SNPs and Small INDELs Calling From NGS Data

李莉

2013.3

工作内容

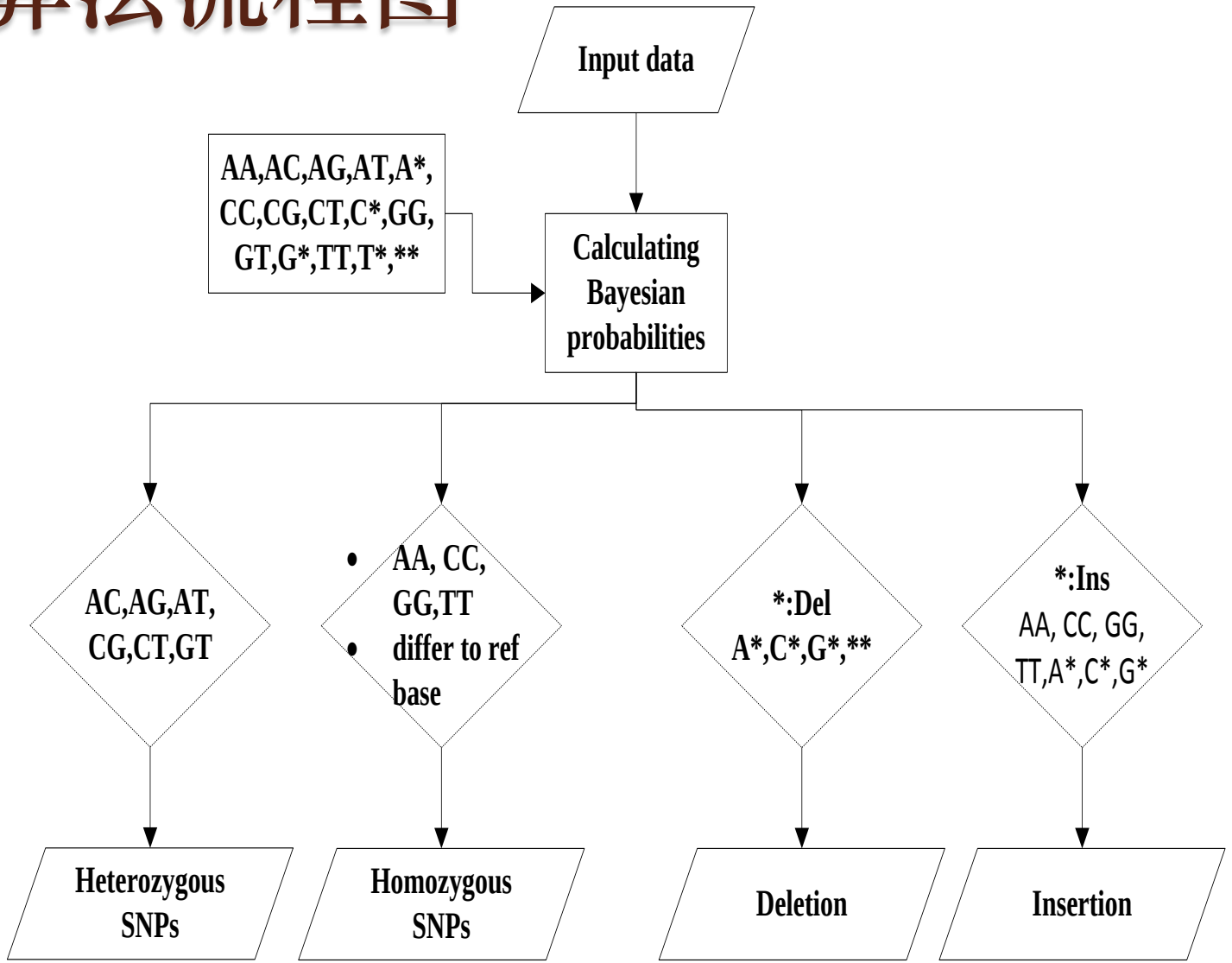
- 基于NGS数据的genomic variation检测(算法及流程)
  - SNP以及small indels(<10bp)的检测
  - large indels检测
  - 其他structural variations检测：inversion、tandem duplication、intra-chromosomal translocation、inter-chromosomal translocation

