Comparing SNP positions from SNPchiMp

Arne B. Gjuvsland

January 4, 2017

Contents

Getting the data

The original data regarding flanking sequences, assembly and Illumina's inferred position can be found at:

- BovineSNP50 v1.0
- BovineSNP50 v2.0
- BovineHD

The NCBI dbSNP database https://www.ncbi.nlm.nih.gov/projects/SNP/ contains the same SNPs, but due to different assemblies and/or mapping procedures the positions are different.

The SNPSNPchiMp database contains both sets of information:

- Nicolazzi et al 2014. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock
- Nicolazzi et al 2014. SNPchiMp v.2: An Open Access Web Tool for SNP Data Management on Bovine, Porcine and Equine Livestock
- Nicolazzi et al 2015. SNPchiMp v.3: integrating and standardizing single nucleotide polymorphism data for livestock species.]
- SNPchiMp website

Data for the Illumina chips 50Kv1, 50Kv2 and 777K was retrieved from http://bioinformatics.tecnoparco.org/SNPchimp/index.php/download/download-cow-data (see figure). Positions from both Illumina (assembly: "Native platform") and dbSNP (assembly: "UMD3.1") was downloaded as gzipped tab-separated files.

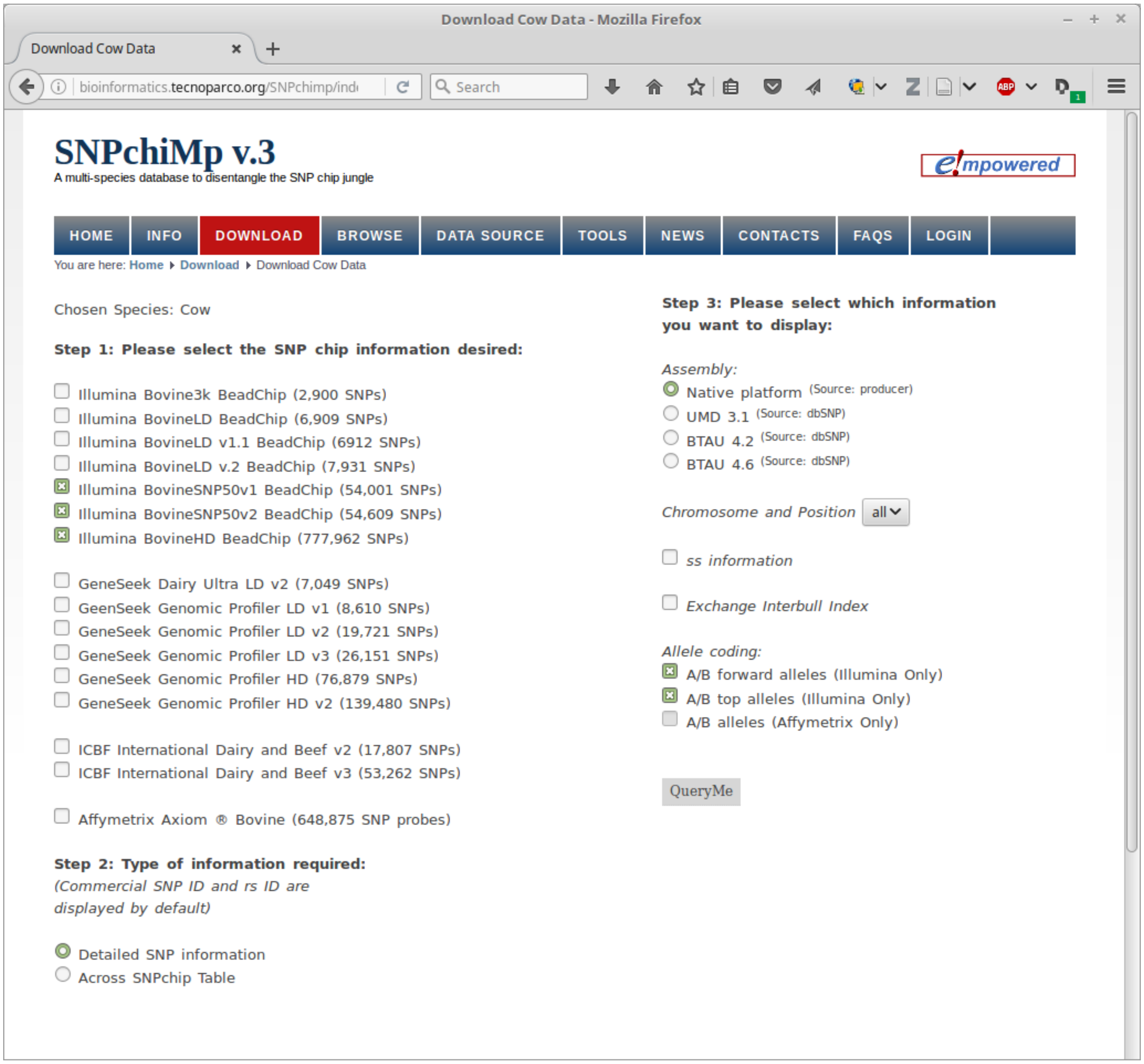
Comparing Native and UMD3.1 positions

First we read in and combine the positions for native and UMD3.1.

```
#read in SNPchimp data with positions from chip provider (assembly: Native platform)
#and from dbSNP (assembly: UMD3.1)
nat <- fread('zcat illumina_50Kv1_50Kv2_777K_native.tsv.gz',showProgress=F)
umd <- fread('zcat illumina_50Kv1_50Kv2_777K_UMD3.1.tsv.gz')
both <- merge(nat,umd,by=c('chip_name','SNP_name'))[,.(chip_name,SNP_name,chr.nat=chromosome.x,pos.nat=
```

Counting cases where the position differes between Native and UMD3.1 we find that:

- Almost every position differs for the Illumina 50Kv1 chip, indicating different assemblies.
- For the 50Kv2 and 777K chips only a few percent of the SNPs differ, indicating the same assembly.



SNPchiMp v.3

A multi-species database to disentangle the SNP chip jungle

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 BovineLD v1.1 BeadChip (6912 SNPs)
□ Illumina BovineLD v.2 BeadChip (7,931 SNPs)
☒ Illumina BovineSNP50v1 BeadChip (54,001 SNPs)
☒ Illumina BovineSNP50v2 BeadChip (54,609 SNPs)
☒ Illumina BovineHD BeadChip (777,962 SNPs)

□ GeneSeek Dairy Ultra LD v2 (7,049 SNPs)
□ GeenSeek Genomic Profiler LD v1 (8,610 SNPs)
□ GeneSeek Genomic Profiler LD v2 (19,721 SNPs)
□ GeneSeek Genomic Profiler LD v3 (26,151 SNPs)
□ GeneSeek Genomic Profiler HD (76,879 SNPs)
□ GeneSeek Genomic Profiler HD v2 (139,480 SNPs)

□ ICBF International Dairy and Beef v2 (17,807 SNPs)
□ ICBF International Dairy and Beef v3 (53,262 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)

Figure 1: Screenshot of download options at SNPchimp website

```
#For Illumina 50Kv1 the native position do not match UMD3.1 at all,
#For the other two chips only a small  number of SNPs between Native and UMD3.1
both <- both[,N_snps:=.N,by=chip_name]
print('Table: Number of SNPs per chip (N_snps), and number of SNPs where the Native platform position di
```

```
## [1] "Table: Number of SNPs per chip (N_snps), and number of SNPs where the Native platform position d
```

```
both[pos.nat!=pos.umd,.(`N_diffpos`=.N),by=c('chip_name','N_snps')]
```

```
##        chip_name N_snps N_diffpos
## 1: Bov_Illu50Kv1  58276     58220
## 2: Bov_Illu50Kv2  58763       927
## 3:    Bov_IlluHD 781797      3392
```

