Supporting Information For:

"A Polymerase Chain Reaction, Nuclease Digestion, and Mass Spectrometry Based Assay for the CGG Trinucleotide Repeat Status of the *Fragile X Mental Retardation 1* Gene"

Eric D. Dodds, Flora Tassone, Paul J. Hagerman, and Carlito B. Lebrilla

In order to allow more rapid calculation of CGG intensity ratios based on mass spectrometry (MS) data, a software utility was designed using IGOR Pro (version 6, WaveMetrics, Lake Oswego, OR, USA). The program, dubbed the CGG Oligonucleotide Recomposition Tool (CORT), carried out four major functions: import of MS peak lists; search of the peak lists for small oligonucleotide (ONT) masses; calculation of metrics related to CGG repeat status based on mass to charge ratio (*m/z*) and relative abundance data; and compilation of reports.

The graphical user interface of the CORT software package is shown in Figure S1. CORT was designed to import mass and intensity (M/I) tables saved in tab delimited text file

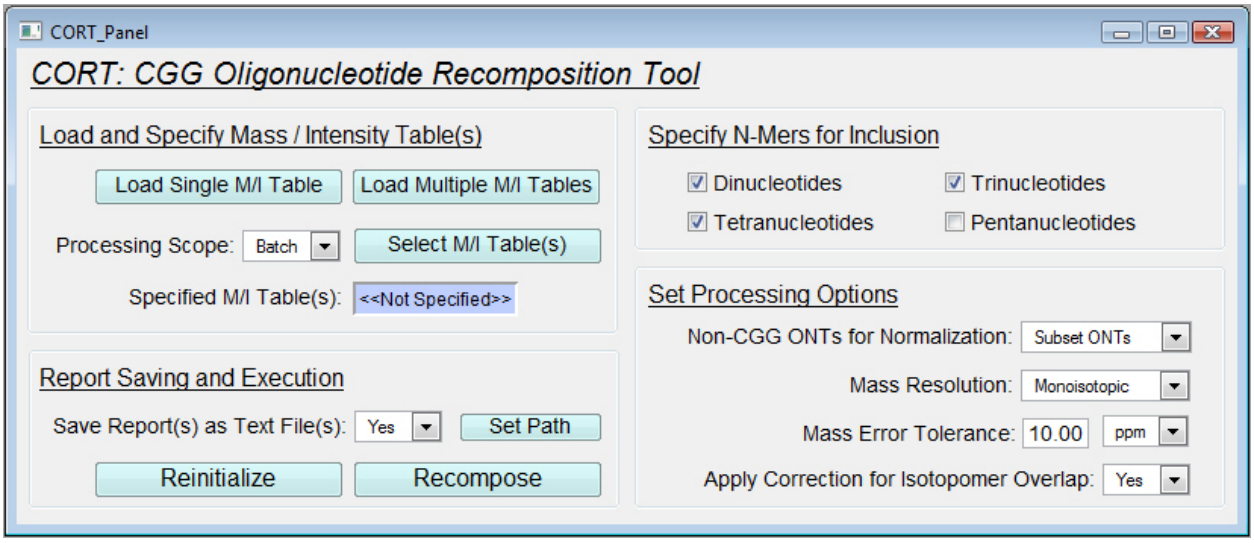


Figure S1. *User interface of the CGG Oligonucleotide Recomposition Tool (CORT) with all controls and options displayed.*

format. Once loaded, the M/I tables were searched for experimental *m/z* values matching known *m/z* values of singly deprotonated ONTs to within a user specified mass error tolerance. The mass error tolerance could be set in terms of Daltons (Da) or parts per million (ppm). Another operator defined parameter allowed the peak lists to be searched for either monoisotopic masses (appropriate for high resolution MS data) or average masses (appropriate for low resolution MS data). Because interest was limited to relatively short ONTs composed of only the four DNA nucleotides (NTs), it was possible to exhaustively search the peak lists for all possible small ONT compositions. The number of possible ONT compositions was determined as the number of combinations of *n* items (*i.e.*, the four DNA NTs) sampled *x* times (*i.e.*, the number of NTs comprising the ONT) with replacement.

