In Silico Target Portal: An Integrated Oncology Target Discovery Web Portal

Ying Li[1], Jyothi Venkatiahgari[1], Liping Jin[1], Donavan T. Cheng[2], James Cai[1]
[1] *pRED Informatics,* [2] *Translational Research Sciences, Pharma Research and Early Development, Hoffmann-La Roche Inc., Nutley, NJ 07110, USA*
Ying_l.li@roche.com

Abstract

With increasing numbers of genomic datasets becoming publicly available, new analysis and filtering tools are needed to serve user communities. Here we describe the development of the Roche In Silico Target Portal (ISTP) for Oncology research, a system that integrates meta-analysis results of gene expression studies with gene-disease, gene-mutation as well as gene-compound (competitor information) relationships into a single entry application. This system uses a three-tier architecture that includes an Oracle relational database, a set of data processing pipelines, and a Java-based web interface. The data processing pipelines were developed in house using Accelrys' Pipeline Pilot. Currently, the ISTP database contains data for seven major tumor types with tens of thousands of records on genes, compounds, mutations, and gene-disease relationship, accessible to Roche internal users via a simple user interface. The ISTP application highlights the potential of such a data integration system for drug discovery and development.

Keywords: gene expression meta-analysis, In Silic target portal, oncology, gene related data extraction.

1. Background

Modern drug discovery is a multi-step process involving target identification, target validation, lead identification and optimization, pre-clinical and clinical studies. Target discovery is a critical step in this process, where scientists seek to identify mechanisms causal for disease pathogenesis, in hope of developing new therapeutic interventions. Over the past decade, microarray technology, which simultaneously measures the mRNA levels of tens of thousands of genes in tissue samples, has been widely used in discovering genes and pathways that define a disease phenotype. Some biologists remain skeptical regarding discoveries obtained via this approach. The heterogeneity and complexity of the disease, and the limited number of samples for each study, make the reproducibility and interpretation of microarray experiments a challenge.

In a simple gene expression analysis, where tumor samples are compared with disease-free samples from the same patients, hundreds or thousands genes can be identified as differentially expressed in tumor relative to normal tissues. It remains a challenge for scientists to pin point the 'driver' genes accountable for the disease. Fortunately, as public gene expression datasets become increasingly available, information can be leveraged in a meta-analysis to identify driver genes amidst lists of differentially expressed genes. A meta-analysis of many studies for a given disease can identify common mechanisms, genes and pathways for the underlying disease. Many articles published in recent years [Ref. 1, 2, 3, 4] have reported using multiple datasets for a specific disease to identify potential drug targets, disease prognosis and diagnosis biomarkers. However, the authors also noted the complexity of microarray meta-analysis; caution must be taken in identifying the suitable studies, processing of raw data, or using pre-processed data. Meta-analysis results may also require further annotation with disease and mutation relationships, to support and help prioritize the genes for drug target and/or disease biomarker identification.

Here we report the development of the *In Silico* Target Portal (ISTP), a web application that combines multiple data types relevant for target and biomarker discovery based on a gene expression meta-analysis framework. In this practice, we first identified a set of publically available microarray datasets for selected tumor types. Meta-analysis was performed on studies from the same tumor type using an internally developed algorithm to produce ranked gene lists. Additionally, Pipeline Pilot protocols were developed to process gene mutations, gene-disease relationships, and gene related competitor information. The processed information was then stored in an Oracle relational database for the web-based java application to access. This integrated information portal is an example of how combining

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

multiple gene properties with gene expression meta-analysis results can yield candidates for target and biomarker discovery.

2. Results and discussions

The development of the Roche *In Silico* Target Portal involved the processing and integration of various data sources into an Oracle relational database, as well as construction of a custom java-based user interface. Data processing and integration formed the foundation of the project, which required both automated and manual data QC and filtering. In this section, we discuss the outcomes of our database and the web application.

2.1. Core gene centric information in ISTP

Gene expression meta-analysis can identify commonly differential expressed genes among a set of similar studies. The number of genes from the meta-analysis can range from hundreds to thousands. In order to support gene prioritization for target and/or biomarker identification, we integrated the gene expression meta-analysis results with additional gene centric information in the ISTP application. As a proof of concept, documents and studies of seven major tumor types were collected for the meta-analysis and information extraction to produce gene expression, gene mutation, gene-disease relationship, and gene-compound relationship. In this section we discuss the detailed results of these processed records in the ISTP database.

