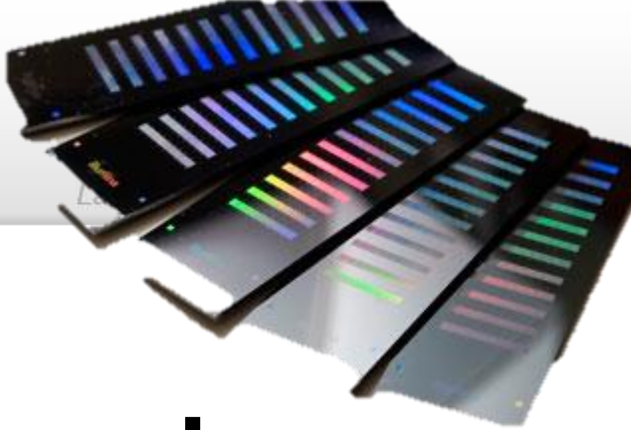


Development of approaches to compare and integrate technologies

(with case scenarios)

Ezequiel L. Nicolazzi

Fondazione Parco Tecnologico Padano

SNP data

Genotypes are only a part…

- Handling genotypes is “easy”, but what about the rest?
- Original files coming from the lab
- Own file recoding and formatting
- Own programming pipeline to get and use data from other sources
- No (or very few and feeble) efforts for standardization
- Genomic analyses rely heavily on this “accessory” information

Need *large* integration with *many* sources of info.
Inefficient use of time and efforts!
Difficult to keep updated
Some steps require knowledge of chip development history
Such large data.. Errors happen

Much work done on developing methods, very little to develop handy tools

Welcome to the (bovine) jungle

Illumina Infinium Bovine SNP50

✓ 1 chip, 1 assembly (BTAU 4.0)

x *output formats (row, matrix, etc)*
x *allele coding (forward, top, A/B)*
x *Illumina SNP names and public DBs*

Illumina Infinium Bovine SNP50 (v.2)
Illumina Golden Gate Bovine3k
Illumina Infinium BovineLD
Illumina Infinium BovineHD
Illumina Infinium Bovine LDv1.1

✓ Improved quality of information
✓ More (less) SNPs

x *output formats (row, matrix,etc)*
x *allele coding (forward,top, A/B)*
x *Illumina SNP names and other. DBs*
x *2 assemblies (BTAU 4.0 and UMD 3.1)*
x *SNPs in common?*
x *SNP names in common?*

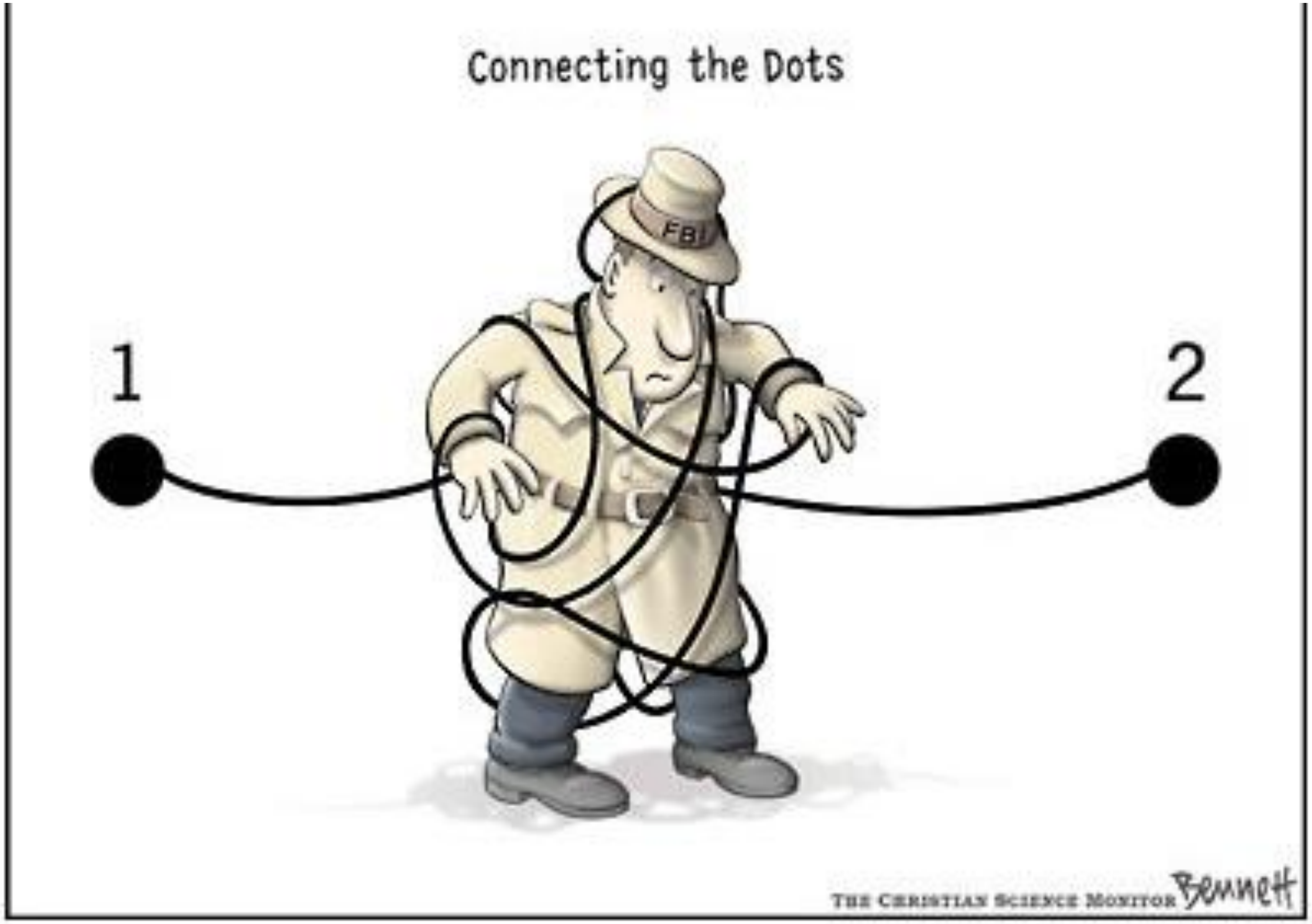
Affymetrix Axiom Bos1 (HD)

✓ New technology
✓ New SNPs

x *New formats and procedures*
x *SNP in common?*
x *SNP names in common?*
x *Concept of probe!*

GeneSeek – NeoGen chips
Many custom SNPchips

Why not?