$$\binom{n}{x} = \frac{(n+x-1)!}{x!(n-1)!} \qquad \textit{[Equation S1]}$$

According to this relationship, there were 10 possible dinucleotide compositions (Table S1), 20 possible trinucleotide compositions (Table S2), 35 possible tetranucleotide compositions (Table S3), and 56 possible pentanucleotide compositions (Table S4). The observation of all possible ONT compositions was not anticipated, since only a specific collection of ONT digestion products could result from the PCR amplicons of interest. Nonetheless, since the total number of small ONTs was finite (only 121 possible compositions for 2-mers through 5-mers), an all-inclusive search was implemented in order to provide maximum extensibility

Table S1. *Compositions of all possible DNA dinucleotides and corresponding mass to charge ratios.*

Dinucleotide Composition	Monoisotopic $[M-H]^-$ (*m/z*)	Average $[M-H]^-$ (*m/z*)
C_2	595.09603	595.377
C_1T_1	610.09569	610.389
C_1A_1	619.10726	619.402
T_2	625.09536	625.401
A_1T_1	634.10693	634.414
C_1G_1	635.10218	635.402
A_2	643.11849	643.427
T_1G_1	650.10184	650.414
A_1G_1	659.11341	659.427
G_2	675.10832	675.427

Table S2. *Compositions of all possible DNA trinucleotides and corresponding mass to charge ratios.*

Trinucleotide Composition	Monoisotopic $[M-H]^-$ (m/z)	Average $[M-H]^-$ (m/z)
C_3	884.14240	884.562
C_2T_1	899.14207	899.574
C_2A_1	908.15363	908.587
C_1T_2	914.14173	914.586
$C_1A_1T_1$	923.15330	923.599
C_2G_1	924.14855	924.587
T_3	929.14140	929.598
C_1A_2	932.16487	932.612
A_1T_2	938.15296	938.611
$C_1T_1G_1$	939.14821	939.599
A_2T_1	947.16453	947.624
$C_1A_1G_1$	948.15978	948.612
T_2G_1	954.14788	954.611
A_3	956.17610	956.637
$A_1T_1G_1$	963.15945	963.624
C_1G_2	964.15470	964.612
A_2G_1	972.17101	972.637
T_1G_2	979.15436	979.624
A_1G_2	988.16593	988.637
G_3	1004.16084	1004.637

and flexibility. Once tabulated, the mass and intensity values were used to calculate the total intensity of ONTs derived from the CGG repeat region (Table 1 of the main article) relative to the total intensity of a normalizing set of non-repeat ONTs. The CGG repeat intensity ratio, R_f, was then calculated according to:

$$R_f = \frac{\sum r_i}{\sum f_j} \qquad \textit{[Equation S2]}$$

where r_i represents the relative abundance of the i-th ONT mass arising from the CGG repeat region, and f_j represents the relative abundance of the j-th ONT arising from the flanking region. The subscript f in R_f denotes that the CGG signals were normalized to all flanking ONT signals (*i.e*, all observed products from benzonase digestion of the flanking region). Alternatively, the ratio R_s was calculated according to:

$$R_s = \frac{\sum r_i}{\sum s_k} \qquad \textit{[Equation S3]}$$

where s_k represents the relative abundance of the k-th ONT from a selected subset of non CGG ONTs. This selected subset of normalizing signals is listed in Table 2 of the main article. These were chosen due to their high intensity (compared to other flank derived ONT

Table S3. *Compositions of all possible DNA tetranucleotides and corresponding mass to charge ratios.*