2.1.1. Cross-study gene expression comparisons

Many studies use microarrays to identify disease specific genes/pathways differentially expressed in the disease condition. Often, studies from different labs with a similar experiment design could result in different gene lists/pathways, making it hard to compare results across studies. Thus, meta-analysis is an important topic in gene expression analysis. Due to the complexity and variability of the gene expression experiments, it is inappropriate to directly compare different studies at the raw expression signal level. Instead, only relative expression changes comparing effect sizes should be used in cross-study comparisons after performing within study normalization. We adopted this practice in ISTP.

The first step of building a tumor gene expression database was to carefully select a set of gene expression studies for the tumor types of interest. Initially, over one hundred publically available gene expression studies were selected. Manual curation identified sixty-four studies, spanning seven major tumor types, as a core set of gene expression studies. These studies contain samples from both tumor and disease-free tissues with minimum sample size greater than three for each group.

Raw-date files and study/sample information of the selected studies were downloaded from GEO or ArrayExpress for further annotation, signal pre-processing, and robust pair-wise comparison analysis [ref. 5]. A non-parametric ranking algorithm was developed to rank the genes (along with their probes) within the studies of each tumor type. The selected studies, their meta-data information, the pair-wise comparison results, as well as the ranking results were stored in an Oracle database. This resulted in ninety disease vs. normal comparisons with thousands of genes for each tumor type. Table 1 shows the distribution of the comparisons and the number of ranked genes per tumor type.

Table 1. Distribution of selected microarray studies by tumor type.

Tumor type	# of comparisons	# of ranked genes
Brain	13	18286
Breast	20	17180
Colon	12	16689
Kidney	20	16500
Liver	7	14079
Lung	8	9990
Head&Neck	10	12898
Total	90	n/a

While traditional array-based gene expression technologies are limited to the detection of known genes and regions, the more recent developed next-generation sequencing (NGS) RNASeq technology permits the digital measurement of the whole transcriptome. Besides increased sensitivity and dynamic range, NGS technology allows us to more precisely measure expression of transcript variants, and even detect novel alternative splicing isoforms. The inclusion of RNASeq data in the ISTP database would greatly enhance our capability to discover novel targets.

2.1.2. Other gene centric annotations

Competitor information (CI) plays an important role in pharmaceutical drug research and development.

Thomson Reuters Integrity is a commonly used CI data source in the pharmaceutical industry. This CI database integrates biological, chemical, and pharmacological data for compounds with biological activity to provide unique knowledge for target assessment. To help scientists seamlessly access the Integrity CI information in ISTP application, we systemically processed and incorporated Integrity's gene-compound competitor information into the database. Currently, the database contains 74,506 records of competitor information records, covering 1,132 unique genes, and 30,898 unique compounds with 3,226 modes of action.

This competitor information was used to assess the ranking results. For each tumor type, the ranked gene list was categorized into four groups: the top 100 ranked genes, 101 to 500 ranked genes, 501 to 1000 ranked genes, and large than 1000 ranked genes. For each category, the percentage of genes that has compounds in ISTP CI database was computed. As showing in Figure 1, for every tumor type, there is a trend of gene enrichment of genes with oncolytic compounds in ISTP database. On average, the top 100 ranked gene group has an average of ~20% of genes with oncolytic compounds; and 14%, 10% and 6% for the 101-500, the 501-1000, and the >1001 ranked gene groups respectively. Evidently, our ranking algorithm has significantly enriched the genes with oncology target indication.

Literature based gene-disease associations can also be leveraged as an addition source of annotation information. Curated gene-disease associations (by co-occurrence) were extracted from PubMed citations using Pipeline Pilot protocols. The information was then stored in ISTP Oracle database for rapid retrieval by ISTP users. Table 2 is a snapshot of the database content for the ISTP gene-disease annotation. The current ISTP database contains over 50,000 total gene-disease relationships, ranging from 2,000 (kidney tumor) to 15,000 (breast cancer) per tumor type.

