Whole Exome Sequencing and Analysis - Illumina

Neuroblastoma (NBL)

*Protocols were performed at the Broad Institute. Please reference Pugh et al. (Published in final edited form as:Nat Genet. 2013 Mar; 45(3): 279–284).

The generation, sequencing, and analysis of 222 pairs of exome libraries at the Broad Institute was performed using a previously described protocol27. Due to the small quantities of DNA available, 81 DNA samples were amplified using Phi29-based multiple-strand displacement whole genome amplification (Repli-g service, QIAgen). Exonic regions were captured by in-solution hybridization using RNA baits similar to those described27 but supplemented with additional probes capturing additional genes listed in ReqSeq78 in addition to the original Consensus Coding Sequence (CCDS)78 set. In total, ~33 Mb of genomic sequence was targeted, consisting of 193,094 exons from 18,863 genes annotated by the CCDS86 and RefSeq86 databases as coding for protein or micro-RNA (accessed November 2010). Sequencing of 76 bp paired-end reads was performed using Illumina Genome Analyzer IIx and HiSeq 2000 instruments. Reads were aligned to the hg19/GRCh37 build of the reference human genome sequence78using BWA70. PCR duplicates were flagged in the bam files for exclusion from further analysis using the Picard MarkDuplicates tool. To confirm sample identity, copy number profiles derived from sequence data were compared with those derived from microarray data when available. Candidate somatic base substitutions were detected using muTect (previously referred to as muTector27) and insertions and deletions were detected using IndelGenotyper27. Segmental copy number ratios were calculated as the ratio of tumor fraction read-depth to the average fractional read-depth observed in normal samples for that region.

Removal of oxoG library preparation artifact

Cases sequenced using WGA and native DNA were sequenced more than eight months apart by the Sequencing Platform at the Broad Institute. Initial comparison of candidate mutation calls from these two data sets identified a preponderance of apparent G>T or C>A substitutions of low allele fraction (<0.15) and within specific sequence contexts (Supplementary Figure 2A). We subsequently characterized this artifact and developed a method to detect and remove these events. In brief, these artifacts are introduced at the DNA shearing step of the library construction process and arise from the oxidation of guanine bases (oxoG) by high-energy sonication. During downstream PCR, oxoG bases preferentially pair with thymine rather than cytosine, resulting in apparent G>T or C>A substitutions of low allele fraction and enriched within specific sequence contexts (Supplementary Figure 2B). Consistent with this mechanism, the intensity of the sonication process was increased with the introduction of a new 150 bp shearing protocol between preparation of the WGA and native DNA samples.

The number of artifacts in a library was apparently sample-dependent (Supplementary Figure 2C) and these events were found in unmatched tumor and normal libraries. In some cases, thousands of candidate mutations were called in cases with a heavily affected tumor sample and an unaffected normal. However, nearly every sample had at least one such artifact and we

have observed similar events in publically available data sets from other centers, suggesting a common artifact mode that was exacerbated in some of our samples. To address this problem, we devised a method to differentiate oxoG artifacts from bona fide mutations.

Due to the modification of only one strand of a G:C base-pair (i.e. only the G base), reads supporting the artifact have characteristic read-orientation conferred upon adapter ligation. Therefore, all reads supporting an artifact were almost exclusively derived from the first or second read of the Illumina HiSeq instrument. Bona fide variants are supported by near-equal numbers of first and second reads. We made use of the skewed read-orientation combinations and low allele fractions characteristic of this artifact to identify and remove oxoG artifacts from mutation calls in our cohort (i.e. removal of all variants with allele fraction <0.1 or exclusively supported by a single read orientation).

Verification of somatic mutations and rearrangements

We used a combination of genotyping and sequencing technologies to verify random candidate mutations (PCR/Sanger and PCR/HiSeq sequencing of candidates from Complete Genomics and BC Cancer Agency Illumina WGS and RNA-seq data), as well as mutations supportive of our significance analyses (Sequenom and PCR/MiSeq of WES and WGS data). Combining all of the validation experiments resulted in overall validation rates of 87% for substitutions (525/605 candidates, 241/282 coding) and 34% for indels (27/79 candidates, 26/41 coding). Some mutations were verified using multiple technologies and therefore the total number of candidate mutations verified is lower than the sum total of mutations described in the Supplementary Note. See Supplementary Note for details and cross-platform comparisons.

Integrated analysis of somatic variation from exome and genome data sets

Somatic mutations detected in WGS, WES, and RNA-seq data sets were annotated using Oncotator (See Broad Institute Cancer Genome Analysis webpage). Genes mutated at a statistically significant frequency were identified using MutSig32, a method that identifies genes with mutation frequencies greater than expected by chance, given detected background mutation rates, gene length and callable sequence in each tumor/normal pair. The relationship between mutation frequency and age of diagnosis was tested using the Spearman rank test. The implementation of the Kolmogorov-Smirnov test in R version 2.11.1 (ks.test) was used to test differences in mutation frequency distributions of several clinical variables (Supplementary Table 4).

Germline variant analysis

Detection of pathogenic germline variation at base-pair resolution in a cohort of cancer patients is complicated by selection of an appropriately matched and sized control population, relatively high carrier frequencies for unrelated disorders, and complex genetics underlying cancer predisposition. To nominate germline variants predisposing to neuroblastoma, we searched for enrichment of putative functional variants in the blood-derived DNA samples from our WES

cohort compared to normal DNAs from 1,974 European American individuals sequenced by the National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project (ESP)56. As indel calls from the ESP cohort were not publically available at the time of our study, we did not include them in our analysis.

To ensure consistency and accuracy of germline variant detection, all neuroblastoma WES cases were called simultaneously with 800 WES cases from the 1000Genomes project using the UnifiedGenotyper from the Genome Analysis Toolkit. A principal component analysis of the genotype calls was performed to determine the ethnic background of our cases (Supplementary Figure 7) with respect to three 1000Genomes populations. As over 80% of our cohort was Caucasian or ad-mixed Caucasian, we downloaded genotyping calls and coverage information from 1,974 European American individuals available on the ESP website to serve as a control population. To focus our analysis on rare variation consistent with the low prevalence of neuroblastoma, we removed from both data sets all variants present in individuals sequenced as part of the 1000 Genomes project. Next, we generated two lists of rare variants: overlaps with clinically-reported variants recorded in ClinVar (downloaded 4/27/2012, 284 variants in neuroblastoma, 2,947 in ESP) and loss-of-function variants in any of 924 genes listed in the Cancer Gene Census53, Familial Cancer database54, or a list of DNA repair genes55 (86 neuroblastoma, 1,068 ESP). We then tested each gene for significant enrichment of variants in the neuroblastoma compared to the ESP cohort (1-tailed Fisher's exact test, Supplementary Tables 7 and 8).

The germline ClinVar analysis uncovered four genes of significance driven by single variants seen at greater frequency in neuroblastoma compared to ESP: CYP2D6, NOD2, SLC34A3, and HPD. All of these variants are present at low frequency in an expanded European American ESP cohort (rs5030865 in 1/8,524 chromosomes, rs104895438 in 5/8600, rs121918239 in 14/8514, and rs137852868 in 11/8600), suggesting they are benign polymorphisms. Note that, while candidates detected by this approach are not significant after correction for multiple testing, we believe there is sufficient biological rationale and supporting evidence for validation in larger cohorts. We also looked for overlap with sites recorded in COSMIC33. This analysis identified a TP53 variant associated with Li-Fraumeni syndrome60.