Title page

Determining correct proportions in single base polymorphisms from direct sequencing chromatograms.

Mats Ensterö, Helene Wahlstedt and Marie Öhman[#].

Department of Molecular Biology and Functional Genomics, Stockholm University, S-106 91 Stockholm, Sweden.

[#]Corresponding author, e-mail: marie.ohman@molbio.su.se

Keywords: chromatogram, SNP, 454, proportions

Word count: 972

Abstract

We present a novel conversion relationship between actual proportions in single site polymorphisms and the proportions determined from chromatograms. Traditional nucleotide sequencing by Sanger result in a fluorescent response represented in chromatograms. Ambiguous peak response at a base-call position is either background noise or single base polymorphisms. When calculating fractions of the intrinsic residues at the heterogeneous base-call positions, a linear relationship between peak heights and the amount of the residues has earlier been presented. By large scale 454 sequencing we produce a vast amount of sequential data showing pattern from known adenosine-to-Inosine RNA editing. Due to the large range of frequencies in which A-to-I editing occurs we can with unprecedented resolution determine actual heterogeneous proportions for site specific editing. Hence, we can thoroughly compare the editing proportions determined by chromatogram peak heights with the very exact proportion determined from the large scale sequencing. Our conclusion is that chromatogram proportions and actual proportions of adenosines and inosines are not at all linear but rather exponential. We present a novel equation to determine real proportions of nucleotides at a base-call position from chromatogram peak heights.

One type of editing is a deamination process of adenosines to inosines catalyzed by a family of proteins called ADARs (Adenosine Deaminases that Act on RNA). ADARs recognize double stranded RNA (dsRNA) structures as targets. Although the characteristics of a dsRNA that is targeted by ADARs are still largely unknown it is clear that some RNA elements have favorable traits (sequential and/or structural) that infer a high degree of editing while other targets are edited to a much lesser degree. Hence, the wide range in which RNA transcripts are subjected to adenosine conversions make editing a suitable mechanism to calculate A and G proportions from very small to very large amounts of Gs. We determine proportions (frequencies) from conventional Sanger sequencing as well as 454 amplicon sequencing. The large scale output (reads) from 454 sequencing give us the means to assign an editing frequency with high accuracy. Comparing frequencies determined from chromatogram peaks and 454 sequencing we can establish an equation that directly convert a calculated chromatogram frequency to an actual frequency within a 95% confidence interval. Previous attempts to perform the same kind of analysis have used DNA concentration measurements as the actual proportion variable versus the chromatogram proportion, {Nakae et al 2008, Nurpeisov et al 2003}. This has yielded a linear relation with a high correlation coefficient but with a small number of data points. There is also an inherent uncertainty in DNA concentration measurements. Our analyses contain 72 data points of the dependent variable (actual proportion) where each proportion is determined from an average of 792 reads. We find that the best fit is a non-linear relationship with a correlation coefficient $R^2 = 0.94$.

From previous work our lab has used 454 sequencing data to determine coupled properties between nearby edited sites and a large scale compilation of editing efficiency regulation through development {Wahlstedt et al 2009}. We use this data to propose a refined model to convert chromatogram A-to-I(G) dual peaks to actual proportions determined by an unprecedented accuracy from this data. Previous work has also established that the ratio of heterogeneous chromatogram peaks are independent of nucleotide composition. I.e., even though the fluorescent response can differ between nucleotides, the relative ratio is always consistent with DNA concentration {Nurpeisov et al., 2003}. We therefore propose that our conversion relationship extends not only to A and G heterogeneity but all possible base compositions.

The 454 amplicon sequencing procedure and the calculations of A and G proportions at the sites of editing is as presented in {Wahlstedt et al 2009}. Briefly, the 454 sequencing give us N number of transcript reads (typically in the range of 500-2000). Each read contain either an A or G at the known sites of editing. In a chromatogram this yields a dual peak response at the base-call position. The actual proportion is in our case calculated directly by counting the number of reads having an A or G respectively, giving us a very exact proportion to compare with the chromatogram response.

