Bioinformatics Journal Club

Ulvi Talas

January 28, 2013

Nucl. Acids Res. (2012) doi: 10.1093/nar/gks836;
Methods Online, First published online: September 10, 2012

From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software

Michael Forster1,*, Peter Forster2,3, Abdou Elsharawy1, Georg Hemmrich1, Benjamin Kreck1, Michael Wittig1, Ingo Thomsen1, Björn Stade1, Matthias Barann1, David Ellinghaus1, Britt-Sabina Petersen1, Sandra May1, Espen Melum4,5, Markus B. Schilhabel1, Andreas Keller6, Stefan Schreiber1,7, Philip Rosenstiel1 and Andre Franke1,*

1 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, 2 Institute of Forensic Genetics, D-48161 Münster, Germany, 3 Murray Edwards College, University of Cambridge, CB3 0DF, UK, 4 Division of Gastroenterology, Hepatology, and Endoscopy, Brigham and Women's Hospital, Harvard Medical School, MA 02115, USA, 5 Norwegian PSC Research Center, Clinic for specialized Medicine and Surgery, Oslo University Hospital, Rikshospitalet, Oslo, Norway, 6 Saarland University, Department of Human Genetics, D-66123 Saarbrücken and 7General Internal Medicine, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany

Subject / to-whom-may-it-concern:

… scientists working with single-nucleotide variants (SNVs), inferred by next-generation sequencing software, often need further information regarding **true variants, artifacts and sequence coverage gaps.** In clinical diagnostics, e.g. SNVs must usually be validated by visual inspection or several independent SNV-callers …

Up to (!) 0.5–60% of relevant SNVs might not be detected due to coverage gaps, or might be misidentified!

.

pibase

Acronym for:

get **P**osition **I**nformation at **BASE** position of interest.

Pitfalls of NGS in applied research …

Unfortunately, there are several challenges when faithfully applying the variation–discovery approaches to other uses, such as clinical diagnostics, forensics and targeted-sequencing-based phylogenetic analyses.

- To begin with, **the filtered SNV-lists generated by these approaches do not include low-confidence genotypes,** e.g. where both-stranded validation is missing, **and the unwary data recipient may interpret missing information as a reference sequence genotype.** Also, the default filters sometimes eliminate obvious genotypes.
- The second problem is that available **variant-calling tools usually do not list sequencing failures**, where there is low coverage or no coverage at all, and the **unwary data recipient may again interpret this omission as a reference sequence genotype.**

Pitfalls of NGS in applied research …

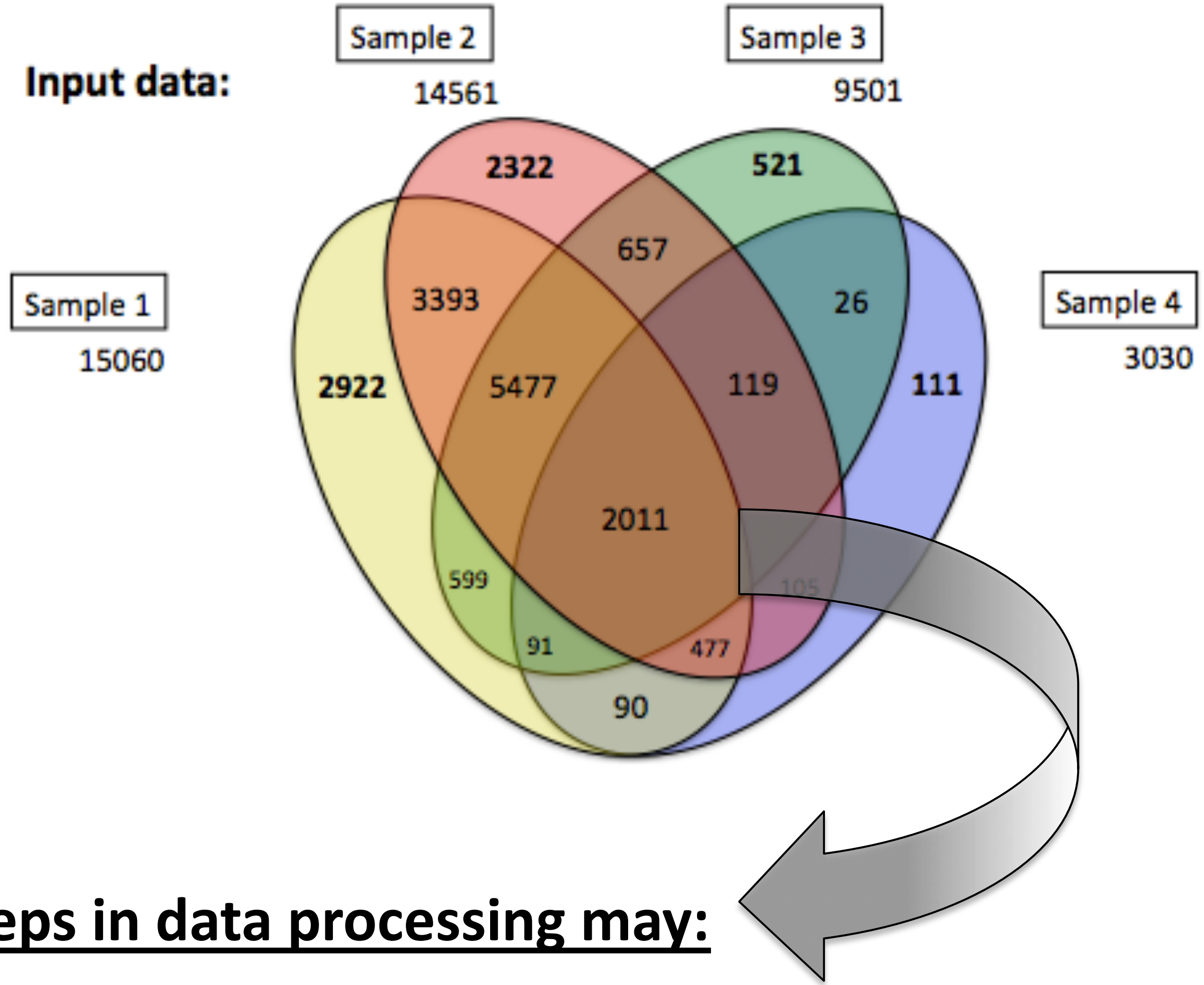
- A third problem is that **SNV-lists usually include incorrectly identified heterozygotes** (prompted by an occasional sequencing error, misalignment or contaminant sequence) **where the pre-set quality filter for machine output or read-alignment is inappropriate.**
- The fourth problem occurs **when the user employs several different SNV-callers to perform a basic validation of the SNV-lists by intersecting the individual SNV-lists to separate cross-validated SNVs from less validated ones.** Because each of these individual tools is prone to filtering away valid SNVs, the **intersected consensus genotypes will exclude even more valid SNVs.**

Pitfalls of NGS in applied research …

- When **performing comparisons between healthy and affected cells/individuals**, a fifth problem surfaces, as each of the first four problems will lead to false differences in the comparative analyses. In other words, **for such comparisons, it may not be advisable to rely on derived SNV-lists.**
- **The sixth and most important problem:** a specific challenge in cell or proband comparisons is to detect significant changes of allelic balance in heterozygous SNVs, e.g. in heterogeneous tumor samples or in the case of copy number variation loci.

