Integrated Genomic Analysis Suggests *MLL3* is a Novel Candidate Susceptibility Gene for Familial Nasopharyngeal Carcinoma

Mark M Sasaki¹, Andrew D Skol¹, Riyue Bao², Lindsay V Rhodes¹, Rachelle Chambers³, Everett E Vokes⁴,⁵, Ezra EW Cohen⁶, Kenan Onel¹,⁵

¹Department of Pediatrics, ²Center for Research Informatics, ⁴Department of Medicine, and ⁵The University of Chicago Comprehensive Cancer Center, The University of Chicago, Chicago, IL 60637
³Laura and Isaac Perlmutter Cancer Center, NYU Langone Medical Center, New York, NY 10016
⁶Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093

Running Title: *MLL3* and Familial Nasopharyngeal Carcinoma Risk (47 characters)

Keywords: Nasopharyngeal carcinoma; cancer genetics; cancer risk; whole exome sequencing; bioinformatics

Financial support: This work was supported by grants from the National Institutes of Health (HD0433871, CA129045 and CA40046 to K Onel); the American Cancer Society – Illinois Division (K Onel); the Cancer Research Foundation (K Onel); and The University of Chicago GREAT KIDS (Genomics for Risk Evaluation and Anticancer Therapy in Kids) Program (K Onel, AD Skol). The Center for Research Informatics is funded by the Biological Science Division and The Institute for Translational Medicine/CTSA (NIH UL1 RR024999) at The University of Chicago.

Correspondence should be directed to:
Kenan Onel, MD, PhD
Section of Hematology/Oncology
Department of Pediatrics
University of Chicago
KCBD 5140
900 East 57th St, MC 4060
Chicago, IL 60637
Phone: 773-702-4919
Fax: 773-834-1329
konel@uchicago.edu

All authors declare no potential conflicts of interest

Abstract: 223 words
Word count: 3085 words
1 Table
2 Figures
6 Supplementary Tables
ABSTRACT

Background: Little is known about genetic factors associated with nasopharyngeal carcinoma (NPC). To gain insight into NPC etiology, we performed whole exome sequencing on germline and tumor DNA from three closely related family members with NPC.

Methods: The family was ascertained through the Pediatric Familial Cancer Clinic at The University of Chicago. The diagnosis of NPC was confirmed pathologically for each individual. For each sample sequenced, 97.3% of the exome was covered at 5x, with an average depth of 44x. Candidate germline and somatic variants associated with NPC were identified and prioritized using a custom pipeline.

Results: We discovered 72 rare deleterious germline variants in 56 genes shared by all three individuals. Of these, only three are in previously identified NPC-associated genes, all of which are located within MLL3, a gene known to be somatically altered in NPC. One variant introduces an early stop codon in MLL3, which predicts complete loss-of-function. Tumor DNA analysis revealed somatic mutations and EBV integration events; none, however, were shared among all three individuals.

Conclusions: These data suggest that inherited mutations in MLL3 may have predisposed these three individuals from a single family to develop NPC, and may cooperate with individually-acquired somatic mutations or EBV integration events in NPC etiology.

Impact: Our finding is the first instance of a plausible candidate high penetrance inherited mutation predisposing to NPC.
INTRODUCTION

Nasopharyngeal carcinoma (NPC) is a rare malignancy arising from epithelial cells of the head and neck. Although worldwide incidence of NPC is under 1:100,000, rates vary by geography and ethnicity. In Southern Italy, Greece, Turkey, Northern Africa, and among Alaskan Eskimos, rates range from 15-20:100,000. Incidence peaks in Southeastern China and Southeast Asia at 25:100,000, whereas NPC is rare in the United States and Western Europe, with an incidence of only 0.5-2:100,000 (1, 2). NPC etiology is multi-factorial, and includes exposure to nitrosamines found in, for example, tobacco, salted fish, and cosmetics and pesticide manufacturing; exposure to formaldehyde; infection with Epstein-Barr Virus (EBV); and genetic susceptibilities (3-5).

Although little is known about the genetic contribution to NPC risk, that there is a genetic component to susceptibility was demonstrated in two separate studies, one performed in South Asian individuals and the other in European individuals from Greenland and Denmark. In both, it was found that individuals with a first-degree relative with NPC were at an 8.0-fold greater relative risk for developing NPC as compared to the general population (6). To identify genetic factors underlying NPC susceptibility, only a small number of studies have thus far been performed. In one candidate gene study, the association between NPC and polymorphic variation in base excision repair genes, the pathway required for repair of nitrosamine-induced DNA damage, was investigated. Variants in XRCC1 and hOGG1 were found to be associated with NPC; these findings, however, await replication (7). In another study, a genome-wide linkage scan of familial NPC in 54 affected individuals from 20 families led to the discovery of a susceptibility locus at chromosome 4p15.1-q12 (8). More recently,
four Genome Wide Association Studies (GWAS) of NPC have been performed and have led to the identification of 20 variants associated with NPC (9-12). The full list of associated variants is in Supplementary Table 1.

