

SNP Analysis and Presentation in the Pharmacogenetics of Membrane Transporters Project

D. Stryke, C.C. Huang, M. Kawamoto, S.J. Johns, E.J. Carlson, J.A. Deyoung, M.K. Leabman, I. Herskowitz, K.M. Giacomini, T.E. Ferrin

Pacific Symposium on Biocomputing 8:535-547(2003)

SNP ANALYSIS AND PRESENTATION IN THE PHARMACOGENETICS OF MEMBRANE TRANSPORTERS PROJECT

DOUG STRYKE, CONRAD C. HUANG, MICHIKO KAWAMOTO,
SUSAN J. JOHNS, ELAINE J. CARLSON, JOSEPH A. DEYOUNG,
MAYA K. LEABMAN, IRA HERSKOWITZ,
KATHLEEN M. GIACOMINI, THOMAS E. FERRIN

*Departments of Pharmaceutical Chemistry,
Biopharmaceutical Sciences, and Biochemistry & Biophysics
University of California San Francisco, 513 Parnassus Avenue,
San Francisco, CA 94143, USA*

The multidisciplinary UCSF Pharmacogenetics of Membrane Transporters project seeks to systematically identify sequence variants in transporters and to determine the functional significance of these variants through evaluation of relevant cellular and clinical phenotypes. The project is structured around four interacting cores: genomics, cellular phenotyping, clinical phenotyping, and bioinformatics. The bioinformatics core is responsible for collecting, storing, and analyzing the information obtained by the other cores and for presenting the results, in particular, for the genomic data. Most of this process is automated using locally developed software written in Python, an open source language well suited for rapid, modular development that meets requirements that are themselves constantly evolving. Here we present the details of transforming ABI trace file data into useful information for project investigators and a description of the types of data analysis and display that we have developed.

1 Introduction

1.1 Motivation

Membrane transporters, a major determinant of pharmacokinetics, are of great pharmacological importance. By controlling the amount of drugs within the body, they determine whether drug levels are sufficient for therapeutic effect. Transporters play a second important role in pharmacology in that about 30% of the most commonly used prescription drugs target transporters.

The UCSF Pharmacogenetics of Membrane Transporters Project (PMT) includes investigators from diverse disciplines who are conducting a series of integrated studies to elucidate the pharmacogenetics of membrane transport proteins. To accomplish this, the investigators are systematically identifying sequence variants in transporters and determining the functional significance of these variants through evaluation of relevant cellular and clinical phenotypes. The goal of the PMT is to understand the genetic basis for variation in drug response for drugs that interact with membrane transporters. As a part of the PMT project, the

goal of the bioinformatics core is to provide support to the genomic, cellular phenotyping, and clinical phenotyping cores.

1.2 Overview

Our multidisciplinary project is structured around four interacting cores: genomics, cellular phenotyping, clinical phenotyping, and bioinformatics. The genomics core identifies sequence variants in the targeted transporters from a collection of DNA samples. The cellular phenotyping core tests the functional significance of the sequence variants in cell-based assays. The clinical phenotyping core tests hypotheses about the functional significance of variation in membrane transporters in clinical drug response. The bioinformatics core develops and maintains a database of the information obtained by the other cores and by individual project investigators. A particularly important additional role of the bioinformatics core is to analyze the genomic data, *e.g.*, determine haplotypes, identify evolutionarily conserved and nonconserved positions in proteins, and calculate population-genetic parameters for mutation rate. The project's web site is <http://pharmacogenetics.ucsf.edu/>.

In the first year, the genomics core screened 367 amplicons, comprising 24 genes. The sample set consisted of 247 ethnically identified DNA samples from the Coriell Institute¹. The bioinformatics core collected, stored, and analyzed these data and presented the results in various formats. Most of this process was automated using software written by the bioinformatics core in Python, an open source programming language². This dynamically typed, object oriented language is well suited for rapid development of software tools in a dynamic, evolving environment. Our software falls roughly into two categories: analysis and presentation. The analysis software we call RefMap and SnpMap; the presentation software, SnpWeb.

