XTENS: a neuroblastoma copy number variation repository at the BIT-Gaslini biobank

Izzo M(1,2), Cangelosi D(1), Pezzolo A(3), Morini M(1), Varesio L(1)

(1) Laboratorio di Biologia Molecolare, Istituto Giannina Gaslini, Genova
(2) Oxford e-Research Centre, University of Oxford, Oxford
(3) Laboratorio di Oncologia, Istituto Giannina Gaslini, Genova

Motivation

Neuroblastoma is the major paediatric solid tumour. Unfortunately, about 50% of high risk patients are refractory to treatment and die, demanding new prognostic indicators for improving and personalising therapy. Biomarkers discovery depends on mining molecular, biological, and clinical data, thus making integrated biobanks the ideal collectors of this heterogeneous information and the essential structure for this task.

The Biobank Integrating Tissue-genomics of Gaslini Institute, Genova, Italy (BIT-Gaslini) has adopted XTENS [1,2], a web-based data management platform to handle the sample management workflow and integrate it with the patients' clinical records and molecular data. As of December 2015 the biobank collected over 3700 primary samples (2140 tissue and 1600 fluid) and 1650 derivatives (1030 DNA and 620 RNA). Besides sample management information - such as specimens characterisation, aliquot deliveries to external laboratories and/or centres, and quality control reports - XTENS stored clinical and molecular details for over 900 neuroblastoma patients as retrieved from the National Neuroblastoma Registry and 175 microarray profiles from primary tumour tissues. We have decided to test the capability of the platform as a repository for Copy Number Variations (CNVs, gain/amplification and loss of DNA) data obtained from oligonucleotide array-Comparative Genomic Hybridisation (aCGH) analyses routinely executed at Gaslini Institute. aCGH data are essential for patients' assignment to specific treatments' protocols and they are a valuable source of information for mining the biology of neuroblastoma and the identification of new prognostic indicators.

Methods

To identify CNVs (numerical and segmental chromosomal alterations) in neuroblastoma samples, we used aCGH, that was performed using the Agilent Human Genome CGH microarrays 180K following the manufacturer's protocol (Agilent Technologies, Santa Clara, California, USA). Slides were scanned using a G2565BA scanner, and analysed using Agilent CGH Analytics software Genomic Workbench 7.0 (Agilent Technologies Inc.) with the statistical algorithm ADM-1 and a sensitivity threshold of 6.0. Probes with a log ratio value greater than 2 were considered as amplified.
To automate the CNV upload to XTENS, we have designed a novel page on the website where the operator can upload the aCGH processed files containing all the analysis metadata and the full list of found aberrations (gain and loss analysis of DNA). A Node.js server-side script parses the uploaded file, and stores all the relevant information in the database. The raw data files are then uploaded using the standard XTENS data management interface and are stored on a distributed file system managed by the iRODS data grid middleware, transparently integrated with XTENS through a REST interface.

Results

As of March 2016, we have uploaded 345 aCGH analyses of neuroblastoma tissue samples run between 2007 to 2015 on XTENS 2, for a total of 6511 aberration calls (3947 gain/amplifications, 2564 losses). Each aberration is characterised by type (amplification or deletion), location (chromosome, cytogenetic band, start and end position in the reference genome), and the list of affected genes and miRNAs. Overall, there are 2.18 million gene annotations from 21,919 different protein-coding genes, and 107,400 miRNAs annotations from 1,040 different miRNA genes. All the metadata concerning the copy number variations are stored using the binary JSON format (JSONB) of PostgreSQL that is natively supported by the XTENS server environment, running on a Node.js server.

JSONB provides a more efficient schemaless solution than traditional metadata storage in relational Entity-Attribute-Value (EAV) catalogues. Queries composed from the website query builder tool to retrieve subjects based on gene annotations for the aberration calls are executed in 260.9 ± 2.1 ms without indexing and 35.9+2.1 ms if a General Inverted (GIN) Index is created for the JSONB metadata field. The execution times when GIN is adopted, below the 50 ms limit usually required by data-intensive web applications, evidence that XTENS is a suitable web-based solution to manage clinical, biological, and genomic metadata for medical research collaborations. The data uploader we have designed can be extended to other data sources. We plan to support automatic upload of circulating miRNA expression profiles obtained by blood plasma samples, in order to provide a broader database for identifying novel prognostic factors in neuroblastoma. Initial data mining applications will be presented.

References

[1] Izzo, Massimiliano, et al. "XTENS-A JSON-Based Digital Repository for Biomedical Data Management." Bioinformatics and Biomedical Engineering. Springer International Publishing, 2015. 123-130.

[2] Izzo, Massimiliano. "Results: XTENS 2, A JSON-Compliant Repository." Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management. Springer International Publishing, 2016. 61-88.