace, Anne Amemiya, Lora JH Bean, Thomas D. Bird, et al. Seattle (WA): University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1269/.

Poon, Kok-Siong, Karen Mei-Ling Tan, and Evelyn Siew-Chuan Koay. 2016. “Targeted next-Generation Sequencing of the ATP7B Gene for Molecular Diagnosis of Wilson Disease.” *Clinical Biochemistry* 49 (1-2): 166–71. doi:10.1016/j.clinbiochem.2015.10.003.

Pourfarzam, Morteza, and Fouzieh Zadhoush. 2013. “Newborn Screening for Inherited Metabolic Disorders; News and Views.” *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences* 18 (9): 801–8.

“PrimerDigital: Biotechnology Solutions.” 2016. Accessed July 9. http://primerdigital.com/.

“Primo.” 2016. Accessed July 9. http://www.changbioscience.com/primo/primoml.html.

Raj, Ankush, Netrapal Singh, Krishna B. Gupta, Dhruva Chaudhary, Aparna Yadav, Anil Chaudhary, Kshitij Agarwal, et al. 2016. “Comparative Evaluation of Several Gene Targets for Designing a Multiplex-PCR for an Early Diagnosis of Extrapulmonary Tuberculosis.” *Yonsei Medical Journal* 57 (1): 88–96. doi:10.3349/ymj.2016.57.1.88.

Reaction buffer composition for nucleic acid replication with packed DNA polymerases. 2016. Accessed July 9. http://www.google.com/patents/US7939645.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26. doi:10.1038/nbt.1754.

Rodríguez, Islay, Caroline Burri, Angel A. Noda, Véronique Douet, and Lise Gern. 2015. “Multiplex PCR for Molecular Screening of Borrelia Burgdorferi Sensu Lato, Anaplasma Spp. and Babesia Spp.” *Annals of Agricultural and Environmental Medicine: AAEM* 22 (4): 642–46. doi:10.5604/12321966.1185767.

Romero-Pastrana, Francisco, and Francisco Romero-Pastrana. 2012. “Detection and Typing of Human Papilloma Virus by Multiplex PCR with Type-Specific Primers, Detection and Typing of Human Papilloma Virus by Multiplex PCR with Type-Specific Primers.” *International Scholarly Research Notices, International Scholarly Research Notices* 2012, 2012 (March): e186915. doi:10.5402/2012/186915, 10.5402/2012/186915.

Roux, Kenneth H. 2009. “Optimization and Troubleshooting in PCR.” *Cold Spring Harbor Protocols* 2009 (4): pdb.ip66. doi:10.1101/pdb.ip66.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. 1988. “Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase.” *Science (New York, N.Y.)* 239 (4839): 487–91.

SantaLucia, J. 1998. “A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (4): 1460–65.

Saul, Robert A., and Jack C. Tarleton. 1993. “FMR1-Related Disorders.” In *GeneReviews(®)*, edited by Roberta A. Pagon, Margaret P. Adam, Holly H. Ardinger, Stephanie E. Wallace, Anne Amemiya, Lora JH Bean, Thomas D. Bird, et al. Seattle (WA): University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1384/.

Schildkraut, Carl, and Shneior Lifson. 1965. “Dependence of the Melting Temperature of DNA on Salt Concentration.” *Biopolymers* 3 (2): 195–208. doi:10.1002/bip.360030207.

Shen, Zhiyong, Wubin Qu, Wen Wang, Yiming Lu, Yonghong Wu, Zhifeng Li, Xingyi Hang, Xiaolei Wang, Dongsheng Zhao, and Chenggang Zhang. 2010. “MPprimer: A Program for Reliable Multiplex PCR Primer Design.” *BMC Bioinformatics* 11: 143. doi:10.1186/1471-2105-11-143.

Shi, Xiaodan, Rui Wu, Ming Shi, Linfu Zhou, Mengli Wu, Yining Yang, Xinyue An, et al. 2016. “Simultaneous Detection of 13 Viruses Involved in Meningoencephalitis Using a Newly Developed Multiplex PCR Mag-Array System.” *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases* 49 (May): 80–86. doi:10.1016/j.ijid.2016.05.023.

Smedley, Damian, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. 2009. “BioMart – Biological Queries Made Easy.” *BMC Genomics* 10: 22. doi:10.1186/1471-2164-10-22.

Stockley, Tracy L., Sarah Akber, Natalie Bulgin, and Peter N. Ray. 2006. “Strategy for Comprehensive Molecular Testing for Duchenne and Becker Muscular Dystrophies.” *Genetic Testing* 10 (4): 229–43. doi:10.1089/gte.2006.10.229.

Teer, Jamie K., Lori L. Bonnycastle, Peter S. Chines, Nancy F. Hansen, Natsuyo Aoyama, Amy J. Swift, Hatice Ozel Abaan, et al. 2010. “Systematic Comparison of Three Genomic Enrichment Methods for Massively Parallel DNA Sequencing.” *Genome Research* 20 (10): 1420–31. doi:10.1101/gr.106716.110.

“Testing Methods.” 2016. Accessed July 9. http://www.nchpeg.org/bssr/index.php?option=com_k2&view=item&id=82:methods&Itemid=111.

Thornhill, Alan R., and Karen Snow. 2002. “Molecular Diagnostics in Preimplantation Genetic Diagnosis.” *The Journal of Molecular Diagnostics : JMD* 4 (1): 11–29.

“UCSC In-Silico PCR.” 2016. Accessed July 9. https://genome.ucsc.edu/cgi-bin/hgPcr.

“Universal Newborn Screening for Cystic Fibrosis in Connecticut.” 2006. *Universal Newborn Screening for Cystic Fibrosis in Connecticut*. July. http://www.ct.gov/dph/lib/dph/genomics/cfinct7_06.pdf.

Universal primer sequence for multiplex DNA amplification. 2016. Accessed July 9. http://www.google.com/patents/US6207372.

Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. 2012. "Primer3—new Capabilities and Interfaces." *Nucleic Acids Research* 40 (15): e115. doi:10.1093/nar/gks596.

Valones, Marcela Agne Alves, Rafael Lima Guimarães, Lucas André Cavalcanti Brandão, Paulo Roberto Eleutério de Souza, Alessandra de Albuquerque Tavares Carvalho, and Sergio Crovela. 2009. "Principles and Applications of Polymerase Chain Reaction in Medical Diagnostic Fields: A Review." *Brazilian Journal of Microbiology* 40 (1): 1–11. doi:10.1590/S1517-83822009000100001.

Wang, Shiyun, Rong Zhang, Guangxin Xiang, Yang Li, Xuhong Hou, Fusong Jiang, Feng Jiang, Cheng Hu, and Weiping Jia. 2015. "Mutation Screening for Thalassaemia in the Jino Ethnic Minority Population of Yunnan Province, Southwest China." *BMJ Open* 5 (12). doi:10.1136/bmjopen-2015-010047.

Wen, Daxing, and Chunqing Zhang. 2012. "Universal Multiplex PCR: A Novel Method of Simultaneous Amplification of Multiple DNA Fragments." *Plant Methods* 8: 32. doi:10.1186/1746-4811-8-32.

Xu, Wentao, Zhifang Zhai, Kunlun Huang, Nan Zhang, Yanfang Yuan, Ying Shang, and Yunbo Luo. 2012. "A Novel Universal Primer-Multiplex-PCR Method with Sequencing Gel Electrophoresis Analysis." *PLOS ONE* 7 (1): e22900. doi:10.1371/journal.pone.0022900.