In the bioinformatics core, one person began working part-time on this project in late 2000. In early 2001, two more people were assigned part-time to the project, bringing the full-time equivalent staff to one-and-one-third. All development is done on clustered HP Alpha servers running Tru64 Unix and TruCluster Server V5.

2 Analysis

2.1 Data Flow

The process begins when the project investigators select a transporter for study (Figure 1). The bioinformatics core retrieves the relevant GenBank entry from NCBI, preferably a curated RefSeq entry³. A Python program extracts gene coding sequence (CDS) and chromosome position from the GenBank entry. Using

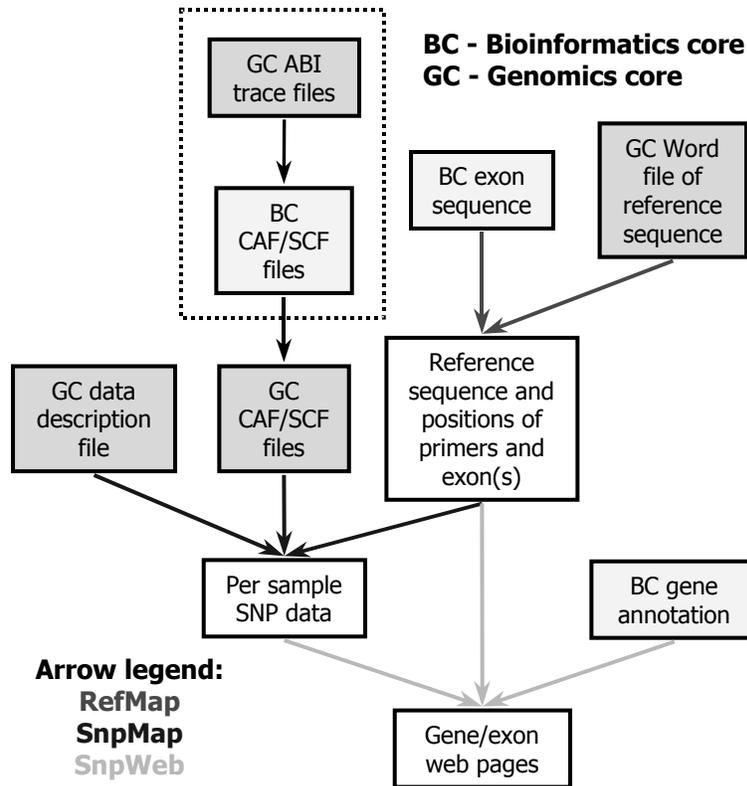


Figure 1 Work flow schematic (dashed box represents planned addition of PolyPhred for base calling).

BLAST⁴, the CDS is queried against the high-throughput genomic sequences to locate exons. If necessary, the nonredundant nucleotide database is also searched. The exon boundaries are then verified by looking for splice junctions.

Whenever multiple genomic locations for the same exon are found, these regions are checked via alignments to determine the nature of the event. All the areas are reported; however, the area that best fits the organization of the rest of the gene is recommended. In the case of an extremely short first exon, we combine the exon with its respective 5' UTR region to aid in its location. Once confirmed, the exon sequences, the genomic sequences, the CDS, and the annotations are forwarded to the genomics core.

The genomics core designs PCR primers to amplify each exon and a minimum of 35 flanking 3' and 5' intronic bases. The amplicons are typically 250-500 bases long. Small, closely spaced exons are combined; larger exons are sequenced using multiple, overlapping amplicons.

Six random samples are sequenced to validate the PCR and to determine the polymorphism frequency. If polymorphisms are detected in more than 3 of 12 chromosomes, all 247 samples are sequenced. Otherwise, the sample set is screened in pools of three using denaturing high-performance liquid chromatography (DHPLC), and only samples that are positive for polymorphisms are sequenced. We assume that the DHPLC “normal” samples match the reference sequence.

When presenting SNP data, the choice of reference sequence is somewhat arbitrary. For example, an ‘A’ to ‘G’ SNP with a 40% frequency is equivalent to a ‘G’ to ‘A’ SNP with a 60% frequency. In theory, an investigator could choose to present SNP data relative to a reference sequence somewhat different from the one used by the genomics core. In practice, however, altering the reference sequence changes the genotype inferred by DHPLC.