Connecting the Dots
FBI
1
2
THE CHRISTIAN SCIENCE MONITOR
Bennett

How?

- The Gene2farm project ***“Next generation European system for <u>cattle</u> improvement and management”***
  - Started Jan 2012, Ends Dec 2015
- *Research for the benefit of SME* funding scheme (19 partners)
- Main objectives (*only small..er breeds*):
  - complete genome information to understand genome structure and design new genotyping panels
  - develop tools to impute data and **to make exchange information easier.**
  - measure a large # of biological variables underlying important commercial traits
  - develop statistical models and applications for using the genomic and phenotypic data
  - disseminate the information to the SMEs, cattle breeding industry & end users.

The perfect excuse: **Task 2.4. SNP panel inter-changeability**

Ok, but HOW?
Not TOO hard, really…

- By connecting people and information (dots)
- **Collect all information from producers [e.g. barely-legal stalking]**
- Download dbSNP database(s) -> all builds from 2012
- **Link the information (get SNP name – rsID link)**
- Put all this into a database.
- Re-check everything independently (Bob Schnabel on cow).
- **Make it easily accessible to users (web-app):**

  http://bioinformatics.tecnoparco.org/SNPchimp

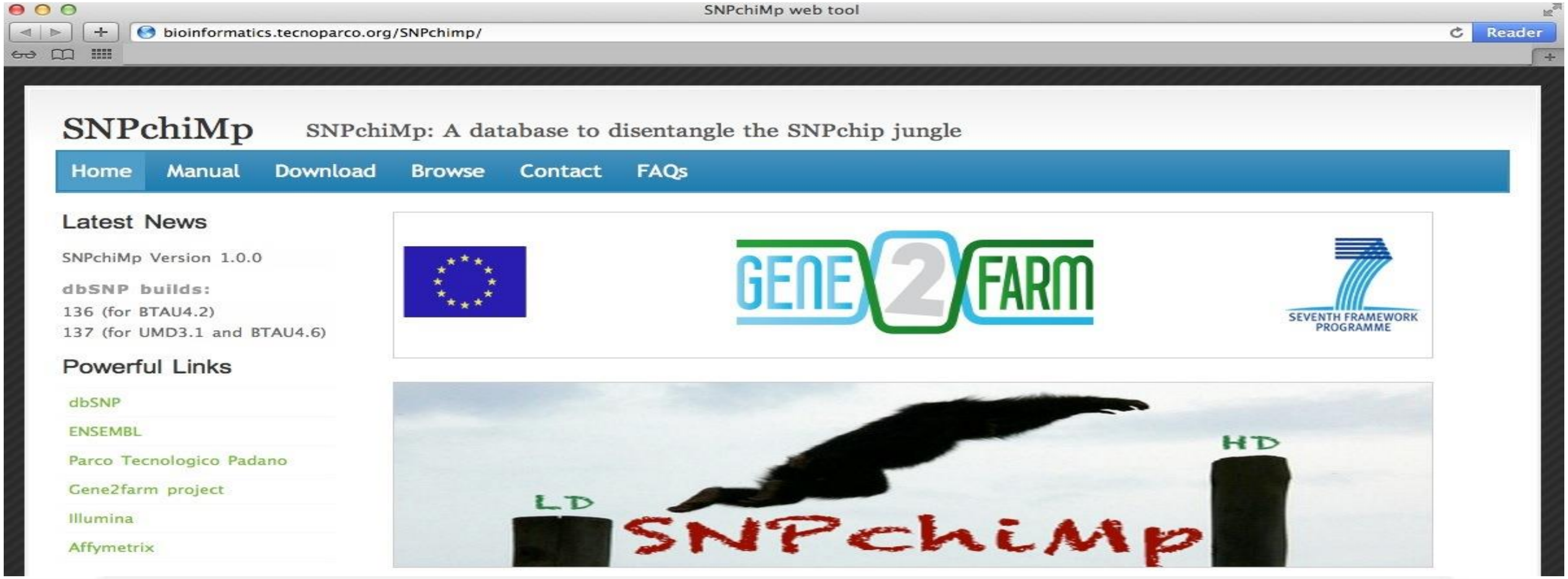
The SNPchiMp is born

Parco Tecnologico Padano — E. Nicolazzi, M. Picciolini, F. Strozzi, A. Stella

Mizzou University of Missouri — B. Schnabel

illumina — C. Lawley

affymetrix Biology for a better world — A. Pirani and F. Brew



SNPchiMp SNPchiMp: A database to disentangle the SNPchip jungle

Latest News

SNPchiMp Version 1.0.0

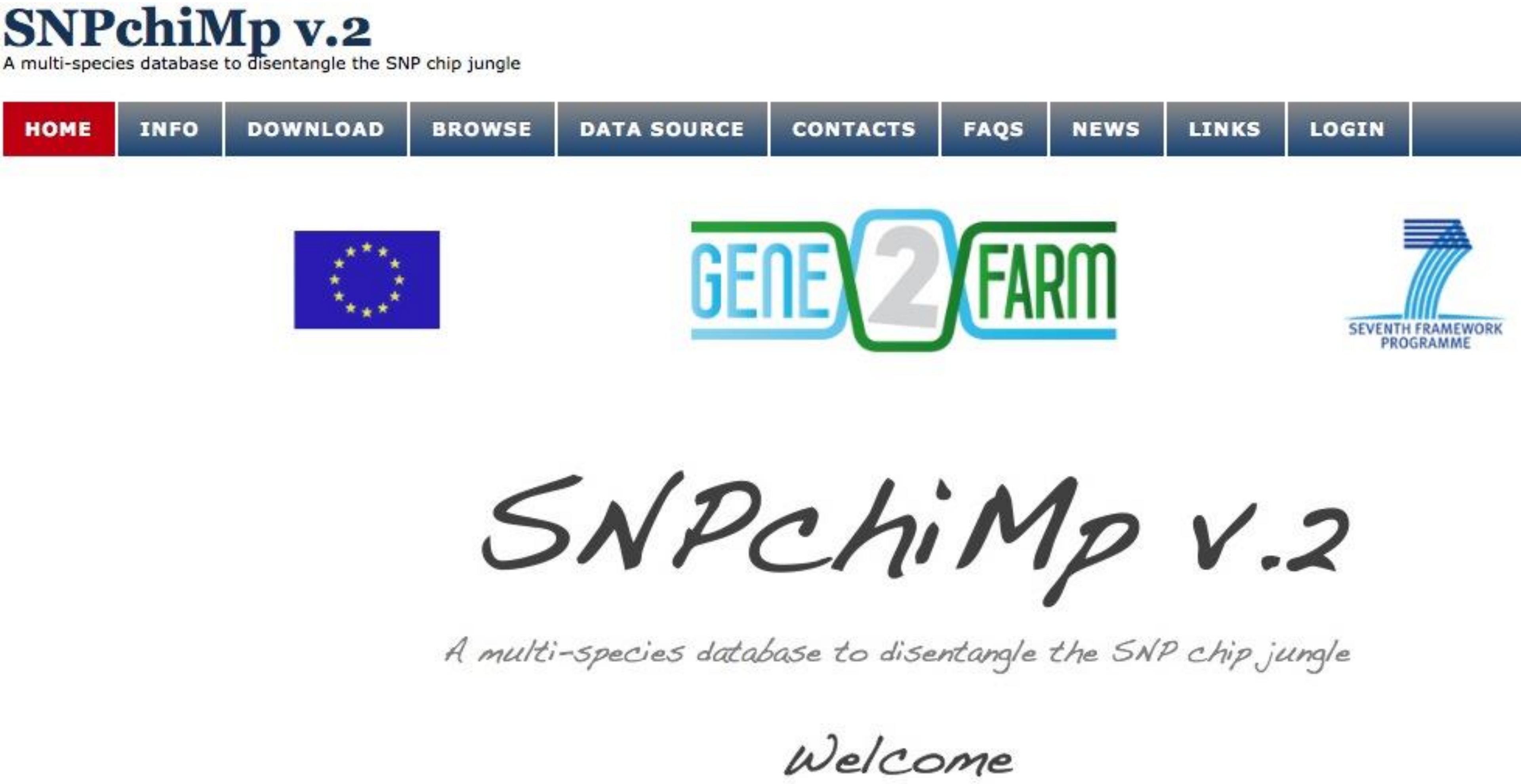
dbSNP builds:
136 (for BTAU4.2)
137 (for UMD3.1 and BTAU4.6)

Powerful Links

- dbSNP
- ENSEMBL
- Parco Tecnologico Padano
- Gene2farm project
- Illumina
- Affymetrix

The SNPchiMp gets updated (v.2)

http://bioinformatics.tecnoparco.org/SNPchimp



SNPchiMp v.2

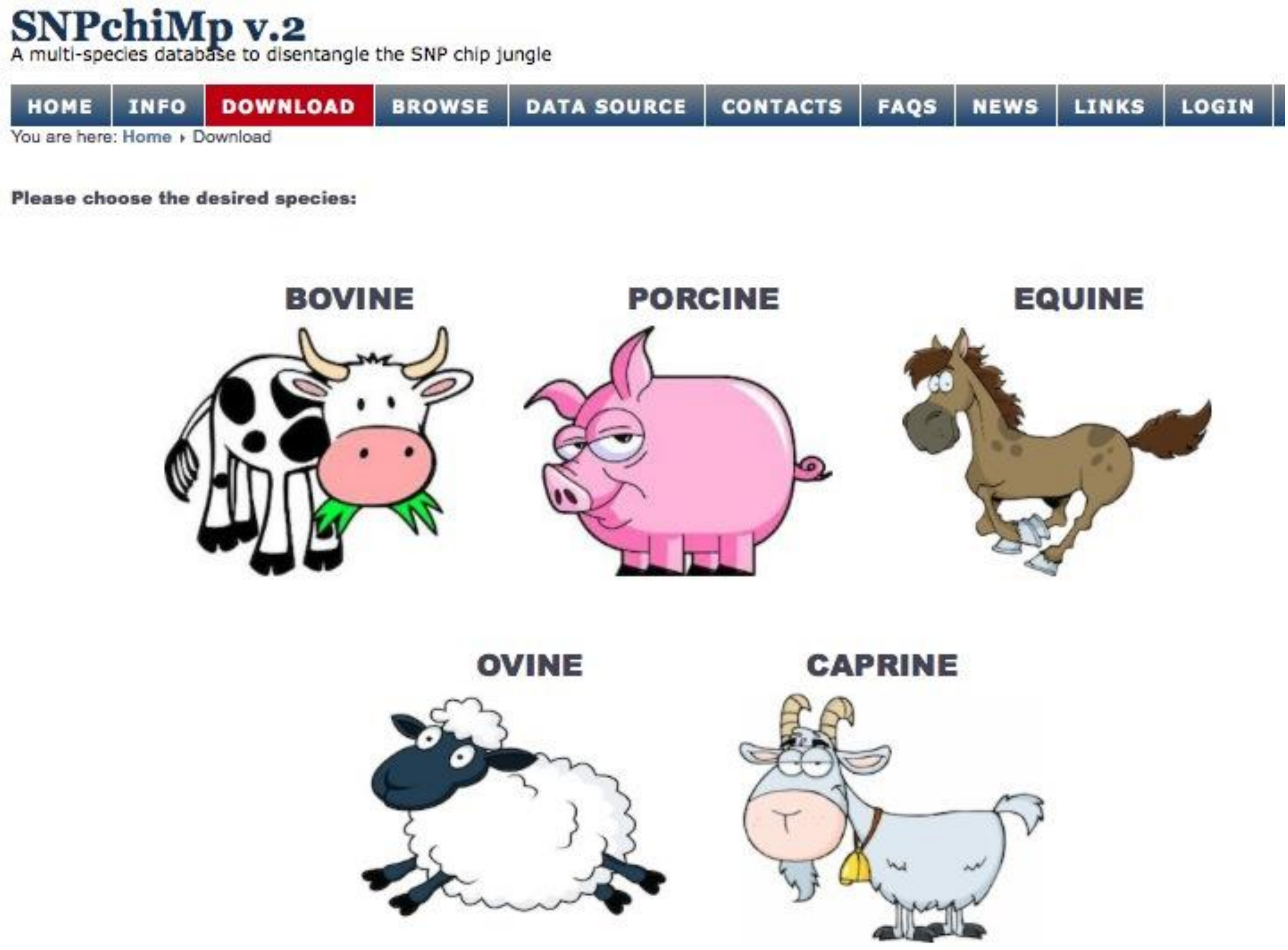
A multi-species database to disentangle the SNP chip jungle

