The American Journal of Human Genetics, Volume *99*

Supplemental Data

Mutations of the Sonic Hedgehog Pathway Underlie Hypothalamic Hamartoma with Gelastic Epilepsy

Michael S. Hildebrand, Nicole G. Griffin, John A. Damiano, Elisa J. Cops, Rosemary Burgess, Ezgi Ozturk, Nigel C. Jones, Richard J. Leventer, Jeremy L. Freeman, A. Simon Harvey, Lynette G. Sadleir, Ingrid E. Scheffer, Heather Major, Benjamin W. Darbro, Andrew S. Allen, David B. Goldstein, John F. Kerrigan, Samuel F. Berkovic, and Erin L. Heinzen

SUPPLEMENTAL INFORMATION

Supplemental Note: ***Additional statistical methods.***
Sequence mutability of genomic regions (genes, pathways) was calculated by summing a trinucleotide based mutation rate estimate [1] (kindly provided by Drs. Shamil Sunyaev and Paz Polak [2]) over each individual's 'callable real estate' [i.e., the regions of the exome, defined by the Consensus Coding Sequence (CCDS) (v14), that had sufficient coverage that a somatic mutation would likely be called if present, or specifically, the part of the exome that was sequenced at least ten times in both leukocyte and hamartoma DNA]. We then investigated the relationship between the presence of somatic mutations in our sample and the trinucleotide-based mutation rate as follows. First, for every base in the consensus coding sequence, we computed the probability of a mutation from the reference to one of the three alternative bases using the trinucleotide-based mutation rate estimate. From this we were able to compute the exome-wide median mutation probability. For each somatic mutation, we calculated the same probability of a mutation occurring at that site based on the trinucleotide mutation rate estimate. If this mutation rate is independent of the presence of a somatic mutation, we would expect that the mutation probabilities we observed would be a random sample from the overall mutation probability distribution and that about half of them would fall above the exome-wide median and half would fall below. However, we observed 43 somatic mutations were below the median and 140 were above, which is highly significant by the sign test (p=2.334e-11). Second, we conditioned on the trinucleotide context in which the somatic mutations occurs, and asked: given the three possible bases that a site could be mutated to, is the trinucleotide-based mutation rate affiliated with the observed somatic mutation more likely to be the highest of the three rates. We found that 64 somatic mutations were affiliated with the highest trinucleotide-based rate, 33 were affiliated with the lowest, and 76 had intermediate values. Again, we find significant enrichment for high mutation rates (sign test, p=0.002152) and conclude that the trinucleotide mutation rate estimates are positively correlated with occurrence of somatic mutations (and/or amplification artifacts). This finding motivated the incorporation of mutatability in our enrichment testing below.

The pathway enrichment analysis was performed by comparing the observed number of candidate variants in a pathway within each individual to that expected based on mutation rate. Specifically, for a given individual, we conditioned on the total number of mutations observed across their exome and computed the difference between the observed number of candidate variants within a gene or pathway to that expected given the proportion of the total mutability found within the pathway. Note that both the mutability of any gene, as well as the exome-wide mutability can vary from individual to individual, since each individual's 'callable real estate' can vary as described above. Also, note that this observed versus expected contrast is calculated within each individual, and therefore, explicitly accounts for differences in overall rates due to amplification. Individual departures from expectation are standardized by a variance estimate and then summed to give the observed test statistic. We estimate the null distribution of the test statistic by randomly distributing (100,000 times) the total number of mutations in each individual in accordance with the proportion of the total mutability found within the pathway and computing the statistic as described above. In calculating a p-value, we compute the proportion of simulated statistics that are as extreme or more extreme than that computed from the observed somatic mutatons. To account for the large number of hypotheses tested in each analysis we computed adjusted p-values using the resampling procedure of Ge et al. (2003) [3]. Gene-level enrichment analyses, comparing the observed number of candidate variants within each individual to that expected based on the mutation rate, were performed identically to the pathway enrichment analysis with the analysis unit being an individual gene rather than a list of genes comprising a pathway. To correct for the ~18,000 protein-coding genes defined in the CCDS, we used the more stringent Bonferroni correction.

