

## An Integrative Model for In-Silico Clinical-Genomics Discovery Science

Yves A. Lussier<sup>§</sup>, M.D., Indra Neil Sarkar, B.Sc., Michael Cantor, M.D.

Department of Medical Informatics, College of Physicians and Surgeons,  
Columbia University, New York, NY, 10032

*Human Genome discovery research has set the pace for Post-Genomic Discovery Research (PGDR). While post-genomic fields focused at the molecular level are intensively pursued, little effort is being deployed in the later stages of molecular medicine discovery research, such as clinical-genomics. The objective of this study is to demonstrate the relevance and significance of integrating mainstream clinical informatics decision support systems to current bioinformatics genomic discovery science. This paper is a feasibility study of an original model enabling novel "in-silico" clinical-genomic discovery science and that demonstrates its feasibility. This model is designed to mediate queries among clinical and genomic knowledge bases with relevant bioinformatic analytic tools (e.g. gene clustering). Briefly, trait-disease-gene relationships were successfully illustrated using QMR<sup>TM</sup>, OMIM<sup>TM</sup>, SNOMED-RT<sup>TM</sup>, GeneCluster<sup>TM</sup> and TreeView<sup>TM</sup>. The analyses were visualized as two-dimensional dendrograms of clinical observations clustered around genes. To our knowledge, this is the first study using knowledge bases of clinical decision support systems for genomic discovery. Although this study is a proof of principle, it provides a framework for the development of clinical decision-support-system driven, high-throughput clinical-genomic technologies which could potentially unveil significant high-level functions of genes.*

### Background and Significance

The Human Genome discovery research has set the pace for postgenomic discovery research (PGDR). While molecular-level post-genomic fields, such as those focused on the transcriptome and the proteome, are intensively pursued, little effort is being deployed in the later stages of molecular medicine discovery research, such as clinical-genomics.

In principle, many post-genomic technologies can interoperate with pre-clinical and clinical information technologies. For example, gene expression clustering algorithms have recently been applied to

higher levels of structures and functions than just genes and proteins. Phenotypic-genotypic relationships have been investigated with clustering algorithms including molecular histopathology<sup>1</sup> and multidimensional gene-drug-traits<sup>2,3</sup>. High levels of functions and structures, such as those found in clinical observations, remain to be explored. In addition, there is no available publication on in-silico clinical-genomics technologies. The availability of in-silico high throughput clinical-genomics system could not only potentially unveil novel phenotypic-genotypic interactions but could also contribute to systems-biology discovery of higher biomodules and molecular-clinical systems interaction.

The objective of this study is to demonstrate the relevance and significance of integrating mainstream clinical informatics decision-support-systems to current bioinformatic genomic discovery science.

The strength of this feasibility study resides in the integrative analyses that it enables: feeding high throughput bioinformatics analytic tools with quality curated knowledge. Specifically, it uses three types of biomedical knowledge meticulously curated by biomedical experts:

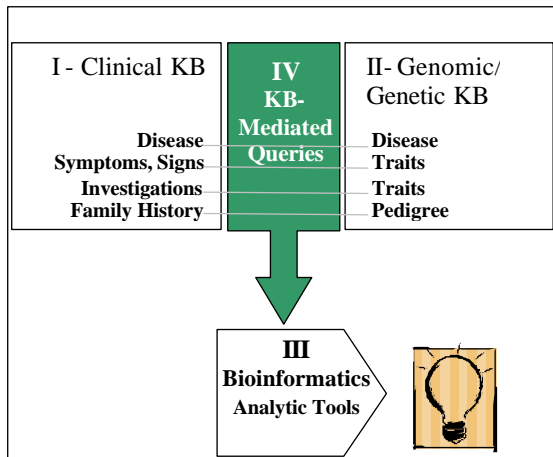
- 1) The **clinical knowledge** of journal publications and expert-clinicians engineered in decision support tools<sup>4,5,6</sup> validated for over three decades in clinical trials<sup>7</sup>. In clinical decision support tools, frequencies of clinical observations are observed in populations of patients with a specific disease. They are clinical analogs to widespread genomic measurements over pooled tissue specimens<sup>8</sup>.
- 2) The **genetic and genomic knowledge** of biomedical journals of publicly available knowledge-bases (e.g. OMIM<sup>9TM</sup>), and
- 3) The **reference terminology knowledge** (e.g. SNOMED-RT<sup>10TM</sup>)

Additionally, this analytical model can also potentially support high-throughput unsupervised gene and protein expression mining driven by

---

<sup>§</sup> Corresponding Author.

