Bioinformatic approaches for next generation sequencing variant calling in matched sensitive-resistant ovarian tumor pairs

Luca Beltrame[1], Mariacristina Di Marino[1], Luca Clivio[1], Brad Chapman[2], Romina Baldo[3], Sonia Magni[3], Maurizio D'Incalci[1],and Sergio Marchini[1]

1. Department of Oncology, IRCCS Istituto di Ricerche Farmacologiche "Mario Negri", Italy; 2. School of Public Health, Harvard University, USA; 3. Clinic of Obstetrics and Gynecology, San Gerardo Hospital, Italy.

Poster number: P16.29-S

ABSTRACT

Introduction: Despite initial response to first line platinum-based chemotherapy, more than 80% of high grade serous ovarian cancer patients relapse and develop resistance. The molecular and genetic features involved in drug resistance are still unknown. Transcrptional profiling, carried out in a cohort of patients from which matched biopsies were taken at primary surgery (PS-O) when tumor was sensitive to chemotherapy and at time of relapse (SCR) when the tumor was resistant, identified the EMT pathway as a key player in tumor relapse (Marchini et al., 2013). Here we describe the development of computational analysis approaches to identify somatic variants (point mutations and insertion/deletions) using targeted DNA resequencing on our cohort of SCR and PS-O samples.
Methods: We improved an existing bioinformatic pipeline, "bcbio-nextgen", which supports most best practices in use within the community. After a thorough examination of existing programs, we added for matched tumor samples in the pipeline, using three major somatic variant callers (MuTect, VarScan2, FreeBayes). The pipeline was then tested on a a high performance cluster computing platform (Cloud4CARE project).
Results: The fitness of the method was first assessed by testing a reduced data set from the Cancer Genome Atlas, then the pipeline was run over the complete data set of matched PS-O – SCR samples. Variants identified by the pipeline were correctly discriminated as being germline or somatic, and external validation confirmed the results.
Conclusions: Our results suggest that our bioinformatic approach is sensitive, robust, reproducible and viable for analysis of matched EOC samples.

BACKGROUND

Epithelial ovarian cancer (EOC), is generally sensitive to first line platinum based therapy, however more than 80% of patients experience relapse within 18 months from the end of therapy, and patients become resistant to subsequent cycles until the disease becomes incurable (Cannistra, 2004). Although many advances have been made in understanding the underlying biology, genetic and molecular mechanisms of drug resistance in EOC have yet to be clearly identified (Bast et al., 2009). It is still not clear whether resistance is due to subpopulations of resistant cells, already existing in the tumor before treatment, or it is induced by mutations or epigenetic changes caused by chemotherapeutic drugs (Lawrenson and Gayther, 2009). Current knowledge is based on the use of cancer cell lines with acquired resistance to chemotherapeutic drugs: although these models have been useful to identify mechanisms of drug resistance, it is unclear whether they recapitulate the actual situation in patients undergoing chemotherapy.
Recently, large-scale deep sequencing projects (Cancer Genome Atlas, 2013) have uncovered important genomic features of EOC, showing the usefulness of the approach to complement existing knowledge. In parallel other studies have been conducted on the genomic changes related to drug resistance (Murtaza et al., 2013; Castellarin et al., 2013) although on small sample sets, thus requiring confirmation in larger cohorts of patients.

MATERIALS AND METHODS

EOC samples were selected from **Pandora**, a tumor tissue collection of more than 1600 snap frozen biopsies from patients recruited at San Gerardo Hospital in Monza (Italy) and stored at -80 C at the Mario Negri Institute: the cohort consists of 33 stage III-IV patients from whom biopsies were taken at primary surgery in the ovary (PS-O) and at for relapse after several lines of chemotherapy (secondary cytoreduction, SCR) (Marchini, Fruscio, Beltrame et al., 2013). Additionally, matched blood samples from patients were used as control.