… and the unnecessary costs:

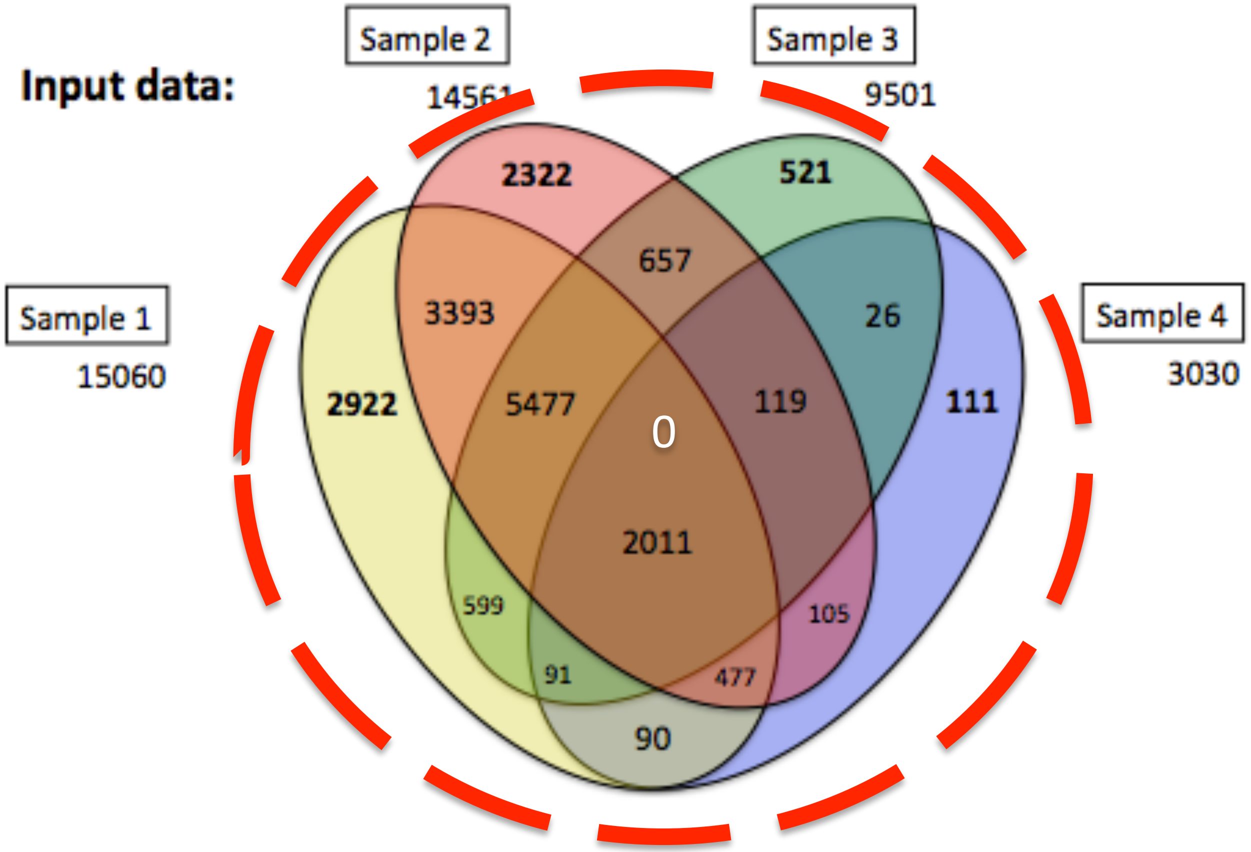
- …, **if there is a communication bottleneck between NGS bioinformaticians** (data producers) **and other scientists/clinicians** (data users), this may result in unnecessary analysis reruns with new work flows or filtering parameters, specifically when new people or new NGS experiments are involved.



The next steps in data processing may:

- Include an overlap (variations called in all or at least two or more analysis runs/samples)
- Soon a hidden assumption of "NO CALL" = "REFERENCE" sequence slips in!

Figure. Prevailing variation calling and phenotype-genotype correlation approach.



For the next steps in data processing:

- Include the union of the variation lists from the initial analysis runs.
- Run pibase on the selected lists to create tables annotating each position in the list with the information on the confidence of the call.

Figure. Accuracy improved variation validation and comparison approach.

pibase

Acronym for: get Position Information at BASE position of interest.

- **Interoperability:**
- pibase reads genomic coordinates of interest from a VCF*, samtools pileup, SOLiD Bioscope gff3, or a tab-separated file.
- **Pibase extracts data at the coordinates of interest from an indexed FASTA reference and from a BAM-file**** generated by BFAST, BWA, SSAHA2, samtools, SOAP (after conversion using soap2sam.pl), and SOLiD Bioscope. To extract the most complete information (including homologous region information and low-coverage genotypes), please use the raw unfiltered BAM-file (which includes non-uniquely mapped reads and duplicate reads).
- **pibase outputs tab-separated text files** which can then be **used in popular spreadsheet software**, or filtered from the linux command line using grep, awk, and cut. pibase can also output variants into VCF, rdf, and snpActs formats.

