Title
Genomic exploration of the peripheral nervous system: Identification of candidate genes for neuroblastoma, hearing loss, and other aspects of neuron biology and tumorigenesis

Author
Hackett, Christopher Sultan

Publication Date
2010

Genomic exploration of the peripheral nervous system: Identification of candidate genes for neuroblastoma, hearing loss, and other aspects of neuron biology and tumorigenesis

by

Christopher S. Hackett

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

This work is dedicated to my parents, who have always fostered the spirit of exploration.

Acknowledgements

I gratefully acknowledge all of the people who have supported me during graduate school and who have contributed to the work in this thesis. First, I would like to thank members of the Weiss Lab, past and present. In addition to providing helpful advice, the lab members created a wonderful working atmosphere that fostered a sprit for scientific exploration and a springboard for a career in science. In particular, I would like to thank QiWen Fan for providing molecular biology mentorship and amusing conversations over the several years we sat together in the lab. Nadia Milshteyn, David Goldenberg, Slava Yakovenko, and Kim Nguyen maintained our mouse colony and provided general lab management and maintenance, without which this project would not have been possible. Christine Cheng provided some very skilled technical assistance for some of the biochemical aspects of the project, and Yvan Chanthery contributed some important *in vivo* techniques. Justin Chen assisted with some of the informatics involved, and provided some useful genomics advice. Anders Persson and Fredrik Swartling provided interesting insight into neuron biology. I am grateful to all lab members for providing a wonderful social atmosphere in the lab. I am also particularly indebted to Clay Gustafson and Theo Nicolaides for their helpful guidance regarding the interface of science and clinical practice.

I would also like to thank our collaborators at UCSF. Pui-Yan Kwok and his lab, in particular Denise Lind and Ludmila Pawlikowska, facilitated all of the SNP genotyping described. Saunak Sen assisted with the interpretation of our genotyping results. Allan Balmain, Jin-Hua Mao, and David Quigley provided direction for our modifier screen, and assisted with the data analysis. David Quigley, in particular, enthusiastically contributed a cutting-edge analysis approach to our data that yielded exciting results. The members of Zena Werb's lab, in particular Pengfei Lu and Aditi Sharma, taught me critical techniques for the breast cancer aspects of this project. Nigel Killeen and his lab, in particular Steve Chmura and DongJi Zhang, assisted with the generation of the two lines of genetically engineered mice described. Graeme Hodgson, Sue Hariono, Greg Hamilton, and Taku Tokuyasu helped to train me in all aspects of array CGH analysis. I am also grateful to the neuropathology tissue bank and Joanna Phillips for fulfilling all of our pathology needs.

I am also grateful to our collaborators outside of UCSF. In particular, David Largaespada has generously assisted our lab in adopting the Sleeping Beauty system he developed. Aron Geurts, Adam Dupuy, Lara Collier, Timothy Starry, and Vincent Keng also provided technical assistance, reagents, and mice for the project. At the NCI, Javed Khan and his lab, in particular Young Song, taught me how to use and analyze expression arrays, and graciously welcomed me into their lab for extended periods of time. Terry Van Dyke very generously provided arrays and advice that were both critical for the project.

I would also like to thank my thesis committee: Zena Werb, Nigel Killeen, Kevin Shannon, and the chair, Allan Balmain, who provided encouragement, helpful advice and keen insight throughout the project, and welcomed me into their own labs to learn many of the necessary skills.

I am tremendously grateful to my thesis advisor, Bill Weiss, who has provided a limitless amount of support, encouragement, inspiration, and insight with regards to science, medicine, careers, and life while I was in his lab. Bill provided me with the opportunity to explore everything I was interested in scientifically and I was blessed to have such a wonderful mentor.

Finally, I would like to thank my family. My parents provided support and encouragement that got me though rough patches from grade school to graduate school and beyond, in addition to serving as occasional scientific collaborators. I would also like to thank my fiancé, Nan Chen, who among many other things gave me constant encouragement, reinforced our mutual love for science, and made the long hours we both spent at UCSF enjoyable.

This work contains published material from the following publications:

Chapter 5: Geurts AM*, Hackett CS*, Bell JB, Bergemann TL, Collier LS, Carlson CM, Largaespada DA, Hackett PB. Structure-based prediction of insertion-site preferences of transposons into chromosomes. Nucleic Acids Res. 2006 May 22;34(9):2803-11. PMID: 16489096

Chapter 6: Hackett, CS, Geurts, AM, Hackett, PB. Predicting preferential DNA vector insertion sites: implications for functional genomics and gene therapy. Genome Biol. 2007;8 Suppl 1:S12. PMID: 18047689

Chapter 7: Collier LS, Adams DJ, Hackett CS, Bendzick LE, Akagi K, Davies MN, Diers MD, Rodriguez FJ, Bender AM, Tieu C, Matise I, Dupuy AJ, Copeland NG, Jenkins NA, Hodgson JG, Weiss WA, Jenkins RB, Largaespada DA. Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality. Cancer Res. 2009 Nov 1;69(21):8429-37. PMID: 19843846

Note on figure and section order: For chapters derived from published manuscripts and manuscripts in preparation, figure order has been preserved, with supplementary figures renumbered and appended to the normal figures. Thus, the text may reference figures seemingly out of order, but this was done to retain the original order of the main figures in the published manuscripts. Additionally, the order of the sections (Introduction, Results, Discussion, and Methods, Figure Legends, References) has been preserved for the published manuscripts and may thus be inconsistent between chapters. Unpublished work follows the order above.

Abstract

Neuroblastoma is a deadly tumor derived from neuronal tissue for which the molecular drivers remain a mystery. Here we have applied classical genetics, analysis of expression quantitative trait loci (eQTL), and forward insertional mutagenesis to uncover novel pathways in the disease. We showed that liver arginase is a candidate susceptibility gene and interacts with component of the GABA pathway both genetically and biochemically to influence tumor susceptibility, and both of these pathways represent potential therapeutic targets. We then constructed a gene coexpression network in tumors and in sympathetic ganglia to explore novel genetic/functional interactions in both neuroblastoma and normal neurons. In particular, we used the coexpression network to identify novel candidate genes for several hereditary hearing loss loci. In a separate project, we focused on forward genetics utilizing the Sleeping Beauty insertional mutagenesis system. We developed a novel algorithm to predict local insertion site preferences of the vector, and show that the transposon system does not cause widespread genomic instability. We then generated a novel transgenic line, TH-SB11, to drive tumors in the peripheral sympathetic nervous system. Finally, we explored methods to drive tumors in the mammary gland, and generated a novel knock-in line capable of driving high-level conditional transposase expression in any tissue. This work illustrates the genetic complexity of neuroblastoma, and has identified novel functional pathways in the disease and a novel therapeutic target. In addition, this work lays the foundation for further gene discovery in neuroblastoma and other tumor types.

Table of contents:

List of Tables

List of Figures

Chapter 1: Introduction

Source: The following contains background on neuroblastoma and genetic screening technology relevant to several of the following chapters.

Contributions: This is an original review of the literature that has been aided by conversations with several individuals.

Introduction

Neuroblastoma is the third most common tumor of childhood and accounts for a disproportionately high share of childhood cancer mortality due to poor survival relative to other pediatric malignancies. Neuroblastomas arise from the developing peripheral neural crest and display cellular features indicating neuronal lineage, distinguishing these tumors from the more common tumors arising from epithelial, glial, hematopoietic, and stromal lineages. Likely due to this unique lineage, most of the common molecular pathways implicated in cancer do not show aberrant activity in neuroblastoma, and the pathways governing tumor development in neuroblastomas remain a mystery.

The most prominent genomic aberration associated with neuroblastoma is amplification of the MYCN proto-oncogene, observed in roughly one-third of cases and strongly associated with poor outcome [1,2]. MYCN is a transcription factor in the *c-myc* family, which has been shown to influence the expression levels of over 15% of genes in the genome[3-6]. As such, the specific mechanism by which MYCN drives neuroblastoma remains a mystery. The canonical model of myc-driven tumors holds that myc family transcription factors simultaneously drive cellular proliferation and apoptosis, and tumors require a second mutation to deactivate the apoptotic pathway. However, the second anti-apoptotic hit in neuroblastoma has not been rigorously established[7]. The vast majority of primary neuroblastomas have wild-type p53, MDM2, and p14/ARF[8,9], suggesting this core tumor suppressor pathway is intact. Similarly, the Rb/p16INK4a tumor suppressor pathway is only rarely mutated in primary neuroblastomas[10]. While studies of familial neuroblastomas identified the ALK receptor tyrosine kinase as a driver for neuroblastomas, this receptor is only implicated in a small subset of cases[11-15]. The neurotrophin receptor TRKB is frequently aberrantly expressed in neuroblastomas and has generated interest as a clinical target [16], however, whether this represents a true receptor tyrosine kinase driving tumor proliferation or is merely a marker for poorly differentiated neuroblasts remains to be determined, as no activating mutations have been identified. Thus, the majority of neuroblastomas are not driven by a known receptor tyrosine kinase. The downstream pathways are also intact; while neuroblastomas have been reported in children with Costello's syndrome, a disorder caused by germline mutations in HRAS[17], these cases are rare and mutation of any of the ras family members is extremely rare in spontaneous neuroblastomas[18]. Similarly, while neuroblastomas have been reported in patients with Noonan's syndrome[19], specifically those carrying mutations of PTPN11, a component of the ras signaling pathway, mutations in this gene are not observed in spontaneous neuroblastomas[20]. While neuroblastomas have been reported in neurofibromatosis cases carrying mutations in the ras-regulating tumor suppressor NF1, the frequency of these tumors is not significantly increased in these patients compared to the general population[21], and while the gene is mutated in neuroblastoma cell lines, it is rarely mutated in spontaneous tumors[22,23] . Thus, while case reports exist of neuroblastomas in children with inherited syndromes caused by mutations in the ras signaling pathway, these cases are exceedingly rare, and mutations in this pathway have been all but excluded from spontaneous primary neuroblastomas. No frequent mutations have been detected in the downstream MAP kinase pathway, nor have any been detected in the PI3 kinase/AKT pathway[23]. The Sonic Hedgehog, Wnt/β-catenin, and Notch signaling pathways are also not associated with neuroblastoma. Inherited mutations in the PHOX2B transcription factor causing congenital central hypoventilation syndrome and Hirschprung's disease have been associated with familial neuroblastomas[24], but the gene does not show mutation or loss of function in spontaneous neuroblastomas[25], limiting its role to cases of rare hereditary syndromes. The primary known molecular lesions strongly associated with spontaneous neuroblastoma are limited to MYCN amplification, ALK mutations, and numerous copy number aberrations for which several potential candidate driver genes exist.

Since most of the common signaling pathways in other tumors are not implicated in neuroblastoma, the molecular mechanisms driving the disease remain poorly defined, and the development of targeted therapeutics (or the use of agents developed for other diseases) has lagged, with a correspondingly modest change in overall survival over the last two decades. Thus, research in neuroblastoma is currently dominated by genetics and genomics strategies to identify novel genes and lay a foundation for further biochemical characterization of the disease.

We have utilized a mouse model for neuroblastoma in which expression of the human MYCN proto-oncogene is targeted to the neural crest via the tyrosine-hydroxylase promoter (as the rate-limiting enzyme in norepinephrine synthesis, the gene is expressed in the sympathetic peripheral nervous tissue that gives rise to neuroblastoma)[26]. Tumors arise in these mice with a median latency of 3 months, possibly preserving some biological aspect of the disease being exclusively pediatric in humans. The tumors display histological features of neurons, arise in thoracic and abdominal paraspinous locations, and secondarily invade the spinal cord; all hallmark features of human neuroblastomas. Since mice develop single, isolated, presumably clonal tumors, we hypothesized that tumor development required secondary genomic events in addition to overexpression of the MYCN oncogene. We used array comparative genomic hybridization (array CGH) to show that these murine neuroblastomas show secondary genomic events that parallel those seen in neuroblastoma[27]. Tumors frequently gain the syntenic region corresponding to human chromosome 17p, gained in 80% of human neuroblastomas (comparative alignment of sub-chromosomal gains was also able to narrow this region by several megabases). Similarly, murine tumors show genomic loss of the distal region of chromosome 4, which corresponds to human 1p, deleted in roughly one-third of human neuroblastomas, and correlating strongly with both MYCN amplification and poor outcome. Murine tumors also demonstrated recurrent combined loss of chromosomes 5, 9, and 16, which together contain regions corresponding to human 3p, 4p, and 11q, lost as a group in a subset of human neuroblastomas. Thus, the mouse neuroblastoma model faithfully recapitulates both the pathology and genetics of human neuroblastomas, and provides a resource for both preclinical testing and the exploration of secondary genomic events. We hypothesized that identifying genes cooperating with the MYCN transgene to drive tumors in our model would provide insight into the molecular pathways driving human neuroblastoma, and may identify novel therapeutic targets for the disease. We adopted two genetics strategies to identify these genes: a genetic linkage mapping approach in which we took advantage of a strain-specific tumor susceptibility in the model to identify a network of genes governing tumor susceptibility (described in Part I), and a forward-genetics approach utilizing transposon-based insertional mutagenesis (described in Part II).

In Part I, we used classical genetics to explore the causes of tumor susceptibility in our mouse model. After establishing that tumor susceptibility was complex (but was nevertheless likely caused by factors relevant to human neuroblastoma), we combined our classical genetics approach with state-of-the-art genomics approach that involved transcriptional profiling on an exon level. We used this approach to identify a possible interaction between the *Arg1* gene (liver arginase) and components of the GABA signaling network, two distinct molecular pathways coupled by both our genetic results and a biochemical synthetic/metabolic pathway. We then showed *in vitro* that inhibition of Arg1 can negatively impact growth of cell lines, suggesting the results of our genetic modifier screen have identified not only a gene, but a possible therapeutic strategy in neuroblastoma. These data are described in Chapter 2.

Our expression array analysis of the peripheral sympathetic nerve ganglia is, we believe, the most extensive transcriptional characterization of this small, diffuse tissue to date. In Chapter 3, we explored this dataset in more detail, using gene expression correlation networks to identify novel genetic interactions in peripheral nervous tissue unrelated to neuroblastoma.

To complement our classical genetics approach, we also attempted to integrate transposon-based insertional mutagenesis to identify genes in neuroblastoma, as well as in breast cancer. We were also involved in the fundamental characterization of the molecular biology of the system as we attempted to utilize it for our specific disease. These efforts are described in the remaining chapters, starting with Chapter 4, which provides further introduction for the forward mutagenesis technology used in Part II.

References

1. Schwab, M. et al. Enhanced expression of the human gene N-myc consequent to amplification of DNA may contribute to malignant progression of neuroblastoma. *Proc Natl Acad Sci U S A* **81**, 4940-4 (1984).
2. Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E. & Bishop, J.M. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science* **224**, 1121-4 (1984).
3. Fernandez, P.C. et al. Genomic targets of the human c-Myc protein. *Genes Dev* **17**, 1115-29 (2003).
4. Li, Z. et al. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci U S A* **100**, 8164-9 (2003).
5. Orian, A. et al. Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. *Genes Dev* **17**, 1101-14 (2003).
6. Patel, J.H., Loboda, A.P., Showe, M.K., Showe, L.C. & McMahon, S.B. Analysis of genomic targets reveals complex functions of MYC. *Nat Rev Cancer* **4**, 562-8 (2004).
7. Hogarty, M.D. The requirement for evasion of programmed cell death in neuroblastomas with MYCN amplification. *Cancer Lett* **197**, 173-9 (2003).
8. Tweddle, D.A. et al. The p53 pathway and its inactivation in neuroblastoma. *Cancer Lett* **197**, 93-8 (2003).
9. Carr, J. et al. Increased frequency of aberrations in the p53/MDM2/p14(ARF) pathway in neuroblastoma cell lines established at relapse. *Cancer Res* **66**, 2138-45 (2006).
10. Easton, J., Wei, T., Lahti, J.M. & Kidd, V.J. Disruption of the cyclin D/cyclin-dependent kinase/INK4/retinoblastoma protein regulatory pathway in human neuroblastoma. *Cancer Res* **58**, 2624-32 (1998).
11. Mosse, Y.P. et al. Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930-5 (2008).
12. George, R.E. et al. Activating mutations in ALK provide a therapeutic target in neuroblastoma. *Nature* **455**, 975-8 (2008).
13. Chen, Y. et al. Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **455**, 971-4 (2008).
14. Janoueix-Lerosey, I. et al. Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455**, 967-70 (2008).
15. Caren, H., Abel, F., Kogner, P. & Martinsson, T. High incidence of DNA mutations and gene amplifications of the ALK gene in advanced sporadic neuroblastoma tumours. *Biochem J* (2008).
16. Brodeur, G.M. et al. Trk receptor expression and inhibition in neuroblastomas. *Clin Cancer Res* **15**, 3244-50 (2009).
17. Gripp, K.W. Tumor predisposition in Costello syndrome. *Am J Med Genet C Semin Med Genet* **137C**, 72-7 (2005).
18. Moley, J.F. et al. Low frequency of ras gene mutations in neuroblastomas, pheochromocytomas, and medullary thyroid cancers. *Cancer Res* **51**, 1596-9 (1991).
19. Cotton, J.L. & Williams, R.G. Noonan syndrome and neuroblastoma. *Arch Pediatr Adolesc Med* **149**, 1280-1 (1995).

20. Bentires-Alj, M. et al. Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. *Cancer Res* **64**, 8816-20 (2004).
21. Brodeur, G.M. Neuroblastoma: biological insights into a clinical enigma. *Nat Rev Cancer* **3**, 203-16 (2003).
22. Holzel, M. et al. NF1 is a tumor suppressor in neuroblastoma that determines retinoic acid response and disease outcome. *Cell* **142**, 218-29.
23. Dam, V., Morgan, B.T., Mazanek, P. & Hogarty, M.D. Mutations in PIK3CA are infrequent in neuroblastoma. *BMC Cancer* **6**, 177 (2006).
24. Mosse, Y.P. et al. Germline PHOX2B mutation in hereditary neuroblastoma. *Am J Hum Genet* **75**, 727-30 (2004).
25. Raabe, E.H. et al. Prevalence and functional consequence of PHOX2B mutations in neuroblastoma. *Oncogene* **27**, 469-76 (2008).
26. Weiss, W.A., Aldape, K., Mohapatra, G., Feuerstein, B.G. & Bishop, J.M. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *EMBO J* **16**, 2985-95 (1997).
27. Hackett, C.S. et al. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-73 (2003).

Chapter 2: eQTL analysis implicates arginine metabolism and GABA signaling as modifiers of susceptibility and therapeutic targets in neuroblastoma

Source: The following chapter contains unpublished data from a manuscript in preparation.

Contributions: I performed essentially all of the experiments listed and wrote the manuscript in its current form. David Quigley performed some of the bioinformatics analysis, and provided software and guidance for much of the rest of the analysis. Christine Cheng assisted with some of the biochemical analysis. Young Song and Javed Khan supervised the exon array hybridizations, which I performed. Ludmila Pawlikowska, Denise Lind, and Pui Kwok supervised the SNP genotyping, all of which was performed by me (additional microsatellite genotyping was done by outside institutions as described in Methods). Yun Bao did the taqman analysis for transgene expression. David Goldenberg, Nadya Milshteyn, Slava Yakovenko, and Kim Nguyen assisted with maintenance of the mice used in these experiments. Terry Van Dyke assisted in the experimental design and provided reagents for the project. Jian-Hua Mao and Saunak Sen assisted with interpretation of the genotyping data. Allan Balmain provided critical advice on several aspects of the project. William Weiss designed, initiated, and supervised the project and helped to write the manuscript.

eQTL analysis implicates arginine metabolism and GABA signaling as modifiers of susceptibility and therapeutic targets in neuroblastoma

Christopher S Hackett, David Quigley, Christine Cheng, Young Song, Ludmila Pawlikowska, Denise Lind, Yun Bao, David Goldenberg, Nadya Milshteyn, Slava Yakovenko, Kim Nguyen, Jian-Hua Mao, Saunak Sen, Terry Van Dyke, Pui-Yan Kwok, Javed Khan, Allan Balmain, and William A. Weiss

Abstract

The neural crest is largely post-mitotic after early development. Thus, molecular pathways governing malignant proliferation of neural crest tissue in neuroblastoma are distinct from those driving other tumor types, and remain largely a mystery. We show that a locus on chromosome 10 governs tumor susceptibility in a transgenic mouse model for this disease. To identify candidate tumor susceptibility genes, we measured gene expression levels in superior cervical ganglia from backcrossed mice and used expression QTL analysis to correlate variation in expression levels with genotype for all genes. *Arg1* (liver arginase) was the chromosome 10 gene under the strongest control of an eQTL at the chromosome 10 susceptibility locus in neural crest derived tissue. Several interacting susceptibility loci also overlapped with eQTL controlling genes in the GABA neurotransmitter signaling pathway. Since *Arg1* lies in the GABA synthesis pathway, the correspondence between the functional linkage and the genetic interaction led us to investigate the role of *Arg1* in human neuroblastoma. Arginase inhibitors significantly decreased viability and growth of neuroblastoma cell lines, suggesting a role for *Arg1* in metabolism and growth signaling in neuroblastoma, and providing insights into neurodegeneration associated with arginemia syndromes. Our observations suggest *Arg1* as a novel neuroblastoma target, and provide a genetic link between the down-regulation of GABA receptors observed in high-risk neuroblastomas and the role of GABA receptors in control of neural crest growth

Introduction

Neuroblastoma, a pediatric tumor of the peripheral nervous system, arises from the embryonal neural crest. In contrast to most tumors of the nervous system that arise from glial cells, neuroblastoma arises from neurons, which terminally differentiate and remain post-mitotic shortly after birth. As such, regulation of proliferation is controlled by different mechanisms than in other cell types, and aberrations in canonical oncogenes found in myeloid, epithelial, and glial tumors are largely unaltered in neuroblastoma. For the most part, the molecular pathways controlling neuroblastoma proliferation remain a mystery.

Among the first genetic lesions linked to neuroblastoma was amplification of the *MYCN* proto-oncogene, a member of the c-myc transcription factor family overexpressed in a wide range of tumors. Amplification of *MYCN* is among the strongest predictors of poor outcome in neuroblastoma[1,2]. However, MYCN itself is a poor drug target, and since myc-family transcription factors have been shown to influence the expression levels of over 15% of the genes in the genome[3], elucidating the specific mechanism by which *MCYN* drives neuroblastoma has proven difficult. As a result, the development of targeted therapeutics for high-risk neuroblastoma has lagged, and improvements in clinical outcomes have been modest over the last several decades.

Mice expressing a MYCN transgene under the control of the rat tyrosine hydroxylase promoter (TH-MYCN) develop tumors with histological and genetic characteristics of human neuroblastoma[4,5], providing a model system to identify molecular pathways necessary for neuroblastoma progression. We observed that tumor incidence in this model was highly dependent on mouse strain background. Here, we took advantage of this phenomenon to map a susceptibility locus on chromosome 10. To identify candidate genes influencing tumor susceptibility, we then profiled superior cervical ganglia (SCG) from backcross mice using Affymetrix Exon arrays to identify expression quantitative trail loci (eQTL) mapping to our susceptibility region. We identified liver arginase (*Arg1*), a component of the urea cycle, as the strongest eQTL in the region, and observed that several interacting susceptibility loci in a 2-QTL model overlapped with multiple genes involved in GABA neurotransmitter signaling.

As *Arg1* is an early component of the GABA synthesis pathway, we hypothesized that differences in *Arg1* expression, combined with modulations in the GABA pathway, influenced tumor susceptibility. To investigate the role of *Arg1* activity in neuronal cells and to test its potential as a neuroblastoma therapeutic target, we treated a panel of neuroblastoma cell lines with the arginase inhibitor nor-NOHA. Treatment with this compound resulted in decreased cell viability and an accumulation of cells in G1. The genetic observation that expression levels of this gene correlate with tumor susceptibility *in vivo* combined with the biochemical observation that inhibition of *Arg1* inhibits cell growth suggests both that *Arg1* plays an important role in neuronal cell metabolism and that *Arg1* represents a therapeutic target for human neuroblastoma.

Results

Tumor penetrance is dependent on strain

Transgenic mice were generated on a Balb/c x C57B6/J background and showed a roughly 10% tumor incidence. As mice were crossed into strain FVB/NJ, tumor incidence decreased to zero by 2 generations (**Figure 2.6A**). Conversely, as mice were crossed into strain 129/SvJ, tumor penetrance steadily increased with each backcross generation, leveling out at roughly 60% incidence[6]. To eliminate the trivial possibility that the difference in tumor incidence was due to different levels of transgene expression between genetic backgrounds, we performed taqman analysis on brain, adrenal gland, and superior cervical ganglia from mice from each strain (**Figure 2.6B**). Expression levels were equivalent between strains in all tissues tested, with brain showing much higher expression levels than adrenal gland and sympathetic ganglia.

We then crossed resistant transgenic FVB/NJ mice to susceptible wild-type 129/SvJ. The resulting F1 mice showed a 4% tumor incidence (N=200), suggesting that tumor resistance was generally genetically dominant. Interestingly, the median age of tumor onset for these mice was 9 months, unusually long compared to the roughly 3-month median age of onset seen in all other genetic backgrounds.

To generate a genetically diverse population for linkage mapping, we then backcrossed transgenic F1 animals to wild-type 129/SvJ mice. The resulting N1 backcross generation showed a 38% tumor incidence, indicating that resistance did not segregate in a Mendelian ratio, suggesting the involvement of multiple susceptibility loci. However, the average age of onset for tumor-prone backcross mice (109 days) was identical to that of mice carrying the transgene in a pure 129/SvJ background[6], suggesting that the genetic variation was affecting tumor incidence but not progression.

Linkage analysis of 225 mice identifies a tumor susceptibility modifier on chromosome 10

To identify genomic loci associated with tumor susceptibility, we genotyped 199 mice using a combination of microsatellite and SNP markers (see Methods). Interval mapping analysis of this data identified a significant locus on chromosome 10 (LOD=4.3, **Figure 2.1A**). Saturation of the region with SNP markers identified SNP RS36323433 as being closest to the maximum LOD score (**Figure 2.1B**). However, the effect of alleles at this locus was opposite of expected: heterozygous mice were tumor prone (~60%), while mice homozygous for the 129/SvJ allele were resistant (~20% incidence), (**Figure 2.1C)**.

Additionally, this effect was strongly influenced by gender. Performing the analysis using sex as an interacting covariate increased the LOD at this locus from 4.3 to 4.9. When mice were segregated by sex and analyzed independently, this locus was not significant in female mice, though it showed the same significant LOD score in males (LOD=4.3, N=82) as in the group overall. The segregation of genotypes at the chromosome 10 locus does not explain tumor susceptibility in the parent strains. Additionally, this locus had no effect in females, and the tumor-prone genotype only had a 60% tumor incidence. Collectively, these data suggested complex genetics governing tumor susceptibility. We next performed a 2-QTL test to identify pairs of genetic loci acting either independently or together to influence tumor susceptibility. We observed several significant loci that interacted with the locus on chromosome 10, many with similar LOD scores (**Figure 2.2,** bottom right, **Table 2.1**). None of these secondary loci were significant using an independent or additive model however (**Figure 2.2, top left),** suggesting a biological interaction between genes at these alleles. Interestingly, while all of these loci interact with the chromosome 10 locus, they do not show significant interactions with each other.

Expression QTL analysis of ganglia and tumors identifies *Arg1* as a candidate modifier

The lod plot of the susceptibility locus on chromosome 10 was broad **(Figure 2.1B).** Dozens of genes fell within an acceptable confidence interval for the locus, complicating direct identification of the candidate gene. Since many heritable phenotypes are driven by variation in gene expression, we hypothesized that tumor susceptibility may be governed in part by differences in gene expression levels of a gene in our chromosome 10 locus. To compare expression levels of neural-crest derived tissue, we isolated superior cervical ganglia (SCG) from pure transgenic male and female 129/SvJ and FVB/NJ mice and profiled RNA expression using Affymetrix Mouse Exon arrays. A SAM[7] analysis with a conservative 0% FDR revealed, surprisingly, that only three genes were significantly differentially regulated between male and female ganglia (**Figure 2.7A**), all of which are located on the Y chromosome, suggesting an overall lack of sexual dimorphism in gene expression in the peripheral sympathetic nervous system. However, the same analysis revealed that roughly 260 genes significantly differentially expressed between strains (**Figure 2.7B**). We detected significant differential expression of several genes (the 10 genes with the most significant difference between strains are shown in **Figure 2.7C**). These genes included the putative oncogene *Eya4*[8], and the candidate breast cancer susceptibility genes *Echdc1* and *Rnf146*[9].

Since regulation of gene expression can be influenced by complex interactions of cis- and trans- acting factors, we next sought to identify the genetic loci responsible for gene expression variation. To achieve this, we analyzed SCG from 116 backcross animals using Affymetrix Mouse Exon Arrays, and treated expression levels of each gene as a quantitative trait (expression QTL or eQTL[10,11]). In backcross animals, the cis- and trans- acting alleles influencing gene expression that are linked in each purebred background are decoupled from each other, allowing us to identify the strongest factors influencing the expression differences between strains for each gene. We hoped to identify strong eQTL within our chromosome 10 locus, as well as at the numerous other secondary loci identified in the 2-QTL tumor susceptibility analysis, since genes with eQTL overlapping with physiological phenotypes have been shown to influence these phenotypes[11,12].

We performed association tests for gene-level expression calls for each gene on the arrays, using the markers for the susceptibility linkage analysis[13]. eQTL with the lowest p-values are given in **Table 2.2**. The *Arg1* gene (liver arginase) was the most significant gene with an eQTL at the chromosome 10 locus. We then validated this association using interval mapping. *Arg1* expression levels had a QTL at chromosome 10 with a LOD score of 17.4 (**Figure 2.3A**). *Arg1* overlapped directly with our tumor susceptibility locus, making *Arg1* our top candidate gene. Segregation of mice by genotype showed that mice heterozygous at that locus had almost 2-fold higher expression of *Arg1* compared to mice homozygous for the 129/SvJ allele (**Figure 2.3B**), consistent with patterns seen in the purebred parental strains (**Figure 2.7C**).

eQTL for GABA-related genes map to secondary susceptibility loci

We next analyzed eQTL mapping to secondary loci that interacted with the chromosome 10 locus to influence tumor susceptibility. We observed that genes related to GABA neurotransmitter signaling were located within multiple secondary susceptibility loci. Most notably, a trans-eQTL on chromosome 2 controlled expression of the *GABRA3* receptor subunit on the X chromosome (**Figure 2.4A-C**). This chromosome 2 locus overlapped directly with the secondary susceptibility locus on chromosome 2, the most significant secondary susceptibility locus in male mice (making the observation that the eQTL controls a gene on the X chromosome more intriguing). Mice harboring alleles resulting in high *Arg1* expression and low *GABRA3* expression were most susceptible to tumors (**Figure 2.4C**). We also noted an eQTL for the *GABRA5* receptor subunit mapping to a secondary susceptibility locus on chromosome 7 (**Figure 2.4D-F**) and an eQTL for the GABA transporter *Slc6a1* at the secondary susceptibility locus on chromosome 4, (**Figure 2.4G-I**). Encouragingly, downregulation of GABA-A receptors is a marker for poor prognosis in human neuroblastomas[14], consistent with our genetic observations.

We then investigated genes located near interacting secondary susceptibility loci that lacked obvious eQTL candidates. The locus on chromosome 1 centered near 73 cM (marker RS50560599, lod 7.83) is 6Mb from *Dbi* (diazepam binding inhibitor), a gene that modulates GABA receptor activity[15]. Similarly, the locus on chromosome 9 centered near 91 cM (marker D9MIT201, lod 7.8) is 4 Mb from the *Trak1* gene, which modulates GABA receptor homeostasis[16]. Finally, the locus on chromosome 17 centered near 38.4 cM (marker D17MIT231, lod 6.85) is 2.5 Mb from the GABA-B receptor 1. Together, at least 6 secondary susceptibility loci co-localized with genes in the GABA pathway, and/or eQTL controlling these genes (**Table 2.1**).

Inhibition of *Arg1* decreases viability of human neuroblastoma cells

Though arginase is typically associated with urea cycle in the liver, expression has been detected in other tissues including sympathetic ganglia[17]. In neurons, *Arg1* is part of the GABA synthesis pathway, producing ornithine, a precursor of glutamate, which is then a precursor of GABA. This biochemical link provides a possible explanation for the genetic interactions between the *Arg1* locus and the numerous secondary susceptibility loci harboring downstream components of the GABA signaling pathway. Since GABA, an inhibitory neurotransmitter, inhibits neuronal cell growth [18], the observation that mice with lower expression of GABA receptor subunits are more tumor prone when they also harbor alleles resulting in higher *Arg1* expression (a component of the GABA synthesis pathway) is not surprising. This may suggest that the observation that the down-regulation of GABA receptors associated with more aggressive human neuroblastomas may be biologically significant beyond being a marker for undifferentiated neuroblasts, though since the receptor can be composed of different combinations of multiple subunits susceptible to different inhibitors, testing this hypothesis is not straightforward.

However, this observation does not explain the primary role of increased *Arg1* expression in predisposing mice to tumors. Several downstream outputs could account for this (**Figure 2.8**). First, *Arg1* competes for the arginine substrate with the nitric oxide synthases (NOSs), and is thought to regulate NOS function through additional mechanisms (reviewed in[19]). Since NOS produces nitric oxide, a molecule that has been shown to inhibit the growth of both tumor cells (including neuroblastoma cell lines[20,21] and can directly antagonize the proliferative effects of MYCN in neurons[22], decreased NO levels as a result of inhibition of the NOS pathway by *Arg1* may have a pro-growth effect. Second, *Arg1* catalyzes the conversion of arginine to urea and ornithine, the substrate for polyamine synthesis. Polyamine production has been linked to tumorigenesis (reviewed in[23]) and plays a role in neural proliferation[24], and the rate-limiting synthetic enzyme, ornithine decarboxylase (ODC) is a well-established MYC transcriptional target[25]. ODC inhibitors inhibit neuroblastoma development in both cell lines and in the TH-MYCN transgenic model[26,27] and are currently in clinical trials for neuroblastoma (http://www.nmtrc.org/phase-i-dfmo/). Third, ornithine can also be converted to glutamate as part of the GABA synthetic pathway. Glutamate is both an excitatory neurotransmitter and a substrate for cellular metabolism; myc-driven tumor cell lines are specifically dependent on glutamate/glutamine metabolism (reviewed in[28]). Additionally, Arginase can act as an immunosuppressant[29,30], thus, increased expression may facilitate tumor formation *in vivo*. While the role of *Arg1* activity in cancer has not been extensively studied, encouragingly, inhibition of arginase has been shown to inhibit growth of a breast cancer cell line[31]

The higher-expressing Arginase allele conferring tumor susceptibility is nested within an overall resistant genetic background in purebred mice (either FVB/NJ or FVB/NJ x 129/SvJ F1), complicating strategies to test this specific allele in vivo. Given the diverse outputs by which Arginase could potentially promote tumor growth, we hypothesize that *Arg1* represents a therapeutic target, and that inhibition of *Arg1* will affect growth of human neuroblastoma. We detected low levels of *Arg1* expression by Western blot in every line in a panel of human neuroblastoma cell lines (**Figure 2.5A**). We then treated this panel with the arginase inhibitor nor-NOHA (N-Omega-hydroxy-nor-arginine) and measured cell viability using a WST-1 assay. Viability was decreased dramatically in chp126 cells, and more modestly but significantly in all cell lines tested near the IC50 dose of 10uM [32] (**Figure 2.5b**). We then tested cell cycle status using flow cytometry. After 24 hours, while no indications of apoptosis were present, we saw an accumulation of cells in G1 phase and a decreased S phase in both Kelly and SY5Y cells (**Figure 2.5C**.

Discussion

The development of the peripheral nervous system involves growth and differentiation regulatory pathways that are distinct from the mitogenic signaling cascades governing growth of epithelial and other cell types. Consequently, common genomic aberrations driving tumorigenesis in epithelial and glial tumors rarely show abnormalities in neuroblastoma. To identify novel pathways contributing to neuroblastoma development, we used a mouse model for the disease driven by the MYCN transcription factor, leveraging the observation that genetic background had a striking influence on tumor penetrance. While we expected simple genetics to govern penetrance for tumors, our linkage analysis revealed a strikingly complex network of secondary susceptibility loci, all with modest individual effects and similar LOD scores (reminiscent of recent human genome-wide association studies), all linked to a central locus on chromosome 10.

Many differences in physiological phenotypes between individuals result from differential gene expression and/or splicing[33], as opposed to coding mutations. Analysis of gene expression as a function of genotype has proven to be a powerful means to identify candidate genes for quantitative trait loci underlying physiological differences in several organisms[10-12]. Genetic control of gene expression (expression quantitative trait loci or eQTL) is more direct than genetic influence over complex physiological phenotypes, providing a very strong signal in linkage analysis calculations and directly implicating specific candidate genes among dozens or hundreds of genes at a quantitative trait locus.

