HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2016 December 06.

Published in final edited form as:

Nat Methods. 2016 July ; 13(7): 584–586. doi:10.1038/nmeth.3893.

Quantitative detection of low-abundance somatic structural variants in normal cells by high throughput sequencing

Wilber Quispe-Tintaya[1], **Tatyana Gorbacheva**[1,2], **Moonsook Lee**[1], **Sergei Makhortov**[3], **Vasily N. Popov**[2], **Jan Vijg**[1,*], and **Alexander Y. Maslov**[1,*]

[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA

[2]Department of Genetics, Cytology, and Bioengineering, Voronezh State University, Voronezh, Russia

[3]Department of Applied and System Software, Voronezh State University, Voronezh, Russia

Abstract

The detection and quantification of low-abundance somatic DNA mutations by high throughput sequencing is challenging because of the difficulty in distinguishing errors from true mutations. While there are several approaches available for analyzing somatic point mutations and small indels, an accurate genome-wide assessment of somatic structural variants (somSVs) in bulk DNA is still not possible. Here we present Structural Variant Search (SVS), a method to accurately detect rare somSVs by low-coverage sequencing. We demonstrate direct quantitative assessment of elevated somSV frequencies induced by known clastogenic compounds in human primary cells.

Genome analysis by high throughput sequencing (HTP-seq) has provided extensive data sets of germline variants in the human and other genomes, as well as detailed information on thousands of somatic mutations in human tumors[1-3]. However, virtually no information is available on somatic mutation frequencies and mutation spectra in normal cells and tissues. This is due to the nature of somatic mutations, which are mostly unique in each cell and virtually indistinguishable from the significant amount of errors associated with every step of HTP-seq, from library preparation to sequencing, sequence alignment, and variant calling[4, 5]. Several approaches were developed for detection of somatic base-pair substitutions and small indels[6-9], but not for somatic structural variants (somSVs), such as large deletions, insertions, inversions, or translocations[5]. Existing computational algorithms for the detection of somSVs in HTP-seq data sets, such as CREST[10], rely on the validation of any variant call by multiple independent supporting sequencing reads spanning the same DNA breakpoint, the junction between two disparate regions of the genome and hallmark of

*Corresponding Authors: Alexander Y. Maslov, alex.maslov@einstein.yu.edu, Jan Vijg, Jan.Vijg@einstein.yu.edu.

Accession Codes: All sequencing data is available on NCBI Sequence Read Archive; accession numbers SRP064739 and SRA304463.

Author Contribution: A.Y.M. and J.V. conceived the idea and wrote the manuscript, W.Q., T.G. and M.L. performed the experiments, A.Y.M., J.V. and V.P. analyzed data, A.Y.M. wrote SVS, S.M. advised on software design.

Competing Financial Interests: The authors declare no competing financial interests.

any SV. However, while this approach can be readily used for the analysis of tumor tissue, — i.e. when somSVs are clonally amplified and therefore present in all or most of the cells — it cannot be applied in detecting ultra-low-abundant somatic SVs, which typically affect only one sequencing read in normal, non-clonal tissue. Here we present Structural Variant Search (SVS) for the quantitative detection of somSVs by ultra-low coverage sequencing.

The key feature of SVS is its ability to definitively call an SV using a single sequencing read that spans the breakpoint without the need for multiple supporting reads. Such high confidence calling of SVs is achieved in two critical steps: a chimera-free library preparation protocol and a novel, non-consensus based SV calling algorithm. Chimeras, i.e., the erroneous concatenation of two genomic fragments during adaptor ligation, occur as unique events spread throughout the sequencing reads and are normally discarded based on the absence of alternative reads covering the same breakpoint. Somatic SVs, however, are themselves spread across the reads as unique events and cannot be distinguished from ligation artifacts. As a method of choice in SVS we use MuPlus, our modification of transposon-based protocol for preparation of sequencing libraries, free from ligation-mediated artifacts[11].