Yamada, Tomoyuki, Haruhiko Soma, and Shinichi Morishita. 2006. "PrimerStation: A Highly Specific Multiplex Genomic PCR Primer Design Server for the Human Genome." *Nucleic Acids Research* 34 (Web Server issue): W665–69. doi:10.1093/nar/gkl297.

Yao, Ruofan, and Katherine R. Goetzinger. 2016. "Genetic Carrier Screening in the Twenty-First Century." *Clinics in Laboratory Medicine* 36 (2): 277–88. doi:10.1016/j.cll.2016.01.003.

Ye, Jian, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, and Thomas L. Madden. 2012. "Primer-BLAST: A Tool to Design Target-Specific Primers for Polymerase Chain Reaction." *BMC Bioinformatics* 13: 134. doi:10.1186/1471-2105-13-134.

Yoon, Hye-Ran. 2015. "Screening Newborns for Metabolic Disorders Based on Targeted Metabolomics Using Tandem Mass Spectrometry." *Annals of Pediatric Endocrinology & Metabolism* 20 (3): 119–24. doi:10.6065/apem.2015.20.3.119.

You, Frank M., Naxin Huo, Yong Qiang Gu, Ming-cheng Luo, Yaqin Ma, Dave Hane, Gerard R. Lazo, Jan Dvorak, and Olin D. Anderson. 2008. "BatchPrimer3: A High Throughput Web Application for PCR and Sequencing Primer Design." *BMC Bioinformatics* 9: 253. doi:10.1186/1471-2105-9-253.

