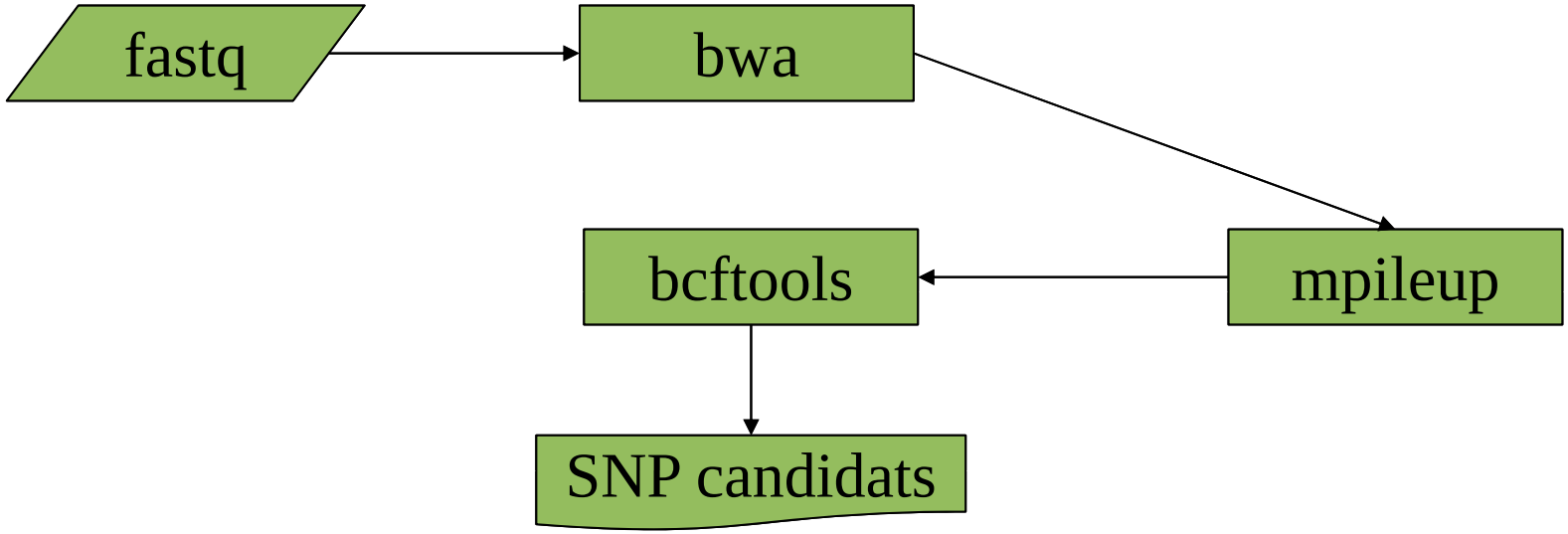
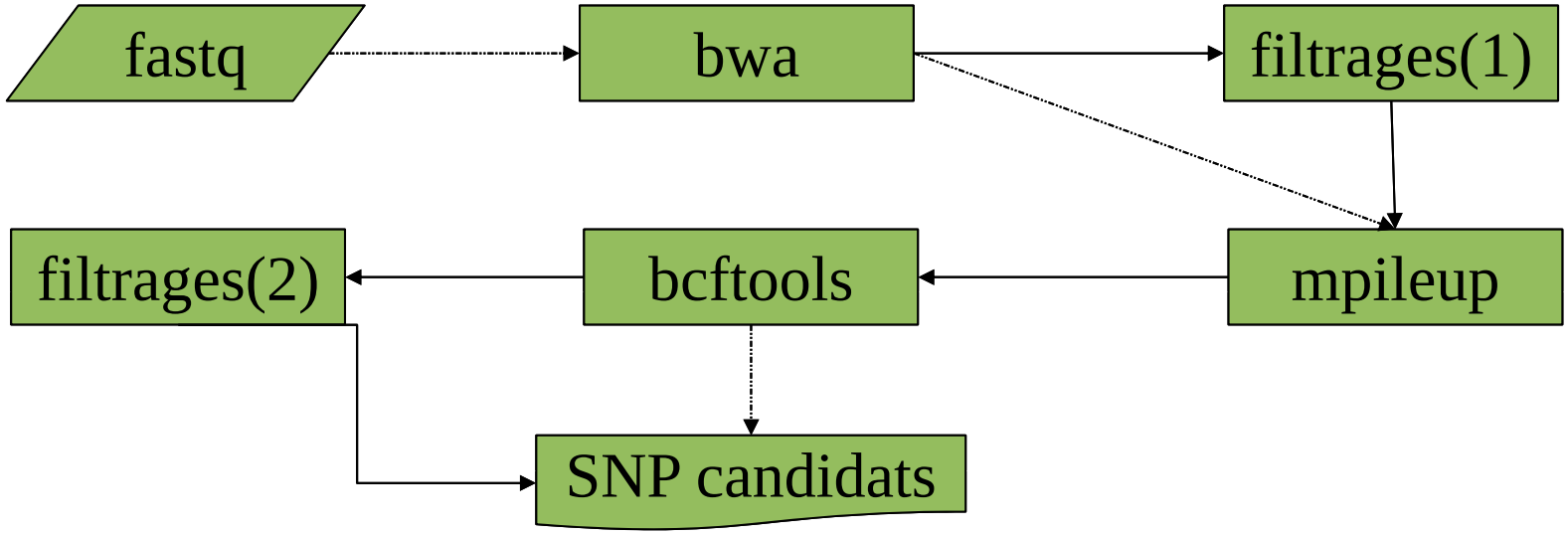
Détection de SNP