Our SV calling algorithm consists of three steps: (1) identification of potential SVs by taking a split-read approach[12]; (2) filtering out potential technical and mapping artifacts, and (3) separation of somatic and germline SVs based on identification of the latter as identical variants repetitively found in independently prepared sequencing libraries (Online Methods and Fig. 1).

To evaluate specificity and sensitivity of SVS we used the CaSki cell line harboring 47 human papillomavirus (HPV) integration events[13], which are in essence structural variants. SVS analysis of CaSki DNA revealed 20 unique HPV integration sites (Supplementary Tables 1 and 2), 17 (85%) of which were previously described[13]. The remaining three were tested by PCR and two of them found to be genuine (Supplementary Fig. 1). Most likely these two novel HPV integration sites had not been detected previously because of their low abundance, underscoring the unique aspect of SVS in being capable of detecting low-frequency SVs. Thus, this experiment demonstrated 95% specificity and 36.2% sensitivity of SVS in the detection of SVs.

Further, we estimated the lower limit of an SV load still measurable by SVS. Assuming that CaSki is completely homogeneous with no subclonal variation, it can be considered as a model system in which every cell has 47 SVs (23.5 SVs per haploid genome). We found ∼2.83 HPV integration sites per library after sequencing 12 independent libraries, each of which was covering ∼0.28 of the genome (Supplementary Table 1). This is less than the expected 6.58 (23.5*0.28) sites per library, most likely due to heterogeneity of the CaSki cell line and the low-coverage sequencing utilized. Indeed, examination of expected but not found HPV integration sites revealed no breakpoints. Thus, SVS is capable of detecting 47 somatic SVs per cell using ∼0.3× sequencing.

Next, to empirically validate SVS for its capacity to detect somSVs human IMR90 fibroblasts were treated with two different clastogens, bleomycin (BLM) and etoposide

(ETO), applied at three different concentrations. Samples were collected at 72 hours and immediately after treatment, and MuPlus libraries were sequenced on the Ion Proton platform; six to twelve samples were multiplexed on each sequencing run. All identified interchromosomal and intrachromosomal rearrangements (larger than 200nt to avoid possible polymerase slippage[14] and homopolymer artifacts), were considered for further analysis (Supplementary Table 3). SVs found in at least two independent samples were counted as germline SVs. Of note, the number of germline SVs reached a plateau after analyzing ∼50 DNA samples (accumulated coverage ∼17×) out of a total of 70, i.e., at ∼4,000 detected total SVs (Fig. 2a), indicating that the majority of germline variants was discovered. After correction for germline SVs we observed a statistically significant dose-dependent increase of SV frequency in samples collected 72 hours after genotoxic insult (Fig. 2b and Supplementary Fig. 2a).

SVs arise as a consequence of erroneous processing of DNA double-strand breaks (DSBs) induced by the clastogens, probably within hours after the beginning of the treatment[15]. It occurred to us that the SVs detected could in reality be not genuine, but represent artifacts generated during library preparation and/or sequencing due to the presence of damaged DNA fragments. Reasoning that such damage should be at a maximum immediately after treatment, we tested for SVs immediately after the six-hour treatment with these clastogens. The results indicated an elevation of SVs after treatment with bleomycin, but not etoposide (except at the highest concentration) (Fig. 2c and Supplementary Fig. 2b). While bleomycin is a DNA-cleaving agent capable of creating DSBs available for repair immediately upon exposure[16], etoposide is a topoisomerase II poison stabilizing the DNA-enzyme complex and initially creating single-strand breaks[17], which can be later transformed into DSBs. Hence, the elevation of SVs after 6 hours of treatment with bleomycin can be explained by the early emergence of errors during DNA double-strand break repair within this time period.

