VCF Miner: A Platform and Tools for Collaborative Information Mining from Next-Generation Sequence Variant Data

Zhi-Liang Hu, James E. Koltes, Eric Fritz-Waters and James M. Reecy

Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, 2255 Kildee Hall, Ames, IA 50011

Abstract

High-throughput next-generation sequencing (NGS) technology has provided powerful tools for genetics/genomics studies. It has also imposed a great challenge on researchers to determine how to efficiently handle and quickly glean information for genotype/phenotype studies. The ability to quickly process NGS data for useful information screening is important before the extensive analysis may be carried out, because the latter is usually time consuming and demands large computation power. The variant call format (VCF) that stores variations against a reference genome, is one of such data formats at the heart of information crossroads of the genotype/phenotype data analysis pipelines. We propose a shared platform for collaborative VCF file storage, handling, information abstraction, querying, and data re-use. A unix-file-system-based data file repository has been built, made available Plink/SEQ, BEDOPS, and BEDtools for VCF file handling as well as a number of in-house scripts to serve various purposes during the data process. An initial collection of 41 VCF files from four laboratories were used as a test case. The goals of this project were to allow VCF data from multiple collaborative projects to be managed in a central location to facilitate pre-computed data abstractions and query. This resource will allow data re-use in subsequent, possibly more diverse types of data analysis in an efficient manner.

Introduction

The amount of new NGS data flowing to genetics lab is phenomenal. While researchers enjoy the increasing amount of genotype data to use, the speed and capacity to handle and possibility to re-use these data are challenging. The purpose of this project was to help researchers to meet these challenges. At this time we focused on pre-abstracting information from VCF files in order to save time for researchers.

Preliminary Results

We have built a UNIX file-system based VCF file repository. The repository is equipped with existing and custom software for information abstraction, and enabled MySQL database for information handling. We have also developed preliminary web interfaces for end users. Figures 1 and 2 show introductory information on how the web site and its front end work. The main idea behind this scheme is to pre-compute some basic but much needed statistics for users to get a summary of each VCF file to guide further analysis.

There are already a few developing as well as mature software available that handle VCF files, such as vcftools, BEDOPS, and PLINK/Seq (1, 2, and 3). They are diverse in terms of functions yet practical to use (Tables 1 and 2). We have adopted these software locally in our pipeline. In Box 1 is shown a conceptual work flow in terms of data processes and jobs done at certain steps.

We were fortunate to have received 41 sample VCF files from five laboratories (Figure 1) for initial test of concept. In Tables 3, 4, and Box 2, are shown some pre-computed statistics, including number of genotypes with a minor allele, homozygotes, heterozygotes, genotyping rate, etc. It's worth pointing out that this is only our start of this work, and the preliminary data shown here is only a very small example of potential features we may facilitate with the repository when it's fully developed.

http://www.animalgenome.org/repository/vcf

Figure 1.

The front page of the VCF file Data Repository showing: (1) User login is required for data contributions and data meta-information management (refer to Figure 2); (2) Automated data repository overview with some basic statistics; (3) Main features offered by the platform.

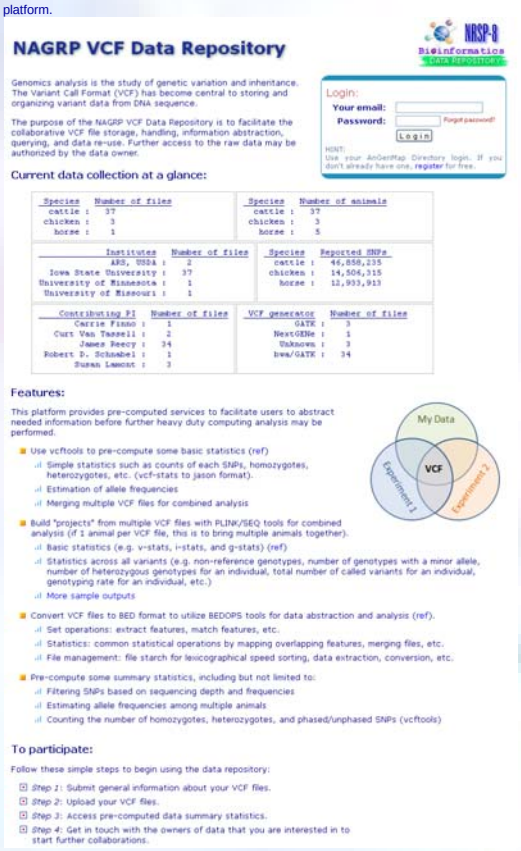
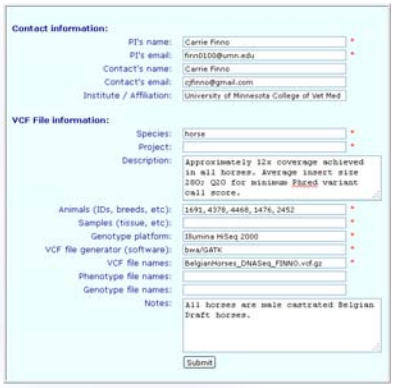


Figure 2.

