OTX2 Duplication Is Implicated in Hemifacial Microsomia

Dina Zielinski1, Barak Markus1, Mona Sheikh1, Melissa Gymrek1,2,3,4, Clement Chu5, Marta Zaks6, Balaji Srinivasan5, Jodi D. Hoffman7, Dror Aizenbud6,* Yaniv Erlich1,*

1 Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, United States of America, 2 Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts, United States of America, 3 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, 4 Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 5 Counsyl, South San Francisco, California, United States of America, 6 Rambam Health Care Campus, Haifa, Israel, 7 Division of Genetics, Tufts Medical Center, Boston, Massachusetts, United States of America

Abstract

Hemifacial microsomia (HFM) is the second most common facial anomaly after cleft lip and palate. The phenotype is highly variable and most cases are sporadic. We investigated the disorder in a large pedigree with five affected individuals spanning eight meioses. Whole-exome sequencing results indicated the absence of a pathogenic coding point mutation. A genome-wide survey of segmental variations identified a 1.3 Mb duplication of chromosome 14q22.3 in all affected individuals that was absent in more than 1000 chromosomes of ethnically matched controls. The duplication was absent in seven additional sporadic HFM cases, which is consistent with the known heterogeneity of the disorder. To find the critical gene in the duplicated region, we analyzed signatures of human craniofacial disease networks, mouse expression data, and predictions of dosage sensitivity. All of these approaches implicated OTX2 as the most likely causal gene. Moreover, OTX2 is a known oncogenic driver in medulloblastoma, a condition that was diagnosed in the proband during the course of the study. Our findings suggest a role for OTX2 dosage sensitivity in human craniofacial development and raise the possibility of a shared etiology between a subtype of hemifacial microsomia and medulloblastoma.

Introduction

Hemifacial microsomia (HFM; also termed oculoauriculovertebral spectrum or Goldenhar syndrome, OMIM: 164210) is a highly heterogeneous condition with an estimated rate of 1 in 5,600 to 20,000 births [1]. The hallmarks of this disorder are marked facial asymmetry due to maxillary and mandibular hypoplasia and ear malformations such as preauricular skin tags, microtia, anotia, and conductive hearing loss. Some cases also present epibulbar dermoids and coloboma of the upper eyelid, cleft lip and palate, as well as cardiac, renal, and vertebral defects. To a lesser extent, the disorder also involves neurological anomalies and developmental delays or mental retardation [1–3]. The characteristic facial anomalies of HFM cases are attributed to disruptions in the first and second pharyngeal arches during days 30–45 of gestation in humans [1]. These arches contribute to the development of muscles of mastication, the maxilla, the mandible, middle ear bones, muscles of facial expression, and the stapedial artery. Animal models suggest embryonic hemorrhage or a deficiency in neural crest cell migration as the pathogenesis, which can disrupt normal development of pharyngeal arch derived structures [4].

The HFM spectrum reflects a complex pathogenesis that presumably includes both extrinsic and genetic risk factors [2]. Several epidemiological surveys suggest a role for environmental factors that affect the vascular system, including use of vasoactive agents, hypoxia, exposure to teratogens, and gestational diabetes [5]. While most HFM cases are sporadic, approximately 2–10% of cases are familial and occur in more than one generation, presumably includes both extrinsic and genetic risk factors [6,7]. Careful examination of seemingly unaffected relatives of a large number of probands revealed familial aggregation of mild craniofacial malformations and preauricular skin tags [8]. These mild features are relatively rare in the general population but do not meet the clinical criteria for HFM, leading to a decreased perception of family history. Segregation analysis of 74 families strongly favored an autosomal dominant mode of inheritance with incomplete penetrance over recessive or polygenic transmission [9]. These results suggest that genetics plays a broad etiological role in the manifestation of the disorder.

