2006

Computational approach to identify deletions or duplications within a gene

Krishna Rani Kalari
University of Iowa

Recommended Citation

Kalari, Krishna Rani. "Computational approach to identify deletions or duplications within a gene." PhD (Doctor of Philosophy) thesis, University of Iowa, 2006.
https://doi.org/10.17077/etd.cmf6apwq

Follow this and additional works at: https://ir.uiowa.edu/etd
Part of the Biomedical Engineering and Bioengineering Commons

COMPUTATIONAL APPROACH TO IDENTIFY DELETIONS OR DUPLICATIONS WITHIN A GENE

by

Krishna Rani Kalari

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

December 2006

Thesis Supervisors: Assistant Professor Todd E. Scheetz
Professor Thomas L. Casavant

ABSTRACT

Although high-throughput methods exist to identify most small disease causing mutations (e.g. substitutions that alter an amino acid), assays to identify larger classes of mutations such as deletions/duplications are time consuming, laborious and expensive. No *in-silico* system exists to identify intragene deletion or duplication candidates. We hypothesize that a computational system, SPeeDD (System to Prioritize Deletion or Duplication candidates), utilizing machine learning techniques can be employed to identify the most likely disease causing deletion or duplication candidates within a gene.

Informative sequence based features were obtained from a set of genes with known intragene deletions or duplications for data mining. Machine learning techniques were applied to this data. The logic model tree (LMT) method, which is a combination of decision tree and logistic regression model, yielded the best results. Sensitivity varied depending on the type of machine learning model used, but specificity exceeded 90% for all methods evaluated. Sensitivity of the system ranged from 20% to 71.6% depending on the type of machine learning method. We were also able to find the new BRCA1 case using our system.

These results suggest that the SPeeDD system provides good sensitivity and specificity and can be used to prioritize candidate genes and gene regions for screening. Focused screening for copy number variations in prioritized regions will reduce the labor and associated costs of the biological assays, and should accelerate the process of mutation discovery.

Abstract Approved: _____________________________________

Thesis Supervisor

____

Title and Department

____

Date

____

Thesis Supervisor

____

Title and Department

____

Date

COMPUTATIONAL APPROACH TO IDENTIFY DELETIONS OR DUPLICATIONS WITHIN A GENE

by

Krishna Rani Kalari

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

December 2006

Thesis Supervisors: Assistant Professor Todd E. Scheetz
Professor Thomas L. Casavant

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Krishna Rani Kalari

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Biomedical Engineering at the December 2006 graduation.

Thesis Committee: ______________________________
Todd E. Scheetz, Thesis Supervisor

____
Thomas L. Casavant, Thesis Supervisor

____
Terry A. Braun

____
Josep M. Comeron

____
Edwin M. Stone

To
my parents

ACKNOWLEDGMENTS

First and foremost I would like to acknowledge my thesis advisors Drs. Scheetz and Casavant for their valuable suggestions through out my study. I have learnt a lot about the field of bioinformatics and about research from them. Next I would like to acknowledge my committee members Dr. Braun, Dr. Comeron and Dr. Stone for giving their valuable suggestions during my study. I would also like to thank all the past and present members of CLCG and CBCB for their support.

I would like to thank my parents, Vasanta and Subbarao who have always loved and encouraged me. Thank you, mom and dad for showing me how important learning is. I would also like to thank my brother Chandra Sekhar for all his encouragement, love and support. Next, my endless thanks go out to my husband Karunya. His encouragement and believe in me are truly amazing. My special thanks to my five-year old daughter Mahathi and two year old daughter Vidushi for their love and support. Finally I would also like to thank rest of my family and friends whose encouragement and wishes were with me through out my study.

ABSTRACT

Although high-throughput methods exist to identify many small disease causing mutations (e.g. substitutions that alter an amino acid), assays to identify classes of larger mutations such as deletions/duplications are time consuming, laborious and expensive. In addition, no *in-silico* system exists to identify intragene deletion or duplication candidates. We hypothesize that a computational system, SPeeDD (System to Prioritize Deletion or Duplication candidates), utilizing machine learning techniques can be employed to identify the most likely disease causing deletion or duplication candidates within a gene.

Informative sequence based features were obtained from a set of genes with known intragene deletions or duplications for data mining. Machine learning techniques were applied to this data. Sensitivity from 20% to 74.2% varied depending on the specific machine learning model used, but specificity exceeded 90% for all methods evaluated. The logic model tree (LMT) method, which is a combination of decision tree and logistic regression model, yielded the best results. The SPeeDD system also succeeded in accurately predicting a recently published novel BRCA1 deletion.

These results suggest that the SPeeDD system provides good sensitivity and specificity and can be used to prioritize candidate genes and gene regions for focused screening. This will reduce the labor and associated costs of the biological assays, and should accelerate the process of mutation discovery.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Statement of problem

Variations in the genomic sequence, known as mutations, can lead to serious conditions that may be passed on to progeny. The most critical mutations are those that alter the coding potential or regulation of a gene. Most commonly these are amino acid substitutions and copy number changes (duplications or deletions of a genomic region). Changes in copy number are most often the result of segmental duplication or deletion of sequences within the genome and in several cases have been found to cause a variety of genetic disorders. Although high-throughput techniques exist for molecularly identifying substitution mutations, the methods available for the identification of deletions or duplications are often more time consuming, labor and resource intensive and expensive. Hence, *in silico* prediction of those deletion or duplication candidates most likely to occur within a particular gene can be an effective strategy to enable investigators to focus on the highest-quality candidates. The objective of this research is to:

Design and implement a high-throughput computational system to identify and prioritize candidate intragene deletions or duplications and to make that information readily available to genetic and biomedical researchers.

The computational system developed as part of this research and described below (SPeeDD; System to Prioritize Deletions and Duplications) evaluates genomic features associated with previously published disease-associated IDDs to predict the most likely set of IDDs for a given gene of interest.

Strategy and objectives of this research

In this study, the most likely recombination events resulting in deletion or duplication mutations are estimated using machine learning techniques. This computational and machine learning system is trained on a set of previously published, biologically verified deletions or duplications. The reliability of machine learning systems is dependent on the size and quality of the training data set. To construct the largest, high-quality training set, I collected information on 1463 previously published gross rearrangements from the Human Genome Mutation Database (HGMD). However, due to the challenges involved in identifying the exact breakpoints of IDDs, many of the publications do not report fully characterized breakpoints, and thus a portion of the published information is ambiguous. To ensure quality in the training set, only those mutations whose breakpoints had been fully-characterized at the sequence level were included in the training set. This study consists of 102 previously published deletions or duplications occurring at different gene loci along with 2338 matched control sequences derived from the same regions as the previously published cases.

Features previously implicated in unequal homologous recombinations are calculated for the breakpoints in the training set. Features include sequence-based features such as GC content of the breakpoints and the distance between the breakpoints, as well as melting temperature features and haplotype block information. These features were used to annotate the training set of biologically verified IDDs, and the matched control set. The annotated training set was then used to train the SPeeDD system to predict likely intragene duplications and deletions (IDDs) within a gene and to prioritize the candidate IDDs with regards to the likelihood of observing an unequal recombination event.

Purpose of the computational system

Rather than conducting biological experiments on the hundreds of candidate IDDs with an exhaustive search, SPeeDD computationally narrows the list of candidates and determines the list of candidates most likely to lead to unequal recombination. Using SPeeDD to prioritize the screening of IDD candidates will aid investigators to focus their research on the most likely candidates, reducing the labor and associated costs of the biological assays, and accelerating the process of mutation discovery.

SPeeDD is an efficient, unbiased method to rank candidate intragene deletion or duplication regions. The performance of the system has been computationally validated, and has also correctly identified a recently published novel duplication. The SPeeDD system is available as a web based system that enables the user to identify high priority candidate regions for a given gene. An implementation of the SPeeDD system is readily available on the web at http://public.eng.uiowa.edu/SPeeDD.

CHAPTER II

BACKGROUND

This thesis presents a novel computational system to identify and prioritize intragene deletions and duplications. This chapter begins with a brief review of the human genome. A summary of mutation categorization is then presented to provide context for the types of mutations under study. It provides information about how recombination leads to deletions or duplications, how significant regions of homology play a role during recombination, types of significant regions available such as the repetitive sequences in the human genome, how abundance of repeats and crossing over leads to human disorders, provide summary of molecular assays that are used to screen deletions or duplications, provide information about how databases and other resources are used to build a computational system. It also provides the details of the how knowledge discovery methods such as machine learning methods are used to study the trends in the data. Finally at the end of the chapter we have discussed about the previous studies done and also discussed about the significance of their work.

Human genome

The human genome contains 23 pairs of chromosomes, of which 22 pairs are autosomes and one pair of sex chromosome (females have two X chromosomes, males have one X and one Y chromosome). A human cell contains around 25,000 to 30,000 genes. Genes carry information from one generation to the next in terms of A, C, G and T nucleotides. In humans as well as most eukaryotes, genes consist of alternating exons and intron only the exons contain information pertaining to protein coding. During the process of transcription the entire region in which a gene resides is transcribed (both exons and introns). The introns are then removed (spliced) from this pre-mRNA to make the final product, a mature messenger RNA or mRNA. The resulting mRNA is then typically utilized in the process of translation to create a specific protein. This process is illustrated in Figure 1.

Changes in the genome are called mutations. Such changes may occur in somatic or germ cells. The important difference being that mutations in germ cells are potentially heritable, where as mutations in somatic cells is not heritable.

Although it is known that mutations can cause a change in an observable characteristic (phenotype), but not all mutations cause a change in the phenotype. Mutations in general fall into several categories. The spectrum of mutations extends from small single-base changes to large-scale changes in copy number or chromosomal structure. Below is a discussion of the types and examples of the mutations available.

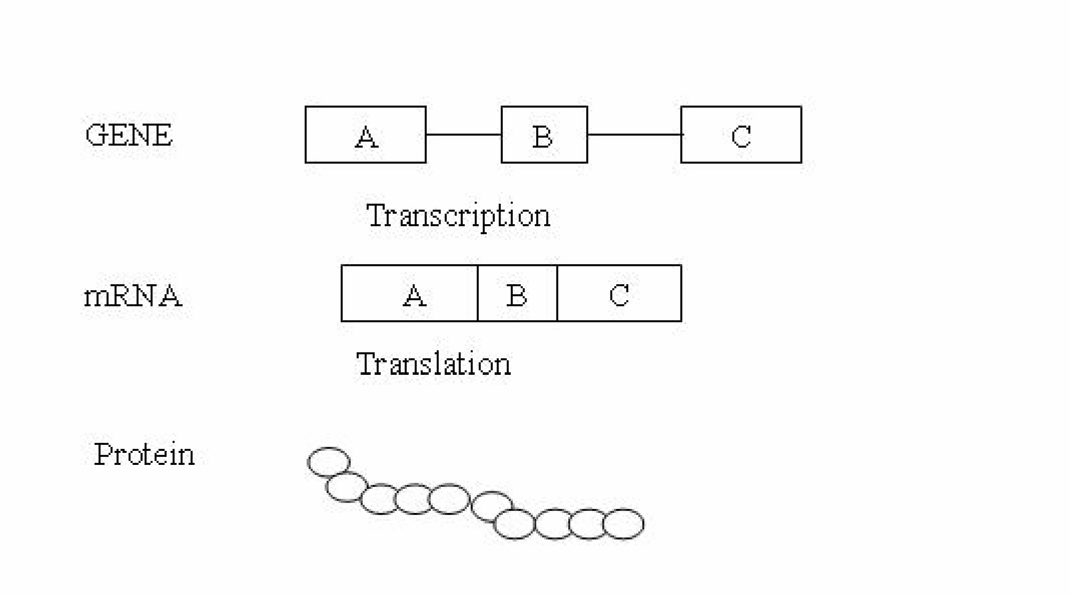


Figure 1 Central Dogma

Types of mutations

Different types of mutations may have no effect on the organism, or may have one or a spectrum of effects. In addition, different types of mutations affect the DNA and corresponding protein sequence differently. Mutations may lead to nutritional or biochemical variation, or changes in a morphological trait, behavior, change in gene regulation or may have no effect.

Single-base substitutions

Mutations affecting a single base are also called as point mutations. As the name indicates only a single nucleotide base gets substituted by another. If a purine (A or G) is replaced by a purine or pyrimidine (C or T) is replaced by a pyrimidine it is called a transition. Similarly, in cases where a purine is replaced by a pyrimidine or vice-versa, the mutation is referred to as a transversion. Single-base pair mutations that lie in coding regions may also be sub-divided into three additional classes. They are missense, nonsense and silent mutations. All these mutations are shown in Table 1.

Missense mutations:

Missense mutation is a change in nucleotide position that causes a change from one amino acid to another amino acid. This mutation results in a different protein product. For example: In sickle-cell disease the replacement of A by T at the 17th position in the gene beta chain of hemoglobin changes the codon from glutamic acid to valine.

Nonsense mutations:

The change in nucleotide position causes an amino acid to change to one of the STOP codons (TAA, TAG, or TGA) and causes the protein to end prematurely. If this mutation occurs in the earlier stage of gene translation then more protein will be lost or truncated. For example: In some of the cystic fibrosis patients a change in C to a T at nucleotide 1609 position changes the codon. It converts a glutamine codon to stop codon and makes an abnormal protein.

Silent mutations

The change in nucleotide does not always cause a change in the amino acid. There are different codons that code for same amino acid base. During the process of translation in silent mutations the amino acid does not change even though there is a change in the nucleotide base. In such cases, we do not see a change in protein.

Table 1. Single base substitutions a) Missense mutations b) Nonsense mutations c) Silent mutations

Mis sense Mutation	Nonsense Mutation	Silent Mutation
TGT → TGG	TGT → TGA	TGT → TGC
Cys → Trp	Cys → Stop	Cys → Cys

Changes in chromosomal structure

Chromosomal mutations refer to a change in the structure of the chromosomes. These mutations occur during the crossing over period of meiosis. There are six different types of structural changes that lead to various types of mutations. They are expansion-contraction type polymorphisms, insertions, deletions, duplications, inversions and translocations which are shown in Figure 2.

Expansion-contraction type polymorphisms

Expansion-contraction type polymorphisms are caused due to slipped strand mispairing in microsatellites and during unequal crossovers in large units of tandemly repeated DNA. For example: Expansion of the CGG triplet in the fragile X syndrome in the FMR-1 gene.

Insertions

This mutation adds extra DNA into the existing genome. One of the reasons for the occurrence of these mutations is due to the presence of transposable elements. Insertions of transposable elements into a gene may cause frameshifts leading to a bad protein. Insertions or deletions caused with multiples of three bases may be less serious because they do not change the reading frame whereas the insertion or deletions that are not multiples of three bases can change the reading frame and may produce an abnormal protein. For example: An addition of 'CAG' nucleotides to the Huntington gene produces a bad protein that interferes with synaptic transmission in parts of the brain and leads to loss of motor control in the Huntington disease.

Deletions

Deletion mutation is defined as the loss of DNA from the genome. The number of bases deleted may range from a few to thousands. There are primarily three different types of deletions, they are unequal crossover, unequal sister chromatid exchange and intrachromatid recombination involving direct repeats. Deletions can be homozygous or heterozygous. But if the deleted region is essential to life then the homozygous deletion would be lethal. Heterozygous deletions can be lethal or nonlethal. For example: Alu-mediated 7.1 deletions found in BRCA1 in breast and ovarian cancer families.

Duplications

Duplication mutation is a doubling of a section of the genome. These may are very important chromosomal changes in evolution because they supply with the additional genetic information which might be capable of having new function. Duplication occurs in directly repeated genes and intergenic direct repeats. Crossing over between sister chromatids during meiosis may cause an out of alignment and lead to a chromatid with a duplication or a deletion. For example: unequal crossing over created a second copy of a gene in the steroid hormone aldosterone.

Inversions

In case of an inversion mutation the whole section of DNA is reversed. Small inversion involves few bases within a gene whereas a longer inversion involves large regions of chromosome. Inversions are caused by intrachromatid recombination between inverted repeats. At times, repairs of chromosome breaks can cause paracentric and pericentric inversions. If the centromere is not included in the inversion, it is called paracentric inversion but if the inversion is spanning the centromere it is called pericentric inversion. For example: Inversion of genomic sequence from exon1 to intron 22 in factor VIII gene causes severe hemophilia.

Translocation

There are three types of translocations they are Robertsonian fusion, reciprocal and insertional translocations. Reciprocal translocation is the most common type of translocation in which a segment from one chromosome is exchanged with a segment from another nonhomologous chromosome. At times during the exchange translocations may alter the position of centromere and size of the chromosome. For example: Offspring of an individual who is a carrier with heterozygous translocation for chromosome 21 leads to downsysndrome.

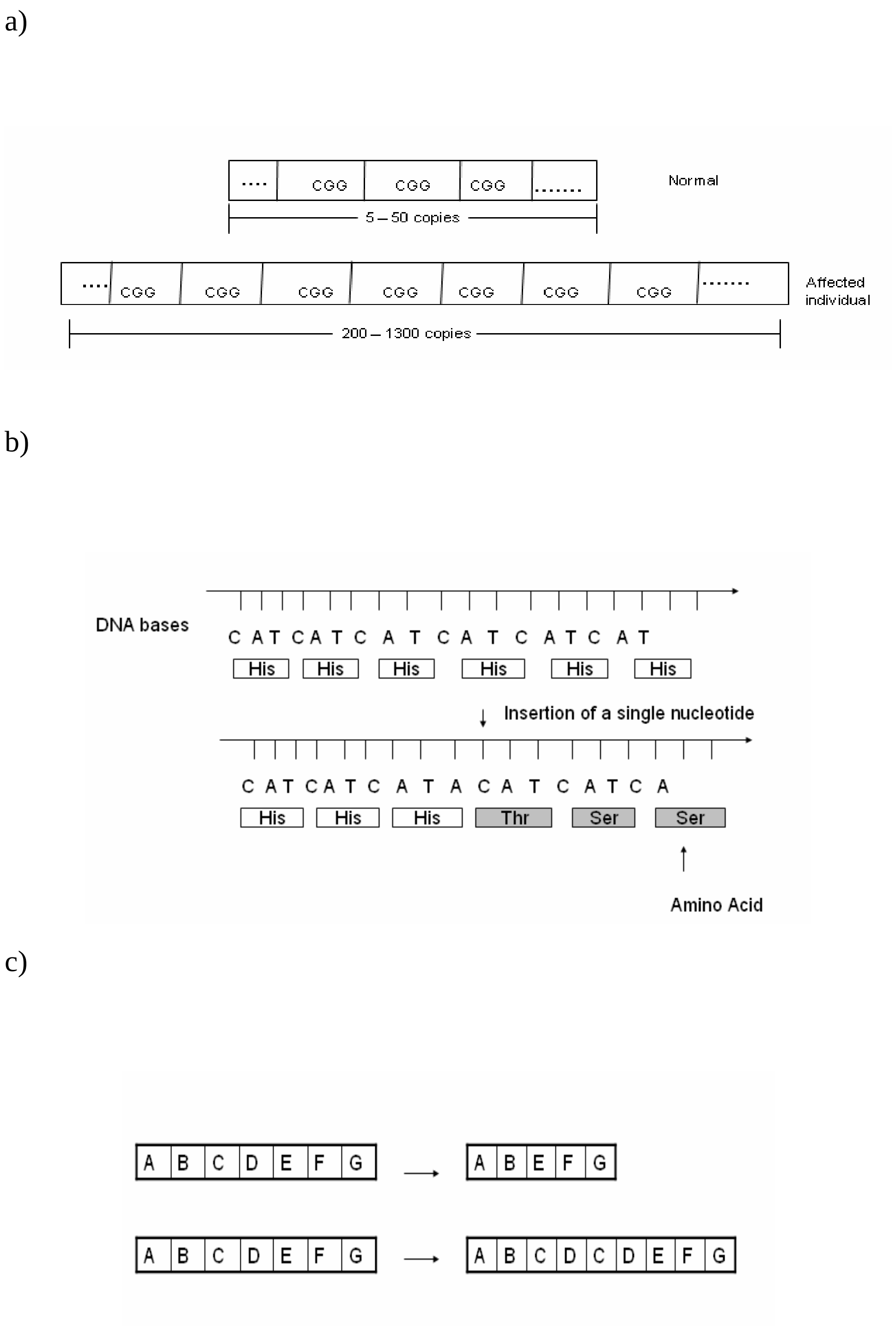
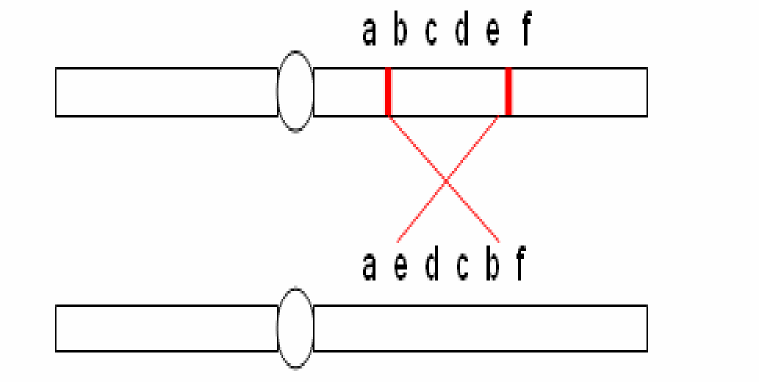


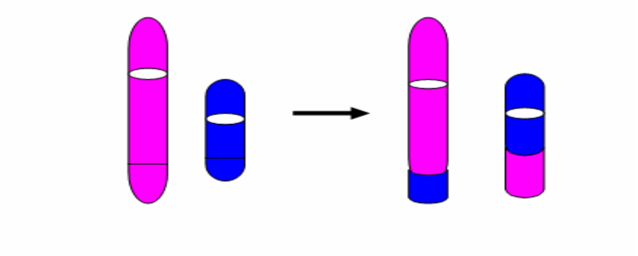
Figure 2 Changes in chromosme structure – a) expansion-contraction type polymorphisms b) Insertions c) Deletions and Duplications d) Inversions e) Translocation

Figure 2 – Continued

d)



e)



Changes in chromosome number

A change in chromosomal number leads to abnormal number of chromosomes or chromosomal sets. In humans, these mutations are known to cause diseases. But in agricultural technology the manipulation of chromosomal numbers are routinely used to grow larger fruits or flowers.

Aberrant euploidy

When changes in chromosome number involve whole set of chromosomes it is called abnormal euploidy. The most common abnormal euploids are polyploids like triploids (3x), tetraploids (4x) etc.. Odd numbers of these chromosomal sets leads to sterility because of unpaired chromosomes during meiosis. But the even sets can produce standard (although abnormal) segregation ratios. For example: Polyploids are formed by combining sets from different species. This can be advantageous in crop breeding. Polyploidy can also result in an organism with greater dimensions and this discovery led to advances in horticulture and in crop breeding.

Aneuploidy

Aneuploidy results in an unbalanced genotype with an abnormal phenotype. Examples of aneuploids are 2n-1 (monosomic) and 2n+1 (trisomic). Aneuploidy is believed to result from chromosomal nondisjunction. Aneuploidy in humans is responsible for several genetic disorders. For example: Aneuploid conditions in humans are Down’s syndrome (trisomy 21) and Klinefelter’s syndrome (XXY).

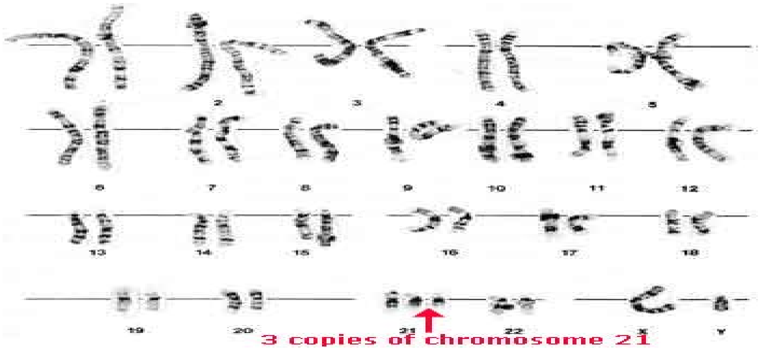


Figure 3 Example of Anueploidy

Figure from Raven and Johnson 1991

DNA mutations, repair and recombination

Mutations can be caused by many different stimuli. Environmental agents such as ultraviolet light, cigarette smoking and chemicals can cause mutations. They may also arise during the process of DNA replication during cell division (mitosis) or production of gametes (meiosis). Most damaged bases are repaired by repair systems such as base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR) and by direct reversal of base damage. But DNA damages such as double strand breaks are repaired by homologous recombination repair (HRR) pathway or by non homologous end-joining (NHEJ) pathway. HRR uses similar sequences to join the broken ends whereas NHEJ prefers some complementary nucleotides but proceeds without it.