The simulation analysis to assess the significance of gene enrichment in the Shh and salivary secretion pathways involved randomly shuffling CNVs or LOH events throughout the genome and assessing how often Shh and salivary secretion pathways are impacted by these events. Specifically, for each simulated dataset, we randomly placed, throughout the genome, CNVs or LOH events that were the same sizes as the CNVs or LOH events found by CMA. We then counted the number of genes in the Shh or salivary pathways that overlapped with simulated CNV or LOH events for each simulated dataset. The proportion

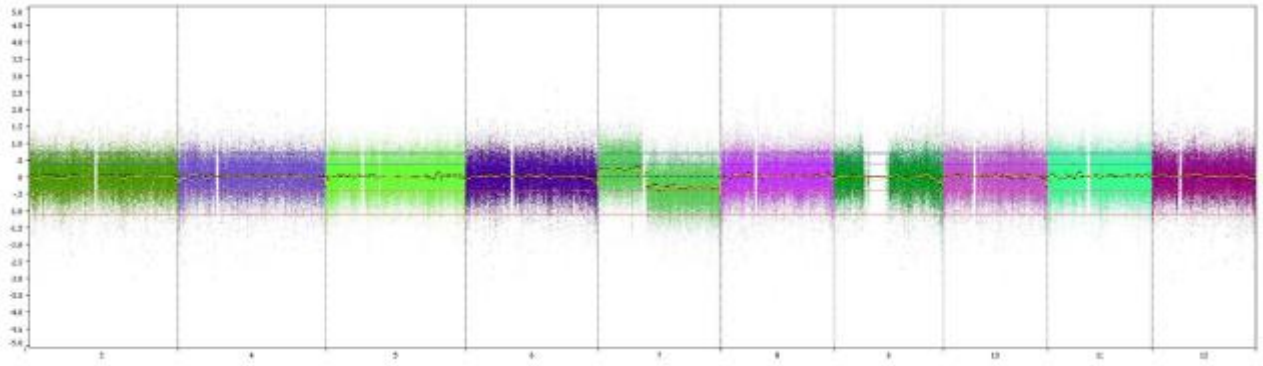
of simulated datasets for which the number of genes hit by the simulated lesions in these pathways greater than or equal to that found by CMA defined an empirical p-value for this assessment.

SUPPLEMENTAL FIGURES

Figure S1: Copy number mutations detected by CMA

For the Affymetrix CytoScan HD microarray experiments, processing of samples was performed by end point PCR amplification using DNA Taq polymerase (Clontech, Inc.; Mountain View, CA). The labeled patient DNA was hybridised to a human whole genome array containing 1.9 million non-polymorphic markers, as well as 750,000 SNP probes (Affymetrix; Santa Clara, CA), according to the manufacturer's instructions. Post-hybridisation procedures were performed according to the manufacturer's instructions. The ChAS (Chromosome Analysis Software) tool (version 1.1.2; Affymetrix) was used for feature extraction, calculation of log2 ratio values, and calculation of several quality control metrics according to the manufacturer's instructions. CNV calling and data interpretation were performed with the .CEL files using the Nexus Copy Number software tool (version 7.5, BioDiscovery; El Segundo, CA) and SNP-FASST2 and SNP-RANK algorithms supplied with the Nexus software suite. A minimum size threshold of 200-kb was used. **A.** Chromosomal microarray of hamartoma tissue in HH patient hht25063 showing copy number gain of chromosome 7p and copy number loss of chromosome 7q. **B.** Chromosomal microarray of blood in HH patient hht25063 showing normal copy number across chromosome 7.

A.



B.

