A MUlti-Dimensional integrative approach to comparing genes for knowledge DIScovery

Maryam Panahiazar
PhD Candidate
Computer Science
Bioinformatics
Oct 2012
Motivation, Background

BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk.


Division of Genetics and Population Health, Queensland Institute of Medical Research, 300 Herston Road, Herston, QLD 4006, Australia; Amanda.Spurdle@qimr.edu.au; Professor David E Goldgar, Department of Dermatology, University of Utah School of Medicine, Salt Lake City, UT 84132, USA; david.goldgar@hsc.utah.edu.

Abstract

BACKGROUND: Clinical classification of rare sequence changes identified in the breast cancer susceptibility genes BRCA1 and BRCA2 is essential for appropriate genetic counselling of individuals carrying these variants. We previously showed that variant BRCA1 c.5096G>A p.Arg1699Gln in the BRCA1 transcriptional transactivation domain demonstrated equivocal results from a series of functional assays, and proposed that this variant may confer low to moderate risk of cancer.

METHODS: Measures of genetic risk (report of family history, segregation) were assessed for 68 BRCA1 c.5096G>A

Related citations in PubMed

- BRCA1 and BRCA2 mutation predictions using the BOI [Breast Cancer Res. 2006]
- Screening for BRCA1, BRCA2, CHEK2, PALB2, BRIP1 [Breast Cancer Res. 2011]
- Selected Aspects of Molecular Diagnos [Hered Cancer Clin Pract. 2006]
- Review Chromosomal mutagen sensitivit [Cytogenet Genome Res. 2004]
- Review Use of association studies to define genetic modifi [Fam Cancer. 2008]
Motivation, Background

Search for specific gene

There are links for information

No citation for similar gene records!
Motivation, Background

search for specific gene

Table of contents
Summary

<table>
<thead>
<tr>
<th>Gene</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Function</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>BRCA1</td>
</tr>
<tr>
<td>vs. M. mulatta</td>
<td>BRCA1</td>
</tr>
<tr>
<td>vs. C. lupus</td>
<td>BRCA1</td>
</tr>
<tr>
<td>vs. B. taurus</td>
<td>BRCA1</td>
</tr>
<tr>
<td>vs. M. musculus</td>
<td>BRCA1</td>
</tr>
</tbody>
</table>

BRCA1 breast cancer 1, early onset [Homo sapiens]
Gene ID: 672, updated on 16-Aug-2012

This gene encodes a nuclear phosphoprotein that plays a role in maintaining genomic
<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>symbol</th>
<th>HomoloGene</th>
<th>Muddis</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H.sapiens</td>
<td>BRCA1</td>
<td>protein</td>
<td>DNA</td>
<td>function</td>
<td>disorder</td>
<td>drug</td>
</tr>
<tr>
<td>vs. M.mulatta</td>
<td>BRCA1</td>
<td>23</td>
<td>45</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs. C.lupus</td>
<td>BRCA1</td>
<td>67</td>
<td>54</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs. B.taurus</td>
<td>BRCA1</td>
<td>33</td>
<td>45</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs. M.musculus</td>
<td>BRCA1</td>
<td>85</td>
<td>90</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motivation

• Similarity between genes from different aspects

• Curated data bases are limited
  • Not all species
  • Not all aspects

• Semantic knowledge discovery
  • Literature
  • Curated data bases
Approach

• Collect data for features
• Define the similarity functions for each feature
• Find the gene-gene similarity based on multiple features
Collect data for features

Gene of interest (geneID)

Function

Disorder/Drug
Collect data for features

Gene of interest (geneID)

- Function
- GO Terms
- Disorder/Drug
Collect data for features

1. Get functional annotation (Go terms) from EntrezGene
2. Get all ancestors for each go term
Collect data for features

1. Get functional annotation (Go terms) from EntrezGene
2. Get all ancestors for each go term
Collect data for features

Gene of interest (geneID)

Function

GO Terms

1. Get functional annotation (Go terms) from EntrezGene

2. Get all ancestors for each go term

Disorder/Drug

Mesh Terms

1. Get PubMedIDs from EntrezGene

2. Get disorder/drug annotation from PubMed abstracts (mesh terms)

3. Get all ancestors for each mesh term
Materials-Features and resources for each feature

Gene of interest (geneID)