A sample VCF file meta-information entry form shows the kind of information needed in order for the Repository to properly handle and process the data file. Note the VCF file contributor has access to update these information once logged in.



Discussions

The VCF Miner will have several potentials to facilitate genetics research. For example, to make breed and species level variant comparisons; to identify novel variants by cross breeds/species comparisons, to visualize variants in the context of QTL and other genome features, etc. Our long term goal is to build tools making the VCF information more readily accessible to bring together useful genotype, phenotype and annotation data for queries and comparisons, thus increasing the re-usability of data and chances for new discoveries.

We will look out and work with other public platforms on using and sharing VCF information that may complement our efforts. For example, we will check with NCBI Variation Database and Ensembl Variation Database for functions they make available and in ways how we may complement their works as well.

Table 1. By adopting BEDOPS tools in the VCF Data Repository, a number of functions will become available.

Set operations

- bedops - apply set operations on multiple BED inputs
- bedextract - efficiently extract BED features
- closest-features - matches nearest features between BED files

Statistics

- bedmap - map overlapping BED elements onto target regions and optionally compute any number of common statistical operations

File manegements

- sort-bed - apply lexicographical sort to BED data
- starch and unstarch - compress and extract BED data
- starchcat - merge compressed archives
- Conversion tools - convert common genomic formats to BED

Source: http://bedops.readthedocs.org

Table 4. Allele frequency estimates by vcftools, showing SNP types substitutions, insertions/deletions (number of animals=5).

CHROM	POS	N_ALLELES	N_CHR	{ALLELE:FREQ}	
chr1	860	2	10	C:0.6	G:0.4
chr1	942	2	10	T:0.2	C:0.8
chr1	1237	2	10	T:0.2	TTTTG:0.8
chr1	1495	2	10	A:0.2	AAGAT:0.8
chr1	1541	2	10	C:0.9	T:0.1
chr1	1939	2	10	C:0.6	G:0.4
chr1	2123	2	10	G:0.9	GT:0.1
chr1	2296	2	10	A:0.9	C:0.1
chr1	3348	2	10	G:0	GT:1
chr1	3688	2	10	C:0.9	T:0.1
chr1	4247	2	10	T:0.2	C:0.8
chr1	4431	2	10	TTTTA:0.6	T:0.4
chr1	4625	2	10	A:0.9	C:0.1

Table 3. Sample output from "pseq" tool to show combined VCF statistics. *(a)* "`i-stats`" on combined VCF file (called "projects" in PSEQ's realm); *(b)* "`i-stats`" on single VCF file with multiple animals.

	Animal ID	NALT	NMIN	NHET	NVAR	RATE	SING	TITV	PASS	PASS_S	QUAL	DP
(a)	1187	4431574	2291355	2290436	4.43E+06	0.191514	2049319	1.67887	1117628	957713	NA	8.87623
	9111	5968417	3662079	3661153	5.97E+06	0.257932	3171831	1.11357	1300156	1068379	NA	13.4638
	9126i	5028157	2967521	2968005	5.03E+06	0.217335	2287842	1.75319	1926875	1569325	NA	21.582
	988a	4920388	2967243	2967786	4.92E+06	0.212677	2526868	1.05168	351893	180693	NA	13.4688
	989	5014611	2929186	2929749	5.02E+06	0.216752	2493733	1.09869	452307	253741	NA	13.9661
	990	4978381	2878071	2878610	4.98E+06	0.215179	2337708	1.03125	413533	189694	NA	14.6209
	991	5244265	2992039	2992639	5.25E+06	0.226675	2512079	0.968322	450253	221759	NA	14.0629
	992	5151410	3081217	3081870	5.15E+06	0.222664	2656056	0.83554	389198	204929	NA	12.8505

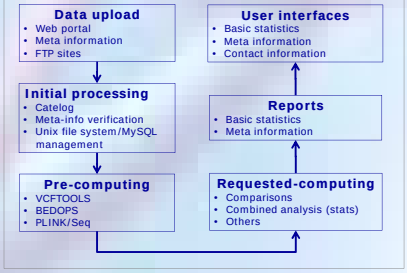
	Animal ID	NALT	NMIN	NHET	NVAR	RATE	SING	TITV	PASS	PASS_S	QUAL	DP
(b)	A1476	7505156	4893843	4228699	1.28E+07	0.986531	1138202	1.47731	4499638	692419	622.393	72.2701
	A1691	7370548	4754034	3996770	1.27E+07	0.981366	1057766	1.46978	4358373	602144	629.885	73.3843
	A2452	7506100	4903176	4236796	1.28E+07	0.986338	1102893	1.48216	4512385	658255	621.247	72.4662
	A4378	7596746	5029035	4355703	1.28E+07	0.986131	1221503	1.47145	4618194	777631	630.65	72.4472
	A4468	7564881	4943141	4273745	1.28E+07	0.985963	1161957	1.46839	4531843	706300	629.982	72.5192

