Characterization of the Korean Genome

Deepali Jhamb*, Meeta Pradhan*, Premkumar Duraiswamy, Akshay Desai, Mathew J. Palakal**
School of Informatics, Indiana University Purdue University Indianapolis, Indiana, USA, 46202
*First Authors, **Corresponding Author
E-mail Addresses: DJ: djhamb@iupui.edu, MP: mpradhan@iupui.edu, PD: premdura@iupui.edu, AD: akdesai@iupui.edu, MJP: mpalakal@iupui.edu

Introduction

Genome wide association studies have identified a large number of common variants associated with complex diseases but despite such efforts a very small fraction of disease heritability and phenotypic variation has been explained by these studies. Whole genome sequencing (WGS) provides the most comprehensive view of an organism's genetic variation. A major challenge is to understand these variants with respect to their functions and interactions with other variants. In this study, we developed an innovative systems biology pipeline for the analysis of next generation sequencing data to discover functional and pathway modules specific for the Korean population. We used the SNV data from the KPGP-38 Human Genomes dataset to understand the interrelationships between the genes with the rare variants. The methodology developed here can also be easily applied to include other variants identified by WGS such as INDELS, CNV, and SV and also other types of next generation sequencing data. Our analysis identified genes related to neurodegeneration, cancer, and hypertension associated with the KPGP-38 Human Genome.

Methodology

SNV data for 37 Korean samples (one sample for Caucasian female, KPGP10, was excluded from our analysis since we focused on the analysis of Korean samples) was extracted in the vcf format. GATK CombineVariants was used to merge these samples into one vcf file (korean_merge). To compare Korean SNVs with other populations, 1000 genome data (1000G) was downloaded. Rare variants in the Korean population were identified (with an allele frequency (AF) less than three in 1000G and those that were unique in the korean_merge were also included) and further analyzed using systems biology approaches. SNPEff was used to annotate the rare variants. The physical interactions among the rare variant genes were extracted using BioGRID. A novel algorithm was designed to identify significant modules from the rare variant network. To initiate this process, a "seed matrix" was constructed which was based on the number of variants in all the regions of a gene. Due to the low representation, we ignored following regions from our analysis: Intragenic, Synonymous Stop, Non Synonymous Start, Stop Lost and Stop Gained. The variants were further normalized based on the remaining nine regions. The gene with highest normalized score was selected as the seed to proceed for constructing the modules. This normalized score was defined as the NodeWeight. The module was expanded based on the NodeWeight, Gene Ontology biological process ≥ 70% and maximum pathway similarity between the seed and the leaf node. We constructed modules for the top 100 seeds from seed matrix based on the above three conditions: The top scoring gene was selected as the seed and the module expansion was performed based on the biological process and pathway match between the connecting nodes. We constructed such modules for the top 100 seeds from the seed matrix. The disease information was overlaid on these modules using the OMIM database.
Below we describe the pseudo code for the identification of seed nodes:

```
For ( i in range (1, total genes))
Normalized i
For (j in range of (1, total genes)
        $score\ j = \sum_{k=1}^{9} region\ of\ k$
sort j extract top 100 genes
```

Results

Figure1a and Figure1b show the distribution of genes and rare variants in the Korean vs. other populations respectively. Rare variants in the 1000G data were calculated based on AF<3 for the respective populations. Twenty five different regions were identified in other populations from 1000G data (26 in European) however genes in the Korean population were distributed across 17 regions. Regions such as "Frame shift" and "Exon deleted" were not found in the Korean rare variants. Even though the number of genes identified were similar for the top chromosome regions, they were relatively less for Koreans in rest of the genomic regions. The number of variants identified in the Korean population were significantly lower than the rest of the populations. This could be due to the low sample size of Korean population as compared to the 1000G data.

