Verification of Structural Variants in Tumor Cells using EGM

Introduction

Structural variants (SVs) play a crucial role in both the onset and progression of cancer, as well as in the development of drug resistance over time. Despite their importance, identifying and validating SVs remains a significant challenge due to the limitations of current sequencing and other genomic technologies. Even when multiple platforms and analysis algorithms are utilized, the accuracy and reliability of SV detection often remain uncertain. While variants smaller than 50 base pairs (bp) are usually identified with confidence, larger SVs—those exceeding 300 bp—become increasingly difficult to characterize as their size grows. The ability to detect these larger SVs depends on their size, type, and genomic context. Cancer genomes add further complexity, often exhibiting abnormal chromosome counts, numerous somatic variants with varying allele frequencies, and intricate rearrangements caused by events like chromothripsis. These challenges underscore the urgent need for novel methods to reliably detect and validate SVs, which are critical for advancing our understanding of cancer biology.

This technical note examines the verification of high-confidence structural variants (hcSVs) in tumor cells using electronic genome mapping (EGM). Previous work by Talsania et al. (2022) identified 1,788 hcSVs through a multi-platform approach. These variants were further analyzed using multiple validation methods, yet fewer than 25% were confirmed, highlighting persistent challenges in SV detection. Given the complexity of structural variants in cancer genomes, where abnormal chromosomal structures and varying allele frequencies complicate analysis, we leveraged EGM to independently verify hcSVs across a wide size range. Our findings demonstrate EGM's utility in confirming SVs that were difficult to validate with conventional methods, reinforcing its role as a complementary technology for SV analysis.

Multi-Platform Approach for SV Detection

A significant effort to address the complexity of SV detection was conducted by Talsania et al. (2022), who employed a multi-platform approach to unravel the SV landscape in two cell lines, the normal HCC1395BL breast cell line and the matching breast tumor cell line HCC1395. Their study integrated five genomic technologies—Illumina short-read sequencing (SRS), 10X Genomics linked reads, PacBio long-read sequencing (LRS), Oxford Nanopore LRS, and Hi-C chromosome conformation capture—coupled with multiple SV detection algorithms for each platform (Figure 1). This comprehensive strategy resulted in a set of 1,788 hcSVs. An SV was included in the hcSV set only if it was detected by at least two of the five platforms. Notably, over two-thirds of the hcSVs were identified by just two platforms, and only 17 of the 1,788 hcSVs were detected across all five platforms.

SVs were classified by the number and type of genomic platforms that identified them. Criteria for high-confidence and validated SVs were defined in Talsania et al. (2022). To

