

Name of Project: The genetic basis of neuroblastoma

SPECIFIC AIM

In the past year, the genetic basis of human neuroblastoma, an important pediatric disease that exacts 15% of mortality attributable to childhood cancer, has come into focus. We have discovered highly penetrant mutations in the *ALK* oncogene as the major familial neuroblastoma gene, and these efforts are being leveraged both diagnostically as well as therapeutically. However, 99% of neuroblastoma cases arise in patients without a family history. To address the problem of discovering the genetic basis of sporadic neuroblastoma, we are conducting a large genome-wide association study (GWAS) in 5,000 sporadic neuroblastoma cases and 10,000 controls to explore the hypothesis that neuroblastoma is a complex disease that results from the interaction of mutant alleles with relatively low to moderate effect on tumor initiation. As described in the research plan, we have had tremendous early success and within a year have identified six loci highly associated with neuroblastoma, and these discoveries have been each replicated. As we continue to build the numbers, it has become clear that we must now begin to understand what these association signals are teaching us, and how the DNA variation and mutation in the germline results in the enhanced propensity for a developing neuroblastic cell to acquire a malignant phenotype. In addition, we have a unique opportunity to study the cancer genome in parallel to understand the somatic event that initiate or propagate tumorigenesis. We think that the GWAS signals we have discovered to date, and the others that will emerge, highlight the critical regions of the genome that are relevant to neuroblastoma biology. We now propose an ambitious one year research plan to quickly define the genomic architecture of all known genomic loci associated with neuroblastoma and to understand the mechanism by which DNA variation and/or mutation within a subset of these regions results in neuroblastoma.

Specific Aim 1. Discover the genomic architecture at validated loci discovered in the neuroblastoma GWAS within both germline and tumor tissues. The major goal of this Aim is to discover rare but more highly penetrant mutations in germline and/or tumor DNAs that will be critical for understanding the biological relevance of the discovered association signals. We will capture all known coding, regulatory and conserved sequences from all validated neuroblastoma associated regions from 80 neuroblastoma cases (both germline and tumor DNAs) and 40 controls using a liquid phase region specific extraction (RSE) procedure employing biotinylated oligonucleotides to enrich the sequences of interest from up to 10 regions (~35 kilobases of total sequence per DNA sample). Libraries for resequencing will be created with unique bar-codes to allow for multiplexing of 96 samples per sequencing run, and data will be generated using the AB SOLiD 3 instrument. Data will be analyzed to parse unique variations associated with known SNP risk alleles in neuroblastoma cases compared to controls, and determine which DNA variations are enriched or extinguished in the tumor samples. While we will computationally map the entire genomic architecture of the captured information from each region, special attention will be focused on discovering rare mutations that will have a more profound effect on region-specific gene regulation and/or function, and these will be prioritized for further study in future research projects.

Background: The Challenge and Potential Impact

Childhood cancers should have a central role in the Nation's cancer research priorities. It is clear that many of the seminal discoveries in the field of cancer have been made by the study of relatively rare pediatric malignancies and cancer syndromes that often present in childhood. We initiated a Genome-Wide Association Study (GWAS) of human neuroblastoma for several reasons (R01-CA124709). First, despite its relative rarity of less than 1000 cases in the U.S. each year, it exacts a devastating toll as it accounts for 15% of childhood cancer mortality. Thus, it is an important human disease. Second, our translational research focus demands that we have a much deeper understanding of the genetic basis of the disease in order to develop therapeutic interventions that are rational and tractable. When we started our GWAS project in 2008, nothing was known about the genes involved in sporadic neuroblastoma tumorigenesis, but we have made substantive progress due to the highly collaborative international team we built (see below). Third, through our efforts in the Children's Oncology Group (COG), we had been steadfastly collecting annotated specimens to allow us to perform a properly powered GWAS in a rare disease. We hypothesized that the effect size of risk alleles would be higher in neuroblastoma compared to those discovered in GWAS's focused on adult malignancies because of the significantly reduced impact of the environment on cancers arising in utero or shortly after birth. This hypothesis has been

realized in our initial discoveries, and should provide further impetus for similar studies in other childhood cancers.

With slightly more than 2,500 of our planned 5,000 cases genotyped, our GWAS has had significant early success. As detailed below, we have not only discovered and validated common SNP variations that tag several regions across the genome that contain genes implicated in cancer, but also copy number variations (CNVs) that also are harbored within biologically relevant genes. As we build our GWAS to the planned accrual goals within the next 12-18 months, it is very clear that we are likely to discover the majority, if not all, of the DNA variations that contribute to the risk of developing neuroblastoma. The major challenge now is to understand the biological and clinical relevance of these association signals. We hypothesize that association signals are marking critical loci in the genome for neuroblastoma tumorigenesis. These loci will be influenced both by common variant effects on regional gene expression, but also by rare and highly penetrant mutations in the germline and/or tumor cells. **We therefore plan a thorough resequencing of the regions discovered to date in two genomes: both the germline and paired tumor from neuroblastoma patients, compared to control DNAs from subject who never developed cancer.** This effort will be sufficiently powered to discover putative rare mutations in these regions as we predict that these will occasionally be present in germline DNA, and will much more commonly be present in tumor DNA. In addition, by leveraging our ongoing efforts through the NCI-funded TARGET project (Maris, PI: <http://target.cancer.gov/>) we also plan a thorough molecular genetic characterization of the two loci that we have discovered to date that show somatic alterations consistent with the underlying genes functioning as oncogenic drivers. Taken together, successful completion of this challenge initiative will provide significant insight and traction on the heritable DNA variations and mutations that cause and propagate an important childhood cancer.

Approach

This section will briefly provide some background and preliminary data, in order to demonstrate feasibility, and then discuss two parallel Specific Aims that inform each other, but are not dependent on one another. The primary goal of this project is to discover mutations in both the germline and cancer genomes at regions marked by our GWAS signals using next generation sequencing (NGS) technology.

Background and preliminary data.