Though the influence of strain background on tumor penetrance is frequently observed in mouse models of cancer, only a handful of genes underlying this susceptibility have been identified (e.g., [34-39]), mostly due to the limited resolution of classical quantitative trait linkage mapping. Recently, traditional QTL analysis has been combined with eQTL analysis to identify genes modifying tumor susceptibility in mouse models for cancer[13,40,41]. In this study, we have used this strategy to identify *Arg1* as a candidate neuroblastoma modifier gene, and to implicate an interaction between up regulation of arginase activity and down-regulation of the the GABA receptor system as a cooperating mechanism driving tumor susceptibility.

While expression levels of several genes at the chromosome 10 locus were different between the two parent strains, expression of the *Arg1* gene, encoding liver arginase, showed the strongest linkage to the genetic background at that locus, revealed by expression profiling of genetically heterogeneous backcrossed animals combined with expression QTL analysis. Implication of *Arg1* as the primary candidate at this locus was strengthened by the observation that genes involved in GABA neurotransmitter signaling co-localized with secondary tumor susceptibility loci that had an interacting (epistatic), but not additive or independent, interaction with the *Arg1* locus. Arginase is an early step in the GABA synthesis pathway, and downregulation of GABA receptors has been reported in high-risk neuroblastoma[14] (in agreement with expression patterns seen at our susceptibility loci, where lower expression correlates with higher tumor incidence). This functional linkage corresponding to our observed genetic interactions encouraged us to pursue *Arg1* as a candidate tumor susceptibility gene. This biochemical link also explains our unique genetic pattern of several secondary loci interacting with a single gene/locus. Only one cytoplasmic arginase gene exists (*Arg1*); the other mammalian arginase, Arg2, is mitochondrial and has an essentially non-redundant function. However, the GABA signaling pathway may be perturbed at several genomic loci, including several components of the GABA-A receptor, the genes for which are dispersed throughout the genome. Thus, in a genetic model involving perturbation of arginase and GABA signaling, many GABA-related genes could interact with the single *Arg1* gene on chromosome 10, with presumably equivalent effect.

In purebred mice, the *Arg1* susceptibility allele (FVB/NJ) is carried on an overall resistant genetic background, presumably due to numerous overriding resistance loci that individually were below the detection threshold of our screen. This made characterizing the specific alleles either *in vivo* or in cultured neurons from purebred mice problematic. To characterize the biochemistry of modulated arginase activity in neurons, as well as to test the possibility that *Arg1* may be a therapeutic target in human neuroblastoma, we turned to established neuroblastoma cell lines as a model system. We showed that treatment of several neuroblastoma cell lines with the competitive arginase inhibitor nor-NOHA decreased cell viability and induced a higher G1 fraction in these cells.

While identification of arginase as a candidate tumor susceptibility gene and therapeutic target was unexpected, the known functions of the pathways linked to arginine in neurons make the observation rational. *Arg1* competes for arginine with NOS, which produces inhibitory nitric oxide. Ornithine, one of the products of the reaction catalyzed by *Arg1*, is a substrate for polyamine synthesis, a process required by tumor cells and validated as a therapeutic target in MYCN driven tumors[26,27]. Ornithine can also be metabolized to glutamate, an excitatory neurotransmitter as well as an intermediate for cellular metabolism specific to myc-driven tumors[28]. Glutamate can also be used to synthesize GABA, an inhibitory neurotransmitter that also inhibits neuronal growth and promotes differentiation[18]. Additionally, arginase can act as an immunosuppressant[29,30], facilitating immune evasion by developing tumors *in vivo*. Thus, *Arg1* represents a central node connecting many diverse pathways that collectively contribute to growth of cells in the neuronal lineage.

The mechanisms governing growth of neuroblastomas remain poorly understood, but are clearly influenced by the pathways governing neuron growth and differentiation. Our results showing differential expression of GABA-A receptor subunits and a GABA transporter are consistent with this hypothesis, and additionally suggest that the down-regulation of GABA receptors in human neuroblastomas may have biological significance beyond serving as markers for the differentiation status of cells. In support of this model, GABA signaling activity has been shown to negatively regulate growth of neural crest stem cells[18].

The observation that many neuroblastomas disregulate the Trk neurotrophin receptors, and that expression patterns of specific receptors correlate with clinical outcome, led to the hypothesis that aberrant neurotrophic signaling (assumed to be mediated by the Trk receptors and their ligands NGF and BDNF) drove proliferation of neuroblasts much in the way mitogenic signaling systems drive other tumor types[42]. While specific functional validation of the Trk receptors in mediating this effect is lacking, our model is conceptually consistent with this hypothesis, suggesting that decreased signaling through the GABA pathway, and possibly nitric oxide pathway (rather than the Trk receptors), may promote neuroblastoma development.

Arginase expression has been previously linked to affects on neuronal growth. In humans, mutations in *Arg1* lead to arginemia, a urea cycle disorder distinguished from other urea cycle disorders by a milder phenotype and a unique neurodegeneration (OMIM 207800). While hepatic metabolic defects in arginemia are speculated to drive neurotoxic metabolic intermediates with associated neurodegeneration[43,44], our data, along with detection of *Arg1* expression in sympathetic ganglia and other neuronal tissues[17,45,46], suggests that *Arg1* may additionally play a role in cell-intrinsic central metabolism of neurons, as well as the growth signaling between neuroblasts. This insight may be used to test whether arginase represents a therapeutic target in neuroblastoma, a common pediatric cancer. This potential is all the more encouraging given that arginase inhibitors are under investigation for highly prevalent diseases such as hypertension[19], making clinical development likely.

Methods

Mice All mice were obtained from the Jackson Labs (Bar Harbor, ME), and were housed and treated following UCSF IACUC guidelines. Mice were sacked prior to palpable tumors reaching 2 cm in diameter. Tumor-negative backcross mice were followed until one year of age (the latest tumor was detected at 342 days). Superior cervical ganglia were surgically isolated and snap-frozen in liquid nitrogen. SCGs were isolated from the parental control groups at 21 days.

Taqman analysis of transgene expression: Taqman expression analysis was performed on 6 mice (3 female, 3 male) from each strain. Proprietary assays for human MYCN and controls L18 and mGUS were obtained from Applied Biosystems (Carlsbad, CA). MYCN relative to mGUS is shown in **Figure 2.6B**.

Genotyping DNA was isolated from spleen tissue using a proteinase K lysis followed by phenol chloroform extraction. Microsatellite marker genotyping was carried out by the Marshfield Clinic (Marshfield, WI), and CIDR (Baltimore, MD). SNP genotyping was performed using FP-TDI[47] and SNPStream[48]. Markers and map positions are shown in **Table 2.3**.

Linkage analysis Interval mapping was performed using the R-QTL[49] package in the R statistical language. Genotypes flagged as probable errors by R-QTL were discarded. The genetic map positions were determined using the physical map positions (NCBI 37/mm9), followed by re-estimation of the map using R-QTL, and likely mis-mapped markers were discarded. Linkage analysis was performed on a 1-cM grid. Genome-wide significance thresholds were determined by running 1000 permutations for each dataset. Interval analysis was performed using the binary mode of the “EM” model.

Expression Arrays RNA was isolated from tumors using a Trizol extraction (Invitrogen, Carlsbad, CA) followed by RNEasy cleanup (QIAGEN, Valencia, CA). RNA from superior cervical ganglia was isolated using only the RNEasy kit, as we found these buffers were more effective at disrupting the ganglia than Trizol. 1µg of RNA was used as a starting template for the RiboMinus rRNA subtraction (Invitrogen, Carlsbad, CA) followed by the ST labeling protocol (Affymetrix, Santa Clara, CA). Labeled samples were hybridized to Affymetrix Mouse Exon 1.0 arrays. Array quality control was performed using the Affymetrix Expression Console.

eQTL analysis Arrays were normalized using RMA in the XPS package (http://www.bioconductor.org/packages/2.6/bioc/html/xps.html). Gene level calls were exported and analyzed in combination with genotyping data using custom software as described[13].

WST-1 assay Neuroblastoma cell lines were grown in RPMI media with 10% serum and antibiotics with the exception of SK-N-BE(2) (DMEM/F12, 10% serum) and IMR-32 (DMEM, 10% serum plus non-essential amino acids). Nor-NOHA was obtained from Bachem (Torrance, CA) and dissolved in DMSO. Cells were grown in 24-well plates and treated for 72 hours prior to the addition of WST-1 reagent (20µl per 1ml media). Absorbance was read after 15 minutes.

FACS cells were plated on 6-well dishes and treated as indicated for 24 hours. They were then harvested, fixed in 70% ethanol for 30 minutes, then stained with a propidium iodide (PI) dye and RNase solution. All samples were analyzed on a FACSCaliber flow cytometer (Becton Dickenson, Franklin Lakes, NJ) and analyzed using ModFit (Verity Software, Topsham ME).

Table 2.1 Two-QTL significant loci and candidate genes.

Chromosomes	pos1 (cM)	pos2 (cM)	LOD(full)	Candidate GABA Gene at secondary locus
c3 :c10	37.4	31.8	9.64	
c1 :c10	73	28.8	7.83	Dbi
c9 :c10	91.2	28.8	7.8	Trak1
c4 :c10	75.3	28.8	7.41	Slc6a1
c10:c13	28.8	8.3	7.37	
c8 :c10	51.1	32.8	7.35	
c10:c10	50.8	80.8	7.08	
c2 :c10	143.4	28.8	7	GABRA3
c12:c16	22	5.5	6.95	
c10:c17	29.8	38.4	6.85	GABA-B receptor 1
c10:c16	32.8	73.5	6.77	
c5 :c10	22.1	29.8	6.75	
c10:c15	31.8	2.1	6.37	
c7 :c10	82.7	28.8	6.33	GABRA5
c10:c12	31.8	22	6.32	

Table 2.2 Most significant eQTL in superior cervical ganglia.

p_val	Gene Name	SNP ID	Perm_Pval	Mean_129/129	Mean_129/FVB	Mean_129/FVB
2.14E-41	Tekt2	D4MIT308	0	5.67014	7.02254	0
7.13E-38	4833420G17Rik	D13MIT78	0	5.70123	6.44677	0
4.00E-36	Mlh1	09.105.291	0	5.02635	5.53984	0
5.57E-34	Gabra3	D2MIT148	0	5.80347	4.52116	0
1.48E-33	Scg5	02.109.360	0	9.25013	10.11152	0
4.30E-31	6977796	D8MIT45	0	7.61935	8.26878	0
1.77E-29	6811260	D13Mit207	0	4.65916	5.10756	0
2.90E-27	Efcab2	01.183.109	0	3.88415	4.63681	0
5.24E-27	Gpt2	D8MIT45	0	7.8877	8.54262	0
1.56E-26	6754519	D1MIT507	0	6.92093	7.79818	0
1.24E-23	Gm13242	D4MIT232	0	2.81229	3.3271	0
3.34E-22	Tshr	D12MIT194	0	6.07144	6.73883	0
3.88E-22	Psmc3ip	11.104.430	0	5.35163	5.95617	0
4.51E-22	Abcb10	D8MIT42	0	6.9635	7.23479	0
3.09E-21	Arg1	RS29316281	0	7.22273	8.02275	0
3.72E-21	Nlrx1	D9MIT247	0	6.98073	7.47301	0
1.26E-20	Aif1	17.034.150	0	5.14008	5.5864	0
2.67E-20	1600012F09Rik	D13Mit207	0	8.94435	8.31097	0
2.68E-20	Cd59a	D2MIT100	0	8.07486	8.7785	0
2.94E-20	Ercc2	07.013.915	0	6.39644	6.70727	0
7.77E-20	Lancl3	RS33457262	0	7.54071	8.01208	8.32673
1.77E-19	Uevld	D7Mit232	0	6.01856	6.66468	0
1.98E-19	Fam103a1	D7MIT350	0	6.79998	7.45781	0
3.20E-19	Npl	D1Mit102	0	5.9148	6.33616	0

Table 2.3 Genotyping markers

marker	chr	bp
chr 1 start	1	1
RS32728630	1	9632792
01.021.731	1	21731000
D1MIT169	1	24019242
D1Mit374	1	34,704,629
D1MIT236	1	45323159
01.061.101	1	61101000
D1MIT24	1	74344887
D1MIT132	1	77029626
D1MIT215	1	78089506
D1MIT134	1	80146900
RS30388122	1	94920500
01.102.953	1	102953000
RS50560599	1	116681037
d1mit340	1	118443768
D1MIT1001	1	130875104
01.136.071	1	136071000
D1Mit102	1	149011738
D1MIT507	1	166884608
01.183.109	1	183109000
chr 1 end	1	197195432
chr 2 start	2	1
D2MIT1	2	3803361
02.021.696	2	21696000
D2MIT81	2	24611112
D2MIT296	2	31146564
D2Mit297	2	42427495
RS27953638	2	50041657
D2MIT61	2	60491107
RS28322831	2	71063776
D2MIT75	2	80385565
RS27416022	2	93628229
D2MIT100	2	106338207
02.109.360	2	109360000
D2Mit274	2	114149035
D2MIT395	2	119216354
RS27258455	2	129951321
RS27267095	2	136652019
RS27267029	2	136669427
RS27265584	2	137006172
D2MIT423	2	148,551,155
D2MIT285	2	152548742
D2MIT411	2	159277868
02.161.464	2	161464000

02.168.990	2	168990000
D2MIT113	2	172997610
D2MIT148	2	178729953
chr 2 end	2	181748087
chr 3 start	3	1
03.016.637	3	16637000
D3MIT304	3	21662328
D3Mit151	3	31429221
03.033.871	3	33871000
D3MIT6	3	48979600
D3MIT67	3	53240491
03.060.525	3	60525000
RS37321647	3	68043880
RS38010777	3	68044559
RS31036560	3	73719554
D3MIT98	3	86267428
D3MIT49	3	89318587
03.106.773	3	106773000
D3MIT57	3	115822396
D3MIT315	3	115833639
RS30160288	3	125981675
D3MIT256	3	136288958
D3MIT351	3	139536842
03.141.220	3	141220000
D3MIT147	3	148682795
D3Mit19	3	157,546,004
chr 3 end	3	159599783
chr 4 start	4	1
04.013.290	4	13290000
RS28262872	4	18026684
D4mit94	4	34193500
D4MIT196	4	39640797
d4mit238	4	45251231
04.053.650	4	53650000
D4MIT164	4	59496294
04.063.977	4	63977000
D4MIT132	4	70158914
D4MIT348	4	82651978
D4Mit166	4	93,441,561
04.098.998	4	98998000
RS27499066	4	114673522
RS27499062	4	114673821
D4MIT308	4	123663603
D4Mit203	4	129074322
04.133.005	4	133005000
D4MIT170	4	137887414
D4MIT232	4	144324343
D4MIT42	4	150413794

chr 4 end	4	155630120
chr 5 start	5	1
D5MIT123	5	6562182
05.018.430	5	18430000
D5MIT294	5	20869141
D5MIT348	5	24429184
D5MIT388	5	33634956
D5MIT352	5	35931828
05.038.809	5	38809000
05.049.898	5	49898000
d5mit233	5	52985474
D5MIT183	5	53625392
RS33623243	5	70546596
D5MIT309	5	80577282
RS33085156	5	90112330
D5MIT10	5	104479307
D5MIT239	5	107653442
d5mit158	5	115224168
D5MIT425	5	120141058
D5MIT95	5	125118214
05.132.979	5	132979000
D5MIT169	5	149692128
D5MIT143	5	151270472
chr 5 end	5	152537259
chr 6 start	6	1
06.016.672	6	16672000
RS50690369	6	19742046
RS51272439	6	19888102
RS49937148	6	22510745
06.036.921	6	36921000
D6Mit272	6	44362455
D6MIT274	6	48656151
d6mit123	6	56,781,170
06.057.998	6	57998000
RS30909511	6	83140362
06.095.876	6	95876000
D6MIT67	6	97717263
D6MIT328	6	112745127
d6mit366	6	115208472
D6MIT194	6	128131104
D6Mit14	6	145613015
chr 6 end	6	149517037
06.149.619	6	149619000
chr 7 start	7	1
07.013.915	7	13915000
07.017.531	7	17531000
D7MIT294	7	26998202
D7MIT267	7	29255706

D7MIT228	7	39798667
D7Mit232	7	52481441
07.056.455	7	56455000
D7MIT248	7	73,384,970
D7MIT350	7	83462274
07.088.976	7	88976000
RS32210051	7	99669474
RS36353338	7	112706514
RS32012407	7	113567948
RS32021248	7	113747158
07.122.234	7	122234000
d7mit109	7	136353406
D7MIT223	7	144419262
D7Mit259	7	144566894
chr 7 end	7	152524553
chr 8 start	8	1
D8MIT155	8	4976574
08.010.585	8	10585000
RS46877379	8	19254138
D8MIT94	8	32807593
D8MIT292	8	36253530
D8Mit191	8	36649302
08.046.718	8	46718000
D8MIT68	8	59883108
08.076.189	8	76189000
D8MIT346	8	85820244
D8MIT45	8	90195480
D8MIT242	8	104648705
D8MIT211	8	105606050
D8MIT47	8	109733298
D8MIT215	8	118746715
D8MIT42	8	129438407
chr 8 end	8	131738871
chr 9 start	9	1
D9MIT250	9	8394192
09.014.560	9	14560000
D9mit90	9	32250020
D9MIT247	9	36882578
D9MIT2	9	37144572
D9MIT285	9	40405563
09.046.588	9	46588000
D9MIT71	9	49951955
D9MIT248	9	58160696
D9MIT336	9	65375870
D9MIT107	9	73264400
D9MIT123	9	73328958
09.079.053	9	79053000
d9mit198	9	91079875

D9MIT24	9	103088551
D9MIT347	9	103115448
09.105.291	9	105291000
D9MIT212	9	108498489
D9MIT201	9	117284864
D9Mit18	9	120138143
D9MIT151	9	121326572
chr 9 end	9	124076172
chr 10 start	10	1
10.002.877	10	2877000
D10Mit123	10	9922711
RS38343005	10	11465792
RS33543047	10	12164362
RS29347557	10	12661713
RS29316898	10	15819840
RS38621064	10	17528671
RS29354311	10	17622431
RS29366730	10	17920176
RS33635595	10	18259087
RS29365246	10	19378741
RS29320979	10	19463275
RS29322393	10	19948509
RS29367295	10	23573544
RS33702022	10	24370362
RS29351336	10	24605158
RS29316281	10	25167321
RS29380418	10	27331665
RS36323433	10	28317634
RS37076985	10	28876470
RS33755224	10	30375003
RS36274062	10	31045127
RS36679837	10	31060665
RS37117129	10	31223706
RS13480581	10	38685357
RS29317824	10	40078017
RS29313239	10	40310336
RS29316185	10	40564059
RS29376554	10	40616211
RS29329200	10	40791505
RS37251794	10	41929010
RS39284379	10	42031801
D10MIT184	10	42057114
RS33837056	10	42287898

RS29330419	10	42417908
RS29363236	10	42525244
RS29325964	10	42582671
RS33849981	10	44004143
RS46067685	10	58432247
D10MIT20	10	66407765
RS36294294	10	66518574
D10MIT31	10	67651017
RS46745265	10	69258223
D10MIT117	10	86994909
d10Mit96	10	98986629
10.113.678	10	113678000
D10Mit14	10	118064252
chr 10 end	10	129993255
chr 11 start	11	1
D11MIT2	11	12218640
RS26845852	11	24370394
D11MIT186	11	35079152
D11MIT51	11	36235173
11.041.143	11	41143000
RS26969123	11	53430698
D11Mit4	11	68425452
D11MIT320	11	70769563
11.072.405	11	72405000
D11MIT285	11	89743879
D11MIT289	11	94696242
11.104.430	11	104430000
D11MIT214	11	114946561
chr 11 end	11	121843856
chr 12 start	12	1
12.007.977	12	7977000
D12MIT182	12	10896611
D12MIT60	12	35375187
12.039.760	12	39760000
D12Mit2	12	42516678
D12MIT285	12	55570751
12.065.348	12	65348000
D12MIT91	12	72661419
D12MIT143	12	80799115
D12MIT194	12	91693003
D12MIT7	12	104,133,530
chr 12 end	12	121257530
chr 13 start	13	1
13.013.314	13	13314000
D13Mit207	13	16225249
RS29514367	13	29499372
13.043.962	13	43962000

D13MIT250	13	56332265
D13MIT13	13	56491058
13.061.624	13	61624000
RS30012306	13	70428413
D13MIT125	13	81186348
13.096.920	13	96920000
D13MIT288	13	108597465
D13MIT213	13	109367904
D13MIT53	13	113415301
d13mit151	13	116672647
D13MIT78	13	119948098
chr 13 end	13	120284312
chr 14 start	14	1
14.008.937	14	8937000
D14MIT98	14	15316978
14.027.409	14	27409000
D14MIT174	14	30475989
14.042.462	14	42462000
D14Mit183	14	50932167
14.067.129	14	67129000
D14MIT39	14	67501372
RS31380922	14	78742431
D14MIT263	14	87808173
D14Mit194	14	92719394
14.095.016	14	95016000
RS31252045	14	111376384
chr 14 end	14	125194864
chr 15 start	15	1
D15MIT13	15	3410212
15.010.846	15	10846000
D15MIT252	15	22565133
15.028.723	15	28723000
15.046.034	15	46034000
D15MIT143	15	51983942
D15MIT103	15	63,603,880
D15MIT67	15	70031534
D15Mit107	15	84214263
D15MIT262	15	87108377
15.088.295	15	88295000
15.090.122	15	90122000
D15MIT44	15	98949317
D15MIT15	15	102821148
chr 15 end	15	103494974
chr 16 start	16	1
D16Mit131	16	7234363
16.010.089	16	10089000
RS4164914	16	15586358
RS4165334	16	23467678

D16MIT60	16	32623838
16.039.061	16	39061000
D16Mit125	16	42296786
RS4187006	16	51575793
D16MIT185	16	60,376,811
D16MIT139	16	65587777
16.065.697	16	65697000
D16MIT188	16	76700185
D16MIT189	16	82416680
16.083.701	16	83701000
chr 16 end	16	98319150
chr 17 start	17	1
17.013.500	17	13500000
D17Mit213	17	16319811
17.021.019	17	21019000
D17MIT231	17	34143405
17.034.150	17	34150000
D17MIT51	17	42969224
D17MIT180	17	50896880
D17MIT20	17	56912782
17.059.041	17	59041000
D17Mit152	17	65240226
D17Mit93	17	73705538
D17MIT76	17	85542217
17.086.091	17	86091000
chr 17 end	17	95272651
chr 18 start	18	1
D18MIT222	18	14730527
D18Mit68	18	21578635
18.038.678	18	38678000
D18MIT202	18	43517266
D18MIT194	18	43786158
D18Mit123	18	56095974
D18MIT208	18	60,985,661
D18MIT152	18	62062136
18.063.800	18	63800000
D18MIT186	18	72145787
RS30267686	18	81658329
chr 18 end	18	90772031
chr 19 start	19	1
19.000.325	19	325000
D19Mit68	19	3645155
19.009.231	19	9231000
19.013.429	19	13429000
D19MIT96	19	21908690
D19MIT13	19	32705020
D19MIT46	19	33001204
D19MIT88	19	37322059

19.046.444	19	46444000
D19MIT103	19	53817478
chr 19 end	19	61342430
chr X start	X	1
X.054.837	X	54837
RS33457262	X	9226705
RS33477935	X	9574173
RS33478059	X	9632292
RS33625666	X	12120156
DXMIT68	X	49567950
DXMit119	X	68662966
RS29086361	X	95902327
DXMIT172	X	118200412
DXMit79	X	126210217
DXMIT132	X	137003182
DXMit216	X	139148521
RS29300656	X	153071529
chr X end	X	166650296

References

1. Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E. & Bishop, J.M. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science* **224**, 1121-4 (1984).
2. Schwab, M. et al. Chromosome localization in normal human cells and neuroblastomas of a gene related to c-myc. *Nature* **308**, 288-91 (1984).
3. Patel, J.H., Loboda, A.P., Showe, M.K., Showe, L.C. & McMahon, S.B. Analysis of genomic targets reveals complex functions of MYC. *Nat Rev Cancer* **4**, 562-8 (2004).
4. Weiss, W.A., Aldape, K., Mohapatra, G., Feuerstein, B.G. & Bishop, J.M. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *EMBO J* **16**, 2985-95 (1997).
5. Hackett, C.S. et al. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-73 (2003).
6. Chesler, L. et al. Malignant progression and blockade of angiogenesis in a murine transgenic model of neuroblastoma. *Cancer Res* **67**, 9435-42 (2007).
7. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001).
8. Miller, S.J. et al. Inhibition of Eyes Absent Homolog 4 expression induces malignant peripheral nerve sheath tumor necrosis. *Oncogene* **29**, 368-79.
9. Menachem, T.D., Laitman, Y., Kaufman, B. & Friedman, E. The RNF146 and ECHDC1 genes as candidates for inherited breast and ovarian cancer in Jewish Ashkenazi women. *Fam Cancer* **8**, 399-402 (2009).
10. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752-5 (2002).
11. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
12. Yang, X. et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet* **41**, 415-23 (2009).
13. Quigley, D.A. et al. Genetic architecture of mouse skin inflammation and tumour susceptibility. *Nature* **458**, 505-8 (2009).
14. Roberts, S.S. et al. GABAergic system gene expression predicts clinical outcome in patients with neuroblastoma. *J Clin Oncol* **22**, 4127-34 (2004).
15. Gray, P.W., Glaister, D., Seeburg, P.H., Guidotti, A. & Costa, E. Cloning and expression of cDNA for human diazepam binding inhibitor, a natural ligand of an allosteric regulatory site of the gamma-aminobutyric acid type A receptor. *Proc Natl Acad Sci U S A* **83**, 7547-51 (1986).
16. Gilbert, S.L. et al. Trak1 mutation disrupts GABA(A) receptor homeostasis in hypertonic mice. *Nat Genet* **38**, 245-50 (2006).
17. Yu, H. et al. Widespread expression of arginase I in mouse tissues. Biochemical and physiological implications. *J Histochem Cytochem* **51**, 1151-60 (2003).
18. Andang, M. et al. Histone H2AX-dependent GABA(A) receptor regulation of stem cell proliferation. *Nature* **451**, 460-4 (2008).

19. Durante, W., Johnson, F.K. & Johnson, R.A. Arginase: a critical regulator of nitric oxide synthesis and vascular function. *Clin Exp Pharmacol Physiol* **34**, 906-11 (2007).

20. Jenkins, D.C. et al. Roles of nitric oxide in tumor growth. *Proc Natl Acad Sci U S A* **92**, 4392-6 (1995).

21. Ortiz-Ortiz, M.A. et al. Nitric oxide-mediated toxicity in paraquat-exposed SH-SY5Y cells: a protective role of 7-nitroindazole. *Neurotox Res* **16**, 160-73 (2009).

22. Ciani, E., Severi, S., Contestabile, A. & Bartesaghi, R. Nitric oxide negatively regulates proliferation and promotes neuronal differentiation through N-Myc downregulation. *J Cell Sci* **117**, 4727-37 (2004).

23. Gerner, E.W. & Meyskens, F.L., Jr. Polyamines and cancer: old molecules, new understanding. *Nat Rev Cancer* **4**, 781-92 (2004).

24. Huang, Y., Higginson, D.S., Hester, L., Park, M.H. & Snyder, S.H. Neuronal growth and survival mediated by eIF5A, a polyamine-modified translation initiation factor. *Proc Natl Acad Sci U S A* **104**, 4194-9 (2007).

25. Bello-Fernandez, C., Packham, G. & Cleveland, J.L. The ornithine decarboxylase gene is a transcriptional target of c-Myc. *Proc Natl Acad Sci U S A* **90**, 7804-8 (1993).

26. Koomoa, D.L., Yco, L.P., Borsics, T., Wallick, C.J. & Bachmann, A.S. Ornithine decarboxylase inhibition by alpha-difluoromethylornithine activates opposing signaling pathways via phosphorylation of both Akt/protein kinase B and p27Kip1 in neuroblastoma. *Cancer Res* **68**, 9825-31 (2008).

27. Hogarty, M.D. et al. ODC1 is a critical determinant of MYCN oncogenesis and a therapeutic target in neuroblastoma. *Cancer Res* **68**, 9735-45 (2008).

28. Dang, C.V. Rethinking the Warburg effect with Myc micromanaging glutamine metabolism. *Cancer Res* **70**, 859-62.

29. Yachimovich-Cohen, N., Even-Ram, S., Shufaro, Y., Rachmilewitz, J. & Reubinoff, B. Human embryonic stem cells suppress T cell responses via arginase I-dependent mechanism. *J Immunol* **184**, 1300-8.

30. Bak, S.P., Alonso, A., Turk, M.J. & Berwin, B. Murine ovarian cancer vascular leukocytes require arginase-1 activity for T cell suppression. *Mol Immunol* **46**, 258-68 (2008).

31. Singh, R., Pervin, S., Karimi, A., Cederbaum, S. & Chaudhuri, G. Arginase activity in human breast cancer cell lines: N(omega)-hydroxy-L-arginine selectively inhibits cell proliferation and induces apoptosis in MDA-MB-468 cells. *Cancer Res* **60**, 3305-12 (2000).

32. Tenu, J.P. et al. Effects of the new arginase inhibitor N(omega)-hydroxy-nor-L-arginine on NO synthase activity in murine macrophages. *Nitric Oxide* **3**, 427-38 (1999).

33. Kwan, T. et al. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40**, 225-31 (2008).

34. MacPhee, M. et al. The secretory phospholipase A2 gene is a candidate for the Mom1 locus, a major modifier of ApcMin-induced intestinal neoplasia. *Cell* **81**, 957-66 (1995).

35. Ewart-Toland, A. et al. Identification of Stk6/STK15 as a candidate low-penetrance tumor-susceptibility gene in mouse and human. *Nat Genet* **34**, 403-12 (2003).

36. Mao, J.H. et al. Genetic variants of Tgfb1 act as context-dependent modifiers of mouse skin tumor susceptibility. *Proc Natl Acad Sci U S A* **103**, 8125-30 (2006).

37. Park, Y.G. et al. Sipa1 is a candidate for underlying the metastasis efficiency modifier locus Mtes1. *Nat Genet* **37**, 1055-62 (2005).

38. Wakabayashi, Y., Mao, J.H., Brown, K., Girardi, M. & Balmain, A. Promotion of Hras-induced squamous carcinomas by a polymorphic variant of the Patched gene in FVB mice. *Nature* **445**, 761-5 (2007).
39. Crawford, N.P. et al. Bromodomain 4 activation predicts breast cancer survival. *Proc Natl Acad Sci U S A* **105**, 6380-5 (2008).
40. La Merrill, M., Gordon, R.R., Hunter, K.W., Threadgill, D.W. & Pomp, D. Dietary fat alters pulmonary metastasis of mammary cancers through cancer autonomous and non-autonomous changes in gene expression. *Clin Exp Metastasis* **27**, 107-16.
41. Crawford, N.P. et al. The Diasporin Pathway: a tumor progression-related transcriptional network that predicts breast cancer survival. *Clin Exp Metastasis* **25**, 357-69 (2008).
42. Brodeur, G.M. et al. Trk receptor expression and inhibition in neuroblastomas. *Clin Cancer Res* **15**, 3244-50 (2009).
43. De Deyn, P.P., Marescau, B. & Macdonald, R.L. Guanidino compounds that are increased in hyperargininemia inhibit GABA and glycine responses on mouse neurons in cell culture. *Epilepsy Res* **8**, 134-41 (1991).
44. Deignan, J.L. et al. Increased plasma and tissue guanidino compounds in a mouse model of hyperargininemia. *Mol Genet Metab* **93**, 172-8 (2008).
45. Yu, H. et al. Arginase expression in mouse embryonic development. *Mech Dev* **115**, 151-5 (2002).
46. Yu, H. et al. Expression of arginase isozymes in mouse brain. *J Neurosci Res* **66**, 406-22 (2001).
47. Hsu, T.M., Chen, X., Duan, S., Miller, R.D. & Kwok, P.Y. Universal SNP genotyping assay with fluorescence polarization detection. *Biotechniques* **31**, 560, 562, 564-8, passim (2001).
48. Bell, P.A. et al. SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Biotechniques* **Suppl**, 70-2, 74, 76-7 (2002).
49. Broman, K.W., Wu, H., Sen, S. & Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-90 (2003).

Figure 2.1. A locus on chromosome 10 is linked to tumor susceptibility. (A) LOD plot for tumor susceptibility shows a single significant locus on chromosome 10. Dotted line indicates genome-wide significance threshold (LOD=2.81, 1000 permutations). (B) LOD plot of chromosome 10 only. Hashmarks on the horizontal axis indicate marker positions. Dotted line indicates genome-wide significance threshold (LOD=2.81, 1000 permutations). **(C)** Effectplot for the marker closest to the maximum LOD score (RS36323433) showing tumor incidence vs genotype. Heterozygous mice show a higher tumor incidence than mice homozygous for the 129/SvJ allele.

Fig 2.1

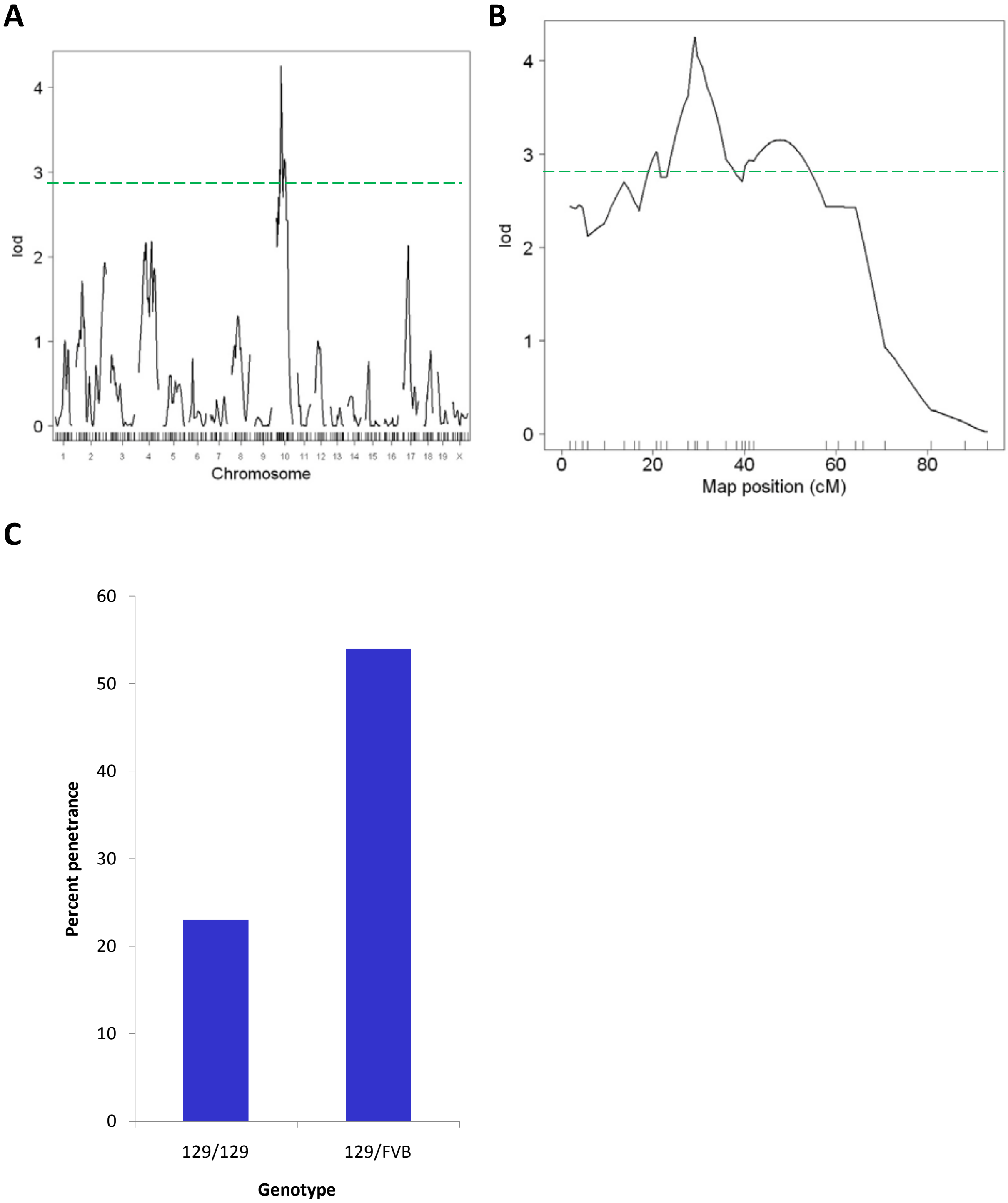

A
4
3
2
1
0
lod
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
B
4
3
2
1
0
lod
0
20
40
60
80
Map position (cM)
C
60
50
40
30
20
10
0
Percent penetrance
129/129
129/FVB
Genotype

Figure 2.2. 2-way QTL analysis reveals multiple loci interacting with a locus on chromosome 10 to influence tumor susceptibility. Top right indicates additive model (no interactions), with no significant loci. Bottom left shows “full” model accounting for epistatic interactions. By convention, LOD scores over 6 are considered statistically significant (see **Table 2.1**).