The process of crossing over is termed as recombination. Recombination plays an important role in evolution. However, sometimes it may lead to genomic rearrangements and disorders. Recombination occurs in germ-line cells where exchange and reassortment of genetic information happens during meiosis. Recombination also happens in somatic cells to repair the damaged or broken regions of chromosomes. Defects in recombination result in unrepaired chromosomal breaks or aberrant gene translocation, the cost of which can lead to cancer. Cell lines derived from patients predisposed to breast cancer through mutations in BRCA1 and BRCA2 exhibited phenotypic properties characteristic of a recombination/repair defect (Moynahan et al., 1999; Snouwaert et al., 1999; Moynahan et al., 2001).

The size of genomic rearrangements varies from small to large. Large rearrangements occur inter chromosomally or intra chromosomally or DNA slippage during replications (Pentao et al., 1992; Krawczak and cooper 1991). Charcot-Marie-Tooth disease type 1A (CMT1A) is an autosomal dominant disorder resulting from unequal crossing-over of misaligned flanking CMT1A-REP elements on chromosome 17p (Patel and Lupski 1994; Chance et al., 1994). Sequence analysis of this large genomic rearrangement revealed flanking CMT1A-REP elements are approximately 30 kb in length, AT-rich and 98% sequence identity (Reiter et al., 1996). Large genomic rearrangements are common in evolution. For e.g. it is known that two genomes may have lots of genes in common but are organized in a different fashion. Small genomic rearrangements lead to a loss of a whole gene or part(s) of a gene. This study focuses on small genomic rearrangements that deletes or duplicates a portion within a gene (IDDs). These small rearrangements deleting a portion of the gene may alter gene function or gene regulation.

There are two main types of genomic rearrangements – homologous recombination and non-homologous recombination. Homologous recombination is a mechanism where DNA exchange occurs between sequences with extensive homology. A hallmark of homologous recombination is the presence of short exact homologous regions at or near the breakpoints - the point at which the sequence switches from one region to another. Breakpoints are typically found in non-coding regions. Often the homology is due to repetitive elements which may be found throughout the genome in high copy number – particularly in intronic and intergenic regions. Examples of homologous event include study of gene function by gene knockout, DNA recombination during meiosis, interchromosomal recombination during mitosis, sister chromatid exchange and non allelic gene conversion (Stahl 1979; Wasmuth et al., 1984; Liskay et al., 1984; Scherer and Davis 1980). Non-homologous recombination occurs between very little or short similar sequences. Some examples of these events include chromosome translocation, the movements of retroviruses and transposable elements, rearrangement of anti body and T-cell receptor of genes (Gerondakis et al., 1984; Stark and wahl 1984; Shapiro 1983; Honjo 1983; Hedrick et al., 1984; Malissen et al., 1984).

Homologous recombination

There are two types of homologous recombination - equal and unequal. Equal homologous recombination showing how recombination leads to exchange of genetic material between genes is shown in Figure 4. The gene in this example (Figure 4) consists of four exons (A, B, C and D boxed elements), intervening three introns (lines) with two repetitive elements in intron 1 and 2. Repeats are a stretch of similar sequences that are repeated some number of times. During equal crossover the structure of genes remains the same. The resultant of equal crossover consists of admixture of same exons form mother to father and vice-versa. Unequal homologous recombination, by comparison occurs when similar sequences recombine resulting in a change in gene structure – the amount of DNA exchanged is unbalanced. There is however, a conservation of genomic material. Unequal recombination results in one chromosomal copy with an insertion, and the other with the complementary deletion. Figure 5 presents an example of unequal homologous recombination starting from the same gene used in Figure 4. The resultant of unequal recombination leads to a deletion of exon B in the first chromosome and a duplication of exon B in the second chromosome (Figure 5).

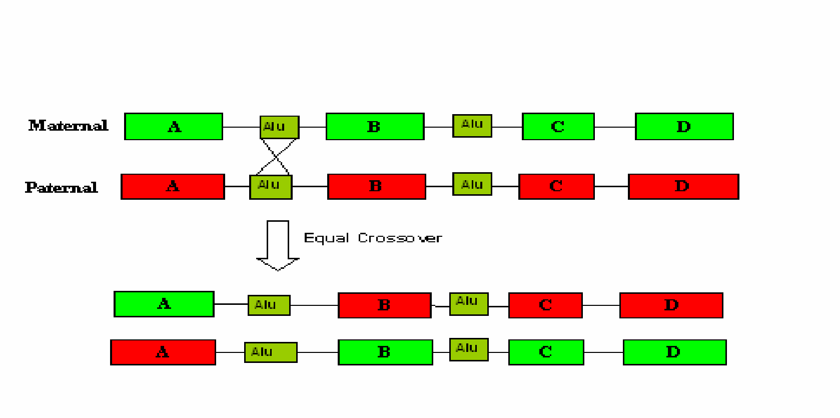


Figure 4 Example of equal homologous recombination between a maternal and paternal copy of a gene.

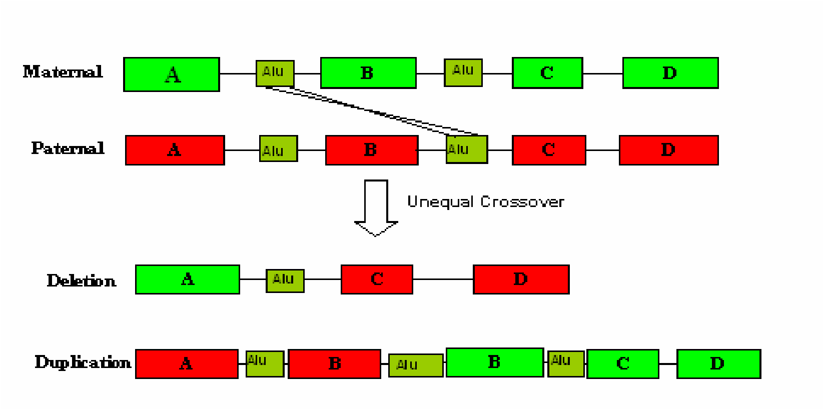


Figure 5 Example of unequal homologous recombination between a maternal and paternal copy of a gene leading to a deletion or duplication

Significant regions of homology

Similar sequences that are repeated some number of times are known as repetitive elements. Features such as length, similarity and distance between sequences play a role during unequal cross over (Deininger and Batzer 2002; Lupski 1998,). Hence this study will mainly focus on regions with high similarity.

Presence of repetitive sequences in the human genome

It is known that human genome sequences consist of approximately 25,000 genes (Ensembl). Most of the DNA sequence is not coding. Apparently the superfluous DNA (approximately 70% in humans) has been termed as junk DNA (Ohno et al., 1972). However, "junk" DNA is not really junk, it might be more appropriately called as "non-coding" DNA, and these contain various repeat elements. Repeat elements are found widely dispersed both among the coding and the non-coding region of the genome. Repeats are found both in prokaryotes and eukaryotes. But are more frequent in eukaryotes particularly those with larger genomes.

Approximately 50% of the Human genome consists of repeats (International Human Genome Sequencing Consortium 2001). Repeats are classified into different types depending on the repeat length. Here is a summary of repeat analysis of the human genome (Sanger Institute).

- 20% consists of LINEs
- 13% consists of SINEs (out of which 11% is Alu sequences)
- 8% retrovirus like and 2% DNA transposons
- 3% is tandem simple sequence repeats (SSR)

Types of repetitive DNA

Repetitive DNA is mainly of two types. They may be tandem (arranged in blocks) or they may be interspersed (distributed in the genome) as shown in Figure 6a and b. Interspersed repeats and tandem repeats are quite common in mammalian genomes. Interspersed repeats are known as mobile or transposable elements they are located at dispersed regions in a genome as shown in Figure 6b.

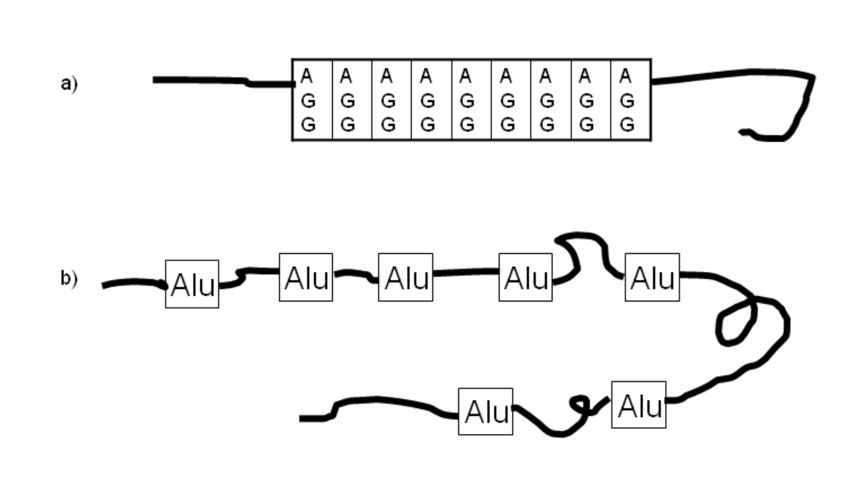


Figure 6 Types of repeats a) Tandem repeats b) Interspersed repeats

Interspersed repeats

In mammals, the most common interspersed repeats are Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs). The average length of LINE is 7 kb. Human genome contains some 850,000 LINEs. Most of these belong to a family called LINE-1 (L1). Length of L1 element ranges from few hundred to 9000 base pairs. Length of SINE element varies from 100-500 bp. Human genome contains about 1,500,000 copies of SINE elements. Alu's are the most abundant of SINES. Alu elements are about 300 bp long and found on average every 3-4 kb in the human genome (Batzer and Deininger 2002).

Though the LINE elements are highly similar it is more observed that SINEs are involved in unequal crossovers. The SINE elements such as ALUs are likely to be close and abundant for unequal recombination. In comparison LINE elements, are big and similar but are far from each other. Hence, reduces the chances of recombination. Mis-pairing between Alu elements and L1 elements has been shown to cause deletion or duplication in several genes.

Tandem repeats

In contrast to interspersed repeated DNA tandem repeats are an array of consecutive repeats. They are sub classified based on the size of blocks; they are satellites, minisatellites and microsatellites. The size of a satellite DNA ranges from 100 kb to over 1 Mb and most of these are located at centromere. The size of the minisatellite ranges from 1 kb to 20 kb. There are two common minisatellites, one found anywhere in the genome (variable number of tandem repeats (VNTR) and the other mostly found at telomeres of a chromosome. Its repeat unit ranges from 9 bp to 80 bp. The telomere contains tandem repeated sequence such as GGGTTA. The size of telomere repeat is about 15 kb. The size of microsatellite ranges from 1 to 6 bp and the whole repetitive region of a microsatellite spans less than 150 base pairs. Microsatellites show a high degree of length polymorphisms and are extremely useful in human genetic studies. These repeats are also called as short tandem repeat polymorphisms (STRPs).

Repeats and Human diseases

Previous studies revealed that length and similarity between sequences correlate positively with the distance between sequences during recombination. Repeats such as Alus are well-known to be involved in unequal crossovers leading to diseases (Deininger and Batzer 2002). There are several examples of unequal crossover between homologous chromosomes leading to a disease. Genes such as LDLR, BRCA1, RB1, ABCA4, C11 have large numbers of repeats, and therefore are likely to be effected by unequal crossovers (Lehrman et al., 1985; Lehrman et al., 1987). The focus of this research is to compile all known cases of deletions or duplications greater than at least 20 base pairs within a gene and to look for any similarities and differences applying machine learning algorithms.

It is known that mutations involving repetitive sequences cause human diseases. Instability of microsatellite, minisatellite repeats leads to human disorders such as Fragile X syndrome (Oberle *et al*., 1991), myotonic dystrophy (Mahadevan *et al*., 1992), Huntington's disease (Gusella et al. 1983) etc. Repetitive sequences such as SINEs and LINEs are also known to cause human disorders. It is very well known that most of the unequal crossovers that lead to diseases are known to occur between Alu (SINE) elements.

Alu repeats leading to human diseases

Alu elements affect the genome by factors like recombination between the elements, insertions, gene conversion and alterations in gene expression (Deininger and Batzer 2002). Insertion of an Alu might change the transcription or disrupt the open reading frame when inserted in exonic region of a gene. Of the diseases related to genetic disorders such as breast cancer and neurofibromatosis ~0.1 % of them are due to Alu insertions (Deininger and Batzer 2002).

Alu elements and recombination

Deletion or duplication can occur due to unequal homologous recombination between Alu elements. It is known that Alu elements hold particular characteristics that make them prone to recombination. These are: 1) The sequence identity is on average greater than 75% between the Alu elements 2) Alu elements are present in close proximity within the genome; hence more chances of recombination events happening between them 3) Due to the large quantity of the number of Alu elements present in the genome there are numerous identical DNA stretches increasing the probability for recombination and 4) A chi-like motif that is present within the Alu sequence may stimulate recombination (Callinan and Batzer 2006).

It is well known that Alu mediated recombination may sometimes lead to disease like breast cancer, Parkinson, diabetes type II etc. Approximately 0.4% of human genetic disorders result from Alu mediated unequal homologous recombination (Deininger and Batzer 2002). This is almost certainly a conservative measurement as this type of mutations is under-surveyed. At the transcript level, these would result in removal or replication of entire exons. The biological assays that are most commonly used to identify mutations this type are described in the section below.

Identification of deletions or duplications using molecular techniques

Deletions and duplications can be molecularly confirmed using several different assays, with each technique having its own advantages and disadvantages. Which assay is best depends on a variety of factors including the expected size of the mutation, the number of samples to be assayed, and how exactly the position of the mutation (s) are known. For example, large-scale mutations such as chromosomal abnormalities are usually detected by performing Fluorescence in-situ hybridization (FISH), Southern blotting, CGH and array CGH. If the number of samples to be assayed is large, then the overhead cost of acquiring and adopting an arrayCGH platform is amortized across the large number of samples. Smaller mutations affecting a few tens or hundreds of bases can be confirmed using techniques such as quantitiative PCR when the sample size is low, or with high-density (potentially custom designed) arrayCGH when the sample size is high. Several of the most common assays are described below.

Southern blotting

Southern blotting (Southern EM 1992) is the most widely used technique to identify deletions and duplications. In this assay, genomic DNA is fragmented and separated by fragment size. The size and quantization of the fragments derived for the locus under study are then used to determine if a deletion or duplication has occurred. The fragmentation of the DNA is performed with one or more restriction enzymes, which cleave the DNA at specific motifs (restriction sites). The fragmented DNA is then electrophoresed on an agarose gel, which separates the fragments by size, with the smaller fragments moving more rapidly through the matrix of the gel. The DNA fragments are then denatured and transferred from the agarose gel into a membrane. Finally, the membrane is probed with a labeled DNA fragment of the gene or genomic region of interest. This will allow visualization of the fragments of genomic DNA from that gene or region.

Polymerase chain reaction

Polymerase chain reaction (PCR; Saiki et al., 1988) is perhaps the most versatile and commonly used assay in molecular biology. It provides a focused and automatable method to amplify the DNA from a specific region. The typical application of a PCR-based assay to detect deletions and duplications is to amplify from both flanks of the deleted region and run the resulting product on a size-separating agarose gel. For a sample with two normal copies, this produces a single band on the gel at a particular molecular weight. However, duplications produce banks of larger molecular weight, and deletions produce lower molecular weight fragments. Thus, with appropriate normal controls, either heterozygous or homozygous deletion and duplication mutations may be distinguished. The PCR process is essentially multiple rounds of geometric replication of a specific region. This procedure requires a thermostable polymerase to duplicate the strands, as well as two oligonucleotide primers flanking the region to be amplified. Each iteration of PCR-based duplication results in a doubling in the representation of the amplified region. Thus 30 rounds of PCR results in approximately a billion-fold amplification (2^{30} = 1,073,741,824). The primary limitations on PCR-based assays are the relatively limited primer size that can be amplified, and the specificity of the PCR primers. Essentially, you have to know exactly what you are looking for.

A more general approach to PCR-based detection of deletions and duplications is quantitative PCR (Q-PCR). This approach requires a much higher degree of control of the temperature than a standard thermocycler, and a better resolution of the exact copy number representation after multiple rounds of PCR. The benefit of Q-PCR over PCR is that the exact flanking regions do not have to be characterized, and the sizes of the mutation do not have to be known. Instead, a small region anywhere within the deleted region is assayed and the quantization of the amplification is compared among cases and controls.

Fluorescence in-situ hybridization (FISH)

Fluorescence in-situ hybridization (FISH) is useful for identifying chromosomal abnormalities. FISH allows researchers to visualize and map where a particular sequence falls within an individuals chromosome. This assay is typically performed on a spread of condensed metaphase chromosomes using one or more fluorescently labeled probes. The location and relative size of the hybridized probe provides evidence for changes in size or location of the underlying locus. The benefit of FISH is that it allows visualization of the entire context surrounding a deletion or duplication mutation. For example, if there have been large-scale changes to the genome resulting in chromosomal abnormalities such as translocations or changes in ploidy. The limitation of FISH is that in practice only the largest of duplications or deletions (100’s of kb) are observable. It is possible to identify smaller mutations, but it requires a more focused set of probes, at which point other methods may be more cost effective.

Comparative genomic hybridization (CGH)

Comparative genomic hybridization (CGH); (Bentz et al., 1998) is a fluorescent cytogenetic technique that identifies gains, losses and amplifications of DNA. During CGH studies the case and control tissues are labeled with different fluorophores. The labeled case and control DNA are then hybridized to a normal metaphase chromosome. The intensity ratio between cases and controls along the length of the chromosome under study is used to evaluate regions of DNA gain or loss. CGH is sensitive to amplifications of 1 Mb or larger, while a single copy loss can be detected if the region is greater than 10 Mb in length. CGH has been applied to identify copy number polymorphisms or genomic rearrangements in human genomes (Sebat et al., 2004) and identifying alterations in breast cancer.

Micro array based CGH

Recently, array-based CGH methods are more commonly used to detect copy number changes. ArrayCGH can be performed at a higher resolution than traditional CGH, and can simultaneously survey the entire genome. Similar to CGH DNA, case and control samples are labeled with different fluorescent colors and hybridized with several hundreds or thousands of DNA probes. The color ratio of case to that of control DNA is then calculated along the chromosomes to evaluate regions of DNA gain or loss in the case sample. Micro and macro deletions can be detected using array CGH method (Pinkel et., 1998, Pinkel and Albertson 2005), with the theoretic ability to detect copy number changes as small as 5 to 10 kb.

Summary of assays available to detect deletions or duplications

Each molecular method described so far to identify deletions or duplications has advantages and disadvantages. Hybridization methods such as blotting and CGH require more DNA (1-5µg) whereas PCR based assays can use 100-1000 fold less DNA. Methods specifically designed for measuring DNA copy number changes such as array CGH also have limitations. Overall, the methods are each relatively expensive and do not have fine resolution across the entire genome.

Databases and other resources

The past decade has seen the completion of the human genome, and a tremendous amount of additional biomedical data including disease descriptions and catalogs of known mutation. Much of this data is available from on-line databases that provide access to those resources. The sections below provide a background on some of these databases and other resources used in the performance of this research.

Mutation databases

Biological mutation databases like OMIM, Human Genome Mutation Database (HGMD), Universal Mutation Database (UMD), the WayStation database and locus specific databases (such as BRCA1 database, LDLR database, etc) provide extensive data on previously identified disease-causing mutations. The availability of these databases offers advantages and disadvantages. For example, the locus-specific databases excel in having the greatest depth of information often having ethno-geographic origin data, population frequency data, detection methodology data and other additional data. However, this comes at the cost of having to incorporate data in a variety of formats from heterogeneous sources. Locus-specific databases often also suffer from poor gene-wise coverage, lack of uniform layout content and quality control (i.e., not all mutations are validated), problems with upkeep and maintenance of data, unreliable URLs and are not publicly available in some cases. Similar kinds of problems also exist with OMIM and UMD. But However, HGMD has the convenience of a central repository with a very high coverage of genes and mutations compared to other databases (Stenson et al., 2003).

Hence we have used HGMD database to gather the list of genes with intragene deletions or duplications.

Gene Annotation repositories

Genomic databases such as those provided from Ensembl, UCSC, and NCBI consist of both the complete genome assembly as well as genomic annotation data. All three of these resources provide access to gene-focused information. Several specific pieces of data are most critical: (1) cross-referencing of a gene with a variety of naming standards (RefSeq, HUGO symbol, gene name, mRNA accessions, etc), (2) the exact position within the genome of the transcript, and (3) the gene structure which maps the positions within the transcript onto genomic coordinates. The UCSC genome database was selected as it provides access to genomic databases through which gene and genomic information can be accessed programmatically. This database consists of information about the gene structures, repetitive sequence elements, and a wide range of additional genomic annotations. Hence, the availability of computer resources, biological mutation databases and genomic databases now permits new approaches to understanding the mutation mechanisms.

Haplotype block data

Genomic rearrangements are often associated with recombination hotspots (Purandare and Patel, 1997; Lupski, 1998; Elliott and Jasin, 2002). It is also known that the regions between the haplotype blocks are recombination hotspots where the sequences tend to recombine (Daly et al. 2001; Patil et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Phillips et al. 2003). A haplotype block is defined as a DNA segment within genetic markers (usually SNPs) exhibit little or no recombination activity (Wall and Pritchard 2003; Hey 2004) within a block. The International HapMap Project is a collaborative study to identify the genetic similarities and disparities among humans by studying various populations (http://www.hapmap.org/). The haplotype block structures identified as part of the International HapMap Project will increase researchers ability to identify genes involved in diseases which will in turn aid in identifying therapeutic drugs (International HapMap Consortium 2005; Deloukas and Bentley 2004). The goal of this project is to compare the genetic sequences of diverse individuals in order to identify regions of chromosomes where genetic variants are shared between populations.

The focus of the HapMap group is to identify haplotype block data based on single nucleotide polymorphisms (SNPs) in various populations. The International HapMap Project was initiated in early 2003. By the end of February 2005, the project had gathered more than one million SNP markers in 270 individuals from four ethnically different study populations (30 trios from Yoruba, 30 trios from CEPH, 45 individuals from Tokyo and 45 individuals from Beijing). Unfortunately, the International HapMap Project does not currently provide access to the haplotype block structures for a given population. They are still in the process of developing algorithms to determine haplotype blocks based on LOD, D' and r^2 values of SNP markers for a given chromosome. Currently the International HapMap Project has an option to download the SNP marker data to haploview software and visualize the haplotype blocks for a max of 250 kbp region. Haploview software allows haplotype analysis of genotype data.

The Encode project and Perelegen Sciences are two additional groups that are studying population variations similar to the International HapMap Project. Perlegen Sciences(http://www.perlegen.com/) genotyped more than 1.5 million SNPs in 71 individuals including those of European American, African American and Chinese ancestry. More than 112 million individual genotypes were obtained. Using data generated by Perlegen Sciences Hinds et al (2005) published a comprehensive study of genetic variation in three different populations. The genotypes were partitioned into haplotype blocks for each chromosome using the HAP phasing software (Halperin and Eskin, 2004). Haplotype block start and end positions for each chromosome are available to download from their website (http://research.calit2.net./hap/wgha/).

