Large scale SNP data management and integration in breeding programs

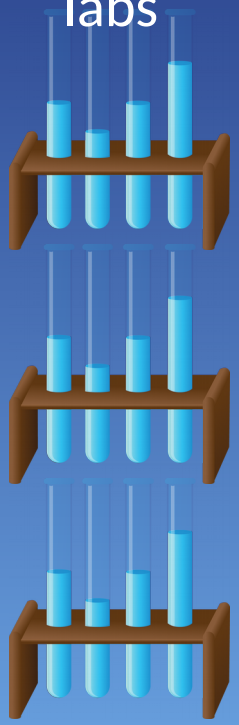
Eildert Groeneveld

Institute of Farm Animal Genetics

TheSNPpit: Who is it for?

- Needs to manage many SNPs
- Uses standardized workflows
- Uses pipelines
- Does some gBLUP

TheSNPpit



Genotyping labs

import

SNPpit database

Define Subsets:

Maf

Nocalls

Chrom

Ind

SNP

export

data use/analysis

-E gs_055 chicken maf.05 plink

-E gs_022 pigs genable

-E gs_123 beef GS run week12

-E gs_086 dairy HF gBLUP

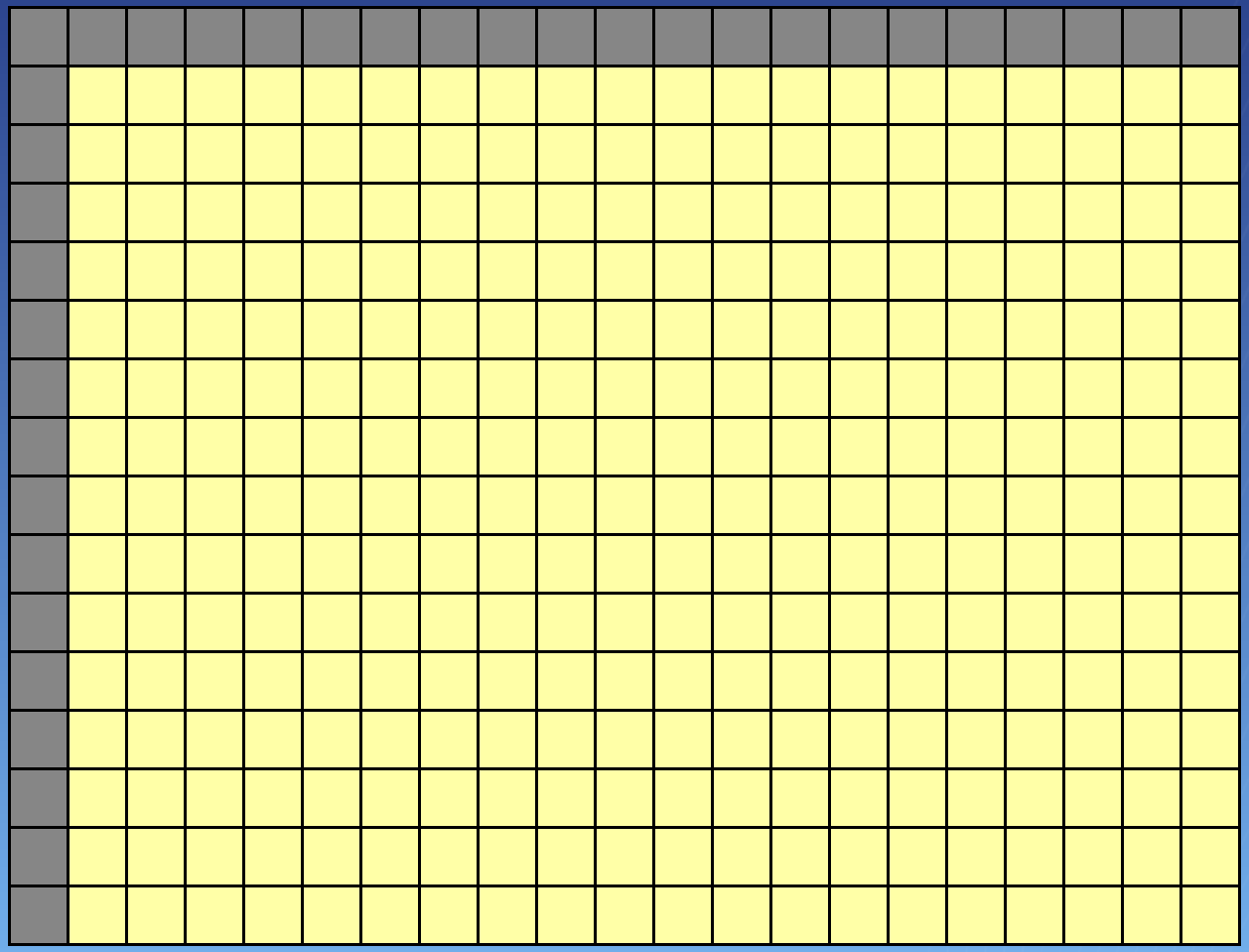
-E gs_009 dairy chrom 12 GWAS

-E gs_031 goat imp 54-> 850K

Genotype sets

