

High-Throughput Identification, Database Storage and Analysis of SNPs in EST Sequences

Francisco José Useche^{1,2}

useche@caps1.udel.edu

Guang Gao^{1,2}

ggao@caps1.udel.edu

Mike Hanafey³

j-antoni.rafalski@USA.dupont.com

Antoni Rafalski³

mike.hanafey@USA.dupont.com

- ¹ Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, DE 19711, USA
- ² Department of Electrical and Computer Engineering, University of Delaware, Evans Hall, Newark, DE 19711
- ³ DuPont Crop Genetics, Delaware Technology Park, 1 Innovation Way, Newark, DE 19711, USA

Abstract

Single nucleotide polymorphisms (SNPs) are the most frequent form of DNA variation and disease-causing mutations in many genes. Due to their abundance and slow mutation rate within generations, they are thought to be the next generation of genetic markers that can be used in a myriad of important biological, genetic, pharmacological, and medical applications [13, 3, 19, 18, 16, 14]. There are several strategies both experimental, and *in-silico* for SNP discovery and mapping. Experimental SNP discovery consists of a number of labourious steps that make this process complex and expensive. *In-silico* discovery has been proposed as an alternative discovery method that makes use and takes advantage of large data sets with potential SNP information that have been generated with other purposes and have not been used as a SNP information source yet. However, in order to successfully apply the *in-silico* method to large data sets, the following challenges need to be addressed: First it is necessary to build an integrated SNP pipeline that handles data processing steps smoothly from the beginning (collecting sequence information) to end (SNPs in the database). Also, SNP detection tool parameters have to be optimized to satisfy specific goals of the project. Finally, SNP data could not be fully used until the *in-silico* method is validated experimentally. In this paper we present a design and implementation of an *in-silico* SNP detection software pipeline that exploits the existence of large EST (expressed sequence tag) data sets and effectively addresses the above challenges. First, the pipeline allows for smooth data transition between its different components by implementing data interfaces that translate the data formats of the different tools in the different stages. Second, we optimized PolyBayes parameters for SNP detection in maize EST. Finally, we implemented a user interface that along with the database structure created allows the scientist to perform preliminary analysis of the data and to perform basic statistics on the SNP data prior to experimental validation. The pipeline works with two different types of sequence assemblers (PHRAP [20] and CAT from DoubleTwist [21]). It uses a Bayesian engine for SNP detection (PolyBayes), selects relevant polymorphism information which is then uploaded into a database. We detected 2439 SNPs and 822 insertion deletions (INDELs) with a PolyBayes probability higher than 0.99 on the public set of 68,000 maize ESTs. The user interface allowed us analyzing the polymorphism information right after discovery in several ways that allowed us to gain insight into the distribution and significance of the newly acquired data.

Keywords: SNP discovery, ESTs, database, point mutation

1 Introduction

Single nucleotide polymorphisms (SNPs) and insertion/deletions (INDELs) are the two types of sequence differences between individuals. A SNP is a single nucleotide change in a molecule of DNA, as opposed to a change of multiple bases at the same time.

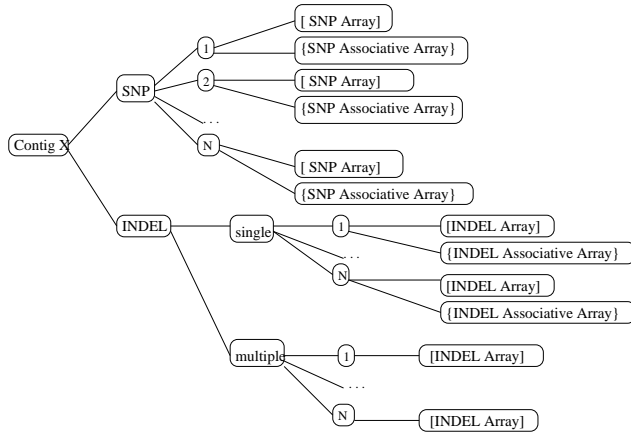


Figure 1: Report data structure.