Genetic investigations of HFM cases have not yet clearly defined the critical genes involved in this disorder. Several studies have reported facial asymmetry and mandibular hypoplasia in cases with gross chromosomal aberrations and trisomies [10–15].
However, these patients exhibited multi-organ pathologies atypical of most HFM cases, suggesting distinct syndromes. Genome-wide linkage analysis of 3 HFM pedigrees revealed potential linkage to 1q43, 11q12–13 [16], and 15q26.2-q26.3 [17] but candidate gene sequencing in these studies failed to find a pathogenic variation. Rooryck et al. [18] performed array CGH on a cohort of 86 HFM patients, most without family history of the disorder. They found 12 copy number variants (CNVs) ranging from 2.7 kb to 2.3 Mb (median: 153 Kb). However, none of these CNVs were recurrent and 9 out of the 10 autosomal CNVs were also present in unaffected individuals. The authors concluded that it is difficult to interpret to what extent these CNVs contribute to the disorder. To date, the field has yet to identify a strong etiological gene that is responsible for the pathogenesis of the disorder.

We conducted a systematic analysis to identify an etiological variant of HFM. To increase the power of the investigation, we focused on a large family with multiple affected individuals. To the best of our knowledge, this family is the largest HFM kinship to date that is described in the literature. We considered both exonic mutations and copy number variations to further increase the probability of identifying the etiological locus while excluding bystander variations [19]. This process revealed a segmental duplication of 8 genes that segregates with the disorder. An unbiased HFM disease network analysis and expression profiling implicate OTX2 as the pathogenic gene in the CNV.

**Results**

**Clinical presentation**

We identified a five generation Ashkenazi kinship that displays variable HFM anomalies in five individuals separated by a total of eight meiosis events (Figure 1, Table 1). In all cases, the family denied consanguinity and the disorder appears to follow an autosomal dominant segregation pattern with incomplete penetrance and variable expressivity.

The proband, subject V.3, was presented to the Craniofacial Department of the Rambam Medical Center in Israel at the age of three. She was born after normal pregnancy (42 weeks) and caesarian delivery. Clinical examination revealed mandibular hypoplasia and facial asymmetry, cleft #7 according to Tessier’s craniofacial classification system, preauricular skin tags, and grade II microtia, all on the right side. Deafness in the right ear was diagnosed at the age of 2 months. She is of normal intelligence and no other abnormalities were noted at the time (Figure 1, Table 1). The proband underwent a combined surgical orthodontic manipulation using the distraction osteogenesis technique to elongate the right mandibular ramus. During the course of this study, at age seven, she was diagnosed with a medulloblastoma in the fourth ventricle. The tumor was completely resected, after which the child received craniospinal radiotherapy and chemotherapy [see a case study on her cancer treatment [20]].

The proband’s mother (IV.3), grandmother (III.1) and cousin (V.2) were also examined at the Craniofacial Department of the Rambam Medical Center. All individuals exhibited milder facial asymmetry with unilateral clefts and preauricular skin tags without ear involvement. Examination of the proband’s uncle (IV.2) did not reveal any facial anomalies, indicating incomplete penetrance of the disorder.

The proband’s first cousin twice removed (III.3) was identified at a later stage of the study. He presented mild facial asymmetry on his left side without auricle involvement and reported that his grandmother (I.1) displayed similar features.

Analysis of exonic variants showed no evidence of causal point mutation

We performed whole exome sequencing of individuals III.1, V.2, and V.3. The average autosomal coverage of the targeted regions in the three samples was 95×-105× reads per base pair. More than 96% of each exome was covered by at least one read (Figure S1, Table S2). Exome sequencing revealed 22,252, 22,746, and 23,175 exonic variants in III.1, V.2, and V.3 respectively. We observed transition/transversion ratios of 2.89–3.00 and homozygous to heterozygous mutation ratios of 0.56–0.58. In parallel, we also conducted genome-wide genotyping of...
these three samples using the Affymetrix SNP Array 6.0. Comparing shared variations between the two platforms showed concordance rates of more than 98% for non-reference loci (Table S1). All of these technical indicators are consistent with the results of previous studies [21–23], supporting the quality of the exome sequencing data.