Genomic DNA was extracted using a commercially available kit, according to the manufacturer's instructions, and enriched DNA libraries (66 genes, encompassing key players of signal transduction, EMT regulation, extracellular matrix interaction, cell cycle and DNA repair) where produced using the TruSeq Custom Amplicon kit (Illumina, USA).
DNA libraries were then sequenced on an Illumina MiSeq (Illumina, USA), using the MiSeq v2 Reagent Kit (Illumina, USA) over 300 cycles (150 bp read length), multiplexing 8 samples for each run. Raw FASTQ files from the instrument were de-multiplexed, then used for downstream processing and analysis.

SELECTION OF SOMATIC VARIANT CALLERS

In order to properly identify mutations in our matched EOC cohort, we aimed at selecting somatic mutations in the tumor, excluding any variant of germline origin. Our initial goal was select multiple programs to run the analysis, as different algorithms, while correct, do not overlap (O'Rawe et al., 2013), and that allelic fractions of mutations can influence the sensitivity of the analysis methods (Xu et al., 2014). As several software is available to perform somatic variant calling, we ran an assessment on the available callers, following these criteria:

- Free availability
- Active maintainership and updates
- Results in standard output formats (Variant Call Format; VCF)
- Compatibility with other tools' inputs and outputs

A summary of the assessment is available in Table 1. **MuTect** (Cibulskis et al., 2013), **VarScan 2** (Koboldt DC et al., 2012) and **FreeBayes** (Garrison and Marth, 2012) met the requirements for inclusion and were then incorporated in the analysis pipeline.

Tool name	Latest release	Output	Compatible
MuTect	2014	VCF, text	yes
VarScan 2	2013	VCF, text	yes
FreeBayes	2014	VCF	yes
Somatic Sniper	2013	VCF	no†
EBCall	2013	text	no
Strelka	2012	text	no
SNVMix	2013	text	no
Virmid	2013	VCF (single sample)	no

