REV GENOMICS

www.revgenomics.com

Welcome to version 1.2.0 of the free Cannabis genetic marker dataset.

This genetic marker dataset was derived by analyzing publically available Cannabis sequence data available for free here. The sequencing data was mapped to the publically available Purple Kush genome resource available here published as part of Laverty et al. 2019. We provide these 23500+ SNP and 2200+ InDel molecular markers from 1358 cultivars requested by the cannabis community to facilitate innovation in breeding efforts and the creation of new cultivars in this amazing plant.

The combined file contains SNPs and InDels that were filtered to have at least 40 sequencing reads and a Phred Score of 40 (base call accuracy of 99.99%) for each cultivar sample relative to the reference genome. The dataset is inclusive for all SNPs and InDels that were called relative to the reference genome. Each molecular marker may not have been called in every individual due to missing data (see technical notes at the end this README for explanation of dataset limitations).

Files contained with data release:
Figure files 1 through 5.
SampleCrossReference- Summary file with cultivar names and corresponding SRR genotype file name
SRR...vcf.gz - the 1358 variant files for each sample
SRR...vcf.gz.tbi - the 1358 index files to view the file in the genome browser in tutorial
Cannabis-SNPs-total-merged-filtered-stringent.vcf - All 1358 samples merged into single file
GCA_000230575.4_ASM23057v4_genomic.fna - genome sequence file
GCA_000230575.4_ASM23057v4_genomic.fai - genome index file

Basic Patterns
The pattern of SNPs and InDels across the Cannabis genome gives each cultivar a unique "fingerprint". While it is not particularly helpful to know a single cultivar's fingerprint (unless for cultivar tracking or specific marker breeding, see below), the aggregated data is quite powerful. For example, aggregated data allows comparison of cultivars based on genetic sequences (see zoomed in Figure 1) and to come up with genetic distance scores. The more closely related a cultivar is to another cultivar the closer they will be on a phylogenetic distance tree like Figure 1. The figure contains the cultivar name associated with the sample and the raw data name if you want to look up a specific cultivars markers (see tutorial below).

If cultivar names and genetic identity had a perfect 1:1 relationship, then all of the cultivars named something would cluster next to one another on this phylogenetic tree (see this excellent explanation of how to read and interpret phylogenetic trees). The fact that they do not imply that the naming conventions currently used in Cannabis are not as helpful as they could be. For example, the cultivar name "Jack Herer" shows up multiple times in the dataset across all of the major axis of genetic variation. Although there are also many examples of local clusters of cultivar names that show up next to each other (see highlighted examples in Figure 1). Names are not useless, but not consistent in the information they convey about the underlying genetics. The main conclusion from this for breeding programs (and end-consumers alike) is that healthy skepticism should be applied to cultivar names and the actual traits the plants produce are much more important. If you open up Figure 1 and zoom in, what patterns do you see?