We passed the exonic variations through a series of filters to find mutations that fit the rare familial pathology (Table 2). First, we excluded synonymous variants. Second, we excluded variations that appear at a frequency greater than 0.1% in large-scale sequencing projects such as the Exome Sequencing Project, 1000 Genomes, and ClinSeq, as documented in dbSNP. In addition, we excluded variations that appeared at least twice in the exome sequencing data of 21 healthy Ashkenazi Jews (provided by Noam Shomron, Tel Aviv University). In the Appendix, we show that these frequency cutoffs are very conservative. Third, we focused only on variants that reside in regions that are identical by descent (IBD) in all individuals. Variants that reside in these haplotypes were transmitted from III.1 to V.2 and V.3. Shared variants outside these regions are from ancient coalescent events and reflect inheritance patterns that do not segregate with the phenotype.

Using genome-wide genotype data, we identified 33 autosomal inheritance patterns that do not segregate with the phenotype. Similar to previous studies, we identified three segments that appear in all three individuals (Figure S2; Table S4). Importantly, these results support the validity of the IBD filtration technique and provide additional evidence supporting the absence of an etiological point mutation in the exome.

To further validate our findings, we performed Sanger sequencing of 37 variants that were identified in the exome sequencing results but excluded after the final IBD filtration step. Four of these variants were located in genes with biological activities that could relate to the disorder (DAB2, IQSEC1, KIAA1456, and ADAM28), such as vascularization, angiogenesis, imprinting, and neurogenesis [24–27]. However, Sanger sequencing of all 37 variations, including these four genes, showed that individual III.3 does not carry the variant, as expected from the IBD analysis (Figure S2; Table S4).

### Copy Number Variation Analysis Identified a Familial Duplication of 14q22.3

Given the absence of point mutations, we turned to copy number analysis using the genotype data from the genome-wide SNP array. Our analysis revealed a 1.3 Mb duplication of 14q22.3 (chr14:57,141,867–58,495,517) in all four individuals that segregated along all 8 meioses (Figure 2a). In general, CNVs of this length are rare and typically deleterious [28]. No other detected CNVs (>10 kb) were found to segregate with the disorder. To increase the sensitivity, we repeated the CNV analysis and inspected only CNVs that are shared in individuals III.1, V.2, and V.3. We excluded individual III.3 from this analysis because the array genotyping was performed separately and showed greater systematic noise. This process revealed seven CNV segments (>10 Kb) in addition to the 14q22.3 duplication. However, all but one were also found in healthy Ashkenazi controls from genome-wide genotyping array data [29]. The one segment that was not present in the Ashkenazi controls was a ~37 kb duplication of a non-coding region (chr3:187,279,170–187,316,070) that overlapped a known duplication found in healthy Asian controls in the Database of Genomic Variants (DGV: nssv1548729). Moreover, we did not see any evidence of this duplication in the array data for III.3. Thus, we concluded that the duplication of 14q22.3 is the only likely CNV that segregates with the disorder.

### Table 1. Clinical features of family members displaying HFM anomalies.

<table>
<thead>
<tr>
<th>Clinical feature</th>
<th>III.1</th>
<th>III.3</th>
<th>IV.3</th>
<th>V.2</th>
<th>V.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial cleft</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Facial asymmetry</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Anotia/microtia</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Preauricular tags</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Mandibular, maxillary hypoplasia</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Retrognathia</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Epibulbar dermoids</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Cardiac anomalies</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Renal anomalies</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Vertebral anomalies</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Medulloblatoma</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

To confirm the expected rarity of this duplication, we evaluated its frequency in the general Ashkenazi population. Analysis of the genome-wide genotyping array data from 942 healthy Ashkenazi chromosomes [29] returned two copies for this
region. In addition, no duplications were found in this region in CNV analysis of deep whole genome sequencing data from 204 chromosomes of Ashkenazi controls sequenced by Complete Genomics that are part of The Ashkenazi Genome Consortium (TAGC) and 1842 chromosomes from phase I of the 1000 Genomes Project [30]. These population-specific results support a familial variant that segregates with the disorder.