Knowledge discovery or machine learning techniques

Data mining algorithms are very useful when exploring large quantities of data and attempting to discover meaningful patterns from that data. Such techniques use statistical analysis, artificial intelligence, and machine learning technologies to identify patterns that are intractable to find by manual analysis alone.

The general structure of a data mining experiment is outlined in Figure 8 in the context of a biological problem. The Biological Understanding phase focuses on understanding the project objectives and requirements from a biological perspective, and then converting this knowledge into a data mining problem. The Data Understanding and Data Preparation phases are the most time consuming. The objectives of these two phases are to understand what features play a role biologically and identify the features that are sensible to obtain computationally. In the Modeling phase, various machine learning algorithms are applied and their parameters are calibrated to optimal values in the Assessment phase. There are many types of machine learning algorithms, several of which are described below.

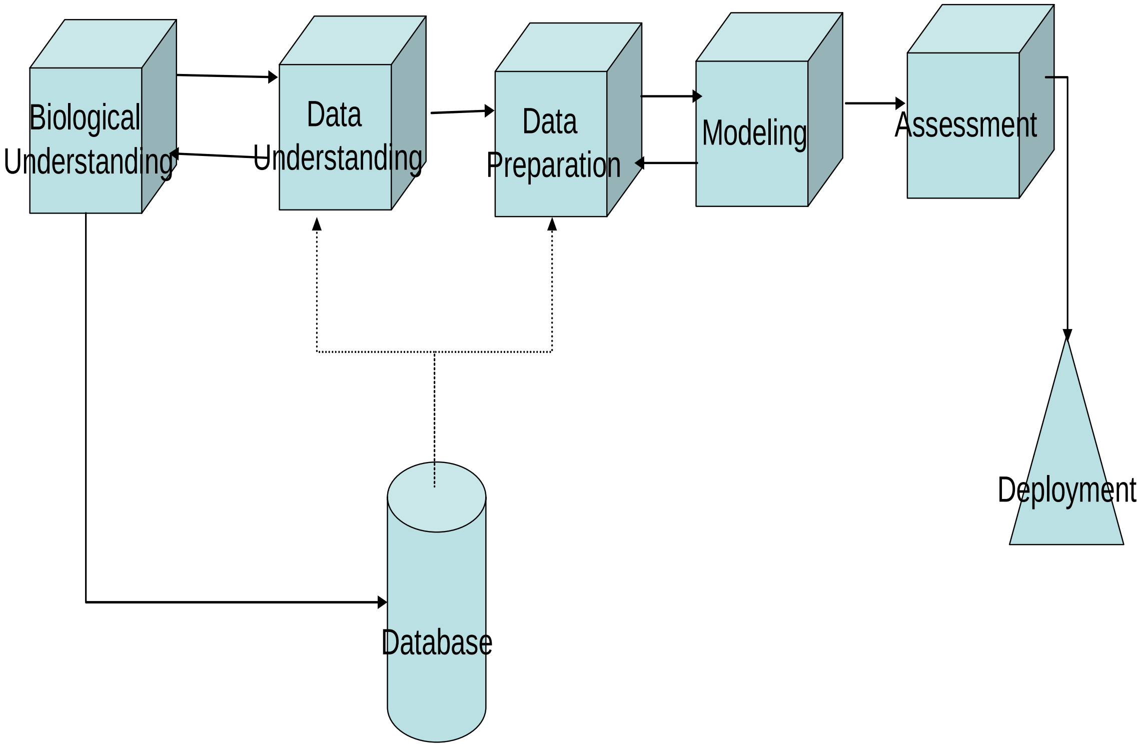


Figure 7 Phases of data mining

Decision trees

The decision tree method is based upon a graph of decisions and their outcomes. This algorithm is used to predict or classify a problem – an easily interpretable white box model. The C4.5 implementation is the most widely used decision tree algorithm. It determines what features to use at a decision point based on information gain (Ross Quinlan 1993). The splitting procedure is repeated in a recursive manner until further splitting is not beneficial. In the resulting tree structure, each inner node in the tree corresponds to a variable, each branch represents a possible value or range of values of that variable and each leaf represents the predicted value of target variable.

Neural networks (ANN)

Neural networks are non-linear statistical data modeling tools. A neural network or artificial neural network (ANN) is an interconnected group of artificial neurons that uses mathematical or computational modeling to identify patterns in data based on the weights of the observed data (Hertz et al., 1990; Havkin 1999; Lawrence 1994). They are used to model complex relations between input and output nodes based on a random function approximation method which 'learns' from data based on the weights.

K-nearest neighbor

The k-nearest neighbor algorithm (k-NN) is a pattern recognition method that classifies the output based on the closest training examples in the feature space (Dasarathy 1991). The space is partitioned into regions. The training phase of the algorithm consists of feature vectors and class labels of the training samples. During the actual classification phase, the same features used in training the system are computed for the test sample (for which the correct classification is not known). The test data point closest to a region is considered to be part of that class. Distances from the new vector to all stored vectors are computed and the k closest samples are selected. The new point is predicted to belong to the class that is nearest to the one within the set.

Logistic model tree (LMT)

Logistic Model regression (LMT) is a combination of tree induction and logistic regression model resulting in a single tree (Landwehr et al., 2003; Landwehr et al., 2005). A logistic model tree consists of a standard decision tree structure with logistic regression functions at the leaves using posterior class probabilities. Therefore, LMTs consists of a tree structure that is made up of a set of inner nodes and a set of leaves or terminal nodes in an instance space. Tree induction identifies subdivisions by recursively splitting the instance space in a divide-and-conquer fashion until further subdivisions are not beneficial. In a study conducted on 36 datasets by Landwehr et al., they concluded that the LMT model outperformed simple logistic, multi logistic, C4.5 and CART methods (Landwehr et al., 2005). In general, LMT model produces smaller trees than the classification trees built by C4.5 or CART.

Support vector machines (SVM)

A support vector machine (SVM) applies linear classification techniques to non-linear classification problems. The support vector machine method is used for classification problems – defining separate features as different dimensions and thus partitioning regions in a multidimensional space corresponding to known outcomes. A hyperplane separates the data points “neatly” with maximum distance to the closest data point from both classes (Cortes and Vapnik 1995).

Summary of data mining

There are several studies that used data mining algorithms such as neural networks, hidden markov models and support vector machines to identify core promoters (Ohler et al., 2002; Pavlidis et al., 2001; Ben-Hur and Brutlag, 2003; Sharan and Myers, 2005; Reese 2001). Other projects, such as the PAR (prioritization of annotated regions) algorithm also used knowledge based methods to identify mutations (Braun et al., 2006). Hence, for more than a decade data mining and machine learning has become popular to identify patterns.

<u>Previous studies and significance of their work</u>

According to HGMD statistics approximately 6% of the mutations available in the database are intragene deletions or duplications (Stenson et al., 2003). It is known that intragene deletions or duplications are nonrandomly distributed in the human genome (Mitelman, 2000) and are responsible to predispose the carrier to breast and colon cancer

(Rabbitts, 1994; Puget et al., 1999; Petrij-Bosch et al., 1997; Wijnen et al., 1998; Mitelman, 2000) and inherited diseases (Stankiewicz and Lupski 2002).

Intragene deletions or duplications are one of many categories of mutations that can cause disease. Different types of mutations occur with varying frequencies in a gene-dependent fashion. For some diseases (e.g., Duchenne and Becker muscular dystrophy), intragene deletions or duplications are common causes of disease. In contrast, they are a rare cause of other diseases (e.g. hemophilia A,Lesch-Nyhan syndrome).

Repetitive elements and similar sequences are known to present an abundant opportunity for genomic rearrangements (Deininger et al., 2003) and mutations involving Alus are actively involved in human diseases (Batzer and Deininger, 2002). Approximately 0.4% of human genetic disorders result from Alu mediated unequal homologous recombination (Deininger and Batzer, 1999). A study revealed that GC content increases in the regions of recombination and the presence of a 26 bp core sequence in Alu mediated unequal homologous recombination (Rudiger et al., 1995). Another study revealed that during rearrangement mechanism features like repeat size, degree of homology, and distance between the similar sequences play a major role (Stankiewicz and Lupski, 2002).

Most of the gross rearrangement analyses have been confined to relatively few genes (Monnat et al., 1992; Janssen et al., 1993; Osterholm et al., 1996). But a recent study on a different set of genes revealed that sequences flanking unequal recombination breakpoints (breakpoint position ± 15base pair) tend to be AT rich. This study also reported that direct repeats are over represented in deletion breakpoints (Abeysinghe et al., 2003). The features that were the focus in the previous studies were length of the similar sequences, percent identity between the sequences, whether repetitive elements were involved and distance between the sequences. We included this set of features used in previous studies and also incorporated novel features such as the GC content of the pairs of similar sequences, GC content of the deleted sequence, melting temperature characteristics and haplotype characteristics of the sequences likely to recombine.

CHAPTER III
COMPUTATIONAL AND MACHINE LEARNING APPROACH

The goal of my thesis is to design and develop a novel method to identify intragene deletions or duplications (IDDs) using computational and machine learning methods. IDDs include a gene and 5 kb upstream and downstream of a gene; which means deletions or duplications within a gene, promoter region and 3’ UTR regions. Computational methods were used to extract key feature data from a set of published IDDs. Machine learning algorithms or data mining techniques were further applied to train on the set of key features to predict IDD candidates for a given gene.

High-level picture of our novel approach

The approach is briefly outlined in Figure8. My strategy for identifying IDD candidates is accomplished in three stages. The First stage is to collect fully-characterized breakpoints for previously published IDD cases. In this context, fully-characterized implies that the exact breakpoint of the IDD is known. The next stage is to identify and select informative features of IDDs, and to compile known case and control sets based on the fully-characterized breakpoints collected in the first stage. Final stage is where analysis is performed utilizing the case and control sets with features to prioritize candidate IDDs. The prioritization is accomplished using machine-learning algorithms which have been trained using the features of the case and control data sets. These three strategies combined lead to a solution to predict deletion or duplication candidate regions for a given gene.

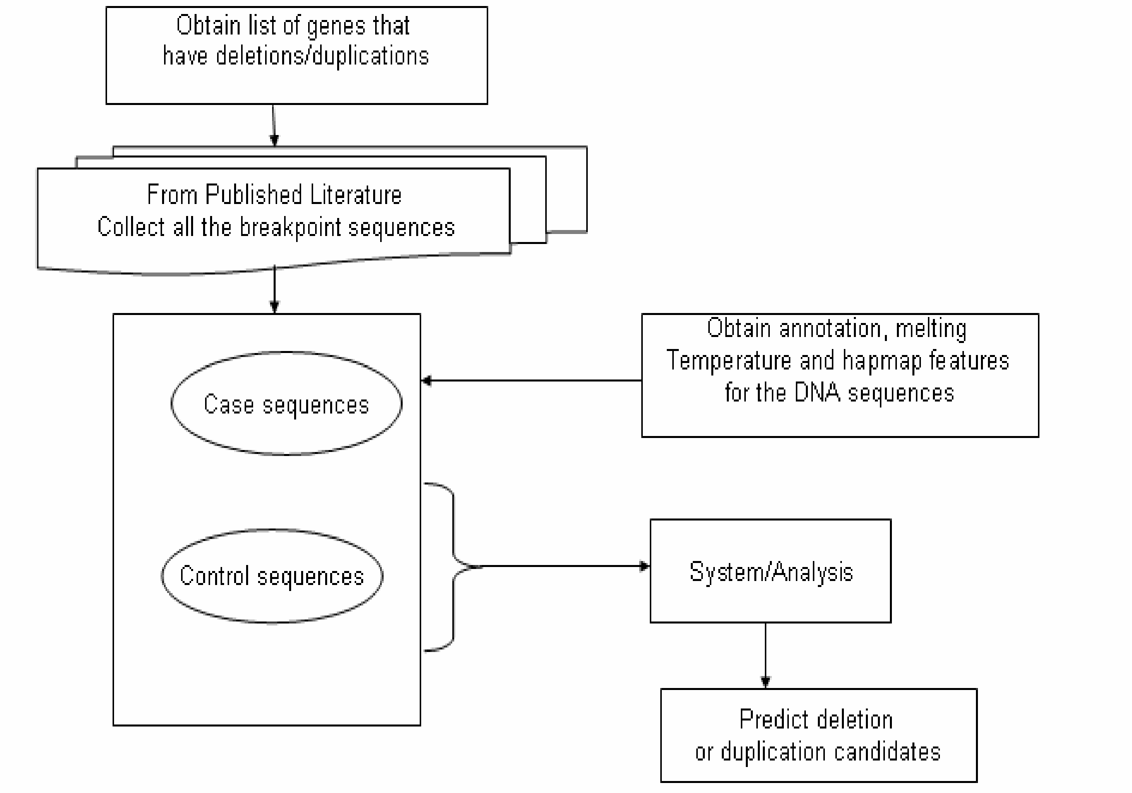

Obtain list of genes that have deletions/duplications
From Published Literature Collect all the breakpoint sequences
Case sequences
Control sequences
Obtain annotation, melting Temperature and hapmap features for the DNA sequences
System/Analysis
Predict deletion or duplication candidates

Figure 8 General outline of the process.

On average the size of an intron is longer than the exon and hence they contain greater number of Alus and therefore more chances of unequal homologous recombinations. But of all the possible unequal recombinations, only few may lead to disease. Others may lead to a variation in the expression of that particular protein, or may have no effect. In addition, assays to identify deletions or duplications are time consuming and expensive. Hence prioritization of candidate IDDs with *in-silico* procedures should be very helpful.

Although studies have been performed to investigate the mechanism underlying homologous recombination, to our knowledge there is no bioinformatics system that will predict deletion or duplication candidates within a gene based on a previous set of known deletions or duplications. SPeeDD – System to Prioritize Deletions or Duplications is such a system, designed to estimate the most probable recombination events resulting in deletions or duplications by applying computational and machine learning techniques on the DNA sequence parameters of known biologically verified deletions or duplications.

Reliability of any machine learning system is based on the type of features used and the quality of the training data set. After surveying the literature, we used the features that are very well understood and proven to be involved during unequal homologous recombination (Deininger and Batzer 1999; Brooks et al., 2001; Shaw et al., 2002; Shaw et al., 2004; Abeysinghe et al., 2003; Lupski and Stankiewicz 2005; Sen et al., 2006). After obtaining the entire feature set data for the cases and controls data mining techniques were applied on the DNA sequence features. Data mining softwares consists of several data pre-processing and machine learning methods like artificial neural networks, decision trees, k-Nearest Neighbor, support vector machines and so on.

The performance of the computational and machine learning system developed is evaluated using cross validation methods. All different types of machine learning methods were applied on the system and efficacy of the models are compared in terms of sensitivity and specificity. The model that yielded high sensitivity and specificity values were considered for the final implementation of the system to predict candidate regions for a given gene of users interest.

SPeeDD is an interactive web based tool, which aids investigators in identifying intragene deletions and duplications. This software with number of valuable options for the scientist may provide a fast and less expensive method to predict deletions or duplications. The application of computational and machine learning methods are complementary to laboratory assays, it helps investigators focus on more likely candidates first with the goal of increasing the pace and efficiency of biological research. Hence our novel method to identify possible deletion or duplication candidate regions likely to cause disease within a gene reduces cost and is effective.

CHAPTER IV

METHODS OF THE COMPUTATIONAL SYSTEM

The computational system developed as part of my thesis research (SPeeDD) incorporates features of previously reported intragene deletions or duplications (IDDs) to predict novel deletion or duplication candidate regions within a gene. This required the collection of two important data sets, and the harnessing of several computational tools. The first data set required was the collection of previously reported IDDs. This was collected from the Human Genome Mutation Database (HGMD). The primary literature was then evaluated for each of the reported IDDs, and the exact break-point sequences were identified when possible. These cases are then annotated with genomic position, and position within the gene using the UCSC genome database. The second data set is the control data set that was derived for each case. It includes a set of potential homologous recombination sequences that have not been observed to be causing disease.

Identification of genomic rearrangements

HGMD maintains a comprehensive database of published human mutations (Cooper and Krawczak 1996; Krawczak et al., 2000). Data from HGMD is freely available, and was obtained by searching the website by gene name. This search was performed for all of the 12,371 gene names with official HUGO symbols (Povey et al., 2001; Wain et al., 2004). From this list of gene symbols, the list of genes in which gross insertions or deletions have been observed was obtained. Gross insertions or deletions are the nomenclature used by the HGMD database to describe deletions or duplications of more than 20 bp. As of September 2006, HGMD consists of 53,208 mutations recorded in 2,056 genes. As shown in Figure 9 the class of mutation in which we are interested, gross insertions or deletions account for about 6.5 % of all mutations in the HGMD database.

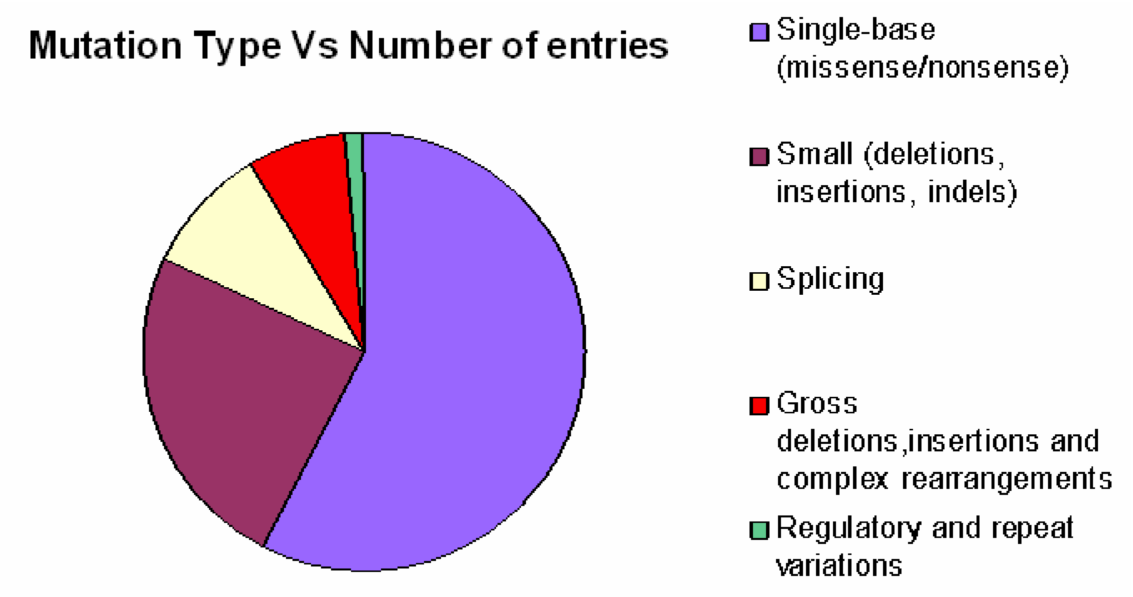


Figure 9 Representation of the HGMD statistics as of September 2006.

Deletions or duplications that occur within a gene the IDDs are one of the least surveyed type of mutation that account for a significant fraction of disease-causing alterations. The mutations in HGMD are classified into small deletions, small insertions, small indels, nucleotide substitutions (missense/nonsense/splicing/regulatory), gross deletions, gross insertions, complex rearrangements and repeat variations. Each mutation is also associated with a gene. As an example, the summary of mutations associated with ABCA4 is shown in Table 2 divided into the total number of mutations for each mutation class. ABCA4 has no mutations reported for nucleotide substitutions (regulatory), gross insertions and duplications and repeat variations (Table 2). Detailed information on the nucleotide substitutions is available via the provided hyperlink, an example of which is shown in Table 3. The exact nucleotide base pair change, its effect on coding, and the associated disease(s) are detailed as shown in Table 3, along with a reference to the journal article. Unfortunately, the hyperlinks provided for the mutations of interest (gross deletions or insertions or complex rearrangements) do not provide the same level of detail. Instead, as shown in Figure Table 4, mutations in the gross deletion or duplication category only provide data on the particular gene regions (e.g. exons 20 through 22 in the Maugeri et al Stargardt deletion) that are affected and a link to the publication in which the mutation was reported.

Construction of local deletion or duplication database

HGMD does not have a computationally accessible programming interface for programmatic interaction, nor do they provide a bulk download option to obtain the entire mutation database. They do however; allow web-based interfaces to access mutations on a per-gene basis. To obtain the complete set of IDDs from HGMD, I therefore developed programs to search HGMD based on gene name. I collected the description, phenotype and publication links of all gross deletions and gross duplications from HGMD and stored it in our local database as shown in Figure 10. Our database (the University of Iowa Human gross deletion or duplication database) has a total of 1463 IDDs in 441 genes with known gross deletions or duplications along with links to scientific publications in which these mutations were published. I manually examined a majority of the publications and collected a set of genes with gross deletions or duplications that have their breakpoint regions sequenced and published. From this literature review and through direct contact with the authors I was able to collect fully characterized break points for 102 intragene deletions or duplications.

Table 2 Number of entries by mutation type for the ABCA4 gene in the HGMD database

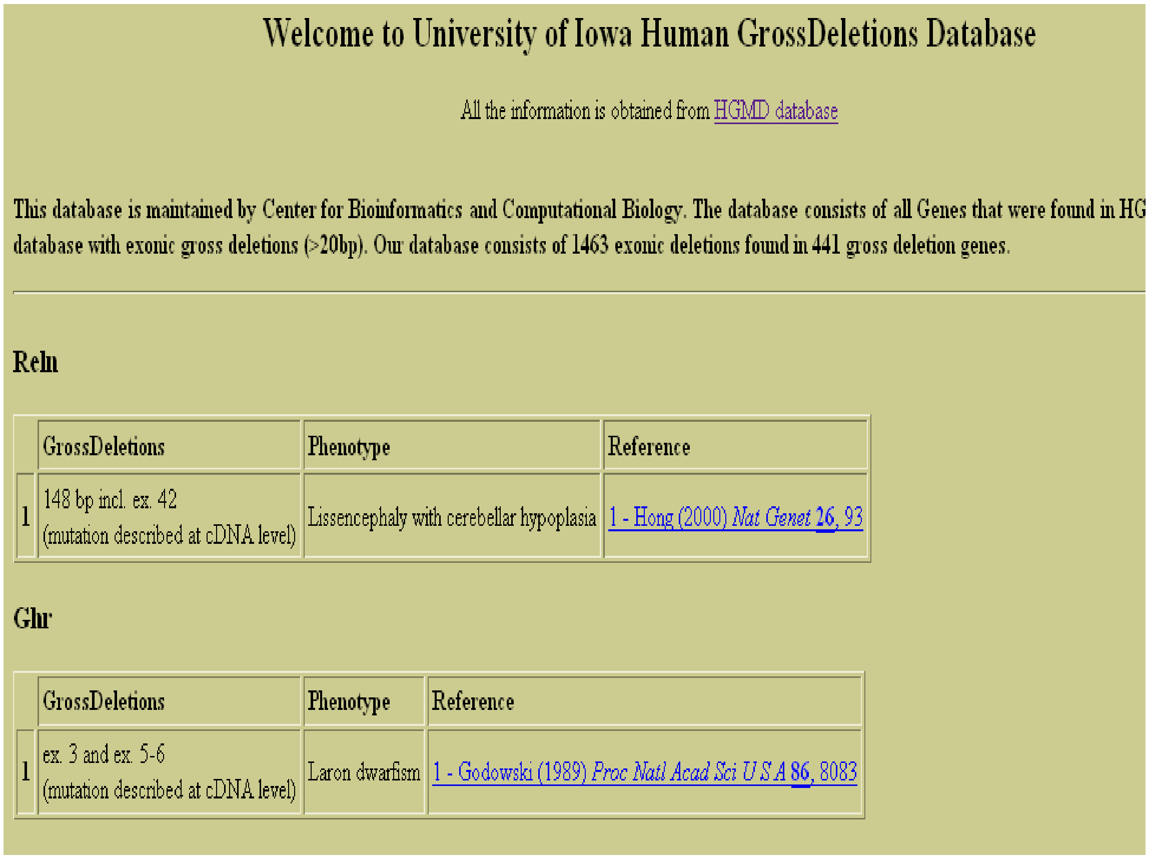
Mutation Type	**Total number of mutations**
Nucleotide substitutions(missense/nonsense)	294
Nucleotide substitutions(splicing)	46
Nucleotide substitutions (regulatory)	0
Small deletions	52
Small insertions	12
Small indels	1
Gross deletions	3
Gross insertions & duplications	0
Complex rearrangements	1
Repeat variations	0

Table 3 HGMD details of the nucleotide substitutions (missense / nonsense) of the ABCA4 gene

Accession Number	**Codon Number**	**Nucleotide**	**Amino acid**	**Phenotype**	**Reference**
CMO14282	1	ATG-GTG	Met-Val	Stargardt disease	Briggs (2001) Invest Ophthalmol Vis Sci **42,** 2229
CM983846	11	CTC-CCC	Leu-Pro	Fundus flavimaculatus	Rozet (1998) Eur J Hum Genet **6,** 291
CM990010	15	TGG-TGA	Trp-Term	Stargardt disease	Maugeri (1999) Am J Hum Genet **64,** 1024
CM980003	18	CGG-TGG	Arg-Trp	Stargardt disease	Gerber (1998) Genomics **48,** 139
CM990011	24	CGC-CAC	Arg-His	Stargardt disease	Lewis (1999) Am J Hum Genet **64,** 422
CM043238	24	CGC-TGC	Arg-Cys	Cone-rod dystrophy	Klevering (2004) Eur J Hum Genet **12,** 1024

Table 4 Details of gross deletions obtained from the HGMD source for the ABCA4 gene

Accession Number	**Description**	**Phenotype**	**Reference**
CG035110	1030 bp incl. ex. 18 (described at genomic DNA level)	Stargardt disease	Yatsenko (2003) Hum Mutat **21,** 636
CG994802	36 bp nt. 6543 (described at genomic DNA level)	Stargardt disease	Lewis (1999) Am J Hum Genet **64,** 422
CG994803	ex. 20-22 (described at genomic DNA level)	Stargardt disease	Maugeri (1999) Am J Hum Genet **64,** 1024



Welcome to University of Iowa Human GrossDeletions Database

All the information is obtained from HGMD database

This database is maintained by Center for Bioinformatics and Computational Biology. The database consists of all Genes that were found in HG database with exonic gross deletions (>20bp). Our database consists of 1463 exonic deletions found in 441 gross deletion genes.