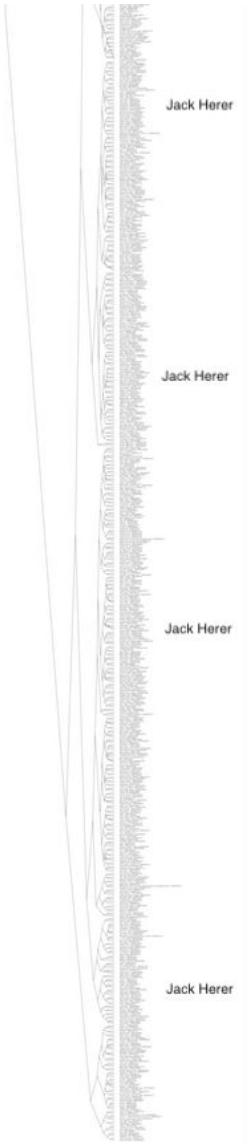
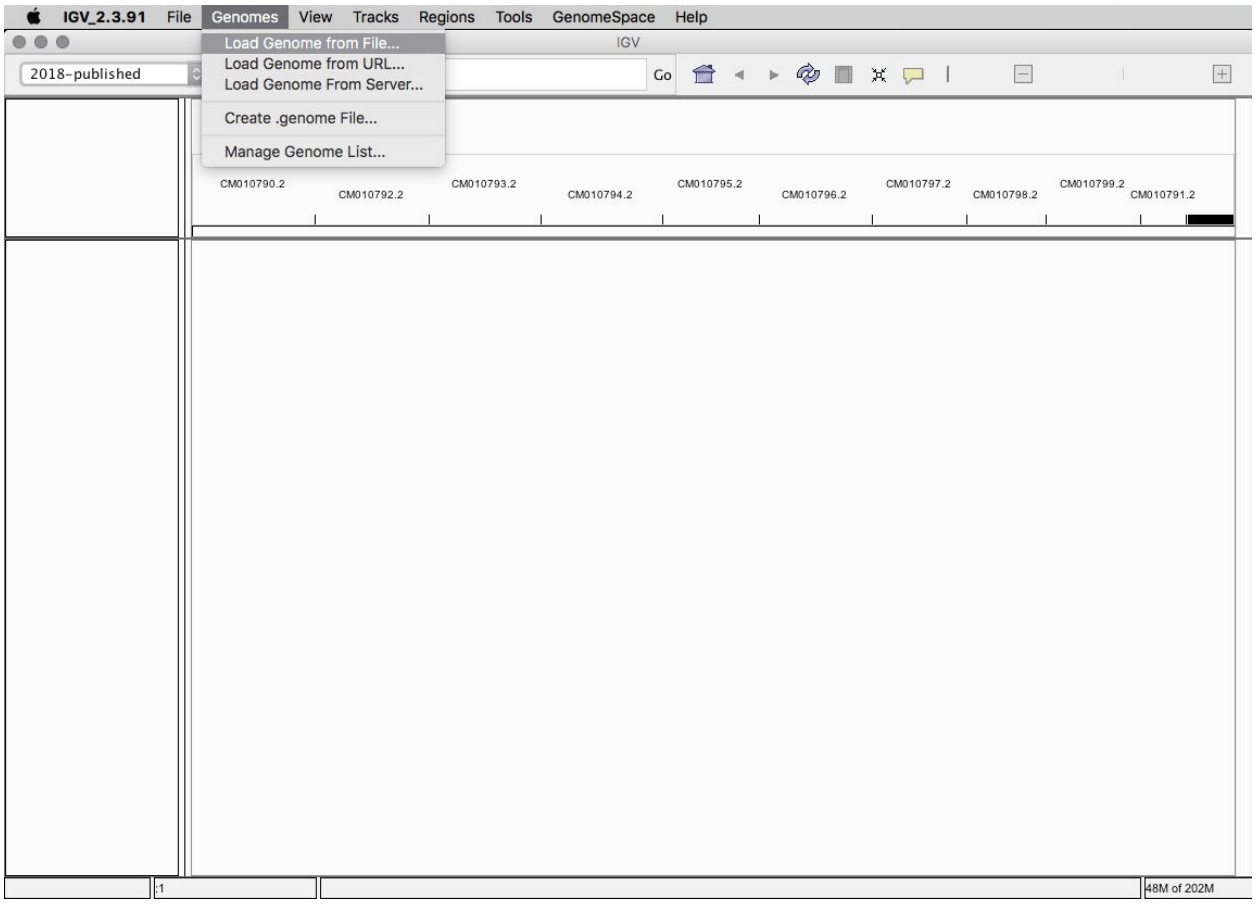


