Sensitive detection of low frequency single nucleotides variants from amplicon and capture sequencing data with Leucippus

Authors: Nikolaos Vasmatzis[1], Chen Wang[1], Sarah E. Kerr[2], Jamie N. Bakkum-Gamez[3], Flora M. Vaccarino[4], Alexej Abyzov[1]

Affiliations: [1]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN; [2]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; Department of Obstetrics & Gynecology, Mayo Clinic, Rochester, MN; [4]Child Study Center, Yale University, New Haven, CT

Abstract

Based on existing overwhelming evidence, the leading paradigm in the field of medical genomics states that the primary cause of all cancers is genomic alteration in the somatic cells of an individual. Such alterations include Copy Number Alterations (CNA), single nucleotide variants (SNVs), small insertions and deletions (indels), and chromosome fusions or translocations. SNV is the most common and best understood type of alteration leading to cancer. Sensitive detection (i.e., when a variant is present in a minority of analyzed cells) of cancer relevant SNVs is important for cancer screening, cancer diagnostics (particularly at early stages), and choosing among various treatment options. However, existing approaches based on Next Generation Sequencing (NGS) data are aimed at detecting variants that are present in all or most cells sequenced during each experimental assay.

We therefore developed an analytical approach for sensitive detection of somatic variants from deep sequencing of captured or amplified genomic regions. The approach is based on reducing multiple hypotheses testing by interrogating only the most relevant sites (sites of interest), which are a small fraction of all sites, designing experiments in such a way that paired reads overlap, correcting read sequencing errors at the overlapping 3'-ends, and empirically estimating sequencing error rate. We implemented this approach in Leucippus software (freely available at GitHub) and experimentally validated it with orthogonal technique digital droplet PCR. Validation demonstrated that SNVs with allele frequency as low as 0.1% can be detected with Leucippus.

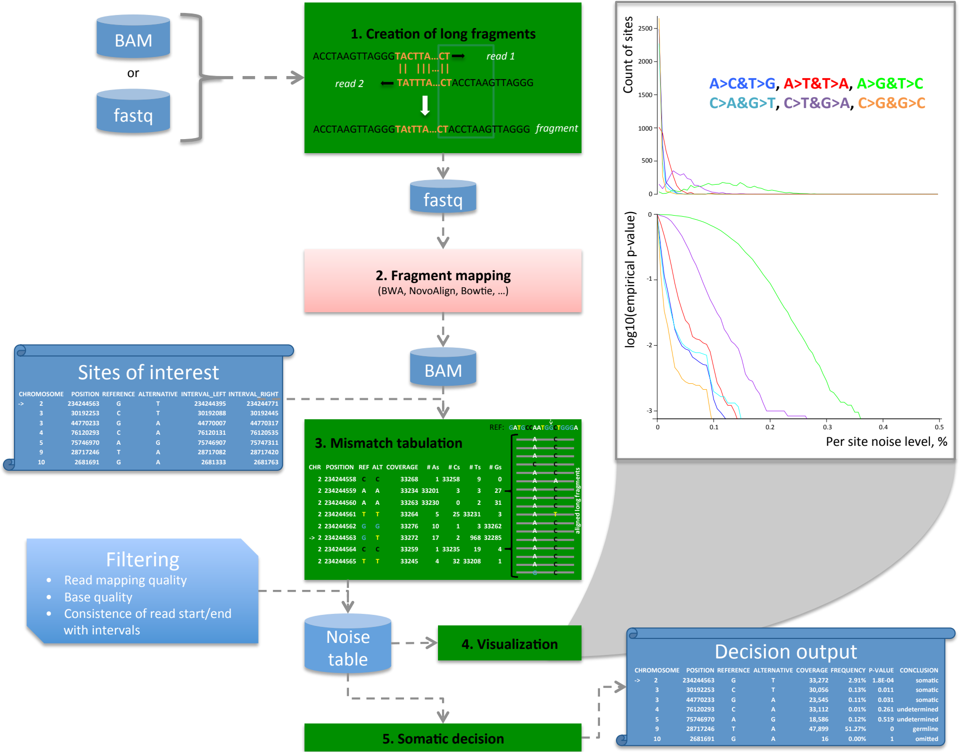
We then applied the approach to clinical samples related to endometrial cancer and detected protein truncating SNVs in genes known to be involved in this cancer. Leucippus is an approach and software that can be used in both research and clinical settings for sensitive detection of somatic SNVs.

