HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2011 March 01.

Published in final edited form as:

Nat Genet. 2010 September ; 42(9): 790–793. doi:10.1038/ng.646.

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

Sarah B. Ng[1,*], **Abigail W. Bigham**[2,*], **Kati J. Buckingham**[2], **Mark C. Hannibal**[2,3], **Margaret McMillin**[2], **Heidi Gildersleeve**[2], **Anita E. Beck**[2,3], **Holly K. Tabor**[2,3], **Greg M. Cooper**[1], **Heather C. Mefford**[2], **Choli Lee**[1], **Emily H. Turner**[1], **Josh D. Smith**[1], **Mark J. Rieder**[1], **Koh-ichiro Yoshiura**[4], **Naomichi Matsumoto**[5], **Tohru Ohta**[6], **Norio Niikawa**[6], **Deborah A. Nickerson**[1], **Michael J. Bamshad**[1,2,3,†], and **Jay Shendure**[1,†]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA

[2]Department of Pediatrics, University of Washington, Seattle, Washington, USA

[3]Seattle Children's Hospital, Seattle, Washington, USA

[4]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan

[5]Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan

[6]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Hokkaido, Japan

Abstract

We demonstrate the successful application of exome sequencing1–3 to discover a gene for an autosomal dominant disorder, Kabuki syndrome (OMIM %147920). The exomes of ten unrelated probands were subjected to massively parallel sequencing. After filtering against SNP databases,

†Corresponding authors: Jay Shendure, MD, PhD, Department of Genome Sciences, University of Washington School of Medicine, Box 355065, 1705 NE Pacific Street, Seattle, WA 98195, Mike Bamshad, MD, Department of Pediatrics, University of Washington School of Medicine, Box 356320, 1959 NE Pacific Street, Seattle, WA 98195.

*These authors contributed equally to this work.

URLs

Refseq 36.3 : ftp://ftp.ncbi.nlm.nih.gov/genomes/MapView/Homo_sapiens/sequence/BUILD.36.3/updates/seq_gene.md.gz

Phaster : http://www.phrap.org

SeattleSeq Annotation : http://gvs.gs.washington.edu/SeattleSeqAnnotation/

1000 Genomes Project: : http://www.1000genomes.org/page.php

Data access

Exome data for the discovery cohort will be available via the NCBI dbGaP repository.

Author contributions

The project was conceived and experiments planned by M.J.B., D.A.N., and J.S. Review of phenotypes and sample collection were performed by M.J.B., M.C.H., M.M., K.Y., N.M., T.O. and N.N. Experiments were performed by S.B.N., K.J.B., A.E.B., C.L., H.C.M. J.D.S., M.J.R., E.H.T., and H.G. Ethical consultation was provided by H.K.T. Data analysis was performed by A.W.B., M.J.B., K.J.B., G.M.C., S.B.N. and J.S. The manuscript was written by M.J.B., S.B.N., and J.S. All aspects of the study were supervised by M.J.B. and J.S.

Competing Interests Statement

The authors declare no competing financial interests.

there was no compelling candidate gene containing novel variants in all affected individuals. Less stringent filtering criteria permitted modest genetic heterogeneity or missing data, but identified multiple candidate genes. However, genotypic and phenotypic stratification highlighted *MLL2*, a *Trithorax*-group histone methyltransferase4, in which seven probands had novel nonsense or frameshift mutations. Follow-up Sanger sequencing detected *MLL2* mutations in two of the three remaining cases, and in 26 of 43 additional cases. In families where parental DNA was available, the mutation was confirmed to be *de novo* (n = 12) or transmitted (n = 2) in concordance with phenotype. Our results strongly suggest that mutations in *MLL2* are a major cause of Kabuki syndrome.