- SNPs常用检测软件：GATK、SAMtools
- INDELs常用检测软件：
  - Small indels检测：GATK，Dindel，Samtools
  - Large indels检测：Delly(del)，Breakdancer
  - Both：Pindel，SOAPindel

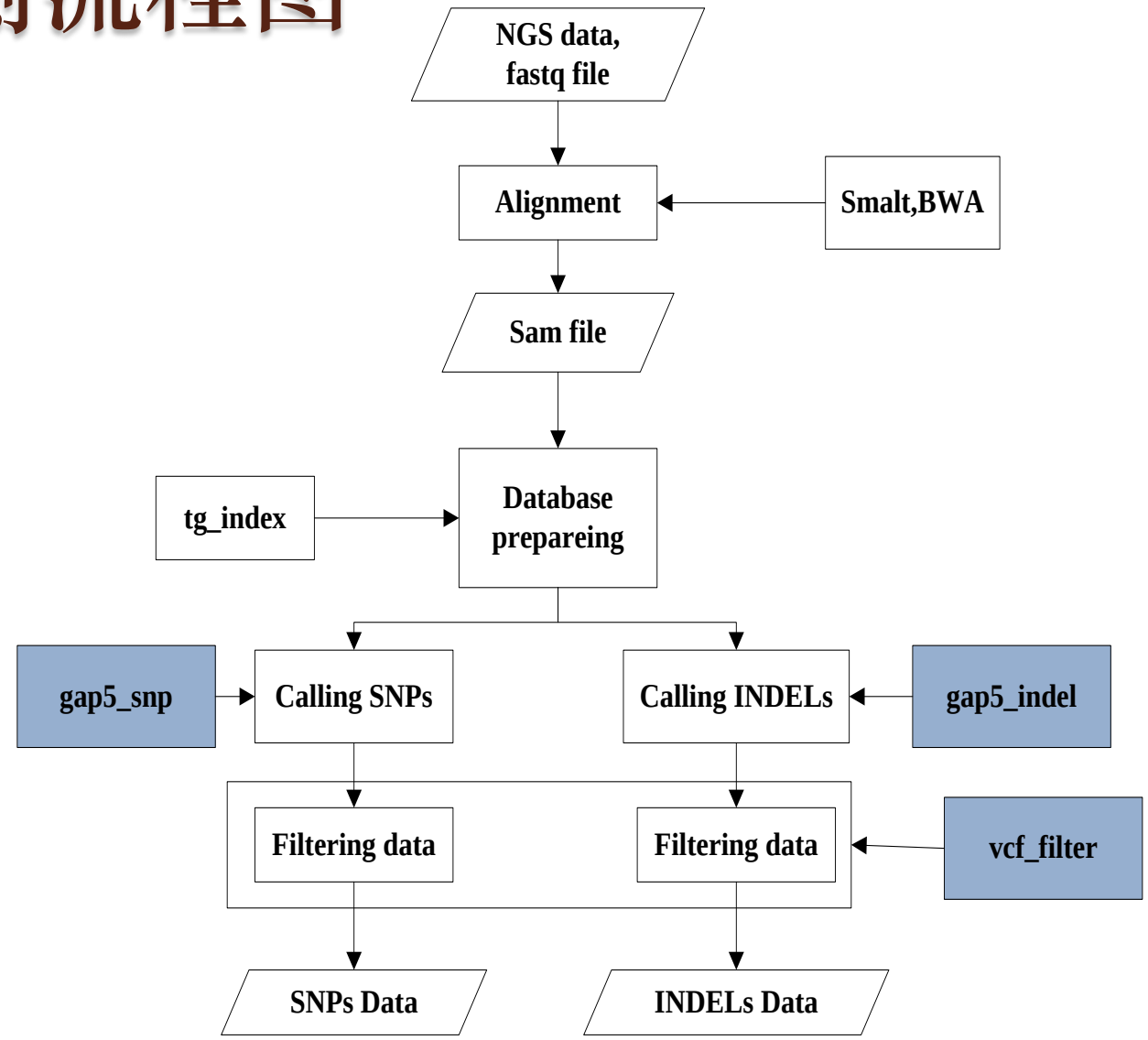
基于NGS数据的SNPs和small indels检测算法流程

- 目的：尽可能检测出存在的SNPs和small indels
- 算法： 基于Bayesian公式

算法流程图


Input data
AA,AC,AG,AT,A*,
CC,CG,CT,C*,GG,
GT,G*,TT,T*,**
Calculating
Bayesian
probabilities
AC,AG,AT,
CG,CT,GT
AA, CC,
GG,TT
differ to ref
base
*:Del
A*,C*,G*,**
*:Ins
AA, CC, GG,
TT,A*,C*,G*
Heterozygous
SNPs
Homozygous
SNPs
Deletion
Insertion

检测流程图


NGS data,
fastq file
Alignment
Smalt,BWA
Sam file
tg_index
Database
prepareing
gap5_snp
Calling SNPs
Calling INDELs
gap5_indel
Filtering data
Filtering data
vcf_filter
SNPs Data
INDELs Data

输出结果例图

```
##fileformat=VCFv4.1
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total number of reads depth">
##INFO=<ID=HP,Number=1,Type=Integer,Description="Reference homopolymer tract length">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles :
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for ger
n">
##contig=chromosome04
##reference=chr4_ref.fasta
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO  FORMAT  VALUE
chromosome04  8 . G C 176 PASS  DP=40;HRun=0;MQ=54.88 GT:AD:GQ:PL 1/1:0,40:99.00:1780,182,0