APPENDICES

APPENDIX A

VCF FILES

Sample 1

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
        3aligned_sorted.bam
chr16   3243888 rs1231122       C       T       221.999 .       DP=1737;VDB=0;SGB=-
0.693147;RPB=0.989109;MQB=1;MQSB=1;BQB=0.130429;MQ0F=0;AF1=1;AC1=2;DP4=6,1,1307,193;M
Q=60;FQ=-
281.989;PV4=1,1,1,0.450828;ASP;CAF=0.6464,.,0.3536;COMMON=1;G5;GENEINFO=MEFV:4210;GNO
;HD;KGPhase1;KGPhase3;LSD;NSM;PM;PMC;REF;RS=1231122;RSPOS=3243888;SAO=1;SLO;SSR=0;SYN
;VC=SNV;VLD;VP=0x050168000b0515053f100101;WGT=1;dbSNPBuildID=87;ANN=T|missense_varian
t|MODERATE|MEFV|MEFV|transcript|NM_001198536.1|protein_coding|8/9|c.1306G>A|p.Gly436A
rg|1346/3041|1306/1338|436/445||,T|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_000
243.2|protein_coding|9/10|c.1764G>A|p.Pro588Pro|1804/3499|1764/2346|588/781||
        GT:PL   1/1:255,255,0
chr16   3243922 rs1231123       A       T       221.999 .       DP=2864;VDB=0;SGB=-
0.693147;RPB=0.897133;MQB=1;MQSB=1;BQB=0.00336729;MQ0F=0;AF1=1;AC1=2;DP4=3,2,1285,283
;MQ=60;FQ=-
281.989;PV4=0.224526,1,1,1;ASP;CAF=0.6887,.,0.3113;COMMON=1;G5;GENEINFO=MEFV:4210;GNO
;HD;INT;KGPhase1;KGPhase3;NSM;REF;RS=1231123;RSPOS=3243922;SAO=0;SLO;SSR=0;VC=SNV;VLD
;VP=0x050100080a0515053f000101;WGT=1;dbSNPBuildID=87;ANN=T|missense_variant|MODERATE|
MEFV|MEFV|transcript|NM_001198536.1|protein_coding|8/9|c.1272T>A|p.Asp424Glu|1312/304
1|1272/1338|424/445||,T|intron_variant|MODIFIER|MEFV|MEFV|transcript|NM_000243.2|prot
ein_coding|8/9|c.1760-30T>A||||||     GT:PL   1/1:255,255,0
chr16   3254626 rs3743930       C       G       225.009 .       DP=5911;VDB=0;SGB=-
0.693147;RPB=0.943861;MQB=1;MQSB=1;BQB=0.420522;MQ0F=0;AF1=0.5;AC1=1;DP4=870,50,2070,
98;MQ=60;FQ=225.007;PV4=0.270706,1,1,0.226029;ASP;CAF=0.8736,0.1264;COMMON=1;G5;GENEI
NFO=MEFV:4210;GNO;HD;INT;KGPhase1;KGPhase3;LSD;NSM;OM;PM;PMC;REF;RS=3743930;RSPOS=325
4626;RV;SAO=1;SLO;SSR=0;VC=SNV;VLD;VP=0x050168080a05150536110100;WGT=1;dbSNPBuildID=1
07;ANN=G|missense_variant|MODERATE|MEFV|MEFV|transcript|NM_000243.2|protein_coding|2/
10|c.442G>C|p.Glu148Gln|482/3499|442/2346|148/781||,G|intron_variant|MODIFIER|MEFV|ME
FV|transcript|NM_001198536.1|protein_coding|1/8|c.277+1685G>C|||||| GT:PL
        0/1:255,0,255
```

Sample 2

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
        12aligned_sorted.bam
chr16   3243888 rs1231122       C       T       225.009 .       DP=1740;VDB=0;SGB=-
0.693147;RPB=0.99936;MQB=1;MQSB=1;BQB=0.517243;MQ0F=0;AF1=0.5;AC1=1;DP4=603,119,572,1
18;MQ=60;FQ=225.007;PV4=0.775977,0.411016,1,0.275513;ASP;CAF=0.6464,.,0.3536;COMMON=1
;G5;GENEINFO=MEFV:4210;GNO;HD;KGPhase1;KGPhase3;LSD;NSM;PM;PMC;REF;RS=1231122;RSPOS=3
243888;SAO=1;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x050168000b0515053f100101;WGT=1;dbSNPBuildI
D=87;ANN=T|missense_variant|MODERATE|MEFV|MEFV|transcript|NM_001198536.1|protein_codi
ng|8/9|c.1306G>A|p.Gly436Arg|1346/3041|1306/1338|436/445||,T|synonymous_variant|LOW|M
EFV|MEFV|transcript|NM_000243.2|protein_coding|9/10|c.1764G>A|p.Pro588Pro|1804/3499|1
764/2346|588/781||     GT:PL   0/1:255,0,255
chr16   3243922 rs1231123       A       T       225.009 .       DP=2635;VDB=0;SGB=-
0.693147;RPB=0.999965;MQB=1;MQSB=1;BQB=0.952023;MQ0F=0;AF1=0.5;AC1=1;DP4=589,148,576,
145;MQ=60;FQ=225.007;PV4=1,1,1,1;ASP;CAF=0.6887,.,0.3113;COMMON=1;G5;GENEINFO=MEFV:42
10;GNO;HD;INT;KGPhase1;KGPhase3;NSM;REF;RS=1231123;RSPOS=3243922;SAO=0;SLO;SSR=0;VC=S
NV;VLD;VP=0x050100080a0515053f000101;WGT=1;dbSNPBuildID=87;ANN=T|missense_variant|MOD
ERATE|MEFV|MEFV|transcript|NM_001198536.1|protein_coding|8/9|c.1272T>A|p.Asp424Glu|13
12/3041|1272/1338|424/445||,T|intron_variant|MODIFIER|MEFV|MEFV|transcript|NM_000243.
2|protein_coding|8/9|c.1760-30T>A||||||       GT:PL   0/1:255,0,255
chr16   3247073 rs224206        A       G       225.009 .       DP=7999;VDB=0;SGB=-
0.693147;RPB=0.986771;MQB=1;MQSB=1;BQB=0.559696;MQ0F=0;AF1=0.5;AC1=1;DP4=2105,5,2230,
6;MQ=60;FQ=225.007;PV4=1,1,1,0.143074;ASP;CAF=0.3369,0.6631;COMMON=1;G5;G5A;GENEINFO=
MEFV:4210;GNO;HD;KGPhase1;KGPhase3;LSD;PM;PMC;REF;RS=224206;RSPOS=3247073;RV;SAO=1;SL
O;SSR=0;SYN;VC=SNV;VLD;VP=0x050168000305170 53f100100;WGT=1;dbSNPBuildID=79;ANN=G|syno
nymous_variant|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_coding|5/10|c.1530T>C|p.A
sp510Asp|1570/3499|1530/2346|510/781||,G|synonymous_variant|LOW|MEFV|MEFV|transcript|
NM_001198536.1|protein_coding|4/9|c.897T>C|p.Asp299Asp|937/3041|897/1338|299/445||
        GT:PL   0/1:255,0,255
chr16   3247175 rs224207        T       C       225.009 .       DP=7948;VDB=0;SGB=-
0.693147;RPB=0.935059;MQB=0.999614;MQSB=0.983993;BQB=0.784707;MQ0F=0;AF1=0.5;AC1=1;DP
4=2026,29,2247,62;MQ=60;FQ=225.007;PV4=0.00393448,1,1,0.206561;ASP;CAF=0.3494,0.6506;
COMMON=1;G5;G5A;GENEINFO=MEFV:4210;GNO;HD;KGPhase1;KGPhase3;PM;PMC;REF;RS=224207;RSPO
S=3247175;RV;SAO=0;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x05012800030517053f000100;WGT=1;dbSNP
BuildID=79;ANN=C|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_codi
ng|5/10|c.1428A>G|p.Gln476Gln|1468/3499|1428/2346|476/781||,C|synonymous_variant|LOW|
MEFV|MEFV|transcript|NM_001198536.1|protein_coding|4/9|c.795A>G|p.Gln265Gln|835/3041|
795/1338|265/445||     GT:PL   0/1:255,0,255
chr16   3247181 rs224208        C       T       225.009 .       DP=7938;VDB=0;SGB=-
0.693147;RPB=0.449413;MQB=0.999618;MQSB=0.98792;BQB=0.0443348;MQ0F=0;AF1=0.5;AC1=1;DP
4=2028,38,2215,82;MQ=60;FQ=225.007;PV4=0.00055145,3.60423e-
08,1,1;ASP;CAF=0.361,0.639;COMMON=1;G5;G5A;GENEINFO=MEFV:4210;GNO;HD;KGPhase1;KGPhase
3;LSD;PM;PMC;REF;RS=224208;RSPOS=3247181;RV;SAO=1;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x05016
800030517053e100100;WGT=1;dbSNPBuildID=79;ANN=T|synonymous_variant|LOW|MEFV|MEFV|tran
script|NM_000243.2|protein_coding|5/10|c.1422G>A|p.Glu474Glu|1462/3499|1422/2346|474/
781||,T|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_001198536.1|protein_coding|4/9
|c.789G>A|p.Glu263Glu|829/3041|789/1338|263/445||    GT:PL   0/1:255,0,255
chr16   3248865 rs224212        T       C       225.009 .       DP=3856;VDB=0;SGB=-
0.693147;RPB=0.679987;MQB=1;MQSB=1;BQB=0.543626;MQ0F=0;AF1=0.5;AC1=1;DP4=965,5,986,23
;MQ=60;FQ=225.007;PV4=0.000894842,0.224615,1,0.294783;ASP;CAF=0.3389,0.6611;COMMON=1;
G5;G5A;GENEINFO=MEFV:4210;GNO;HD;INT;KGPhase1;KGPhase3;RS=224212;RSPOS=3248865;RV;SAO
=0;SLO;SSR=0;VC=SNV;VLD;VP=0x05010008000517053f000100;WGT=1;dbSNPBuildID=79;ANN=C|int
ron_variant|MODIFIER|MEFV|MEFV|transcript|NM_000243.2|protein_coding|4/9|c.1356+44A>G
||||||,C|intron_variant|MODIFIER|MEFV|MEFV|transcript|NM_001198536.1|protein_coding|3
/8|c.723+44A>G||||||   GT:PL   0/1:255,0,255
chr16   3249749 rs224213        G       A       188.009 .       DP=240;VDB=0;SGB=-
0.693147;RPB=0.994399;MQB=1;BQB=0.965263;MQ0F=0;AF1=0.5;AC1=1;DP4=0,120,0,120;MQ=60;F
Q=187.478;PV4=1,1,1,1;ASP;CAF=0.3093,0.6907;COMMON=1;G5;G5A;GENEINFO=MEFV:4210;GNO;HD
;KGPhase1;KGPhase3;LSD;PM;PMC;REF;RS=224213;RSPOS=3249749;RV;S3D;SAO=1;SLO;SSR=0;SYN;
VC=SNV;VLD;VP=0x05036800030517053f100100;WGT=1;dbSNPBuildID=79;ANN=A|synonymous_varia
nt|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_coding|3/10|c.942C>T|p.Arg314Arg|982/
3499|942/2346|314/781||,A|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_001198536.1|
protein_coding|2/9|c.309C>T|p.Arg103Arg|349/3041|309/1338|103/445|| GT:PL
        0/1:218,0,217
chr16   3254654 rs224224        T       C       225.009 .       DP=7381;VDB=0;SGB=-
0.693147;RPB=0.368831;MQB=1;MQSB=1;BQB=0.358463;MQ0F=0;AF1=0.5;AC1=1;DP4=1657,105,191
3,113;MQ=60;FQ=225.007;PV4=0.625015,1,1,0.302438;ASP;CAF=0.6134,.,0.3866;COMMON=1;G5;
GENEINFO=MEFV:4210;GNO;INT;KGPhase1;KGPhase3;LSD;PM;PMC;REF;RS=224224;RSPOS=3254654;R
V;SAO=1;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x05016808030515013e100100;WGT=1;dbSNPBuildID=79;
```

```
ANN=C|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_coding|2/10|c.4
14A>G|p.Gly138Gly|454/3499|414/2346|138/781||,C|intron_variant|MODIFIER|MEFV|MEFV|tra
nscript|NM_001198536.1|protein_coding|1/8|c.277+1657A>G||||||      GT:PL
        0/1:255,0,255
chr16   3254762 rs224225        A       G       225.009 .       DP=7432;VDB=0;SGB=-
0.693147;RPB=0.312989;MQB=1;MQSB=1;BQB=0.0219872;MQ0F=0;AF1=0.5;AC1=1;DP4=1590,149,19
58,158;MQ=60;FQ=225.007;PV4=0.21001,1,1,0.483176;ASP;CAF=0.6106,0.3894;COMMON=1;G5;G5
A;GENEINFO=MEFV:4210;GNO;HD;INT;KGPhase1;KGPhase3;LSD;PM;PMC;REF;RS=224225;RSPOS=3254
762;RV;SAO=1;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x05016808030517053f100101;WGT=1;dbSNPBuildI
D=79;ANN=G|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_coding|2/1
0|c.306T>C|p.Asp102Asp|346/3499|306/2346|102/781||,G|intron_variant|MODIFIER|MEFV|MEF
V|transcript|NM_001198536.1|protein_coding|1/8|c.277+1549T>C||||||  GT:PL
        0/1:255,0,255
```

