Institute of
Medical Genetics and
Applied Genomics

megSAP
a Medical Genetics Sequence Analysis Pipeline

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Marc Sturm[1*], Christopher Schroeder[1], Tobias Haack[1]
[1] Institute of Medical Genetics and Applied Genomics, University of Tuebingen, Germany.
*marc.sturm@med.uni-tuebingen.de

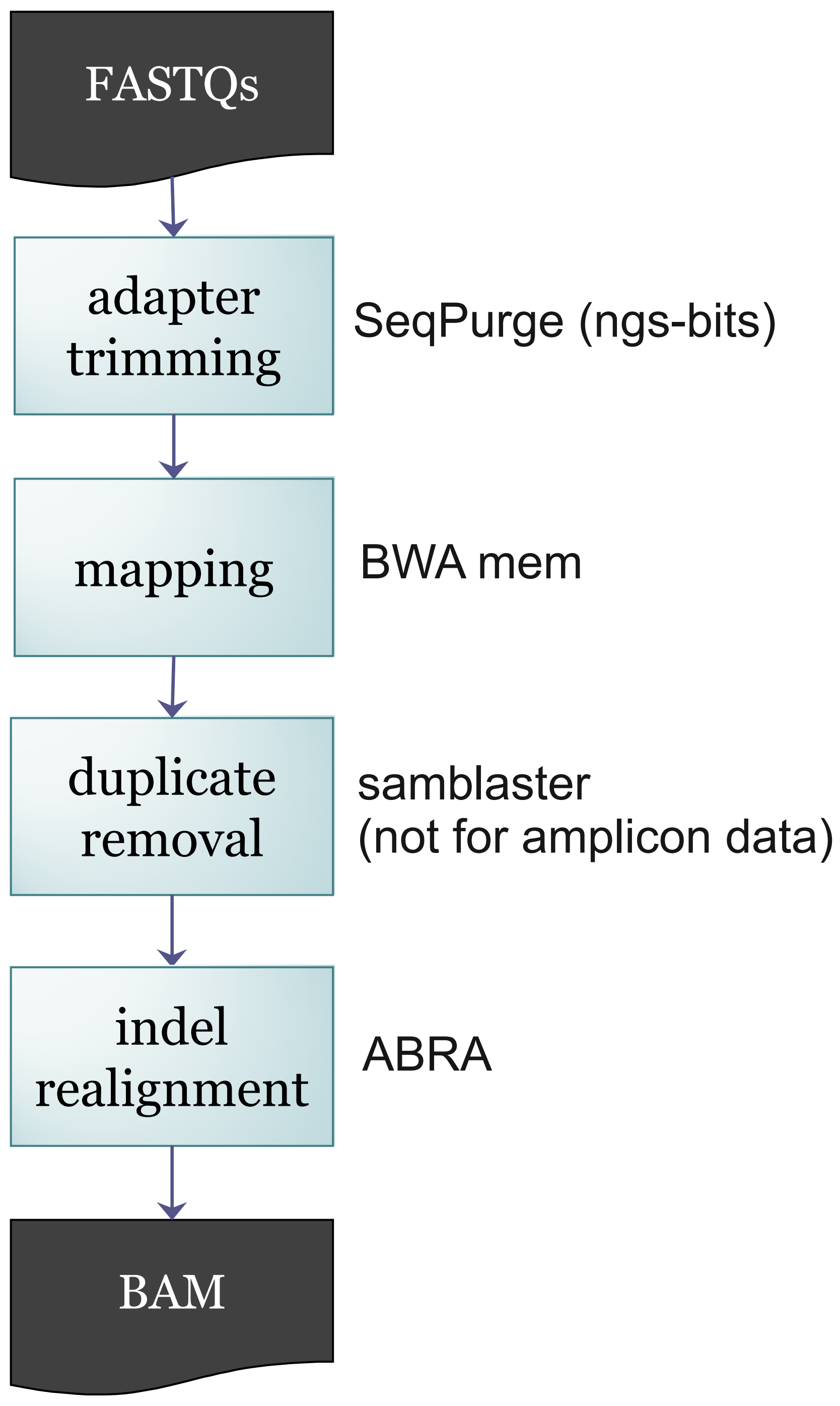
Abstract

Today, NGS is widely used in clinical diagnostics and translational research to identify disease-causing variants. While several commercial software suites for short-read NGS data analysis are available, they are normally quite costly and not easy to automate in a high-throughput setting.