To permit flexibility in presentation, we implemented a second, independent reference sequence for presentation. Differences from the analysis reference sequence in the form of base substitutions are correctly interpreted and presented so long as the positions changed contain SNPs, since our SNP data file format contains genotypes for all samples at those positions. However, as all other positions are assumed to be reference, care still must be exercised to avoid changing the presentation reference sequence at non-SNP positions.

The genomics core sequences the samples using BigDye Terminator cycle sequencing and an ABI PRISM 3700 DNA Analyzer. The sequences are aligned and edited, and heterozygous bases are assigned IUPAC-IUB ambiguity codes using Sequencher⁵. The genomics core trims each sample sequence of primer and poor quality base calls, both of which would confound subsequent analysis. From Sequencher, aligned contigs are exported as a Common Assembly Format⁶ (CAF) file and all sample sequence data as Standard Chromatogram Format⁷ (SCF) files.

The plain text CAF file contains all of the predominant sample base calls aligned into one or more contigs, each with its own consensus sequence. Each consensus sequence entry also contains a list associating each sample sequence with start and stop positions in two coordinate systems: the sample’s own and that of the consensus sequence. The genomics core adds a copy of the amplicon reference sequence to each contig, thereby aligning it to the consensus as well. This creates a two-step alignment from each sample sequence to the reference sequence via the contig consensus sequences. The Sequencher-scored heterozygous base calls are contained in the binary SCF files. Optimally, each sample is sequenced in both sense (forward) and antisense (reverse) directions.

In addition to the CAF file’s inherent data structure, we have defined a naming scheme for contigs and samples that allows most polymorphisms to be interpreted using automated software analyses. Heterozygous insertions and deletions (indels), however, throw chromatograms out of phase, making their interpretation beyond the

indel difficult. The genomics core places heterozygous indel sequences into separate CAF file contigs, one contig for each unique event. They encode the number of bases involved and the SNP types in the contig name. For example, "EXON_7_HET_2_DEL" indicates a two-base heterozygous deletion in exon seven. To further simplify analysis, the genomics core removes the phase-shifted base calls beginning immediately beyond the last indel base in each direction.

Assuming the availability of data from both forward and reverse directions, the above indel protocol provides for automated analysis of up to two heterozygous indels in a single sample by separating the forward and reverse reads into distinct contigs. This proved sufficiently robust for all but one SNP in the first set of 367 amplicons.

More complicated samples require manual intervention. The genomics core gathers these sequences into a contig named, appropriately, "nonstandard." They also describe the indel events in a text file already included with each amplicon. The nonstandard contig signals the need to build a data file by hand to augment the automatically derived data.

The genomics core transmits data to the bioinformatics core in amplicon bundles. Each bundle includes the reference sequences for the amplicon and primers, a CAF file, and the SCF and ABI trace files. Also included is a text file describing information not available in the sequence data. For example, a "no SNPs" flag is used to distinguish an amplicon without trace files for which DHPLC screening found no SNPs from one that is simply missing data.

The bioinformatics core program RefMap begins with the amplicon reference sequence, received as a Microsoft Word document and formatted to delineate exon and PCR primer boundaries. RefMap transforms this file into an HTML file by running it through a program called mswordview⁸. The HTML format preserves the relevant formatting contained in the Microsoft Word document yet is easily parsed. RefMap confirms the exon sequence and adjusts the exon boundaries using the exon sequence originally identified by the bioinformatics core. The amplicon sequence and the primer and exon locations are written to a text file in a modified FASTA format.

Next, SnpMap parses the CAF file. The Sequencher-scored heterozygous base calls contained in the SCF files are substituted for the predominant base calls obtained from the ABI 3700. Using CAF file contig alignment information, sample sequences are aligned relative to the reference sequence. Sample alleles are collected by contig for each reference position. Each multi-base insertion is collated into the appropriate reference position and the downstream positions adjusted for the offset created by the inserted bases. The per-contig SNP sites are then collated into a single file-wide set.

SnpMap subjects the data to quality checks at each step. As stated, almost all aspects of the CAF file follow a defined naming convention. Nonconforming items

are flagged. The original reference amplicon is compared against the reference sequence contained in each CAF file contig. Multiple sample reads for a single ID of the same direction are compared for consistency, as are the forward and reverse reads for each sample.