- Méthodologie, Données & outils
- Alignement sur génome de référence
- Détection de SNP
- Sélection de SNP

Méthodologie



Méthodologie



Données volumineuses

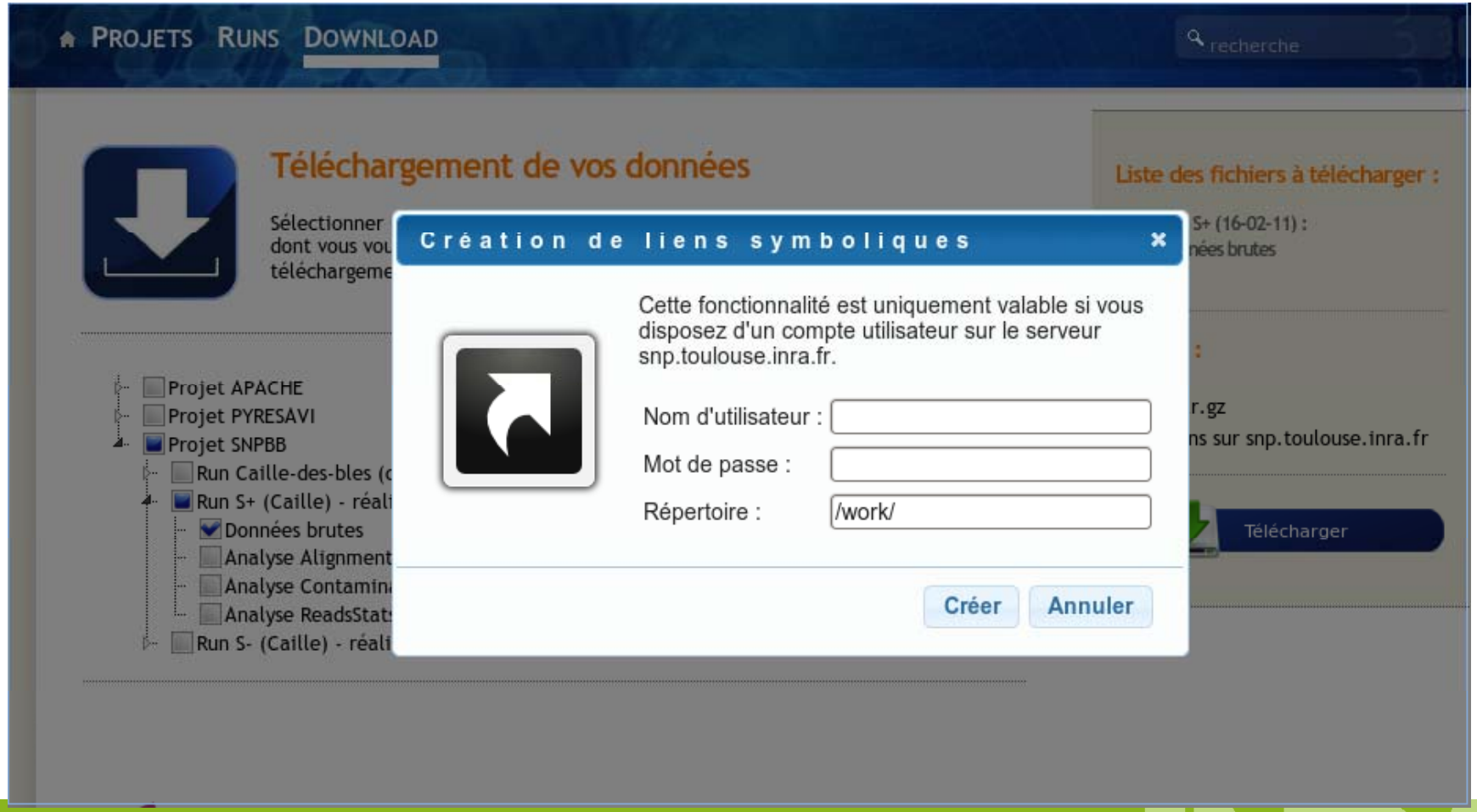
Projet SNPBB :

