High-Throughput, Large-Scale SNP Genotyping: Bioinformatics Considerations

Nino Margetic

Centre National de Génotypage, Evry, France

In order to provide a high-throughput, large-scale genotyping facility at the national level we have developed a set of inter-dependent information systems. A combination of commercial, publicly-available and in-house developed tools links a series of data repositories based both on flat files and relational databases providing an almost complete semi-automated pipeline.

Keywords: computing, biology, genotyping.

1. Introduction

It has already been acknowledged [14] that biology has become a data-driven science. One of the simplest definitions of the human genome, in computer science terms, describes it as a string of approximately 3×10^9 letters [13]. The letters A, C, T and G represent the four nucleic acids, constituents of the DNA, which are strung together to make the chromosomes in our cells. When combined into one string, the chromosomes contain the blueprint for a human being – the human genome.

The success of the Human Genome Program, the international initiative to obtain the aforementioned string of letters, together with high-throughput data acquisition systems, has led to unprecedented opportunities to study the aetiology of diseases at the molecular level, leading to the rapid expansion of basic information on genetic variation [17].

Centre National de Génotypage (CNG) was established in 1997 by the French Ministry of Research, with an annual budget of about US$10 million [8], as a national resource with a mission to provide the academic community with large-scale research infrastructure to support applications of genomics in the fields of biology and medicine. The Centre, currently numbering around one hundred staff, develops and perfects methods for genotyping and related techniques for investigation of the genetic basis of human diseases with the principal activity centred on identification of implicated DNA variants. Such knowledge has provided new tools for diagnosing and understanding the causes of such diseases and provides the basis for gene therapy and the development of other treatments.

2. Materials and Methods

2.1. Genotyping

Genetic variability in human populations has arisen from mutations that have accumulated at a slow rate in each generation since the origin of the species. The genome of every individual is variable due to the presence of two copies (maternal and paternal) of most of the genome (the 22 autosomal chromosomes occur in pairs and the sex chromosomes occur as X/X in females and X/Y in males).

Genotyping is the characterisation of this sequence variation within individuals, populations or families and it represents a key approach for exploiting the results of the Human Genome Project — program that initiated the research for identification of the complete base-line sequence of the human genome. It has been used with great success over the last 15 years in investigations leading to the identification of genes responsible for many monogenic orders and has emerged as one of the key tools for the study of multifactorial diseases. Genotyping a large number of individuals rapidly at many different variant sites in the genome is the key step in discovering which of the variants are implicated in a disease.

2.2. Single Nucleotide Polymorphisms

The majority of human sequence variation is due to substitutions that occurred once in the history of mankind, called single nucleotide polymorphisms or SNPs [3, 15, 27]. As a single-point mutation or change in a single base at a specific position, SNPs appear in most cases with only two alleles [2]. Today, the expression is typically applied to variable sites for which the rarer base is present within the population at $> 1\%$ frequency. Consequently, there has been much focus on SNPs as a source of markers in the analysis of complex traits in human genetics [6].

There are several reasons for this: SNPs include most of the DNA variants responsible for effects such as the disease risk or variable drug response. They are abundant and easy to characterize (i.e. genotype), which makes them a plentiful source of markers for association studies and linkage disequilibrium (LD) mapping [7]. Finally, being binary by nature, they naturally lend themselves to automatic analysis.

2.3. Large-Scale SNP Genotyping

In order to be used in the above context, SNP genotyping will, in many cases (e.g. when used for the elucidation of complex diseases where large DNA collections have to be scanned for a given set of markers) require a method of analysis that can provide high-quality and low-cost methodology that is amenable to very high throughput [4]. Although many methods such as DNA microarrays, gel-based and plate-reader based assays for SNP genotyping have been introduced [11, 16, 18, 21], mass spectrometry has been employed as one of the attractive solutions for SNP genotyping, because it can be used to obtain direct and rapid measurement of DNA. At the Centre National de Génotypage in particular, the matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI) has been successfully developed and used for high-throughput SNP genotyping [10, 12, 26].