Two selected targets of site selective editing were chosen for conventional sequencing by Sanger – Gabra3 and 5-HT$_{2C}$ (gamma-aminobutyric acid A receptor, subunit alpha 3 and the serotonin receptor 2C). The data from the 454 sequencing cover four different developmental stages, embryonic days 15 and 19 as well as post natal days 2 and 21. The conventional sequencing was made for the same four developmental stages and for three different mouse individuals. Importantly, one of the individuals was the same that was used in the 454 sequencing. This mean that we can establish that the chromatogram response does not differ significantly between that individual and the remaining two. Hence, ruling out the possibility that the chromatogram response for the first individual is not atypical in any way. We also uncontroversially assume independency between developmental regulation of editing and the suggested relationship between chromatogram proportions and actual proportion. The targets were chosen with respect to optimize the resolution (i.e number of reads) and to cover as much as possible of the proportion spectra (0 – 100%).

The serotonin receptor contain 5 selectively edited sites (A, B, E, C and D sites) {Burns et al 1997} and Gabra3 contain one site (I/M) {Ohlson et al 2007}. We use Chromas lite to determine A and G peaks heights in pixels (H) (Figure 1). The proportion P(%) is:

$$P_G = \frac{H_G}{H_A + H_G}$$

We use the MATLAB® Statistics toolbox to compile our data. First, we propose different model functions for our data and see that an exponential relationship give the best correlation fit ($R^2 = 0.94$). The β vector contain the different constants in the function that read $y_{fitted} = b(0) + b(1)x + b(3)x^2$ where $\beta = [b(0)\ b(1)\ b(3)]$. Since each data point of the actual proportion is determined from different number of reads we also put a weight to these (i.e the more reads the more weight that data point has on the determination of β. We find that the equation reads:

$$y = 4.42 + 0.43x + 0.0058x^2$$ (Figure 2).

Raw data and conversion table can be retrieved from the online supplementary material (url). In Figure 2 the confidence boundaries is also plotted. In addition to the correlation coefficient we also diagnose the fit by plotting the residuals against the independent variable (residuals = $y - y_{\text{fitted}}$). The residuals should be independently and evenly distributed around zero (Figure 3).

Overall, we have been able to use a costly large scale sequencing method (454 amplicon sequencing) to determine exact proportions of As and Gs at sites of selective editing. The template data has then been compared to proportions calculated from chromatogram peak heights (conventional method). We see an exponential relationship with a good fit to the data. Our function can bee used to accurately determine single site polymorphism proportions directly from chromatograms without the need for the more expensive sequencing methods.

References

1. Nakae A, Tanaka T, Miyake K, Hase M, Mashimo T. 2008. Comparing methods of detection and quantitation of RNA editing of rat glycine receptor alpha3^{P185L}. *Int J Biol Sci* 4:397-405.

2. Nurpeisov, Viktoria, Hurwitz, Selwyn J., Sharma, Prem L. 2003. Fluorescent Dye Terminator Sequencing Methods for Quantitative Determination of Replication Fitness of Human Immunodeficiency Virus Type 1 Containing the Codon 74 and 184 Mutations in Reverse Transcriptase. *J. Clin. Microbiol.* 41: 3306-3311.

3. Helene Wahlstedt, Chammiran Daniel, Mats Ensterö and Marie Öhman. 2009. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res*. 19: 978-986.

4. Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 15;387(6630):303-8.

5. Johan Ohlson, Jakob Skou Pedersen, David Haussler and Marie Öhman. 2007. Editing modifies the GABAA receptor subunit α3. *RNA* 13: 698-703.

6. MATLAB 7, 2005. Mathworks, Sweden.

7. Chromas lite 2001. Technelysium Pty Ltd.

Figure 1.

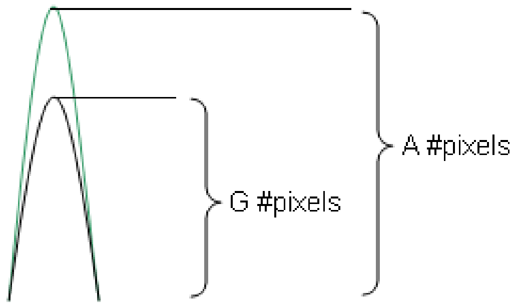

Figure 2.

