

1 **Re-analysis of genetic risks for Chronic Fatigue Syndrome from** 2 **23andMe data finds few remain**

3 **Felice L. Bedford^{1*}, Bastian Greshake Tzovaras²**

4 ¹University of Arizona, Tucson, AZ, USA

5 ²Université de Paris, Paris, France

6 *** Correspondence:**

7 Corresponding Author

8 bedford@u.arizona.edu

9

10 **Keywords: Chronic Fatigue Syndrome¹, Myalgic Encephalomyelitis², single nuclear**
11 **polymorphism³, 23 and Me⁴, genetic predispositions⁵**

12 **July 31, 2020 Word count: 1999 Tables: 1 Supplementary Figures: 1**

13 **Abstract**

14 It is tempting to mine the abundance of DNA data that is now available from direct-to-consumer genetic
15 tests but this approach also has its pitfalls A recent study put forth a list of 50 single nucleotide
16 polymorphisms (SNPs) that predispose to Chronic Fatigue Syndrome (CFS), a potentially major advance in
17 understanding this still mysterious condition. However, only the patient cohort data came from a
18 commercial company (23andMe) while the control was from a genetic database. The extent to which
19 23andMe data agree with genetic reference databases is unknown. We reanalyzed the 50 purported CFS
20 SNPs by comparing to control data specifically from 23andMe which are available through public platform
21 OpenSNP. In addition, large high-quality database ALFA was used as an additional control. The analysis lead
22 to dramatic change with the top of the leaderboard for CFS risk reduced and reversed from an astronomical
23 129,000 times to 0.8. Errors were found both within 23andMe data and the original study-reported Kaviar
24 database control. Only 3 of 50 SNPs survived initial study criterion of at least twice as prevalent in patients,
25 EFCAB4B, involved in calcium ion channel activation, LINC01171, and MORN2 genes. We conclude that the
26 reported top-50 deleterious polymorphisms for Chronic Fatigue Syndrome were more likely the top-50 errors
27 in the 23andMe and Kaviar databases. In general, however, correlation of 23andMe control with ALFA was a
28 respectable 0.93, suggesting an overall usefulness of 23andMe results for research purposes but only if
29 caution is taken with chips and SNPs.

30

31 **1 Introduction**

32 As part of a growing number of researchers that advocate using the plentiful genetic data from direct
33 to consumer testing (1), we are also aware of its pitfalls. Recently, a study in Frontiers in Pediatrics
34 (2) described an ambitious project to elucidate genetic predispositions for Chronic Fatigue Syndrome
35 (CFS), a not-uncommon condition of debilitating fatigue, immune dysregulation, and central nervous
36 system impairment. The study analyzed the approximate 500,000 genetic single nuclear

37 polymorphisms (snps) resulted by the commercial 23andMe genetic testing company in people with
38 CFS.

39 Two items from the Perez et al. results were immediate red flags. They showed the top 50 deleterious
40 snps with the greatest difference in frequency between CFS patients and control data. At the very top
41 of the leaderboard was a snp on the gene GPBAR1 that was 129,000 times more prevalent in CFS
42 patients. If this finding is accurate then the authors may well have discovered THE genetic cause of
43 CFS rather than just a predisposition. In addition, the CYP2D6 gene on their list is recognizable as
44 one that in the past, for a different snp, had the majority of 23andMe customers believing they had a
45 poor xenobiotic metabolizer phenotype that only actually affects less than one percent of the
46 population (from Snpedia 3).

47 In the Perez et al. study, the data used to compare to the 23andMe CFS participants did not come
48 from 23andMe. Instead, they relied on published frequencies in the Kaviar database, a compilation
49 from multiple projects (4). Unless the control data comes from the identical source as the
50 experimental data, any differences in quality or population constituency between the two datasets
51 may lead to inaccurate conclusions when compared. How many of the reported frequency ratios
52 between CFS patients and controls would remain noteworthy if both data come from the same
53 source? The study established a criterion that a snp should be at least twice as prevalent in CFS
54 patients, a ratio of 2, to be of note. In addition, how closely do 23andMe data match published
55 genetic reference data? Errors in direct-to-consumer genetic tests have been reported (5).

56 To address these questions, we accessed a publicly available control set of genomes from 23andMe
57 participants. We reanalyzed the frequency ratio of the highly prevalent snps in CFS from the Perez et
58 al. study to this more appropriate control. We also incorporated an additional new large high-quality
59 online control dataset for further comparison among control datasets.