Neuroblastoma is a cancer of early childhood that arises from the developing autonomic nervous system. It is the most common malignancy diagnosed in the first year of life and shows a wide range of clinical phenotypes with some patients having tumors that regress spontaneously, whereas the majority of patients have aggressive metastatic disease (1). These latter neuroblastoma cases have survival probabilities of less than 40% despite intensive chemoradiotherapy, and the disease continues to account for 15% of childhood cancer mortality (1). Until recently, the genetic etiology of neuroblastoma was not known. We have now discovered the genetic basis of hereditary neuroblastoma (1% of all cases). The majority of hereditary neuroblastomas arise due to activating mutations in the *ALK* oncogene (2), and we have also shown that these activating mutations occur sporadically. A clinical trial of *ALK* inhibition therapy for children with refractory neuroblastoma is ongoing under our leadership. In addition, we have shown that a smaller subset of familial neuroblastoma cases is due to mutation in *PHOX2B* (3). Together, mutations in these genes explain most, if not all, of the highly penetrant (Mendelian) risk to develop neuroblastoma, and we have established a genetic testing program at our institution for suspected familial cases.

To address the problem of genetic susceptibility to neuroblastoma for the 99% of cases in which there is no family history, we designed a genome-wide association study (GWAS). Very briefly, through our decades-long efforts of collecting biospecimens for the vast majority of neuroblastoma patients diagnosed across North America in the 238 institution Children's Oncology Group, we have accrued enough cases to properly power a GWAS study. In a disease that afflicts about 700 children in the U.S. each year, we have over 5,000 cases available for study, and thus are planning to compare these neuroblastoma cases to 10,000 children without neuroblastoma accrued through the Children's Hospital of Philadelphia network. We hypothesized that effect sizes might be larger in this embryonal tumor that occurs early in life due to the likely lesser impact of the environment, compared to the majority of malignancy affecting adults. This project has been highly successful since the start date of 04/01/08, and

we have already identified and validated six loci that are highly associated with neuroblastoma. Currently, we have genotyped over 2,500 neuroblastoma cases, and our Center for Applied Genomics has already delivered on the 10,000 controls. We have restricted our early analyses to individuals of European descent due to the simple fact that over 70% of cases occur in this population, but now will begin to address susceptibility in African Americans. We will briefly highlight each of our major biological discoveries below, followed by our methodological advancements that are critical for enabling our research.

Initial discovery of a susceptibility locus to clinically aggressive neuroblastoma at 6p22 (4). We performed a preliminary analysis of our GWAS in the first 1,032 neuroblastoma patients and 2,043 controls of European descent. We observed highly significant association between neuroblastoma and the common minor alleles of three single nucleotide polymorphisms (SNPs) within a 94.2 kilobase (Kb) linkage disequilibrium block at chromosome band 6p22 containing the predicted genes *FLJ22536* and *FLJ44180* (P -value range = 1.71×10^{-9} – 7.01×10^{-10} ; allelic odds ratio range 1.39–1.40). Homozygosity for the at-risk G allele of the most significantly associated SNP, rs6939340, resulted in an increased likelihood of developing neuroblastoma of 1.97 (95% CI 1.58–2.44). Subsequent genotyping of these 6p22 SNPs in the three independent case series (N=720 cases and N=2128 controls) confirmed our observation of association ($P=9.33 \times 10^{-15}$ at rs6939340 for joint analysis). Unexpectedly and demonstrating the power of our rich phenotypic information in our GWAS, neuroblastoma patients homozygous for the risk alleles at 6p22 were more likely to develop metastatic (Stage 4) disease ($P=0.02$), show amplification of the *MYCN* oncogene in the tumor cells ($P=0.006$), and to have disease relapse ($P=0.01$).

Common variations in *BARD1* influence susceptibility to high-risk neuroblastoma (5). Having shown that the 6p22 signal was enriched in the high-risk subset of case, we performed a second GWAS restricted to this most clinically relevant group of patients. In a study underpowered by most estimations, we showed that by focusing on only 397 high-risk cases (2,043 controls) we not only enriched the previously discovered 6p22 signal, but also detected new significant association of six SNPs at 2q35 within the *BARD1* (*BRCA1*-associated RING domain-1) gene locus ($P_{\text{allelic}} = 2.35 \times 10^{-9} - 2.25 \times 10^{-8}$). Each SNP association was confirmed in a second series of 189 high-risk cases and 1,178 controls ($P_{\text{allelic}} = 7.90 \times 10^{-7} - 2.77 \times 10^{-4}$). The two most significant SNPs (rs6435862, rs3768716) were also tested in two additional independent high-risk neuroblastoma case series, yielding combined allelic odds-ratios of 1.68 each ($P = 8.65 \times 10^{-18}$ and 2.74×10^{-16} , respectively). Finally, significant association was also found with known nsSNPs from the coding and regulatory regions of *BARD1*. These data show that common variation in the *BARD1* tumor suppressor gene contributes to the etiology of the aggressive and most clinically relevant subset of human neuroblastoma.

To follow-up this GWAS signal since publication, we have initiated multiple parallel recent investigations. We fine-mapped 2q35 with HapMap CEU data and found that the six genome-wide significant 2q35 SNPs are all in strong linkage ($r^2 > 0.60$) with 14 genomic regions densely occupied with regulatory domains, suggesting that common variation at 2q35 may alter *BARD1* expression and/or *BARD1* mRNA processing. We then showed that *BARD1* mRNA and protein expression is increased in neuroblastoma cell lines with a homozygous risk allele genotype at the disease associated SNP, rs3768716 (**Figure 1**). We subsequently identified fifteen alternatively spliced *BARD1* transcripts in neuroblastoma tissues, but only a subset are translated into detectable levels of protein (**Figure 1c**). More provocatively, the specific expression levels of certain *BARD1* mRNA variants appears to be strongly influenced by

