A web-server framework to explore and visualize large genomic variation data in lab and its applications to wheat and its progenitors

Wenxi Wang†, Zihao Wang†, Xintong Li, Zhongfu Ni, Zhaorong Hu, Mingming Xin, Huiru Peng, Yingyin Yao, Qixin Sun and Weilong Guo*

Key Laboratory of Crop Heterosis and Utilization, State Key Laboratory for Agrobiotechnology, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing 100193, China.

*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 Auhtors.

Contact: guoweilong@cau.edu.cn

Abstract

Decreased sequencing costs now allow a laboratory to investigate genomic varia-tions in hundreds or thousands of samples by resequencing or genotype-by-sequencing (GBS). Managing and exploring the large genomic variation data require programming skills. The public genotype querying databases for many species may not be frequently updated and are still lim-ited in samples. Many individual samples have unique genomic variations in either cancer or crop studies.

Here we present SnpHub, a Shiny/R-based server framework for retrieving, analyzing and visualizing the large genomic variation data within a lab. After the pre-building process based on the provided VCF files and genome annotations, the local server allows users to instantly access the SNP/INDELs and annotation information by locus or gene, and for user-defined sam-ple sets, without any programming background. The users can also easily analysis and visualize the genomics variations in heatmap, phylogenetic tree, haplotype-network, or geographically, and get genomic sequences replaced

by sample-specific SNPs and INDELs. SnpHub can be applied to any species, and here we provide demo-servers for wheat progenitors using the public GBS data.

SnpHub and its tutorial are available as http://guoweilong.github.io/SnpHub/.

Introduction

High-throughput sequencing competition largely bring benefits for the reduction of sequencing cost. Nowadays one thousand dollars can be used to sequence about 5 human genomes, 1 hexaploid wheat genome, 6 maize genomes or 50 rice genomes at 10X. Whole genome sequencing is commonly used for mid-size genome species such as human and maize. Genotyping-by-Sequencing (GBS) or exon-capture technologies are also frequently used for large genome species such as wheat(Chapman et al., 2015). To investigate the genetic diversity among individuals, one laboratory is now able to sequencing individual genomes from dozens to hundreds, and even in thousands. The raw data mapping and SNP/INDEL calling to generate VCF files are becoming routine bioinformatics analysis steps. However, with the fast accumulation of genomic sequencing data, managing and analyzing the large VCF files requires computational skills. Unfortunately, most researchers lack lacking programming skill and have difficulty in exploring the large genomic variation data.

Some public databases are available for querying sample-specific genomic variations, such as cBioPortal (Cerami et al., 2012) for cancer study, and IC4R database (IC4R Project Consortium et al., 2016) for rice study. While these databases are implemented independently, and cannot meet the need for accessing new data locally in a lab. Also, samples such as the TILLING

population in crop study usually carry special mutations which are rare in nature. Here we developed SnpHub as a framework, can be easily applied for building an interactive local server, so as to help researchers to navigate and explore the genomic diversity in their own lab.

Implementations

SnpHub is designed to be installed in Linux system, utilizing Shiny/R and integrating several bioinformatics softwares. To build a local instant, the variant call files (VCF format), reference genome sequence file (FASTA format), gene annotation file (gff3) and metadata files defining sample information are needed. A shell wrapper program is provided for the pre-building process. Once an instance-server is built, users can access the data through webpage directly (Figure 1A).

SnpHub provides user-friendly functions for navigating the genomic variation data (Figure 1B). (1) One can search and download the SNP/INDEL tables for specific region or gene in a group of samples. Minimum allele frequency and missing data frequency thresholds are provided for filtering sites. (2) The SNPs within a specific region can be visualized by different type in heatmap, or be explored in haplotype in network. (3) Users can also explore the genetic distances among samples for a specific region as phylogenetic trees or PCoA analysis. (4) The SNP/INDELs can be visualized with functions and frequency in lollipop graph with transcript-tracks; (5) The genotypes can be visualized geographically in map for exploring the distributions of different haplotypes. (6) Moreover, users can also retrieve a sample-specific sequence by replacing the genomic variation according to the reference sequence, which will be useful for works such as primer-design. The tables can be

downloaded as CSV files, and figures can be downloaded in either PDF or PNG formats.