60 **2 Method**

61 To provide control data from 23andMe for the 50 top-predisposing CFS snps, we accessed publicly
62 available genome files on openSNP¹, a platform co-founded and maintained by one of us (BGT).
63 OpenSNP allows individuals to upload their own genetic results from a variety of commercial test
64 companies for public sharing (6). The allele frequencies and genotypes for these SNPs were
65 calculated for all 23andMe data sets present in openSNP on 2020-06-19. While self-selection of
66 participants can skew a dataset, the platform also allows phenotypes of interest to be added by
67 participants. We noted the absence of CFS and Myalgic Encephalomyelitis (ME) on the list of
68 phenotypes which provided an initial confidence that the dataset does not contain an
69 overrepresentation of CFS patients compared to the general population.

70 For the additional control: The allele frequency aggregation project (ALFA) project was developed as
71 part of the National Center for Biotechnology Information (NCBI) database of genotypes and
72 phenotypes (dbGAP) and had their inaugural release on March 10, 2020 (7). It contains a high-quality
73 aggregate of over 1200 studies with a goal of one million dbGAP subjects. ALFA (build 154, release
74 date April 21, 2020) was accessed through NCBI dbSNP². ALFA Europe was selected when

¹ <https://opensnp.org>

² <https://www.ncbi.nlm.nih.gov/snp>

75 available to match the population of 23andMe, primarily Americans of European decent. When
76 unavailable, ALFA Global was the second choice, followed by GnomAD – exome and then 1000
77 Genome project if absent from ALFA entirely. When ALFA is referred to subsequently, it is a
78 shortcut notation for the totality of this procedure.

79 To recalculate the ratio with the new controls, for the CFS data we used the frequencies provided by
80 Perez et al. in Table 1. Since neither the table nor supplementary materials explicitly listed the alleles,
81 we assumed that the frequencies always referred to the derived allele. In the event of multiple derived
82 alleles at a position, we further assumed the most prevalent one was used. Spot checks of their Kaviar
83 control supported that this was the study’s intended listing. The new ratios were recalculated for the
84 23andMe control and for the ALFA control and subsequently compared to the original ratios. The
85 three control datasets at the 50 snps (Kaviar, 23andMe, ALFA) were compared to each other.

86 **3 Results**

87 The recalculated ratio of allele frequency in CFS patients to control subjects using 23andMe control
88 data, or where unavailable, ALFA frequencies, along with the elimination of 2 duplicates, found that
89 only 11 of the 50 polymorphisms now exceeded a ratio of 2. That is, only 22 percent of the originally
90 reported polymorphisms remained at the original study criterion that notable snps were at least
91 double the frequency in CFS patients compared to unaffected controls.

92 Of the 11 remaining polymorphisms with a ratio that met the criterion, the majority, 7, could be
93 based only on ALFA frequencies; and 1 was not reported on ALFA either. These 8 snps were only
94 present in the 23andMe control data set in very few samples, ranging from 17 to 0, in contrast to a
95 median of nearly 3000 samples in the 23andMe control data overall. These further 8 snps therefore
96 were also not shown to have a higher prevalence in 23andMe CFS patients than in 23andMe controls.
97 Dates of upload to openSNP hint that the early days of the 23andMe v5 chip could be a source of
98 error. All 8 of these come from the top of the original ratio leaderboard (Table 1).

99 Only 3 snps of 50 remained, on genes EFCAB4B, LINC01171, and MORN2, that could be shown to
100 meet the original criterion of at least double in CFS patients compared to a comparable 23andMe
101 control, all hovering at a ratio of about 2.0. None of the astronomically high ratios of patients to
102 controls could be shown to remain.

103 Table 1 presents the ratios of allele frequency in CFS patients to control subjects (original,
104 recalculated with the new 23andMe control, recalculated with ALFA), the frequency of allele
105 occurrence (original CFS patients, original Kaviar, new 23andMe control, new ALFA control), and
106 the number of samples (23andMe control) for each of the snps reported in the original table of the
107 study. Genotype frequencies found in the 23andMe control samples for each snp is in Supplementary
108 Figure 1.

109 Comparison of the control datasets (Kaviar, 23andMe, ALFA) found 2 primary patterns. Most
110 prevalent, 29 out of the 48 of the 23andMe control frequencies were in good agreement with ALFA
111 with both being substantially higher than the reported Kaviar values, suggesting a Kaviar-related
112 error. We call this error Type A. For 8 of the 48 snps, the 23andMe control frequencies were instead
113 different (higher) than both ALFA and Kaviar, which were in good agreement with each other and
114 point to a 23andMe error (Type B error). Like the missing 23andMe snps, Type B errors came from
115 the top of the leaderboard and further accounts for the original astronomic reported ratios.