SNPchiMp v.2

A multi-species database to disentangle the SNP chip jungle

Welcome

1° step - Selection of species



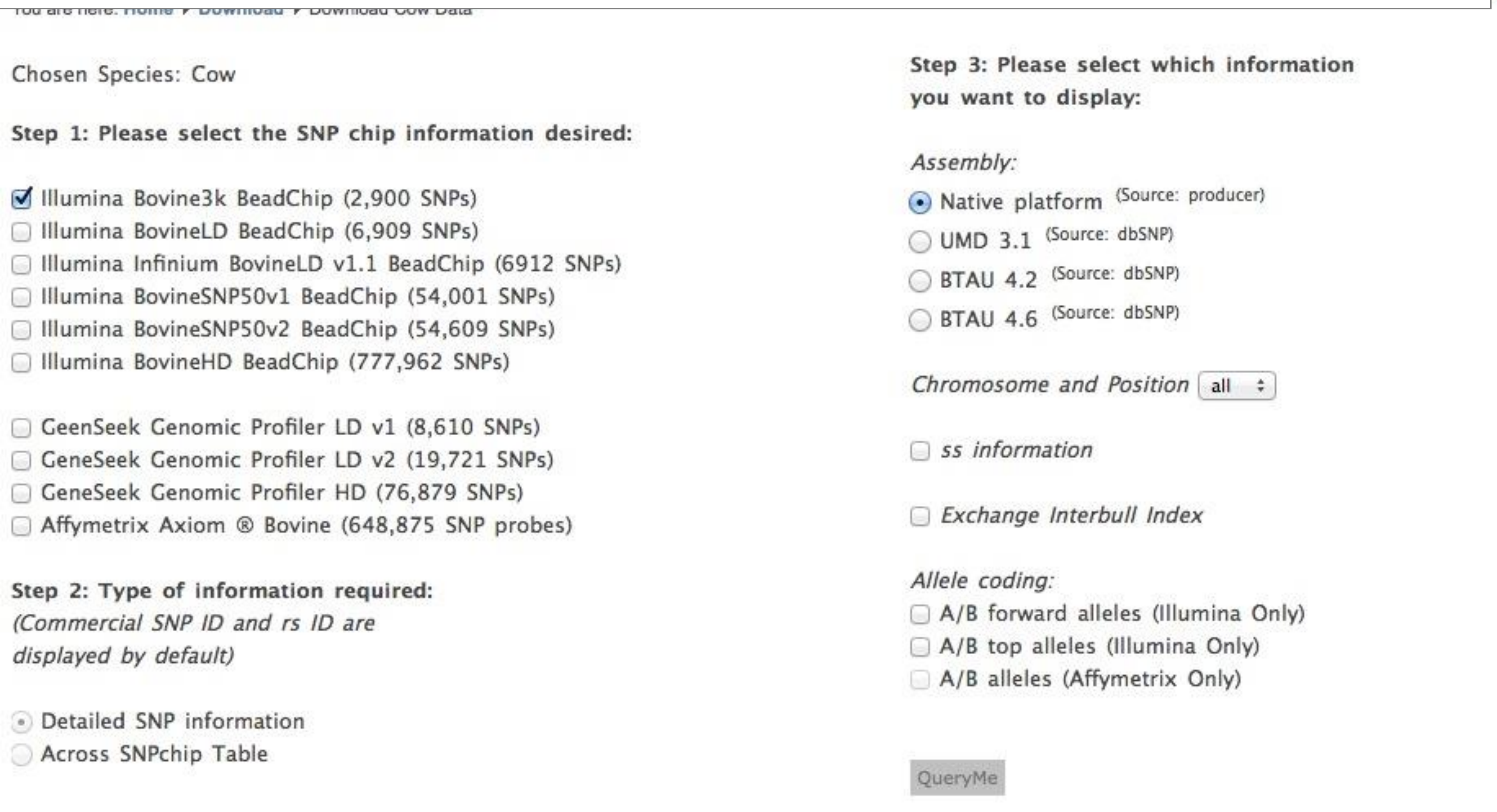
SNPchiMp v.2

A multi-species database to disentangle the SNP chip jungle

You are here: Home › Download

Please choose the desired species:

- *Get coordinates in a different assembly? 3 clicks (think of sheep and goats!)*
- *Latest Interbull index for your chip? 3 clicks*
- *Get allele codings for a chip? 4 clicks*
- *Convert genotype allele coding into forward strand and UMD3.1 assembly (for imputation from chip to full sequence)? 4 clicks*
- *Know which SNPs are in **<u>any</u>** SNP chip combination? At least 5 clicks...*



You are here: Home ▸ Download ▸ Download Cow Data

Chosen Species: Cow

Step 1: Please select the SNP chip information desired:

☑ Illumina Bovine3k BeadChip (2,900 SNPs)
☐ Illumina BovineLD BeadChip (6,909 SNPs)
☐ Illumina Infinium BovineLD v1.1 BeadChip (6912 SNPs)
☐ Illumina BovineSNP50v1 BeadChip (54,001 SNPs)
☐ Illumina BovineSNP50v2 BeadChip (54,609 SNPs)
☐ Illumina BovineHD BeadChip (777,962 SNPs)

☐ GeenSeek Genomic Profiler LD v1 (8,610 SNPs)
☐ GeneSeek Genomic Profiler LD v2 (19,721 SNPs)
☐ GeneSeek Genomic Profiler HD (76,879 SNPs)
☐ Affymetrix Axiom ® Bovine (648,875 SNP probes)

Step 2: Type of information required:
(Commercial SNP ID and rs ID are displayed by default)

● Detailed SNP information
○ Across SNPchip Table

Step 3: Please select which information you want to display:

Assembly:

● Native platform (Source: producer)
○ UMD 3.1 (Source: dbSNP)
○ BTAU 4.2 (Source: dbSNP)
○ BTAU 4.6 (Source: dbSNP)