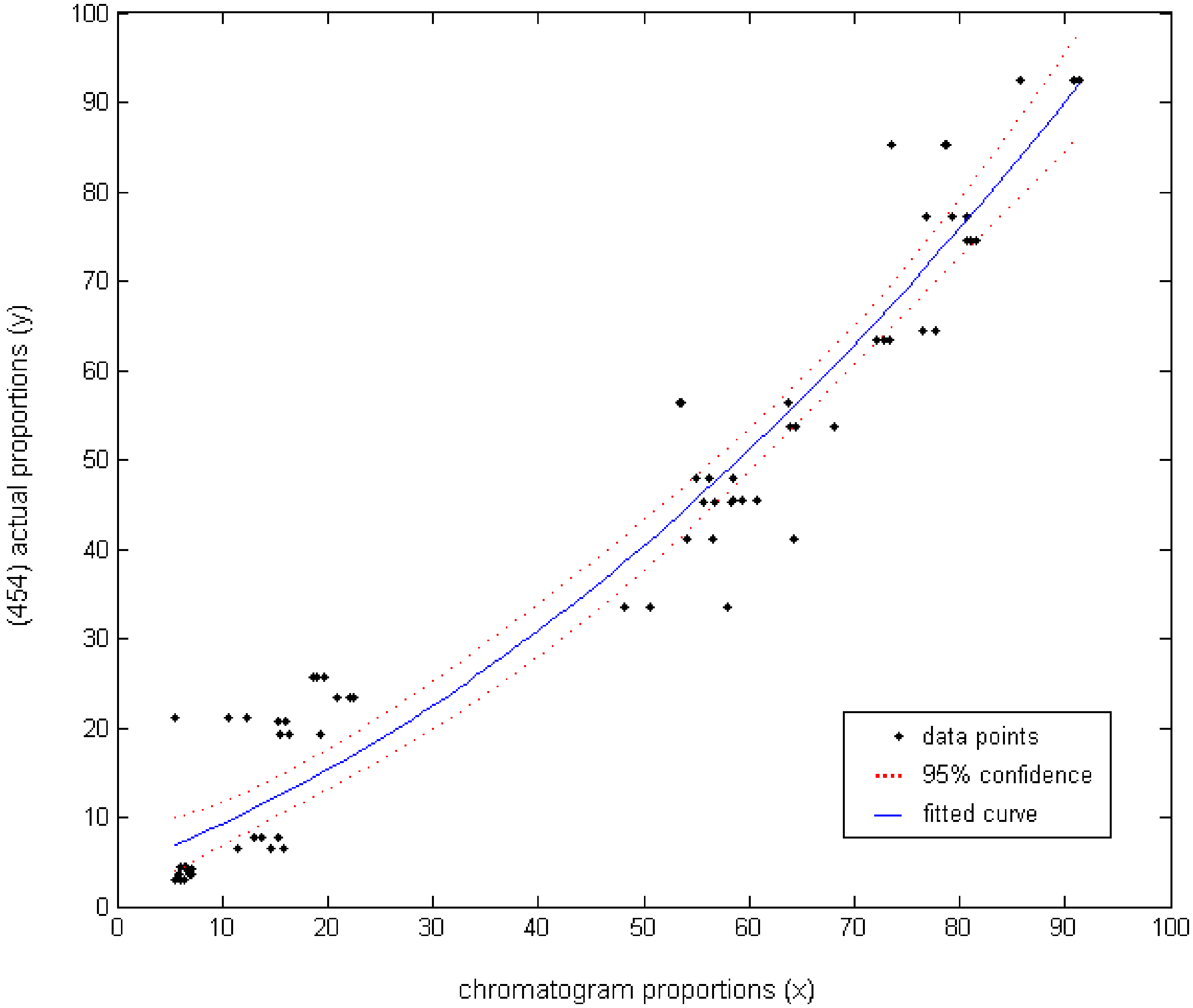
100
90
80
70
60
50
40
30
20
10
0
(454) actual proportions (y)
0
10
20
30
40
50
60
70
80
90
100
chromatogram proportions (x)
data points
95% confidence
fitted curve

Figure 3.

