Cytogenet Genome Res 115:205–214 (2006)
DOI: 10.1159/000095916

Cytogenetic and Genome Research

Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome

J. Zhang[a] L. Feuk[a] G.E. Duggan[a] R. Khaja[a] S.W. Scherer[a, b]

[a]The Centre for Applied Genomics and the Program in Genetics and Genomic Biology, The Hospital for Sick Children and [b]Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario (Canada)

Manuscript received 23 March 2006; accepted in revised form for publication by A. Geurts van Kessel, 15 May 2006.

Abstract. The discovery of an abundance of copy number variants (CNVs; gains and losses of DNA sequences >1 kb) and other structural variants in the human genome is influencing the way research and diagnostic analyses are being designed and interpreted. As such, comprehensive databases with the most relevant information will be critical to fully understand the results and have impact in a diverse range of disciplines ranging from molecular biology to clinical genetics. Here, we describe the development of bioinformatics resources to facilitate these studies. The Database of Genomic Variants (http://projects.tcag.ca/variation/) is a comprehensive catalogue of structural variation in the human genome. The database currently contains 1,267 regions reported to contain copy number variation or inversions in apparently healthy human cases. We describe the current contents of the database and how it can serve as a resource for interpretation of array comparative genomic hybridization (array CGH) and other DNA copy imbalance data. We also present the structure of the database, which was built using a new data modeling methodology termed Cross-Referenced Tables (XRT). This is a generic and easy-to-use platform, which is strong in handling textual data and complex relationships. Web-based presentation tools have been built allowing publication of XRT data to the web immediately along with rapid sharing of files with other databases and genome browsers. We also describe a novel tool named eFISH (electronic fluorescence in situ hybridization) (http://projects.tcag.ca/efish/), a BLAST-based program that was developed to facilitate the choice of appropriate clones for FISH and CGH experiments, as well as interpretation of results in which genomic DNA probes are used in hybridization-based experiments.
Copyright © 2006 S. Karger AG, Basel

During the last few years numerous studies have identified a large number of copy-number variants (CNVs) and other structural variants in the human genome (Iafrate et al., 2004; Sebat et al., 2004; Sharp et al., 2005; Tuzun et al., 2005; Conrad et al., 2006; Hinds et al., 2006; McCarroll et al., 2006). A CNV is a term collectively used to describe gains and losses of DNA sequences >1 kb in length (reviewed in Feuk et al., 2006a) and the relative high frequency of CNVs in the human genome has generated considerable excitement in the field (Carter, 2004; Check, 2005; Lee, 2005; Eichler, 2006). Since these changes can be hundreds of kilobases in size they can have a direct effect on transcription and transcriptional regulation, which in turn may be a cause for disease susceptibility and phenotypic variation. There are currently more than 1,000 CNVs described in literature, but this represents only a small fraction of all CNVs expected to exist in the human population.

The main approach used to identify CNVs to date has been array-based comparative genomic hybridization (CGH) (Kallioniemi et al., 1992; Pinkel et al., 1998). The

Supported by Genome Canada, the McLaughlin Centre for Molecular Medicine, and the Hospital for Sick Children Foundation. L.F. is supported by the Swedish Medical Research Council. S.W.S. is an investigator of the Canadian Institutes of Health Research and an International Scholar of the Howard Hughes Medical Institute.

Request reprints from Stephen W. Scherer
The Centre for Applied Genomics
Program in Genetics and Genomic Biology
The Hospital for Sick Children
MaRS Centre – East Tower, 101 College Street, Room 14-701
Toronto, Ontario, M5G 1L7 (Canada)
telephone: +1-416-813-7613; fax: +1-416-813-8319
e-mail: steve@genet.sickkids.on.ca

J.Z. and L.F. contributed equally to this work.

© 2006 S. Karger AG, Basel
1424–8581/06/1154–0205$23.50/0

Accessible online at:
www.karger.com/cgr

development of the array-CGH technology and other oligonucleotide-based platforms (Feuk et al., 2006a) has important implications for both research and clinical diagnostics laboratories. Specific arrays targeting the micro-deletion and duplication syndrome regions are now commercially available for diagnostic purposes, alongside whole-genome coverage arrays which give a global view of genome imbalances. The introduction of these high-throughput technologies into diagnostic and clinical settings, and possibly all genetic research studies (Feuk et al., 2006b), allows scanning for rearrangements at an unprecedented resolution, but at the same time creates challenges in terms of data handling, interpretation, and validation.

For each sample screened using a whole-genome coverage array-based platform, anywhere between 5 and 300 variants might be found (depending on which of the currently available platforms was used and the stringency of cutoffs applied). This data must then be stored in an appropriate way, and regions should be validated in some way and then prioritized for further analysis. In order to deal with some of these issues we have developed new bioinformatics resources. The first is the public database called 'The Database of Genomic Variants', with the aim of cataloguing all CNVs described in the literature in a format accessible to medical geneticists and molecular biologists alike. The database was built using a new platform for data handling and sharing called BioXRT, which in turn is based on the Cross-Referenced Tables (XRT) data model. For maximum translational impact, it is necessary to establish online databases to facilitate information sharing within a research community. For example, for collections of locus-specific disease mutations alone, there were 262 databases as of 2002 (Claustres et al., 2002); and the 2005 updated Nucleic Acids Research online Molecular Biology Database Collection included 719 databases, an increase of 171 over the previous year, and this listing was far from exhaustive (Galperin, 2005). Online databases provide many advantages, such as wide-accessibility, advanced querying, fast retrieval and persistent referencing. Despite varied content and architectures, most of these databases are functionally similar; instead of in-house development of such databases, if a generic and easy-to-use platform could be used a significant amount of duplicate effort would be avoided. The BioXRT platform was developed with the aim to provide a lightweight generic solution for housing and publishing biological data. Besides the prototypic Database of Genomic Variants described here, BioXRT has now been adapted to many other online databases that are widely utilized by the genetics community.