SNP vector

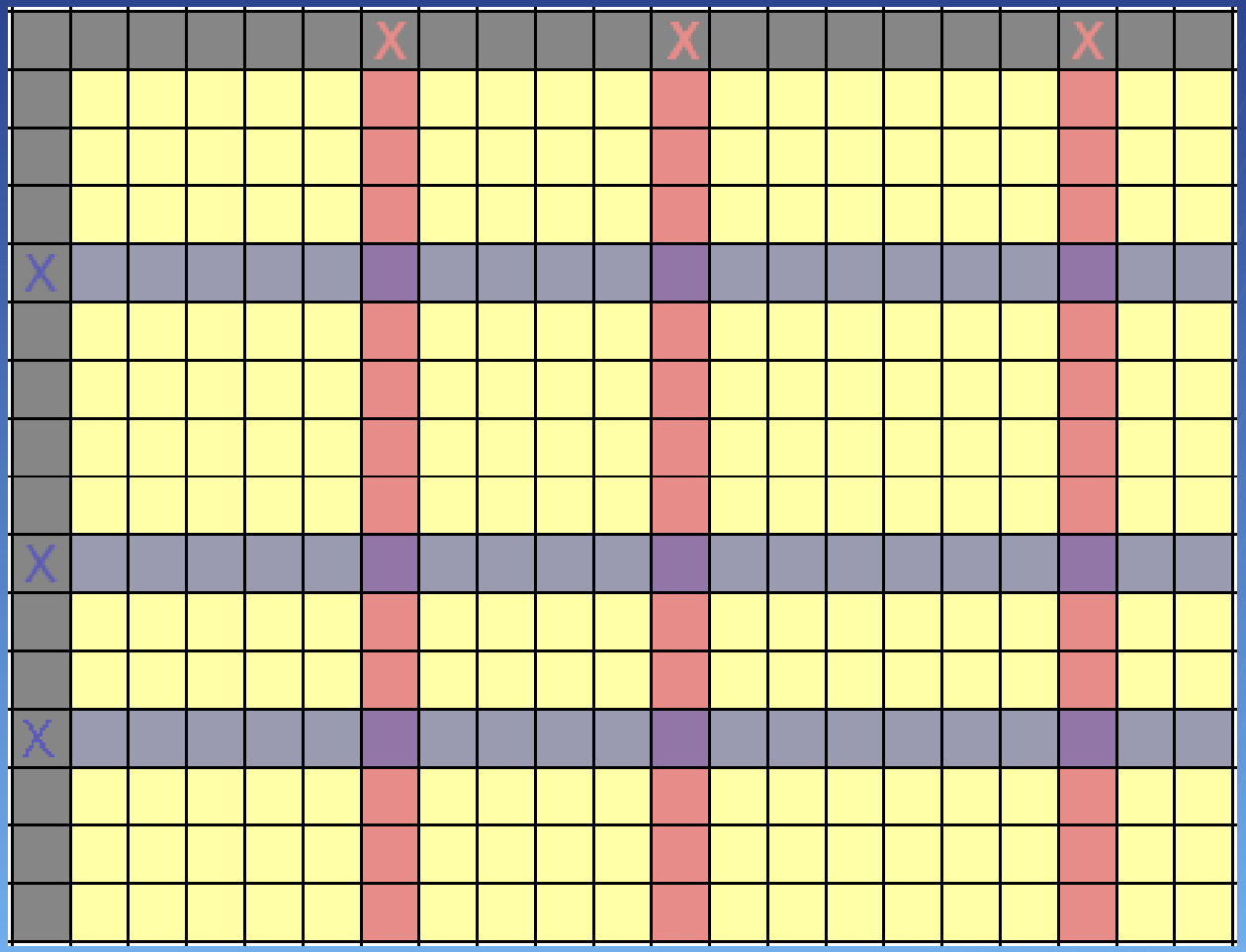
Sample vector



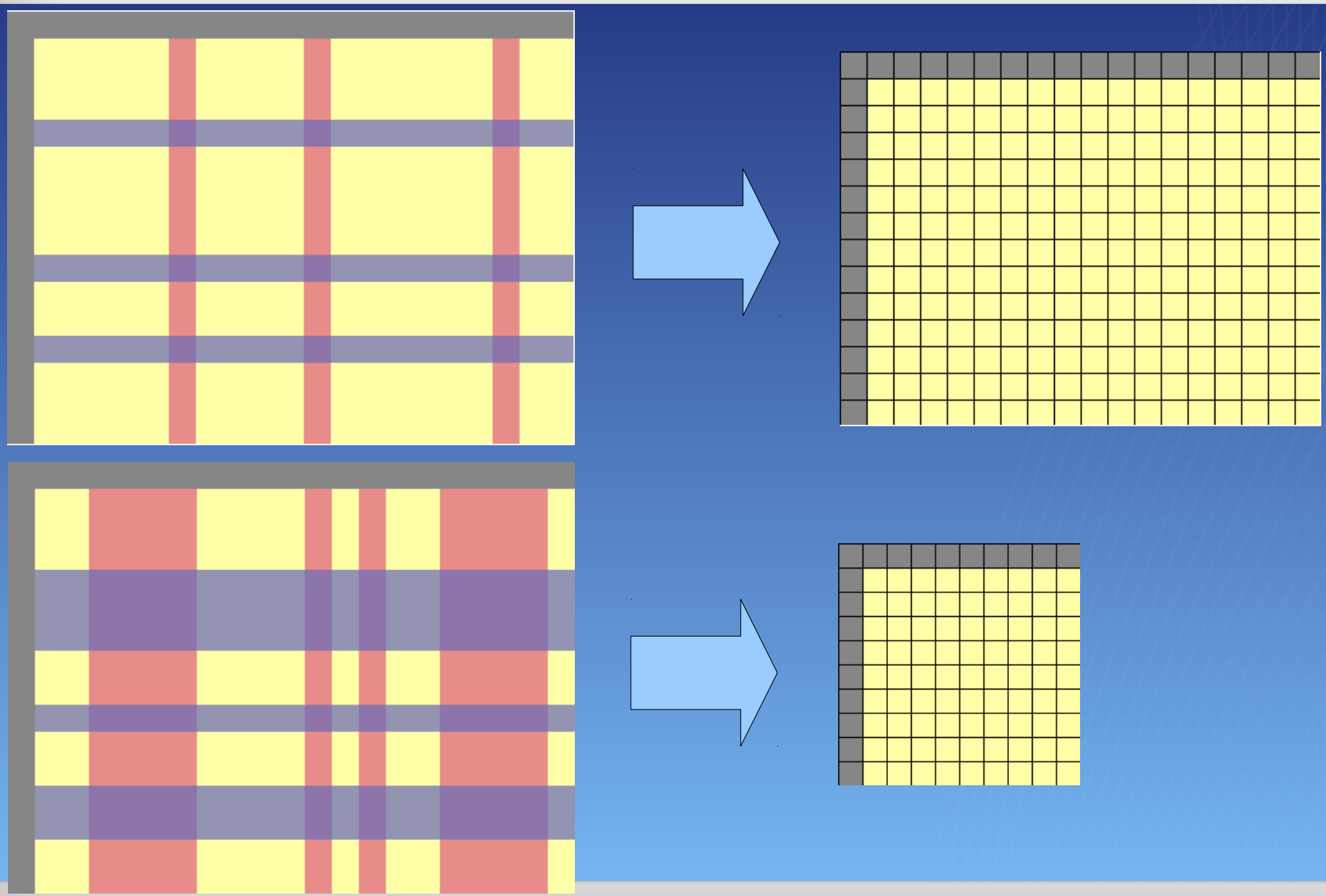
Genotype sets

SNP vector

Sample vector



Genotype sets



TheSNPpit command line

Map info for each chip

```
>snppit --import panel ctl_800K  800k.map

>snppit --import panel shp_57K   shp_57k.map
```