Reln

	GrossDeletions	Phenotype	Reference
1	148 bp incl. ex. 42 (mutation described at cDNA level)	Lissencephaly with cerebellar hypoplasia	1 - Hong (2000) *Nat Genet* **26**, 93

Ghr

	GrossDeletions	Phenotype	Reference
1	ex. 3 and ex. 5-6 (mutation described at cDNA level)	Laron dwarfism	1 - Godowski (1989) *Proc Natl Acad Sci U S A* **86**, 8083

Figure 10 Screen shot of our local deletion database.

Computational method to identify candidate breakpoint sequences

Previous studies revealed that sequence length and similarity correlates positively with the distance between sequences during homologous recombination; features such as length, similarity and distance between sequences play a role during unequal crossing over (Lupski, 1998). This thesis focuses on “significant regions of homology” that are required for recombination. To identify candidate breakpoints a computational system was developed to identify similar sequences within a genomic neighborhood. This system uses the UCSC genome annotation database to identify the transcribed region associated with a gene, and their genome build to obtain the gene’s sequence along with flanking genomic region (Tatusova and Madden, 1999). The BL2SEQ program (Tatusova and Madden, 1999) was used to find pairs of similar sequences within the gene’s locus. Bl2SEQ was used to perform reciprocal comparison between the gene’s sequence and itself using blastn algorithm - a heuristic approximation to the Smith-Waterman local alignment algorithm.

For each gene under study, the genomic sequence including the entire transcribed region and 5 kb of flanking sequences on each end were obtained and blasted against it self using the BL2SEQ software. The filtering criterion of the BL2SEQ output is shown in Figure 11. BL2SEQ output with duplicate hits (e.g., subsequence A aligned with subsequence B and subsequence B aligned with subsequence A) represent a single unique alignment, and only a non-redundant set is maintained. Out of all similar sequence pairs obtained from the BL2SEQ analysis, only those pairs with at least 80% identify and less than 50 kb away were considered for future study. The exact context (intron, exon, promoter or UTR) of these similar sequences was obtained using the gene structure annotation from the refGene table in the UCSC genome database (Tatusova and Madden, 1999). Details of similar sequences that span exons were obtained for future study, as shown in computational pipeline (Figure 11). For each gene we were able to identify the known IDD pair of sequence from the filtered BL2SEQ output.

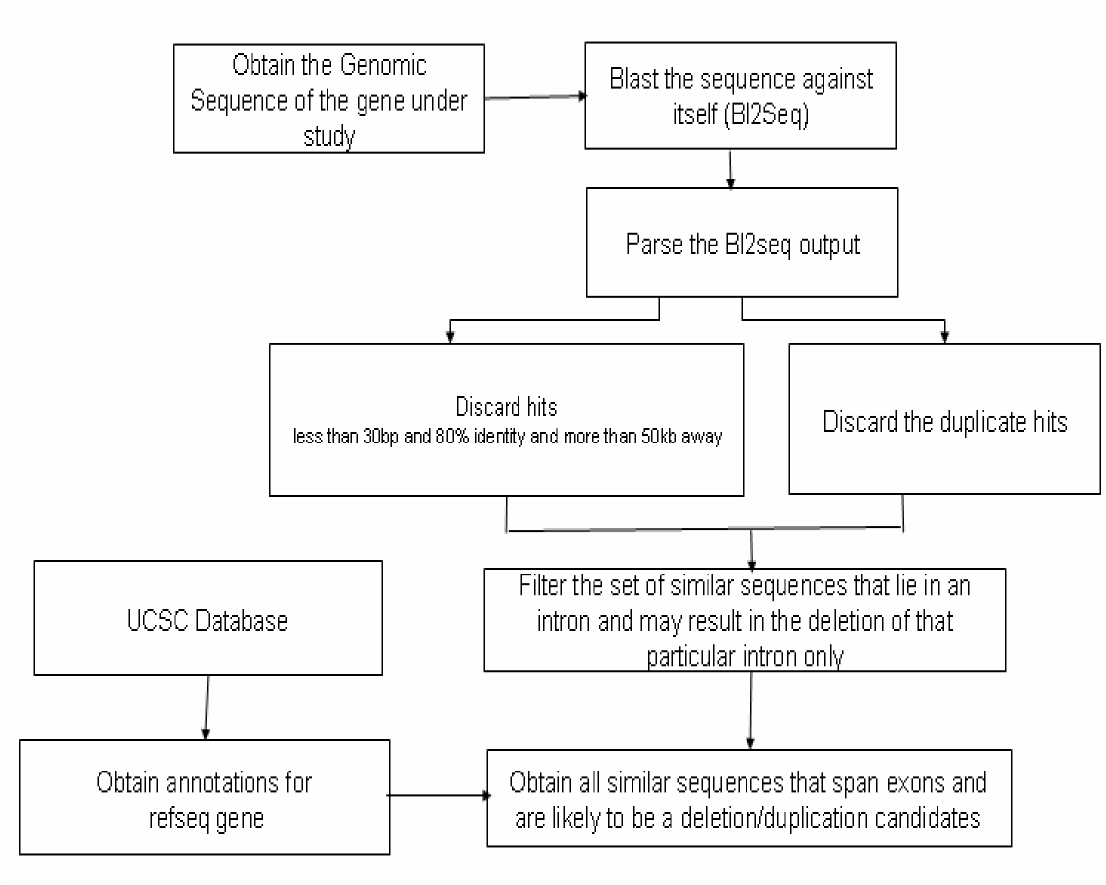

Obtain the Genomic Sequence of the gene under study
Blast the sequence against itself (Bl2Seq)
Parse the Bl2seq output
Discard hits
less than 30bp and 80% identity and more than 50kb away
Discard the duplicate hits
UCSC Database
Filter the set of similar sequences that lie in an intron and may result in the deletion of that particular intron only
Obtain annotations for refseq gene
Obtain all similar sequences that span exons and are likely to be a deletion/duplication candidates

Figure 11 Computational pipeline of the BL2SEQ approach

Our studies to gather the intragene deletion or duplication break points and application of computational method to identify the sequence pairs that are prone to rearrangement in their respective genes from the BL2SEQ output have established the validity of our approach. We extended the BL2SEQ data with other informative features likely to be useful using other data and computational resources. For every candidate IDDs within a gene, the sequence specific, hapmap and melting temperature features are obtained as described below.

Feature Selection for the analysis

The results of any automated analysis depend on the nature and quality of the data analyzed. Therefore, the selection of appropriate DNA features is a very important criterion. There are several features to be examined for each potential pair of recombination. For example Stankiewicz and Lupski demonstrated (Stankiewicz and Lupski, 2002) that during the process of recombination, the more identical and the closer together the repetitive elements are the greater the chances of recombination. Hence, parameters like repeat length, percent identity between repeats and the distance between them play a vital role. It is also known that GC-rich and GC-poor regions in the genome play a role in the recombination (Fullerton et al., 2001). Hence, GC content parameters of the DNA sequences are included. In the same way melting temperature and haplotype characteristic features were also included in the analysis. Below is a list of the features, description of the features and how we computationally obtained them in Table 5.

Table 5 List of features which were used by the machine learning application

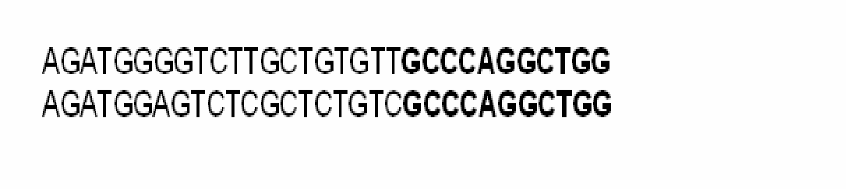
Feature	**Source**	**Description**
Length	BL2SEQ	Length of the sequences potentially required for recombination
Percent Identity	BL2SEQ	Percent identity between the pair of sequences
Score	BL2SEQ	BL2score of the similar sequences
Distance	UCSC database	Distance between the similar DNA elements
GC content of sequences	Simple perl program	GC content of sequence elements
Repeat involved	UCSC database	Repeat Characteristics
Tm	TmAlign (IDT Software)	Melting temperature of the pair of sequences likely to recombine
TmExact	TmAlign(IDT Software)	Melting temperature of the longest exact match of the pair of sequences likely to recombine
HapType	Haplotype block data	Haplotype features (sequences located inside or outside or span the block)
HapDist1, HapDist2	Haplotype block data	Other haplotype characteristics like - the distance to the nearest neighboring block from beginning to ending of the sequences likely to recombine.

Sequence-specific features

A Perl program was developed to obtain the sequence-specific features such as the sequence length, identify, distance directly from the BL2SEQ output file or with a simple analysis of the sequence files themselves. For example, the sequence length refers to the length of match for a given pair of homologous sequences. Similarly, the percent identity, blast score and the distance between the pairs were extracted from the blast output. The GC content calculations were performed for the paired sequences and the intervening sequences based upon the location of the aligned sequence specified in the BL2SEQ output. The human_annot_jul03 release of the UCSC annotation database was used to obtain gene structure information (Tatusova and Madden, 1999).

Melting temperature features

The thermo align (TmAlign) program was obtained from IDT (Integrated DNA Technologies) to calculate melting temperature (Tm) of the two sequences flanking the candidate IDDs (IDT unpublished software). Thermo align uses energy files, hybridization temperature, oligo concentration and salt concentration during alignment. The thermo align program was used to obtain melting temperature data for the candidate IDD breakpoint sub sequences. It is known in homologous recombination events that the longer the exact base pair sequence match the higher the chances of recombination (Deininger and Batzer 1999; email communication with Deininger) Hence, longest exact subsequence (i.e., not interrupted by any base mismatch, as indicated in Figure 12) properties are also utilized in the analysis. The DNA concentration, salt concentration and hybridization temperature parameters we used to align sequences are 0.0001µM, 140mM and 37°C respectively.



AGATGGGGTCTTGCTGTGTT**GCCCAGGCTGG**
AGATGGAGTCTCGCTCTGTC**GCCCAGGCTGG**

Figure 12 Example of pair of similar sequence with longest exact sub sequence in bold format

Haplotype block features

It is known that genomic rearrangements (deletions or duplications) are often associated with recombination hotspots (Purandare and Patel, 1997; Lupski, 1998; Elliott and Jasin 2002). As described in the background chapter of my thesis Perlegen Science group genotyped over 1.5 million SNPs in populations of European, African American and Han Chinese ancestry and partitioned them into haplotype blocks using the HAP program (Halperin and Eskin, 2004; Hinds et al., 2005). The start and end positions of each block was obtained from their files and stored in our local database. HaplotypeBlock feature has three options depending where the IDD candidate sequences lie (INSAME, OUT, SPAN). Distance to the nearest haplotype block nearestBlock1 and nearestBlock2 (nearestLeft, nearestRight) also are obtained for all candidate IDD breakpoints.

For example as shown in Figure 19 there are two haplotype blocks HAPBLK1 and HAPBLK2 in the region of interest. If the IDD candidate breakpoints are found in between the blocks then the HaplotypeBlock feature is determined as outside the block (Out). In case, where IDD candidate is located in the same block the HaplotypeBlock feature is determined as present in the same block (InSame). But if one of the IDD candidate breakpoint sequence is located inside the block and the other is found to be outside the block then it is determined as spanning the block (Span). NearestBlock1 attribute is the shortest distance between the IDD candidate breakpoint sequences to the end of the haplotype block HAPBLK1; as shown in Figure 13 it is the value of d1.

NearestBlock2 is the longest distance between the IDD candidate breakpoint sequences to the end of HAPBLK; as shown in Figure 13 it is the value of d2.

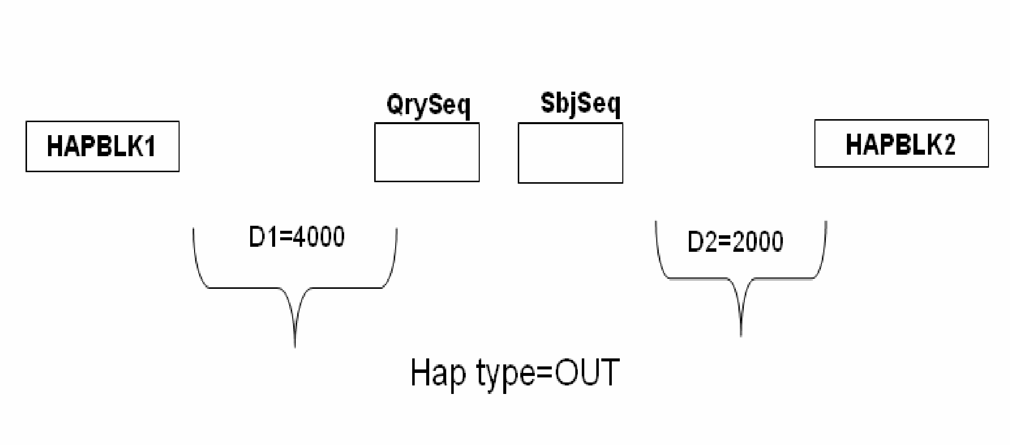


Figure 13 An example to explain haplotype block feature

Summary of the data sets used for data mining

Figure 14 is used to explain the datasets that were used in this study (cases and controls). Figure 14 shows a gene with several repetitive elements in introns 3 and 4. Theoretically, all the repeats in introns 3 and 4 have equal likely chance to be the candidates for deleting exon 4. If an unequal crossover between repetitive elements r3 and r9 has been observed then the DNA features of r3 and r9 sequences are considered as a case set in our study. The other repetitive sequences in introns 3 and 4 could recombine, however these other crossovers have never observed. Hence, all other possible recombinants are defined as the control sequences in this research.

Both case and matching control sequences for each gene are obtained from the exon spanning recombinants found in the BL2SEQ output as shown in Figure 15. Each case with a fully characterized breakpoint sequence is manually validated. Similarly, controls sequences that are in the same gene and exhibit the same potential gene rearrangement as the individual cases are also dedeidentified. Sequence specific features, melting temperature and haplotype block characteristic features that are known or implicated to play a role during unequal homologous recombination are obtained for the case and control sequences that are identified. To summarize, we have collected the feature set data for all the 102 fully characterized IDDs (case data). Additionally, we also have collected feature set data for an extra 2338 potential homology-based candidates from the same set of genes for which IDDs have never been observed (control data).

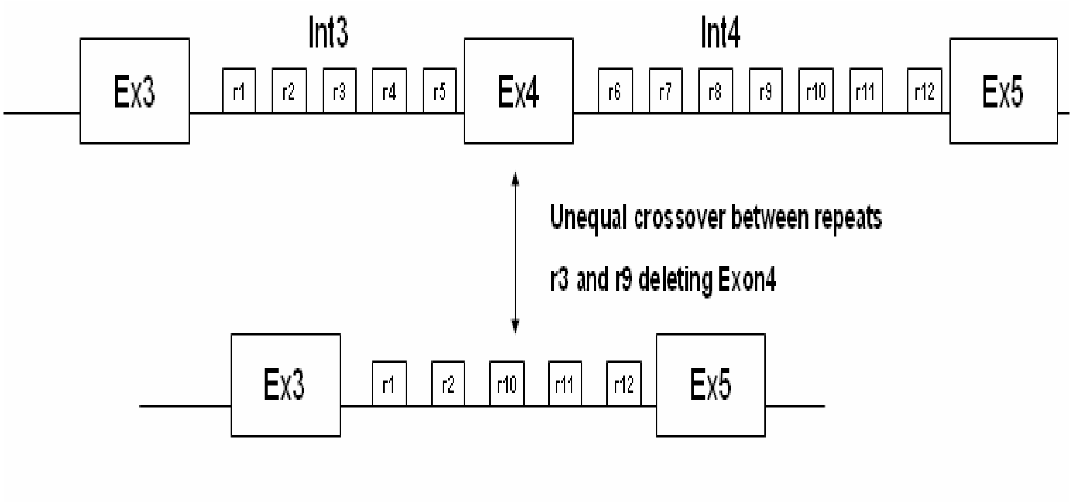


Figure 14 Example of unequal crossover within a gene. Diagram consists of exons 3-5 and intervening lines intron 3 and 4. Small boxed elements in introns 3 and 4 are the repetitive elements r1 thru r12. Resultant of unequal crossover between repeats r3 and r9 deletes exon 4.

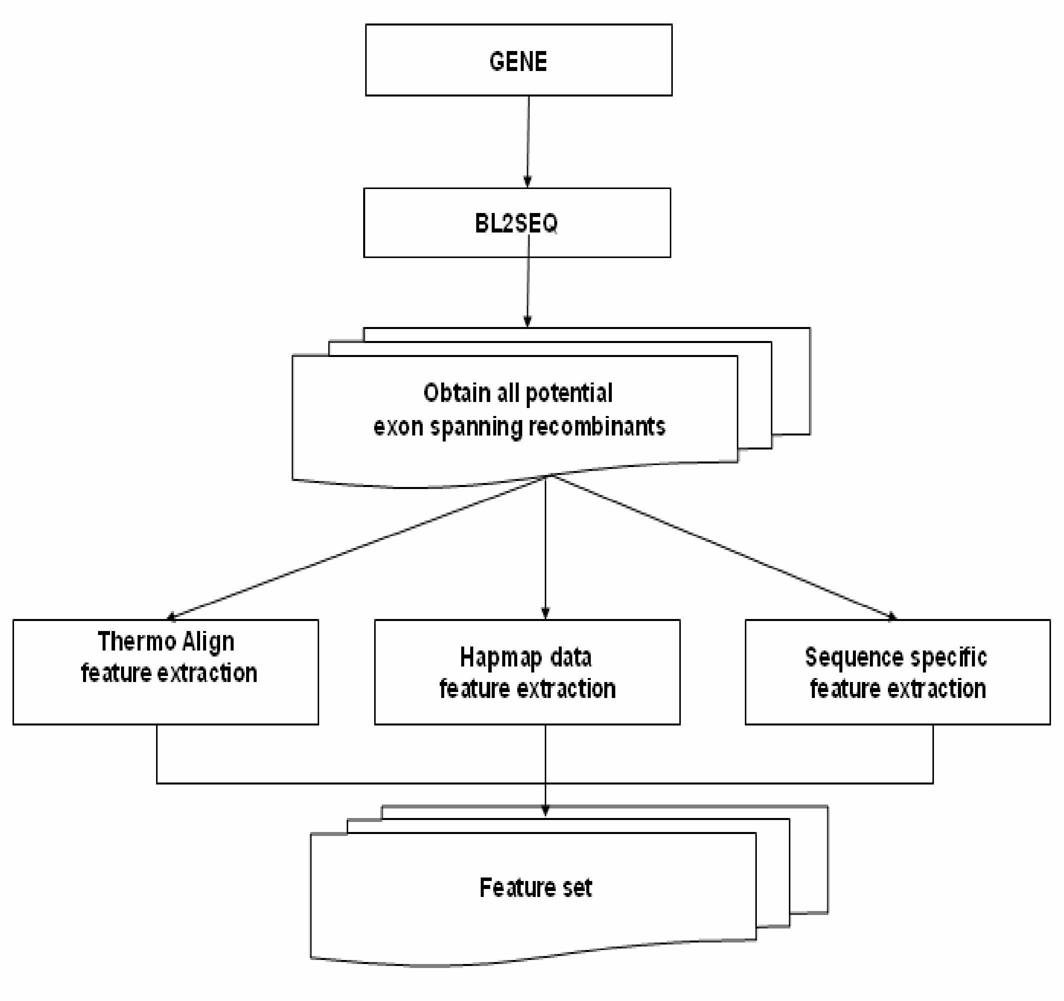

GENE
BL2SEQ
Obtain all potential exon spanning recombinants
Thermo Align feature extraction
Hapmap data feature extraction
Sequence specific feature extraction
Feature set

Figure 15 Computational pipeline to obtain feature set for a given gene

Computational and machine learning system

After collecting the feature information I designed and developed a computational and machine learning system to compile case and control data sets and to explore for general characteristics in features by applying machine-learning algorithms. This new system prioritized candidate deletions and duplications within a gene and is named SPeeDD (System to Prioritize Deletions and Duplications). For any gene under study, SPeeDD obtains the features described in Table 5 by using several Perl programs, databases and softwares. The candidate IDDs with annotated features is fed into the

SPeeDD classifier (Figure 16). Case and control sequences with annotated features are also used by the SPeeDD classifier as shown in Figure 16. SPeeDD then differentiates the candidates based on the training data set and divides the candidate IDDs for the gene under study into the most likely candidates and those that are less likely to undergo unequal recombination. SPeeDD predicts and ranks the candidates IDDs. For a given gene, SPeeDD calculates a confidence score for every compatible pair of homologous sub-sequences and ranks them based on that score.

Data mining or machine learning in the SPeeDD system is performed using Weka – an open source machine-learning software package (Witten and Frank 2005). It consists of several data pre-processing tools and supports several machine learning methods including those used in this research: artificial neural networks, decision trees, k-Nearest Neighbor, support vector machines, logistic model tree. A variety of different machine learning methods were examined using the collected dataset. The performance and error rates among the methods were compared. The best method was selected based upon error rate, sensitivity and specificity.