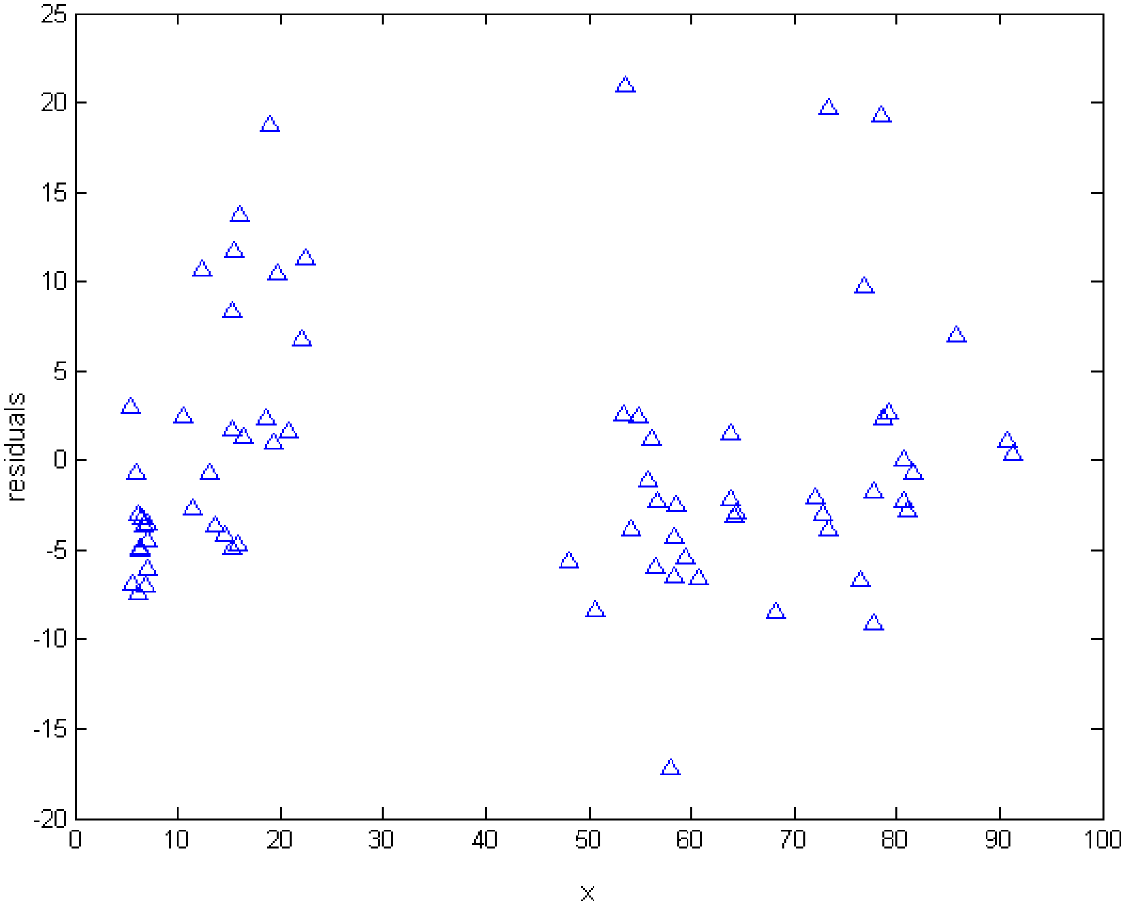

Figure legends.

Figure 1.
Show a typical chromatogram dual (A and G) peak. We determine chromatogram proportions by measuring the peak heights in pixels for A and G.

Figure 2.
Shows the scatter plot for the actual proportions versus chromatogram proportions (black dots). The blue line depicts the fitted curve for our model relationship and the red dotted line is the boundaries for the 95% confidence interval.

Figure 3.
The residuals for each x data point is shown. The residuals is the difference between the actual proportion (y) and the fitted value (y_{fitted}) fore each x.

Supplementary table 1.