In the MALDI technique [11, 16, 18, 22] allele-specific products are deposited with a matrix on the metal surface of a target plate. The matrix and the analyte are desorbed into the gas phase with a short laser pulse. Analytes are ionised by collision and accelerated towards the detector. The time-of-flight of a product is directly related to its mass. In this way, each molecule yields a distinct signal. SNP genotyping using MALDI aims to distinguish the different alleles of an SNP by their respective masses. Data collection is done serially (one sample after another), but at a very high rate (less than one second per a measurement).

Although the flexibility (i.e. easy introduction of new SNPs and/or DNAs into the process) and speed of data collection in MALDI satisfy the necessary conditions for the large-scale, low-cost genotyping, the improvement in genotyping methodology in itself is not sufficient. There are many other important questions such as the logistics, sample preparation and liquid handling that need to be addressed and the GOOD assay [20] is routinely used in the SNP production at the CNG.

2.4. Workflow

After obtaining the target DNA collection, a typical SNP study in human cohorts proceeds as follows. Human genome sequence is investigated in the selected target gene in order to validate known (i.e. existing) and/or identify new SNPs during an initial sequencing screen of 32–96 individuals. Alternatively, validated random SNPs may be chosen from public databases to study the relevant genomic region. From the complete set of validated SNPs a representative subset is chosen for the study. However, before the commencement of the proper study, an initial screen is performed. A known set of DNA samples ("CEPH panel" of 384 well-defined family-related individuals that allow the assessment of allele transmission using statistical genetics methodology) is genotyped for the chosen subset of SNPs and the results of genotyping are compared with the results of the SNP validation by sequencing. The consistency of genotyping results is checked in order to detect any recurring problems of genotype scoring (e.g. preferential amplification of an allele). At this point, adjustments in the protocol necessary for production are made. It should be noted that all experimental steps are handled in 384-well DNA plates using the liquid handling robotics and the automatic accumulation mode of the MALDI mass spectrometers.

The proper production, i.e. genotyping for the representative set of validated SNP markers on the DNA collection of interest (i.e. for the disease to be investigated) is then initiated. All data generated by the mass spectrometers is transferred into a database for genetic and statistical analysis.

2.5. Resources

The resources required for this operation are separated into several categories. Bruker Daltronics MALDI mass spectrometers come attached to either a Sun workstation (older models) or a Windows NT/2000 workstation. Data generated by the spectrometers are initially stored by the proprietary software on the network file system in a proprietary file format. Raw data is regularly transferred into the relational database (Sybase System 12.5) on a Sun Enterprise E450 (4CPU, 2GB RAM) running Solaris 8 for further analysis.

Data storage is centralized for the whole network and kept on a fibre-channel storage area network of 2TB capacity. All file systems are in RAID5 disk configuration, apart from the Sybase which uses RAID 0+1. Raw data is kept on the network for a certain period of time in order to be checked, migrated into the database, archived onto DVD-R disks and finally backed up on tape.

Network architecture is central and based on a few servers providing service to a large number of thin-clients, whilst the numerically intensive calculations are performed on a 20-node dual-CPU Linux cluster. Main software components have been written in Java, and are typically run on Windows (DELL PowerEdge 8450, 8 CPU, 4GB RAM running Windows 2000 Advanced Server with Terminal Services), although they are available and run also on any platform for which the Java run-time environment exists (e.g. Linux, Solaris, MacOS).

3. Results and Discussion

In order to manage and track the information flows for the SNP genotyping, several information systems comprising of databases and a number of software packages have been designed and developed. Due to large amounts of data produced on a daily basis, automatic data analysis procedures have been introduced and parallel approaches to data processing procedures have been investigated and partially implemented.

3.1. Laboratory Information Management System

It has long been recognized that the laboratory processes in a typical genotyping environment are highly complex in several dimensions [9]: they give rise to complex datasets, the laboratory processes are very complex and they experience a very rapid rate of change. However, such projects have reasonable operational requirements; the databases are not too big; database activity is moderate; it is not a disaster for the system to crash occasionally; security is not a major concern.

In order to track the laboratory production processes (a task often called *sample tracking* when simple and *workflow management* when complex) strategic decision to employ a Laboratory Information Management System (LIMS) was taken very early on. The key challenge was a decision whether to develop the system using internal resources and reap the benefits of in-depth knowledge of the local work processes or to procure a commercial system which will never be completely adapted to the laboratory procedures in place. We chose a hybrid solution comprising a co-development process with a commercial partner who provided a rapid-prototyping approach to the final system.