Sample 3

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
        17aligned_sorted.bam
chr16   3243888 rs1231122       C       T       221.999 .       DP=960;VDB=0;SGB=-
0.693147;RPB=0.847579;MQB=1;MQSB=1;BQB=0.434994;MQ0F=0;AF1=1;AC1=2;DP4=29,2,689,159;M
Q=60;FQ=-
281.989;PV4=0.0978225,1,1,0.497719;ASP;CAF=0.6464,.,0.3536;COMMON=1;G5;GENEINFO=MEFV:
4210;GNO;HD;KGPhase1;KGPhase3;LSD;NSM;PM;PMC;REF;RS=1231122;RSPOS=3243888;SAO=1;SLO;S
SR=0;SYN;VC=SNV;VLD;VP=0x050168000b0515053f100101;WGT=1;dbSNPBuildID=87;ANN=T|missens
e_variant|MODERATE|MEFV|MEFV|transcript|NM_001198536.1|protein_coding|8/9|c.1306G>A|p
.Gly436Arg|1346/3041|1306/1338|436/445||,T|synonymous_variant|LOW|MEFV|MEFV|transcrip
t|NM_000243.2|protein_coding|9/10|c.1764G>A|p.Pro588Pro|1804/3499|1764/2346|588/781||
        GT:PL   1/1:255,255,0
chr16   3243922 rs1231123       A       T       221.999 .       DP=1632;VDB=0;SGB=-
0.693147;RPB=0.592421;MQB=0.952814;MQSB=0.995052;BQB=0.84938;MQ0F=0.000612745;AF1=1;A
C1=2;DP4=26,4,676,224;MQ=60;FQ=-
281.989;PV4=0.195472,1,1,0.407338;ASP;CAF=0.6887,.,0.3113;COMMON=1;G5;GENEINFO=MEFV:4
210;GNO;HD;INT;KGPhase1;KGPhase3;NSM;REF;RS=1231123;RSPOS=3243922;SAO=0;SLO;SSR=0;VC=
SNV;VLD;VP=0x050100080a0515053f000101;WGT=1;dbSNPBuildID=87;ANN=T|missense_variant|MO
DERATE|MEFV|MEFV|transcript|NM_001198536.1|protein_coding|8/9|c.1272T>A|p.Asp424Glu|1
312/3041|1272/1338|424/445||,T|intron_variant|MODIFIER|MEFV|MEFV|transcript|NM_000243
.2|protein_coding|8/9|c.1760-30T>A||||||      GT:PL   1/1:255,255,0
chr16   3254463 rs224222        C       T       128.076 .       DP=39;VDB=1.51284e-
19;SGB=-
0.693127;RPB=0.981043;MQB=1;BQB=0.86525;MQ0F=0;AF1=0.508032;AC1=1;DP4=4,0,33,0;MQ=60;
FQ=-
12.2521;PV4=1,0.439573,1,1;ASP;CAF=0.864,0.136;COMMON=1;G5;GENEINFO=MEFV:4210;GNO;HD;
INT;KGPhase1;KGPhase3;LSD;NSM;PM;PMC;REF;RS=224222;RSPOS=3254463;RV;SAO=1;SLO;SSR=0;V
C=SNV;VLD;VP=0x050168080a05150536100100;WGT=1;dbSNPBuildID=79;ANN=T|missense_variant|
MODERATE|MEFV|MEFV|transcript|NM_000243.2|protein_coding|2/10|c.605G>A|p.Arg202Gln|64
5/3499|605/2346|202/781||,T|intron_variant|MODIFIER|MEFV|MEFV|transcript|NM_001198536
.1|protein_coding|1/8|c.277+1848G>A||||||     GT:PL   0/1:158,0,15
chr16   3254654 rs224224        T       C       225.009 .       DP=3796;VDB=0;SGB=-
0.693147;RPB=0.995193;MQB=1;MQSB=1;BQB=0.996552;MQ0F=0;AF1=0.5;AC1=1;DP4=1010,10,879,
24;MQ=60;FQ=225.007;PV4=0.00832428,0.267356,1,1;ASP;CAF=0.6134,.,0.3866;COMMON=1;G5;G
ENEINFO=MEFV:4210;GNO;INT;KGPhase1;KGPhase3;LSD;PM;PMC;REF;RS=224224;RSPOS=3254654;RV
;SAO=1;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x05016808030515013e100100;WGT=1;dbSNPBuildID=79;A
NN=C|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_coding|2/10|c.41
4A>G|p.Gly138Gly|454/3499|414/2346|138/781||,C|intron_variant|MODIFIER|MEFV|MEFV|tran
script|NM_001198536.1|protein_coding|1/8|c.277+1657A>G||||||       GT:PL
        0/1:255,0,255
chr16   3254762 rs224225        A       G       225.009 .       DP=3819;VDB=0;SGB=-
0.693147;RPB=0.109525;MQB=1;MQSB=1;BQB=0.0954619;MQ0F=0;AF1=0.5;AC1=1;DP4=960,43,896,
41;MQ=60;FQ=225.007;PV4=1,1,1,0.103684;ASP;CAF=0.6106,0.3894;COMMON=1;G5;G5A;GENEINFO
=MEFV:4210;GNO;HD;INT;KGPhase1;KGPhase3;LSD;PM;PMC;REF;RS=224225;RSPOS=3254762;RV;SAO
=1;SLO;SSR=0;SYN;VC=SNV;VLD;VP=0x05016808030517053f100101;WGT=1;dbSNPBuildID=79;ANN=G
|synonymous_variant|LOW|MEFV|MEFV|transcript|NM_000243.2|protein_coding|2/10|c.306T>C
|p.Asp102Asp|346/3499|306/2346|102/781||,G|intron_variant|MODIFIER|MEFV|MEFV|transcri
pt|NM_001198536.1|protein_coding|1/8|c.277+1549T>C||||||    GT:PL   0/1:255,0,255
```

APPENDIX B

PYTHON IMPLEMENTATION

```
#coding: utf-8
#batch MPCR primer pair finder

import os.path
import time
from Bio.Blast.Applications import NcbiblastnCommandline
from Bio.Blast import NCBIXML

import redis
from PrimerPair import PrimerPair
from MultiplexPairs import MultiplexPairs


import re
import primer3
from Bio.SeqUtils import MeltingTemp as mt
from Bio.Seq import Seq

productSizeLimit = 500
extremeLen = 120
tmLimit = 1
productSizeLowerLimit = 300

crossDimerLimit = 0
threePrimeEndLimit = -3
hairpinLimit = -3
homodimerLimit = -3
primerMin = 23
primerMax = 30
optimumGC = .5
optimumTm = 60


from tempDict import tempDict
r = tempDict()

print extremeLen
print productSizeLimit
print productSizeLowerLimit
print tmLimit
print primerMin
print primerMax
print optimumGC
print optimumTm

finishedSeq = []


def batchAnalysis(batchSeqFile, outputFile):
```

```
        finished = open(outputFile, 'r')
        line = ''
        for line in finished.readlines():
                finishedSeq.append( line.split()[0] )
        finished.close()


        batch = open(batchSeqFile, 'r')

        for line in batch.readlines():
                seq = line.split()[0]
                geneName = line.split()[1]
                if seq in finishedSeq:
                        pass
                else:
                        doAnalysis(seq, geneName, outputFile)



