# Using relational databases to analyze Microarray probes and single nucleotide Polymorphisms

Abhijit W. Phatak[1] and Eric C. Rouchka[1]
TR-ULBL-2007-04

December 14, 2007

[1]University of Louisville
Speed School of Engineering
Department of Computer Engineering and Computer Science
123 JB Speed Building
Louisville, Kentucky, USA 40292

Abhi2766@yahoo.com; eric.rouchka@louisville.edu

UNIVERSITY of LOUISVILLE dare to be great
Bioinformatics Laboratory Technical Report Series

*Bioinformatics Research*

# Using relational databases to analyze microarray probes and single nucleotide polymorphisms

Abhijit W. Phatak[1], and Eric C. Rouchka[1,*]

[1]Department of Computer Engineering and Computer Science, University of Louisville, 123 JB Speed Building, Louisville, KY, USA

UNIVERSITY OF LOUISVILLE BIOINFORMATICS LABORATORY TECHNICAL REPORT SERIES REPORT NUMBER **TR-ULBL-2007-04**

**ABSTRACT**

**Motivation:** Microarrays such as those from the Affymetrix® Inc. provide a very useful means of studying thousands of genes for DNA analysis and expression levels and are also valuable in the study of single nucleotide polymorphisms (**SNP**s). Aside from the physical use of microarrays for the assessment of gene expression levels, the data in them can be used in various research efforts. Our objective is primarily to study information from microarray experiments that is typically discarded during analyses of the results from such experiments for potentially useful answers to genetic research questions.

**Results:** We focused on creating a relational database of sequence alignment searches of probe data from an Affymetrix® microarray against sequence data from publicly available dbSNP database as well as setting the process of analyzing the results of these searches into motion.

## 1 INTRODUCTION

Single nucleotide polymorphisms (SNPs, often pronounced as *snips*) are the focus of an increasing number of research efforts due to the potential they have in providing clues to various diseases and abnormalities, including cystic fibrosis [1], sickle cell anemia [2], muscular dystrophy [3], Type II diabetes [4] and migraine headaches [5]. In sickle cell anemia, for example, the change of an adenosine base in a healthy individual's hemoglobin to a thymine base causes the inserted protein to become valine instead of glutamine (Figures 1 and 2). This results in the diseased individual's blood containing cells that look adopt a sickle shape, thus giving the disease its name.

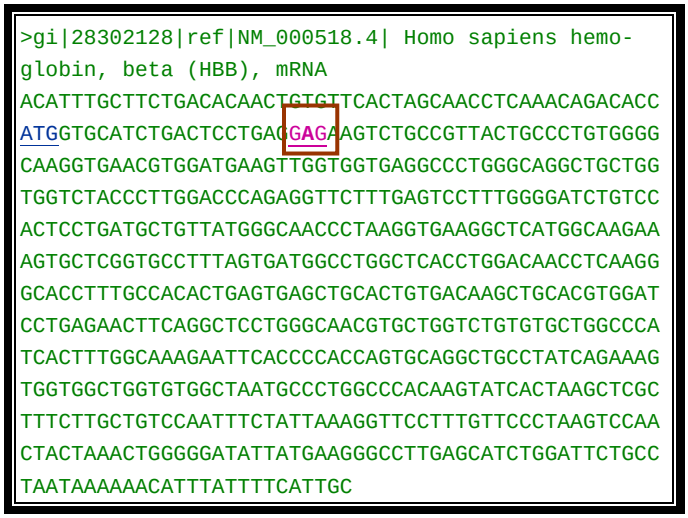


**Figure 1: Hemoglobin sequence showing the codon for adenosine in a healthy individual.**

Microarray technology has proven to be a potent tool in furthering a wide range of genetic research including SNPs. In particular, microarrays produced by Affymetrix® Inc. provide a very useful way of conducting research in SNPs. The typical Affymetrix® microarray consists of hundreds of thousands of probes representing thousands of genes. The probes are each just twenty-five bases long. Since a SNP relates to a variation in a single base within a segment of DNA, the relatively small size of the probe allows the researcher to focus his/her attention close to the locus of a known SNP.

Our research utilizes this specificity of probes from an Affymetrix® microarray to find highly similar sequence segments in the most authoritative public SNP database currently available, namely the dbSNP database [6] maintained

*To whom correspondence should be addressed.

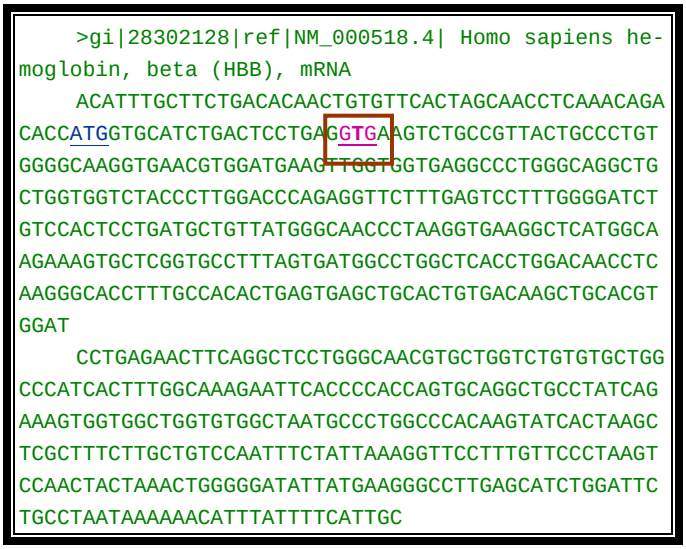


**Figure 2: Hemoglobin sequence showing the codon for thymine in place of adenosine for a diseased individual.**

by the National Center for Biotechnology Information (NCBI). The project focuses on searching human SNPs from dbSNP build 124 available from the NCBI website (`http://www.ncbi.nlm.nih.gov/projects/SNP/`).