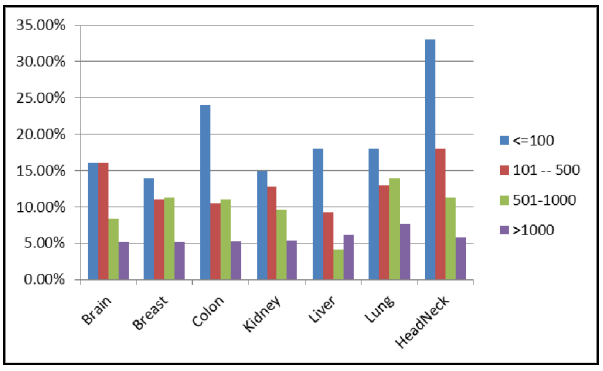

Figure 1. A bar graph of the analysis result of ranked gene lists using competitor information. For each tumor type, the ranked gene list was categorized into 4 groups: top 100, 101 – 500, 501-1000, and >1000 ranked gene group; the percentage of genes in each group that have oncolytic compounds was calculated and plotted.

Table 2. Distribution of gene-disease information per tumor type.

Tumor type	# of records	# of genes
Brain	2761	1215
Breast	15496	3376
Colon	9603	2444
Kidney	1809	901
Liver	5723	2528
Lung	7802	2188
Head&Neck	7502	1967
Total	50696	14622

This gene-disease association strength can be simply measured by the number of PubMed citations for each gene-disease pair. This information can be used to identify the less-well studied but significantly regulated genes in tumor samples. Table 3 shows an example outcome of using the breast cancer gene-disease information. In this example, we identified 22 highly regulated genes in breast cancer, with less than two publication citations (citation <2). These genes would potentially be novel targets for breast cancer drug discovery. In comparison, only 2 highly cited genes were commonly differentially expressed, reflecting the fact that traditionally, genes studied in association with cancer were not chosen on the basis of differential expression in whole genome transcriptomic approaches. We applied this and similar workflows to the ISTP records to prioritize expression gene lists to a small number of interesting genes for each tumor type.

Additionally, gene-disease association results from automated text-mining technologies could be another source for ISTP database. Many institutes, including ourselves [ref. 6, 7], are adopting these text-mining tools to systemically extract the gene centric information from massive amount of publications. However, relationships obtained from automated text mining approaches tend to suffer from high numbers of false positives. More sophisticated algorithms with increased specificity will be needed to reduce the numbers of false positive associations. The remaining information can then be integrated into the ISTP system to broaden the disease knowledge database.

Table 3. An example of analysis results using ISTP breast cancer information. Top 100 genes from the breast cancer meta-analysis (see Section 3.1 in Methods for gene ranking algorithm) were categorized by the number of disease-association annotations from literature. Twenty-two genes have 1 or no public citation for disease association.

Citation Category	Total # of genes	Genes in top 100 meta-rank
<2	1847	22
2-9	1277	17
10-29	189	8
30>	66	2

Traditionally, cancer is considered as a mutation-driven disease with abnormal cell proliferation. Frequently, mutations affect oncogenes or tumor suppressor genes which in turn affect cell growth and survival. Many cancer mutation databases have been built to capture the mutations in cancer related genes [8]. Among them, the Catalogue of Somatic Mutations in Cancer (COSMIC) is highly regarded in the cancer research field. This database contains somatic mutations of benign and malignant tumor tissues and tumor-derived cell lines [9]. This COSMIC database is the primary source for the gene-mutation information in ISTP. Pipeline Pilot protocols were developed to process the COSMIC mutation information of selected tumor type tissues. At present, we have over 40,000 gene-mutation records for the 7 tumor tissues, largely from breast, brain, and colon samples.

ISTP provides an integrated platform of multiple cancer-related gene-centric data sources to allow scientists to leverage the diverse information simultaneously. Meta-analysis of gene expression studies allows us to bring similar studies to compare. While the resulting gene lists from meta-analysis were then enriched with gene-mutation, gene-disease, and competitor information. The web interface enables end users to search and explore genes in each category of the gene-information (not restricted to gene expression data) for target and biomarker identification.

We are continuously reviewing additional data sources to improve the ISTP database. Much cancer mutation information is now available for public use, such as, analyzed results of NGS whole genome and exome sequencing data from the NCI Cancer Genome Atlas project (TCGA). However, careful consideration will be needed to integrate these data with existing annotations. Questions regarding how mutations should be categorized, how non-functional mutations should be filtered, and how loss of function mutations can be identified, should all be put into consideration.

