DGV Newsletter May 2016

Hello!

The *Database of Genomic Variants* has recently been updated. In this newsletter, we will give an overview of the data added, and the changes that have been made to the website. The latest updates include five new datasets and we have updated the DGV Gold Standard curated set of variants.

New Studies and New Datasets Added to the Database of Genomic Variants

1. Suktitipat2014. Study Accession = estd224

Copy number variation in Thai population. Suktitipat B, Naktang C, Mhuantong W, Tularak T, Artiwet P, Pasomsap E, Jongjaroenprasert W, Fuchareon S, Mahasirimongkol S, Chantratita W, Yimwadsana B, Charoensawan V, Jinawath N. *PLoS One*. 2014 Aug 13;9(8):e104355

Copy number variation (CNV) is a major genetic polymorphism contributing to genetic diversity and human evolution. Clinical application of CNVs for diagnostic purposes largely depends on sufficient population CNV data for accurate interpretation. CNVs from general population in currently available databases help classify CNVs of uncertain clinical significance, and benign CNVs. Earlier studies of CNV distribution in several populations worldwide showed that a significant fraction of CNVs are population specific. In this study, we characterized and analyzed CNVs in 3,017 unrelated Thai individuals genotyped with the Illumina Human610, Illumina Human Omni Express, or Illumina HapMap550v3 platform. We employed hidden Markov model and circular binary segmentation methods to identify CNVs, extracted 23,458 CNVs consistently identified by both algorithms, and cataloged these high confident CNVs into our publicly available Thai CNV database. Analysis of CNVs in the Thai population identified a median of eight autosomal CNVs per individual. Most CNVs (96.73%) did not overlap with any known chromosomal imbalance syndromes documented in the DECIPHER database. When compared with CNVs in the 11 HapMap3 populations, CNVs found in the Thai population shared several characteristics with CNVs characterized in HapMap3. Common CNVs in Thais had similar frequencies to those in the HapMap3 populations, and all high frequency CNVs (>20%) found in

Thai individuals could also be identified in HapMap3. The majorities of CNVs discovered in the Thai population, however, were of low frequency, or uniquely identified in Thais. When performing hierarchical clustering using CNV frequencies, the CNV data were clustered into Africans, Europeans, and Asians, in line with the clustering performed with single nucleotide polymorphism (SNP) data. As CNV data are specific to origin of population, our population-specific reference database will serve as a valuable addition to the existing resources for the investigation of clinical significance of CNVs in Thais and related ethnicities.

2. Thareja2015. Study Accession = nstd99

Sequence and analysis of a whole genome from Kuwaiti population subgroup of Persian ancestry. Thareja G, John SE, Hebbar P, Behbehani K, Thanaraj TA, Alsmadi O. *BMC Genomics*. 2015 Feb 18;16:92

The Kuwait Genome Project an initiative to sequence individual genomes from the three subgroups of Kuwaiti population namely, Saudi Arabian tribe; "tent-dwelling" Bedouin; and Persian, attributing their ancestry to different regions in Arabian Peninsula and to modern-day Iran (West Asia). These subgroups were in line with settlement history and are confirmed by genetic studies. In this work, we report whole genome sequence of a Kuwaiti native from Persian subgroup at >37X coverage. RESULTS: We document 3,573,824 SNPs, 404,090 insertions/deletions, and 11,138 structural variations. Out of the reported SNPs and indels, 85,939 are novel. We identify 295 'loss-of-function' and 2,314 'deleterious' coding variants, some of which carry homozygous genotypes in the sequenced genome; the associated phenotypes include pharmacogenomic traits such as greater triglyceride lowering ability with fenofibrate treatment, and requirement of high warfarin dosage to elicit anticoagulation response. 6,328 non-coding SNPs associate with 811 phenotype traits: in congruence with medical history of the participant for Type 2 diabetes and ß-Thalassemia, and of participant's family for migraine, 72 (of 159 known) Type 2 diabetes, 3 (of 4) ß-Thalassemia, and 76 (of 169) migraine variants are seen in the genome. Intergenome comparisons based on shared disease-causing variants, positions the sequenced genome between Asian and European genomes in congruence with geographical location of the region. On comparison, bead arrays perform better than sequencing platforms in correctly calling genotypes in low-coverage sequenced genome regions however in the event of novel SNP or indel near genotype calling position can lead to false calls using bead arrays. CONCLUSIONS: We report, for the first time, reference genome resource for the population of Persian ancestry. The resource provides a starting point for designing large-scale genetic studies in Peninsula including Kuwait, and Persian population. Such efforts on populations under-represented in global genome variation surveys help augment current knowledge on human genome diversity.

3. Alsmadi2014. Study Accession = nstd106

Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from Kuwaiti population subgroup of inferred Saudi Arabian tribe ancestry. Alsmadi O, John SE, Thareja G, Hebbar P, Antony D, Behbehani K, Thanaraj TA. *PLoS One*. 2014 Jun 4;9(6):e99069.