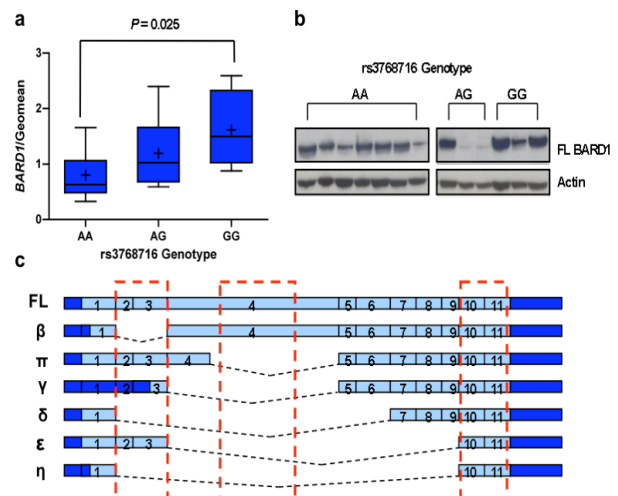


Figure 1. Common SNP variation at 2q35 is associated with alterations in *BARD1* expression. Neuroblastoma cell lines with a homozygous risk allele genotype have increased *BARD1* mRNA (a) and protein (b) expression. (c) A subset of the 15 identified *BARD1* mRNA variants are expressed at the protein level consistent with their open reading frame (light blue) across alternatively spliced boundaries. Antibodies available for use selectively bind N-terminal, C-terminal, and exon 4 *BARD1* epitopes (red boxes) allowing for identification of specific isoforms. The β isoforms appears to be preferentially expressed in high-risk neuroblastomas (data not shown).

common variation at 2q35, including the β isoforms that splices out the BRCA1 binding RING domain and has been proposed as an oncogenic driver in other malignancies (6). Further, consistent with a possible role for *BARD1* mRNA variants in the predisposition of neuroblastoma, we showed that the β isoforms and a subset of other isoforms are preferentially expressed in human developing sympathetic nervous tissues harvested from fetal autopsies. **Taken together, these data support a growth promoting (oncogenic) rather than a tumor suppressive role for *BARD1* in neuroblastoma tumorigenesis.**

As a pilot study for the work proposed here, we are in the midst of performing Sanger-based resequencing of all *BARD1* coding exons and 1000 bp of flanking sequence in 188 primary neuroblastomas. At the time of writing, we have completed approximately 30% of the work (unfortunately, much of the 5' region of the gene where the association signal was strongest is still in the pipeline), and have identified three heterozygous new mutations (S103N, N326D and S761N). The S103N and the S761N variants are in the RING and BRCT domains of the *BARD1* protein, respectively. The S761N variant has been described as a mutation in breast and uterine cancers, but the other two are apparently novel (none of these three are in SNP databases or in a panel of 228 control alleles). These initial results support our hypothesis that rare, potential disease contributing, mutant alleles will be found in GWAS regions. Ongoing work in neuroblastoma cell line models will investigate the phenotypic consequences of both transient and stable *BARD1* knockdown with pooled and isoform-specific siRNA and shRNA reagents. In addition, the design of mammalian expression constructs is underway to allow for the investigation of the over expression of a subset of *BARD1* variants and isoforms in these, and other, models. Lastly, investigations are underway to determine how common variation at 2q35 influences *BARD1*'s DNA damage response.

Copy number variation at 1q21.1 associated with neuroblastoma (7). To address the problem of common copy number variations (CNVs) also representing a significant source of genetic diversity, we have completed the first thorough CNV-based GWAS in a human cancer. This required significant methodological and computational innovation, and we compared 846 Caucasian neuroblastoma cases to 803 healthy Caucasian controls at 550,000 single nucleotide polymorphisms, and performed a CNV-based test for association. We then replicated significant observations in two independent sample sets comprised of a total of 595 cases and 3,357 controls genotyped on the same platform. We identified a common CNV at 1q21.1 associated with predisposition to neuroblastoma ($P = 2.97 \times 10^{-17}$; OR = 2.49, 95% CI: 2.02 to 3.05). This CNV was validated by quantitative PCR, fluorescent *in situ* hybridization, and analysis of matched tumor specimens. The CNV was shown to be heritable in an independent set of 713 trios without reported incidence or history of cancer. Finally, we identified a novel transcript that maps within the CNV; this transcript showed high sequence similarity to several neuroblastoma breakpoint family (NBPF) genes and mRNA expression of this gene was correlated with DNA copy number at the CNV. These data demonstrate that inherited copy number variation at 1q21.1 is associated with susceptibility to neuroblastoma and provide strong evidence for a novel NBPF gene family member influencing susceptibility to neuroblastoma. To our knowledge, this is the first example of a CNV being associated with susceptibility to a human cancer.

Integrative genomics identifies *LMO1* as a neuroblastoma predisposition gene (8). A large subset of our cases also has tumor tissue genotyped on the same array platform (as well as expression profiled and methylation status determined genome-wide), providing a rich resource for integrative genomics. As we have built the numbers, additional signals have emerged and have been replicated, and a recent comparison of 1,817 cases to 4,761 control subjects showed a highly significant association with *LMO1* (LIM domain only 1) at 11p15.4 that has been replicated in two independent case series (rs110419, combined $P=1.4 \times 10^{-13}$, OR=0.74; **Figure 2**). *LMO1* encodes a cysteine-rich transcriptional regulator, and its paralogues (*LMO2*, *LMO3* and *LMO4*) have been previously implicated in cancer. Copy number alteration analysis of tumor specimens indicated that 17% of primary tumors harbor gain encompassing *LMO1*, and that another 13% show LOH with uniparental isodisomy. The SNP risk alleles and somatic *LMO1* copy number gains, and LOH with isodisomy, are each independently associated with increased *LMO1* expression in primary tumors, indicating that gain-of-function mutations may influence neuroblastoma tumorigenesis. Whole-genome transcriptional genomics further identified *PTK7*, *NCBP2* and *MLLT3* as potential downstream targets of *LMO1*, and their expression levels correlate with both *LMO1* risk genotypes and copy number gains. Together, these data illustrate the importance of applying integrative molecular and genomic approaches in both the germline and cancer genomes to identify novel oncogenes.

Additional discoveries, methodological achievements and data sharing. At the time of this writing we have one additional SNP- and one additional CNV-based association discovery that have each been replicated once, but we are awaiting independent replication by our international collaborators. Finally, in terms of progress, we have published several methods papers that enabled our research and are utilized by the research community (9-12). As an additional resource to the research community, all data are submitted to dbGAP: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000124.v1.p1 and phenotype information (including clinical covariates and outcome) are available to academically qualified petitioners. Several such projects are ongoing, and we are also collaborating with Dr. Andy Olshan on a new R01 to overlay environmental information to our genome-wide and his candidate gene SNP genotyping data via a national case-parent triad study (R01-CA132887).