The technological basis of the current LIMS is a three-tier object-oriented model built on top of a relational database. The implementation was done in Java with TopLink providing the object-relational connection and Sybase providing the database backend. In addition to managing workflow, the current system holds the complete DNA sample information in a relational database and we are able to access and use the required sample information, using either standard SQL procedures or an object-oriented Java API.

3.2. SNP Discovery

In order to carry out SNP discovery, DNA samples from different individuals must be compared for sequence differences. Sequencing random clones from mixed genomes of many individuals gives an accurate method for large scale SNP discovery, because the sequence is haploid. On the other hand, when performing the SNP discovery in a particular gene, locus-specific polymerase chain reaction (PCR) amplification is very useful. By selecting the appropriate primers, biologically meaningful regions (e.g. exons and promoter regions) can be analysed in a short time and relatively rare SNPs can be identified. However, the analysis of diploid sequences is more complex.

Assuming an easy and accurate method of nucleotide sequence peak value detection and quantification, identification of heterozygous nucleotide peaks becomes possible. Moreover, under such circumstances, SNP identification in pooled (i.e. mixed) DNA samples becomes also a possibility, thereby significantly reducing the workload and allowing simultaneous estimation of allele frequencies. In other words, screening of disease susceptible genes by comparing the genotype of patient and control groups before commencing large scale study becomes a real possibility.

We have developed methodology [24, 25] that incorporates sequence data, finds mixed nucleotides using its own base calling algorithm, tests if these correspond to a real SNP or not, by reading through the nucleotide bases automatically and marks a position as an SNP. At the same time, Genalys, in-house developed software package that runs both on Windows and MacOS, draws the peak ratio of polymorphic alleles of all the samples, so that the variation and pattern of the peak ratio can be easily recognized. The allele frequency is calculated by measuring and drawing the peak ratios of the alleles of each sample, thereby rendering the SNP verification in an easy graphical manner.

3.3. SNP Discovery Database

The aim of the SNP discovery database is twofold. On one hand, it needs to store the results of the SNP discovery process. On the other hand, it has to generate and provide access to the average size sequence-tagged site (STS) markers based on internally generated, as well as publicly available data (i.e. dbSNP, GenBank, LocusLink and UniSTS databases [28]) aligned onto the most recent genome draft sequence.

Our annotated sequence database provides information on SNPs and their effects by means of searching by gene name, accession number, gene definition or a SNP identifier, as well as a graphical representation of the area of interest, with relevant annotations such as SNPs, exons, primers, PCR products, sequencing coverage,...) represented on an auto-resizable image within a web page. It also provides detailed information about all known polymorphisms present in a region of interest, such as the public id of an SNP (if existing), its position and orientation along chromosome, contig and assay fragment, its allele values and amino-acid codon changes (if relevant).

The markers, which contain both internally discovered as well as all presently known public polymorphisms and sequence discordances found during consecutive pair-wise alignment (by using BLAST [1] to position markers along the most recent genome draft sequence), will serve as a template for generation of genotyping probes.

All data is organized according to projects (both internal and collaborative) with appropriate access control lists for different users and groups who can connect to the database over the Internet [5]. Each connection to the database is authenticated and managed by a session with fixed time-out intervals. Software modules used in the above processes are a combination of Perl [23] and C/C++ code typically using high-level ESQL language of Sybase with a web-based user interface for queries and presentation of results [5].

3.4. SNP Production Database

The experimental results of the spectral analysis of mass for each and every DNA sample that is analysed using MALDI are stored in a database as a compressed vector of integer values, together with a set of experimental parameters, such as the expected masses of alleles that define a particular SNP.

Since the labs currently generate around 15000 experiments (thereby generating at least the same amount of genotypes) on a daily basis, it quickly becomes obvious that one needs automatic methods of data analysis. Using knowledge about employed SNPs, several automatic methods for allele calling have been developed [19]. Namely, given that each SNP marker used in the genotyping experiment will have well known alleles, it is simple to calculate the expected masses for a particular experiment.