This project is remarkable in view of the following facts. There is growing interest in the fields of SNP research and microarray research. It is also well understood that microarray technology provides a very useful means of studying SNPs. There has already been a considerable amount of research conducted in the area of the study of SNPs through the physical use of microarrays [7-9]. However, analyzing data across the two sources of SNP research, namely the SNP databases and the microarrays that provide one of the sources for creating these databases is not yet a widely used research technique. A quick search through PubMed [10;11] for the key terms, microarray, SNP and database yielded some interesting references [12-14]. However, these are still not as numerous as are reported in many other areas and techniques of genetic research. There is, therefore, scope for such projects, not because it is uncommon but because there is potential in this approach for answering some pertinent genetic research questions. This project attempts to lay the foundation for such studies by turning the comparative data generated through the use of tools such as BLAST [15] into a relational database.

This relational database will, it is hoped, be helpful in furthering the study of both sources in a better manner. Indeed, time constraints on this project have meant that this generated database is limited to a search of data only from the human genome. However, the methods and the knowledge gained from conducting this project will certainly help facilitate similar and more ambitious studies of data from other genomes as well.

## 2 METHODS

In essence, the steps involved in conducting the project consisted of the following:

- Acquiring and downloading the source data.
- Preprocessing source data.
- Running the BLAST alignment searches on processed source data.
- Parsing and storing search results in formats appropriate for the project.
- Translating the stored results into MySQL database tables.
- Analyzing the results.

### 2.1 Data Acquisition

The sources of the data were:

- Sequence data representing 247,965 twenty-five base oligomers from the HG-U133A microarray manufactured by Affymetrix® Inc.
- Referenced sequences for known SNPs in the human genome available in twenty-seven compressed FASTA files from the NCBI dbSNP database.
- Twenty-five compressed FASTA files for the human genome from the UCSC goldenpath assembly.

Microarrays from Affymetrix® typically consist of probes that are twenty-five bases long designed to be complementary in sequence to known genes. These probes will then hybridize with the complementary mRNA samples that are washed over the array to yield a measurement of the expression levels of the genes on the array. These microarrays also contain a paired set of 25-base oligomers identical to the first set except for the middle (thirteenth) base, which is a complement of the base in the first set. This paired set is intended as a control measure to trace possible cross-hybridization during the testing process. The actual sequence segment is loosely referred to as a ‘match’ and the other half of the pair with the complement of thirteenth base as a ‘mismatch’. Since these pairs of probes differ only in a known specific location, the mismatch half of each pair can be inferred from the match data. Thus, it was sufficient to use only the 247,965 oligomers representing the ‘match’ half of the microarray for this project. HG-U133A is a gene expression microarray representing nearly 20,000 well-documented genes from the human genome.

The data for the SNPs consists of the compressed FASTA format (Figure 3) files available for download from the NCBI dbSNP site. As noted earlier, the data used corresponds to Build 124 of the dbSNP database, which is itself based on Build 33 of the NCBI Human Genome build 33. The data consists of referenced sequences containing SNPs (*refSNPs*) that have been verified and consolidated from multiple submission sources. Thus, each sequence in this database is annotated as 'rsNNNNNN' where N represents the sequence number. This data is available, unlike for other organisms, in files comprising SNPs found on each of the 23 human chromosomes. Chromosome 23 is appropriately broken into two parts, X and Y. Additionally, a file consisting of sequences that are likely to occur across multiple chromosomes and a file containing sequences that have not yet been mapped to any specific chromosome. These dbSNP files contain a total of just over 10 million sequences containing an important allele each.

```
>gnl|dbSNP|rs17105379_allelePos=101totallen=
201|taxid=9606|snpClass=1|alleles='C/T'|mol=
genomic|build=123
AAAGGTCACA ATTTAAGCAC TAATTGCATA TAGTTTTTCT
TGACTTGGCA TTCAAGGGAT GGGAAAACTC AATAGAAGAC
TCTTGCAATA GCCCAGATAA
Y
GTGTAGATAG TTAGCAGAGG GAATGAACAG TAGTGAACAA
AACCCAAAGA CACATCACAG GCAAAAATCA ATTGGGTCTG
GAAATACATT TAAGTTATGG
```

**Figure 3: Sample sequence from dbSNP database in FASTA format.**

The other part of the project used data from the entire human genome for comparison against the Affymetrix® microarray probes. This data is also available in compressed FASTA format on the Kybrin Bioinformatics cluster at the University of Louisville with one file for each of the first twenty-two chromosomes and separate files for the X and Y chromosomes. The latest available version of the data dated July 2004, named *hg17*, was used.

### 2.2 Preprocessing

The source data mentioned above was preprocessed for running the focused alignment searches using WUBLAST on the FASTA sequences. The microarray data was in a tab-delimited text file (Figure 4) and had to be turned into the FASTA format using a simple Java program. The dbSNP and UCSC data was already in FASTA format but still needed to be transformed for the purpose of the project.

Sequences in the dbSNP database contain a single instance of an important SNP each, which is denoted by the standard IUPAC[16] code. The sequences typically range from a few hundred to a few thousand bases in length. However, the location of interest for the project was the segment surrounding the SNP for alignment with the twenty-five base oligomers from the microarray. As such, a Perl script was used to create a subset of forty-nine bases from each dnSNP sequence spanning twenty-four bases to the left and to the right of the locus of the allele in the majority of the cases (Figure 5). This meant that all possible alignments containing the SNP could be found when matched against the microarray probes. In cases where the allele position or the length of the original sequence did not leave twenty-four bases on either side of the allele, the resulting segment be-