Applications

As an important staple crop, bread wheat is an allohexaploid, with genome size of ~16Gbp. Previous study published 62 lines of bread wheat (BBAADD) using WEC and GBS methods (Jordan et al., 2015). The population genomics data of wheat progenitors were also available, including 65 wild and domesticated emmers (BBAA) (Avni et al., 2017) and 549 accessions of Aegilops tauschii (DD) (Singh et al., 2019), are publicly available. We these data, generated VCF files, and setup three demonstration servers for wheats, emmers and aegilops using SnpHub, which can be accessed at http://wheat.cau.edu.cn/snphub_demos/.

With the decreased sequencing cost, even more samples and species will be sequenced, and universal database will not satisfy many needs. SnpHub can be applied to any species with genome assembly and annotations, and be instantly setup based on variation calls. SnpHub can serve as lab-level servers for navigating and visualizing the genomic diversity or individual line or lineage, and will be useful for different occasions. For example, investigators can infer trait-associated genes with population structure information and variation function annotations from specific sample set; breeders can access their genetic diversity for specific locus for designing the breeds. Moreover, SnpHub provides uniform server framework for navigating genomic variations, and the future population genetic study can also easily setup a SnpHub-based querying sever, and publish it beside the raw data, making the data to be more easily accessed by the community.

Acknowledgements

We thank Xiaoming Xie, Yongming Chen, Zhengzhao Yang for exploring technology, and thank Kuohai Yu for IT support.

Funding

This work has been supported by National Key Research and Development Program of China [grant number 2016YFD0100801 and 2018YFD0100803].

Conflict of Interest: none declared.

References

Avni,R. et al. (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science, 357, 93–97.

Cerami,E. et al. (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. Cancer Discovery.

Chapman,J.A. et al. (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biology, 16, 26.

IC4R Project Consortium et al. (2016) Information Commons for Rice (IC4R). Nucleic Acids Research, 44, D1172–D1180.

Jordan,K.W. et al. (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biology, 16, 48.

Singh,N. et al. (2019) Genomic Analysis Confirms Population Structure and Identifies Inter-Lineage Hybrids in Aegilops tauschii. Frontiers in Plant Science, 10.

Figures
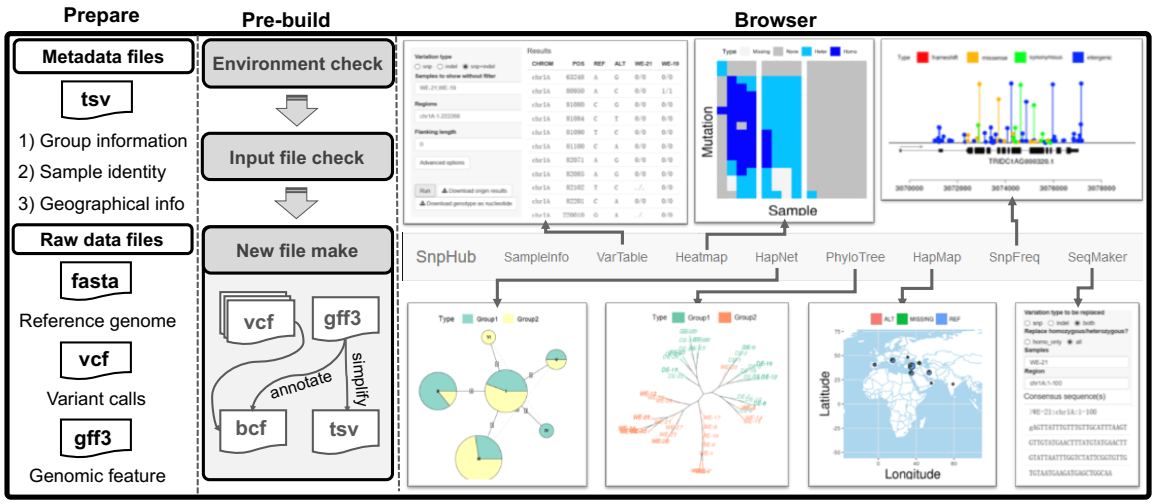
Figure 1 - The design, building step and functions of SnpHub server.