Armed with that knowledge, one can model, assuming normal distribution, the expected distribution of time of flight for unknown DNA samples. Using such a model, estimation of the DNA sample masses, and therefore the values of the unknown alleles, becomes feasible. The database stores the raw data, genotypes calculated/obtained by various allele calling algorithms, as well as the final (i.e. consensus) genotype after the quality control.

3.5. Parallel Approaches to Data Analysis

Given the organization of the information in the SNP production database (i.e. raw data grouped in “plates” i.e. sets of 384 sample results, i.e. spectra), we have identified four levels of parallel data analysis:

a) “inter-plate”, providing each computing node with a complete plate of data to analyse;

b) “intra-plate”, dividing a single plate into a number of subsets of individual spectra and distributing each subset to a single node for subsequent analysis;

c) “inter-spectrum” providing each node a single spectrum of data for analysis at a time; and

d) “intra-spectrum” providing each node the same spectrum or part of the same spectrum for analysis.

Different parallel paradigms will warrant different approaches: for example, the shared memory machine would probably benefit from the “intra-spectrum” approach and a cluster with a dedicated high speed network is reasonably efficient using “intra-plate” and “inter-spectrum” strategies [19].

4. Conclusion

Automated pipelines and management of downstream data require considerable investment in software engineering. Although not designed from the outset with the ultimate goal of portability, large parts of the infrastructure we developed may provide a core set of utilities to a large genotyping facility that shares at least parts of the technologies and strategies employed at the CNG. Exactly how portable our infrastructure is, remains to be seen due to tradeoffs between customisation and ease of use. Nevertheless, the wealth of experiences gained, as well as the tools developed in the process of building a national genotyping facility, will be a valuable resource to any group wishing to undertake a similar exercise.

References

[1] ALTSCHUL S. F., GISH W., MILLER W., MYERS E. W., AND LIPMAN D. J., Basic local alignment search tool, *J Mol Biol* 1990; 215 (3):403–10.

[2] BROOKES A. J., The essence of SNPs, *Gene* 1999; 234 (2):177–86.

[3] CARGILL M., ALTSHULER D., IRELAND J., SKLAR P., ARDLIE K., PATIL N., SHAW N., LANE C. R., LIM E. P., KALYANARAMAN N., ET AL., Characterization of single-nucleotide polymorphisms in coding regions of human genes, *Nat Genet* 1999; 22 (3):231–8.

[4] CHICUREL M., Faster, better, cheaper genotyping, *Nature* 2001; 412 (6847):580–2.

[5] CNG, The SNP discovery database, `https://db.cng.fr/`, 2003.

[6] COLLINS F. S., GUYER M. S., AND CHARKRAVARTI A., Variations on a theme: cataloging human DNA sequence variation, *Science* 1997; 278 (5343):1580–1.

[7] DAWSON E., ABECASIS G. R., BUMPSTEAD S., CHEN Y., HUNT S., BEARE D. M., PABIAL J., DIBLING T., TINSLEY E., KIRBY S., ET AL., A first-generation linkage disequilibrium map of human chromosome 22, *Nature* 2002; 418 (6897):544–8.

[8] France gives go-ahead for genomics centre on complex diseases, *Nature* 1997; 389 (10):10.

[9] GOODMAN N., ROZEN S., STEIN L. D., AND SMITH A. G., The LabBase system for data management in large-scale biology research laboratories, *Bioinformatics* 1998; 14 (7): 562–74.

[10] GRIFFIN T. J., HALL J. G., PRUDENT J. R., AND SMITH L. M., Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry, *Proc Natl Acad Sci USA* 1999; 96 (11):6301–6.

[11] GUT I. G., Automation in genotyping of single nucleotide polymorphisms, *Hum Mutat* 2001; 17 (6):475–92.

[12] HAFF L. A. AND SMIRNOV I. P., Multiplex genotyping of PCR products with MassTag-labeled primers, *Nucleic Acids Res* 1997; 25 (18):3749–50.

[13] HAUSSLER D., A Brief Look at Some Machine Learning Problems in Genomics, *Proceedings of the tenth annual conference on computational learning theory*; ACM Press; 1997, pp. 109–13.

[14] HEATH L. S., RAMAKRISHNAN, N., The emerging landscape of bioinformatics software systems, *Computer* 2002; 35 (7):41–5.

[15] KRUGLYAK L., Prospects for whole-genome linkage disequilibrium mapping of common disease genes, *Nat Genet* 1999; 22 (2):139–44.