Due to the unique nature of somatic SVs it is not possible to confirm them independently. However, germline SVs are mostly in databases and can, therefore, function as internal positive controls. We found that 925 out of 1,012 germline SVs in the IMR90 cells (91.4%) (Supplementary Table 4) are listed in the available database of human structural variants (ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/). The remaining 87 were tested using size separation after PCR with SV-specific primer pairs. We confirmed 17 out of 18 germline interchromosomal rearrangements (94.4%) and 66 out of 69 (95.6%) intrachromosomal SVs specific for this particular cell line (Supplementary Fig. 3 and Supplementary Table 5). Of note, only 3.7% of SVs identified by SVS as somatic were found in the database and were therefore false positives.

Analysis for the presence of microhomology (MH) (5nt or larger) at the junction points of the SVs revealed that while germline SVs and background SVs have approximately equal fractions of rearrangements containing MH (1.8% and 1.4% respectively), the BLM and ETO induced SVs are substantially enriched for MHs (4.9% and 3.9% respectively, Supplementary Fig. 4). This suggests that microhomology-mediated end joining is involved in the repair of clastogen-induced DNA DSBs. Next, we analyzed the relative distance between observed SVs and centrosomes of corresponding chromosomes and found that

breakpoints were distributed evenly along the chromosomes (Supplementary Fig. 5). Finally, we analyzed the distribution of SVs' breakpoints across different genomic features and found an ∼2.5 times higher probability for somatic SVs than for germline SVs to reside in transcription factor binding sites and in exonic regions (Supplementary Figs. 6a and 6b). A similar distribution of germline and somatic SVs was also found for DNAse sensitive sites (Supplementary Fig. 6c). Further analysis revealed that germline SVs are depleted from functionally active genomic regions, whereas the frequency of somatic SVs in these regions is higher (∼66% on average) than would be expected assuming random distribution (Supplementary Fig. 6d). This suggests that although euchromatic regions of genome are generally more prone to DNA breakage, the germline SVs, unlike somatic SVs, are under selective pressure that eliminates variants with negative functional consequences.

Further analysis of SV spectra revealed that the fraction of intrachromosomal rearrangements for bleomycin was substantially smaller than for etoposide (22% and 39% of all SVs, respectively; Fig. 2d). Among intrachromosomal rearrangements the fraction of inversions was substantially higher for etoposide than for bleomycin (45% and 30% respectively; Supplementary Fig. 7). This preferential production of translocations by bleomycin and inversions by etoposide may reflect a different mechanism of action between these two clastogens. Notably, the frequency of germline SVs was approximately equal in all tested samples — control and treated with clastogens (Supplementary Fig. 8). Interestingly, we found that 69% of somSVs in control, non-treated samples were interchromosomal rearrangements, compared with 2% among germline SVs (Fig. 2e). This is in agreement with findings by others comparing SVs in tumor with germline SVs[18].

SVS assay, is cost-effective since it does not require high-coverage sequencing and it should enable the characterization of tissue- and age-specific landscapes of SVs in humans and experimental animals. Of note, the assay should also be applicable as a routine genetic toxicology tool to assess clastogenicity of new drugs and chemicals. Finally, SVS will be useful for assessing genome instability as biomarker in aging and disease.

Online Methods

Cell culture and treatment

Human normal lung IMR90 (ATCC CCL-186) fibroblasts and the CaSki (ATCC CRL-1550) cell line were obtained from Einstein Cell Culture Core (Albert Einstein College of Medicine, Bronx, NY, USA) and were not further authenticated. Cells were routinely tested for mycoplasma contamination using MycoAlert PLUS Mycoplasma Detection Kit (Lonza Inc., Allendale, NJ, USA). Cells were maintained in 10% CO_2 and 3% O_2 atmosphere at 37 °C in DMEM (GIBCO, Grand Island, NY, USA) supplemented with 10% FBS (GIBCO). For clastogen treatment experiments serum-free medium with bleomycin (CALBIOCHEM, San Diego, CA, USA) or etoposide (SIGMA, San Louis, MO, USA) was prepared at the time of application from stock solutions of the drugs (10 mg/ml and 25 mg/ml in water or DMSO respectively) and applied for six hours. At the end of application cells were washed with PBS and either harvested or cultured for additional 72 hours in complete medium before harvesting.