Reference positions lacking polymorphisms are then dropped. The remaining polymorphic sites are written to a tab-delimited file containing sample number, read direction, reference position, genotype, and the source of the base calls, *i.e.*, observed by direct sequencing or inferred via DHPLC. Subsequent analyses are based upon this file.

Not every amplicon successfully passes this process the first time. SnpMap classifies problems into two categories: warnings and errors. Warnings typically apply to individual samples, which can be dropped from the amplicon pending correction. Errors, on the other hand, halt the analysis. Descriptive messages are written to log files, which are linked to and summarized by a color-coded web page that provides at-a-glance status of all PMT amplicons.

2.2 Products

The data are analyzed over four scopes: by SNP site, by gene, by gene family, and over the entire gene set. Each SNP is categorized by type, including exonic vs. intronic, coding vs. noncoding, cytoplasmic vs. extracellular vs. transmembrane, and substitution vs. insertion vs. deletion. SNP statistics, including the chi-square probability that the difference between the observed allele distributions and the predicted Hardy-Weinberg equilibrium could be due to chance alone⁹, are calculated overall and by ethnic group.

Gene analyses include molecular genetics statistics over categories such as transitions, transversions, and specific variations within CpG islands. We compare our SNPs to those reported in dbSNP¹⁰. Haplotypes are estimated using PHASE¹¹, a program for reconstructing haplotypes (Figure 2). Haplotype-based genetic diversity is calculated by treating each gene haplotype as an allele, and counting the number of differences for all pairs of sample haplotypes¹².

If at least two mammalian homologs of the human transporter can be found, we use multiple sequence alignments to determine evolutionarily conserved amino acid positions. Transmembrane domains and SNP locations are also indicated.

We calculate population genetics statistics, such as nucleotide polymorphism, nucleotide diversity, and the Tajima test of the neutral mutation hypothesis¹³ for each gene. These statistics are also computed for gene families or across the entire gene set.

Results are currently written out as tab-delimited files. Some of these files are reformatted into Extensible Markup Language¹⁴ (XML), conforming to schemas defined by the Pharmacogenetics and Pharmacogenomics Knowledge Base¹⁵

(PharmGKB), and then transmitted to the PharmGKB. As part of the Pharmacogenetics Research Network¹⁶, our resource disseminates data publicly via the PharmGKB. The results also become the input for the suite of programs that we call SnpWeb, which presents these results in a scientifically useful manner to the PMT investigators.

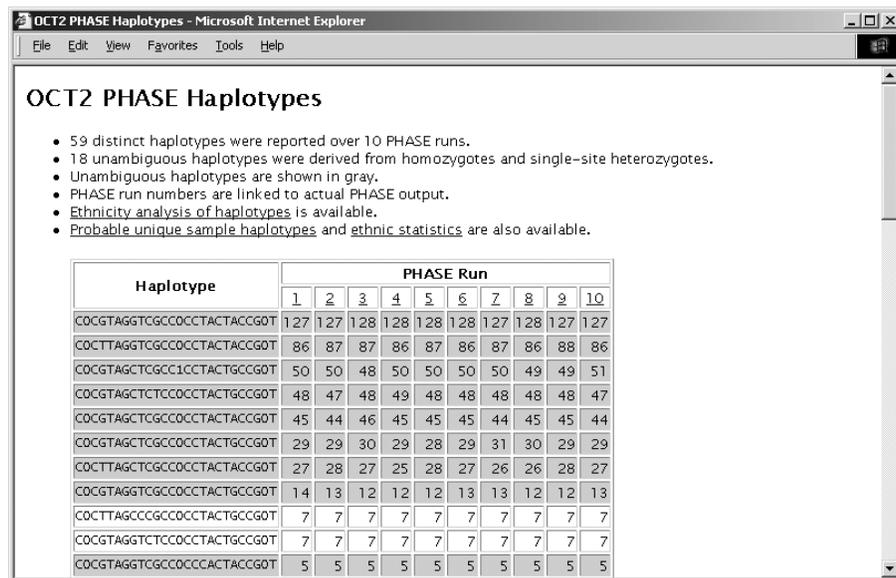


Figure 2 Partial PHASE haplotype estimation for OCT2

3 Presentation

3.1 Overview

SnpWeb presents the analysis results in two formats. Tab-delimited files are made available to project investigators and researchers for further analysis. These files are easily imported into desktop spreadsheet or database programs. Basic SNP statistics, per sample data, and summary population genetics statistics are made available this way.