116 Concerning the two-red flag genes mentioned at the outset, both are Type B errors. The GPBAR1
117 snp for derived allele A was found in 97% of the 23andMe control sample compared to practically 0
118 in the other control data leading to a recalculation of the ratio of prevalence between CFS patients
119 and controls from the reported 129,000 to 0.8, a reversal. Based on upload dates, this erroneous base
120 A call may be traceable specifically to the v4 chip but would require exact test dates for
121 confirmation. The CYP2D6 had 2 snps, one of which had more than a third erroneous 23andMe call
122 for base A.

123 The correlation between the 23andMe control and ALFA control frequencies without the genes
124 above. and without any 23andMe data having fewer than 50 participants, was a respectable 0.93.

125 Discussion

126 The genetic predispositions reported for Chronic Fatigue Syndrome are not supported when
127 reanalyzed with more appropriate control data including those drawn from the same 23andMe pool as
128 the CFS patients. Out of the original 50 genomic positions presented to have the most prevalent
129 deleterious polymorphisms among CFS sufferers, only three remained that met the original study
130 criterion of at least twice as frequent as healthy individuals. The top-ranked risk factor on gene
131 GPBAR1 with an astronomical ratio of 129,000 was reduced to the more sensible 0.8, which would,
132 if anything, be a protective snp against CFS.

133 The erroneous odds ratios were found to originate from a mixture of errors in both 23andMe and in
134 the reported Kaviar control dataset. The more dramatic frequencies that had been listed as dozens,
135 hundreds, and thousands of times higher in CFS patients were due to seeming 23andMe peculiarities
136 of very high frequencies for minor alleles. We found these 23andMe errors to either also be present
137 in high numbers in the 23andMe controls (Genes GPBAR1, CYP2D6, PLA2G4D, CYP2A6, DDX5)
138 or quite often were simply missing from most samples. The number of CFS subjects from Perez et al.
139 at each of these snps, and overall, is unknown. The majority of the errors with less-striking ratio
140 inflations arose from the reported Kaviar control data where we found these to be lower frequencies
141 than both ALFA and 23andMe control datasets for many of the snps.

142 The Perez et al. discussion goes through, spelling out one by one, the function of each of the genes
143 from the top 10 in the table by summarizing what is known and including speculation for how these
144 factors may tie into CFS. For example, they suggest that decreased metabolism of xenobiotics may
145 be relevant to multiple chemical sensitivity disorder which is in turn relevant for some CFS cases and
146 a gene that is downregulated by sleep deprivation which in turn is a factor in chronic fatigue states.
147 The present reanalysis finds at the very least that such discussion and speculation are premature as
148 there is no evidence that any of those genes are relevant.

149 The three genes with polymorphisms that remained with the original study criteria of occurring at
150 least twice as often in CFS patients were EFCAB4B, MORN2 and LINC01171 which involve
151 calcium ion channels, cell differentiation, and a long non-coding RNA transcript respectively. It is
152 tempting to fish for connections such as to other ion channel polymorphisms that have been found
153 relevant in CFS (8) but here too, it is premature to speculate; reanalysis of the full 23andMe data for
154 CFS patients, as is now clearly warranted, may produce an entirely different top 50 leaderboard and
155 functional analysis. Likewise, the criterion of a ratio of at least double may also prove too stringent
156 which may then put some of the snps back into consideration but this too is unknown until a full
157 reanalysis. It is beyond the scope of this article to raise that high CADD scores also may not always
158 be an appropriate filter (e.g. 9). We conclude that the top-50 table presented in the CFS study does not

159 reflect the top 50 deleterious differences between Chronic Fatigue Syndrome and unaffected
160 individuals as intended but rather the top 50 errors in the 23andMe and Kaviar databases.

161 The present reanalysis highlights the need to use control data from the same commercial direct-to-
162 consumer genetic testing company when used for research. On a positive note, quirks aside, there is
163 generally high agreement between 23andMe and scientific genetic database ALFA. There are 10
164 million direct-to-consumer genetic test results which positively dwarfs the data collected in scientific
165 studies. Using the abundant commercial DNA results to find genetic predispositions is very appealing
166 especially for disorders without known cause, like Chronic Fatigue Syndrome The promise for
167 successful continued mining of public data for research purposes remains but with caution over select
168 snps and chips.