DNA isolation, sequencing library preparation and sequencing

DNA from harvested cells was isolated using Quick-gDNA™ Blood MiniPrep (Zymo Research Corporation, Irvine, CA, USA) according to the manufacturer instructions. The barcoded sequencing libraries were prepared using transposon-based MuPlus method, as we described previously[11]. In short, after transposase-mediated fragmentation and tagmentation the first sequencing adaptor was integrated by PCR; the second adaptor was introduced as a single-stranded oligonucleotide after enzymatic cleavage of a complementary part on one strand of the transposon tag. Libraries were size selected on PippinHT apparatus (Sage Science, Inc., Beverly, MA, USA) and quantified using KAPA Library Quantification kit for Ion Torrent (Kapa Biosystems, Inc., Wilmington, MA, USA). Sequencing was performed on the Ion Proton platform (Life Technologies Corporation, Grand Island, NY, USA) using PI sequencing chip (Life Technologies) and Ion PI Sequencing 200 Kit v3 (Life Technologies). Six samples per sequencing chip were multiplexed for etoposide treatment experiment; 12 samples per chip for bleomycin treatment. In the last case the sequencing was performed twice on two different chips with independently prepared libraries (technical replicates).

Data processing and variant calling

Raw sequencing data was aligned to the hg19 human reference genome using the TMAP aligner. Only aligned reads with length more than 120nt were considered for variant calling. Candidate SVs were identified as soft-clipped reads, i.e., reads for which only part (anchor) was successfully aligned to the reference genome. Soft-clipped portions were excised from the original reads and independently aligned to the same reference genome using the TMAP aligner. If the second round of alignment was successful and the length of each aligned portion of the original read exceeded 50nt, this read was considered as candidate SV. The 50nt cut-off was applied to minimize possible aberrant mapping — due to repetitive elements and regions with low complexity, the fraction of the human genome that is uniquely mappable is 79.6% for 30nt sequence tags and 86.7% for 50nt sequence tags[19]. Further filtering of identified candidate SV was performed by comparing mapping results of anchors and corresponding soft-clipped portions of the read. In the case where these two alignments were in agreement with each other, i.e., formed uninterrupted alignment for the entire original read when combined, the read was considered normal. Next, assuming alignment of the anchor was incorrect, local realignment of the anchor was performed to a position on the reference genome defined by alignment of the soft-clipped portion; if successful, such a read was also considered normal. The remaining candidate SVs were accepted as true variants if the mapping quality score of both the anchor and soft-clipped part were not less than 30 (probability of misalignment <0.001 for each, hence probability that both parts of candidate SV are misaligned is less than 10^{-6}) and if these reads were not marked by RepeatMasker[20]. Candidate SVs that did not pass these filters were considered false positives. The frequency of identified SVs was expressed as a number of identified SVs per 1 million of total sequencing reads.

The transposon-based MuPlus library preparation protocol rules out ligation-mediated artificial chimeric sequences. However, this protocol includes a PCR amplification step, which is prone to errors mimicking true SVs, particularly in multi-template settings, such as a sequencing library. Possible mispriming of the transposon-tagged DNA fragments will

lead to the formation of chimeric fragments carrying tag sequences in the middle of the sequencing read (Supplementary Fig. 9a and 9b). The SVS filtering algorithm was designed to eliminate PCR artifacts produced by erroneous priming of transposon-tagged DNA fragments. This is done by finding and eliminating sequencing reads containing the transposon tag TTCGTGCGTCAGTTCA in the middle of the candidate SV.