The results described above provide several important discoveries, but also present multiple significant challenges. While we were the first to show that childhood cancers indeed can be influenced by common variation in genome, and also the first to show that CNVs are definitively associated with cancer susceptibility, we still have very little understanding of the biology underlying these association signals. This application is designed to address this void.

General strategy

This grant will have the primary goal of rapid discovery of oncogenic mutations at all neuroblastoma-associated loci discovered at the time the project launches. In addition and in parallel, we will thoroughly

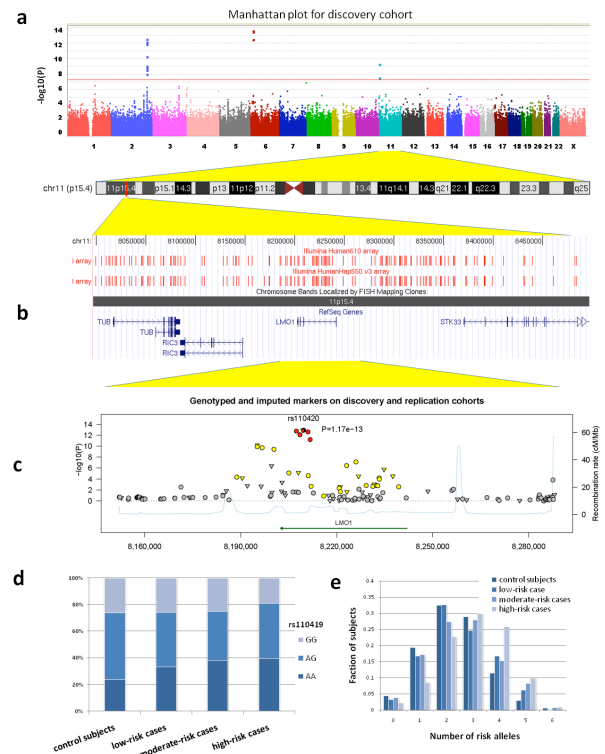


Figure 2. *LMO1* GWAS discovery at 11p15.4. (a) Three genomic loci reached genome-wide significance. (b) 11p15 locus encompasses *LMO1*. (c) Both genotyped (diamonds) and imputed SNPs (grey circles) are plotted with their combined P-values. (d) Stratifying the disease by risk groups illustrate that the frequency of risk genotypes of rs110419 tends to be higher in high-risk group. (e) Combining 3 known neuroblastoma susceptibility loci, subjects carrying 4 or more risk alleles are >2 fold more likely to develop a high-risk neuroblastoma than control subjects.

dissect the functional consequence of both common DNA variation and rare mutations at two loci harboring strong candidates for neuroblastoma oncogenes. Key to our strategy is the parallel assessment of germline and tumor genomes in both Aims, and tying our results to rich genomics and clinical datasets that have been created through our work with the Children's Oncology Group.

Specific Aim 1

Rationale for Aim 1. The motivation for this Aim is the expectation that our highly robust association signals are marking points in the human genome that are involved in neuroblastoma tumorigenesis via both common variation and acquired mutations. While we plan to understand how the association signals themselves translate into variations in regional gene expression in Aim 2 (trans effects across the genome will not be explored in this grant, but are likely also critically important), we think it is also very likely that rare variations are also present in germline DNA samples that may have a more profound impact on regional gene expression. As was recently demonstrated in type 1 diabetes (13), these rare variants can be particularly informative in understanding how gene expression variations may influence phenotype. For example, an association signal within a gene in the context of cancer may mark the presence of an oncogene or tumor suppressor. Discovery of rare sequence variations that encode nonsense alterations or, on the other hand, predict for constitutive activation of a kinase function, would be particularly informative for future studies. In addition, we expect that these rare sequence alterations may be enriched for somatically, and/or that genomic copy number alterations may occur to amplify the effect of the common variation association. For example, we have shown that the risk alleles associated with neuroblastoma at the *LMO1* locus are also associated with increased *LMO1* expression (**Figure 3A**). In addition, tumors arising in patients with the *LMO1* risk allele and more likely to show genomic gain or LOH with uniparental isodisomy, and this also is correlated with increased *LMO1* expression (**Figure 3B**). Therefore, we plan parallel deep resequencing of a large number of germline and paired tumor samples with the primary goal of discovering rare but highly penetrant DNA sequence alterations. Paired with ongoing genomic analyses of these same tumor samples in the TARGET project, we will have a rich data set to understand the biological meaning of the genetic susceptibility signals we have discovered to date, and a platform to extend these studies to future discoveries.

Approach for Aim 1.

Samples. The COG has developed a tremendous biological bank of highly annotated human neuroblastoma specimens. Dr. Maris led the effort through over the last decade to develop this resource, and we accrue over 650 primary tumor samples with matched blood to our bank annually. This has enabled our multiple genomics projects and has helped establish new prognostic criteria based on tumor genomics (14), as well as new discoveries that are being translated into new targeted therapies in phase 1 clinical trials currently (2,15). For this project, we propose to study 80 neuroblastoma germline DNAs, 80 matched neuroblastoma tumor DNAs, and 40 control germline DNAs. We have developed a method to enrich our sample set for cases with haplotypes showing the highest odds ratios, allowing us to keep the sample size relatively small. We have extensive experience in building case-series in the manner (14,16-18), and all samples will undergo rigorous quality control in our central reference laboratory including direct estimation of %tumor content by histopathology. For this project, only cases with tumor samples showing >80% tumor content will be eligible for study.

