Bioinformatics Tailored to Meet Cancer Biomarker Research Needs

Sudhir Srivastava, Ph.D., MPH
Chief, Cancer Biomarkers Research Group, NCI
In collaboration with Dan Crichton, JPL
INVESTIGATOR-DRIVEN CONSORTIUM

An “infrastructure” to support collaborative research on molecular, genetic & other biomarkers in human cancer detection & risk assessment
EDRN’s Mission

To implement biomarker research through strategic and systematic, evidence-based discovery, development and validation of biomarkers for:

- cancer risk assessment,
- early detection,
- diagnosis and
- prognosis of cancer
• Discover, develop and validate biomarkers for cancer detection, diagnosis and risk assessment

• Conduct correlative studies/trials to validate biomarkers as indicators of early cancer, pre-invasive cancer, risk, or as surrogate endpoints

• Develop quality assurance programs for biomarker testing and evaluation

• Forge public-private partnerships
Key Objectives of EDRN Informatics

- Structured data collection, storage and curation;
- Organization of data that are searchable and informative;
- Dissemination of data;
- Knowledge integration of biomarker research

All activities are tailored toward making data verifiable, consistent and informative to community.
Building Bioinformatics Infrastructure

In 2000, EDRN investigators identified the following needs:

- Ability to develop informatics standards such as common data elements (CDEs) for biomarker research to allow for data to be linked together and analyzed
- Informatics to support multi-institutional validation studies
- Create tools to facilitate the use of the informatics standards for data collection (e.g. CDE Form Tool, EDRN Data Model, Public Portal)
- Ability to access specimens across EDRN Clinical validation Centers
- Ability to support secure data transfer, data analysis and communication
- Ability to support EDRN-wide data storage, curation and retrieval of multidimensional, multi-format data
High level *ontology* model of EDRN which describes
- Core “object” concepts of EDRN data
- Relationships between those objects

Specific models are derived from this high level model
- Model of biospecimens
- Model for each class of science data

EDRN is specifically focusing on a granular model for annotating biomarkers and their studies using CDEs

But, the high level model allows us to move towards an integrated information environment

The model is independent of the software component architecture

Captured in an ontology modeling tool called Protégé
Implemented Informatics Systems

Public Portal – Access to the EDRN data enterprise

Secure Site -

eSIS – Capture of protocol information

VSIMS – Capture of validation study information

ERNE (EDRN Specimen System) – Access to distributed specimen inventory

eCAS – Capture, access, sharing of biomarker data results

Biomarker Database – Capture, sharing of biomarker annotations

http://cancer.gov/edrn (operational)
http://edrn.jpl.nasa.gov (beta; emerging capabilities)
Common Data Elements provide a set of standard terms and values for a domain

- Used to annotate the data in EDRN systems
- Integrated with the Biomarker Ontology
- They are classified into organ (Breast and GYN, G.I., Lung and Upper Aerodigestive, Prostate and Urologic), epidemiological and specimen CDEs
- Based on ISO/IEC 11179 (standard for data elements)
- More than 2600 CDEs developed

Captured by EDRN and maintained by the EDRN Data Management and Coordinating Center in Seattle
EDRN CDEs Cover the Following Areas

• Pre-clinical cancer biomarkers
• Study management
• Specimen
• Instrument data (including proteomics, genomic data)
• Publications
• Investigators
• Other data that integrates the EDRN into a semantic knowledge system.
EDRN CDE Details

Classified into the following for EDRN Sites:

- Core (all sites collect basic agreed upon elements)
  - Study
  - Demographic
  - Medical History
  - Smoking History
  - Cancer History
- Organ Specific
  - Developed and agreed upon by EDRN Organ Collaborative groups, relative to organ studied
    - Demographic
    - Medical History
    - Smoking History
    - Cancer History
Using CDEs to publish Biomarker Information to the Portal