Approach

The schematic of the approach is depicted in **Figure 1**. At first step one constructs long fragments from overlapping paired reads. To find overlap, reads are sled against each other by one base pair at a time. Matched and mismatched nucleotides in the overlap are counted. Then a statistical test, using binomial coefficients, is performed to calculate the probability that such or a larger count of matches for the overlap length can occur by

adfa, p. 1, 2011.

chance. Here a random chance for a base match is set to 0.25. The best overlap is the one with the smallest such probability. The pair of reads with the best overlap of at least 50 nucleotides in length and at least 75% of matched nucleotides is used to construct a long fragment. Nucleotides in the overlapping part of both reads are constructed in the following way: i) if nucleotides match, then this nucleotide is assigned at the current position, and its quality is equal to the sum of the nucleotide qualities at this position in each read; ii) if nucleotides mismatch, then the one with the higher quality is assigned at the current position, while its quality is reduced by the quality of the other nucleotide. If the qualities are equal, one of the two nucleotides is chosen randomly and its quality is assigned zero. At step 2 a user can use their favorite mapping alignment software to align created long fragments to the reference genome.



Sites of interest

	CHROMOSOME	POSITION	REFERENCE	ALTERNATIVE	INTERVAL_LEFT	INTERVAL_RIGHT
->	2	234244563	G	T	234244395	234244771
	3	30192253	C	T	30192088	30192445
	3	44770233	G	A	44770007	44770317
	4	76120293	C	A	76120131	76120535
	5	75746970	A	G	75746907	75747311
	9	28717246	T	A	28717082	28717420
	10	2681691	G	A	2681333	2681763

3. Mismatch tabulation

	CHR	POSITION	REF	ALT	COVERAGE	# As	# Cs	# Ts	# Gs
	2	234244558	C	C	33268	1	33258	9	0
	2	234244559	A	A	33234	33201	3	3	27
	2	234244560	A	A	33263	33230	0	2	31
	2	234244561	T	T	33264	5	25	33231	3
	2	234244562	G	G	33276	10	1	3	33262
->	2	234244563	G	T	33272	17	2	968	32285
	2	234244564	C	C	33259	1	33235	19	4
	2	234244565	T	T	33245	4	32	33208	1

Decision output

	CHROMOSOME	POSITION	REFERENCE	ALTERNATIVE	COVERAGE	FREQUENCY	P-VALUE	CONCLUSION
->	2	234244563	G	T	33,272	2.91%	1.8E-04	somatic
	3	30192253	C	T	30,056	0.13%	0.011	somatic
	3	44770233	G	A	23,545	0.11%	0.031	somatic
	4	76120293	C	A	33,112	0.01%	0.261	undetermined
	5	75746970	A	G	18,586	0.12%	0.519	undetermined
	9	28717246	T	A	47,899	51.27%	0	germline
	10	2681691	G	A	16	0.00%	1	omitted

Figure 1. Schematics of the approach. At step 1, the overlapping algorithm is used to construct long reads obtained from bam or fastq files. The initial read pairs are sled against each other with a gradually increased overlapping window to find best overlap. Positions with mismatches are assigned lower base quality score. At step 2, a user can use their favorite mapping alignment software to align created long fragments to the reference genome. At step 3, mismatch tabulation is done for all sites of interest and surrounding intervals provided in the input files. The resulting table lists counts and coverage for each position and nucleotide type. At step 4, one can create graphs visualizing mutation frequency and empirical p-values for noise level. At this and the next steps sites of interest are excluded and mismatches for the remaining sites are considered as background noise. At step 5, decision (somatic, germline, or undetermined) about SNVs at sites of interest is made by comparing frequency of the SNVs with background noise. Low coverage sites are omitted

At third step one tabulates sequencing error (e.g., "noise") for sites targeted by amplicon or capture sequencing. This information is used later to call mosaic sites. In essence, the information tabulated is a comprehensive count of nucleotides covering a particular position. Therefore, the required inputs are a list of targeted sites and .bam files with aligned reads. The table includes all necessary information that is used to visualize noise and make a decision about sites of interest being mosaic.

At fourth step, to facilitate data quality control and interpretation, one can create graphs for substitution rates and p-values of background noise reaching particular values, for each nucleotide substitution type (i.e., C to T, A to G, etc.). The graphs are created from the noise table generated at the previous step, and excluding the sites of interest (where nucleotide expected and nucleotide alternative are not the same). Mismatches to the reference base in aligned long reads are assumed to be noise, either from sample preparation (e.g., amplification) or sequencing. A mutation rate graph shows the frequency of sites versus rate of a particular substitution type (e.g., C to T). A p-value graph shows the proportion of sites having substitution rate larger than a value. For example, for the expected nucleotide C, substitution type C to T, and substitution frequency of 0.01 the p-value is the fraction of C-sites having T substitution frequency larger than 0.01.