Identification of germline and clonally expanded somatic structural variants

The key feature of germline structural variants is that unlike somatic SVs, germline SVs are present in each DNA sample of any particular biological subject. In principle it is possible to identify all germline SVs by one round of high-coverage sequencing. However, low-coverage sequencing of multiple samples by SVS will eventually achieve the same result. The SVS algorithm identifies germline SVs as identical variants repetitively found in independently prepared sequencing libraries; only SVs unique for each sample are accepted as somatic variants.

Identification of microhomology in structural variants

We applied an in-house Python script to identify a microhomology sequence at the breakpoints. Microhomology was defined as a stretch of 5 nucleotides or longer flanking or spanning over the breakpoint, and shared between both fragments, forming a particular structural variant (Supplementary Fig. 4a).

PCR validation of structural variants

The PCR primer pairs specific for selected germline SVs and HPV integration sites discovered during analysis were designed using Prime-BLAST software (http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome; Supplementary Table 5). The size of expected PCR products was in the range 100-150bp. The PCR was performed using GoTaq Green Master Mix (Promega Corporation, Madison, WI, USA) and following program: 95°C – 2 min, (95°C – 30 sec, 50°C -30 sec, 72°C – 30 sec)×35, 72°C – 5 min, 4°C – forever. The products of PCR were separated on 2% agarose gel, stained with SYBR Gold nucleic acid gel stain (Life Technologies) and photographed.

Statistical analysis

The descriptive statistics and the P-values were calculated using a two-sample t-test (Microsoft Excel software package).

Code availability

SVS variant caller Python code is available for downloading.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the NIH, AG017242, AG047200, AG038072, the Glenn Foundation for Medical Research (JV); Albert Einstein College of Medicine Human Genome Program Pilot project grant and the

Einstein-Nathan Shock Center of Excellence Pilot and feasibility grant 5P30AG038072–05 (AYM); Ministry of Education and Science of the Russian Federation grant 6.149.2014/K (VNP).

References for Main Text

1. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015; 347:78–81. [PubMed: 25554788]
2. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends Genet. 2013; 29:575–584. [PubMed: 23684843]
3. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015; 349:1483–1489. [PubMed: 26404825]
4. Gundry M, Vijg J. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. Mutat Res. 2012; 729:1–15. [PubMed: 22016070]
5. Maslov AY, Quispe-Tintaya W, Gorbacheva T, White RR, Vijg J. High-throughput sequencing in mutation detection: A new generation of genotoxicity tests? Mutat Res. 2015
6. Gundry M, Li W, Maqbool SB, Vijg J. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. 2012; 40:2032–2040. [PubMed: 22086961]
7. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011; 108:9530–9535. [PubMed: 21586637]
8. Lou DI, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A. 2013; 110:19872–19877. [PubMed: 24243955]
9. Schmitt MW, et al. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012; 109:14508–14513. [PubMed: 22853953]
10. Wang J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods. 2011; 8:652–654. [PubMed: 21666668]
11. Gorbacheva T, Quispe-Tintaya W, Popov VN, Vijg J, Maslov AY. Improved transposon-based library preparation for the Ion Torrent platform. Biotechniques. 2015; 58:200–202. [PubMed: 25861933]
12. Zhang ZD, et al. Identification of genomic indels and structural variations using split reads. BMC Genomics. 2011; 12:375. [PubMed: 21787423]
13. Akagi K, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. 2014; 24:185–199. [PubMed: 24201445]
14. Viguera E, Canceill D, Ehrlich SD. In vitro replication slippage by DNA polymerases from thermophilic organisms. J Mol Biol. 2001; 312:323–333. [PubMed: 11554789]
15. Scarpato R, et al. Kinetics of nuclear phosphorylation (gamma-H2AX) in human lymphocytes treated in vitro with UVB, bleomycin and mitomycin C. Mutagenesis. 2013; 28:465–473. [PubMed: 23696313]
16. Roy B, Hecht SM. Hairpin DNA sequences bound strongly by bleomycin exhibit enhanced double-strand cleavage. J Am Chem Soc. 2014; 136:4382–4393. [PubMed: 24548300]
17. Baranello L, et al. DNA break mapping reveals topoisomerase II activity genome-wide. Int J Mol Sci. 2014; 15:13111–13122. [PubMed: 25056547]
18. Campbell PJ, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008; 40:722–729. [PubMed: 18438408]

Methods-Only References

19. Rozowsky J, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009; 27:66–75. [PubMed: 19122651]
20. Smit, A., Hubley, R. & Green, P. (2013-2015).