Lastly, we present a tool named eFISH (electronic fluorescent in situ hybridization), which is a BLAST-based approach to predict the results of FISH and other hybridization-based assays. The eFISH program was created to facilitate the selection of genomic probes and analysis of results for FISH experiments, but can be used in the interpretation of results from any DNA hybridization-based approaches.

Results and discussion

The Database of Genomic Variants

Following the initial reports on global distribution of CNVs in the human genome (Iafrate et al., 2004; Sebat et al., 2004), it was apparent that the ~300 regions described represented only a small fraction of all the CNVs in the human genome. Clearly, there was demand for a database where information on structural variants in general, and CNVs in particular, could be stored and accessed by the research community. Not only would this simplify the comparison of new datasets to what has already been published, but would also allow the compilation of up to date summary statistics and analysis of this type of variation. There are currently two existing databases which focus on collecting data on submicroscopic structural variation; The Database of Genomic Variants (http://projects.tcag.ca/variation/) (Iafrate et al., 2004) described for the first time in detail here, and the Human Structural Variation Database (http://humanparalogy.gs.washington.edu/structuralvariation/) (Sharp et al., 2005).

The Database of Genomic Variants currently has the aim of cataloguing all submicroscopic structural variants >1 kb in size identified in control individuals that have been documented in peer-reviewed literature. The majority of these are CNVs, but there are also inversion breakpoint regions. The main goal of the database is to provide a user-friendly resource for the scientific and medical genetics community. The DECIPHER (DatabasE of Chromosomal Imbalances and Phenotype in Human using Ensembl Resources; see http://www.sanger.ac.uk/PostGenomics/decipher/) initiative, for example, uses this as the source of its genomic variant data track.

The database can be searched by a genomic feature, such as gene name, clone name or DNA sequence (Fig. 1). Alternatively, the contents of the database can be browsed in either table format or in a genome browser displaying relevant information (e.g. gene, cytogenetic location, segmental duplication, genomic clone, etc). Each entry in the table is linked to a page that contains more detailed information about the locus in question (Fig. 2). One recent update to the database is that for any variant identified in the HapMap sample set, we also include information on which samples that were found to carry a specific variant. This serves two purposes; first, knowing that a specific sample carries a specific variant makes it useful as a control sample when testing new methods or trying to find the sensitivity and accuracy of a certain method, and second, it facilitates further analysis of structural variants in relation to other data available for the HapMap samples, including SNP data and gene expression data.

Data currently in The Database of Genomic Variants

There are currently 1,267 regions of structural variation in the database. Of these, 1,207 have been reported as CNVs, 37 as inversion breakpoints and the remaining 23 as regions containing both CNVs and inversion breakpoints. In total, the 1,267 variants cover 143 Mb of genomic sequence. The

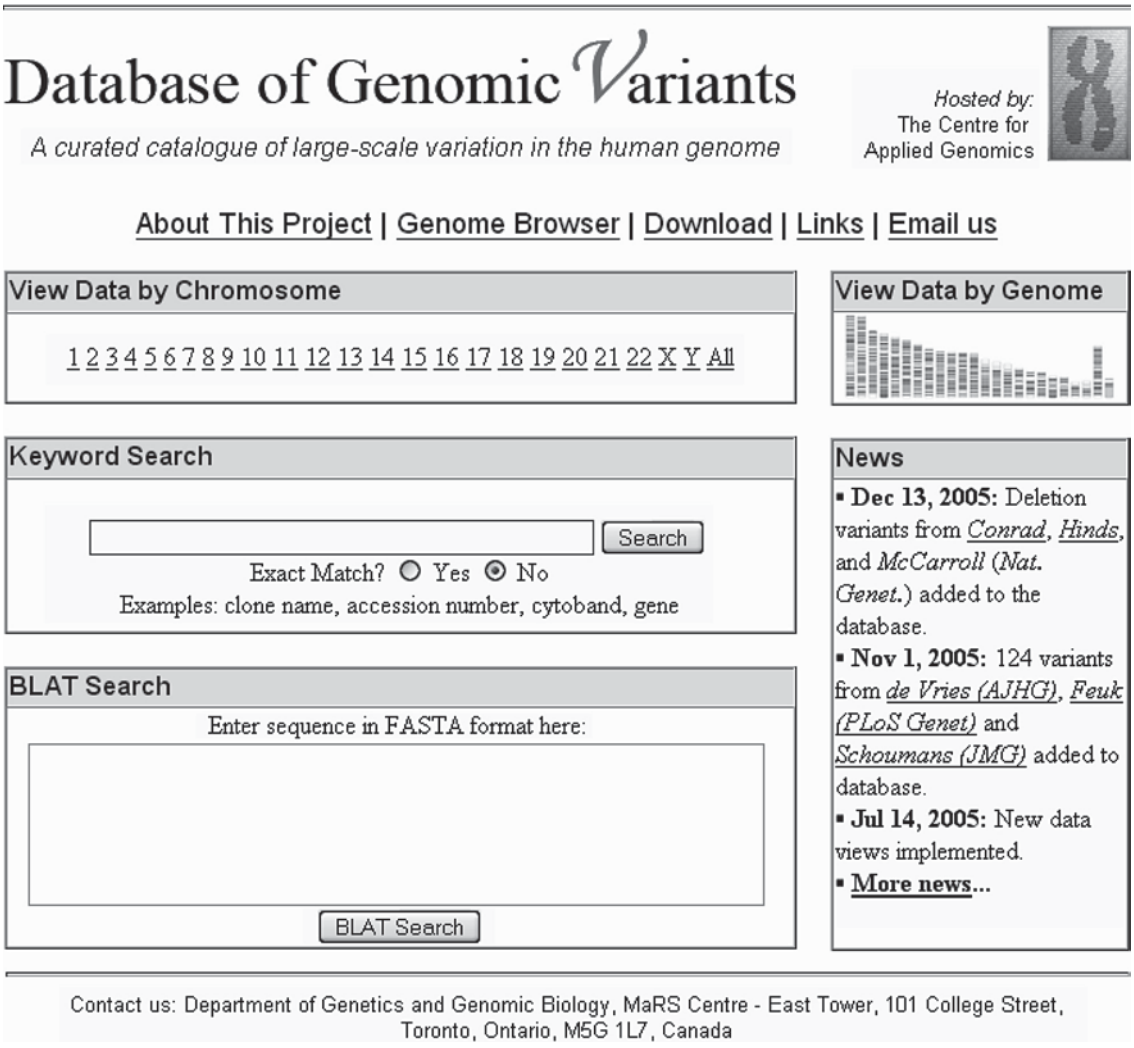
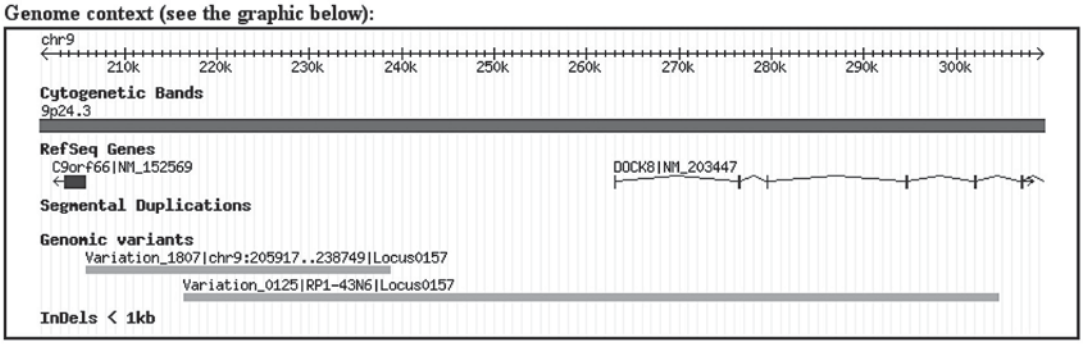