def doAnalysis(seq, geneName, outputFile):
        forwardCandidates, reverseCandidates = getSuitablePairs(seq)
        print
        print geneName
        print len(forwardCandidates)
        print len(reverseCandidates)

        result = seq + '\t' + geneName + '\t'
        for noOfTubes in range(2,6):  #2,3,4, and 5 tubes
                #print seq
                passingTime, noOfPairs = growPairs(seq, noOfTubes, forwardCandidates,
reverseCandidates, startPosition = 0)
                print str(noOfTubes) + '\t' + str(passingTime) + '\t' + str(noOfPairs)
                result += str(noOfTubes) + '\t' + str(passingTime) + '\t' +
str(noOfPairs) + '\t'
        result += '\n'
        output = open(outputFile, 'a')
        output.write(result)
        output.close()



def growPairs(seq, noOfTubes, forwardCandidates, reverseCandidates, startPosition =
0):
        start = time.time()
        timeout = 240

        multiplexStack = []
        listed = []
        multiplexCount = 0

        #for the beginning, choose primers starting from the leftmost
        for pairs in pairsInInterval(seq, forwardCandidates, reverseCandidates, 0,
extremeLen, 0):        #Zero for the 1st tube
                currMultiplexPrimers = MultiplexPairs()
                currMultiplexPrimers.add(pairs)
                multiplexStack.append(currMultiplexPrimers)


        count = 0
        newLimit = 0
        while len(multiplexStack):

                #check the time to exit if cannot find anything until now and report
the time
                if time.time() - start > timeout:
                        return time.time() - start, 0


                currPairs = multiplexStack.pop(0)
                #we are looking at the multiplex pairs, and decide if the last pair is
even or odd
```

```
                #isNextPairEven = not currPairs.getLastPair().isEven()
                #we identify the next tube to be continue with
                tubeNo = currPairs.getLastPair().getTubeNo()


                if currPairs.lastIndex() < len(seq) - extremeLen:
                        newCandidatePairs = getPossibleCombinations(seq,
forwardCandidates, reverseCandidates, currPairs, tubeNo,noOfTubes)
                        nextTubeNo = (tubeNo + 1) % noOfTubes
                        suitablePairs =
findSuitableMultiplexPairs(currPairs.getPrimersInTube(nextTubeNo), newCandidatePairs)
                        for pair in suitablePairs:
                                extendedPairs = MultiplexPairs()
                                extendedPairs.copy(currPairs)
                                extendedPairs.add(pair)
                                multiplexStack.insert(0, extendedPairs)
                else:
                        if any(currentListed == currPairs for currentListed in listed):
                                pass
                        else:   #found a multiplex! congrats :)
                                return time.time() - start, currPairs.len()
                                listed.append(currPairs)
                                multiplexCount += 1

                                print '>MULTIPLEX-'+ str(multiplexCount)

                                print currPairs
        return time.time() - start, 0



def getSuitablePairs(seq):
        forwardCandidates = getForwardCandidates(seq, primerMin, primerMax, 0)
        forwardCandidates = filterPrimers(forwardCandidates)
        reverseCandidates = getReverseCandidates(seq, primerMin, primerMax, 0)
        reverseCandidates = filterPrimers(reverseCandidates)
        return forwardCandidates, reverseCandidates



def pairsInInterval(seq, forwardCandidates, reverseCandidates, start, end, tubeNo):
        #print 'pairsInInterval' + str(tubeNo)
        pairs = []
        for fSeq, fPos in forwardCandidates:
                if fPos <= end and fPos >= start:
                        for rSeq, rPos in reverseCandidates:
                                #if rPos > fPos+len(fSeq)+1:
                                if rPos > end+primerMax:
                                        if checkPrimerPair(fSeq, fPos, rSeq, rPos):
                                                pair = PrimerPair(fSeq, fPos,
float(r.get(fSeq + ':tm')), rSeq, rPos, float(r.get(rSeq + ':tm')), tubeNo,
seq[fPos:rPos+len(rSeq)])
                                                pairs.append(pair)

        return pairs



def checkPrimerPair(forwardPrimer, forwardPosition, reversePrimer, reversePosition):
        forwardTm = float(r.get(forwardPrimer + ':tm'))
        productSize = reversePosition - forwardPosition + len(reversePrimer)
        if productSize <= productSizeLimit and productSize >= productSizeLowerLimit:
                reverseTm = float(r.get(reversePrimer + ':tm'))
                if abs(forwardTm - reverseTm) <= tmLimit:
                        crossDimerScore = checkCrossDimer(forwardPrimer, reversePrimer)
                        if crossDimerScore >= crossDimerLimit:
                                threePrimeEndScore = checkThreePrimeEnd(forwardPrimer,
reversePrimer)
                                if threePrimeEndScore:
```

```
                                        return True

        return False


def findSuitableMultiplexPairs(currPrimers, candidatePairs):
        #start = time.time()
        #print 'suitable multiplex basladi'
        suitablePairs = []
        if len(currPrimers) < 1:
                return candidatePairs

        for pair in candidatePairs:
                for seq1 in [pair.fwd(), pair.rev()]:
                        for seq2 in currPrimers:
                                crossDimerDg = checkCrossDimer(seq1, seq2)
                                if crossDimerDg >= crossDimerLimit:
                                        suitablePairs.append(pair)
        #print time.time() - start
        return suitablePairs


def getPossibleCombinations(seq, forwardCandidates, reverseCandidates, currPairs,
tubeNo,noOfTubes):
        #print 'getPossibleCombinations' + str(tubeNo)
        #start = time.time()
        #print 'suitable combinations basladi'
        suitablePairs = []
        #print currPairs
        currTubeLastPair = currPairs.getLastPairInTube( tubeNo)
        prevTubeLastPair = currPairs.getLastPairInTube( (tubeNo-1) % noOfTubes)
        newTubeLastPair = currPairs.getLastPairInTube( (tubeNo+1) % noOfTubes)
        candidatePairs = pairsInInterval(seq, forwardCandidates, reverseCandidates,
currTubeLastPair.fPos()+len(currTubeLastPair.fwd()), currTubeLastPair.rPos(),
((tubeNo+1) % noOfTubes) )
        for pair in candidatePairs:
                if prevTubeLastPair:
                        if pair.fPos() > prevTubeLastPair.lPos():
                                if newTubeLastPair:
                                        if pair.fPos() > newTubeLastPair.lPos() and
pair.rPos() > currTubeLastPair.lPos() and pair.fPos() <  currTubeLastPair.rPos()-
primerMax:
                                                suitablePairs.append(pair)
                                else:
                                        if pair.rPos() > currTubeLastPair.lPos() and
pair.fPos() <  currTubeLastPair.rPos()-primerMax:
                                                suitablePairs.append(pair)
                else:
                        if newTubeLastPair:
                                if pair.fPos() > newTubeLastPair.lPos() and pair.rPos()
> currTubeLastPair.lPos() and pair.fPos() <  currTubeLastPair.rPos()-primerMax:
                                        suitablePairs.append(pair)
                        else:
                                if pair.rPos() > currTubeLastPair.lPos() and
pair.fPos() <  currTubeLastPair.rPos()-primerMax:
                                        suitablePairs.append(pair)

        return suitablePairs



def checkCrossDimer(seq1, seq2):
        (seq1, seq2) = sorted([seq1, seq2])
        crossDimer = float(r.get(seq1 + ':' + seq2 + ':crossDimer') or 1000)
        if crossDimer > 999:
                crossDimer = primer3.calcHeterodimer(seq1, seq2, mv_conc=50,
dv_conc=1.5, dntp_conc=0.6, dna_conc=50, temp_c=(optimumTm-5)).dg / 1000
                r.set(seq1 + ':' + seq2 + ':crossDimer', crossDimer + 0.0000001)
        if crossDimer < 0:
```