The primary output format that SnpWeb generates is World Wide Web content, including HTML, graphics, and plots. The advantages of web pages include cross-platform availability, familiarity to scientists, and availability from multiple

indicates how many variants of that type were found. The exons are numbered with hyperlinks, which lead to the amplicon web pages.

The second gene page graphic is generated by TOPO2¹⁷, which uses transmembrane predictions to create a secondary-structure graphic of the amino acid locations relative to the cell membrane (Figure 4). We obtain transmembrane predictions from SwissProt annotations and from published papers. Coding SNP locations are indicated. A TOPO2 image is also generated for each amplicon, which additionally highlights exon boundaries.

Exon	SNP #	CDS Pos	Exon Pos	Nucleotide Change	Amino Acid Position	Amino Acid Change	Total Freq	AA Freq	CA Freq	AS Freq	ME Freq	PA Freq
							n=494 o=492 i=0	n=200 o=200 i=0	n=200 o=198 i=0	n=60 o=60 i=0	n=20 o=20 i=0	n=14 o=14 i=0
1	1		(-47)	C -> T	-	-	0.002 (0.975) n=492	0.000 (n/a) n=200	0.000 (n/a) n=198	0.017 (0.926) n=60	0.000 (n/a) n=20	0.000 (n/a) n=14
1	2	134	134	T -> +A	45	Phe -> INS	0.002 (0.975) n=492	0.000 (n/a) n=200	0.005 (0.960) n=198	0.000 (n/a) n=60	0.000 (n/a) n=20	0.000 (n/a) n=14
1	3	160	160	C -> T	54	Pro -> Ser	0.002 (0.975) n=492	0.005 (0.960) n=200	0.000 (n/a) n=198	0.000 (n/a) n=60	0.000 (n/a) n=20	0.000 (n/a) n=14
1	4	390	390	G -> T	130	syn	0.280 (0.036) n=492	0.205 (0.270) n=200	0.394 (0.126) n=198	0.200 (0.819) n=60	0.250 (0.292) n=20	0.143 (0.659) n=14

Figure 5 SNP statistics by ethnic group and overall for OCT2 exon 1.

We present the basic SNP statistics on another set of web pages, one for each amplicon, in tabular form (Figure 5). Three different position numbers are presented: the CDS position, the exon relative position, and the amino acid position. The amino acid change, if any, is shown. The number of chromosomes constituting the nominal sample set (n), the number observed by sequencing (o), and those inferred by DHPLC (i) are listed by ethnic group and overall. Frequency figures are broken down by ethnic group, as are Hardy-Weinberg probability figures. High frequency SNPs and presumed out-of-equilibrium distributions are highlighted in red. The full amplicon sequence is presented in color to indicate primer and exon locations (not shown). SNPs are mapped onto the amplicon in red and keyed to the table.

Per-sample diplotype data are presented in tabular format (not shown). One table uses colored squares to indicate whether each sample has data at each SNP

position and, if so, whether it is homozygous, heterozygous, or reference. Reference samples are further differentiated as to whether they were observed from sequencing or inferred by DHPLC screening. A second table presents similar data in a predominantly textual format.

We present a summary of population genetics statistics in tabular form. This summary provides statistics by ethnicity for various combinations of SNP characteristics, such as coding vs. noncoding, synonymous vs. nonsynonymous, conserved vs. unconserved, and transmembrane vs. loop. A web-based plotter allows researchers to plot the level of nucleotide diversity (θ), average heterozygosity (π), and Tajima's D statistic (Figure 6). Plots can be generated for any combination of genes or gene families over all samples or by ethnic group and for various sequence regions.