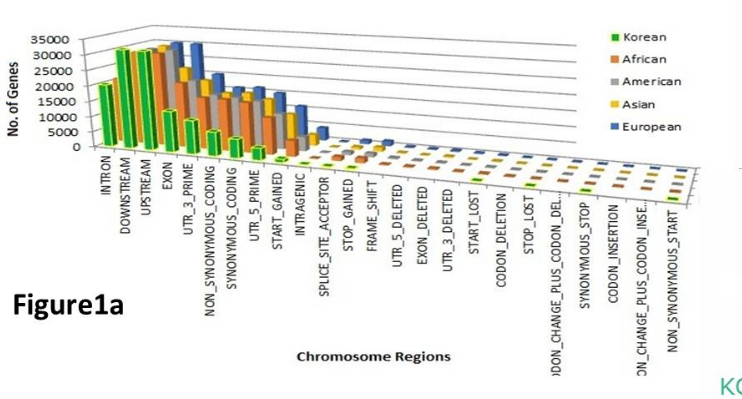

Figure1a

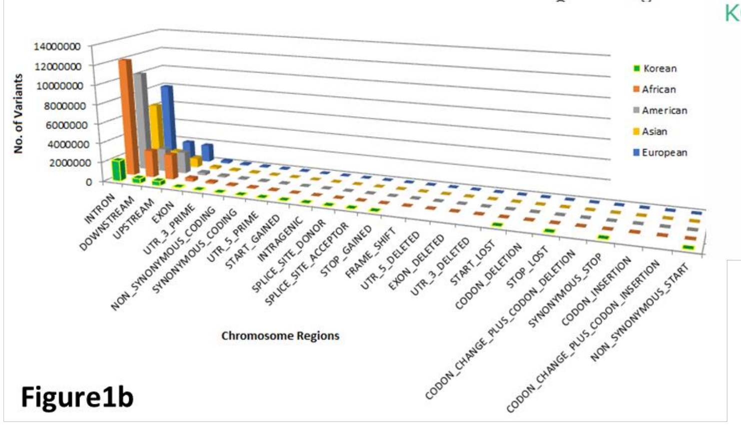

Figure1b

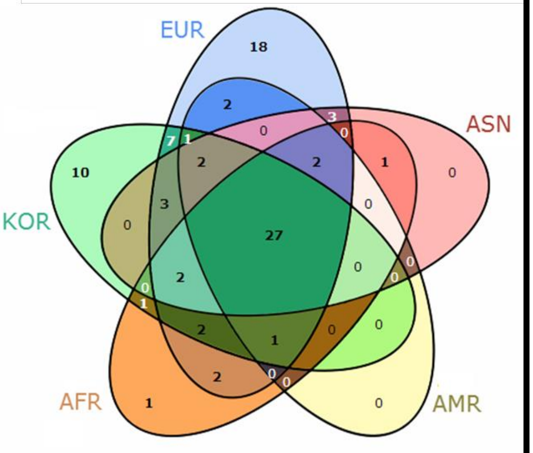

Figure2

Biological processes (BPs) for genes of all the populations were extracted using DAVID and the comparison of top 10 biological processes is shown in **Figure2**. A total of ten unique BPs were identified for the Korean population (KOR) however no unique BPs were identified in the Asian (ASN) and American (AMR) population. Eighteen unique BPs were identified for the European (EUR) population and one for the African (AFR) population. Twenty seven BPs were common to all the populations. Interestingly, the BPs unique to KORs were all related to brain such as cerebellum morphogenesis, and cerebellar cortex development among others. A recent study shows the prevalence of autism in South Korea which is estimated to be 2.6 percent and 1child in every 38 children is affected with this disorder [1]. It has also been previously reported that widely known dysbindin gene, DTNBP1, in schizophrenia is not associated in the Korean population [2]. Our analysis also did not identify this gene.