```
                return crossDimer
        else:
                return 0


#gets the list of forward primer candidates for a given sequence
def getForwardCandidates(seq, primerMin, primerMax, distance):
    forwardList=[]

    for primerLen in range(primerMin, primerMax+1):
        for position in range(0,(len(seq)-primerLen+1)):
            candidate = seq[position:position+primerLen]
            forwardList.append( (candidate, distance+position) )
    return sorted(forwardList, reverse=False, key=lambda item: item[1])



#gets the list of reverse primer candidates for a given sequence
def getReverseCandidates(seq, primerMin, primerMax, distance):
    reverseList=[]
    revCompSeq = revComp(seq)

    for primerLen in range(primerMin, primerMax+1):
        for position in range(0,(len(revCompSeq)-primerLen+1)):
            candidate = revCompSeq[position:position+primerLen]
            reverseList.append( (candidate, len(revCompSeq) - position - primerLen +
distance)  )
    return sorted(reverseList, reverse=False, key=lambda item: item[1])



#returns the reverse complement of given sequence
def revComp(seq):
    tr = {'A':'T', 'T':'A', 'G':'C', 'C':'G'}
    revComp = ''
    for base in seq[::-1]:
        revComp += tr[base] if base in tr else 'X'

    return revComp



def getRegionSeq(gene):
        regionSequence = open(gene+'.txt', "r").read()
        return ''.join(regionSequence.split("\n"))



#################
#   filtering   #


#filtering given primer candidates for gc content, tm, and other criteria
def filterPrimers(primerCandidates):
        passingFilters = []
        for seq, position in primerCandidates:
                primerCheck = getPrimerScore(seq, optimumGC, optimumTm)
                if primerCheck:
                        passingFilters.append((seq, position))


        return passingFilters



#Calculates a score for a given primer sequence
def getPrimerScore(seq, optimumGC, optimumTm): #main function in filtering

        baseCount = {}
        baseCount['G'] = seq.count('G')
```

```
        baseCount['C'] = seq.count('C')
        baseCount['T'] = seq.count('T')
        baseCount['A'] = seq.count('A')
        baseCount['total'] = len(seq)

        gcScore = getGcScore(baseCount, optimumGC)

        tmScore = getTmScore(seq, optimumTm)

        hpScore = getHomopolymerScore(seq)

        lbScore = checkLastBases(seq)

        hrScore = checkHairpin(seq)

        hdScore = getHomodimerScore(seq)

        tpScore = checkThreePrimeEnd(seq, seq)

        #return totalScore
        return gcScore and tmScore and hpScore and lbScore and hrScore and hdScore and
tpScore



#penalize deviation from optimum gc rate
def getGcScore(baseCount, optimumGC):
        gcRate = 1.0 * (baseCount['G'] + baseCount['C']) / baseCount['total']

        if abs(gcRate - optimumGC) <= 0.1:
                return True

        return False



#penalize deviation from optimum tm
def getTmScore(seq, optimumTm):
        tm = float(r.get(seq + ':tm') or 1000)
        if tm > 999:
                tm = primer3.calcTm(seq, mv_conc=50, dv_conc=1.5, dntp_conc=0.6,
dna_conc=50)
                r.set(seq + ':tm', tm)

        if abs(tm-optimumTm) <= 0.5:
                return True

        return False



#penalize homopolymer existence harshly
#optimizing function to get rid of regex, which results in 200 fold performance
increase!
def getHomopolymerScore(seq):
        if seq.count('AAAA'):
                return False
        if seq.count('TTTT'):
                return False
        if seq.count('GGGG'):
                return False
        if seq.count('CCCC'):
                return False

        return True



#penalize if last 3 of 5 bases are G or C
def checkLastBases(seq):
        totalGC = seq[-5:].count('G') + seq[-5:].count('C')
```

```
        if totalGC >3:
                return False

        return True



#penalize with hairpin deltaG
def checkHairpin(seq):
        hairpinDg = float(r.get(seq + ':hairpin') or 1000)
        if hairpinDg > 999:
                hairpinDg = primer3.calcHairpin(seq, mv_conc=50, dv_conc=1.5,
dntp_conc=0.6, dna_conc=50, temp_c=(optimumTm-5)).dg / 1000
                r.set(seq + ':hairpin', hairpinDg + 0.0000001)

        if hairpinDg < -3:
                return False
        return True



#penalize if a homodimer exists
def getHomodimerScore(seq):
        dimerDg = float(r.get(seq + ':homodimer') or 1000)
        if dimerDg > 999:
                dimerDg = primer3.calcHomodimer(seq, mv_conc=50, dv_conc=1.5,
dntp_conc=0.6, dna_conc=50, temp_c=(optimumTm-5)).dg / 1000
                r.set(seq + ':homodimer', dimerDg)
        if dimerDg < -3:
                return False

        return True



def checkThreePrimeEnd(seq1, seq2):
        (seq1, seq2) = sorted([seq1, seq2])
        end_dG = float(r.get(seq1 + ':' + seq2 + ':end_dG') or 1000)
        if end_dG > 999:
                end_dG = primer3.bindings.calcEndStability(seq1, seq2, mv_conc=50,
dv_conc=1.5, dntp_conc=0.6, dna_conc=50, temp_c=(optimumTm-5)).dg / 1000
                r.set(seq1 + ':' + seq2 + ':end_dG', end_dG)

        if end_dG < -3:
                return False

        return True



def showSuitablePositions(seq):
        forwardCandidates = getForwardCandidates(seq, primerMin, primerMax, 0)
        forwardCandidates = filterPrimers(forwardCandidates)
        reverseCandidates = getReverseCandidates(seq, primerMin, primerMax, 0)
        reverseCandidates = filterPrimers(reverseCandidates)
        fTemp = []
        rTemp = []
        out = ''
        for f, p in forwardCandidates:
                fTemp.append(p)
        for r, p in reverseCandidates:
                rTemp.append(p)
        for i in range(len(seq)):
                if i % 100 == 0:
                        out +=  '\n' + str(i) + '\t'
                if i % 20 == 0:
                        out +=  ' '
                if i in fTemp and i in rTemp:
                        out += '@'
                elif i in fTemp:
                        out += '+'
```

```
                elif i in rTemp:
                        out += '-'
                else:
                        out += '.'