Because the genetic architecture of familial disease is vastly simplified relative to that of sporadic disease, studying the genetics of NPC in families with multiple affected individuals is an attractive strategy for discovering high penetrance susceptibility variants. Towards this end, we performed WES on germline DNA from three related individuals of Italian descent, two full siblings and a half nephew, all of whom developed NPC (Figure 1). Additionally, we performed WES on their tumor DNA to determine whether their shared predispositions are associated with the acquisition of shared somatic mutations. Finally, we scanned their germline and tumor exomes for sites of EBV integration to investigate the possibility that patterns of EBV insertion were common among all three individuals. This is the first study of the genetic etiology of NPC undertaken in individuals of European ancestry.
MATERIALS AND METHODS

Study subjects

The family investigated was ascertained by the Pediatric Familial Cancer Clinic at The University of Chicago. All study subjects provided written informed consent to participate in a study of NPC genetics that was approved by the local institutional review board. The pedigree is presented in Figure 1. To protect the anonymity of the study subjects, the family pedigree was altered in ways that did not affect the genetic analysis.

Exome capture and sequencing

Germline DNA for WES was obtained from whole blood. Tumor DNA was isolated from FFPE scrolls after evaluation by a pathologist (>80% tumor). At least 1 ug of DNA was used for whole exome capture using SureSelect Human All Exon V4 50 Mb kit (Agilent Technologies, Santa Clara, USA). Sequence reads were generated on an Illumina HiSeq2000 instrument (Illumina, San Diego, USA). An average of 63 million 2x100bp paired-end (PE) reads were generated for each sample.

Variant calling and quality control

The quality of raw reads was assessed by FastQC (13), followed by adapter clipping and 3' overlap mate merging. Processed reads were aligned to the human reference genome assembly (hg19) using three short-read aligners: BWA (14), Bowtie2 (15), and Novoalign (16). Exon coverage was calculated using BEDTools (17). Read duplicates were removed using the Picardtools MarkDuplicates program (18). The alignment was post-processed by GATK v1.6 (19) for InDel realignment and base
quality score calibration. For each alignment, GATK UnifiedGenotyper (19, 20), FreeBayes (21), Atlas2 (22), and SAMtools mpileup/bcftools (23) were used to detect variants. Variant calls passing the internal quality filters of each caller were then filtered to remove potential false positives based upon: 1) variant quality score <50; 2) read coverage ≤5; or 3) location within a single nucleotide variant (SNV) cluster in which >3 SNVs were called within a 10bp window. After combining results from the three aligners and four callers, variants called by at least two callers using the aligned sequence from at least two aligners were carried forward for annotation using ANNOVAR (24).

Population minor allele frequencies (MAF) were derived from The 1000 Genomes Project database (25) (phase 1, release v3, 20101123) and the Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP, version ESP6500-V2-SSA137, Seattle, WA; 06/2012 accessed) (26).

Each variant was annotated for pathogenicity using: SIFT (27); Polyphen-2 (28); MutationTaster (29); MutationAssessor (30); FATHMM (31); LR and LRT (32); and Radial SVM (24, 33). They were assessed for multispecies conservation using GERP++ (34), and PhyloP (35).

Germline mutations were defined as variants with a population allele frequency of <0.01 or which were unobserved in either The 1000 Genomes Project or The ESP databases.

**Prioritization of candidate rare germline variants associated with NPC**

To investigate rare variation, we required that variants passing our QC pipeline: 1) have a population MAF ≤0.01 in the European subset from either The ESP or The
1000 Genomes Project; 2) be either non-synonymous, a splice site modifier, an insertion or deletion creating a frameshift, or create a stop codon; 3) and be deleterious as predicted by one of the pathogenicity prediction algorithms. For each variant identified, we confirmed its presence in the tumor sample of each individual.

We compiled a list of NPC-associated genes from a thorough literature review, a previously published catalog of genes associated with NPC (36), and 4 NPC GWAS listed in the NHGRI resource (9, 11, 12, 37). Variants were prioritized as likely to be NPC-associated if they were found in genes either associated with NPC risk or somatically mutated in NPC (Supplementary Table 1).