E-MAIL: Lussier@dmi.columbia.edu



**Figure 1 – In-Silico Clinical-Genomics Knowledge Discovery Model**

automated clinical-genomics-based hypothesis generation.

**Methods**

**In-Silico Clinical-Genomics Discovery Model**

As shown in Figure 1, the information model for in-silico clinical-genomic knowledge discovery is designed I) to mediate queries among clinical knowledge bases (KBs) of decision support systems, II) to compare these queries across genetic, genomic and post-genomics KBs, and III) to integrate relevant bioinformatics analytic tools, such as gene clustering, to classify the results of these in-silico studies.

As illustrated in Table 1, terminology classes of online genetic and genomic knowledge bases have been modeled to interoperate with clinical decision support systems.

A specific instance of the clinical-genomic information model (CGM) was developed using the following well-validated components:

- I. *Clinical KB:* Quick Medical Reference™ (QMR™)<sup>4</sup> Clinical Decision Support System, version 2.03.
- II. *Genetic/Genomic KB:* There is increasing evidence that even monogenic diseases are not simple and that other genes and their associated phenotype are subjected to modifications by other genes and the environment<sup>11</sup>. Mendelian diseases provide some of the simplest gene-disease-trait expression models. Thus, the Online Mendelian Inheritance in Man<sup>9</sup> (OMIM™) database was selected.
- III. *Bioinformatic Analytic Tool:* GeneCluster™<sup>12</sup>

**Table 1 Genomic Research and related Clinical Knowledge Terminology**

Genetic / Genomic		Clinical Observations
Traits	Gene expression - DNA - mRNA - Protein	Objective personal history of illnesses (e.g. symptoms, past history) Signs (e.g. cough) Results of clinical investigation: - tests and their interpretation (e.g. ECGs, X-Rays) - Labs & Pathology, (including Molecular Markers)
		Family history
Pedigree		Occupational exposures (e.g. radiation, Asbestos) Chemical treatments & substance use/abuse (e.g. medication, smoking, alcohol) Physical procedures (Radiotherapy, Surgery, etc.)
Environment		
<b>Disease</b>		

version 2.11 and TreeView™ Error! Bookmark not defined. version 1.50.

IV. *Mediating Methodology:* The Systematized Nomenclature of Medicine<sup>10</sup> (SNOMED-RT™) was used as a reference terminology between the KBs. In this feasibility study, the process required to determine each relationship between the terms of QMR™ and the ones of OMIM™ using SNOMED-RT™ was a) manually established by a domain expert using their respective user-interfaces and b) recorded for ulterior analysis according to a previously validated methodology<sup>13</sup>. The identified relationships and their interacting concepts were then structured in the CGM.

**In-Silico Gene-Traits Expression Study**

As shown in Table 2, nine diseases with well-established Mendelian heredity and three characteristic cancers, previously studied by gene expression arrays<sup>14</sup>, were selected from a set of diseases common to OMIM™ and QMR™. Their corresponding gene(s) and traits (as defined in Table 1) were ascertained from OMIM™ and QMR™ respectively. The purpose of this analysis is to elucidate traits-gene associations, rather than emulate clinical DSS. Therefore, the clinical knowledge was retained (e.g., signs, symptoms, results from laboratories and other investigations). Furthermore,

the genetic and environmental data was discarded (e.g., family history, occupational exposures, medication, drug and substance abuse, etc.).