| Probe Set Name | Probe X | Probe Y | Probe Interrogation Position | Probe Sequence Target | Strandedness |
|---|---|---|---|---|---|
| 1007_s_at | 467 | 181 | 3330 | CACCCAGCTGGTCCTGTGGATGGGA | Antisense |
| 1007_s_at | 531 | 299 | 3443 | GCCCCACTGGACAACACTGATTCCT | Antisense |
| 1007_s_at | 86 | 557 | 3512 | TGGACCCCACTGGCTGAGAATCTGG | Antisense |
| 1007_s_at | 365 | 115 | 3563 | AAATGTTTCCTTGTGCCTGCTCCTG | Antisense |
| 1007_s_at | 207 | 605 | 3570 | TCCTTGTGCCTGCTCCTGTACTTGT | Antisense |
| 1007_s_at | 593 | 599 | 3576 | TGCCTGCTCCTGTACTTGTCCTCAG | Antisense |
| 1007_s_at | 425 | 607 | 3583 | TCCTGTACTTGTCCTCAGCTTGGGC | Antisense |
| 1007_s_at | 552 | 101 | 3589 | ACTTGTCCTCAGCTTGGGCTTCTTC | Antisense |
| 1007_s_at | 680 | 607 | 3615 | TCCTCCATCACCTGAAACACTGGAC | Antisense |
| 1007_s_at | 532 | 139 | 3713 | AAGCCTATACGTTTCTGTGGAGTAA | Antisense |
| 1007_s_at | 143 | 709 | 3786 | TTGGACATCTCTAGTGTAGCTGCCA | Antisense |
| 1007_s_at | 285 | 623 | 3793 | TCTCTAGTGTAGCTGCCACATTGAT | Antisense |
| 1007_s_at | 383 | 479 | 3799 | GTGTAGCTGCCACATTGATTTTTCT | Antisense |

**Figure 4: Sample source data from HG-U133A microarray in tab-delimited format.**

came less than forty-nine bases long. So, for instance, if a dbSNP sequence was 101 bases long and the allele was noted to occur at base 95 in the sequence, the truncated sequence of interest was taken at twenty-four bases before the SNP and the remaining 6 bases in the sequence after the SNP position. The truncated sequence would therefore comprise the 31 base segment starting from base 71 and ending in base 101. Using such shortened segments from the original data greatly reduced the complexity of the project while enabling the research to focus on its essential purpose.Additionally, the original dbSNP sequences are soft-masked for low-complexity regions and tandem repeats of low significance by using lowercase symbols for the nucleotides. Although masking of this sort is often advisable for removing known non-informative regions and focusing on regions with potentially more information, it was necessary to unmask the source data to accurately align all SNPs. Hence, the truncated segments from the original dbSNP data were restored to uppercase symbols so they would not be ignored in the BLAST searches.

### 2.3 Running BLAST

The tool of choice for generating the alignments between the microarray oligomers and the dbSNP sequence segments was BLAST. Over the course of many years, BLAST (Basic Local Alignment Search Tool) has proven to be the most popular sequence-matching tool available on account of its robustness and efficiency. BLAST was originally developed by the NCBI [17] and remains one of the most frequently used tools for sequence alignments. However, the version of BLAST developed at the Washington University at St. Louis, MO, popularly known as WUBLAST is equally popular and offers many enhancements over the NCBI version. In fact, version 2.0 of WUBLAST contains many improvements over its own previous versions (http://blast.wustl.edu/blast/cparms.html).

NCBI BLAST and WUBLAST have multiple variants geared towards finding alignments for different purposes. The entire set of BLAST tools from the Washington University is referred to as BLASTA, of which BLASTN is the appropriate program for the current project.

After creation of the reduced dbSNP dataset it was formatted into a BLAST database using the xdformat utility, which is part of the BLASTA package. This step is necessary since both NCBI BLAST (which uses formatdb in place of xdformat) and WUBLAST create their own formats for the target database for improved efficiency.

The next important issue was deciding on the parameters to use for running the BLASTN searches. To create a database of alignments between the microarray probes and the two other databases for analysis the need was to find matches that were identical to the probes or very nearly so.

By default, the program looks for gapped as well as ungapped alignments. In a comprehensive exercise, these alignments would include gapped as well as ungapped alignments. Since this is a new research, the scope of this project was intentionally limited to ungapped alignments. The expanded search for gapped alignments was deferred to future efforts that can be based on this initial research. The other default parameters for BLASTN include a word size of 11 for nucleotide sequences and a scoring scheme where matching base pairs get assigned a score of 5, mismatches -4 and gap opening and extending penalties are both set to 10. Using this scoring scheme, a perfectly matching alignment of 25 base pairs would score 125, an alignment of 25 base pairs with two mismatches would score 107 and an alignment of 25 base pairs with two gaps would score 95.

The word size is used as a seed to look for matches against the database. When a match is found the algorithm works in a greedy fashion by trying to extend the matched word on either side until the overall score of the alignment drops below a threshold value. A smaller word size is expected to increase the sensitivity, yielding potentially more alignments than the default while a larger word size is expected to reduce sensitivity. However, the gains in sensitivity with a smaller word size are most likely to be achieved at a cost of longer search time in databases of any considerable length.

There are a number of other parameters that can be set to fine tune the search process. A complete list is available at the WUBLAST website (http://blast.wustl.edu/blast/parameters.html) Experimentation with with different word sizes and score thresholds determined that a word size of 8 proved to be appropriate for the desired output. The word size of 8 would allow for alignments with up to two mismatches on the maximum of twenty-five base-pair alignments that would be obtained by running the searches of the microarray probes against the dbSNP and the genome databases.

### 2.4 Parsing and storing the results

BLAST searches were conducted chromosome-wise, keeping the structure of the source data intact. However, the output from BLAST searches contains much more information than was practically required for creating the database tables. As such, the normal BLAST output was piped to Perl scripts to filter out the basic statistical information required for a database table. BPlite [18] is a Perl module, originally developed by Ian Korf, to parse and filter the output from all types of BLAST searches.