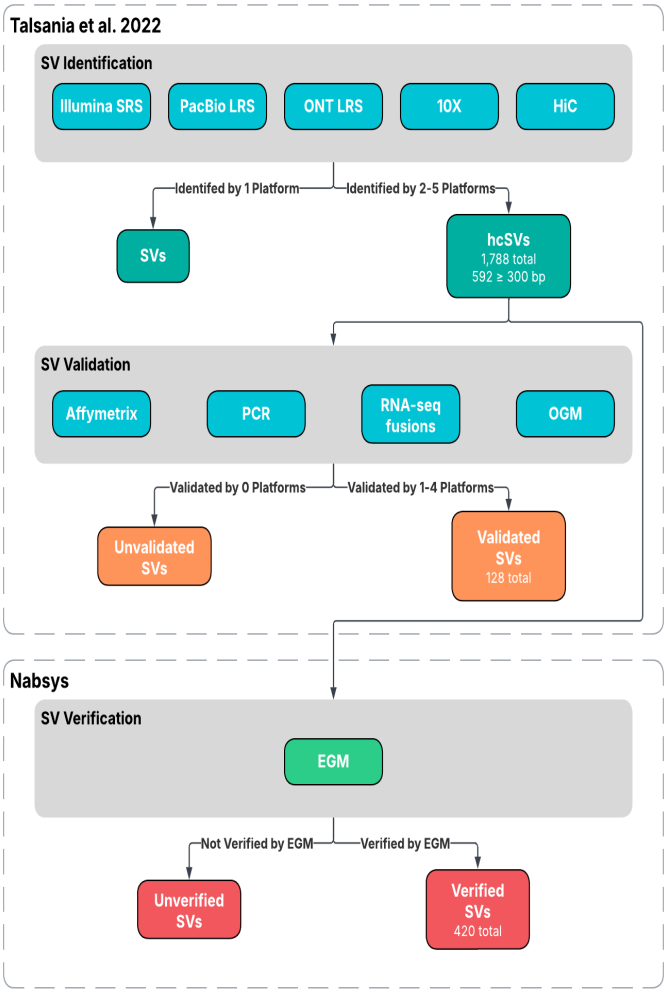

Figure 1: Multi-technology analysis of SVs.

categorize an SV as validated, at least one of the second set of platforms was required to concur with the original call. We defined verified SVs as those SVs identified as positive after additional examination by EGM. Only results from hcSVs were reported here.

Once the hcSVs were identified, the researchers sought to validate them using additional technologies, including PCR, Affymetrix microarrays, Optical Genome Mapping (OGM), and RNA-seq for fusion detection. Validation was defined as confirmation by at least one additional platform. Each technology comes with inherent strengths and limitations depending on the size and type of SV being analyzed. Despite leveraging this diverse array of tools, only 128 of the 1,788 hcSVs were successfully validated, highlighting the persistent technical challenges in SV detection and verification. This study underscores the complexity of accurately identifying and validating SVs, especially in cancer genomes, and highlights the pressing need for more advanced methodologies to overcome current technological barriers.

The technologies used to identify hcSVs in this study all provide data at base-pair resolution. However, all technologies used in the subsequent analyses, with the exception of PCR, provide lower resolution for the position and size of SVs, which poses a challenge for comparing SVs across technologies. This issue compounds the already complex task of defining SVs using a single technology.

Verifying SV Calls Using EGM

EGM's current interval size resolution limit is approximately 300 bp, so only high-confidence deletions, insertions, and duplications > 300 bp were evaluated. This reduced the hcSV set from 1,788 to 592 insertions and deletions for examination. Using the publication-defined matching parameters, these insertions and deletions were analyzed using EGM. Since the researchers focused on a tumor cell line, allele frequencies (AFs) for variants varied widely due to the somatic nature of the SVs, potentially leading to missed variants with low AFs. Additionally, only a subset of hcSVs could be validated using the secondary alternative technologies, leaving the possibility that some hcSVs may be artifacts.

After the hcSV set was identified, they were examined with multiple technologies including OGM. The OGM rare variant pipeline used in this analysis is not intended for SVs less than 5,000 bp, reducing the set of hcSVs that can be compared to 260 hcSVs. Of these, 94 were validated by OGM, 24 were validated by other technologies but not OGM, and 142 were not validated by any of these secondary technologies. In contrast, 203 of the 260 insertions/deletions larger than 5,000 bp were verified by EGM. In the bigger set of 592 hcSVs greater than 300 bp, 420 were verified by EGM while 128 were validated by the other four secondary technologies.

With EGM, insertions generally exhibit higher verification rates than deletions, with shorter deletions having the lowest verification rate. Longer deletions, however, are verified at rates comparable to insertions. Further detail for the concordance of EGM with the LRS technologies is summarized in Table 2. Among the 592 hcSVs > 300 bp, only 72% were identified by both PacBio and ONT, while 21% of the hcSVs received discordant calls between PacBio and ONT LRS technologies. Another 7% of hcSVs were flagged by neither LRS method but were identified by at least two other technologies. The highest EGM verification rate was with hcSVs that were identified with both LRS technologies with lower verification rates among the other categories. Overall, EGM demonstrates strong performance in verifying validated SVs across most size ranges, with performance similar to other LRS technologies.