x	y	x_{grid}	y_{fitted}	residuals	Δ (95% confidence)
19.2308	19.3000	5.4795	6.9468	2.0503	2.9805
53.4483	56.3000	6.6883	7.5512	5.6549	2.8065
80.7692	77.3000	7.8971	8.1726	0.1488	2.6524
78.6885	85.2000	9.1059	8.8109	5.0588	2.5191
16.4179	19.3000	10.3147	9.4661	2.8615	2.4074
53.5714	56.3000	11.5235	10.1384	16.0107	2.3177
76.7857	77.3000	12.7323	10.8276	7.4431	2.2499
78.5714	85.2000	13.9411	11.5338	14.6752	2.2033
15.4930	19.3000	15.1499	12.2569	8.9539	2.1769
63.7931	56.3000	16.3587	12.9970	1.1234	2.1688
79.2453	77.3000	17.5675	13.7541	2.5071	2.1770
73.4375	85.2000	18.7763	14.5282	18.7763	2.1989
15.2174	7.8000	19.9851	15.3192	-4.6864	2.2322
54.1667	41.2000	21.1939	16.1272	-3.6584	2.2744
77.6699	64.4000	22.4027	16.9522	-8.7452	2.3233
80.7018	74.6000	23.6115	17.7941	-2.3013	2.3766
13.6364	7.8000	24.8203	18.6530	-3.5827	2.4328
56.5217	41.2000	26.0291	19.5289	-6.0234	2.4900
76.4151	64.4000	27.2379	20.4217	-6.7180	2.5471
81.0811	74.6000	28.4467	21.3315	-2.7963	2.6029
13.0435	7.8000	29.6555	22.2583	-1.4693	2.6564
64.2857	41.2000	30.8643	23.2020	-6.7858	2.7069
77.6699	64.4000	32.0731	24.1627	-3.8444	2.7536
81.5789	74.6000	33.2819	25.1404	-1.5967	2.7960
5.9000	3.6000	34.4907	26.1350	-1.6235	2.8338
6.1000	2.9000	35.6995	27.1467	-5.7074	2.8664
6.4000	4.5000	36.9083	28.1752	-3.8078	2.8937
7.0000	4.2000	38.1171	29.2208	-4.5981	2.9153
6.9000	3.6000	39.3259	30.2833	-5.3173	2.9312
5.5000	2.9000	40.5347	31.3628	-5.2835	2.9412
6.5000	4.5000	41.7435	32.4593	-3.0942	2.9452
6.6000	4.2000	42.9523	33.5727	-3.4611	2.9432
7.0000	3.6000	44.1611	34.7031	-4.2821	2.9352
6.4000	2.9000	45.3699	35.8504	-4.6945	2.9213
6.1000	4.5000	46.5787	37.0148	-2.8705	2.9015
7.0000	4.2000	47.7875	38.1961	-3.5380	2.8761
12.3596	21.1000	48.9963	39.3943	10.5709	2.8452
15.3846	20.8000	50.2051	40.6096	8.3834	2.8091
22.0588	23.5000	51.4139	41.8418	6.7703	2.7681
19.7368	25.6000	52.6227	43.0910	10.4232	2.7227
10.5263	21.1000	53.8315	44.3571	5.2754	2.6733
15.3846	20.8000	55.0403	45.6402	3.8478	2.6205
20.8333	23.5000	56.2491	46.9403	3.4882	2.5652
18.5714	25.6000	57.4579	48.2573	5.1319	2.5082
5.4795	21.1000	58.6667	49.5914	6.4631	2.4505
16.0494	20.8000	59.8755	50.9423	10.4765	2.3935
22.3881	23.5000	61.0843	52.3103	8.5944	2.3385
18.9873	25.6000	62.2931	53.6952	14.3255	2.2875
58.4746	47.9000	63.5019	55.0971	-1.9366	2.2422
58.2524	45.3000	64.7107	56.5160	-4.9908	2.2050
58.4158	45.4000	65.9195	57.9518	-4.0961	2.1783
73.2759	63.5000	67.1283	59.4046	-3.7220	2.1644
56.1404	47.9000	68.3371	60.8744	1.1225	2.1660
56.7010	45.3000	69.5459	62.3611	-2.2200	2.1852
59.4340	45.4000	70.7547	63.8648	-5.2586	2.2241
72.8070	63.5000	71.9635	65.3855	-2.9804	2.2838
54.9020	47.9000	73.1723	66.9231	2.4268	2.3654
55.7692	45.3000	74.3811	68.4777	-1.1198	2.4691
60.7843	45.4000	75.5899	70.0493	-6.5557	2.5946
72.1311	63.5000	76.7987	71.6378	-2.0933	2.7412
14.6067	6.4000	78.0075	73.2433	-4.7832	2.9082
57.8947	33.6000	79.2163	74.8658	-16.1532	3.0943
63.8554	53.7000	80.4251	76.5053	-2.0054	3.2986
91.3043	92.5000	81.6339	78.1617	0.4275	3.5199
15.7895	6.4000	82.8427	79.8351	-5.4030	3.7573
48.1481	33.6000	84.0515	81.5254	-5.2840	4.0099
68.1319	53.7000	85.2603	83.2328	-7.6699	4.2770
85.7143	92.5000	86.4691	84.9571	7.7418	4.5576
11.3924	6.4000	87.6779	86.6983	-3.1698	4.8514
50.5882	33.6000	88.8867	88.4565	-7.8947	5.1577
64.3678	53.7000	90.0955	90.2317	-2.6717	5.4761
90.7563	92.5000	91.3043	92.0239	1.1589	5.8062

Table S1.
Showing the datapoints for the exponential relationship between chromatogram proportions (x) and actual proportions determined by large scale 454 sequencing (y). y_{fitted} contain the calculated data points for the x_{grid} variable that contain 72 evenly spaced pints between x_{min} and x_{max}. Δ is the ± boundaries to adjust y_{fitted} to reach within a 95% confidence interval.