Fig 2.2

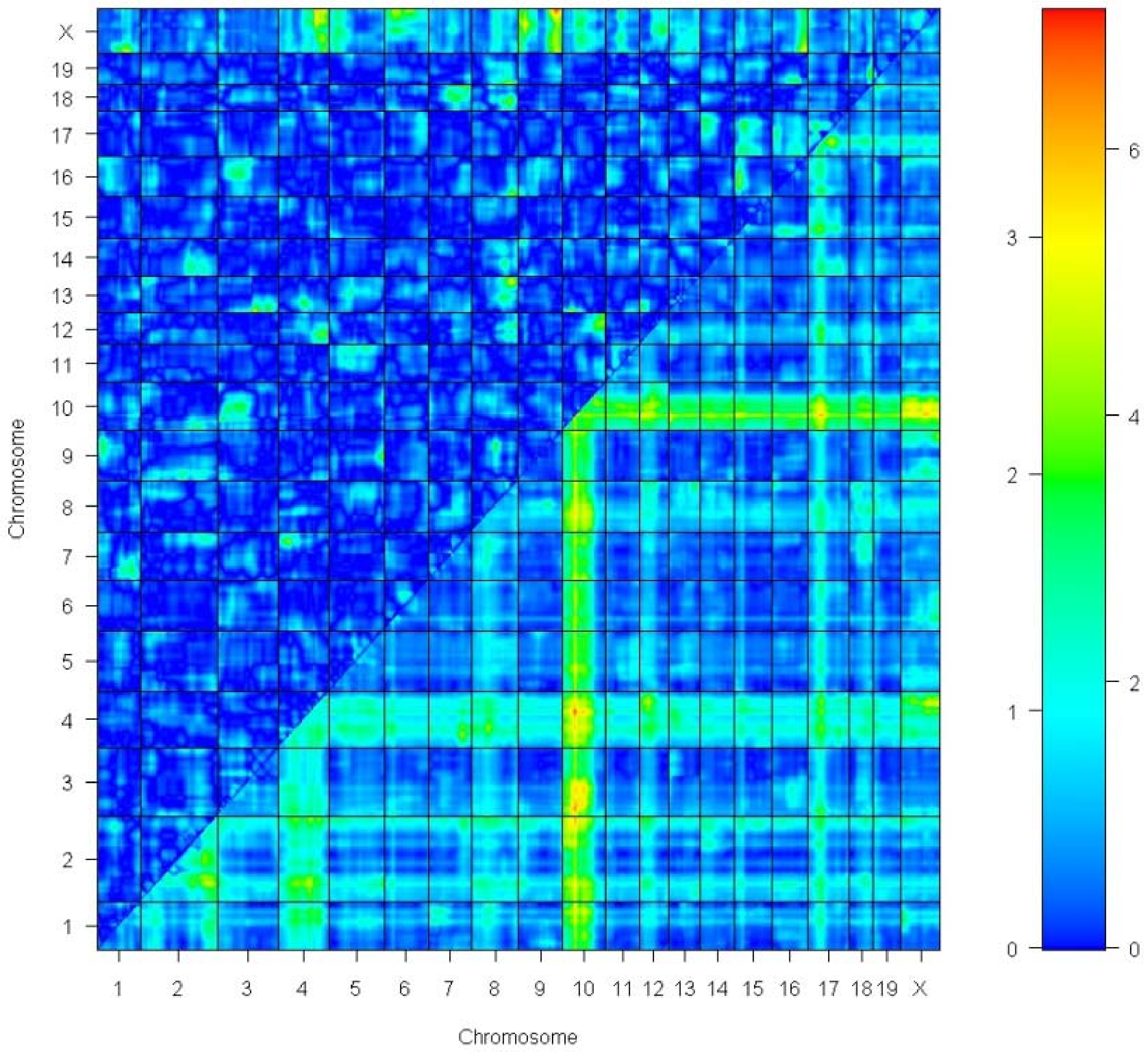

X
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
3
2
1
0
6
4
2
0

Figure 2.3. An eQTL for Arg1 colocalizes with the tumor susceptibility locus on chromosome 10. (A) Interval mapping for *Arg1* expression, the most significant eQTL in the chromosome 10 region, showing a LOD score of 17.4 on chromosome 10, centered at the physical location of the *Arg1* gene. **(B)** Log_2 Expression level vs genotype plot shows that tumor-prone heterozygous mice have almost 2-fold higher expression than mice homozygous for the 129/SvJ allele.

Fig 2.3

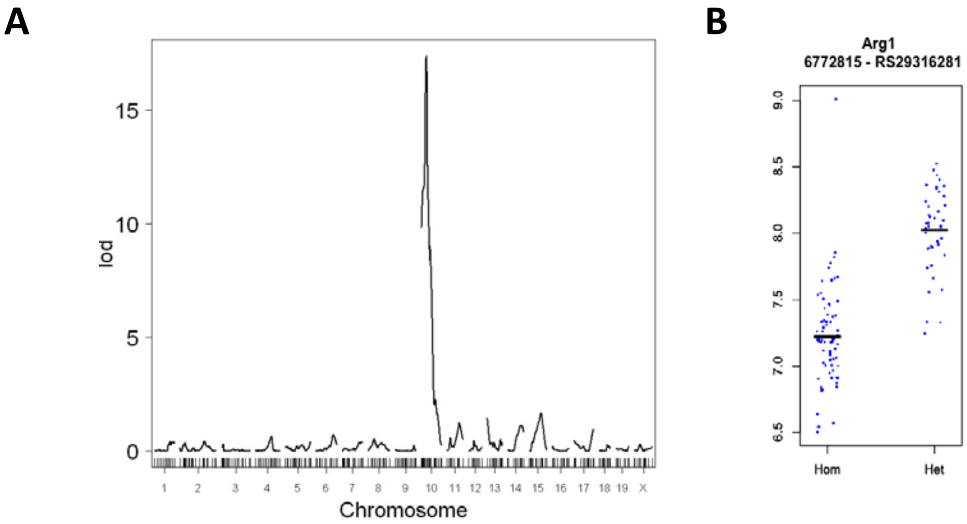

A
15
10
5
0
lod
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
B
Arg1
6772815 - RS29316281
9.0
8.5
8.0
7.5
7.0
6.5
Hom
Het

Figure 2.4. Secondary susceptibility loci co-localize with eQTL for GABA-related genes. **(A)** A trans-eQTL on chromosome 2 controls expression of the GABA-A receptor subunit 3 (*GABRA3*) on the X chromosome (LOD=30.9). **(B)** Expression vs genotype analysis homozygous mice have higher expression of the receptor subunit than heterozygous mice. **(C)** Plot of tumor incidence as a function of genotype at the chromosome 2 and 10 loci. Chromosome 10 homozygotes (tumor resistant) are largely unaffected by the genotype of the locus on chromosome 2. In contrast, among chromosome 10 heterozygous (tumor-prone) mice, the genotype at the chromosome 2 locus has a significant effect on tumor incidence, with doubly-heterozygous mice (high *Arg1* expression, low *GABRA3* expression) showing the highest tumor penetrance. **(D)** An expression QTL at the GABA-A receptor subunit 5 (*GABRA5*) locus (LOD=7.1) **(E)** Homozygous mice have higher expression of the gene than heterozygous mice. **(F)** Tumor susceptibility as a function of genotype shows a modest increase in tumor incidence among doubly-heterozygous mice vs. mice homozygous for the 129 allele on chromosome 7 but heterozygous at the locus on chromosome 10. 2 **(G)** Expression QTL for the *slc6a1* GABA transporter on chromosome 4 (LOD=4.1). Arrow indicates the physical location of *slc6a1* on chromosome 6. **(H)** Homozygous mice have lower expression of the gene than heterozygous mice. **(I)** Tumor incidence as a function of genotypes at the chromosome 10 and 4 loci. Mice with low expression of slc6a1 and high *Arg1* expression are the most tumor prone; patterns show a non-additive effect on tumor incidence at this locus.

Fig 2.4

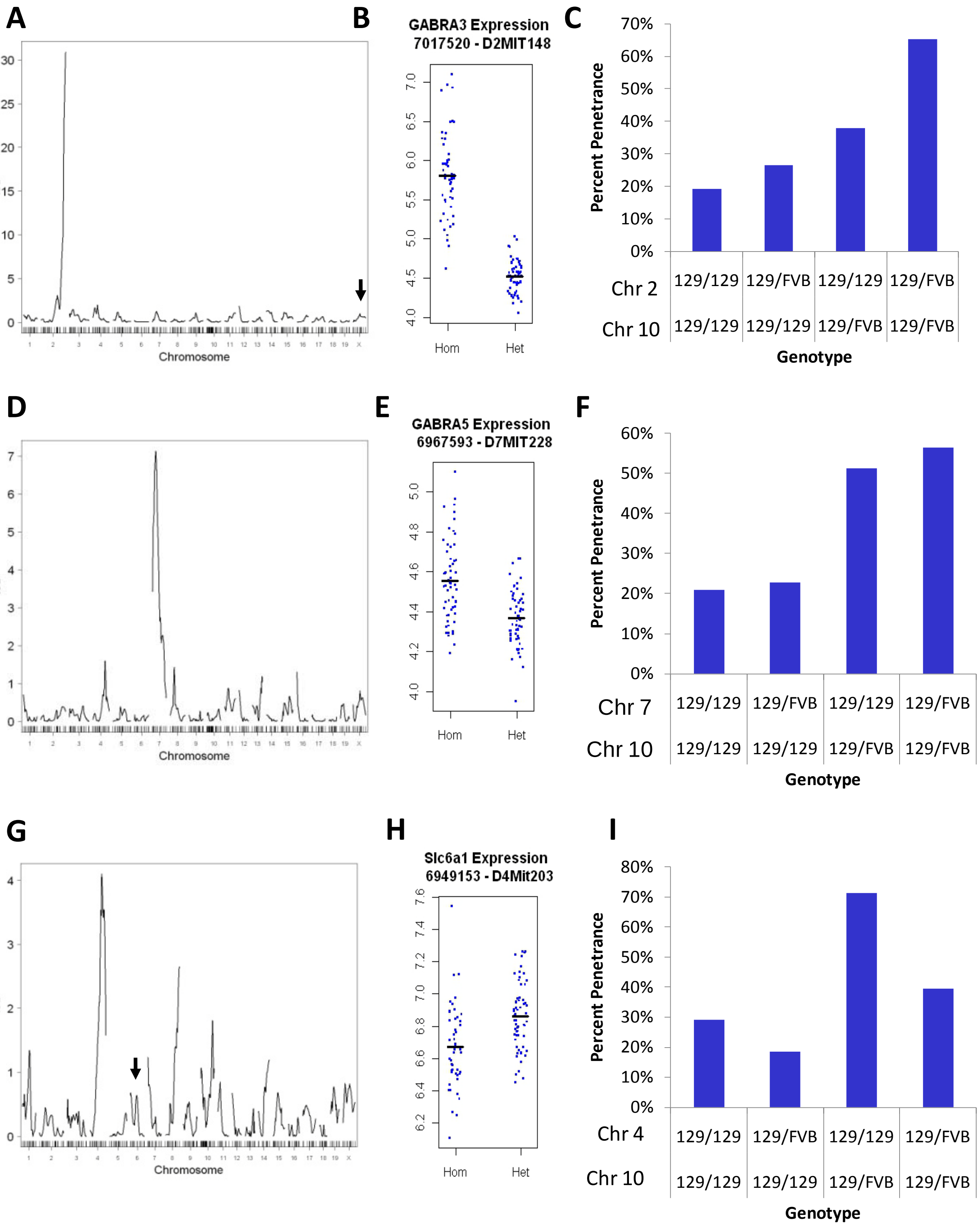

A
lod
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Chromosome
B
GABRA3 Expression
7017520 - D2MIT148
7.0
6.5
6.0
5.5
5.0
4.5
4.0
Hom
Het
C
Percent Penetrance
70%
60%
50%
40%
30%
20%
10%
0%
Chr 2
129/129
129/FVB
129/129
129/FVB
Chr 10
129/129
129/129
129/FVB
129/FVB
Genotype
D
lod
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Chromosome
E
GABRA5 Expression
6967593 - D7MIT228
5.0
4.8
4.6
4.4
4.2
4.0
Hom
Het
F
Percent Penetrance
60%
50%
40%
30%
20%
10%
0%
Chr 7
129/129
129/FVB
129/129
129/FVB
Chr 10
129/129
129/129
129/FVB
129/FVB
Genotype
G
lod
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Chromosome
H
Slc6a1 Expression
6949153 - D4Mit203
7.6
7.4
7.2
7.0
6.8
6.6
6.4
6.2
Hom
Het
I
Percent Penetrance
80%
70%
60%
50%
40%
30%
20%
10%
0%
Chr 4
129/129
129/FVB
129/129
129/FVB
Chr 10
129/129
129/129
129/FVB
129/FVB
Genotype

Figure 2.5. Inhibition of *Arg1* in neuroblastoma cells decreases cell viability and inhibits growth. (A) Western blot for *Arg1* in neuroblastoma cell lines. 293T cells transduced with viral constructs expressing either human or mouse *Arg1* are loaded as controls. **(B)** WST-1 assay showing a dose-dependent decrease in viability following treatment of neuroblastoma cells with varying doses of nor-NOHA for 72 hours. **(C)** FACS plot showing increased G1 phase in Kelly and SY5Y cells treated with nor-NOHA for 24 hours.

Fig 2.5

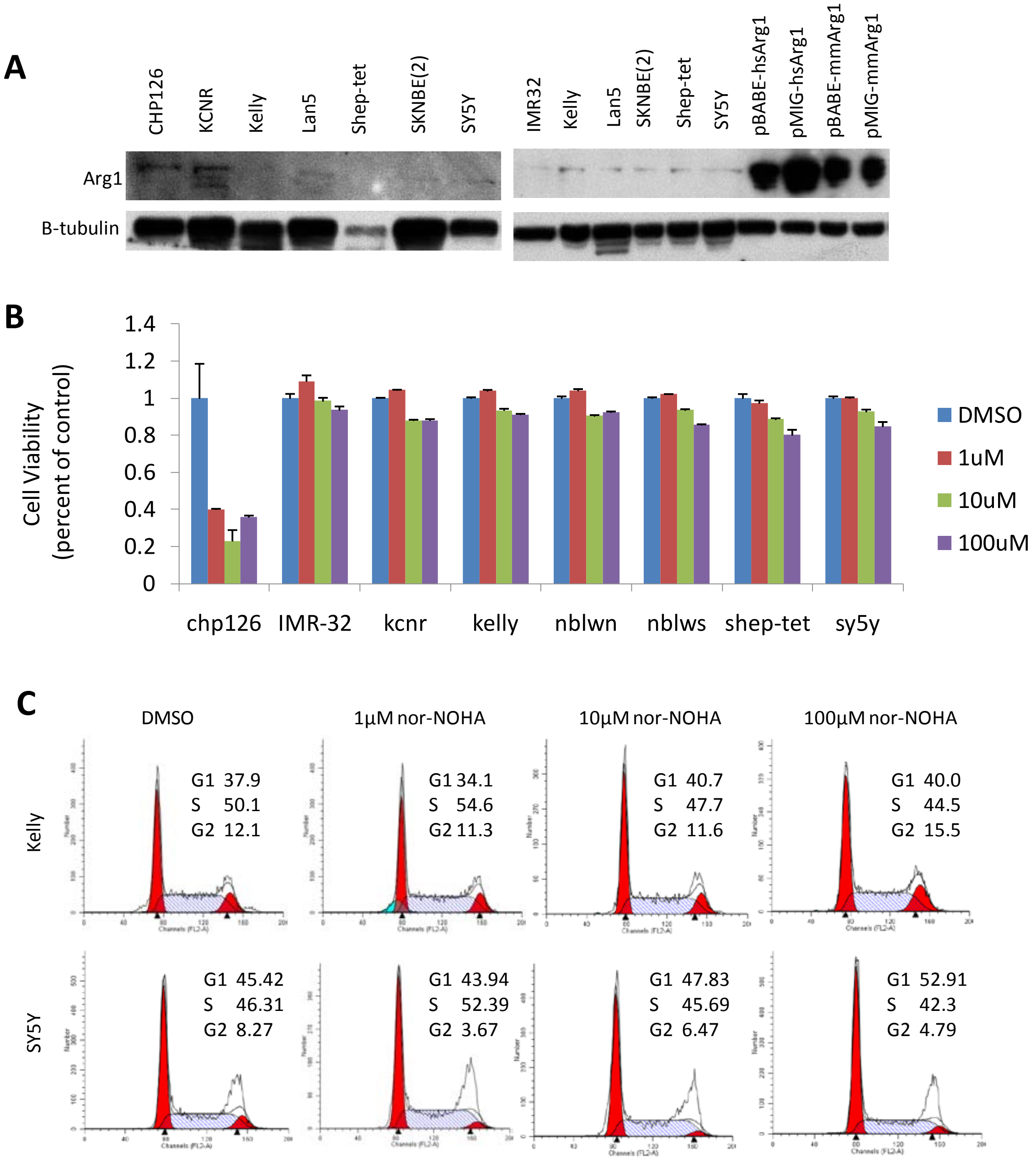

A
CHP126
KCNR
Kelly
Lan5
Shep-tet
SKNBE(2)
SY5Y
IMR32
Kelly
Lan5
SKNBE(2)
Shep-tet
SY5Y
pBABE-hsArg1
pMIG-hsArg1
pBABE-mmArg1
pMIG-mmArg1
Arg1
B-tubulin
B
Cell Viability
(percent of control)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
chp126
IMR-32
kcnr
kelly
nblwn
nblws
shep-tet
sy5y
DMSO
1uM
10uM
100uM
C
DMSO
1μM nor-NOHA
10μM nor-NOHA
100μM nor-NOHA
Kelly
G1 37.9
S 50.1
G2 12.1
G1 34.1
S 54.6
G2 11.3
G1 40.7
S 47.7
G2 11.6
G1 40.0
S 44.5
G2 15.5
SY5Y
G1 45.42
S 46.31
G2 8.27
G1 43.94
S 52.39
G2 3.67
G1 47.83
S 45.69
G2 6.47
G1 52.91
S 42.3
G2 4.79

Figure 2.6. Neuroblastoma incidence is dependent on strain. (A) Incidence of tumors in FVB/NJ and 129/SvJ as a function of backcross number. **(B)** Taqman expression showing equivalent levels of transgene expression in all tissues tested.

Fig 2.6

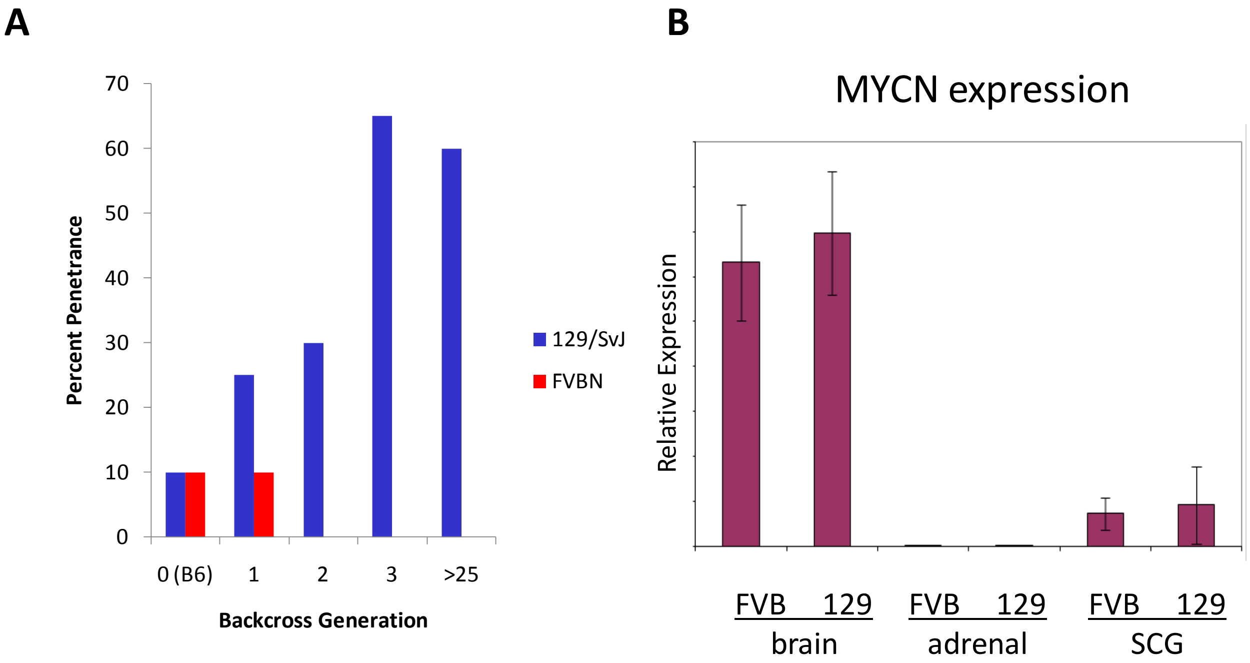

A
70
60
50
40
30
20
10
0
Percent Penetrance
0 (B6)
1
2
3
>25
Backcross Generation
🟦 129/SvJ
🟥 FVBN
B
MYCN expression
Relative Expression
FVB 129
brain
FVB 129
adrenal
FVB 129
SCG

Figure 2.7. Expression analysis of superior cervical ganglia from pure strains reveals differences in gene expression as a function of strain but not sex. (A) SAM analysis showing 3 genes differentially regulated between male and female FVB/NJ mice (N=5 in each group) **(B)** SAM analysis showing 260 genes differentially expressed between strains in male ganglia **(C)** Top 10 most differentially expressed genes in the chromosome 10 locus (**Figure 1B**) as measured in pure-bred transgenic 129/SvJ and FVB/NJ mice (N=5 in each group), as well as tumors arising in 129/SvJ animals (N=4).

Fig 2.7

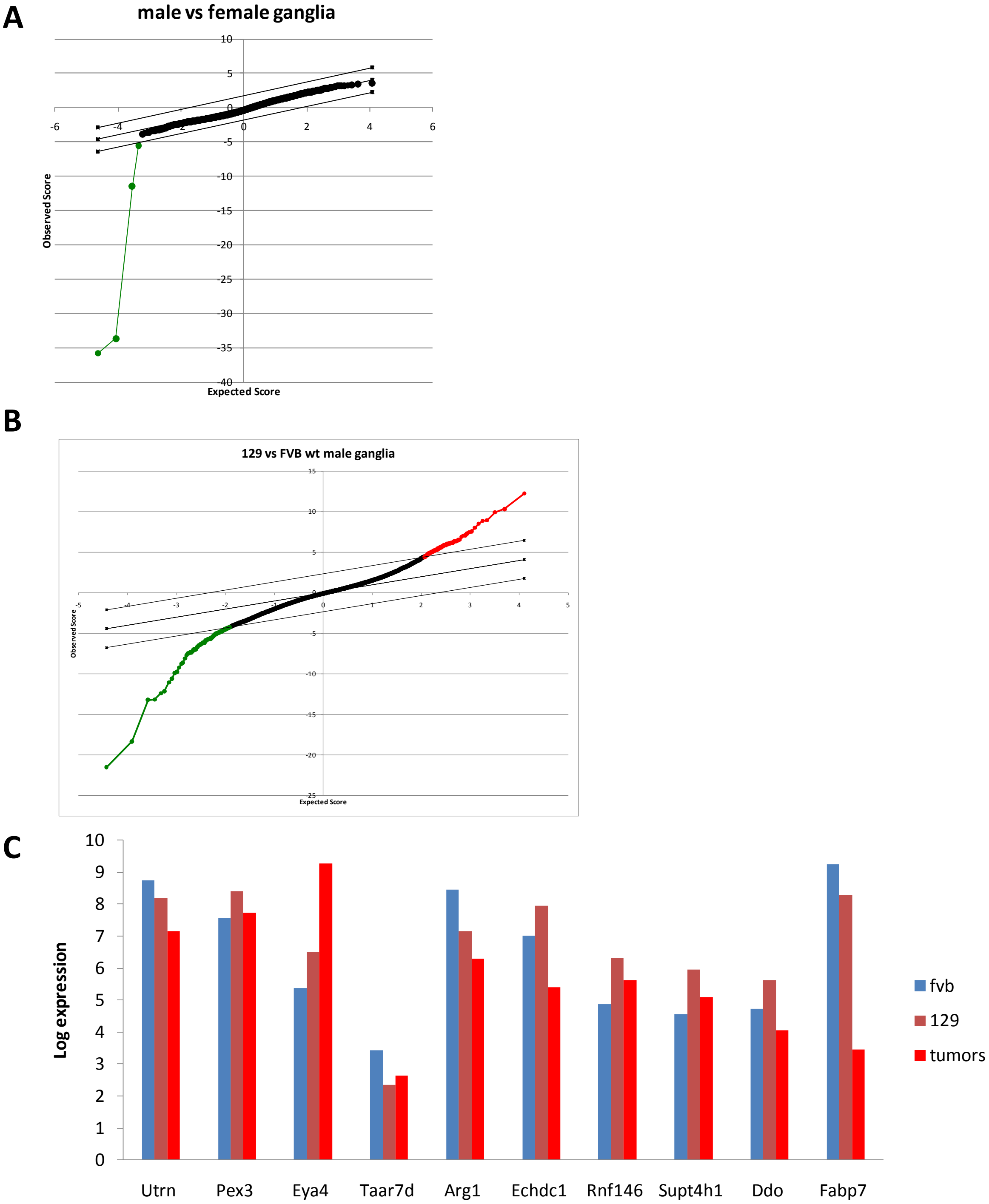
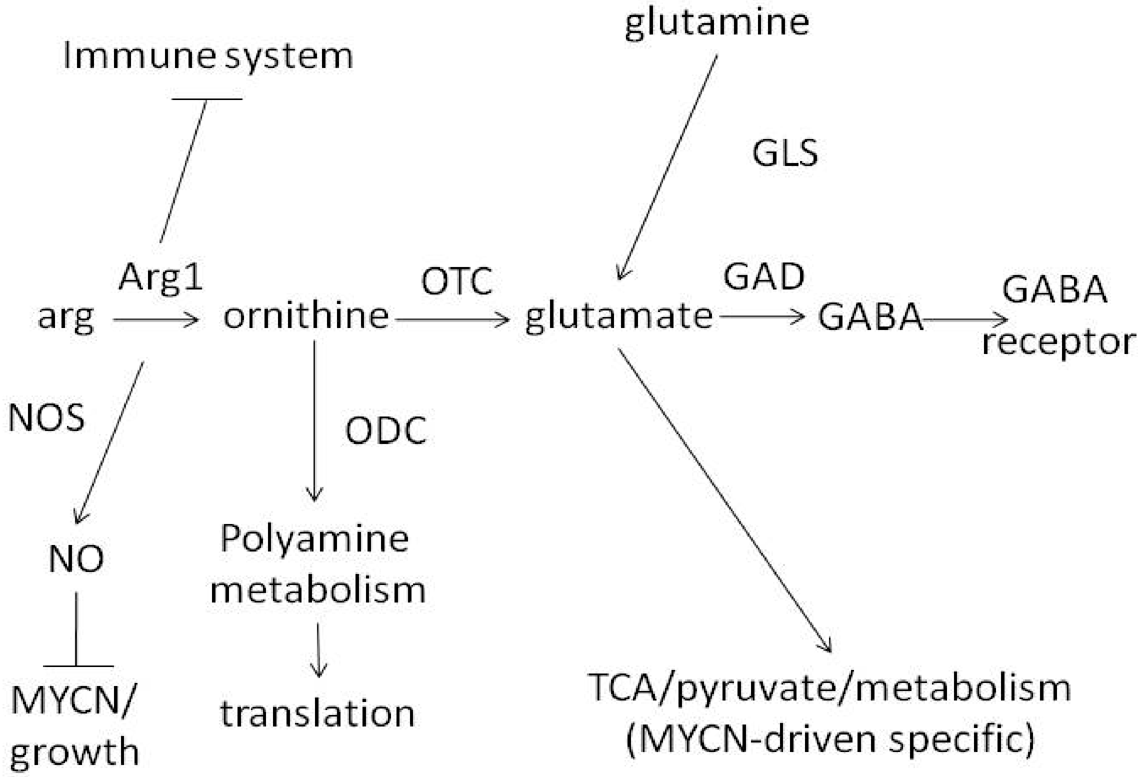

A
male vs female ganglia
10
5
0
-5
-10
-15
-20
-25
-30
-35
-40
-6
-4
-2
0
2
4
6
Observed Score
Expected Score
B
129 vs FVB wt male ganglia
15
10
5
0
-5
-10
-15
-20
-25
-5
-4
-3
-2
-1
0
1
2
3
4
5
Observed Score
Expected Score
C
10
9
8
7
6
5
4
3
2
1
0
Log expression
Utrn
Pex3
Eya4
Taar7d
Arg1
Echdc1
Rnf146
Supt4h1
Ddo
Fabp7
fvb
129
tumors

Figure 2.8. Arg1 is a node with several potential pro-growth outputs, but upstream of the inhibitory GABA pathway.

Fig 2.8


Immune system
glutamine
GLS
Arg1
arg
ornithine
OTC
glutamate
GAD
GABA
GABA
receptor
NOS
ODC
NO
Polyamine
metabolism
MYCN/
growth
translation
TCA/pyruvate/metabolism
(MYCN-driven specific)

Chapter 3: A gene co-expression network for the peripheral nervous system identifies candidate genes for inherited sensory syndromes and ciliary diseases.

Source: The following chapter contains unpublished data.

Contributions: I performed the data acquisition and primary analysis. David Quigley performed data analysis, supervised data analysis, and provided software, expert advice, and direction. Young Song assisted with array data acquisition. Terry Van Dyke provided reagents and expert advice. Javed Khan provided expert advice and supervised the array data acquisition. Larry Lustig provided expert advice on hearing. Allan Balmain provided expert advice on concepts and experimental design. William A. Weiss supervised the project.

A gene co-expression network for the peripheral nervous system identifies candidate genes for inherited sensory syndromes and ciliary diseases.

Christopher S. Hackett, David Quigley, Young Song, Terry Van Dyke, Javed Khan, Larry Lustig, Allan Balmain, and William A. Weiss

Abstract

The molecular biology of the peripheral sympathetic nervous system has been relatively poorly characterized due to the small, diffuse nature of the tissue. Here we have utilized a dataset generated to identify tumor susceptibility genes to explore the genetic aspects of gene expression in peripheral nerves. Using exon array data from 117 superior cervical ganglia and 46 neuroblastoma tumors from genetically heterogeneous mice, we have built gene correlation networks encompassing the majority of the genes in the genome. We used this network to identify novel genetic interactions with genes involved with neuroblastoma and neurofibromatosis. As the dataset was generated from a relatively pure population of neurons and Schwann cells, we could also use this network to explore genetic interactions related to other neuronal cell types, including cerebellar Purkinje cells. We were also able to use our network to suggest candidate genes for hereditary hearing loss syndromes. Finally, our network revealed uncharacterized genes involved in global regulation of expression and splicing of thousands of genes, illustrating that the network could be used to interrogate basic biology as well as human disease.

Introduction

The peripheral nervous system is a complex, diffuse organ that has remained relatively poorly characterized on the molecular level due to the lack of accessible, pure source material for study. The sympathetic nervous system is a component of the autonomic nervous system that mediates the body’s response to stress by innervating numerous target organs and controlling physiological responses. As such, the sympathetic system influences numerous biological processes (e.g. blood pressure, heart and breathing rate) and can be involved in several diseases. Additionally, the neuronal precursors of the sympathetic nervous system are the cells of origin for neuroblastoma, and the supporting Schwann cells give rise to the usually-benign peripheral nerve sheath tumors (schwannomas and neurofibromas). Together, the anatomical and physiological significance of the system, as well as its role in a number of tumor types, including an often deadly pediatric tumor, make it an important tissue for study.

Unfortunately, the small and diffuse nature of the sympathetic nervous system make it difficult to study on a molecular level. Since the neurons are mostly post-mitotic, *in vitro* systems are essentially limited to small, short-lived primary cultures, and cultures of immortalized neuroblastoma cell lines. *In vivo*, peripheral sympathetic nerves synapse with central nerves in the brain stem and spinal cord and carry signals to the major organs. The sympathetic system interfaces with the endocrine system in the adrenal medulla, the inner layer of the adrenal gland. The cell bodies of the peripheral sympathetic nerves are housed in the extra-spinal sympathetic ganglia, the largest of which are the superior cervical ganglia (SCG, **Figure 3.1**), composed of the neurons that innervate the face.

In this study, we have isolated SCG from 117 genetically heterogeneous mice and measured transcriptional levels at the exon level for all genes using Affymetrix exon arrays. The genetic heterogeneity provided subtle variations in gene regulatory systems, allowing us to identify several groups of independently co-regulated genes using gene correlation analysis. We performed the same analysis on 46 murine neuroblastomas, identifying several gene networks unique to tumors. We show several applications of this dataset, including prediction of function for specific genes, and identification of candidate genes for hereditary hearing loss.

Results

Gene Co-expression networks from sympathetic ganglia and tumors are highly connected.

As described in Chapter 2, we performed Affymetrix exon array analysis on 117 mouse SCG isolated from TH-MYCN[1] transgenic mice. Mice were a heterogeneous backcross from an F1 (129/SvJ x FVB/NJ) crossed to 129/SvJ. Due to meiotic recombination in the F1 germ cells, the resulting backcross mice could be heterozygous or homozygous for the 129/SvJ alleles at every locus. Genome wide, each mouse was on average 50% homozygous and 50% heterozygous across all autosomes, and population wide, for any given autosomal locus, 50% of mice were heterozygous and 50% were homozygous. This genetic heterogeneity provided a subtle perturbation in the genetic regulation of gene expression in vivo that allowed us to identify groups of genes under common transcriptional control.

Sample isolation and array hybridization were described in Chapter 2. It should be noted that while the small nature of the ganglia (**Figure 3.1A**) presented difficulties for sample isolation, preservation, and yield, this small size also made the tissue particularly homogenous compared to larger tissues. As shown in **Figure 3.1B**, ganglia are composed of mainly neurons and Schwann cells, with little vascularization or stroma. By comparison, the neuroblastomas were complex, heterogeneous, and highly vascularized (for illustrations, see[1,2]).

We constructed gene correlation networks for both SCG and tumors. To calculate correlation networks, the expression patterns for each gene across all samples were compared to patterns of all other genes using a Spearman rank correlation (to provide robustness against outliers). This calculation produced a rho value, with values approaching 1 suggesting greater similarity in gene expression patterns. Genome-wide error rates were calculated to establish a rho cutoff using 1000 permutations (see[3]). For SCG, a highly-conservative significance cutoff of 0.01% corresponds to a rho value of 0.617 (any gene-gene connection with a greater rho values is considered significant). Using a cutoff of 0.7 (higher than the minimum cutoff), our SCG network contains 12,801 genes connected by 8,272,081 edges (connections). Similar analysis of the tumor arrays produces a network of 14,468 genes connected by 13,905,067 edges. The most highly-correlated genes are shown in **Figure 3.2**. For SCG (**Figure 3.2A**), a rho cutoff of 0.9 was used, producing a network of 2557 genes connected by 177432 edges, revealing totally distinct regulatory groups that showed significant enrichment for specific functional groups (for example the large group of olfactory receptors, and a cluster of muscle/actin genes). The tumor network (**Figure 3.2B**), set at a higher cutoff (rho=0.95) shows a much more uniform, centralized “hairball”, with seven individual satellite genes connected to a large number of genes in the network. An independent cluster of striated muscle and collagen genes appears at the top middle of the network diagram. In total, this network contains 1,251 genes connected by 28,405 edges at this cutoff.

Network connections to the Bard1 gene suggest novel molecular functions.

To illustrate the utility of gene coexpression analysis, we analyzed the *Bard1* gene, which was implicated in neuroblastoma by a genome-wide association study[4]. In our tumor network, using a rho of 0.8, the Bard1 gene is correlated with 42 other genes (**Figure 3.3**). *Bard1* is a *BRCA1*-associated gene involved in the DNA damage repair system. The gene coexpression network shows connections to several other genes involved in DNA damage and repair (e.g. *Check2*, *Rad51*, *Rad54*, and the Fanconi’s anemia genes *Fanci* and *Fancc*, DNA excision repair genes *Ercc6l* and *Xrcc6*). The network also contains, somewhat unexpectedly, several mitotic genes (*Zwilch*, *Ttk*, *Bub1*, *MTBP*, *Esco2*, and the kinesins *Kif20b, Kif11*, *Kif15*), suggesting common gene regulation governing both DNA repair and mitosis (in the same direction). The network also contains ubiquitin ligases (*Ube2t*, *Fbxo5*), as well as the DNA methyltransferase *Dnmt1*, suggesting a functional linkage with these genes. Interestingly*, Atg9a*, a gene involved in autophagy, has an inverse correlation (red edges) with many of the genes in this network, suggesting an inverse relationship between this gene and the processes listed above. Together, this network illustrated the connection between gene regulation and molecular function, as well as the utility of this analysis to reveal novel molecular connections.