To validate our results, we performed qPCR analysis of the duplicated region using Taqman assays (Figure 2b). Three probes targeting genes in the duplication (OTX2OS1, EXOC5, and NAA30) were confirmed as CN = 3 (copy number) in individuals IV.2, IV.3, and III.3. We also observed duplication of OTX2OS1 and NAA30 in V.3 and of NAA30 in III.1, confirming segregation of this CNV along all informative meioses of the family. Assays targeting OTX2OS1, EXOC5, and NAA30 returned CN = 2 in all HapMap controls and OTX2OS1 and NAA30 were both CN = 2 in 45 Ashkenazi control samples (Figure S4). To validate the boundaries of the CNV, we also targeted KTN1 and PSMA3, upstream and downstream of the predicted CNV. Both probes returned CN = 2 in affected family members and HapMap controls (Figure 2b).

In order to evaluate the presence of the duplication in additional HFM cases in Israel, the Craniofacial Department of Rambam Medical Center collected DNA from 7 families that consisted of one affected offspring and unaffected parents. Interrogation of 2 genes in the duplicated region (NAA30 and OTX2OS1) by qPCR did not reveal any copy number changes in the seven additional HFM cases (Figure S3). These findings suggest a distinct genetic etiology of the disorder in our family and are consistent with previous studies that described genetic heterogeneity [10]. However, a literature search revealed that a spectrum of genetic lesions in the 14q22 region have been associated with various developmental anomalies. Ou et al. [31] reported a complex event of a translocation involving OTX2 as the gene with the closest molecular signature to other facial malformations. We and others have successfully identified etiological CNV using this guilt-by-association approach in previous studies of rare human disorders [34–36]. The basis of this technique is that similar phenotypes are caused by genes that reside in close biological modules such as the same pathway, co-expression cluster, and shared regulatory control (Goh et al 2007). To identify a set of disorders similar to HFM in an unbiased manner, we used MimMiner, which ranks clinical conditions in OMIM based on phenotypic resemblance [37]. The top three phenotypes with similar features to HFM were CHARGE syndrome (OMIM: 214800), VACTERL association (OMIM: 314390), and Townes-Brocks syndrome (OMIM: 107480). In fact, HFM and TBS are both characterized by first and second arch defects, including ear, jaw, and kidney malformations [38]. Interestingly, a previous study also cited the commonalities between HFM, CHARGE, and VACTERL [39], adding additional support to the MimMiner prediction. We then compared the biological signatures of all coding genes in the duplicated region to CHD7, ZIC3, and SALL1, the corresponding genes of the three syndromes. To increase the robustness of our analysis, we tested these similarities using two gene prioritization tools: Endeavour [40] and ToppGene [41]. These algorithms utilize different biological datasets and employ distinct prioritization procedures. These two algorithms independently ranked OTX2 as the gene with the closest molecular signature to other facial anomalies (Figure 3a).

Candidate Gene Prioritization in the Duplicated Segment

We sought to predict the etiological gene that contributes most to the phenotype in an unbiased manner among the eight genes (OTX2, OTX2OS1, EXOC5, AP5M1, NAA30, C14orf105, SLC35F4, and C14orf37 [partial]) that reside in the duplicated region.

First, we prioritized the genes in the duplicated region based on the similarity of their molecular signatures to known etiological genes of other facial malformations. We and others have successfully identified etiological CNV using this guilt-by-association approach in previous studies of rare human disorders [34–36]. The basis of this technique is that similar phenotypes are caused by genes that reside in close biological modules such as the same pathway, co-expression cluster, and shared regulatory control (Goh et al 2007). To identify a set of disorders similar to HFM in an unbiased manner, we used MimMiner, which ranks clinical conditions in OMIM based on phenotypic resemblance [37]. The top three phenotypes with similar features to HFM were CHARGE syndrome (OMIM: 214800), VACTERL association (OMIM: 314390), and Townes-Brocks syndrome (OMIM: 107480). In fact, HFM and TBS are both characterized by first and second arch defects, including ear, jaw, and kidney malformations [38]. Interestingly, a previous study also cited the commonalities between HFM, CHARGE, and VACTERL [39], adding additional support to the MimMiner prediction. We then compared the biological signatures of all coding genes in the duplicated region to CHD7, ZIC3, and SALL1, the corresponding genes of the three syndromes. To increase the robustness of our analysis, we tested these similarities using two gene prioritization tools: Endeavour [40] and ToppGene [41]. These algorithms utilize different biological datasets and employ distinct prioritization procedures. These two algorithms independently ranked OTX2 as the gene with the closest molecular signature to other facial anomalies (Figure 3a).

