RESEARCH STRATEGY

A. SIGNIFICANCE

In this proposal we plan to develop novel computational techniques that will identify genotypes and structural variations (SVs) responsible for various diseases. Traditional techniques employ all of the SNPs and SVs in building a model correlating SNPs and SVs with various phenotypes. Given that there are millions of SNPs and many diseases are polygenic, these models are impractical since the time needed is humongous. In this proposal we propose a novel way of reducing the number of attributes in model building. The idea is to employ only those SNPs that occur in minimotifs (we refer to these as minimotif SNPs) and relevant SVs. The reasons for picking these are: Minimotifs are a principal means of connecting proteins in the protein-protein interaction network, because the majority are under strong negative selection, and because they are mutated in a number of diseases [44]. SVs have been found to be associated with many diseases [154-156]. Minimotifs are short contiguous peptide sequences generally confined to a single secondary structure element and less than 15 residues in length. Examples are the Pro-x-x-Pro minimotif for binding SH3 domains (binding), the Asn-x-Ser/Thr N-glycosylation motif (covalent modification of minimotif), and the Ser-Lys-Leu peroxisomal targeting motif (protein trafficking). While several other groups have built focused databases on select groups of minimotifs (e.g., Merops and CutDB contain protease sites [40] [99]), our group has built Minimotif Miner (MnM), a comprehensive database with the goal of including all minimotifs that are published and supported by experimental data. MnM now contains ~300,000 minimotifs. Existing approaches to GWAS can only identify single variants (or attributes) associated with diseases. However, the real cause for a disease could be a combination of (10s or even 100s) of variants. In this project we address this extremely challenging problem.

The proposed project is significant for three primary reasons:

(1) *A new paradigm in Genome-Wide Association Study (GWAS).*

An important milestone has nearly been reached – cost-effective human genome sequencing! As the costs to sequence an individual human exome plummet, this technology will undoubtedly have a huge influence on health care and personalized medicine in the near future. One major barrier is yet to be addressed – How are we going to analyze and use the data? While recent progress has yielded computational applications to generate and compare exomes and predicted proteomes from the DNA sequence data [34] [122] [120] [28] [21], the ability to connect patient genomes with changes in protein function, design clinical treatments, and predict outcomes still presents a major challenge. In this project we propose a novel way of reducing the number of attributes necessary for the identification of genotypes corresponding to various diseases. Thus the theme of this project is in the heart of translational science and personalized medicine.

The minimotif and structural variant (SV) analysis we do in Aim 1 enables us to reduce the number of SNPs and SVs needed for the genotype-phenotype correlational analysis. The minimotif field has significantly expanded since our original publication of MnM [6]. Several less-comprehensive minimotif databases and *de novo* minimotif prediction tools have emerged [6] [98] [113] [31] [117] [120] [15] [22] [23] [24] [70] [37] [77] [71] [103] [98]. Given the comprehensive database in MnM, we are extremely qualified to conduct this analysis. Existing approaches for GWAS have several limitations in effectiveness as well as computational cost. Existing approaches in GWAS focus on identifying SNPs and their role in variations in phenotypes. Specifically, the search space consists of all the (millions of) SNPs. For example, one of the current approaches is to detect interactive SNP × SNP effects. This problem is known as the k-locus association problem (see e.g., [2]). Unfortunately, the best-known algorithms for solving this problem take time that is $\Omega(n^k)$, where n is the number of SNPs, severely limiting their applicability in practice (especially when n is large). Given that there are more than 4 million common SNPs in humans, identifying a small subset of these (of size 5 or 6) that is responsible for the phenotype is a big challenge. The idea of narrowing down this search to within the minimotifs is thus very attractive. We believe that the proposed technique of studying minimotif variability will be more effective in identifying variations in functions and phenotypes and computationally efficient. SVs also play an important role in genomic diseases such as cancer [154]. In this Aim we will also study the combined variability of minimotifs and structures in individuals and populations. Information on structural variations will be obtained from the literature using novel text mining. No such work has been conducted in GWAS.

(2) Determining if there is a global role of minimotifs and SVs in disease and human variation.

In Aim 1 we will examine how minimotifs are changed by polymorphisms and disease-associated mutations. This will provide researchers an important tool for both studying the causes of disease, helping to analyze novel variants, and in the near future analyzing patient genomes/exomes. We will also study SVs in this Aim. The tool will also allow us to globally address how functional minimotifs change in individuals, as in our recent work [62]. We hypothesize that there will be many changes of minimotifs associated with disease. This hypothesis is supported by the observations that there are a number of known cases where minimotifs are mutated in disease, that a single missense mutation can introduce a new minimotif or disable an existing minimotif, and that ~85% of disease causing mutations are thought to exist in protein coding regions [17]. Furthermore, analysis of select posttranslational modification minimotifs suggests that 5% of nonsynonymous SNPs are in regions of posttranslational modification [102] [59] [62]. Our recent work with Drs. Edwin Wang and Shawn Li (University of Western Ontario) showed that acquiring new minimotifs is a central determinant in the evolutionary expansion of tyrosine kinase signaling pathways, networks, and intracellular communication in multicellular organisms [57]; other studies have also supported a role for minimotifs in evolution [62] [15] [57] [16] [22] [71] [124]. There are ~20,000 minimotifs with known polymorphic loci and newly identified rare variants are likely to damage minimotifs [62]. A role for minimotifs in disease was first summarized in our 2007 review of the literature [44], but we still do not know to what extent minimotifs influence disease, which will be studied as part of Aim 1. From an evolutionary perspective, our analysis will help us understand key branches in phylogenetic trees where different types of minimotifs first originate.

(3) *Validation of our approach using text mining and cross validation*.

Another novel aspect of our project is the employment of text mining and cross validation to validate our computational approaches. In general, there are two ways to validate computational predictions, namely, biological experiments and confirmation with known experimental results. We feel that the biological study itself is a huge project and that the computational and biological studies cannot be combined into a single R01 project. Thus, we have chosen the second approach. We will apply our algorithms to identify the most relevant genotypes and structural variations (SVs) corresponding to various diseases. Followed by this, we will use text mining to collect all the experimental results (especially genotypes and SVs) reported for these diseases. Finally, we will compare our computational predictions with the experimental results reported in the literature. Once we complete this project, we plan to apply for a NIH grant that focuses on validating the computational predictions biologically.

B. INNOVATION

There are three innovative aspects of this proposal:

(1) *Narrowing the search space in GWAS.*

Existing approaches look for genomic variations by focusing on all of the SNPs, for example. This makes the search problem a computationally intensive task. The *k*-locus association problem introduced in this context takes time that is exponential in *k*. When *k*=2, most of the existing algorithms take quadratic time [2]. If the number of SNPs is 4 millions, then the number of operations needed for analysis will be 16 x 10^{12}. If it takes one second per billion operations, then this time will be 4.44 hours. If *k*=3, then this time will be more than 2000 years! If we can narrow down the search to 10,000 variant non synonymous minimotifs with an SNP [62], then these two times will be 0.1 second and 16.67 minutes, respectively! Adapting MnM to identify functional minimotifs that are introduced or eliminated by disease and variation is an innovative idea. **No other groups have a comprehensive minimotif database, thus we are the only group that is well positioned to build this important tool for analyzing exomes and genomes for experimentally identified minimotif functions.**

The improved MnM would work synergistically with the existing open source genome/exome analysis tools. There are three general types of existing tools: (1) Databases that contain SNP and systems that consolidate and federate SNP data that we will analyze in Aim 1 [115] [123] [14] [111] [104]; (2) Several computational tools that **predict** the effect of a nonsynonymous SNP on protein function. The majority of these tools predict functions based upon the function of the gene, the molecular pathways of the gene, the structure or biophysical characteristics of the protein, pre-mRNA splice sites, phenotypes, etc. [36] [115] [123] [14] [104] [20] [54] [106] [60] [58] [32] [27] [112] [118] [126]. For example, Polyphen, SIFT and SNPeffect, provide predictions for effects of SNP on protein functions and structures [104] [32] [26] [72]. Predictions have limitations, as evidenced by a benchmark study of four different SNP-function predictors where only 11% of the predictions were consistent among the applications [41]. A complementary approach that we plan to implement is to determine if an SNP affects an experimentally verified minimotif, and thus a known function of the protein; and (3) Applications that analyze SNPs in genomes/exomes [123] [56] [58] [32] [47]. The MnM tool would be unique by analyzing genomes/exomes, using a comprehensive minimotif/PTM database, and directly linking results with MnM to enable interpretation in the context of structure, amino acid conservation, and other functions of the protein. Our proposed tool would be similar to identifying splice site changes in SNPNexus in that it assesses mutation of a nucleotide that has a known functional association [14]. We will develop collaborations with some of these groups working on exomes and SNPs during the course of this project. We expect to get very good accuracies by focusing on only minimotif SNPs. We propose to employ (a minimal set of) SVs that together with these SNPs will increase the accuracy further. A large number of research teams are working on different types of SVs. We will utilize the results from these efforts by obtaining SV data from the literature via text mining.

(2) Analyzing >1,000,000 available exomes/genomes to generate a broad view of minimotif frequencies. This is indeed a big data analytics problem. We propose to develop efficient parallel and out-of-core algorithms for this analysis. Big data analytics demands huge computational times. One of the main reasons for this has been that core memories of computers are not large enough to hold all the data to be analyzed, and hence most of the data have to be stored in secondary storages (SSs) such as solid state drives (SSDs) and (rotating) disks. Data access times from SSs are several orders of magnitude more than from core memories. Tremendous speedups can be obtained by minimizing the number of data accesses from SSs. Also, although there has been much recent research in the development of multicore and GPU algorithms for biological problems, for many of the problems only sequential in-core algorithms are known. Algorithms that explicitly minimize the number of data I/Os (input/output) from/to slower storages (such as SSs) in any memory hierarchy are called *out-of-core algorithms*. The PI has extensive experience in the development of parallel and out-of-core algorithms. As evidence, in September 2014, NSF has funded the PI's project that focuses on parallel and out-of-core algorithms for biological big data.

(3) *Developing novel algorithms for text mining and the k-locus association problem.* We propose to employ text mining techniques to validate our computational approaches (Aim 3). In our prior work we have employed text mining to populate the MnM database. In this project we will start from these algorithms and develop variants suitable for the identification of experimentally validated results for various diseases and phenotypes. Text mining will also be used to retrieve SV data from the literature. We will also develop efficient algorithms for the *k*-locus association problem for the identification of phenotype-genotype correlations in Aim 2. Most of the existing algorithms find only single variants (i.e., *k*=1) associated with diseases. We are addressing the much more difficult and relevant problem of identifying combinations of variants (i.e., the case of *k*>1).