A gene correlation network for the NF1 gene reveals known and potentially novel functional interactions.

We next examined the network for NF1, a tumor suppressor in the ras signaling pathway responsible for the inherited syndrome neurofibromatosis, one of the most common diseases of the peripheral nervous system. Again using a cutoff of rho=0.8, 112 genes are significantly correlated with NF1 (**Figure 3.4**). Of the top 10 genes most significantly correlated with NF1 by gene coexpression analysis, five have a known functional connection to the gene (*Bmpr2, Nrxn1*, *Dst, Htt,* and *Kif1b*), again illustrating the connection between high-level gene co-expression correlation and functional relationships. Interestingly*, Apba3*, an amyloid beta interacting protein and candidate Alzheimer's gene, has a negative association with *Nf1* and several genes in the network, suggesting a possible relationship between two genes involved neuronal disease.

A gene correlation network for the cis-eQTL-controlled GRID2 orphan glutamate receptor reveals connections illuminating its molecular function.

We next attempted to go beyond functional relationships to infer mechanisms. As described in Chapter 2, we have genotyped all of the mice used in the microarray study and performed eQTL analysis to identify genetic linkage to control of gene expression for every gene in the genome. Genes under the control of a detectable cis-eQTL (that is, control of expression is linked to the location of the gene) are thought to be primarily under local genetic control. When these genes are wired into a gene co-expression network (meaning these genes have highly similar expression patterns and are thus under the same transcriptional control), the inference can be made that the cis-eQTL controlled gene then controls the expression of other genes in the network.

The *Grid2* gene (also called *GluD2*) is an orphan glutamate receptor for which the ligand, Cbln1, was only recently discovered[5,6]. *Grid2* is thought to play a role in neuronal apoptosis, as evidenced by a gain-of-function point mutation in this gene in “lurcher” mice that causes apoptosis of cerebellar Purkinje cells and ataxia in heterozygotes, and death soon after birth in homozygotes due to massive loss of multiple neuronal populations in the brain[7]. While the role of this gene in neuronal apoptosis is well established, the exact mechanism remains a mystery.

We identified *Grid2* as a gene with a significant cis-eQTL that was also connected to numerous genes in the gene co-expression network. **Figure 3.5** shows the *Grid2* gene coexpression network. This network was build using a rho cutoff of 0.7 that, while still significant, is lower than the previous networks. We observed a general trend that genes under the control of eQTL (thus, expression levels were under genetic control) tended to have fewer connections, and the connections they did have tended to have lower rho values.

The *Grid2* network displays an interestingly complex balance of positive and negative correlations. Interestingly, the *Apba3* amyloid-beta interacting gene from the *Nf1* network is similarly inversely correlated with the *Grid2* network. As evidenced by the numerous yellow edges on the right side of the network diagram, this gene has a complex inverse interaction with a number of genes. The overall network is rich in various ion channels, signal transduction molecules, neuronal cell adhesion molecules, and mitotic genes, each presenting testable models for the mechanism by which *Grid2* mediates neuronal apoptosis.

A network of hearing associated genes reveals candidate hereditary hearing loss genes.

In analyzing the network, we frequently noticed a tightly-correlated cluster of unrecognized genes that investigation revealed to be involved in ear formation and hearing. This observation, in conjunction with the observation of a tightly-clustered group of olfactory receptors in the gene correlation network (**Figure 3.2**), was surprising, considering our source material was isolated from paraspinal nerve ganglia in the neck. We speculate that “leaky” transcriptional control leads to low, but detectable, levels of expression for genes in the sensory systems that are usually expressed in distinct, but related, neuronal cell types. Since the inner ear is no more accessible experimentally than the peripheral sympathetic nervous system, we sought to take advantage of this observation to explore novel genetic and molecular aspects of hearing. In particular, numerous genomic loci have been linked to hereditary hearing loss (http://hereditaryhearingloss.org/), with causal genes/mutations linked to only a subset of loci. We sought to determine whether genes wired with known hearing genes in our gene coexpression network may co-localize with any of the hereditary hearing loss loci. There are 66 genes associated with hearing in the Gene Ontology (GO) database (http://www.geneontology.org/, see Methods), of which 12 were present in our network and wired to at least one other gene. At a rho of 0.95, 223 genes are connected with this group of 12 seed genes (**Figure 3.6A**). Interestingly, near the center of the network is the gene *Nphs1*. The gene encodes the nephrin protein, which is involved in kidney function. Mutations in this gene cause congenital nephrosis syndrome (OMIM 256300). While this gene is not identified as a hearing gene by GO, patients with congenital nephrosis can present with hearing defects, along with other neurological deficits[8]. Thus, our hearing network is capable of identifying genes involved in hearing loss not currently classified as hearing genes by GO. We next matched the locations of the genes in our network with loci for hereditary hearing loss for which no genes have been identified. **Table 3.1** shows the hereditary hearing loss loci with STS marker positions (note that marker positions do not necessarily denote the central peak of the LOD score, and, as evidenced by the genomic cytoband location, some loci are quiet large). We have listed genes from our SCG hearing coexpression network mapping within roughly 15 Mb of hearing loss loci, signifying novel candidate hearing loss genes. We noticed several trends from this pool of genes. Like *Nphs1*, the *Slc34a3* gene (chromosome 9) plays a role in kidney biology. *Abca4* (chromosome 1) is involved in photoreceptor signaling; photoreceptor degeneration and hearing loss are linked as diseases involving dysfunctional cilia, and the two processes are caused by many of the known hereditary hearing loss syndrome genes[9]. The *Zan/zonadhesin* gene (chromosome 7), a gene primarily expressed in sperm (which share a possible flagellar/ciliary trait with photoreceptors and components of the ear), and involved in zona pellucid binding during fertilization, is wired closely into the center of our network (**Figure 3.6B**). Interestingly, mutations in the zonadhesin-like domain of *Tecta*, another gene in our network, causes hereditary hearing loss[10]. We also noted that while several collagen and myosin genes have been linked with hereditary hearing loss, several novel members of these families are present in our network, and co-localize with hereditary hearing loss loci. Thus, the collection of novel candidate hereditary hearing loss genes derived from our network represent genes from both known and novel functional groups with respect to existing hearing loss genes.

Master node genes under the control of eQTL but connected to thousands of genes in the coexpression network identify putative targets for orphan zinc finger proteins.

We next explored the genes in our network under the control of eQTL that were connected to the largest number of other genes in the coexpression network. As shown **in Table 3.2**, many genes were connected to several thousand other genes in the gene co-expression network. These include multiple orphan zinc finger proteins (*Zfp758, Zfp280a*), putative transcription factors that have not been characterized. The thousands of genes co-regulated by these genes are possible target genes for these potential transcription factors. Interestingly, *Nphs2*, a nephrosis gene like *Nphs1* discussed above, has a cis-eQTL and is connected to 1313 other genes in the network, suggesting it may play a larger role than its canonical role in glomerular permeability in the kidney. Three genes in the network (*Otud6b, Asb3, Stam2*) are involve in signal transduction through ubiquitination, the latter two coupling this with cytokine signaling. Finally, several genes on this list (bold) are involved in splicing and RNA processing. We speculated in Chapter 2 that very subtle differences in gene expression that nevertheless register as significant in eQTL (and here in gene coexpression) analyses may in fact be due to splice variation, rather than differences in overall gene expression levels. A gene consistently differentially-spliced as a result of genetic factors may produce an overall difference in gene-level transcription calls that, while small (as little as 20%), is robust enough to be picked up in analysis. The presence of splicing and RNA processing factors at the top of our list of genes influencing the largest number of other genes suggests that differential splicing may play a large role in the genetic bases of variation in gene expression.

Discussion

We have utilized a dataset used to identify neuroblastoma susceptibility genes to explore the molecular biology of peripheral sympathetic nerves, a tissue that has been poorly characterized on the molecular level due to its inaccessibility. A common problem with genome-wide transcriptional characterization is the need to compare a specific tissue with some control or reference, making all observations relative to an often arbitrary external value. Here, we have utilized an approach that allows us to make comparisons within the dataset. We have profiled tissue from genetically heterogeneous mice, thus perturbing the regulation of gene expression in a subtle, natural way that reveals mechanisms of genetic regulation, but does not introduce any factors that are physiologically artificial. This approach allowed us to build a gene correlation network in sympathetic ganglia with a strikingly high number of significant correlations: in sympathetic ganglia, over 12,000 genes are wired together with over 8 million connections. Since we analyzed fewer than 17,500 well-characterized genes on the array, this network should represent the majority of all genes expressed in peripheral sympathetic neurons. We have built a similar network with murine neuroblastomas, tumors arising from peripheral sympathetic nervous tissue. We have used this network to identify both known and novel functional connections with genes known to be involved with neuroblastoma (*Bard1*) and diseases of the peripheral nervous system (*Nf1*).

Surprisingly, though we analyzed peripheral nerve ganglia, we found clusters of co-expressed genes involved in olfaction, hearing, and cerebellar Purkinje cell function, as well as photoreceptor biology (not shown). We speculate that low-level expression levels of these genes from distinct but related neuronal populations was perturbed enough by genetic heterogeneity that subgroups of coordinately-regulated genes could emerge. We also speculate that the small, pure nature of the tissue we profiled, consisting mostly of neuron cell bodies and supporting Schwann cells, provided a clean “neuron” signal that was not diluted by supporting cells that would be present in larger organs. Thus, our dataset may be useful to explore transcriptional connections for a wide variety of neuronal cell types, including those in the central nervous system. We have used this to identify genes linked to the *Grid2* receptor, which plays a poorly-understood role in neuronal apoptosis in the central nervous system. We have also used this to identify candidate hereditary hearing loss loci. Finally, by exploring the genes that are most highly connected to the network, we have identified groups of genes (e.g., zinc finger proteins, splicing elements, ubiquitinating enzymes) that play a fundamental role in the expression and splicing of thousands of genes. Further analysis using novel approaches to explore genetic regulation of gene expression at the exon (rather than gene level) may facilitate the identification of novel mechanisms underlying global control of differential splicing. In the meantime, this gene correlation network provides a useful resource to identify novel genetic and functional interactions for genes involved in neuron biology and neuroblastoma.

Materials and Methods

Mouse breeding, sample isolation, genotyping, and array analysis: The methods for these aspects of the work presented here were described in Chapter 2.

Histology: Freshly-isolated SCG were enlarged on a water droplet and photographed with a dissecting scope (**Figure 1A**). Freshly-isolated samples were fixed overnight in 10% buffered formalin, fixed in paraffin. Sectioning and H&E staining were performed by the UCSF Neuropathology Core.

Gene Correlation network analysis: Gene correlation networks were constructed essentially as described[3]. Networks were computed using a Spearman Rank correlation and visualized using Cytoscape (www.cytoscape.org**).** The eQTL data used in conjunction with the analysis here was described in Chapter 2.

Hereditary Hearing loss network: The network was seeded using the following GO categories: "Detection of mechanical stimulus involved in the sensory perception of sound", Sensory perception of sound", "Vibrational conductance of sound to the inner ear". The genomic locations of both candidate hearing loss genes and STS markers for hereditary hearing loci were obtained from the UCSC genome browser (hg19 assembly).

Table 3.1: Candidate Hereditary Hearing Loss Genes

Locus Name	Marker	Location	Location (MB)	Candidate Genes	Gene Start
DFNB32	D1S2819	1p13.3-22.1	95525577	**Abca4**	94458395
DFNA37	D1S495	1p21	102561336		
DFNA49	D1S1167	1q21-q23	160009054	Vsig8, Igsf9	159824106
DFNA7	D1S196	1q21-q23	167604127	Vsig8, Igsf9	159824106
DFNM1	D1S2815	1q24	171714918	Vsig8, Igsf9	159824106
?	D1S238	1q31	188146177	Ptprv, Chit1, Golt1a, Slc26a9	203185206
DFNA34	D1S1609	1q44	244065856		
DFNB47	D2S2952	2p25.1-p24.3	8077969	Aox4	2104600
DFNA43	D2S2116	2p12	76649076	Sftpb	85884440
DFNB58	D2S2970	2q14.1-q21.2	118948332	**Myo7b**	**128293377**
DFNA16	D2S2380	2q24	163815473		
DFNB27	D2S2307	2q23-q31	175456288		
USH2B	D3S1619	3p23-24.2.	34115982	**Col7a1**, Tessp2	48601505
DFNA18	D3S3606	3q22	127200205	Uroc1, Sema5b	126200007
DFNB15	D3S1764	3q21-q25	139188287	Uroc1, Sema5b	126200007
OTSC5	D3S3554	3q22-q24	139605936	Uroc1, Sema5b	126200007
DFNB25	D4S2632	4p15.3-q12	35704112	Corin	47596019
DFNA27	D4S428	4q12	55634405	Corin	47596019
DFNB26	D4S424	4q31	142197645	Etfdh, **Col25a1**	159593276
DFNA24	D4S426	4q	189107651		
DFNB60	D5S404	5q22-q31	116847170		
DFNA42	D5S2115	5q31.1-q32	134719247	Slc23a1, Pura	138702890
DFNA54	D5S1972	5q31	142342662	Slc23a1, Pura	138702890
DFNA21	D6S422	6p21	20370036	Ggnbp1	33154223
OTSC3	D6S1588	6p21.3-22.3	22045333	Ggnbp1	33154223
DFNA31	D6S276	6p21.3	24185801	Ggnbp1	33154223
OTSC7	D6S1036	6q13-16.1	73409804	**Col19a1**	70576447
DFNB38	D6S1599	6q26-q27	162839594	Sod2	160102754
DFNB39	D7S2516	7q11.22-q21.12	71114419	Vps37d, Wbscr28	73082173
DFNB14	D7S554	7q31	97324854	**Zan**	100331248
DFNB17	D7S501	7q31	106440365	**Zan**	100331248
DFNA50	D7S461	7q32	128351601		
OTSC2	D7S2560	7q34-q36	138193292	Atp6v0a4	138391039
DFNB13	D7S1824	7q34-36	140012471	Atp6v0a4	138391039
DFNA47	D9S285	9p21-22	16077944	Gldc	6532465
DFNB33	D9S1826	9q34.3	138448267	**Slc34a3**, Adamts13	140125208
DFNA32	D11S1984	11p15	1566685	Muc6, Syt8	1012823
DFNB20	D11S968	11q25-qter	133818375	Adamts8	130274819
DFNA25	D12S1063	12q21-24	98699399	Stab2	103981068
DFNA41	D12S343	12q24-qter	130804608		
DFNA53	D14S581	14q11-q12	24298493	**Myh7**, Rec8	23881947
DFNB5	D14S286	14q12	38858761		
DFNA23	D14S592	14q21-q22	61396849	Papln	73704204
DFNB48	D15S216	15q23-q25.1	70003754	Lbxcor1	68117941
DFNA30	D15S127	15q25-26	91397599	Acan	89346673
OTSC1	D15S652	15q26.1-qter	92517334	Acan	89346673

DFNB19	D18S452	18p11	5829472	Lama1	6941887
USH1E	D21S1922	21q21	22298744		
DFNB40	D22S686	22q	23068515		
DFN6	DXS8036	Xp22	17048390		
DFN4	DXS997	Xp21.2	31880022		
DFN2	DXS8020	Xq22	99568667		

Table 3.2: Most highly-connected genes with eQTL

eQTL pval	Gene Name	Edges (connected genes)	Function
5.31E-05	Otud6b	6551	Deubiquitinating enzyme
1.17E-07	Zfp758	5133	Zinc finger protein
6.12E-07	Rps13	4666	Ribosomal protein
7.62E-07	Asb3	3364	Cytokine signaling/Ubiquitination complex
3.90E-05	Mfsd7a	3072	?
5.22E-05	Stam2	2456	Cytokine signaling/Ubiquitination complex
2.51E-06	6330416L07Rik	2197	?
1.37E-06	**Skiv2l2**	2196	May be involved in pre-mRNA splicing
7.55E-05	**Paip1**	2114	Translation/RNA binding
2.00E-05	Oas1f	2083	oligoadenylate synthetase
2.52E-05	Zfp280c	1920	Zinc finger protein
1.69E-05	**Aqr**	1817	spliceosome-associated protein
2.72E-05	**Hnrnpr**	1422	precursor mRNA processing
3.38E-05	Nphs2	1313	regulation of glomerular permeability
1.46E-05	Stxbp4	1153	Glucose transport

References

1. Weiss, W.A., Aldape, K., Mohapatra, G., Feuerstein, B.G. & Bishop, J.M. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *EMBO J* **16**, 2985-95 (1997).
2. Chesler, L. et al. Malignant progression and blockade of angiogenesis in a murine transgenic model of neuroblastoma. *Cancer Res* **67**, 9435-42 (2007).
3. Quigley, D.A. et al. Genetic architecture of mouse skin inflammation and tumour susceptibility. *Nature* **458**, 505-8 (2009).
4. Capasso, M. et al. Common variations in BARD1 influence susceptibility to high-risk neuroblastoma. *Nat Genet* **41**, 718-723 (2009).
5. Matsuda, K. et al. Cbln1 is a ligand for an orphan glutamate receptor delta2, a bidirectional synapse organizer. *Science* **328**, 363-8.
6. Uemura, T. et al. Trans-synaptic interaction of GluRdelta2 and Neurexin through Cbln1 mediates synapse formation in the cerebellum. *Cell* **141**, 1068-79.
7. Zuo, J. et al. Neurodegeneration in Lurcher mice caused by mutation in delta2 glutamate receptor gene. *Nature* **388**, 769-73 (1997).
8. Laakkonen, H. et al. Muscular dystonia and athetosis in six patients with congenital nephrotic syndrome of the Finnish type (NPHS1). *Pediatr Nephrol* **21**, 182-9 (2006).
9. Wright, A.F., Chakarova, C.F., Abd El-Aziz, M.M. & Bhattacharya, S.S. Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. *Nat Rev Genet* **11**, 273-284.
10. Alloisio, N. et al. Mutation in the zonadhesin-like domain of alpha-tectorin associated with autosomal dominant non-syndromic hearing loss. *Eur J Hum Genet* **7**, 255-8 (1999).

Figure 3.1. Superior Cervical Ganglia. A. Murine superior cervical ganglia (SCG) (right) compared to a sesame seed (left) and a poppy seed (middle) for scale. **B.** H&E stain of a fixed SCG, showing a balance of neuronal cell bodies and Schwann cells. Picture was taken with a 10x objective.

Fig 3.1

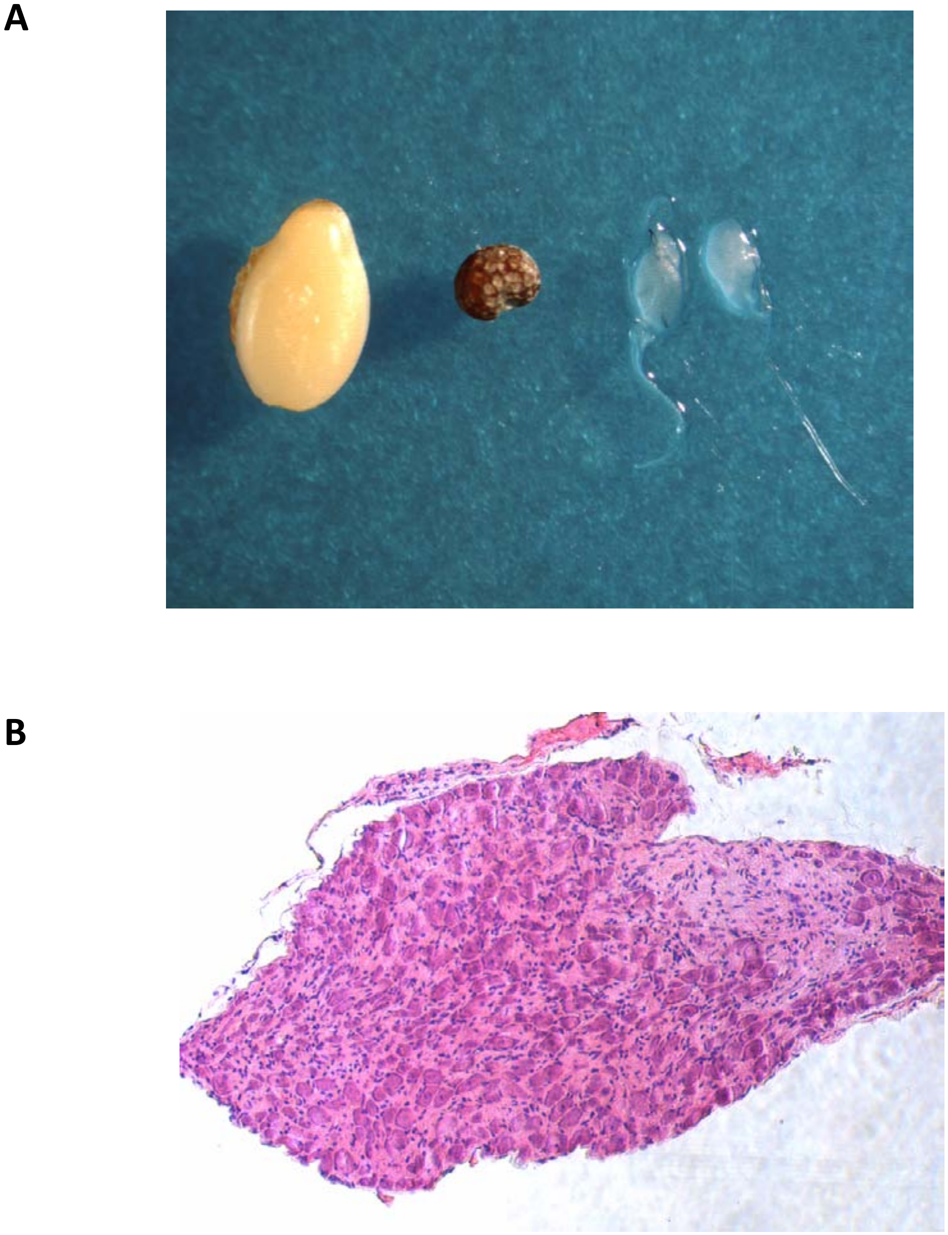


Figure 3.2. Gene coexpression networks from SCG and neuroblastomas. A. Gene coexpression network from SCG samples. Genes are displayed with gray boxes. Two genes are connected if their absolute rho value (reflecting degree of similarity in expression patterns) is greater than 0.9. 2577 genes are wired into the network with 177432 connections. Labels indicate functionally enriched gene groups within the distinct sub-networks. **B.** Tumor co-expression network constructed similarly to the network in (A), using a rho cutoff of 0.95. 1251 genes are connected into this network. The separated sub-network at the top is enriched for collagens and genes involved in striated muscle function.

Fig 3.2

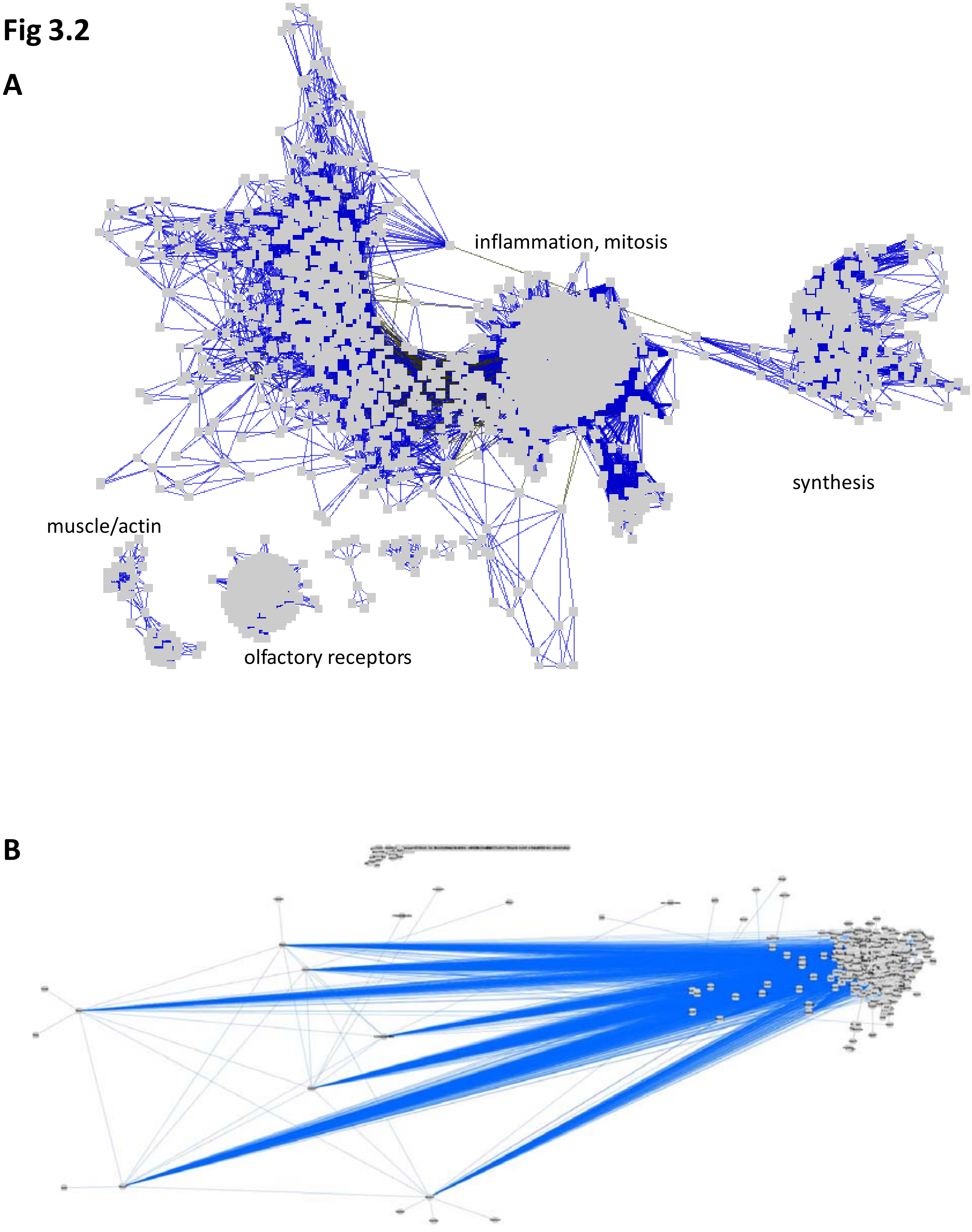

A
inflammation, mitosis
synthesis
muscle/actin
olfactory receptors
B

Figure 3.3. The Bard1 coexpression network from neuroblastoma tumors. The network of genes connected to the Bard1 gene (and to each other) by coexpression in tumors at a rho of greater than 0.8. Blue lines represent positive correlations, red lines indicate inverse correlations. The network contains genes from multiple DNA damage response and repair functional groups, as well as mitotic genes and genes from other functional groups.

Fig 3.3

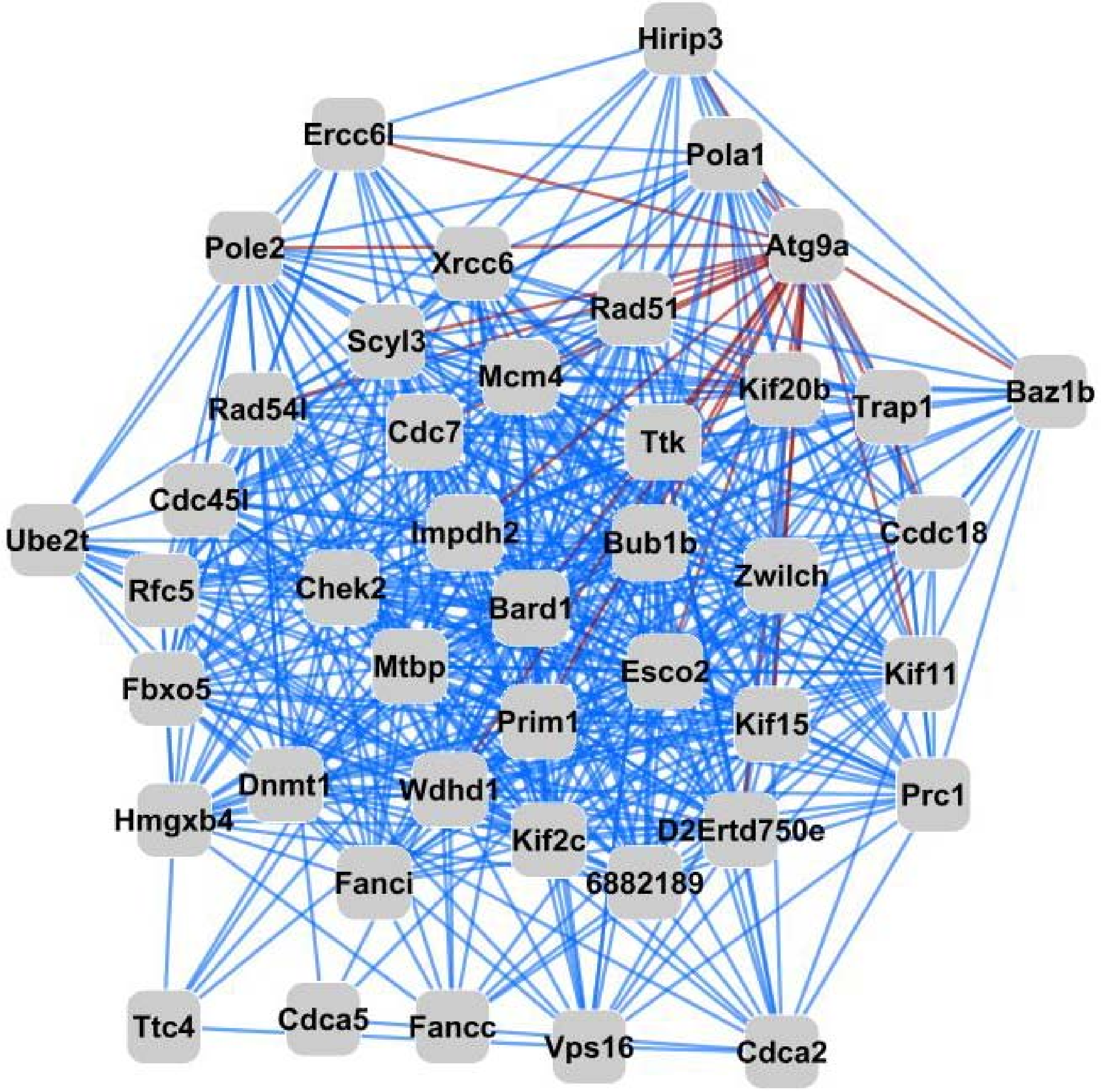

Hirip3
Ercc6l
Pola1
Pole2
Xrcc6
Atg9a
Rad51
Scyl3
Mcm4
Kif20b
Baz1b
Rad54l
Cdc7
Trap1
Ttk
Cdc45l
Ube2t
Impdh2
Bub1b
Ccdc18
Rfc5
Chek2
Zwilch
Bard1
Fbxo5
Mtbp
Esco2
Kif11
Prim1
Kif15
Dnmt1
Wdhd1
Prc1
Hmgxb4
Kif2c
D2Ertd750e
Fanci
6882189
Ttc4
Cdca5
Fancc
Vps16
Cdca2

Figure 3.4. The NF1 gene coexpression network from SCG. Genes connected to *NF1* (and their interconnections) with a rho greater than 0.8 are shown. Blue lines represent positive correlation, red lines indicate inverse correlations. The *Apba3* gene, a candidate Alzheimer's gene, shows a striking inverse relationship with many genes in the network, including *NF1*.

Fig 3.4

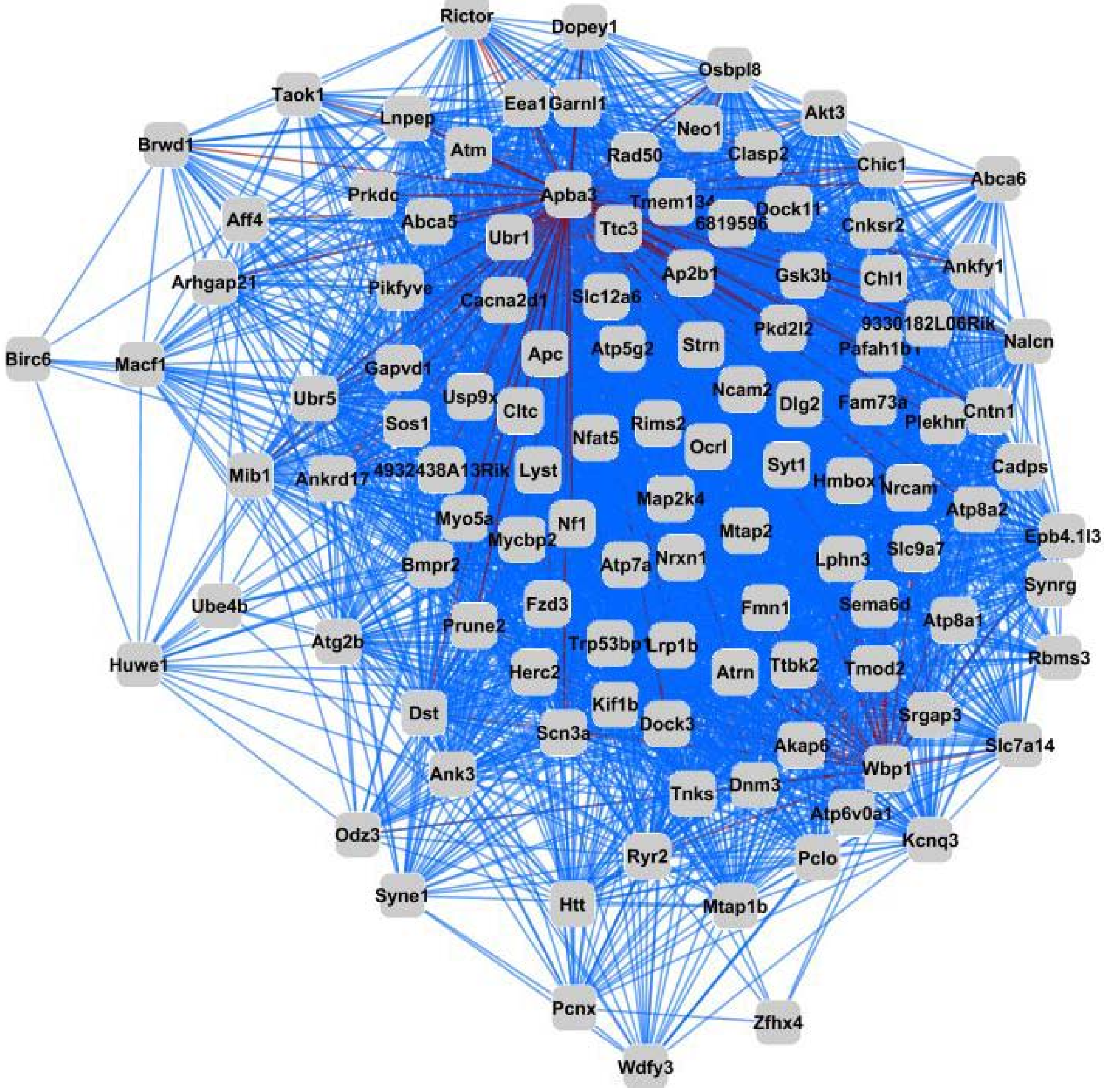

Rictor
Dopey1
Osbpl8
Taok1
Eea1
Garnl1
Akt3
Lnpep
Neo1
Brwd1
Atm
Rad50
Clasp2
Chic1
Abca6
Prkdc
Apba3
Tmem134
Dock11
Aff4
Abca5
6819596
Ttc3
Cnksr2
Ubr1
Arhgap21
Pikfyve
Ap2b1
Gsk3b
Chl1
Ankfy1
Cacna2d1
Slc12a6
Pkd2l2
9330182L06Rik
Nalcn
Birc6
Macf1
Apc
Atp5g2
Strn
Pafah1b1
Gapvd1
Ubr5
Usp9x
Ncam2
Dlg2
Fam73a
Cntn1
Cltc
Sos1
Rims2
Plekhm
Nfat5
Ocrl
Cadps
Syt1
Mib1
Ankrd17
4932438A13Rik
Lyst
Hmbox1
Nrcam
Map2k4
Atp8a2
Myo5a
Nf1
Epb4.1l3
Mtap2
Mycbp2
Slc9a7
Bmpr2
Atp7a
Nrxn1
Lphn3
Synrg
Fzd3
Ube4b
Fmn1
Sema6d
Atp8a1
Prune2
Atg2b
Trp53bp1
Lrp1b
Rbms3
Huwe1
Herc2
Atrn
Ttbk2
Tmod2
Kif1b
Dst
Srgap3
Scn3a
Dock3
Akap6
Slc7a14
Wbp1
Ank3
Tnks
Dnm3
Atp6v0a1
Kcnq3
Odz3
Ryr2
Pclo
Syne1
Htt
Mtap1b
Pcnx
Zfhx4
Wdfy3

Figure 3.5. A network of genes connected to the *GRID2* orphan glutamate receptor suggest mechanisms for its role in cellular functions. *GRID2* has a strong cis eQTL, suggesting it is under local genetic control. However, several genes in the network have expression patterns correlated with the *GRID2* expression pattern (at a rho greater than 0.7), indicating they are under related genetic control (in this case, likely *GRID2*). Blue lines represent positive correlation, yellow lines indicate inverse correlations. Functions of these connected genes may illuminate the role of *GRID2* in neuronal apoptosis.