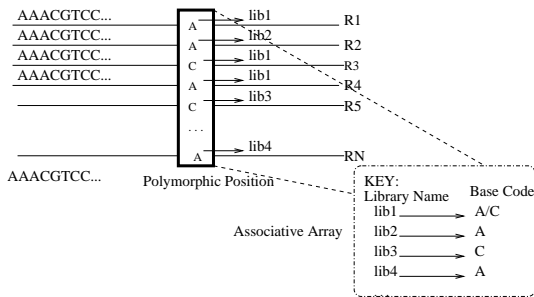


Figure 3: Associative array data structure.

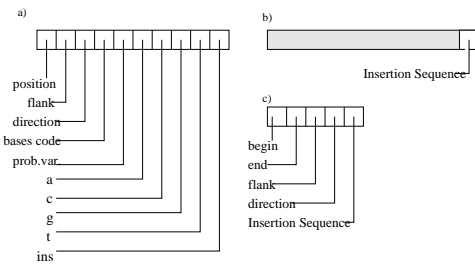


Figure 2: Array data structure: a) SNPs, b) Single indel, c) Multiple indel.

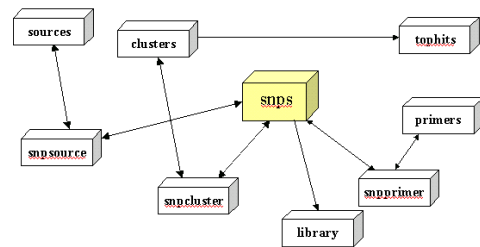


Figure 4: Main database structure.

A rough draft of the human genome was presented in the summer of the year 2000 [8]. The completion of the human genome opened up the race to sequence more genomes, including model organisms, farm animals, and crop plants. Moreover, this important step and improvements in speed and price of sequencing technology [12], increased interest in resequencing or comparative sequencing: the search for genetic differences between individuals, that would ultimately relate an individual's phenotype with his genotype.

There are two main methods to detect SNPs: experimental [4, 17, 5, 7, 9], and *in-silico* [10, 11, 1]. The experimental method in general refers to the polymorphism screening by DNA sequencing. There are plenty of experimental techniques to discover and detect SNPs. This experimental techniques must take into account the possibility of sequencing errors in the identification of polymorphisms. The *in-silico* method refers to the polymorphism screening done by computer analysis in sequences from different individuals. This method attempts to discern between true polymorphisms and sequencing errors by establishing a likelihood of a particular locus being polymorphic.

We built an *in-silico* SNP discovery pipeline that takes advantage of existing EST data, with the following benefits: lower costs, shorter time required for the SNP discovery process and no need for specialized equipment once the primary sequence data are available. The pipeline is high-throughput oriented, and capable of analyzing for SNPs large numbers of sequences.

First, in the pipeline we implemented two data interfaces that allow for smooth data transition. The first interface adapts the data output of the assembly tools to meet the data input requirements of PolyBayes. For PHRAP the interface uses the ace-file data input mode and for CAT the interface uses the anchored multiple sequence alignment input mode. The second data interface parses PolyBayes

output and builds data structures that are uploaded into a database. Secondly, in order to fine tune the pipeline for SNP detection in maize ESTs, we optimized PolyBayes detection parameters by using a test data set of several hundred contigs in which SNPs were visually identified by an experienced sequence analyst. We then maximized the agreement between the visually identified SNPs and the SNPs detected by the pipeline. Finally, via the user interface at “the end” of the pipeline and the information present in the database we managed to make important observations like SNP frequency, distribution of SNPs with respect to mutation type, alignment depth and minor allele occurrences, etc.

To summarize, we used in the assembly stage two different assemblers (modified version of PHRAP and CAT). The output of these was fully integrated into the pipeline with an interface to the SNP detection stage (PolyBayes [10]), a Bayesian engine that assigns a significance to each candidate SNP detected via a probability score. We used PolyBayes for the SNP detection stage in the pipeline in order to overcome the issue of variable sequence quality in EST data. Through a data interface the most relevant SNP information is extracted from PolyBayes output and stored in a relational database. The database was designed to store SNP information from several sources. Finally, a web-based user interface allows one to browse through the data and analyze the results. All the interfaces between the elements of the pipeline were built in Perl. The pipeline smoothly assembles existing tools to build a unified SNP discovery tool.