Recherche de polymorphisme chez la caille, QTL de comportement
3 run(s) et 0 analyse(s) ont été réalisées sur le projet SNPBB.
L'ensemble des données brutes et des résultats d'analyses occupent 47.37 Gb d'espace disque pour l'ensemble du projet.

Projet SWANPORC :

Etude de maladies congenitales chez le porc
11 run(s) et 0 analyse(s) ont été réalisées sur le projet SWANPORC.
L'ensemble des données brutes et des résultats d'analyses occupent 123.16 Gb d'espace disque pour l'ensemble du projet.

Portail NG6


PROJETS
RUNS
DOWNLOAD
recherche
Téléchargement de vos données
Sélectionner
dont vous vou
téléchargeme
Liste des fichiers à télécharger :
S+ (16-02-11) :
nées brutes
Projet APACHE
Projet PYRESAVI
Projet SNPBB
Run Caille-des-bles (c
Run S+ (Caille) - réali
Données brutes
Analyse Alignment
Analyse Contamina
Analyse ReadsStat
Run S- (Caille) - réali
r.gz
ns sur snp.toulouse.inra.fr
Télécharger
Création de liens symboliques
Cette fonctionnalité est uniquement valable si vous disposez d'un compte utilisateur sur le serveur snp.toulouse.inra.fr.
Nom d'utilisateur :
Mot de passe :
Répertoire :
/work/
Créer
Annuler

Données sur snp.toulouse.inra.fr

- Emplacement des bam

  http://ng6.toulouse.inra.fr/fileadmin/data/run/...

  /save/ng6/data/run/...

- Emplacement des génomes de référence

  /save/ng6/TODO/HiSeqIndexedGenomes/ :

  arabidopsis canard cheval chevre mouton porc poule taureau vigne

Pré-traitements NG6 : bwa aln données brutes

Swanporc

Maladie	HS				IS	
Animal	29MQ1081403		44UJ5062010		35HJ3070331	49UAZ080365
Date run	02-02-11	01-03-11	11-05-11	12-07-11	02-02-11	16-02-11
Raw mapped	85478425	150560510	75843154	184420256	54857356	158205686
% Raw mapped	80.28%	77.04%	73.83%	75.95%	69.04%	74.65%
Raw well paired	0	145974614	0	177695072	52216680	153124668
% Raw well paired	0.00%	74.70%	0.00%	73.18%	65.72%	72.25%
Nb Seq	106480566	195422034	102721977	242826148	79457908	211933260

SNPbb

Animal	S+		S-		caille des blés	
	read1	read2	read1	read2	read1	read2
Date run	16-02-11		16-02-11		16-02-11	
Raw mapped	26545363		31636999		27847370	
% Raw mapped	15.13%		13.71%		14.64%	
Raw well paired	25405990		30442382		26836332	
% Raw well paired	14.48%		13.19%		14.11%	
Nb Seq	87730296	87730296	115402866	115402866	95079542	95079542

Choix de l'algorithme (SNPbb)

- BWA aln : 14.5% lectures alignées
- Bowtie : 13.6%
- BWA bwasw : 42%
- Glint : 70%
- Blat, Blast, Fasta36 ... ?