Fig. 1. The Database of Genomic Variants. The home page of the Database of Genomic Variants is shown. The data can be viewed in table format by clicking on a chromosome of interest. Alternatively, the database can be searched using a keyword, which could be a gene name, clone name, or cytogenetic band. If the region entered overlaps with a genomic variant, it can be viewed in a genome browser or in the context of CNVs overlapping the region. The database can also be queried using DNA sequence. The search is based on BLAT to identify matching regions. In the top right corner a genome-wide overview of structural variants in the genome can be viewed. The 'download' link can be used for downloading the entire contents of the database for incorporation into other browsers.

average size of entries is 118 kb. This is likely not a reflection of the true size distribution of CNVs, as it is currently influenced by the bias in how they were assessed. The majority of the variants in the database have been identified by either CGH arrays or by using SNP data to detect deletions by identifying regions showing Mendelian inconsistencies, null genotypes or Hardy-Weinberg disequilibrium. The highest resolution of the array-CGH studies published to date is ~35 kb, but most arrays do not reach that level of resolution. Looking specifically at regions in the database identified by array-CGH, the average size of regions is 289 kb. The use of SNP data is less biased in terms of the size of regions that can be detected, but instead is biased in that only deletions can be detected (and not duplications). At present, the database contains data from 36 research papers. It is important to point out that the database only reports on the regions described in each study, regardless of whether they have been validated by independent approaches. All methods currently used for identification of structural variation will generate some false-positive regions, and since these are included in the published datasets some regions represented in the database are not true structural variants. As more data is published, it will become more clear which of the variants are common polymorphisms and which regions are rare mutations or false positives. Of all 1,267 regions, 280 have been reported by two or more separate stud-

Locus: Locus0157

Genome context (see the graphic below):



Variation: Variation_0125

Landmark: RP1-43N6 (Genome Browsers: TCAG Segmental Duplication , UCSC , Ensembl)
Variation Type: CopyNumber
Overlap with TCAG Segmental Duplication: No
Gap within 100k: No
Known Genes: DOCK8
Method: Array CGH
Reference: Iafrate et al. (2004)
Pub Med ID: 15286789
Frequency Information:
Subject Cohort: Control
Sample Size: 55 in total (39 unrelated healthy individuals and 16 individuals with previously characterized chromosomal imbalances)
Normal Gain: 1
Normal Loss: 1
Total Gain/Loss: 2

Variation: Variation_1807

Landmark: chr9:205,917..238,749 (Genome Browsers: TCAG Segmental Duplication , UCSC , Ensembl)
Variation Type: CopyNumber
Overlap with TCAG Segmental Duplication: No
Gap within 100k: No
Method: Null genotypes
Individual: NA18576
Reference: McCarroll et al. (2005)
Frequency Information:
Subject Cohort: Control
Sample Size: 269 HapMap individuals

Fig. 2. Detailed information about a specific variant. An example of a page displaying detailed information for a locus harboring a CNV is shown. A simple graphical overview is provided for genomic context information. This includes chromosomal position, genes and segmental duplications. If the same region has been identified in several studies, each finding is assigned a unique variation ID. For each entry, there is a link to Pubmed to see the abstract of the article from which the information is extracted. There is also detailed information about study cohort, sample size and methodology.

ies. The region reported in most papers is the *defensin* gene cluster on chromosome 8, which has been identified as polymorphic in nine different studies.

The most obvious link between copy number changes and their effect on gene expression is when the CNV directly overlaps a gene. The 1,267 CNVs currently in the database overlap with a total of 1,298 genes, and of these 846 are contained entirely within the boundaries of the regions reported to contain CNVs. It is important to point out that in cases where genomic clones on array-CGH experiments are reported to show copy number variation, it is impossible to determine the exact boundaries of the variant without performing further experiments. A more detailed analysis of the genes present in CNV regions shows that certain gene ontology categories are found at higher frequencies than expected by chance. The biological processes most significantly overrepresented in CNV regions are shown in Table 1. Genes important for interaction with the environment and defense against pathogens seem to be very variable in copy number between individuals. These genes may be amenable to copy number variation as a means to quickly adapt to external threat and changing surroundings. There are several examples where gene copy number affects response to exposure to common drugs, e.g. increased copy

