Abstract of 'late breaking research' talk given at ISMB, 2012:

MalaCards – the integrated Human Malady Compendium

Marilyn Safran[1*], Noam Nativ[1], Yaron Golan[2], Irina Dalah[1], Tsippi Iny Stein[1], Gil Stelzer[1] and Doron Lancet[1] [1]Dept of Molecular Genetics, Crown Human Genome Center, The Weizmann Institute of Science, Rehovot, Israel and [2]XenneX, Inc, Cambridge, Ma, USA

Motivation: Comprehensive disease classification, integration and annotation are sorely needed for biomedical discovery, but at present, disease compilation is incomplete, heterogeneous and often lacking systematic inquiry mechanisms.

Preliminary results: We introduce MalaCards, an integrated database of human maladies and their annotations (malacards.weizmann.ac.il), modeled on the architecture and richness of the popular GeneCards database of human genes, (www.genecards.org). MalaCards similarly mines varied web data sources to generate a computerized web card for each human disease via: **1.** Identifying 7 initial sources of human disease nomenclature and annotation, targets for disease data mining (Fig. 1); **2.** Developing algorithms for judiciously merging the heterogeneous disease names in these sources, and defining unique MalaCards identifiers. For example, alzheimer's disease, ad, dat - dementia alzheimer's type, are merged under Alzheimer Disease, acronym AD, ID=ALZ001, with others listed as aliases (see malacards.weizmann.ac.il/card/index/ALZ001); **3.** Engineering scripts to automatically mine annotative information; **4.** Building the MalaCards V1.01 (*alpha*) website, with thousands of user-friendly 'cards' for all incorporated maladies, containing a variety of explicated sections (Fig. 1). As in GeneCards, the left side of the MalaCard lists the relevant mined data sources; the right side contains malady-specific information; **5.** Implementing a strategy in which detailed gene-disease relationships within GeneCards are used to create disease-specific content, leveraging the GeneCards relational database and search engine. A simple example is the Related Genes section, showing gene symbols and descriptions for all genes found to be associated textually with the key disease; **6.** Constructing a second-tier annotator, based on the Set Distiller tool of GeneDecks, a GeneCards suite member [2]. For example, other diseases related to the key disease are constructed to be those maximally associated with the set of genes found in step 5. Similarly, we obtain drugs/compounds, publications and mouse phenotypes contextually related to the key disease. **7.** Formulating scores for prioritizing the derived annotations. **8.** Initiating quality assurance based on extensive knowledge within the Crown Human Genome Center.

Algorithms: Figure 1 denotes the flow of the MalaCards system. An offline process is responsible for generating the comprehensive integrated list of diseases by mining heterogeneous, partially overlapping sources, unifying names and acronyms, and organizing characterizations. Disease name unification is effected by transforming each name to a canonical form (e.g. by lowercasing, removing special characters, and words like disease, syndrome), which is hashed and used for comparison against transformed new names. Each disease is then fed into the GeneCards search engine to find the relevant gene set, as well as publications, disease-gene associations, and the corresponding contexts wherein the match occurred (e.g. in a Gene Ontology term). The gene-set is then forwarded to GeneDecks,

which distills statistically significant descriptors (e.g. “cardiovascular system phenotype”, “apoptosis”). Finally, the online system offers a MalaCard for each disease, with sections featuring these shared descriptors, sorted by relevance. We define a new score to be the log of the rank of the GeneDecks p-value and gene-association score, multiplied by the product of the logs of the ranks of the GeneCards search hit scores for each of the genes.

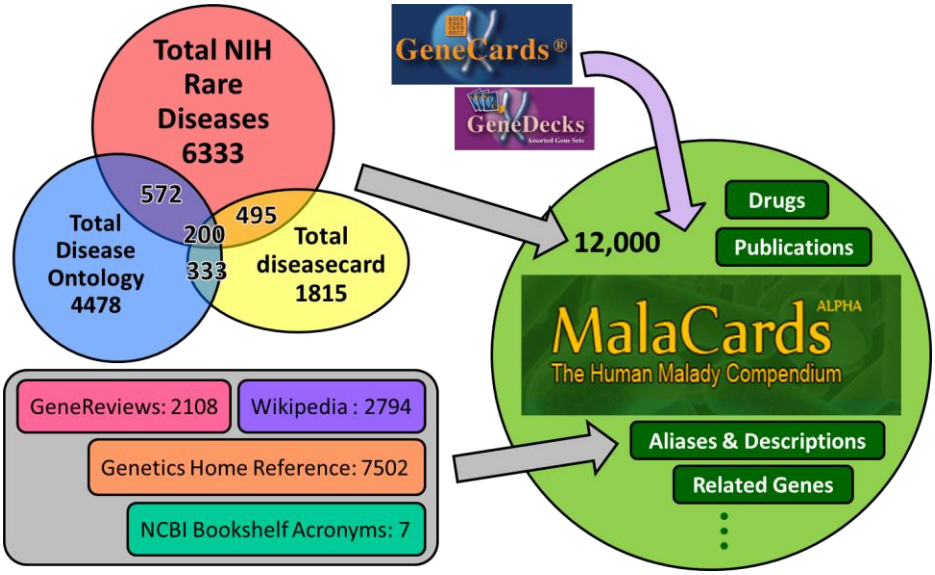

Figure 1: MalaCards architectural pipeline, showing integration of 7 heterogeneous disease lists and the leveraging of GeneCards annotations, facilitating navigation and insight generation

Discussion: MalaCards now has ~12,000 disease entries, ~5,300 with genes associated by GeneCards, many of which are putative candidates for future research. As proofs of concept of the search/distill/infer pipeline we are gratified to find expected elucidations in the *Alzheimer Disease* MalaCard, as well as potentially novel ones. Mouse phenotypes “mortality/aging” and “behavior/neurological” are enriched in 754 and 562 alzheimer-related genes respectively; disease-related pathways, such as “Parkinson’s disease” and “Huntington’s disease signaling”, which are strongly neurologically oriented, are also enriched. Since axonal and mitochondrial perturbations are widely documented as being involved with AD, it is reasonable that “The Axonal Guidance Signaling” and “Mitochondrial Dysfunction” pathways also arise. We note “Cardiac Hypertrophy Signaling” associated with 90 genes; this exemplifies an entry that merits further study.

Browsing through the list of diseases mined from 7 initial sources, the need for dealing with disease subtypes has become acutely apparent. We are exploring several options, including (as is now the case) leaving the subtypes as separate diseases. As our R&D continues, we plan to expand the list of annotation sources and sections, and include genetic variation details. This will be enhanced by collaborations with researchers outside of our group, and expanded by the initiation of systems biology tools for batch queries and smart clustering, towards the goal of enabling novel biomedical discoveries.

References : 1. Safran et al *GeneCards Version 3: the human gene integrator* Database 2010 **2.** Stelzer et al *GeneDecks: paralog hunting & gene-set distillation with GeneCards annotation* OMICS, 2009

Supported by grant from XenneX, Inc, Cambridge, Ma, USA