chromosome04  26  . A G 220 PASS  DP=55;HRun=0;MQ=54.96 GT:AD:GQ:PL 1/1:0,55:99.00:2171,225,0
chromosome04  34  . T C 204 PASS  DP=64;HRun=0;MQ=55.10 GT:AD:GQ:PL 1/1:1,63:99.00:2171,204,0
chromosome04  162 . C T 208 PASS  DP=80;HRun=0;MQ=54.55 GT:AD:GQ:PL 1/1:0,71:99.00:2171,264,0
chromosome04  170 . A G 285 PASS  DP=80;HRun=0;MQ=54.36 GT:AD:GQ:PL 1/1:0,80:99.00:3076,290,0
chromosome04  176 . G A 275 PASS  DP=78;HRun=1;MQ=54.79 GT:AD:GQ:PL 1/1:0,77:99.00:2171,282,0
chromosome04  184 . A G 276 PASS  DP=78;HRun=0;MQ=55.18 GT:AD:GQ:PL 1/1:0,78:99.00:2171,282,0
chromosome04  192 . C T 255 PASS  DP=70;HRun=1;MQ=55.66 GT:AD:GQ:PL 1/1:0,70:99.00:2171,260,0
```

结果比较(与GATK)

- SNPs calling

  Test data：GLA4, chromsome04,
  100bp, coverage of 70x；

  Total number of SNPs reference data：140726

  SNPs calling tools：GAP5 and GATK

Case	Total number	TP	FP
GAP5	323588	118495	205093
GATK	351910	120820	231090

结果比较(与GATK)

- INDELs calling

  Test data：GLA4, chromsome04, 100bp, coverage of 70x；

  Total number of indels reference data：19231 (<=10bp)

  Indels calling tools：GAP5 and GATK

  Insertion: the addition of one or more bases into the reference sequence, on the contrary is deletion.

