CRISPR/Cas9 Customization of Electronic Genome Mapping Repeat Expansion Assays

John F. Thompson,[1] Syndi Koltz,[1] Lindsay Opsahl,[1] Wentao Huang,[1] Dong Zhang,[1] Xu Tan,[1] Shuk Shukor,[2] Mike Kaiser[1]
[1]Nabsys LLC, Applications, Providence, RI; [2]Hitachi High Tech America

Abstract

Background: Many genetic disorders are caused by expansions of simple sequence repeats, including Fragile X syndrome (FXS) and Friedreich's ataxia (FA). Accurate sizing of these expansions is critical. However, conventional PCR and sequencing methods have limited sensitivity and are unreliable for long repeat regions. Electronic genome mapping (EGM) uses solid-state nanodetectors to survey long DNA molecules and detect structural variants (SVs), including those caused by repeat expansions and contractions.

Material and Methods: We assessed EGM's ability to detect *FMR1* and *FXN* repeat expansions using Coriell cell lines. Single injections and replicates were analyzed using two bioinformatics pipelines: Human Chromosome Explorer® (HCE) and the RepX™ Repeat Expansion Analysis (RepX) pipeline were optimized by using CRISPR/Cas9 to add and remove tag sites as needed.

Results: A single detector with a single sample loading generated sufficient data to call both *FMR1* and *FXN* repeats using the RepX pipeline. Multiple replicates of multiple samples were concordant with expectations. These findings confirm the accuracy and robustness of EGM for repeat expansions that are challenging to analyze by standard techniques. The pipeline should also be able to detect many other repeat expansions. Some repeat expansion disorders, such as FA, result in mosaicism and we will discuss our progress with its detection. For repeat expansions whose tagging patterns are not optimal, the assay can be customized using CRISPR/Cas9 to create or remove tag sites as needed.

Conclusion: EGM offers integrated workflows for high-resolution analysis of repeat regions in either whole-genome or gene-specific mode. This approach enables efficient, scalable detection of repeat expansions.

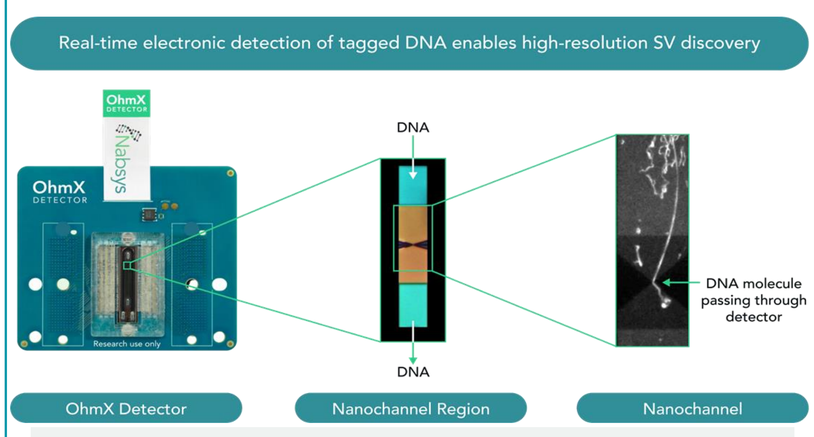

Figure 1: EGM detector. EGM is enabled by the novel, solid-state nanodetectors developed by Nabsys. The OhmX Detector houses 256 parallel nanochannels, each with its own electronic sensor. Nabsys has developed DNA tagging chemistry that provides a high signal-to-noise. The high signal and spatial resolution allows tags to be closely spaced on the high molecular weight DNA molecules. This yields optimal tag density such that structural variations as small as 300 bp can be routinely identified and mapped genome-wide.