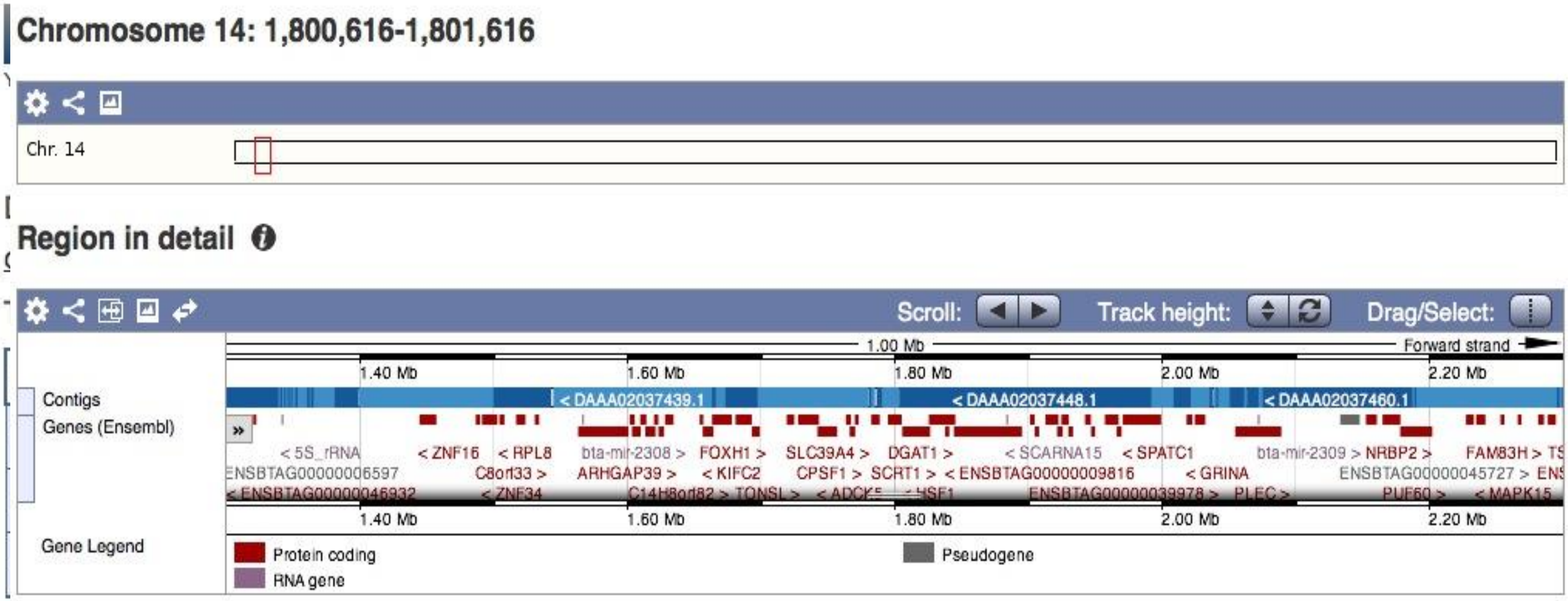
Chromosome and Position all

☐ *ss information*

☐ *Exchange Interbull Index*

Allele coding:
☐ A/B forward alleles (Illumina Only)
☐ A/B top alleles (Illumina Only)
☐ A/B alleles (Affymetrix Only)

Browse menu



Chromosome 14: 1,800,616-1,801,616

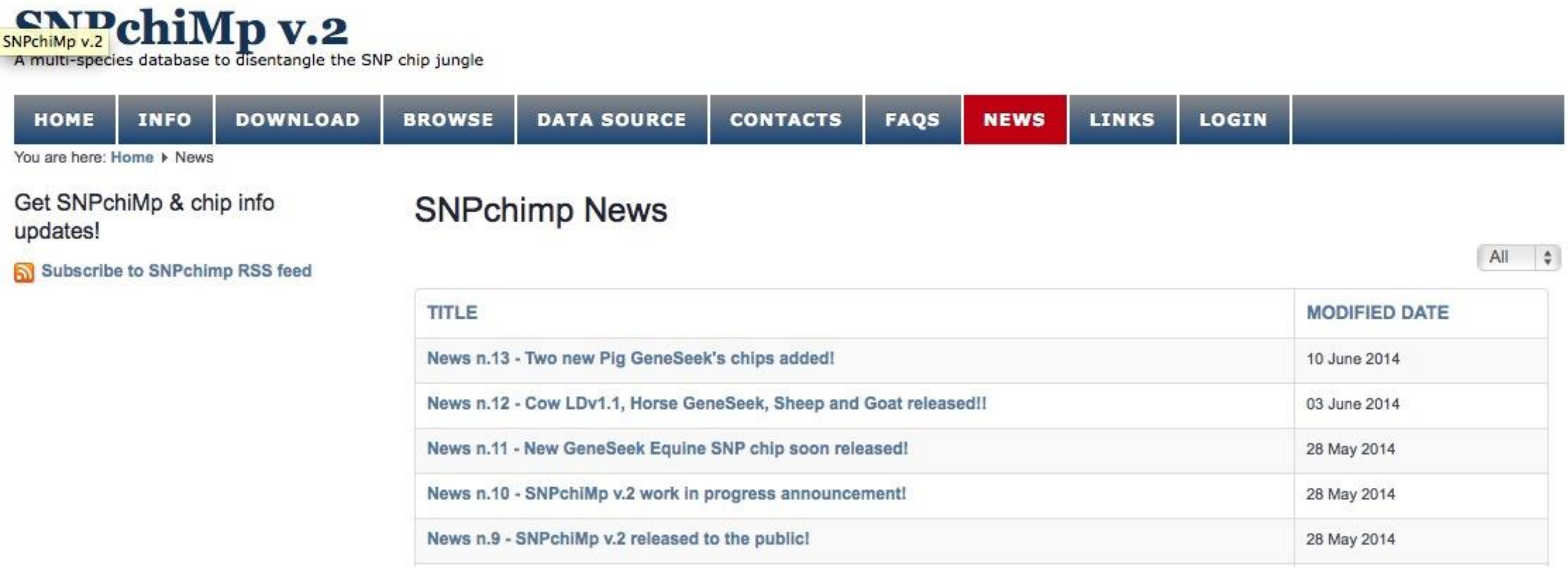
Region in detail

Current status and consistency

- Information received by producers, linked to dbSNP and updated to the database regularly.
- 19 SNP chips available on 5 species (14 mln record):
  - 10 COW (10,028,386 records)
  - 4 PIG (809,992 records)
  - 2 HORSE (479,038 records)
  - 2 SHEEP (2,640,989 records)
  - 1 GOAT (266,736 records)
- SNPchiMP can now be queryed directly from URL! (makes it accessible to external tools!)

http://bioinformatics.tecnoparco.org/SNPchimp/snpchimp/downloadSNP.php?animal=cow&force_distinct=true&action=browse&assembly=bta4_2&info_rs=on&info_ss=on&query_pos=1:1..1000000000

Keeping updated is now easy(er)!



