TheSNPpitGroup - managing large scale genotyping data in Open Source

E.Groeneveld [1] H.Schwarzenbacher [2] J. McEwan [3]
F. Seefried [4] M. Jafarikia [5] P. von Rohr [4] J. Jakobsen[6]

[1]Institute of Farm Animal Genetics (FLI), Höltystrasse 10, D-31535 Neustadt, Germany, [2]ZuchtData GmbH, Dresdner Straße 89/19, A-1200 Vienna, Austria, [3]AgResearch, Invermay, Mosgiel 9053, New Zealand, [4]Qualitas AG, Chamerstr 56, 6300 Zug, Switzerland, [5]CCSI, 960 Carling Ave, Ottawa, Canada, [6]NSG, Box 104, 1431 Ås, Norway,

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data
- handles millions of SNP records with millions of SNP

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data
- handles millions of SNP records with millions of SNP
- export around 100mio SNP/second

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data
- handles millions of SNP records with millions of SNP
- export around 100mio SNP/second
- backend data storage used in SNP pipelines, e.g. in gBLUP

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data
- handles millions of SNP records with millions of SNP
- export around 100mio SNP/second
- backend data storage used in SNP pipelines, e.g. in gBLUP
- across species – any biallelic system

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data
- handles millions of SNP records with millions of SNP
- export around 100mio SNP/second
- backend data storage used in SNP pipelines, e.g. in gBLUP
- across species – any biallelic system
- being used in breeding programs, and one stop genotype repos

What are we talking about

- TheSNPpit is an ultrafast database for large scale SNP data
- handles millions of SNP records with millions of SNP
- export around 100mio SNP/second
- backend data storage used in SNP pipelines, e.g. in gBLUP
- across species – any biallelic system
- being used in breeding programs, and one stop genotype repos
- released as Open Source

Interfacing with the DB: CLI, pipelining, API

command line commands:

```
eg(eno,~)snppit --import panel -p sheep_780K -i sheep_2014-12wk.map
eg(eno,~)snppit --append --panel --name sheep_780K -i sheep_2014-13wk.ped
eg(eno,~)snppit --append --panel --name sheep_780K -i sheep_2014-14wk.ped
eg(eno,~)snppit --append --panel --name sheep_780K -i sheep_2014-15wk.ped
```

Interfacing with the DB: CLI, pipelining, API

command line commands:

```
eg(eno,˜)snppit --import panel -p sheep_780K -i sheep_2014-12wk.map
eg(eno,˜)snppit --append --panel --name sheep_780K -i sheep_2014-13wk.ped
eg(eno,˜)snppit --append --panel --name sheep_780K -i sheep_2014-14wk.ped
eg(eno,˜)snppit --append --panel --name sheep_780K -i sheep_2014-15wk.ped
```

building a pipeline:

```
#!/bin/bash
SS = $(snppit -q -C snp_selection -x ss_019 -p amido -i Wsnp.txt -c ’W chrom’)
GS = $(snppit -q -C genotype_set -n $SS -s is_009 -c ’WChrom. & 388 samples’)
```

Interfacing with the DB: CLI, pipelining, API

command line commands:

```
eg(eno,~)snppit --import panel -p sheep_780K -i sheep_2014-12wk.map
eg(eno,~)snppit --append --panel --name sheep_780K -i sheep_2014-13wk.ped
eg(eno,~)snppit --append --panel --name sheep_780K -i sheep_2014-14wk.ped
eg(eno,~)snppit --append --panel --name sheep_780K -i sheep_2014-15wk.ped
```

building a pipeline:

```
#!/bin/bash
SS = $(snppit -q -C snp_selection -x ss_019 -p amido -i Wsnp.txt -c ’W chrom’)
GS = $(snppit -q -C genotype_set -n $SS -s is_009 -c ’WChrom. & 388 samples’)
```

direct database access through API:

```
my $genotype = SNPcontrib_1::snp_per_sample($sample,$panel,$snp);
```

TheSNPpit space requirements

```
            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
```

It's Open Source, so what needs consideration?

- TheSNPpit is a critical component in workflow
- components: PostgreSQL, Perl, C, bash, git, installation, documentation
- database access: CLI, bash/* pipeline, direct access PERL/Python

It’s Open Source, so what needs consideration?

- TheSNPpit is a critical component in workflow
- components: PostgreSQL, Perl, C, bash, git, installation, documentation
- database access: CLI, bash/* pipeline, direct access PERL/Python

resulting issues:

- software/system maintenance
- creating new releases
- further development
- support / advice

What needs consideration / promises

- similar problems everywhere: pigs, cattle, tomatoes ...
- $\Longrightarrow$ same programs/scripts everywhere
- e.g. pedigree check
- ...

What needs consideration / promises

- similar problems everywhere: pigs, cattle, tomatoes ...
- $\Longrightarrow$ same programs/scripts everywhere
- e.g. pedigree check
- ...

issues:

What needs consideration / promises

- similar problems everywhere: pigs, cattle, tomatoes ...
- $\Longrightarrow$ same programs/scripts everywhere
- e.g. pedigree check
- ...

issues:

- infrastructure to support contributed software
- private new development

What needs consideration / promises

- similar problems everywhere: pigs, cattle, tomatoes ...
- $\Longrightarrow$ same programs/scripts everywhere
- e.g. pedigree check
- ...

issues:

- infrastructure to support contributed software
- private new development
- further development
- GBS under consideration

TheSNPpitGroup

TheSNPpitGroup setup during 2nd workshop

The Objectives of the Group

1. It serves as a platform for the maintenance of the Open Source TheSNPpit software.
2. It coordinates bug fixes and further development of the core software as well as contributed modules.
3. It creates and maintains an infrastructure for mutual support and fosters discussion of ideas around TheSNPpit.

Organizational structure

Coordinator just that

Users run TheSNPpit software on the basis of the SNPpit release tar balls. They want to be included in discussions around TheSNPpit.

Organizational structure

Coordinator just that

Users run TheSNPpit software on the basis of the SNPpit release tar balls. They want to be included in discussions around TheSNPpit.

Developers are active in coding of all aspects around the TheSNPpit system, the regression tests, creation of releases, installaton procedures and documentaton.

Organizational structure

Coordinator just that

Users run TheSNPpit software on the basis of the SNPpit release tar balls. They want to be included in discussions around TheSNPpit.

Developers are active in coding of all aspects around the TheSNPpit system, the regression tests, creation of releases, installaton procedures and documentaton.

Contributers support the objectives of TheSNPpitGroup through in-kind donations, like providing and running the GIT server and the WEB infrastructure and any other support.

Organizational structure

Coordinator just that

Users run TheSNPpit software on the basis of the SNPpit release tar balls. They want to be included in discussions around TheSNPpit.

Developers are active in coding of all aspects around the TheSNPpit system, the regression tests, creation of releases, installaton procedures and documentaton.

Contributers support the objectives of TheSNPpitGroup through in-kind donations, like providing and running the GIT server and the WEB infrastructure and any other support.

There is no free lunch not even in Open Source

TheSNPpit Group web site `https://thesnppit.net/`

- development area itemize
- development environment
- regression test
- tools for creation of new releases
- public area
  - core software download
  - contributed library
  - BLOG
  - user communication

https://thesnppit.net

Thank you for your attention!