SEQUENCE DOWNLOAD

What exactly am I downloading when I use Sequence Download Feature?

The FASTA formatted file you get from this tool is what may be called a "pseudo-sequence"; it is a result of *in silico* replacing reference bases at variant positions with the variant bases found in VCF. It is not a biologically validated sequence, and in many cases it won't be faithful to the actual DNA sequence. One reason for that is that the SNPs present in the VCFs do not exhaust all sequence variants - the indels and larger structural variants are all missing from this.If your region under study is longer than, say 150 nt, there is a very high chance that in the 3000 samples there are variants in this region missing from VCFs and unaccounted for in this tool.Thus, proceed with caution, and do not expect this output to represent actual DNA data of varieties. The output may be acceptable for some population-genetic analyses that are tolerant to some degree of omission, but it is not guaranteed to work for analyses requiring precise sequence such as primer design.

Why did my dataset go missing?

We used to provide SNP data for 4 lower-quality references (93-11, Kasalath, IR 64, DJ123).
One problem with that data was that the SNPs were not exhaustive, but incremental relative to other dataset. So that when you need to find all SNPs for a region in 93-11, you would first need to get all SNPs from all aligning subregions in Nipponbare, then get the 93-11 SNPs. We realize this is suboptimal, and adds extra burden on researchers.

As new higher quality references become available, we are preparing new SNP datasets for 3,010 genomes aligned to those references. These will replace the data that was removed. Stay tuned.

If you really need SNP data for the previous references, the original single-sample VCF files are available on Amazon Web Services public 3k dataset.

Why can't I download all varieties? Should I submit multiple download queries?

The tool has a limit of N varieties, and the maximal size of the region is 50kb. If you need more, please download from AWS directly and use GATK tool.
We limited the tool, as it is taxing our cloud resources.

Please also refrain from submitting many queries at once. There is a storage limit on the AWS instance running the tool, and in case it gets exhausted, all ongoing jobs will be cancelled.

HAPLOTYPE ANALYSIS

Why do group alleles in haplotype view keep changing?

This behaviour had to do with the internal stochasticity of the K-means algorithm, which is employed when deciding on the "best" number of clusters. As such, changing results indicate there is no clear-cut "best" number of clusters.

This behaviour will be changed in future, but please read the answer to the next Q.

What does the output of this tool represent, actually?

The main goal of the haplotype viewer tool is to help visualize the structure and relationship between haplotypes at a locus.

What it does:

1. Cluster haplotypes in the region into "haplotype groups"
2. Visualize the genotypes, showing all genotypes from the same group together.
   - It does so in a way that can account for "whole-genome" relatedness that is based on subpopulation classification by admixture.
3. Output total number of varieties and that of each population, for each haplotype group cluster
4. Output some intermediate results useful for QC, such as ...

What it does not

1. Report all unique haplotypes in the region. (This task is easier, and may be done using the summary table itself)

REPORTING BUGS AND COMMUNICATION WITH SNP-SEEK TEAM

When will my request be looked into?

We assure users that we take note every request given by users.

How do I report a bug or request a new feature?

Use the bug report form at Help - > Report bug menu.

OTHER QUESTIONS

Are assembled sequences from the 3K RG dataset available in SNP-seek?

Assembled sequences are not in SNP-Seek, but check out this link for our collaborator's website, the 3,000 Rice Pan-Genome browser http://cgm.sjtu.edu.cn/3kricedb/

How to use the GWAS SNP dataset in TASSEL?

The PLINK files available form Downloads page are in "binary PLINK" format (bed, bim, fam).To use with TASSEL, one needs a "text PLINK" format (ped and map files). Thus, to use our files with TASSEL, one needs to convert binary plink to text plink. Here is how to do it on Windows. (If you use Mac or Linux, it is much easier and follows the same lines)

1. Download PLINK 1.90 from https://www.cog-genomics.org/plink2/ (choose the file suitable for your operating system).
2. Unpack the archive.
3. Create a folder on your desktop, named plinkdata
4. Put plink.exe (or plink file) AND the dataset you want to convert into this folder. I will use the example of the 1M GWAS dataset which contains 3 files (pruned_v2.1.bed, pruned_v2.1.bim, pruned_v2.1.fam).
5. Open PowerShell . Type

   $ cd Desktop

   $ cd plinkdata

(press Enter after typing to run each command)
You should be inside the PLINK3K folder. Type:
$ ls
and press Enter to check that files are there.

Type
$.\plink --bfile pruned_v2.1 --recode tab --out 1Mdata
and press Enter.
- This will create files **1Mdata.ped** and **1Mdata.map** that can be used to be input in TASSEL via **"Data -> Load PLINK"**