Disease-associated genes can be more uniquely expressed in affected tissues than in those that are unaffected [42,43]. Thus, analysis of expression patterns could help to stratify the contribution of the genes in the region to the pathology. We used publicly available expression array profiles of mouse embryonic tissue to compare the expression of the duplicated genes in affected versus unaffected tissues. Specifically, we analyzed expression levels in the pharyngeal arches at embryonic day 10.5 and in the entire head at E13.5. These developmental stages approximately overlap with the suggested critical periods for the HFM developmental perturbation in humans [1]. We contrasted these expression levels with the expression profiles of liver, heart, and lung (E13.5) and heart and urogenital epithelium (E10.5) since these tissues are rarely implicated in HFM. At E10.5, the arrays contained data for Otx2, Ap5m1, Naa30, and Scl35f4. At E13.5,
OTX2 Duplication in Hemifacial Microsomia

A

B

Duplicated region

KTN1  OTX2OS1  EXOC5  NAA30  PSMA3

Calculated Copy Number

Sample
the arrays contained data for Otx2, Otx2os1, Exoc5, Ap5m1, Naa30, and Slc35f4. The expression profiles showed that Otx2 tends to be more highly expressed in the affected tissues than other duplicated genes at E10.5 and E13.5 compared to any of the unaffected tissues (Figure 3b).

Finally, we also evaluated the general sensitivity of the genes in the region to duplication. Huang et al. [44] developed a gene-level classifier that compares evolutionary, functional, gene-structure, and interaction patterns between haplosufficient and haploinsufficient genes. Interestingly, they found higher expression and tissue specificity of haploinsufficient genes early in development. Although the classifier predicts the probability of haploinsufficiency, it is also useful for detecting genes with increased dosage sensitivity [M. Hurles, personal communication, August 2013]. Three of the duplicated genes were included in their classifier: OTX2 had the highest sensitivity score (0.9) followed by NAA30 (0.474) and SLC35F4 (0.418) (Figure 3c). To summarize, all of our in silico analysis techniques suggested that duplicated OTX2 is the most likely pathological gene in our HFM cases.

Discussion

We conducted a systematic study of familial HFM that implicates OTX2 dosage sensitivity in the disorder. OTX2 encodes a transcription factor that plays a critical role in craniofacial development and anterior brain morphogenesis. OTX2 homologs in model organisms are expressed in a complex spatial, temporal, and gradient-specific manner that is required for correct antero-posterior patterning and craniofacial development [45]. These expression patterns are influenced by tissue-specific feedback from other genes and by auto-regulation, which may introduce compensatory mechanisms that depend on the activity of other genes. Perturbations in relative expression levels could explain the tissue-specific pathologies as well as the highly variable phenotype and incomplete penetrance of the disorder in our cases. Loss-of-function studies in mice showed that null embryos fail to develop the anterior head and die during embryogenesis while Otx2+/2 mice exhibit a range of severe craniofacial anomalies, including micrognathia, agnathia, anophthalmia, and head narrowing with no involvement of the auricle [46]. The severity of the phenotype depends on the genetic background [47].
consistent with the wide spectrum of phenotypes associated with loss of function in humans. Temporal loss of one copy of Otx2 during mouse embryogenesis up to E12.5 results in haploinsufficiency that leads to significantly low survival rates and abnormal head development, including reduction or absence of the forebrain, eyes, and jaw [43]. OTX2 hemizygous deletions and non-synonymous point mutations have been reported in patients with severe ocular malformations, developmental delays, and hypopituitarism, symptoms that are not seen in our pedigree [48–50]. OTX2 loss-of-function mutations are associated with a wide phenotypic spectrum and the absence of such anomalies in our subjects suggests a different set of pathologies resulting from OTX2 duplications.