- **Pitfalls and alternatives.** Our published work documents our ability to garner the patient resources

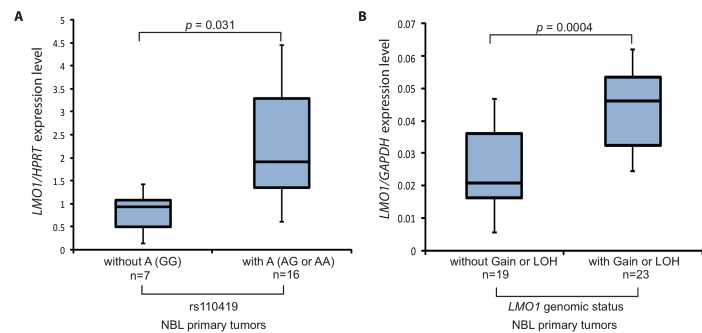


Figure 3. *LMO1* risk genotypes and copy number gains are associated with increased mRNA expression. *LMO1* expression in primary neuroblastomas by quantitative RT-PCR, normalized to housekeeping gene *HPRT*, in primary neuroblastomas segregated by either (A) genotype at rs110419 (most significantly associated SNP with adenosine being the risk allele); or (B) copy number and allelic status as defined on the Illumina (HH550) SNP array.

necessary for a project of this size and scope (4,5,7-9,14,16,17,20). The biggest potential problem is that we generally receive a relative small blood sample (from often very small children) at diagnosis, making our stock supply of germline DNA limiting in many cases. Tumor DNA quantity can also be a problem, especially in high-risk disease when the diagnostic sample may be from a metastatic site (bone marrow) and/or of very small size. While we frequently perform genome-wide amplification (GWA) of DNA stocks, and have used this successfully in Sanger sequencing projects, because the effect of GWA in deep resequencing experiments is as yet untested, we will only use unmanipulated DNA in this project. In addition, with well over 6,000 cases of matched germline and tumor samples in the bank currently, and with close to 3,000 cases now studied in our GWAS, we expect no problem in identifying the cases and controls for this study.

Regions and targeted sequencing strategy. We predict that by the time this project starts, ten loci will be amendable for study. These regions include each of our four discovered, validated and published association loci in our ongoing GWAS (4,5,7,8), two recently discovered loci that are awaiting final replication (*NME7* and *DUSP12*), as well as the two hereditary neuroblastoma predisposition loci we have discovered (2,3). **Table 1** (next page) shows these eight loci, and includes completed analyses of somatic DNA copy number alterations at each locus in our set of nearly 600 primary neuroblastoma specimens. These data show that somatic gain of these loci are common events, and as we discussed above, somatic mutation of *BARD1* further supports gain of function alterations at this locus. The rationale for including the hereditary loci is that we do not think we have discovered all of the highly penetrant mutational events as yet, and also that we have not excluded the possibility that common variation at these loci also influence tumor initiation in cases without a family history. We have budgeted for 10 total loci because we think it is highly likely, given the pace of discovery and multiple additional tentatively identified association signals, that at least two other regions will be validated by the time we set forth on DNA capture as described below. We will use standard genome analysis software tools to identify all open reading frames, putative exons, conserved sequence elements as well as 1000 bp flanking all genes within these regions. The definition of “region” is purposely conservative (**Table 1**). For GWAS signals, we will define the region by all SNPs within an r^2 of 0.1, and for the *ALK* and *PHOX2B* loci, we will include 1 kb flanking each gene. Copy number association regions are more difficult to define, but we again will err on the side of being inclusive (see Table legend).

- **Pitfalls and alternatives.** We expect no problem in mapping the regions to be studied, although we are acutely aware that our targeted approach may miss DNA variations that are relevant to neuroblastoma. As a very viable alternative, we could adapt an agnostic approach and simply resequencing the entire region in linkage disequilibrium with the associated SNPs or CNVs (within an r^2 of 0.1-0.2). This would increase the amount of DNA to be sequenced by several

Chr	Association Type	Gene ¹	Gene Type	Exons	Region size ²	Target Gene Deletion ³	Target Gene Gain or Amplification ³
6p22.3	Common SNP	FLJ22536	Non-coding RNA	12	110.8	10/599 (1.7%)	158/599 (26.4%)
2q35	Common SNP	BARD1	Coding, cancer gene	11	278.6	14/599 (2.3%)	136/599 (22.7%)
1q21.1	Common CNV	NBPF8	Coding, neurodevelopment		79.1	11/599 (1.8%)	131/599 (21.9%)
11p15.4	Common SNP	LMO1	Coding, cancer gene	4	227.8	75/599 (12.5%)	102/599 (17.0%)
1q24.2	Common CNV	NME7	Coding, cell cycle kinase	7	94.1	7/599 (1.2%)	141/599 (23.5%)
1q23.3	Common CNV	DUSP12	Coding, phosphatase	6	219.4	9/599 (1.5%)	138/599 (23.0%)
4p13	Rare mutation	PHOX2B	Coding, tumor suppressor	3	733.8	90/599 (15.0%)	38/599 (6.3%)
2p23.1-2	Rare mutation	ALK	Coding, oncogene	29	9.9	0/599 (0.0%)	234/599 (39.1%)

Table 1. Eight regions identified to date as neuroblastoma predisposition loci via our GWAS screen or family-based linkage studies. ¹Association loci often contain more than one gene, but only the largest RefSeq gene is shown here (the 6p22, 1q21 and 1q24 loci each have other genes to be considered). ²Region size calculated for SNP associations by LD structure and includes all SNPs with $r^2 > 0.1$. This is more difficult for CNV associations as the SNPs within the CNVs are not in Hardy-Weinberg equilibrium, and the sizes shown include the maximum region containing all exons for genes mapping across the CNV. The regions containing the genes with high penetrance mutations are sized based on the coding region. We calculate that there are approximately 25-35 kB of target material within the eight regions, thus reducing the amount of material to be sequenced to 2% of the genomic landscape, thus substantially increasing the number of samples that can be studied. ³The final two columns show the frequency of genomic alteration at these loci discovered in a set of 599 human neuroblastoma samples studied on the same SNP array as used in the GWAS study. These data are consistent with most loci harboring an oncogenic effect on tumorigenesis, except the known tumor suppressor locus *PHOX2B*. Of note, the “deletions” at *LMO1* are almost exclusively LOH with isodisomy, with resultant overexpression of *LMO1*.

orders of magnitude, and would definitely identify multiple new SNPs, as was the case for the resequencing of 8q34 in prostate and colon cancers (21). By including germline and tumor DNA in parallel, acquired sequence alterations would be readily determined, but many of these could simply be passengers. The biggest disadvantage of the agnostic approach of resequencing the entire LD block is potentially cost (in capture and on the informatics end), and this could ultimately result in the need for a reduction in sample size. While we will remain open to this possibility as technology evolves, we currently plan to focus on a strategy that is gene- and regulatory region-centric as the primary goal is to find highly penetrant mutation events.