C. APPROACH

C.1. APPROACH – PROGRESS REPORT (Start date - 10/01/10; End date – 9/30/15)

Progress from the previous award: Before we address what we want to do in the renewal, we first briefly summarize what we have accomplished in the previous award as requested in the application instructions. There were three major aims for the previous period of support: 1) to develop efficient algorithms for three versions of motif search, namely, Simple Motif Search (SMS), Planted Motif Search (PMS), and Edit-distance-based Motif Search (EMS); 2) to evaluate these algorithms; and 3) to develop parallel algorithms and a web site. We have achieved these goals with a number of outstanding results. Thus far, we have the following publications: 50 published, 1 accepted, and 2 submitted. MnM has had a significant impact on protein research. It is the largest minimotif database. The MnM website has had over 108,000 external searches and more than 225 citations since its initial publication; accelerating increase in its impact on discovery [6] [98] [113]. Several of our published papers on this project have been in relatively high impact journals such as Science Signaling, Nature Methods, Scientific Reports, Bioinformatics, Nucleic Acids Research, and the Nucleic Acids Research database issue.

In previous Aim 1, we have developed a series of novel algorithms for motif search. In the past 13 years, our Lab has contributed some of the best-known algorithms for motif search. For the problem of planted motif search (or (*l, d*)-motif search) we have developed the following algorithms: PMS5 [29], PMS6 [8] [10], PMS6MC [9], qPMS7 [30], PMS8 [74], and qPMS9 [75]. PMS5 solved the challenging instances (21, 8) and (23, 9) for the first time. One of the time and memory consuming steps in the development of motif search algorithms is that of generating all the neighbors of a given *l*-mer. In [29], we introduced a novel algorithm for doing this. Specifically, we formulated an Integer Linear Program (ILP) to check if a given *l*-mer has any motifs in its neighborhood. We solve many such ILPs in advance and store the results so as to optimize the run time. This idea has been subsequently employed in PMS6 and PMS6MC as well. In PMS6 and PMS6C, the Bloom filter is used in an elegant manner to speedup the preprocessing step of PMS5. For instance, the preprocessing time for PMS6 is 34 times faster than that for PMS5 for the challenging instance (23, 9). In [30], we addressed the quorum version of the planted motif search problem. PMS is known to be NP-hard and hence is difficult. The quorum version is even more difficult. In [30], we presented a novel algorithm for this version that was on an average 5 times faster than the state-of-the-art algorithm at that time. In [74] we showed how we can eliminate the preprocessing step performed in the algorithms PMS5, PMS6, and PMS6C. The basic idea is to derive necessary and sufficient conditions for three *l*-mers to share a common *d*-neighbor. The algorithm in [74], known as PMS8, solved the challenging instances (25, 10) and (26, 11) for the first time. It was also able to solve large instances such as (50, 21). The algorithm qPMS9 presented in [75] is faster for PMS as well as quorum PMS problems than any of the known algorithms as of this writing. It solves, for the first time, the challenging instances (28, 12) and (30, 13). In [85] we offer a general technique that can be used to speedup any motif search algorithm. Motif search algorithms from our Lab have been cited more than 370 times.

With respect to our previous Aim 2, we have evaluated our motif search algorithms on synthetic as well as real data sets. Specifically, we have used data sets mentioned in [12], [119], ChIP_Chip data (http://bdtnp.lbl.gov/Fly-Net/browseChipper.jsp), ChIP-Seq, and ABS datasets. On all of these data sets we have compared our algorithms with other algorithms in the literature with respect to measures such as sensitivity, specificity, etc. A comprehensive report on our experiments can be found in [114]. Experimental comparisons are also presented in most of our papers on motif algorithms (see e.g., [8] [10] [9] [30], [29], [66], [74], [78], and [85]). Our previous Aim 3 pertains to parallel algorithms. We report parallel algorithms for planted motif search in [9], [74], and [75]. For instance, PMS8 [74] has been implemented in C++ using OpenMPI on the Hornet cluster available at UConn. With up to 48 processors, the speedup obtained was close to linear [74]. PMS6MC is a multicore algorithm [9]. For edit-distance based motif search (EMS), we have developed two efficient algorithms in [80] and [78]. In [80], the novel idea was to compute edit distance neighbors using neighbors that are at a distance of one at a time. In [78] we improved the performance of [80] using novel pruning techniques and a variant of the trie data structure. We have developed a web system for motif search at pms.engr.uconn.edu [29]. This system can be used to perform motif search on DNA as well as protein sequences. This is also a repository for the motif search algorithms developed in our lab.

On the MnM front also, we have made numerous enhancements. We have released Version 3.0 in 2011 [113]. We have developed a series of filtering (i.e., scoring) algorithms for improving the predictions of MnM. In [65] we report a filter that is based on genetic interactions. A protein structure based scoring mechanism is offered in [79]. In [69] we combined these filters using linear regression, neural network, and support vector machine models and tested them on a large minimotif dataset with cross validation. MnM 3.0 now has a comprehensive motif score, with a threshold that nearly eliminates all false positive predictions, while maintaining high sensitivity. In addition to the above motif related work, we have also contributed to some fundamental problems in bioinformatics including sequence assembly [49] [50], biological data compression [76] [107] [109] [108], out-of-core computing [48] [67], records linkage [64] [68] [63], sequence alignment [55], etc.

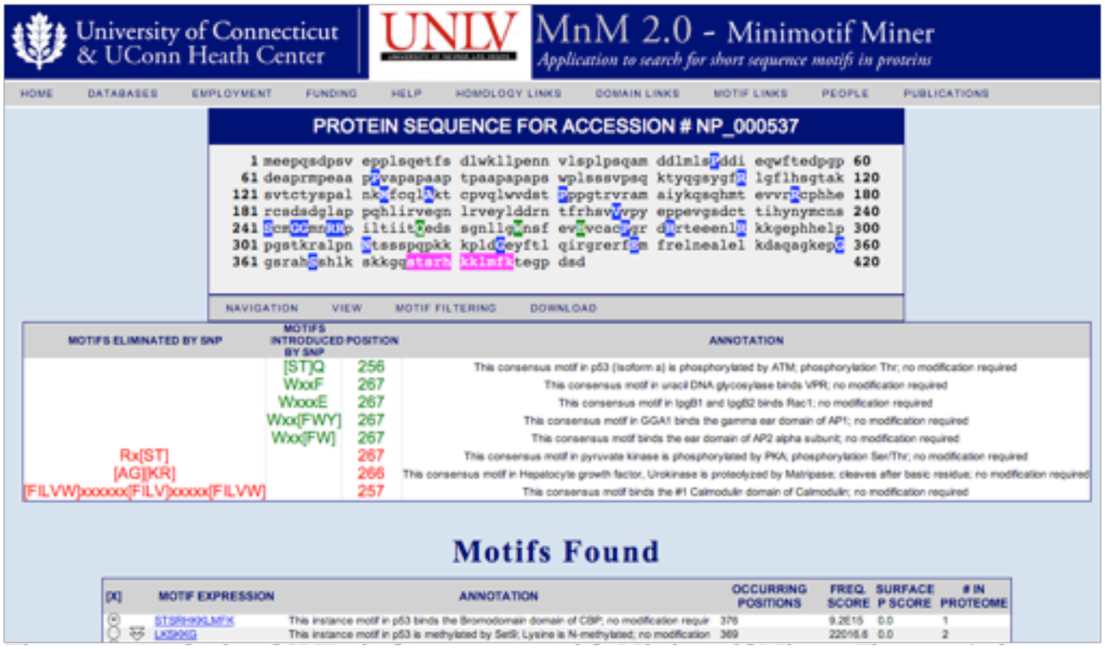
C.2. APPROACH – Specific Aims

C.2.1. Aim 1: To develop computational tools for the identification of minimotif changes and SVs in individuals and populations

Rationale. Companies such as *Navigenics*, *Decode Genetics*, *23andMe*, etc., offer a glimpse of the future of personalized genomics and the beginning of how it will affect personalized medicine. *23andMe is now offering complete exome sequencing*. As the costs of next generation sequencing (NGS) plummet there is a rapidly-increasing publishing stream of complete human exomes [73] [11] [33] [53] [83] [125]. There are more than one million complete human exomes and genomes already published and the 1000 Genomes project has sequenced (and released in Phase III) many more genomes [82]. What has become evident is that even with these tremendous advances in data acquisition, there are significant roadblocks in applying the knowledge of genome sequences to medicine [73]. One of the major challenges in the realization of personalized medicine is connecting genotypes with phenotypes. Only 1% of 682 SNPs in the coding region of disease-associated genes in Craig Venter's genome are well-characterized and the authors suggest "We are only at the beginning of relating genotypes to phenotypes, even for the well-characterized disease genes" [73]. In most cases the direct link between the polymorphism and function of the afflicted protein is not likely to be known. This is a vital problem in connecting basic research data with human phenotypes and health. This is also the problem our group is well positioned to address and the focus of this Aim.

We propose to build a suite of computational and statistical analysis tools that can be used to identify ~240,000 experimentally validated human minimotif functions (600,000 total) that are altered by polymorphisms and mutations. A prototype has already been built [62] and this will be refined to handle big data. We will also be able to extend this to predictions.

Preliminary progress. There are several pieces of preliminary data that support this aim. We hypothesize that mutation of minimotifs is a significant contributor to human phenotypes. We have reviewed the literature and found that many minimotifs are mutated in, or close to mutations that cause disease; pathogens also use minimotifs [44]. Furthermore, our collaborative studies with Drs. Shawn Li and Edwin Wang show that the network expansion observed in evolution of the phosphotyrosine signaling network in complex multicellular organisms is driven much more so by acquiring new targets through minimotifs than by gene duplication [57]. Nevertheless, the global role of minimotifs in disease has never been evaluated at the genomic level.