Fig 3.5

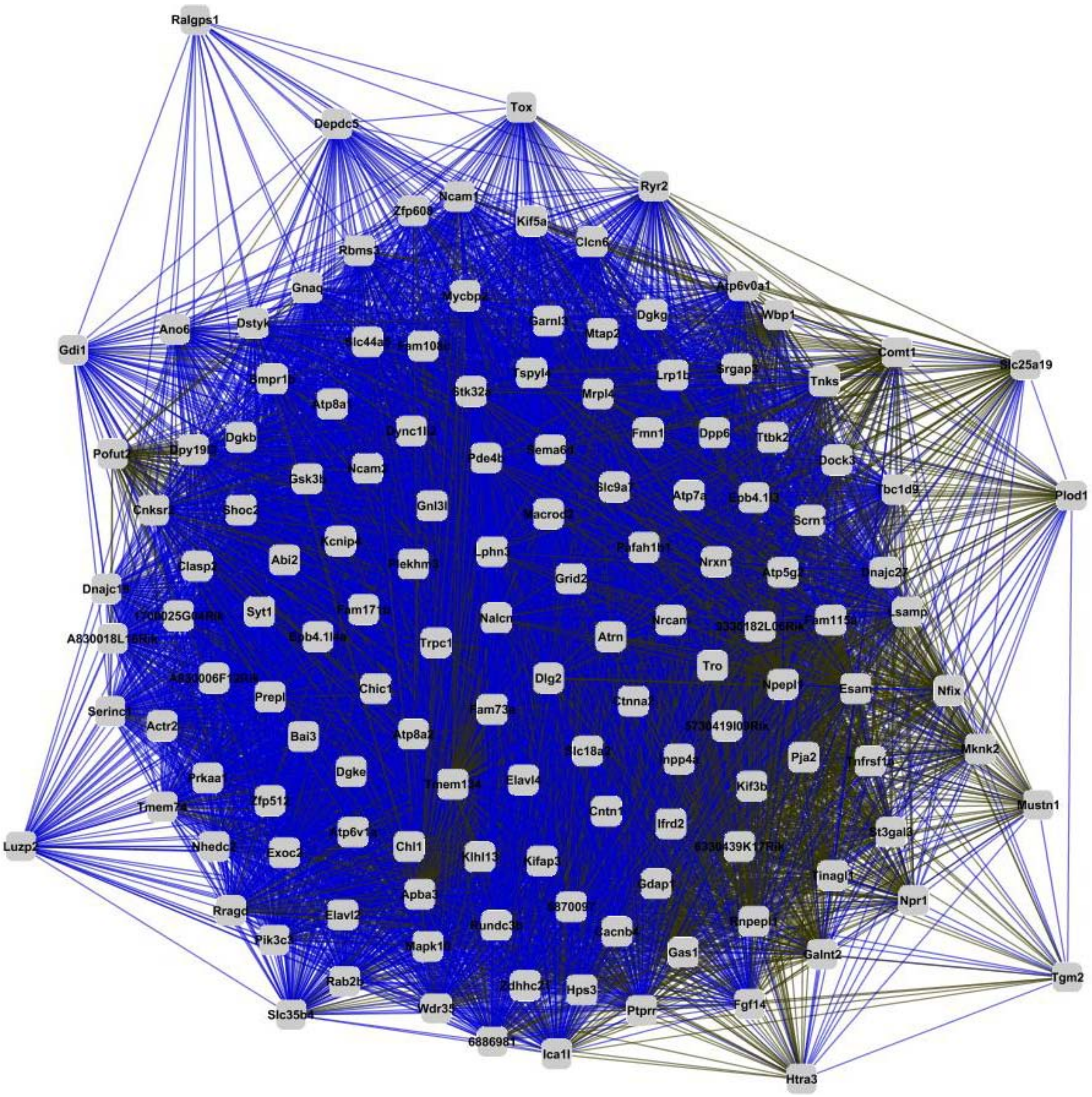
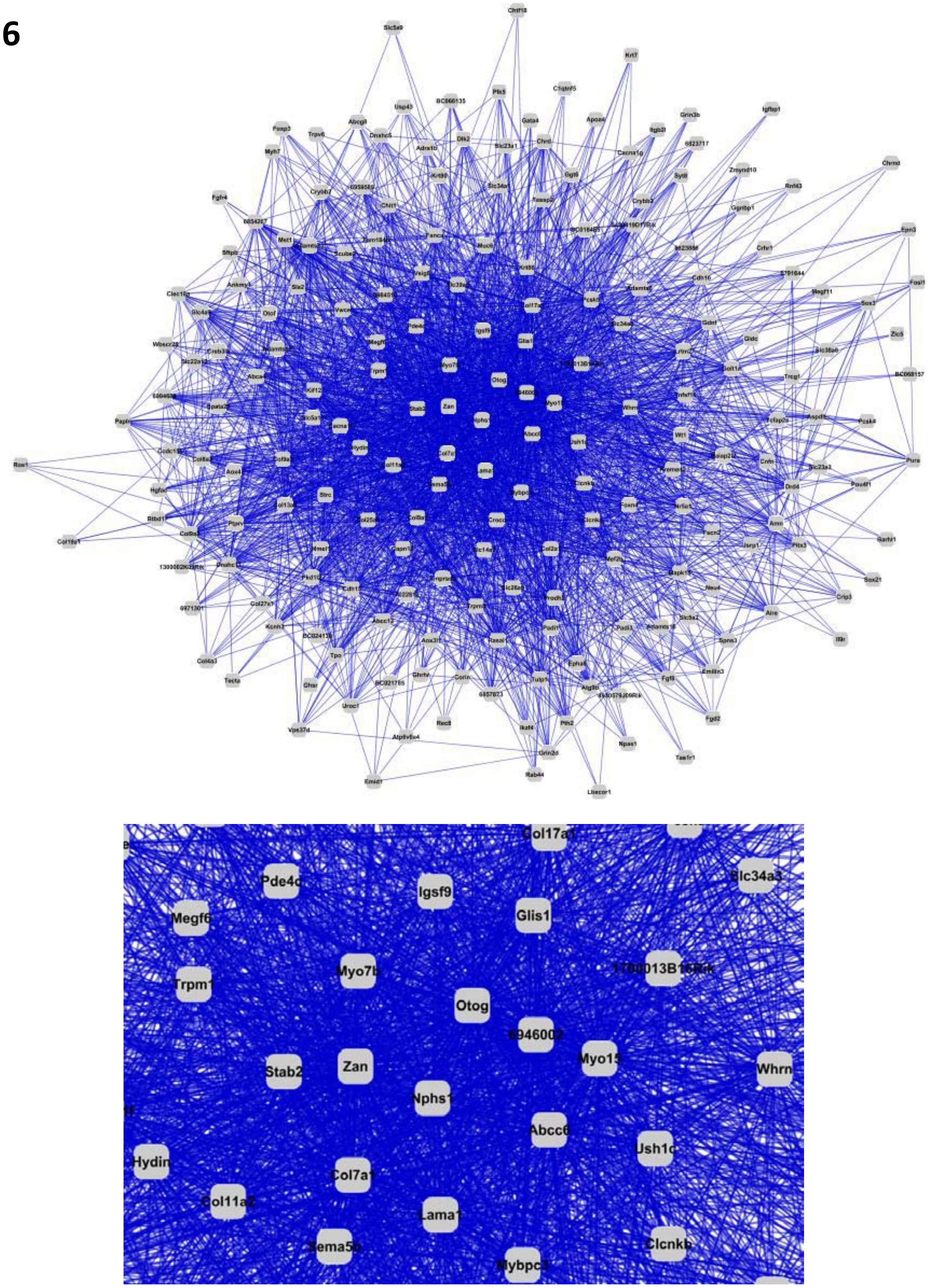

Ralgps1
Tox
Depdc5
Ryr2
Ncam1
Zfp608
Kif5a
Clcn6
Rbms3
Gnaq
Mycbp2
Atp6v0a1
Wbp1
Gdi1
Ano6
Dstyk
Slc44a
Fam108
Garnl3
Mtap2
Dgkg
Comt1
Slc25a19
Bmpr1b
Tspyl4
Srgap3
Lrp1b
Tnks
Atp8a1
Stk32a
Mrpl4
Dgkb
Dync1li2
Pofut2
Dpy19l
Fmn1
Dpp6
Ttbk2
Sema6d
Pde4b
Dock3
Gsk3b
Ncam2
Tbc1d9
Plod1
Slc9a1
Atp7a
Epb4.1l3
Cnksr2
Shoc2
Gnl3l
Macrod2
Scrn1
Kcnip4
Pafah1b1
Abi2
Plekhm3
Lphn3
Nrxn1
Atp5g2
Dnajc27
Clasp2
Grid2
Dnajc18
Syt1
Fam171a
1700025G04Rik
Nalcn
Nrcam
3330182L06Rik
Fam115a
Lsamp
A830018L16Rik
Epb4.1l4a
Trpc1
Atrn
Tro
A830006F12Rik
Dlg2
Npepl1
Esam
Nfix
Chic1
Prepl
Serinc1
Ctnna2
Fam73a
Actr2
5730419I09Rik
Bai3
Atp8a2
Slc18a2
Mknk2
Inpp4a
Pja2
Tnfrsf1a
Prkaa1
Dgke
Tmem134
Elavl4
Kif3b
Tmem74
Zfp512
Mustn1
Cntn1
Ifrd2
Atp6v1a
St3gal3
Nhedc2
Exoc2
Chl1
Klhl13
6330439K17Rik
Luzp2
Kifap3
Tinagl1
Apba3
Gdap1
Npr1
Rragd
6870097
Elavl2
Rnpepl1
Rundc3b
Pik3c3
Cacnb4
Mapk10
Gas1
Galnt2
Tgm2
Rab2b
Zdhhc21
Hps3
Fgf14
Slc35b4
Wdr35
Ptprr
6886981
Ica1l
Htra3

Figure 3.6. A network seeded by hearing-related genes reveals novel candidate hereditary hearing loss genes. A. 12 (of 66) GO-defined hearing-related genes were present in our overall network. Their connections (at a rho cutoff of 0.95), as well as the interconnections of the connected genes (some of which may share a functional role in hearing), result in a network of 225 genes. **B.** At the center of the network lie genes such as *Nphs1*, a nephrosis gene, and *Zan*, a gene thought to be expressed primarily in sperm. Both genes represent novel candidates that may play a role in the biology of hearing.

Fig 3.6


A
B
Col17a1
Pde4c
Igsf9
Glis1
Slc34a3
Megf6
Trpm1
Myo7b
Otog
1700013B16Rik
6946002
Stab2
Zan
Myo15
Whrn
Nphs1
Abcc6
Ush1c
Hydin
Col7a1
Col11a2
Lama1
Sema5b
Clcnkb
Mybpc3

Chapter 4: Introduction and Background to Part II

Source: The following contains background on forward genetic screens relevant to chapters in Part II

Contributions: This is an original review of the literature that has been aided by conversations with several individuals.

Introduction and Background to Part II

Insertional mutagenesis as a strategy to identify gene involved in tumor development

Forward genetic screens provide a powerful complement to linkage mapping as a means to unravel the molecular biology underlying disease. Forward genetic screens in cancer using integrating vectors such as retroviruses and transposons provide two advantages over QTL mapping. First, while QTL mapping is dependent on a limited set of pre-existing, naturally occurring genetic differences affecting a trait of interest, forward genetic screens can generate a vast amount of relatively unbiased novel mutations, providing the potential to identify many more genes underlying a phenotype. Second, the resolution of QTL mapping is inherently limited by a finite set of markers and a limited number of meiotic recombinations, and even large screens are rarely able to resolve a QTL to a single gene or a manageable group of candidate genes. In contrast, forward genetic screens utilizing insertional mutagens (as opposed to radiation or chemical mutagens) have an advantage in that insertion sites of integrating vectors can be localized to the exact base position in the genome. In combination with the fully sequenced mouse genome, this allows for the specific identification of mutated genes in forward genetic insertional mutagenesis screens in mice.

While transposon vectors such as the drosophila P-element have been used in mutagenesis screens in invertebrates, the lack of a functional transposon in mammals initially limited the possible insertional mutagenesis vectors in mouse model systems to integrating retroviruses. While gene-trapping retroviruses have been used to conduct forward genetic screens in the germline, mutagenesis of somatic tissue was limited to chemical mutagens and radiation, both of which presented difficulties in pinpointing causal mutations, and retroviruses, which provided an easy means to identify disrupted genes, but were limited to tissues accessible by viral vectors. Nevertheless, retroviral vectors have been particularly useful for insertional mutagenesis screens for cancer-causing genes. In these screens, viruses infect and randomly integrate into the host genomes of a large number of cells from a given tissue. In individual cells in which a retrovirus inserts into or near a dormant oncogene and activates it, the cell can become transformed, proliferate, and give rise to a tumor in which virtually all of the cells harbor a copy of the oncogenic insertion. Though the same process can deactivate tumor suppressors, in practice this is less frequently observed due to the necessity of deactivating both alleles of a gene in a diploid genome. The chances of deactivating both alleles independently is much lower than activating a single allele (specifically, roughly the square of the probability), without invoking another mechanism such as haploinsufficiency or loss of the other allele through genomic instability. When several independent tumors are analyzed, genomic loci harboring retroviral insertions in multiple tumors are called common insertion sites (CIS), a phenomenon statistically unlikely to happen by chance and suggestive of clonal selection for cells harboring insertions into genes at the locus.

While this process was initially recognized in chickens infected with the ALV virus, the approach has been of particular use in mice. In particular, the Murine Leukemia Viruses (MLV) and mouse mammary tumor virus (MMTV) have been used to identify a large number of cancer-promoting genes in the hematopoietic system and mammary glands, respectively (reviewed in[1]). While the system is often used to identify genes capable of independently initiating oncogenic transformation, the approach can also be applied to mouse model systems carrying oncogenic transgenes or tumor suppressor knockouts to identify secondary mutations cooperating with the primary lesion (for example, see[2]), as well as resistance to therapy (for example, see[3]). While these vectors have been tremendously useful in identifying genes driving leukemia and breast cancer, the vectors are limited by several technical drawbacks. Most notably, the use of retroviral vectors for somatic insertional mutagenesis screens in mice is limited by the range of tissues that can be infected by viruses *in vivo*. This has effectively limited large-scale insertional mutagenesis screens to leukemia and breast cancer, though the chicken Tva receptor for the ALV virus was expressed as a transgene and used to perform viral insertional mutagenesis in glioma[4]. Second, while cancer is a process requiring multiple genetic lesions[5], retroviral insertional mutagens can be limited in the number of insertions possible per cell, as replication-competent viruses can express envelope proteins that bind receptors on the cell surface and prevent secondary re-infection[1] and thus limit the number of retrovirally-induced mutations per cell. Third, all integrating vectors display preferences for particular characteristics of their insertion sites, whether large-scale (e.g. preferences for genes and promoters) or small-scale (e. g. preferences for specific sequences) that bias the spectrum of genes mutated in these screens.

Several of these shortcomings and biases were overcome with the development of the Sleeping Beauty transposon system[6] in the late 1990's. The Sleeping Beauty transposon is a synthetic mammalian transposon that was developed using site-directed mutagenesis to restore consensus sequences derived from several dormant Tc1/mariner DNA transposons identified in salmanoid fish. Restoration of an active, nuclear-localized transposase and indirect-repeat/direct-repeat (IR/DR) bounding regions recognized by the transposase resulted in the first cut-and-paste DNA transposon with activity in mammalian cells. This proof-of concept paved the way for the development of other DNA transposons with activity in mammalian cells, notably Tol2[7], Frog Prince[8] and piggyBac[9].

The advent of DNA transposons active in mammalian cells provided an opportunity to expand the possibilities of somatic mutagenesis screens in mouse cancer model systems. Pioneering work by the Largaespada group resulted in a functional Sleeping Beauty based insertional mutagenesis system in mice[10,11]. In the first iteration of this technology[10], an oncogenic transposon, T2/Onc, was constructed by inserting an MSCV promoter upstream of a splice donor element capable of activating genes surrounding the transposon insertion site, generating gain-of-function mutations. Additionally, splice acceptor sites upstream of transcriptional stop elements were engineered in both directions, capable of terminating/truncating upstream transcripts, generating loss-of-function mutations. These elements were nested between the SB flanking IR/DR sequences. These dormant transposons were activated by a trans-acting SB transposase (SB10), in this case driven by a (presumed) constitutively-active CAGGS (CMV enhancer/chicken β actin promoter) element. Transgenic mice carrying dormant concatemers of T2/Onc transposons and CAGGS-SB10 transposase (at a separate, trans-acting locus) were then created. While the mice showed no phenotypes by themselves, the CAGGS-SB10;T2/Onc2 combination decreased the latency of sarcoma formation in knockout mice lacking the *p19/Arf* tumor suppressor . Virtually all tumor-prone mice were shown to have a T2/Onc insertion in intron 9 of the *Braf* oncogene, generating a gain-of-function transcript, among other transposon insertions. This observation validated the Sleeping Beauty insertional mutagenesis system as a method for gene discovery in cancer.

While this proof-of-principle experiment demonstrated the potential for somatic transposon-based insertional mutagenesis, the development of an antibody for Sleeping Beauty suitable for immunohistochemistry (IHC) revealed that transposase expression in adult CAGGS-SB10 mice was limited primarily to muscle tissue (though a PCR assay had detected transposon excision from the donor concatemer all tissues examined, presumably due to expression at some early developmental window). A more potent transposase mouse was then generated by knocking the enhanced SB11 transposase[12] into the *Rosa26* locus to generate (again, presumably) ubiquitous expression[11]. In addition, new donor transposon transgenic lines were made that carried high copy numbers (200-400, compared to roughly 20 for the original T2/Onc lines). The Rosa26-SB11;T2/Onc2(high-copy) combination was capable of generating tumors without a predisposing oncogene or tumor suppressor knock-out, as all doubly-transgenic mice succumbed to tumors by four months of age. While a diverse range of tumors were reported, the majority of mice succumbed to leukemia. Additionally, there was an estimated 50% embryonic lethality among doubly-transgenic mice. Analysis of insertion sites showed that among 10 leukemias studied, 6 harbored activating insertions in the *Notch1* locus, a gene activated in 50% of human T-cell acute lymphoblastic leukemia, again validating the system for finding relevant cancer-causing genes. Additionally, a subset of tumors with *Notch1* insertions also harbored insertions in *Rasgrp1*, a regulator of the ras oncogenic signaling pathway, suggesting an interaction between these two genes and demonstrating the power of the SB system to uncover multiple genetic events in tumor development.

The rapid tumor onset, complete penetrance, and diverse range of tumors observed with the Rosa26-SB11;T2/Onc2(high-copy) mice represented a technical improvement over the CAGGS-SB10;T2Onc2 mice, and demonstrated that the oncogenic power of the system was coupled to transposase activity/expression and transposon copy-number. However, these traits also introduced a practical limitation for the system. The rapid onset of lethal leukemias precluded the use of the system to identify genes accelerating tumorigenesis in existing model systems in which tumor onset was later than the first few weeks of life, essentially eliminating the possibility of screens for tumors in adult tissue. To circumvent this limitation, a third generation transposase construct was created in which a floxed stop element was inserted upstream of the SB11 transposase, and the construct was again knocked into the *Rosa26* locus (Rosa26-lsl-SB11), allowing for tissue-specific transposase expression activated by the Cre recombinase[13]. Mice carrying this construct and the high-copy T2/Onc2 transposon donor concatemer do not develop tumors in the absence of Cre. This construct was used to identify genes promoting hepatocellular carcinoma[13] and colorectal cancer[14].

To date, the SB system has been used, with varying degrees of success, to identify genes driving sarcomas[10], leukemias[11], prostate cancer[15], hepatocellular carcinoma[13], colorectal cancer[14], glioma/astrocytoma[16,17], and medulloblastoma ([11], M. Taylor, unpublished results). Additionally, a new mutagenic transposon, T2/onc3, in which the MSCV LTR has been replaced with the CAGGS element, has been used to generate an even wider range of tumors when combined with a constitutively-expressed transposase[18]. However, the system is not universal; for every tissue in which tumors have been successfully generated or accelerated by SB insertional mutagenesis, there are several tissues in which the strategy has been tried and has failed, though these results generally go unreported (numerous postdocs and grad students, personal communication).

As mentioned in Part I, neuroblastoma is a disease for which little is known about the molecular biology underlying tumor development, and, as such, a forward genetics screen to uncover new genes involved in tumor pathology could be of great benefit to our understanding of the disease. We sought to utilize the Sleeping Beauty system in our TH-MYCN driven mouse model of neuroblastoma to identify genes that cooperated with the MYCN transgene and accelerated tumorigenesis in our model, potentially identifying genes from regions of copy number abnormalities seen in both mouse and human tumors[19], as well as genes mapping to our susceptibility modifier loci (Chapter 2). This project proceeded through three stages. First, thought our ultimate goal was to use the existing Sleeping Beauty constructs to identify genes in neuroblastoma, as well as *c-myc* and HER2-driven breast cancer, this emerging technology was still relatively uncharacterized in terms of basic molecular biology, biochemistry, and genetics, and we thus contributed to the characterization of the system. We characterized the insertion site preferences of the transposon, ultimately generating a bioinformatics tool to predict likely insertion sites, and compared these characteristics with other vector systems[20,21] (Chapters 5 and 6). We also analyzed the large-scale genomic changes in SB-accelerated tumors, showing that the transposon-transposase system generated small genomic copy number abnormalities surrounding the donor transposon concatemer, but did not cause overall genomic instability and in fact generated tumors with fewer overall copy number abnormalities than control tumors lacking mobilized transposons[16] (Chapter 7). Finally, after establishing that the existing transposase mice were insufficient for mutagenesis in the peripheral sympathetic nervous system, certain brain tumors, and in breast cancer, we developed novel tools and approaches to perform insertional mutagenesis in these tissues (Chapters 8, 9, and 10).

References

1. Jonkers, J. & Berns, A. Retroviral insertional mutagenesis as a strategy to identify cancer genes. *Biochim Biophys Acta* **1287**, 29-57 (1996).
2. Yin, B. et al. A retroviral mutagenesis screen reveals strong cooperation between Bcl11a overexpression and loss of the Nf1 tumor suppressor gene. *Blood* **113**, 1075-85 (2009).
3. Lauchle, J.O. et al. Response and resistance to MEK inhibition in leukaemias initiated by hyperactive Ras. *Nature* **461**, 411-4 (2009).
4. Johansson, F.K. et al. Identification of candidate cancer-causing genes in mouse brain tumors by retroviral tagging. *Proc Natl Acad Sci U S A* **101**, 11334-7 (2004).
5. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
6. Ivics, Z., Hackett, P.B., Plasterk, R.H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-10 (1997).
7. Kawakami, K., Shima, A. & Kawakami, N. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. *Proc Natl Acad Sci U S A* **97**, 11403-8 (2000).
8. Miskey, C., Izsvak, Z., Plasterk, R.H. & Ivics, Z. The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. *Nucleic Acids Res* **31**, 6873-81 (2003).
9. Ding, S. et al. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-83 (2005).
10. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6 (2005).
11. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6 (2005).
12. Geurts, A.M. et al. Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther* **8**, 108-17 (2003).
13. Keng, V.W. et al. A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-74 (2009).
14. Starr, T.K. et al. A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323**, 1747-50 (2009).
15. Rahrmann, E.P. et al. Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping Beauty transposon-based somatic mutagenesis screen. *Cancer Res* **69**, 4388-97 (2009).
16. Collier, L.S. et al. Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality. *Cancer Res* **69**, 8429-37 (2009).
17. Bender, A.M. et al. Sleeping beauty-mediated somatic mutagenesis implicates CSF1 in the formation of high-grade astrocytomas. *Cancer Res* **70**, 3557-65.
18. Dupuy, A.J. et al. A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. *Cancer Res* **69**, 8150-6 (2009).

19. Hackett, C.S. et al. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-73 (2003).
20. Geurts, A.M. et al. Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Res* **34**, 2803-11 (2006).
21. Hackett, C.S., Geurts, A.M. & Hackett, P.B. Predicting preferential DNA vector insertion sites: implications for functional genomics and gene therapy. *Genome Biol* **8 Suppl 1**, S12 (2007).

Chapter 5: Structure-based prediction of insertion-site preferences of transposons into chromosomes.

Source: The following chapter was published as a manuscript in Nucleic Acids Research in 2006 (PMID 16717285).

Contribution: I shared first-authorship with Aron Geurts. I designed the computational approaches for the manuscript and wrote all data analysis scripts. Aron perfomed data analysis PB Hackett supervised the project, and the other authors provided data, statistical analysis, and insight into experimental design.

Structure-Based Prediction of Insertion-Site Preferences of Transposons into Chromosomes

Aron M. Geurts*[1], Christopher S. Hackett*[2], Jason B. Bell[1], Tracy L. Bergemann[3], Lara S. Collier[4], Corey M. Carlson[4], David A. Largaespada[1,4] & Perry B. Hackett[1,4‡]

[1]Department of Genetics, Cell Biology, and Development, The Arnold and Mabel Beckman
Center for Transposon Research, University of Minnesota, Minneapolis, MN 55455

[2]Biomedical Sciences Graduate Program, University of California San Francisco, San Francisco, CA 94143-0452

[3]Biostatistics Core, University of Minnesota Cancer Center, Minneapolis, MN 55455

[4]University of Minnesota Cancer Center, Minneapolis, MN 55455

* These authors contributed equally to the work
‡ To whom correspondence should be addressed E-mail: perry@cbs.umn.edu

Running Title: Integration-site Preferences of SB Transposons

Key words: gene therapy, insertional mutagenesis, mouse, ProTIS, transposons

Abbreviations: AAV, adeno-associated virus; HIV-1, human immunodeficiency virus-1, MLV, murine leukemia virus; SB, Sleeping Beauty

For Correspondence:

Perry B. Hackett

Professor

Department of Genetics, Cell Biology, and Development, University of Minnesota

6-160 Jackson Hall

321 Church St SE

Minneapolis, MN 55455

Tel: 612-624-6736

Fax: 612-625-6140

Abstract

Mobile genetic elements with the ability to integrate genetic information into chromosomes can cause disease over short periods of time and shape genomes over eons. These elements can be used for functional genomics, gene transfer, and human gene therapy. However, their integration-site preferences, which are critically important for these uses, are poorly understood. We analyzed the insertion sites of several transposons and retroviruses to detect patterns of integration that might be useful for prediction of preferred integration sites. Initially we found that a mathematical description of DNA-deformability, called V_{step}, could be used to distinguish preferential integration sites for *Sleeping Beauty* (SB) transposons into a particular 100-bp region of a plasmid [1]. Based on these findings, we extended our examination of integration of SB transposons into whole plasmids and chromosomal DNA. To accommodate sequences up to 3 Mbp for these analyses, we developed an automated method, *ProTIS*©, that can generate profiles of predicted integration events. However, a similar approach did not reveal any structural patterns of DNA that could be used to predict favored integration sites for other transposons as well as retroviruses and lentiviruses due to a limitation of available data sets. Nonetheless, *ProTIS*© has utility for predicting likely SB transposon integration sites in investigator-selected regions of genomes and our general strategy may be useful for other mobile elements once a sufficiently high density of sites in a single region are obtained. *ProTIS* analysis can be useful for functional genomic, gene transfer and human gene therapy applications using the SB system.

Introduction

Mobile genetic vectors have been harnessed for genetic studies in model organisms and are being developed as agents for gene-therapy in humans [2-4]. For example, the awakening of the *Sleeping Beauty* (SB) transposon system as a powerful tool for insertional mutagenesis to identify oncogenes [5,6] and other classes of genes [7,8] complements retroviral vectors, which have been used for decades [9]. Importantly, understanding the parameters that affect integration of vectors is required to appreciate fully the results of their applications.

Although transposons and some retroviruses integrate in virtually all regions of host genomes, their integration is not random [10-19]. Weak consensuses sequences have been described surrounding the sites of integration for retroviruses [17,20] and transposable elements [21-23]. However, the most-favored integration sites do not always conform to these sequences [1]. In addition to specific-sequence recognition, DNA structural characteristics, including protein-induced deformability, A-philicity and bendability, have been shown to influence binding of proteins [24]. Although these structural characteristics are sequence-dependent, two dissimilar sequences can have similar structural patterns. As a result, distinct preferred integration sites may not match consensus sequences, but rather share similar structural patterns. Unique patterns of these DNA structural characteristics at integration sites have been reported for retroviruses and lentiviruses [20], *P*-elements [21] and SB transposons [1,22] that may contribute to mechanisms that differentiate potential loci for integration of mobile genetic elements. We previously used a mathematical description of DNA 'deformability' called V_{step}, to identify shared structural patterns among several preferred integration sites for SB transposons into a short 100-bp region of a target plasmid[1]. DNA deformation is characterized by a non-uniform twisting of the double helix, alteration in the spacing between the base pairs at the integration site, and localized tilting of the target site such that the axis around the insertion site is off center. This initial analysis did not answer the question of whether these parameters can be used to effectively predict integration site preferences into chromosomal DNA in mammalian genomes nor whether other integrating vectors followed similar rules.

Here we describe our strategy of using a small dataset of high-density integrations into a defined region of DNA to formulate rules that govern integration-site preferences in lengths of chromatin of more than 3 Mbp. To analyze such long stretches of DNA, we developed an algorithm for rapidly scanning DNA sequences to predict favored sites of integration of mobile elements into mammalian chromosomes. We used SB transposons as a model element to establish a method for finding and testing rules that govern integration-site preferences. We used two datasets from forward-genetic studies to verify the predictions made by our algorithms and then examined potential integration preferences for two other transposons as well as retroviral and lentiviral vectors.

Materials and Methods

Algorithm for determining the V_{step} profile of SB transposon integration sites

Figure 5.1 illustrates the steps in establishing V_{step} profiles for a given TA site. We developed a Perl script, called *ProTIS* © (*Profiler of Transposon Insertion Sites*), to analyze automatically every TA site in an input sequence file (up to 20 Mb tested). For each TA dinucleotide in the input sequence, the script extracts 5 bases on each side of the TA dinucleotide and translates the 12-base sequence into a series of V_{step} values for the 11 transitions between consecutive base pairs (referred to as dimer steps) within the sequence (**Figure 5.1**). Then, using a series of less-than (<) or greater-than (>) comparisons, the program uses the ordered V_{step} values to classify each TA site and its flanking nucleotides into one of five classes (4^+-peak, 3.5-peak, 3-peak, 2.5 peak, and basal). For each schematic, the TA-peak is shown in bold and is generated from the [4], [5], and [6] V_{step} values, where [5] always has a V_{step} value of 6.3. For the 4^+-peak and 3.5-peak patterns, the light lines represent peaks or half-peaks that must be on at least one or can be on both sides of the central TA-peak. The patterns shown in **Figure 5.1** show mirror images because the transposon can integrate in either direction into the symmetrical TA dinucleotide basepairs. For the 3-peak and 2.5-peak patterns, the peaks flanking the TA-peak can be to the left or the right of the TA-peak. Experimentally, 2.5-, 3-, and 3.5-peak pattern target sites exhibited negligible differences in their abilities to attract SB transposons, so they were grouped together. This allowed us to simplify our TA site classification into three groups, *preferred* sites (4^+-peak), *semi-preferred* sites (3.5/3/ 2.5-peak), and *basal* sites. With these definitions, for every bin of a given length of DNA, the total number of each class of TA site per bin is tallied for the input sequence. Using weighted coefficients (shown in bold in the equation below) for each class of TA site from the pFV/Luc data in **Table 5.1**, a Total V_{step} score can be calculated for each bin using the equation:

$$\text{Total } V_{step} = \sum_{N \to N+(\text{bin size})} [\mathbf{13}(\#\text{ preferred sites}) + \mathbf{5}(\#\text{ semi-preferred sites}) + \mathbf{1}(\#\text{ basal sites})]$$

The script produces a tab-delineated table output that is then conveniently analyzed and graphed using Microsoft Excel (Microsoft, Redmond, WA).

Analyzing V_{step} and A-philicity profiles of insect transposons and retroviruses

An additional script was generated to accept tabulated integration site data from different sources. The V_{step} classifier script can accept sequence information accumulated in integration-site studies. The script takes each line of the tabulated data, extracts the pertinent sequence information, assigns both the V_{step} and A-philicity values to each dimer step, and generates tab-delineated output files similar to that of *ProTIS* ©. *ProTIS* ©, including further instructions, is available for download on the Hackett lab website http://www.cbs.umn.edu/labs/perry/ as open-source code. Control sequences for *piggyBac* and *P*-element analyses were obtained from three separate one-megabase regions of the *Drosophila* genome (4.2 BDGP release), chromosome 2L from position 10-11 Mb, chromosome 2L from position 17-18 Mb, and chromosome 3L from position 11-12 Mb.

Statistical Analysis

To examine the relationship between the *ProTIS*© prediction, based on the Total *Vstep* score and known insertion sites, we fit a Poisson regression model. This model takes the number of insertions into each bin as a measurement of "insertion activity" in that region and compares it to the predicted score for that bin made by *ProTIS*©. To take into account that the incoming transposon does not define a target sequence in terms of 100-bp bins, we fit a *lag-1* autocorrelation structure. The autocorrelation structure assumes that neighboring bins are correlated at some estimated level r, and that the correlation disappears exponentially with increasing genetic distance, r^d. These combined methods measure the relationship of the insertion activity and Total V_{step} scores across the entire target sequence and calculate regression coefficients $\ddot{\beta}$ using generalized estimating equations. The robust standard errors associated with this analysis were used to derive *p*-values (25). The regression coefficient, in turn, can be used to derive a relative risk value $e^{\hat{\beta}}$. In this case, $e^{\hat{\beta}}$ correlates an increase in Total V_{step} score *for any bin*, compared to any particular TA site, with the likelihood that a transposon will insert in the bin. For the pFV/Luc integrations, (Results) this fit yielded a regression coefficient $\ddot{\beta} = 0.05$ and corresponds to a relative risk of $e^{\ddot{\beta}} = 1.05$. A comparison of the Total V_{step} values of bins 17 and 22 in **Figure 5.2**, which have nearly the same number of TA sites, 15 and 16 respectively, but different Total V_{step} scores of 72 and 27, respectively, gives a difference of 45. The difference of 45 corresponds to 9.5-fold ($e^{45\hat{\beta}}$) increase in the likelihood that an insertion will occur in bin 17 compared to 22. Likewise, fitting an equivalent model for the *Braf* data (Results) yields a slightly smaller regression coefficient $\hat{\beta} = 0.045$. Interpreting this coefficient means that an increase in the Total *Vstep* score of 20 raises the probability of an insertion by $e^{20\hat{\beta}} = 2.46$, while an increase of 40 raises the probability of an insertion by $e^{40\hat{\beta}} = 6.05$.

Results

Development of an algorithm for V_{step} profiles of transposon-integration sites

Our analysis began with SB transposons, which always integrate into the simple dinucleotide sequence TA[25]. The DNA structural parameter V_{step} is a measurement of the protein-induced deformability of DNA sequences, gathered from the analysis of DNA molecules bound and unbound by proteins [26]. A V_{step} value correlates with the level of deformability of the DNA double helix at the transition between two consecutive base pairs in a sequence (dimer steps, **Figure 5.1** top panel) [26]. Using an intra-plasmid transposition analysis that examined 100 bp of the 7758-bp pFV/Luc plasmid, we found that potential TA-integration sites could be divided into three groups with a 16-fold range for integration preference based upon V_{step} patterns of base pairs flanking the target TA dinucleotide [1].

Based on this very special case, we extended our analysis of target-site selection to refine our ability to predict preferred SB integration sites. Since establishing a V_{step} profile for extended regions is extremely tedious, we generated a Perl script that analyzes every TA site in a DNA sequence and assigns a V_{step} value to consecutive transitions between base pairs flanking the site. The series of V_{step} values for each dimer step in the sequence can be graphed to establish a pattern that can be used to distinguish various integration sites. Each TA site is then classified in terms of likelihood of transposon integration based on which of the three categories of V_{step} patterns it mimics (Materials and Methods).