Done once

TheSNPpit command line

Import SNP data

- `snppit --import data ctl_800K  wk13.ped`
- `snppit --import data ctl_800K  wk14.ped`
- `snppit --import data shp_57K   wk-01.0125`

Export SNP data

- `snppit --export gs_012`
- `snppit --export gs_023`

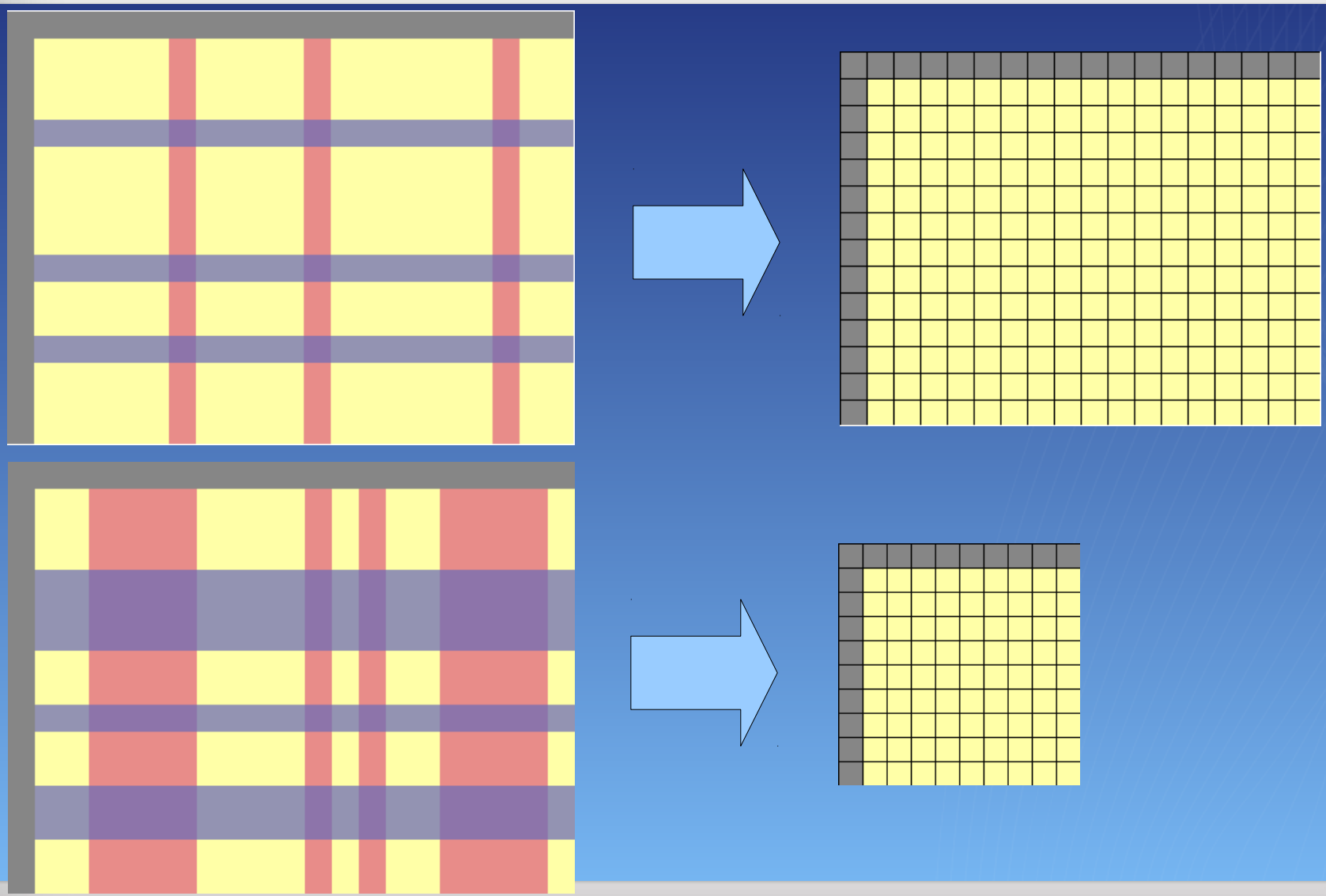
Breeding prg: workflow comp

- Phenotype database
- SNPpit database

Breeding prg: workflow

- Phenotype database
  - Extract pedigree
  - Extract phenotype data
  - Extract list of ID with SNP data
  - Extract list of SNPname/panel
- SNPpit database
  - Provide SNP data for ID/SNPs
- Run gBLUP

Genotype sets - reminder

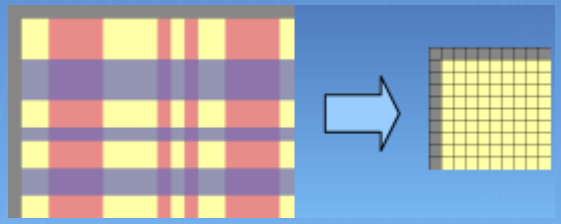


Breeding prg: workflow cc

- **Phenotype database**
  - On your table: `wk21_ANIs.txt, wk21_SNPs.txt, pheno.dat,pedi.dat`
- **SNPpit database**

Breeding prg: workflow cc

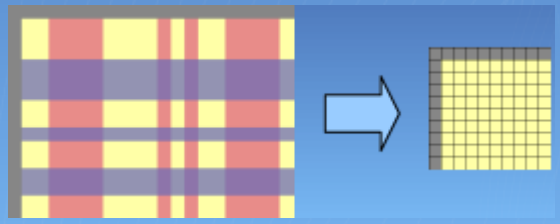
- Phenotype database
  - On your table: `wk21_ANIs.txt, wk21_SNPs.txt, pheno.dat,pedi.dat`
- SNPpit database



Breeding prg: workflow cc

- **Phenotype database**