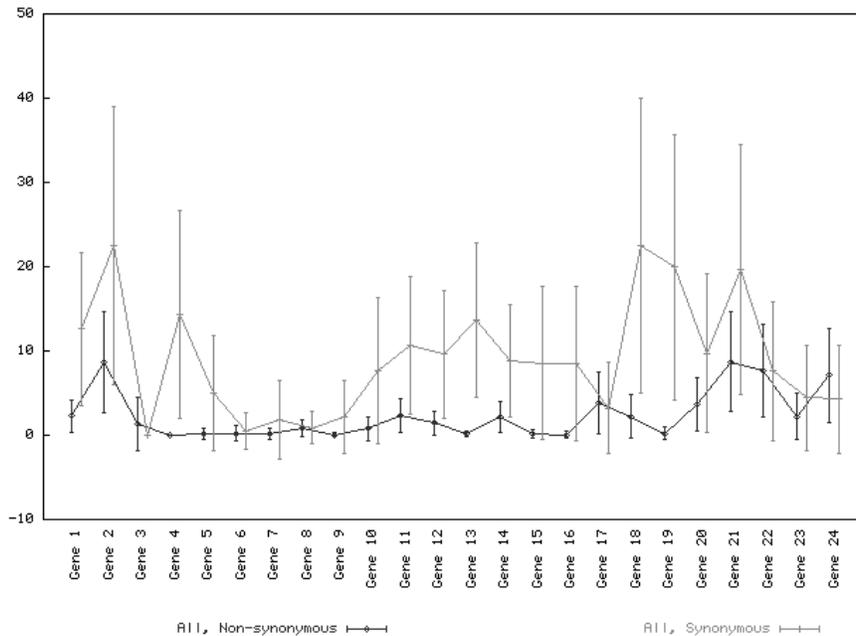


Figure 6 Average heterozygosity (π) for synonymous and non-synonymous SNPs for all genes.

Alignments to other mammalian species are displayed in a color-coded multiple alignment (Figure 7). Gray shading indicates conserved regions. Orange bars above the alignments highlight transmembrane domains. The positions of the coding SNPs are indicated on the alignment, making it readily apparent whether a nonsynonymous SNP affects an evolutionarily conserved or nonconserved position.

In an identical manner, transporter gene families are also presented in graphical multiple alignments.

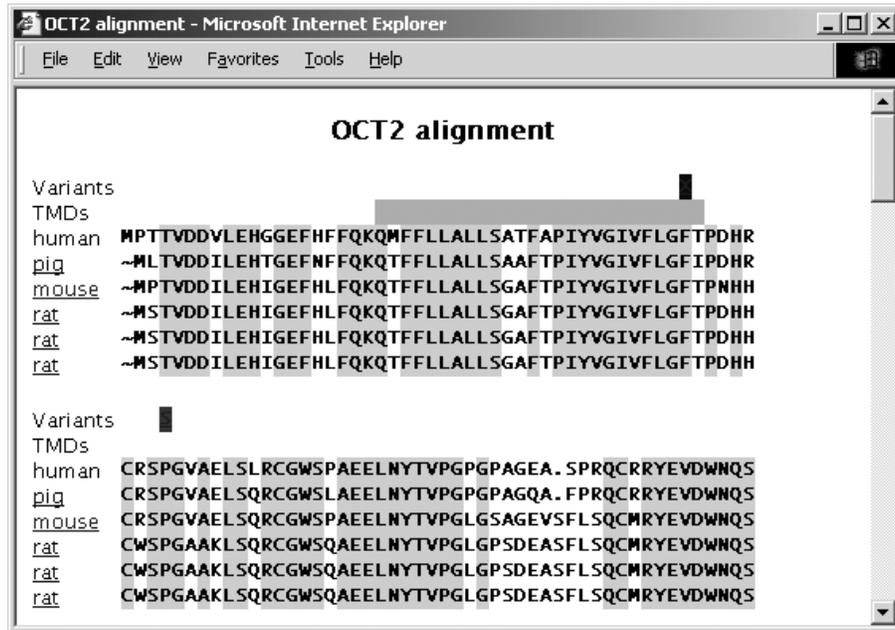


Figure 7 Consensus alignment with mammalian homologs showing conserved locations, predicted transmembrane domains, and SNP locations for a portion of OCT2.