Looking at the 53099 SNPs that are both on the 50Kv2 and 777K chips they all have the same position for the same assembly.

```
#Compare positions of SNPs with same name in Illumina 50Kv2 and 777K
#all SNP with same name have same position for same assembly
snpchimp <- rbind(cbind(nat,assembly='Native'),cbind(umd,assembly='UMD3.1'))
pos_per_name <- snpchimp[chip_name!='Bov_Illu50Kv1',.(positions=length(unique(position)),chips_with_snp=
pos_per_name[,.N,by=c('assembly','positions','chips_with_snp')]
```

```
##    assembly positions chips_with_snp      N
## 1:   Native         1              1 734362
## 2:   Native         1              2  53099
## 3:   UMD3.1         1              1 734362
## 4:   UMD3.1         1              2  53099
```

Looking at the differences in position we see that the main changes are for unplaced SNPs (chr 0 in the figures below). Many SNPs are unplaced by Illumina but placed on chromosomes by dbSNP and vice versa. A few SNP are also moved from one chromsome to another chromosome.

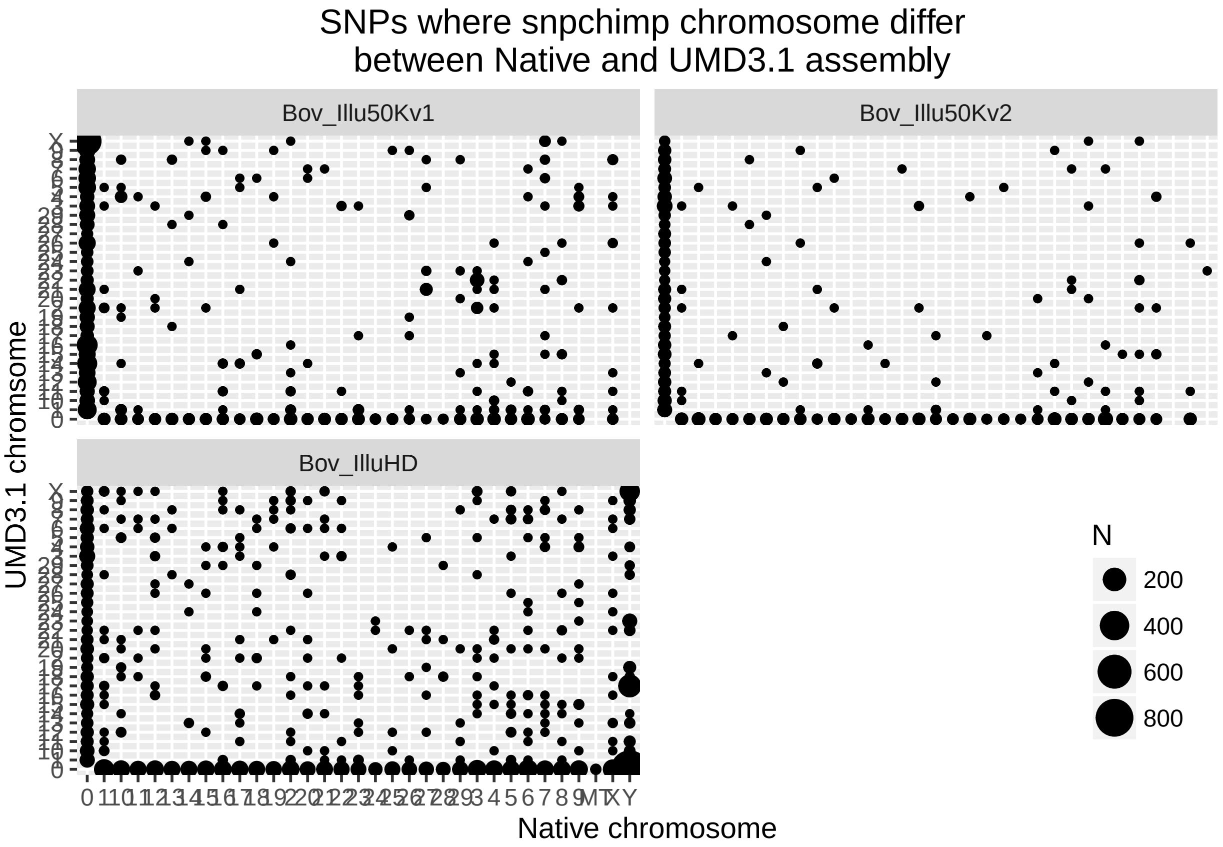
```
both[chr.nat!=chr.umd][,.N,by=chip_name]
```

```
##        chip_name    N
## 1: Bov_Illu50Kv1 2347
## 2: Bov_Illu50Kv2 1030
## 3:    Bov_IlluHD 5188
```

```
both[chr.nat==chr.umd&pos.nat!=pos.umd][,.N,by=chip_name]
```

```
##        chip_name     N
## 1: Bov_Illu50Kv1 55929
## 2: Bov_Illu50Kv2    82
## 3:    Bov_IlluHD   330
```

```
both[chr.umd=='99']$chr.umd<-'0'
p <- ggplot(both[chr.nat!=chr.umd,.N,by=c('chr.nat','chr.umd','chip_name')]) + geom_point(aes(x=chr.nat
p <- p + ggtitle('SNPs where snpchimp chromosome differ \n between Native and UMD3.1 assembly') + labs(x
p + facet_wrap(~chip_name,nrow=2) + theme(legend.justification=c(1,0), legend.position=c(1,0))
```