Region Specific Extraction (RSE): We will use a solution phase methodology for region-specific capture of DNA from our samples that we have dubbed Region Specific Extraction (RSE). RSE is being co-developed at CHOP with our paid consultant Generation Biotech, primary in collaboration with Co-I Dr.

Monos. RSE is being developed as a relatively simple and automated methodology for capture of predefined regions across the genome. As discussed in the *Pitfalls and Alternatives* section below, we have prioritized this methodology above the myriad of others being developed because it most specifically meets our Project-specific requirements of focusing on a relatively small amount of genomic territory (25-35 kilobases) and being amenable to nanogram quantities of input DNA. The methodology is published (22), and will only briefly be discussed here. RSE is a three-stage process that involves 1) targeting a specific region with sequence specific oligonucleotides, 2) discrimination of the targeted region from other fragments by enzymatic incorporation of biotinylated tags, and 3) purification of the targeted DNA. Oligonucleotides are designed with our in house RSE OligoTool, and these are hybridized to targeted areas of the genome by exploiting unique sequence elements. An enzymatic step then adds biotin depending on whether the target sequence is present or not. Streptavidin coated magnetic microparticles are added to the reaction mix to isolate the targeted DNA along with flanking regions; the non-targeted DNA is washed away, and capture efficiency monitored by quantitative PCR (copy number/input copy number). Standard operation procedures including automation on a Qiagen EZ-1 robot have been established in the Monos lab, and we now have extensive experience with regional DNA enrichment (**Figure 4**). Important to this project, oligonucleotides will be designed to incorporate known SNPs, allowing for allele /haplotype-specific extraction of nucleic acids (22), and this may provide a significant advantage in downstream applications (understanding if any newly discovered sequence alteration in germline or tumor DNA is linked to the risk allele discovered in our GWAS).

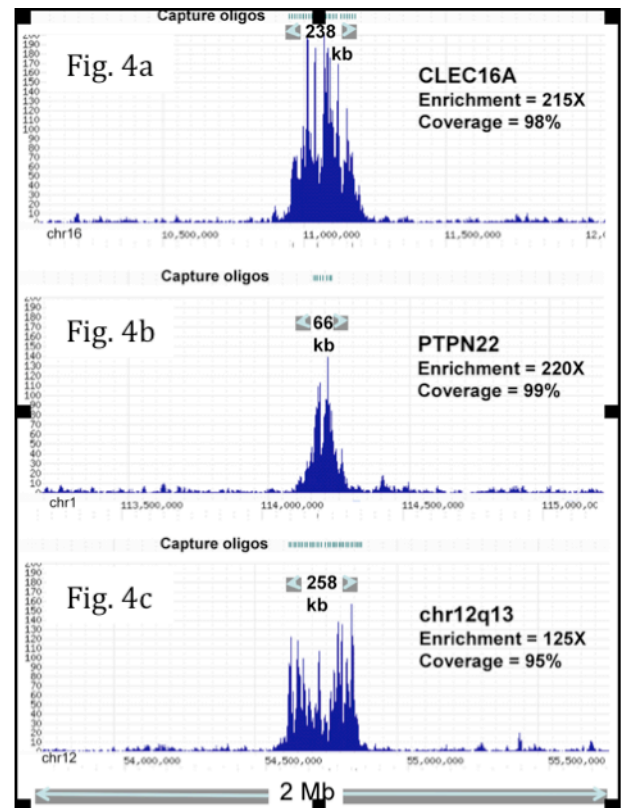


Figure 4. Mapped sequencing reads (Genome Analyzer platform) after RSE. Three separate gene regions of ~0.5Mb were targeted with 67 oligos spaced at 8 kb intervals in a single RSE. Coverage and enrichment ranged from 95-99% and 125-220X, respectively.

- *Pitfalls and alternatives.* The field of NGS is rapidly evolving, and this is especially true for methodologies used for enrichment of regions of the genome. We have followed this field intently, and have experience with each of the major technologies being advanced. Some of the commercial approaches include hybridization, washing and recovery of target from planar microarrays (Agilent, Roche-Nimblegen, Febit) or bead-based methods, which utilize libraries of probes to capture regions in solution (Agilent Oligonucleotide Library Synthesis - OLS). The scale of capture varies from 125 kb to 30 Mb (for whole exome capture) and from \$125 to \$2000 per sample. It is clear that most of the focus has been on capturing large regions of the genome, making chip-based methods cost-inefficient for our needs. A very interesting and different approach has been developed by RainDance Technologies, which uses microfluidics to control and mix small droplets of primers and template to generate libraries of stabilized microreactors for

use in emulsion PCR. This process makes the PCR workflow less labor-intensive, more cost-effective and may require minimal normalization while providing sequence coverage without major variation across the desired region. We do not yet have experience with RainDance, but plan pilot experiments in the near future. **Due to low input DNA requirements (maximum 400 ng), flexibility in oligonucleotide design, haplotype-specific nature of the capture, throughput, simplicity, low cost and based on our pilot data, RSE is currently optimal for our needs, resulting in a maximization of our sample number, and potentially also improve depth of sequence coverage, compared to other technologies currently available.** As we showed when evaluating the various SNP-based array technologies prior to launching our GWAS, we think we are well positioned to monitor this rapidly evolving field and make the best decision at the time funding is available, but if there are no significant changes in the field we are ready to move forward with RSE for this project. In addition, the other major challenge for all enrichment strategies, including RSE, is maximizing signal-to-noise. As show in Figure 4, background reads do exist, and it is not clear how these arise (we have experimental data that these are not due to unintended probe performance during the enrichment process). These artifacts are present in all methodologies currently available for targeted resequencing applications, and it is not clear whether or not RSE compares favorably with other techniques. This could be a problem in regions of high complexity (such as our NME7 locus), and this problem will need to be carefully monitored and evaluated. Importantly, our consultant Generation Biotech is working on methods to maximize signal-to-noise, and any technological advancements will be made available to this project.