Our analysis of the entire 7758-bp plasmid revealed that a 12-bp window, including five base pairs flanking each side of a target TA dinucleotide, was sufficient to distinguish four TA-site V_{step} profiles that differed in their integration potentials when compared to a non-preferred, or *basal*, TA site (**Figure 5.1**, bottom panel). To facilitate our analyses, we combined several profiles of TA sites that have similar V_{step} patterns into a single category (**Figure 5.1**, Semi-preferred), so that any TA site in a target falls into one of three groups - *preferred, semi-preferred*, or *basal*. As shown in **Table 5.1**, these groups vary more than ten-fold in integration preference. We next sought to test whether "weighing" each TA site based on the observed integration frequencies and summing the weighted scores of all TA sites in a given region could be used to predict the likelihood of integration into that region. For this we modified our script to bin the input sequence, tally each class of TA site, sum their relative weights using the preferences in **Table 5.1** as coefficients, and generate a "Total V_{step} score" for each bin. We called this Perl script *ProTIS*© (*Profiler of Transposon Insertion Sites*).

The distribution of TA sites for the entire pFV/Luc sequence is shown in **Figure 5.2A**. The sequence is divided into 100-bp bins and the numbers of each type of TA site within each bin are enumerated. The theoretical plot for the Total V_{step} scores for pFV/Luc is shown in **Figure 5.2B** and the actual distribution of integrations [1] is shown in **Figure 5.2C**. Two regions, Amp and Ori of pFV/Luc, are underrepresented (shaded regions) because insertions into these regions can disrupt the selection method for recovering events. When the entire sequence is divided into 100-bp bins, and the numbers of insertions sites into each bin are treated as events from a Poisson distribution, the experimental data, outside of Amp and Ori, show a statistically significant overlap with the Total V_{step} scores plot ($p<0.0001$).

As an alternative approach based on the apparent overlap in the distribution of TA dinucleotides in **Figure 5.2A** and the integration profile in 2c, we tested whether the TA-dinucleotide distribution alone would be an equally faithful predictor of integration sites. Similar significance of an overlap between the TA distribution and integration pattern was found using the aforementioned statistical method (*p<0.0001*). The residual deviance, however, is larger in this model and so the regression fit is inferior to the use of Total V_{step} scores when using the number of TAs. The Akaike Information Criterion (AIC) formally compares two model fits based on their likelihoods [27]. Fitting the model using a TA tally results in a larger AIC, 328.2, than using the Total V_{step}, 312.9. These results suggest that the Total V_{step} is the better predictor of insertion sites. Accordingly, using the training set of interplasmid transposition events and the Total V_{step} score, we identified a method that could potentially predict the outcomes of applied genetic studies using SB transposons.

Remobilization of transposons into the ninth intron of the mouse *Braf* gene

The key to identifying preferred sites in chromatin is to examine multiple integrations into a limited genomic region and quantify variations from Poisson statistics. Such data became available from a study in which the SB transposon, T2/Onc, was engineered to elicit gain-of-function mutations and accelerate tumor formation in somatic tissues of mice lacking the p19*Arf* tumor suppressor [5]. The most frequent oncogenic insertion site was intron-9 of the *Braf* gene. All of the 25 analyzed insertions in intron-9 were oriented toward the tenth exon (**Figure 5.3A**), resulting in a transcript encoding the kinase domain of Braf that acts as a dominant oncogene. Of the 347 potential TA-integration sites in the 4069-bp intron, 22 were targets and three sites (marked by asterisks) were hit twice. In this case, the probability of two insertions into a single TA site is 0.07 and the odds of this happening three times are 0.0004, which strongly suggested the existence of preferential insertion sites.

Because translation of the amino-terminally truncated Braf polypeptide is initiated from an internal start codon in exon-10, we assumed any T2/Onc insertion regardless of location or reading frame in *Braf* intron-9 would lead to oncogenic selection, and that the uneven distributions of insertions were the result of preferential target site selection. We thus identified these events as a dataset with which to test our method, and ran *ProTIS* © on the intron-9 sequence. The individual V_{step} profiles for T2/Onc-targeted sites in intron-9 are shown in **Figure 5.7**. **Table 5.1** shows the distribution of integrations into the various categories of profiles for the 347 sites. The 33 sites with preferred-site profiles had a 20% hit rate, the 105 predicted semi-preferred sites had an 11% hit rate, and the 209 basal sites showed only a 2% hit rate. These data strongly suggest a 5- to 10-fold preference for integrations at semi-preferred sites and preferred sites in intron-9 compared to basal sites. **Figure 5.3A** shows that the distribution of T2/Onc insertions into intron-9 matches the plot of Total V_{step} scores (**Figure 5.3B**). Using the same statistical procedure described for **Figure 5.2**, this overlap between the experimental data and the theoretical prediction are highly significant (*$p<0.0001$)*.

V_{step}-profiling of an extended chromosomal region

SB transposons resident in a mouse chromosome can be remobilized to new sites, most often within about 10 Mbp of their original locus [23,28-31], providing another source of densely localized transposon integrations. We thus examined 3.2 Mbp of mouse chromosome 1 (position 158,550,000 bp to 161,750,000 bp according to NCBI m33 build) in which 34 remobilized transposition events were mapped in the vicinity of a transgenic donor concatemer of SB transposons [5]. As shown in **Figure 5.4A**, this region (asterisk) is about 15 Mbp from the concatemer (arrow). In this region there are 208,299 TA sites corresponding to approximately one TA site per 15 bp. Of these TA sites, *ProTIS* © predicts 117,454 basal, 67,070 semi-preferred, and 23,775 preferred TA sites. The distribution of sites was divided into 32,000 100-bp bins, in terms of either map position (**Figure 5.4B**) or Total V_{step} score (**Figure 5.4C**). The average Total V_{step} score per 100-bp bin over the entire region is 23 (range from 0 to 435) and transposons inserted into intervals with an average score of 50 (range from 9 to 250). Thus, the insertions clearly are skewed towards the higher V_{step} values ($p<0.0001$). **Table 5.1** shows that the distribution of integrations into each V_{step} profile category is similar to the integration preferences observed in pFV/Luc and *Braf* intron-9.

Overall, the data from insertions into an active gene (*Braf*), a region of chromosome 1 comprising about 0.1% of the mouse genome, and a plasmid are remarkably consistent despite a 1,000-fold range in insertion density between pFV/Luc and 3.2 Mbp in chromosome 1. These results indicate that *ProTIS* © and its future derivatives will be valuable predictors of vector integration sites into genomes.

Application of the *ProTIS* © method to a genomic target of therapeutic vectors

The randomness of integration sites is an area under discussion with regard to vectors for gene therapy, including SB-based vectors [3,4,32]. However, the potential of severe adverse effects following random integration has been a concern [33,34]. In particular, two cases of acute lymphoblastic leukemia in children followed transfer of the IL-2 $\square_c$ gene in retrovirus-based vectors. Each apparently resulted from an insertional activation of the *LMO2* oncogene followed by selective outgrowth of the treated cells [35,36]. Because SB transposons are being developed as gene therapy vectors [4] and *LMO2*, out of more than 291 identified cancer genes [37], is associated with all of the severe adverse events in the IL-2 $\square_c$ trials, we examined the *LMO2* locus using *ProTIS*. Consequently, we analyzed the *LMO2* locus for potential transposon integration hotspots as a model for how the program can be applied to any genetic locus of interest. **Figure 5.5** shows the plot of Total V_{step} scores of 100 Kbp of genomic sequence containing the *LMO2* gene, with 50 kbp of upstream sequence, and the relative positions of the two activating retroviruses (P4, P5). *ProTIS*© predicts two sequences with prominent V_{step} scores, labeled 1 and 2, that derive from a simple tandem repeat, $(TCTA)_n$ and a 165-bp sequence that is replete with tandem $(TA)_n$ repeats, respectively. SB apparently has a ten-fold preference to land in microsatellite repeat regions containing TA dinucleotides [19], which is consistent with our findings that preferred sites such as $(TA)_n$ repeats have a 13-fold predicted preference using *ProTIS*© profiling. The *ProTIS* © plot of the *LMO2* locus suggests that SB vectors would target regions 1 and 2, which are more than 10 kbp from the transcriptional initiation site, and three times the distances of the activating proviruses, P4 and P5. Similar analyses can be done for any gene of interest.

Profiling other transposable elements and retroviruses.

Although SB was the first DNA-based transposable element developed to deliver DNA sequences into mammalian genomes, lepidopteran *piggyBac* transposons and *Drosophila P*-elements are powerful germline-transformation tools in insects [38,39]. Although both of these vectors have significantly strong preferences for transcriptional units, we hypothesized that they might exhibit target-site selection patterns related to DNA structure that would further define sites of integration within genes. Accordingly, we examined the integration-site sequence-tags deposited in Genbank from multiple investigations. The single largest deposit of integration sites was generated by Exelixis and comprises over 18,000 *piggyBac* and 6500 *P*-element insertions [21,40]. We refined the *piggyBac* data to 11,791 integrations that could be identified by the TTAA sequence recognized by *piggyBac* transposase and 5070 *P*-element integrations into validated genomic sequences. For both transposons we used the same procedure to identify preferential integration sites as we did for SB integrations: 1) find insertion hotspots, 2) develop rules based on these sequences and 3) test the rules against a much larger set of integrations. In contrast with what we found for SB transposons, there was no consistent V_{step} pattern shared amongst either the *piggyBac* or *P*-element integration sites (**Figures 5.8 and 5.9**).

Retroviruses have been utilized in genetic screens and for germline and somatic transgenesis in vertebrates for decades. Weak consensus sequences are found at the integration sites of several retroviruses [17,20], based upon the examination of relatively few integration sites scattered across a target genome. Using curated data kindly provided by Drs. Xiaolin Wu and Alex Holman, we examined 695 murine leukemia virus [20], 1371 human immunodeficiency virus-1[11,14], 148 simian immunodeficiency virus [14] and 551 avian sarcoma-leukosis virus [14,15] integration sites for V_{step} patterns that would aid in predicting integration preferences (**Figure 5.6**). As with *P*-elements, we found symmetric patterns that overlap with the base pairs involved in the target site duplication for most family members. Importantly, these patterns are based on the same compilations used to identify unique, weak consensus sequences for the various viruses [17,20] and cannot be used to generate algorithms alone. The indicated patterns shown in **Figure 5.6** suggest that V_{step} rules for identifying preferential integration sites might exist, but adequately dense sources of *in vivo* integration sites for these vectors, along with the identification of hotspots, are still required to generate appropriate algorithms.

Discussion

The observation that hotspots for SB transposon integration do not always match the published consensus sequence from different studies [1] led us to investigate other properties of sequences surrounding target sites. The data presented in this report confirm our hypothesis that SB transposase recognizes distinct structural features in DNA sequences, regardless of primary DNA sequence, that can be described by the V_{step} DNA-deformation parameter. Preferential TA-integration sites can be identified by specific V_{step} profiles of the DNA sequences flanking a TA site, regardless of whether the target sequence is a 100-bp segment of a plasmid, an entire plasmid, a portion of an actively transcribed gene, or bulk chromatin (**Table 5.1**). This method is more accurate than a simple distribution of TA sites in a target sequence. As transposon insertions approach saturation of a target genome, the *ProTIS* © algorithm will provide a closer approximation, in part, because some simple repeat sequences containing TA dinucleotides have a greater ability to attract SB transposons than other repeats containing TA dinucleotides. For instance, a 100-bp target consisting of the repeat $(TATC)_{25}$ translates into a Total V_{step} score of 325, whereas a 100-bp sequence consisting of the repeat $(TACT)_{25}$ has a Total V_{step} score of only 25. Each sequence represents an equal number of TA target sites and the same base composition, but when compared, translate into a 13-fold difference in the number of integrations that would be observed. Nevertheless, genomes are vast and non-uniform in terms of structure, protein associations, methylation, compaction, etc. Thus, it would not be surprising that in some cases predictions made by ProTIS will fail.

The application of SB to forward-genetic studies [5,6] has opened possibilities for the identification of novel genes that influence the formation of various tumor types. Repeated observation of transposon-induced mutations in the same gene in several different tumor samples identifies that gene as a candidate cancer gene. *ProTIS* © will be a valuable tool in this field, helping geneticists to distinguish between those events that are truly biologically significant common sites of integration from those events that are biased to be repeatedly tagged because of an abundance of preferred integration sites.

SB transposase has catalytic properties that are shared by other DDE-type recombinases, including retroviral integrases [41]. Consequently, we reasoned that these other enzymes might also have integration site preferences that are based on local DNA structure. However, even though thousands of integration sites have been recorded for various viral vectors, there are no reports of regions of chromatin that harbor densities of integrations that result in multiple integrations into a single site, a requirement for defining V_{step}-based rules for preferential integration sites. Thus, quantitative measurements that generate rules for prediction of structure-based preferential integration sites for *piggyBac* and *P*-element transposons as well as for retroviruses are not possible using this approach with currently available datasets. Although sequence-based assays for examining some retroviral integration patterns in defined targets have been developed [42-44], hundreds of integration sites for any vector will likely have to be generated to generate rules for predicting preferred sites. Many factors have been shown to influence the integration of retroviral and lentiviral integration including preferences to integrate into transcription units, gene expression profiles of the target cell genome, nucleosome packing of chromatin, sequence motifs such as CpG islands [45] and growth arrest of cultured cells in the case of HIV integration [46]. Our understanding of the contributions of these factors is insufficient for prediction of retroviral integration sites. Perhaps local DNA structure, as we have shown for *Sleeping Beauty* transposons, plays yet an additional role in defining preferential sequences for integration. For example, it may provide a mechanism by which HIV prefers to avoid integration into or near

CpG islands because the structure of dimer steps in the CpG sequence is not favorable to integration. Validation of this hypothesis requires a substantial dataset of numerous integrations into a small, defined target sequence to identify specific V_{step} patterns common to the most preferred insertion sites. Otherwise, V_{step} analyses provide essentially the same information as a consensus sequence.

Our examination of SB transposon integrations in about 0.1% of the euchromatic genome (**Table 5.1**) suggest that of the approximate 200 million TA sites in the mouse genome, about 10% (20 million) will be preferred sites that would account for 55% of transposon insertions, whereas 120 million (60%) basal TA sites would attract only 5% of transposon insertions. We expect the same results in humans. Thus, although SB transposons can integrate into practically any TA site, within a given region about half will go to only 10% of the available sites. This information is important for evaluating SB transposons for both insertional mutagenesis and as a vector for gene therapy.

Our analysis of integration sites is applicable to understanding the biology of other transposons whose consensus preferences are already known. For example, the *Tc1* transposon in *C. elegans* that integrates into TA sites has a consensus sequence GA(G/T)(A/G)**TA**(T/C)(G/C)T [47,48]. One hotspot, TGGTG**TA**TGTCT, was hit 51 times in 166 mapped insertions [49]. V_{step} analyses of the consensus and hot spot match the most preferred category for SB transposition. In contrast, the integration consensus sequence for a related *C. elegans* transposon, *Tc3,* does not match that of *Tc1* and the V_{step} profile of both its consensus and most preferred integration site, ACTAA**TA**TTATG, are distinctly different from *Tc1* and SB [49,50]. Specifically, there is extra spacing in the most preferred *Tc3* profile on both sides of the TA peak compared to the profiles for *Tc1* and SB (**Figure 5.10**). Likewise, some of the hottest sites for *Drosophila Himar1* integration [51] also match the V_{step} profiles of SB and *Tc1*.

Repetitive (mobile) elements play a significant role in genome evolution [52-54]. For instance, the most prominent differences in the human and chimpanzee genomes are rates of transposable element insertions and new insertions of novel retroviral elements [55]. Until now, parameters governing the integration of transposons and proviruses have been ignored. By identifying preferences for the different classes of repetitive elements, it should be possible to determine the role(s) of natural selection on newly introduced elements by comparing their observed distributions compared with the theoretical expectations. Because viral elements comprise a significant proportion of mammalian genomes, further work in identifying the rules for their integration preferences will be of interest to those studying evolution as well as those interested introducing new genetic sequences into genomes for functional genomic studies and therapeutic purposes.

Acknowledgements

We thank the Arnold and Mabel Beckman Foundation for support of our work and all members of the Beckman Center for Transposon Research for a long history of contributions of ideas and results. We are especially grateful to Drs Michael Simmons, Reuben Harris, William A. Weiss (UCSF), and Elena Aronovich for careful reading of the manuscript, Dr. Keiko Akagi (NCI) for mapping the T2/Onc integrations, Michael H.L Lee for technical assistance, and Joe Carlson (Berkeley Drosophila Genome Project) for sending *P*-element datasets. The authors were also supported by NIH grants 1PO1 HD32652-07 and R43 HL076908-01 (PBH) and DA014764

(DAL) and T32 HD007480 (AMG), NIGMS GM008568 (CSH) and NCI T32CA09138-28 (LSC & CMC) and F32CA10619-01 (LSC).

References

1. Liu, G. et al. Target-site preference for Sleeping Beauty transposons. *J. Mol. Biol.* **346**, 161-173 (2005).
2. Ivics, Z. & Izsvak, Z. Transposable elements for transgenesis and insertional mutagenesis in vertebrates: a contemporary review of experimental strategies. *Meth. Mol. Biol.* **260**, 255-276 (2004).
3. Izsvak, Z. & Ivics, Z. Sleeping Beauty transposition: biology and applications for molecular therapy. *Mol. Therap.* **9**, 147-156. (2004).
4. Hackett, P.B., Ekker, S.C., Largaespada, D.A. & McIvor, R.S. *Sleeping Beauty* transposon-mediated gene therapy for prolonged expression. *Adv. Genet.* **54**, 187-229 (2005).
5. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6 (2005).
6. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6 (2005).
7. Carlson, C.M., Frandsen, J.L., Kirchhof, N., McIvor, R.S. & Largaespada, D.A. Somatic integration of an oncogene-harboring Sleeping Beauty transposon models liver tumor development in the mouse. *Proc Natl Acad Sci U S A* **102**, 17059-64 (2005).
8. Keng, V.W. et al. Region-specific saturation germline mutagenesis in mice using the Sleeping Beauty transposon system. *Nat Methods* **2**, 763-9 (2005).
9. Mikkers, H. & Berns, A. Retroviral insertional mutagenesis: tagging cancer pathways. *Adv. Cancer Res.* **88**, 53-99 (2003).
10. Zhang, P. & Spradling, A.C. Insertional mutagenesis of Drosophila heterochromatin with single P elements. *Proc Natl Acad Sci U S A* **91**, 3539-43 (1994).
11. Schroder, A.R.W. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-529 (2002).
12. Wu, X., Li, Y., Crise, B. & Burgess, S.M. Transcription start regions in human genome are favored targets for MLV integration. *Science* **300**, 1749-1751 (2003).
13. Nakai, H. et al. AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nature Genet.* **34**, 297-302 (2003).
14. Mitchell, R.S. et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLOS* **2**, 1127-1136 (2004).
15. Narezkina, A. et al. Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78**, 11656-11663 (2004).
16. Maxfield, L.F., Fraize, C.D. & Coffin, J.M. Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc. Natl. Acad. Sci. USA* **102**, 1436-1441 (2005).
17. Holman, A.G. & Coffin, J.M. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc. Natl. Acad. Sci. USA* **102**, 6103-6107 (2005).
18. Hematti, P. et al. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol* **2**, e423 (2004).

19. Yant, S.R. et al. High-resolution genome-wide mapping of transposon integration in mammals. *Mol. Cell. Biol.* **25**, 2085-2094 (2005).
20. Wu, X., Li, Y., Crise, B., Burgess, S.M. & Munroe, D.J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79**, 5211-5214. (2005).
21. Liao, G.C., Rehm, E.J. & Rubin, G.M. Insertion site preferences of the P transposable element in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **97**, 3347-51 (2000).
22. Vigdal, T.J., Kaufman, C.D., Izsvak, Z., Voytas, D.F. & Ivics, Z. Common physical properties of DNA affecting target site selection of Sleeping Beauty and other Tc1/mariner transposable elements. *J. Mol. Biol.* **323**, 411-452 (2002).
23. Carlson, C.M. et al. Transposon mutagenesis of the mouse germline. *Genetics* **165**, 243-256 (2003).
24. Olson, W.K. & Zhurkin, V.B. Modeling DNA deformations. *Curr. Opin. Struct. Biol.* **10**, 286-297. (2000).
25. Plasterk, R.H.A., Izsvák, Z. & Ivics, Z. Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet.* **15**, 326-332 (1999).
26. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. & Zhurkin, V.B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Nat. Acad. Sci. USA* **95**, 11163-11168. (1998).
27. Akaike, H. Prediction and entropy. *A Celebration of Statistics (A.C. Atkinson and S.E. Fienberg, eds.)*, 1-24 (1985).
28. Dupuy, A.J., Fritz, S. & Largaespada, D.A. Transposition and gene disruption using a mutagenic transposon vector in the male germline of the mouse. *Genesis* **30**, 82-88 (2001).
29. Dupuy, A.J. et al. Mammalian germ-line transgenesis by transposition. *Proc. Natl. Acad. Sci. USA* **99**, 4495-4499 (2002).
30. Horie, K. et al. Efficient chromosomal transposition of a Tc1/mariner- like transposon Sleeping Beauty in mice. *Proc. Nat. Acad. Sci. USA* **98**, 9191-9196 (2001).
31. Horie, K. et al. Characterization of *Sleeping Beauty* transposition and its application to genetic screening in mice. *Mol. Cell. Biol.* **23**, 9189-9207 (2003).
32. Essner, J.J., McIvor, R.S. & Hackett, P.B. Awakening of gene therapy with *Sleeping Beauty* transposons. *Curr. Opin. Pharmacol.* **5**, (in press) (2005).
33. Yi, Y., Hahm, S.H. & Lee, K.H. Retroviral gene therapy: safety issues and possible solutions. *Curr. Gene Therap.* **5**, 25-35 (2005).
34. Baum, C. et al. Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. *Mol. Therap.* **9**, 5-13. (2004).
35. Dave', U.P., Jenkins, N.A. & Copeland, N.G. Gene therapy insertional mutagenesis insights. *Science* **303**, 33. (2004).
36. Hacein-Bey-Abina, S. et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415-419. (2003).
37. Futreal, P.A. et al. A census of human cancer genes. *Nature Rev. Cancer* **4**, 177-183 (2004).
38. Ryder, E. & Russell, S. Transposable elements as tools for genomics and genetics in Drosophila. *Brief Funct Genomic Proteomic* **2**, 57-71 (2003).
39. Handler, A.M. Use of the piggyBac transposon for germ-line transformation of insects. *Insect Biochem Mol Biol* **32**, 1211-20 (2002).

40. Thibault, S.T. et al. A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. *Nat Genet* **36**, 283-7 (2004).
41. Craig, N.L. Target site selection in transposition. *Annu Rev Biochem* **66**, 437-74 (1997).
42. Pryciak, P.M., Muller, H.P. & Varmus, H.E. Simian virus 40 minichromosomes as targets for retroviral integration in vivo. *Proc Natl Acad Sci U S A* **89**, 9237-41 (1992).
43. Pryciak, P.M., Sil, A. & Varmus, H.E. Retroviral integration into minichromosomes in vitro. *Embo J* **11**, 291-303 (1992).
44. Fitzgerald, M.L. & Grandgenett, D.P. Retroviral integration: in vitro host site selection by avian integrase. *J Virol* **68**, 4314-21 (1994).
45. Bushman, F. et al. Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3**, 848-58 (2005).
46. Ciuffi, A. et al. Integration Site Selection by HIV-Based Vectors in Dividing and Growth-Arrested IMR-90 Lung Fibroblasts. *Mol Ther* **13**, 366-73 (2006).
47. Eide, D. & Anderson, P. Insertion and excision of *Caenorhabditis elegans* transposable element *Tc1*. *Mol. Cell biol.* **8**, 737-746 (1988).
48. Mori, I., Benian, G.M., Moerman, D.G. & Waterston, R.H. Transposable element *Tc1* of *Caenorhabditis elegans* recognizes specific target sequences for integration. *Proc. Natl. Acad. Sci. USA* **85**, 861-864 (1985).
49. van Luenen, H.G.A.M. & Plasterk, R.H.A. Target site choice of the related transposable elements *Tc1* and *Tc3* of *Caenorhabditis elegans*. *Nucl. Acids Res.* **22**, 262-269 (1994).
50. Preclin, V., Martin, E. & Segalat, L. Target sequences of *Tc1, Tc3* and *Tc5* transposons of *Caenorhabditis elegans. Genet. Res.* **82**, 85-88 (2003).
51. Lampe, D.J., Grant , T.E. & Robertson, H. Factors affecting transposition of the *Himar1 mariner* transposon in vitro. *Genetics* **149**, 179-187 (2006).
52. Britten, R.J. Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc Natl Acad Sci U S A* **101**, 16825-30 (2004).
53. Charlesworth, B., Sniegowski, P. & Stephan, W. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**, 215-20 (1994).
54. Shapiro, J.A. & von Sternberg, R. Why repetitive DNA is essential to genome function. *Biol Rev Camb Philos Soc* **80**, 227-50 (2005).
55. Mosse, Y.P. et al. High-resolution detection and mapping of genomic DNA alterations in neuroblastoma. *Genes Chromosomes Cancer* **43**, 390-403 (2005).

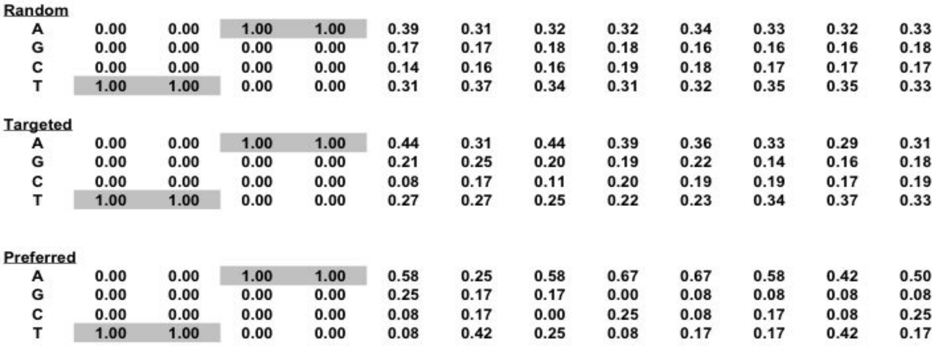
Table 5.1. SB transposition-site preferences as a function of V_{step} profiles.

V_{step} Pattern	(% of total)	Hit	Site	Preference
pFV/Luc:				
Basal	299 (61%)	39	0.13	**1X**
Semi-Preferred	154 (31%)	92	0.60	**5X**
Preferred	36 (7%)	62	1.7	**13X**
Braf Intron-9:				
Basal	209 (60%)	5	0.02	**1X**
Semi-Preferred	105 (19%)	12*	0.11	**6X**
Preferred	33 (10%)	8**	0.2	**10X**
3.2 Mbp Chromosome 1:				
Basal	117,454 (56%)	5	0.00004	**1X**
2.5-peak	67,070 (32%)	15	0.00022	**6X**
Preferred	23,775 (11%)	14	0.00059	**15X**

* 11 sites were hit; one was hit twice for a total of 12 hits.

**6 sites were hit; two were hit twice for a total of 8 hits.

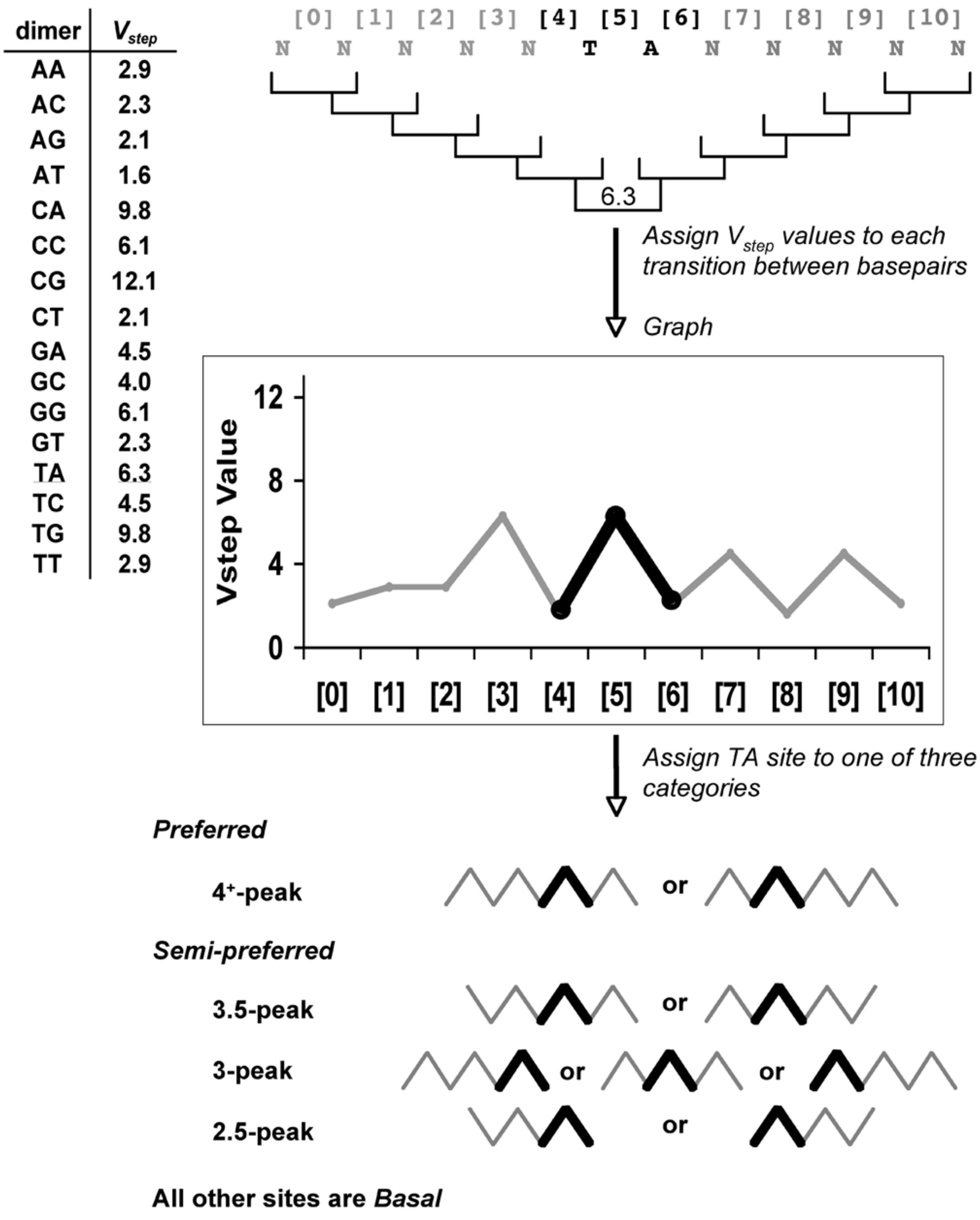
Table 5.2: Sequences of random, targeted and preferred piggyBac sites



Random											
A	0.00	0.00	1.00	1.00	0.39	0.31	0.32	0.32	0.34	0.33	0.32	0.33
G	0.00	0.00	0.00	0.00	0.17	0.17	0.18	0.18	0.16	0.16	0.16	0.18
C	0.00	0.00	0.00	0.00	0.14	0.16	0.16	0.19	0.18	0.17	0.17	0.17
T	1.00	1.00	0.00	0.00	0.31	0.37	0.34	0.31	0.32	0.35	0.35	0.33
Targeted												
A	0.00	0.00	1.00	1.00	0.44	0.31	0.44	0.39	0.36	0.33	0.29	0.31
G	0.00	0.00	0.00	0.00	0.21	0.25	0.20	0.19	0.22	0.14	0.16	0.18
C	0.00	0.00	0.00	0.00	0.08	0.17	0.11	0.20	0.19	0.19	0.17	0.19
T	1.00	1.00	0.00	0.00	0.27	0.27	0.25	0.22	0.23	0.34	0.37	0.33
Preferred												
A	0.00	0.00	1.00	1.00	0.58	0.25	0.58	0.67	0.67	0.58	0.42	0.50
G	0.00	0.00	0.00	0.00	0.25	0.17	0.17	0.00	0.08	0.08	0.08	0.08
C	0.00	0.00	0.00	0.00	0.08	0.17	0.00	0.25	0.08	0.17	0.08	0.25
T	1.00	1.00	0.00	0.00	0.08	0.42	0.25	0.08	0.17	0.17	0.42	0.17

Figure 5.1. Profiling TA sites using the V_{step} algorithm. Sequences of twelve base pairs (N) with TA sites at positions six and seven were analyzed with respect to the eleven V_{step} values ([0]-[10]) for transitions from one base pair to the next (brackets). Profiles are charted and subsequently assigned to one of three categories, *preferred*, *semi-preferred*, or *basal*, based upon the graphical pattern. In all profiles there is a "TA-peak" that always exists in such profiles because the T-to-A V_{step} value is 6.3 and all steps from N to T and from A to N (N = any base) are always less than 3.0, as shown on the left side of the figure. The "TA peak" formed by the two lines that connect the three V_{step} values for the N-to-T, T-to-A, and A-to-N steps are shown in bold.

Fig 5.1



dimer	V_{step}
AA	2.9
AC	2.3
AG	2.1
AT	1.6
CA	9.8
CC	6.1
CG	12.1
CT	2.1
GA	4.5
GC	4.0
GG	6.1
GT	2.3
TA	6.3
TC	4.5
TG	9.8
TT	2.9

Figure 5.2. Total V_{step} profile of the 7758 bp plasmid pFV/Luc. The sequence is divided into 78 100-bp bins. (a) Plot of the number of each type of TA site per bin. The hexagon indicates the Chinook salmon poly(A) addition motif and the following square indicates an M13 origin of replication. (b) Plot of Total V_{step} score per bin. **(c)** Distribution of observed insertion sites (adapted from Liu *et al.* [ref. 1]). Shaded areas are regions required for selection and thus unlikely to be scored. The asterisks indicate the three most likely regions for integration based on *ProTIS* analysis and the arrow indicates a region that has a high number of TA sites, but relatively few integrations.

Fig 5.2

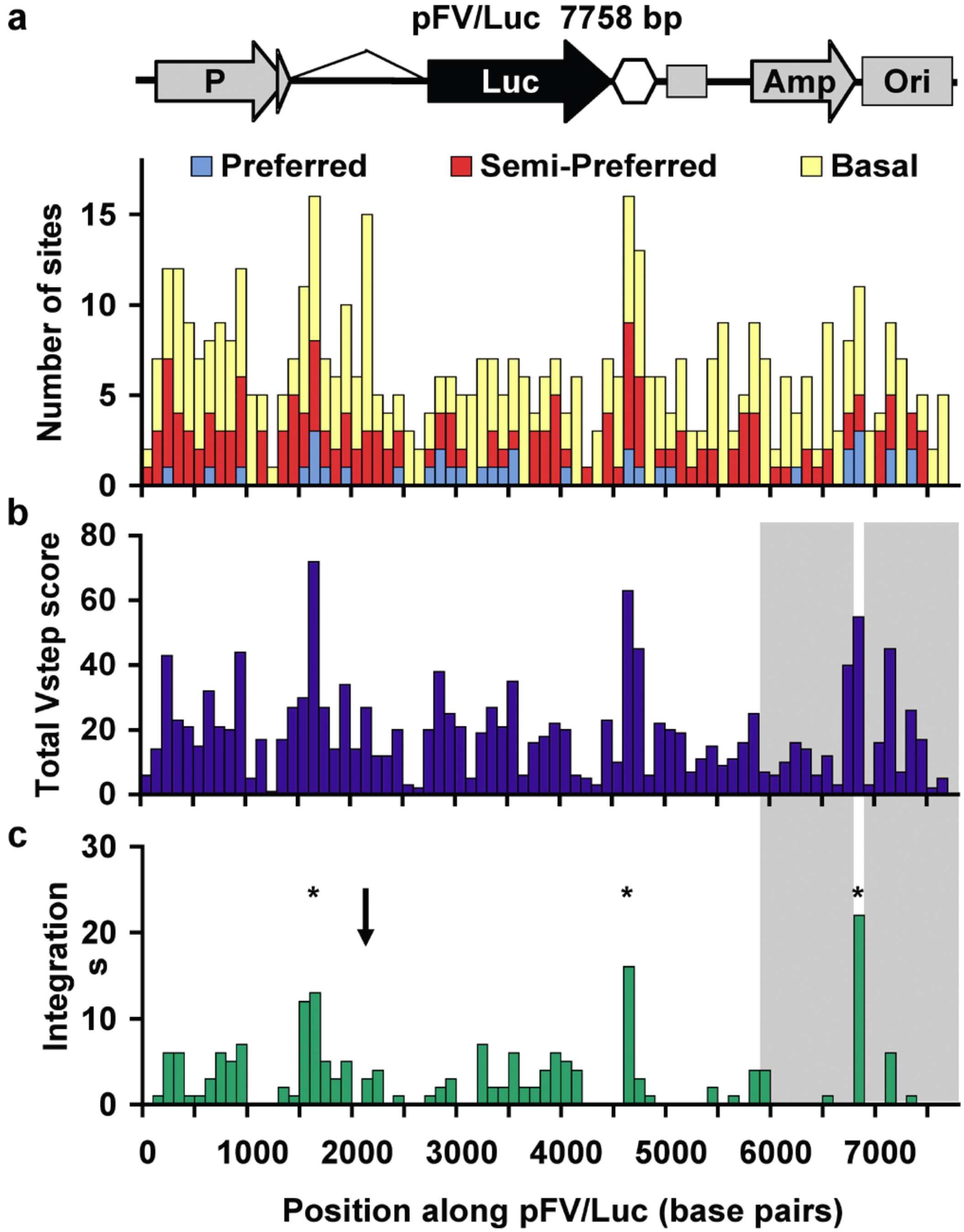

a
pFV/Luc 7758 bp
P
Luc
Amp
Ori
Preferred
Semi-Preferred
Basal
Number of sites
0
5
10
15
b
Total Vstep score
0
20
40
60
80
c
Integrations
0
10
20
30
*
*
*
0
1000
2000
3000
4000
5000
6000
7000
Position along pFV/Luc (base pairs)

Figure 5.3. V_{step} analysis of insertion sites of T2/Onc into the mouse *Braf* gene. (a) Schematic of mapped insertions into *Braf* (exons shown as tall vertical lines) with an expanded intron-9. Only T2/Onc transposons that integrated in a left-to-right orientation would be identified in the genetic screen. SA, splice-acceptor site, SD, splice-donor site, LTR, retroviral long terminal repeat, double arrowheads, inverted terminal repeats of the integrating transposon. The long arrow represents the direction of transcription from the LTR promoter within T2/Onc. (b) Total V_{step} profile of intron-9 in terms of 82 50-bp bins.

