SNP data management in genomics pipelines

Eildert Groeneveld
(eildert.groeneveld@fli.de)

EAAP Belfast September 2016

Advances in Genotyping

- Rapid development of genotyping
- Decreasing costs
- Increasing panel sizes
- Massive increase in data volume

Data volume issues

- Increasing numbers
- Multiple panels
- Different sizes:
- 6K, 54K, 80K, 200K, 700K, 5000K
- From whole genome sequences:
- 20mio SNPs

Data volume issues (CDCB)

Genotype counts by Chip Type, Breed Code, and Sex Code in database as of 2016-07-24

	AY F	AY M	BS F	BS M	GU F	GU M	HO F	HO M	JE F	JE M	MS F	MS M	XX F	XX M	Total
50K V	0	20	91	5652	0	0	20757	37583	913	4938	0	0	0	0	69954
50K V2	157	405	135	10786	18	178	32728	40140	885	3079	2	0	5	0	88518
3K	3	0	473	11	5	0	49024	3912	9678	195	0	0	0	0	63301
HD	12	520	3	182	26	121	1009	2011	31	212	0	0	0	0	4127
AFFY	0	0	0	0	0	0	0	19	0	0	0	0	0	0	19
LD	969	14	529	162	0	0	157307	5356	10459	211	0	1	1	0	175009
GGP	56	5	482	288	0	4	40341	13424	12513	1509	5	1	1	0	68629
GHD	447	429	175	578	330	251	16052	13306	972	1330	2	2	0	0	33874
GGP2	358	81	622	1211	161	8	58444	24908	17128	3180	0	8	2	0	106111
ZLD	0	0	84	20	0	0	106860	1143	10667	151	0	0	3	0	118928
ZMD	566	1	2	0	0	0	2708	588	2	21	0	0	0	0	3888
ELD	0	0	0	0	0	0	712	98	5	2	0	0	0	0	817
LD2	0	0	16	21	0	0	13163	2328	1996	15	0	0	0	0	17539
GP3	832	128	1189	2513	1301	2	87097	35744	28838	5306	0	2	1	0	162953
ZL2	3	0	420	23	0	0	313711	7839	23320	385	3	0	26	0	345730
ZM2	0	3	2	5	0	0	9513	874	86	129	0	0	0	0	10612
GH2	29	176	1	592	2	133	3931	4858	126	512	0	15	0	0	10375
G7K	1	0	19	0	2	0	14286	177	7064	6	1	0	1	0	21557
GP4	754	25	543	373	285	19	21131	10133	9172	1810	0	1	2	0	44248
ZL4	87	5	235	25	0	0	120256	3616	6380	110	0	0	1	0	130715
AMD	0	0	0	0	0	0	452	6	9	0	0	0	0	0	467
Total	4274	1812	5021	22442	2130	716	1069482	208063	140244	23101	13	30	43	0	1477371

Data management solution: database

- One pot for all data
- Not 100s of files
- Uniform access
- Handles inputs/outputs
- But: no downstream data processing

Pipelines

- Based on lab files
- Cumulative delivery
- Subset definition:
- animals>chromo>maf>nc>SNPsel.
- Chain of programs: Plink>R>GS..

What do we need to handle?

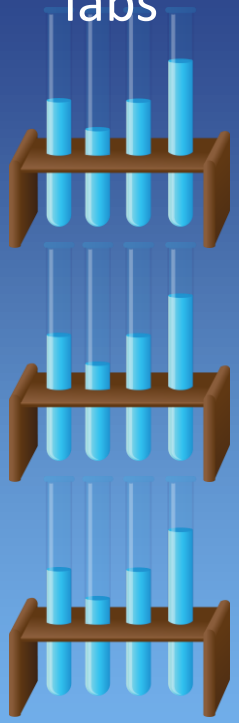
- Assume 50000 animals
- 700000 SNPs
- Plink ped file size: 140GB
- Exporting 35 billion SNPs:
- DB 1: ~116 hours
- DB 2: ~30 hours

Will not fly!!

- And we can do better

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 GWAS

-E gs_009 dairy chrom 12 GWAS

-E gs_031 goat imp 54-> 850K

Import map and SNP data

snppit --import panel -p chk-57K -i picken.map

snppit --append chk-57K -i picken-wk32.ped

snppit --append chk-57K -i picken-wk33.ped

snppit --import panel -p ctl-770K -i aff770K.map

snppit --append ctl-770K -i HF-2016-wk35.0125

snppit --append chk-57K -i picken-wk35.ped

Genotype sets

List of SNP panels

Panel	nSNP	nSample	SNP(mio)	Timestamp
Chk-57K	57543	4321	249	2015-10-31 11:44:58
Ctl-770K	770432	12022	9261	2015-10-31 11:27:35

List of Genotype Sets

SetName	IndSel	SNPSel	Comments
gs_004	is_002	ss_002	initial load from picken-wk32.ped
gs_005	is_005	ss_003	initial load from HF-2016-wk35.0125

Create subsets

snppit --subset gs_002 --ncsnp .03
snppit --subset gs_006 --maf .01
Snppit --subset gs_007 –chromosom -X w

List of Genotype Sets

SetName	IndSel	SNPSel	Comments
gs_002	is_002	ss_002	initial load from picken-wk32.ped
gs_006	is_006	ss_003	NCsnp: 37593/36357 0.030
gs_007	is_006	ss_004	MAF: 57636/38873 0.010
gs_008	is_002	ss_005	Chroms: all but W with 54000/54000 SNP

Export genotype set

List of Genotype Sets

```
SetName|IndSel|SNPSel|          Comments
-------|------|------|--------------------------------
gs_002 |is_002|ss_002|initial load from picken-wk32.ped
gs_006 |is_006|ss_003|NCsnp: 37593/36357 0.030
gs_007 |is_006|ss_004|MAF: 57636/38873 0.010
gs_008 |is_0026ss_005|Chroms: all but W with 54000/54000 SNP
```

```
snppit --export gs_002 -f plink
snppit --export gs_006 -f bped
snppit -E gs_008 -f 0125
```

What do we need to handle

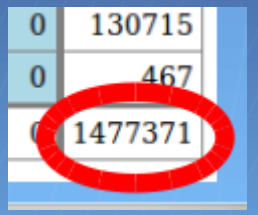
- Assume 50000 animals
- 700000 SNPs
- Plink ped file size: 140GB
- Exporting 35 billion SNPs:
- DB 1: ~ 116 hours
- DB 2: ~ 30 hours
- TheSNPpit: ~500 seconds

Export–analyze–delete cycle

- Subset definition – 'no' space in DB
- Avoid file clutter through fast exports: export analyze delete

scaling

- 18 mio samples 1K -- 20000K
- Totalling 3.4 trillion SNPs
- Stored in 840GB
- On a 4 year old laptop



Availability of TheSNPpit

- Open source
- Runs on Linux
- 70 page user guide
- If interested: get in touch

Acknowledgements

- Helmut Lichtenberg
- Troung Van Chi Cong

Thank you for your attention