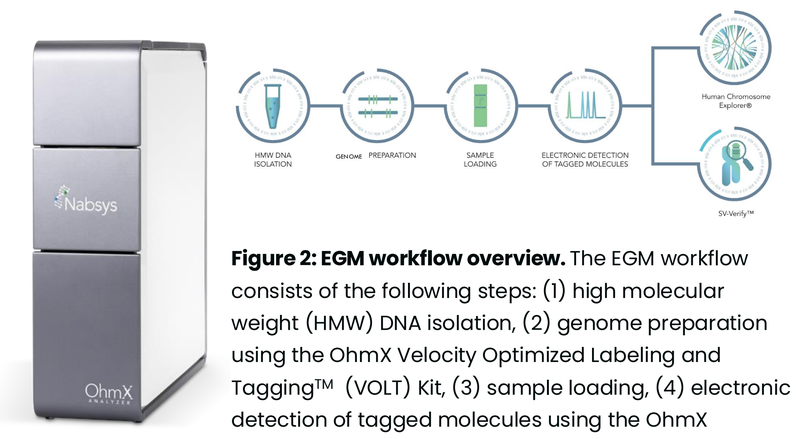

Figure 2: EGM workflow overview. The EGM workflow consists of the following steps: (1) high molecular weight (HMW) DNA isolation, (2) genome preparation using the OhmX Velocity Optimized Labeling and Tagging™ (VOLT) Kit, (3) sample loading, (4) electronic detection of tagged molecules using the OhmX Analyzer, and (5) data analysis.

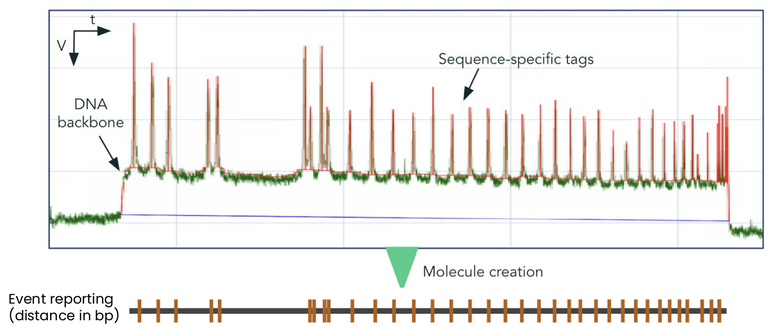

Figure 3: Electronic profile for tagged DNA molecules over 150 kb. The baseline voltage is shown on the left. When the DNA molecules enter the nanochannel, a change in voltage is observed. Whenever a site-specific tag on the DNA is also present, the voltage changes again. The time between tags is converted to distance in base pairs.

RepX Analysis for Repeat Disorders

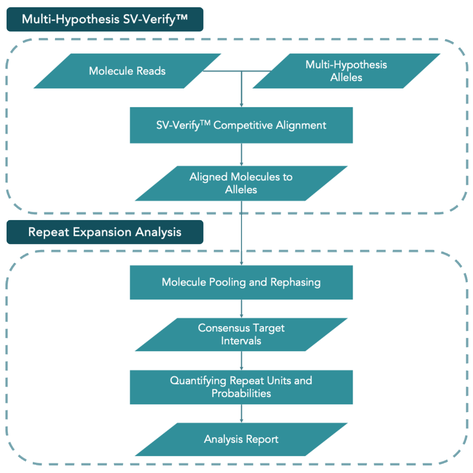

Figure 4: Methodology for RepX pipeline. In the first stage, DNA molecule reads are analyzed for competitive alignment to multiple alleles with various repeat expansion lengths (multi-hypothesis alleles). In the second stage, aligned molecules are pooled, phased, and quantified for repeat expansion metrics. Output provides repeat unit counts, 95% confidence intervals, as well as probabilities of elevated and full mutation. This pipeline is customizable for other target genes and regions suspected to contain repeat expansions.