[16] LANDEGREN U., KAISER R., CASKEY C. T., AND HOOD L., DNA diagnostics–molecular techniques and automation, *Science* 1988; 242 (4876):229–37.

[17] LANDER E. S., LINTON L. M., BIRREN B., NUSBAUM C., ZODY M. C., BALDWIN J., DEVON K., DEWAR K., DOYLE M., FITZHUGH W., ET AL., Initial sequencing and analysis of the human genome, *Nature* 2001; 409 (6822):860–921.

[18] MEIN C. A., BARRATT B. J., DUNN M. G., SIEGMUND T., SMITH A. N., ESPOSITO L., NUTLAND S., STEVENS H. E., WILSON A. J., PHILLIPS M. S., ET AL., Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation, *Genome Res* 2000; 10 (3):330–43.

[19] PARROT C., RENAULT E., AND MARGETIC N., Coarse Grain Parallelization of Single Nucleotide Polymorphism Identification on the Grid, *In Preparation* 2004.

[20] SAUER S., LECHNER D., BERLIN K., LEHRACH H., ESCARY J. L., FOX N., AND GUT I. G., A novel procedure for efficient genotyping of single nucleotide polymorphisms, *Nucleic Acids Res* 2000; 28 (5):E13.

[21] SMITH L. M., Automated DNA sequencing: a look into the future, *Cancer Detect Prev* 1993; 17 (2):283–8.

[22] SMITH L. M., The future of DNA sequencing, *Science* 1993; 262 (5133):530–2.

[23] STAJICH J. E., BLOCK D., BOULEZ K., BRENNER S. E., CHERVITZ S. A., DAGDIGIAN C., FUELLEN G., GILBERT J. G., KORF I., LAPP H., ET AL., The Bioperl toolkit: Perl modules for the life sciences, *Genome Res* 2002; 12 (10):1611–8.

[24] TAKAHASHI M., MATSUDA F., MARGETIC N., AND LATHROP M., Automated identification of single nucleotide polymorphisms from sequencing data, *Proceedings of IEEE Computer Society Bioinformatics Conference*; Aug 14–16, 2002; Stanford, CA; IEEE Computer Society; 2002. pp. 87–96.

[25] TAKAHASHI M., MATSUDA F., MARGETIC N., AND LATHROP M., Automated identification of single nucleotide polymorphisms from sequencing data, *Journal of Bioinformatics and Computational Biology* 2003; 1 (2):253–65.

[26] TANG K., FU D. J., JULIEN D., BRAUN A., CANTOR C. R., AND KOSTER H., Chip-based genotyping by mass spectrometry, *Proc Natl Acad Sci USA* 1999; 96 (18):10016–20.

[27] WANG D. G., FAN J. B., SIAO C. J., BERNO A., YOUNG P., SAPOLSKY R., GHANDOUR G., PERKINS N., WINCHESTER E., SPENCER J., ET AL., Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, *Science* 1998; 280 (5366):1077–82.

[28] WHEELER D. L., CHURCH D. M., FEDERHEN S., LASH A. E., MADDEN T. L., PONTIUS J. U., SCHULER G. D., SCHRIML L. M., SEQUEIRA E., TATUSOVA T. A., ET AL., Database resources of the National Center for Biotechnology, *Nucleic Acids Res* 2003; 31 (1):28–33.

Received: June, 2004
Accepted: June, 2004

Contact address:
Nino Margetic
Centre National de Génotypage
2 rue Gaston Cremieux
CP5721
91057 Evry
France
Phone: +33 1 60 87 84 20
Fax: +33 1 60 87 84 85
e-mail: nino@cng.fr

NINO MARGETIC graduated in theoretical physics at the University of Zagreb (Croatia) but his interest was always in the cross-section of information technology and biomedicine. In 1988 he came to UK and spent next few years working on various research projects at the Dept. of Medical Physics and Bio-Engineering of the University College London. In 1993 he moved to Oxford University to take the position of the Head of Computing at The Wellcome Trust Centre for Human Genetics. In 2000 he joined the French National Centre for Genotyping in Evry as the Head of Information Technology. His current interests lie in bioinformatics, high performance computing, systems analysis and security.