Another Perl script that used BPlite was written to further parse and filter the output to only store alignments of 22 or more base pairs and only information about these alignments sufficient for the purpose of this research. This included the following:

- Reference numbers of the query and target sequences.
- Locations of sequences on the microarray and on the dbSNP and genomic databases.
- Location of the SNP within a dbSNP segment.
- Lengths of the query and target sequences.
- Start and end positions of the alignments found.
- Aligned segment pairs.
- The alignment string itself (figure 5), which is a means of depicting the matched and mismatched base pairs within a sequence. Matching pairs have a '|' between them, mismatches remain blank and where a base in the query sequence matches any one of the possible variations of the SNP, a plus sign ('+') is used to show this 'partial' match.
- Length of the alignment found.
- Number of matches within the sequence.
- Percent identity of the matches (number of matches divided by the alignment length).
- Raw score of the alignment from standard scoring scheme of WUBLAST.

The parsed output was maintained in plain text files on the Kybrin cluster hard drives with one file corresponding to each input file.

```
1520637 1 22 23 25 49 0 0

QUERY:
Probe_Set_Name:201060_x_at|Probe_X:71|Probe_Y
:35|Probe_Pos:1596|Strand:Antisense
SBJCT:
gnl|dbSNP|rs10839991_allelePos=298totallen=699|taxid=9606|snp
Class=1|alleles='

Score:  103 Bits:30.90 PctID:88.00 Length: 25
Q:   25 AGTTCGTGACTAGCCTGGCCAACAT          1
     25 |||||| ||| ||||||||||||+|         25
S:    2 AGTTCGAGACCAGCCTGGCCAACRT          26
1520637 1 22 23 25 49 0 0
```

**Figure 5: Sample of the parsed output from BLAST for a dbSNP sequence with a '+' at SNP position.**

### 2.5 Creating the MySQL database tables

From the parsed output it was a fairly straightforward, if tedious task to further split the lines of each alignment into formats suitable for a MySQL database and conduct the transfer of data from the text files to the database. In this effort the Active State version of Perl and the DBI and DBD modules that help establish the connection between the files and the tables were used.

A new database was created in MySQL with a schema that had master tables for the source of the database sequences such as the dbSNP database and the UCSC genome database (Figure 6). Another master table was made to store the name of the organism involved in the searches. These measures were done in view of the likely extension of the scope of the research to other databases and organisms in the future. The identification numbers for the source databases and organisms were used as foreign keys in the tables that were created to hold the results from the BLAST searches.

The database tables' schema followed the information obtained from the alignments. Some additional fields were added by calculating values from other fields (Figures 7-9). For instance, a field was added to store the position of an SNP in the alignment itself from the knowledge of its position on the original string. The alignment string itself was also stored besides the aligned sequence segments for easy searching. The empty space that exists in the original string from the BLAST output representing a mismatch between a base pair was replaced with the character '*m*' for easier viewing and searching (Figure 9).

One table holds the entire data for the alignments obtained from matching the microarray probes to dbSNP sequences. The fields in this table are as shown in Figures 7-9.

The results from the searches involving the probes and the genome database sequences were stored in three tables with an identical schema mainly for convenience. The schema was similar to that for the dbSNP results except that specific fields for the allele (SNP) positions and the subject lengths were not relevant for these tables. One table holds all alignments of exact matches between the query and target sequences. The other two tables hold the less than exact matches. One table has data from chromosome 1 to 10 and the other from chromosomes 11 through X and Y. The reason for this division is that a combined table for all chromosomes would have required storage space of over 2 GB. Tables larger than that size need to be treated differently by the database program and require slightly different handling than regular tables. So, the entire dataset was divided into two parts to avoid the complexity that would have been introduced by one table of size larger than 2 GB. When required, using the appropriate join syntax in queries and stored procedures can accumulate data from both tables.

All tables have an auto generated primary key field since it is possible that the identifiers for the query and the target sequences, may be repeated in the table due to multiple alignments found for the same sequence.

The table for the dbSNP search results occupies 1.3 GB for its over 6 million records. If required, this can be broken down into small tables as was done for the results from the genomic sequence searches if handling the entire data in one table proves to be unwieldy or overly time-consuming.