Qualité d'un alignement

- Bonne vieille école : e value, score, % id, nb hits

Qualité d'un alignement

- Bonne vieille école : e value, score, % id, nb hits

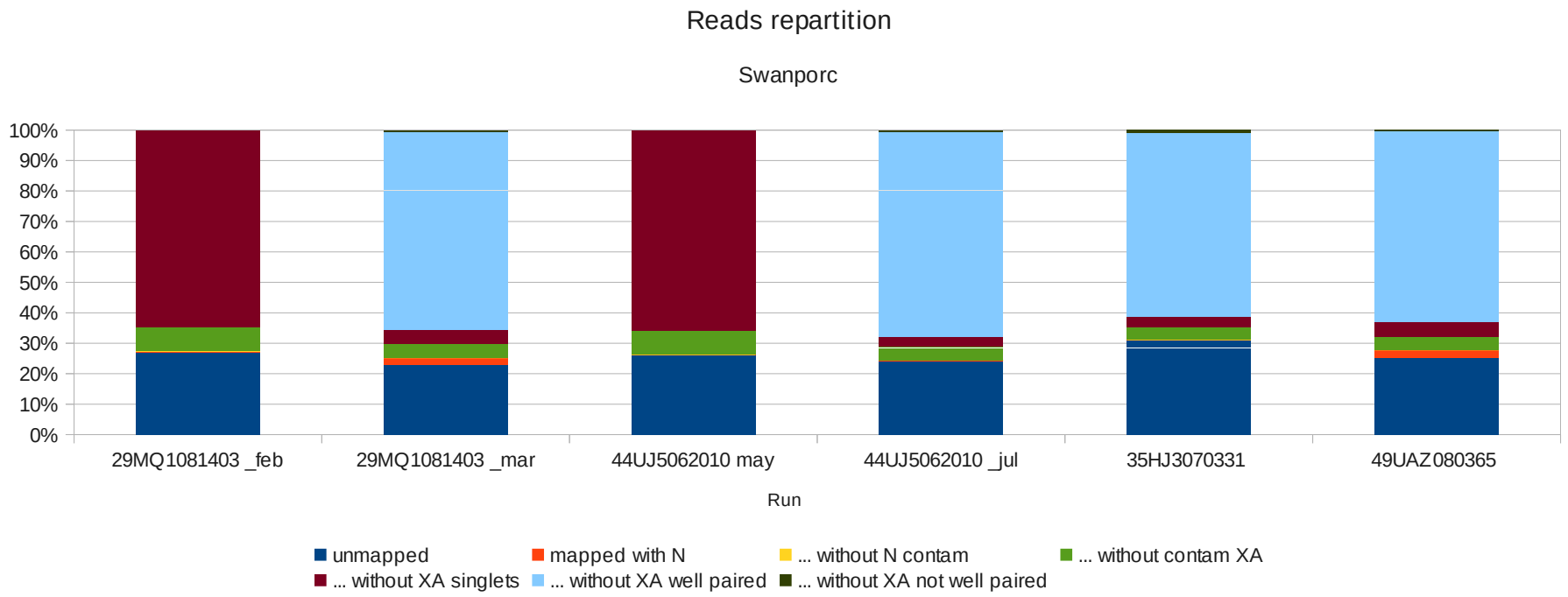
```
HWI-S...681#0   163     CHRMT   153     60      101M    =201    149     CTAT...CGT      ggg...BBB       \
XT:A:U  NM:i:0  SM:i:23 AM:i:23 X0:i:1  X1:i:1  XM:i:0  X0:i:0  XG:i:0  MD:Z:101        XA:Z:CHR7,-90650540,101M,1;

HWI-S...498#0   163     CHRMT   154     29      101M    =222    169     TATA...GTA      ggg...Tc]       \
XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:2  X1:i:0  XM:i:2  X0:i:0  XG:i:0  MD:Z:15G2G82    XA:Z:CHRMT,+154,101M,2;
```