University of Connecticut & UConn Heath Center
UNLV
MnM 2.0 - Minimotif Miner
Application to search for short sequence motifs in proteins
HOME
DATABASES
EMPLOYMENT
FUNDING
HELP
HOMOLOGY LINKS
DOMAIN LINKS
MOTIF LINKS
PEOPLE
PUBLICATIONS
NAVIGATION
VIEW
MOTIF FILTERING
DOWNLOAD

PROTEIN SEQUENCE FOR ACCESSION # NP_000537

```
  1 meepqsdpsv epplsqetfs dlwkllpenn vlsplpsqam ddlmlspddi eqwftedpgp  60
 61 deaprmpeaa ppvapapaap tpaapapaps wplsssvpsq ktyqgsygfr lgflhsgtak 120
121 svtctyspal nkmfcqlakt cpvqlwvdst pppgtrvram aiykqsqhmt evvrrcphhe 180
181 rcsdsdglap pqhlirvegn lrveylddrn tfrhsvvvpy eppevgsdct tihynymcns 240
241 scmggmnrrp iltiitleds sgnllgrnsf evrvcacpgr drrteeenlr kkgephhelp 300
301 pgstkralpn ntssspqpkk kpldgeyftl qirgrerfem frelnealel kdaqagkepg 360
361 gsrahsshlk skkgqstsrh kklmfktegp dsd                              420
```

MOTIFS ELIMINATED BY SNP	MOTIFS INTRODUCED BY SNP	POSITION	ANNOTATION
	[ST]Q	256	This consensus motif in p53 (isoform a) is phosphorylated by ATM; phosphorylation Thr; no modification required
	WxxF	267	This consensus motif in uracil DNA glycosylase binds VPR; no modification required
	WxxxE	267	This consensus motif in Ipgb1 and Ipgb2 binds Rac1; no modification required
	Wxx[FWY]	267	This consensus motif in GGA1 binds the gamma ear domain of AP1; no modification required
	Wxx[FW]	267	This consensus motif binds the ear domain of AP2 alpha subunit; no modification required
Rx[ST]		267	This consensus motif in pyruvate kinase is phosphorylated by PKA; phosphorylation Ser/Thr; no modification required
[AG][KR]		266	This consensus motif in Hepatocyte growth factor, Urokinase is proteolyzed by Matripase; cleaves after basic residue; no modification required
[FILVW]xxxxxx[FILV]xxxxx[FILVW]		257	This consensus motif binds the #1 Calmodulin domain of Calmodulin; no modification required

Motifs Found

[X]	MOTIF EXPRESSION	ANNOTATION	OCCURRING POSITIONS	FREQ. SCORE	SURFACE P SCORE	# IN PROTEOME
◉	STSRHKKLMFK	This instance motif in p53 binds the Bromodomain domain of CBP; no modification requir	376	9.2E15	0.0	1
○	LKSKKG	This instance motif in p53 is methylated by Set9; Lysine is N-methylated; no modification	369	22016.6	0.0	2

Figure 1: Analysis of SNPs in human p53 with Minimotif Miner. *The top window shows the protein sequence of the query protein (p53 in this case). When "show SNPs" is selected from the "VIEW" menu, all SNPs are highlighted blue. Any set of these SNPs can be selected with a mouse click, which changes the highlighted color to green (e.g. SNPs L257Q, R267W, and R273H were selected here). Initiating the analysis of this set of SNPs from the "VIEW" menu reveals a new table showing five new minimotifs (green font) and three eliminated minimotifs that would be introduced by a p53 allele containing these SNPs.*

To explore the role the minimotifs in disease we have built a function into our existing Minimotif Miner website that is capable of examining how SNPs change minimotifs for one query protein at a time [98]. This tool allows investigators to generate a new hypothesis for the cause of almost any disease. An example analysis of the human p53 protein is shown in **Figure 1**. A user can select any set of the 22 known polymorphisms in this gene derived from an older build of dbSNP. In this example, three polymorphisms were selected (colored green). After selecting the polymorphisms, the tool reveals a new table with any new minimotifs that are introduced (green font) or eliminated (red font) by this set of SNPs. This analysis uses ~240,000 experimentally verified minimotifs and ~1500 consensus sequences. Although this tool can be used for examining any one protein, in this Aim we will expand upon this concept and build a tool that can analyze complete genomes and exomes, as well as sets of multiple genomes and exomes.

As preliminary progress toward genome/exome wide analysis of SNPs and minimotifs, we have built a Java application that reads multiple BAM exome files, accesses an internal SNP database, indexes the chromosomal positions to the same reference genome, and calculates the allele frequencies of known SNPs and minimotifs. The analysis done using this tool is primitive and based mainly on simple statistics such as the mean and variance. In this project we propose to develop enhanced analysis tools based on solid statistical techniques as explained below.

Analysis for the role of minimotifs in disease and evolution. The proposed analysis tool will reduce the search space by providing different staged services: 1) Analyze just the minimotifs in the 4 million common variants; 2) Analyze just rare variants in minimotifs; and 3) Analyze only new rare variants in the exome that are not previously observed. This tool will be used to test our central hypothesis: minimotifs are commonly mutated at SNPs that have a known association with disease. All the information gathered in Aim 1 will be utilized in Aim 2 to conduct phenotype-genotype analyses based on novel learning techniques. This analysis tool will become a part of MnM. One of the basic functions supported will be the following: A user can input any population data. This data will be analyzed and a report will be generated. The report will contain a list of motifs or combinations of motifs that reach genome-wide significance. A list will also be provided for candidates approaching significance above a slightly lower threshold. For each motif reported, it will contain the minimotif, its location, whether or not it is a minimotif loss of a function variant, the minimotif source and target genes, minimotif activity and any PubMed ids for papers associated with the minimotif. In the backend, MnM will periodically obtain all the newly released genome data and conduct a minimotif analysis. Details on the statistical analysis to be conducted follow. Note that the statistical techniques described below will also apply to Aim 2.

Statistical Analysis: In Aim 1 the analysis we conduct is to identify the difference between two groups of subjects in terms of minimotif (or SV) distributions. A minimotif can be thought of as a string of characters from an alphabet Σ. Any subject can be represented as an ordered concatenation of the corresponding minimotifs which again is a string. Let m be the length of this string. For example, the string corresponding to a subject will take the form $\sigma_1\sigma_2 \cdots \sigma_m$, where each σ_i is a character from Σ. The problem is to check how this string changes between two groups. For instance we can consider m distributions, one for each character in this string. We could measure how the distribution of σ_i differs between the two groups of subjects. The two groups could be case and control, for example. One group could have an illness and the other does not. Also the size of the group could be 1 or more. We can think of each σ_i as an attribute or variable. These attributes will be analyzed using a mixed-effects model by regressing on them adjusting for baseline severity of illness and covariates and accounting for within-subject correlation among repeated measures within each subject. If necessary, missing values will be estimated using regression-type imputations. The initial data analysis will include univariate explorations of all the variables, using histograms, statistical summaries, and other graphical techniques. Expected ranges for all the variables will be defined a priori; and out-of-range values, or outliers, will be checked for errors, including going back to original data forms. This initial analysis will also serve the purpose of describing the study characteristics and identifying skewed variables that need transformation. Outliers that appear to be real will be kept in the data sets, but the Multivariate Outlier Detection Algorithm will be used to eliminate outliers in all variables before constructing the final analytical datasets [132, 133]. This will eliminate the possibility that any associations seen are driven by a few extreme observations.

A paired t-test will be used to assess the difference between the two groups. A two-sample t-test and one-way ANOVA will be used to assess the difference between subjects within the two different groups. A linear regression model and a mixed-effects model (for repeated measurements) will be used in the multivariate analysis, adjusting for other covariates. We will adjust for multiple comparisons using methods such as a Bonferroni correction or FDR-based adjustment [135, 136]. Also, we will use dimension reduction tools such as hierarchical clustering and principal component analysis to reduce the number of relevant variables. We will also analyze the genotyping data using SNP-sets analysis, which is a logistic kernel association test [134]. Standard analysis of a GWAS, which focuses on assessing the association between each individually genotyped SNP and disease risk, sometimes suffers from limited reproducibility and difficulties in detecting multi-SNP and epistatic effects. In contrast, SNP-set analysis, which borrows information from different but correlated SNPs grouped on the basis of genomic structure or prior biological knowledge, e.g., exons, introns, promoter regions, genes, and pathways, offers improved reproducibility and increased power. This will be of interest due to the use of the minimotif SNP sets.

Analysis of the structural variants: Mark Gerstein's lab has extensive experience in large-scale variant calling and interpretation through being active members of the 1000 Genomes Consortium, especially in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups of the consortium. We have a lot of experience in large-scale structural variant calling [137, 138, 139, 140]. We have developed a number of SV calling algorithms, including BreakSeq, which compares raw reads with a breakpoint library (junction mapping) [141], CNVnator, which measures read depth and estimates copy number variation [142], AGE, which refines local alignment [143], PEMer, which uses paired ends [144].

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets [145, 146, 147]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations [145]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region [148]. Gerstein lab also used allelic variability to prioritize regions of the genome. Our variant analysis work includes AlleleSeq [149], a computational pipeline to identify allele-specific events, and AlleleDB, our database connecting single nucleotide variants with allele-specific binding and expression. We also used networks as a framework for integrating a great variety of genomic variation/mutation data across individuals and organisms and studying their impact on biological systems [150, 151].

In this project we will build on the variant annotation and characterization tools to develop novel computational pipelines that identify variants that are significantly associated phenotypic variation and disease in humans. We will first utilize the BreakSeq algorithm's output on breakpoint analysis and combine it with the minimotif information. This will enable us to stratify the breakpoints with respect to the nearby minimotifs. We will next build a model that correlates the variation in the minimotifs and the existence of breakpoint among 1000 Genomes individuals. Followed by this we will combine this model with the gene expression levels from the GEUVADIS project to identify the minimotifs that are associated with breakpoints and gene expression levels. We will also use the GTex project's tissue specific expression data to classify the eQTLs with respect to tissue specificity. Likewise, we will use 1000 Genomes data to identify the population specific eQTLs in minimotifs.

We will next focus on allele specific analysis of the minimotifs. For this, we will first query the AlleleDB web site for the minimotifs to evaluate whether there is enrichment for certain minimotifs in the database. We will then extend the AlleleSeq pipeline for identifying the allele specific effects in relation to the minimotifs. For this, we will first build the diploid personal genomes, then map the RNA sequencing reads to the diploid genomes. We will then use the heterozygous SNV calls within minimotifs and estimate the unbalanced allelic events using the AlleleSeq's statistical machinery. In particular, we will model the dispersion in RNA-seq read counts so as not to bias the allelic calls because of the reference bias. We will finally focus on the structured non-coding regulatory elements and evaluate how they are effected by the minomotifs. Many regulatory non-coding RNAs have been shown to have motifs with stable secondary structures that are vital for their function. For this, we will use our RNA secondary structure prediction pipeline, incRNA [152], and modify it to identify the structured RNAs around the minimotifs. We will then focus on these regions and evaluate how the variation in the secondary structure around the minimotifs may be related to changes in expression of nearby genes. We will build a multivariate computational model that accounts for many factors like existence of nearby variants.

C.2.2. Aim 2: Phenotype-Genotype Correlational Study:

A main problem of interest to biologists is to study the correlation between phenotypes and genotypes. This problem takes as input (say) two groups of individuals separated based on some phenotype, together with their genotypes and SV information and the task is to identify the most relevant SNPs and SVs that can explain the groupings. In this context, SNPs or SVs can be substituted with any type of genetic variations. In Aim 2 we propose to develop novel algorithms for the study of phenotype-genotype correlations. Aim 1 helps in identifying a small subset of SNPs and SVs that are very likely to be the most important causes for phenotypes. In Aim 2 we will work with this reduced set of attributes. It is conceivable that this set can be reduced further. Our prior approach [110] is based on two paradigms: gene selection [116] and random projections [1] to identify a subset of SNPs from a set of SNPs that can altogether differentiate two groups of individuals efficiently and reliably within a short amount of time. In the first approach, we employ gene selection (GS), a feature selection algorithm, to identify the k most relevant SNPs (where k can be chosen by the user) to differentiate a group of individuals from another. To validate this approach, we computed the p-value for each of the SNPs. It is found that a significant number of SNPs selected by GS has a very low p-value. In the second approach, we employ random projections to project the original data into a space of dimension d (where d can be chosen by the user). We then compute a subset of dimensions which can together differentiate two groups of individuals. We have done this in two steps. We first use GS to pick the best m SNPs (for some suitable value of m) and then project the points onto a k-dimensional space for various values of k. GS is then employed to identify a subset of dimensions that can best predict a particular class of subjects. Both of these approaches have yielded very good outcomes in the context of opium addiction analysis [110]. They also outperform one of the currently best performing algorithms [105] in terms of predicted accuracy and runtime. In this project we plan to employ these (with suitable modifications). We also propose to employ the k-locus association problem formulation for phenotype-genotype correlational study. We provide details on these approaches next.