Thus, we have developed megSAP, a free-to-use open-source data analysis pipeline tailored towards research and diagnostics in medical genetics. megSAP offers a complete NGS data analysis pipeline (adapter trimming, mapping, duplicate removal if applicable, indel realignment, variant calling, variant normalization and variant annotation) that is complemented with quality control on several levels (raw reads, mapped reads and variant). It is entirely based on open-source tools that are free for commercial use, which rules out popular tools like GATK and Annovar. It also integrates free-to-use databases like 1000 Genomes, ExAC, Kaviar and ClinVar for annotation of variants. Optionally, commercial databases which are important for diagnostics (OMIM, HGMD and COSMIC) can be used if a license is available. Due to the comprehensive annotation, the variant lists (produced in VCF and TSV format) can be easily filtered to identify disease variants.

megSAP is regularly updated (both tool and annotation databases). Each release is validated using the GiaB NA12878 gold-standard dataset, inter-laboratory comparisons and EMQN test schemes. Currently, megSAP is readily usable to analyze single-sample NGS data from whole-genome sequencing, whole-exome sequencing and panel sequencing (both shotgun and amplicon-based data). Several other applications (RNA-Seq, tumor-normal pairs, trios, and molecular barcodes) are already implemented and the corresponding documentation will be added shortly. To facilitate the installation of megSAP and thereby improve usability, we are working on a first containerized release using Docker.

(1) Mapping


FASTQs
adapter trimming — SeqPurge (ngs-bits)
mapping — BWA mem
duplicate removal — samblaster (not for amplicon data)
indel realignment — ABRA
BAM

This box shows the main steps of our mapping pipeline which uses hg19 as a reference.

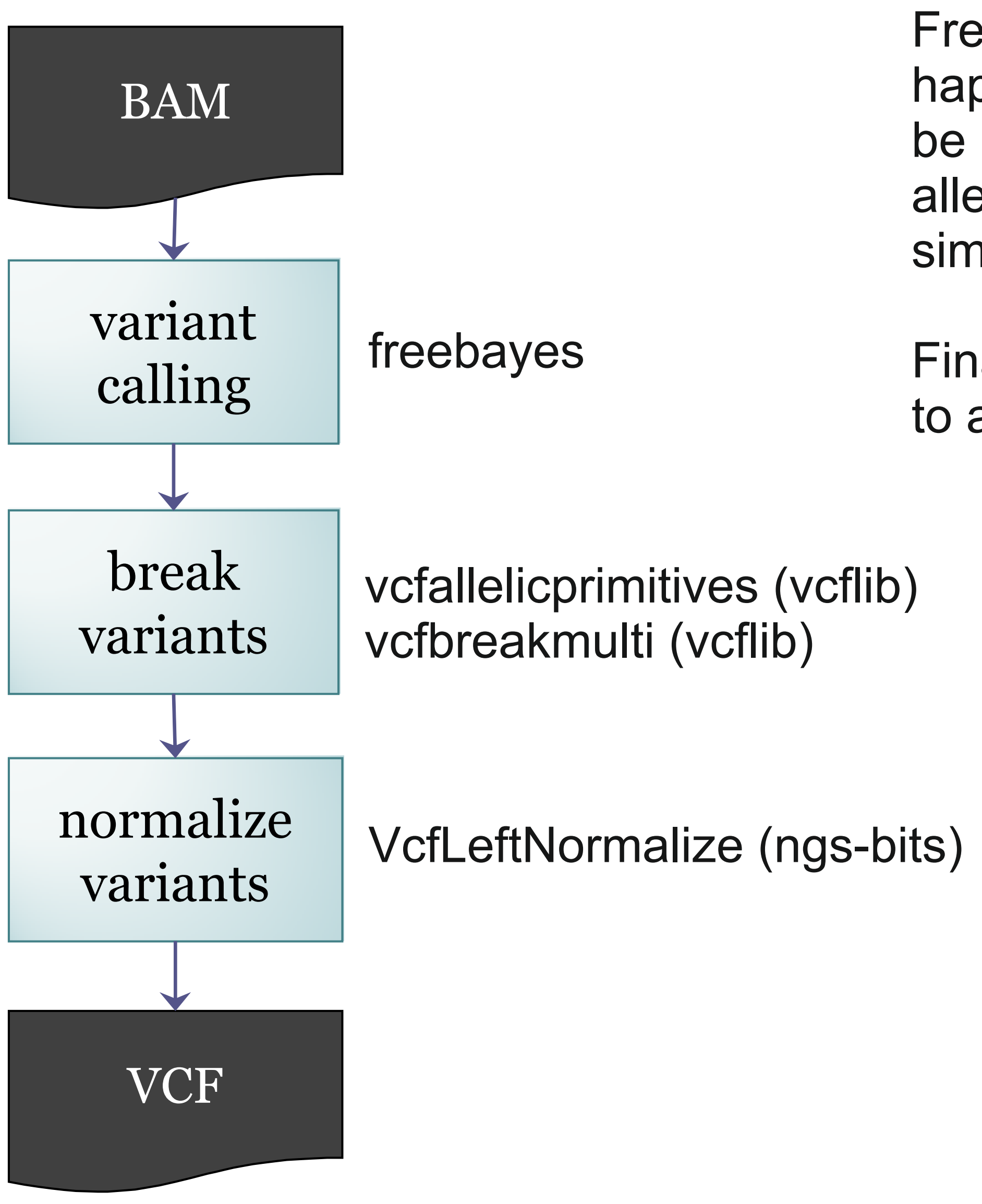
Adapter and quality trimming reduces both mapping runtime and errors. Thus it is always applied.