At fifth step one decides whether sites of interest are germline or somatic. By default, sites with at least 35% non-reference nucleotide coverage are termed germline. The decision to call a variant as somatic is done by comparing the frequency of expected non reference nucleotide with the background substitution rate of reference to expected nucleotide. The background rate is derived from all sites in the noise table, excluding sites of interest and likely germline variants that have more than 10% of non-reference nucleotide coverage. By default (but changeable), sites with p-value of less than 0.05 are deemed somatic. Also, by default (but changeable), sites with read coverage less than 100 reads are deemed as having insufficient data and omitted from decision process. The remaining sites are called undetermined, i.e., they are not germline, but frequency of alternative nucleotide is consistent with background.

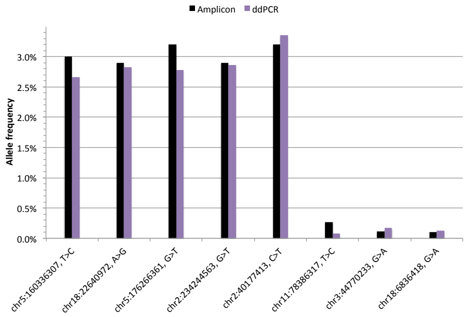


Figure 2. Experimental validation of the approach. Site amplification with ultra-deep sequencing (amplicon-seq) and ddPCR reactions revealed excellent concordance in allele frequency estimates. The customized ddPCR assays confirmed as somatic all SNVs defined as such by amplicon-seq.

Leucippus software

Leucippus is a Java program aimed at determining mosaic SNVs and estimating their allelic frequency at specified sites of interest from amplicon and capture sequencing data. The software implements analytical part for the approach outlined in **Figure 1**. Each step is implemented as a separate command. The software is freely available to GitHub (https://github.com/abyzovlab/Leucippus).

Validation

To test the approach we applied it to confirm candidate somatic SNVs discovered from clonal expansion of human induced Pluripotent Stem Cell lines (hiPSC) [4]. We conducted site amplification with ultra-deep sequencing (amplcon-seq) experiment for 69 SNVs. 16 of these SNVs were defined as somatic with cell frequency as low as 0.1%. Next, we used digital droplet PCR (ddPCR) as a mean of validation of our results for 8 randomly selected somatic SNVs with low cell frequency. Remarkably, all the SNVs were validated (**Figure 2**). Allele frequencies from amplicon-seq was also in good agreement with those from ddPCR but were more deviant for low frequency SNVs (AF ~0.1%), as those have few supporting counts in both experiments and are estimated with larger statistical error.

Application

To demonstrate clinical utility of the approach we applied it to amplicon data aimed at detecting low frequency variants likely disrupting 19 genes implicated in endometrial cancer. For those genes we predicted all possible mutations that will cause a stop codon and consider them as sites of interest. Applying Leucippus to 45 samples with the data revealed 90 SNVs (with allele frequency >1%) causing stop codons in 10 genes (FBXW7, ARID1A, ARID5B, PTEN, POLD1, PIK3R1, POLE, PIK3CA, FGFR2, and CDKN2A) of 30 samples. Remarkably, 17 stop mutations that were known to exist in these samples were all detected. Additionally, the list of mutated genes was highly concordant with the one of highly mutated genes in endometrial cancer reported previously [5].

References

1. Crowley, E., Di Nicolantonio, F., Loupakis, F., Bardelli, A.: Liquid biopsy: monitoring cancer-genetics in the blood. Nat Rev Clin Oncol. 10, 472–484 (2013).
2. Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C.W., Modlin, L.A., Liu, C.L., Neal, J.W., Wakelee, H.A., Merritt, R.E., Shrager, J.B., Loo, B.W., Alizadeh, A.A., Diehn, M.: An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat. Med. (2014).
3. Murtaza, M., Dawson, S.-J., Pogrebniak, K., Rueda, O.M., Provenzano, E., Grant, J., Chin, S.-F., Tsui, D.W.Y., Marass, F., Gale, D., Ali, H.R., Shah, P., Contente-Cuomo, T., Farahani, H., Shumansky, K., Kingsbury, Z., Humphray, S., Bentley, D., Shah, S.P., Wallis, M., Rosenfeld, N., Caldas, C.: Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. Nat Commun. 6, 8760 (2015).
4. Abyzov, Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., Kocabas, A., Calixto, N.E., Grigorenko, E.L., Huttner, A., Chawarska, K., Weissman, S., Urban, A.E., Gerstein, M., Vaccarino, F.M.: Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature. 492, 438–442 (2012).
5. Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., Yau, C., Laird, P.W., Ding, L., Zhang, W., Mills, G.B., Kucherlapati, R., Mardis, E.R., Levine, D.A.: Integrated genomic characterization of endometrial carcinoma. Nature. 497, 67–73 (2013).