Kabuki syndrome is a rare, multiple malformation disorder characterized by a distinctive facial appearance (Supplementary Fig. 1), cardiac anomalies, skeletal abnormalities, immunological defects, and mild to moderate mental retardation. Originally described by Niikawa *et al*.5 and Kuroki *et al*.6 in 1981, Kabuki syndrome has an estimated incidence of 1 in 32,0007 and about 400 cases have been reported worldwide. The vast majority of reported cases have been sporadic, but parent-to-child transmission in more than a half a dozen instances8 suggests that Kabuki syndrome is an autosomal dominant disorder. The relatively low number of cases, the lack of multiplex families, and the phenotypic variability of Kabuki syndrome have made the identification of the gene(s) underlying Kabuki syndrome intractable to conventional approaches of gene discovery, despite aggressive efforts.

We sequenced the exomes of ten unrelated individuals with Kabuki syndrome, seven of European ancestry, two of Hispanic ancestry, and one of mixed European/Haitian ancestry (Supplementary Fig. 1, Supplementary Table 1). Enrichment was performed by hybridization of shotgun fragment libraries to custom microarrays, followed by massively parallel sequencing1–3. On average, 6.3 gigabases of sequence were generated per sample to achieve 40× coverage of the mappable, targeted exome (31 megabases). As previously, our analyses focused primarily on nonsynonymous (NS) variants, splice acceptor and donor site mutations (SS) and coding indels (I), anticipating that synonymous variants were far less likely to be pathogenic. We also predicted that variants underlying Kabuki syndrome are rare, and therefore likely to be novel. Novelty was defined here by absence from all datasets used for comparison, including dbSNP129, the 1000 Genomes Project, exome data from sixteen individuals previously reported by us2,3, and ten exomes sequenced as part of the Environmental Genome Project (EGP).

Under a dominant model in which each case was required to have at least one novel NS/SS/I variant in the same gene, only a single candidate gene (*MUC16*) was shared by all ten exomes (Table 1, row 4; Supplementary Table 2). However, *MUC16* was considered likely to be a false positive due to its extremely large size (14,507 aa). Potential explanations for our failure to find a compelling candidate gene in which novel variants are observed in all affected individuals included: (a) that Kabuki syndrome is genetically heterogeneous, and therefore not all affected individuals will have mutations in the same gene; (b) that we failed to identify all mutations in the targeted exome; (c) that some or all causative mutations were outside of the targeted exome, e.g., in non-coding regions or unannotated genes. To allow

for a modest degree of genetic heterogeneity and/or missing data, we conducted a less stringent analysis by looking for candidate genes shared among subsets of affected individuals. Specifically, we searched for subsets of *x* out of 10 exomes having ≥ 1 novel variant in the same gene, for *x* = 1 to 10. For *x* = 9, 8, and 7, novel variants were shared in three genes, six genes, and sixteen genes, respectively (Table 1, row 4). However, there was no obvious way to rank these candidates.

We speculated that genotypic and/or phenotypic stratification would facilitate the prioritization of candidate genes identified by subset analysis. Specifically, we assigned a categorical rank to each Kabuki case based on a subjective assessment of the presence of, or similarity to, the canonical facial characteristics of Kabuki syndrome (Supplementary Fig. 1) and the presence of developmental delay and/or major birth defects (Supplementary Table 1). The highest ranked case was one of a pair of monozygotic twins with Kabuki syndrome. We then categorized the functional impact (i.e. nonsense versus nonsynonymous substitution, splice-site disruption, frameshift versus in-frame indel) of each novel variant in candidate genes shared by each subset of two or more ranked cases. Manual review of these data highlighted distinct, novel nonsense variants in *MLL2* in each of the four highest ranked cases. On sequential analysis of phenotype-ranked cases with a loss-of-function filter, *MLL2* is the only candidate gene remaining after addition of the second individual (Table 2, row 5, column "+2"). No novel variant in *MLL2* was found in the Kabuki case ranked 5th, such that the number of candidate genes drops to zero after the fourth individual (Table 2, row 5). However, a 4-bp deletion was found in the case ranked 6th and nonsense variants in the cases ranked 7th and 9th. Thus, exome sequencing identified a nonsense substitution or frameshift indel in *MLL2* in seven of the ten Kabuki cases.

