SNPchiMp v.2: An Open Access Web Tool for SNP Data Management on Bovine, Porcine and Equine Livestock

E.L. Nicolazzi[1], **N. Nazzicari**[1], **A. Caprera**[1], **I. Fojadelli**[1], **F. Strozzi**[1], **R.D. Schnabel**[2], **C. Lawley**[3], **A. Pirani**[4], **F. Brew**[5], **C. Soans**[3], **H. Jorjani**[6], **G. Evans**[7], **B. Simpson**[8], **J.L.Williams**[1], **A.Stella**[1,9].

[1]Fondazione Parco Tecnologico Padano, Lodi, Italy, [2]University of Missouri, Columbia, MO, USA, [3]Illumina Inc, San Diego, CA, USA, [4]Affymetrix Inc., Santa Clara, CA, USA, [5]Affymetrix UK Ltd., High Wycombe, UK, [6]Interbull center, Uppsala, Sweden, [7] GeneSeek, a Neogen Company, Scotland, UK [8] GeneSeek, a Neogen Company, Lincoln, NE, USA [9]IBBA, CNR, Lodi, Italy

ABSTRACT: Since the beginning of the genomic era, the SNP chip market in livestock species has grown almost exponentially. Today, researchers are asked to deal with many SNP chips on daily basis, and this requires having the (general and specific) information on the SNPs available and at hand. However, the information is often difficult to obtain (e.g. data on chips no longer on the market), integrate and standardise. Here we present SNPchiMp v.2, a multi-species database linked to an open-access web-based interface that solves many of these problems. This second version of the tool includes 9 bovine, 2 porcine and 1 equine SNP chips, marketed by Illumina, Affymetrix and GeneSeek, and genomic exchange indexes from Interbull (for bovine data). The interest of the animal genetics community on the first version of this tool has rapidly increased as have the number of chips and species included.
Keywords: SNP chip; database; web-interface

Introduction

Currently, a large number of SNP chips are available for genomic analysis in all major livestock species. Although genome-wide SNP genotyping technology has mainly been massively applied mainly in dairy cattle and pigs, the number of SNP chips available for other species is steadily increasing. The two main manufacturers (Illumina and Affymetrix), in fact, continuously upgrade their products to fit market needs. Moreover, there are a number of companies or public/private consortia producing "custom" SNP chips, which include a specific selection of SNPs.

In a scenario common to all major livestock species, different chips and densities of SNPs are routinely exchanged within collaborations between projects and countries (VanRaden et al. (2012); Lund et al. (2011)). However, many difficulties can be found when combining genomic data coming from different sources, such as: identifying SNPs in common among different chips and referencing them to a common assembly, correcting for SNP IDs cross-references (e.g. different names for same SNPs), and retrieving other useful information including allelic coding formats, consensus sequences, orientation and strand of the SNP. Much of this information is available but spread across different publicly accessible databases. However, the retrieval of the information from the different sources is not straightforward, as there is no common searchable key.

We here present SNPchiMp v.2, the second version of a MySQL database linked to an open access web-based interface that links the information from different sources in order to efficiently manage SNP chip data. The first version included only official manufacturers SNP chips for bovine cattle (Nicolazzi et al. (2014)). This new version, also developed within the Gene2Farm project (G.A. 289592), currently includes information for chips from Bovine, Porcine and Equine species, marketed by Illumina, Affymetrix and GeneSeek, for a total of 12 SNP chips. The number of chips and species is increasing rapidly and information will be updated as it becomes available. For cattle, as across-countries genomic evaluations are now in place, SNPchiMp v.2 includes Interbull within and across official SNP indexing.

Materials and Methods

Information on the methods used for the collection, management and retrieval of data are fully described in Nicolazzi et al. (2014). Integrations and new features for the current version of the SNPchiMp tool are summarized below.

Data collection and new species. Procedures applied during the construction of the first version, were followed here. SNP data were collected from the following sources:

1- All Bovine, Porcine and Equine dbSNP database builds (dbSNP (2014)) were downloaded from June 2012 and January 2013, for bovine and the other species, respectively. This resulted in the current availability of builds 136, 137 and 138 for the Bovine species, 137 and 138 for the Porcine species and 138 and 139 for the Equine species.
2- Illumina SNP chip data (five Bovine, two Porcine and one Equine SNP chips) were collected from manifest files of the Illumina GenomeStudio software (Illumina (2014)) for SNP chips currently marketed, and from data obtained directly from the manufacturer for "historic" SNP chips.
3- Annotation files for the Affymetrix SNP chip (Axiom Genome-Wide BOS 1 array) were downloaded from the Affymetrix website support section (versions 33 and 34; Affymetrix (2014)). The latter version of the annotation file contained approximately 30% more rsIDs than the previous release, allowing a greater overlap with Illumina SNPs.