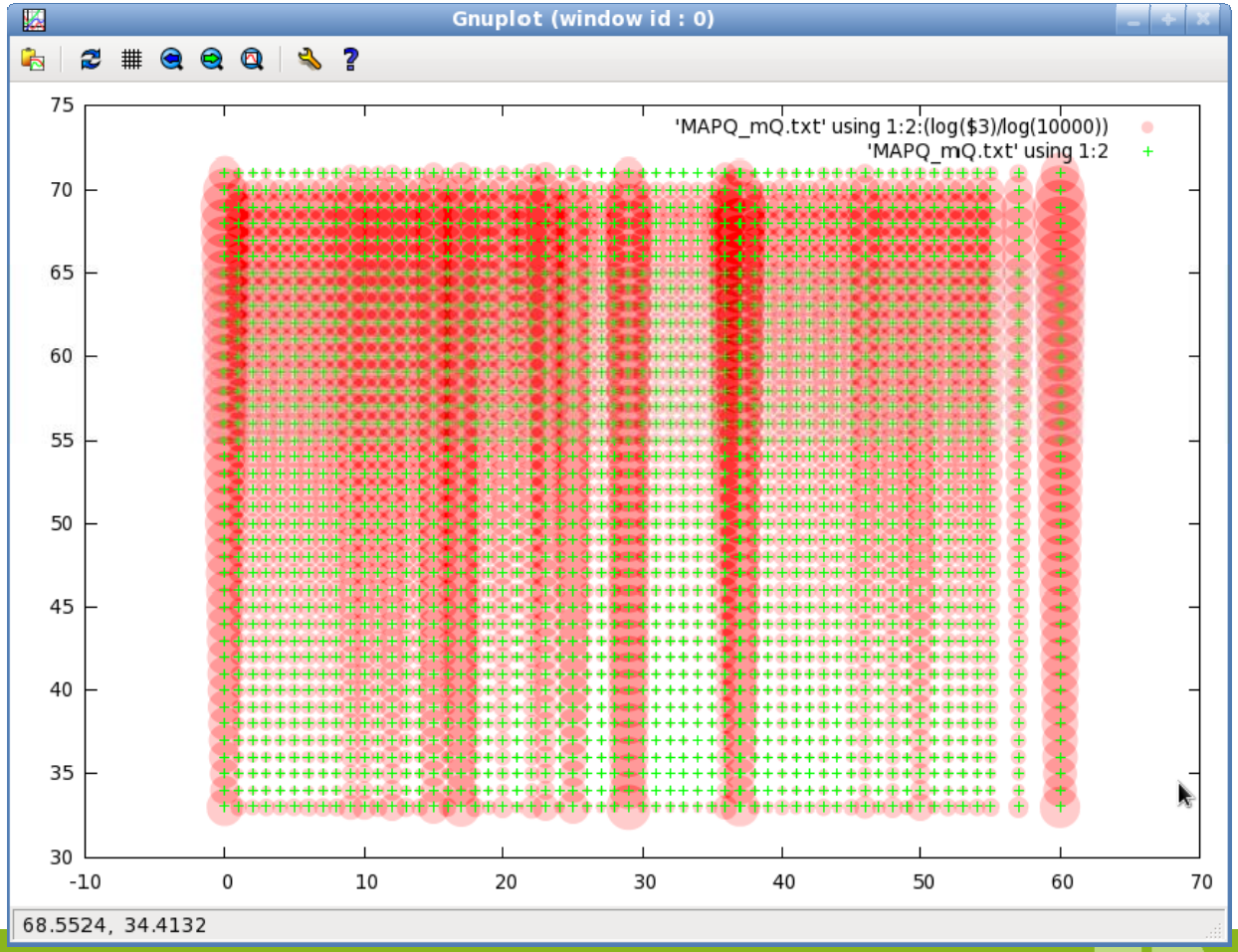
- MAPQ
- CIGAR
- NM(distance) MD(mismatches)
- X? ex : XA

XA

Step 8 XA	9474839	16938343	9242914	20725798	6513547	17398024
Step 8 maxIDXA	8306146	8707190	8122316	10956219	3364183	9181211