With the pipeline we analyzed around 380,000 ESTs, including a publicly available set of *Zea* maize ESTs [22], with approximately 68,000 sequences. Polymorphism frequency information, distribution of different types of mutations, and SNP count as a function of PolyBayes probability were obtained for the public maize EST set. The subset of maize SNPs will be experimentally validated, and then all SNPs will be released to the public (M. McMullen and A. Rafalski, in preparation).

2 Pipeline Description

2.1 cDNA Assembly

The assembly process is based on sequence similarity. All the members of a group are multiple aligned and a consensus sequence that represents the whole group is derived. We used two different assemblers: PHRAP and CAT.

2.2 Interface between Assembly and SNP Detection

This interface is responsible of adapting the data output of both assembly tools to be used as the input for the SNP detection stage. For PHRAP the interface simply divides PHRAP multicontig files into single contig files that are fed to PolyBayes. For CAT the interface keeps the cluster information given by the assembly but it is necessary to realign the members of a contig using PolyBayes’ anchored multiple sequence alignment.

2.3 SNP Detection

For the SNP detection stage we used PolyBayes, a Bayesian detection engine. Both data input modes were used in the analysis: ace file data input mode and anchored multiple sequence alignment input mode (see Section 3.1).

2.4 Interface between SNP Detection and DB Data Upload

The database structure was designed so that multiple sources of SNP data could be uploaded and identified for retrieval. For each source of data, the interface between the SNP detection software and the database works as follows:

- Data source characterization must be recorded prior to the SNP data upload into the database.
- For each contig, PolyBayes generates a text file containing SNPs found in a contig, along with several fields of related information. Additionally, after SNP detection, PolyBayes produces an ace file where the assembly information for a contig is contained.
- For each contig, we built two data structures by parsing the information in these files. These data structures contain all the information needed for the polymorphisms present in a contig (Fig. 1). The polymorphisms are divided into two groups, SNPs and INDELS. The INDELS can be single (occupying only one position within the consensus sequence) or multiple. Contigs containing no polymorphisms are discarded. Every polymorphism (SNP or insertion/deletion) has two associated sub-structures. The first of these sub-structures is an array (Fig. 2) where every element is part of the information that will be stored in the database. The second sub-structure is an associative array (an array where each element is associated with a key) that pairs library name with allele code (see Fig. 3). The information in the associative array data structure is important for later validation of the SNP candidates. The EST library name provides a link to the identification of individual genotype from which a sequence is derived.
- The polymorphisms data structure is traversed and the information for each particular polymorphism is uploaded into the appropriate tables of the database. At this point redundancy checks are done in order not to upload the same SNP twice. Each SNP is uniquely characterized by its flanking sequence. The flanking sequence is defined as the sequence formed by the union of the 20 nucleotides to the left of the polymorphic position in the consensus sequence, and the 20 nucleotides to the right. If the SNP is already in the database, it is not uploaded again. However the tables that relate this SNP to the particular source and contig are updated. In very rare cases, the same SNP may be detected in two different places of the same contig. In such a case, only the table that relates the SNP to the data source is updated.

2.5 SNPdb

The relational database structure has 9 tables (Fig. 4).

1. The **sources** table stores the information that characterizes each SNP source. For example a new assembly of an EST set would be considered a different source.
2. The **clusters** table holds the information that characterizes each contig within a SNP data source.
3. The **snps** table is the main table of the database, it stores all the information related to a particular SNP.
4. The **snpsource** table holds one-to-many relationships between SNP data sources and SNPs in both directions. E.g. one source is related to several SNPs, but one SNP can also be detected by different data sources.
5. The **snpccluster** table holds one-to-many relationships between contigs and SNPs in both directions. One contig could have multiple SNPs and one SNP can be present in different contigs.
6. The **library** table stores the information about library source of alleles for each SNP. The library alleles are encoded using standard IUPAC-IUB ambiguity codes.
7. The **tophits** table holds the information of the 5 tophits of a particular consensus sequence when searching this sequence in GeneBank with BLAST.
8. The **primers** table holds the information about PCR primers associated with a particular SNP.
9. The **snpprimer** table holds one-to-many relationships between primers and SNPs, since one SNP can have different types of primers associated with it, also several SNP can have the same primer set.