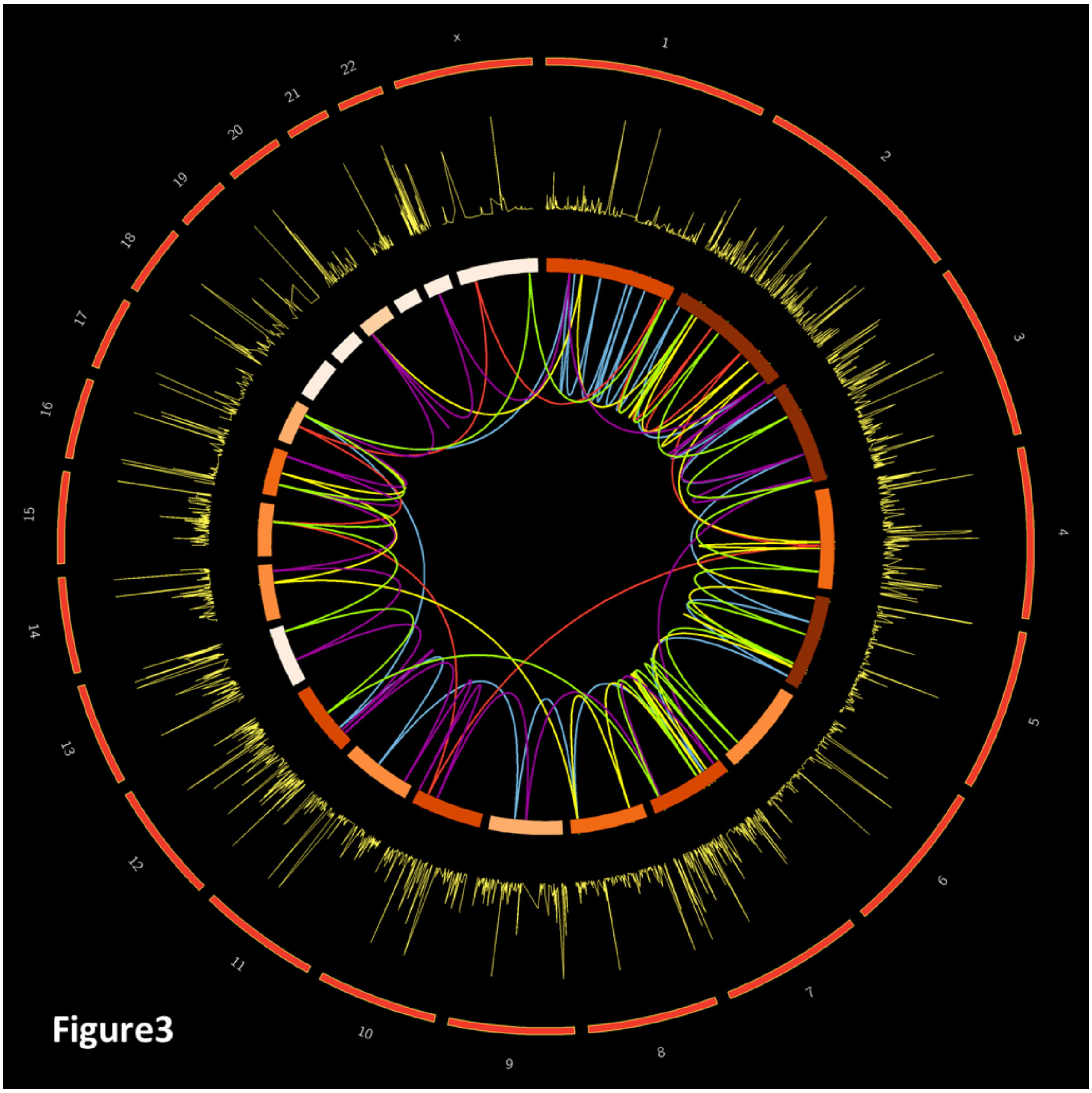

Figure3

Figure3 depicts the overall distribution of the rare variants and KEGG pathways across the chromosomes. Only genes with greater than 200 variants were included in this analysis. The outermost circle in the figure represents the chromosomes, second circle shows the number of variants for genes in each chromosome, third circle indicates the frequency of pathway distribution (deeper orange color – higher number of variants), and fourth circle in the center shows the interconnected genes from different chromosomes participating in a pathway.

