Disputation vid Karolinska Institutet

Fredman, David

Computational exploration of human genome variation

Fredagen den 24 september 2004, kl. 10.00.
Institutionen för Cell- och Molekylärbiologis auditorium, Berzelius väg 21, Karolinska Institutet.

ISBN: 91-7140-025-7 Diss: 04:397

http://diss.kib.ki.se/2004/91-7140-025-7/

Abstract:

In studies of human genome variation, researchers attempt to identify the DNA sequence differences between our genomes that contribute substantially to variation that can be observed on a physical level (phenotype). Genetic variation can also be used to study population dynamics in human history, and how evolutionary forces have shaped the human genome. These endeavors require comprehensive data resources and computational tools to facilitate directed data generation on the scale necessary for detection of biologically relevant signals. Data management systems are also necessary to store, compare and interpret the collective data masses created by researchers in the field.

This thesis describes the development of computational resources and algorithms for improving the efficiency of studies into human genome variation with a focus on single nucleotide polymorphism (SNP). Issues addressed are: i) *databases* emphasizing quality assurance, improved annotation and portable formats, ii) *assay design* for improved PCR and hybridization reactions through a consideration of DNA secondary structure, iii) SNP *selection* using a combination of *in silico* methods for prediction of the functional impact of SNPs and evidence of positive selection to identify sequence differences that may be disruptive to a living cell, and possibly cause disease, iv) *genome structure* in a comprehensive study into the dynamics of duplicated segments and how they affect SNP genotyping.

Building on previous work around a database for human sequence variation within genes (Hgbase), a scalable database and accompanying portable data formats for all human sequence variation was developed (HGVbase). Following the availability of the complete human genome draft, the sequence variations were layered on top of this scaffold and annotation and search capabilities were vastly improved. Going forward, systems were constructed to capture genotypes, haplotypes, phenotypes and the (complex) relations between them. This new information increases our ability to extract and prioritize biologically interesting subsets of data. In conjunction with new genotyping technology, high-throughput genotyping assay design software facilitated studies encompassing large numbers of SNPs. This capacity was leveraged in creating a set of validated SNP markers in coding regions of genes for subsequent use in studies into disease genetics and population genetics. The genotyping technology was developed further to enable a genome-wide study of single base variants in duplicated segments. The study led to the discovery of a form of sequence variation that we termed "Multi-site Variation" (MSV). This explains a large fraction of the observed increase of predicted polymorphism in duplicated sequence and indicates considerable copy number variation in the human genome. MSVs are able to masquerade as SNPs when genotyped with standard methods in population samples. Unfortunately this is not always revealed by Hardy-Weinberg disequilibrium considerations or mendelian inheritance tests.

Keywords: SNP, polymorphism, database, genotyping, duplication, haplotype

List of papers

- HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources.
  Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ
  Nucleic Acids Res, 2002; 30(1): 387-91
- HGVbase: a curated resource describing human DNA variation and phenotype relationships.
  Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ
  Nucleic Acids Res, 2004; 32: Database issue:D516-9
- DFold: PCR design that minimizes secondary structure and optimizes downstream genotyping applications.
  Fredman D, Jobs M, Stromqvist L, Brookes AJ
  Hum Mutat, 2004; 24(1): 1-8
- Tag n tell: Software that aids haplotype marker selection.
  Fredman D, Brookes AJ
  Manuscript
- The variability of worldwide allele frequencies in functional polymorphism in the human genome.
  Sawyer SL, Fredman D, Mottagui-Tabar S, Kidd KK, Wahlestedt C, Chanock S, Brookes AJ
  Manuscript
- Complex SNP-related sequence variation in segmental genome duplications.
  Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ
  Nat Genet, 2004; 36(8): 861-6. Epub 2004 Jul 11