Retrospectively, if we apply a loss-of-function filter to the subset analysis of exome data (Table 1, row 5), at *x* = 7, *MLL2* is the only candidate gene. We also developed a *post hoc* ranking of candidate genes based on functional impact of variants present (“variant score”) and the rank of the cases in which each variant was observed (“case score”). When applied to the exome data as a combined metric, *MLL2* emerges as the top candidate (Supplementary Fig. 2).

In parallel with these analyses, we applied genomic evolutionary rate profiling (GERP)9 to exome data. GERP uses mammalian genome alignments to define a rejected substitution (RS) score for each variant, regardless of functional class. We have previously shown that the quantitative ranking of candidate genes by the RS scores of their novel variants can facilitate the exome-based analysis of Mendelian disorders10. In subset analysis with GERP-based ranking, *MLL2* remains on the candidate list up to *x* = 8, ranking 3rd in a list of 11 candidate genes at this threshold (Table 3, Supplementary Fig. 3). Interestingly, the additional *MLL2* variant contributing to this analysis (such that *MLL2* is still considered at *x* = 8) is a synonymous substitution with an RS score of 0.368 in the 5th ranked case.

We sought to confirm all novel variants identified in *MLL2*, particularly because loss-of-function variants identified through massively parallel sequencing have a higher prior probability of being false positives. All seven loss-of-function variants in *MLL2* were validated by Sanger sequencing. We further analyzed the three cases in which we did not

initially find a loss-of-function variant in *MLL2*, first by array comparative genomic hybridization (aCGH) to determine any gross structural changes, and then by Sanger sequencing of all exons of *MLL2* in case of false negatives by exome sequencing. Since an average of 96% of coding bases in *MLL2* were called at sufficient quality and coverage for single-nucleotide variant detection, we anticipated that any missed variants were more likely to be indels instead, because of the higher coverage required for confident indel detection in short-read sequence data. Indeed, although aCGH did not find any structural variants in the region, Sanger sequencing did identify frameshift indels in two of these three cases (ranked 8^{th} and 10^{th}).

Ultimately, loss-of-function mutations in *MLL2* were identified in nine out of ten cases in the discovery cohort (Fig. 1), making it a compelling candidate for Kabuki syndrome. For validation, we screened all 54 exons of *MLL2* in 43 additional cases by Sanger sequencing. Novel nonsynonymous, nonsense or frameshift mutations in *MLL2* were found in 26 of these 43 cases (Fig. 1 and Supplementary Table 3). In total, through either exome sequencing or targeted sequencing of *MLL2*, 33 distinct *MLL2* mutations were identified in 35 of 53 families (66%) with Kabuki syndrome (Fig. 1 and Supplementary Table 3). In each of twelve cases for which DNA from both parents was available, the *MLL2* variant was found to have occurred *de novo*. Three mutations were found in two cases each: one mutation was confirmed to have arisen *de novo* in one of the cases, indicating that some mutations are recurrent. Novel *MLL2* mutations (K4527X and T5464M) were also identified in each of two families in which Kabuki syndrome was transmitted from parent-to-child. None of the additional *MLL2* mutations were found in 190 control chromosomes from individuals of matched geographical ancestry.

Our results strongly suggest that mutations in *MLL2* are a major cause of Kabuki syndrome. *MLL2* encodes a large 5,262 residue protein that is part of the SET family of proteins, of which *Trithorax*, the *Drosophila* homologue of *MLL*, is the best characterized11. The SET domain of *MLL2* confers strong histone 3 lysine 4 methyltransferase activity and is important in the epigenetic control of active chromatin states12. Murine loss of *Mll2* on a mixed 129Sv/C57BL/6 background slows growth, increases apoptosis and retards development leading to early embryonic lethality, due in part to mis-regulation of homeobox gene expression13. However, no morphological defects have been reported in $Mll2^{+/-}$ mice13.