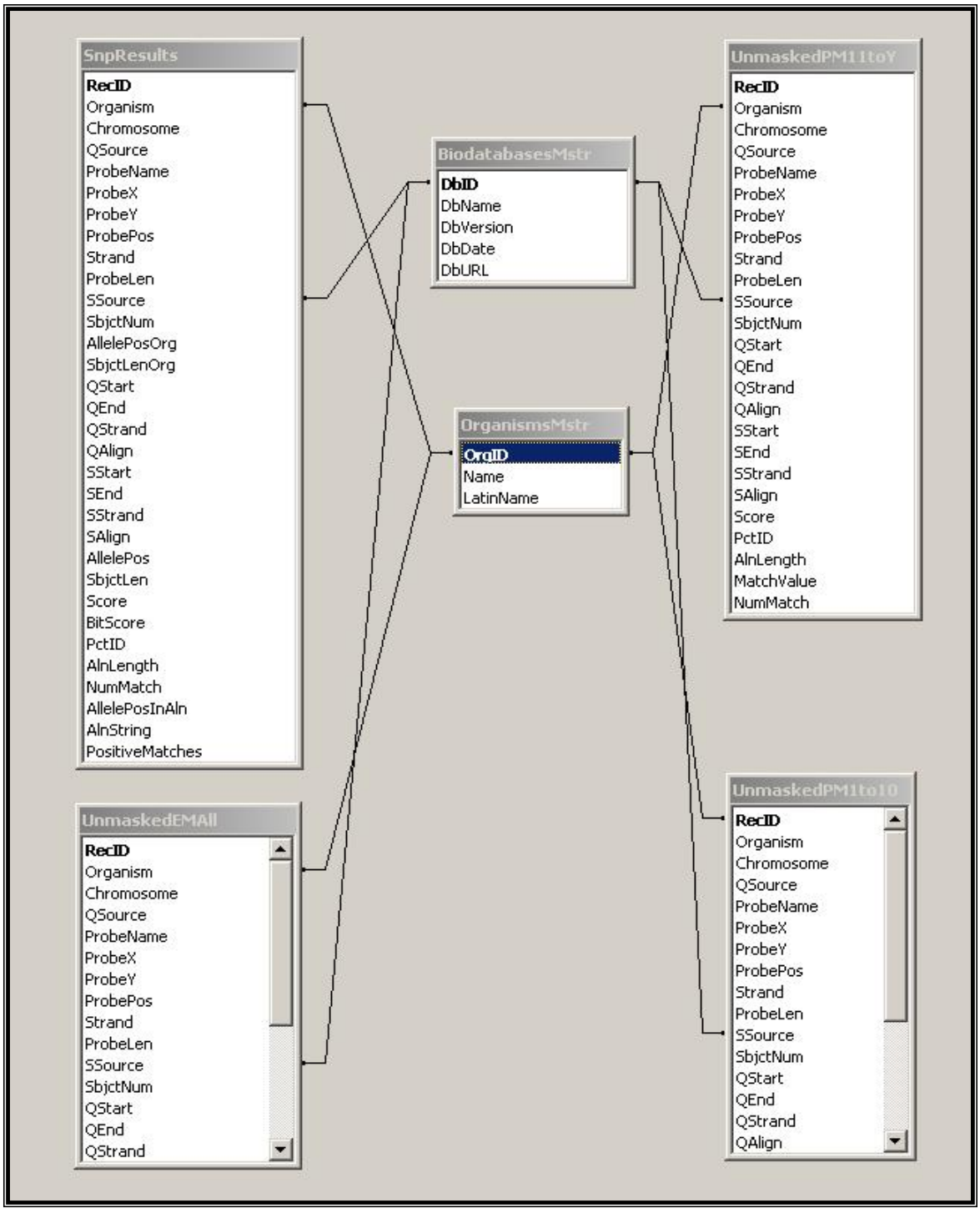


**Figure 6: Database schema for the search results.** Foreign keys for source database and organism reference the master tables for the values in those fields.

| RecID | Organism | Chromosome | QSource | ProbeName | ProbeX | ProbeY | ProbePos | Strand | ProbeLen |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | Affy_HG-U133A | 117_at | 510 | 685 | 1691 | Antisense | 25 |
| 2 | 1 | 1 | Affy_HG-U133A | 117_at | 222 | 111 | 1697 | Antisense | 25 |
| 3 | 1 | 1 | Affy_HG-U133A | 117_at | 370 | 699 | 1703 | Antisense | 25 |
| 4 | 1 | 1 | Affy_HG-U133A | 117_at | 185 | 581 | 1727 | Antisense | 25 |
| 5 | 1 | 1 | Affy_HG-U133A | 117_at | 185 | 581 | 1727 | Antisense | 25 |
| 6 | 1 | 1 | Affy_HG-U133A | 117_at | 539 | 111 | 1733 | Antisense | 25 |
| 7 | 1 | 1 | Affy_HG-U133A | 117_at | 452 | 459 | 1739 | Antisense | 25 |
| 8 | 1 | 1 | Affy_HG-U133A | 117_at | 341 | 501 | 1871 | Antisense | 25 |
| 9 | 1 | 1 | Affy_HG-U133A | 117_at | 341 | 501 | 1871 | Antisense | 25 |
| 10 | 1 | 1 | Affy_HG-U133A | 117_at | 341 | 501 | 1871 | Antisense | 25 |
| 11 | 1 | 1 | Affy_HG-U133A | 117_at | 278 | 325 | 1877 | Antisense | 25 |
| 12 | 1 | 1 | Affy_HG-U133A | 117_at | 278 | 325 | 1877 | Antisense | 25 |
| 13 | 1 | 1 | Affy_HG-U133A | 117_at | 278 | 325 | 1877 | Antisense | 25 |

**Figure 7: Snapshot of records in the table containing alignment data for the dbSNP sequences.**

| ProbeLen | SSource | SbjctNum | AllelePosOrg | SbjctLenOrg | QStart | QEnd | QStrand | QAlign |
|---|---|---|---|---|---|---|---|---|
| 25 | 2 | rs753856 | 73 | 777 | 25 | 1 | minus | TCCTCTTCGGGAATCTTGTCCCTAA |
| 25 | 2 | rs753856 | 73 | 777 | 25 | 1 | minus | CGCCTGTCCTCTTCGGGAATCTTGT |
| 25 | 2 | rs452004 | 201 | 401 | 25 | 1 | minus | ATTTTGCGCCTGTCCTCTTCGGGAA |
| 25 | 2 | rs452004 | 201 | 401 | 24 | 1 | minus | GGACTTCCCGACACTTGTCTTGCA |
| 25 | 2 | rs368844 | 29 | 610 | 25 | 1 | minus | AGGACTTCCCGACACTTGTCTTGCA |
| 25 | 2 | rs368844 | 29 | 610 | 25 | 1 | minus | CAGGCAAGGACTTCCCGACACTTGT |
| 25 | 2 | rs368844 | 29 | 610 | 25 | 1 | minus | TCCAGCCAGGCAAGGACTTCCCGAC |
| 25 | 2 | rs439078 | 201 | 401 | 25 | 1 | minus | GCTTGAGTGCCACAACTGCTGCCCC |
| 25 | 2 | rs394965 | 201 | 401 | 4 | 25 | plus | GCAGCAGTTGTGGCACTCAAGC |
| 25 | 2 | rs391125 | 201 | 401 | 1 | 25 | plus | GGGGCAGCAGTTGTGGCACTCAAGC |
| 25 | 2 | rs394965 | 201 | 401 | 1 | 25 | plus | GCAGTTGTGGCACTCAAGCCCGCCA |
| 25 | 2 | rs439078 | 201 | 401 | 25 | 1 | minus | TGGCGGGCTTGAGTGCCACAACTGC |
| 25 | 2 | rs391125 | 201 | 401 | 1 | 25 | plus | GCAGTTGTGGCACTCAAGCCCGCCA |
| 25 | 2 | rs394965 | 201 | 401 | 1 | 25 | plus | GTGGCACTCAAGCCCGCCAGGGGGA |