4 Conclusion

Current Status

We currently store data and analysis results as files, using the file system to keep everything organized. The first year files number over 500,000 and consume approximately 30 gigabytes of disk space. The web pages, though created by Python scripts, are static. Programmatically, the gene is the unit of analysis, and the average time to rerun the analysis for a gene is 11 minutes. Software maintenance has been kept manageable by distributing the programming code over more than 30 Python modules. Evolving requirements for analysis have encouraged some modules to grow overly complex, however. This fact and the continued growth of the data set have led us to consider using a relational database to store intermediate results. Storing intermediate results in a database would facilitate modular

programming, speed some analysis updates, and greatly enhance our ability to mine the data.

Another planned change is to use the PolyPhred^{18,19} suite of programs from the University of Washington for primary base calling and SNP detection (Figure 1). Preliminary testing with version 3 of the program has largely confirmed the accuracy of our SNP detection protocol, but with the availability of version 4 of the program we opted to put off implementation in favor of further testing.

To date, the bioinformatics core has interacted primarily with the genomics core and, to a lesser extent, the cellular phenotyping core. The clinical phenotyping core has been primarily engaged in recruitment of test subjects. That part of the project should begin producing data in the next year, bringing new challenges for analysis and presentation.

Acknowledgments

The UCSF Pharmacogenetics of Membrane Transporters (PMT) Project is sponsored by the National Institutes of Health's National Institute of General Medical Sciences (grant U01 GM61390). Support for this project also comes from NIH P41-RR01081.

References

1. Coriell Cell Repository, <http://locus.umdj.edu/nigms/>
2. <http://www.python.org/>
3. K.D. Pruitt and D.R. Maglott, "RefSeq and LocusLink: NCBI gene-centered resources" *Nucleic Acids Research* **1**, 137 (2001)
4. S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, "Basic local alignment search tool" *Journal of Molecular Biology* **215**, 403 (1990)
5. <http://www.genecodes.com/>
6. S. Dear, R. Durbin, L. Hillier, G. Marth, J. Thierry-Mieg, and R. Mott, "Sequence assembly with CAFTOOLS" *Genome Research* **3**, 260 (1998)
7. S. Dear and R. Staden, "A standard file format for data from DNA sequencing instruments" *DNA Sequence* **3**, 107 (1992)
8. <http://www.wvware.com/>
9. D.L. Hartl and A.G. Clark, *Principles of Population Genetics, Third Edition*, 80-82 (1997)
10. <http://www.ncbi.nlm.nih.gov/SNP/>
11. M. Stephens, N. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data" *American Journal of Human Genetics* **68**, 978 (2001)

12. J.C. Stephens, J.A. Schneider, D.A. Tanguay, J. Choi, T. Acharya, S.E. Stanley, R. Jiang, C.J. Messer, A. Chew, J. Han, J. Duan, J.L. Carr, M.S. Lee, B. Koshy, A.M. Kurnar, G. Zhang, W.R. Newell, A. Windemuth, C. Xu, T.S. Kalbfleisch, S.L. Shaner, K. Arnold, V. Schulz, C.M. Drysdale, K. Nandabalan, R.S. Judson, G. Ruano, and G.F. Vovis, "Haplotype variation and linkage disequilibrium in 313 human genes" *Science* **293**, 489 (2001)
13. F. Tajima, "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism" *Genetics* **123**, 585 (1989)
14. <http://www.w3.org/XML/>
15. M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman and T. E. Klein, "PharmGKB: the Pharmacogenetics Knowledge Base" *Nucleic Acids Res.* **30**, 163 (2002)
16. <http://www.nigms.nih.gov/pharmacogenetics>
17. S.J. Johns and R.C. Speth, "TOPO, Transmembrane protein display software" <http://www.sacs.ucsf.edu/TOPO/topo.html>
18. B. Ewing, L. Hillier, M.C. Wendl and P. Green, "Base-calling of automated sequencer traces using Phred. I. Accuracy assessment" *Genome Research* **8**, 186 (1998)
19. B. Ewing and P. Green, "Base-calling of automated sequencer traces using Phred. II. Error probabilities" *Genome Research* **8**, 186 (1998)