2.2. Architecture and user interface

The design of ISTP platform allows users to easily retrieve meta-analyzed gene expression results with additional gene-centric information. This information was pre-processed and stored in an Oracle relational database. A Java-based interface was built for easy data access and retrieval. In this section, we will discuss the ISTP system architecture and the Java-based web interface.

2.2.1 ISTP system architecture

The in-house developed ISTP system was based on a typical three-tier system architecture as shown in Figure 2. The first tier (data tier), for ISTP data storage and retrieval, is based on an Oracle (11g) relational database on Red Hat Linux. It contains integrated data records processed from multiple data sources. The second tier (logic tier) is developed in Java 1.6 within the framework of Java Spring and Hibernate to manage the application Java beans and database access. The third tier (presentation tier) utilizes Java RichFaces and JSF for the quick development and interactive (AJAX) web front-end presentation. This decoupled design allows us to demeanor the data-process and update independently from front end data pretention.

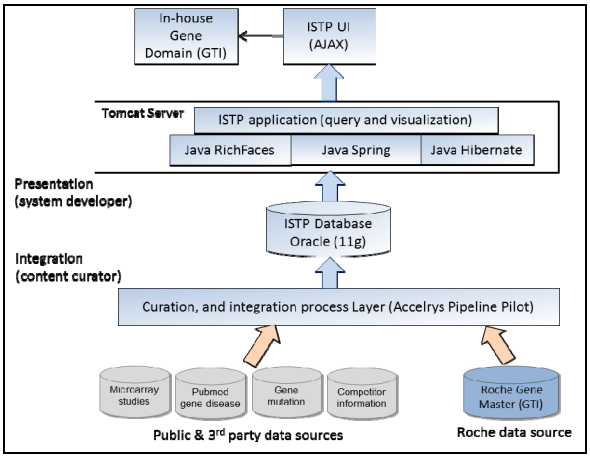

Figure 2. ISTP three-tier system architecture

2.2.2. Web interface

To allow scientists to leverage meta-analysis results of tumor gene expression studies in the context of enriched gene information, we designed a web

interface with a simple gene search and a view specific to the tumor type of interest. Figure 3 is a screen shot of the tumor type view of the application. It has a table report of the ranked genes with the number of studies a given gene has appeared in; gene mutation frequency in the given tumor tissue (COSMIC), the number of gene-disease citations for the given tumor type, and number of competitor compounds for the oncolytic indication. From here, users can either select a different tumor type to view (left panel on Figure 3) or drill down to more detailed information for each type of data (the navigation bars on the right panel on Figure 3).

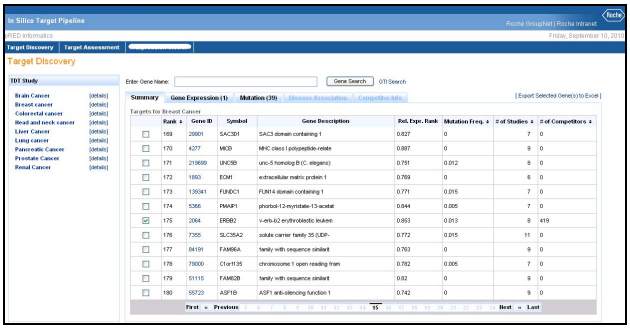

Figure 3. A screen shot of the web-based ISTP interface. The left panel of the interface is the tumor type selection navigation; the right side has both a Google-like gene search, and the detailed tumor view with gene rankings, mutation, disease association, and CI information. Users can select genes to drill down (right panel, select boxes, and navigation bars for detailed information).

To aid data exploration, visualization images were generated for each data type as shown in Figure 4. The top of the image is an expression heatmap, where the y-axis represents the number of studies color coded for the direction of expression change comparing tumor vs. normal sample (red indicates up-regulation in the tumor samples, green indicates down-regulation); and x-axis represents the ranked genes ordered from left to right in decreasing order (refer to Methods 3.1 for the ranking algorithm). The top ranked gene is most consistently up-regulated across the most number of studies, whereas the bottom ranked gene is most consistently down-regulated across the most number of studies. Genes in the middle of the plot are not significantly differentially expressed and have support from few studies.