C.2.2.1. Gene Selection: Gene selection is a classification algorithm based on support vector machines (SVMs). The aim of gene selection algorithm is to identify the (smallest) subset of genes responsible for certain event(s) [116]. Please note that even though in the gene selection algorithm we refer to genes, the algorithm is generic and in general a 'gene' should be thought of as an arbitrary feature. Gene selection is based on SVMs and it takes as input n genes $g_1, g_2, \cdots, g_n$, and l vectors $v_1, v_2, \cdots, v_l$. In our study, each gene corresponds to a genetic attribute (a particular SNP, for example) and each vector corresponds to a subject. Each vector could be of the following form: $v_i = x_i^1, x_i^2, \cdots, x_i^n, y_i$. Here x_i^j is the value of the j^{th} gene g_j for subject i. The value of y_i is either *+1* or *-1* based on whether the subject i has a specific phenotype or not. The problem is to identify a subset of genes $g_i^1, g_i^2, \cdots, g_i^m$ sufficient to predict the value of y_i for any subject i. Given a set of vectors, the gene selection algorithm learns to identify the minimum subset of genes needed to predict the event of interest and the prediction function. These vectors form the training set for the algorithm. Once trained, the algorithm is provided with a new set of data which is called the test set. The accuracy of gene selection is measured in the test set as a percentage of subjects for whom the algorithm correctly predicts the phenotype of interest. The procedure solely relies on the concept of SVM.

[38] introduced a naive gene selection algorithm called sort-SVM. Here the genes were sorted according to their corresponding weights and a subset of genes was selected from the sorted sequence and thus discarded the redundant information. The authors also developed an algorithm called Recursive Feature Elimination (RFE) which is based on the sensitivity analysis proposed by [52]. In [116] we have presented a more efficient algorithm called the Greedy Correlation Incorporated Support Vector Machine (GCI-SVM) algorithm. The main idea is to sort the genes

according to the norm of the weight vector corresponding to these genes. GCI-SVM brings the concept of sort-SVM and RFE-SVM together which makes it more efficient. We will employ GS on the attributes identified in Aim 1.

C.2.2.2. Random Projections: Given that the number of genetic attributes could be very large, we can obtain huge computational efficiencies (without sacrificing much on the quality) by reducing the data dimension. Mapping a set of points from a higher dimensional space to a lower dimensional space in such a way that the pair-wise distances are closely preserved is a problem that has been studied widely. A finite set of n points in a d-dimensional Euclidean space can be represented by a matrix $A_{n\times d}$, where each row represents a point in d dimensions. The objective is to identify a mapping $f: \mathbb{R}^d \rightarrow \mathbb{R}^k$ with negligible distortion in the distance between any pair of points. Here k is the dimension of the reduced space. [43] have given an elegant randomized mapping such that the original pairwise distances are ϵ-preserved in the k-dimensional space. Their theorem establishes the existence of a $O(\log n)$-dimensional space that preserves distances closely with a high probability.

We can use random projections in conjunction with any feature selection algorithm. We expect Aim 1 to identify around 10,000 attributes. This is because in our prior analysis we have found that the number of minimotifs found in the human genome is around 10,000. This itself is a huge reduction (from the millions of SNPs that have been identified for the human genome, for example). Random projections will be employed to further reduce this dimension. There will be two phases. In the first phase we will reduce the data dimension and in the second phase we will employ a feature selection algorithm such as GS on the reduced data set. The quality of output will be judged using p-values and other statistical analysis techniques described in Section C.2.1.

C.2.2.3. *k*-locus Association Problem Formulation: The problem of identifying the most relevant variations responsible for a specific phenotype can also be formulated as the k-locus association problem. This problem can be thought of as the problem of identifying the most relevant k variants responsible for a disease. Most of the techniques in the literature only work when k=1. The problem is much more difficult when k>1. The real cause for a disease could be a combination of (10s or even 100s) of variants. The two-locus association problem is defined as follows. Input is a matrix M of size $(m_1 + m_2)\times n$ where $m_1 + m_2$ is the number of subjects each with n attributes (SNPs, e.g.). Here m_1 is the number of cases and m_2 is the number of controls. There are three possible values for each SNP, namely, 0, 1, or 2. The cases are of phenotype 1 and the controls are of phenotype 0. Rows 1 through m_1 of M correspond to cases. Let this submatrix be called A. Rows $m_1 + 1$ through $m_1 + m_2$ of M correspond to controls and let this submatrix be called B. Each column of M corresponds to an SNP. The two-locus association problem is to identify the pair of SNPs whose statistical correlation with phenotype is maximally different between cases and controls. As mentioned in [2], the goal is to identify the pair: $\underset{i,j}{\arg max} |P_A(i,j) - P_B(i,j)|$. If Q is any matrix, then, $P_Q(i,j)$ denotes the correlation between the columns i and j of Q.

The algorithm of [2] exploits the light bulb algorithm of [81] and locality sensitive hashing (LSH) [13]. They use LSH to transform matrices A and B to A' and B', respectively. In particular, each column c_i of A is converted to a column c_i' of zeros and ones. The size of c_i is $1\times m_1$ and the size of c_i' is chosen to be $u = \max\{m_1, m_2\}$. The matrix B is also transformed into B' in a similar manner using LSH. Followed by this, the pair of interest is identified. To be precise, using A' and B', the matrix D is formed where $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A'} & \boldsymbol{A'} \\ \boldsymbol{B'} & \overline{\boldsymbol{B'}} \end{bmatrix}$. Here $\overline{B'}$ is obtained from B' by complementing every element of B'. Note that D is of size $2u\times 2n$. Let $D_1 = \{1, 2, \cdots, n\}$ and $D_2 = \{n+1, n+2, \cdots n, 2n\}$. Consider all the pairs of columns (i,j) such that $i \in D_1$ and $j \in D_2$. From out of these pairs, identify the pair (i', j') of columns with the maximum number of matches. If $i' = a$ and $j' = n + b$, then (a, b) is the pair of interest. This pair can be found using the light bulb algorithm of [81]. In our recent work [86] we have introduced some novel mapping schemes to avoid LSH and speedup the solution of the 2-locus problem. Experimental results reveal that our algorithm is faster than that of [2].

Table 1: Run times Comparison

n	Brute force	Our algorithm
50K	4.97	0.13
100K	19.00	0.22
300K	178.23	0.87
500K	482.30	1.35
1,000K	1867.05	3.48

Table 1 presents a comparison of our algorithm and the brute force algorithm for the 2-locus problem. Times are in minutes. Our algorithm is also more than 2 orders of magnitude faster than the best known algorithms published in the literature [86]. When k is more than 2, the technique of forming the above matrix D does not work. Specifically, it is not clear how to identify the *least correlated k bulbs* from out of n given bulbs except for doing a brute force search (that will take $O(n^k)$ time). We have developed the following greedy algorithm. First use our 2-locus algorithm to identify the most correlated m pairs of SNPs (for some suitable value of m). Let S be the set of these pairs. For each pair (a, b) in S we do the following: Add (a, b, c) as a candidate triplet to a set Q, for every SNP c

other than *a* and *b*. After forming the set *C* in this manner, evaluate each triplet in *C* and identify the best triplets. We are in the process of testing this algorithm. We also believe that random sampling can play a major role for this problem. A simple strategy is to pick a random sample and analyze this sample to identify the best triplets. We also plan to develop randomized algorithms.

Table 2 shows the run times of the brute force algorithm for solving the 3-locus problem. This algorithm takes cubic time. In this table *n* stands for the number of SNPs, s for seconds, m for minutes, h for hours, d for days and y for years. Run times in bold face are estimates. For one million SNPs the expected run time is 628 years! Thus it is very important for us to develop novel algorithms. Given the complexity of the problem, parallelism is needed to reduce the run times. Also, the brute force algorithm will end up generating a cubic number of triplets not all of which can be stored in the core memory. This means that these intermediate data will have to be stored in secondary storages such as disks. Data access from disks is very time consuming and we need efficient out-of-core algorithms. For instance if we have one million SNPs, the number of triplets is 1.67×10^{17}. If each triplet takes 9 bytes, then the amount of secondary storage needed will be 1500 petabytes!

Table 2: Runtimes

n	**Run time**
1,000	28 s
2,000	223 s
3,000	763 s
4,000	21 m
5,000	58 m
10,000	5.5 h
100,000	**229 d**
1000,000	**628 y**

Handling structural variants: SNPs corresponding to minimotifs are expected to correlate well with phenotypes (such as diseases). It is conceivable that for some specific phenotypes this may not be the case. In this cases minimotif SNPs in conjunction with other structural variants (SVs) could display a high correlation with phenotypes. We will also expand our minimotif database in MnM using the tools that we have already developed for populating MnM. The combination of SNPs and SVs can be handled using all the algorithms discussed above for correlations finding. As an example, consider the 2 locus problem. The input for this problem can be thought of as matrices where the elements are 0, 1, or 2. We can convert the SVs data into this format. For instance, if we have categorical SVs they can be easily represented in this format. If a SV is real valued, it can be discretized and mapped to binary strings. We will employ this approach to combine SNPs and SVs in building correlation models.

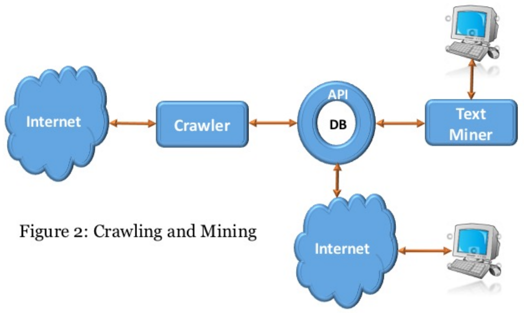


Figure 2: Crawling and Mining

C.2.3. Aim 3: Validation of our approach using text mining and cross validation

C.2.3.1. Rationale: Validation of our approach is a critical task. In Aim3 we will evaluate and validate our computational approaches using publicly available datasets and the literature. If there is an algorithm that predicts the genotypes responsible for a phenotype, there are two ways of validating the predictions. The first way is to employ biological experiments involving human subjects. The other approach is to check if the computational predictions coincide with what scientists have reported in the literature using biological experiments. It is conceivable that most of the papers in the literature report only single variants associated with diseases. Once an algorithm is validated in this manner, then our confidence in the algorithm will be established and we could then use it for unknown phenotypes. Our project being mainly computational, we plan to employ this approach. The employment of text mining in this manner is novel. PubMed will be one of our primary sources.