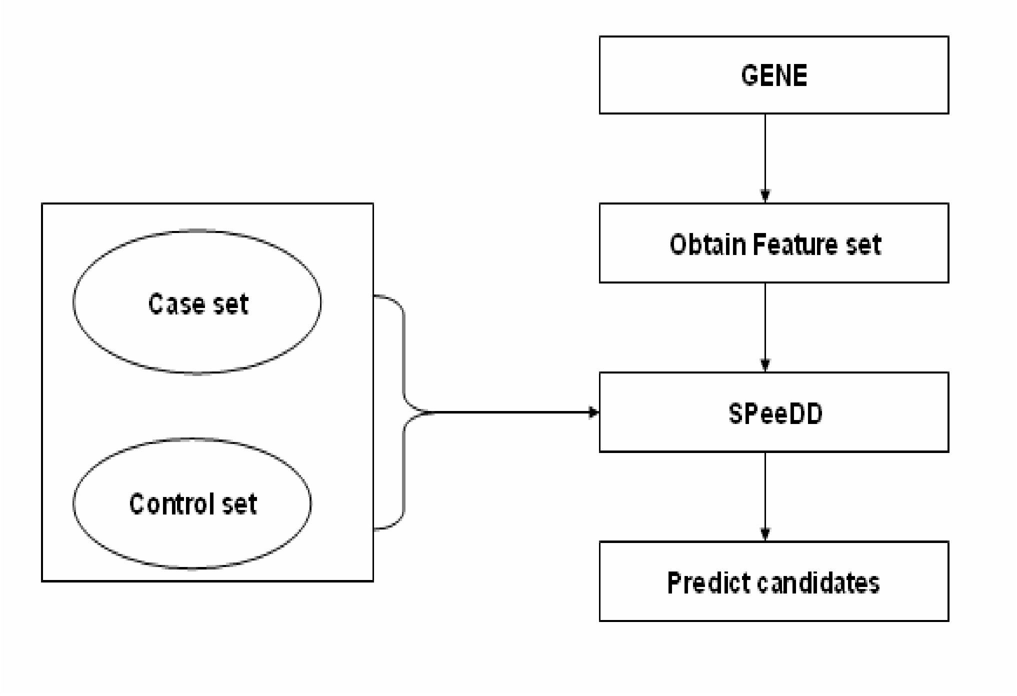


Figure 16 Flow diagram of the computational system (SPeeDD). For any gene of interest the feature set data as shown in Table 5. The case and control feature set data for all known intragene deletions or duplication is obtained as the training data set input for the SPeeDD system. SPeeDD performs analysis based on the training data and predicts the candidates for the gene under study.

Validation of the system

Cross-validation and hold-out (split-sample) methods are commonly used for estimating the performance and generalization error. We used N-fold cross-validation method on the collected dataset to train and evaluate the SPeeDD system. This method trains on N-1 subsets and holds out one data set for testing. The basic structure of the N-fold cross-validation procedure is as follows

1. Shuffle the data randomly
2. For loop i = 1 to N
    - a. Reserve one of the N subsets as the validation set
    - b. Train with other N-1 subsets
    - c. Test on the validation set

3. Report the average performance across N trails

A value of 10 for N is frequently used for estimating generalization error. Hence we used 10-cross fold validation to evaluate methods. Efficacy of a method is measured in terms of sensitivity and the specificity.

Two class confusion matrix

Performance of the classification systems is typically presented in the form of a matrix called a confusion matrix. A confusion matrix contains information about correct and predicted classification. Table 6 shows a confusion matrix for a two class classifier. The performance of a method can be described in terms of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) predictions. TP, TN, FP and FN are the four different possible outcomes for a two class classifier as shown in the table. True positives and the true negatives denote the two correct classifications. Conversely, when a case input is incorrectly classified as control candidate it is termed a false negative; and when a control input is incorrectly classified as a case candidate it is termed a false positive.

Table 6 Confusion Matrix of a two class classifier.

	Predicted Case	Predicted Control
Actual case	TP	FN
Actual Control	FP	TN

Performance Evaluation in terms of sensitivity and specificity

The performance of the classification is measured in terms of sensitivity and specificity. The sensitivity of a set of predictions is described as the percent of positives that are correctly classified and the specificity is defined as the percent of negatives that are correctly predicted (Weiss and Provost 2001).

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

CHAPTER V
RESULTS AND DISCUSSION

Despite the advances to identify small insertions or deletions (1-50 bp) using the sequence-based methodologies and cytogenetics methods (utilizing light microscopy with high-resolution chromosome banding techniques) to identify the large rearrangements (50 kb – 5 Mb) it is still hard to identify the intermediate level variations (50 bp – 50 kb) (Weber et al., 2002; Iafrate et al., 2004; Conrad et al., 2006; Bhangale et al., 2005; Sharp 2006). It is laborious and expensive to screen all the possible candidates to identify the intermediated level variations within a gene. Hence, the goal of this research is to design and implement a high-throughput computational system (SPeeDD) to identify and prioritize intermediate level genomic variations with possible deletion or duplication candidates within a gene. This was accomplished by computing and mining informative sequence features to build a model for the genes with previously published disease causing deletions or duplications. Using this data a machine learning model is trained and used to predict deletion or duplication candidate regions for genes of interest.

Construction of the testing set

HGMD database consists of about 6.5% of gross deletions or duplications (Cooper and Krawczak 1996; Krawczak et al., 2000). However, only non-lethal mutations that cause observed phenotypes might have been captured by this statistic. Gross deletions or duplications that are lethal and those that cause mild phenotypes might have been missed by the above statistics; thereby under estimated the true number of mutations.

Few reasons why we could not use all the deletion or duplication mutations provided by HGMD are: 1) The HGMD mutation description of gross deletions or duplications does not have the clarity in specifying the exact regions of the mutation in the genes. 2) They do not have précised information about where exactly sequences recombined; they have the information of what exons were lost or inserted and provided a link to the paper that published the mutation.

Two reasons why HGMD has less clarity in describing their genomic deletions or duplications are:

1) Assays to detect these mutations are different; some researchers would identify the mutations in a cDNA level and some in a genomic level.
2) Mutations are not fully characterized (some researchers would not sequence the region that was deleted to identify the exact break point regions).

All of the gross rearrangements available from HGMD with publications were obtained and stored in our University of Iowa human gross deletion or duplication database as mentioned in the methods chapter IV of this thesis. The scientific publications were searched to obtain the exact breakpoints of the previously published IDDs. A total of 1463 publications were examined, and the set of fully characterized gross deletions or duplications were collected. In a few cases, we also contacted the corresponding author of the publication for more information on breakpoint sequences. Through literature mining and direct contact with the authors I collected 102 fully characterized intragene deletions or duplications. Most of these cases are deletions with only two instances of duplications.

This is unusual because, in general, unequal recombination events are known to result in one copy containing a deletion and the other copy containing duplication; but there are more cases of deletions causing a phenotype than duplications. One reason for this apparent bias would be if duplications are more lethal than deletions. Another possible reason is that assays used to detect IDDs may have greater ability to resolve/detect deletions than duplications. This is likely true of hybridization-based methods such as comparative genome hybridization, where the ratio of a deletion to a control is 1:2 and the ratio of duplication is 3:2. They are both a 50% change in comparison to a normal control, but the deletion is a 2x difference rather than 1.5 x differences and is therefore easier to detect.

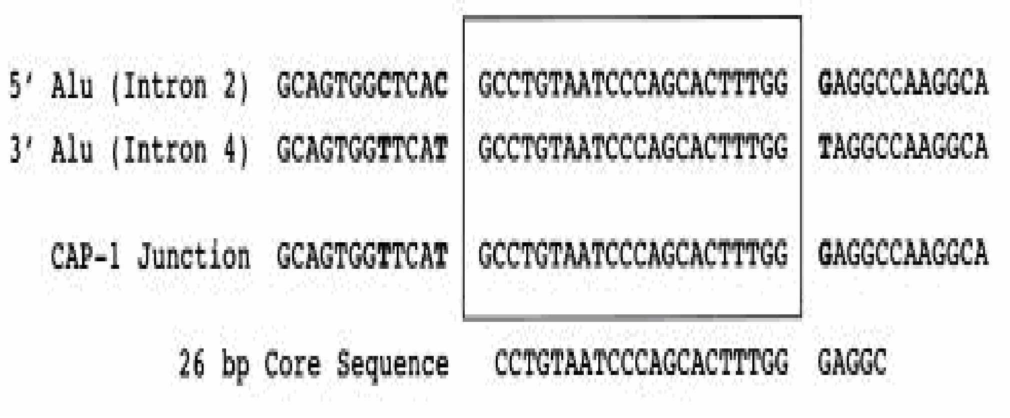
After collecting 102 different cases of previously identified deletions or duplications, we identified the exact pair of homologous sequence from potential pair of recombinatory sequences. These are the breakpoints of the IDDs. For each case there are typically numerous combinations of similar sequence that could recombine, resulting in IDDs with identical consequence to the gene structure (e.g. deletion of a specific exon or exons). One would reasonably expect that these variants would be indistinguishable in their phenotypic consequence. The case and control DNA sequences were obtained from the BL2SEQ output (The BL2SEQ approach used to obtain the exon spanning recombinatory sequences is described in detail in the methods chapter IV of the thesis).

Construction of the case and control sets

The case and control sequences were obtained for the 102 fully characterized mutations. Identification of case and control sequences for one of these 102 is described in detail for clarity. In the gene BBS4, deletion of exons 3 and 4 have been shown to be a causative mutation for Bardet-Biedl syndrome. An example of the fully-characterized mutation is shown in Figure 17. Interestingly, this deletion of exons 3- 4 was reported in two different families from different ethnic populations (Mykytyn et al., 2001; Nishimura et al., 2005). When molecularly characterized, both instances of the deletion shared the exact same Alu-based breakpoints in introns 2 and 3. When the deleted region was sequenced they found that the sequences share high similarity and that the deletion breakpoint region consists of 26 bp sequence (Figure 17) that was reported to be a possibly recombinogenic sequence (Rudiger et al., 1995); this short well conserved region of Alu-sequences is known to be involved in human gene rearrangements and is known to have homology with the prokaryotic chi (Rudiger et al., 1995). Below is a description of how we identified these fully characterized mutations from the BL2SEQ output of the BBS4 gene.

After obtaining the article, the deletion with the exact breakpoint sequences was identified for the BBS4 gene (Nishimura et al., 2005), we obtained the genomic sequence of the gene (accession number NM_033028), 5 kb upstream and 5kb downstream. The sequence acquired was aligned to itself using the BL2SEQ software. Perl programs were used to parse and obtain the sequences from the BL2SEQ output. As described in the methods the blast hits that were redundant or failed to meet our significance criteria were removed, as shown in Figure 11 to obtain a list of potential candidate breakpoints. The output of the BL2SEQ consisted of sub sequences that were similar and in same (plus/plus) and different (plus/minus) orientations. In general, the Plus/plus hits are higher than plus/minus hits. The 62 kb sequence (BBS4 gene with the 5 kb upstream and downstream sequence) had 5457 pairs of subsequences from the BL2SEQ output prior to the removal of redundant and low-quality candidate breakpoints (Figure 18). After filtering the blast hits based on the criteria shown in Figure 18, 1270 candidate breakpoints remained. Out of the 1270 candidate breakpoints only 32 pairs were consistent with deletion of exons 3-4. I went through these manually and identified the homologous sequence pair that was found in the BBS4 families from the publication in the 32 pairs of sequences likely to delete exons3-4. Based upon this analysis, the candidate breakpoint that matched the fully characterized breakpoint sequence for the BBS4 was defined as the control set and the remaining 31 candidate breakpoints were defined as the control set.

Similarly we collected 102 fully characterized intragene deletions or duplications (IDDs) from which the case data set was constructed. Furthermore, we have also collected an extra 2338 potential homology-based candidates from the same set of genes for which IDDs have never been reported and called it as control data set in this study. In other words, the control data consists of surrounding sequences where there is an a priori equal chance of generating the same exonic loss or gain but which have not been previously observed.



5' Alu (Intron 2)	GCAGTGGCTCAC	GCCTGTAATCCCAGCACTTTGG	GAGGCCAAGGCA
3' Alu (Intron 4)	GCAGTGGTTCAT	GCCTGTAATCCCAGCACTTTGG	TAGGCCAAGGCA
CAP-1 Junction	GCAGTGGTTCAT	GCCTGTAATCCCAGCACTTTGG	GAGGCCAAGGCA
26 bp Core Sequence		CCTGTAATCCCAGCACTTTGG	GAGGC

Figure 17 The sequence flanking of 5' and 3' Alu elements for the *BBS4* exons 3 and 4 deletions (Nishimura et al., 2005). The region in which the breakpoint has occurred within each *Alu* element is shown in a box. The 26-bp core sequence in the *Alu* element identified by Rudiger et al. (1995) is also shown in the figure.

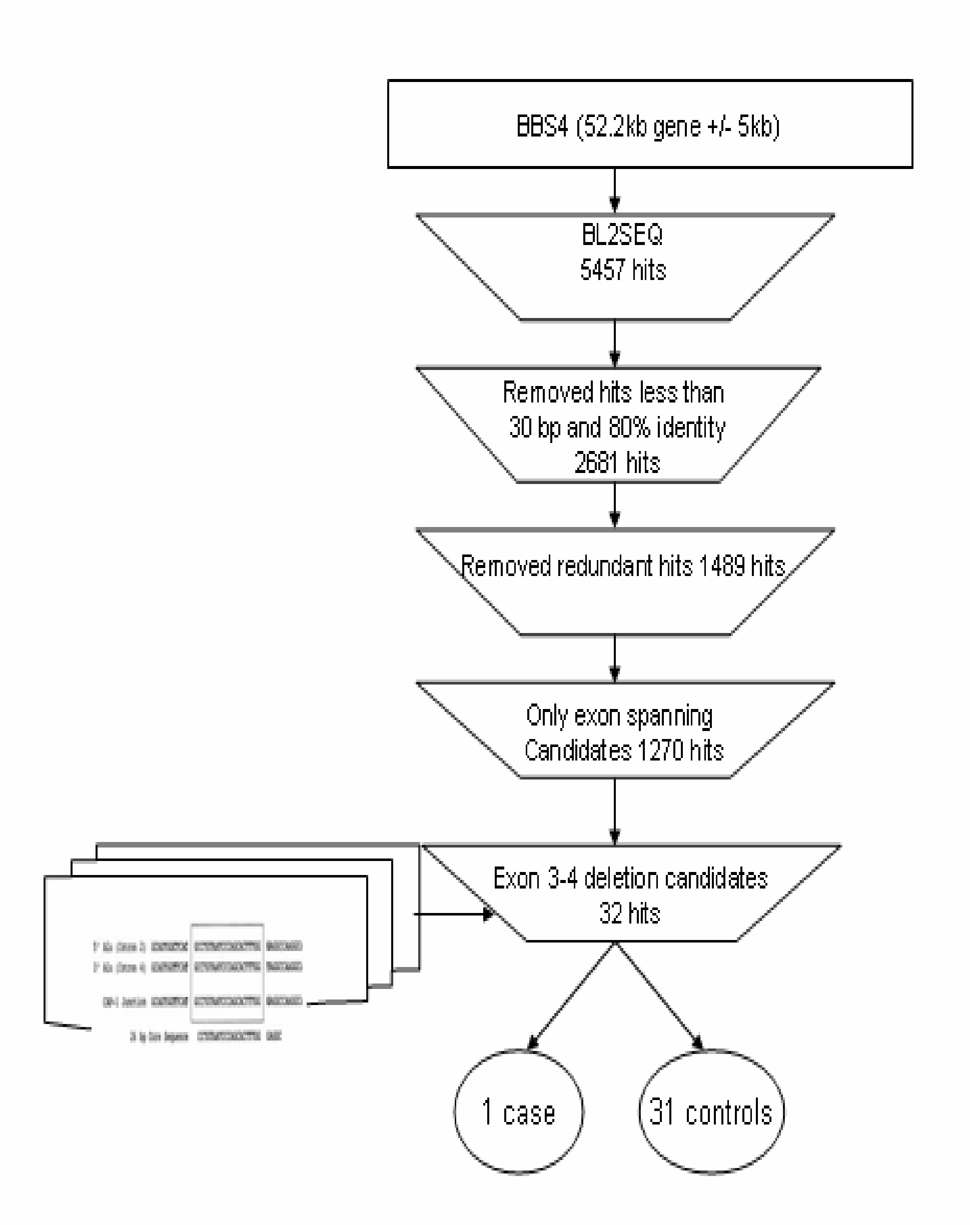
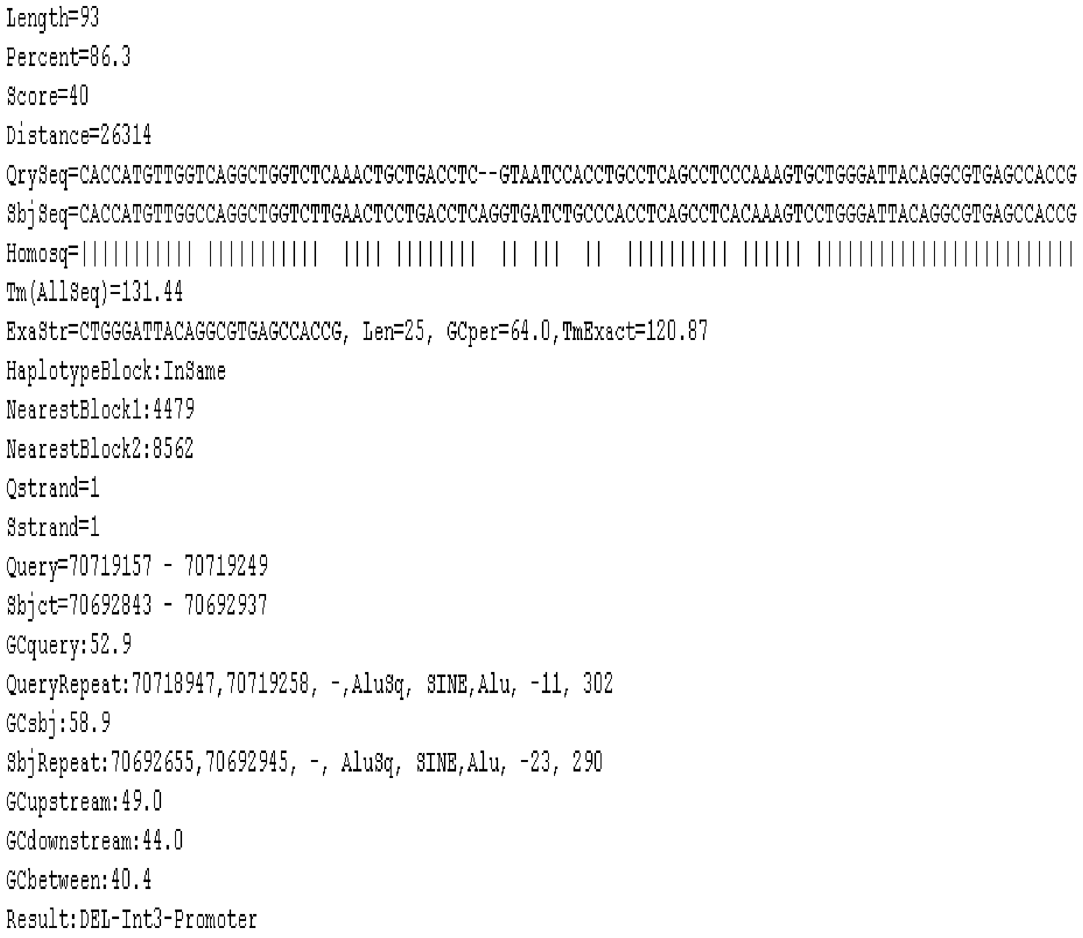


Figure 18 Flowchart of the case and control sequences in BBS4 exons 3-4 deletion

Computation of the pertinent features

Data mining is a widely used technology in biomedical research, and has been used to identify patterns in microarray expression data, promoter discovery, promoter modeling, disease causing mutations based on sequence annotation and discovery of regulatory elements. In this research I have applied data mining techniques to identify and prioritize candidate IDDs. Identification and collection of the data to be mined is often the most challenging component in data mining. After data collection, preprocessing and cleaning the data are also the toughest jobs. In order to eliminate the limitations of a single data source the data used in this research was collected from multiple sources. At times there were difficulties including various sources such as the genomic build they considered while doing their analyses. Although the current human genome assembly is NCBI’s build 36, to be consistent with our analysis we used NCBI’s build 34 across all various sources and data. This was necessary to incorporate the haplotype data set from Perlegen Sciences, which is currently only available on build 34.

Below is an example of a IDD candidate that is selected from an actual file of the BBS4 gene obtained with details of the features (Figure 19).



```
Length=93
Percent=86.3
Score=40
Distance=26314
QrySeq=CACCATGTTGGTCAGGCTGGTCTCAAACTGCTGACCTC--GTAATCCACCTGCCTCAGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCG
SbjSeq=CACCATGTTGGCCAGGCTGGTCTTGAACTCCTGACCTCAGGTGATCTGCCCACCTCAGCCTCACAAAGTCCTGGGATTACAGGCGTGAGCCACCG
Homosq=||||||||||| |||||||||||  |||| ||||||||  || |||  ||  |||||||||| |||||| |||||||||||||||||||||||||
Tm(AllSeq)=131.44
ExaStr=CTGGGATTACAGGCGTGAGCCACCG, Len=25, GCper=64.0,TmExact=120.87
HaplotypeBlock:InSame
NearestBlock1:4479
NearestBlock2:8562
Qstrand=1
Sstrand=1
Query=70719157 - 70719249
Sbjct=70692843 - 70692937
GCquery:52.9
QueryRepeat:70718947,70719258, -,AluSq, SINE,Alu, -11, 302
GCsbj:58.9
SbjRepeat:70692655,70692945, -, AluSq, SINE,Alu, -23, 290
GCupstream:49.0
GCdownstream:44.0
GCbetween:40.4
Result:DEL-Int3-Promoter
```

Figure 19 Example of DNA sequence features collected for a possible deletion candidate

The features listed in figure 19 are collated from several different sources: sequence features based on the BL2SEQ identified breakpoints, haplotype block features and melting temperature features. Many of these features are interdependent. For example, there is a positive relationship between the length and GC content of the breakpoint sequence and the Tm of the breakpoint sequence. We have removed the features that are highly correlated because they add no extra information to the model, however if the variables are dependent but are obtained from different sources and measured in a different way then we have retained those variables.

Length, percent identity, score, strands, starting and ending positions of candidate IDDs breakpoint sequences are obtained using BL2SEQ program. Length attribute is the number of basepairs that are aligned between the query and subject sequences; as shown in Figure 19 the length of the query (QrySeq) and subject sequence (SbjSeq) is 93. Percent identity indicates the similarity between the sequences. The percent identify (Percent) is 86.3 between the query and subject sequences (Figure 19). The BL2SEQ output also returns the alignment of the query and subject sequences. The starting and ending positions of the query and subject sequence are converted into genomic positions as shown in Figure 19 (Query and Sbjct attributes). The starting genomic position of the query sequence (Query) 70719157 is subtracted from the ending genomic position of the subject sequence (Sbjct) 70692937 to obtain the distance 26314 between the IDD candidate breakpoints (Figure 19). Query sequence (Qstrand) and subject sequence (Sstrand) strands are the orientations obtained from the BL2SEQ output.

Few parameters shown in Figure 19 are obtained using computational methods. Below is the detailed description of the features obtained using computational methods as shown in Figure 19. Repeat characteristics of the IDD candidate breakpoints are obtained by using UCSC database and simple Perl programs (QryRepeat and SbjRepeat). As shown in Figure 19 Perl program is used to obtain the length and sequence of the longest uninterrupted exact matching string (ExaStr). Sequences of the longest uninterrupted string (ExaStr), query sequence (GCquery), subject sequence (GCsbj), upstream of the query sequence (GCupstream), downstream of the subject sequence (GCdownstream) and the likely deleted sequence (GC between) are obtained and the GC content of the sequences (Figure 19) are calculated using Perl programs. Gene structure is obtained using the UCSC database and possible result of the IDD candidate is programmatically determined (Result); as shown in Figure 19 the result describes the possible deletion of part of promoter region, exons 1, 2 and 3.

The melting temperature characteristics that were used in this study were obtained from the TmAlign software (unpublished software). Below is the detailed description of the features obtained with that program as shown in Figure 19. The TmAlign features were calculated for the breakpoint pairs identified by BL2SEQ, Tm(AllSeq), or the longest exact matching sequence, TmExact.

Haplotype features were obtained from the haplotype block information that was developed by Perlegen Sciences. These features included a qualitative assessment of where the breakpoints are located with respect to haplotype blocks (HaplotypeBlock), and distances upstream (NearestBlock1) and downstream (NearestBlock2) to the nearest haplotype block boundary. In Figure 19 the breakpoints are located in the same haplotype block (InSame) and are 4479 and 8562 bp from the upstream and downstream haplotype block junctions.