  - On your table: `wk21_ANIs.txt, wk21_SNPs.txt, pheno.dat,pedi.dat`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_800 -i wk21_ANIs.txt > is_123
```

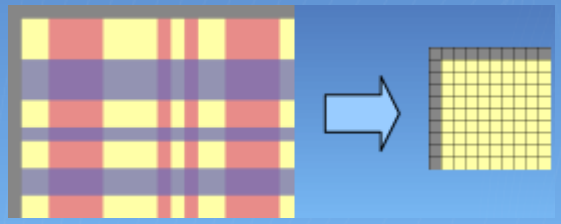


Breeding prg: workflow cc

- **Phenotype database**
  - On your table: `wk21_ANIs.txt, wk21_SNPs.txt, pheno.dat,pedi.dat`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_800 -i wk21_ANIs.txt > is_123

>snppit --create snp_sel -p ctl_800 -i wk21_SNPs.txt > ss_243
```



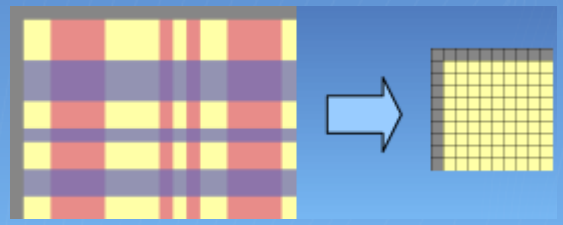
Breeding prg: workflow cc

- **Phenotype database**
  - On your table: `wk21_ANIs.txt, wk21_SNPs.txt, pheno.dat,pedi.dat`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_800 -i wk21_ANIs.txt > is_123

>snppit --create snp_sel -p ctl_800 -i wk21_SNPs.txt > ss_243

>snppit --create genotype_set is_123 ss_243 > gs_434
```



Breeding prg: workflow cc

- **Phenotype database**
  - On your table: `wk21_ANIs.txt, wk21_SNPs.txt, pheno.dat,pedi.dat`
- **SNPpit database**

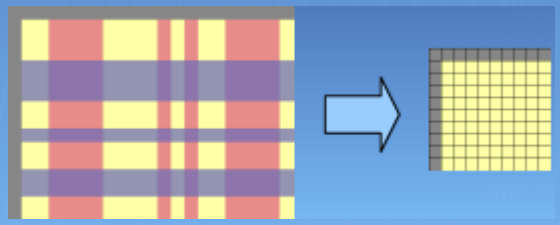
```
>snppit --create ind_sel -p ctl_800 -i wk21_ANIs.txt > is_123