A two dimensional hierarchical clustering was applied to 635 distinct trait expressions classes [Figure 2, B] (frequencies of clinical observations) determined on 12 Gene-Disease relationships [Figure 2, A] (from OMIM<sup>TM</sup>). In QMR<sup>TM</sup>, frequencies of clinical observations are coded as numbers (from 1 to 5) representing ranges. For this analysis, these codes and ranges (%) were transformed as the mean of the range (in parenthesis): 1 <6%(.030), 2= 6-35%(.205), 3=36-65%(.505), 4=66-96%(.810), 5>97%(.985). The non-filtered data were analyzed in GeneCluster<sup>TM</sup> as an average hierarchical clustering with uncentered correlation without weights (a form of Pearson correlation). The results were then visualized and printed with TreeView as shown in Figure 2.

## Results

As illustrated in Table 2, the twelve diseases common to OMIM<sup>TM</sup> and QMR<sup>TM</sup> provided an anchor to relate the genes of the former with the traits of the latter. This produced a matrix of 635 distinct traits against twelve gene sets, containing 1255 different positive trait-gene relationships. As described in the methods, this matrix was loaded in GeneCluster as a Gene-Trait expression table and analyzed. Figure 2 contains selected views of the output of GeneCluster as visualized in TreeView. The leftmost column of Figure 2 shows a thumbnail image of the total number of genes and their clustering. Three vertical close-up views are provided on the right-hand side with legible clustered traits. The topmost group exhibits traits common to both leukemia-related genes and PKD1, while the traits of middle group pertain mainly to F8, F9, and VWF (Table 2, factors associated with hereditary bleeding disorders). Finally, the bottom group depicts traits common to all genes associated with blood dyscrasia (hereditary diseases and non-hereditary leukemia alike). As the traits were clustered around genes, their clusters go beyond just the received views of traditional medicine. Hence, with larger sets, clinicians alone would be hard-pressed to reorganize clinical knowledge in similar ways, demonstrating the originality of the clinical-genomic model.

Beyond the scope of this feasibility study, addi-

**Table 2 – Summary of the Data Set**

OMIM <sup>TM</sup>		Disease (abbreviation)	Traits QMR (#)
Gene(s), Protein(s)	MI		
F8, F8C	XL	Hemophilia A (HemA)	90
F9	XL	Hemophilia B (HemB)	89
VWF, F8VWF	AD	von Willebrand disease (vWD)	86
ENG, END, HHT1, ORW	AD	Hereditary Hemorrhagic Telangiectasia (HHT)	40
HBB	AD	Sickle cell anemia (SC)	161
CFTR, ABCC7, MRP7	AR	Cystic Fibrosis (CF)	152
PKD1	AD	Polycystic Kidney Disease (PKD)	86
UGT1A1, UGT1, GNT1	AD	Gilberts Syndrome (GS)	24
IT15	AD	Huntington Disease (HD)	84
<i>TCF3, E2A</i>	<i>NH</i>	<i>Acute Lymphoblastic Leukemia (ALL)</i>	<i>169</i>
<i>RUNX1, CBFA2, AML1</i>	<i>NH</i>	<i>Acute Myeloid Leukemia (AML)</i>	<i>175</i>
ABL1	<i>NH</i>	<i>Chronic Myeloid Leukemia (CML)</i>	<i>99</i>

MI= Mode of Inheritance, AD = Autosomal Dominant, AR = Autosomal Recessive, XL=X-linked, NH = Not Hereditary