169 **Conflict of Interest**

170 *The authors declare that the research was conducted in the absence of any commercial or financial*
171 *relationships that could be construed as a potential conflict of interest.*

172 **Author Contributions**

173 Conceived the project (FB), extraction from openSNP (BGT), data analysis (FB, BGT), manuscript
174 writing (FB), manuscript edit (BGT).

175 **References**

- 176 1. Bedford FL. Sephardic signature in haplogroup T mitochondrial DNA. *Eur J Hum Genet.* 2012
177 Apr;20(4):441–8.
- 178 2. Perez M, Jaundoo R, Hilton K, Del Alamo A, Gemayel K, Klimas NG, et al. Genetic
179 Predisposition for Immune System, Hormone, and Metabolic Dysfunction in Myalgic
180 Encephalomyelitis/Chronic Fatigue Syndrome: A Pilot Study. *Front Pediatr [Internet].* 2019
181 [cited 2020 Jul 24];7. Available from:
182 <https://www.frontiersin.org/articles/10.3389/fped.2019.00206/full>
- 183 3. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation
184 and analysis. *Nucleic Acids Res.* 2012 Jan 1;40(D1):D1308–12.
- 185 4. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for
186 testing SNV novelty. *Bioinformatics.* 2011 Nov 15;27(22):3216–7.
- 187 5. Tandy-Connor S, Guiltinan J, Krempely K, LaDuca H, Reineke P, Gutierrez S, et al. False-
188 positive results released by direct-to-consumer genetic tests highlight the importance of clinical
189 confirmation testing for appropriate patient care. *Genetics in Medicine.* 2018 Dec;20(12):1515–
190 21.
- 191 6. Greshake B, Bayer PE, Rausch H, Reda J. openSNP—A Crowdsourced Web Resource for
192 Personal Genomics. *PLOS ONE.* 2014 Mar 19;9(3):e89204.
- 193 7. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI’s Database of Genotypes
194 and Phenotypes: dbGaP. *Nucleic Acids Res.* 2014 Jan;42(Database issue):D975-979.

195 8. Marshall-Gradisnik S, Smith P, Brenu E, Nilius B, Ramos S, Staines D. Examination of Single
 196 Nucleotide Polymorphisms (SNPs) in Transient Receptor Potential (TRP) Ion Channels in
 197 Chronic Fatigue Syndrome Patients. *Immunology and Immunogenetics Insights*. 2015 Apr 15;7.

198 9. Mather CA, Mooney SD, Salipante SJ, Scroggins S, Wu D, Pritchard CC, et al. CADD score has
 199 limited clinical validity for the identification of pathogenic variants in non-coding regions in a
 200 hereditary cancer panel. *Genet Med*. 2016 Dec;18(12):1269–75.

201

202 **Table 1** Recalculated ratios of polymorphism frequencies of CFS patients to controls

Gene	rsID	Frequency				PolyM	Ratio of CFS to Control			N	Flag
		ME/CFS	Kaviar control	23and Me control	Alfa control		Original (with Kaviar)	with 23and Me	with Alfa		
GPBAR1	rs199986029	77.3	0.0006	99.73	0	G to A/C	129,000	0.78	infinite	1835	
HLA-C	rs41560916	62.7	0.0013	-	0	C to A/G/T	48,200	-	infinite	0	*
BCAM	rs3810141	10.2	0.0006	-	6.3	C to A/T	17,000	-	1.62	2	*
AAAS	rs150511103	19.3	0.0013	8.33	0.01	C to A/G/T	14,900	2.32	1930.00	12	*
FGA	rs146387238	19.3	0.0013	0.00	0.03	C to A/G	14,900	infinite	643.33	17	*
SLC25A13	rs80338723	19.3	0.0013	0.00	0	C to A/G/T	14,900	infinite	infinite	17	*
MYBPC3	rs112738974	19.3	0.0019	3.13	0	C to A/G/T	10,200	6.18	infinite	16	*
PEX6	rs112298166	19.3	0.0019	-	-	C to G/T	10,200	-	-	-	*
CYP2D6	rs1135830	45.4	0.0097	35.31	0.01	G to A/T	4,680	1.29	4540.00	1892	
HLA-DRB1	rs112796209	41.5	0.0109	0.00	10.2	T to C	3,810	-	4.07	1	*
PLA2G4D	rs147516345	15.9	0.0103	18.20	0.5	T to C	1,550	0.87	31.80	1865	
CYP2A6	rs5031017	38.6	0.0264	30.50	0.1	C to A	1,460	1.27	386.00	1983	
CYP2D6	rs199535154	94.3	0.231	50.00	0.5	A to G	408	1.89	188.60	4	*
DDX51	rs201101053	15.9	0.0708	15.08	0	G to A	225	1.05	infinite	1873	
LHB	rs34349826	74.2	0.644	27.27	7	A to G	115	2.72	10.60	6	*
HLA-A	rs1137110	13.8	0.249	-	0.13	T to G	56	-	106.15	0	*
HLA-DRB1	rs1136756	43.9	1	50.00	30	T to C/G	44	-	1.46	1	*
HLA-DRB1	rs9269744	40.5	1.3	-	29.8	G to C	31	-	1.36	0	*
TPTE	rs1810540	34.5	1.16	30.29	34.8	C to A/T	30	1.14	0.99	1835	
HLA-DQA1	rs1061172	15.7	1.33	46.07	16.8	A to G	12	0.34	0.93	1922	
C6orf183	rs399561	63.2	6.46	40.68	40.7	G to A	10	1.55	1.55	3027	
C14orf37	rs3829765	81.5	9.75	51.54	54.5	G to A/T	8	1.58	1.50	3826	
EFCAB4B	rs11062745	27.9	3.39	13.78	15.5	T to C	8	2.02	1.80	3783	
PLD5	rs2810008	55.4	6.71	32.30	32.8	G to A/C/T	8	1.72	1.69	3024	