>snppit --create snp_sel -p ctl_800 -i wk21_SNPs.txt > ss_243

>snppit --create genotype_set is_123 ss_243 > gs_434

>snppit --export gs_434 -o SNPs_wk21.0125

>gBLUP_prg pheno.dat pedi.dat SNPs_wk21.0125
```



Breeding prg: workflow script

Put the following in file run_newexport.sh :

```
#!/bin/bash
IS=$(snppit -q -C individual_selection -p 57K -i $1_ANIs.txt)
SS=$(snppit -q -C snp_selection        -p 57K -i $1_SNPs.txt)
GS=$(snppit -q -C genotype_set --snp $SS --individual $IS
echo '------ new genotype set: '$GS
snppit -E genotype_set --name $GS -f 0125 -o $1.0125
```

Run as:

```
>bash run_newexport.sh wk21
```

Breeding prg: workflow timings

- **Assumptions**
  - 100000 animals with SNPs > `wk21_ANIs.txt`
  - 40000 SNP from 57K panel > `wk21_SNPs.txt`

Breeding prg: workflow timings

- **Assumptions**
  - 100000 animals with SNPs > `wk21_ANIs.txt`
  - 40000 SNP from 57K panel > `wk21_SNPs.txt`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_57 -i wk21_ANIs.txt :    .57sec
```

Breeding prg: workflow timings

- **Assumptions**
  - 100000 animals with SNPs > `wk21_ANIs.txt`
  - 40000 SNP from 57K panel > `wk21_SNPs.txt`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_57 -i wk21_ANIs.txt :    .57sec

>snppit --create snp_sel -p ctl_57 -i wk21_SNPs.txt :    .56sec
```

Breeding prg: workflow timings

- **Assumptions**
  - 100000 animals with SNPs > `wk21_ANIs.txt`
  - 40000 SNP from 57K panel > `wk21_SNPs.txt`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_57 -i wk21_ANIs.txt :    .57sec

>snppit --create snp_sel -p ctl_57 -i wk21_SNPs.txt :    .56sec

>snppit --create genotype_set is_123 ss_243          :    .28sec
```

Breeding prg: workflow timings

- **Assumptions**
  - 100000 animals with SNPs > `wk21_ANIs.txt`
  - 40000 SNP from 57K panel > `wk21_SNPs.txt`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_57 -i wk21_ANIs.txt :    .57sec

>snppit --create snp_sel -p ctl_57 -i wk21_SNPs.txt :    .56sec

>snppit --create genotype_set is_123 ss_243          :    .28sec

>snppit --export gs_434 -o SNPs_wk21.0125            : 51.07sec
```

Breeding prg: workflow timings

- **Assumptions**
  - 100000 animals with SNPs > `wk21_ANIs.txt`
  - 40000 SNP from 57K panel > `wk21_SNPs.txt`
- **SNPpit database**