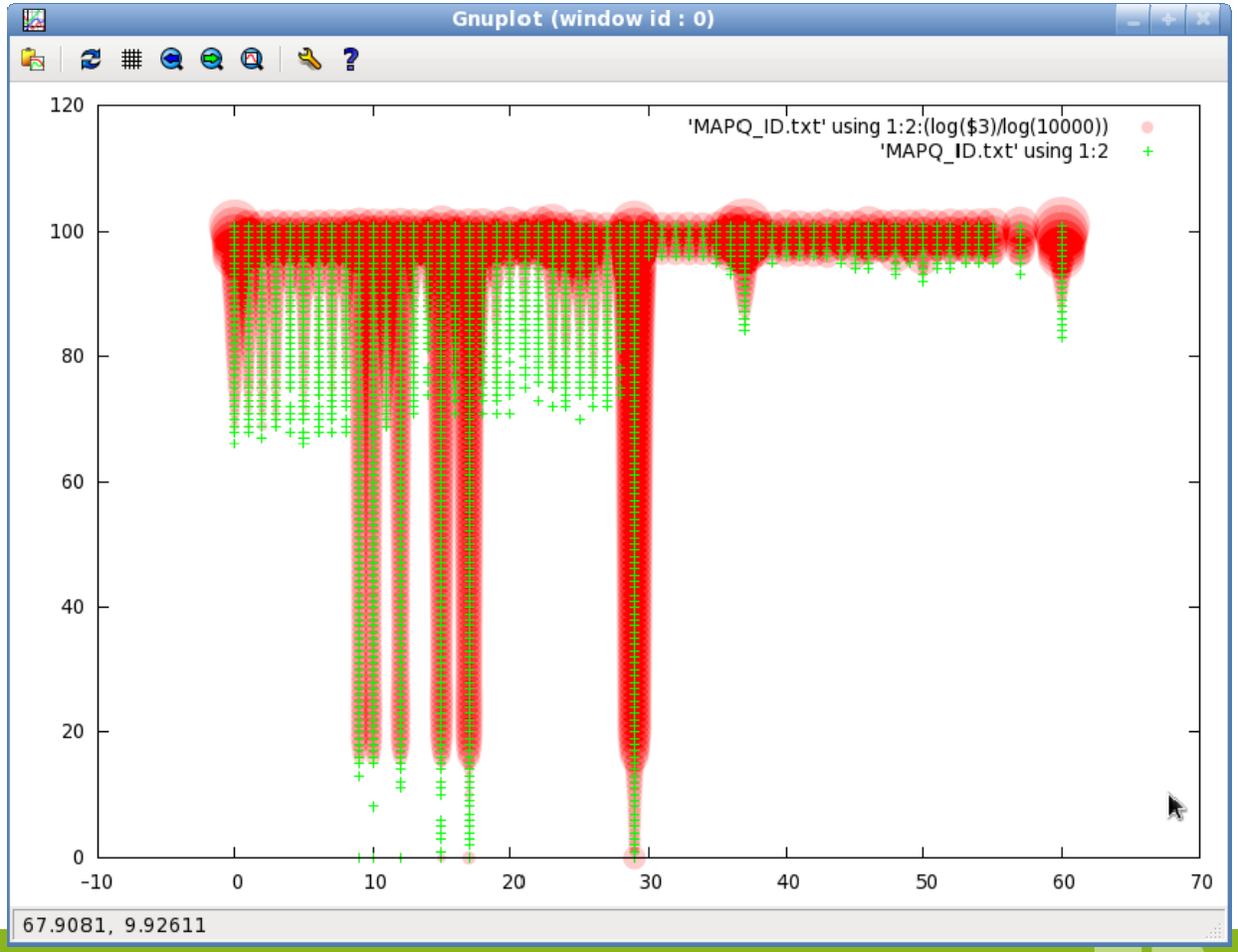


MAPQ / mean qual (swanporc)



MAPQ / ID

(swanporc)



Bwa aln bugs ?

- Séquences bien mappées ... dans le vide

```
HWI-ST...964#0        163     CHR3    50772519        29      25S25M51S       =       50772562        144     ATA...CGA        cWd...BBB
XT:A:N  NM:i:25 XN:i:25 SM:i:29 AM:i:29 XM:i:6  XO:i:0  XG:i:0  MD:Z:N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N  XI:i:0

HWI-ST...387#0       99      CHR2    90615073        29      37S13M1I7M43S   =       90615097        125     TTT...CAT        ggg...ggg
XT:A:N  NM:i:21 XN:i:20 SM:i:29 AM:i:29 XM:i:3  XO:i:1  XG:i:1  MD:Z:N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N0N    XI:i:0
```