Legend:

ID	*Individual ID*
NALT	*Number of non-reference genotypes*
NMIN	*Number of genotypes with a minor allele*
NHET	*Number of heterozygous genotypes for individual*
NVAR	*Total number of called variants for individual*
RATE	*Genotyping rate for individual*
SING	*Number of singletons individual has*
TITV	*Mean Ti/Tv for variants for which individual has a nonreference genotype*
PASS	*Number of variants PASS'ing for which individual has a nonreference genotype*
PASS_S	*Number of singletons PASS'ing for which individual has a (singleton) nonreference genotype*
QUAL	*Mean QUAL for variants for which individual has a nonreference genotype*
DP	*Mean variant DP for variants for which individual has a nonreference genotype*

Box 1.

A conceptual data flow chart to show how the VCF file and associated meta information is handled.



Box 2.

An overview of a VCF file is shown in this vcf-stats output in jason format.

```
'all' => {
        'mait_1' => 5979847,
        'snp_count' => 5979847,
        'shared' => {
                    '1' => 5979847
                  },
        'snp' => {
                    'G>C' => 224703,
                    'C>T' => 832147,
                    'T>C' => 911945,
                    'C>G' => 224336,
                    'C>A' => 276406,
                    'G>T' => 278651,
                    'T>G' => 516897,
                    'T>A' => 225646,
                    'A>G' => 912859,
                    'T>C' => 516897,
                    'G>A' => 835895,
                    'A>T' => 225504,
                    'A>C' => 514858
                  },
        'count' => 5979847
      },
'samples' => {
             '9111' => {
                         'snp' => {
                                     'C>T' => 832147,
                                     'G>C' => 224703,
                                     'C>G' => 224336,
                                     'C>A' => 276406,
                                     'G>T' => 278651,
                                     'T>C' => 911945,
                                     'T>A' => 225646,
                                     'T>G' => 516897,
                                     'A>G' => 912859,
                                     'A>C' => 514858,
                                     'G>A' => 835895,
                                     'A>T' => 225504
                                   },
                         'ref' => 3668064,
                         'snp_count' => 5979847,
                         'ref_count' => 3668064,
                         'unphased' => 5979847,
                         'count' => 5979847,
                         'hom_AA_count' => 2311783,
                         'private' => 5979847,
                         'het_RA_count' => 3668064
                       }
           }
```

Table 2. Function sets from vcftools

Name of Tools	Functions
vcftools	To analyse VCF files (multiple optoins available)
vcf-stats	Obtain basic statistics of a VCF file
vcf-to-tab	Generate tab-delimited list of genotype metrix
vcf-convert	Convert between VCF versions
vcf-merge	Merge multiple VCF files
vcf-compare	Compare two or more VCF files for differences
vcf-query	Query specific chromosome region for information

References

1. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics. 2011 Aug 1;27(15):2156-8. doi: 10.1093/bioinformatics/btr330. Epub 2011 Jun 7. Web site http://vcftools.sourceforge.net/ last accessed on December 1, 2013.*
2. Shaun Purcell et al. (2011). PLINK/Seq: Analysis of genetic variation data from large-scale, population-based medical sequencing studies. *Workshop at the Analytic and Translational Genetics Unit, MGH; Center for Human Genetic Research, MGH. Web site https://atgu.mgh.harvard.edu/plinkseq/ last accessed on December 21, 2013.*
3. Shane Neph, M. Scott Kuehn, Alex P. Reynolds, et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics (2012) 28 (14): 1919-1920. doi: 10.1093/bioinformatics/bts277. Web site https://bedops.readthedocs.org last accessed on December 11, 2013.*

Acknowledgements

We sincerely thank Dr. Carrie Finno from University of Minnesota, Dr. Robert D. Schnabel from University of Missouri, Dr. Curt Van Tassell from USDA-ARS, and Dr. Susan Lamont from Iowa State University for kindly sharing their VCF files for the experimental works testing various tools on the NAGRP VCF Data Repository to proof the concept. We also thank Xue-feng Zhao and James Coyle from the Baker & HPC Centers at Iowa State University for their assistance using the ISU computing facilities.