The OTX2 germline duplication in our subjects suggests a potential link to the medulloblastoma of the proband. OTX2 is a known oncogenic driver of medulloblastoma [51]. Focal duplications and overexpression of this gene are prevalent in subclasses C and D of medulloblastoma [52]. Analysis of her tumor revealed an additional loss of heterozygosity on chromosome 17q [20] that is exclusively associated with subclasses C and D [52]. The potential biological link between OTX2 duplications in hemifacial microsomia and medulloblastoma raises the possibility of their comorbidity. While confirming this hypothesis will require the analysis of a large number of cases, we suggest clinicians be aware of the possibility of increased risk for medulloblastoma in HFM cases with OTX2 duplications.

Our study adds to the existing literature in multiple ways. First, we investigated the largest HFM pedigree to date, increasing the confidence of our genetic analysis. Second, it is the first HFM study to combine whole exome sequencing analysis with the scanning of copy number variants. This approach increases the likelihood that the duplicated region is indeed the etiological site. Third, we present data from more than 1000 chromosomes of unaffected controls, which strongly diminishes the likelihood that the duplication is a polymorphism that segregates in the population. Fourth, we report an unbiased search using different systems biology approaches to find the most likely pathological gene in the region. These analyses implicated OTX2 as the most likely causal gene. Fifth, our findings suggest a potential shared etiology for HFM and medulloblastoma.

Determining the causative gene for HFM can promote stratification of cases based on the molecular pathology, guide clinical care, offer reproductive alternatives to families that carry an OTX2 duplication, and facilitate definitive diagnosis, which is currently inadequate for HFM. Importantly, implicating OTX2 in this disorder can improve understanding of the basic molecular processes that underlie normal and pathological craniofacial development.

Materials and Methods

Human Subject Research

This study was approved by the Helsinki Committee at the Rambam Medical Center (Haifa, Israel), the Israeli Ministry of Health, and MIT’s Committee on the Use of Humans as Experimental Subjects. Written consent for sample collection and use in the study was approved by both committees and obtained from all participants. Informed consent was obtained from guardians on behalf of minors enrolled in the study. Subjects were informed of the terms of the PLOS open-access license and use in the study was approved by both committees and obtained from all participants. Informed consent was obtained from guardians on behalf of minors enrolled in the study. Subjects were informed of the terms of the PLOS open-access license and subject IV.3 gave written informed consent to publish photographs. Additionally, photos of the proband (V.3) were previously published [20]. MIT’s IRB (COUHES) approved this consent procedure.

Coordinate System

All alignment and genomic coordinates in this manuscript are reported according to hg19. All coverage values are reported after removing PCR duplicates.

DNA Collection

All DNA was derived from whole blood using standard procedures.

Exome Sequencing

Paired-end library preparation and exome enrichment were done following a streamlined protocol written by Blumenstiel et al. [53], using Agilent’s SureSelect All Exon V.2 kit, which covers 98.2% of exons and splice sites, according to the Consensus CDS (CCDS) database [54]. Sequencing was performed at Counsyl (South San Francisco, USA) on a single flow cell on the Illumina HiSeq2000 with 100 bp paired end reads (V.2 and V.3 on 3 lanes and III.1 on 2 lanes).

To increase the accuracy of our analysis, we processed the sequencing data with two distinct pipelines. First, we iteratively aligned the sequence reads with Bowtie [55] and with BWA [56]. Multi-mappers were excluded. Reads that failed to align were repeatedly trimmed by 10 bp down to a minimum of 36 bp and were processed in an additional round of alignment. The BAM files of all unique mappers from the different alignment rounds were merged and PCR duplicates were removed using SAMtools [57]. Variant calling of Bowtie-aligned reads was done using VarScan v2.3.8 [58] with mpileup2cns and the following options: —min-coverage 5 —min-freq-for-hom 0.9 —p-value 0.97 —strand-filter 1. After alignment using BWA, variant calling was done using the Genome Analysis Toolkit (GATK) [59], following the recommended workflow and filtering of low quality variant calls. In addition, we used lobSTR 1.0.6 [60] to examine short tandem repeat variations in the exomes of III.1, V.2, and V.3. We filtered for STRs genotyped in all three samples with at least 5× coverage in each, that fell within regions shared by all samples with IBD = 1, and falling within annotated Refseq genes. Six loci were called as non-reference in all three samples. For each locus, the non-reference allele was found in at least one healthy control from a panel of more than 30 healthy controls, mainly of European descent.