Table 1: Software evaluated. †, depends on incompatible software

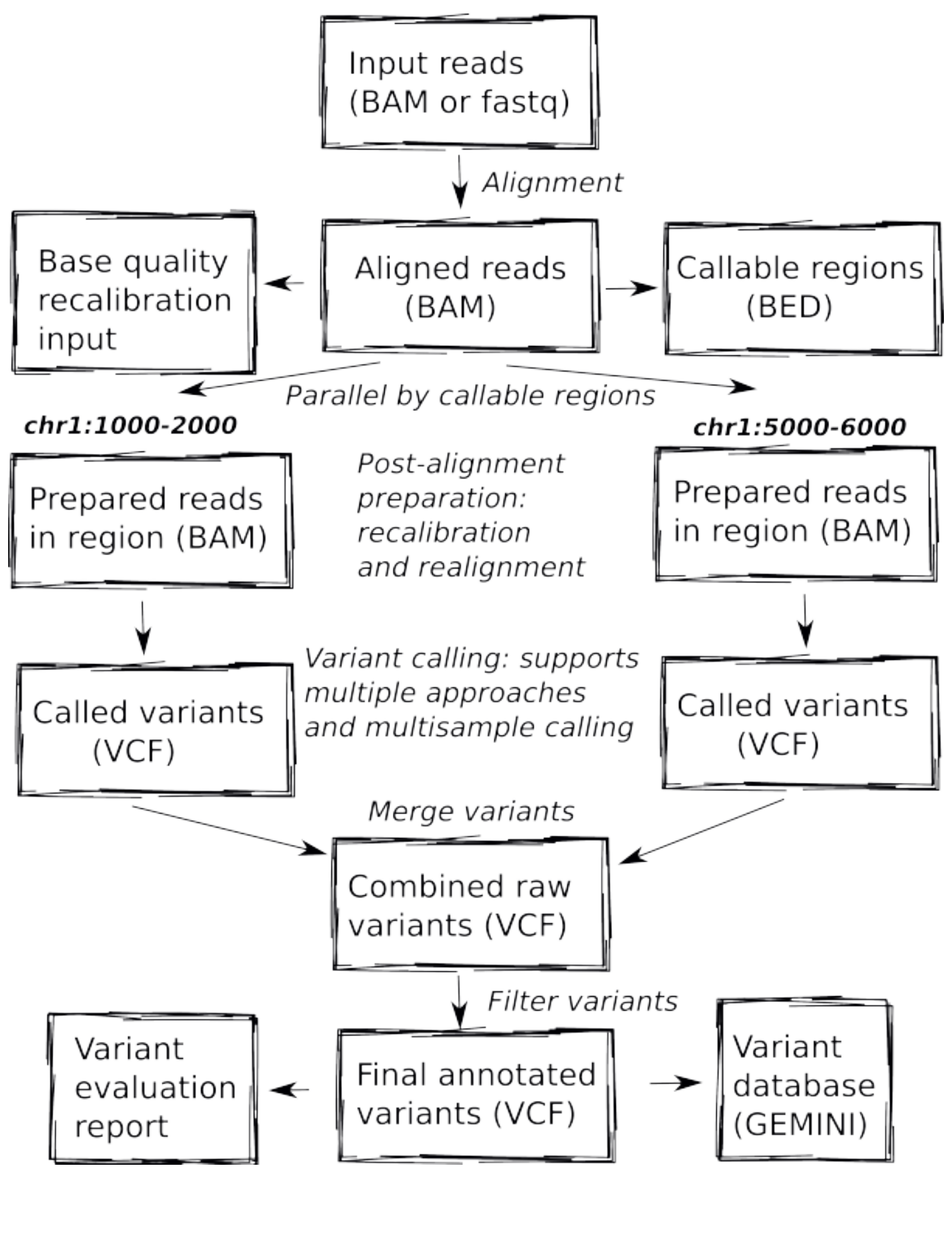
Input reads (BAM or fastq)
Alignment
Base quality recalibration input
Aligned reads (BAM)
Callable regions (BED)
Parallel by callable regions
chr1:1000-2000
chr1:5000-6000
Prepared reads in region (BAM)
Post-alignment preparation: recalibration and realignment
Prepared reads in region (BAM)
Called variants (VCF)
Variant calling: supports multiple approaches and multisample calling
Called variants (VCF)
Merge variants
Combined raw variants (VCF)
Filter variants
Variant evaluation report
Final annotated variants (VCF)
Variant database (GEMINI)

Figure 1: Pipeline workflow.

PIPELINE SET-UP AND IMPLEMENTATION

Our goal for data analysis was to use an approach that could provide a rigorous, well-tested environment for the necessary steps, to ensure the best data quality as possible and reduce the number of possible false positives. We chose bcbio-nextgen (https://github.com/chapmanb/bcbio-nextgen) due to its support of high performance computing (HPC) platforms and the complete implementation of "best practice" quality control and data preprocessing following established standards by the data analysis community. The mode of operation of bcbio-nextgen is outlined in Figure 1.

We then modified bcbio-nextgen in order to support somatic variant calling, where a tumor sample is compared against either a paired reference, or a panel of normal references. Firstly, we added support for preprocessing of paired samples at the same time, to compensate for slight differences between the reference and the tumor. Secondly, support for somatic variant callers was added on top of bcbio-nextgen's existing variant calling framework.

The specific analysis workflow is depicted in Figure 2. Briefly, raw variant calls from the somatic callers are subsequently cleaned from calling artifacts and then reinserted in bcbio-nextgen's downstream processing. Once the pipeline run is complete, calls from different samples are merged together, excluding non-somatic or low quality loci, and annotated using GEMINI (Paila U et al., 2013).

The pipeline was installed on two HPC clusters (208 and 816 CPU cores, respectively), part of the Cloud4CaRE (Cloud for Cancer REsearch) project, a joint effort between the Mario Negri Institute, the ACTO foundation and the IT departments of two major banks to provide computational resources for NGS data analysis.