Most of the *MLL2* variants identified in Kabuki cases are predicted to truncate the polypeptide chain before translation of the SET domain. Accordingly, though it is not certain whether Kabuki syndrome results from haploinsufficiency or a gain-of function at *MLL2*, haploinsufficiency seems to be the more likely mechanism. Deletion of chromosome 12q12-q13.2, which encompasses *MLL2*, has been reported in a child with characteristics of Noonan syndrome14. However, we re-analyzed this case using oligo aCGH (including 21 probes that cover *MLL2*) and found the distal breakpoint to be located ~700 kb proximal of *MLL2* (data not shown). Interestingly, all of the pathogenic missense variants identified herein are located in regions of *MLL2* that encode C-terminal domains. This suggests that missense variants elsewhere in *MLL2* could be better tolerated or, alternatively, are embryonic lethal.

For the 18 of 53 cases for which no novel protein-altering variant was found, it is possible that non-coding or other missed mutations in *MLL2* are responsible instead. Alternatively, Kabuki syndrome could be genetically heterogeneous, and further analysis of these cases by exome sequencing may elucidate additional genes for Kabuki syndrome and potentially explain some of the phenotypic heterogeneity seen in this disease. Notably, 9 of 10 individuals in the discovery cohort (90%), but only 26 of 43 individuals in the replication cohort (60%), were ultimately found to have mutations in *MLL2*. It is therefore possible that the careful selection of canonical Kabuki cases for the discovery cohort enriched for a shared genetic basis. This underscores the importance of access to deeply phenotyped and well-characterized cases.

In summary, we applied exome sequencing of a small number of unrelated cases to discover that mutations in *MLL2* underlie Kabuki syndrome. As predicted in previous analyses2,3, allowing for even a small degree of genetic heterogeneity or missing data significantly confounds exome analysis by increasing the number of candidate genes consistent with the model of inheritance. To facilitate the prioritization of genes under such criteria, we stratified data by ranked phenotypes and found that *MLL2* was prominent in the higher ranked cases. However, nine of the ten Kabuki cases in the discovery cohort were ultimately found to have *MLL2* mutations, such that stratification by phenotype was of less importance than originally appeared to be the case. Nonetheless, the sequential analysis of ranked cases may have reduced the probability of confounding due to genetic heterogeneity. All of the *MLL2* mutations found in the discovery set via exome sequencing were loss-of-function variants. As a result, *MLL2* ranked highly among candidates assessed by predicted functional impact. Such a pattern will likely occur for some, but not all, Mendelian phenotypes subjected to this approach. We anticipate that the further development of strategies to stratify data at both the genotypic and phenotypic level will be critical for exome and whole genome sequencing to reach their full potential as tools for discovery of genes underlying Mendelian and complex diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the families for their participation and the Kabuki Syndrome Network for their support. We thank J. Allanson, J. Carey, and M. Golabi for referral of cases and M. Emond for helpful discussion. We thank the 1000 Genomes Project for early data release that proved useful for filtering out common variants. Our work was supported in part by grants from the National Institutes of Health/National Heart Lung and Blood Institute (5R01HL094976 to D.A.N. and J.S.), the National Institutes of Health/National Human Genome Research Institute (5R21HG004749 to J.S., 1RC2HG005608 to M.J.B., D.A.N., and J.S.; and 5RO1HG004316 to H.K.T.), National Institute of Heath/National Institute of Environmental Health Sciences (HHSN273200800010C to D.N. and M.R.), Ministry of Health, Labour and Welfare (K.Y., N.M., T.O., and N.N.), Japan Science and Technology Agency (N.M.), Society for the Promotion of Science (N.M.), the Life Sciences Discovery Fund (2065508 and 0905001), the Washington Research Foundation, and the National Institutes of Health/National Institute of Child Health and Human Development (1R01HD048895 to M.J.B.). S.B.N. is supported by the Agency for Science, Technology and Research, Singapore. A.W.B. is supported by a training fellowship from the National Institutes of Health/National Human Genome Research Institute (T32HG00035).