Case	Total number	TP	FP
GAP5	39356	17219	22137
GATK	36417	17927	18490

存在问题

- 算法对杂合的敏感度较高，导致SNPs检测假阳性或者将纯合SNP检测为杂合SNP。
- 不适合平均覆盖度低(低于5)的检测。
- 假阳性偏高

过滤方法

- 目的：平衡准确性与检出率，在减少假阳性的同时，保持较高检出率。

  - Raw data (reads) filtering
    - Alignment quality
    - Mapping quality(MQ)
  - Coverage depth(DP)
  - Calling score(QUAL)

DP对结果的影响

GAP5

DP	total	TP	FP
0-10	1955	444	1511
10-20	3014	955	2059
20-30	3514	1470	2044
30-40	4127	1995	2132
40-50	4919	2722	2197
50-60	5708	3447	2261
60-70	4408	2926	1482
70-80	1950	1314	636
80-90	548	308	240
90-100	152	72	80
100-110	79	26	53
110-120	50	20	30
120-130	52	22	30
130-	274	84	190
total	30750	15805	14945

GATK

DP	total	TP	FP
0-10	857	58	799
10-20	2247	318	1929
20-30	2791	643	2148
30-40	3561	1199	2362
40-50	5024	2104	2920
50-60	6924	3656	3268
60-70	7240	4307	2933
70-80	4493	2841	1652
80-90	1754	1067	687
90-100	587	282	305
100-110	294	112	182
110-120	200	67	133
120-130	159	58	101
130-	723	226	497
total	36854	16938	19916

QUAL对结果的影响

GAP5

QUAL	total	TP	FP
0-1	2049	1140	909
1-10	4234	2312	1922
10-50	6298	3278	3020
50-100	6291	2818	3473
100-150	5005	2450	2555
150-200	3262	1724	1538
200-250	2165	1245	920
250-300	755	465	290
300-350	254	141	113
350-400	136	80	56
400-450	74	47	27
450-500	57	32	25
500-	170	73	97
total	30750	15805	14945

GATK

QUAL	total	TP	FP
0-500	3982	441	3541
500-1000	4621	1073	3548
1000-1500	5930	2353	3577
1500-2000	7556	3905	3651
2000-2500	6480	3771	2709
2500-3000	4090	2575	1515
3000-3500	1984	1320	664
3500-4000	1016	710	306
4000-4500	451	317	134
4500-5000	256	181	75
5000-	488	292	196
total	36854	16938	19916

QUAL/DP对结果的影响

GAP5

QUAL/DP	total	TP	FP
0-2	18659	10493	8166
>=2	15233	7117	8116
>=3	12091	5312	6779
>=4	8782	3441	5341
>=5	4721	1580	3141
>=6	2790	817	1973
>=7	1935	488	1447
>=8	1399	329	1070
>=9	1038	227	811
>=10	860	193	667
>=15	0	0	0

GATK

QUAL/DP	total	TP	FP
0-2	143	13	130
2-10	1977	257	1720
>=10	34691	16665	18026
>=20	31022	15734	15288
>=30	21380	11019	10361
>=40	9573	5222	4351
>=50	3800	2170	1630
>=60	1881	1128	753
>=70	874	542	332
>=80	482	322	160
>=90	271	188	83
>=100	142	104	38

- 总结： GATK在这两个过滤条件上均表现出一定的规律，即低阈值得到的FP会大于TP。GAP5这方面表现出异常，有待修正。

不同过滤结果比较（GATK）

- SNPs

Filter	total	TP	FP	TP:FP
--	351910	120820	231090	0.5
Mismatch<=2	165778	101828	63950	1.6
Mismatch<=1	100379	75565	24814	3.0
MQ>=60	88711	72697	16014	4.5
Cluster filter	87557	72616	14941	4.9
QUAL/DP>=20	82333	70300	12033	5. 8
QUAL >=1500	20016	18917	1099	17. 2

不同过滤结果比较(GATK)