Validation by Sanger Sequencing

We used Primer3 [61] to design primers flanking candidate variants (+/-100 bp upstream and downstream). We excluded primers that generated more than one in silico PCR product on the UCSC Genome Browser [62]. Sanger sequencing was done on an ABI 3730 DNA Analyzer.

Genome-Wide Human SNP Array 6.0

Genomic DNA was extracted from peripheral blood leukocytes using standard methods. We performed genotyping of subjects III.1, III.3, V.2, and V.3 using the Affymetrix SNP 6.0 Array. We analyzed the 4 cases together with 471 unrelated Ashkenazi controls [29] (NCBI GEO GSE23636) using the Affymetrix genotyping console (v 4.1.3) and Birdsuite [63] for genotype calling.

Investigating exonic variations

Annotation of exonic variations was done using SeattleSeq Annotation 137 [23] and minor allele frequencies in dbSNP were
IBD Calculations

We used the Affymetrix genotyping console (v 4.1.3) for genotype calling of our 4 subjects together with 50 randomly selected individuals from the Ashkenazi controls (Bray 2010). Initial data analysis and selection of SNPs were carried out using PLINK [66]. We selected subsets of SNPs with MAF > 0.1 that are in approximate linkage equilibrium. This was carried out using the pairwise correlation method for LD pruning implemented in PLINK. We used the following parameters: window size = 50, pairwise correlation method for LD pruning implemented in PLINK. We selected subsets of SNPs with MAF > 0.1 that are in approximate linkage equilibrium. This was carried out using the pairwise correlation method for LD pruning implemented in PLINK. We used the following parameters: window size = 50, step = 5, r^2 threshold = 0.35. The pruned data contained 129209 SNPs.

We used the pruned data as input to MERLIN [67] for pairwise IBD inference, with genetic map positions of 1 Mbp = 1 cM. Candidate IBD regions were selected based on pair-wise IBD probabilities. We marked all regions for which IBD probabilities for sharing an allele for all pairs of cases in the data were inferred to be higher than 0.5. We then extended the IBD region to include the tips of the chromosomes for cases when IBD = 1 was detected in the first or the last SNP on the chromosome.

Taqman CNV Assays

We purchased custom Taqman probes to interrogate the CNV and flanking regions (probe start locations in NCBI build 37: chr14:20811565, chr14:56099993, chr14:57267695, chr14:572709, 23, chr14:57272149, chr14:57277101, chr14:57328402, chr14:57476529, chr14:57597148, chr14:57700715, chr14:57868427, chr14:58725357, chr17:44203062). Reactions were carried out in 10 ul, with 10 ng genomic DNA and 10 ng reference DNA (RnaseP), in 4 replicates. Copy number was determined using the delta delta Ct method and CopyCaller v2.0 with HapMap samples NA06991 and NA11832 as calibrators. The OTX2 probes that were purchased from ABI failed to work despite repeated attempts. They produced non-Mendelian inheritance patterns for trios and reported deletions of the region in normal healthy controls. We therefore excluded these probes from the analysis.

Prioritization using Biological Signatures


Expression analysis of genes in the region

Expression profiles were derived from the following experiments in GEO [68]: Pharyngeal arches E10.5: experiment GDS3803 with subjects GSM448013, GSM448014, GSM448015, GSM448016, and GSM448017. Urogenital epithelium E10.5: experiment GDS3173 with subjects GSM257875, GSM257932, and GSM257933. Heart E10.5: experiment GDS6267 with subjects GSM25150, GSM25151, GSM25152. Head E13.5: experiment GDS2 874 with subjects GSM212558, GSM212560, GSM212562, and GSM212564. Liver E13.5: experiment GDS2693 with subjects GSM177034, GSM177035, and GSM177036. Lung E13.5: experiment GSM290632 with subject GSE11539.

All experiments were carried out on the Affymetrix Mouse Expression Array 430. The pharyngeal arches experiment reported results only from the A array and all the others reported both the A and B arrays. Therefore, in all E10.5 comparisons, we restricted the analysis only to genes that are on the A array.

Based on experimental details in GEO or associated publications, the genetic background of all mice was concluded to be C57BL/6, with the exception of GDS3173 (E10.5 urogenital epithelium), the background of which was not documented.