By the use of linux pipes, the mapping is optimized both in terms of speed and IO load. Because samblaster supports pipes, we use it instead of the popular Picard tools.

The indel realigner provided by GATK is likely the most commonly used one, however a GATK license is needed for diagnostics. We found that the indel realigner ABRA is faster and has a very similar performance. For large deletions it even performs better than GATK.

(2) Variant calling

Because GATK is not free for diagnostics, we cannot use the very widely used HaplotypeCaller. We use freebayes, a variant caller that is also haplotype-based and shows a very similar performance in recent reviews.

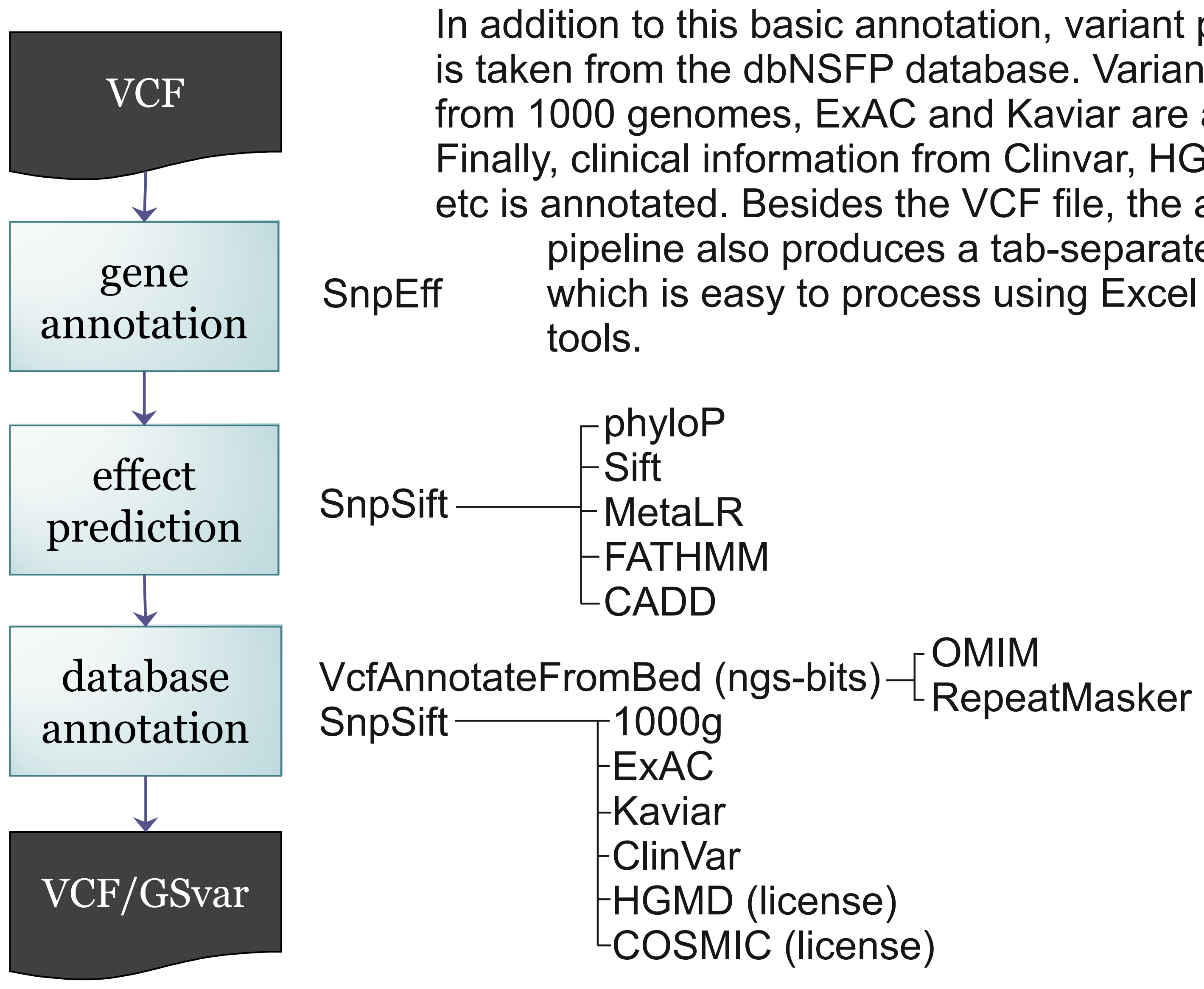

BAM
variant calling — freebayes
break variants — vcfallelicprimitives (vcflib)
vcfbreakmulti (vcflib)
normalize variants — VcfLeftNormalize (ngs-bits)
VCF

Freebayes calls and reports variants as haplotype blocks. These blocks need to be broken into allelic privitives and multi-allelic variants need to be converted to simple variants.

Finally, indel left-alignment and is applied to allow proper annotation.