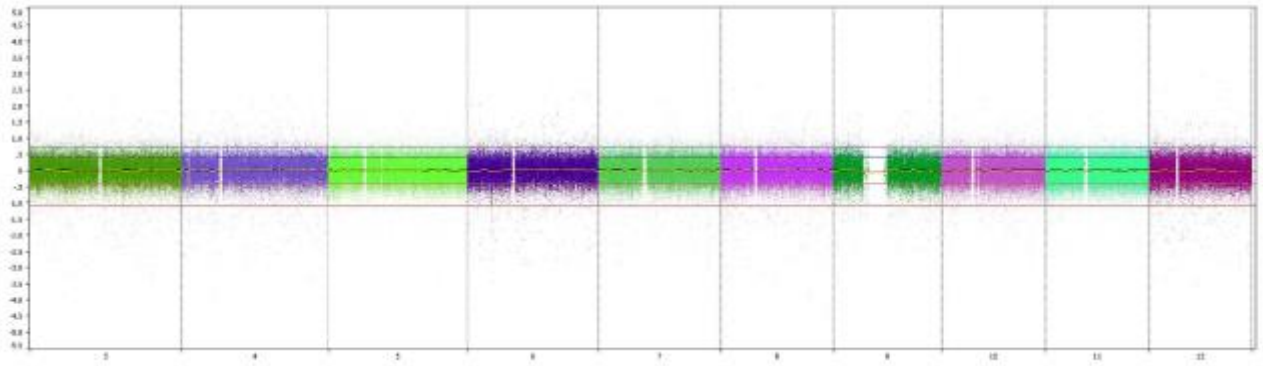
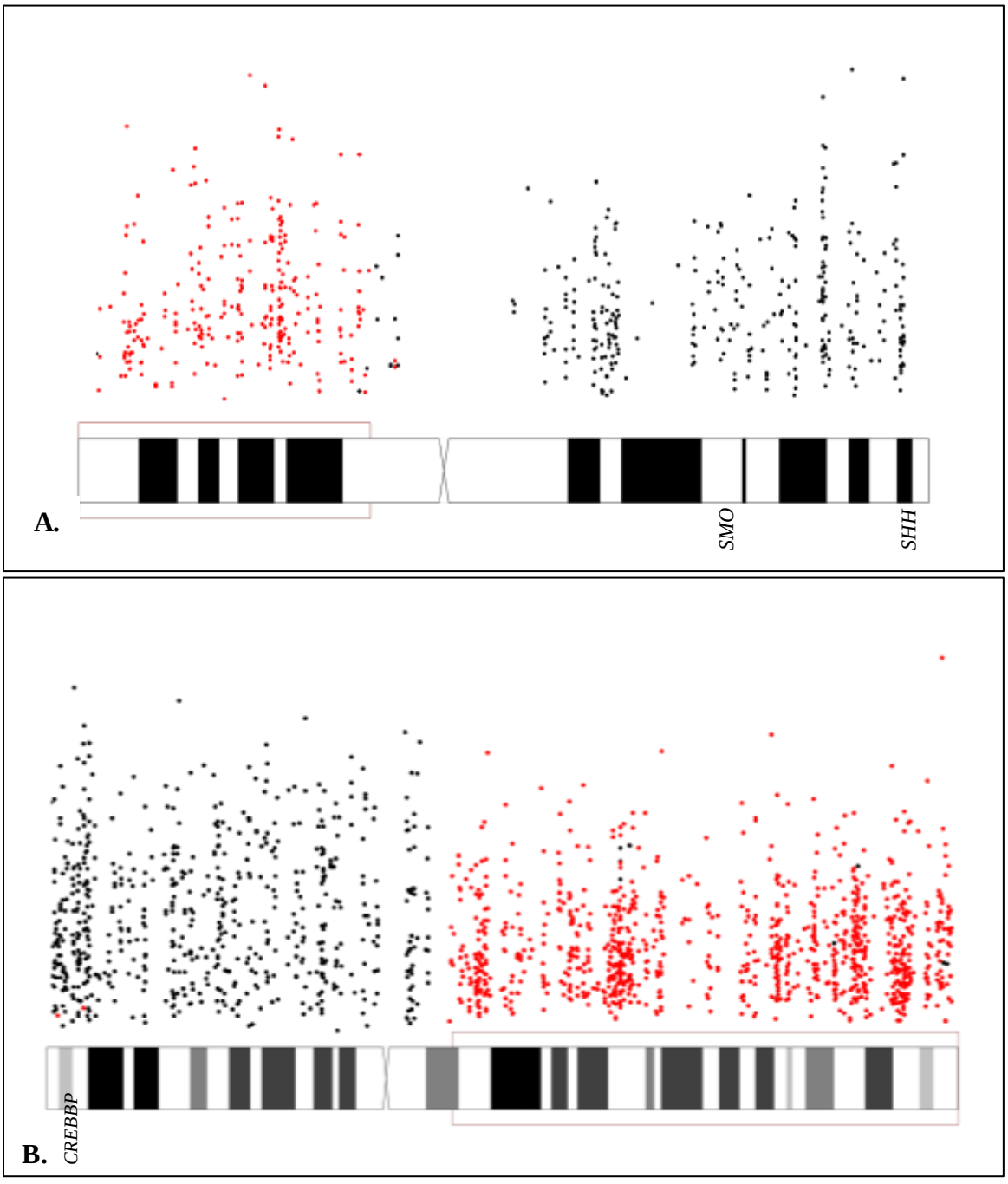


