

Department Seminars

Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data

Hector Corrada Bravo
Postdoctoral Fellow
Biostatistics Department
Johns Hopkins



Abstract:

Second-generation sequencing (sec-gen) technology is capable of sequencing millions of short fragments of DNA in parallel and can be used to assemble complex genomes for a fraction of the price and time of previous technologies. In fact, a recently formed international consortium, the 1,000 Genomes Project, plans to sequence the genomes of approximately 1,200 people. The possibility of comparative analysis at the sequence level of a large number of samples across multiple populations may be achievable within the next five years.

These data present unprecedented challenges in statistical analysis. For instance, analysis operates on millions of short nucleotide sequences, or reads, which are the result of complex processing of noisy continuous fluorescence intensity data. This complex processing, known as base-calling, results in discretized sequence reads of widely varying quality. Furthermore, this variation in processing quality results in infrequent but systematic errors that we have found to be misleading in downstream analysis of the discretized data at the sequence read level. For instance, a central goal of the 1000 Genomes Project is to quantify across-sample variation at the single nucleotide level. At this resolution, small error rates in sequencing prove significant, especially for rare variants. Therefore, modeling and quantifying the uncertainty inherent in the generation of sequence reads is important. We present a simple model to capture uncertainty arising in the base-calling procedure of the Illumina platform. Model parameters have a straightforward interpretation in terms of the bio-chemistry of base-calling, which allow for informative and easily interpretable metrics that capture the variability in sequencing quality. In contrast to other recently proposed methods for improved base-calling in the Illumina platform, our model provides informative estimates readily usable in quality assessment tools while retaining base-calling performance.

Tuesday,*

May 12, 2009

5275 MSC

11:30-12:30 p.m.*

***note changed
day & time**