SNPs that might need flipping of alleles

The ANPchimp data also give an easy way to identify the SNPs that will be flipped if the wrong allel-coding is used during file conversion.

```
umd[,flip:=Alleles_A_B_FORWARD!=Alleles_A_B_TOP,by=SNP_name]
```

```
##          chip_name         rs Alleles_A_B_FORWARD Alleles_A_B_TOP
##      1: Bov_IlluHD rs17870340                 A/G             A/G
##      2: Bov_IlluHD rs17870417                 T/C             A/G
##      3: Bov_IlluHD rs17870546                 A/G             A/G
##      4: Bov_IlluHD rs17870550                 A/G             A/G
##      5: Bov_IlluHD rs17870946                 T/C             A/G
##     ---
## 898832: Bov_IlluHD       NULL                 T/G             A/C
## 898833: Bov_IlluHD       NULL                 T/G             A/C
## 898834: Bov_IlluHD       NULL                 T/C             A/G
## 898835: Bov_IlluHD       NULL                 T/C             A/G
## 898836: Bov_IlluHD       NULL                 A/G             A/G
##         chromosome  position               SNP_name  flip
##      1:          1  98367573     BovineHD4100000577 FALSE
##      2:          1  79326737     BovineHD4100000457  TRUE
##      3:          1 144579256     BovineHD0100041712 FALSE
##      4:          1 144587013     BovineHD4100000819 FALSE
##      5:          1 153282696     BovineHD0100044630  TRUE
##     ---
## 898832:         99         0  Hapmap38311-BTA-39536  TRUE
## 898833:         99         0 Hapmap39460-BTA-109014  TRUE
## 898834:         99         0          UA-IFASA-2402  TRUE
## 898835:         99         0          UA-IFASA-5520  TRUE
## 898836:         99         0          UA-IFASA-7534 FALSE
```

```
umd[,.N,by=c('chip_name','flip')]
```

```
##        chip_name  flip      N
## 1:    Bov_IlluHD FALSE 391761
## 2:    Bov_IlluHD  TRUE 390036
## 3: Bov_Illu50Kv1 FALSE  29291
## 4: Bov_Illu50Kv2 FALSE  29527
## 5: Bov_Illu50Kv1  TRUE  28985
## 6: Bov_Illu50Kv2  TRUE  29236
```

Summary / TODOs

- For consistent positions across all three chips we should use dbSNP positions.
- TODO: Check which positions have been used for converting Illumina files
- TODO: Compare Tims list of SNPs to be flipped with this data
- TODO: Compare Tims remapped positions with the dbSNP positions