The first bar graph presents the mutation information (y-axis represents frequency of the mutation, x-axis the ranked genes); the middle bar graph shows gene-disease information (y-axis represents the number of unique PubMed citation counts, x-axis represents the ranked genes); the bottom bar graph presents the CI information (y-axis represents the number of unique compound, x-axis represents the ranked genes).

If the user selects one gene to drill down on, red arrows would appear in the images to locate the selected gene (Figure 4, the red arrows). This visual tool can help user simultaneously view all the four types of information for the selected gene. Users can quickly tell if their gene of interest is highly over- or under- expressed in the tumor studies, what the mutation frequency of the gene reported in COSMIC is for this tumor tissue; if there is any reported association of this gene with the given tumor type; and if any oncolytic compounds are currently in development for this gene.

3. Methods

Four data process pipelines were created in-house for public microarray gene expression data, gene-disease information, gene mutation information, and gene-related competitor information. The processed gene-centric information was then stored into an Oracle relational database for easy retrieval. In the next few sections, we will discuss the details of each data process.

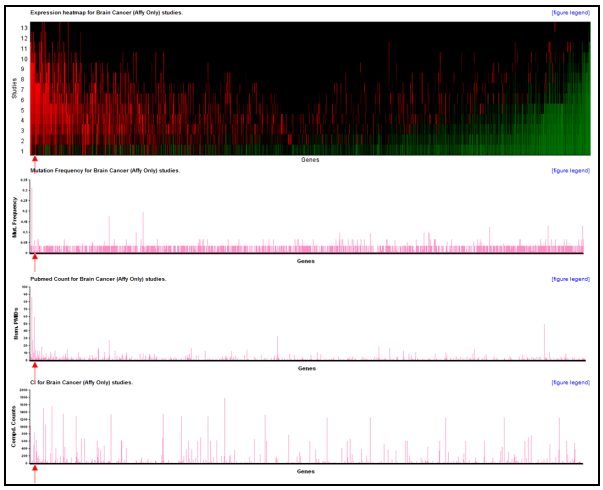

Figure 4. A screen shot of visual integration of ISTP data content.

3.1. Gene expression data process

The raw data files and study/sample information for the selected public gene expression studies were downloaded from GEO and ArrayExpress, manually annotated and curated. T-test was applied to each tumor vs. normal pair group to obtain the pair-wise

comparison results. The meta-data information as well as the pair-wise comparison results was processed into Oracle database tables. For all the comparisons in each tumor type, a non-parametric ranking algorithm was developed to rank the genes across multiple studies. Figure 5 is a flow-diagram shown the process at high level.

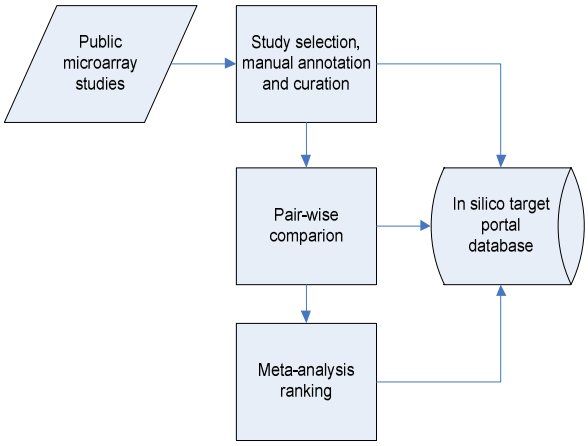

Figure 5. Gene expression meta-analysis flow diagram.

More specifically, pair-wise comparisons for gene expression studies (disease vs. normal) were obtained using Omicsoft's Robust High-dimensional Linear Model [ref. 5]. Each pair-wise comparison generated a list of genes with fold changes and p-values. This comparison test was performed on all the studies, and resulted in 90 gene datasets.

To compute genes persistently over- or under-expressed in different studies of the same tumor type, a non-parametric ranking algorithm was developed. The input of this algorithm was the subset of differentially expressed genes for each study (absolute fold change > 1.2 and p-value <= 0.05), one tumor type at a time.

For each gene in a tumor type, the algorithm computed a rank based on 3 metrics: 1) the frequency at which the gene was differentially expressed across the various studies of the given tumor type; 2) the consistency of its change in gene expression among the studies; 3) the median normalized expression rank within the studies.