Tetranucleotide Composition	Monoisotopic $[M-H]^-$ (m/z)	Average $[M-H]^-$ (m/z)
C_4	1173.18877	1173.747
C_3T_1	1188.18844	1188.759
C_3A_1	1197.20000	1197.772
C_2T_2	1203.18810	1203.771
$C_2A_1T_1$	1212.19967	1212.784
C_3G_1	1213.19492	1213.772
C_1T_3	1218.18777	1218.783
C_2A_2	1221.21124	1221.797
$C_1A_1T_2$	1227.19934	1227.796
$C_2T_1G_1$	1228.19458	1228.784
T_4	1233.18743	1233.795
$C_1A_2T_1$	1236.21090	1236.809
$C_2A_1G_1$	1237.20615	1237.797
A_1T_3	1242.19900	1242.808
$C_1T_2G_1$	1243.20325	1243.796
C_1A_3	1245.22247	1245.822
A_2T_2	1251.21057	1251.821
$C_1A_1T_1G_1$	1252.20582	1252.809
C_2G_2	1253.20107	1253.797
T_3G_1	1258.19392	1258.808
A_3T_1	1260.22214	1260.834
$C_1A_2G_1$	1261.21739	1261.822
$A_1T_2G_1$	1267.20548	1267.821
$C_1T_1G_2$	1268.20073	1268.809
A_4	1269.23370	1269.847
$A_2T_1G_1$	1276.21705	1276.834
$C_1A_1G_2$	1277.21230	1277.822
T_2G_2	1283.20040	1283.821
A_3G_1	1285.22862	1285.847
$A_1T_1G_2$	1292.21197	1292.834
C_1G_3	1293.20722	1293.822
A_2G_2	1301.22353	1301.847
T_1G_3	1308.20688	1308.834
A_1G_3	1317.21845	1317.847
G_4	1333.21336	1333.847

compositions) and appropriate distribution across the mass range of interest. In addition, the two ratios above were calculated using weighting factors $w_{i,j,k}$ such that the intensity of each ONT was multiplied by the number of NTs comprising the ONT. Thus, the intensity of a tetranucleotide was given twice the weight of a dinucleotide. The repeat intensity ratios with weighting, $R_{f(w)}$ and $R_{s(w)}$, were thus defined as shown below.

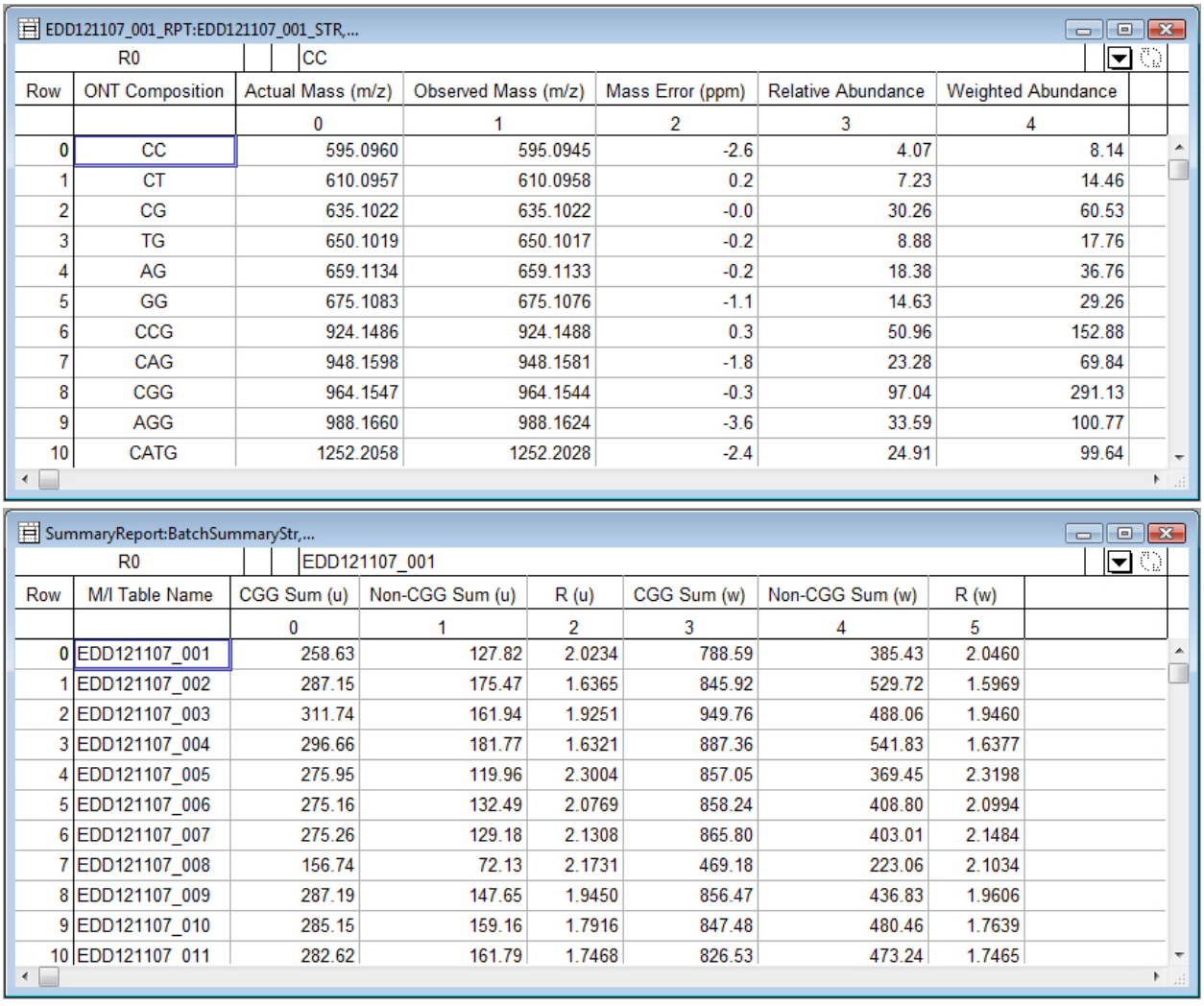
$$R_{f(w)} = \frac{\sum w_i r_i}{\sum w_j f_j} \qquad \textit{[Equation S4]}$$

