Text mining for improving the prioritization, curation, and integration of knowledge for clinically relevant variants

Zhiyong Lu, 陆致用

NCBI, NLM, NIH
Highly pathogenic Alzheimer's disease presenilin 1 P117R mutation causes a specific increase in p53 and p21 protein levels and cell cycle dysregulation in human lymphocytes.

Bottleneck in Database Curation

Over 1 million new articles per year

Two articles per minute!
**Homozygously deleted gene DACH1 regulates tumor-initiating activity of glioma cells.**

Aiko Watanabe, Akane Ogawa, Shinya Ishida, Atsuki Morikawa, Shigetoshi Fujii, Kazuaki Ueda, Yosuke Matsuo, Hirotaka Akutsu, Yumiko Kaneko, and Masaaki Tsuchida.

**A novel DNVR31 mutation associates with variable degrees of auditory and Portuguese family.**

Isabelli Anto, Reina Fujiki, Hideto Kato, Akiko Morita, Shinya Yoneda, Masaaki Tsuchida, Kenji Hashimoto, and Kiyohiko Ishida.

**Genetic and Epigenetic Alterations of the NF2 Gene in Sporadic Vestibular Schwannomas.**

Jong Hae Lee, Tae Jun Kwon, Hye Jung Kwon, and Byung-Hee Kwon.

**Variation in the CXCR1 gene is associated with chronic rhinosinusitis.**

Curtin, Rivette, Unnasch, and O'Brien.

**Conclusions:**

1. The chemokine receptor CCR1 is upregulated in chronic rhinosinusitis with nasal polyps compared to controls.
2. The CCR1 polymorphism rs1421280 is associated with chronic rhinosinusitis with nasal polyps.

**Introduction:**

Chronic rhinosinusitis (CRS) is a chronic inflammatory disease of the nasal cavity and paranasal sinuses. It is characterized by symptoms including nasal obstruction, rhinorrhea, and facial pain.

**Methods:**

A genomewide association study was conducted on 300 patients with CRS with nasal polyps and 300 healthy controls.

**Results:**

The CCR1 polymorphism rs1421280 was associated with CRS with nasal polyps (p = 0.007).

**Conclusion:**

The CCR1 polymorphism rs1421280 may be involved in the pathogenesis of CRS with nasal polyps.
Mutation Nomenclature

Human Genome Variation Society (HGVS)
- Recommended standards
- c.77A>C

Possible variations
- c.77A→C
- 77A>C
- A>C
- A77C

Mutation mentions in corpus
- 23% Fit HGVS
- 77% Do not fit
A novel missense mutation Asp506Gly in codon 506 of the F11 gene in an asymptomatic Korean woman with mild factor XI deficiency

Lee JH, Cho HS, Hyun MS, Kim HY, Kim HJ.
Department of Laboratory Medicine, Yeungnam University College of Medicine, 317-1 Daemyeong-dong, Nam-gu, Daegu, Korea.

Abstract
Factor XI (FXI) deficiency is a rare bleeding disorder, but it is also found in other ethnic groups. It is a trauma or surgery-related bleeding disorder, but spontaneous bleeding is rarely seen. The clinical manifestation of bleeding in FXI deficiency cases is variable and seems to poorly correlate with plasma FXI levels. The molecular pathology of FXI deficiency is mutation in the F11 gene on the chromosome band 4q35. We report a novel mutation of the F11 gene in an 18-year-old asymptomatic Korean woman with mild FXI deficiency who was born to a family with FXI deficiency. The A to G substitution in nucleotide 1517 (c.1517A>G) results in substitution of aspartic acid with glycine in codon 506 (p.As506Gly). To the best of our knowledge, the Asp506Gly is a novel missense mutation, and this is the first genetically confirmed case of mild FXI deficiency in Korea.
Variant name normalization

- **tmVar**
  - Mutation detection
  - Wei et al., *bioinformatics*, 2013

- **tmVar 2.0**
  - Mutation normalization
  - Wei et al., *bioinformatics*, 2017

Various mentions in literature

HGVS names
tmVar

- Approach: machine learning (CRF) + patterns
- High performance (90%+ accuracy)
- Easy access through PubTator RESTful APIs
Enriching literature links (PMID-RSID pairs)

Overlap: 68,969

"Freestyle" mutation mention(s) in abstract

tmVar (auto): 230,147

dbSNP (Annotated): 308,517

33,192 RS have no PubMed links in dbSNP
American College of Medical Genetics (ACMG)