Our validations will be done for two specific diseases, namely, smoking addiction and schizophrenia. The approach is generic and can be extended for any disease. As a part of the GWAS, various research groups have conducted case-control studies for various diseases. These datasets are available in such databases as dbGaP. We will run our algorithms on these datasets to identify the most significant genotypes that can be associated with the different diseases. In order to validate the outputs of our algorithms, we will employ novel text mining techniques. The idea is to collect all the information published about the different diseases in the literature. Specifically, we will be interested in biological studies that report findings on biomarkers, genetic mutations, genotypes, etc. corresponding to the diseases. We will compare the computational predictions with the experimentally validated findings reported in the literature. A summary of the process is shown in Figure 2. The crawler will be collecting articles from various sources and populate a database. The text miner will use this database to mine for relevant articles. We can also think of the crawler as a filter that is more lenient than the text miner. We plan to make the database created public and it can be accessed by other interested researchers.

C.2.3.2. Preliminary Progress: Research in medical and biological sciences has resulted in the generation of voluminous datasets. As a result, retrieving information relevant to a specific topic has become a big challenge. In our recent work we have presented generic computational techniques that can classify articles efficiently [128, 129, 130, 131]. Our algorithms are based on a number of elegant ideas (summarized below) and perform better than the other known algorithms in accuracy and speed. Our algorithms have been tested in the domain of motif search. We have used text mining to populate the database of MnM. In this project we will specialize them for genotype-disease correlations. Depending on the performance of these algorithms in the new domain, we will modify them to obtain optimal performance. If these algorithms do not perform well, we will develop classification algorithms based on other learning algorithms such as probably approximately correct (PAC) learners and neural networks.

C.2.3.3. Our TextMine Algorithm

The basic problem of text mining can be stated as follows: given a research article (or an abstract), automatically rank the article by its likelihood of containing information of relevance. One of the early algorithms for text mining in biology is that of Goh, *et al.* [127]. This algorithm has been employed to characterize unknown microorganisms. In [131] we have presented a paper scoring (PS) algorithm called TextMine that has a better performance than the algorithm of [127]. TextMine has been used to identify papers containing information on minimotifs. We used a subset of papers as a training set for training the PS algorithm. Each article in a research article collection A, which is used for training, is read by hand and given a score of either 0, indicating that the paper does not contain relevant information, or 1, indicating that the paper has some relevant information. A crucial difference between our PS algorithm and that of Goh, *et al.*, is that PS algorithm provides an ordering of the papers instead of using a threshold.

The workflow for this phase consists of the following steps: We start with disjoint sets P, N, and T of abstracts (or full papers), which are positive, negative, or not reviewed for minimotif content, respectively. Let W be the ordered term vector found by taking all significant words from the documents of sets P and N. Words like "the", "of", "new", etc., that have no discriminatory value between P and N will not be in W. For each word w in W and each article a in P we divide the number of instances of w by the size of a: this is the enrichment of w in a. Then, we sum these enrichments over all articles in P and divide by the size of P to obtain an overall enrichment of w. We repeat this over the set N, and subtract the result from wp to arrive upon a "score" for word w which ranges from -1 to 1. Higher values indicate more positive association with minimotif content. We now have a vector of decimal "scores", which has the same dimension as W, with one entry per term in the term vector. Call this vector S.

Now, we compute a score for each unknown paper by combining word scores. This phase has the following steps:
1) Scan through the paper (or abstract) to count how many times each word w of W occurs in this article.
2) Construct a vector v of enrichment values of words in the paper in which the order corresponds with S.
3) Compute the correlation between v and S and obtain a Pearson's correlation coefficient pc for each paper. If X and Y are any two random variables, then the Pearson's correlation coefficient between X and Y is computed as $\frac{E(X-\mu_X)E(Y-\mu_Y)}{\sigma_X\sigma_Y}$, where μ_X is the expected value of X, μ_Y is the expected value of Y, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.
4) Thus, we have now computed a "score" of the article, which is the Pearson's correlation coefficient between the scored words from the training set W and respective enrichments of those words in the article a.

The correlation coefficients for the lexemes range from -1.000 to 1.000. This score positively correlates with the presence of minimotif content, as expected.

Improvements to TextMine:

We have improved the performance of TextMine further with the employment of a number of techniques including 1) selecting the keywords in a careful way, 2) deterministic and randomized sampling, 3) Support Vector Machine (SVM) [130], and 4) the Gene Selection algorithm [129]. In this project we will start with these algorithms and specialize them for genotype-disease correlations. In [153] we have introduced a novel randomized technique for feature selection. This technique can be used in the context of any learning algorithm. We plan to use this as well.

Table 3: Representative Datasets

Disease	Data ID	No. of subjects
Breast Cancer	phs000812.v1.p1	12,501
Breast Cancer	phs000851.v1.p1	5,367
Breast Cancer	phs000799.v1.p1	5,152
Parkinson's Disease	phs000918.v1.p1	11,402
Parkinson's Disease	phs000196.v3.p1	4,011
Parkinson's Disease	phs.000126.v1.p1	1,991
Cardiovascular Disorders	phs000963.v1.p1	5,890
Cardiovascular Disorders	phs001013.v1.p1	357

C.2.3.4. Cross Validations

Another computational technique we plan to employ for the validation of our computational predictions is using the cross validation approach. The idea here is to employ independent datasets. For the same disease independent teams have published (case-control) datasets. See e.g., Table 3. We will use one of these datasets (call it D1) and run our algorithms to produce predictions. For instance these predictions will include a list of the *k* most relevant variants associated with the disease. We will run our algorithms on a dataset (call it D2) published by a different group for the same disease to get predictions and compare these with the predictions obtained for D1. In some sense we use D1 as training set and D2 as test set. A close matching between the two predictions will add to the confidence in the predictions and the algorithms. We will do this validation for different diseases using DbGaP.

C.2.3.5. Datasets to be used

We plan to employ all the genotype datasets publicly available for various phenotypes (especially diseases). dbGaP will be one of the important databases that we plan to employ. We will also employ our crawler to identify other relevant databases. dbGaP is a rich database that has large case-control datasets for various diseases. We have identified several datasets for the diseases of our interest. Some of these datasets are summarized in Table 3. All of these datasets are of the Case-Control type.

Most of the experimental results in the literature report single SNPs correlated with diseases. For these results, we will run our algorithms for the case of *k*=1 and compare the predictions with the experimental results.

C.2.4. Algorithmic Techniques

C.2.4.1. Random Sampling

Our prior study has revealed that random sampling can be very effective for addressing motif search and text mining. Sampling has served as an effective tool in the design of algorithms over a variety of models and for varying problems. The idea of sampling is to pick a small subset of the input, process this subset, and infer certain global properties of the input so as to be able to efficiently solve the problem under concern. An early example of sampling-based algorithm is that of [35] for sorting. The idea here was to pick a random sample of keys, sort the sample, use the sorted sample keys to partition the original input, and recursively sort the resultant independent parts. This idea has since then been used to develop optimal randomized sorting algorithms on a variety of models (see e.g., [101], [100], [45], [89], [91], [92], and [93]). Rajasekaran and Ramaswami [87] have proposed an elegant technique called *multi-sampling* for the design of randomized algorithms. The idea is to perform random sampling at different levels in the algorithm. We plan to use sampling extensively in this project.

C.2.4.2. Parallel Techniques

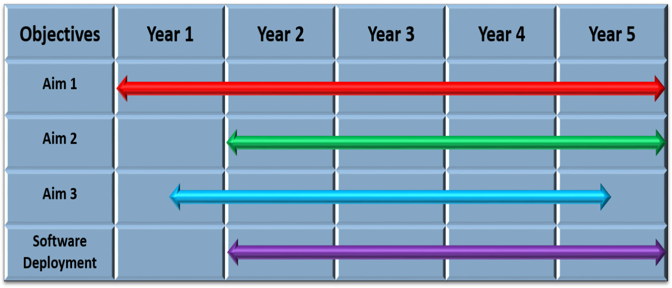
Given the high time complexities associated with the existing algorithms for solving the *k*-locus association problem, parallelism could help tremendously. Parallelism could also help in Aim 1 where we have to deal with voluminous datasets. The PI has an extensive experience in the design of efficient parallel algorithms for numerous fundamental problems of computing such as sorting, selection, clustering, sequence assembly, motif search, Voronoi diagram, etc. In this project we propose to design novel parallel algorithms for computational problems involved in Aims 1, 2, and 3. Emphasis will be placed for employing multicore architectures such as GPUs (e.g., [7]). The Booth Engineering Center for Advanced Technologies (BECAT) at UConn houses a large CPU/GPU cluster with 6,248 cores and 14,336 GPU cores. A brief summary of our approach follows.

We have developed two general techniques for speeding up parallel computations, called LessTalk and micro kernels. LessTalk enables one to reduce inter-processor communications and micro kernels help in reducing the data access times. We have demonstrated the effectiveness of LessTalk in coastal wave simulations, biological cells modeling, and fuel cells modeling [61] [94] [95]. LessTalk has yielded close to linear (sometimes super-linear) speedups for these problems. The idea of LessTalk is to perform some redundant computations in order to reduce the number of communication steps. Multicore achitectures such as GPUs will be utilized extensively in this project. In our prior work we have developed a novel speedup technique for multicore machines. We have developed a micro-threading framework realized by a nano-kernel implemented on top of each core [3] [4] [5]. We have tested this framework on different kinds of algorithms and these results indicate that our framework yields a speedup of up to 500%. Micro-threads are small threads that context switch inside each core's local store whenever there is a memory access wait time. Its context switching cost is relatively small compared to context switching the whole core's state to main memory. This framework should hide memory latency and make it easier to program any multicore machine. In this project we plan to exploit our nano-kernel framework on GPUs to solve compute-intensive problems.

C.2.4.3. Out-of-core Computing: Given the volume of data to be processed, the core memory of a typical desktop, a server, or even a parallel computer may not be enough to hold all the data to be processed. Therefore, efficient out-of-core computing techniques are necessary. Out-of-core computing refers to the case of computing where not all of the data to be processed can be stored in the core memory of the computer used.