1. Get functional annotation (Go terms) from EntrezGene
   - GO terms
     - GO:0000151
     - GO:0001726
     - GO:0005634
     - GO:0005654
     - GO:0008274
     - GO:0030529
     - GO:0043234

2. Get all ancestors for each Go term
   - Ancestor
     - GO:0005622
     - GO:0043234
     - GO:0032991
     - GO:0044464
     - GO:0005623
     - GO:0000151
     - GO:0044424
     - GO:0005575

1. Get disorder/drug annotation from PubMed abstract (mesh terms)
   - 1. Get PubMedIDs from EntrezGene
      - 2. Get disorder/drug annotation from PubMed abstract (mesh terms)
         - 3. Get all ancestors for each mesh term

Function → GO Terms → Disorder/Drug → Mesh Terms

BRCA1(627)
Define similarity function between 2 genes

Sim(g1, g2) = ?

- Gene of interest: BRAC1

- Species
  - g1: geneID: 672 from H.Sapiens
  - g2: geneID: 12189 from M.musculus

- Features
  - f1: Function of the gene
  - f2: Disorder
  - f3: Drug

<table>
<thead>
<tr>
<th>Sim(f1)</th>
<th>Sim(f2)</th>
<th>Sim(f3)</th>
<th>--------</th>
<th>Sim(fn)</th>
</tr>
</thead>
</table>

\[
sim(g_1, g_2) = \sum_{i=1}^{n} w_i \cdot sim_i(g_1, g_2)
\]

\[w_1 + w_2 + \ldots + w_i = 1\]
Define similarity function between 2 genes

<table>
<thead>
<tr>
<th>Sim(f1)</th>
<th>Sim(f2)</th>
<th>Sim(f3)</th>
<th>--------</th>
<th>Sim(fn)</th>
</tr>
</thead>
</table>

Feature-Feature similarity

- BRCA1(geneID672) - function
- BRCA1(geneID12189) - protein binding

- enzym binding
- protein binding
- breast cancer
- ovarian cancer
- Bevacizumab
- trastuzumab
- lapatinib

Simf1(g1,g2)
Define similarity function between 2 genes

\[
sim_i(g_1, g_2) = \frac{1}{m+n} \times (\sum_k \max(\text{Sim}(t_k, t_p)) + \sum_p \max(\text{Sim}(t_k, t_p)))
\]
Define similarity functions $H. Sapiens(672) \text{ vs. } M. musculus(12189)$

<table>
<thead>
<tr>
<th>Sim(g1, g2)</th>
<th>Sim(f1)</th>
<th>Sim(f2)</th>
<th>Sim(f3)</th>
<th>--------</th>
<th>Sim(fn)</th>
</tr>
</thead>
</table>

**Feature-Feature similarity**
- Sim(f1)
- Sim(f2)
- Sim(f3)
- --------
- Sim(fn)

**Set-Set similarity for each feature**
- Sim(t1, t2)

**Term-Term similarity**
- Ancestors
  - GO:0044464
  - GO:0005623
  - GO:0001726
- Ancestors
  - GO:0005622
  - GO:0044424
  - GO:0005575

**Drug**
- Bevacizumab

**Protein binding**
- Breast cancer
- Ovarian cancer

**Disorder**
- Trastuzumab
- Lapatinib

**Function**
- Enzym binding
- Protein binding
- Breast cancer

**Examples**
- BRCA1(geneID672)
- BRCA1(geneID12189)
**Similarity formula**

**Term-Term similarity**

\[
sim(t_1, t_2) = -\log \frac{|T(t_1) \cup T(t_2)| - |T(t_1) \cap T(t_2)|}{|T(t_1) \cup T(t_2)|}
\]

Francisco Azuaje, Haiying Wang and Olivier Bodenreider 2005
Similarity formula

Aggregation of Feature-Feature similarity

\[
sim(g_1, g_2) = \sum_{i=1}^{n} w_i \cdot \text{sim}_i(g_1, g_2)
\]

where \( w_1 + w_2 + \ldots + w_i = 1 \)

Set-Set similarity for each feature \( i \)

\[
sim_i(g_1, g_2) = \frac{1}{m + n} \times (\sum_k \max(\text{sim}(t_k, t_p)) + \sum_p \max(\text{sim}(t_k, t_p)))
\]

Term-Term similarity

\[
sim(t_1, t_2) = -\log \left( \frac{|T(t_1) \cup T(t_2)| - |T(t_1) \cap T(t_2)|}{|T(t_1) \cup T(t_2)|} \right)
\]

Francisco Azuaje, Haiying Wang and Olivier Bodenreider 2005
Use Case

Find the similar genes to BRCA1

BRCA1 -- P -- O (?)