Running Title

MUC19	rs11564109	24	2.95	14.18	15	G to A	8	1.69	1.60	3825
ARHGAP42	rs17647207	14.4	1.82	8.83	9.5	G to A	8	1.63	1.52	3018
ADAMTS19	rs30645	76.5	9.75	51.01	51.5	T to A/C	8	1.50	1.49	3791
LINC01171	rs11605546	23	2.97	10.68	10.4	G to A	8	2.15	2.21	3826
ANKDD1B	rs34358	83.3	10.9	62.93	63.6	G to A/T	8	1.32	1.31	2990
ZBED5	rs2232919	12	1.61	6.56	7.4	T to C/G	7	1.83	1.62	2979
CTC-441N14.4	rs9112	60.3	8.44	40.47	41.8	G to A/C	7	1.49	1.44	2987
SLC35B2	rs3187	13.1	1.85	10.22	9	G to A	7	1.28	1.46	2887
PRSS41	rs61747737	11.5	1.63	7.01	7.9	T to A/G	7	1.64	1.46	1879
OTOG	rs12422210	26.4	3.76	15.09	17.3	G to A	7	1.75	1.53	2941
MTCH2	rs1064608	45.7	6.58	39.77	34.26	G to C/T	7	1.15	1.33	45
SULF1	rs6990375	51.2	7.49	30.51	30	G to A/T	7	1.68	1.71	3832
OTOG	rs11024333	29.5	4.34	16.20	16.3	G to A/C/T	7	1.82	1.81	3795
ART3	rs14773	43.3	6.41	28.20	26.7	C to A	7	1.54	1.62	2950
PPHLN1	rs12658	36.3	5.45	23.14	22.4	C to A./T	7	1.57	1.62	2991
PRICKLE1	rs12658	36.3	5.45	D	D		7	D	D	D
VARS2	rs2249464	74.7	11.4	55.74	54.4	T to C	7	1.34	1.37	3736
MORN2	rs3099950	21.9	3.37	11.08	12.1	G to A	7	1.98	1.81	2990
AC007956.1	rs2270424	36.8	5.99	21.37	20.4		6	1.72	1.80	3793
AREL1	rs2270424	36.8	5.99	dup	20.4	G to A	6	D	D	D
PRRT4	rs359642	95	15.5	80.00	82.3	G to A	6	1.19	1.15	3751
HUS1	rs2307252	16.7	2.76	11.45	9.9	G to A	6	1.46	1.69	2990
PRSS56	rs1550094	92.2	16.2	69.83	69.2	G to A/C/T	6	1.32	1.33	3674
C5orf52	rs10051838	24	4.35	13.38	13.3	G to A	6	1.79	1.80	3024
ZNHIT1	rs17319250	40.5	7.41	24.14	23.5	T to C	5	1.68	1.72	3798
CPLX2	rs3822674	70.5	12.9	49.23	48	T to A/C	5	1.43	1.47	2903

203 D = duplicate; - = missing value; N = number of participants; PolyM = polymorphisms

204 * Ratios obtained with fewer than twenty 23andMe control subjects were considered invalid