$$R_{s(w)} = \frac{\sum w_i r_i}{\sum w_k s_k} \qquad \textit{[Equation S5]}$$

Table S4. *Compositions of all possible DNA pentanucleotides and corresponding mass to charge ratios.*

Pentanucleotide Composition	Monoisotopic $[M-H]^-$ (m/z)	Average $[M-H]^-$ (m/z)
C_5	1462.23514	1462.932
C_4T_1	1477.23481	1477.944
C_4A_1	1486.24638	1486.957
C_3T_2	1492.23447	1492.956
$C_3A_1T_1$	1501.24604	1501.969
C_4G_1	1502.24129	1502.957
C_2T_3	1507.23414	1507.968
C_3A_2	1510.25761	1510.982
$C_2A_1T_2$	1516.24571	1516.981
$C_3T_1G_1$	1517.24096	1517.969
C_1T_4	1522.23381	1522.980
$C_2A_2T_1$	1525.25728	1525.994
$C_3A_1G_1$	1526.25252	1526.982
$C_1A_1T_3$	1531.24537	1531.993
$C_2T_2G_1$	1532.24062	1532.981
C_2A_3	1534.26884	1535.007
T_5	1537.23347	1537.992
$C_1A_2T_2$	1540.25694	1541.006
$C_2A_1T_1G_1$	1541.25219	1541.994
C_3G_2	1542.24744	1542.982
A_1T_4	1546.24504	1547.005
$C_1T_3G_1$	1547.24029	1547.993
$C_1A_3T_1$	1549.26851	1550.019
$C_2A_2G_1$	1550.26376	1551.007
A_2T_3	1555.25661	1556.018
$C_1A_1T_2G_1$	1556.25186	1557.006
$C_2T_1G_2$	1557.24710	1557.994
C_1A_4	1558.28008	1559.032
T_4G_1	1562.23995	1563.005
A_3T_2	1564.26817	1565.031
$C_1A_2T_1G_1$	1565.26342	1566.019
$C_2A_1G_2$	1566.25867	1567.007
$A_1T_3G_1$	1571.25152	1572.018
$C_1T_2G_2$	1572.24677	1573.006
A_4T_1	1573.27974	1574.044
$C_1A_3G_1$	1574.27499	1575.032
$A_2T_2G_1$	1580.26309	1581.031
$C_1A_1T_1G_2$	1581.25834	1582.019
C_2G_3	1582.25359	1583.007
A_5	1582.29131	1583.057
T_3G_2	1587.24644	1588.018
$A_3T_1G_1$	1589.27466	1590.044
$C_1A_2G_2$	1590.26991	1591.032
$A_1T_2G_2$	1596.25800	1597.031
$C_1T_1G_3$	1597.25325	1598.019
A_4G_1	1598.28622	1599.057
$A_2T_1G_2$	1605.26957	1606.044
$C_1A_1G_3$	1606.26482	1607.032
T_2G_3	1612.25292	1613.031
A_3G_2	1614.28114	1615.057
$A_1T_1G_3$	1621.26449	1622.044
C_1G_4	1622.25973	1623.032
A_2G_3	1630.27605	1631.057
T_1G_4	1637.25940	1638.044
A_1G_4	1646.27097	1647.057
G_5	1662.26588	1663.057