After collecting the data, preparation of the data for actual analysis is done by determining the statistical variability in the data, validating the data values in the context of the biology, creating a high level picture of the nature and the content of data to be mined, dealing with missing data values and transforming data values from one representation to another when necessary.

Out of all the features collected, only the following features were utilized in the final model to classify the control and case data sets efficiently. The features used are length, percent identity, GC content, melting temperature and bl2 score (bl2score) of the candidate breakpoints, the distance and GC content between the breakpoints, the haplotype block status, and the nearest haplotype block boundary to both ends of the IDD. One final feature is included that specifies the classification of a given entry. This feature takes the value of YES for cases and NO for controls.

Building the classifier models

The classification model was built using the feature data for the case and control set described above. In our study, the classification system was trained on case and control data sets to predict intragene deletion or duplication (IDD) candidate regions within a gene and to prioritize them based on a confidence score. A knowledge flow diagram of the classification system is shown in Figure 20. As shown in Figure 20 the classifier was built on top of the Weka machine learning system, requiring that the feature data for the training set and candidates to be evaluated be converted into Arff format (the data format used by Weka). Each record in an Arff data set describes the set of features for an IDD or candidate IDD. Each record is assigned to a class or outcome as shown in Figure 20. The outcome (classification feature) of the training data set is set to TRUE for the case data set and FALSE for the control data set in which there is no evidence of deletion or duplication occurrences. Based on the training data, the classifier will determine the properties to distinguish the case and control data sets. A Variety of classifiers are applied and the machine learning method that performed best will be selected. The performance of various methods was evaluated using cross validation methods on the training data set. The information of the classification function (best performed machine learning method) is stored in a classification model and further applied to predict the outcome in case of an unknown set.

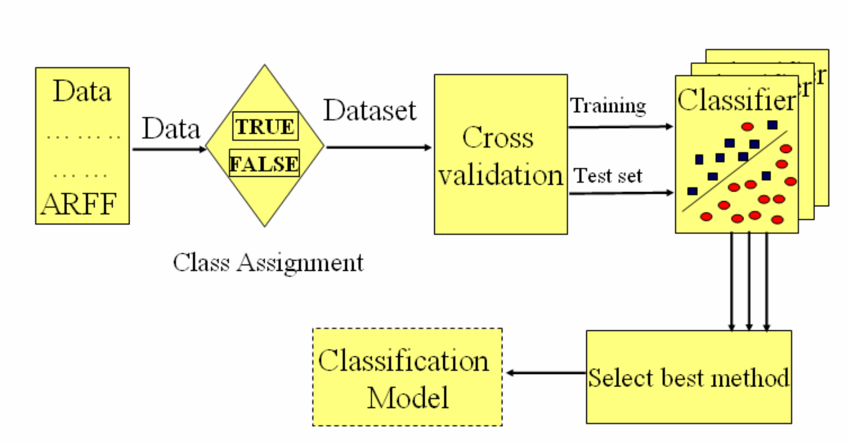


Figure 20 Knowledge flow diagram of the classification model

Validation

Fitness of a model is evaluated using cross validation methods. Three cross-validation techniques that are commonly used are hold out method, N-fold cross validation and the leave-one-out method. Cross-validation methods predict the performance of a machine learning method on unseen data by dividing the training set into two parts, training on the first part, and evaluating the performance on the second part. This allows us to estimate the amount of error based on what is being learned by the given model from the training data subset.

N-fold cross validation

I used 10-fold cross validation to evaluate several machine learning methods available in the Weka system (Witten and Frank 2005). In 10-fold cross-validation method the method is trained on 9 subsets and performance is tested on the one remaining data set; this process is repeated 10 times and the average performance is reported across 10 trails. The advantage with the 10-fold cross validation method is it estimates the generalization error on 10 different subsets as an alternative of only a single subset that is used during the hold out method validation (Weiss and Kulikowski 1991; Hjorth 1994; Plutowski et al., 1994; Shao and Tu 1995).

Selecting the appropriate machine learning system

Multiple machine learning models including artificial neural networks, decision trees, logistic model tree, simple logistic, simple Naive Bayes, k-nearest neighbor and support vector machines were applied using the 10-fold cross validation method on the case and control datasets. The performance of the algorithms on various data sets was compared to predict the best model. The determination of the best model was performed based on the performance of predicting the true positives, false positives, ability to deal with the missing data, noisy data and the ability to explain the classification. Sensitivity of the system varied from 20% to 74.2% but the specificity exceeded 90% for all the methods that were assessed (Figures 21 and 22). One reason, for the high specificity is the utilization of an unbalanced data set. The majority of the training data set consisted of the control sequences. Only 4.2% of the data set consisted of the cases. This however is the case in most of the real-world problems; in which there are always fewer cases than controls.

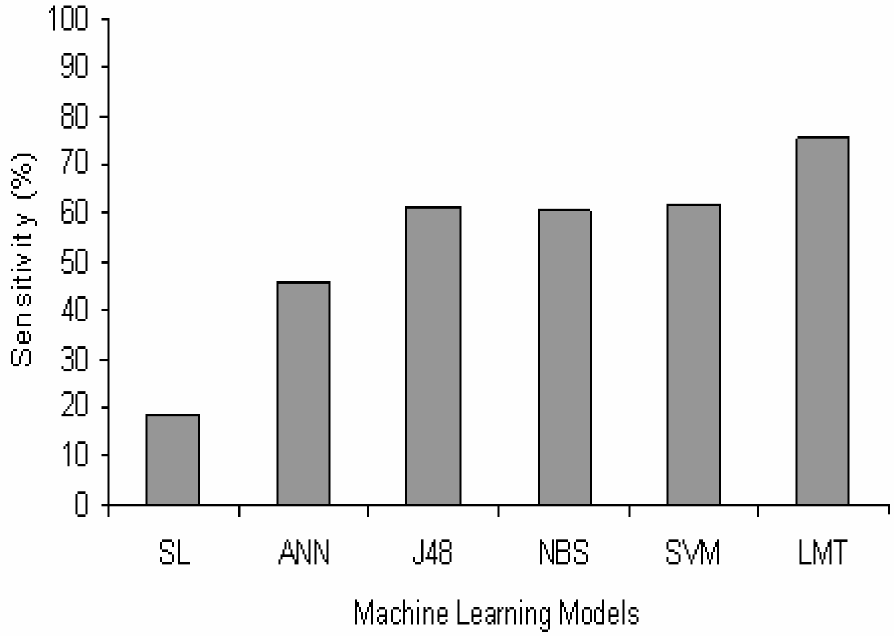


Figure 21 Sensitivity of various models using 10-fold cross validation method. The models tested include simple logistic regression models (SL), artificial neural networks (ANN), a decision tree (J48), simple naïve Bayes (NBS), support vector machine (SVM) and logistic model tree (LMT).

Logistic models uses logit boost with simple regression functions for fitting (Rice 1994). The classifier to build simple logistic (SL) regression models gave low sensitivity value. Reason why SL gave low sensitivity value could be due to the fact that the features are highly inter dependent. Due to low sensitivity value this method was not used as final classification function for our model.

As shown in Figure 21 the sensitivity of the artificial neural network (ANN) model is greater than simple logistic method. Unfortunately, in general, the decisions made with the ANN methods are not easily interpreted and its knowledge representation is also poor. Few reasons why we did not choose this as our final model to train the system is because of its low sensitivity and also due to the fact that the decisions made by this method are not easily interpretable in the biological context.

The performance of simple naive bayes (NBS), decision tree (J48) and support vector machine (SVM) methods were similar in terms of sensitivity and specificity. Specificity of all three methods exceeded 95. Sensitivity is between 60-65 for SVM, NBS and J48 methods (Figure 21 and 22).

Logistic model trees (LMT) are classification trees with logistic regression at the leaves. LMT which is a combination of the logistic regression and decision tree method yielded the best results with sensitivity and specificity of 74.2 and 97.2 respectively (Figure 21 and 22). In case of the LMT model the rules implicated in making decisions are easily interpretable which gives an insight into how the classifier works. The decision tree that is being generated by LMT is an easy to interpret white box model. The challenge with a 10-fold cross validation is that for every iteration you are throwing out 10% of your training set to evaluate. This reduction in training set size can have severe consequences to the performance of a classifier. Therefore, I also used the leave-one-out validation strategy to evaluate the various machine learning models using the maximum amount of data for use in training the methods.

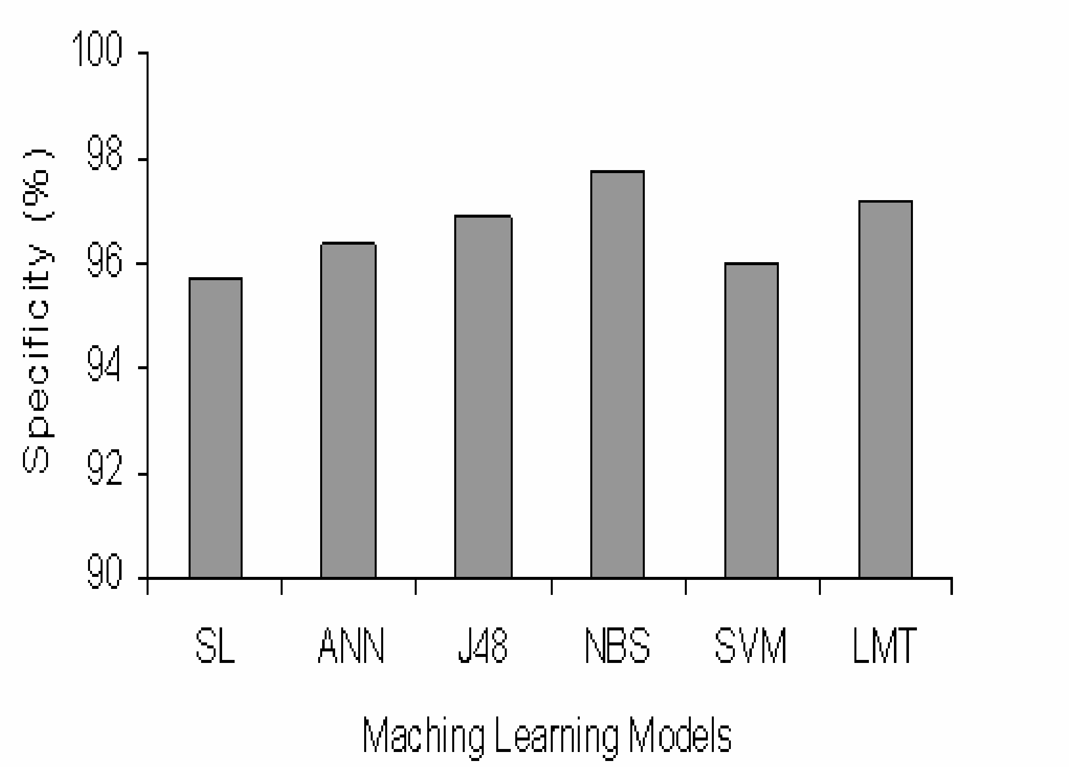


Figure 22 Sensitivity of various methods using 10-fold cross validation method. The models tested include simple logistic regression models (SL), artificial nueral networks (ANN), a decision tree (J48), simple naïve Bayes (NBS), support vector machine (SVM) and logistic model tree (LMT).

Leave-one-out method

In the leave-one-out approach each entry in the case data set was evaluated based upon training with features of the remaining 101 cases. This maximized the size of the training set, and also provided an estimate of the expected performance for the final system. Following the construction of the classifier and the evaluation with the 10-fold cross validation it is clear that the LMT algorithm performed best of my data set. Hence we used the best performed method (LMT) to train and predict the outcome of the test case. Based on the training data set the model predicts the test case; it makes a decision if it is a disease causing deletion candidate or not. Leave-one-out method is a different way of evaluating the performance of the model. The advantage of leave-one-out method is it allows to train on a bigger dataset than in N-cross validation. Leave-one-out method identified the left out mutations with an accuracy of 81%.

Sensitivity to training set size (expected performance)

We evaluated the performance of the system using variety of input sizes to estimate the benefit of larger models. In other words, how much can we expect the performance to increase as more IDDs are identified and included in the training set. To address this question, a Monte Carlo simulation was performed. In this example, training sets of decreased size were randomly selected from the complete set of 102 fully characterized cases. In this experiment we varied the size of the randomly constructed training set between 50 and 100 in steps o f 5 to study the impact on the classifier of varying the size of the training set. For each input case data set size 100 random training sets were generated and evaluated with the 10-fold cross validation to assess performance. The results of the model with respect to varying sizes and sensitivity are shown in Figure 23. The slope of the curve gradually increased throughout. As expected larger training sets result in better performance – here expressed as the sensitivity. Results indicate that there is a steady improvement in the beginning rather towards the end of the curve. In input sizes 85 to 100 we did not see any tremendous improvement in terms of sensitivity of the model. This experiment also proved that by increasing the training input size the sensitivity will also increase, the sensitivity has increased from 35% to 70% for input sizes 50 to 100. This shows that the larger the case data set positively correlates with the sensitivity.

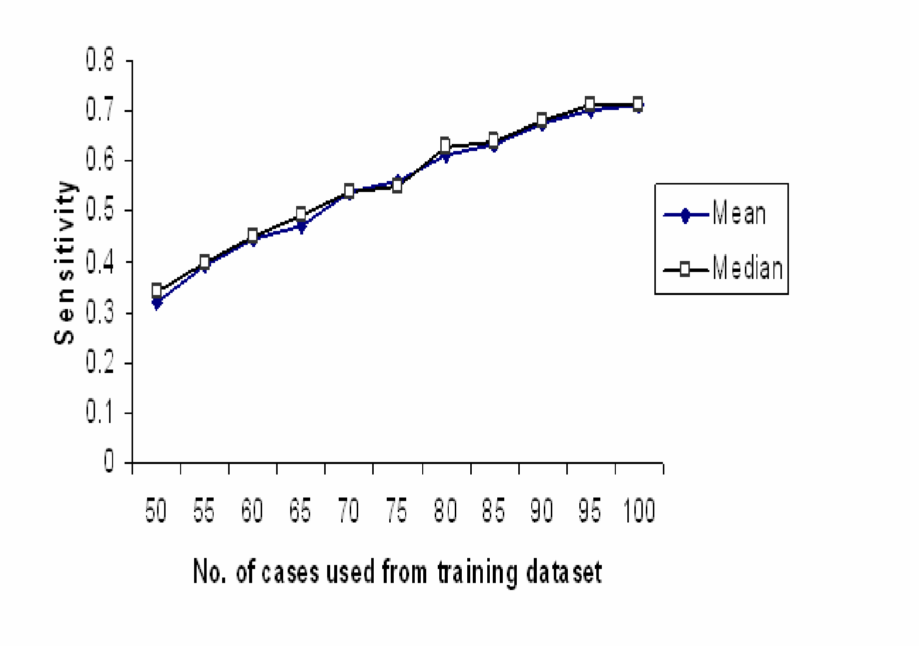


Figure 23 Monte-carlo simulation of various input datasets

Role of each feature in predicting the candidates

Importance of each feature in predicting the true positives was evaluated. At a time only one feature was removed from the features collected and 10-fold cross validation was applied to see how well the system predicts the IDD candidates. Influence of each feature was estimated based on the performance of predicting the true positives from the set of IDDs after removing the feature from the case and control data sets.

As shown in Figure 24 removal of GCbwn (GC between the IDD candidate breakpoints) feature has a big impact on the model and reduces the power of true prediction drastically. Percent identity, distance between the IDD breakpoint sequences also has a impact on the model if it is removed (Figure 24).

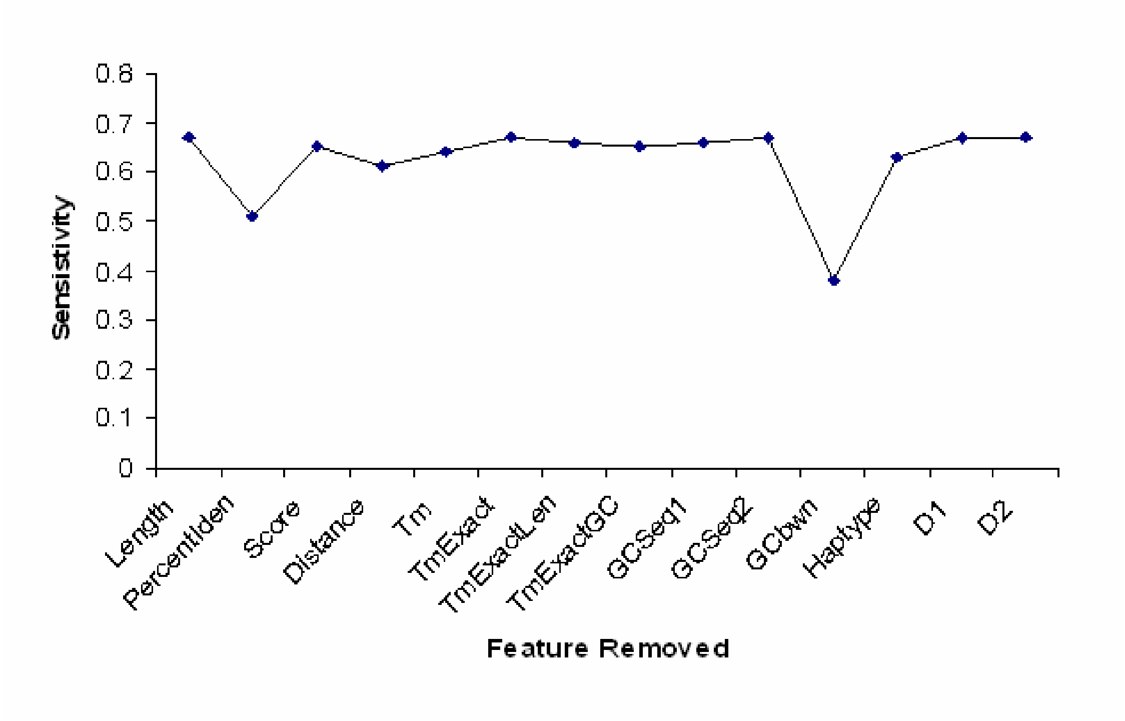


Figure 24 Influence of each feature on predicting the IDD candidates. This experiment was conducted by removing each parameter at a time and observing how well the system predicts the true positive candidates.

System implementation (SPeeDD)

Recent studies have made progress to understand the trends in the homologous recombination mechanism (Abeysinghe et al., 2004; Sen et al., 2006). But to our knowledge there is no system that predicts and prioritizes deletion candidates within a gene. SPeeDD – System to Prioritize Deletions or Duplications is a bioinformatics system that has been developed to predict the most likely IDD candidates by applying data mining techniques to the set of features identified above of previously identified deletions or duplications (i.e., the training data set). SPeeDD is an interactive visualization tool, designed to aid researchers to generate hypotheses and focus their efforts in identifying intragene deletions and duplications. The system can predict the prioritized deletion or duplication candidates for any specified gene of interest. This software with number of valuable options for the scientist may provide a fast and less expensive method to identify deletions or duplications.

SPeeDD Computational Pipeline

The computational pipeline of the SPeeDD system is shown in Figure 25. SPeeDD consists of both bioinformatics and machine learning methods. The system diagram in Figure 25 is divided into two sections. The section on the left presents the path taken to obtain the case and control data sets. The section on the right of Figure 25 shows how IDD candidates are identified for a specified gene by the SPeeDD system. Note that he output of the left hand side (the case and control data sets) is an input to the machine learning component on the right hand side.

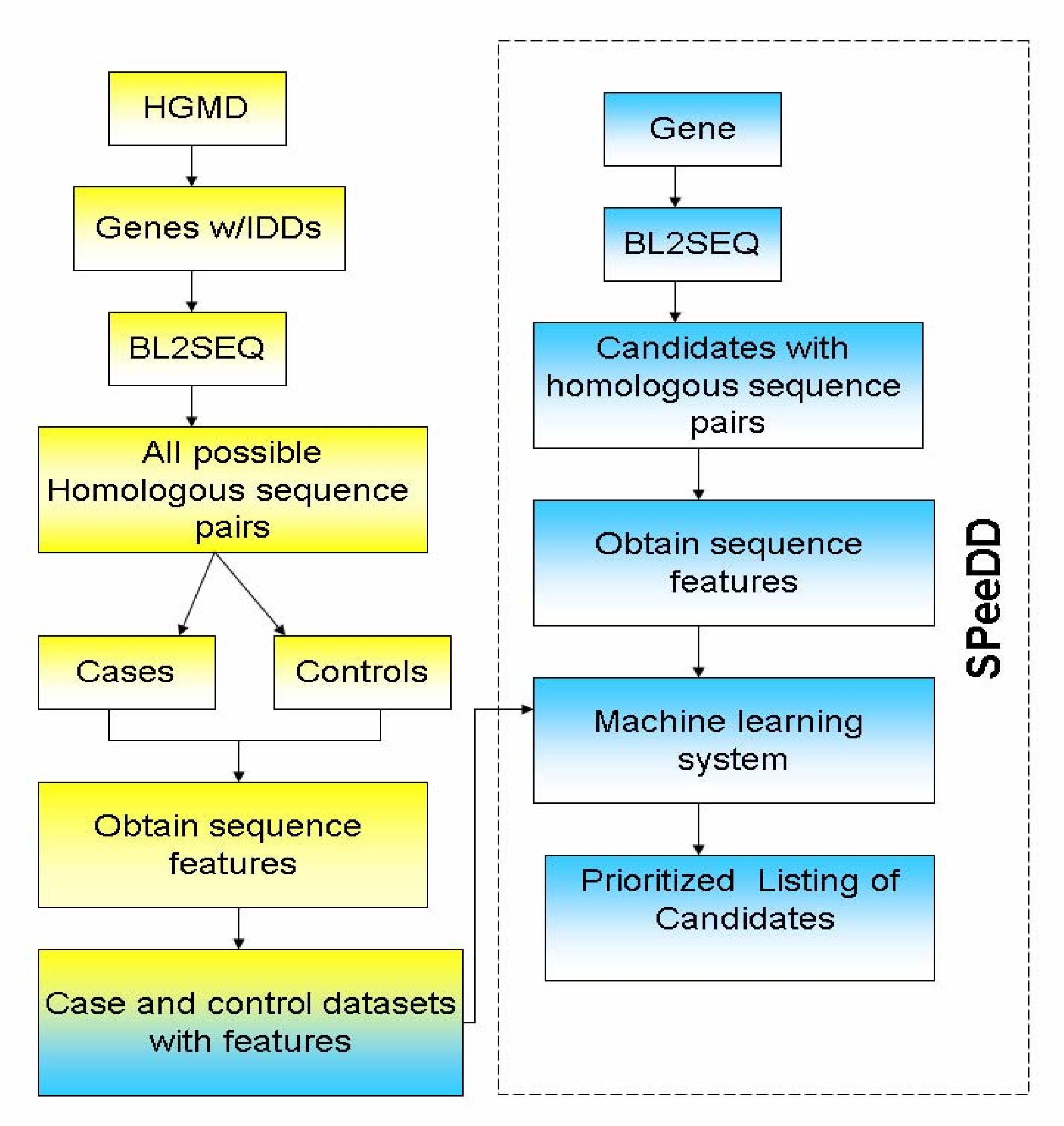

HGMD
Genes w/IDDs
BL2SEQ
All possible Homologous sequence pairs
Cases
Controls
Obtain sequence features
Case and control datasets with features
Gene
BL2SEQ
Candidates with homologous sequence pairs
Obtain sequence features
Machine learning system
Prioritized Listing of Candidates
SPeeDD

Figure 25 Computational pipeline of the SPeeDD system

Case and control data set flow diagram

The left hand side (boxes in yellow color) of the flowchart illustrates the steps that were followed to obtain case and control data sets to train the SPeeDD system. As shown in Figure 25 HGMD was used to obtain all links to previous journal publications with known deletions or duplications. These publications were then manually searched to obtain the set of genes with fully characterized breakpoints. 102 previously published intragene deletions or duplications (IDDs) with exact breakpoint sequences were collected for the study. Potential IDDs were identified from the set of homologous sequence pairs obtained by blasting the gene sequence against itself (BL2SEQ). Published breakpoints for the 102 fully characterized IDDs (case set) were identified from the potential recombinant set obtained from the BL2SEQ output. Surrounding potential sequences that could delete or duplicate same exon(s) in the BL2SEQ output, but were never observed in the patient population were utilized as a control set. For all the case and control data sets the DNA sequence features were obtained using various data resources and software.