**Figure 8: Fields showing database and query information for each alignment.**

| Score | BitScore | PctID | AlnLength | NumMatch | AllelePosInAln | AlnString |
|---|---|---|---|---|---|---|
| 121 | 35.9 | 96 | 25 | 24 | 18 | |
| 121 | 35.9 | 96 | 25 | 24 | 24 | |
| 121 | 35.9 | 96 | 25 | 24 | 2 | |
| 120 | 35.6 | 100 | 24 | 24 | 25 | |
| 121 | 35.9 | 96 | 25 | 24 | 10 | |
| 121 | 35.9 | 96 | 25 | 24 | 16 | |
| 121 | 35.9 | 96 | 25 | 24 | 22 | |
| 121 | 35.9 | 96 | 25 | 24 | 11 | |
| 92 | 27.9 | 90.9 | 22 | 20 | 25 | |
| 112 | 33.4 | 92 | 25 | 23 | 10 | |
| 103 | 30.9 | 88 | 25 | 22 | 22 | |
| 121 | 35.9 | 96 | 25 | 24 | 17 | |
| 103 | 30.9 | 88 | 25 | 22 | 12 | |
| 103 | 30.9 | 88 | 25 | 22 | 16 | |
| 121 | 35.9 | 96 | 25 | 24 | 23 | |

**Figure 9: Remaining fields of dbSNP results table showing the statistics and alignment string for each alignment.**

## 3 RESULTS AND ANALYSIS

Since this is a first effort at searching such large volumes of input data, reliable estimates of the amount of time and resources that would be required for these searches were not available. As such the search for ungapped alignments was taken up as a starting point. With estimates of time, computing resources and storage requirements from this process some basis for projecting the same could be estimated when looking for the expanded result sets that would include alignments with gaps as well as those without them. Of course, it must be borne in mind that being a new research process, some amount of time and effort that was spent in coming to terms with the various issues involved in the implementation of the research will, hopefully serve as a capital investment for future projects that build upon this one. Too, the over six million and over fifteen million ungapped alignments found between the microarray probes and the SNP segments and between the probes and the genomic sequences respectively should provide a strong base for beginning the analysis even as the search for a wider result set including gapped alignments is undertaken in future projects.

The major objective of this exercise in creating databases of alignments between microarray probe data and other databases is to provide the basis for conducting deeper analyses of microarray assays than those that focus mainly on gathering information from the intensity of the hybridizations. In cases such as that of Affymetrix® microarrays, with their short, 25-mer match/mismatch structure, when physical experiments are analyzed, the entire result sets showing hybridization with the mismatch probes is completely discarded. A major hypothesis of this project is that it is likely that in at least a few cases, this approach amounts to throwing the baby out with the bathwater, so to speak. So, a prominent question a project of this nature seeks to answer is ‘what if there is valuable information in what is thrown away?’ Since a SNP is just that, a variation in a single base among different genotypes, it is worth mining the habitually discarded information and studying it more closely to find out if any of that is not just a case of random cross hybridization but is actually a site that contains a real SNP.

One possible offshoot of such research is gaining the detailed information necessary to make possible projections about the genotype of the test subjects. If samples from a particular individual or race show a sustained tendency to hybridize differently than the expected norm, it would be worth investigating whether the hybridization information is more than a testing issue and whether it actually reveals characteristics about that individual, that race or perhaps a disease that individual or race may be prone to. For example, individuals can differ in their zygosity on the same genes. A person may be homozygous for the gene but that may indicate presence of a disease while another person with the other variation of the allele may have no signs of the disease.

Another outcome of analysis of microarray data using the relational databases that may help to potentially map patterns of unusual hybridizations to certain diseases or genotypes is the design of more specialized arrays. These arrays would contain probes that hybridize with certain sequences known to be involved in diseases or abnormal mutations and thus furthering the possibility of what is referred to as ‘personalized medicine’ – the ability to find a remedy for an individual problem instead of a more generalized solution.

As much as the search for SNPs that may be the cause of particular diseases and abnormalities is like searching for a needle in a haystack, it is still a worthwhile effort when even one such variation is finally identified positively.

With these possibilities in mind, a result set that consisted of a little over six million ungapped alignments of twenty-two base pairs or more between the microarray and the dbSNP database was found. Some other interesting figures are shown in Table I. In particular, the 1,656 alignments with a mismatch only at the thirteenth base are of interest. Although it is not a large number, that is in line with expectations. It must be noted that several thousand sequences from the human repeat regions of the genome having annotations starting with 'Affx-hum' were not considered in these totals since they are known to be of little value for the purpose of this particular analysis.

**Table I: Summary statistics from database tables.**

| | |
|---|---|
| SNP alignments | **Over 6,000,000** |
| Perfect matches | **45,984** |
| Mismatch in 13th base | **1,656** |
| Mismatch with allele position | **58,505** |
| Genome sequences hits | **Over 15,000,000** |

An analysis of a little over 1000 alignments covering the entire length of the 25-mer probe in each case where the only mismatch is at the thirteenth base showed some interesting results. Bearing in mind that Affymetrix® microarrays have pairs of probes where one half of the pair has the complement of the other's middle (thirteenth) base, a reason to study such a set of results is to find out the relative hybridizations between the match and mismatch pairs. If a higher percentage of hybridizations with the mismatch probes compared to the match probes appears to consistently occur that may raise questions as to whether those results are merely caused by random testing error or if they are indicative of possible variations or markers involved in a disease or harmful mutation.

Accordingly the test results of 14 experiments with 3 replicates each involving 1114 of these alignment probes with a mismatch at base 13 were plotted by taking the log-odds ratio of the matches to the mismatches (Figure 10). The data used was actual sample data for past experiments involving those probes made available by Affymetrix® Inc, the producer of the microarray. The scatter diagram for these experiments shows a fairly high percentage of values at or below 0, indicating the region where the mismatches outweigh the matches. Another plot of these probes (Figure 11) representing the frequency of matches greater than mismatches and vice versa also shows an interesting pattern. Both graphs intersect smoothly and 263 of 1114 probes consistently have the mismatches higher than the matches.