SNPchiMp v.2

A multi-species database to disentangle the SNP chip jungle

You are here: Home ▸ News

Get SNPchiMp & chip info updates!

Subscribe to SNPchimp RSS feed

SNPchimp News

TITLE	MODIFIED DATE
News n.13 - Two new Pig GeneSeek's chips added!	10 June 2014
News n.12 - Cow LDv1.1, Horse GeneSeek, Sheep and Goat released!!	03 June 2014
News n.11 - New GeneSeek Equine SNP chip soon released!	28 May 2014
News n.10 - SNPchiMp v.2 work in progress announcement!	28 May 2014
News n.9 - SNPchiMp v.2 released to the public!	28 May 2014

Real case scenarios

- Getting goat chip coordinates in Chinese assembly v.2 and convert alleles from Forward to Top strand
- Imputation accuracy across reference assemblies
- Integration of information across platforms (Illumina – Affymetrix)
- Imputation from HD to full sequence (tips)

Real case scenarios

- Getting goat chip coordinates in Chinese assembly v.2 and convert alleles from Forward to Top strand
- Imputation accuracy across reference assemblies
- Integration of information across platforms (Illumina – Affymetrix)
- Imputation from HD to full sequence (tips)

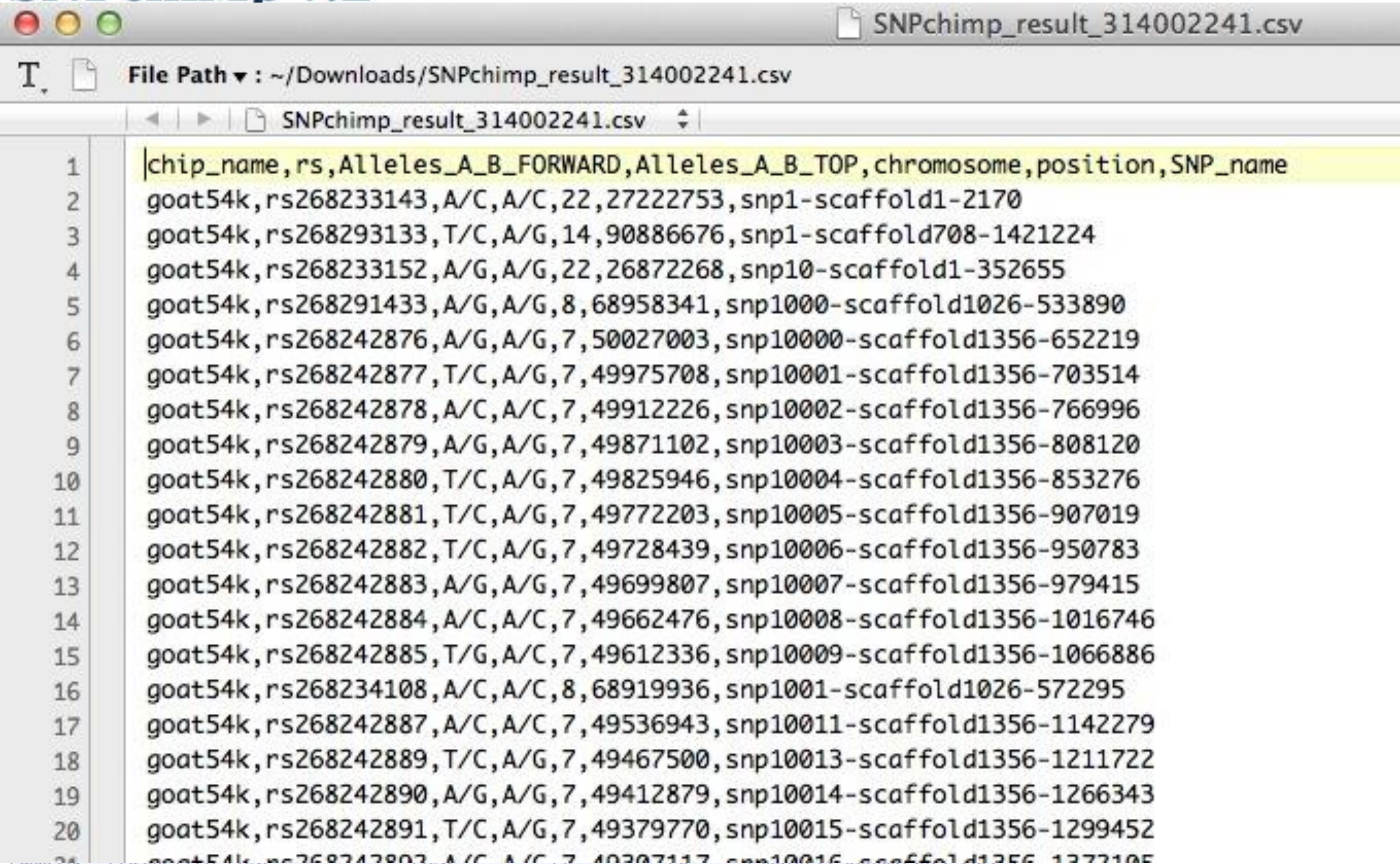
Goat ADAPTmap initiative

- Goat HapMap, coordinated by Alessandra Stella (PTP)
- Collecting goat genotypes from projects all over the world.

- SNPs in IGGC SNP chip are *natively* unmapped.
- IGGC, however, mapped the SNPs against 3 different reference assemblies (goat chinese assembly v.2, sheep assembly v.2 , cow assembly UMD3.1).
- Many researchers prefer to use FORWARD strand allele coding
  - Public databases usually show alleles in the FORWARD strand