BRCA2

BRCA1

ZBTB7A
## Intermediate results

**BRCA1 (Gene ID: 672) in Homo sapiens**

<table>
<thead>
<tr>
<th>Species</th>
<th>GeneID</th>
<th>Sequence similarity</th>
<th>Muddis similarity**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Protein*</td>
<td>DNA*</td>
</tr>
<tr>
<td>vs. M.musculus</td>
<td>12189</td>
<td>58.1</td>
<td>74.4</td>
</tr>
<tr>
<td>vs. C.lupus</td>
<td>403437</td>
<td>74.8</td>
<td>84.2</td>
</tr>
<tr>
<td>vs. R.norvegicus</td>
<td>24227</td>
<td>58.3</td>
<td>75.2</td>
</tr>
<tr>
<td>vs. B.taurus</td>
<td>353120</td>
<td>72.6</td>
<td>83.8</td>
</tr>
<tr>
<td>vs. G.gallus</td>
<td>373983</td>
<td>34.3</td>
<td>50.0</td>
</tr>
<tr>
<td>vs. P.troglodytes</td>
<td>449497</td>
<td>98.2</td>
<td>99.3</td>
</tr>
<tr>
<td>vs. M.mulatta</td>
<td>712634</td>
<td>93.1</td>
<td>96.1</td>
</tr>
</tbody>
</table>

*Pairwise Alignment Scores from HomoloGene

** Results from defined similarity functions
Evaluation frame work

The hypothesis is:
“The results of our similarity functions go with the control similarity”

- Control: Similarity score from
  Find the **Correlation** between feature based similarity to control

- Control: Similarity from Curated data such as
  Choose control from **orthologous** genes
Find the correlation between score of Muddis similarity and the score of HomoloGene

- High control similarity & low Muddis similarity: False negative
- Low control similarity & High Muddis similarity: False positive
Find the correlation between score of Muddis similarity and the score of HomoloGene

Correlation \{\text{score(muddis)}, \text{score(homologene)}\} = 0.67

• High control similarity & low Muddis similarity: False negative
• Low control similarity & High Muddis similarity: False positive
Discussion

Significance of findings:
• False positives: Candidates for missed gene pairs with high similarity
• False negatives: No correlation between sequence and functional similarity

The results of this research are useful in:
• Model organism research
• Drug target discovery research in human
Future work

- Apply framework in different use cases and evaluate them.
- Use the results for knowledge discovery.
- Add or remove dimensions for similarity comparison.
- Develop an interactive user interface.
Acknowledgment

Mentor at knoesis: Dr. Amit Sheth
Mentor at NIH: Dr. Olivier Bodenreider

Dr. Bastien Rance
Dr. Rainer Winnenburg
Dr. Marcelo Fiszman
May Cheh
Use Case A

Find the similar genes to BRCA1

Use Case B

How similar they are?
Thanks for your attention

Questions?
(A) Term-term similarity

\[ \text{sim}(t_1, t_2) = -\log \frac{|T(t_1) \cup T(t_2)| - |T(t_1) \cap T(t_2)|}{|T(t_1) \cup T(t_2)|} \]

(B) Set-set similarity for each feature \( i \)

\[ \text{sim}_i(g_1, g_2) = \frac{1}{m + n} \times \left( \sum_k \max(\text{sim}(t_k, t_p)) + \sum_p \max(\text{sim}(t_k, t_p)) \right) \]

(C) Feature-feature similarity

\[ \text{sim}(g_1, g_2) = \sum_{i=1}^{n} w_i \text{sim}_i(g_1, g_2) \]