Upon completion of the calculations, two types of reports were generated (Figure S2): a detailed report for each sample which provided compositions, masses, and intensities of all matched ONT signals; and a simple batch report which listed the sums of CGG and non CGG ONT intensities and the corresponding *R* values for each sample. While the report tables could be maintained within the IGOR experiment file, the reports could also be exported as tab delimited text files at the option of the user. The reports provided by CORT were generated such that both the weighted and unweighted values of *R* were given side by side,



EDD121107_001_RPT:EDD121107_001_STR,...

R0 | CC

Row	ONT Composition	Actual Mass (m/z)	Observed Mass (m/z)	Mass Error (ppm)	Relative Abundance	Weighted Abundance
		0	1	2	3	4
0	CC	595.0960	595.0945	-2.6	4.07	8.14
1	CT	610.0957	610.0958	0.2	7.23	14.46
2	CG	635.1022	635.1022	-0.0	30.26	60.53
3	TG	650.1019	650.1017	-0.2	8.88	17.76
4	AG	659.1134	659.1133	-0.2	18.38	36.76
5	GG	675.1083	675.1076	-1.1	14.63	29.26
6	CCG	924.1486	924.1488	0.3	50.96	152.88
7	CAG	948.1598	948.1581	-1.8	23.28	69.84
8	CGG	964.1547	964.1544	-0.3	97.04	291.13
9	AGG	988.1660	988.1624	-3.6	33.59	100.77
10	CATG	1252.2058	1252.2028	-2.4	24.91	99.64

SummaryReport:BatchSummaryStr,...

R0 | EDD121107_001

Row	M/I Table Name	CGG Sum (u)	Non-CGG Sum (u)	R (u)	CGG Sum (w)	Non-CGG Sum (w)	R (w)
		0	1	2	3	4	5
0	EDD121107_001	258.63	127.82	2.0234	788.59	385.43	2.0460
1	EDD121107_002	287.15	175.47	1.6365	845.92	529.72	1.5969
2	EDD121107_003	311.74	161.94	1.9251	949.76	488.06	1.9460
3	EDD121107_004	296.66	181.77	1.6321	887.36	541.83	1.6377
4	EDD121107_005	275.95	119.96	2.3004	857.05	369.45	2.3198
5	EDD121107_006	275.16	132.49	2.0769	858.24	408.80	2.0994
6	EDD121107_007	275.26	129.18	2.1308	865.80	403.01	2.1484
7	EDD121107_008	156.74	72.13	2.1731	469.18	223.06	2.1034
8	EDD121107_009	287.19	147.65	1.9450	856.47	436.83	1.9606
9	EDD121107_010	285.15	159.16	1.7916	847.48	480.46	1.7639
10	EDD121107_011	282.62	161.79	1.7468	826.53	473.24	1.7465

Figure S2. *Excerpts from an individual sample report (upper panel) and a batch summary report (lower panel) generated by CORT.*

While the option to calculate R_f or R_s was provided as a user-defined parameter. For all results shown in the main article, $R_{s(w)}$ was calculated according to Equation S5.

In the range of dinucleotides through pentanucleotides, each ONT composition corresponded to a unique mass. However, an examination of Tables S1 through S4 revealed a number of cases in which the monoisotopic mass *A* of a particular ONT coincided with an *A*+1 or *A*+2 isotopomer of different ONT composition. The simplest example of this is encountered with the dinucleotide compositions A_1T_1 (monoisotopic *m/z* 634.1069; [*A*+1] isotopomer *m/z* 635.1103) and C_1G_1 (monoisotopic *m/z* 635.1022). In this case, the intensity of the C_1G_1 monoisotopic *A* peak would be inflated by the overlap from the A_1T_1 *A*+1 peak. Another example of isotopomer overlap is illustrated in Figure S3, wherein the tetranucleotides $C_1A_1T_1G_1$ and C_2G_2 overlap in a similar manner. Indeed, this is the case for any two ONT compositions related by a CG for AT substitution. Correction of the C_2G_2 *A* peak intensity was accomplished according to:

$$I_{C_2G_2(c)} = I_{C_2G_2(u)} - I_{C_1A_1T_1G_1}F \qquad \textit{[Equation S6]}$$

where $I_{C_2G_2(c)}$ is the corrected monoisotopic C_2G_2 intensity, $I_{C_2G_2(u)}$ is the uncorrected monoisotopic C_2G_2 intensity, $I_{C_1A_1T_1G_1}$ is the monoisotopic intensity of $C_1A_1T_1G_1$, and *F* is a

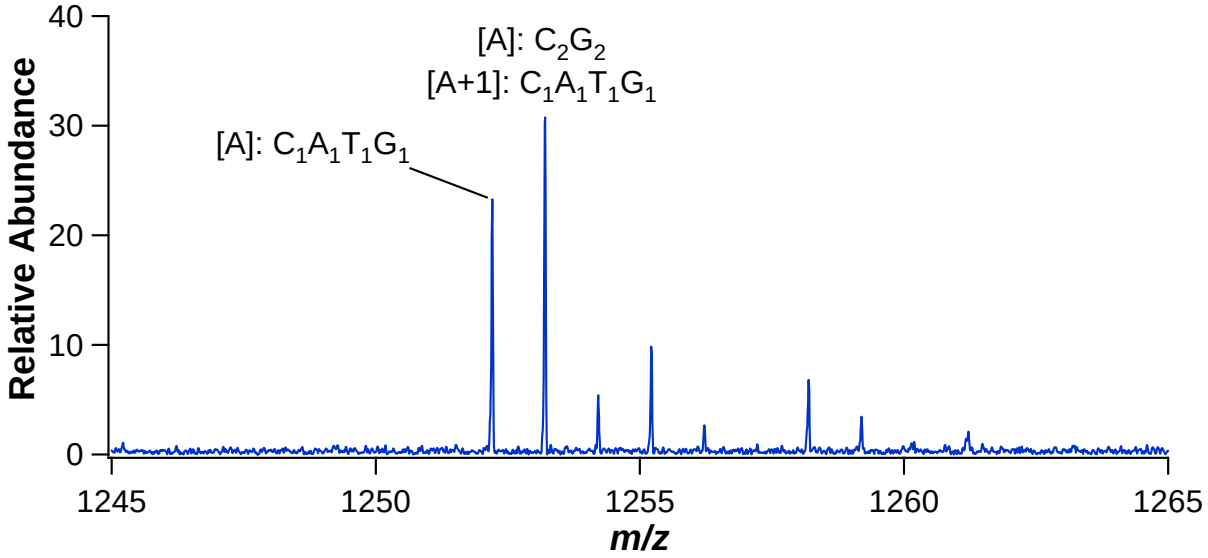


Figure S3. *Example of isotopomer overlap involving the tetranucleotide compositions $C_1A_1T_1G_1$ and C_2G_2.*

correction factor derived from the expected relative intensities of the *A* and *A*+1 isotopomers of $C_1A_1T_1G_1$. In the case of the $C_1A_1T_1G_1$ ONT, the intensity of the *A*+1 peak was predicted to be 48.8% of the *A* peak intensity based on elemental composition and the natural abundance of elemental isotopes.

The option to subtractively correct for all such isotopomer overlaps according to Equation S6 was incorporated into CORT and was provided as a user defined setting. All *F* correction factors were obtained by simulating the isotopomer distributions for each overlap contributing ONT at a mass resolution of 30,000 using the Varian IonSpec Exact Mass Calculator (version 8, Lake Forest, CA). In practice, the isotopomer overlap correction was not found to significantly affect the values of *R* because most of the AT type signals were of much lower abundance than the CG type signals. However, in other potential applications, isotopomer overlap correction could be crucial to obtaining meaningful relative intensity data.

Other operator selectable options in CORT include the ability to import and / or process single peak lists or peak list batches, and the option to independently include or exclude any particular ONT length from the calculations.