Annexin 1 belongs to a family of calcium-dependent phospholipid binding proteins which are preferentially located on the cytosolic face of the plasma membrane. Annexin 1 protein has an apparent relative molecular mass of 40 kDa, with phospholipase A2 inhibitory activity. Since phospholipase A2 is required for the biosynthesis of the potent mediators of inflammation, prostaglandins and leukotrienes, annexin 1 may have potential anti-inflammatory activity. Annexin 1 also promotes membrane fusion and is involved in exocytosis. It seems to bind from two to four calcium ions with high affinity.
Gateway to information
Information both within and outside of EDRN
Initial starting point for community to get to EDRN research
Google-like search to access the wealth of data
Multi-level Security

http://cancer.gov/edrn
http://edrn.jpl.nasa.gov (beta site for emerging capabilities)
Ability to support secure data transfer, data analysis and communication: EDRN Secure Site

EDRN Secure Website

- A unified portal allowing PIs to access shared information
- Restricted to EDRN registered users
- Uses the Internet as the primary means of access to the data
- A collaborative website for sharing of information among PIs
Validation Study Management System (VSIMS) was developed to allow:

- Use of EDRN Developed CDEs
- Forms Entry
- Data Transfer
- Eligibility Confirmation
- Specimen Tracking
- Real-time Data Reporting
- Issue Tracking
- Portal Interface
VSIMS (provide an informative snapshot of VSIMS screen)
Ability to locate specimens across EDRN Clinical Centers: ERNE

- Specimen Locator System, dubbed ERNE, EDRN Resource Network Exchange was developed to query data across EDRN’s Clinical Validation Centers (CVC)

- The system is based on NASA JPL’s Object-oriented Data module which can be easily tailored to the CVC’s institutional informatics system
  - Same software module used to share earth and planetary science data

- ERNE allows the user to query the availability of specimens in real-time

- This is a first-ever system developed to query specimen on disparately distributed specimens across the country
  - “ERNE” has been a model studied by many groups
Deployed EDRN Specimen Locator System

- Highly diverse sites
- Geographically distributed
- New centers continually added
- Multi-center data system infrastructure
- Sites with different data systems allowing access
- Integrated based on common data standards
- Secure (e.g. encryption, authentication, authorization)
- IRB and HIPAA Compliant
EDRN Resource Network Exchange (ERNE)

- An *infrastructure* for sharing data resources across EDRN
- Supports *real time* (on demand) *distribution* of data to users
- EDRN CDE Mapping Tool
Serving Up Specimens: NASA-NCI Project Links Databases Across the Country

David Tenenbaum
EDRN Portal: Provides one-stop access to various resources in EDRN

- Gateway to information
- Information both within and outside of EDRN
- Initial starting point for community to get to EDRN research
- Google-like search to access the wealth of data
- Multi-level Security

http://cancer.gov/edrn
http://edrn.jpl.nasa.gov (beta site for emerging capabilities)
To Capture and Share Biomarker Annotations Provides connection to the following:

- Protocol
- Scientific Data
- Publications
- Additional Biomarker Resources
Improve Access to Data Analysis

- EDRN has created a seamless informatics platform that access data from various sources without having the user to worry about informatics jargon.
- Data can be accessed through secure and public sites.
- User ID and password are required to maintain the Confidentiality Disclosure Agreement and prevent unauthorized use.
Access to Data

Early Detection Research Network
Research and development of biomarkers and technologies for the clinical application of early cancer detection strategies

Autoantibody Biomarkers

Dataset Metadata:

<table>
<thead>
<tr>
<th>Protocol Name</th>
<th>Validation of Protein Markers for Lung Cancer Using CARET Sera and Proteomics Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol ID</td>
<td>138</td>
</tr>
<tr>
<td>Dataset Name</td>
<td>Autoantibody Biomarkers</td>
</tr>
<tr>
<td>Principal Investigator</td>
<td>Samir Hanash</td>
</tr>
<tr>
<td>Site Name</td>
<td>Fred Hutchinson Cancer Research Center (Biomarker Developmental Laboratories)</td>
</tr>
<tr>
<td>Data Custodian</td>
<td>Ji Qiu</td>
</tr>
<tr>
<td>Data Custodian Email</td>
<td><a href="mailto:jiqiu@fhcrc.org">jiqiu@fhcrc.org</a></td>
</tr>
<tr>
<td>Organ Site</td>
<td>Lung</td>
</tr>
<tr>
<td>Organ Groups</td>
<td>Collaborative Lung and Upper Aerodigestive</td>
</tr>
<tr>
<td>Method</td>
<td>Proteins from human lung adenocarcinoma cell line A549 lysates were</td>
</tr>
</tbody>
</table>