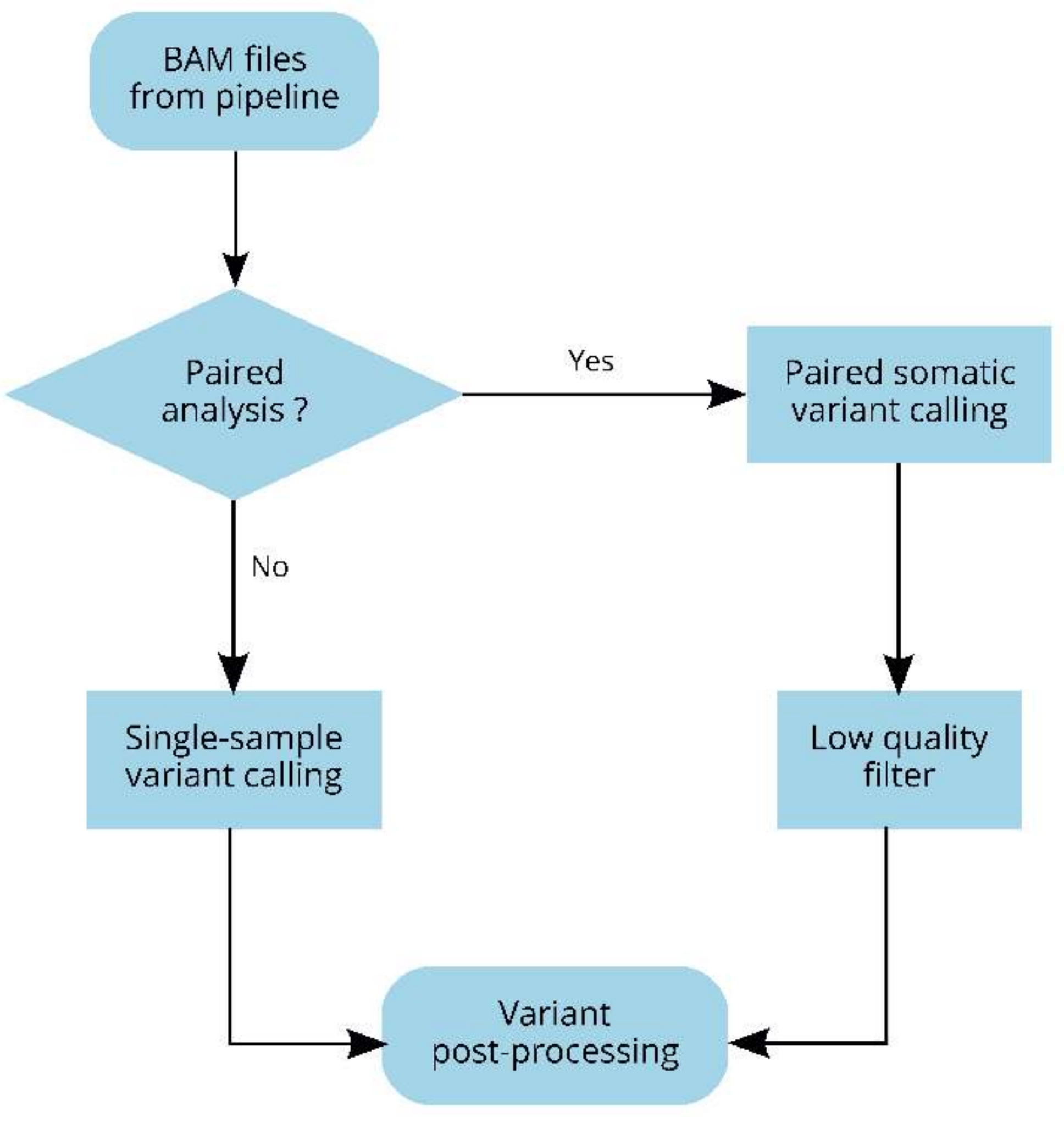

Figure 2: Somatic variant calling workflow.