The frequency at which a gene was differentially expressed was calculated by dividing the number of studies that the gene appeared (differentially expressed) by the total number of studies. The genes were then ranked in descending order based on their frequencies to generate the frequency rank.

The consistency of the direction of change in gene expression was measured using an information theory derived entropy metric, which calculated the mutual information (MI) of the gene expression direction within a tumor type.

$$MI = 1 - H = 1 - (-\log_2 P1 - \log_2(1-P1))$$

Where P1 = number of studies where the gene was differentially expressed divided by total number of studies. This calculation was done for all the genes in a tumor type (exception: when P1 = 0 or 1-P1 = 0, MI = 1). If a gene has P1 = 1, it is consistently differentially expressed in all the studies, so MI = 1; while P1 = 0.5 means half of the time the gene is up-regulated and the other half time it was down-regulated in the stuides, so MI = 0. Genes were ranked in descending order based on consistency (MI), so each gene received a rank number in consistency ranking.

The median normalized gene expression rank was calculated as follows: First, a normalized rank was assigned to each gene in a study where the gene with the minimum fold change was given a rank score of -1 and the gene with maximum fold change was given a rank score of 1. This was performed for all studies for a given tumor type. The median value of the normalized rank scores was computed for each gene and was called its "median normalized gene expression rank".

These three rank scores were combined to yield an overall rank by the following procedure. First, a combined rank was calculated using the gene's frequency rank and consistency rank. A sign was applied to the combined rank (either positive or negative) using the sign of the gene's median normalized expression rank. Genes were then ranked in descending order to result in the final gene ranking for the tumor type. In this final ranking, the most up-regulated gene appeared as the top ranked gene (starting from 1), while the most down-regulated gene appeared as the lowest ranked gene. The genes with ranks in the middle were the ones that did not show significant differential expression across studies.

This ranking algorithm was applied to each of the seven tumor types to generate an overall ranked gene list for each tumor type. Data update schedule of the gene expression meta-analysis is at user's request.

3.2. Competitor information process pipeline

The flowchart in Figure 6 shows the steps involved in the competitor information annotation pipeline. XML formatted compounds and targets (genes or proteins) documents of the Integrity Drugs and

Biological knowledge Area were periodically downloaded to Roche server by Thomson Reuters. Accelrys' Pipeline Pilot protocols were developed to extract the target information (gene or proteins), as well as compound-target relationships from these documents. The target information from Integrity was then manually curated and mapped into Roche standard gene symbols. The drug indication of the compounds was filtered to select for compounds with oncolytic indications. Due to the limited numbers of oncology compounds, and the inter-connection of different tumor treatments, we did not further divide the compounds by specific oncology subtypes. These gene-compound relationships were then stored in ISTP Oracle database for later use.

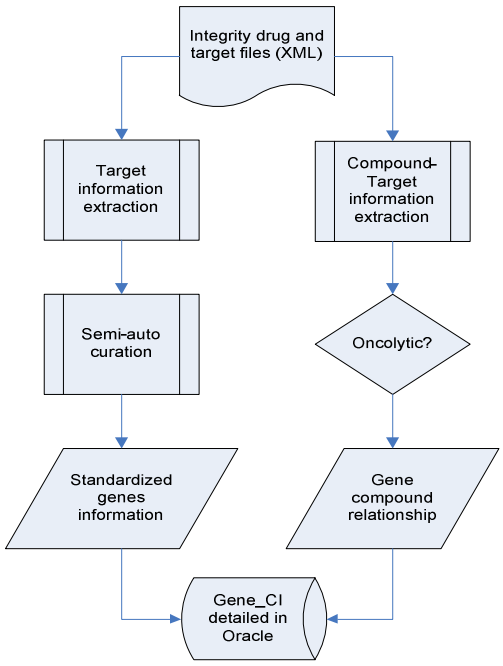

Figure 6. Flowchart of Competitor Information process pipeline.