2.6 User Interface

A user interface has been developed in order to facilitate access the SNP information present in the database.

The flow diagram of the interface can be seen in Figure 5. From the main query page, the user can query the database for different items or introduce a new sequence that may be searched against the SNPs present in the database. The query given by the user is processed, and produces a list of hits. The user can browse through the record of each hit in the **record information** page. Each record (SNP) is associated with one or more contigs. This information can be checked in the **contig information** page. Existing records, may be changed in the **update record** page. If the user entered a new DNA sequence to be searched against the SNPs in the database, the resulting matches against the consensi sequences are presented in the **top matches** page. The SNPs that belong to a certain consensus match are listed in the **contig SNP content** page. Thus the user is able to associate a new sequence with a subset of the SNPs present in the database.

Our implementation of the SNP discovery pipeline relied heavily on Perl. Perl is a particularly suitable language for bioinformatics applications which are not solely based on numerical calculations, but heavily rely on the need of data parsing, database interface development and web-interface development. Perl has XML extensions which assure data transaction standardization.

3 Results

We present here some observations gathered from the main data sets that were processed and stored in the database.

The SNP pipeline has the following main features:

- Smooth data transition from beginning to end, this is, modify the different output formats from the pipelines involved in the pipeline to match the format of the next stage in the process.
- High-throughput capabilities. With the pipeline we have analyzed for SNPs more than 380,000 EST sequences.
- SNP detection parameter optimization scheme, which allows to fit a generic SNP discovery pipeline to a particular set of ESTs. We used a test data set of maize ESTs, but this scheme could be used for any EST collection for which a similar test data set to “train” the parameters is available.
- Transparent interface between assembly tools and SNP detection stage. In order to do this, we took advantage of both types of data input modes available in PolyBayes. The interface built for CAT assembler used the clustering information given by CAT, but it realigned the contig members to the consensus sequence of the contig with the help of PolyBayes anchored multiple sequence alignment.
- Interface between PolyBayes and SNP database, which builds data structures based on PolyBayes output and subsequently updates the SNP database.
- DB structure. The database structure implemented here was designed to store multiple sources of SNP data. Since we stored all the consensus sequences of the contigs of an assembly, in order

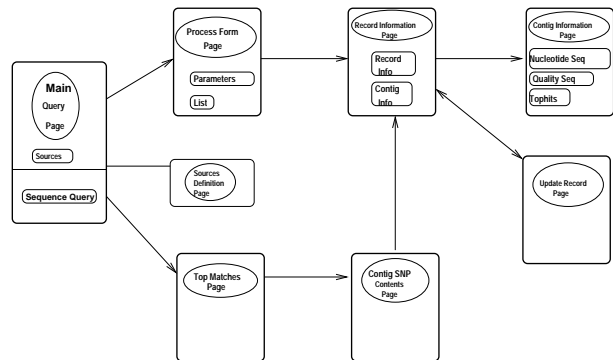


Figure 5: Flow diagram of user interface.

to associate sequences with SNPs, this could lead to very high storage requirements. Therefore we compressed the consensus sequences before storage and decompress them before presenting them in the user interface. This represents significant reduction of storage space for big data sets.

- User interface. We implemented a user interface that allows a user, not familiar with the details of SNP discovery, to use the resulting information.

3.1 SNP Detection

In the SNP detection stage we used PolyBayes, a Bayesian detection engine. We optimized PolyBayes SNP detection parameters using a test data set. Originally, parameters were given certain defaults suitable for SNP detection in humans. We expect some genome characteristics to be different in maize for the following reasons:

- Maize has a high rate of polymorphism. The expected frequency is 1 SNP per 60-120 bp vs 1 SNP per 1000-1200bp in humans [2].
- The frequency of indels present in maize is high.