Editorial summary

Structural Variant Search, a combination of a chimera-free library preparation and a non-consensus-based SV-calling algorithm, enables the quantitative detection or rare somatic variants.

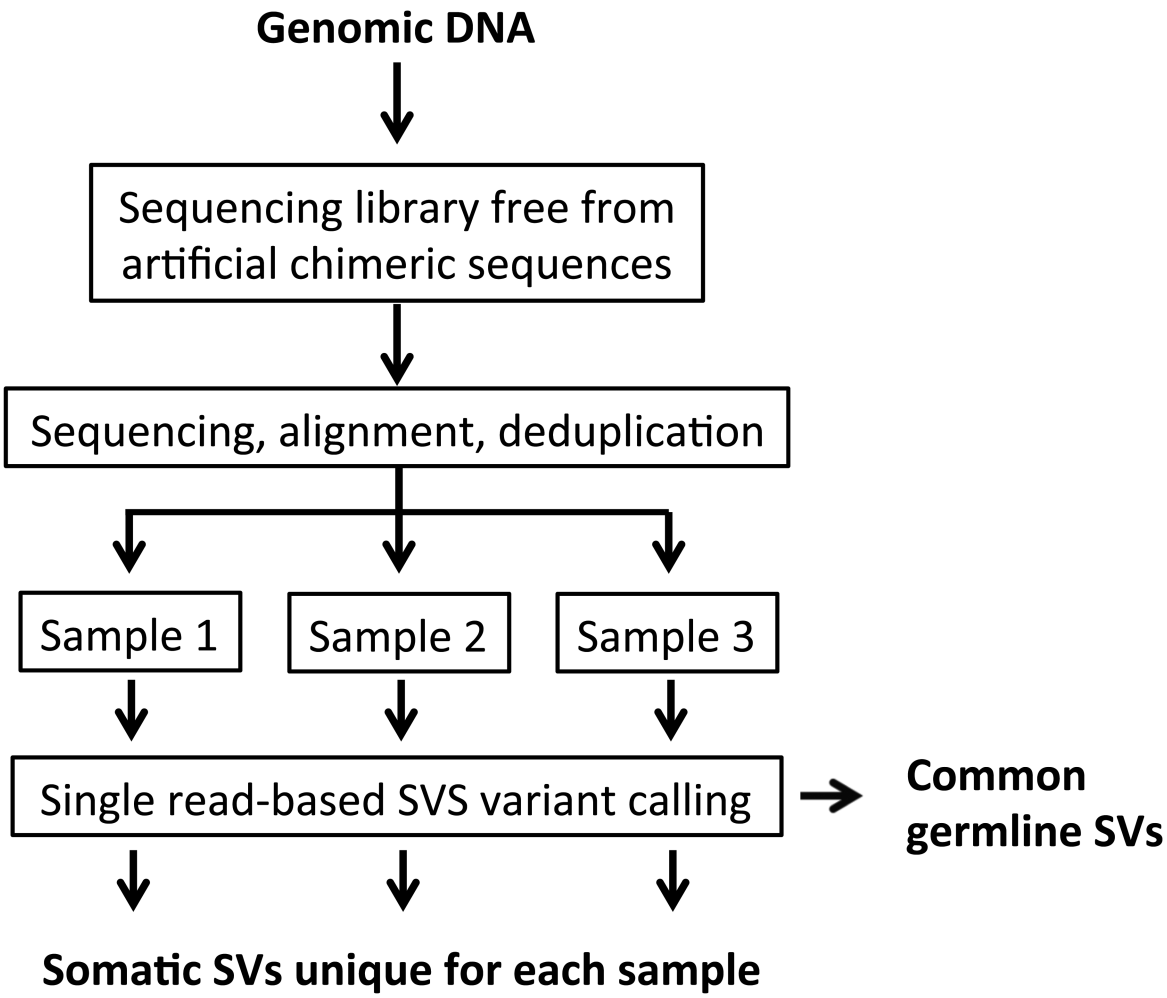

Figure 1.
SVS workflow.

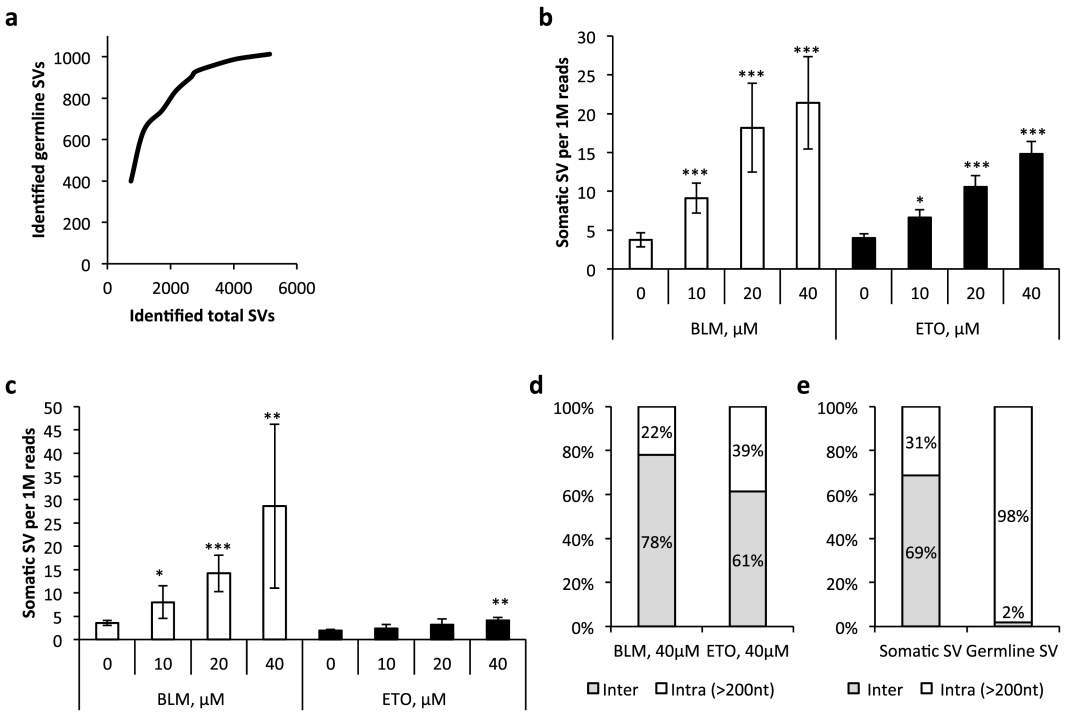

Figure 2.

Quantitative detection of induced structural variants in IMR90 cells. (**a**) Accumulation of discovered germline SVs as a function of identified total SVs. (**b**) Frequency of somatic SVs 72 hours after treatment with bleomycin (BLM) and etoposide (ETO). (**c**) Frequency of somatic SVs immediately after six-hour treatment with clastogens. (**d**) Spectra of somatic SVs induced by different clastogens. (**e**) Spectra of background somatic SVs and germline SVs. All data points represent three biological replicates; data shown as average ± SD; asterisk (*) designates statistically significant difference with corresponding control as determined by two-tail t-test (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).