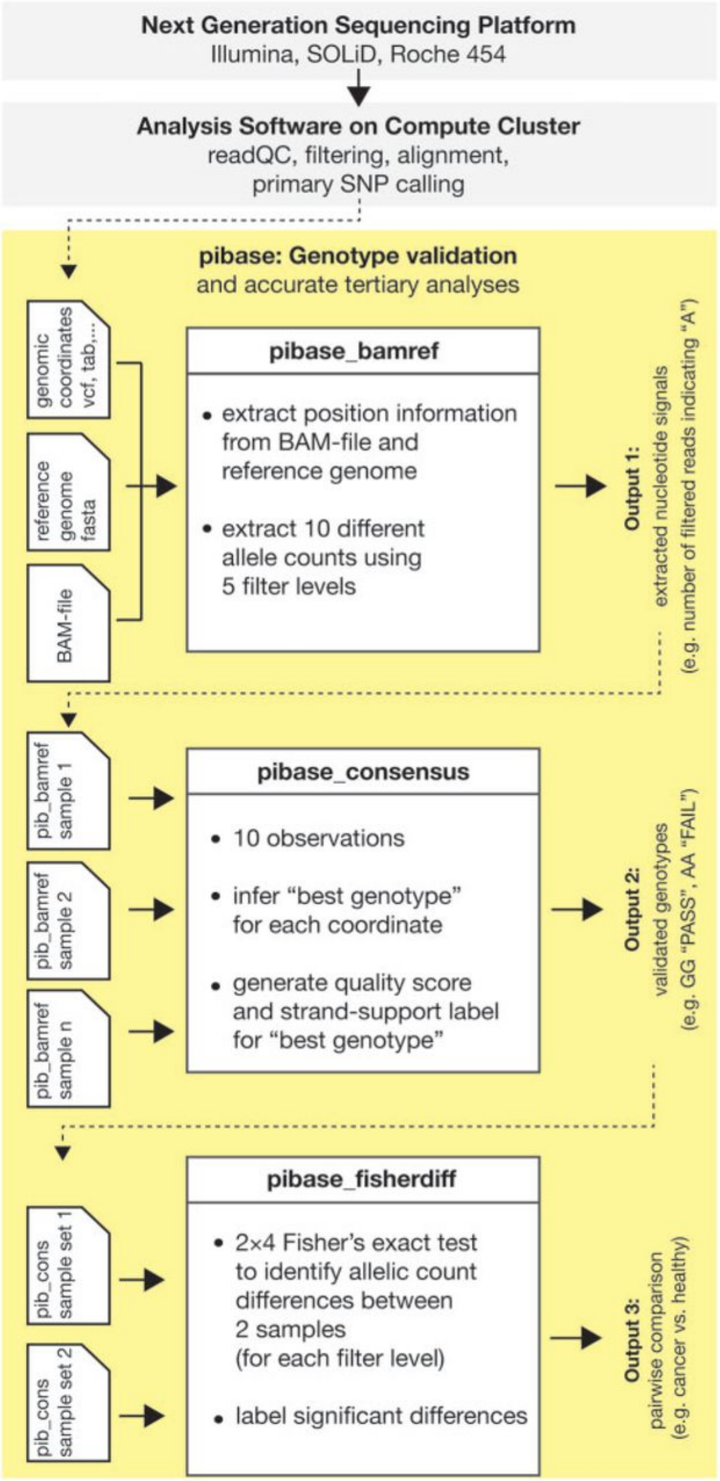
piBASE pre-requisites / system requirements:

- Linux operating system (the authors use CentOS 5.5 / linux 2.6.18-194.32.1.el5 on a linux cluster and Ubuntu 8, 9, or 10 on our PCs.)
- python ~~v2.4.3~~ or v2.6.5 or v2.7.2 (v2.7 recommended for speed!!) http://www.python.org/download/
- pysam v0.6
  http://code.google.com/p/pysam/downloads/list
- GNU Fortran (installable using the Synaptics package manager under Ubuntu PCs)
  http://gcc.gnu.org/wiki/GFortran
- 1GB of RAM (2GB for pibase_fisherdiff)
- Bash command line, or a linux cluster job scheduler such as PBS.

Pibase workflow:

- **pibase_bamref :** extract position info from BAM file and reference sequence file.
- **pibase_consensus over single run:** infer multi-filter-level genotypes from a single pibase_bamref-file and classify the genotypes into stable or dubious genotypes (BestQual flag).
- **pibase_consensus over multiple runs:** infer multi-filter-level ''consensus'' genotypes from pibase_bamref-files from multiple runs and classify the genotypes into stable or dubious genotypes (BestQual flag).
- **[Optional: pibase_fisherdiff :** compare two samples by unique start point counts (Fisher's exact test 2x4), using the pibase_consensus-files**]**

Flow chart showing the standard NGS sequencing and bioinformatic analysis (gray).



Next Generation Sequencing Platform
Illumina, SOLiD, Roche 454

Analysis Software on Compute Cluster
readQC, filtering, alignment,
primary SNP calling

pibase: Genotype validation
and accurate tertiary analyses

genomic coordinates vcf, tab,...

reference genome fasta

BAM-file

pibase_bamref

- extract position information from BAM-file and reference genome
- extract 10 different allele counts using 5 filter levels

Output 1: extracted nucleotide signals (e.g. number of filtered reads indicating "A")

pib_bamref sample 1

pib_bamref sample 2

pib_bamref sample n

pibase_consensus

- 10 observations
- infer "best genotype" for each coordinate
- generate quality score and strand-support label for "best genotype"

Output 2: validated genotypes (e.g. GG "PASS", AA "FAIL")

pib_cons sample set 1

pib_cons sample set 2

pibase_fisherdiff

- 2×4 Fisher's exact test to identify allelic count differences between 2 samples (for each filter level)
- label significant differences

Output 3: pairwise comparison (e.g. cancer vs. healthy)

