BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 11 2013, pages 1481–1483
doi:10.1093/bioinformatics/btt173

Databases and ontologies

Advance Access publication April 26, 2013

KGVDB: a population-based genomic map of CNVs tagged by SNPs in Koreans

Sanghoon Moon[1,†], Kwang Su Jung[2,†], Young Jin Kim[1], Mi Yeong Hwang[1], Kyungsook Han[3], Jong-Young Lee[1], Kiejung Park[2,4] and Bong-Jo Kim[1,*]

[1]Division of Structural and Functional Genomics, [2]Division of Bio-Medical informatics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, 363-951, Korea, [3]School of Computer Science and Engineering, Inha University, Inchon, 402-751, Korea and [4]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Despite a growing interest in a correlation between copy number variations (CNVs) and flanking single nucleotide polymorphisms, few databases provide such information. In particular, most information on CNV available so far was obtained in Caucasian and Yoruba populations, and little is known about CNV in Asian populations. This article presents a database that provides CNV regions tagged by single nucleotide polymorphisms in about 4700 Koreans, which were detected under strict quality control, manually curated and experimentally validated.

Availability: KGVDB is freely available for non-commercial use at http://biomi.cdc.go.kr/KGVDB.

Contact: kbj6181@cdc.go.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 20, 2012; revised on December 31, 2012; accepted on January 23, 2013

1 INTRODUCTION

To date, single-marker association analysis in genome-wide association studies (GWAS) has identified a large number of single nucleotide polymorphisms (SNPs) that are highly associated with complex diseases, but only a small portion of genetic heritability is explained by these variants. A copy number variation (CNV) is a physical change of genomic segment ranging from a kilobase to several megabases. CNV may influence disease susceptibility and has been proposed to be one potential source of missing heritability (Jarick *et al.*, 2011). A recent study also found a strong functional relevance of CNV with complex diseases (Gamazon *et al.*, 2011). However, the extent to which CNV impacts disease susceptibility and underlies complex traits has yet to be fully determined. The CNV association study conducted by the Wellcome Trust Case Control Consortium (WTCCC) reported that common CNVs typed on existing platforms are unlikely to contribute to the genetic basis of common diseases (Wellcome Trust Case Control Consortium, 2010). However, only about 40% of the identified CNVs (3432 multi-class CNVs) were well separated enough to be genotyped. In contrast, 60% of the CNVs could not be tested for association (Supplementary Fig. S1). Moreover, well-defined polymorphic CNVs tagged by SNPs are more likely to affect multiple expression traits than frequency-matched variants (Gamazon *et al.*, 2011). CNVs encompassing single genes or a set of genes may be more likely to be the causative variants of a given genetic disease than tagged SNPs. Therefore, SNPs correlated with CNVs are a valuable resource for GWAS.

Most CNV databases do not consider multi-copy number classes (Gamazon *et al.*, 2010; Iafrate *et al.*, 2004; Shaikh *et al.*, 2009). The SCAN database is an exception in that it includes the latter, but it only contains data from Caucasian and Yoruban populations, and Asian populations are completely absent. Owing to the difference in CNVs between distinct ethnic groups, providing polymorphic CNVs and allele frequency of each genotype in Asian populations will help investigate CNV association with diseases and ethnic differences.

Recently, we developed a database called Korean Genomic Variant Database (KGVDB), which provides multi-class CNV regions and well-tagged SNP information. The data were obtained from 4694 individuals using two different genotyping platforms and publicly available CNV data. The large dataset of KGVDB will provide a rich public resource for the study of CNV and SNP.

2 CONTENTS AND FEATURES

2.1 Resources

Data on CNV regions and breakpoints were constructed using two types of resources: data from a large-scale Korean CNV study including SNP information from GWAS (Cho *et al.*, 2009) and publicly available CNV data (Supplementary Table S1). To define exact polymorphic CNV regions in the large-scale Korean CNV study, we used two different genotyping platforms (Fig. 1):

- A total of 4694 Korean individuals that were genotyped on both the NimbleGen HD2 3 × 720K aCGH assay with the HapMap sample (NA10851) as a reference and the Affymetrix genome-wide human SNP array 5.0 (Supplementary Fig. S2)

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

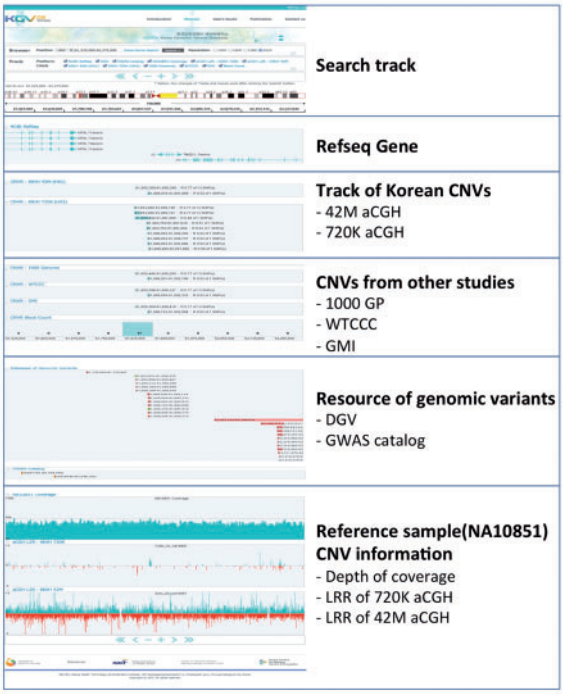

Fig. 1. A screenshot of the KGVDB browser. It provides CNV data from the Korean CNV study and other reliable CNV studies. The last three tracks show the depth of coverage and log2 ratio of the reference sample

- A total of 48 Korean individuals that were genotyped on the NimbleGen HD2 42M aCGH

For CNV discovery, the Genome Alteration Detection Analysis software was used to ascertain CNV regions from the two platforms (Pique-Regi *et al*., 2008). To assign individuals to copy number classes (CNV genotyping), all the detected CNV regions were examined by the CNVtools software (Barnes *et al*., 2008). Among these regions, only multi-copy number classes were selected. Moreover, all the results of CNVtools were manually curated by visual inspection to discriminate false positive CNV calls. Finally, 3601 multi-class CNVs and tagging SNPs were defined (see Supplementary Fig. S3).