- Séquences bien appairées … chr différents

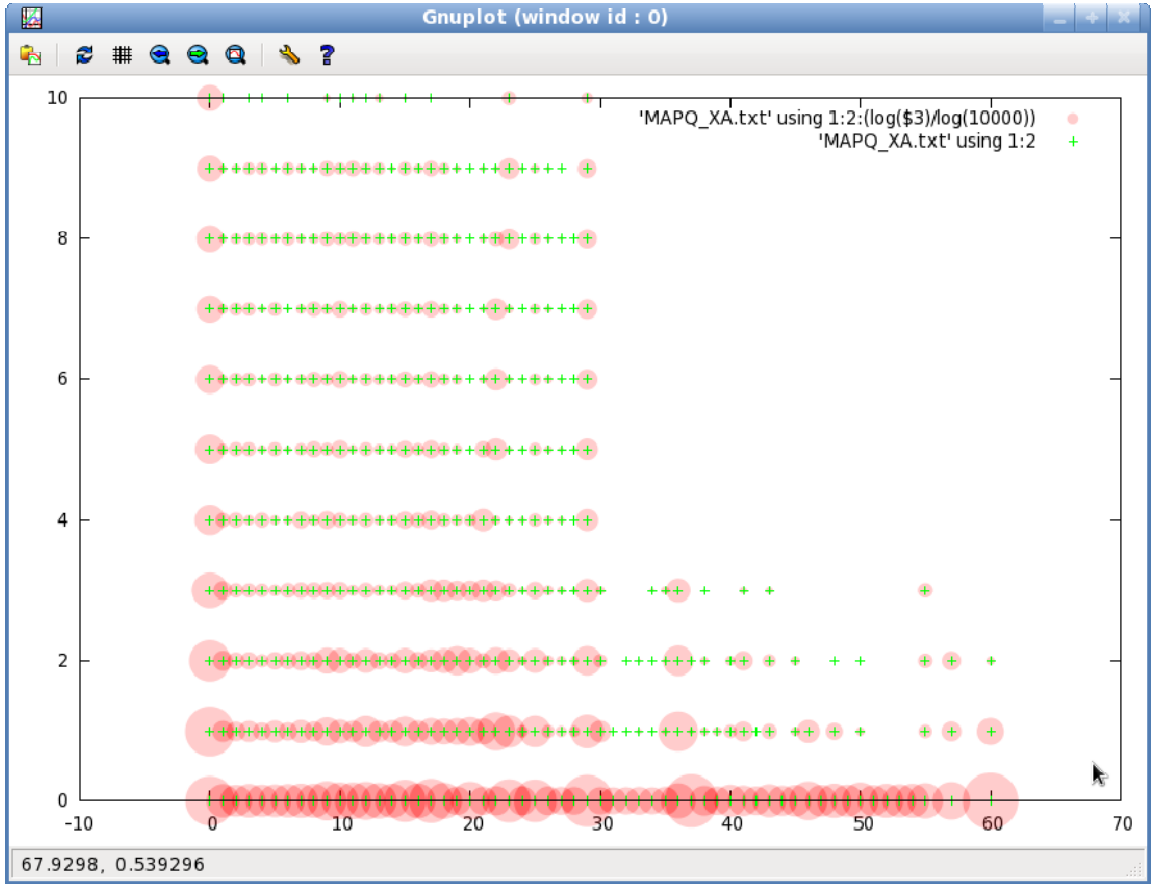
```
HWI-ST314_0079:5:1:2552:137885#0      99      CHR12_random    6959    29      55S11M2D12M23S CHR18_random    19      ...
HWI-ST314_0079:5:1:2552:137885#0      147     CHR18_random    19      29      101M    CHR12_random    6959    ...
HWI-ST314_0079:5:66:17594:200750#0    163     CHR13_random    35634   17      101M    CHRE64  50      ...
HWI-ST314_0079:5:66:17594:200750#0    83      CHRE64  50      17      60S5M1I21M14S   CHR13_random    35634   ...
```

- Flags étranges

flag	read paired	read unmapped	read reverse strand	mate reverse strand	fragment
117	oui	**oui**	**oui**	oui	premier
181	oui	**oui**	**oui**	oui	second

MAPQ / #XA

(swanporc)



MAPQ / #XA

- Paramétrage du nombre de XA

```
Usage:   bwa sampe [options] <prefix> <in1.sai> <in2.sai> <in1.fq> <in2.fq>
Options: -a INT   maximum insert size [500]
         -o INT   maximum occurrences for one end [100000]
         -n INT   maximum hits to output for paired reads [3]
```

Manual page bwa :

-n INT Maximum number of alignments to output in the XA tag for reads paired properly. **If a read has more than INT hits, the XA tag will not be written**. [3]

n	aln	-f 2	XA
1	9926	9926	928
2	9926	9926	1259
3	9926	9926	1272
4	9926	9926	1276

Filtrage (1)

- Sans contaminations
- Sans N
- MAPQ >= 30
- Localisation unique ?
- %id ?
- Séquences bien appairées ?