Nucleic Acids Research

Forster M et al. Nucl. Acids Res. 2013;41:e16-e16

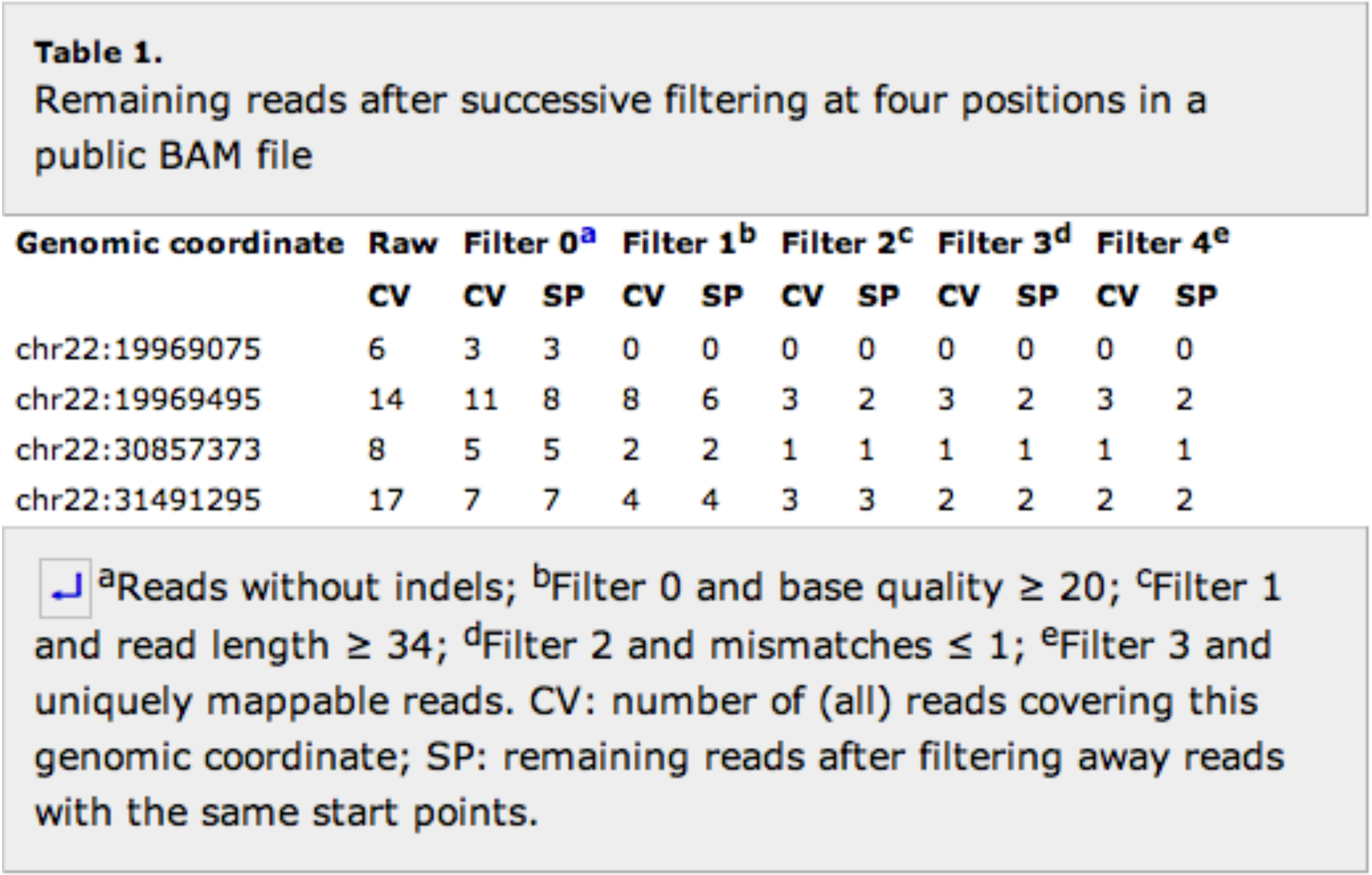


Table 1.

Remaining reads after successive filtering at four positions in a public BAM file

Genomic coordinate	Raw	Filter 0[a]		Filter 1[b]		Filter 2[c]		Filter 3[d]		Filter 4[e]	
	CV	**CV**	**SP**	**CV**	**SP**	**CV**	**SP**	**CV**	**SP**	**CV**	**SP**
chr22:19969075	6	3	3	0	0	0	0	0	0	0	0
chr22:19969495	14	11	8	8	6	3	2	3	2	3	2
chr22:30857373	8	5	5	2	2	1	1	1	1	1	1
chr22:31491295	17	7	7	4	4	3	3	2	2	2	2

[a] Reads without indels; [b] Filter 0 and base quality ≥ 20; [c] Filter 1 and read length ≥ 34; [d] Filter 2 and mismatches ≤ 1; [e] Filter 3 and uniquely mappable reads. CV: number of (all) reads covering this genomic coordinate; SP: remaining reads after filtering away reads with the same start points.

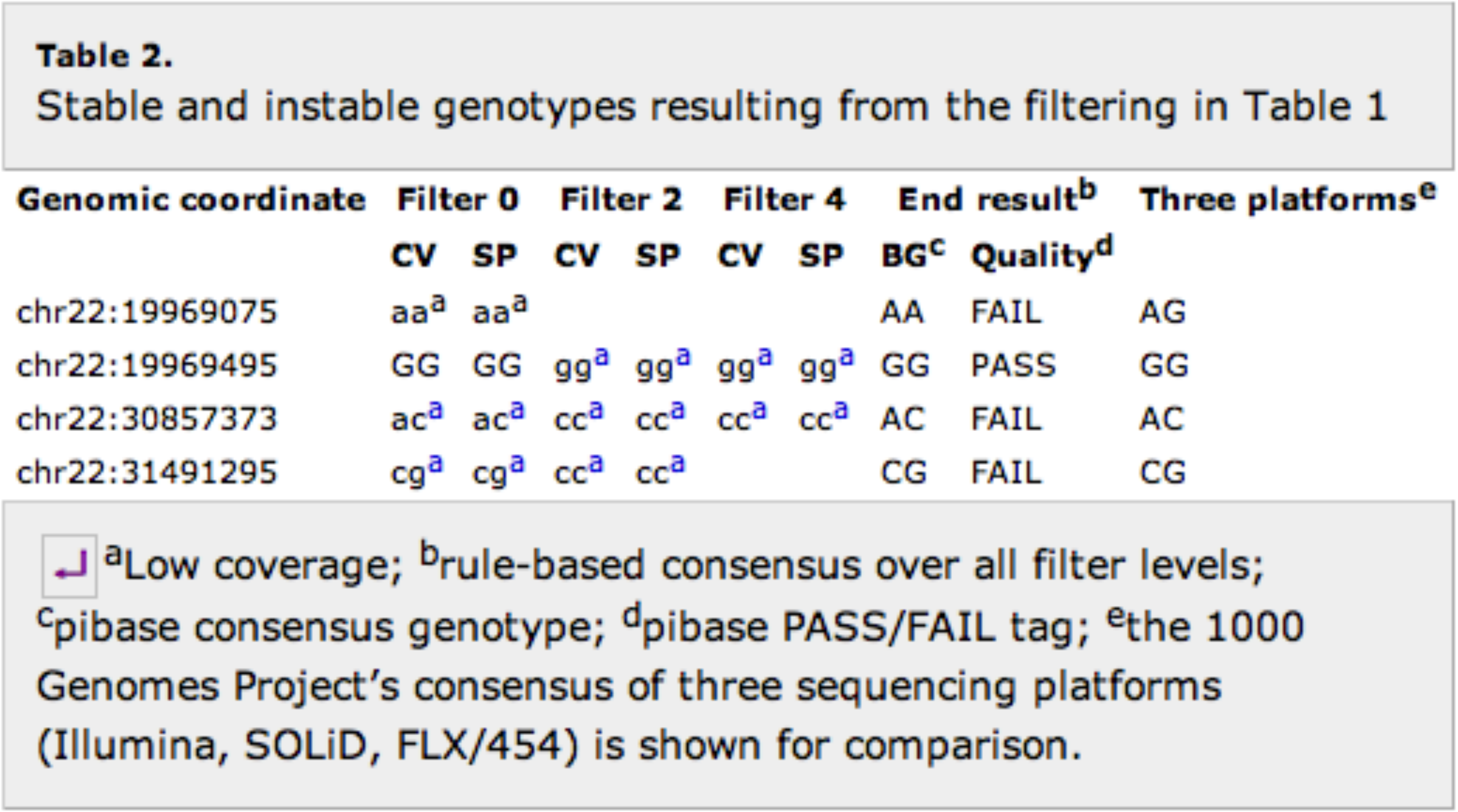


Table 2.