	Insertions						Deletions						Ins/Dels
SV size >	300-1,000 bp		1,001-5,000 bp		>5,000 bp		300-1,000 bp		1,001-5,000 bp		>5,000 bp		All Sizes >300 bp
Technology	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number
All hcSVs	113	100	70	100	186	100	94	100	55	100	74	100	592
EGM	103	91	57	81	146	78	28	30	29	53	57	77	420
OGM	NA	NA	NA	NA	39	21	NA	NA	NA	NA	55	74	94
PacBio	106	94	53	76	119	64	89	95	46	84	58	78	471
ONT	112	99	67	96	142	76	89	95	49	89	51	69	510
Illumina	9	8	36	51	178	96	39	41	43	78	72	97	377
HiC	0	0	0	0	6	3	0	0	0	0	8	11	14
10X	0	0	0	0	86	46	0	0	0	0	42	57	128

Table 1: High-confidence structural variant calls by size and technology. The ability of different technologies to identify SVs of different sizes and types is shown for hcSVs. The five technologies used for identification comprise the bottom five rows and EGM and OGM are shown above with the number of SVs identified and the percentage of each class. The Bionano rare variant analysis pipeline is not intended to be used to identify SVs smaller than 5,000 bp.

Effects of Allele Frequency

To evaluate whether AF influenced hcSV verification by EGM, AFs were estimated using ONT (Oxford Nanopore Technology) data from the original study. This included read counts generated by the SV-calling program Sniffles. For hcSVs identified by Sniffles, AFs were calculated for both verified and unverified hcSVs. Among the hcSVs, AF data was available for 110 of ONT calls not verified by EGM and 339 of 420 calls verified by EGM. The other hcSVs were not identified by Sniffles.[2] The median AF was 36% for unverified calls and 41% for verified calls, with both groups including SVs spanning AFs from less than 10% to 100%. These results suggest that AF had minimal impact on the verification of hcSVs.

Summary

EGM demonstrates its capacity to verify somatic SVs from tumor-derived cell lines across a wide size range, spanning from 300 bp to several megabases (Mb). While EGM's resolution becomes limited for SVs smaller than 300 bp, it can still verify variants as small as 100 bp, depending on the sequence context. As outlined here, EGM serves as a valuable complementary technology for confirming SVs across diverse size ranges with performance similar to other LRS technologies. Its straightforward workflow and cost-effectiveness make it an appealing option today. Future advancements in algorithms and data quality are expected to enhance EGM's performance further, solidifying its role as a crucial tool in cancer genomics research.

SV Size	Insertion											Deletion									
LRS Calls	Both LRS		PacBio Only		ONT Only		Neither		All Calls		% Called	Both LRS		PacBio Only		ONT Only		Neither		All Calls		% Called
EGM	+	-	+	-	+	-	+	-	+	-		+	-	+	-	+	-	+	-	+	-	
300-1,000	96	9	1	0	6	1	0	0	103	10	91	27	57	0	5	1	4	0	0	28	66	30
1,001-4,000	47	3	1	2	9	8	0	0	57	13	81	23	17	3	3	3	6	0	0	29	26	53
4,001-10,000	37	2	1	1	14	4	0	0	52	7	88	8	3	0	0	0	0	0	0	8	3	73
10,001-100K	26	9	4	0	11	7	10	0	51	16	76	18	5	3	3	2	1	4	3	27	12	69
>100K	25	4	5	5	3	0	10	8	43	17	72	10	2	6	0	2	0	4	0	22	2	92
All	231	27	12	8	43	20	20	8	306	63	83	86	84	12	11	8	11	8	3	114	109	51

Table 2: Verification of high-confidence structural variants by technology. To analyze hcSVs, the variants were categorized by type (insertions or deletions) and further divided into five size-based bins containing similar numbers of hcSVs. All hcSVs were initially identified by two or more technologies, including 72% found via both PacBio and ONT. However, there was a subset found by only one (14%) or neither (7%) LRS technologies, but were also found by one or more other technologies. These SVs were then assessed using EGM. + indicates the number of calls that EGM was able to confirm, while - indicates calls that EGM did not confirm. Overall, EGM showed high concordance with LRS calls.

References

1. Talsania K, Shen TW, Chen X, et al. *Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies. Genome Biol*. 2022;23(1):255. Published 2022 Dec 13. doi:10.1186/s13059-022-02816-6
2. Sedlazeck, F.J., Rescheneder, P., Smolka, M. et al. *Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods* 15, 461–468 (2018). https://doi.org/10.1038/s41592-018-0001-7