SPeeDD flow diagram

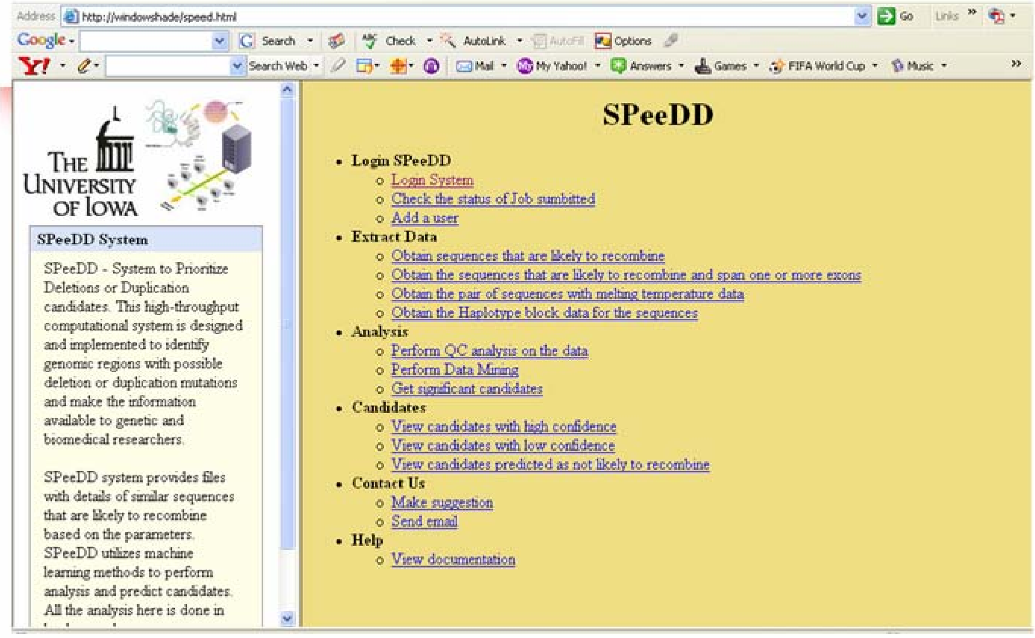
The right hand side (boxes in blue color) of the Figure 25 summarizes the analytic steps performed in the SPeeDD system follows to predict the deletion or duplication regions. As shown in Figure 25, the initial step is to obtain the genomic sequence for the specified gene and perform a self-self alignment with BL2SEQ to identify the homologous sequence pairs that are the candidate IDDs. The sequence, melting temperature and haplotype features are the computationally obtained for those candidate IDDs as described in the Methods chapter. The classifier, having been trained on the case and control datasets as shown in Figure 24, then prioritizes the resulting IDD candidates. Based on the knowledge gained with the trained data the machine learning model differentiates predicts the outcome of the homologous sub-sequences of the gene under study. SPeeDD predicts and prioritizes the list of candidates based on the confidence score that the machine learning model calculates.

SPeeDD Implementation

SPeeDD is implemented using several computational and bioinformatics platforms including Perl, Bioperl, MySQL and CGI. SPeeDD uses a collection of softwares such as BL2SEQ (NCBI's alignment software), TmAlign (IDT's melting temperature software), WEKA (data mining software in Java; Witten and Frank 2005) and databases such as UCSC, HUGO, HGMD, Perlegens Sciences database with haplotype blocks.

SPeeDD web-interface

Our implementation of SPeeDD is readily available on the web at http://public.eng.uiowa.edu/SPeeDD. As demonstrated by the validation results presented above, SPeeDD successfully identifies and enriches lists of deletion or duplication candidate regions for a given gene. The system provides an automated, unbiased method to save investigators time and effort when examining a gene for potential deletion or duplication candidate regions. Snap shots of our implementation is shown in the below Figures from 26- 31. Figure 26 is the snapshot of the home page of the SPeeDD system. SPeeDD system is password protected. It consists of a login page where user enters authorized login and password. After logging in successfully the SPeeDD system provides the user an option of entering a gene name or refseq id of interest. As shown in Figure 27, SPeeDD uploads and displays the fasta file of a gene of interest with the 5kb upstream and downstream of the gene. Following Figure 27, a page with various blast filtering options as shown in Figure 28 appears. This page has list of options to filter the BL2SEQ output. Default parameters for length, percent identity and distance are set to 30, 80 and less than 50 kb respectively. The system provides an option to change this if user desires by choosing one of the options in drop down list boxes. After selecting the options, various features (sequence specific, melting temperature and haplotype characteristics) are obtained for IDD candidates of that gene. Once the features are collected the data is converted into arff format to be handled by the Weka machine learning system (Witten and Frank 2005). The unknown outcomes of the IDD candidates are predicted by the machine learning model and prioritized based on the confidence score and displayed as shown in Figure 29. Details of the candidates are obtained by using the detail link in Figure 29. The output of this link is as shown in the Figure 30 with the details of the resultant of the deletion. It displays the total number of bases affected and details of the frame shift mutation. Genomic view for the candidate with a positive outcome is obtained thru the link in Figure 29. The genomic view of the true positive deletion candidate is displayed by utilizing the dynamically created custom track in the UCSC browser (Figure 31).



SPeeDD System

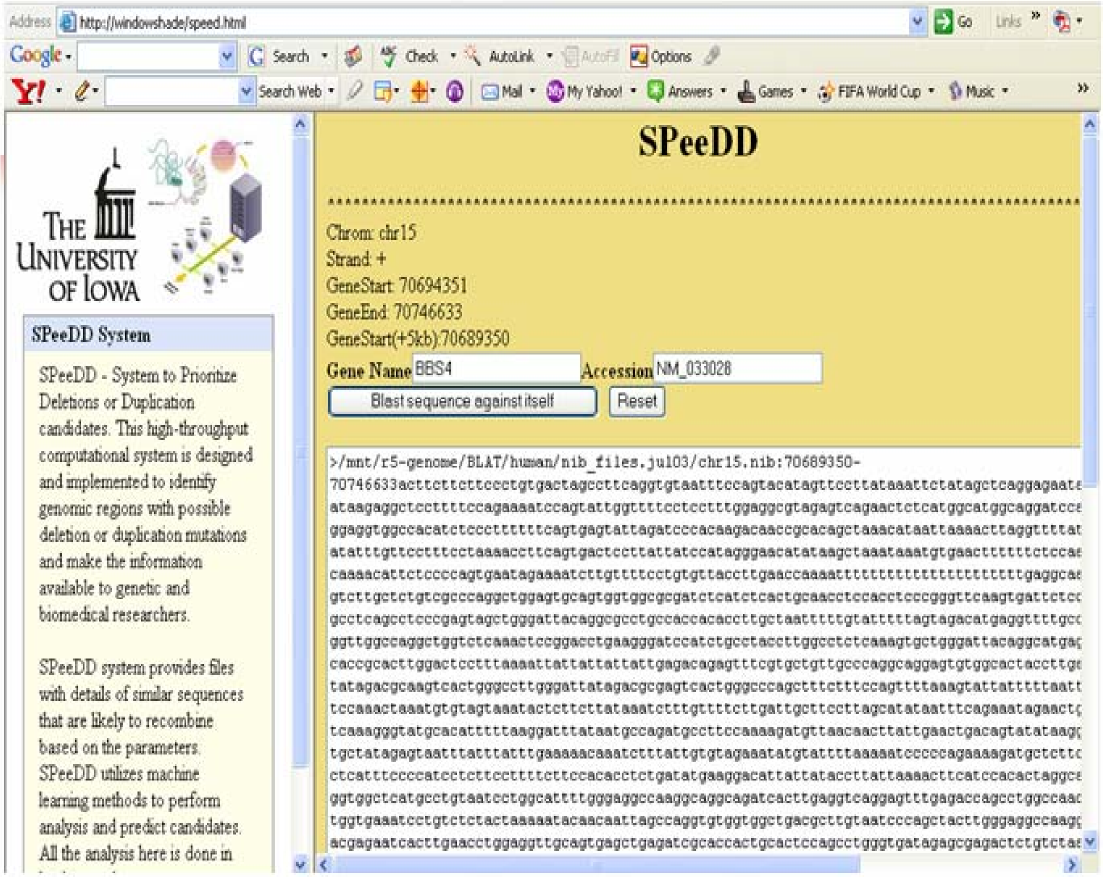
SPeeDD - System to Prioritize Deletions or Duplication candidates. This high-throughput computational system is designed and implemented to identify genomic regions with possible deletion or duplication mutations and make the information available to genetic and biomedical researchers.

SPeeDD system provides files with details of similar sequences that are likely to recombine based on the parameters. SPeeDD utilizes machine learning methods to perform analysis and predict candidates. All the analysis here is done in

SPeeDD

- Login SPeeDD
  - Login System
  - Check the status of Job sumbitted
  - Add a user
- Extract Data
  - Obtain sequences that are likely to recombine
  - Obtain the sequences that are likely to recombine and span one or more exons
  - Obtain the pair of sequences with melting temperature data
  - Obtain the Haplotype block data for the sequences
- Analysis
  - Perform QC analysis on the data
  - Perform Data Mining
  - Get significant candidates
- Candidates
  - View candidates with high confidence
  - View candidates with low confidence
  - View candidates predicted as not likely to recombine
- Contact Us
  - Make suggestion
  - Send email
- Help
  - View documentation

Figure 26 Screen shot of the SPeeDD system. This is the home of the system. On the left hand side it consists of a description of the SPeeDD. On the right hand side there are several options such as the login, analysis and help.



SPeeDD

****

Chrom: chr15
Strand: +
GeneStart: 70694351
GeneEnd: 70746633
GeneStart(+5kb):70689350

Gene Name BBS4 **Accession** NM_033028

Blast sequence against itself   Reset

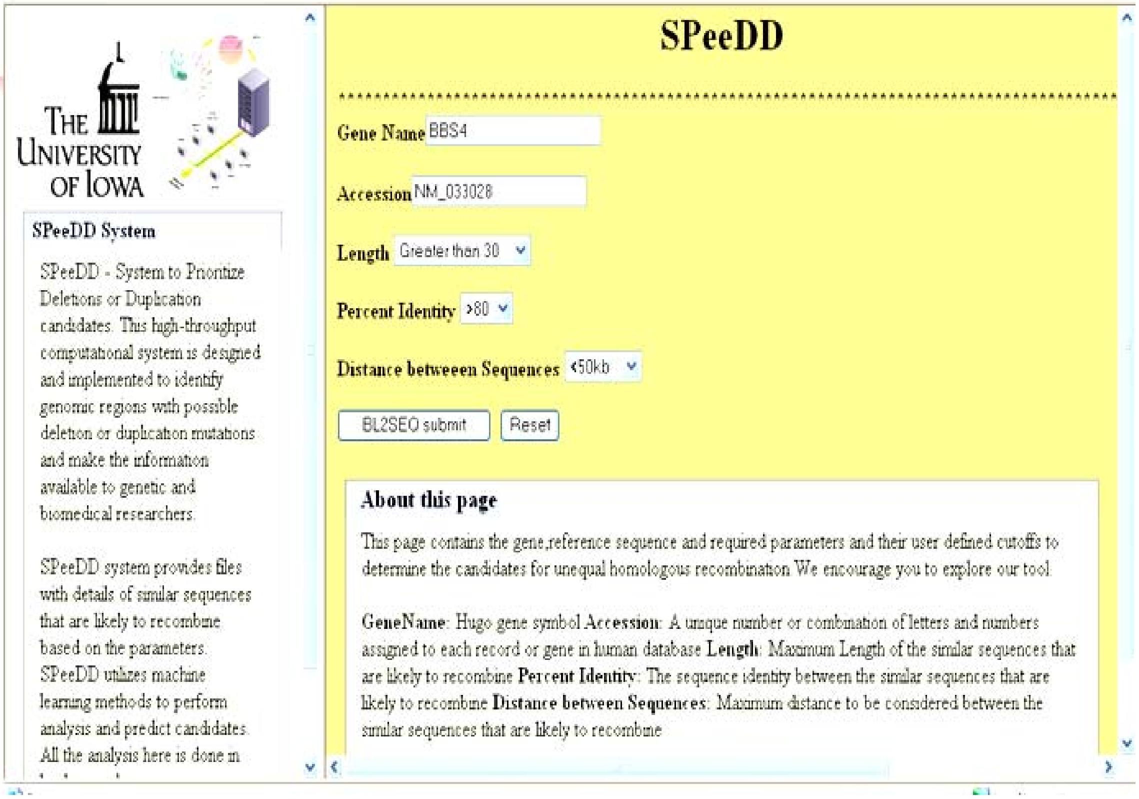
```
>/mnt/r5-genome/BLAT/human/nib_files.jul03/chr15.nib:70689350-
70746633acttcttcttccctgtgactagccttcaggtgtaatttccagtacatagttccttataaattctatagctcaggagaat
ataagaggctccttttccagaaaatccagtattggttttcctcctttggaggcgtagagtcagaactctcatggcatggcaggatcc
ggaggtggccacatctcccttttttcagtgagtattagatcccacaagacaaccgcacagctaaacataattaaaacttaggttttat
atatttgttcctttcctaaaaccttcagtgactccttattatccatagggaacatataagctaaataaatgtgaactttttttctcca
caaaacattctccccagtgaatagaaaatcttgttttcctgtgttaccttgaaccaaaatttttttttttttttttttttgaggca
gtcttgctctgtcgcccaggctggagtgcagtggtggcgcgatctcatctcactgcaacctccacctcccgggttcaagtgattctc
gcctcagcctcccgagtagctgggattacaggcgcctgccaccacaccttgctaatttttgtatttttagtagacatgaggttttgcc
ggttggccaggctggtctcaaactccggacctgaagggatccatctgcctaccttggcctctcaaagtgctgggattacaggcatga
caccgcacttggactcctttaaaattattattattattgagacagagtttcgtgctgttgcccaggcaggagtgtggcactaccttg
tatagacgcaagtcactgggccttgggattatagacgcgagtcactgggcccagctttctttccagttttaaagtattatttttaatt
tccaaactaaatgtgtagtaaatactcttcttataaatctttgttttcttgattgcttccttagcatataatttcagaaatagaact
tcaaagggtatgcacatttttaaggatttataatgccagatgccttccaaaagatgttaacaacttattgaactgacagtatataag
tgctatagagtaatttatttatttgaaaaacaaatctttattgtgtagaaatatgtattttaaaaatcccccagaaaagatgctcttc
ctcatttccccatcctcttcctttttcttccacacctctgatatgaaggacattattataccttattaaaacttcatccacactaggc
ggtggctcatgcctgtaatcctggcattttgggaggccaaggcaggcagatcacttgaggtcaggagtttgagaccagcctggccaa
tggtgaaatcctgtctctactaaaaatacaacaattagccaggtgtggtggctgacgcttgtaatcccagctacttgggaggccaag
acgagaatcacttgaacctggaggttgcagtgagctgagatcgcaccactgcactccagcctgggtgatagagcgagactctgtcta
```

SPeeDD System

SPeeDD - System to Prioritize Deletions or Duplication candidates. This high-throughput computational system is designed and implemented to identify genomic regions with possible deletion or duplication mutations and make the information available to genetic and biomedical researchers.

SPeeDD system provides files with details of similar sequences that are likely to recombine based on the parameters. SPeeDD utilizes machine learning methods to perform analysis and predict candidates. All the analysis here is done in

Figure 27 Screen shot of SPeeDD system. For the gene of interest the +/- 5kb gene is obtained in order to blast. The above screen displays the genomic positions of the gene and the fasta sequence retrieved to blast the sequence. This pages also provides an option to blast the sequence against itself.



THE UNIVERSITY OF IOWA

SPeeDD System

SPeeDD - System to Prioritize Deletions or Duplication candidates. This high-throughput computational system is designed and implemented to identify genomic regions with possible deletion or duplication mutations and make the information available to genetic and biomedical researchers.

SPeeDD system provides files with details of similar sequences that are likely to recombine based on the parameters. SPeeDD utilizes machine learning methods to perform analysis and predict candidates. All the analysis here is done in

SPeeDD

****

Gene Name BBS4

Accession NM_033028

Length Greater than 30

Percent Identity >80

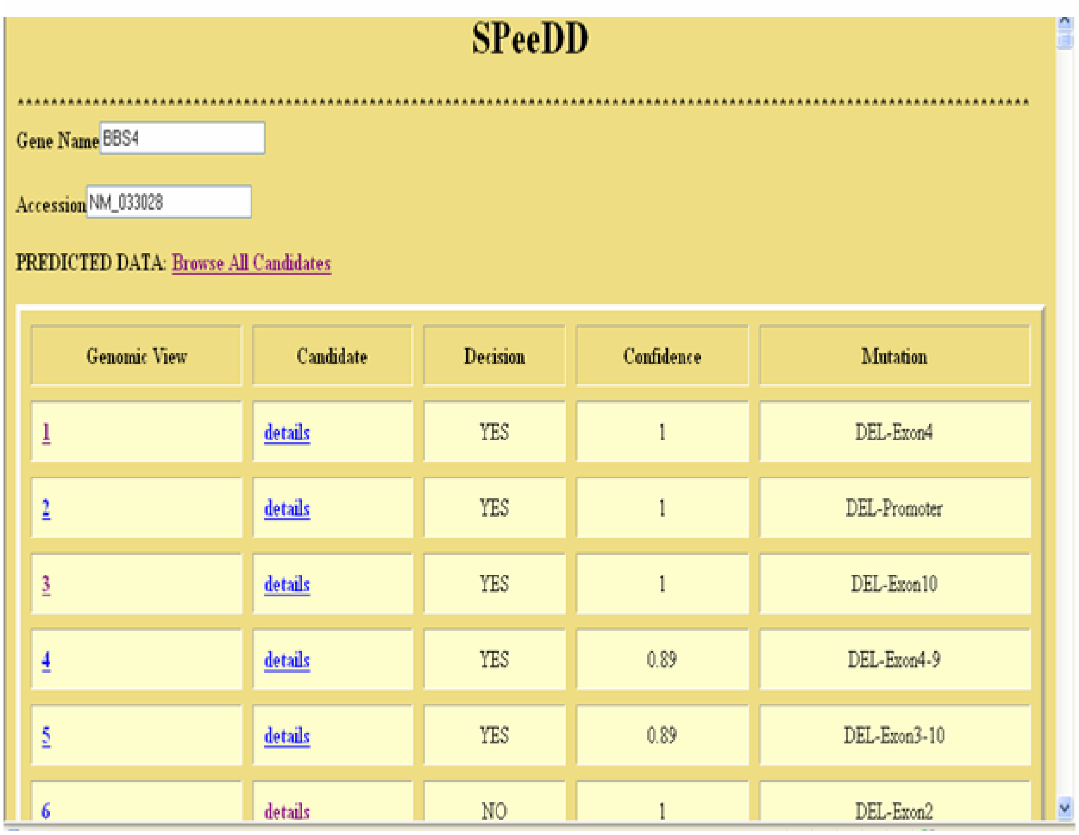
Distance betweeen Sequences <50kb

About this page

This page contains the gene,reference sequence and required parameters and their user defined cutoffs to determine the candidates for unequal homologous recombination We encourage you to explore our tool

GeneName: Hugo gene symbol **Accession**: A unique number or combination of letters and numbers assigned to each record or gene in human database **Length**: Maximum Length of the similar sequences that are likely to recombine **Percent Identity**: The sequence identity between the similar sequences that are likely to recombine **Distance between Sequences**: Maximum distance to be considered between the similar sequences that are likely to recombine

Figure 28 Screen shot of SPeeDD system. This page consists of various options to filter the BL2SEQ output. The default parameters are shown in the figure. Drop down boxes are provided to change the length, percent identity and distance options for the pairs of homologous sub-sequences.



SPeeDD

****

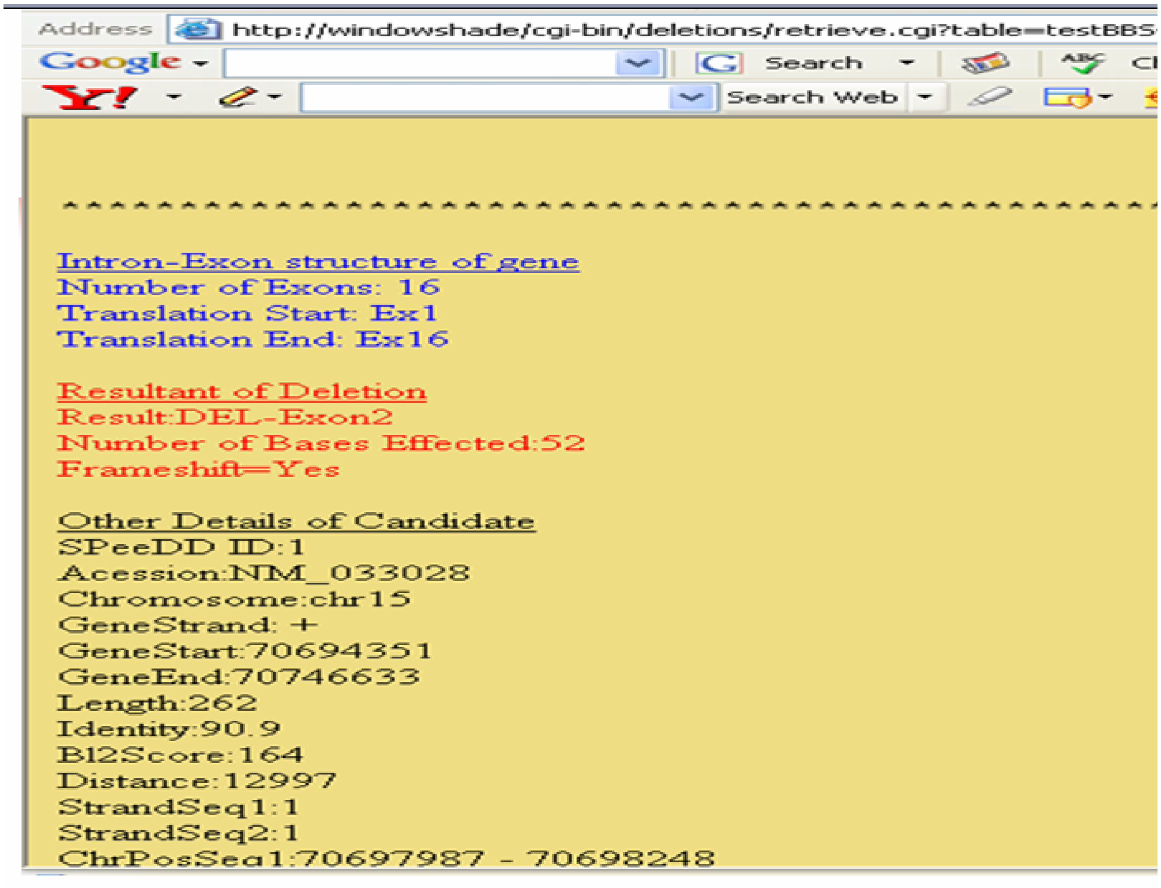
Gene Name BBS4

Accession NM_033028

PREDICTED DATA: Browse All Candidates

Genomic View	Candidate	Decision	Confidence	Mutation
1	details	YES	1	DEL-Exon4
2	details	YES	1	DEL-Promoter
3	details	YES	1	DEL-Exon10
4	details	YES	0.89	DEL-Exon4-9
5	details	YES	0.89	DEL-Exon3-10
6	details	NO	1	DEL-Exon2

Figure 29 Screen shot of SPeeDD system. Data mining is performed and it displays the list of candidates with a outcome (Decision) yes or no and a confidence value associated to it. It displays the outcome of the deletion or duplication in terms of exon or exons.