All the changes made to bcbio-nextgen have been submitted to the upstream source repository and are part of its regular release.

RESULTS

In order to test the proper operation of the pipeline, we initially tested our analysis method on a reduced synthetic data set (2 samples, HCC1143, from The Cancer Genome Atlas benchmark), ensuring that the process ran in a robust and reproducible way.

After the initial tests were complete, we ran the complete pipeline on our EOC data set. The pipeline called a total of **13582** putative somatic variants across 66 samples (PS-O and their matched SCR counterparts), of which **13121** were point mutations (SNPs) and **461** indels. We then filtered out variants with low depth (less than 10 reads) and allelic fraction (less than 1%), yielding **3467** SNPs and **298** indels passing the thresholds.

To verify correct calling from the pipeline, we selected mutations marked as germline and as somatic by the analysis software and verified them with independent methods, either using alternative next generation sequencing platforms (Ion Torrent, Life Technologies), pyrosequencing (Pyromark Q24, QIAGEN) or high-resolution droplet digital PCR (QX200™ Droplet Digital™ PCR System, Bio-Rad) (Table 2).

Gene	Chromosome	Position	Mutation	Type	Mutated fraction	Validation
RB1	chr10	48919236	T → TA	Somatic	25%	Ion Torrent
TP53	chr17	7577506	C → A	Somatic	39%	Ion Torrent
TP53	chr17	7578391	TC → T	Somatic	50%	Ion Torrent
TP53	chr17	7577115	T → C	Germline	10%	Pyrosequencing
BRCA1	chr17	41243941	G → A	Germline	50%	Pyrosequencing
BRCA1	chr17	41249261	G → A	Somatic	3%	Droplet digital PCR

Table 2: Mutations found and validated by the pipeline.

CONCLUSIONS

Identifying somatic mutations from sequencing experiments on highly heterogeneous samples such as solid tumors is a challenging task. We based our approach on using readily available software from the scientific community, to build upon existing expertise. We used multiple variant callers to increase the sensitivity of the analysis, and we were able to successfully validate selected mutations using independent methods. Our results suggest that this approach is sensitive, robust, reproducible and viable for analysis of matched EOC samples.

REFERENCES

Bast RC Jr, Hennessy B, Mills GB. The biology of ovarian cancer: new opportunities for translation. *Nat Rev Cancer*. **2009**; 9(6): 415-28.
Camistra SA. Cancer of the ovary. N Engl J Med. **2004**;351(24):2519-29.
Cancer Genome Atlas Research Network, *et al.* Integrated genomic characterization of endometrial carcinoma. Nature. **2013**; 497(7447): 67-73.
Castellarin M, Milne K, Zeng T, Tse K, Mayo M, Zhao Y, Webb JR, Watson PH, Nelson BH, Holt RA. Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease. J Pathol. **2013**;229(4):515-24.
Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. **2013**; 31(3) :213-9.
Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* [q-bio.GN] **2012**
Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **2012**; 22(3): 568-76.
Lawrenson K, Gayther SA. Ovarian cancer: a clinical challenge that needs some basic answers. *PLoS Med.* **2009**; 6(2):e25.
Marchini S, Fruscio R, Clivio L, Beltrame L, *et al.*. Resistance to platinum-based chemotherapy is associated with epithelial to mesenchymal transition in epithelial ovarian cancer. *Eur J Cancer.* **2013**; 49(2): 520-30.
Murtaza M, Dawson SJ, Tsui DW, Gale D, Forshew T, Piskorz AM, *et al.*. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature.* **2013**; 497(7447):108-12.
O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **2013**; 5(3):28.
Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput Biol*. **2013**; 9(7): e100315
Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics.* **2014;** 28;15(1):244.

IRCCS Istituto di Ricerche Farmacologiche "Mario Negri" Mail: luca.beltrame@marionegri.it Web: http://www.marionegri.it