Chromosomes 3, 1, 2, 7, and 5 were found to have the highest number of rare variants in the Korean population. However, chromosome 16 has the maximum no. of genes (57) with variants above average followed by chromosomes 5, 2, 3, and 1. Chromosomes 2, 3, 5, 1, and 7 have the highest no. of pathways. Top ten pathways of these genes were plotted in the center of this image. It can be seen that several genes from different chromosomes interact together to participate in pathways. Neuroactive ligand-receptor interaction (colored blue in **Figure3**), and axon guidance (colored red) pathways have genes distributed across 12 chromosomes each. It has been indicated in the literature that genes associated to these chromosomes are correlated with neurodegenerative diseases in Koreans as well as other populations [3, 4, and 5].

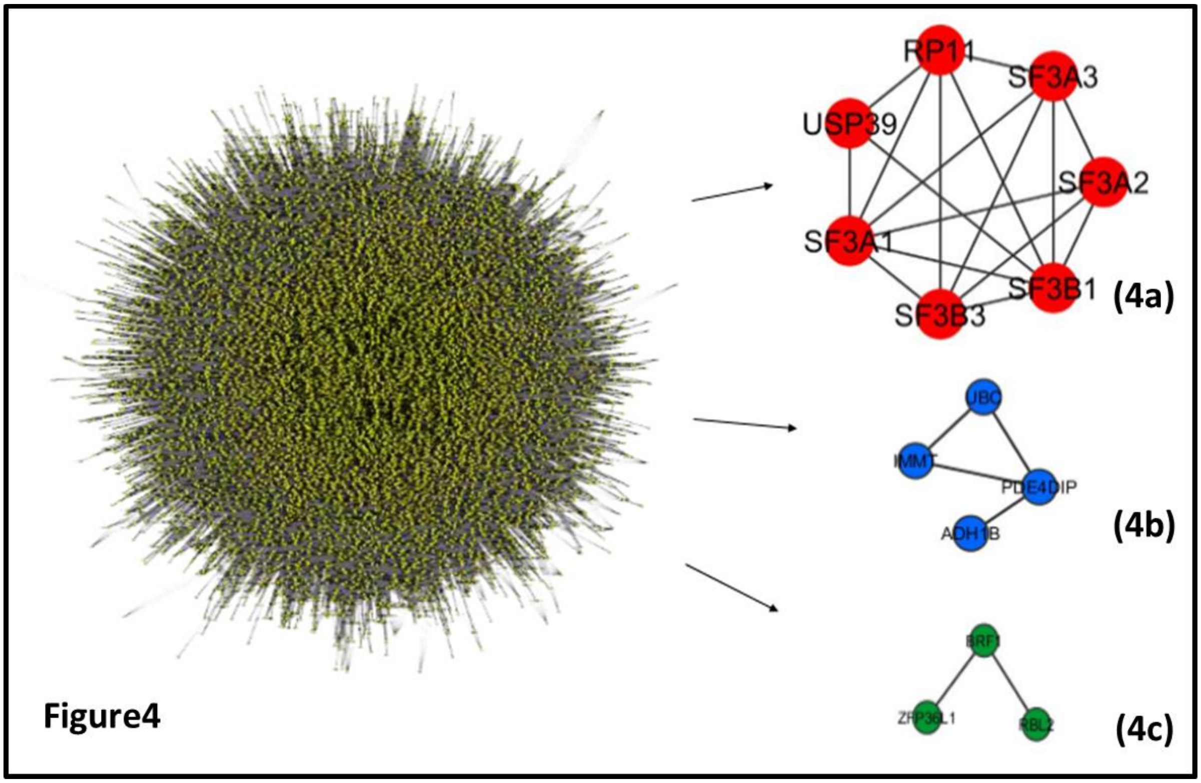

Figure4

Systems biology analysis was further performed to identify significant genes and functional modules which are present in the Korean population. Functional modules were extracted as described in the methodology section. **Figure4** represents the network and few important modules of different sizes. We obtained modules of size 3-9 and these were ranked based on their node property and edge property. The genes in the modules were analyzed for their association with any disease or a disease reported in literature associated with Korean population. One of the most important modules was RP11 module (**Figure 4a**): RP11 SF3A1 SF3A3 SF3A2 USP39 SF3B3 SF3B1 PRPF3. A novel mutation in the RP11 gene is known to cause retinitis pigmentosa in the Chinese population [6], SF3A1 is known to be associated with myleodysplasia and retinoblastoma associated binding protein [7], SF3A2 with CNV aberrations in Korean contribute to AML [8], USP9 with gastric cancer [9], SF3B3 Gastric cancer cell [10], SF3B1 with tumors [11]. This analysis shows that the genes in this module are all correlated with tumor. Another top scored module identified in this work consisted of PDE4DIP, IMMT, UBC ADH1B genes (**Figure 4b**). Detailed analysis of genes in this module associated PDE4DIP with psychiatric disorders [12], IMMT with spinocerebellar ataxia [13], UBC with hungtinton disease [14], ADH1B as ethnic variant in Asian population [15], correlating this module with neurodegenerative diseases. In addition, genes associated with the height of Korean population and hypertensions