References

1. Choi M, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009; 106:19096–19101. [PubMed: 19861545]
2. Ng SB, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42:30–35. [PubMed: 19915526]
3. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]
4. FitzGerald KT, Diaz MO. MLL2: A new mammalian member of the trx/MLL family of genes. Genomics. 1999; 59:187–192. [PubMed: 10409430]
5. Niikawa N, Matsuura N, Fukushima Y, Ohsawa T, Kajii T. Kabuki make-up syndrome: a syndrome of mental retardation, unusual facies, large and protruding ears, and postnatal growth deficiency. J Pediatr. 1981; 99:565–569. [PubMed: 7277096]
6. Kuroki Y, Suzuki Y, Chyo H, Hata A, Matsui I. A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. J Pediatr. 1981; 99:570–573. [PubMed: 7277097]
7. Niikawa N, et al. Kabuki make-up (Niikawa-Kuroki) syndrome: a study of 62 patients. Am J Med Genet. 1988; 31:565–589. [PubMed: 3067577]
8. Courtens W, Rassart A, Stene JJ, Vamos E. Further evidence for autosomal dominant inheritance and ectodermal abnormalities in Kabuki syndrome. Am J Med Genet. 2000; 93:244–249. [PubMed: 10925391]
9. Cooper GM, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005; 15:901–913. [PubMed: 15965027]
10. Cooper GM, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods. 7:250–251. [PubMed: 20354513]
11. Prasad R, et al. Structure and expression pattern of human ALR, a novel gene with strong homology to ALL-1 involved in acute leukemia and to Drosophila trithorax. Oncogene. 1997; 15:549–560. [PubMed: 9247308]
12. Issaeva I, et al. Knockdown of ALR (MLL2) reveals ALR target genes and leads to alterations in cell adhesion and growth. Mol Cell Biol. 2007; 27:1889–1903. [PubMed: 17178841]
13. Glaser S, et al. Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. Development. 2006; 133:1423–1432. [PubMed: 16540515]
14. Tonoki H, Saitoh S, Kobayashi K. Patient with del(12)(q12q13.12) manifesting abnormalities compatible with Noonan syndrome. Am J Med Genet. 1998; 75:416–418. [PubMed: 9482650]