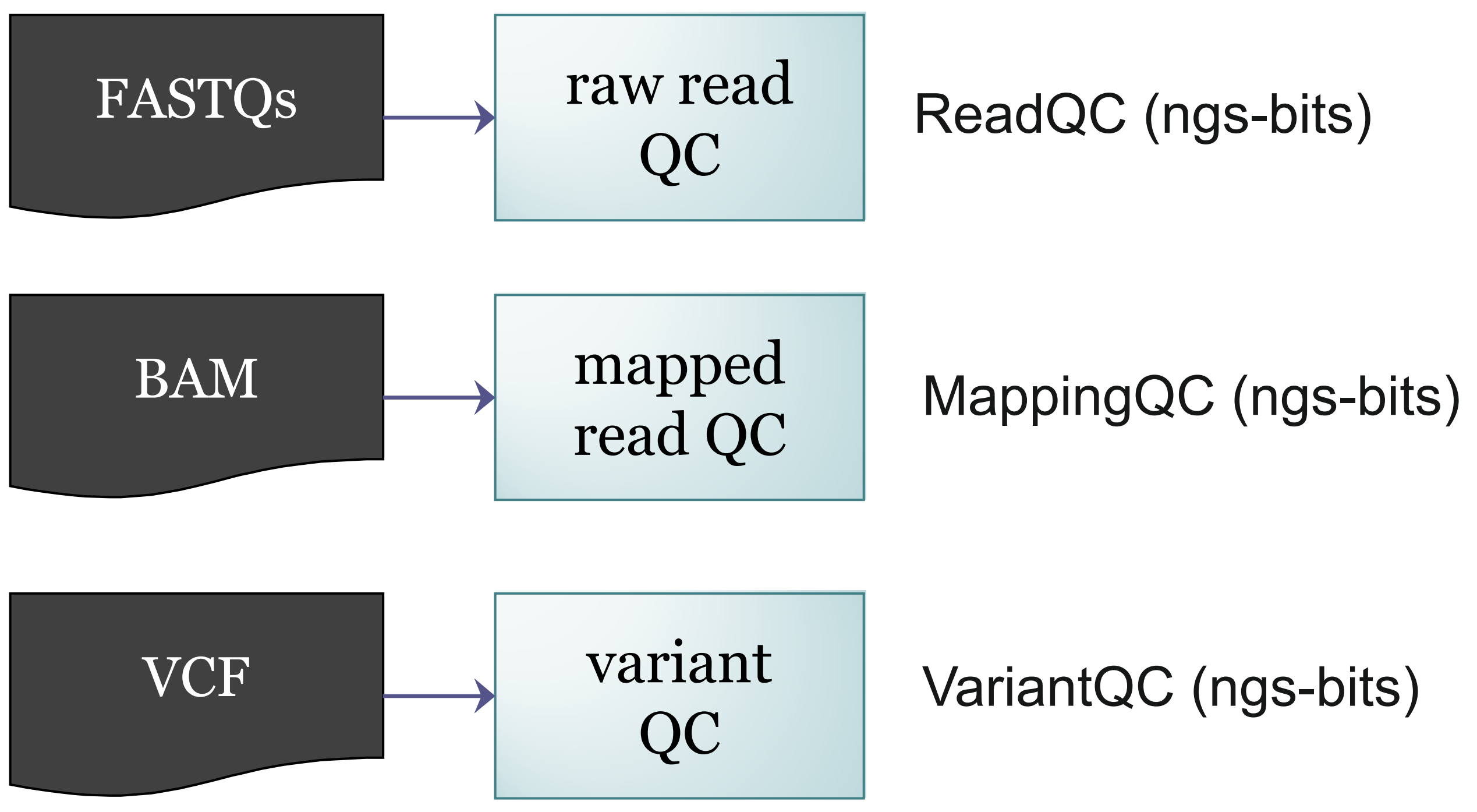
(3) Annotation

For annotation of variants, we use the free tool SnpEff, which annotates the effect of a variant on all transcripts. In addition to this basic annotation, variant pathogenicity is taken from the dbNSFP database. Variant frequencies from 1000 genomes, ExAC and Kaviar are annotated. Finally, clinical information from Clinvar, HGMD, OMIM, etc is annotated. Besides the VCF file, the annotation pipeline also produces a tab-separated GSvar file, which is easy to process using Excel or similar tools.


VCF
gene annotation — SnpEff
effect prediction — SnpSift — phyloP, Sift, MetaLR, FATHMM, CADD
database annotation — VcfAnnotateFromBed (ngs-bits) — OMIM, RepeatMasker
SnpSift — 1000g, ExAC, Kaviar, ClinVar, HGMD (license), COSMIC (license)
VCF/GSvar

(4) Quality control

Quality control is a crucial step when using NGS in clinical diagnostics. Thus we perform QC for individual samples using all three layers of information. Unprocessed raw data, mapped reads and variant list are used to calculate a comprehensive list of quality metrics. The metrics are stored in qcML format. These qcML files are easy to process computationally because they are XML-based. They can also be opened and rendered in a web browser through an embedded style sheet.


FASTQs → raw read QC — ReadQC (ngs-bits)
BAM → mapped read QC — MappingQC (ngs-bits)