RepX *FMR1* Analysis Concordance

Sample	Sex	Repeat Annotation	Run #	Coverage (X)	Group	Allele A					Allele B					Concordance to Annotation*
						P(Elevated)	P(Full Mutation)	Repeat Count	95% C.I.	Repeat Coverage (X)	P(Elevated)	P(Full Mutation)	Repeat Count	95% C.I.	Repeat Coverage (X)	
HG003	Male	<54	1	49	Negative Control	44.50%	0.02%	38	[0,114]	16						
			2	65		29.80%	0.00%	13	[0,97]	15						
GM12891	Male	<54	1	80		44.20%	0.00%	41	[0,108]	20						
GM12892	Female	<54	1	56		55.90%	0.11%	51	[0,131]	14	56.60%	0.17%	51	[0,135]	13	
			2	65		64.20%	0.24%	63	[0,142]	14	64.20%	0.24%	63	[0,142]	14	
HG004	Female	<54	1	91		45.20%	0.02%	39	[0,114]	16	46.10%	0.03%	39	[0,117]	15	
			2	60		19.30%	0.00%	0	[0,82]	14	19.30%	0.00%	0	[0,82]	14	
GM20233	Male	117	1	88	Elevated	95.00%	12.20%	136	[37,240]	11						Yes
			2	76		96.00%	1.55%	117	[46,190]	22						Yes
GM06968	Female	107 &32	1	146		100.00%	10.30%	165	[114,217]	44	5.53%	0.00%	0	[0,54]	29	Yes
GM06897	Male	477	1	50	Full Mutation	100.00%	100.00%	471	[384,562]	17						Yes
			2	75		100.00%	100.00%	577	[471,687]	12						Yes
			3	86		100.00%	100.00%	471	[396,548]	23						Yes
GM07861	Male	351-400	1	49		100.00%	99.90%	371	[265,482]	11						Yes
			2	56		100.00%	100.00%	409	[320,501]	16						Yes
			3	64		100.00%	100.00%	372	[289,458]	18						Yes
GM07862	Male	501-550	1	42		100.00%	100.00%	569	[463,679]	12						Yes
			2	80		100.00%	100.00%	581	[500,663]	21						Yes

Table 1: RepX-derived repeat counts and pathogenic probabilities for FXS. The RepX Analysis output summary for control, elevated, and full mutation samples. Reported values include the observed repeat count (maximum *a posteriori* estimate), repeat-spanning molecule coverage, the probability that the count exceeds the full mutation threshold (P(Full Mutation)), the probability that the count exceeds the elevated threshold (P(Elevated)), and the 95% confidence interval. Allele B for male samples are greyed out to indicate the lack of a second chrX allele. Columns P(Elevated) and P(Full Mutation) refer to probabilities of observed repeat counts being between 55–150 units and > 200 units, respectively. *Annotation from Coriell website used to determine concordance.

FXN Repeat Length Variability

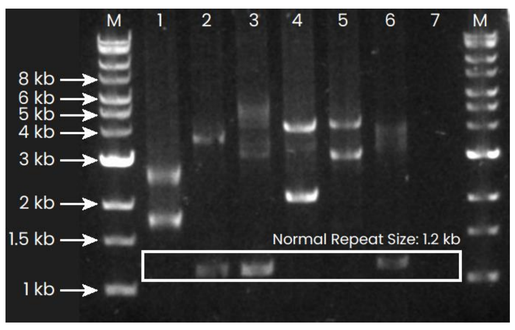

Figure 5: PCR confirming *FXN* GAA repeat expansion sizes. PCR amplification products spanning the *FXN* GAA repeat region were resolved by gel electrophoresis. Lane order (left to right): M: Marker, 1: GM16216, 2: GM16219, 3: GM16237, 4: GM16798, 5: GM16243, 6: GM15847, 7: Negative Control, M: Marker. The 1 Kb Extend DNA Ladder was used to estimate allele size. Band migration patters correspond to normal, carrier, and expanded alleles, where the normal alleles are 1,192 bp with the primers used (white box). Broader distributions in samples reflects variability in repeat size.