Library construction and sequencing. We will construct fragment (i.e., single-read) libraries as our strategy for enriching only regions of the genome that are likely to be targets for mutation (exons, regulatory and conserved regions) and will not require the mapping advantages of paired-end libraries. We will utilize the AB SOLiD platform, and the captured target will be ligated to adapters (P1 and P2), which contain sequences for subsequent emulsion PCR and bead capture steps. The P2 adapter will also contain a barcode (5-8mers) to maximize sample throughput. Currently, we are routinely multiplexing 20 barcoded samples per sequencing run, but AB plans to deliver on the methods to barcode 96 samples per run by the fall of 2009. **Thus, we have designed the experiments with plans to individually and uniquely barcode each sample within a 96 well plate.** Applied Biosystems uses a ligation-mediated approach (Sequencing by Oligonucleotide Ligation and Detection) and “Two-base Encoding” to generate its extensions and signal, and the details of signal generation and image processing will not be discussed here. Once the sequence reads have been generated, the barcodes will be deconvoluted in our informatics pipeline before the more complex steps of sequence assembly and polymorphism assessment occur.

- *Pitfalls and alternatives.* We will use the AB SOLiD system for these experiments, but also have access to Solexa and 454 instruments on campus if a clear advantage of one method emerges. In our opinion, each of these technologies provide outstanding data, and there are pros and cons of each platform, as has been recently described (23,24). In order to use the capacity of NGS instruments efficiently and in a cost-effective manner, we plan to use high dimensional barcoding, but it is also possible to use a gasket to subdivide the physical space of the surface into multiple separate areas for different applications or samples. However, there are limitations to the number or regions, which can be generated in this fashion, and the areas covered by the separating material represent loss of potential throughput. We have developed a distributed model for NGS experiments at CHOP, with expertly trained users for library construction and sequencing working with individual investigators on a one-on-one basis. The extensive experience of Co-I Dr. Rappaport with conventional and next generation sequencing will improve the likelihood of 100% efficiency with little or no need for sample re-runs.

Data Analysis and statistical considerations for Aim 1. Our Bioinformatics Core (BiC) Facility, under the direction of Dr. Xiaowu Gai, has extensive experience with all of the major sequencing platforms and has installed that computational infrastructure to support the institutions NGS needs. Sequence alignment and analysis will be performed off instrument on a high performance compute (HPC) cluster. The system is composed of 8 compute nodes, each equipped with two quad-core Intel CPU's (a total of 64 cores). Each system contains 32GB of RAM to support loading of entire reference genomes when necessary. Local hard disk space will support up to 15 Terabytes of data initially. The current system has the ability

to align 17 million reads to a 135MB region on a single CPU in 48 minutes. Parallelization of this process across this HPC system reduces this time to less than 2 minutes. Sequencing data will be processed through an in-house developed analysis pipeline provided by the Bioinformatics Core. Sequencing data will be deposited on a shared file system, which is accessible to the HPC cluster. The web based analysis pipeline will allow a user to design a desired workflow through a variety of alignment algorithms and downstream tools. Information about experiments will be obtained and stored at every stage using a relational database. Alignments can be performed using a variety of algorithms available through the core, followed by automated analyses for subsequent SNP discovery, indel detection and comparison of related genomes (tumor versus blood).

Specifically, we will be interested in comparisons of the neuroblastoma germline genomes to the reference (control) genomes to determine if sequence variations are 1) unique in the neuroblastoma set (candidate mutations) or 2) enriched (associated with neuroblastoma). Standard procedure for single marker chi-square testing will be performed to test for association, but this study is not powered to prove new associations. We will be much more interested in the candidate mutations, and these will be prioritized for future analyses. In addition, we will also compare the neuroblastoma germline genome to the neuroblastoma tumor genome. Again, we will be interested in sequence variations that are unique and even perhaps enriched in the tumor DNAs compared to control. We have extensive experience with computationally assessing tumor-derived sequence variations as being potential mutations in the TARGET project (<http://target.cancer.gov/>) and our other areas of interest in the lab (2,25). The BiC at CHOP has developed tools for sensitive SNP detection, haplotype assembly and reference-guided de novo assembly of NGS sequence data.

As mentioned previously, the sample size is driven by primarily by cost, but scientifically by our desire to maximize the number of individuals studied. We are using a similar design to that published recently to discover rare variants in type 1 diabetes to drive our study design (13), but the “power” of our approach is likely higher by studying paired cancer genomes and a large number of control genomes in parallel. Based on our ongoing experience with neuroblastoma genomics related to *ALK* (2) and the TARGET projects, it is likely that the mutation frequency could be as low as 1-3% at some of our loci, but we should detect mutations at this frequency with our current sample size. As a matter of fact, if mutations in one of the targeted regions occurred in neuroblastoma even at a frequency of only 1%, there is a 99% probability of detecting at least one such mutation in our sample of 480 neuroblastoma germline DNAs. On the other hand, our control sample size of 240 germline DNA of unaffected children will give us a >90% probability of detecting any rare polymorphism with a frequency of 1% or more, and >70% for a frequency as low as 0.5%. In addition, as there may be tumor heterogeneity, deep resequencing could identify rare mutant alleles. As discussed below, we already have preliminary data validating this approach at the *BARD1* locus, and it is our expectation that the large sample size proposed here (rather than the alternative design of reduced sample size but increased genomic landscape) will help define mutations that can be used as tractable tools to dissect underlying biology and relevance to neuroblastoma tumorigenesis.

