UpSetR Korean Papers

Jake Conway

November 3, 2016

De *novo* assembly and phasing of a Korean human genome (Nature)

- Identified 18,210 structural variants by comparing assembly with reference genome
- Many **insertions** are reflected in transcriptome and are **shared across the Asian population**
- Specifically, identified 7,358 deletions, 10,077 insertions, 71 inversions, and 704 complex variants
- Of the 10,077 insertions, 7,710 were novel
  - Of the novel insertions most were mobile elements and tandem repeats
- DAVID tool showed that 77% of non-synonymous variants were predominantly related to *ion binding*, *epidermal growth factor*, and *fibronectin*
- Focused on gene AK1, and noted that it contained repeats, duplications, and *unique sequences that are not found in the reference genome*
- Other genes that had very large insertions (bp wise) that were specific to the Asian population were POU2F3, HRASLS2, and ANO2

Identification of Korean-specific SNP markers from whole-exome sequencing data (Int J Legal Med)

- Identified 300 Korean-specific SNPs from 306 Korean individuals
  - DNA was extracted from peripheral blood samples
- Selected Korean SNPs that had minor allele frequency (MAF) between 0.3 and 0.7
- Filtered by removing non-synonymous variants, and then selecting SNPs that had a MAF < 0.3 in 1000 Genomes project
  - Basically, kept synonymous variants that were common in Koreans but not common in global population
- Eventually selected 300 SNPs
- Found that the MAF of these SNPs were correlated with their MAF in Chinese and Japanese populations
- Also found the MAF of these SNPs were not correlated with their MAF American, Western European, Puerto Rican populations
- It appears that using their method they were not looking for insertions and deletions

Conclusions

- Nothing to really take away from papers
- Megan didn't observe the same pathways referenced in the Nature paper using DAVID tool
- Second paper limited to base changes. Did not look at insertions and deletions
- Looking back through Megan's final presentation, it was neat to see how both Korean cancer cohorts overlapped a lot until we filtered for only SNPs in protein coding regions