Stable and instable genotypes resulting from the filtering in Table 1

Genomic coordinate	Filter 0		Filter 2		Filter 4		End result[b]		Three platforms[e]
	CV	SP	CV	SP	CV	SP	BG[c]	Quality[d]	
chr22:19969075	aa[a]	aa[a]					AA	FAIL	AG
chr22:19969495	GG	GG	gg[a]	gg[a]	gg[a]	gg[a]	GG	PASS	GG
chr22:30857373	ac[a]	ac[a]	cc[a]	cc[a]	cc[a]	cc[a]	AC	FAIL	AC
chr22:31491295	cg[a]	cg[a]	cc[a]	cc[a]			CG	FAIL	CG

↵ [a] Low coverage; [b] rule-based consensus over all filter levels; [c] pibase consensus genotype; [d] pibase PASS/FAIL tag; [e] the 1000 Genomes Project’s consensus of three sequencing platforms (Illumina, SOLiD, FLX/454) is shown for comparison.

Table 5.

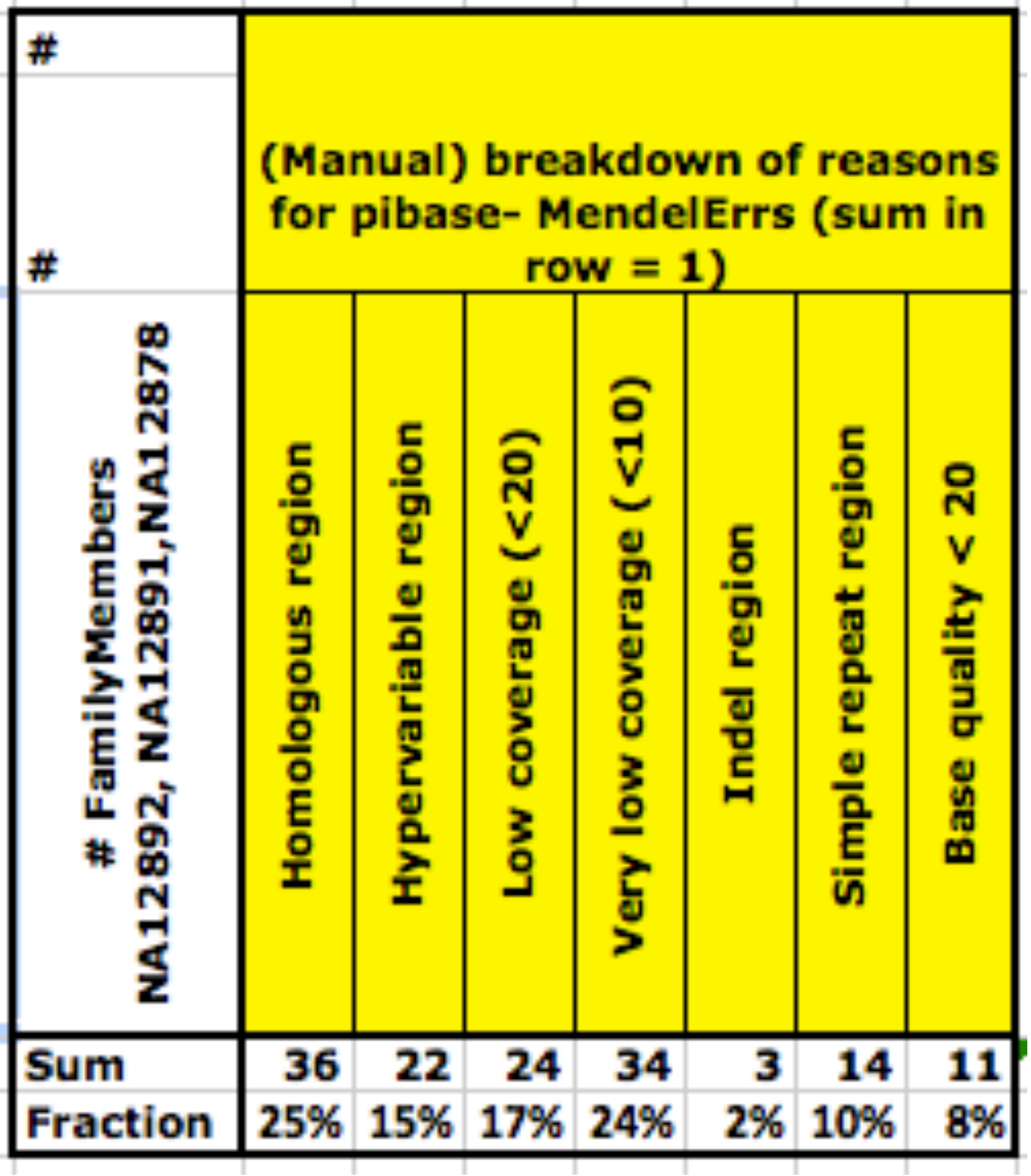
Categorization of instable SNV-calls using SNV label (BestQual)

Label	Explanation
?1	Mapping stringency versus reference sequence context class is good. Not all 10 genotyping filter stages lead to the same genotype. However, for the high mapping stringency filter stages, at least n_1 unique start points and at least n_2 reads support this genotype (defaults: $n_1 = 4$, $n_2 = 8$).
?2	Mapping stringency versus reference sequence context class is good. This genotype is supported by less than five filter stages, but by at least two filter stages, of which one stage is in the unique start points category, and the other stage is in the coverage category.
?3	Poor quality. Low complex reference sequence context (homopolymeric run > 4, or STRs) and low mapping stringency, but at least one stringent filter supports this genotype.
?4	Very poor quality. Low complex reference sequence context (homopolymeric run > 4, or STRs) and mapping stringency was low. But at least one of the unique-start-point filters supports this genotype.
?5	Highly problematic quality. The best unique-start-point derived genotype is in conflict with the best coverage-derived genotype.
?6	Highly problematic quality. The best unique-start-point-derived genotype is in conflict to the best coverage-derived genotype, and the best coverage-derived genotype is 'superior' to the best unique-start-point-derived genotype.
?7	Low-coverage guess. The coverage is less than n_2 (default: $n_2 = 8$).
?8	Low-coverage guess. The coverage is less than n_2 (default: $n_2 = 8$), low complex reference sequence context (homopolymeric run > 4, or STRs), and there are no stringently mappable reads.