- NOT a wise choice… we'll see why in a moment

SNPchiMp v.2



SNPchimp_result_314002241.csv
File Path ▾ : ~/Downloads/SNPchimp_result_314002241.csv
SNPchimp_result_314002241.csv

```
chip_name,rs,Alleles_A_B_FORWARD,Alleles_A_B_TOP,chromosome,position,SNP_name
goat54k,rs268233143,A/C,A/C,22,27222753,snp1-scaffold1-2170
goat54k,rs268293133,T/C,A/G,14,90886676,snp1-scaffold708-1421224
goat54k,rs268233152,A/G,A/G,22,26872268,snp10-scaffold1-352655
goat54k,rs268291433,A/G,A/G,8,68958341,snp1000-scaffold1026-533890
goat54k,rs268242876,A/G,A/G,7,50027003,snp10000-scaffold1356-652219
goat54k,rs268242877,T/C,A/G,7,49975708,snp10001-scaffold1356-703514
goat54k,rs268242878,A/C,A/C,7,49912226,snp10002-scaffold1356-766996
goat54k,rs268242879,A/G,A/G,7,49871102,snp10003-scaffold1356-808120
goat54k,rs268242880,T/C,A/G,7,49825946,snp10004-scaffold1356-853276
goat54k,rs268242881,T/C,A/G,7,49772203,snp10005-scaffold1356-907019
goat54k,rs268242882,T/C,A/G,7,49728439,snp10006-scaffold1356-950783
goat54k,rs268242883,A/G,A/G,7,49699807,snp10007-scaffold1356-979415
goat54k,rs268242884,A/C,A/C,7,49662476,snp10008-scaffold1356-1016746
goat54k,rs268242885,T/G,A/C,7,49612336,snp10009-scaffold1356-1066886
goat54k,rs268234108,A/C,A/C,8,68919936,snp1001-scaffold1026-572295
goat54k,rs268242887,A/C,A/C,7,49536943,snp10011-scaffold1356-1142279
goat54k,rs268242889,T/C,A/G,7,49467500,snp10013-scaffold1356-1211722
goat54k,rs268242890,A/G,A/G,7,49412879,snp10014-scaffold1356-1266343
goat54k,rs268242891,T/C,A/G,7,49379770,snp10015-scaffold1356-1299452
```

Why store Illumina data in TOP and not in FOR strand?

- TOP/BOT allele coding format is a way Illumina has to call consistently alleles irrespectively of the reference assembly (or actual strand). It is sequence based and has NOTHING to do with FOR/REV allele coding.
- (at least some) FOR/REV allele codings will change over time, as assemblies are updated.
- **Warning**: This is already happening in COW: LDv1.1 and GeneSeek-Neogen SNPchips have different allele coding for FOR strand!

Hapmap30759-BTA-123220

ANY Illumina chip: FORWARD: A/G TOP: A/G

Illumina LDv1.1 : FORWARD: T/C TOP: A/G

... and many others like this!

- A/B coding format? Good enough, but less powerful in terms of error checks.

Real case scenarios

- Getting goat chip coordinates in Chinese assembly v.2 and convert alleles from Forward to Top strand
- Imputation accuracy across reference assemblies
- Integration of information across platforms (Illumina – Affymetrix)
- Imputation from HD to full sequence (tips)

Imputation accuracy across reference assemblies

- A lot of research on imputation methods.
- Many methods available, some specifically developed on livestock species (bovine cattle, mostly)

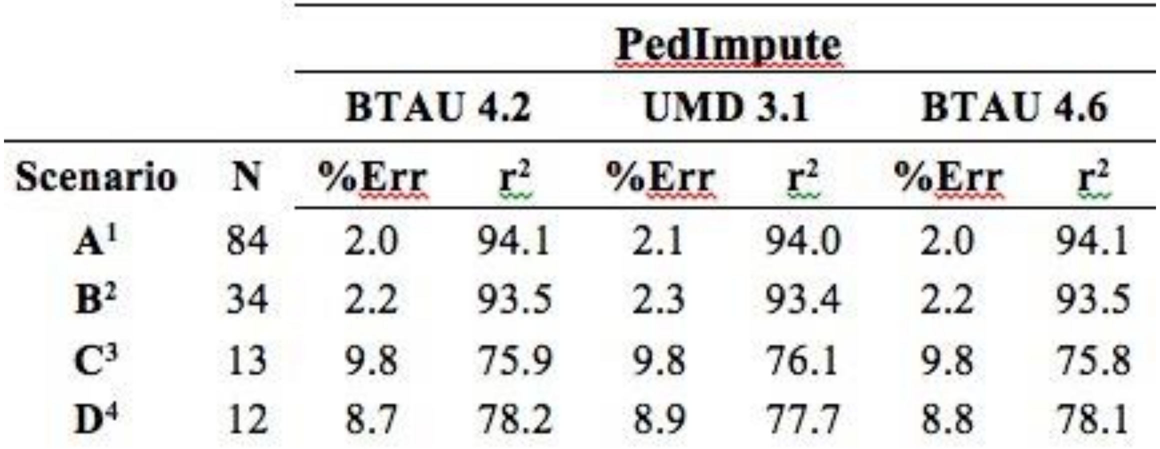
- Imputation methods can be divided into:
  - Population-based methods (use LD)
  - Pedigree-based methods (usually use also LD)

- What will happen when we update the reference assembly?
  - SNPs change position, sometimes chromosomes..

- Will imputation accuracy be better? Worse?

Imputation accuracy across reference assemblies

- *Milanesi et al. (in prep.)*



		PedImpute					
		BTAU 4.2		UMD 3.1		BTAU 4.6	
Scenario	N	%Err	r^2	%Err	r^2	%Err	r^2
A[1]	84	2.0	94.1	2.1	94.0	2.0	94.1
B[2]	34	2.2	93.5	2.3	93.4	2.2	93.5
C[3]	13	9.8	75.9	9.8	76.1	9.8	75.8
D[4]	12	8.7	78.2	8.9	77.7	8.8	78.1

Great variability across scenarios (expected)
Some variability across methods (not shown)
Very little variability across assemblies (in all methods)…

Real case scenarios

- Getting goat chip coordinates in Chinese assembly v.2 and convert alleles from Forward to Top strand
- Imputation accuracy across reference assemblies
- Imputation across platforms (Illumina – Affymetrix)
- Imputation from HD to full sequence (tips)