Identification of somatic mutations

To identify somatic variants, we analyzed matched normal/tumor pairs for all three individuals using MuTect (38), Strelka (39), Virmid (40) and VarScan2 (41). All four programs detect somatic SNVs. Strelka and VarScan2 also detect somatic InDels. Variants passing the internal quality control of each caller were retrieved and filtered for high-confidence calls based upon: 1) variant quality score ≥20; 2) sequencing read depth ≥8 and; 3) allele fraction in the tumor sample of >0.20 and allele fraction in the germline sample of <0.05. We then combined somatic variants identified by any of the four calling algorithms for downstream analysis. Somatic mutations in each tumor were manually inspected using Integrative Genome Viewer (42) to confirm that the variant allele was not present in the matched normal sample.

Identification of EBV insertion sites
Using Novoalign, sequencing reads were mapped to the human reference genome assembly hg19, and to the type I, type I-HKNPC1 and type II EBV reference genomes (GenBank Accession IDs: NC_007605.1, JQ009376.1 and NC_009334.1). Read duplicates were removed, and alignments with mapping quality scores <20 were excluded. EBV insertion sites were identified using chimeric read pairs, which have one mate mapped to the human genome and the other mapped to at least one of the three EBV reference genomes. The insertion sites were approximated as the 3' end of the chimeric mate mapped to the human genome, and annotated for nearby genes.
RESULTS

Familial WES pipeline

We performed WES on germline and tumor DNA from three members of a single family of Italian ancestry, all of whom were affected with NPC. For all individuals, the tumors were EBV-positive, as confirmed by EBER in situ hybridization. Two individuals are full siblings and the third is their half nephew (Figure 1). Individual I-1 was diagnosed with renal clear cell carcinoma at age 45 and with NPC at age 49. He lived in Argentina, worked as a beautician for 30 years, and smoked regularly (half a pack for over 30 years). Individual I-2 is the sister of Individual I-1 and was diagnosed with NPC at age 29. She is 12 years younger than Individual I-1, worked in a nail salon for 7 years prior to her diagnosis, and did not smoke or chew tobacco. Individual II-1 is the maternal half nephew of Individuals I-1 and I-2, and was diagnosed with NPC at age 39. Over a period of 8 years, he worked in construction and then went on to be a line cook at a restaurant. He did not smoke or chew tobacco. He died from metastatic NPC at 41 years of age. Of note, two siblings who were first cousins of the NPC-affected sibling pair and in the same blood lineage as Individual II-1 also developed cancer, one a brain tumor at 18 months, and the other early-onset breast cancer (at age 44). No samples from these individuals were available for WES.

Following WES and quality control, for each sample sequenced, 97.3% of the exome was covered at 5x and 86.4% of the exome was covered at 20x, with an average coverage depth of 44x or greater across the exome (Supplementary Table 2). Variants of low quality score, coverage depth less than 5, and those that were not called by at least two genotype callers within the aligned sequence generated by at least two
aligners were removed. After filtering out non-exonic variants and those leading to synonymous amino acid changes, an average of 8,767 variants in each germline sample and 8,832 variants in each tumor sample were identified. The mutational spectrum of the variants for each sample (germline and tumor) for each individual is summarized in **Supplementary Table 3**.

### Identification of candidate familial NPC-predisposing germline mutations

To discover candidate NPC-predisposing mutations in this family, we first identified rare exonic germline variants (MAF ≤0.01) shared by all three individuals. We found an average of 780 rare or novel variants in each individual, of which 190 were shared by all three individuals. Variants were categorized as: non-frameshift insertion/deletions (n = 34), non-synonymous single nucleotide variants (SNVs, n = 113), frameshift insertion/deletions (n = 3), stop gain (n = 1), and unannotated (n = 39). Of these, we predicted 72 variants in 56 genes to be deleterious (**Supplementary Table 4**).

To prioritize among these 72 candidates, we filtered them against a list of 76 previously identified NPC-associated genes (**Supplementary Table 1**) (9, 11, 12, 36, 37). This reduced the list to only three variants in a single gene, *MLL3* (rs150073007; rs4024453; and rs10454320) (**Table 1**). All three deleterious variants are unique to this family and not found in any of the 9007 individuals sequenced as part of The 1000 Genomes or Exome Sequencing Projects. rs4024453 (c.2315c>t), results in a serine to leucine change that occurs in a serine rich domain, and therefore may affect protein-protein interactions. rs10454320 (c.946a>t), is a threonine to serine change at amino acid 316, upstream of any known functional domain. Most compellingly, rs150073007
results in the introduction of a premature stop codon at position 816, N-terminal to most functional domains within the gene product.