RepX *FXN* Analysis Concordance

Disease Status	Sample	Repeat Annotation (by gel)	Run #	Genome Coverage (X)	Allele A				Allele B			
					Observed Repeat Count	Repeat coverage	P(Full Mutation)	95% C.I	Observed Repeat Count	Repeat coverage	P(Full Mutation)	95% C.I
Control Samples (No expansion)	HG003	< 66	1	98	22	25	0.04	[0,63]	22	24	0.05	[0,63]
			2	84	30	19	0.11	[0,76]	30	20	0.11	[0,75]
	GM12891	< 66	1	102	7	17	0.03	[0,60]	7	17	0.03	[0,60]
			2	98	13	23	0.03	[0,57]	13	23	0.03	[0,57]
	GM12892	< 66	1	53	1	22	0.01	[0,49]	1	22	0.01	[0,49]
			2	66	18	19	0.05	[0,65]	18	20	0.05	[0,64]
	HG004	< 66	1	127	4	33	0.00	[0,42]	4	33	0.00	[0,42]
			2	81	1	16	0.03	[0,58]	1	17	0.02	[0,56]
Carriers (1 allele > 66 repeats)	GM15847	< 66, 609-942, & 942-1275	1	70	963	12	1.00	[893,1077]	25	19	0.08	[0,72]
			2	120	808	37	1.00	[759,858]	11	23	0.02	[0,55]
	GM16219	< 66 & 609-942	1	82	916	20	1.00	[848,987]	0	26	0.00	[0,43]
			2	91	872	33	1.00	[819,926]	0	27	0.00	[0,43]
	GM16237	< 66, 609-942, & 942-1609	1	98	1278	27	1.00	[1214,1345]	32	24	0.09	[0,72]
			2	100	1174	32	1.00	[1116,1234]	42	25	0.17	[3,84]
Affected (2 alleles > 66 repeats)	GM16216	109-275 & 275-609	1	95	498	20	1.00	[438,560]	242	26	1.00	[195,292]
			2	63	619	15	1.00	[547,695]	286	17	1.00	[226,348]
	GM16243	609-942 & 942-1275	1	87	1001	14	1.00	[917,1088]	529	13	1.00	[454,608]
			2	93	1122	25	1.00	[1058,1189]	661	28	1.00	[607,716]
	GM16798	275-609 & 942-1275	1	57	1041	14	1.00	[956,1129]	329	21	1.00	[274,386]
			2	79	1011	22	1.00	[944,1080]	307	25	1.00	[257,359]

Table 2: RepX-derived repeat counts and pathogenic probabilities for FA samples. Reported values include the observed repeat count (maximum *a posteriori* estimate), repeat-spanning molecule coverage, the probability that the repeat count exceeds the pathogenic threshold (≥ 66 repeats), and the 95% confidence interval.

Repeat Length Variability

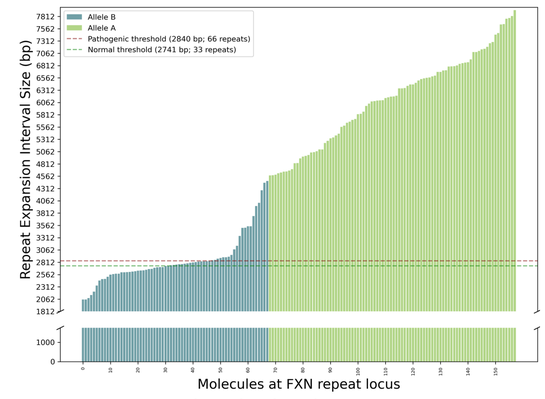

Figure 6: Molecule-level size distributions at the *FXN* locus. RepX waterfall plot resolves two distinct *FXN* allele populations in carrier sample GM16237, revealing repeat-length heterogeneity in the expanded allele. Each bar represents a single molecule spanning the *FXN* repeat locus, ranked by repeat count. Allele A (green) molecules span a wide range of repeat lengths well above the pathogenic threshold (red dashed line), consistent with somatic instability or mosaicism of the expanded allele. Allele B (blue) molecules cluster around a central value, consistent with somatic stability. In this example, we see both expanded and normal molecules.