We downloaded the full soft file of each experiment from GEO, extracted the data from the relevant subjects, and normalized the expression data to range from zero to one for each subject. Experiments with multiple sets were averaged inside the same condition. Then, genes with more than one probe were averaged inside the same condition. Finally, we divided the expression of each gene in the affected tissue (pharyngeal arches and head) by expression in the control tissues (liver, lung, heart, and urogenital epithelium) and ranked the expression levels.

Dosage sensitivity analysis

Data was taken from Dataset_S1.txt of Huang et al. [44].

Access to the sequencing and array data

The sequencing and genotyping data from this study are available on dbGAP (http://www.ncbi.nlm.nih.gov/gap) and http://erlichlab.wi.mit.edu/hfm (please refer to the website for Terms and Conditions).

Appendix

Our working hypothesis was that any point mutation that causes HFM will have a minor allele frequency (MAF) of less than 0.1% in large sequencing projects. We based our hypothesis on the fact that HFM is estimated to occur at a frequency of 1:5,000–1:20,000 births in the general population. Segregation analysis by Kaye et al. (1992) predicted that the sum of minor allele frequencies of all HFM causative genes is 1:3000 (after taking into account penetrance levels). The MAF of a single etiological variant is even smaller, since previous linkage analysis identified at least three non-overlapping segments.

Moreover, the affected family is of Ashkenazi heritage. With the limited gene flow between the Ashkenazi population and other European populations, the causal mutation in our family is expected to be at even smaller frequencies in these large sequencing projects due to the low sampling rates of Ashkenazi Jews. To confirm this assumption, we compared the MAFs of more than 50 recessive mutations associated with Ashkenazi genetic disorders to the Exome Sequencing Project where we obtained most of the control chromosomes used in our analysis. These mutations are found at frequencies of 1/25 to 1/70 in the Ashkenazi population. Even if the causal mutation is found at a very unlikely rate of 1% in Ashkenazim, we expect it to be < 0.05% in ESP. Thus, a 0.1% threshold is highly unlikely to miss the causative mutation.

Similarly, we excluded variants that were seen at least twice in 42 unaffected Ashkenazi chromosomes. The probability to
sec a mutation with a true MAF of 0.1% in two individuals from this cohort is \(<1 \times 10^{-5}\). Therefore, there is a very small risk of excluding the causative mutation using this MAF cutoff.

Supporting Information

Figure S1 Distributions of exome sequencing coverage for the three datasets. (EPS)

Figure S2 Sanger traces of the four genes with biological activity that could be associated with HFM. (EPS)

Figure S3 qPCR results of the probands in the seven families. Both tested probes showed copy number 2 of the critical region. NA06991 is a HapMap control. (EPS)

Figure S4 qPCR results of 2 genes in the duplicated region in 45 Ashkenazi control samples. (EPS)

Table S1 Array and sequencing concordance. The probabilities to observe genetic variants in the sequencing data conditioned on the array data status and collapsed in all three individuals. (DOXC)

Table S2 Exome sequencing quality metrics. A summary of the quality control indicators from the three exome sequencing datasets. (DOCX)

Table S3 Candidate variants. We retained only variations shared IBD in individuals III.1, V.2, and V.3. (DOCX)

Table S4 Sanger sequencing results of the four genes with biological activity that could be attributed to HFM. (DOCX)

Acknowledgments

We gratefully acknowledge the study participants and thank Noam Shomron, Isik Pe’er, Brian Handsaker, and Sara Selig for providing valuable information about the duplicated region in control samples. We also thank Helen Skakel, Tatjana Pyykkö, and Natalia Koutseva for translations and Whitehead Institute’s Genome Technology Core for assistance in producing the array datasets. YE is an Audria and Paul Heafy Family Fellow and holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was also supported by generous gifts from Cathy and Jim Stone and Ron Casty.

Author Contributions

Conceived and designed the experiments: DZ YE. Performed the experiments: DZ MZ DA. Analyzed the data: DZ BM MS MG YE JH. Contributed reagents/materials/analysis tools: CC BS. Wrote the paper: DZ YE.

References