The location of these three variants, in particular the early stop-gain mutation, suggests that they completely abrogate MLL3 protein function, leading to the hypothesis that one or more of these mutations predisposed these three family members to develop NPC. Supporting this contention is the observation that somatic mutations in MLL3 have been previously reported in 4% of NPC cases, including a recurrent mutation introducing a premature stop codon at amino acid 728, near the site of the germline premature stop codon observed in this family (Figure 2) (36).

Somatic mutation analysis of familial NPC

We then investigated somatic mutations in the exome of the tumor DNA from each individual to determine whether a common “second hit” had occurred in all three individuals in MLL3 or any of the 56 genes containing a candidate deleterious germline mutation.

Individual I-1 had 90 somatic mutations of which 36 were predicted to be deleterious. Individual I-2 had 68 somatic mutations of which 24 were predicted to be deleterious. Individual II-1 had 110 somatic mutations of which 47 were predicted to be deleterious. No individual had acquired a somatic mutation in MLL3. Among the three individuals, six of the overall set of 56 genes with deleterious germline mutations had also acquired somatic mutations. Specifically, Individual I-1 acquired somatic mutations in MUC2 and MUC6; Individual I-2 acquired somatic mutations in MUC6; and Individual II-1 acquired somatic mutations in HRNR, KCNJ12, PABPC1, and PCMTD1. Notably,
mutations in these genes have not previously been implicated in NPC (36). Additionally, mutations in *MUC* genes are frequently observed as false-positives in next-generation sequencing studies and must be interpreted with caution (43). There were, however, two *de novo* somatic mutations in genes previously implicated in NPC. Individual I-1 acquired a somatic mutation in *NRAS* (Q61R) (44-48), a well-characterized mutation observed in numerous cancers, and Individual II-1 acquired a somatic mutation in *PIK3CG* (X87Y), a mutation not previously reported.

Thus, we did not find overlap in the spectrum of somatic mutations among the three individuals. Results are summarized in Supplementary Table 5.

**EBV integration analysis**

There is a strong association between NPC and EBV integration (3). To determine whether there are shared patterns of EBV integration among the three NPC-affected family members, we mapped EBV integration sites in the germline and tumor DNA of each individual. We found that the germline exome from the two siblings, Individuals I-1 and I-2, did not contain any EBV DNA. Individual I-1 had only a single somatic insertion event in his tumor, while Individual I-2 had nine somatic insertion events. In contrast, Individual II-1 had one EBV insertion event in his germline exome and 42 somatic insertions in his tumor. EBV integration events were not found in any of the 56 genes with deleterious germline mutations or in any gene previously associated with NPC in any individual. Results are summarized in Supplementary Table 6.
DISCUSSION

In this study, we employed a family-based WES strategy to discover germline variants predisposing to NPC. We hypothesized that the three affected individuals in this family were predisposed to NPC through the shared inheritance of a single or small number of highly penetrant mutations. We found 72 rare deleterious germline variants in 56 genes shared by all three family members, three of which are in known NPC-associated genes. All three are located within a single gene, *MLL3*, which is recurrently mutated somatically in NPC. While all germline *MLL3* mutations are predicted to attenuate *MLL3* function, one mutation, rs150073007, is a particularly compelling candidate as causative because it results in the introduction of a stop codon near the N-terminus of the protein. The observation that none of these three variants is reported in large population databases such as The 1000 Genomes Project and The ESP leads us to speculate that they originated in this family. Based on our analysis, we propose *MLL3* as the candidate NPC-predisposition gene in this family.

*MLL3* is a histone lysine methyltransferase that functions in transcriptional co-activation of nuclear receptor targets. It is mutated not only in NPC, but in a variety of other cancers as well (49-53). As a component of the ASCOM complex, *MLL3* is a co-activator of *p53*, and deletion of its catalytic domain results in the development of kidney and ureter epithelial tumors in mice (54). Additionally, *MLL3* functions to regulate enhancer activity. Since enhancers play an important role in the tissue-specific expression of genes, mutations in *MLL3* may affect tumorigenesis in a tissue-dependent manner (53). Functional studies will be necessary to determine the consequences of the mutations identified here on *MLL3* activity.
We did not find any combination of germline variants, somatic mutations and/or EBV integration events common among all three individuals that altered the function of any gene other than \textit{MLL3}. The fact that we did not observe any recurrent acquired genetic changes in the tumor DNA of the three individuals suggests that either: 1) shared acquired mutations may have occurred outside of the exome; 2) the mutations we did identify are unique to each individual but converge on and deregulate common pathways; or 3) other factors such as differences in exposures, variation in regulatory molecules such as miRNAs or lncRNAs, or epigenetic changes may also have contributed to the excess of NPC in this family. The complexity of the mutational landscape and lack of concordant acquired changes among all three individuals underscores the difficulties inherent in the genomic analysis of even highly penetrant families.