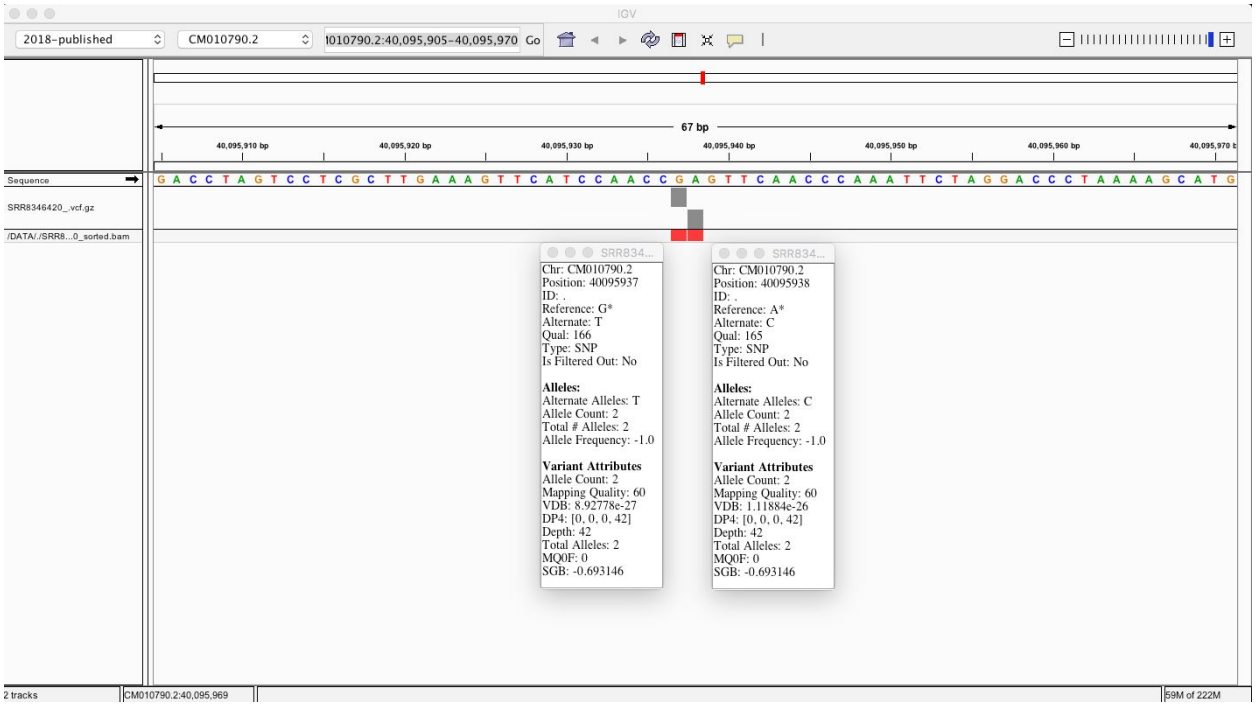
Figure 1 zoomed in.