Previous to this work, PCR-amplified 3'-untranslated regions of maize genes from many corn genotypes were sequenced and SNPs were visually identified (D. Bhatramakki and A. Rafalski unpublished data). We set out to maximize the agreement between this test data set of SNPs, identified by experienced researchers, and the set of SNPs reported by PolyBayes. The PolyBayes parameters with most influence on the SNP detection process are:

1. `memberBaseQualityDefault`: This is the default base quality that is used by PolyBayes when a nucleotide sequence has no corresponding base quality sequence information. For some of the sequences that were analyzed in this study, there were no trace or quality values available. In order to use them in the assembly process, we set the `memberBaseQualityDefault` value to 5, so that if one of these sequences produced a polymorphism, then the relevance given to such a candidate SNP would be very low.
2. `preScreenSnpsMinimumBaseQuality`: In order to reduce the complexity of the analysis, PolyBayes has a screening process, prior to the SNP detection algorithm, that screens out sequence slices with no observed discrepancies or sites where the aggregate base quality value of an alternative allele falls below a settable threshold value. This value is the `preScreenSnpsMinimumBaseQuality`. We set this parameter to 40.
3. `priorPoly`: This is the total expected polymorphism rate. This rate value is used in the calculation of the P_{SNP} , the probability of a cross-section in an alignment of being polymorphic. It is therefore very important for the final polymorphism analysis. We set the `priorPoly` to 1/60.
4. `thresholdSNP`: The `thresholdSNP` is the threshold of calculated SNP probability (P_{SNP}) score above which a cross-section is considered candidate SNP site. This parameter provides a balance between true positive identification rate and recovery of low-frequency alleles. Using higher values for `thresholdSNP` reduces the number of false positives, but also discards more polymorphic sites. We chose to collect all candidate SNP information and select the desired P_{SNP} range at a later step via the user interface.

3.1.1 Parameter Optimization to Match Maize EST Test Data

PolyBayes was used to analyze 873 contigs that were earlier inspected visually for SNPs. Both type of detections had the SNP detection information in an alignment file (ace file-PHRAP). These files were produced by a previous alignment process in the case of the visually identified SNPs and by PolyBayes in the case of the automatically detected SNPs. For each contig both ace files were parsed in order to

assign each detected SNP to one of the three classification groups: detected only experimentally, detected both experimentally and automatically, and detected automatically. The agreement coefficient was defined as:

$$AC = \sum_{i=1}^n \frac{Ag_i}{Exp_i + Aut_i}, \quad (1)$$

where Ag_i is the number of SNPs in agreement in the i th contig, Exp_i is the number of SNPs detected only experimentally in the i th contig, and Aut_i is the number of SNPs detected only in an automated fashion (PolyBayes) in the i th contig. Also, $1 \leq i \leq n$, where n is the total number of contigs. Higher values of the agreement coefficient, mean better correspondance between experimental and automated SNP detection. The parameter `thresholdSNP` set to 0.8 and `preScreenSnpsMinimumBaseQuality` set to 40 yielded the highest agreement coefficient.

3.2 Insights Gained with PolyBayes

PolyBayes is the commonly tool for SNP detection, for example to analyze human genome sequence variation [15], but it is not well documented.

- PolyBayes is a very powerful SNP detection engine but is poorly documented.
- When using sequences coming from multiple genomic locations, as is the case with ESTs, it is very important to use the duplicate sequence identification filter, the purpose of this filter is to separate sequences that come from different genomic locations but are very similar. No assembly process is perfect and therefore some sequences due to their high degree of similarity are classified in the wrong contig. This filter allows to partially recover from this type of error. If the filter is not used, the number of candidate SNPs reported by PolyBayes is too high.
- PolyBayes chooses candidate SNPs without taking into account the minor allele frequency information and it makes its decision based on the probability threshold. In the cases when the minor allele is present only once, even if it has been classified as a candidate SNP due to its statistical significance, biologists may not accept such a SNP as reliable.
- The two parameters to which the SNP detection is most sensitive are `thresholdSnps` and `preScreenSnpsMinimumBaseQuality`.
- When running PolyBayes over a significant number of contigs, time of analysis becomes an issue. For instance, in a set of 68,000 ESTs, the SNP detection by itself took 10 hours in a 450 SUN workstation with at 300 MHz.
- PolyBayes output can be read and displayed with `Consed`. In the initial development stage of the pipeline this was a very useful feature which allowed facile comparisons between SNPs detected visually and SNPs detected by PolyBayes.