```
>snppit --create ind_sel -p ctl_57 -i wk21_ANIs.txt :    .57sec

>snppit --create snp_sel -p ctl_57 -i wk21_SNPs.txt :    .56sec

>snppit --create genotype_set is_123 ss_243          :    .28sec

>snppit --export gs_434 -o SNPs_wk21.0125            : 51.07sec

>gBLUP_prg pheno.dat pedi.dat SNPs_wk21.0125         : ?????sec
```

Scaling: Storage

2 bits/SNP - 4 SNP/byte
4096 mio SNP / 1 MB

List of SNP panels

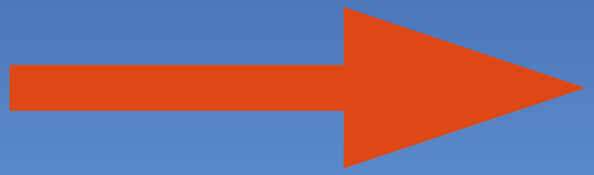
Panel	nSNP	nSample	SNP(mio)
P.01000K	1000000	1000800	1000800
P.00200K	200000	4525000	905000
P.00100K	100000	5533000	553300
P.00700K	700000	525000	367500
P.00054K	54000	5525899	298398
P.00500K	500000	526600	263300
P.00010K	10000	80000	800
P.20000K	20000000	40	800
P.00001K	1000	800000	800
P.05000K	5000000	160	800
P.10000K	10000000	80	800
P.03000K	3000000	264	792
P.056000	56000	10000	560
Total	-	18526843	3393650

Database size

Tables	total_size
public.genotype_data	840 GB
public.snp	5095 MB
public.individual_selection	3359 MB
public.individual	1023 MB
public.genotype_set	216 KB
public.snp_selection	144 KB
public.panel	48 KB
public.phenotype	16 KB

Scaling: Storage

2 bits/SNP - 4 SNP/byte
4096 mio SNP / 1 MB



List of SNP panels

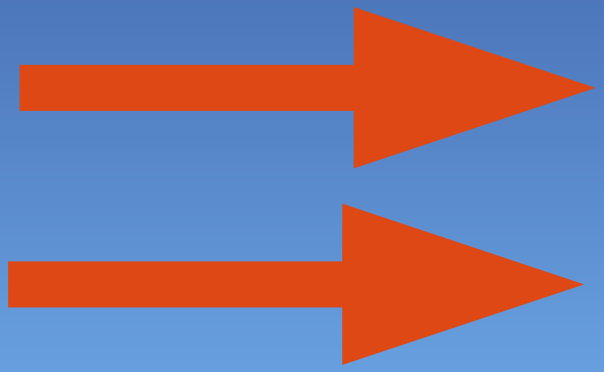
Panel	nSNP	nSample	SNP(mio)
P.01000K	1000000	1000800	1000800
P.00200K	200000	4525000	905000
P.00100K	100000	5533000	553300
P.00700K	700000	525000	367500
P.00054K	54000	5525899	298398
P.00500K	500000	526600	263300
P.00010K	10000	80000	800
P.20000K	20000000	40	800
P.00001K	1000	800000	800
P.05000K	5000000	160	800
P.10000K	10000000	80	800
P.03000K	3000000	264	792
P.056000	56000	10000	560
Total	-	18526843	3393650

Database size

Tables	total_size
public.genotype_data	840 GB
public.snp	5095 MB
public.individual_selection	3359 MB
public.individual	1023 MB
public.genotype_set	216 KB
public.snp_selection	144 KB
public.panel	48 KB
public.phenotype	16 KB

Scaling: Storage

2 bits/SNP - 4 SNP/byte
4096 mio SNP / 1 MB



List of SNP panels

Panel	nSNP	nSample	SNP(mio)
P.01000K	1000000	1000800	1000800
P.00200K	200000	4525000	905000
P.00100K	100000	5533000	553300
P.00700K	700000	525000	367500
P.00054K	54000	5525899	298398
P.00500K	500000	526600	263300
P.00010K	10000	80000	800
P.20000K	20000000	40	800
P.00001K	1000	800000	800
P.05000K	5000000	160	800
P.10000K	10000000	80	800
P.03000K	3000000	264	792
P.056000	56000	10000	560
Total	-	18526843	3393650

Database size

Tables	total_size
public.genotype_data	840 GB
public.snp	5095 MB
public.individual_selection	3359 MB
public.individual	1023 MB
public.genotype_set	216 KB
public.snp_selection	144 KB
public.panel	48 KB
public.phenotype	16 KB

Ressources

- SNPpit server
- Linux
- TheSNPpit software
- User Manual

Ressources

- SNPpit server
- Linux
- TheSNPpit software

Ressources

- SNPpit server
- Linux
- TheSNPpit software

- Open Source
- Open Source

PLOS | ONE

RESEARCH ARTICLE

TheSNPpit—A High Performance Database System for Managing Large Scale SNP Data

Eildert Groeneveld*, Helmut Lichtenberg

Institute of Farm Animal Genetics, Friedrich-Loeffler-Institute, 31535 Mariensee, Germany

* eildert.groeneveld@fli.de

Thank you for your attention