STR, short tandem repeats

http://nar.oxfordjournals.org/content/41/1/e16/suppl/DC1

Worksheet



#							
#	(Manual) breakdown of reasons for pibase- MendelErrs (sum in row = 1)						
# FamilyMembers NA12892, NA12891,NA12878	Homologous region	Hypervariable region	Low coverage (<20)	Very low coverage (<10)	Indel region	Simple repeat region	Base quality < 20
Sum	36	22	24	34	3	14	11
Fraction	25%	15%	17%	24%	2%	10%	8%

Run times:

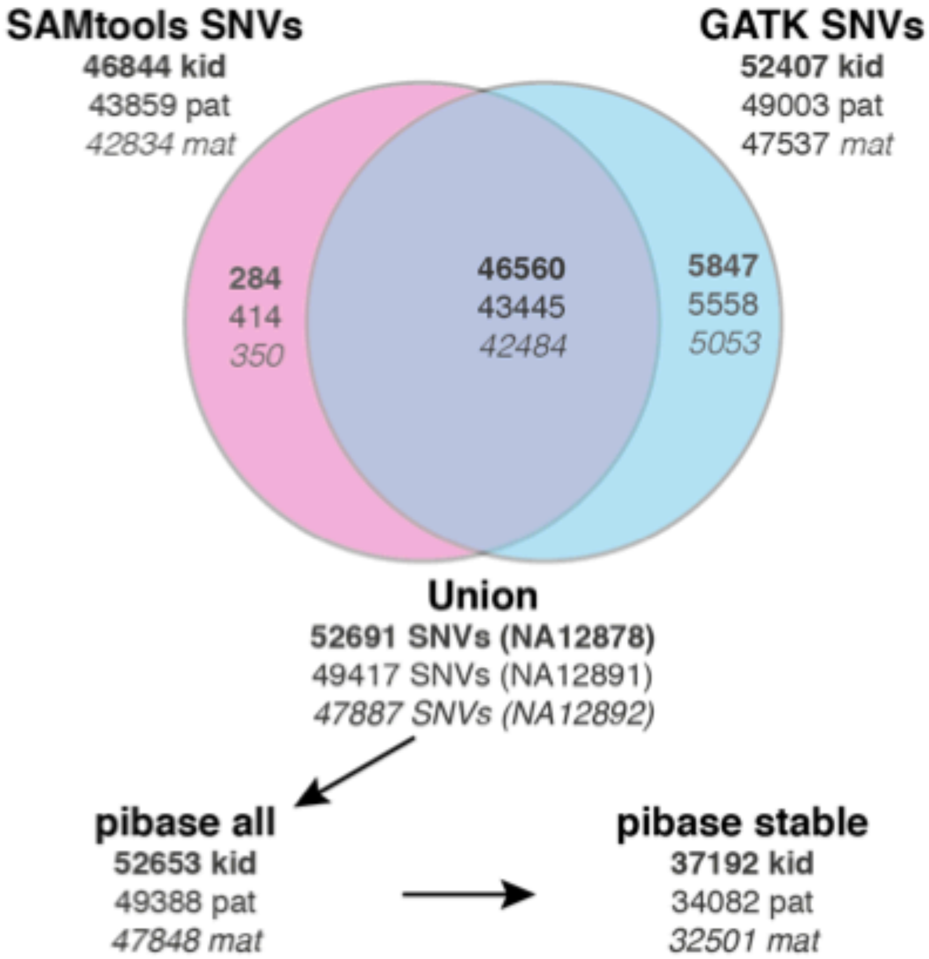
Each sample was analyzed for 19 600 HapMap SNPs on human chr22 on a linux cluster, requiring only a single CPU per run:

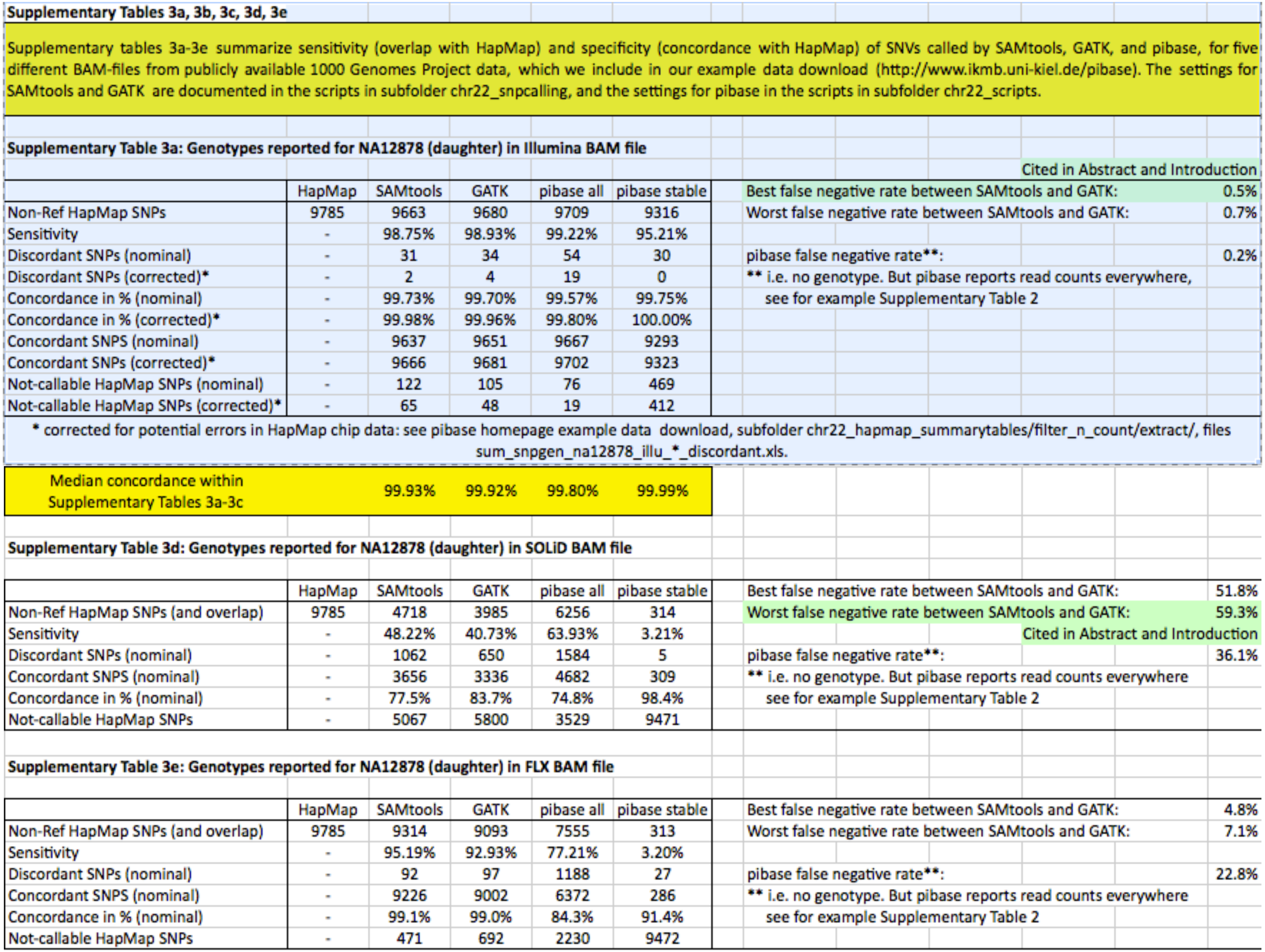
- 4–10 min per sample using pibase,
- 17–55 min per sample using SAMtools and
- about 5 h per sample using GATK.