Table 1. GO terms describing biological processes for genes within CNVs

GO ID	GO term	Observed	All gene	Expected	Ratio
GO:0007565	pregnancy	13	43	1.1	11.881
GO:0006805	xenobiotic metabolism	8	28	0.7	11.228
GO:0009613	response to pest, pathogen or parasite	5	25	0.6	7.859
GO:0006952	defense response	12	77	2.0	6.124
GO:0007600	sensory perception	38	486	12.4	3.073
GO:0006968	cellular defense response	5	67	1.7	2.933
GO:0008152	metabolism	15	371	9.4	1.589
GO:0005975	carbohydrate metabolism	9	224	5.7	1.579
GO:0006955	immune response	12	308	7.8	1.531

GO terms for biological processes that are significantly overrepresented for genes in CNV regions are shown. Only categories with more than five genes observed and GO level 2–5 are included.

number pf CYP2D6 leads to faster metabolizing of debrisoquine (Ingelman-Sundberg, 2002). A more recent example shows how CNVs can play important roles in defense against pathogens, as exemplified by carriers of extra copies of *CCL3L1* having increased resistance against HIV (Gonzalez et al., 2005).

Scanning for copy number changes will become a routine part of many monogenic disease studies, as well as part of the study design to identify complex disease genes (Feuk et al., 2006b). The number of CNVs identified will therefore increase on a regular basis. There are also on-going efforts to identify all large CNVs in the HapMap samples, using multiple array-based platforms (Freeman et al., 2006). Once a good dataset exists for control samples, it will facilitate interpretation of data from studies in patient cohorts. The Database of Genomic Variants will continue to be updated as new studies are published, and will aim to provide the best possible resource for researchers working in the field of structural variation. All data will also continue to be made available in standardized files for incorporation into other genome databases.

The Cross-Referenced Tables (XRT) data model

The Database of Genomic Variants is based on an open source database platform called BioXRT (http://projects.tcag.ca/bioxrt). It was designed to be a generally applicable platform for databases in the biomedical research field. Although the content and the architecture among most databases in biomedicine are quite different, many of them do share a common cycle of tasks including: (i) collecting and curating data from different sources such as public databases, scientific literature and internal laboratory results; (ii) integrating this information using an appropriate model; (iii) loading data into a relational database, and (iv) providing a web-based interface for users to query and browse the data in a read-only fashion. When updating with new data, they follow a repetitive cycling pattern. These common tasks make it practically feasible to build and maintain a biology database using a generic platform. By avoiding in-house development, a generic approach can prevent unnecessary duplication of effort.

A data model is a description of the data for a particular subject area, how they are defined and organized, and how they relate to one another. It includes the data items and their relationship. Taking the large diversity and fast emerging pace of biological data (in this case structural variation data) into account, a broadly applicable and extensible data model is essential for a generic approach to build biological databases. With this in mind, we developed a data modeling system termed Cross-Referenced Tables (XRT). For simplicity, the XRT model uses tab-delimited flat files (i.e. text tables) as a basic modeling unit to keep data items. A text table structured by a field/value convention is the most natural format and is commonly used in many public biological data sources. It is capable of storing arbitrary data items by simply adding new fields, and is generally applicable to any textual information. Additionally, no special tool is needed to prepare or parse a text table. However, it also has some caveats, such as lack of referencing and constraints, and difficulty in modeling complex data with a single table. To overcome these limitations, we applied several rules to the text tables, basically, injecting mechanisms to handle relationships among data items.

XRT is a simple file schema, which encapsulates data in an object hierarchy with arbitrary attributes and relationships. It organizes data into different classes according to its biological meaning (e.g. gene, Gene Ontology term and OMIM entry). Each class has as many attributes as necessary to describe the properties of its elements, and its attributes can be settled in one or more XRT tables. An XRT table is a tab-delimited flat file. The first line specifies the attribute names, while the following lines contain the actual attribute values for elements with each element having a unique identifier (ID, primary key in database terminology). Special attributes called P_ID for parent ID and C_ID for child ID keep track of references between data elements (in database terms, these relationships are known as foreign keys). The relationship can be one to one, one to many, many to one, or many to many. The XRT class name is defined as the string before the first dot (.) of the XRT table file name, for example, the class name of XRT table Transcript.main.xrt is Transcript. Online documentation on de-

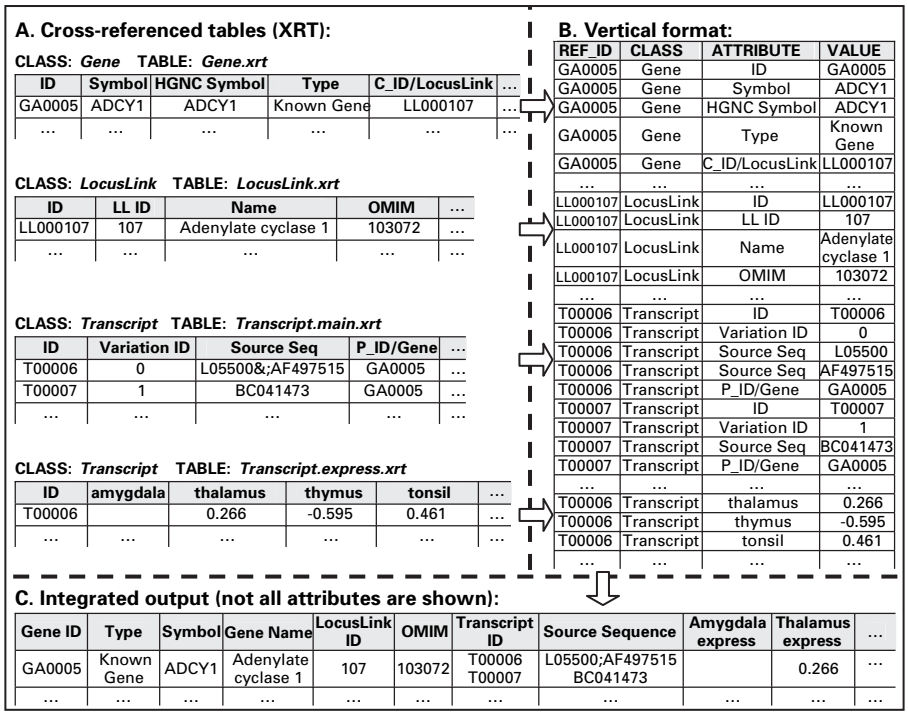


Fig. 3. XRT example and its format transformation. (**A**) Four cross-referenced tables, XRTs; (**B**) the vertical format of the original XRTs; (**C**) integrated output of the source XRT tables. The XRT model is very flexible for changes and adding new data types. New table (with new attributes) can be added to an existing class later on without touching any existing table(s), and different tables (even for the same class) can be generated and maintained separately as long as appropriate referencing is kept. This feature allows database expansion of new data and facilitates integration of data from scattered sources.