Figure S2: Copy number mutations detected by WES

Somatic LOH variants were filtered by requiring that there be at least five consecutive exonic LOH calls that spanned at least 1-kb (including noncoding sequence) with at least 10-fold sequencing coverage. To estimate the boundaries of the LOH event we plotted all germline variant calls with at least 25-fold coverage in hamartoma and leukocytes with a variant allele frequency between 40% and 60% and all LOH calls meeting the above criteria. **A.** A large region of homozygosity (ROH) on chromosome 7q indicates a loss of heterozygosity (LOH) in HH patient hht1198 that includes the *SMO* and *SHH* genes. **B.** ROH on chromosome 16p indicates a large somatic region of LOH in HH patient hht735 that includes the *CREBBP* gene. In **A** and **B** black dots indicate germline variant calls, and red dots indicate LOH variant calls.



SUPPLEMENTAL TABLES

Table S1: Summary of clinical data from 38 patients

Patient Number	Shh Gene (KEGG)	Gender	Seizure types	Seizure frequency	Refractory (Y/N)	Age of onset	Age at surgery	IQ<70 (Y/N)	Precocious puberty	HH type [4]	Volume (cm^3)
hht25057	*CREBBP*[a]	M	Multi	11-20/day	Y	1 month	16 years	Y	Y	2	1.26
hht25063	*GLI3, SHH, SMO, WNT16, WNT2*	F	Multi	6-10/day	Y	1 month	2 years	N	N	2	0.74
hht25085	*PRKACA*	F	Multi	1-5/day	Y	5 years	17 years	N	N	3	1.06
hht25086	*PRKACA*	M	Gelastic	11-20/day	Y	1 month	3 years	Y	N	3	4.14
hht25077	*GLI3*	F	Multi	1-5/day	Y	3 months	10 years	N	N	2	0.24
hht25094	*WNT11*	F	Gelastic	1-5/day	Y	4 years	8 years	Y	Y	2	0.07
hht209	*GLI3*	M	Multi	1-5/day	Y	Birth	10 years	Y	N	2	0.33
hht26139	*GLI3*	M	Multi	1-5/day	Y	Birth	4 years	Y	N	2	0.58
hht238a	*PRKACA*	F	Multi	1-5/day	Y	Birth	8 years	N	Y	2	1.94
hht1198c	*SHH, SMO, WNT16, WNT2*	F	Multi	11-20/day	Y	Birth	13 years	Y	Y	3	3.39
hht735	*CREBBP*[a]	M	Multi	1-5/day	Y	Birth	9 years	Y	Y	3	14.21