3.3 Main Observations on the Data Set Used

Table 1 summarizes the results of SNP discovery in the public maize EST collection. A detailed discussion of maize SNPs will be presented elsewhere (A. Rafalski in preparation), but here we present some general observations:

SNP frequency: We obtained an overall average nucleotide substitution frequency of 1 per 410 bp and an average indel frequency of 1 per 1168 bp. The expected single nucleotide substitution rate was 1 per 130.5 bp [2] according to the test data set. The value of 1 per 410bp obtained which constitutes a lower frequency than expected might be explained as follows: The

Table 1: Results.

	Public collection
Total No. of bases	1806556
Total No. of ESTs	68654
Total No. of contigs	12179
Contigs with more than 1 read	9838
Contigs containing polymorphisms	1438
Total No. of polymorphisms	4307
Total No. of SNPs	2875
Total No. of Indels	1432

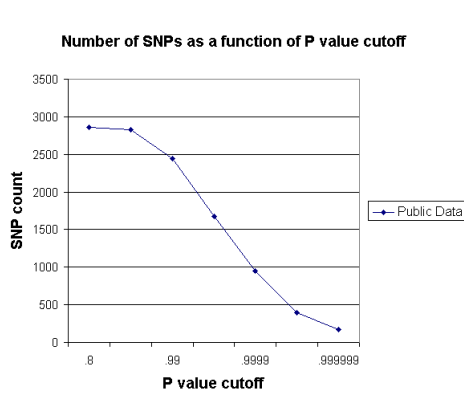


Figure 7: Number of SNPs as a function of probability value cutoff.

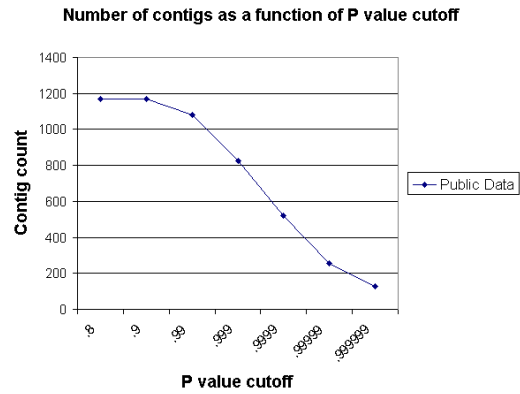


Figure 8: Number of contigs as a function of probability value cutoff.

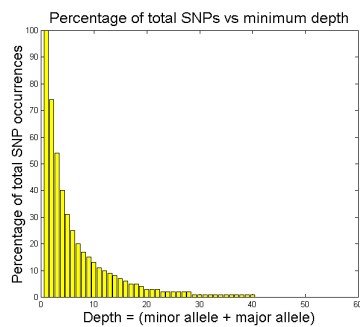


Figure 9: Depth distribution of alignments. The first bar represents a depth of two sequences.

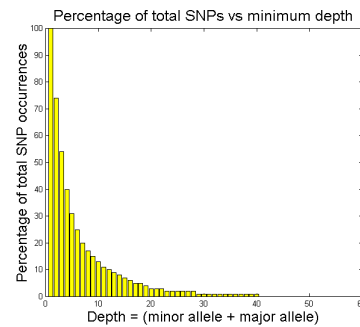


Figure 10: Percentage of total SNPs vs. minimum depth.

frequency calculation takes into account the whole length of the contigs including segments containing only a single sequence. This lowers the frequency, because the SNPs within a contig can only be found in the regions where at least two member sequences are aligned.