Fig 5.3

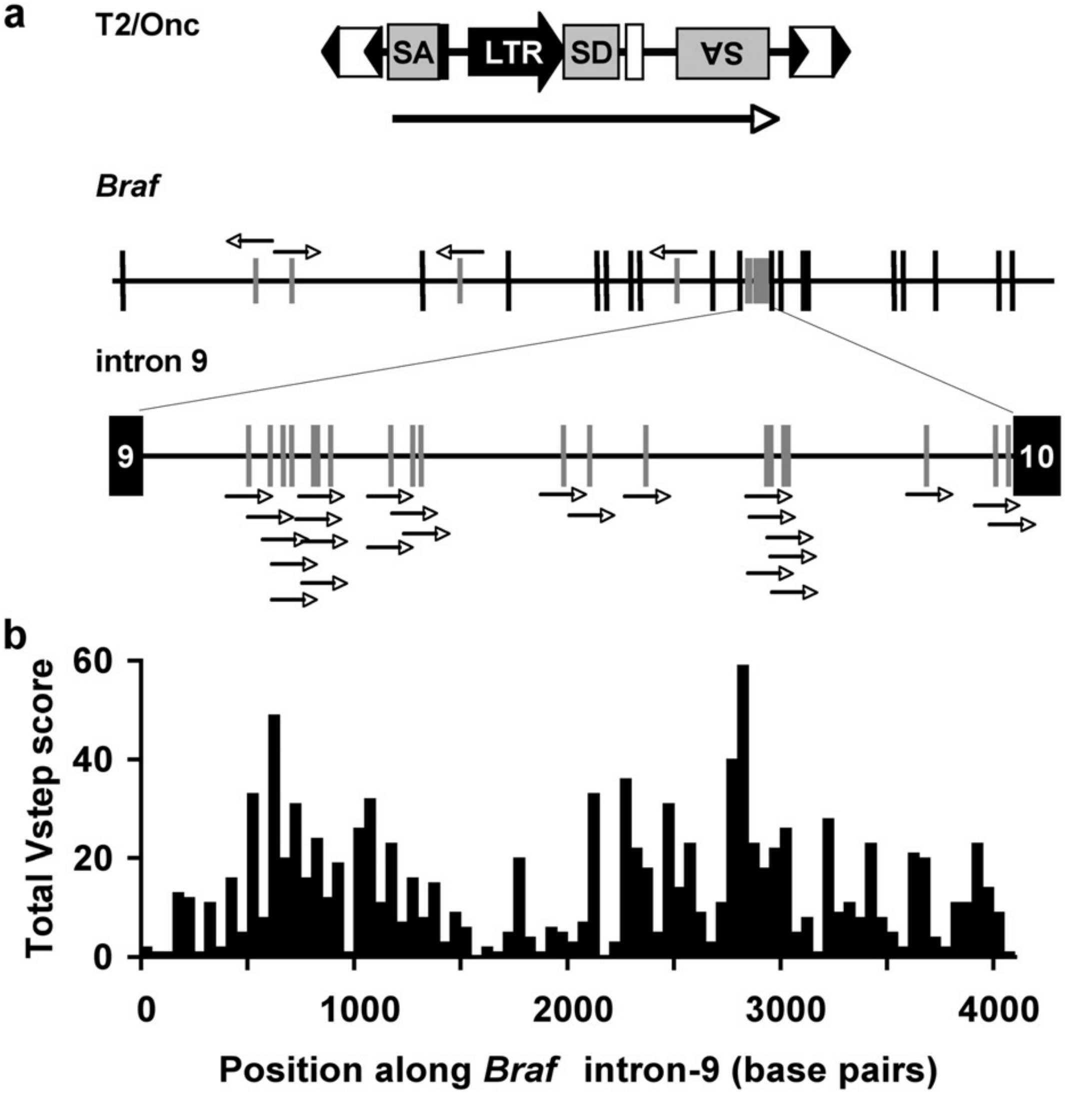

a
T2/Onc
SA
LTR
SD
SA
Braf
intron 9
9
10
b
Total Vstep score
0
20
40
60
0
1000
2000
3000
4000
Position along *Braf* intron-9 (base pairs)

Figure 5.4. Transposon insertion sites in 3.2 Mbp of mouse chromosome 1. a) SB integration sites in Chromosome 1, the locations of the concatemer from which the transposons were remobilized (↓) and the 3.2 Mbp region that had the highest density of integrations is marked with an asterisk. Region *(b)* was divided into 32,000 100-bp bins and the Total V_{step} scores for each bin calculated as described in Fig. 3. The average Total V_{step} value per bin is 23. b) Blue bars, Total V_{step} scores/bin; red bars, insertion sites mapped as a function of position. c) Insertion sites (red) displayed as a function of Total V_{step} score/bin (blue).

Fig 5.4

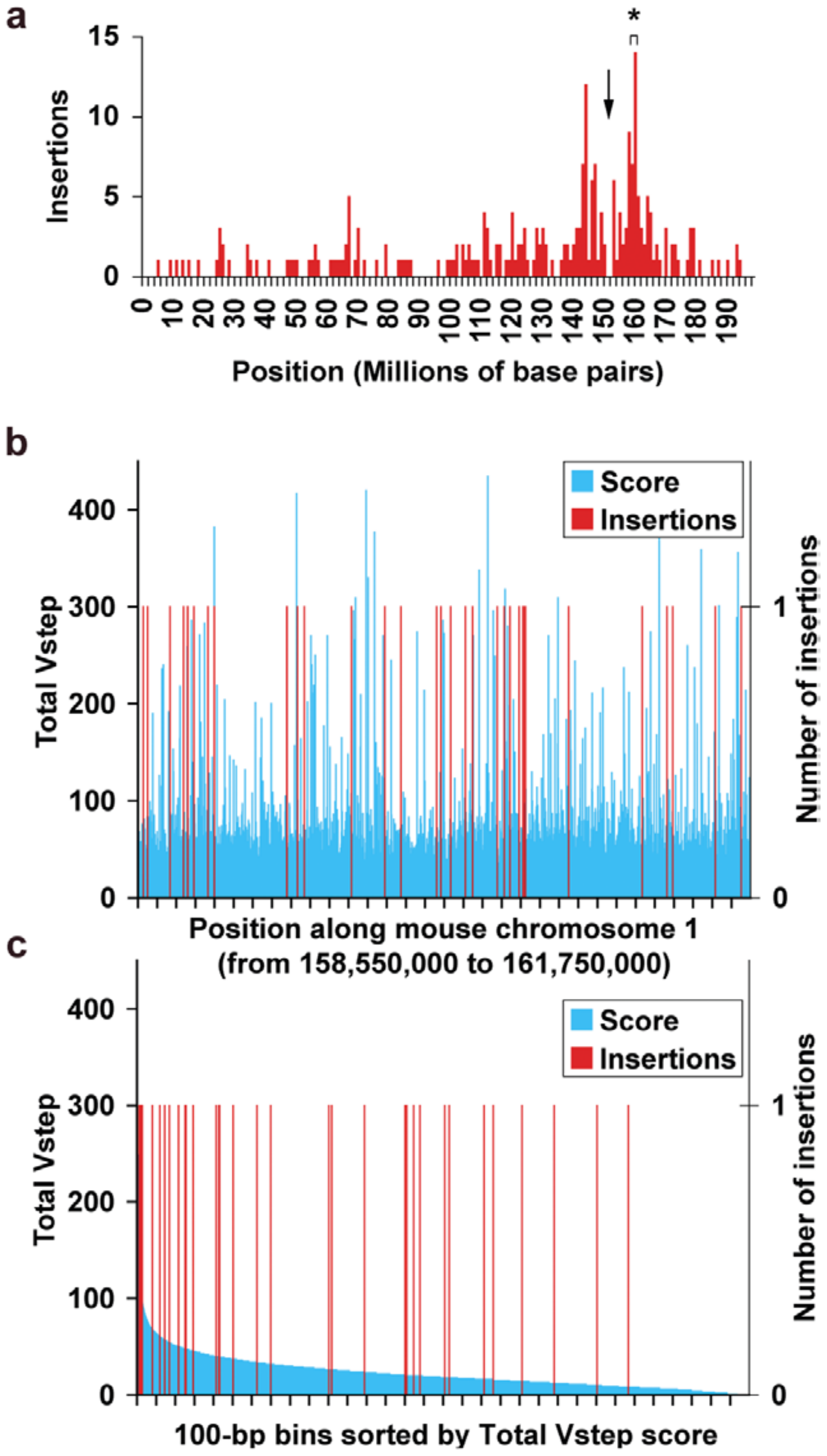
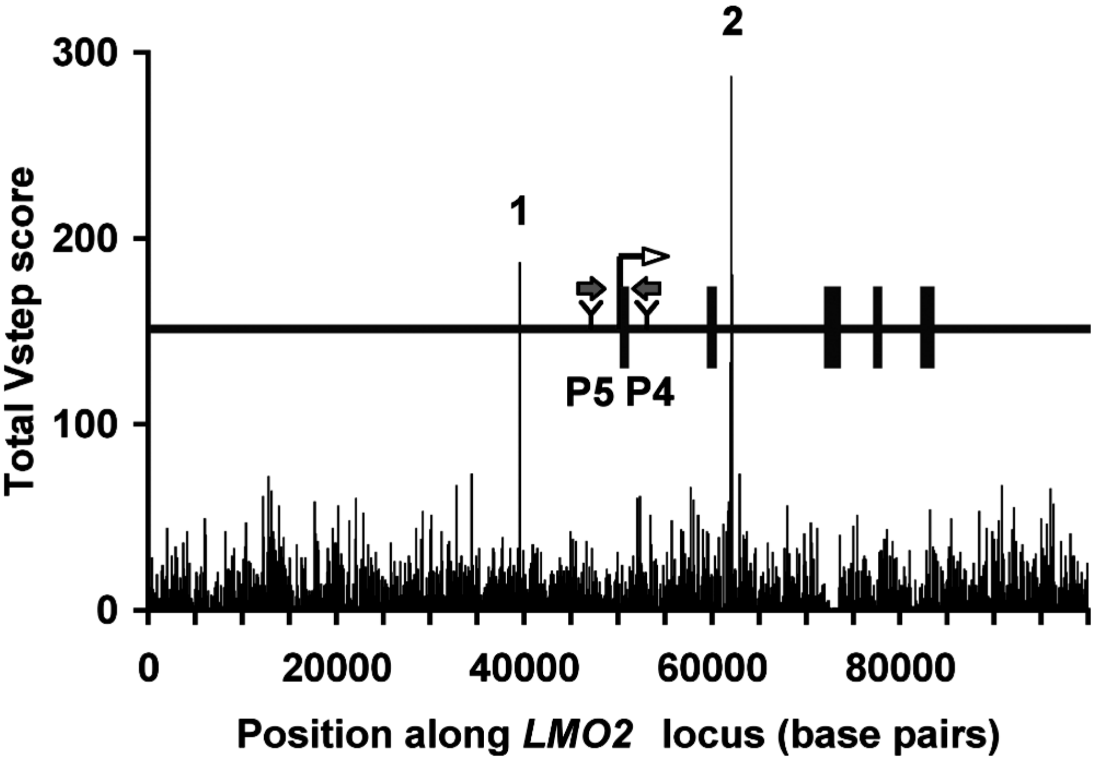

a
15
10
5
0
Insertions
*
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
Position (Millions of base pairs)
b
400
300
200
100
0
Total Vstep
Score
Insertions
1
0
Number of insertions
Position along mouse chromosome 1
(from 158,550,000 to 161,750,000)
c
400
300
200
100
0
Total Vstep
Score
Insertions
1
0
Number of insertions
100-bp bins sorted by Total Vstep score

Figure. 5.5. Total V_{step} Profile for the human *LMO2* gene plotted as 100-bp bins. The map of 100 Kbp of the *LMO2* locus is shown above the center of the V_{step} profile. Rectangles, exons; red arrowheads, sites of two activating retroviral insertions (P4 and P5 [ref 36]). Spikes 1 and 2 in the Total V_{step} profile correspond to short tandem repeats of $(TCTA)_n$ and $(TA)_n$ respectively.

Fig 5.5



2
1
300
200
100
0
Total Vstep score
P5 P4
0
20000
40000
60000
80000
Position along *LMO2* locus (base pairs)

Figure. 5.6. V_{step} analysis of insertion sites of proviruses and transposons. The arrows in the profiles indicate the boundaries of the TSD sequence that occurs with the staggered cuts made by the various integrase enzymes. (a) Average V_{step} profiles for 573 SB transposon integrations, (b) Average V_{step} profiles for murine leukemia virus, (c) Average V_{step} profiles for human immunodeficiency virus, (d) Average V_{step} profiles for simian immunodeficiency virus, (e) Average V_{step} profiles for avian sarcoma/leucosis virus. (f) Average V_{step} profiles for 1,006 random DNA 20-mer sequences.

Fig 5.6

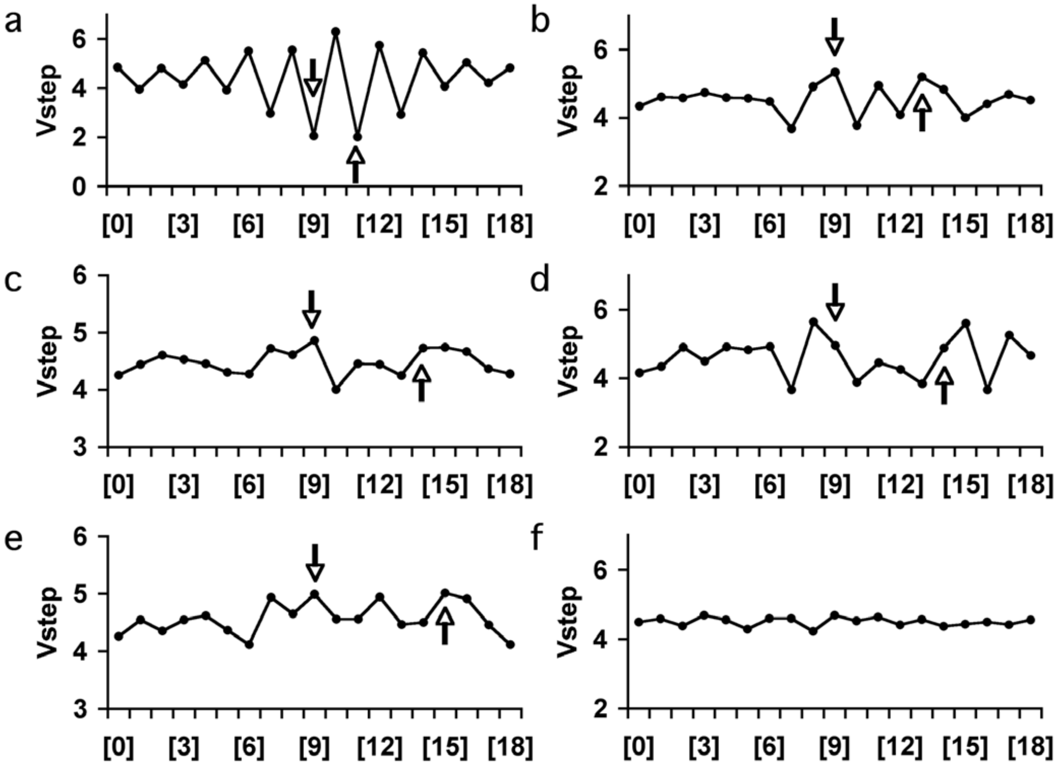


Figure 5.7. V_{step} profiles of sites in *Braf* intron-9 into which T2/Onc transposons integrated. The sites of integration, TA-peaks that are in the center (with a peak value of 6.3) are shaded. The first column shows the profiles of the six preferred sites that were hit, two of which had two insertions (marked with asterisks in the upper right-hand corners). Eleven semi-preferred sites were hit, one of which was hit twice (asterisk in the upper right-hand corner). The third column shows the V_{step} profiles of the five basal sites that were hit. The basal site at 3626 does not meet the criteria for a preferred or semi-preferred site due to the absence of an additional half or full peak on either end of the center pattern of three peaks. Each profile is identified by the base pair position of the *T* in the *TA* integration site.

Fig 5.7

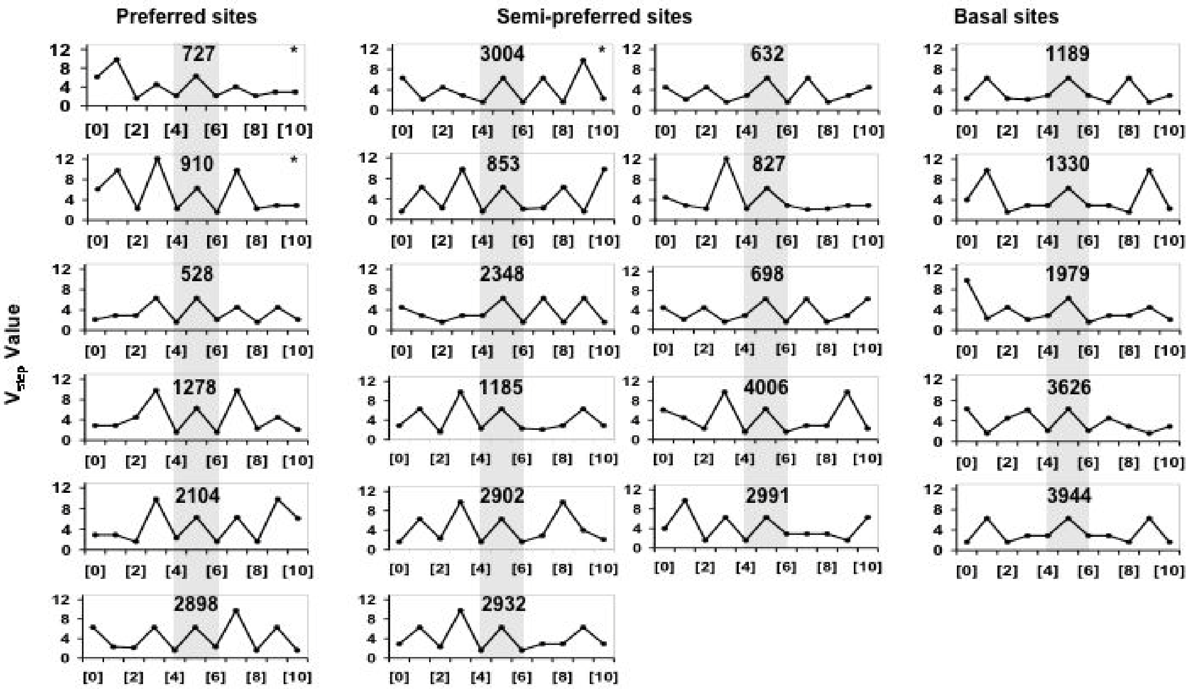

Preferred sites
Semi-preferred sites
Basal sites
V_{step} Value
727
*
3004
*
632
1189
910
*
853
827
1330
528
2348
698
1979
1278
1185
4006
3626
2104
2902
2991
3944
2898
2932
12
8
4
0
[0]
[2]
[4]
[6]
[8]
[10]

Figure 5.8. *piggyBac* transposon insertion sites. (a) The V_{step} profiles of the twelve Drosophila loci most frequently hit by the *piggyBac* transposon are shown with their corresponding number of integrations in *italics*. The sequence of each twelve base pair site is shown on the abscissa. The boxed region encloses the constant profile of the central TTAA-integration site into which the transposon integrated. (b) V_{step} profiles of 11,796 characterized integrations of *piggyBac* transposons (half-profiles) from the Exelixis dataset and 25,794 random TTAA sites (full profiles). The TTAA transposase recognition sequence is boxed in each profile. (c) A-philicity profiles of the same characterized integration and random sites.

Fig 5.8

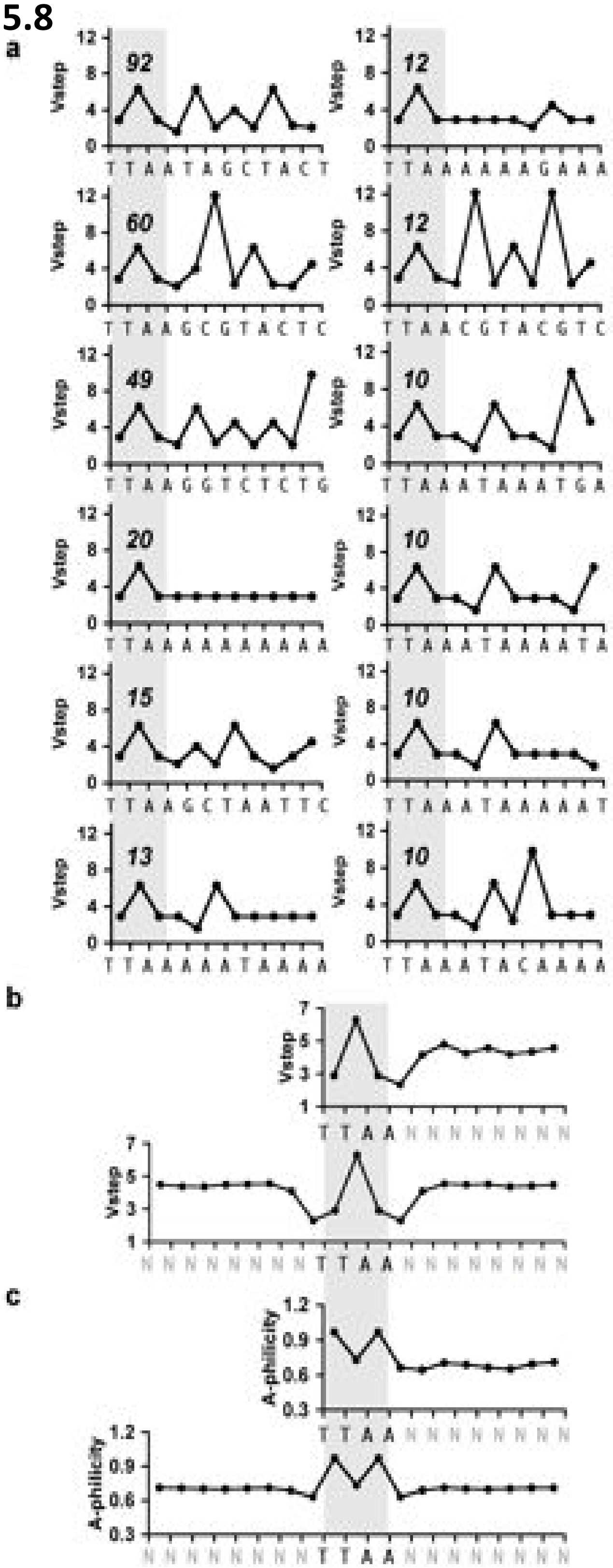

a
Vstep
0
4
8
12
92
T T A A T A G C T A C T
12
T T A A A A A A G A A A
60
T T A A G C G T A C T C
12
T T A A C G T A C G T C
49
T T A A G G T C T C T G
10
T T A A A T A A A T G A
20
T T A A A A A A A A A A
10
T T A A A T A A A A T A
15
T T A A G C T A A T T C
10
T T A A A T A A A A A T
13
T T A A A A A T A A A A
10
T T A A A T A C A A A A
b
Vstep
1
3
5
7
T T A A N N N N N N N N
N N N N N N N N T T A A N N N N N N N N
c
A-philicity
0.3
0.6
0.9
1.2
T T A A N N N N N N N N
N N N N N N N N T T A A N N N N N N N N

Figure 5.9. *P*-element Transposon insertion sites. a) The V_{step} profiles of the twenty most frequently hit Drosophila loci by the *P*-based vectors from the Exelixis dataset are shown with their corresponding number of integrations in *italics*. The sequence of the target site duplication is shown on the abscissa. b) The average V_{step} and A-philicity profiles of 3,752 *P* integration target site duplications. c) The average V_{step} and A-philicity profiles of 1,318 *P* integration target site duplications

Fig 5.9

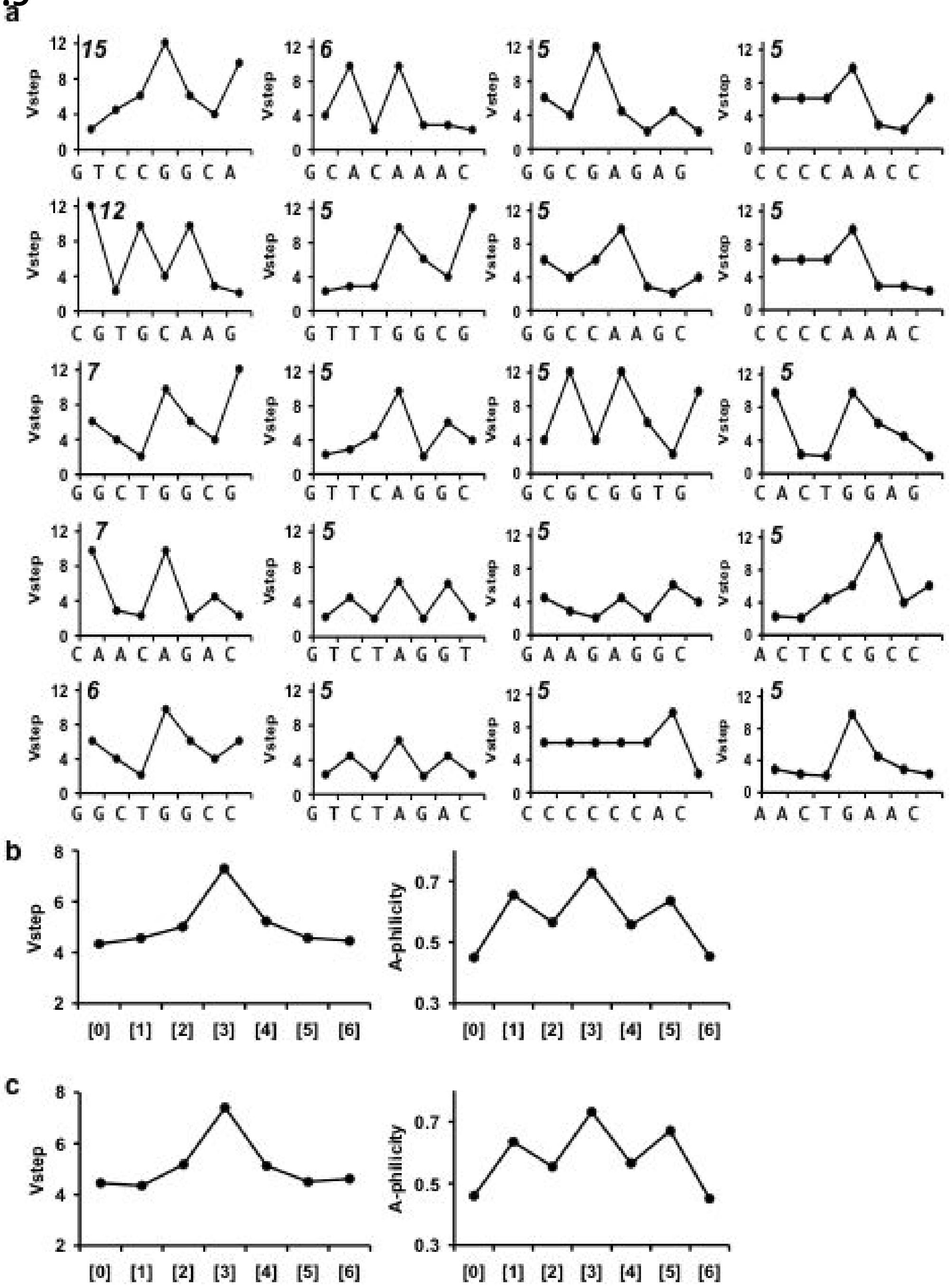
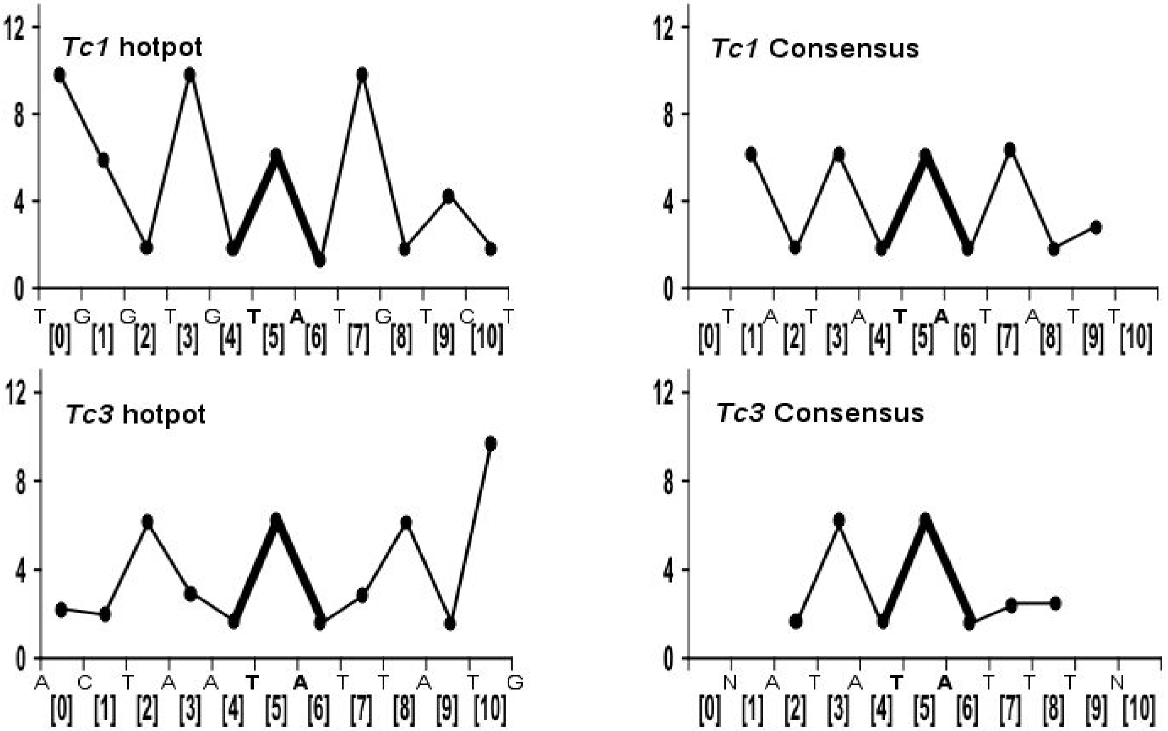

a
15
G T C C G G C A
6
G C A C A A A C
5
G G C G A G A G
5
C C C C A A C C
12
C G T G C A A G
5
G T T T G G C G
5
G G C C A A G C
5
C C C C A A A C
7
G G C T G G C G
5
G T T C A G G C
5
G C G C G G T G
5
C A C T G G A G
7
C A A C A G A C
5
G T C T A G G T
5
G A A G A G G C
5
A C T C C G C C
6
G G C T G G C C
5
G T C T A G A C
5
C C C C C C A C
5
A A C T G A A C
Vstep
0
4
8
12
b
Vstep
2
4
6
8
A-philicity
0.3
0.5
0.7
[0] [1] [2] [3] [4] [5] [6]
c
Vstep
2
4
6
8
A-philicity
0.3
0.5
0.7
[0] [1] [2] [3] [4] [5] [6]

Figure 5.10. V_{step} profiles of preferred and consensus *Tc1 and Tc3* Transposon insertion sites [refs 46, 47]. The sequences are listed below each profile and the dimer step numbers are in brackets. The TA integration sites are indicated by the bold lines. *N* indicates no strong preferential nucleotide in the *Tc3* consensus. The reported consensus sequences are two basepairs shorter than the sequences used for most of the analyses in this report.

Fig 5.10


12
8
4
0
Tc1 hotpot
T G G T G T A T G T C T
[0] [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]
12
8
4
0
Tc1 Consensus
T A T A T A T A T T
[0] [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]
12
8
4
0
Tc3 hotpot
A C T A A T A T T A T G
[0] [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]
12
8
4
0
Tc3 Consensus
N A T A T A T T T N
[0] [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

Chapter 6: Predicting Preferential DNA Vector Insertion Sites: Implications for Functional Genomics and Gene Therapy

Source: The following chapter was published in a special issue of Genome Biology in 2007 (PMID: 18047689)

Contributions: I performed most of the data analysis and literature review in the paper, and wrote most of the manuscript. Aron Geurts performed some of the data collection and analysis, reviewed all of the data analysis, and helped write the manuscript. PB Hackett supervised the project and wrote the section of the manuscript on delivery methods.

Predicting Preferential DNA Vector Insertion Sites: Implications for Functional Genomics and Gene Therapy

Christopher S. Hackett[1], Aron M. Geurts[2] & Perry B. Hackett[3]

[1]Biomedical Sciences Graduate Program and Department of Neurology, University of California San Francisco, San Francisco, CA 94143-0663

[2]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI 53226

[3]Department of Genetics, Cell Biology, and Development, The Arnold and Mabel Beckman Center for Transposon Research, Gene Therapy Program, and Cancer Center, University of Minnesota, Minneapolis, MN 55455

Key words: DNA structure, gene therapy, insertional mutagenesis, P-elements, *piggyBac*, *Sleeping Beauty*, Tol2, transposons

Abbreviations: SB, *Sleeping Beauty*; PB *piggyBac*; PE, P-elements

For Correspondence:

Perry B. Hackett

Department of Genetics, Cell Biology, and Development

University of Minnesota

6-160 Jackson Hall

321 Church St SE

Minneapolis, MN 55455

Tel: 612-624-6736

Fax: 612-625-6140

Abstract

Viral and transposon vectors, which are able to introduce genetic constructs into mammalian chromatin, have been employed for gene therapy as well as functional genomics studies. However, the goals of gene therapy and functional genomics are entirely different – gene therapists hope to avoid altering endogenous gene expression (and the resulting risk in activation of oncogenes), while those studying functional genomics *do* want to alter expression of chromosomal genes. The odds of either outcome depend on a vector's preference to integrate into genes and/or their transcriptional control regions, and these preferences are variable between vectors. This variability in insertion preference has generated interest in the use of DNA transposons as vectors for gene therapy and functional genomics. Here we discuss the relative strengths of DNA vectors over viral vectors, and review methods to overcome barriers to delivery inherent to DNA vectors. We also review the tendencies of several classes of retroviral and transposon vectors to target DNA sequences, genes and genetic elements with respect to the balance between insertion preferences and oncogenic selection. Theoretically, knowing the variables that affect integration for various vectors will allow researchers to choose the vector with the most utility for their specific purposes. There are three principle benefits from elucidating factors that affect preferences in integration: 1) in gene therapy, to assess the overall risks of activating an oncogene or inactivating a tumor suppressor gene that could lead to severe adverse effects years after treatment; 2) in genomic studies, to discern random from selected integration events, which is important for determining function; 3) in gene therapy as well as functional genomics, to design vectors and integrases that have greater targeting to specific sequences, which would be a significant advancement in the art of transgenesis.

Introduction

Elements such as viruses and transposons, through evolution with their host organisms, have acquired the ability to integrate into host genomes and ultimately shuffle genetic material between organisms. These elements have an established history in molecular biology and genetics research due to their ability to deliver specific genetic cargo, randomly disrupt host genomes for genetic screens, and serve as delivery vectors for delivery of therapeutic expression cassettes to treat human diseases. Viral vectors have been the predominant tools for these applications for three reasons: 1) the ease and efficiency with which specific viral genetic cassettes can be introduced into cells, 2) the vast accumulated knowledge of viruses and their mechanisms of gene transfer into chromosomes, and 3) the large number of sites in genomes into which they can integrate. Retroviruses in particular have been used for random insertion into chromatin to interrupt host genes (insertional mutagenesis) and thereby identify their function [1-3] as well as for delivery of therapeutic genes [4-6]. In particular, viral activation of oncogenes and, more recently, inactivation of tumor suppressors has been used to discover several novel genes involved in cancer progression [7-12]. The consequence of insertional activation of host-cell oncogenes by viral vectors, however, has presented itself as a major risk/obstacle in gene therapy, with a few cases of leukemia arising from oncogene activation by therapeutic vectors [13,14]. The potential genetic consequences of insertions of integrating vectors are summarized in **Figure 6.1**.