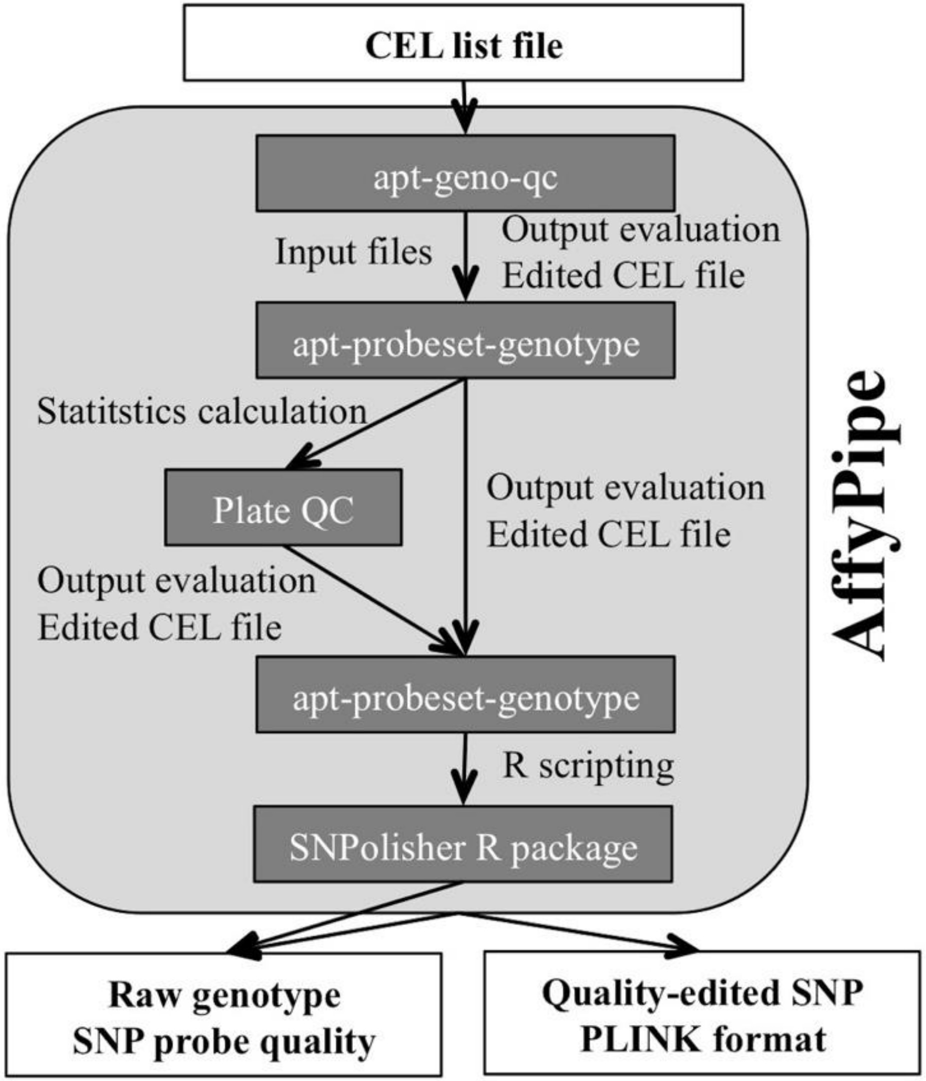
Imputation across platforms (Illumina – Affymetrix)

- A new case of lack of handy tools.
- No all-in-one tool for Linux/Mac users for Affymetrix workflow (raw data → genotypes - and its a long road!)
- Many users (human & livestock). No ready-to-go tool for extraction of genotypes in user-friendly format.

- https://github.com/nicolazzie/AffyPipe

(free, source codes available, no need to program anything)

- *Under review @Bioinformatics.* Received great comments and suggestions from reviewers, new version soon).



Imputation across platforms (Illumina – Affymetrix)

- Still waiting for the Affy data (w.i.p)
- Challenges:
  - Different SNP names ✓
  - Different technology ~✓
  - SNP against SNPprobes ✓
  - SNP calling formats? ~✓
  - Allele coding? ✗

Affymetrix allele coding is ALWAYS FORWARD but…

chip	rsID	FORWARD allele	Chromosome	Position	SNPname
IlluHD	rs42146684	T/G	28	35294673	BTB-00987935
AffyHD	rs42146684	A/C	28	35294673	AX-24625366

Not the only one… ~20k of these!

Will contact both Affymetrix and Illumina on this…

Real case scenarios

- Getting goat chip coordinates in Chinese assembly v.2 and convert alleles from Forward to Top strand
- Imputation accuracy across reference assemblies
- Integration of information across platforms (Illumina – Affymetrix)
- Imputation from HD to full sequence (tips)

Imputation from HD to full sequence (w.i.p)

In collaboration with Qualitas A.G.
Holstein bull. Full sequence (11x – PTP pipeline – Freebayes caller) + HD chip
Assessment before attempting imputation

451652	SNPs total found in sequence (by position)
10.43	avg. coverage for those SNPS.
3275	SNPs skipped since missing in genotypes.
10683	HETEROZYGOUS in genotypes but HOMOZYGOUS in sequence
616	HOMOZYGOUS in genotypes but HETEROZYGOUS in sequence

RESULTS for file: vcf_HD_**ILLUMFOR**.tsv

8114	HOMOZYGOUS in both seq and geno but for a different allele
8287	HOMO or HETERO ok, but different alleles!
420672	**SNPs ok**

RESULTS for file: vcf_HD_**DBSNPFOR**.tsv (from dbSNP → SNPchimp)

50841 HOMOZYGOUS in both seq and geno but for a different allele
33544 HOMO or HETERO ok, but different alleles!
352688 SNPs ok

WHY?!?!?!

Conclusion

- SNPchimp will continue to grow. Plans include linking it to other tools & extend it to more species, more chips, more tools.
- Still a long road from an unified, standardized information.
- We started discovering (& raising) problems to commercial companies
  - Request should come from whole AG Community!
- Many issues could be easily solved if open collaboration was possible (sometimes a "*super-partes*" figure helps..)
- Links across databases are much easier now (through rsID), but not enough!
- Long road to integrate SNP chip data with full sequence. And we're heading that way!

- **WE NEED MORE TOOLS**. We need consolidated, good information. And more tools able to make this information EASILY available.

Thank **you** for your attention

Acknowledgments:

- **Genomic Resources**
- **FPTP**
  - Nelson Nazzicari
  - Andrea Caprera
  - Stefano Biffani
  - Filippo Biscarini
  - Ilaria Fojadelli
  - John Williams
  - Daniela Iamartino
  - Francesco Strozzi
  - Alessandra Stella
- **SNPchimp v.2 team + contributing:**
  - Bob Schnabel (MISSOU)
  - Cindy Lawley (Illumina)
  - Chandrasen Soans (Illumina)
  - Ali Pirani (Affymetrix)
  - Fiona Brew (Affymetrix)
  - Hossein Jorjani (Interbull)
  - Barry Simpson (GeneSeek)
  - Gary Evans (GeneSeek)
  - John McEwan (AgResearch)
  - Rudiger Brauning (AgResearch)
  - Gwenola Tosser-Clopp (INRA)
- **UNICATT**
  - Marco Milanesi
  - Paolo Ajmone-M.
- **Partners from:**
  - Gene2Farm
  - NextGen