Number of contigs containing SNPs: In both EST collections, the percentage, of contigs where polymorphisms were detected, is relatively small. This is because the selection of genotypes from which the ESTs are derived is not optimally suited for SNP detection.

Distribution of SNPs according to mutation type (Fig. 6): As a general trend we observed that the most frequent type of mutation is that having a base change of either A/G or C/T. This is consistent with the experimental observation that cytosine demethylation is the most common mutational event. The least common type of base change is A/T. SNPs with three alleles are very rare, and those found will still have to be verified experimentally. The most common type of INDELS are single insertions deletions (A/T plus G/C).

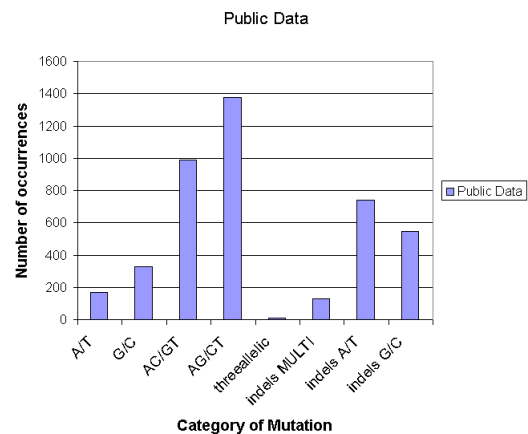


Figure 6: SNP and indel distribution.

Number of SNPs identified as a function of PolyBayes probability cutoff is shown in Fig 7. It remains to be established what fractions of SNPs can be experimentally confirmed at each probability cutoff. In Figure 7 we present a graph of SNP occurrences vs. PolyBayes probability cutoff.

In Figure 8 we also present the relationship between the number of contigs containing SNPs and probability value cutoff. As expected, this relationship is similar to the one between the total number of SNPs detected and the probability value cutoff. This is because a significant number of the total number of contigs contain only one SNP.

Figure 10 represents a graph of percentage of the number of SNPs detected that correspond to a particular minimum alignment depth. Although a significant number of SNPs are detected in shallow alignments with depths of 2 or 3 member sequences (Fig. 9), most of the SNPs detected are found in alignments of 4 or more member sequences. Alignments with a depth of four or more represent 54% of the total number of SNPs. 40% of the SNPs are in alignments of the depth of five or more.

Distribution of SNPs according to minor allele occurrences (Fig. 11): In 46% of the SNPs detected, the minor allele was found only once, therefore many of these SNPs may have lower reliability. However, this assumption needs to be tested experimentally by comparing confirmation rate of single occurrence SNPs vs. multiple occurrence SNPs

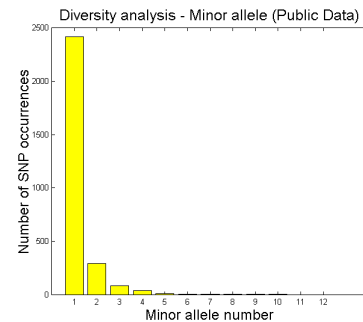


Figure 11: Minor allele distributions in alignments.

4 Conclusions

The SNP detection heavily relies on the quality values of the sequences, therefore is very important to obtain the base quality values or trace files for as many sequences as possible. So, if too many sequences do not have quality values the SNP analysis decreases its reliability.

Although PolyBayes has a sound statistical basis, it would be interesting to analyze the same set of ESTs with two different SNP detection tools based on different detection principles and then reconfirm the SNPs experimentally. For instance DoubleTwist recently released SNPTwist, which is based on the principle of maximum entropy.

In order to be able to analyze millions of sequences, automated SNP discovery *in-silico* will be the predominant SNP discovery approach in the future.

Today's bioinformatics landscape is a "glued" mosaic of experimental tools, of which the work reported is an example. At this point of time this is unavoidable, but the aim of future bioinformatics tool development and database construction is to build systems that follow a minimum level of common structure. There is an urgent need of standardization in the field.