- 58 Genes
- Disease-causing: cancers, cardiac diseases, etc.
- Medically actionable
- Newly discovered variants (expected pathogenic) to be reported
Challenges of linking genes and diseases

**TITLE:**
Factors associated with oxidative stress and cancer risk in the Breast and Prostate Cancer Cohort Consortium.

**ABSTRACT:**
Both endogenous factors (genomic variations) and exogenous factors (environmental exposures, lifestyle) impact the balance of reactive oxygen species (ROS). Variants of the ND3 (rs285346, G10398A) gene of the mitochondrial genome, manganese superoxide dismutase (MnSOD; rs4880 Val16Ala) and glutathione peroxidase (GPX-1; rs1050450 Pro198Leu), are purported to have functional effects on regulation of ROS balance. In this study, we examined associations of breast and prostate cancer risks and survival with these variants, and interactions between rs4880-rs1050450, and alcohol consumption-rs2853826. Nested case-control studies were conducted in the Breast and Prostate Cancer Cohort Consortium (BPC3), consisting of nine cohorts. The analyses included over 10726 post-menopausal breast and 7532 prostate cancer cases with matched controls. Logistic regression models were used to evaluate associations with risk, and proportional hazard models were used for survival outcomes. We did not observe significant interactions between polymorphisms in MnSOD and GPX-1, or between mitochondrial polymorphisms and alcohol intake and risk of either breast (p-interaction of 0.34 and 0.98, respectively) or prostate cancer (p-interaction of 0.49 and 0.50, respectively). We observed a weak inverse association between prostate cancer risk and GPX-1 Leu198Leu carriers (OR 0.87, 95% CI 0.79-0.97, p = 0.01). Overall survival among women with breast cancer was inversely associated with G10398 carriers who consumed alcohol (HR 0.66 95% CI 0.49-0.88). Given the high power in our study, it is unlikely that interactions tested have more than moderate effects on breast or prostate cancer risk. Observed associations need both further epidemiological and biological confirmation.

**Positive Assertion**

**Negative Finding**
Table 2. Precision computation based on human annotation of random samples

<table>
<thead>
<tr>
<th>Frequency Group</th>
<th>High (138)</th>
<th>Medium (343)</th>
<th>Low (4903)</th>
<th>Total (5384)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct triplets</td>
<td>47</td>
<td>89</td>
<td>195</td>
<td>331</td>
</tr>
<tr>
<td>Triplets evaluated</td>
<td>58</td>
<td>112</td>
<td>260</td>
<td>430</td>
</tr>
<tr>
<td>Accuracy (precision)</td>
<td>0.81</td>
<td>0.80</td>
<td>0.75</td>
<td>0.77</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pcbi.1005017.t002
On expert curation and scalability: UniProtKB/Swiss-Prot as a case study

Sylvain Poux, Cecilia N. Arighi, Michele Magrane, Alex Bateman, Chih-Hsuan Wei, Zhiyong Lu, Emmanuel Boutet, Hema Bye-A-Jee, Maria Livia Famiglietti, Bernd Rocchet. The UniProt Consortium

Bioinformatics, btx439, https://doi.org/10.1093/bioinformatics/btx439
Published: 13 July 2017 Article history
Related work

- **Mutation extraction tools & corpora**
  - MutationFinder (Caporaso et al. 2007), EMU (Doughty et al. 2011), SETH (Thomas et al 2016) Nala (Cejuela et al 2017), etc
  - Tool evaluations (e.g. Jimeno & Verspoor, 2014)

- **Mutation-related relation extraction**
  - MuteXt (2004); MuGeX (2007); EnzyMiner (2009); OSIRIS (2006) ...
  - PhenoMiner (Collier, Groza et al., 2015)
  - EMU (Doughty, et al., 2011)
  - DiMeX (Mahmood, et al., 2016)
Acknowledgments

Panel Organizers: Cecilia Arighi & Fabio Rinaldi

NCBI: Alexis Allot, Kyubum Lee, Yifan Peng, Chih-Hsuan Wei, Michael Simmons, Ayush Singhal

dbSNP Collaborators: Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon

UniProt Collaborators: Alex Bateman, Cathy Wu, Ioannis Xenarios, Sylvain Poux, Cecila Arighi, Michele Magrane et al.
Questions?

Thank you!

zhiyong.lu@nih.gov

tmVar & PubTator available at

Limitations & Future Work

- Full text analysis
  - Tables, figures
  - Supplementary materials

- Public release
  - Text-mined variants
  - New & improved software tools

- Complex relationships