Processor speeds in computers have increased steadily over the years. However, disk access speeds have not seen similar improvements. As a result, the I/O bottleneck has grown over time. Loading a register can take a fraction of a nanosecond (10^{-9} seconds), and accessing internal (core) memory takes several nanoseconds, but the latency of accessing data on a disk is multiple milliseconds (10^{-3} seconds), which is about one million times slower [121]! Even the fastest solid state drives (that are very expensive) are much slower than core memory. The latency of solid state disks is several microseconds. Thus any innovations in minimizing the number of I/O operations could yield great time improvements. Not many scientists have access to machines with core memories of the magnitude that biological data require. Thus it is extremely crucial for the bioinformatics community to develop a library of out-of-core algorithms and software that can be utilized by the community. In today's High Performance Computing (HPC) systems, there exists a deep on-chip memory hierarchy. Future architectures are expected to have even more levels of memory hierarchy. As a result, even in the context of HPC, out-of-core algorithms play a vital role. In this project we propose to develop out-of-core algorithms for minimotif analysis and phenotype-genotype correlations.

In the past, we have developed novel out-of-core algorithms for sorting ([84], [93], [48]) and sequence assembly [50]. This experience will be very valuable in this project. Out-of-core computing is vital in Aim 1 and Aim 2 because of the large sizes of the datasets and in Aim 2 because of the large sizes of the intermediate results generated.



C.2.5. Project Management

This is a multi-disciplinary project involving a computer scientist (Rajasekaran–SR), a structural biologist (Schiller–MS), a geneticist specializing in GWAS (Chen-XC), a geneticist specializing in SVs (Gerstein-MG), and a statistician (Harel-OH). Subsets of the team members have collaborated in the past and published together. The PI of the project is Sanguthevar Rajasekaran. He will be responsible for the overall management of the project and all the sequential, parallel, and out-of-core algorithmic aspects. Specifically, he will work on the following problems: out-of-core algorithms for the minimotif analysis, incremental analysis of new genome sequences, parallel and out-of-core algorithms for the *k*-locus problem, and text mining algorithms. Schiller will work on the biological aspects of the project. He will be in-charge of identifying the types of analysis in Aim 1, and coming up with keywords for text mining. Schiller will also work on augmenting MnM with functionalities relevant for GWAS. He will also work with Chen on the genetics part. Chen will work on interpreting the results from the learning algorithms, and suggesting modifications to the approaches. He will analyze the predictions of the computational algorithms with respect to various diseases, especially for smoking addiction and schizophrenia. Gerstein will provide feedback on computational techniques, identify datasets for validation, and provide guidance on all the genetics aspects. Harel will be in-charge of all the statistical analyses especially in Aims 1 and 2. MS, XC, and MG will identify all the datasets needed for the project.

There are two post-doctoral fellows (PD1 and PD2), a Research Scientist (RS) and a graduate student (GA) in this project. PD1 will work with Schiller and Chen and focus on the biological aspects of the project. (S)he will be in-charge of implementation algorithms and methods pertaining to MnM and the statistical analyses involved. PD1 will also help on the analysis methods in Aim 1. PD2 and GA will work closely with the PI and OH on all aspects of the project including algorithms, data collection, implementation of algorithms, analyses, and evaluating them. PD2 will be responsible for Aim 2 and GA will be responsible for Aim 3. PD2 and GA will be responsible for developing all the professional quality software, preparing documents, disseminating the software, and addressing the needs of the users. RS will work with Gerstein on SVs. Specifically, he will be in-charge of building variant annotation and characterization tools to develop novel computational pipelines that identify variants that are significantly associated phenotypic variation and disease in humans. MG will guide RS closely. Close interactions among the investigators will always be kept. These interactions will be vital in sharing techniques and conducting cross-disciplinary activities. The team will meet weekly once via video teleconferencing. Communication via email and telephone will happen constantly as and when needed. A summary of the project timeline that is necessary to accomplish the proposed work over the 5 year timeframe is shown in the above Figure. We expect that some seniors will also participate in implementation aspects as a part of their senior projects.

References

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66 (4): 671–687, 2003.

[2] P. Achlioptas, B. Schölkopf, and K. Borgwardt. Two-locus association mapping in subquadratic time. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 726–734. ACM, 2011.

[3] M.F. Ahmed, R.A. Ammar, and S. Rajasekaran. SPENK: adding another level of parallelism on the cell broadband engine. In *Proceedings of the 1st international forum on Next-generation multicore/manycore technologies*, page 2. ACM, 2008.

[4] M.F. Ahmed, R. Ammar, S. Rajasekaran, et al. Novel micro-threading techniques on the Cell Broadband Engine. In *Computers and Communications, 2009. ISCC 2009. IEEE Symposium on*, pages 570–575. IEEE, 2009.

[5] M.F. Ahmed, S. Rajasekaran, and R.A. Ammar. FFTI: Fast In-Place FFT on the Cell Broadband Engine. In *CATA*, pages 167–173, 2010.

[6] S. Balla, V. Thapar, S. Verma, T. Luong, T. Faghri, C.-H. Huang, S. Rajasekaran, J.J. Del Campo, J.H. Shinn, W.A. Mohler, et al. Minimotif Miner: a tool for investigating protein function. *Nature methods*, 3 (3): 175–177, 2006.

[7] S. Bandyopadhyay and S. Sahni. Sorting on a cell broadband engine SPU. In *Computers and Communications, 2009. ISCC 2009. IEEE Symposium on*, pages 218–223. IEEE, 2009.

[8] S. Bandyopadhyay, S. Sahni, and S. Rajasekaran. PMS6: A Fast Algorithm for Motif Discovery. In *IEEE... International Conference on Computational Advances in Bio and Medical Sciences:[proceedings]. IEEE International Conference on Computational Advances in Bio and Medical Sciences*, page 1. NIH Public Access, 2012.

[9] S. Bandyopadhyay, S. Sahni, and S. Rajasekaran. PMS6MC: A multicore algorithm for motif discovery. *Algorithms*, 6 (4): 805–823, 2013.

[10] S. Bandyopadhyay, S. Sahni, and S. Rajasekaran. PMS6: A fast algorithm for motif discovery. *International Journal of Bioinformatics Research and Applications 2*, 10 (4-5): 369–383, 2014.

[11] L.G. Biesecker and R.C. Green. Diagnostic clinical genome and exome sequencing. *New England Journal of Medicine*, 370 (25): 2418–2425, 2014.

[12] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for Phylogenetic Footprinting. *Journal of Computational Biology*, 9 (2): 211–223, 2002.

[13] M.S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

[14] C. Chelala, A. Khan, and N.R. Lemoine. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25 (5): 655–661, 2009.

[15] C. Chica, A. Labarga, C.M. Gould, R. López, and T.J. Gibson. A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC bioinformatics*, 9 (1): 229, 2008.

[16] C. Chica, F. Diella, and T.J. Gibson. Evidence for the concerted evolution between short linear protein motifs and their flanking regions. *PLoS One*, 4 (7): e6052, 2009.

[17] M. Choi, U.I. Scholl, W. Ji, T. Liu, I.R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Özen, S. Sanjad, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, 106 (45): 19096–19101, 2009.

[18] R. Cole. Parallel merge sort. *SIAM Journal on Computing*, 17 (4): 770–785, 1988.

[19] F.S. Collins, M.S. Guyer, and A. Chakravarti. Variations on a theme: cataloging human DNA sequence variation. *Science*, 278 (5343): 1580, 1997.

[20] L. Conde, J.M. Vaquerizas, H. Dopazo, L. Arbiza, J. Reumers, F. Rousseau, J. Schymkowitz, and J. Dopazo. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic acids research*, 34 (suppl 2): W621–W625, 2006.

[21] G.P. Consortium, et al., An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491 (7422): 56–65, 2012.

[22] N.E. Davey, D.C. Shields, and R.J. Edwards. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*, 25 (4): 443–450, 2009.

[23] N.E. Davey, R.J. Edwards, and D.C. Shields. Computational identification and analysis of protein short linear motifs. *Frontiers in Bioscience*, 15: 801–825, 2010.

[24] N.E. Davey, N.J. Haslam, D.C. Shields, and R.J. Edwards. SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic acids research*, page gkq440, 2010.

[25] J. Davila, S. Balla, and S. Rajasekaran. Fast and practical algorithms for planted (l, d) motif search. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 4 (4): 544–552, 2007.

[26] G. De Baets, J. Van Durme, J. Reumers, S. Maurer-Stroh, P. Vanhee, J. Dopazo, J. Schymkowitz, and F. Rousseau. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Research*, page gkr996, 2011.

[27] X. Deng. SeqGene: a comprehensive software solution for mining exome and transcriptome sequencing data. *BMC bioinformatics*, 12 (1): 267, 2011.

[28] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43 (5): 491–498, 2011.

[29] H. Dinh, S. Rajasekaran, and V. Kundeti, PMS5: an efficient exact algorithm for the (*l, d*)-motif finding problem, *BMC Bioinformatics* 12: 410, 2011.

[30] H. Dinh, S. Rajasekaran, and J. Davila. qPMS7: a fast algorithm for finding (l,d)-motifs in DNA and protein sequences. *PloS one*, 7 (7), 2012.

[31] H. Dinkel and H. Sticht. A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, 23 (24): 3297–3303, 2007.

[32] S. Flanagan, A. Patch, and S. Ellard. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers*, 14 (4): 533, 2010.

[33] B.L. Fogel, H. Lee, J.L. Deignan, S.P. Strom, S. Kantarci, X. Wang, F. Quintero-Rivera, E. Vilain, W.W. Grody, S. Perlman, et al., Exome sequencing in the clinical diagnosis of sporadic or familial cerebellar ataxia. *JAMA neurology*, 71 (10): 1237–1246, 2014.

[34] K.A. Frazer, S.S. Murray, N.J. Schork, and E.J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10 (4): 241–251, 2009.

[35] W.D. Frazer and A. McKellar. Samplesort: A sampling approach to minimal storage tree sorting. *Journal of the ACM (JACM)*, 17 (3): 496–507, 1970.

[36] S.J. Goodswen, C. Gondro, N S. Watson-Haigh, and H.N. Kadarmideen. FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. *BMC bioinformatics*, 11 (1): 311, 2010.

[37] C.M. Gould, F. Diella, A. Via, P. Puntervoll, C. Gemünd, S. Chabanis-Davidson, S. Michael, A. Sayadi, J.C. Bryne, C. Chica, et al., ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic acids research*, page gkp1016, 2009.

[38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46 (1-3): 389–422, 2002.

[39] E. Horowitz, S. Sahni, and S. Rajasekaran. *Computer Algorithms*. Silicon Press, Summit, NJ, USA, 2nd edition, 2007. ISBN 0929306414, 9780929306414.

[40] Y. Igarashi, A. Eroshkin, S. Gramatikova, K. Gramatikoff, Y. Zhang, J.W. Smith, A.L. Osterman, and A. Godzik. CutDB: a proteolytic event database. *Nucleic acids research*, 35 (suppl 1): D546–D549, 2007.