CRISPR/Cas9 Optimization

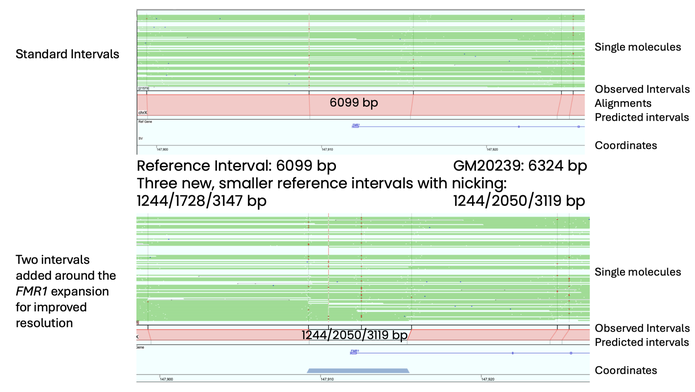

Figure 7: CRISPR/Cas9 nickase used to decrease expansion interval size. With the standard protocol, the interval containing the *FMR1* repeat expansion is 6,099 bp. This makes premutation expansions challenging to detect. Using CRISPR/Cas9, the expansion-containing interval can be reduced to 1,728 bp so that expansions of 150–200 bp can be more easily detected. The top alignment shows the normal interval patterns on chrX while the bottom alignment has two additional intervals for higher resolution. The same strategy can be used with other repeat expansions across the genome.

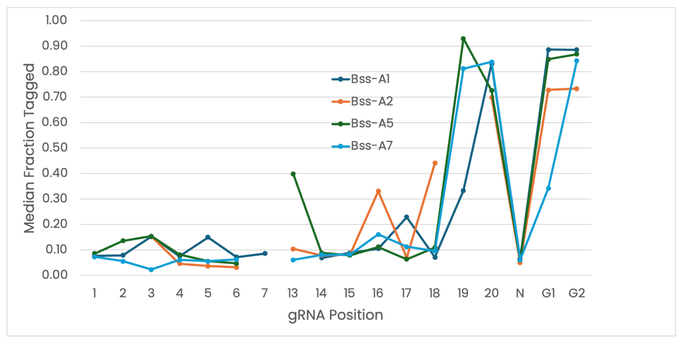

Figure 8: EGM can be used to assess effect of mismatches on CRISPR/Cas9 binding. Four repetitive gRNAs (Bss-A1, Bss-A2, Bss-A5, and Bss-A7) were examined at hundreds of sites around the human genome. Single mismatches had little effect on binding except in the PAM and nearby sequences, where gRNA Position represents the distance from the PAM.

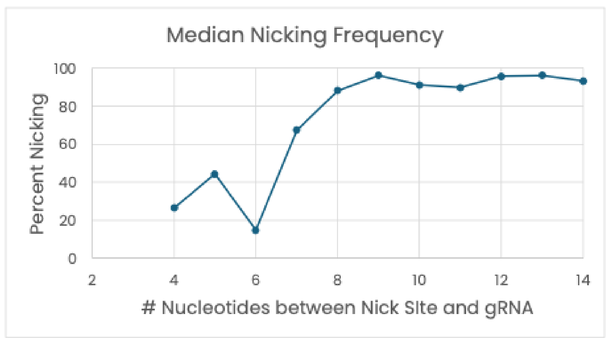

Figure 9: Interactions between CRISPR/Cas9 and other DNA modifying enzymes determined at base pair resolution. To determine median tagging efficiency, we examined percent nicking as a function of distance between the gRNA sequence and the recognition sequence for commercial nickases.

Conclusions

- **Accurate detection:** EGM RepX pipeline determined *FMR1* repeat lengths in healthy controls and Fragile X cell lines, including expansions up to ~600 CGG units, and *FXN* repeat lengths in healthy controls and FA cell lines, including expansions up to ~1200 GAA units.
- **Workflow advantages:** No amplification or ligation required, thus reducing false positives; scalable and higher throughput compared to Southern blots and PCR.
- **Current developments:** Assay optimization using CRISPR to create or remove tag sites for repeat expansions whose tagging patterns are not optimal for OhmX detection. EGM is also being used to optimize CRISPR/Cas9 and gRNA functions by providing a whole-genome approach to monitoring their behavior.

Learn more at Nabsys.com