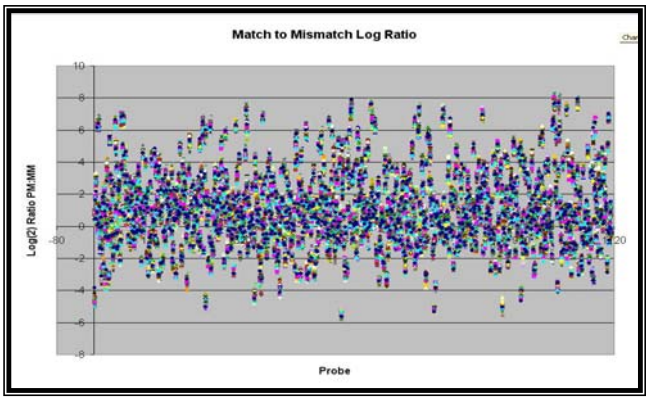


**Figure 10: Chart showing plot of log-odds ratio of perfect matches to mismatches for 1114 query probes involved in alignments with 1 mismatch at 13th base (14 exp., 3 replicates each)**

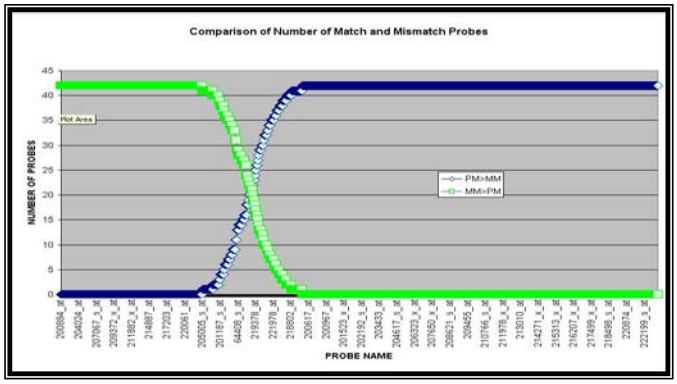


**Figure 11: Line chart showing number of times matches greater than mismatches and vice-versa for same experiments as for Figure 10.**

Although not high, it is still significantly more than expected. Also, a greater number of values would have been expected in the middle and with a greater scatter than is seen. As a first step in beginning to use the information from the database tables it is certainly encouraging, in a sense, to come across such results that seem to very from the norm. Although far from being conclusive, they do evoke curiosity and the desire to dig deeper into the maze for more answers.

As a test of these seemingly unusual results, a test of 765 alignments where the alignments were still 25 bases long and containing only one mismatch but where the mismatch was noted to be at positions other than the thirteenth base was also run. Figures 12 and 13 show the plots for this second set of tests. This time, the scatter of the log-odds values on the mismatch side i.e. below zero, is much less than in the previous set. However, the second plot shows a very even distribution between matches and mismatches,

when the matches would have been expected to outnumber the mismatches more definitely.

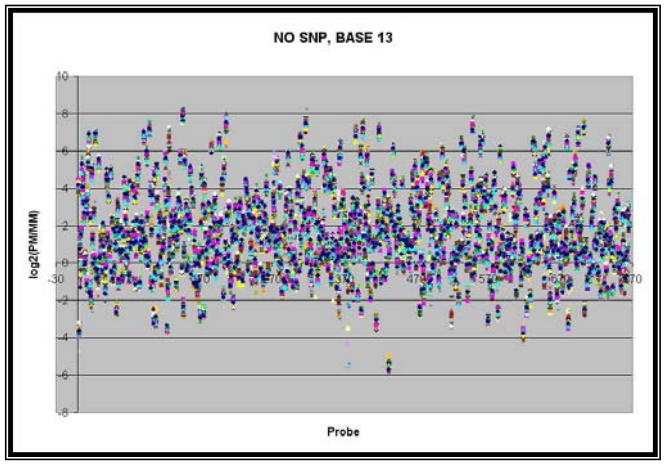


**Figure 12:** ***Plot similar to Fig. 10, of values for 765 probes with alignments having 1 mismatch, but not at 13th base.***

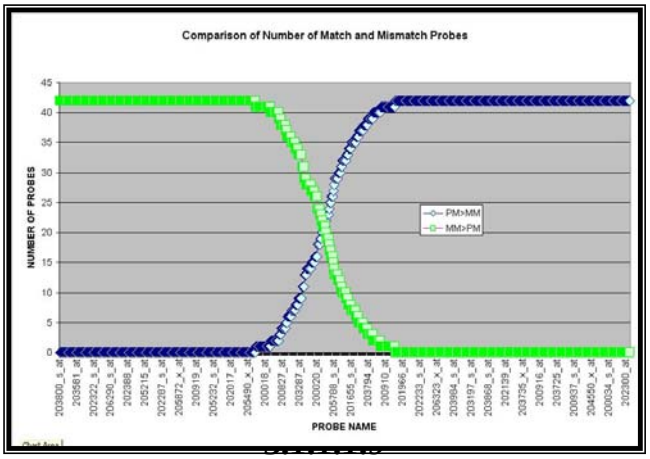


**Figure 13: Line graph similar to Figure 11 for second dataset of 765 probes.**

As noted before, although such results are very far from being conclusive of any kind of significant deviation from the normal they do highlight the value of generating the kind of information this project was intended for.

## 4 DISCUSSION

The project provided the author with an opportunity to delve a little deeper into the areas of microarray analysis and SNPs. Although the time frame for the project was fairly small, the author did get to understand a great many of the issues involved in efforts on those aspects of genetic research.

Perhaps the greatest concern in such a project is the handling of large volumes of data in as efficient a manner as possible. This is not a new realization in terms of genetic research but is a significant issue for this kind of a project. With a longer timeframe, there will be better scope to test alternate tools and methods for better computational performance as well as for finding more focused results.