While NPC etiology depends on multiple factors such as environmental exposures, geography, diet, and EBV, the co-occurrence of the disease in three closely related individuals from a single family is strongly suggestive of a common genetic etiology. Recently, a germline mutation in \textit{MLL3} that introduced a premature stop codon at amino acid 827 was reported in a Chinese family with colorectal cancer and acute myeloid leukemia (55). This mutation is located very close to the premature stop codon at amino acid 816 identified in all three NPC-affected family members described in this study. Importantly, in addition to the three NPC-affected individuals we sequenced, there are two other closely related cancer-affected individuals following the same blood lineage in this family; one is an individual with early onset breast cancer (diagnosed at age 44), and the other is a baby with a brain tumor who died at age 18.
months. Unfortunately, hospital records and samples from these two individuals are not available. Taken together, the finding of familial mutations predicted to abolish MLL3 function in two unrelated families with multiple cancer-affected members leads us to the intriguing hypothesis that inactivating mutations of MLL3 may be associated with a highly penetrant and previously unsuspected cancer-predisposition syndrome. In other studies, the familial aggregation of other cancers with NPC remains controversial (6, 56). It will be of interest to determine whether inactivating mutations in MLL3 are found in other families in which NPC is one of several cancer types observed.

In summary, we have identified the first instance of a plausible high penetrance inherited mutation predisposing to NPC. This study indicates that by performing WES on just a few affected individuals from a single well-chosen family, it is possible to generate a small list of highly likely disease-causing germline mutations that are amenable to future functional investigation.
ACKNOWLEDGEMENTS

We thank E Bartom for development of the WES analysis pipelines; M Jarsulic for technical assistance with job execution on high-performance computing clusters.
REFERENCES


**TABLE 1.** Familial deleterious germline mutations in MLL3 identified by WES. All three variants were observed in all three NPC-affected family members.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Position</th>
<th>Reference Allele</th>
<th>Variant Allele</th>
<th>Exonic Function</th>
<th>rsID</th>
<th>Nucleotide Change</th>
<th>Protein Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLL3</td>
<td>7</td>
<td>151945071</td>
<td>-</td>
<td>T</td>
<td>stopgain SNV</td>
<td>rs150073007</td>
<td>2447dupA</td>
<td>Y816*</td>
</tr>
<tr>
<td>MLL3</td>
<td>7</td>
<td>151945204</td>
<td>G</td>
<td>A</td>
<td>ns SNV</td>
<td>rs4024453</td>
<td>C2315T</td>
<td>S772L</td>
</tr>
<tr>
<td>MLL3</td>
<td>7</td>
<td>151970856</td>
<td>T</td>
<td>A</td>
<td>ns SNV</td>
<td>rs10454320</td>
<td>A946T</td>
<td>T316S</td>
</tr>
</tbody>
</table>

Chr: Chromosome; SNV: single nucleotide variant; ns: nonsynonymous
FIGURE LEGENDS

Figure 1. Nasopharyngeal Carcinoma family pedigree. Shown is a four-generation pedigree of the family. Germline and tumor DNA of individuals with NPC (Individuals I-1, I-2, and II-1) were analyzed by WES.

Figure 2. MLL3 mutations in NPC. Diagram of MLL3 with functional domains. Black lollipops indicate germline mutations identified in this study. White lollipops indicate previously identified somatic mutations.
Figure 1

Colon dx 65  Prostate dx 75  Stomach dx 92  Breast dx 44

I-1
NPC dx 49
Renal Cell Carcinoma dx 45

I-2
NPC dx 29

II-1
NPC dx 39

d. 18 mo Brain Tumor
Breast dx 44  Cervical dx 29

Nasopharyngeal Carcinoma

Other Cancers
Integrated Genomic Analysis Suggests MLL3 is a Novel Candidate Susceptibility Gene for Familial Nasopharyngeal Carcinoma

Mark M. Sasaki, Andrew D. Skol, Riyue Bao, et al.

Cancer Epidemiol Biomarkers Prev  Published OnlineFirst May 26, 2015.

Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-15-0275

Supplementary Material
Access the most recent supplemental material at:
http://cebp.aacrjournals.org/content/suppl/2015/05/28/1055-9965.EPI-15-0275.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.