# CURATING GENETIC ASSOCIATION LITERATURE FOR COMMON DISEASES

Elana Silver, Eran Halperin, Kord M. Kober, Badri Padhukasahasram, Nila Patil, Michelle Sommargren, Dietrich A. Stephan, Daryl J. Thomas, Heather Trumbower, Michele Cargill

## INTRODUCTION

Papers describing genetic associations with common diseases are currently being published at a rapid rate. These new papers add to an already large body of literature which includes candidate gene studies, genome wide association studies, review papers, and meta-analyses. Related papers describe the basic epidemiology of these common diseases, gene-environment interactions, gene-gene interactions, and pharmacogenomics (gene-drug interactions), all of which may affect disease predisposition and management.

## GOALS

- **Track gene-disease associations over time to see if they are replicated**
  - Ideally different populations by different authors
  - Lack of replication is one historical limitation of genetic association studies and can be a major barrier to the adoption of personalized genomics
- **Systematically collect data**
  - Magnitude of the gene-disease effect (ie odds ratio)
  - Variant identifier
  - Allele frequency
  - Risk and non-risk alleles
  - Other key information
- **Challenges**
  - Data presentation by authors
  - Which is the risk allele?
  - DNA strand issues
  - Different coordinate systems, incomplete specification of variant location

## NAVIGENICS CURATION PROCESS

**Scouting**
- Top 7 journals as published
- RSS feed
- Directed searches
  - PubMed
  - HugeNet
  - NHGRI GWAS catalog

**Evaluation**
- Curation criteria
  - Replication
  - Study, statistical, and laboratory methods
  - Sample size and significance level
- Product fit
  - Actionability
  - Utility and relevance

**Curation**
- Genetic information
- Condition specific information
- Team of epidemiologists and human geneticists
- QC measures include ≥2 independent curations

**Testmaking**
- Affymetrix 6.0 array
- TaqMan (Applied Biosystems)

## RESULTS

We have built a literature curation database with a web interface that addresses these key needs, using Ruby on Rails with MySQL. Lack of consistent standards for reporting gene-disease associations, either by journal editors or other consortia or agencies, make automated curation infeasible at this time.

### Features

- Compare multiple independent curations
- Allow remote data entry
- Track changes
- PubMed linkage
  - Journal, author, etc. fill from PubMed id
- HapMap and dbSNP linkage
  - Check alleles and their frequency in ancestral group, strandedness

**NAVIGENICS SCIENCE BROWSER**

PUBS: NEW BROWSE | CURATED LOCUS: BROWSE | LOCUS REPLICATES: BROWSE
POSSIBLE NEW COND: NEW BROWSE | CSI: BROWSE | TMG
PUB SNPS: NEW BROWSE | TEST SNPS: NEW BROWSE | TABLE EXPORT | ADMIN | LOGOUT (5)

Show publication

Curated Locus Associated with this Publication

| CurlID | NaviCurlID | PubMed ID | Published SNP | Population |
|---|---|---|---|---|
| 178 | AAAAAR | 18650507 | rs4149056 | CAU |

| | |
|---|---|
| curlDate | 2009-02-19 |
| pubPmid | 18650507 |
| pubDoi | 10.1056/NEJMoa0801936 |
| pubTitle | SLCO1B1 variants and statin-induced myopathy--a genomewide study. |
| pubAuthors | SEARCH: Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R |
| pubJournal | N Engl J Med |
| pubDate | 2008 Aug 21 |
| pubGenes | SLCO1B1 |
| naviCondition | PGx_statin_myopathy |

*Curation browser. Publication title, author, journal, and date fill automatically from PubMed id. Publication is linked to curated locus.*

Editing Replicate Locus

Replicates Associated with this Curated Locus

| CurlRep ID | PubMed ID | Published SNP | Population |
|---|---|---|---|
| 1 | 18794853 | rs3890745 | CAU |
| 2 | 18794853 | rs3890745 | CAU |

§ Designates fields are system assigned and should not be edited.

| | Curated Locus 164 | Locus Replicate 2 |
|---|---|---|
| id § | 164 | 2 |
| naviCurlId § | AAAAAD | AAAAAD |
| naviCondition | RA | RA |
| subtype | | |
| gene | MMEL1 | MMEL |
| gender | B | B |
| pubSnp | rs3890745 | rs3890745 |
| pubRiskAllele | T | T |

*Two independent reviews are compared side by side and discrepancies are highlighted in blue*

Editing Curated Locus

Curated Publication (Pubs) Result

| PubMed ID | Authors |
|---|---|
| 18794853 | Raychaudhuri S, Remmers EF, Lee AT, Hackett R, ... |

Replicates Associated with this Curated Locus

| CurlRep ID | PubMed ID | Published SNP | Population | Notes |
|---|---|---|---|---|
| 1 | 18794853 | rs3890745 | CAU | |
| 2 | 18794853 | rs3890745 | CAU | |

Basic pubSnp Test Results

| | |
|---|---|
| Failure | pubMajorAllele (T) mismatch to HapMap2 allele 2 (G) |
| Failure | pubMinorAllele (C) mismatch to HapMap2 allele 1 (A) |
| Success | Matched population 'CAU' in HapMap2 as CEU. |
| Warning | dbSnp bilallelic status not checked. |
| Warning | dbSnp entry count not checked. |

*A variety of QC checks minimize data entry errors*

## CONCLUSIONS

- A custom built curation database with web interface can facilitate curating genetic association literature for common diseases
- QC checks are very helpful, and more could be incorporated, both interally and against external sources

 | 1001 E. Hillsdale Blvd, Suite 550, Foster City, CA USA 94404 | Telephone: (650) 585-7700 | www.navigenics.com

Presented at: Third International Biocuration Conference, Berlin, Germany, April 16-19, 2009

Nature Precedings : doi:10.1038/npre.2009.3174.1 : Posted 24 Apr 2009