The section on running the BLAST searches notes some of the important obstacles that had to be overcome in the conduct of the project. The recording of those hurdles and the measures taken to overcome them will, it is hoped, prove useful in making future efforts less laborious, less time-consuming and easier to manage.

Some future objectives include testing data from microarrays from sources other than Affymetrix® such as from Agilent® Technologies, another major producer of microarrays. Agilent's microarrays typically contain sixty-base probes as opposed to the shorter, twenty-five base ones from Affymetrix®.

Also, expanding the current database to include alignments with gaps is a logical step in future projects. SNPs can be related to insertions/deletions as well as alternate bases at a locus. Alignments with gaps will help to study such variations better.

Other future goals, as suggested earlier include creating such databases for data from other species besides humans. This will allow comparisons across organisms for specific mutations and gene expression levels.

Despite the limited scope of the project though, the creation of the database is an important first step towards conducting larger and more in-depth research in future.

## ACKNOWLEDGEMENTS

ER is supported by NIH-NCRR grant P20RR16481 and NIH-NIEHS grant P30ES014443. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH. AWP would also like to acknowledge Nathan Johnson, I. Elizabeth Cha, and Tim Hardin for their gracious assistance.

## REFERENCES

[1] E. Mateu, F. Calafell, O. Lao, B. Bonne-Tamir, J. R. Kidd, A. Pakstis, K. K. Kidd, and J. Bertranpetit, "Worldwide genetic analysis of the CFTR region," *Am. J. Hum. Genet.*, vol. 68, no. 1, pp. 103-117, Jan.2001.

[2] K. Y. Chang JC, "Antenatal diagnosis of sickle cell anaemia by direct analysis of the sickle mutation," *Lancet*, 2005.

[3] M. Koenig, A. H. Beggs, M. Moyer, S. Scherpf, K. Heindrich, T. Bettecken, G. Meng, C. R. Muller, M. Lindlof, H. Kaariainen, A. Delachapelle, A. Kiuru, M. L. Savontaus, H.

Gilgenkrantz, D. Recan, J. Chelly, J. C. Kaplan, A. E. Covone, N. Archidiacono, G. Romeo, S. Liechtigallati, V. Schneider, S. Braga, H. Moser, B. T. Darras, P. Murphy, U. Francke, J. D. Chen, G. Morgan, M. Denton, C. R. Greenberg, K. Wrogemann, L. A. J. Blonden, H. M. B. Vanpaassen, G. J. B. Vanommen, and L. M. Kunkel, "The Molecular-Basis for Duchenne Versus Becker Muscular-Dystrophy - Correlation of Severity with Type of Deletion," *American Journal of Human Genetics*, vol. 45, no. 4, pp. 498-506, Oct.1989.

[4] N. Vionnet, M. Stoffel, J. Takeda, K. Yasuda, G. I. Bell, H. Zouali, S. Lesage, G. Velho, F. Iris, P. Passa, P. Froguel, and D. Cohen, "Nonsense Mutation in the Glucokinase Gene Causes Early-Onset Non-Insulin-Dependent Diabetes-Mellitus," *Nature*, vol. 356, no. 6371, pp. 721-722, Apr.1992.

[5] M. Wessman, M. Kallela, M. A. Kaunisto, P. Marttila, E. Sobel, J. Hartiala, G. Oswell, S. M. Leal, J. C. Papp, E. Hamalainen, P. Broas, G. Joslyn, I. Hovatta, T. Hiekkalinna, J. Kaprio, J. Ott, R. M. Cantor, J. A. Zwart, M. Ilmavirta, H. Havanka, M. Farkkila, L. Peltonen, and A. Palotie, "A susceptibility locus for migraine with aura, on chromosome 4q24," *American Journal of Human Genetics*, vol. 70, no. 3, pp. 652-662, Mar.2002.

[6] S. T. Sherry, M. Ward, and K. Sirotkin, "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation," *Genome Res.*, vol. 9, no. 8, pp. 677-679, Aug.1999.

[7] M. Wirtenberger, K. Hemminki, B. Chen, and B. Burwinkel, "SNP microarray analysis for genome-wide detection of crossover regions," *Hum. Genet.*, June2005.

[8] Ann-Christine Syvänen, "Toward genome-wide SNP genotyping,", 37 ed 2005, p. S5-S10.

[9] Liu S, Li Y, and et al, "Analysis of the factors affecting the accuracy of detection for single base alterations by oligonucleotide microarray,", 37 ed 2005, pp. 71-77.

[10] R. J. Roberts, "PubMed Central: The GenBank of the published literature," *Proc. Natl. Acad. Sci. U. S. A*, vol. 98, no. 2, pp. 381-382, Jan.2001.

[11] "PubMed Central,".

[12] S. J. Tebbutt, I. V. Opushnyev, B. W. Tripp, A. M. Kassamali, W. L. Alexander, and M. I. Andersen, "SNP Chart: an integrated platform for visualization and interpretation of microarray genotyping data," *Bioinformatics.*, vol. 21, no. 1, pp. 124-127, Jan.2005.

[13] Pérez-Encisoa M, "In silico study of transcriptome genetic variation in outbred populations,", 166 ed 2004, p. 554.

[14] Itoshi Nikaido and et al, "EICO (Expression-based Imprint Candidate Organizer): finding disease-related imprinted genes,", 32 ed 2004, p. D548-D551.

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403-410, Oct.1990.

[16] [Anon], "Iupac-Iub Commission on Biochemical Nomenclature (Cbn) - Abbreviations and Symbols for Nucleic Acids, Polynucleotides and Their Constituents," *Virology*, vol. 45, no. 1, p. 326-&, 1971.

[17] W. G. W. M. E. W. M. a. D. J. L. Stephen F.Altschul, "Basic Local Alignment Search Tool," 2005, pp. 403-410.

[18] I. Korf, "BPLite," 1999.