3.3. Gene-disease & gene-mutation relationship extracting process pipelines

For gene-disease association information, Entrez genes with PubMed IDs and PubMed documents with disease MeSH terms were downloaded from public sources. Pipeline pilot protocols were developed to process these files. Figure 7 is a flow diagram that illustrates this process. Curated gene-PubMed ID relationships were extracted from Entrez Gene files. For PubMed ID and MeSH disease term relationships, we first generated 7 subsets of documents for the 7 tumor types of interest. For each tumor type, the PubMed ID – MeSH disease term relationships were extracted from the corresponding documents. The join of these 2 relationships resulted in gene-disease (MeSH) information. The later information was then stored in ISTP database tables.

Similarly, internally developed Pipeline Pilot protocols were applied to the COSMIC mutation files to obtain gene mutation information for selected tumor tissues. The gene mutation frequency in each tumor type was calculated and stored in ISTP database.

CI, gene-disease, and gene-mutation data update is scheduled bi-annually.

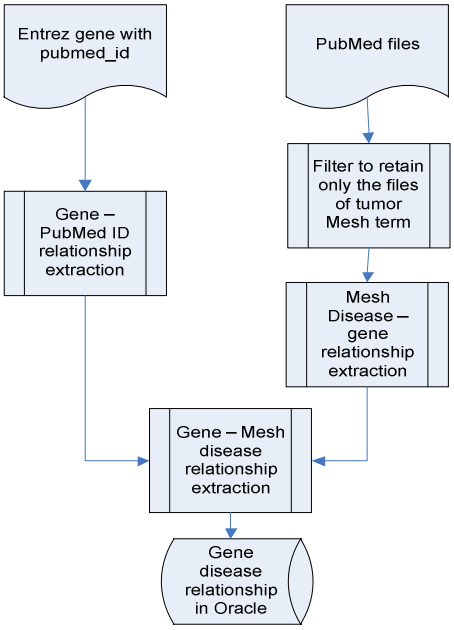

Figure 7. Flowchart of gene-disease relationship process pipeline.

5. Conclusion

An integrated *In Silico* Target Portal (ISTP) was created to facilitate the search and reporting of tumor specific gene expression meta-analysis results along with other gene-centric annotations. From multiple data sources, ISTP pipelines uniformly QC and processed these data. The non-parametric meta-ranking protocol significantly increased the quality and produced comparable microarray data cross studies from different laboratories. The single entry web interface offers Roche scientists the capability to access the integrated multi-dimensional data information in a user-friendly interface. Incorporation

of gene-compound information allows scientists to quickly assess the competitor landscape for the genes of interest. The gene-disease and gene-mutation information provides scientists a quick glance at the most relevant knowledge within a tumor-specific context. This application is facilitating Roche oncology research efforts in target and biomarker discovery applications.

6. Acknowledgements

We greatly appreciate Dr. Kathryn Packman and Dr. Helmut Burtscher for the scientific guidance on the project data acquisition and user interface design. We are grateful to Dr. W. Venus So for reviewing the manuscript.

7. References

[1] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA* 2004, 101:9309-9314.

[2] Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 2002, 62:4427-4433.

[3] Xu L, Geman D, Winslow RL: Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics* 2007, 8:275.

[4] Rasche A, Al-Hasani H, Herwig R: Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. *BMC Genomics* 2008, 9:310.

[5] Ni X, Liu K, Young S, Gaido K: Robust High-Dimensional Linear Model Method in Microarray Data Analysis. *Omicsoft.com whitepaper*

[6] S. Yilmaz1, P. Jonveaux1, C. Bicep, L. Pierron, M. Smaïl-Tabbone and M. D. Devignes. Gene–disease relationship discovery based on model-driven data integration and database view definition. *Bioinformatics* 2009, Vol 25:230-236

[7] Luis Tari, Jagruti Patel, Ying Li, Zhengwei Peng, Jan Kuntzer, Yuan Wang, Laura Aguiar and James Cai. Mining gene-centric relationships from literature: the roles of gene-mutation and gene expression in supporting drug discovery. *Int. J. Data Mining and Bioinformatics* (in print)

[8] Olivier M, Petitjean A, Teague J, Forbes S, Dunnick JK, den Dunnen JT, Langerød A, Wilkinson JM, Vihinen M, Cotton RG, Hainaut P. Somatic mutation databases as tools for molecular epidemiology and molecular pathology of cancer: proposed guidelines for improving data collection, distribution, and integration. *Hum Mutat* 2009; 30(3): 275–282

[9] Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* 2004; 91(2): 355-358 (and the COSMIC database website)