NB! *The intended use of pibase is to extract in-depth information at selected coordinates of interest (e.g. at coordinates from the National Center for Biotechnology Information database of SNPs (dbSNP), HapMap coordinates or SNV-call coordinates), rather than to scan the entire chromosome for potential non-reference genotypes.*

Supplementary Table 8a: Overlap between samtools and GATK SNV-calls in chr22 of 1000G CEU Trio Illumina BAM-files

Genotyping results	SAMtools SNVs	GATK SNVs	Overlap	Union	pibase all	pibase stable
NA12878 (daughter)	46844	52407	46560	52691	52653	37192
NA12891 (father)	43859	49003	43445	49417	49388	34082
NA12892 (mother)	42834	47537	42484	47887	47848	32501





Supplementary Tables 3a, 3b, 3c, 3d, 3e

Supplementary tables 3a-3e summarize sensitivity (overlap with HapMap) and specificity (concordance with HapMap) of SNVs called by SAMtools, GATK, and pibase, for five different BAM-files from publicly available 1000 Genomes Project data, which we include in our example data download (http://www.ikmb.uni-kiel.de/pibase). The settings for SAMtools and GATK are documented in the scripts in subfolder chr22_snpcalling, and the settings for pibase in the scripts in subfolder chr22_scripts.

Supplementary Table 3a: Genotypes reported for NA12878 (daughter) in Illumina BAM file

	HapMap	SAMtools	GATK	pibase all	pibase stable
Non-Ref HapMap SNPs	9785	9663	9680	9709	9316
Sensitivity	-	98.75%	98.93%	99.22%	95.21%
Discordant SNPs (nominal)	-	31	34	54	30
Discordant SNPs (corrected)*	-	2	4	19	0
Concordance in % (nominal)	-	99.73%	99.70%	99.57%	99.75%
Concordance in % (corrected)*	-	99.98%	99.96%	99.80%	100.00%
Concordant SNPS (nominal)	-	9637	9651	9667	9293
Concordant SNPs (corrected)*	-	9666	9681	9702	9323
Not-callable HapMap SNPs (nominal)	-	122	105	76	469
Not-callable HapMap SNPs (corrected)*	-	65	48	19	412

Cited in Abstract and Introduction

Best false negative rate between SAMtools and GATK:	0.5%
Worst false negative rate between SAMtools and GATK:	0.7%
pibase false negative rate**:	0.2%

** i.e. no genotype. But pibase reports read counts everywhere, see for example Supplementary Table 2

* corrected for potential errors in HapMap chip data: see pibase homepage example data download, subfolder chr22_hapmap_summarytables/filter_n_count/extract/, files sum_snpgen_na12878_illu_*_discordant.xls.

	SAMtools	GATK	pibase all	pibase stable
Median concordance within Supplementary Tables 3a-3c	99.93%	99.92%	99.80%	99.99%

Supplementary Table 3d: Genotypes reported for NA12878 (daughter) in SOLiD BAM file

	HapMap	SAMtools	GATK	pibase all	pibase stable
Non-Ref HapMap SNPs (and overlap)	9785	4718	3985	6256	314
Sensitivity	-	48.22%	40.73%	63.93%	3.21%
Discordant SNPs (nominal)	-	1062	650	1584	5
Concordant SNPS (nominal)	-	3656	3336	4682	309
Concordance in % (nominal)	-	77.5%	83.7%	74.8%	98.4%
Not-callable HapMap SNPs	-	5067	5800	3529	9471

Best false negative rate between SAMtools and GATK:	51.8%
Worst false negative rate between SAMtools and GATK:	59.3%

Cited in Abstract and Introduction

pibase false negative rate**:	36.1%

** i.e. no genotype. But pibase reports read counts everywhere see for example Supplementary Table 2

Supplementary Table 3e: Genotypes reported for NA12878 (daughter) in FLX BAM file

	HapMap	SAMtools	GATK	pibase all	pibase stable
Non-Ref HapMap SNPs (and overlap)	9785	9314	9093	7555	313
Sensitivity	-	95.19%	92.93%	77.21%	3.20%
Discordant SNPs (nominal)	-	92	97	1188	27
Concordant SNPS (nominal)	-	9226	9002	6372	286
Concordance in % (nominal)	-	99.1%	99.0%	84.3%	91.4%
Not-callable HapMap SNPs	-	471	692	2230	9472

Best false negative rate between SAMtools and GATK:	4.8%
Worst false negative rate between SAMtools and GATK:	7.1%
pibase false negative rate**:	22.8%

** i.e. no genotype. But pibase reports read counts everywhere see for example Supplementary Table 2

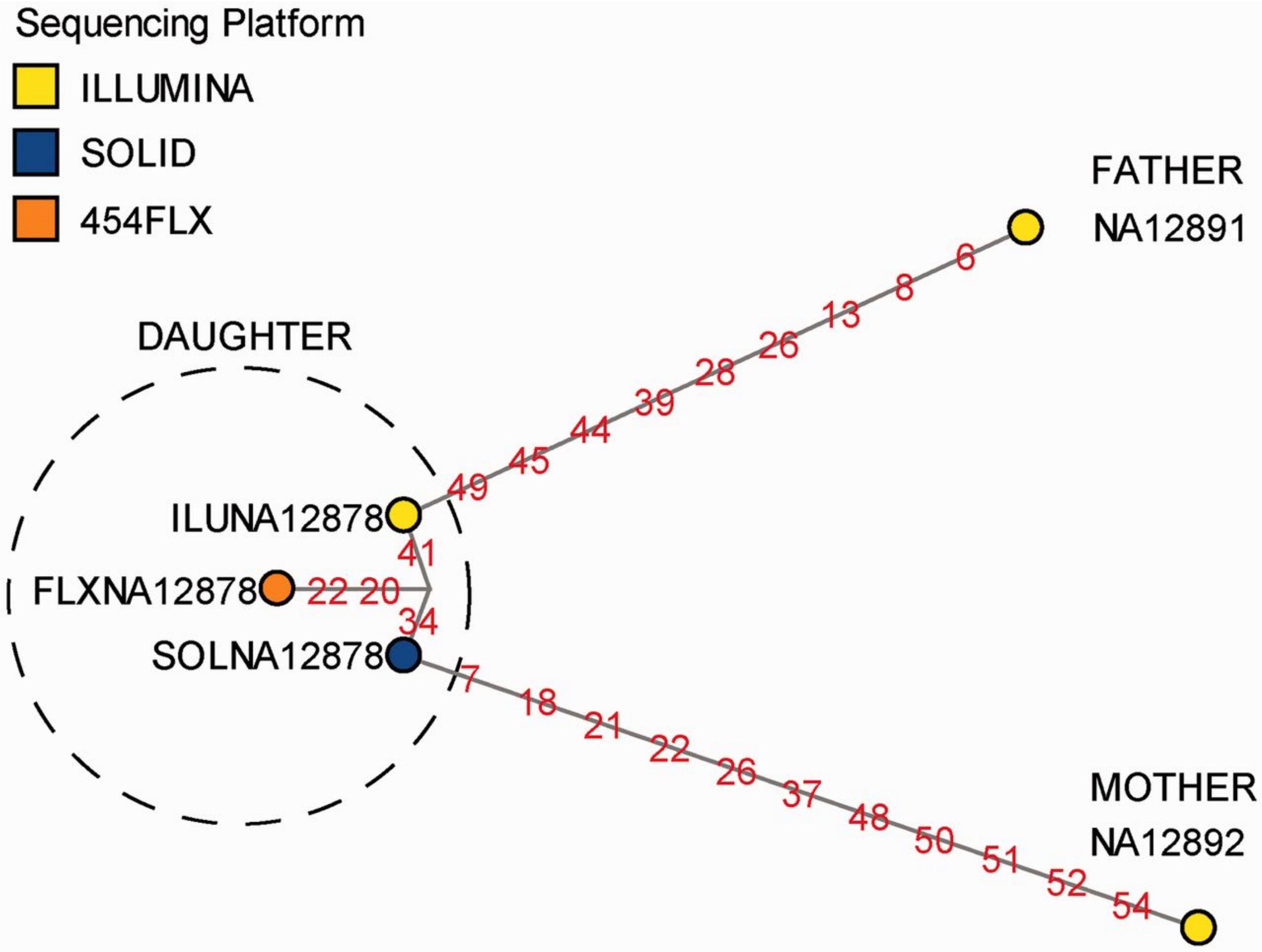
Optional complementary workflows & utilities:

The **‘phylogenetics workflow’** provides a link from NGS data to *median joining network* analysis. Can also be **used to:**

- compute the evolutionary network of heterogeneous tumor cells within a single patient
- compute SNV differences in identical twins
- phylogenetic screening for sample confusion

Limited **‘annotation workflow’**

Median joining network showing the differences between the five examples of BAM files of the CEU trio.



Forster M et al. Nucl. Acids Res. 2013;41:e16-e16

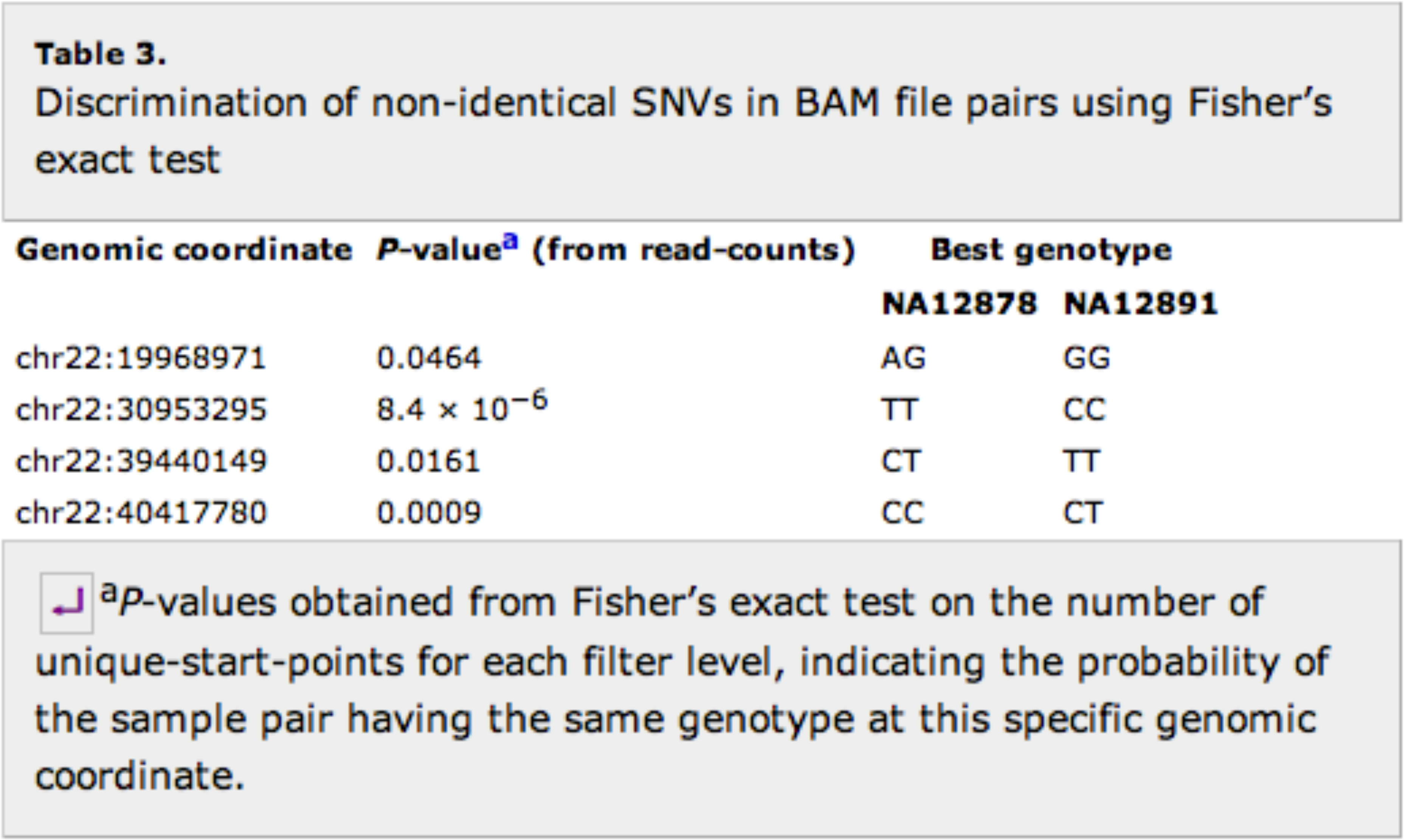


Table 3.

Discrimination of non-identical SNVs in BAM file pairs using Fisher's exact test

Genomic coordinate	*P*-value[a] (from read-counts)	Best genotype	
		NA12878	NA12891
chr22:19968971	0.0464	AG	GG
chr22:30953295	8.4×10^{-6}	TT	CC
chr22:39440149	0.0161	CT	TT
chr22:40417780	0.0009	CC	CT

[a] *P*-values obtained from Fisher's exact test on the number of unique-start-points for each filter level, indicating the probability of the sample pair having the same genotype at this specific genomic coordinate.

In summary,
pibase addresses major problems pertaining to the quality control, validation and accurate comparison of NGS variant data, which are a bottleneck in currently emerging translational uses of NGS.

Furthermore, the pibase data tables facilitate the practical use of NGS data by non-bioinformaticians such as archaeogeneticists, biologists, clinicians and forensic scientists.