Address
http://windowshade/cgi-bin/deletions/retrieve.cgi?table=testBBS
Google
Search
Search Web

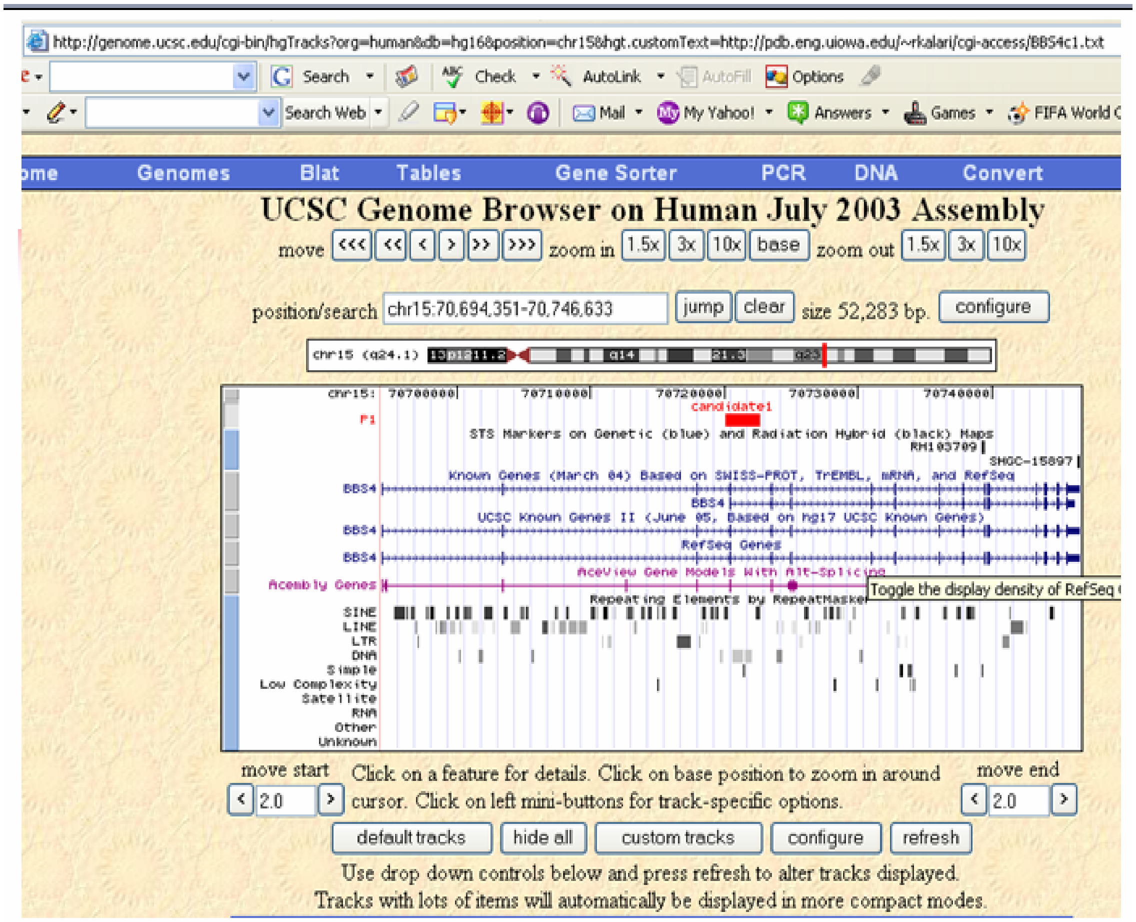
****

Intron-Exon structure of gene
Number of Exons: 16
Translation Start: Ex1
Translation End: Ex16

Resultant of Deletion
Result:DEL-Exon2
Number of Bases Effected:52
Frameshift=Yes

Other Details of Candidate
SPeeDD ID:1
Acession:NM_033028
Chromosome:chr15
GeneStrand: +
GeneStart:70694351
GeneEnd:70746633
Length:262
Identity:90.9
Bl2Score:164
Distance:12997
StrandSeq1:1
StrandSeq2:1
ChrPosSeq1:70697987 - 70698248

Figure 30 Screen shot of SPeeDD system. Candidates link of Figure 29 displays this figure 30. It consists of more details of mutation description such as the resultant of deletion.



http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&db=hg16&position=chr15&hgt.customText=http://pdb.eng.uiowa.edu/~rkalari/cgi-access/BBS4c1.txt

UCSC Genome Browser on Human July 2003 Assembly

position/search chr15:70,694,351-70,746,633 size 52,283 bp.

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

Use drop down controls below and press refresh to alter tracks displayed.

Tracks with lots of items will automatically be displayed in more compact modes.

Figure 31 Screen shot of SPeeDD system. Genomic view of a candidate with decision yes in Figure 29 leads to a custom track of UCSC browser. The red block above the STS markers in the figure indicates the genomic starting ending positions of the block that may potentially be deleted

Novel applications

The theoretic utility of the SPeeDD system has been demonstrated in the Validation section above. Additional opportunities to validate the system's prioritizations are available as new IDDs are reported. To date, one novel duplication that was never included in any training set has been validated with the SPeeDD system.

BRCA1 mutation identification

In July of 2006, a novel duplication within the BRCA1 gene was identified in a population of Chinese ethnicity (Yap et al., 2006). This mutation is a duplication of 8.4 kb, resulting in an additional copy of exon 13. Because the exon is not a multiple of three bp long (127 bp) and is in the middle of the coding region, including two copies would induce a frame shift mutation. Therefore, a duplication of exon 13 would be expected to adversely affect the protein. Because this mutation has been only recently reported, and had never been included in our training set, it is a perfect candidate to validate the performance of the SPeeDD system.

The BRCA1 gene contains several (how many) repetitive elements all of which could potentially recombine resulting in the deletion or duplication of one or more exons. SPeeDD collected all homologous sequences that could recombine with the feature data and performed data mining to predict the possible outcome of the candidates. In particular, SPeeDD identified 44 exon spanning IDD candidates.

The system analyzed and prioritized each candidate IDD. Based on the patterns of the data the model predicted 44 candidates to delete or duplicate exons and 5601 candidates that may not recombine and delete or duplicate exons. SPeeDD was successfully able to identify the novel BRCA1 duplication mutation as a predicted, highly prioritized candidate IDD. This illustrates that SPeeDD, and the method underlying the system can be used to identify new deletions or duplications.

Summary

The goal of the SPeeDD system is to enhance a researcher’s ability to identify and validate intragene deletions and duplications, an under-represented class of mutations capable of causing human disease. The strategy through which this is accomplished is to increase the efficiency of the mutation discovery process through the use of a computational system. The SPeeDD system utilizes information from a set of previously identified and validated intragene deletions and duplications to train a computational system which can then identify and prioritize candidate intragene deletions and duplications.

Based upon both computational and biological validation, SPeeDD successfully identifies those candidates that are most likely to be involved in an intragene deletions or duplication. The growing availability of both sequence and functional annotation has greatly improved the quality of computational predictions. Due to the abundance of repeats within a gene it is laborious to look for deletion or duplication candidates comprehensively across a gene. SPeeDD provides a quick, unbiased method to rank candidate deletion or duplication regions. This allows investigators to focus their research on those candidate IDDs that are most likely to be deleted, thereby reducing the labor and associated costs of the biological assays and accelerating the process of mutation discovery. Our implementation of SPeeDD is readily available on the web at http://public.eng.uiowa.edu/SPeeDD.

REFERENCES

Abeysinghe, S. S., Chuzhanova, N., Krawczak, M., Ball, E. V. and Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat 22, 229-244.

Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. Nat Rev Genet 3, 370-379.

Ben-Hur, A. and Brutlag, D. (2003). Remote homology detection: a motif based approach. Bioinformatics 19 Suppl 1, i26-33.

Bentz, M., Plesch, A., Stilgenbauer, S., Dohner, H. and Lichter, P. (1998). Minimal sizes of deletions detected by comparative genomic hybridization. Genes Chromosomes Cancer 21, 172-175.

Bhangale, T. R., Rieder, M. J., Livingston, R. J. and Nickerson, D. A. (2005). Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14, 59-69.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev 16, 6-21.

Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., Curwen, V., Cutts, T., Down, T., Durbin, R., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Kahari, A., Jekosch, K., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., Mcvicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, C., Clamp, M. and Hubbard, T. (2004). Ensembl 2004. Nucleic Acids Res 32, D468-470.

Braun, T. A., Shankar, S. P., Davis, S., O'leary, B., Scheetz, T. E., Clark, A. F., Sheffield, V. C., Casavant, T. L. and Stone, E. M. (2006). Prioritizing regions of candidate genes for efficient mutation screening. Hum Mutat 27, 195-200.

Brooks, E. M., Branda, R. F., Nicklas, J. A. and O'neill, J. P. (2001). Molecular description of three macro-deletions and an Alu-Alu recombination-mediated duplication in the HPRT gene in four patients with Lesch-Nyhan disease. Mutat Res 476, 43-54.

Callinan, P.A. and Batzer, M.A. (2006) *Retrotransposable elements and human disease.* Genome Dynamics **1**,104-115

Chance, P. F. and Lupski, J. R. (1994). Inherited neuropathies: Charcot-Marie-Tooth disease and related disorders. Baillieres Clin Neurol 3, 373-385.

Chen, J. M., Chuzhanova, N., Stenson, P. D., Ferec, C. and Cooper, D. N. (2005). Meta-analysis of gross insertions causing human genetic disease: Novel mutational mechanisms and the role of replication slippage. Hum Mutat 25, 318.

Chuzhanova, N. A., Anassis, E. J., Ball, E. V., Krawczak, M. and Cooper, D. N. (2003). Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 21, 28-44.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. and Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38, 75-81.

Cooper, D. N. and Krawczak, M. (1996). Human Gene Mutation Database. Hum Genet 98, 629.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. Nat Genet 29, 229-232.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R. and Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature 418, 544-548.

Deininger, P. L. and Batzer, M. A. (1999). Alu repeats and human disease. Mol Genet Metab 67, 183-193.

Deininger, P. L., Moran, J. V., Batzer, M. A. and Kazazian, H. H., Jr. (2003). Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13, 651-658.

Deloukas, P. and Bentley, D. (2004). The HapMap project and its application to genetic studies of drug response. Pharmacogenomics J 4, 88-90.

Elliott, B. and Jasin, M. (2002). Double-strand breaks and translocations in cancer. Cell Mol Life Sci 59, 373-385.

Feil, R. and Khosla, S. (1999). Genomic imprinting in mammals: an interplay between chromatin and DNA methylation? Trends Genet 15, 431-435.

Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics 20, 2479-2481.

Fullerton, S. M., Bernardo Carvalho, A. and Clark, A. G. (2001). Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18, 1139-1142.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., Defelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. Science 296, 2225-2229.

Gerondakis, S., Cory, S. and Adams, J. M. (1984). Translocation of the myc cellular oncogene to the immunoglobulin heavy chain locus in murine plasmacytomas is an imprecise reciprocal exchange. Cell 36, 973-982.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y. and Et Al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306, 234-238.

Halperin, E. and Eskin, E. (2004). Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics 20, 1842-1849.

Hedrick, S. M., Cohen, D. I., Nielsen, E. A. and Davis, M. M. (1984). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature 308, 149-153.

Hey J (2004) What's So Hot about Recombination Hotspots? PLoS Biol 2(6): e190

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. and Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science 307, 1072-1079.

Honjo, T. (1983). Immunoglobulin genes. Annu Rev Immunol 1, 499-528.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat Genet 36, 949-951.

Janssen, J. W., Ludwig, W. D., Sterry, W. and Bartram, C. R. (1993). SIL-TAL1 deletion in T-cell acute lymphoblastic leukemia. Leukemia 7, 1204-1210.

Jones, P. A. and Laird, P. W. (1999). Cancer epigenetics comes of age. Nat Genet 21, 163-167.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32, D493-496.

Krawczak, M., Ball, E. V. and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63, 474-488.

Krawczak, M., Ball, E. V., Fenton, I., Stenson, P. D., Abeysinghe, S., Thomas, N. and Cooper, D. N. (2000). Human gene mutation database-a biomedical information and research resource. Hum Mutat 15, 45-51.

Krawczak, M. and Cooper, D. N. (1991). Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum Genet 86, 425-441.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., W_terston, R. H., Wilson, R. K., Hillier, L. W., Mcpherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic Model Trees. *Proceedings of the 16th European Conference on Machine learning.*

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Proceedings of the 16th European Conference on Ma-*

Lehrman, M. A., Goldstein, J. L., Russell, D. W. and Brown, M. S. (1987). Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholesterolemia. Cell 48, 827-835.

Lehrman, M. A., Schneider, W. J., Sudhof, T. C., Brown, M. S., Goldstein, J. L. and Russell, D. W. (1985). Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. Science 227, 140-146.

Liskay, R. M., Stachelek, J. L. and Letsou, A. (1984). Homologous recombination between repeated chromosomal sequences in mouse cells. Cold Spring Harb Symp Quant Biol 49, 183-189.

Lupski, J. R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14, 417-422.

Lupski, J. R. and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 1, e49.

Macilwain, C. (2000). World leaders heap praise on human genome landmark. Nature 405, 983-984.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'hoy, K. and Et Al. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. Science 255, 1253-1255.

Malissen, M., Minard, K., Mjolsness, S., Kronenberg, M., Goverman, J., Hunkapiller, T., Prystowsky, M. B., Yoshikai, Y., Fitch, F., Mak, T. W. and Et Al. (1984). Mouse T cell antigen receptor: structure and organization of constant and joining gene segments encoding the beta polypeptide. Cell 37, 1101-1110.

Mitelman, F. (2000). Recurrent chromosome aberrations in cancer. Mutat Res 462, 247-253.

Monnat, R. J., Jr., Hackmann, A. F. and Chiaverotti, T. A. (1992). Nucleotide sequence analysis of human hypoxanthine phosphoribosyltransferase (HPRT) gene deletions. Genomics 13, 777-787.

Moynahan, M. E., Chiu, J. W., Koller, B. H. and Jasin, M. (1999). Brca1 controls homology-directed DNA repair. Mol Cell 4, 511-518.

Moynahan, M. E., Pierce, A. J. and Jasin, M. (2001). BRCA2 is required for homology-directed repair of chromosomal breaks. Mol Cell 7, 263-272.

Mykytyn, K., Braun, T., Carmi, R., Haider, N. B., Searby, C. C., Shastri, M., Beck, G., Wright, A. F., Iannaccone, A., Elbedour, K., Riise, R., Baldi, A., Raas-Rothschild, A., Gorman, S. W., Duhl, D. M., Jacobson, S. G., Casavant, T., Stone, E. M. and Sheffield, V. C. (2001). Identification of the gene that, when mutated, causes the human obesity syndrome BBS4. Nat Genet 28, 188-191.

Newman, T. L., Rieder, M. J., Morrison, V. A., Sharp, A. J., Smith, J. D., Sprague, L. J., Kaul, R., Carlson, C. S., Olson, M. V., Nickerson, D. A. and Eichler, E. E. (2006). High-throughput genotyping of intermediate-size structural variation. Hum Mol Genet 15, 1159-1167.

Nishimura, D. Y., Swiderski, R. E., Searby, C. C., Berg, E. M., Ferguson, A. L., Hennekam, R., Merin, S., Weleber, R. G., Biesecker, L. G., Stone, E. M. and Sheffield, V. C. (2005). Comparative genomics and gene expression analysis identifies BBS9, a new Bardet-Biedl syndrome gene. Am J Hum Genet 77, 1021-1033.

Oberle, I., Vincent, A., Abbadi, N., Rousseau, F., Hupkes, P. E., Hors-Cayla, M. C., Gilgenkrantz, S., Oostra, B. A. and Mandel, J. L. (1991). New polymorphism and a new chromosome breakpoint establish the physical and genetic mapping of DXS369 in the DXS98-FRAXA interval. Am J Med Genet 38, 336-342.

Ohler, U., Liao, G. C., Niemann, _. and Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. Genome Biol 3, RESEARCH0087.

Ohno, S. (1972). So much "junk" DNA in our genome. Brookhaven Symp Biol 23, 366-370.

Osterholm, A. M., Bastlova, T., Meijer, A., Podlutsky, A., Zanesi, N. and Hou, S. M. (1996). Sequence analysis of deletion mutations at the HPRT locus of human T-lymphocytes: association of a palindromic structure with a breakpoint cluster in exon 2. Mutagenesis 11, 511-517.

Panning, B. and Jaenisch, R. (1998). RNA and the epigenetic regulation of X chromosome inactivation. Cell 93, 305-308.

Patel, P. I. and Lupski, J. R. (1994). Charcot-Marie-Tooth disease: a new paradigm for the mechanism of inherited disease. Trends Genet 10, 128-133.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., Mcdonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294, 1719-1723.

Pavlidis, P., Furey, T. S., Liberto, M., Haussler, D. and Grundy, W. N. (2001). Promoter region-based classification of genes. Pac Symp Biocomput, 151-163.

Pentao, L., Wise, C. A., Chinault, A. C., Patel, P. I. and Lupski, J. R. (1992). Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet 2, 292-300.

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F. S., Camisa, A. L., Pazorov, V., Scott, K. E., Carey, B. J., Faith, J., Katari, G., Bhatti, H. A., Cyr, J. M., Derohannessian, V., Elosua, C., Forman, A. M., Grecco, N. M., Hock, C. R., Kuebler, J. M., Lathrop, J. A., Mockler, M. A., Nachtman, E. P., Restine, S. L., Varde, S. A., Hozza, M. J., Gelfand, C. A., Broxholme, J., Abecasis, G. R., Boyce-Jacino, M. T. and Cardon, L. R. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33, 382-387.

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. Nat Genet 37 Suppl, S11-17.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W. and Albertson, D. G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20, 207-211.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M. and Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 109, 678-680.

Puget, N., Stoppa-Lyonnet, D., Sinilnikova, O. M., Pages, S., Lynch, H. T., Lenoir, G. M. and Mazoyer, S. (1999). Screening for germ-line rearrangements and regulatory mutations in BRCA1 led to the identification of four new deletions. Cancer Res 59, 455-461.

Purandare, S. M. and Patel, P. I. (1997). Recombination hot spots and human disease. Genome Res 7, 773-786.

Rabbitts, T. H. (1994). Chromosomal translocations in human cancer. Nature 372, 143-149.

Reese, M. G. (2001). Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem 26, 51-56.

Reiter, L. T., Murakami, T., Koeuth, T., Pentao, L., Muzny, D. M., Gibbs, R. A. and Lupski, J. R. (1996). A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. Nat Genet 12, 288-297.

Rudiger, N. S., Gregersen, N. and Kielland-Brandt, M. C. (1995). One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. Nucleic Acids Res 23, 256-260.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich,_H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487-491.

Scherer, S. and Davis, R. W. (1980). Recombination of dispersed repeated DNA sequences in yeast. Science 209, 1380-1384.

Schouten, J. P., Mcelgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F. and Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30, e57.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. Science 305, 525-528.

Sen, S. K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P. A., Dyer, M., Cordaux, R., Liang, P. and Batzer, M. A. (2006). Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet 79, 41-53.

Sharan, R. and Myers, E. W. (2005). A motif-based framework for recognizing sequence families. Bioinformatics 21 Suppl 1, i387-393.

Sharp, A. J., Cheng, Z. and Eichler, E. E. (2006). Structural Variation of the Human Genome. Annu Rev Genomics Hum Genet.

Sharp, A. J., Locke, D. P., Mcgrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D. and Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77, 78-88.

Shaw, C. J., Bi, W. and Lupski, J. R. (2002). Genetic proof of unequal meiotic crossovers in reciprocal deletion and duplication of 17p11.2. Am J Hum Genet 71, 1072-1081.

Shaw, C. J. and Lupski, J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet 13 Spec No 1, R57-64.

Snouwaert, J. N., Gowen, L. C., Latour, A. M., Mohn, A. R., Xiao, A., Dibiase, L. and Koller, B. H. (1999). BRCA1 deficient embryonic stem cells display a decreased homologous recombination frequency and an increased frequency of non-homologous recombination that is corrected by expression of a brca1 transgene. Oncogene 18, 7900-7907.

Southern, E. M. (1992). Detection of specific sequences among DNA fragments separated by gel electrophoresis. 1975. Biotechnology 24, 122-139.

Stankiewicz, P. and Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. Trends Genet 18, 74-82.

Stark, G. R. and Wahl, G. M. (1984). Gene amplification. Annu Rev Biochem 53, 447-491.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., Abeysinghe, S., Krawczak, M. and Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21, 577-581.

Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99, 3740-3745.

Tatusova, T. A. and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174, 247-250.

Wain, H. M., Lush, M. J., Ducluzeau, F., Khodiyar, V. K. and Povey, S. (2004). Genew: the Human Gene Nomenclature Database, 2004 updates. Nucleic Acids Res 32, D255-257.

Wall, J. D. and Pritchard, J. K. (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73, 502-515.

Wasmuth, J. J. and Vock Hall, L. (1984). Genetic demonstration of mitotic recombination in cultured Chinese hamster cell hybrids. Cell 36, 697-707.

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C. and Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. Am J Hum Genet 71, 854-862.

Yabe, T., Kawamura, S., Sato, M., Kashiwase, K., Tanaka, H., Ishikawa, Y., Asao, Y., Oyama, J., Tsuruta, K., Tokunaga, K., Tadokoro, K. and Juji, T. (2002). A subject with a novel type I bare lymphocyte syndrome has tapasin deficiency due to deletion of 4 exons by Alu-med_ated recombination. Blood 100, 1496-1498.

Yamey, G. (2000). Scientists unveil first draft of human genome. Bmj 321, 7.

Yap, K. P., Ang, P., Lim, I. H., Ho, G. H. and Lee, A. S. (2006). Detection of a novel Alu-mediated BRCA1 exon 13 duplication in Chinese breast cancer patients and implications for genetic testing. Clin Genet 70, 80-82.

Yoder, J. A., Walsh, C. P. and Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13, 335-340.

Books and conference papers

Bishop, C.M. (1995) Neural Networks for Pattern Recognition, Oxford: Oxford University Press. ISBN 0-19-853849-9 (hardback) or ISBN 0-19-853864-2.

Breiman, L.,Friedman, R.A.,Olshen and Stone,C.J.(1984) "Classification and regression trees". Wadsworth.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning. 20, 273-297.

Dasarathy,B.V. (1991) Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, ISBN 0-8186-8930-7

Duda, R.O., Hart, P.E., Stork, D.G. (2001) Pattern classification (2nd edition), Wiley, ISBN 0471056693

Gurney, K. (1997) An Introduction to Neural Networks London: Routledge. ISBN 1-85728-673-1 (hardback) or ISBN 1-85728-503-4.

Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, Prentice Hall, ISBN 0-13-273350-1

Hertz, J., Palmer, R.G., Krogh. A.S. (1990) Introduction to the theory of neural computation, Perseus Books. ISBN 0201515601

Hjorth, J. (1994). Computer Intensive Statistical Methods Validation, Model Selection, and Bootstrap. Chapman & Hall, London.

Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.

Lawrence, Jeanette (1994) Introduction to Neural Networks, California Scientific Software Press. ISBN 1-883157-00-5

Plutowski, M., Sakata, S. and White, H. (1994). Cross-validation estimates integrated mean square error. Advances in Neural Information Processing Systems, 6.

Ross Quinlan.(1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, ca,22.

Shapiro,J.A (1983) Mobile genetic elements (Academic New York).

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap, Springer, New York.

Stahl (1979), Genetic Recombination : Thinking about it in phage and fungi (Freeman, San Francisco).

Swets,J.,Dawes,R.and Monaban,J.(2000) "Better decisions through science"Scientific American, October,82-87.

Weiss, G.M. and Provost, F. (2001) The effect of class distribution on classifier learning. Technical Report ML-TR 43, Department of Computer Science, Rutgers University.

Weiss, S.M. and Kulikowski, C.A.(1994).Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann Publishers Inc.