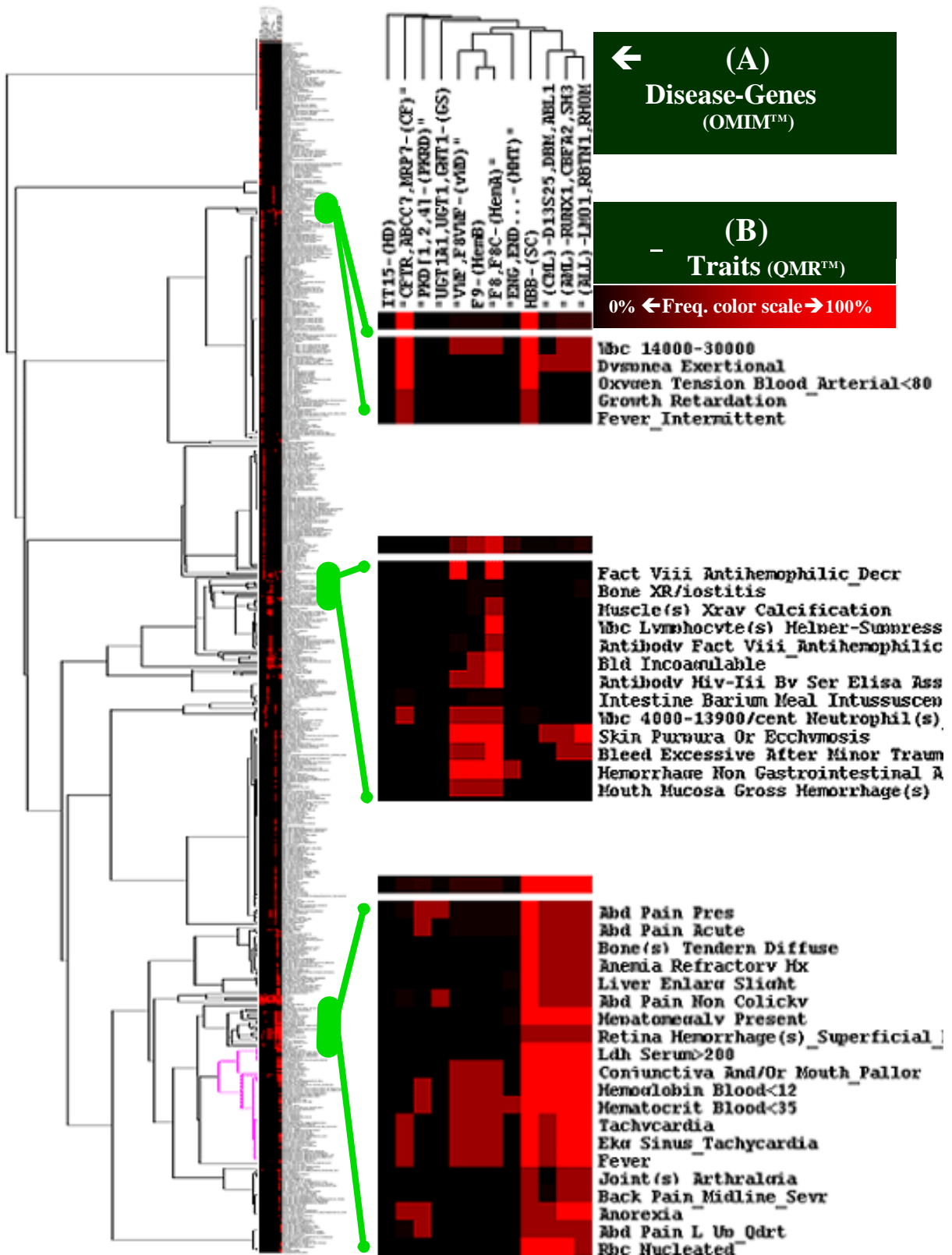
tional efforts have been deployed to corroborate the valuable relationships between the clustered traits and genes with PubMed entries (unpublished data)<sup>15</sup>.

## Discussion:

This feasibility study demonstrates that clinical knowledge can be obtained from decision support systems, subsequently associated with genetic-genomic knowledge, and finally analyzed with standard bioinformatic systems. The two dimensional clustering presented in Figure 2 has grouped the small sample of disease-genes relationships according to a classical taxonomy of diseases (e.g., hemophilia A and B were grouped together, then with von Willebrand disease, etc.). These results are consistent with another in-silico study using OMIM<sup>TM</sup> paired with textbooks which was designed to test the hypothesis that diseases can be classified according to genes<sup>16</sup>. In-vitro molecular taxonomies have also been built using clustering and neural networks with gene expression<sup>14,17</sup>, and have yielded similar results. More importantly, the proposed CGM contrasts with the later two studies in several ways: it provides 1) additional information from which new knowledge can be uncovered such as clustering traits, 2) novel gene-trait relationships that can be further

### Figure 2 – Relative Traits Expression Levels of Disease-Genes

Two dimensional hierarchical clustering was applied to 635 trait expressions data [B] (frequencies of clinical observations in the QMR™ decision support-system; 0%=black, 100%=gray or red) determined on 12 Disease-Genes [A] (from OMIM™).



explored with signal regulation pathway databases, 3) potential environment-gene relationships that may merit further future investigation.