VCF → variant QC — VariantQC (ngs-bits)

(5) Development progress

This poster only shows the megSAP single sample pipeline for Panel, WES or WGS data, which is fully tested and documented. megSAP also supports several other applications based on similar pipelines. Here the current status of implemented and planned applications is listed:

	implemented	tested	documented
DNA single sample - molecular barcodes	yes	in progress	no
DNA trio analysis	yes	yes	no
DNA multi-sample analysis	in progress	no	no
DNA somatic (tumor-normal pair)	yes	yes	no
RNA expression	yes	yes	no
RNA variant calling	no	no	no

Besides calling of small variants, we also work on copy-number calling for targeted sequencing (panel/WES) and WGS, and on structural variant calling for WGS.

(6) Availability

megSAP is freely available under the „GPL Version 3" license from GitHub:
https://github.com/imgag/megSAP

If you want to join the effort and contribute, please contact Marc Sturm:
marc.sturm@med.uni-tuebingen.de

References:

ngs-bits	https://github.com/imgag/ngs-bits
BWA	http://bio-bwa.sourceforge.net/
samblaster	https://github.com/GregoryFaust/samblaster
ABRA	https://github.com/mozack/abra
freebayes	https://github.com/ekg/freebayes
vcflib	https://github.com/vcflib/vcflib
SnpEff	http://snpeff.sourceforge.net/index.html
SnpSift	http://snpeff.sourceforge.net/SnpSift.html