1-7 of 7 Products Associated With This Dataset:

- (By Batch)ANXA1(FR–00–84).pdf
- (By Batch)14±3±3.pdf
- CARET_lung_cancer.xls
- (By Batch)ID–1.pdf
- (By Batch)LMR1.pdf
- (By Batch)PP9.5.pdf
- annexin+lmr1±14±3±3.pdf

Result Page 1

Credit: Sam Hanash (Validation of Protein Markers for Lung Cancer Using CARET Sera and Proteomics Techniques)
Accessing EDRN Data: Data may come any of these sources:

- Biomarker Annotations
- Protocols
- Biomarker Data Results
- Specimens
- Linked through Public Portal
- Access to download data
Joint Architecture Diagram: EDRN and Canary Foundation

Sample Sources
- BCCA + UTSW

Pls and Laboratories
- Intuitive web upload

Shared Data Store (Amazon S3)
- Cloud raw data storage

EDRN EKE (Knowledge System Environment & Portal)
- Final Data; Biomarkers (curated)
- RDF - Automated

Canary Portal (GLS System)
- Web Service APIs
- Connection to data

FHCRC DMCC
- EPI Data sent to DMCC
- DMCC Harmonized, De-IDs, acts as Honest Broker

Data Analysis
- Download data sets to use local analysis tools

EDRN Study Information (via RDF)
- Working research, Biomolecular queries

Public Data Access
- EDRN PI Access

Biomarker Study Information and Results
EDRN Informatics Goals: Positioning Informatics for the Future

- Continue to support a robust *information architecture* based on EDRN CDEs
- Comprehensive capture of data across EDRN
- Increase collaboration among the groups through informatics
- Push tools all the way up the pipeline to support data sharing as early as possible
- Improved access to data to enable analysis
- Increase collaboration between science and informatics
- Partnerships
Increase Collaboration Between Science and Informatics

• It is EDRN’s philosophy to drive informatics development based on scientific needs and not the vice-versa

• The EDRN’s Subcommittee on Data Sharing continuously evaluate the informatics to meet scientific needs; active input is solicited from the users and EDRN members alike

• A close interaction with EDRN scientific leaders and informatic developers are fostered through regular consultations
• Information Reuse and Integration

- Explore strategies for integrating biomarker knowledge into novel systems and applications:
  - Database Integration using unifying data models and knowledge system ontology
- Develop strategies and algorithms for applying data integration approaches in biomarker research, discovery and development
- Enable knowledge generation about biomarkers and effective decision making on clinical application and prioritization

(continues)
CrowdSourcing

Expand IT support and new tool development for improved and biologically more relevant biomarker discovery

• In the era of “Big Data” increase capabilities of relational data bases to ensure higher performance and easy adaptability to CLOUD computing

• Information reuse, information integration, and reusable systems

  ➢ Comprehensive implementation of CDEs across entire biomarker development pipeline

  ➢ Development of flexible schemas for ease in adapting annotation and analysis pipelines to the needs of new biomarker discovery platform data outputs (NextGen Sequencing, nc/miRNA, metabolomic, imaging analyses, etc)
• Crowdsourcing’s biggest benefit is to receive better quality results of multidimensional, high throughput data (sequencing, profiling data) on biomarker discovery,
• Since several experts will offer their best ideas, skills, & support, results can be delivered much quicker than traditional methods, since crowdsourcing is a form of freelancing.
• However, clear instructions are essential in crowdsourcing. If the instructions are not clearly understood, analysis of data could be complicated. Quality can be difficult to judge if proper expectations are not clearly stated.
Using EDRN Informatics Infrastructure, Canary Foundation of America was able to:

1. Integrate clinical and research data, supporting “multi-omic” data sets.
2. Upload, management and access in cloud for all project stakeholders.
3. Develop query interface for quickly finding and accessing data sets for analysis.
4. Deliver low cost and scalable data storage provided by Amazon.
5. Support secure data transfer, data analysis and communication.
Globally Connected Biomarkers Knowledge System