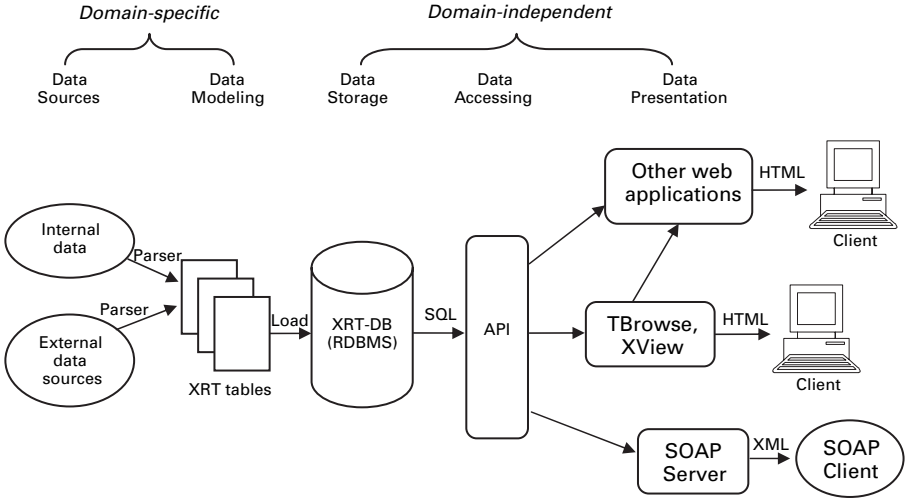


Fig. 4. Overview and data flow of the BioXRT platform. To build an online database using the BioXRT platform, one starts with modeling domain specific data into XRT. XRT tables can then be loaded into a relational database. Data accessing API serves as a bridge between web applications and the XRT-DB. Two standard web tools (TBrowse and XView) provide user-friendly interfaces for data querying and browsing. With relative ease, users with special needs can develop their own web applications which access XRT-DB via API, TBrowse or XView. XRT data accessing through a SOAP server provides a program-friendly interface for robust data integration.

tailed XRT specification is available at http://projects.tcag.ca/bioxrt/xrt_spec.html. An example XRT model of the gene-centric data is shown in Fig. 3A, where data is organized into three classes, and is physically contained in four XRT tables. Figure 3 also illustrates how XRT tables can be transformed into a unified vertical format (Fig. 3B) for easier data storage and manipulation, and data in this vertical format can later be converted back to a human readable table (Fig. 3C), which integrates the original XRT tables. The XRT model used in the Database of Genomic Variants is available at http://projects.tcag.ca/variation/download.html.

Implementation of the BioXRT platform. While having data modeled in XRT, we developed the BioXRT platform to provide XRT data storage and web presentation. An overview of the BioXRT platform is shown in Fig. 4. It is implemented in Perl, and is built exclusively upon open source components such as MySQL and BioPerl (http://www.bioperl.org). These choices reflect the easiest configuration to install, however, the schema and configuration are applicable to any permutation of platform factors and support will be provided to labs choosing to implement it in a different setup. To build a particular online database using

BioXRT, first of all, we needed to model the data in XRT, i.e. define classes and their relationships, such as the example model shown in Fig. 3. Then, data from either internal results or external sources is converted into XRT tables. A Perl script named 'bulk_load_xrt.pl' can later transform all XRT tables to the vertical format, and load them into the database and build the requisite indices. For the default implementation, the MySQL database management system was used to host XRT data because of its open source status, and its superior performance in read-mostly environments. Any SQL92 compliant database engine could be used with relative ease.

In order to provide an efficient, user-friendly and widely accessible interface to an XRT database, we have implemented a web application called TBrowse. The browser accesses the XRT database via a standard connection such as the Perl DBI, with an XRT-specific API which translates the data requests into appropriate SQL queries, and converts results into HTML tables. Several options can be customized to the output table (e.g. table title, column headers, and hyperlinks). For additional details about the table configuration, an online tutorial is provided at http://projects.tcag.ca/bioxrt/tutorial. In addition to browsing pre-defined tables, TBrowse also functions as a data retrieval tool, users can perform keyword searches, select columns to show and filter records on certain column(s) to obtain their data of interest. Output of TBrowse can be exported and downloaded in several formats: tab-delimited flat file, XML and Microsoft Excel file. Besides the interactive web interface, URL-based access to the XRT database is also supported in TBrowse. Due to the simplicity of a two-dimensional table, TBrowse is not entirely ideal in displaying data of complex structure. Another web application called XView was implemented, which can recursively handle (theoretically) unlimited levels of XRT relationship in a hierarchical structure. XView presents data in an easy-to-understand hierarchical tree reflecting the logical relationship of XRT data items (an example of XView output is shown in Fig. 2). Similar to TBrowse, the tree structure is defined in a user-managed configuration file.