[41] A. Jaffe, G. Wojcik, A. Chu, A. Golozar, A. Maroo, P. Duggal, and A.P. Klein. Identification of functional genetic variation in exome sequence analysis. In *BMC proceedings*, volume 5, page S13. BioMed Central Ltd, 2011.

[42] P.C. Johnson and T.H. Cormen. Networks beat pipelines: the design of FG 2.0. In *Proceedings of the 2012 International Workshop on Programming Models and Applications for Multicores and Manycores*, pages 168–175. ACM, 2012.

[43] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26 (189-206): 1, 1984.

[44] K. Kadaveru, J. Vyas, and M.R. Schiller. Viral infection and human disease-insights from minimotifs. *Frontiers in bioscience: a journal and virtual library*, 13: 6455, 2008.

[45] C. Kaklamanis, D. Krizanc, L. Narayanan, and T. Tsantilas. Randomized sorting and selection on mesh-connected processor arrays. In *Proceedings of the third annual ACM symposium on Parallel algorithms and architectures*, pages 17–28. ACM, 1991.

[46] R.J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011: bar030, 2011.

[47] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4 (7): 1073–1081, 2009.

[48] V. Kundeti and S. Rajasekaran. Efficient out-of-core sorting algorithms for the Parallel Disks Model. *Journal of parallel and distributed computing*, 71 (11): 1427–1433, 2011.

[49] V. Kundeti, S. Rajasekaran, and H. Dinh. An efficient algorithm for Chinese postman walk on bi-directed deBruijn graphs. *Discrete Mathematics, Algorithms and Applications*, 4 (02): 1250019, 2012.

[50] V.K. Kundeti, S. Rajasekaran, H. Dinh, M. Vaughn, and V. Thapar. Efficient parallel and out of core algorithms for constructing large bi-directed de bruijn graphs. *BMC bioinformatics*, 11 (1): 560, 2010.

[51] A. Labarga, F. Valentin, M. Anderson, and R. Lopez. Web services at the European bioinformatics institute. *Nucleic acids research*, 35 (suppl 2): W6–W11, 2007.

[52] Y. LeCun, J. S. Denker, and S. A. Solla. Optimal Brain Damage. In *Advances in Neural Information Processing Systems*, pages 598–605, 1990.

[53] H. Lee, J.L. Deignan, N. Dorrani, S.P. Strom, S. Kantarci, F. Quintero-Rivera, K. Das, T. Toy, B. Harry, M. Yourshaw, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAmA*, 312 (18): 1880–1887, 2014.

[54] P.H. Lee and H. Shatkay. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic acids research*, 36 (suppl 1): D820–D824, 2008.

[55] J. Li, S. Ranka, and S. Sahni. Pairwise sequence alignment for very long sequences on GPUs. *International Journal of Bioinformatics Research and Applications 2*, 10 (4-5): 345–368, 2014.

[56] K. Li and T.B. Stockwell. VariantClassifier: a hierarchical variant classifier for annotated genomes. *BMC research notes*, 3 (1): 191, 2010.

[57] L. Li, C. Tibiche, C. Fu, T. Kaneko, M.F. Moran, M.R. Schiller, S.S.-C. Li, and E. Wang. The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. *Genome research*, 22 (7): 1222–1230, 2012.

[58] S. Li, L. Ma, H. Li, S. Vang, Y. Hu, L. Bolund, and J. Wang. Snap: an integrated SNP annotation platform. *Nucleic acids research*, 35 (suppl 1): D707–D710, 2007.

[59] S. Li, L.M. Iakoucheva, S.D. Mooney, and P. Radivojac. Loss of post-translational modification sites in disease. In *Pacific Symposium on Biocomputing*, volume 15, pages 337–347. World Scientific, 2010.

[60] C.-K. Liu, Y.-H. Chen, C.-Y. Tang, S.-C. Chang, Y.-J. Lin, M.-F. Tsai, Y.-T. Chen, and A. Yao. Functional analysis of novel SNPs and mutations in human and mouse genomes. *BMC bioinformatics*, 9 (Suppl 12): S10, 2008.

[61] J. Luo and S. Rajasekaran. Parallizing 1-dimensional estuarine model. *International Journal of Foundations of Computer Science*, 15 (06): 809–821, 2004.

[62] K.F. Lyon, C.L. Strong, S.G. Schooler, R.J. Young, N. Roy, B. Ozar, M. Bachmeier, S. Rajasekaran, and M. R. Schiller. Natural variability of minimotifs in 1092 people indicates that minimotifs are targets of evolution. *Nucleic acids research*, page gkv580, 2015.

[63] A.-A. Mamun, R. Aseltine, and S. Rajasekaran. Efficient Record Linkage Algorithms using Complete Linkage Clustering. *In submission: Nature Scientific Reports*, 2015.

[64] A.-A. Mamun, R. Aseltine, and S. Rajasekaran, RLT-S: A web system for record linkage. *PLoS ONE*, 10 (5), 2015.

[65] J.C. Merlin, S. Rajasekaran, T. Mi, and M.R. Schiller, Reducing false-positive prediction of minimotifs with a genetic interaction filter. *PloS one*, 7 (3): e32630, 2012.

[66] T. Mi and S. Rajasekaran. Efficient algorithms for biological stems search. *BMC bioinformatics*, 14 (1): 161, 2013.

[67] T. Mi and S. Rajasekaran. A two-pass exact algorithm for selection on Parallel Disk Systems. In *Computers and Communications (ISCC), 2013 IEEE Symposium on*, pages 000612–000617. IEEE, 2013.

[68] T. Mi, S. Rajasekaran, and R. Aseltine. Efficient algorithms for fast integration on large data sets from multiple sources. *BMC medical informatics and decision making*, 12 (1): 59, 2012.

[69] T. Mi, S. Rajasekaran, J. C. Merlin, M. Gryk, and M.R. Schiller. Achieving High Accuracy Prediction of Minimotifs. *PLoS ONE*, 7 (9), 2012.

[70] V. Neduva and R.B. Russell. DILIMOT: discovery of linear motifs in proteins. *Nucleic acids research*, 34 (suppl 2): W350–W355, 2006.

[71] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. De Masi, T.J. Gibson, J. Lewis, L. Serrano, and R.B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS biology*, 3 (12): 2090, 2005.

[72] P.C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31 (13): 3812–3814, 2003.

[73] P.C. Ng, S. Levy, J. Huang, T.B. Stockwell, B.P. Walenz, K. Li, N. Axelrod, D.A. Busam, R.L. Strausberg, and J.C. Venter. Genetic variation in an individual human exome. *PLoS Genet*, 4 (8): e1000160, 2008.

[74] M. Nicolae and S. Rajasekaran. Efficient Sequential and Parallel Algorithms for Planted Motif Search. *BMC bioinformatics*, 15 (1): 34, 2014.

[75] M. Nicolae and S. Rajasekaran. qPMS9: An Efficient Algorithm for Quorum Planted Motif Search. *Nature Scientific Reports*, 5, 2015.

[76] M. Nicolae, S. Pathak, and S. Rajasekaran. LFQC: A lossless compression algorithm for FASTQ files. *Bioinformatics*, page btv384, 2015.

[77] J.C. Obenauer, L.C. Cantley, and M.B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research*, 31 (13): 3635–3641, 2003.

[78] S. Pal and S. Rajasekaran. Improved Algorithms for Finding Edit Distance Based Motifs. *IEEE BIBM*, 2015, 537-542.

[79] S. Pathak, V. K. Kundeti, M.R. Schiller, and S. Rajasekaran. A Structure Based Algorithm for Improving Motifs Prediction. In *Pattern Recognition in Bioinformatics*, pages 242–252. Springer, 2013.

[80] S. Pathak, S. Rajasekaran, and M. Nicolae. EMS1: An Elegant Algorithm for Edit Distance Based Motif Search. *International Journal of Foundations of Computer Science*, 24 (04): 473–486, 2013.

[81] R. Paturi, S. Rajasekaran, and J. Reif. The light bulb problem. *Information and Computation*, 117 (2): 187–192, 1995.

[82] E. Pennisi. 1000 Genomes Project gives new map of genetic diversity. *Science*, 330 (6004): 574–575, 2010.

[83] B. Rabbani, M. Tekin, and N. Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59 (1): 5–15, 2014.

[84] S. Rajasekaran. A framework for simple sorting algorithms on parallel disk systems. *Theory of Computing Systems*, 34 (2): 101–114, 2001.

[85] S. Rajasekaran and H. Dinh. A speedup technique for (l, d)-motif finding algorithms. *BMC research notes*, 4 (1): 54, 2011.

[86] S. Rajasekaran, S. Saha, and S. Pathak, Efficient Algorithms for Genome-wide Association Study and Related Problems, submitted for publication, May 2016.

[87] S. Rajasekaran and S. Ramaswami. Optimal parallel randomized algorithms for the Voronoi diagram of line segments in the plane and related problems. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 57–66. ACM, 1994.

[88] S. Rajasekaran and J.H. Reif. Optimal and sublogarithmic time randomized parallel sorting algorithms. *SIAM Journal on Computing*, 18 (3): 594–607, 1989.

[89] S. Rajasekaran and S. Sahni. Randomized routing, selection, and sorting on the OTIS-mesh. *Parallel and Distributed Systems, IEEE Transactions on*, 9 (9): 833–840, 1998.

[90] S. Rajasekaran and S. Sen. Random sampling techniques and parallel algorithms design. *Synthesis of Parallel Algorithms*, pages 411–451, 1993.

[91] S. Rajasekaran and S. Sen. PDM sorting algorithms that take a small number of passes. In *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International*, pages 10–10. IEEE, 2005.

[92] S. Rajasekaran and S. Sen. A Simple Optimal Randomized Sorting Algorithm for the PDM. In *Proc. International Symposium on Algorithms and Computation (ISAAC)*, pages 543–552, 2005.

[93] S. Rajasekaran and S. Sen. Optimal and Practical Algorithms for Sorting on the PDM. *Computers, IEEE Transactions on*, 57 (4): 547–561, 2008.

[94] S. Rajasekaran, R. Ammar, B. Cheriyan, and L. Loew. Parallel techniques for Virtual Cell. In *Signal Processing and Information Technology, 2004. Proceedings of the Fourth IEEE International Symposium on*, pages 391–395. IEEE, 2004.

[95] S. Rajasekaran, R. Ammar, K. Reifsnider, L. Achenie, A. Mohamed, G. Zhang, and M. Ahmed. Efficient parallel simulation of direct methanol fuel cell models. *Journal of Fuel Cell Science and Technology*, 2 (2): 141–144, 2005.

[96] S. Rajasekaran, S. Balla, and C.-H. Huang. Exact algorithms for planted motif problems. *Journal of Computational Biology*, 12 (8): 1117–1128, 2005.

[97] S. Rajasekaran, S. Balla, C.-H. Huang, V. Thapar, M. Gryk, M. Maciejewski, and M. Schiller. High-performance exact algorithms for motif search. *Journal of clinical monitoring and computing*, 19 (4-5): 319–328, 2005.