Détection de SNP

- samtools mpileup -uID -q (-B)

```
-q INT      filter out alignment with MQ smaller than INT [0]
-u          do not compress BCF output
-B          disable BAQ computation
-D          output per-sample DP
-I          do not perform indel calling
```

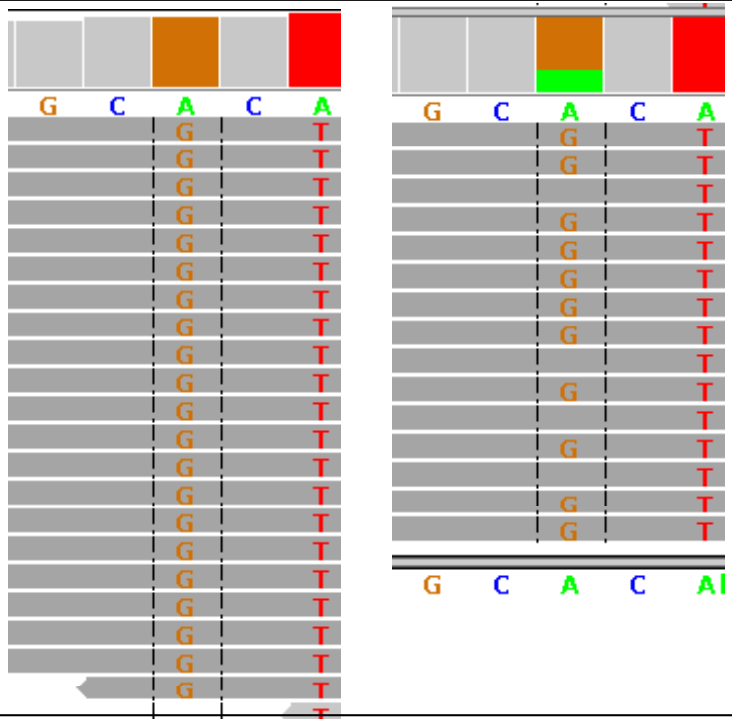
- bcftools view -cbvg

```
Options: -c        SNP calling
         -v        output potential variant sites only (force -c)
         -g        call genotypes at variant sites (force -c)
         -b        output BCF instead of VCF
```

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  S-_well-paired_90id_noXA_sorted.bam     S+_well-
paired_90id_noXA_sorted.bam
CHR10   5738    .       G       A       24.8    .       DP=6;AF1=1;CI95=0.5,1;DP4=0,0,3,1;MQ=17;FQ=-33.3  GT:PL:DP:GQ
1/1:42,9,0:3:52 1/1:17,3,0:1:46
```

mpileup

```
Bam complet :
CHR21    5909579    .    A    G    36.4    .    DP=36;AF1=0.5156;CI95=0.25,0.75;DP4=2,2,0,5;MQ=49;FQ=31;PV4=0.17,1,1,0.0097 …
GT:PL:DP:GQ    1/1:76,15,0:5:13    0/0:0,12,68:4:9
```



```
CHR21 5909579 A       21     GgGggGgGgGgGggGGggggg
CHR21 5909579 A       15     gGGg.GgG.g,G,GG
```

```
Bam sur zone restreinte ou option -B sur mpileup :
CHR21  5909579      .      A      G      181    .      DP=36;AF1=0.7487;CI95=0.75,0.75;DP4=2,2,2,7;MQ=48;FQ=20.3;PV4=0.53,1,1,0.0074
… GT:PL:DP:GQ       1/1:179,21,0:7:23    0/1:40,0,52:6:46
```

Sélection des SNP

- Allèle minoritaire
- Fréquences alléliques
- Profondeur minimale ou maximale
- Qualité du SNP
- Spécificité
- Répétition/environnement
- Conséquence
- Score illumina

Sélection des SNP

- Allèle minoritaire => comptage à partir du format pileup
- Fréquences alléliques => déduit du comptage
- Profondeur minimale ou maximale => samtools.pl varFilter
- Qualité du SNP => samtools.pl varFilter
- Spécificité => génotype ou calculs sur fréquences alléliques ?
- Répétition/environnement => filtrage en base de données
- Conséquence => Ensembl variant_effect_predictor.pl
- Score illumina => service web

```
CHR21    5909579    .    A    G    36.4    .    DP=36;AF1=0.5156;CI95=0.25,0.75;DP4=2,2,0,5;MQ=49;FQ=31;PV4=0.17,1,1,0.0097 …
GT:PL:DP:GQ    1/1:76,15,0:5:13    0/0:0,12,68:4:9
```

```
CHR21 5909579 A       21     GgGggGgGgGgGggGGggggg
CHR21 5909579 A       15     gGGg.GgG.g,G,GG
```

Conclusion

- Les outils sont « neufs », tout comme la techno
- Ils sont nombreux ; parfois hermétiques
- Les données en entrée sont grosses
  => difficile de tâtonner sans perte de temps