- *Pitfalls and alternatives.* This project will generate an enormous amount of data, but the institution has invested heavily in NGS technology, and quite appropriately an equal investment has been made in the BiC to support this work. Importantly, ongoing work in diabetes, autism, mitochondrial diseases and immunity (HLA loci) involving NGS will dramatically inform our work. The BiC is the central knowledge warehouse for NGS technology and informatics. Our team will be able to access these resources, but could reach out into the Penn community where other NGS applications are ongoing if we find barriers that are currently not anticipated. As mentioned above, we are cognizant that our approach may miss important sequence variations outside of the 2% of the genome from our ~10 regions that we will sequence. Likewise, we will be focused mainly on discovery of classic mutations that would predict a significant deleterious effect on normal protein function. This means that we could miss some more subtle mutations, even those arising in the noncoding regions that we sequence, if we are not careful. By comparing our results to a large number of controls run in parallel, and by constantly surveying results from other projects at CHOP and in the community, we should minimize this problem. Finally, we note that

despite selective targeting and sequencing of well-defined regions, all regional capture experiments that we are aware of continue to show artifacts outside of the intended target regions. We think these are problems with assembly, not capture, due mainly to the relatively short reads, especially in regions with repetitive elements. We are cognizant of this issue, and point out that our experimental plan with capture of coding and regulatory elements should minimize this problem compared to approaches that pull down entire regions.

Timeline and Milestones

Aim 1. We have the samples in hand and the pilot work has been completed. The computational work for defining the exact regions to be sequenced and oligonucleotide primer design for RCE is ongoing. Thus, year 1 will focus on capturing the predefined regions of interest in the samples. This should be complete by month 6, and we are exploring alternative approaches to RSE in parallel in case we run into trouble. Ongoing work in the lab makes changing capture strategy a not too disruptive alteration in experimental plan, and should not affect timeline dramatically. Sequence analysis will be complete by the end of year 1.

Literature Cited

1. Maris JM, Hogarty MD, Bagatell R, Cohn SL. Neuroblastoma. *Lancet*. 2007 Jun 23;369(9579):2106-20.
2. Mosse YP, Laudenslager M, Longo L, Cole KA, Wood A, Attiyeh EF, et al, Maris JM. Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature*. 2008 Oct 16;455(7215):930-5.
3. Mosse YP, Laudenslager M, Khazi D, Carlisle AJ, Winter CL, Rappaport E, et al, Maris JM. Germline PHOX2B Mutation in Hereditary Neuroblastoma. *Am J Hum Genet*. 2004 Oct;75(4):727-30.
4. Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH, et al. Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N Engl J Med*. 2008 Jun 12;358(24):2585-93.
5. Capasso M, Devoto M, Hou C, Asgharzadeh S, Attiyeh EF, Mosse YP, et al., Maris JM. A genome-wide association study identifies common variations in the BARD1 tumor suppressor gene predisposing to high-risk neuroblastoma. *Nature Genetics*. 2009;In Press.
6. Ryser S, Dizin E, Jefford CE, Delaval B, Gagos S, Christodoulidou A, et al. Distinct roles of BARD1 isoforms in mitosis: full-length BARD1 mediates Aurora B degradation, cancer-associated BARD1beta scaffolds Aurora B and BRCA2. *Cancer Res*. 2009 Feb 1;69(3):1125-34.
7. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al., Maris JM. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*. 2009;In press.
8. Wang K, Zhang H, Hou C, Diskin SJ, Winter C, Bosse K, et al., Maris JM. Integrative genomics identifies LMO1 as a neuroblastoma predisposition gene. Submitted.
9. Attiyeh EF, Diskin SJ, Attiyeh MA, Mosse YP, Hou C, Jackson EM, et al., Maris JM. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res*. 2009 Feb;19(2):276-83.
10. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*. 2008 Nov;36(19):e126.
11. Li H, Wei Z, Maris JM. A Hidden Markov Random Field Model for Genome-wide Association Studies *Biostatistics*. In press.
12. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007 Nov;17(11):1665-74.
13. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*. 2009 Apr 17;324(5925):387-9.
14. Attiyeh EF, London WB, Mosse YP, Wang Q, Winter C, Khazi D, et al., Maris JM. Chromosome 1p and 11q deletions and outcome in neuroblastoma. *N Engl J Med*. 2005 Nov 24;353(21):2243-53.
15. Maris JM. Unholy matrimony: Aurora A and N-Myc as malignant partners in neuroblastoma. *Cancer Cell*. 2009 Jan 6;15(1):5-6.
16. Mosse YP, Greshock J, Margolin A, Naylor T, Cole K, Khazi D, et al., Maris JM. High-resolution detection and mapping of genomic DNA alterations in neuroblastoma. *Genes Chromosomes Cancer*. 2005 Aug;43(4):390-403.
17. Wang Q, Diskin S, Rappaport E, Attiyeh E, Mosse Y, Shue D, et al. Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. *Cancer Res*. 2006 Jun 15;66(12):6050-62.

18. Cole KA, Attiyeh EF, Mosse YP, Laquaglia MJ, Diskin SJ, Brodeur GM, et al., Maris, JM. A Functional Screen Identifies miR-34a as a Candidate Neuroblastoma Tumor Suppressor Gene. *Mol Cancer Res.* 2008 6(5):735-42.
19. Huang W, He Y, Wang H, Wang Y, Liu Y, Wang Y, et al. Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc Natl Acad Sci U S A.* 2006 Jan 31;103(5):1418-21.
20. Schlisio S, Kenchappa RS, Vredeveld LC, George RE, Stewart R, Greulich H, et al. The kinesin KIF1Bbeta acts downstream from EglN3 to induce apoptosis and is a potential 1p36 tumor suppressor. *Genes Dev.* 2008 Apr 1;22(7):884-93.
21. Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum Genet.* 2008 124(2):161-70.
22. Dapprich J, Ferriola D, Magira EE, Kunkel M, Monos D. SNP-specific extraction of haplotype-resolved targeted genomic regions. *Nucleic Acids Res.* 2008 Sep;36(15):e94.
23. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 2009 Mar 27;10(3):R32.
24. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008 Oct;26(10):1135-45.
25. Raabe EH, Laudenslager M, Winter C, Wasserman N, Cole K, LaQuaglia M, et al. Prevalence and functional consequence of PHOX2B mutations in neuroblastoma. *Oncogene.* 2008 Jan 17;27(4):469-76.
26. Irminger-Finger I, Jefford CE. Is there more to BARD1 than BRCA1? *Nat Rev Cancer.* 2006 May;6(5):382-91.