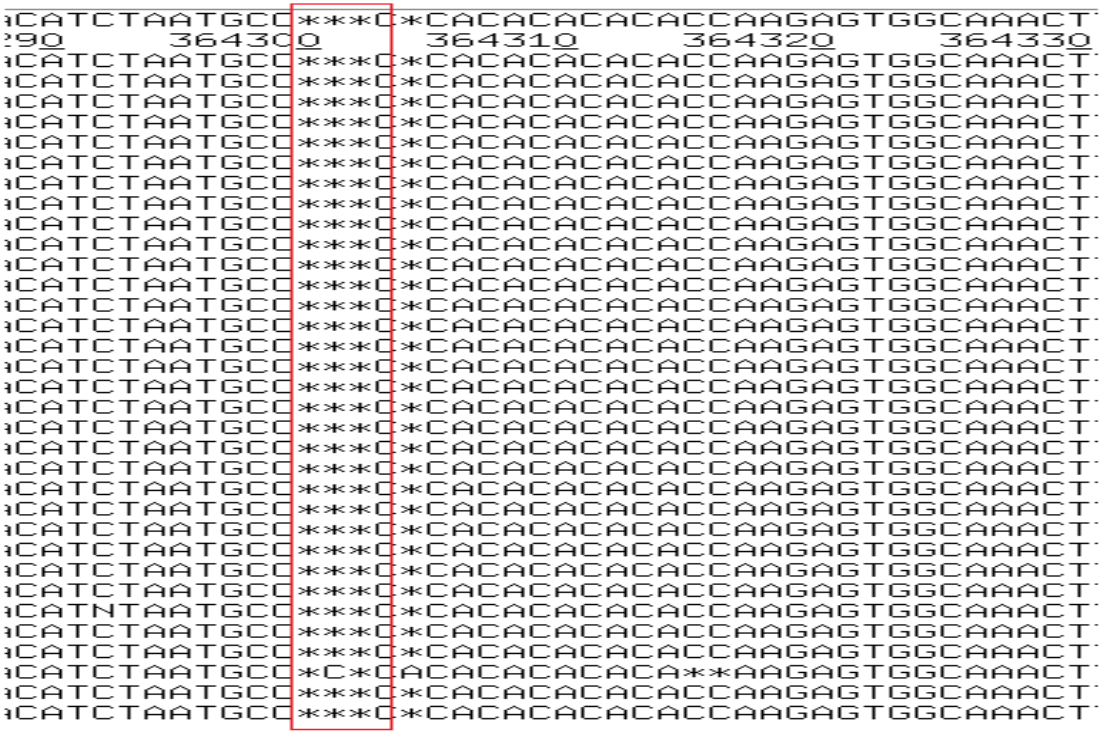
- small indels

Filter	total	TP	FP	TP:FP
--	36417	17927	18490	0.9
Mismatch=0	14447	11304	3143	3.6
Cluster filter	12489	10955	1534	7.1
QUAL/DP>=20	11424	10280	1144	9
QUAL >=1500	3778	3503	275	12.7
MQ>=60	13584	11420	2164	5.3
FP pool	14400	11795	2605	4.5
Mismatch<=1	22324	14583	7741	1.9
Mismatch<=2	26956	15807	11149	1.4
Mismatch<=3	30038	16289	13749	1.2

Examples of FP and undetected indels in gap5_indel

- Example of FP indels

tools	CHROM	POS	REF	ALT	QUAL	INFO	GT:AD
GAP5	chromosome04	364299	ACA	-	159	DP=25; MQ=60	1/1:0,25
GATK		364299	ACA	-	1274.55	DP=26; MQ=60	1/1:0,24