4- GeneSeek GGP SNP chip data (currently three Bovine SNP chips) were provided directly by the company.
5- Interbull within and across chip exchange indexes were provided directly by Interbull.

A total of 1,663,871, 123,728 and 54,602 original SNP data were stored in the database for Bovine, Porcine and Equine species, respectively. These data were linked to the 48,676,943, 29,051,427 and 5,452,501 records extracted from the dbSNP database for the same species, respectively. The integration of the Porcine and Equine species with GeneSeek SNP data is expected soon.

As for the first version of the SNPchiMp, associations between commercial SNP IDs and rsIDs were found by merging the SNP IDs against the "Submitter ID" section of the dbSNP database for Illumina and GeneSeek SNP chips. These associations were re-checked independently and confirmed by blasting a random selection of consensus sequences. For Affymetrix SNP array, 360,274 SNP IDs versus rsIDs association were supplied in the annotation file provided by the manufacturer.

Changes in the database architecture and web-interface. In this version 2 of the SNPchiMp, general data structure has been rationalised and data redundancy has been resolved in favour of orthogonality. In particular, the concept of "native" assembly is not considered as a special data structure as in version one, and it is now treated as a standard assembly. As a consequence, all the SNP coordinates have been moved in the unique ALLmap table, with most SNP referenced to a specific assembly, and all to a virtual "native" assembly.

Multi-species support has been achieved through a system of table prefixes. The current database contains a replica of the whole data structure for each species - there is, thus, a cow_SNPall table, a pig_SNPall table, and a horse_SNPall table. In this way the tool can easily extend the support to new species, and guarantee data isolation when new chips are introduced. A new PHP based query engine has been developed. This helps to separate data storage from data presentation, introducing an abstraction layer between web interface code and database.

This mechanism allows for database evolution, limiting the impact of a data structure modification on the code. Moreover, optimization techniques including controlled redundancy can now be handled without disrupting other SNPchiMp functions.

Implementation of web-based tool.

The User interface has been modified and extended to allow access to new data included in the SNPchiMp v.2 database (e.g. data retrieval and display functionality have been modified to accommodate the new chips and species). The accessibility and the "user friendliness" of the interface have also been improved. In this new version of the tool, client side scripts in JQuery maximize responsiveness and graphical appeal of the pages, which will help the user to surf the data in a more effective way (JQuery (2014)). Ruby on Rails was substituted with a more flexible and simple system linked to the PHP query functions in order to improve coding ease, modularity and flexibility.

Results and Discussion

The amount of information contained in the database, and the simplicity of access to the information make SNPchiMp v.2 a valuable tool for users analysing SNP data for the three species currently included in the database. The SNPchiMp tool can be used for:

- the integration of genomic data coming from different sources, including different SNP chip releases, densities and platforms;
- the imputation of SNP data from lower to higher densities (and even from high density to full sequence, as ssID information from dbSNP is linked to all SNPs with a SNP ID vs. rsID association);
- the direct update of genomic coordinates to any of the available assemblies, for species with more than one reference assembly, and the alignment with the latest build of the assembly from the original manufacturer coordinates. This is particularly important for species in which genome assemblies are at the early stages and rapidly evolving;
- the integration of post-hoc genomic analyses, e.g. of a subset of significant SNPs. A dedicated menu "Browse" is available to interrogate the database for a subset of SNPs, instead of downloading the whole database. In this case, during the preview of results, a click of the mouse on the appropriate logo will redirect the user to the corresponding SNP information on dbSNP or ENSEMBL web-pages.

The data contained in the database has increased considerably from the first version of the tool. For cattle, three new chips (with total of new ~100k new SNPs) and the Interbull exchange indexing are now stored in the database. The number of SNPs with SNP ID vs. rsID associations found varies, depending on the chip considered. For cattle, with the exception of the Affymetrix array, all chips have >99% of rsIDs identified, with more than 99.3% of the rsIDs having genomic coordinates on UMD3.1 (Table 1). Two Illumina Porcine chips are currently included, with nearly 70% of the SNPs mapped on the reference assembly (SusScrofa 10.2; Table 2). The Equine chip available has all SNP IDs associated to a rsID, and more than 99.9% of these have mapped coordinates on the reference assembly (EquCab 2.0; Table 3).