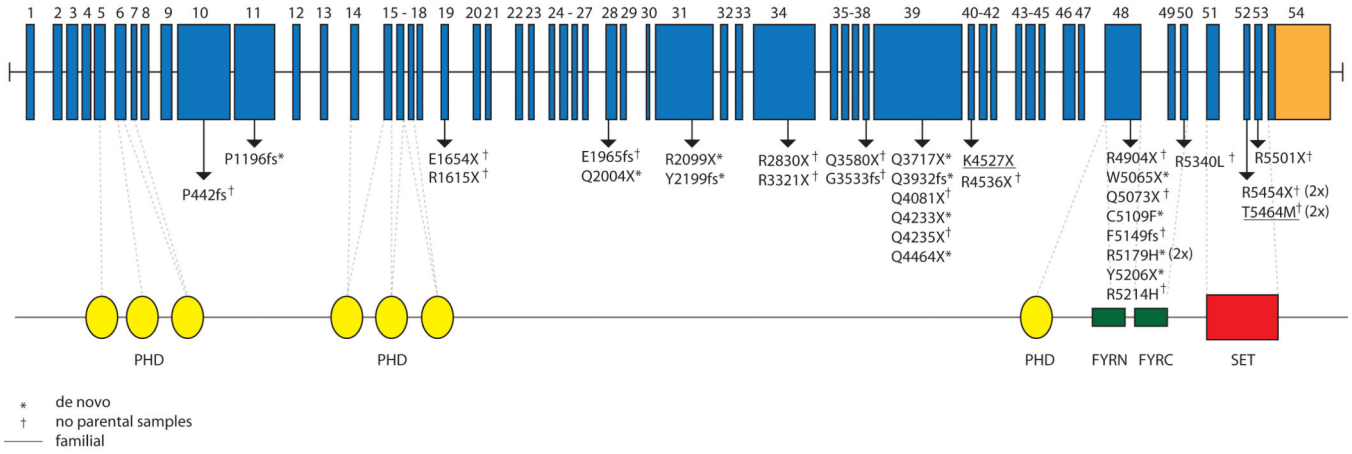

Figure 1. Genomic structure and allelic spectrum of *MLL2* mutations that cause Kabuki syndrome

MLL2 is composed of 54 exons that encode untranslated regions (orange) and protein coding sequence (blue) including 7 PHD fingers (yellow), FYRN (green), FYRC (green), and a SET domain (red). Arrows indicate the locations of 32 different mutations found in 53 families with Kabuki syndrome including: 20 nonsense, 7 indels, and 5 amino acid substitutions. Asterisks indicate mutations that were confirmed to be *de novo* and crosses indicate cases for which parental DNA was unavailable.

Table 1

Number of genes common to any subset of *x* affected individuals

The number of genes with at least one non-synonymous variant (NS), splice-site acceptor/donor variants (SS) or coding indel (I) are listed under various filters. Variants were filtered by presence in dbSNP or 1000 genomes ("Not in dbSNP129 or 1000 genomes") and control exomes ("Not in control exomes") or both ("Not in either"); control exomes refer to those from 8 Hapmap3 4 FSS3, 4 Miller2 and 10 EGP samples. The number of genes found using the union of the intersection of x individuals is given.

a. Subset analysis (any *x* of 10)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
NS/SS/I	12,042	8,722	7,084	6,049	5,289	4,581	3,940	3,244	2,486	1,459
Not in dbSNP129 or 1000 genomes	7,419	2697	1057	488	288	192	128	88	60	34
Not in control exomes	7,827	2865	1025	399	184	90	50	22	7	2
Not in either	6,935	2227	701	242	104	44	16	6	3	1
Is loss-of-function (nonsense/frameshift indel)	753	49	7	3	2	2	1	0	0	0

Table 2

Number of genes common in sequential analysis of phenotypically ranked individuals

Variants were filtered as in Table 1. Exomes were added sequentially to the analysis by ranked phenotype, e.g. column "+ 3" shows the number of genes at the intersection of the three top ranked cases. (Supplementary Fig. 1). The gene with at least one NS/SS/I in all individuals is *MUC16* which is very likely to be a false positive due to its extreme length (14,507 aa).

b. Sequential analysis	1	+2	+3	+4	+5	+6	+7	+8	+9	+10
NS/SS/I	5,282	3,850	3,250	2,354	2,028	1,899	1,772	1,686	1,600	1,459
Not in dbSNP129 or 1000 genomes	687	214	145	84	63	54	42	40	39	34
Not in control exomes	675	134	50	26	13	13	8	5	4	2
Not in either	467	89	34	18	9	8	4	4	3	1
Is loss-of-function (nonsense/frameshift indel)	25	1	1	1	0	0	0	0	0	0

Table 3

Analysis of exome variants using genomic evolutionary rate profiling

The number of genes with at least a single novel variant with an rejected substitution (RS) score10 > 0 in at least *x* individuals is given. A gene rank is assigned based on the average GERP score9 over all observed novel variants with RS score > 0 in all affected individuals.

c. GERP Score analysis (at least *x* of 10)	1	2	3	4	5	6	7	8	9	10
Variant RS score > 0	7,176	2,360	754	269	106	39	20	11	3	1
MLL2 Rank	3,732	1,232	399	136	47	14	6	3	NA	NA