## Conclusion

In this study, we have shown that the integration of the knowledge embedded in clinical decision support systems and in genetic/genomic knowledge bases is methodologically feasible and can provide clinical-genomic expression knowledge that can be analyzed with conventional gene clustering tools. Furthermore, there is an essential distinction between traits of expression arrays and traits of clinical knowledge bases: the former pertain to samples, while the latter are calculated over populations, not individuals.

To our knowledge, this is the first study using knowledge bases of clinical decision-support-systems for genomic discovery. Although this study is a proof of principle, it provides a framework to the development of clinical decision-support-system driven, high-throughput clinical-genomic technologies which could potentially unveil significant high-level functions of genes.

However, additional studies are required to demonstrate the significance of this method in genomic research. Additionally, this study did not address the essential functionalities required for unsupervised interoperability of the knowledge bases. We are currently investigating the problems arising from the absence of common terminologies and common semantic network, as well as the challenging queries of textual biomedical databases as OMIM<sup>TM</sup>.

## Acknowledgements

This project was partially supported by grant 528753/PO P417322 from the National Aeronautics and Space Administration (HRSA/UVC) and the 1 D1B TM 00043-01 grant from the Health Resources & Services Administration (HRSA/OAT). Additional support was provided NLM Medical Informatics Training Grant LM07079-07. We are also grateful to Dr. Philip R. Alper for his grateful suggestions and encouragements.

## References

<sup>1</sup> Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification

of cancer: class discovery and class prediction by gene expression monitoring. *Science* 199;286(5439):531-7.

<sup>2</sup> Dan S, Tsunoda T, Kitahara O, et al. . An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res.* 200;62(4):1139-47.

<sup>3</sup> Zembutsu H, Ohnishi Y, Tsunoda T, et al. Genome-wide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. *Cancer Res.* 2002 15;62(2):518-27.

<sup>4</sup> Miller R, Massarie FE, Myers JD. Quick Medical Reference (QMR) for diagnostic assistance. *MD Comput* 1986;3(5):34-48.

<sup>5</sup> Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. *N Engl J Med.* 1994 Jun 23;330(25):1792-6.

<sup>6</sup> Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. *JAMA.* 1987 Jul 3;258(1):67-74.

<sup>7</sup> Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med.* 1982 Aug 19;307(8):468-76.

<sup>8</sup> Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;291(5507):1304-51

<sup>9</sup> Online Mendelian Inheritance in Man, OMIM<sup>TM</sup>. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University and National Center for Biotechnology Information, National Library of Medicine, 2000. <http://www.ncbi.nlm.nih.gov/omim/>

<sup>10</sup> Spackman K.A., Campbell K.E., Cote R.A. SNOMED RT: A Reference Terminology for Health Care. *Proc AMIA* 1997;640-644.

<sup>11</sup> Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science.* 2001 Feb 16;291(5507):1224-9.

<sup>12</sup> Eisen MB, Spellman PT, Brown PO and Botstein D. Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc Natl Acad Sci U S A* 1998; 95, 14863-8. <http://rana.lbl.gov/EisenSoftware.htm>

<sup>13</sup> Cantor M, Lussier YA. A Knowledge Framework for Computational Molecular-Disease Relationships in Cancer. Submitted to AMIA Symposium 2002.

<sup>14</sup> Ben-Dor A, Bruhn L, Friedman N et al. Tissue Classification with gene Expression Profiles. *J Comput Biol* 2000;7(3/4):559-583.

<sup>15</sup> <http://www4.ncbi.nlm.nih.gov/PubMed/>

<sup>16</sup> Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature.* 2001;409(6822):853-5.

<sup>17</sup> Khan J, Wei JS, Ringner M, Saal LH et al. . Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med.* 2001;7(6):673-9.