hht953	*BMP4*	M	Multi	6-10/day	Y	Birth	23 years	N	N	2	1.21
hht880	*GLI2, IHH, LRP2, STK36, WNT10A, WNT6*	M	Multi	6-10/day	Y	2 years	9 years	Y	N	2	0.69
hh31536	*GLI3*	M	Gelastic	6-10/day	Y	Birth	22 months	N	N	2	0.13
hht25056		M	Multi	6-10/day	Y	3 years	12 years	Y	N	3	0.42
hht25059		F	Gelastic	11-20/day	Y	1 month	5 years	N	N	3	2.91
hht25050		M	Multi	1-5/day	Y	7 years	31 years	Y	N	2	0.13
hht25092		M	Multi	1-5/day	Y	7 years	15 years	Y	Y	1	0.04
hht25186		F	Multi	1-5/day	Y	1 month	8 years	N	N	2	0.66
hht25093g		M	Multi	6-10/day	Y	1 month	18 years	Y	Y	2	1.91
hht25079		M	Multi	1-5/day	Y	9 months	29 years	Y	N	4	4.51
hht25097		M	Multi	1-5/day	Y	6 months	13 years	Y	N	2	0.91
hht25098		M	Multi	1-5/day	Y	1 month	19 years	Y	N	2	1.10
hht25080		M	Multi	1-5/day	Y	2 years	22 years	N	Y	2	0.19
hht25099		M	Multi	1-5/day	Y	8 months	9 years	Y	N	2	0.57
hht25082		M	Multi	1-5/day	Y	1 month	14 months	Y	N	2	0.15

hht25132h		M	Multi	6-10/day	Y	3 months	11 years	N	N	2	0.20
hht322b		F	Multi	1-5/day	Y	3 months	4 years	N	N	2	0.56
hht786		F	Multi	6-10/day	Y	Birth	8 years	N	Y	2	1.12
hht929		M	Multi	6-10/day	Y	Birth	5 years	Y	N	2	2.14
hht1276d		M	Multi	1-5/day	Y	12 months	10 years	Y	N	2	1.43
hht20138		F	Multi	11-20/day	Y	9 months	13 years	N	Y	3	1.04
hht25052		M	Multi	11-20/day	Y	1 month	2 years	Y	N	2	0.34
hht25054		M	Multi	1-5/day	Y	1 month	13 years	Y	Y	4	8.23
hht25060		M	Multi	1-5/day	Y	1 month	24 years	Y	Y	2	1.29
hht25064		M	Gelastic	2/week	Y	1 month	11 years	N	Y	2	1.34
hht25066		M	Gelastic	11-20/day	Y	1 month	2 years	Y	Y	3	2.05
hht25072		M	Multi	11-20/day	Y	2 years	4 years	Y	Y	2	2.05

[a]***Transcriptional regulator of the Shh pathway;*** **F: female; M: male; Multi: multiple seizure types; N: no; Y: yes**

Table S2: Coverage and mutational burden summary of somatic variants from whole exome sequencing of tumor and leucocyte-derived DNA in 15 HH patients. sSNV= somatic single nucleotide variant; sindel= somatic indel.