[98] S. Rajasekaran, S. Balla, P. Gradie, M.R. Gryk, K. Kadaveru, V. Kundeti, M.W. Maciejewski, T. Mi, N. Rubino, J. Vyas, et al. Minimotif miner 2nd release: a database and web system for motif search. *Nucleic acids research*, 37 (suppl 1): D185–D190, 2009.

[99] N.D. Rawlings, F.R. Morton, C.Y. Kok, J. Kong, and A.J. Barrett. MEROPS: the peptidase database. *Nucleic acids research*, 36 (suppl 1): D320–D325, 2008.

[100] J.H. Reif and L.G. Valiant. A logarithmic time sort for linear size networks. *Journal of the ACM (JACM)*, 34 (1): 60–76, 1987.

[101] R. Reischuk. Probabilistic parallel algorithms for sorting and selection. *SIAM Journal on Computing*, 14 (2): 396–409, 1985.

[102] J. Ren, C. Jiang, X. Gao, Z. Liu, Z. Yuan, C. Jin, L. Wen, Z. Zhang, Y. Xue, and X. Yao. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Molecular & Cellular Proteomics*, 9 (4): 623–634, 2010.

[103] S. Ren, G. Yang, Y. He, Y. Wang, Y. Li, and Z. Chen. The conservation pattern of short linear motifs is highly correlated with the function of interacting protein domains. *BMC genomics*, 9 (1): 452, 2008.

[104] J. Reumers, S. Maurer-Stroh, J. Schymkowitz, and F. Rousseau. SNPeffect v2. 0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*, 22 (17): 2183–2185, 2006.

[105] M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, and J.H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69 (1): 138–147, 2001.

[106] M. Ryan, M. Diekhans, S. Lien, Y. Liu, and R. Karchin. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics*, 25 (11): 1431–1432, 2009.

[107] S. Saha and S. Rajasekaran. Efficient algorithms for the compression of FASTQ files, *Proc. IEEE BIBM*, 2014, pp. 82-85.

[108] S. Saha and S. Rajasekaran. ERGC: An efficient referential genome compression algorithm. *Bioinformatics*, 2015.

[109] S. Saha and S. Rajasekaran. NRRC: A Non-referential Reads Compression Algorithm. In *Bioinformatics Research and Applications*, pages 297–308. Springer, 2015.

[110] S. Saha, S. Rajasekaran, J. Bi, and S. Pathak. Efficient techniques for genotype-phenotype correlational analysis. *BMC medical informatics and decision making*, 13 (1): 41, 2013.

[111] E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, page gkq1172, 2010.

[112] C. Schaefer, A. Meier, B. Rost, and Y. Bromberg. SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics*, 28 (4): 601–602, 2012.

[113] M.R. Schiller, T. Mi, J.C. Merlin, S. Deverasetty, M.R. Gryk, T.J. Bill, A.W. Brooks, L.Y. Lee, V. Rathnayake, C.A. Ross, et al. Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic acids research*, page gkr1189, 2011.

[114] D. Sharma, S. Rajasekaran, and S. Pathak. An experimental comparison of PMSprune and other algorithms for motif search. *International journal of bioinformatics research and applications*, 10 (6): 559–573, 2014.

[115] T.H. Shen, C.S. Carlson, and P. Tarczy-Hornoch. SNPit: a federated data integration system for the purpose of functional SNP annotation. *Computer methods and programs in biomedicine*, 95 (2): 181–189, 2009.

[116] M. Song and S. Rajasekaran. A greedy correlation-incorporated SVM-based algorithm for gene selection. In *Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on*, volume 1, pages 657–661. IEEE, 2007.

[117] S.-H. Tan, W. Hugo, W.-K. Sung, and S.-K. Ng. A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC bioinformatics*, 7 (1): 502, 2006.

[118] S. Teng, E. Michonova-Alexova, and E. Alexov. Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Current pharmaceutical biotechnology*, 9 (2): 123–133, 2008.

[119] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23 (1): 137–144, 2005.

[120] A. Via, C.M. Gould, C. Gemünd, T.J. Gibson, and M. Helmer-Citterich. A structure filter for the Eukaryotic Linear Motif Resource. *Bmc Bioinformatics*, 10 (1): 351, 2009.

[121] J.S. Vitter. Algorithms and data structures for external memory. *Foundations and Trends in Theoretical Computer Science*, 2 (4): 305–474, 2008.

[122] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38 (16): e164–e164, 2010.

[123] P. Wang, M. Dai, W. Xuan, R.C. McEachin, A.U. Jackson, L.J. Scott, B. Athey, S.J. Watson, and F. Meng. SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, 22 (14): e523–e529, 2006.

[124] S. Wuchty, Z.N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature genetics*, 35 (2): 176–179, 2003.

[125] Y. Yang, D.M. Muzny, F. Xia, Z. Niu, R. Person, Y. Ding, P. Ward, A. Braxton, M. Wang, C. Buhay, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312 (18): 1870–1879, 2014.

[126] Z. Zhang, M.A. Miteva, L. Wang, and E. Alexov. Analyzing effects of naturally occurring missense mutations. *Computational and mathematical methods in medicine*, 2012, 2012.

[127] C.S. Goh, T.A. Gianoulis, Y. Liu, J. Li, A. Paccanaro, Y.A. Lussier, and M. Gerstein. Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics* 7:257, 2006.

[128] R. Kilany, R.A. Ammar, and S. Rajasekaran. A correlation-based algorithm for classifying technical articles. *Proc. 11th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec. 14-17, Bilbao, Spain, 2011, pp. 50-53.

[129] R. Kilany, R.A. Ammar, and S. Rajasekaran. A novel algorithm for technical articles classification based on gene selection. *Proc. 12th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),* 2012, pp. 234-238.

[130] R. Kilany, R.A. Ammar, and S. Rajasekaran. Document classification: a novel approach based on SVM. *Proc. 5th International Conference on Bioinformatics and Computational Biology (BICoB),* March 2-6, Honolulu, HI, 2013.

[131] J. Vyas, R.J. Nowling, T. Meusburger, D. Sargeant, K. Kadaveru, M.R. Gryk, V. Kundeti, S. Rajasekaran, and M.R. Schiller. MimoSA: a system for minimotif annotation. *BMC Bioinformatics*, 2010, 11: 328.

[132] B.A. Coull, D. Ruppert, M.P. Wand. Simple incorporation of interactions into additive models. *Biometrics* 2001;57:539-45.

[133] A. Houseman, L.M. Ryan, and B.A. Coull. Cholesky Residuals for Assessing Normal Errors in a Linear Model With Correlated Outcomes. *Journal of the American Statistical Association* 2004;99:383-94.

[134] M.C. Wu, P. Kraft, M.P. Epstein, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 2010;86:929-42.

[135] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research,* 125, 2001, 279-284.

[136] Y. Benjamini and Y. Yekutieli. False discovery rate controlling confidence intervals for selected parameters. *Journal of the American Statistical Association* **100** (469): 71 80. doi:10.1198/016214504000001907.

[137] Z.D. Zhang, H. Lam, A. Abyzov, A.E. Urban, M. Snyder, and M. Gerstein. Identification of genomic indels and structural variations using split reads. *BMC Genomics*, 2011 Jul 25;12:375. doi: 10.1186/1471-2164-12-375.

[138] R.E. Mills, K. Walter, K. Handsaker, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 2011 Feb 3;470(7332):59-65. doi: 10.1038/nature09708.

[139] 1000 Genomes Project Consortium, et al. A map of human genome variation from population-scale sequencing. *Nature*, 2010 Oct 28;467(7319):1061-73. doi: 10.1038/nature09534.

[140] 1000 Genomes Project Consortium, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632.

[141] H.Y. Lam, X.J. Mu, A.M. Stutz, A. Tanzer, P.D. Cayting, M. Snyder, P.M. Kim, J.O. Korbel, and M.B. Gerstein. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology*, 2010 Jan;28(1):47-55. doi: 10.1038/nbt.1600. Epub 2009 Dec 27.

[142] A. Abyzov, A.E. Urban, M. Snyder, and M.B. Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 2011 Jun;21(6):974-84. doi: 10.1101/gr.114876.110. Epub 2011 Feb 7

[143] A. Abyzov and M.B. Gerstein. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics*, 2011 Mar 1;27(5):595-603. doi: 10.1093/bioinformatics/btq713. Epub 2011 Jan 13

[144] J.O. Korbel, A. Abyzov, X.J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M.B. Gerstein. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 2009 Feb 23;10(2):R23. doi: 10.1186/gb-2009-10-2-r23.

[145] X.J. Mu, Z.J. Lu, Y. Kong, H.Y. Lam, and M.B. Gerstein. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Research*, 2011 Sep 1;39(16):7058-76. doi: 10.1093/nar/gkr342. Epub 2011 May 19.

[146] K.Y. Yip, C. Cheng, J.B. Bharadwaj, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 2012 Sep 26;13(9):R48. doi: 10.1186/gb-2012-13-9-r48.

[147] M.B. Gerstein, A. Kundaje, M. Hariharan, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012 Sep 6;489(7414):91-100. doi: 10.1038/nature11245.

[148] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.

[149] J. Rozowsky, A. Abyzov, J. Wang, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 2011 Aug 2;7:522. doi: 10.1038/msb.2011.54.

[150] E. Khurana, Y. Fu, J. Chen, and M.B. Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Computational Biology*, 2013;9(3):e1002886. doi: 10.1371/journal.pcbi.1002886. Epub 2013 Mar 7.

[151] Y. Xia, E.A. Franzosa, and M.B. Gerstein. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Computational Biology*, 2009 Jun;5(6):e1000413. doi: 10.1371/journal.pcbi.1000413. Epub 2009 Jun 12.

[152] Z.J. Lu, K.Y. Yip, G. Wang, et al. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Research,* 2011 Feb;21(2):276-85. doi: 10.1101/gr.110189.110. Epub 2010 Dec 22.

[153] S. Saha, S. Rajasekaran, and R. Ramprasad. Novel Randomized Feature Selection Algorithms. *International Journal of Foundations of Computer Science,* 26(3), 2015, 321-342.

[154] A. Abyzov, S. Li, and M.B. Gerstein, Understanding genome structural variations, *Oncotarget,* 7(7):7370-1, 2016, Feb. 16, 2016, doi: 10.18632/oncotarget.6485.

[155] P.H. Sudmant, T. Rausch, E.J. Gardner, et al., An integrated map of structural variation in 2,504 human genomes, *Nature,* Oct. 1, 2015, 526(7571):75-81, doi: 10.1038/nature15394.

[156] J. Chen, J. Rozowsky, T.R. Galeev, et al., A uniform survey of allele-specific binding and expression over 1000-Genomes-Project Individuals, *Nature Communications,* April 18, 2016, 7:11101, doi: 10.1038/ncomms11101.