Three public resources of CNVs and tagging SNPs were used to build KGVDB (see Supplementary Fig. S4 and Supplementary Table S1). Because CNV tagging SNPs of SCAN database are based on WTCCC CNV results, CNVs information of WTCCC study will help compare with our CNVs.

- A total of 20 Yoruba and 20 Caucasian individuals from the WTCCC CNV study (Conrad *et al*., 2010)
- A total of 30 Asian individuals including 10 Koreans from the Genomic Medicine Institute CNV study (Park *et al*., 2010)
- 1000 genomes project deletion regions (Mills *et al*., 2011)

To find well-tagging CNVs with SNP, we calculated the squared Pearson's *r* value between all CNVs (i.e. curated polymorphic CNVs from large-scale Korean study and CNVs from public online resources) and SNPs from the Affymetrix 5.0 array (SNP call rate >0.98) (Cho *et al*., 2009; see Supplementary Fig. S4). We considered all SNPs within 1 Mb of the estimated 2 CNV breakpoints (i.e. start and end points). KGVDB allows users to get all of the tagging SNP information clicking on each CNV region (Supplementary Fig. S5 and Supplementary Fig. S6). Moreover, frequencies of the HapMap population and those of Korean population corresponding tagging SNPs were also provided.

Even small differences of quality in the underlying CNV measurements could lead to an artifactual shift in the copy number distribution (Aldhous *et al*., 2010). Especially, it is important to eliminate incorrect CNVs, which were affected by reference-induced CNV calling bias. To provide copy number state of the reference, KGVDB provides $\log_2$ ratio value of the NA10851 using two platforms (NimbleGen 42M and 720K aCGH) with 48 Korean pooled samples as a reference as well as the depth-of-coverage of the NA10851 from the whole-genome sequencing study (Park *et al*., 2010). This information enables users to consider copy number state of the reference sample at the corresponding coordinate to a CNV region. KGVDB also allows users to get information from GWAS catalog and Database of Genomic Variants (Hindorff *et al*., 2009; Iafrate *et al*., 2004).

2.2 Web interface and example

KGVDB has been implemented using MySQL database with Java Server Page. Users can access KGVDB in any web browser with simple queries such as coordinate of genomic site and gene name.

Example: In a previous study by Gamazon *et al*. (2011), overlap between trait-associated SNPs and its tagging CNVs of WTCCC study has been observed. We compared the squared Pearson's *r* value of tagging SNP of WTCCC study with those of our data (see Supplementary Table S2 and Supplementary Fig. S3). Most squared Pearson's *r* values of tagging SNPs agree with ours. However, in the case of rs12191877 tagging the CNVR2841.6 (chr6: 31384505-31397416), which is associated with psoriasis and AIDS progression, the squared Pearson's *r* value of CEU/YRI is 0.90, whereas those of Koreans is 0.51 (Supplementary Table S2), suggesting that this discrepancy may reflect ethnic differences.

3 CONCLUSION

We constructed a database called KGVDB, which provides polymorphic CNVs tagged with SNPs. The major features of KGVDB that are different from others include the following: (i) polymorphic CNV regions identified under strict quality controls and manual curation; (ii) CNV information from Korean populations to supplement currently biased ethnic information; (iii) large dataset of CNVs tagged with SNPs from 4694 individuals using two different genotyping platforms (SNP array and CGH array); (iv) rich information on tagging SNPs, including frequencies in HapMap populations; and (v) copy number states of the reference sample using $\log_2$ ratios from two kinds of CGH data and the depth of coverage from whole-genome sequencing data.

In conclusion, KGVDB is a rich resource of the genomic variants, which will complement the lack of Asian CNV data. In particular, correlation data like rs12191877 tagging the

CNVR2841.6 will help understand ethnicity-specific genetic changes.

Funding: This work was supported by an intramural grant from the Korean National Institute of Health (2012-N73004-00, 2011-N72001-00) and grants from the Korean Centers for Disease Control and Prevention (4845-301, 4851-302, 4851-307).

Conflict of Interest: none declared.

REFERENCES

Aldhous,M. *et al.* (2010) Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. *Hum. Mol. Genet.*, **19**, 4930–4938.

Barnes,C. *et al.* (2008) A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.*, **40**, 1245–1252.

Cho,Y.S. *et al.* (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.*, **41**, 527–534.

Conrad,D.F. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Gamazon,E.R. *et al.* (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.

Gamazon,E.R. *et al.* (2011) A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet.*, **7**, e1001292.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Jarick,I. *et al.* (2011) Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum. Mol. Genet.*, **20**, 840–852.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Park,H. *et al.* (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.*, **42**, 400–405.

Pique-Regi,R. *et al.* (2008) Sparse representation and Bayesian detection of the genome copy number alterations from microarray data. *Bioinformatics*, **24**, 309–318.

Shaikh,T.H. *et al.* (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.*, **19**, 1682–1690.

Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.