BioXRT sample and proof of concept databases. Besides The Database of Genomic Variants mentioned above, BioXRT has also been successfully applied in several of our online projects in a wide range including: the Human Chromosome 7 Annotation Project (Scherer et al., 2003) (http://www.chr7.org), the Genome Segmental Duplication Project (Cheung et al., 2003) (http://projects.tcag.ca/humandup), the Autism Chromosome Rearrangement Database (Xu et al., 2004) (http://projects.tcag.ca/autism/), the Genomic Clone Database (http://projects.tcag.ca/gcd/, Zhang et al., unpublished), and the Lafora Progressive Myoclonus Epilepsy Mutation Database (Ianzano et al., 2005) (http://projects.tcag.ca/lafora/). Within the chromosome 7 database, BioXRT is the primary harness for gene-centric data that are derived from diverse sources. There are currently 21 XRT tables representing 18 classes. Each table can be maintained individually, even by different curators. When new data needs to be integrated, it is simply converted into XRT format while referencing existing data correctly, and the configuration file is updated. After being uploaded, the new data gets integrated automatically, with no need for database structure changes or program modifications.

With the BioXRT platform available, setting up an online biological database becomes significantly easier, with solutions for database schema design programs for data query and presentation already built in. The only thing users need to do is to model their data in XRT, which is like a simplified version of the relational database schema design, since only the logical design phase is involved, and no normalization or other physical design concerns are required.

Biological data is rarely static due to the fast pace of new data emergence, change is unavoidable no matter which modeling tool has been used. This means that considerable effort is needed for data re-modeling. The advantages of XRT's simplicity stand out while handling data model changes, which actually was the initial motivation of the BioXRT project. Modification (adding, changing and deleting) of the XRT classes and/or their attributes can be easily done through the updating of XRT tables. More importantly, due to the content-independency of the BioXRT platform, no effort is needed for database or program re-engineering to accommodate the updated XRT model. Thus, the XRT model is highly flexible and broadly applicable, and the reusability of the BioXRT platform is maximized.

We believe the light-weight approach presented here is an attractive solution for biological data sharing. This open source initiative was developed with two missions; first, to allow biologists the ability to quickly bring their research data online, where data is widely accessible throughout the world, and secondly, to provide outside developers the opportunity to contribute their own ideas and requirements to enhance BioXRT's ability to accomplish biological goals.

eFISH

With the exception for regions commonly interrogated by FISH in diagnostic labs and in targeted research studies, the majority of genomic clones including those used in the sequencing and assembly of the human genome reference sequence, have not been mapped in a standardized way. Choosing suitable clones for FISH experiments can therefore be problematic, as many clones give rise to multiple hybridization signals, making the FISH results difficult or impossible to interpret. One of the problems with the large amounts of data being generated using array-CGH is validation of the results. FISH is one common approach for validation of clone based array results, and ideally the same clone as the one giving rise to a signal on the array should be used. The fact that many CNVs overlap regions of segmental duplications (low copy repeats) (Fredman et al., 2004; Iafrate et al., 2004; Sharp et al., 2005) further complicates analysis of FISH results for these regions.

In order to simplify the choice of clones for FISH experiments and facilitate the interpretation of results where multiple hybridization signals appear, we have developed an in silico FISH simulation program called eFISH. The input sequence can be any clone or region that can be anchored to

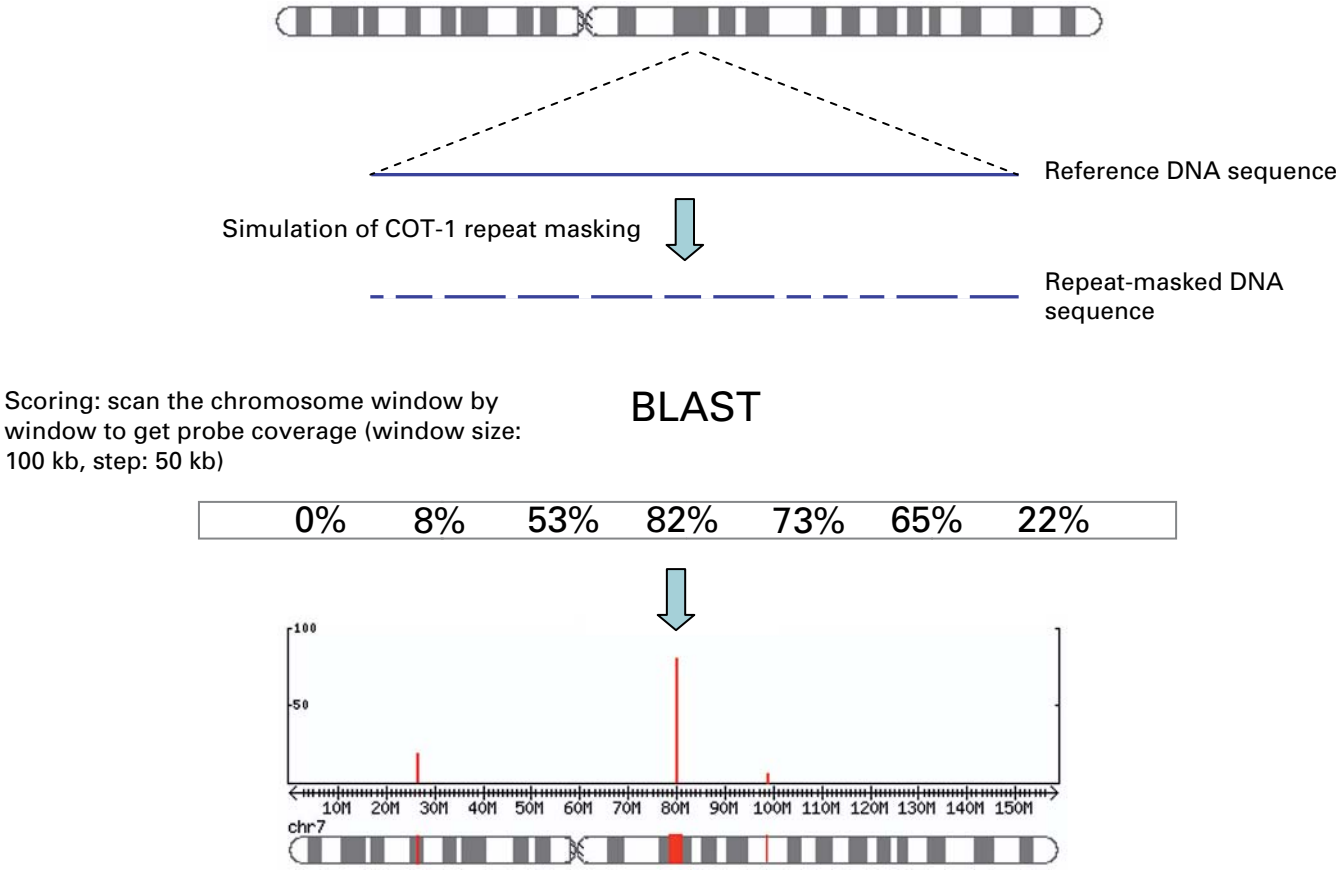


Fig. 5. Overview of the eFISH analysis process. Input sequences are first repeat-masked and then BLASTed against the genome in 100-kb sliding windows. When significant alignment is found (>2 kb of sequence within a window) it will be indicated by a peak in the result output. The height of the peak is relative to the amount of matched sequence within the 100-kb window.