        print out


batchAnalysis('exonSequences_2240_to_2340.txt', 'bothNormal.txt')
```

```
#CLASSES

#Class 1

import numpy

class MPCRExperiment:

        def __init__(self, method,  seq, gene, sec2, tube2, sec3, tube3, sec4, tube4,
sec5, tube5):
                self.data = {}
                self.data['method'] = method
                self.data['seq'] = seq
                self.data['gene'] = gene
                self.data['sec2'] = sec2
                self.data['tube2'] = tube2
                self.data['sec3'] = sec3
                self.data['tube3'] = tube3
                self.data['sec4'] = sec4
                self.data['tube4'] = tube4
                self.data['sec5'] = sec5
                self.data['tube5'] = tube5


        def new(self, method,  seq, gene, sec2, tube2, sec3, tube3, sec4, tube4, sec5,
tube5):
                pass


        def getParameter(self, parameter):
                return self.data[parameter]


        def getMethod(self):
                return self.data['method']


        def getSafeSec(self, tube):
                if self.data['tube'+str(tube)]>0:
                        return self.data['sec'+str(tube)]
                return -1


        def checkTubeImprovement(self):
                if numpy.std(self.getPairs()) > 0:
                        return True
                return False


        def getPairs(self):
                return [self.data['tube'+str(i)] for i in range(2,6)]


        def getPair(self, tube):
                if self.data['tube'+str(tube)] > 0 :
                        return self.data['tube'+str(tube)]
                return -1


        def getSeconds(self):
                return [self.getSafeSec(i) for i in range(2,6)]


        def getOnlyFoundSeconds(self):
                return [self.getSafeSec(i) for i in range(2,6) if
self.getSafeSec('tube'+str(i)) > 0]


        def getSecImprovementCV(self):
                return 100.0 * numpy.std(self.getOnlyFoundSeconds()) /
(numpy.mean(self.getOnlyFoundSeconds())+0.0001)
```

```
#Class 2

class MultiplexPairs:
	from PrimerPair import PrimerPair


	def __init__(self):
		self.primerPairs = []
		self.lastPosition = 0
		self.lastPair = ''


	def add(self, primerPair):
		if primerPair.rPos() > self.lastPosition:
			self.lastPosition = primerPair.rPos()
			self.lastPair = primerPair
		self.primerPairs.append(primerPair)


	def len(self):
		return len(self.primerPairs)


	def pairs(self):
		return self.primerPairs


	def primers(self):
		allPrimers = []
		productSizes = []
		for pair in self.primerPairs:
			allPrimers.append(pair.fwd())
			allPrimers.append(pair.rev())
			productSizes.append(pair.productSize())
		return allPrimers, productSizes


	def getPrimersInTube(self, tubeNo):
		allPrimers = []
		for pair in self.primerPairs:
			if pair.getTubeNo() == tubeNo:
				allPrimers.append(pair.fwd())
				allPrimers.append(pair.rev())
		return allPrimers


	def getPairsInTube(self, tubeNo):
		allPairs = []
		for pair in self.primerPairs:
			if pair.getTubeNo() == tubeNo:
				allPairs.append(pair)
		return allPairs


	def lastIndex(self):
		return self.lastPosition


	def copy(self, multiplexPairs):
		for pair in multiplexPairs.pairs():
			self.add(pair)


	def getLastPair(self):
		return self.lastPair


	def primerPairList(self):
		return self.primerPairs
```

```
        def __str__ (self):
                output = 'Number of suitable primer pairs \t_' + str(self.len()) +
'_\n'
                json = '>'+str(self.len())+'\n'
                for pair in self.primerPairList():
                        output += str(pair.getTubeNo()) + ': ' + pair.fwd() + ' ('+
str(pair.fTm()) + ')\t -> ' + pair.rev() + ' (' + str(pair.rTm()) +  ')\t( ' +
str(pair.fPos()) + '-' + str(pair.rPos()) + ')\t( ' + str(pair.productSize()) + '
bp)\n'
                        json += '{\n"tube":'+str(pair.getTubeNo()+1) + ',\n"forward":"'
+ pair.fwd()+'",\n"forwardTm":'+ str(pair.fTm())
                        json += ',\n"reverse":"' + pair.rev()+'",\n"reverseTm":'+
str(pair.rTm())
                        json += ',\n"forwardPos":' + str(pair.fPos()) +
',\n"forwardLen":' + str(len(pair.fwd()))
                        json += ',\n"reversePos":' + str(pair.rPos()) +
',\n"reverseLen":' + str(len(pair.rev()))
                        json += '\n},'
                output += 'Last index: \t' + str(self.getLastPair().fPos()) + ', ' +
str(self.getLastPair().rPos()) + '\n'
                return json#output


        def __eq__ (self, other):
                selfList = self.primerPairList()
                otherList = other.primerPairList()
                if len(selfList) == len(otherList):
                        for i in range( len(selfList) ):
                                if selfList[i] == otherList[i]:
                                        pass
                                else:
                                        return False
                else:
                        return False

                return True


        def getLastPairInTube(self, tubeNo):
                pairs = self.getPairsInTube(tubeNo)
                lastPair = None
                if len(pairs) > 0:
                        maxRPos = 0
                        for pair in pairs:
                                if pair.rPos() > maxRPos:
                                        maxRPos = pair.rPos()
                                        lastPair = pair
                        return lastPair
                else:
                        return pairs
```

```
#Class 3

class PrimerPair:

	def __init__(self, forwardPrimer, forwardPosition, forwardTm, reversePrimer,
reversePosition, reverseTm, tubeNo, productSeq):
		self.forwardPrimer = forwardPrimer
		self.forwardPosition = forwardPosition
		self.forwardTm = forwardTm
		self.reversePrimer = reversePrimer
		self.reversePosition = reversePosition
		self.reverseTm = reverseTm
		self.tubeNo = tubeNo
		self.productSeq = productSeq


	def fwd(self):
		return self.forwardPrimer


	def rev(self):
		return self.reversePrimer


	def fPos(self):
		return self.forwardPosition


	def rPos(self):
		return self.reversePosition


	def lPos(self):		#last position
		return self.reversePosition + len(self.reversePrimer)


	def __str__ (self):
		return self.forwardPrimer + ' -> ' + self.reversePrimer + ' (' +
str(self.forwardPosition) + ', ' + str(self.reversePosition) + ')'


	def productSize(self):
		return self.reversePosition - self.forwardPosition +
len(self.reversePrimer)


	def getTubeNo(self):
		return self.tubeNo


	def fTm(self):
		return self.forwardTm


	def rTm(self):
		return self.reverseTm


	def getProductSeq(self):
		return self.productSeq
```

```
#Class 4

class tempDict:

        def __init__(self):
                self.temp = {}


        def set(self, key, value):
                self.temp[key] = value


        def get(self, key):
                if self.temp.has_key(key):
                        return self.temp[key]
                else:
                        return False
```

CURRICULUM VITAE

WORK EXPERIENCE

April 2015 — **Diagnosis Diagnostic Systems Inc., Turkey**
- Founder & CEO, Bioinformatics & Biotechnology Company
- EU funded "Newborn Genetic Check-Up" project

September 2014 — **Bielefeld University, Bielefeld, Germany**
- Visiting Instructor (Gastdozent)
- Data Visualization with D3 Intensive Course

Februrary 2014 – ... (http://en.genkok.com/) — **Genome and Stem Cell Center, Kayseri, Turkey**

- Bioinformatics Group
- Managing high throughput data analysis procedures, training, and team building

2013 – ... — **Blogger**
- Turkish blog on Bioinformatics, to promote knowledge in Turkish: http://biyoinformatiktr.blogspot.com
- English blog on business, data, and visualization: http://bizdataviz.blogspot.com

2011 – … — **AG Bioinformatics Technopreneurship Ltd. Co., Turkey**
- Founder and Bioinformatician, the first bioinformatics company in

Turkey

- The analysis, interpretation and visualization of biological high-throughput data

2009 – 2011 — **The Scientific and Technological Research Council of Turkey**

- Scientific Programs Deputy Expert in Health Sciences Group
- Managing research project applications and funding procedures
- Project: Network Analysis and Visualization of Previous Funding Applications and Collaborations
- Project: Online bureau-informatics tools for the routine legal letters

2008 – 2009 — **AG Bioinformatics Technopreneurship Ltd. Co., TR**
- Founder and Bioinformatician, the first bioinformatics company in

Turkey

- The analysis, interpretation and visualization of biological high-throughput data

2003 – 2007 — **Bilkent University Computer Center, TR**
- Volunteer work for network administration
- Development of scripts for the control and management of the internet usage in dormitories
- Development of various scripts for the in-house IT infrastructure of Bilkent University

EDUCATIONAL BACKGROUND

2011 – **PhD, Middle East Technical University, Informatics Institute, Ankara, TR**
- Medical Informatics
- Ongoing thesis: Network Analysis of Next Generation Sequencing Data
- Expected to graduate: June-2016

2007 – 2010 **MSc, Bilkent University, Engineering and Science Institute, Ankara, TR**
- Molecular Biology and Genetics Department, full scholarship
- Thesis: Effects of Microarray Data Normalization in the Context of Network Analysis

2001 – 2007 **BSc, Bilkent University, Science Faculty, Ankara, TR**
- Molecular Biology and Genetics Department, full scholarship
- 490th in university entrance exam among 1.5 million students

1998 - 2001 **Kayseri Science High School, Kayseri, TR**
- 3rd best high school in Turkey
- Science project: Effects of Urtica dioica (stinging nettle) seeds on blood paramaters of healthy rats

TEACHING EXPERIENCE

2014 June-August **Bioinformatics Summer School**
- Transformation of non-programmer molecular biology students to Python, Django, and D3 programmers for bioinformatics, 13 fellows
- Extensive documentation using a daily blog, a wiki, and iPython Notebook.

2014 May **Erasmus IP Program, Hungary**
- Instructor, Bioinformatics in Stem Cell Research

2013 September **Interactive Data Visualizations**
- Instructor

2011 - **Professional Trainer**
- Bioinformatics, Programming, and Data Visualization
- Perl, Python, R, Bioinformatics Web Tools

2007 – 2010 **Teaching Assistant (TA)**
- Bioinformatics Lab, 2009-2010 Fall in Bilkent University
- Genomics Lab, 2007-2008 & 2008-2009 Spring in Bilkent University

Seminars for Undergraduates
- Data Analysis and Interpretation Approaches in Molecular Biology
- More than 20 seminars

Training Undergraduates for Data Analysis
- Data Analysis and Interpretation

SCIENTIFIC INTERNSHIPS

2006 — **Bioinformatics Internship in Emili Lab at Toronto University Best and Banting Institute, ON, CA**

- Mouse Heart Disease Biomarker Project
  - Creating annotation data,
  - Developing Cytoscape plug-ins
- Converting BIND protein-protein interaction data into PSIMI1.0 format
- Installing and building local cPath database

2005 — **Bioinformatics Internship in Sabanci University, TR**

- Transcription Factor Binding Determination using Classification Algorithms

2004 – 2007 — **Undergraduate bioinformatics research in Bilkent University, TR**

- Annotation, regulation and compartmentalization of gene expression complexes
- Development of various on-line biological scripts for the needs of the department.

2000 — **Research in DEKAM (Experimental and Clinic Research Center) in Erciyes University, TR**

- Research on the effects of Urtica dioica (stinging nettle) on blood parameters of healthy rats
- Familiar with treating rats and taking various tissue samples.
- Experienced with data parsing/filtering from online databases

SKILLS AND INTERESTS

Languages (beginner) — Turkish (native), English (fluent), French (beginner), German

Computing skills

OS: Experienced with various Windows and Linux OS

Programming languages:

Experienced: Python, D3, Perl, JavaScript, SQL, HTML, CSS

Moderate: Matlab, R, Django

Experienced with:

- Data Analysis and Visualization
- Cytoscape
- vi editor, Microsoft Excel, bioinformatics web tools
- Software Engineering Best Practices
- Text Mining

Molecular Biology

Bioinformatics

High-throughput data analysis (SNP and expression microarrays, NGS, qPCR), gene regulatory networks, comparative genomics, biological data mining, graph visualization, clustering and classification

Wet Lab Techniques

qPCR, DNA and RNA purification, Western Blotting, gene cloning, recombinant DNA technology, Zebrafish

General Interests

Bioinformatics, Molecular Interactions and Network Analysis, High-throughput Data Analysis, Visualization, and Interpretation, Systems Biology, Artificial Intelligence, Text Mining, NLP, Social Psychology, Photography

CONFERENCES

*Participated as an **Invited Speaker**, selected events:*

CELLmicrocosmos neXt Workshop in context of the "German Conference on Bioinformatics", Bielefeld, Germany (28/09/2014) Web-based Visualization with D3

Multidisciplinar R&D and Innovation in Life Sciences, Istanbul University, Istanbul (15-17 June 2012) Bioinformatics

HIBIT 2010 (International Symposium on Health Informatics and Bioinformatics), Antalya (20 – 22 April 2010) Bioinformatics Workshop

2009 Summer Computer Simulation Conference, Grand Cevahir Hotel and Convention Center, Istanbul (13-16 July 2009) Bioinformatics Workshop

4. Molecular Biotechnology Summer School, Karadeniz Technical University, Trabzon (30 April-3 May 2009) Microarray Technology and Data Analysis

A story of a technopreneurship: AG Bioinformatics, Ankara University Biotechnology Institute, Ankara (3 December 2008)

Participated events, selected:

HIBIT conference 09 (International Symposium on Health Informatics and Bioinformatics) METU, Ankara, Turkey (16-17 April 2009)

HIBIT conference 08 (International Symposium on Health Informatics and Bioinformatics) Sabanci University, Istanbul, Turkey (18-20 May 2008)

“Statistical Methods in Bioinformatics” Conference; Max Planck Institute, Germany (22-23 November 2007)

“Practical DNA Microarray Analysis 2007” Practical Course; Max Planck Institute, Germany (26-29 November 2007)

“Systems Biology: Will It Work?” Meeting; organized by Biochemical Society in Sheffield University, UK (12-14 January 2005).

HIBIT conference 05 (International Symposium on Health Informatics and Bioinformatics, Turkey); paper presentation (10-12 November 2005)

Laboratory Animals Workshop, Erciyes University – Kayseri/Turkey (2000)

PUBLICATIONS

Guclu S, **Ozturk AR**, Atalay A. Analysis of global microRNAome profiles of Caenorhabditis elegans oocytes and embryos. Turkish J. Biol. 2016 Mar 25

Bayrakli F, Guclu B, Yakicier C, Balaban H, Kartal U, Erguner B, Sagiroglu MS, Yuksel S, **Ozturk AR**, Kazanci B, Ozum U, Kars HZ. Mutation in MEOX1 gene causes a recessive Klippel-Feil syndrome subtype. BMC Genet. 2013 Sep 28;14:95.

Gur-Dedeoglu, B, Konu O, Kır S, **Ozturk AR**, Bozkurt B, Gulusan E, Yulug I. A resampling-based meta-analysis for detection of differential gene expression in breast cancer. BMC Cancer 8, 396(2008).

Workshop Proceeding: Web-based Visualization with D3 (**Ahmet R. Ozturk**)

+ Accepted for German Conference on Bioinformatics, published on September 28th, 2014.

Conference Proceeding: Development of A Web-Interface For Construction and Processing of Microarray Meta Datasets (**Ahmet R. Ozturk**, Can U. Ayfer, Ozlen Konu. Bilkent University / Turkey)

+ Accepted for HIBIT conference, published on November 12th, 2005.