For continuity, the functionality of the SNPchiMp v.2 web-interface was maintained with very similar appearance to that of the first version. Small changes have provided the user with more options. The inclusion of new species and the increase in the memory requirements, however, necessitated a thorough revision of the technical aspects of the database and web-interface, which have been completely re-built. The new web-interface now allows greater flexibility, which is a key aspect of the SNPchiMp. This version currently contains SNP chip information for

three species, but has the technical capability of handling an indefinite number of species. Data collection is underway to incorporate new species, including sheep, goat and chicken.

Table 1. Proportion, in % of the total number of SNP (SNP #), of Bovine SNPs with genomic coordinates on BTA4.2 (BTA1), UMD3.1 (UMD) or BTA4.6 (BTA2) assemblies or without rsID (NOrs) assigned.

Chip[1]	SNPs #	BTA1	UMD	BTA2	NOrs
B1	2,900	98.8	99.3	99.1	0.5
B2	6,909	98.4	99.8	98.7	0.1
B3	54,001	96.2	99.6	96.9	<0.1
B4	54,609	95.8	99.3	96.6	0.3
B5	777,962	90.3	99.6	92.5	<0.1
B6	648,875	41.8	74.3	70.4	25.7
B7	8,610	96.6	99.6	97.4	0.2
B8	19,721	90.8	99.8	92.2	0.1
B9	76,879	91.2	99.7	93.2	0.1

[1] SNP chips code: B1- Illumina Golden Gate Bovine3K BeadChip; B2- Illumina Infinium BovineLD BeadChip; B3- Illumina Infinium BovineSNP50 v1 BeadChip; B4- Illumina Infinium BovineSNP50 v2 BeadChip; B5- Illumina Infinium BovineHD BeadChip; B6- Affymetrix Axiom Genome-Wide BOS 1; B7- GeneSeek Genomic Profiler (GGP) LD v.1 (unique SNPs); B8- GeneSeek Genomic Profiler (GGP) LD v.2 (unique SNPs); B9- GeneSeek Genomic Profiler (GGP) HD (unique SNPs).

Table 2. Proportion, in % of the total number of SNP (SNP #), of Porcine SNPs with genomic coordinates on SScrofa10.2 (SS10.2) assembly or without rsID (NOrs) assigned.

Chip[1]	SNP #	SS10.2	NOrs
P1	62,163	69.4	26.65
P2	61,565	69.3	26.68

[1] SNP chips code: P1- Illumina Infinium PorcineSNP60 v1 BeadChip; P2- Illumina Infinium PorcineSNP60 v2 BeadChip.

Table 3. Proportion, in % of the total number of SNP (SNP #), of Equine SNPs with genomic coordinates on EquCab2.0 (EC2.0) assembly or without rsID (NOrs) assigned.

Chip[1]	SNP #	EC2.0	NOrs
E1	54,602	99.9	0.0

[1] SNP chips code: E1- Illumina Infinium EquineSNP50 BeadChip.

Conclusion

The SNPchiMp v.2 is a very powerful tool to obtain, integrate, analyse and standardise SNP chip data in three of the major livestock species (cow, pig and horse). The database will soon be further updated to include SNP chip data for at least three more species: sheep, goat and chicken.

Although a thorough revision of the technical aspects were required to handle the substantial increase of data, the user experience of the tool remains very similar to the first version.

SNPchiMp v.2 is freely available at http://bioinformatics.tecnoparco.org/SNPchimp. This tool is open access and multiplatform (accessible from any browser in any operating system).

SNPchiMp wiki

SNP: Single nucleotide polymorphism;
dbSNP: NCBI database of SNPs and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants;
rsID: Reference SNP (from dbSNP);
ssID: Submitted SNP (from dbSNP).

Acknowledgments

Authors wish to acknowledge Matteo Picciolini for his contribution to the first version of this tool. Franz Seefried and Birgit Gredler are kindly acknowledged for their important feedback on the early stages of this new version of the tool. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 289592 – Gene2Farm.

Literature Cited

Affymetrix (2014). http://www.affymetrix.com/analysis/downloads/na33/genotyping/Axiom_GW_Bos_SNP_1.na33.1.annot.csv.zip Accessed on Feb. 2014.

dbSNP (2014). http://www.ncbi.nlm.nih.gov/snp/ Accessed on Feb. 2014.

Illumina (2013). http://support.illumina.com/array/downloads.ilmn Accessed on Feb. 2014.

JQuery (2014). http://jquery.com Accessed on Feb. 2014.

Lund M. S., de Roos A. P. W., de Vries A. G., et al. (2011). Genet. Sel. Evol. 43:43.

Nicolazzi, E. L., Picciolini, M., Strozzi F. et al. (2014). BMC genomics 15:123.

VanRaden P. M., O'Connell J. R., Wiggans G.R., et al. (2012). Genet. Sel. Evol. 43:10.