specific coordinates in the human genome reference assembly, e.g. a BAC clone, a fosmid or chromosomal coordinates. This sequence is first repeat masked in an effort to mimic the COT-1 blocking of repeats commonly used in FISH experiments. The repeat-masked sequence is next compared to the reference human genome sequence using MegaBLAST (Zhang et al., 2000). A sliding-window approach is used, and the input sequence is compared to a 100-kb window from the genome at a time, sliding 50 kb per window (Fig. 5). If the BLAST results within a window show a total unique alignment length of 2% or higher (i.e. at least 2 kb in the 100-kb window), it will be shown in the output. The result would always be expected to give the best match for the region the sequence was taken from. Any additional peaks in the output represent regions of high identity in other parts of the genome, which may give rise to multiple hybridization signals in FISH experiments (Fig. 6).

In order to test how well the eFISH simulations reflect actual results, a number of test assays were designed which were run using both FISH and eFISH. In all cases where one or two regions were indicated by eFISH, those regions were also detected in the FISH experiment. When multiple regions were indicated by eFISH, regions that gave a very low score were sometimes not seen in the actual FISH result. However, this seems to be due to variability between experiments and may to some extent depend on the composition of the underlying sequence. In certain experiments, the hybridization intensities are stronger overall, and then also the signals just above the threshold in the eFISH tool gave rise to weak signals in the FISH experiment. eFISH is implemented as a widely accessible web tool. DNA sequence BLASTing has been pre-computed, which substantially speeds up the performance. Usually it takes only one or two seconds to give the prediction for one probe. eFISH is freely accessible at http://projects.tcag.ca/efish.

In our experience, eFISH is an accurate predictor of the outcome of FISH experiments. We routinely check all potential probes in eFISH before they are ordered, and it is a helpful part of the process for choosing the optimal probe for a specific region. Applying this as a step in the experi-

A

eFISH Result

Back to main page

Genome:	Human Genome - May 2004 Assembly (hg17)
Mapped Probes:	RP11-365N20 (chr15:36,800,725..36,968,084|BAC_End) RP11-24O14 (chr5:69,802,531..69,966,791|BAC_End)
Show Chromosomes:	☑chr5 ☑chr6 ☑chr15 ☑chr20 ☑chr22 Check All Uncheck All
Show Probes:	☑ RP11-365N20 [c15] ☑ RP11-24O14 [c5,c6,c20,c22]
Max Score (Y axis):	◉100 ○50 ○10
Image Width:	◉1000 ○2000 ○4000
	Refresh

eFISH image

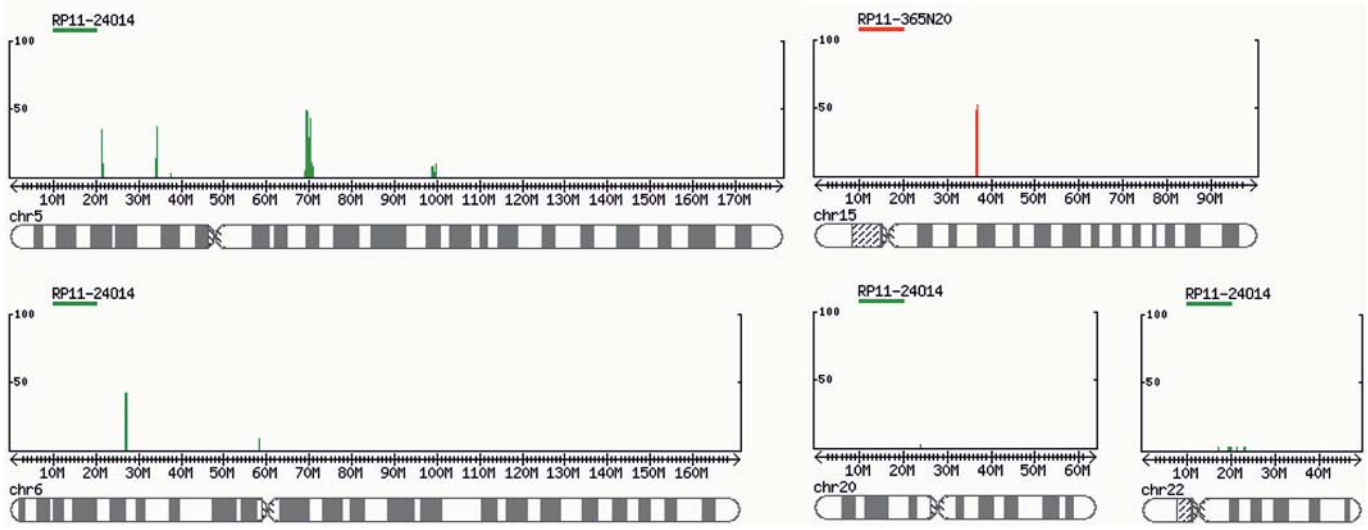
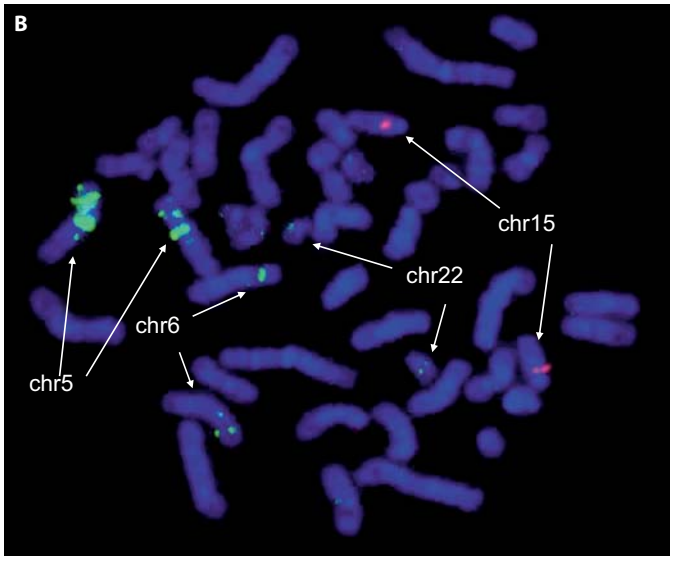


