

Building a Laboratory Information Management System for FP-TDI Genotyping Research

Daniel C. Koboldt¹, Pui-Yan Kwok², and Raymond D. Miller¹

Keywords: SNP, FP-TDI, Laboratory Information Management System, LIMS, HapMap Project, open source, XML, MySQL, Perl, assay design, quality control

1 Introduction

Single-nucleotide polymorphisms (SNPs) can change the function or the regulation of a protein. They also are useful as genetic markers that can be used to find the actual DNA sequence variants that cause differences in gene function or regulation; many such differences directly contribute to disease processes.¹ The ability to genotype SNPs rapidly and cost-effectively is essential for many applications such as mutation detection and linkage mapping.²

Our laboratory uses a genotyping method for SNPs that combines the specificity of nucleotide incorporation by DNA polymerase and the sensitivity of fluorescence polarization, known as Template-directed Dye Terminator Incorporation assay with Fluorescent Polarization detection (FP-TDI). Following PCR amplification of the target region, a SNP-specific primer anneals immediately upstream of the polymorphic site in the target DNA. Appropriate dye-labeled terminators extend the SNP-specific primer by one base; by determining which terminator is incorporated, the allele present in the target DNA can be inferred. The highly specific and sensitive nature of FP-TDI allows us to perform high-throughput SNP genotyping in a 384-well plate format, generating about 10,000 genotypes per day.³

The development of a laboratory information system became especially critical with our participation in the International HapMap Project, the goal of which is to develop a haplotype map of the human genome. The freely-available information produced by the Project is expected to be an important resource for researchers who want to find genes related to health, drug response, and disease. As a genotyping center for the HapMap project, our laboratory must have the capability to receive large incremental SNP allocations from a data control center (DCC) over a thousand miles away. Our FP-TDI machines output data in a raw format which has to be processed, organized, put through quality control, and then submitted back to the center.

2 Data Requirements

Participation in the HapMap project requires a data pipeline that begins at SNP allocation. On a regular basis the data control center sends us massive compressed XML files that contain the sequence and alleles for each SNP that has been found in our region, chromosome 7p. We must parse, sort, and store all of that information. With each new allocation, we must determine the SNPs for which we can successfully design assays, a computationally-intensive process that takes advantage of freely available databases including dbSNP and RepBase⁴. We select from among assayable SNPs those which meet our criteria for genotyping. When technicians run the assays, a

¹ Washington University School of Medicine, Campus Box 8123, St. Louis, Missouri 63110 USA.
E-mail: dkoboldt@psts.wustl.edu

² University of California, San Francisco, 505 Parnassus Ave, Long 1332A, Box 0130, San Francisco, CA 94143-0130 USA.

E-mail: kwok@cvrmail.ucsf.edu

raw text file is the only output. Those raw files must be processed so that our quality control team can view the results in PerkinElmer's SNPScorer software. An interface is necessary for the team to view call rates and update QC assessments for each assay without making direct changes to the database records. In preparation for submission, called data from our laboratory must be combined with that from collaborators located in San Francisco. Finally, the data must be exported into a precise XML format and delivered to the data control center.

3 LIMS Solution

We felt that an open-source platform offered the flexibility and capabilities that best suited the vast quantities of data in our pipeline. A customized UNIX environment provides the backbone for our data processing. Three relational MySQL databases store our laboratory data in every stage of the data pipeline; they also provide the storage and organization capabilities required for complex informatics tasks including assay design, SNP choosing, primer orders, data storage, analysis, quality control, and data submission. Once DCC files are downloaded and decompressed, Perl scripts parse out the requisite SNP information and update our database. In the ensuing assay design pipeline, the most ideal PCR and SNP-specific primers are selected and ranked for every SNP possible. A mapping program written in Perl selects the SNPs in our region that have assays, have not been ordered, and meet the distribution requirements set by the HapMap Project Steering Committee. Additional software generates the order files for mapped SNPs and parses the raw output files into our database once the assays have been run. Our intranet provides web interfaces built in Perl CGI that allow our staff to analyze, call, and assess the quality of each assay. Additional web interfaces chart our progress on 7p, help staff members to update plate ordering information, and display calculated assay errors (such as Mendelian failures). Data is exported into XML format for transmission between laboratories and submission to the DCC.

4 SUMMARY

We used open-source technology to build a Laboratory Information Management System (LIMS) to store, manipulate, and share the data produced by our FP-TDI technology.

References

- [1] Brooks, Lisa D. 2003. SNPs: Why Do We Care?. *Methods in Molecular Biology, vol. 212: Single Nucleotide Polymorphisms: Methods and Protocols.*, Humana Press Inc., Totowa, NJ, USA 1-14.
- [2] Gibson, Greg and Muse, Spencer V. 2002. *A Primer of Genome Science*. Sunderland, Mass.: Sinauer Associates, Inc.
- [3] Hsu, Tony M. and Kwok, P.Y. 2003. Homogeneous Primer Extension Assay With Fluorescence Polarization Detection. *Methods in Molecular Biology, vol. 212: Single Nucleotide Polymorphisms: Methods and Protocols.*, Humana Press Inc., Totowa, NJ, USA 177-187.
- [4] Vieux, E.F., Kwok, P.Y., and Miller, R.D. 2002. Primer Design for PCR and Sequencing in High-Throughput Analysis of SNPs. *Biotechniques*, USA Suppl:28-30, 32.