	Amplified	Brain tissue coverage	Blood coverage	% of Shh genes sequenced at least 10-fold in paired DNA samples	All sSNVs	Candidate sSNVs	All sindels	Candidate sindels
hht238a	no	122.73	225.92	87.9%	193	2	221	1
hht322b	no	67.07	184.13	86.6%	269	3	200	0
hht1198c	no	92.16	82.01	85.9%	136	2	167	0
hht1276d	no	103.48	202.19	88.2%	154	2	229	0
hht25093g	no	62.54	101.87	84.8%	164	4	162	0
hht25132h	no	92.01	107.06	85.1%	135	7	138	0
hht209	yes	134.33	148.97	83.2%	681	77	577	13
hht735	yes	146.43	108.65	77.6%	1755	9	163	3
hht786	yes	135.49	107.78	79.8%	254	5	163	3
hht929	yes	78.84	173.07	69.5%	180	6	762	26
hht25080	yes	150.64	124.72	83.4%	242	4	198	4
hht25082	yes	140.54	127.65	79.1%	238	5	199	3
hht25086	yes	178.71	95.72	68.8%	398	9	174	2
hht25099	yes	180.15	36.7	50.2%	665	33	115	0
hht26139	yes	120.19	153.16	74.7%	148	5	347	0
Mean		**120**	**132**	**79.0%**	**374**	**12**	**254**	**4**
Standard Deviation		**37**	**50**	**10.0%**	**421**	**20**	**180**	**7**

Table S3: Candidate somatic variants called from WES and TRS of paired harmatoma-leukocyte DNA samples from individuals with HH

See excel sheet attached.

SUPPLEMENTAL REFERENCES

1. Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80, 727-739.
2. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G., et al. (2015). Genome-wide patterns and properties of de novo mutations in humans. Nat Genet 47, 822-826.
3. Ge, Y., Dudroit, S., and Speed, T.P. (2003). Resampling-based multiple testing for microarray data hypothesis. Test 12, 1-77.

SUPPLEMENTAL ACKNOWLEDGEMENTS

We would like to acknowledge the following individuals or groups for the contribution of control samples: D. Daskalakis; R Buckley; M. Hauser; J. Hoover-Fong, N. L. Sobreira and D. Valle; A. Poduri; T. Young and K. Whisenhunt; Z. Farfel, D. Lancet, and E. Pras; R. Gbadegesin and M. Winn; K. Schmader, S. McDonald, H. K. White and M. Yanamadala; R. Brown; S. H. Appel; E. Simpson; S. Halton, L. Lay; A. Holden; E. Behr; C. Moylan; A. M. Diehl and M. Abdelmalek; S. Palmer; G. Cavalleri; N. Delanty; G. Nestadt; D. Marchuk; V. Shashi; M. Carrington; R. Bedlack,; M. Harms; T. Miller; A. Pestronk; R. Bedlack; R. Brown; N. Shneider; S. Gibson; J. Ravits; A. Gilter; J. Glass; F. Baas; E. Simpson; and G. Rouleau; The ALS Sequencing Consortium; The Murdock Study Community Registry and Biorepository; the Carol Woods and Crosdaile Retirement Communities; Washington University Neuromuscular Genetics Project; the Utah Foundation for Biomedical Research; and DUHS (Duke University Health System) Nonalcoholic Fatty Liver Disease Research Database and Specimen Repository. The collection of control samples and data was funded in part by: Biogen Idec.; Gilead Sciences, Inc.; New York-Presbyterian Hospital; The Columbia University College of Physicians and Surgeons; The Columbia University Medical Center; The Duke Chancellor's Discovery Program Research Fund 2014; Bill and Melinda Gates Foundation; The Stanley Institute for Cognitive Genomics at Cold Spring Harbor Laboratory; B57 SAIC-Fredrick Inc M11-074; The Ellison Medical Foundation New Scholar award AG-NS-0441-08; National Institute of Mental Health (K01MH098126, R01MH099216, R01MH097993); National Institute of Allergy and Infectious Diseases (1R56AI098588-01A1); National Human Genome Research Institute (U01HG007672); National Institute of Neurological Disorders and Stroke (U01-NS077303, U01-NS053998); and National Institute of Allergy and Infectious Diseases Center (U19-AI067854, UM1-AI100645).