One important lesson learnt is that many of the currently available bioinformatics tools fail or underperform when handling significant amounts of data. New bioinformatics tools have to be created that support fast and easy access to an ever growing amount of data.

Acknowledgements

We would like to thank DuPont Crop Genetics in Newark (Delaware) for their invaluable support. We also would like to thank Delaware Biotechnology Institute in Newark (Delaware) for their invaluable support and for opening plenty of research opportunities for scientists as well as for graduate students in the field of bioinformatics.

References

- [1] Buetow, K.H., Edmonson, M.N., and Cassidy, A.B., Reliable identification of large numbers of candidate SNPs from public EST data, *Nat. Genet.*, 21:323–325, 1999.
- [2] Ching, A. *et al.*, SNP frequency and haplotype structure of 18 maize genes, submitted.
- [3] Davis, G.P. and DeNise, S.K., The impact of genetic markers on selection, *J. Anim. Sci.*, 76(9):2331–2339, 1998.
- [4] Grompe, M., The rapid detection of unknown mutations in nucleic acids, *Nat. Genet.*, 5:111–117, 1993.
- [5] Grompe, M., Muzny, D.M., and Caskey, C.T., Scanning detection of mutations in human ornithine transcarbamoylase by chemical mismatch cleavage, *Proc. Natl. Acad. Sci. USA*, 86(15):5888–5892, 1989.
- [6] Gut, I.G., Automation in genotyping of single nucleotide polymorphisms, *Hum. Mutat.*, 17(6):475–492, 2001.
- [7] Hacia, J.G., Brody, L.C., Chee, M.S., Fodor, S.P., and Collins, F.S., Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis, *Nat. Genet.*, 14(4):441–447, 1996.
- [8] Lander, E.S. *et al.*, Initial sequencing and analysis of the human genome, *Nature*, 409(6822):860–921, 2001.
- [9] Leren, T.P., Rodningen, O.K., Rosby, O., Solberg, K., and Berg, K., Screening for point mutations by semi-automated DNA sequencing using sequenase and magnetic beads, *Biotechniques*, 14(4):618–623, 1993.
- [10] Marth, G.T. *et al.*, A general approach to single-nucleotide polymorphism discovery, *Nat. Genet.*, 23(4):452–456, 1999.
- [11] Picoult-Newberg, L. *et al.*, Mining SNPs from EST databases, *Genome Res.*, 9(2):167–174, 1999.
- [12] Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., and Baumeister, K., A system for rapid DNA sequencing with fluorescent chain-terminating dydeoxynucleotides, *Science*, 238(4825):336–341, 1987.
- [13] Risch, N. and Merikangas, K., The future of genetic studies of complex human diseases, *Science*, 273(5281):1516–1517, 1996.
- [14] Roses, A.D., Pharmacogenetics and the practice of medicine, *Nature*, 405:857–865, 2000.
- [15] Sachidanandam, R. *et al.*, A map of human genome sequence variation containinig 1.42 million single nucleotide polymorphisms, *Nature*, 409(6822):928–933, 2001.
- [16] Shattuck-Eidens D. *et al.*, A collaborative survey of 80 mutations in the BRCA1 breast and ovarian cancer susceptibility gene, *Jour. Am. Med. Ass.*, 273(7):535–541, 1995.
- [17] Sheffield, V.C., Cox, D.R., Lerman, L.S., and Myers, R.M., Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes, *Proc. Natl. Acad. Sci. USA*, 86(1):232–236, 1989.
- [18] Soller M. and Andersson L., Genomic approaches to the improvement of disease resistance in farm animals, *Rev. Sci. Tech.*, 17(1):329–345, 1998.
- [19] Solomon-Blackburn, R.M. and Barker, H., Breeding virus resistant potatoes (*Solanum tuberosum*): a review of traditional and molecular approaches, *Heredity*, 86(Pt1):17–35, 2001.
- [20] <http://www.phrap.org/>
- [21] <http://www.doubletivist.com/>
- [22] <http://www.zmldb.iastate.edu/>