Examples of FP and undetected indels in gap5_indel

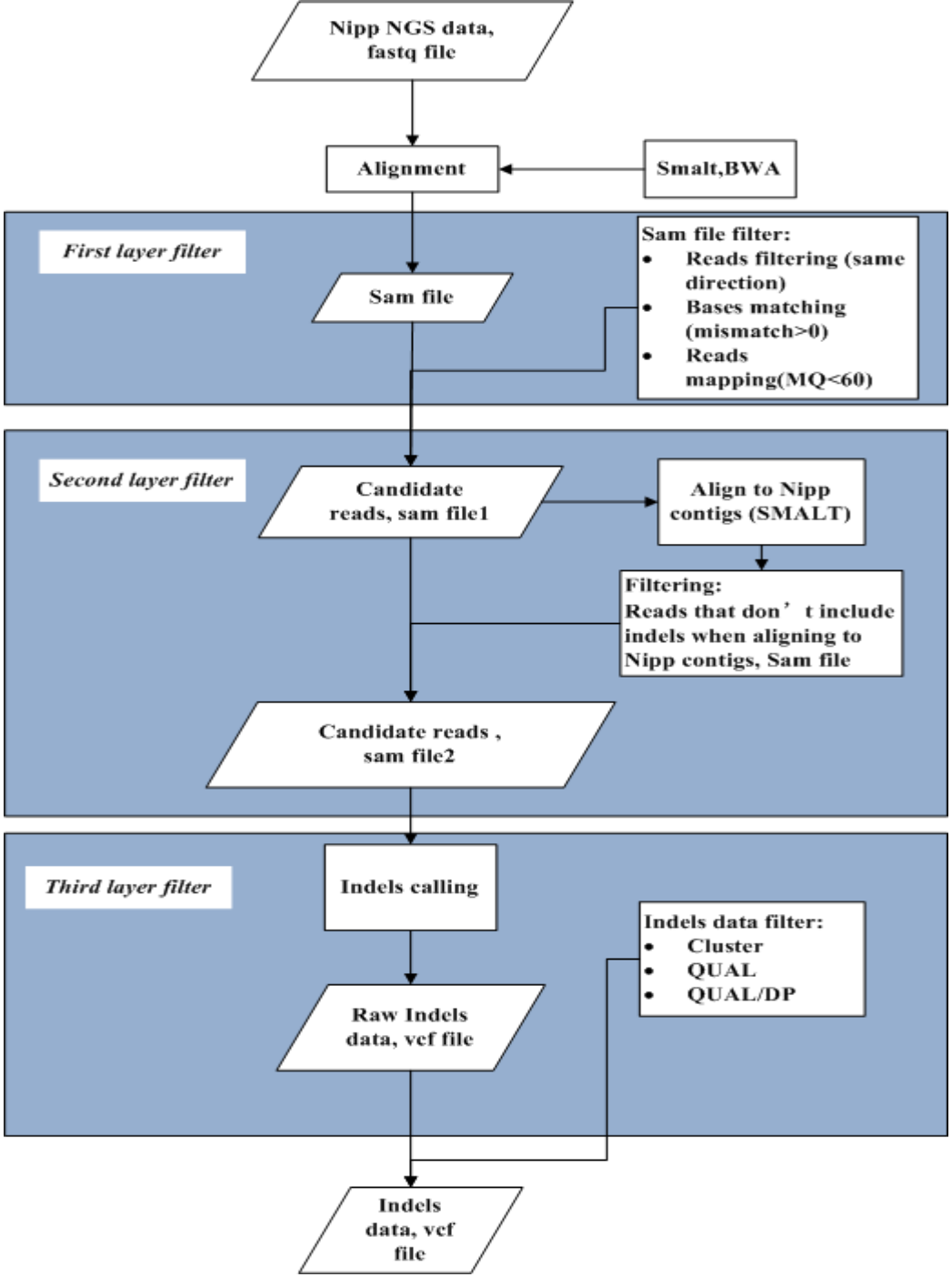
- **Example of undetected indels**

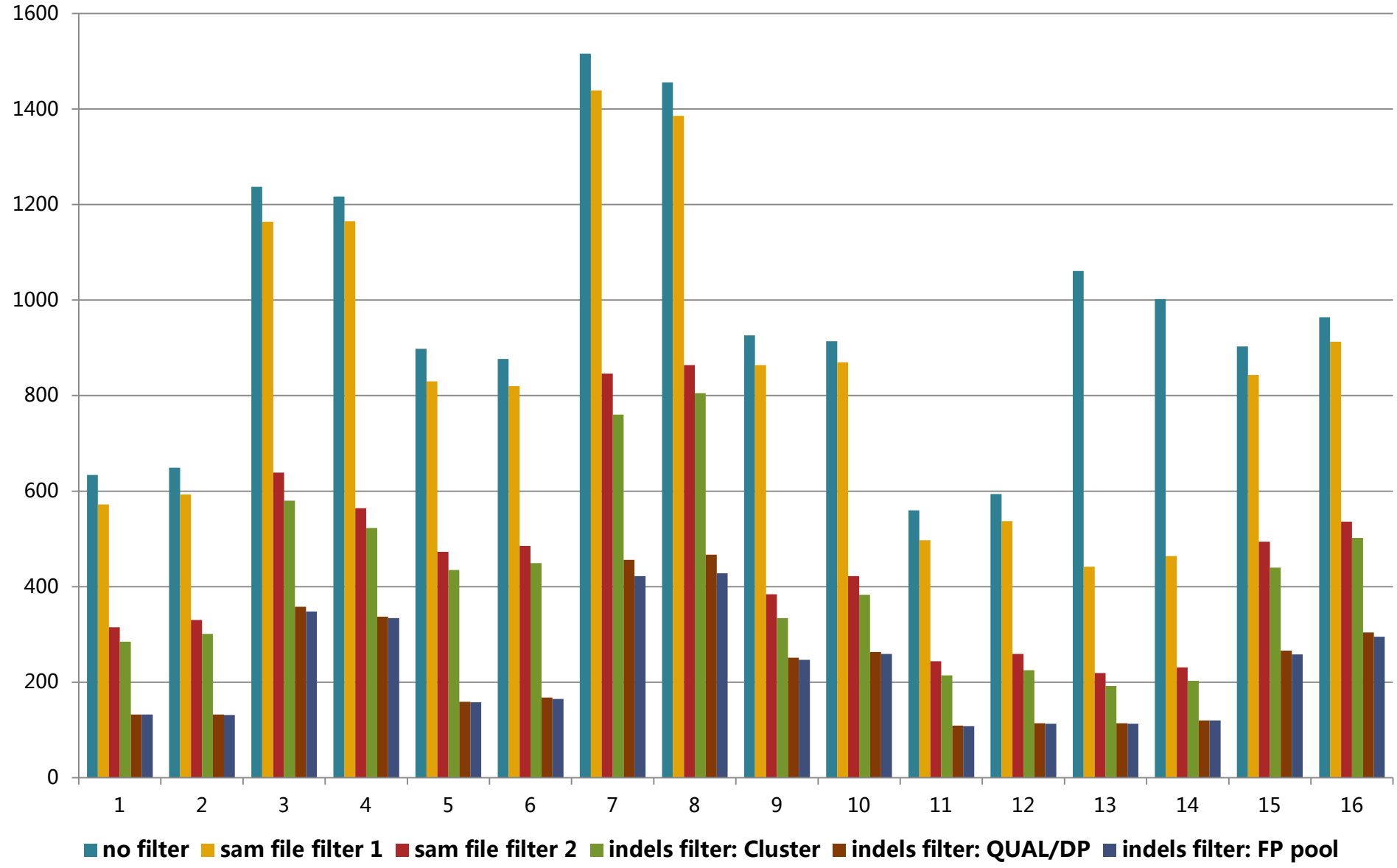
INDEL reference:

CHROM	POS	REF	ALT
chromosome04	**40792**	**ACGTTC**	**-**

日本晴small indels检测流程

- **目的：**尽可能准确地检测small indels


Nipp NGS data, fastq file
Alignment
Smalt,BWA
First layer filter
Sam file
Sam file filter:
Reads filtering (same direction)
Bases matching (mismatch>0)
Reads mapping(MQ<60)
Second layer filter
Candidate reads, sam file1
Align to Nipp contigs (SMALT)
Filtering:
Reads that don’ t include indels when aligning to Nipp contigs, Sam file
Candidate reads , sam file2
Third layer filter
Indels calling
Indels data filter:
Cluster
QUAL
QUAL/DP
Raw Indels data, vcf file
Indels data, vcf file



X-coordinate: Sample ID

Y-coordinate: Numbers of INDELs

下一步工作

- Detective method of large indels (Structural Variations) from PE read
  - 方法：
    - STEP1：identifying breaking point
    - STEP2：assembly & alignment
  - 重点：breaking point detection
  - 参考算法：
    - Clustering-based algorithm：PINDEL, SOAPindel
    - Distribution-based algorithm：MoDIL
    - Combination algorithm：BreakDancer

谢 谢!