Population of the State of Kuwait is composed of three genetic subgroups of inferred Persian, Saudi Arabian tribe and Bedouin ancestry. The Saudi Arabian tribe subgroup traces its origin to the Najd region of Saudi Arabia. By sequencing two whole genomes and thirteen exomes from this subgroup at high coverage (>40X), we identify 4,950,724 Single Nucleotide Polymorphisms (SNPs), 515,802 indels and 39,762 structural variations. Of the identified variants, 10,098 (8.3%) exomic SNPs, 139,923 (2.9%) non-exomic SNPs, 5,256 (54.3%) exomic indels, and 374,959 (74.08%) non-exomic indels are 'novel'. Up to 8,070 (79.9%) of the reported novel biallelic exomic SNPs are seen in low frequency (minor allele frequency <5%). We observe 5,462 known and 1,004 novel potentially deleterious nonsynonymous SNPs. Allele frequencies of common SNPs from the 15 exomes is significantly correlated with those from genotype data of a larger cohort of 48 individuals (Pearson correlation coefficient, 0.91; p <2.2×10-16). A set of 2,485 SNPs show significantly different allele frequencies when compared to populations from other continents. Two notable variants having risk alleles in high frequencies in this subgroup are: a nonsynonymous deleterious SNP (rs2108622 [19:g.15990431C>T] from CYP4F2 gene [MIM:*604426]) associated with warfarin dosage levels [MIM:#122700] required to elicit normal anticoagulant response; and a 3' UTR SNP (rs6151429 [22:g.51063477T>C]) from ARSA gene [MIM:*607574]) associated with Metachromatic Leukodystrophy [MIM:#250100]. Hemoglobin Riyadh variant (identified for the first time in a Saudi Arabian woman) is observed in the exome data. The mitochondrial haplogroup profiles of the 15 individuals are consistent with the haplogroup diversity seen in Saudi Arabian natives, who are believed to have received substantial gene flow from Africa and eastern provenance. We present the first genome resource imperative for designing future genetic studies in Saudi Arabian tribe subgroup.

4. John2014. Study Accession = nstd107

Kuwaiti population subgroup of nomadic Bedouin ancestry-Whole genome sequence and analysis. John SE, Thareja G, Hebbar P, Behbehani K, Thanaraj TA, Alsmadi O. *Genom Data*. 2014 Dec 18;3:116-27.

Kuwaiti native population comprises three distinct genetic subgroups of Persian, "city-dwelling" Saudi Arabian tribe, and nomadic "tent-dwelling" Bedouin ancestry. Bedouin subgroup is characterized by presence of 17% African ancestry; it owes it origin to nomadic tribes of the deserts of Arabian Peninsula and North Africa. By sequencing whole genome of a Kuwaiti male from this subgroup at 41X coverage, we report 3,752,878 SNPs, 411,839 indels, and 8451 structural variations. Neighbor-joining tree, based on shared variant positions carrying disease-risk alleles between the Bedouin and other continental genomes, places Bedouin genome at

the nexus of African, Asian, and European genomes in concordance with geographical location of Kuwait and Peninsula.

5. Lou2015. Study Accession = nstd111

A 3.4-kb Copy-Number Deletion near EPAS1 Is Significantly Enriched in High-Altitude Tibetans but Absent from the Denisovan Sequence. Lou H, Lu Y, Lu D, Fu R, Wang X, Feng Q, Wu S, Yang Y, Li S, Kang L, Guan Y, Hoh BP, Chung YJ, Jin L, Su B, Xu S. *Am J Hum Genet*. 2015 Jul 2;97(1):54-66.

Tibetan high-altitude adaptation (HAA) has been studied extensively, and many candidate genes have been reported. Subsequent efforts targeting HAA functional variants, however, have not been that successful (e.g., no functional variant has been suggested for the top candidate HAA gene, EPAS1). With WinXPCNVer, a method developed in this study, we detected in microarray data a Tibetan-enriched deletion (TED) carried by 90% of Tibetans; 50% were homozygous for the deletion, whereas only 3% carried the TED and 0% carried the homozygous deletion in 2,792 worldwide samples ($p < 10(-15)$). We employed long PCR and Sanger sequencing technologies to determine the exact copy number and breakpoints of the TED in 70 additional Tibetan and 182 diverse samples. The TED had identical boundaries (chr2: 46,694,276-46,697,683; hg19) and was 80 kb downstream of EPAS1. Notably, the TED was in strong linkage disequilibrium (LD; $r(2) = 0.8$) with EPAS1 variants associated with reduced blood concentrations of hemoglobin. It was also in complete LD with the 5-SNP motif, which was suspected to be introgressed from Denisovans, but the deletion itself was absent from the Denisovan sequence. Correspondingly, we detected that footprints of positive selection for the TED occurred 12,803 (95% confidence interval = 12,075-14,725) years ago. We further whole-genome deep sequenced (>60×) seven Tibetans and verified the TED but failed to identify any other copy-number variations with comparable patterns, giving this TED top priority for further study. We speculate that the specific patterns of the TED resulted from its own functionality in HAA of Tibetans or LD with a functional variant of EPAS1.

Summary

If you have any questions or comments, please feel free to contact us by email at dgv-contact@sickkids.ca

Sincerely,

The DGV team