Fig. 6. Comparing eFISH to FISH. Shown in **A** is an example of the results reported from the eFISH program. In this case two probes, RP11-365N20 (clone end accession numbers are AQ543813 and AQ543816) and RP11-24O14 (clone end accession numbers are B89373 and B89383), were entered as search terms. In these two instances the entire clone sequence is not known so their end-sequences are used to identify the intervening sequence from the human genome reference assembly, and this is used for the BLAST analysis. All chromosomes where any of the two probes give a significant hit are shown. RP11-365N20, shown in red, gives a single signal on chromosome 15. RP11-24O14, shown in green, generates hits on four different chromosomes, with several signals on chromosome 5. In **B**, the results from an actual FISH experiment using the same two clones are shown. As expected, RP11-365N20 yields a single signal on chromosome 15, while RP11-24O14 shows signals on chromosomes 5, 6 and 22. All these hits, including the multiple signals detected on chromosomes 5 and 6, were predicted using the eFISH program. eFISH also predicted very weak hybridization to chromosome 20, but these cannot be seen in this experiment. The signal intensity often varies between experiments, and signals predicted by eFISH to be very low may not be detected in all experiments. Genomic clones for FISH hybridization experiments are available from several sources including BACPAC Resource Center (http://bacpac.chori.org/) and The Centre for Applied Genomics (www.tcag.ca), to name a few.



mental design also increases the success rate for our experiments and therefore decreases cost for failed or un-interpretable assays. In difficult regions containing segmental duplications, the results of eFISH are also helpful for interpretation of the data.

Summary

The complexity of genomic variation data requires the development of special databases, bioinformatics tools, and algorithms compatible to a diverse range of users, and to other databases. The resources described here provide a relevant and reliable information source for the study of structural variants in the human genome. The Database of Genomic Variants will continue to be improved including the curation and updating of new material, as it becomes available. The overall project will be considered a success when an equal number of molecular biologists, medical geneticists, physicians, and diagnostic laboratories utilize the information for a better understanding of the role of genomic variation in development and disease.

Acknowledgements

The authors would like to thank Weimin Zhu and Charles Lee of the European Bioinformatics Institute, and Jeffrey MacDonald, Cheng Qian, Terence Tang, Ying Qi, and the bioinformatics support staff of The Centre for Applied Genomics (www.tcag.ca). We also acknowledge Dr. Charles Lee of Harvard University and Brigham and Women's Hospital, Drs. Nigel Carter and Matthew Hurles (Wellcome Trust Sanger Institute), Dr. Keith Jones (Affymetrix), and Dr. Hiroyuki Aburatani (University of Tokyo) for ongoing contributions to the Copy Number and Structural Variation Project (http://www.sanger.ac.uk/humgen/cnv/).

References

Carter NP: As normal as normal can be? Nat Genet 36:931–932 (2004).

Check E: Human genome: patchwork people. Nature 437:1084–1086 (2005).

Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, et al: Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4: R25 (2003).

Claustres M, Horaitis O, Vanevski M, Cotton RG: Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12:680–688 (2002).

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK: A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81 (2006).

Eichler EE: Widening the spectrum of human genetic variation. Nat Genet 38:9–11 (2006).

Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. Nat Rev Genet 7: 85–97 (2006a).

Feuk L, Marshall CR, Wintle RF, Scherer SW: Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1:R57–66 (2006b).

Fredman D, White SJ, Potter S, Eichler EE, Dunnen JT, Brookes AJ: Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861–866 (2004).

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, et al: Copy number variation: new insights in genome diversity. Genome Res 16:949–961 (2006).

Galperin MY: The Molecular Biology Database Collection: 2005 update. Nucleic Acids Res 33: D5–24 (2005).

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al: The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440 (2005).

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA: Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82–85 (2006).

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. Nat Genet 36:949–951 (2004).

Ianzano L, Zhang J, Chan EM, Zhao XC, Lohi H, et al: Lafora progressive Myoclonus Epilepsy mutation database-*EPM2A* and *NHLRC1* (*EMP2B*) genes. Hum Mutat 26:397 (2005).

Ingelman-Sundberg M: Polymorphism of cytochrome P450 and xenobiotic toxicity. Toxicology 181–182:447–452 (2002).

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258:818–821 (1992).

Lee C: Vive la difference! Nat Genet 37:660–661 (2005).

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al: Common deletion polymorphisms in the human genome. Nat Genet 38: 86–92 (2006).

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211 (1998).

Scherer SW, Cheung J, MacDonald JR, Osborne LR, Nakabayashi K, et al: Human chromosome 7: DNA sequence and biology. Science 300:767–772 (2003).

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al: Large-scale copy number polymorphism in the human genome. Science 305:525–528 (2004).

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al: Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88 (2005).

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al: Fine-scale structural variation of the human genome. Nat Genet 37:727–732 (2005).

Xu J, Zwaigenbaum L, Szatmari P, Scherer SW: Molecular cytogenetics of autism. Curr Genomics 5:347–364 (2004).

Zhang Z, Schwartz S, Wagner L, Miller W: A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214 (2000).