Now let's view some of this data so it makes more sense.

1) Download the Cannabis genetic data.
2) Download and follow the installation instructions of the open-source genome viewer IGV. Alternatively, if you have the data in a dropbox or google drive folder of your own, you can use IGV in a web browser following instructions here. Once installed launch the IGV browser.
3) In IGV select the Genomes bar and Select Load Genome from file.
4) Navigate to where you saved the genetic data package on your computer
5) Select the file: "2018-published.genome". It will load the genome and index.

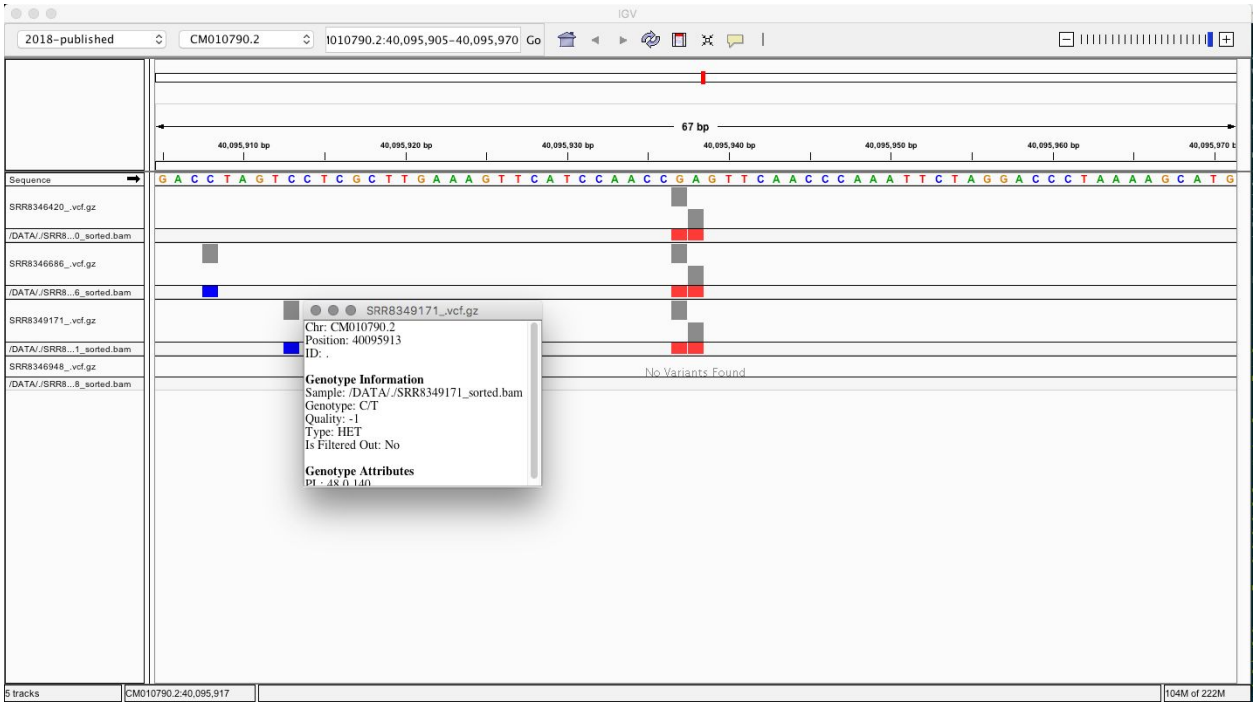


6) In the white search bar at the top of IGV copy and paste CM010790.2:40,095,905-40,095,970
7) This will zoom into a single assembled chromosome region between 40,095,905-40,095,970 DNA base pairs.
8) These basepairs will be displayed as the sequence of A,T,C,G of this section of the chromosome. A's will be green, G's will be orange, T's will be red, and C's will be blue.
9) Next load a few samples that are named "Jack Herer".
10) Click on File --> then Load from file.
11) Navigate to the folder where the data is stored and select SRR8346420_vcf.gz, this will load the variant file containing SNPs and InDels for this sample.



12) In this region of the genome, there are 2 SNPs side by side. Instead of G and A in the reference genome, the SNPs are both red T's in this sample.
13) Now load three other "Jack Herer" samples. SRR8346686_vcf.gz, SRR8349171_vcf.gz, and SRR8346948_vcf.gz
14) Your IGV should now look like below.

15) The first three samples displayed share the same T, T SNPs compared to the reference. These samples are also homozygous (2 copies) for these SNPs.
16) However, the last sample (SRR8346948_vcf.gz) has no variation in this region of the genome compared to the reference.
17) Also noticeable is that samples SRR8346686_vcf.gz and SRR8349171_vcf.gz both have additional SNPs in this region that they are heterozygous (below).



18) Heterozygous would mean that the parents of these cultivars are different in this part of the genome from one another.
19) This variation in a population would allow molecular markers for this region of the genome to be developed and individuals to be selected based on their genotype at this location. Selectable markers like this are used in a marker assisted breeding program.

There is a great deal of data to explore and literally tens of thousands of selectable molecular markers that could be developed from this dataset. Enjoy!

Technical Notes and Data Limitations:
A comprehensive review of all the technical limitations of this data and everything this data can be used for are outside the scope of this release.

Missing data is common with the amplicon based sequencing approach used for creating the raw data for this dataset. Missing data is so common in large scale genomic experiments that an entire subfield statistical genetics are devoted to dealing with missing data (Li et al. 2009, Shi et al. 2018). Differentiating samples with missing data are also generally not a problem and has been demonstrated using simulations (Huang and Knowles 2016) and empirical data (Tripp et al. 2017). However, determining large scale population genetic parameter estimation (Arnold et al. 2013) and branch length support from phylogenetic trees (Roure et al. 2013, Leaché et al. 2015) can be harder with missing data and potentially introduce artifacts.

Additional Resources:
www.kannapedia.net - opensource .vcf files from samples mapped to 2011 (CanSat3, VanBakel et al. 2011) version of cannabis genome. Note that the CanSat3 version of the genome does not include the ten assembled chromosomes making it less useful than the newest version used in this data release for molecular breeding.

Another recent genome assembly and genetic map by Grassa et al. 2019 are available in pre-print form. The raw data are available here: http://cannabisgenome.org/

References:

Arnold, B., Corbett-Detig, R.B., Hartl, D., and Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. *22*, 3179–3190.

Grassa, C.J., Wenger, J.P., Dabney, C., Poplawski, S.G., Timothy Motley, S., Michael, T.P., Schwartz, C.J., and Weiblen, G.D. (2018). A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. BioRxiv 458083.

Huang, H., and Knowles, L.L. (2016). Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. Syst. Biol. *65*, 357–365.

Laverty, K.U., Stout, J.M., Sullivan, M.J., Shah, H., Gill, N., Holbrook, L., Deikus, G., Sebra, R., Hughes, T.R., Page, J.E., et al. (2019). A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci. Genome Res. *29*, 146–156.

Leaché, A.D., Banbury, B.L., Felsenstein, J., de Oca, A.N.-M., and Stamatakis, A. (2015). Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. Syst. Biol. *64*, 1032–1047.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. Annu Rev Genomics Hum Genet *10*, 387–406.

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., and Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. HHE *83*, 107–116.

Tripp, E.A., Tsai, Y.-H.E., Zhuang, Y., and Dexter, K.G. (2017). RADseq dataset with 90% missing data fully resolves recent radiation of Petalidium (Acanthaceae) in the ultra-arid deserts of Namibia. Ecol Evol *7*, 7920–7936.