were also found in these highly ranked modules. Several other modules which were identified with the GO biological processes unique to Koreans (**Figure2**) were also analyzed and found to correlate to the brain or neuro diseases. Overall it was observed that modules extracted in our study using the system's biology approach could be classified into two major domains: (i) Neurodegenerative diseases and (ii) Tumor related genes. This observation suggests that the Korean population might be more susceptible to these two major classes of complex diseases.

Conclusions

In this work, we identified rare variants unique to the Korean population which were mainly distributed across chromosomes 1, 2, 3, 5, and 7. The novel systems biology approach developed here identified modules whose genes were reported in the literature to be correlated as markers of neurodegenerative diseases and tumor. Systems biology can help elucidate the major variants in a population. The power of systems biology should be exploited to reduce the "big data" generated by next generation sequencing studies and identify significant variants.

Acknowledgements

We would like to acknowledge National Center for Genome Analysis Support, Indiana University, for providing the supercomputing support for the data analysis. We would also like to acknowledge Sammed Mandape for annotating the genes with PharmGKB knowledge base and Shruti Sakhare for literature analysis.

References

1. http://www.autismspeaks.org/about-us/press-releases/new-study-reveals-autism-prevalence-south-korea-estimated-be-26-or-1-38-chil.
2. Joo, E.Y. et. al., Neurosci lett. 2006 Oct 23; 407(2):101-6. Epub 2006 Sep 7.
3. Shin, J. H. et. al., Chest. 2005 Oct; 128(4):2999-3003.
4. Lee, H.R. et. al., Cancer Genet Cytogenet. 2010 Dec; 203(2):193-202.
5. Genetics Home reference (http://medline.gov/)
6. Xia et al., Mol.Vis. 2004, May 20; 10:361-5
7. Wang et. al., J. Proteome Res. 2009 Oct;8(10):4428-40
8. Yun H.J. et. al., Bio Chem. 2008 Oct; 389(10):1313-8
9. Yang, S. et. al., Genomics 2007 Apr;89(4):451-9.
10. Kang, H.C. et. ,al. Clin. Cancer Res. 2004 Jan 1:10(1 Pt 1):272-84
11. Eun Mi Je et al., IJC 2013, Vol. 133, Issue no. 1, pages 260-265.
12. Kim, S. et. al., Transl Psychiatry 2012 May; 2(5)e113.
13. Lee, L.C. et. al., Clin Chim Acta. 2009 Feb; 400(1-2):56-62.
14. Bett, J.S. et. al., J Cell Mol. Med. 2009 Aug;13(8B):2645-57
15. Li, H. et. al., PLoS One. 2008 Apr 2;3(4):e1881