The risk of oncogene activation in gene therapy

The success of intentional insertional mutagenesis by endogenous retroviruses to activate oncogenes in mice suggested the possibility of inadvertent oncogene activation by relatively benign therapeutic vectors as a risk for gene therapy. While gene therapy vectors are extensively minimized to eliminate their replicative potential and reduce their collateral effects on the target genome [15], extensive testing in animals demonstrated that the risk of oncogenic activation was real, although variable and dependent both on the viral vector used, the genetic cargo, and the background genetics of the model system [16-22]. Given what was assumed to be acceptable risk, retroviral gene therapy trials have been conducted in human patients. Nearly 1000 clinical gene therapy trials have been initiated, more than half with retroviral vectors [4], but as yet no vectors have been approved in the US for clinical gene therapy outside of clinical trials (http://www.fda.gov/cber/gene.html). (Gendicine, an adenovirus designed to restore p53 function in cancerous cells, has been approved for commercial human gene therapy in China[23], although this vector is essentially non-integrating and thus confers a decreased risk of oncogene activation via vector insertion.)

The worst fears of the gene therapy field, oncogene activation, were realized when three of more than 20 patients treated for X-linked severe combined immunodeficiency disease (X-SCID) developed leukemia. These adverse results, including one death, occurred three or more years after administration of therapeutic murine leukemia virus (MLV)-derived retrovirus vectors [24,25]. The linkage between treatment and leukemias could be inferred because the expanded transformed cell populations harbored clonal integrations of the therapeutic vector that suggested a biological selection for the retrovirus-induced mutation [26-29]. However, these studies also indicated that clonal expansions in some cases appeared to be temporary and did not always lead to adverse effects – features that could actually improve the likelihood of successful gene therapy. The cause of at least two of the leukemias appears to be insertion of the MLV vector close to the LMO2 oncogene, which led to LMO2's activation by enhancers in the long terminal repeat sequences (LTRs) of the vector [30-32]. Retrospective examination of the role in LMO2 during development supported this conclusion [33,34]. Subsequent studies in which the cargo gene IL2γc was overexpressed in mice (albeit at levels higher than in the X-SCID leukemia patients) suggested that this gene could itself act as an oncogene in T-cells [35], and observation of simultaneous activation of Il2γc and LMO2 by oncogenic retroviruses had been observed in one mouse, suggesting a possible genetic interaction between the cargo IL2γc gene and LMO2[32]. The relevance of these observations to the clinical cases, however, is highly debatable [36,37].

In contrast, other gene therapy trials that employed retroviral vectors for adenosine deaminase (ADA) deficiency [38-40] and chronic granulomatosis disease (CGD) [41] have not yet reported any equivalent adverse events. In the CGD study, there appeared to be a powerful selection for integration events of the spleen focus-forming virus vector, which also was used as a vector for X-SCID [42], into the neighborhoods of three previously identified genes, *MDS-EVI1, PRDM16* and *SETBP1*, that have been associated with enhanced proliferation following integration of retroviruses with activating LTRs [43-45]. As noted earlier, findings of preferential integration around certain genes is not necessarily due to a preference for these genes, but rather a consequence of clonal expansion that can be transient and thereby beneficial for enhancing the number of therapeutic cells. A similar effect has also been observed in non-human primate studies, indicating this result may not be unique [19]. Despite the striking incidence of common integration sites (CISs) that are often associated with tumor or leukemia formation [9,46,47] there has been no report of adverse events in the CGD patients and no indication that the corrective gene, gp91^{phox}, synergizes with any of the three CIS genes to promote growth. Likewise, a murine stem cell retrovirus has been used to deliver the alpha and beta chains of the anti-MART-1-T-cell-receptor genes *ex vivo* into peripheral blood lymphocytes to treat melanoma without any apparent adverse effects, although integration sites were not examined and the patient population had very low odds for survival even with the treatment (2/15) for more than 1 year [48].

Taken together, the results of the CGD and X-linked plus ADA SCID trials demonstrate that oncogenesis is not necessarily an inherent, inevitable side effect of gene therapy; in more than 20 patients, the genetic deficiencies of more than 80% have been fully corrected allowing them to lead normal lives. However, tumors and leukemias can take years to manifest and these trials are in their early years. Obviously, a clearer understanding of variables underlying oncogenesis is needed to increase the safety of these trials. These variables include insertion-site preferences of therapeutic vectors, their abilities to activate nearby genes, and interactions between specific genetic cargos and activated host genes. Although cargo-host interactions will be specific for each gene therapy approach, the vectors themselves govern other parameters of insertion preference and neighboring gene activation. Analyses of insertion preferences, in particular, have received much recent attention, and have sparked interest in the use of transposons as an alternative to viruses as gene therapy vectors.

Non-viral vectors for introduction of genetic cassettes into mammalian genomes

Transposable elements also have been used for insertional mutagenesis and genetic studies in model organisms and are being developed as gene-therapy agents in humans [49-52]. The most well-characterized DNA transposon vector used in mammals is the synthetic *Sleeping Beauty* (SB) transposon system [53], which has become a powerful tool for functional genomics to identify genes in vertebrates, including fish and mammals [54-60] over the past decade. Transposon-mediated gene transfer has been explored for gene therapy in order to avoid several disadvantages of viral delivery systems including: 1) their preferences for integrating into genes [61-64], 2) the difficulty of purification to eliminate toxic or infectious agents [65], 3) their potential to elicit unwanted immune and/or inflammatory responses [66,67], 4) the constraint on therapeutic cargo size, and 5) the difficulty and expense to produce in large quantities [68,69]. In contrast to viral vectors, preparations of non-viral plasmid-based transposon vectors are relatively inexpensive to purify, are largely non-immunogenic, and have no hard constraints on genetic sequences that can be delivered.

A negative tradeoff for DNA vectors is an increased difficulty in delivery. Delivery of non-viral DNA into mammalian genomes involves avoiding and/or traversing numerous barriers, including enzymes in the blood and cellular environments, the endothelial lining of vessel walls, cellular plasma membranes, endosomal membranes, nuclear membranes and chromosomal integrity [70]. There are three delivery approaches that work across the nano, micro and macro scales [71]. Nanoscale delivery involves particles or complexes that are most often designed to be about 100 nm or less in diameter, although sizes up to 1 µm fit into this category. The nanoscale approach comprises delivery of single or small numbers of DNA molecules that most often are collapsed by polycationic polymers (e.g., polylysine and other modified amino acids, various linear and branched forms of polyethylenimine (PEI), etc.) or lipids with or without various ligands (reviewed in [70]). Some polycationic complexes are cytotoxic or unstable in the blood, which can be circumvented by encasing the complexes in polyethylene glycol (PEG) [72]. Alternative delivery routes are those at the micro and macroscale, in which DNA in packages up to 10 µm are phagocytized (microscale) or enter cells via fusions with other cells or entities larger than 10 µm (macroscale). In mice, the most effective method for *in vivo* gene transfer and expression has been demonstrated in hepatocytes using simple infusion of naked plasmid DNA under increased pressure. This can be accomplished by *hydrodynamic delivery* of DNA using high pressure/high volume injection [73,74], a procedure that in the mouse involves the injection of a large volume (10% vol/wt) of DNA/saline solution through the tail vein in less than 10 seconds. This procedure results in the uptake of infused DNA into up to 10% of the hepatocytes in test animals [73,74] by expanding and rupturing liver endothelium, which in mice heals within 24-48 hrs [75]. Achieving a clinically feasible method of local delivery to liver in large animals, including humans, is a challenge that is being addressed by more localized hydrodynamic delivery using specialized catheters or pressure cuffs [76,77]. On the microscale, condensing DNA with polyamines such as PEI to a complex small enough to be taken up by cells into endosomes has been studied intensively [78,79]. Our results (unpublished) suggest that gene expression following hydrodynamic delivery is about 100-fold more effective than delivery using polyethylenimine [80,81] and only about 10 to 100-fold less effective than viral delivery to the liver [71]. Alternative delivery *ex vivo* using electroporation is under development and has been achieved in hematopoietic stem cells [82]

Since the development of the SB system, non-viral, integrating DNAs have established themselves as potential vectors for gene therapy. Following hydrodynamic delivery, transposons have been used in mice to cure hemophilias A and B [83-86], and tyrosinemia Type I [87,88]. Other somatic delivery methods were used to ameliorate blistering skin disease (junctional epidermolysis bullosa) [89], retard glioma xenographs [90,91], produce Huntingtin protein in a model of Huntington disease [92], and as a preventative treatment of lung allograft fibrosis [93]. Based on the results reported above, we estimate that only about 1 in 10,000 SB transposons that are delivered to liver or lung actually transpose into chromatin (PH, unpub.). Although this is a small fraction, it is possible to deliver more than 10^8 therapeutic cassettes to an animal in order to treat as many as 10-20% of liver cells with a single injection of plasmids [83,87,94]. This procedure is sufficient to cure diseases such as hemophilia and tyrosinemia type 1, as well as ameliorate other diseases such as mucopolysaccharidoses Type I and Type VII. Although quantifying the number of transposon insertions per cell has not been done due to the difficulty of cloning insertion sites in mostly non-dividing cells in most organs of animals, the expression data are consistent with a single integration in most, if not all, transgene-expressing cells.

In addition to SB, several other transposon vectors and phage integrase-based vectors have been tested for their potential to deliver therapeutic genes, including *Frog Prince* [95], Tol2 [88] and *piggyBac* [96], as well as other well-characterized transposons such as the Drosophila P-elements that are not mobilized very efficiently in mammalian cells [97]. These vectors differ in their efficiency of gene insertion, genetic cargo capacity, integration-site preferences, and effects on chromosomal stability. Among other advantages these systems have over retroviruses as gene therapy vectors, transposons present a wide variety of insertion site preferences that differ from those of retroviruses, with possible consequences for oncogene activation. The characteristics of these vectors are summarized in **Table 6.1**. The remainder of this review will discuss these differences as they relate to gene therapy and functional genomics.

Factors governing insertion site preferences and their variation among vectors

Although most vectors will integrate into a vast number of sites scattered throughout the genome, numerous studies have shown that these integrations are not random with respect to several variables. Global preferences for vector integration can be governed by large-scale genomic context such as coding and regulatory regions of genes, and their transcriptional status, as compared to intragenic regions [98]. The fine-tuning that determines specific sites of integration is governed by smaller scale, physical features, such as the specific sequences of nucleotides surrounding insertion sites and DNA structural characteristics derived from these sequences. **Figure 6.2** illustrates some of the physical features of DNA that are influenced by local sequence.

Viruses and transposons display a wide range of variability with respect to preference for genes and transcriptional units. Several studies have mapped hundreds to thousands of insertions into human or mouse genomes, and correlated insertion positions with known genes. Many retroviruses display a non-random preference for genes [64]. This could be due to greater accessibility of the DNA in ‘open’ chromatin or interaction of integrase (IN) enzymes with cellular factors bound to transcriptional regulatory elements. In the case of HIV, the LEDGF/p75 transcriptional factor may act as a tether between the IN and transcriptionally activated chromatin [99-101], which is similar to an idea that was proposed earlier for designer-targeting of integrating vectors [102-104]. In a similar approach using the SB transposon, Yant et al. [105] found that SB displayed a much lower (although non-random) preference for genes. Although a preference for transcriptional units might seem beneficial for functional genomics studies, the myriad of recently identified non-coding RNA (ncRNA) genes [106] (as well as other RNA-product genes such as those encoding rRNA and tRNAs) involved in gene regulation may not be targeted by viral vectors that preferentially integrate into or near protein-encoding genes. Conversely, targeting of various vectors to these ncRNAs in gene therapy, and any resulting deleterious effects, has not been extensively examined.

Many vectors appear to display a preference for specific genes. In insertional mutagenesis studies, the identification of recurrent viral insertions into a specific group of genes was taken to mean that viral activation of these putative oncogenes in individual cells led to clonal expansion among a pool of cells where every host gene was an equal target for integration (as discussed above for LMO2). However, when MLV insertions were mapped in normal HeLa cells that did not undergo any type of selection, oncogenic or otherwise, many of these same genes harbored recurrent integrations, suggesting that vectors may inherently target specific genes [47]. The basis of this selection is not understood, but may be similar to that discussed above for HIV.

In addition to general preferences for genes, many viral vectors including retroviruses, lentiviruses, and adeno-associated virus preferentially target transcriptional units and/or their promoters. MLV retroviruses have a preference for integration proximal to transcriptional initiation sites [63,64,107-110], a problematic trait considering MLV-based vectors are the most commonly used vectors in human gene therapy [4]. HIV and adeno-associated viruses have preferences for entire transcriptional units [99,107,110-112], in contrast to MLV that targets only the region proximal to promoters. Additionally, expression array studies have shown that HIV has a preference for transcriptionally active genes [64] as well as an avoidance of chromatin regions in which transcription is repressed [113]. In contrast to these viral vectors, SB transposons and avian leukosis virus (a retrovirus) apparently have only a slight preference for either transcriptional units or their regulatory elements [105,114], with little or no preference for transcriptionally active genes [64]. In one survey, SB displayed an overall preference for microsatellite repeats, found primarily in noncoding regions [105], possibly due to the preferred target sites found in TA repeats [115]. A study correlating insertions sites with hundreds of genome annotations illustrated the degree to which genomic features and primary sequence influenced vector integration preferences for several vectors (for example, the L1 and SB transposon insertions were much more influenced by primary sequence than retroviral vectors) [98]. This study also found variable preferences between vectors for elements such as CpG islands, DNase I sensitive sites, and transcription factor binding sites. The recent identification of a periodic sequence encoding nucleosome positioning [116] may also correlate with vector integration patterns, as nucleosomes have been shown to affect patterns of retroviral integration [117]. Similar studies to identify trends for *piggyBac* and Tol2 with respect to genome-wide integration preferences will be valuable to assess the relative safety of these vectors for gene therapy.

Local insertional preferences: DNA sequence and structure

Although many vectors display a preference for genes, and even specific genes, few vectors repeatedly integrate into the same precise position with any significant frequency. Rather, most genes harboring frequent insertions show a distribution of insertions into several positions within the same gene. Some vector integrases, such as those for phages □C31[118-120], □BT1[121] as well as the *E. coli* Tn7 transposon [122], recognize specific DNA sequences or degenerate sequences that exist in mammalian genomes. SB integrates specifically at a TA dinucleotide, and the *piggyBac* transposon integrates into the sequence TTAA. Because the oncogenic potential of a vector is related to its propensity to integrate in or near a select few genes, understanding local parameters that affect integration may contribute to our ability to assess the risk of these vectors in gene therapy.

For retroviruses and the *Sleeping Beauty* transposon, consensuses sequences have been described surrounding the sites of integration [110,123-126]. Although retroviruses do not display a strong consensus sequence, the non-random pattern of integrations and the observation that frequently-hit sites did not match the consensus sequences led investigators to examine other properties of DNA sequences surrounding target sites, including structural characteristics of the DNA itself. DNA structural characteristics are based on non-Watson and Crick interactions between nucleotides and encompass deformations to the regular double helix structure caused by interactions between adjacent, planar bases (**Figure 6.2**). Originally characterized from analysis of crystal structures of DNA bound to histones and other proteins, these characteristics include "protein-induced DNA deformability", "A-philicity," and trinucleotide "bendability," properties that underlie local variations in DNA structure likely relevant to recognition of DNA by transposases and integrases. Early investigations into insertion preferences showed that viruses preferred "bent" DNA [117,127,128], and several groups have investigated secondary DNA structural patterns in sequences flanking mapped insertion sites for both transposons [123,129-131] and retroviruses [110,125] to determine general characteristics of the flanking sequence of "preferred" integration sites. Similarly, the RAG1/2 protein complex, which has properties akin to the cut-and-paste transposases, recognizes a specific sequence/structure for recombination of antigen receptor genes [132].

Different DNA sequences may produce highly similar patterns of DNA secondary structure, and thus common structural patterns preferred for integration may be obscured by approaches that analyze sequence alone. Analysis of secondary structure for a DNA sequence is based on translation of a sliding window of two or three bases into structural values for each "step." For example, the tendency of a *B*-form helix to adopt the *A*-form (A-philicity, **Figure 6.2**) can be predicted by translating each consecutive (overlapping) dinucleotide into one of 10 A-philicity values for the 16 combinations of base-pair transitions [133-135]. Similarly, protein-induced deformability encompasses several changes in base-pair orientation from a "perfect *B*-form double helix" in a transition between two consecutive base pairs (**Figure 6.2C**). All of these changes can be expressed as a single composite parameter of protein-induced DNA deformability known as V_{step} [136-138]. V_{step} represents the physical relationships of any two planar base pairs in terms of their relative shifts and angular orientation. In contrast to A-philicity and protein-induced deformability, DNA bendability is best modeled using a sliding window of three bases, with 64 possible trinucleotide bendability values [139].

An example of DNA structural analysis for the Tol2 transposon is shown in **Figure 6.3**, where average structural values for each position flanking an insertion site are plotted and compared to a plot of random sequences. In the case of Tol2, weak preferences in V_{step} and A-philicity values at specific coordinates are apparent by the peaks in the heavy black lines in panels A and B (left sides) in contrast to the same averages derived from random sequences (right sides). Overall, the bendability around Tol2 insertion sites shows little deviation from a random sequence (**Figure 6.3C**), unlike those preferred by SB transposase (**Figure 6.3C**). Analysis of hundreds of integration sites for potential gene therapy vectors, including viruses as well as transposons, shows that many have subtle preferences for these variables (**Figure 6.4**). For example, the *piggyBac* transposon may favor sites with slightly higher A-philicity, lower bendability, and lower V_{step} values than random sequences. In contrast, “preferred” SB insertion sites (see below) clearly display a jagged V_{step} pattern and higher bendability. Interestingly, although retroviruses (ASV, HIV, MLV, SIV) integrate into bent DNA [127] such as that in nucleosomes, our analyses of sequences around viral insertion sites do indicate a particular preference for bendable DNA (**Figure 6.4**). A similar, more rigorous approach has been utilized to characterize Drosophila P-elements [129] and non-LTR retrotransposons in *E. histolytica* [140], demonstrating that DNA structural characteristics at insertion sites for both elements are significantly different from collections of random sequences.

For SB, the observation of general structural trends surrounding insertion sites eventually led to the identification of a specific DNA structural pattern governing insertion preference. Vigdal et al [123] observed that increased DNA deformability and A-philicity were features of a consensus sequence that flanked *Sleeping Beauty* TA insertion sites. Subsequently, Liu et al. [130] mapped roughly 200 integrations into a relatively small 7-kb plasmid sequence and observed that some common integration sites did not share the consensus sequence. These results identified several "preferred" TA dinucleotides that harbored recurrent integrations. These preferred integration sites exhibited a striking specific pattern of alternating high-and-low deformability (V_{step}) values that were absent in TA sites that were rarely, if ever, used. This led to the conclusion that SB transposase prefers a "zigzag" V_{step} pattern of DNA deformability [130], which was later confirmed on a larger, genomic scale [131]. It remains unknown whether these patterns influence the recognition and binding of the SB transposase, catalysis of the transposon integration, or some other mechanistic factor.

This analysis was repeated for other vectors including *piggyBac*, P-elements, and several retroviruses [131]. However, only weak structural signatures were detected, which were no more informative than the weak consensus sequences previously identified. A key difference in the SB screen was the level of saturation of a small target, which allowed for the identification of highly preferred sites over non-preferred TA dinucleotides. In contrast, the datasets for the other vectors were derived from a relatively small number of insertions into mammalian genomes, which was insufficient to obtain an initial set of preferred sequences. Since non-preferred sites are quite likely to vastly outnumber preferred sites in the genome for most vectors, any genome-wide screen will produce a mix of indistinguishable preferred and non-preferred sites. For example, we have estimated that of the approximately 200,000,000 TA sites in a human genome, only about 10% fall into the *preferred* category [131] although in the Yant screen [105] 189/573 (33%) of genomic SB insertions were classified as preferred sites. Analysis of the bendability of all SB sites mapped in Yant's screen shows a peak at the center of the insertion site that is defined by the central TA dinucleotide. However, when only the preferred sites are analyzed, the surrounding nucleotides display a much higher level of bendability (**Figure 6.3D**). This effect is in spite of the fact that the preferred sites were identified based on protein-induced deformability, V_{step}, which is distinct from DNA bendability. The lesson from these studies is that most genome-wide datasets (particularly from experiments involving some form of genetic selection) will likely show a similar dilution effect of preferred sites by greater numbers of non-preferred sites.

There is a caveat to the analyses discussed to this point – they all assume that the structures around integration sites have an absolute center of reference defined by the site into which the vector integrated. Such analyses could miss structural patterns that are not strictly position specific. For instance, an integrase may have preference for a local region that is highly bendable or deformable, but not have a requirement for a particular pattern (or sequence). To account for this, we have examined a parameter called "jaggedness", which we define as the degree to which V_{step} values alternate from high to low, as in the preferred "zigzag" sites for SB. We calculated jaggedness by taking the absolute value of the sum of the differences between adjacent V_{step} values across a sequence – so that a jagged/zigzag site would have a higher total value than a flat, basal site, which should have a jaggedness value close to 0. Jaggedness values for several vectors are shown in **Figure 6.4**. Although jaggedness values at insertion sites are similar to V_{step} values for most vectors (with the possible exception of To2), the jaggedness patterns show a high degree of variability across genomic sequences and are somewhat independent of V_{step} patterns, e.g. the *c-myc* gene (**Figure 6.6**).

Integration preference vs. oncogenic selection

We see two uses for profiling the insertion-site preferences for integrating vectors. First, in functional genomics screens, insertion profiles that emerge can be compared to expected profiles that are only structure-based rather than genetics-based. A striking example of this is evident in the oncogene screens conducted with the SB transposon [57,58] that is illustrated in **Figure 6.5** with respect to the *Braf* gene. Integration sites that emerged from the screen are shown across the entire locus (panel B) and in a selected region comprising exons 10-13/introns 10-12 (panel D) where most of the integrations were selected due to induced expression of a truncated gain-of-function kinase polypeptide. Panels A and C show insertion-site preference scores across the region obtained using an automated script (*ProTIS*) that counts and scores preferred TA dinucleotide insertion sites based on V_{step} values [131]. The results shown in **Figure 6.5** make two strong points 1) The frequency of oncogenic insertions in a select region correspond to that predicted on the basis of preference profiling (panels C and D), i.e., micro-scale structure can be a good predictor of integration-site preference. 2) Many predicted hotspots (panels A and B) were not sites that lead to oncogenesis. The combination of these two observations enhances the biological importance of the integrations into introns 11 and 12.

The second application of predicting profiles of vector insertions may be as part of a risk-assessment program. While current understanding of integration-site preferences for most vectors is still inadequate to predict the probability of integration into specific genes, genome-wide integration datasets may suggest the likelihood of a vector to integrate within the general vicinity of a specific gene. Similarly, analysis of DNA structural characteristics may be used to assess the likelihood of each vector to integrate within specific regions of genes. For example, though *Braf* can act as a potent oncogene, the pattern of SB integrations into *Braf* suggest that integrations into a relatively small region of the gene (introns 11 and 12) are the most highly-selected for oncogenesis, in spite of the presence of hotspots across the entire gene. Thus, the range of possible insertions capable of generating an oncogenic transcript, combined with the relative “attractiveness” of the sequence across these regions, will dictate the chances of insertional activation.

An analysis of several structural characteristics is presented for mouse *c-myc* gene (**Figure 6.6**), the human ortholog of which is activated in many cancers [141]. The figure highlights the 3-kb region encompassing the promoter that harbors the bulk of oncogenic retroviral integrations at this locus that have been deposited in the Retroviral-Tagged Cancer Gene Database (RTCGD, http://rtcgd.ncifcrf.gov/). The sequence was divided into 50-bp bins, and the total values for V_{step}, A-philicity, jaggedness and bendability were summed across each bin. Measured in 50-bp bins, these structural parameters are highly variable across the sequence, and vary independently from each other. Actual oncogenic retroviral insertions observed in insertional mutagenesis screens and deposited into the RTGCD are shown for comparison in panel A. The profiles indicate two features of transposons under consideration for gene therapy. First, the most likely sites for SB transposons to integrate (panel G) are shifted away from the most commonly found activation sites as revealed by retroviral integrations (panel A). Second, the profile of TTAA sites, required by the *piggyBac* transposon (panel F), is similar to the preferred SB sites, and further shows that some regions harboring retroviral integrations contain no TTAA sequences, making *piggyBac* insertions into these sites impossible. Thus, at first approximation, it would appear that the transposons have a lower chance of insertion close to the *c-myc* promoter than retroviral vectors. In support of this, *c-myc* is infrequently hit in SB-based insertional mutagenesis screens; to date, only one *c-myc* integration has been deposited into the RTCGD. In contrast, many retroviral insertions into *c-myc* have been mapped, although the number of deposited retroviral insertions is much higher than the number of transposons. The relative lack of SB insertions into *c-myc* may be due to either a paucity of favorable SB insertion sites in regions of the gene competent for oncogenic activation, or an overall lack of oncogenic selection for insertions into this gene. In support of the former, transposon-free amplification of *c-myc* was one of the few genomic aberrations observed in tumors harboring mobile transposons (D.A. Largaespada, L.C. Collier and CSH, unpublished observations), suggesting that activation of *c-myc* plays a role in the biology of these tumors, i.e., there was likely oncogenic selection for the genomic amplicon. Similar *ProTIS* analysis of the *LMO2* locus revealed the most preferential integration sites for SB transposons that were considerably farther away from the *LMO2* promoter than mapped integrations by activating retroviruses [131]. That said, it is evident that prediction of vector integration is not precise and even rare integrations into unfavorable sites have a potential to promote oncogenic expansion, as indicated in **Figure 6.5**.

Vector behavior in risk/outcome assessment: Lessons from intentional oncogenic insertional mutagenesis

In spite of the inherent behavior of each integrating vector, existing evidence suggests the oncogenic potential of any given vector can be attenuated depending on how it is used. As with retroviruses, the *Sleeping Beauty* transposon has been used for functional genomics as well as for delivery of therapeutic genes in mouse models of inherited disease. These studies were motivated by two limitations of retroviruses for insertional mutagenesis - the limitation of viruses to infect specific cell types, and the tendency of many viral vectors to insert near and activate a possibly limited number of genes [47]. In two recent SB mutagenesis screens, a transgenic concatemer of T2/Onc transposons carried in the germlines of mice was remobilized in somatic cells by a *trans*-acting, transgenic SB transposase. The two screens differed in the expression level, domains of expression, and activity of the SB transposase as well as the copy number of the transposon concatemers [57,58]. An important finding from the two studies was that the oncogenic potential of the same T2/Onc transposon vector, which was engineered specifically to activate oncogenes and cause cancers in mice, varied between no observable phenotype on one end and rapid development of severe cancer at birth on the other. The oncogenic effect was directly related to the number and types of cells at risk for transposon-induced mutations and perhaps the re-mobilization rates. The same properties may be relevant for a wide range of other gene therapy vectors.

Coupled with the lack of a preference to integrate near genes, the chances of an SB insertion of a therapeutic gene (in contrast to a genetic cassette designed to wreak havoc on transcriptional units) activating a gene would seem to be lower than for vectors that have an affinity to integrate into genes [64,96]. This observation may be a disadvantage for SB-based functional genomics studies aimed at mutating genes but may be advantageous for gene therapy.

Engineering safer vectors

As an alternative to finding vectors that don’t target genes, several groups are attempting to target vector integration to a specific region of the genome by generating integrase and SB transposase molecules that are fused to DNA binding domains that recognize specific DNA sequences [142,143]. It appears that targeting introduces a reduction in activity, without much increase in specificity of integration into specific sites in a mammalian genome (Yant et al.

submitted). This is not surprising if the ability of SB transposase to integrate promiscuously into TA sites is not abridged. There are about $2x10^8$ potential TA-dinucleotide SB integration sites into which SB transposons can integrate, of which $2x10^7$ are estimated to be preferred integration sites [131]. Consequently, the chances of a sequence-specific targeting motif added to SB transposase actually guiding transposition to a specific, low-copy target sequence is expected to be extremely low compared to the chances of integrating into any of the millions of other available TA sites. Similarly, to overcome the risk of activation of neighboring genes following vector integration, self-inactivating (SIN) vectors are being engineered to have diminished ability to activate genes over long distances [144,145], although it is not clear that these vectors will be safer [146]. The φC31 phage integrase system targets relatively few sites in mammalian genomes [118,147] but it appears to introduce a relatively high level of chromosomal recombination [147-149]. Thus, further development of safer vectors remains a wide-open area.

Conclusion: Assessing risks of randomly integrating vectors

Ultimately, functional genomics and gene therapy would like to answer the same question for any given vector, although hoping for opposite outcomes - what are the chances of activating genes? There are four major factors influencing the answer, with each retroviral and transposon having different characteristics for each factor. First, what is the overall tendency of the vector to integrate into genes and/or promoters? Second, are there adequate local target sites around genes of interest to attract the vector? Third, over what distance can the vector activate a gene? Fourth, to what end can the integration activity be modulated to control the overall likelihood of hitting specific insertion sites close enough for activation of specific genes? Theoretically, knowing each of these variables for every vector would allow researchers to choose the vector with the most utility, and lowest risk, for the specific purpose intended. In gene therapy, these parameters translate to the risk of hitting a specific oncogene or tumor suppressor gene that could lead to a severe adverse effect. If, in the future, hotspots for integration of SB and other potential gene therapy vectors can be predicted, we should be able to more accurately assess and modify the various risks of adverse effects from therapeutic vectors. This goal should be within reach in the coming years.

Acknowledgements. We thank the Arnold and Mabel Beckman Foundation for support of our work and all members of the Beckman Center for Transposon Research for a long history of contributions of ideas and results. We appreciate the help of Drs. Nik Somia and Marina O'Reilly in determining the number of gene therapy trials reviewed by the RAC. We are especially grateful to Dr. Darius Balciunas and Kirk Wangensteen for sharing their Tol2 dataset, and to Drs. David Largaespada and Lara Collier as well as two reviewers for discussions about the manuscript. The authors were supported by DOD fellowship BC050930 (CSH) and NIH grants T32 HD007480 (AMG) and 1PO1 HD32652-07 and R43 HL076908-01 (PBH).

References:

1. Zambrowicz, B.P. et al. Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature* **392**, 608-611 (1998).
2. Mitchell, K.J. et al. Functional analysis of secreted and transmembrane proteins critical to mouse development. *Nature Genet.* **28**, 198-200 (2001).
3. Mikkers, H. & Berns, A. Retroviral insertional mutagenesis: tagging cancer pathways. *Adv. Cancer Res.* **88**, 53-99 (2003).
4. Edelstein, M.L., Abedi, M.R., Wixon, J. & Edelstein, R.M. Gene therapy clinical trials worldwide 1989-2004-an overview. *Gene Med.* **6**, 597-602 (2004).
5. Sinn, P.L., Sauter, S.L. & McCray Jr, P.B. Gene therapy progress and prospects: Development of improved lentiviral and retroviral vectors - design, biosafety, and production. *Gene Ther.* **12**, 1089-1098 (2005).
6. Connelly, J.B. Lentiviruses in gene therapy clinical research. *Gene Ther.* **9**, 1730-1743 (2002).

7. Jonkers, J. & Berns, A. Retroviral insertional mutagenesis as a strategy to identify cancer genes. *Biochim. Biophys. Acta* **1287**, 29-57 (1996).
8. Largaespada, D.A. Genetic heterogeneity in acute myeloid leukemia: maximizing information flow from MuLV mutagenesis studies. *Leukemia* **14**, 1174-1184 (2000).
9. Lund, A.H. et al. Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nature Genet.* **32**, 160-165 (2002).
10. Suzuki, T. et al. New genes involved in cancer identified by retroviral tagging. *Nature Genet.* **32**, 166-174 (2002).
11. Kim, R. et al. Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J. Virol.* **77**, 2056-2062 (2003).
12. Suzuki, T., Minehata, K., Akagi, K., Jenkins, N.A. & Copeland, N.G. Tumor suppressor gene identification using retroviral insertional mutagenesis in *Blm*-deficient mice. *EMBO J.* **25**, 3422-3431 (2006).
13. Yi, Y., Hahm, S.H. & Lee, K.H. Retroviral gene therapy: safety issues and possible solutions. *Curr. Gene Therap.* **5**, 25-35 (2005).
14. Berns, A. Good news for gene therapy. *New Eng. J. Med.* **350**, 1679-1680 (2004).
15. Pages, J.C. & Bru, T. Toolbox for retrovectorologists. *J. Gene Med.* **Suppl 1**, S67-S82 (2004).
16. Kohn, D. et al. American Society of Gene Therapy (ASGT) ad hoc subcommittee on retroviral-mediated gene transfer to hematopoietic stem cells. *Mol. Ther.* **8**, 180-187 (2003).
17. Hematti, P. et al. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol* **2**, e423 (2004).
18. Kiem, H.P. et al. Long-term clinical and molecular follow-up of large animals receiving retrovirally transduced stem and progenitor cells: no progression to clonal hematopoiesis or leukemia. *Mol. Ther.* **9**, 389-395 (2004).
19. Calmels, B. et al. Recurrent retroviral vector integration at the Mds1/Evi1 locus in nonhuman primate hematopoietic cells. *Blood* **106**, 2530-2533 (2005).
20. Bell, P. et al. No evidence for tumorigenesis of AAV vectors in a large-scale study in mice. *Mol. Ther.* **12**, 299-306 (2005).
21. Themis, M. et al. Oncogenesis following delivery of a non-primate lentiviral gene therapy vector to fetal and neonatal mice. *Mol. Ther.* **12**, 763-771 (2005).
22. Du, Y., Spence, S.E., Jenkins, N.A. & Copeland, N.G. Cooperating cancer-gene identification through oncogenic-retrovirus-induced insertional mutagenesis. *Blood* **106**, 2498-2505 (2005).
23. Peng, S. Current status of gendicine in China: recombinant human Ad-p53 agent for treatment of cancers. *Hum. Gene Ther.* **16**, 1016-1027 (2005).
24. Hacein-Bey-Abina, S. et al. Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *New Eng. J. Med.* **346**, 1185-1193 (2002).
25. Hacein-Bey-Abina, S. et al. A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *New Engl. J. Med.* **348**, 255-256 (2003).
26. Baum, C. et al. Side effects of retroviral gene transfer into hematopoietic stem cells. *Blood* **101**, 2099-2114 (2003).
27. Baum, C. & Fehse, B. Mutagenesis by retroviral transgene insertion: risk assessment and potential alternatives. *Curr. Opin. Mol. Ther.* **5**, 458-462 (2003).

28. Baum, C. et al. Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. *Mol. Ther.* **9**, 5-13. (2004).

29. Kustikova, O. et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science* **308**, 1171-1174 (2005).

30. Hacein-Bey-Abina, S. et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415-419. (2003).

31. McCormack, M.P., Forster, A., Drynan, L., Pannell, R. & Rabbitts, T.H. The LMO2 T-cell oncogene is activated via chromosomal translocations or retroviral insertion during gene therapy but has no mandatory role in normal T-cell development. *Mol. Cell. Biol.* **23**, 9003-9013 (2003).

32. Dave', U.P., Jenkins, N.A. & Copeland, N.G. Gene therapy insertional mutagenesis insights. *Science* **303**, 33. (2004).

33. McCormack, M.P. & Rabbitts, T.H. Activation of the T-cell oncogene LMO2 after gene therapy for X-linked severe combined immunodeficiency. *New Eng. J. Med.* **350**, 913-922 (2004).

34. Nam, C.H. & Rabbitts, T.H. The role of LMO2 in development and in T cell leukemia after chromosomal translocation or retroviral insertion. *Mol. Ther.* **13**, 15-25 (2006).

35. Woods, N.B., Bottero, V., Schmidt, M., von Kalle, C. & Verma, I.M. Gene therapy: therapeutic gene causing lymphoma. *Nature* **440**, 1123 (2006).

36. Pike-Overzet, K. et al. Gene therapy: is IL2RG oncogenic in T-cell development? *Nature* **443**, E5 (2006).

37. Thrasher, A.J. et al. Gene therapy: X-SCID transgene leukaemogenicity. *Nature* **443**, E5 (2006).

38. Schmidt, M. et al. Clonality analysis after retroviral-mediated gene transfer to CD34+ cells from the cord blood of ADA-deficient SCID neonates. *Nature Med.* **9**, 463-468 (2003).

39. Aiuti, A., Ficara, F., Cattaneo, F., Bordignon, C. & Roncarolo, M.G. Gene therapy for adenosine deaminase deficiency. *Curr. Opin. Allergy Clin. Immunol.* **3**, 461-466 (2004).

40. Gaspar, H.B. et al. Successful reconstitution of immunity in ADA-SCID by stem cell gene therapy following cessation of PEG-ADA and use of mild preconditioning. *Mol. Ther.* **14**, 505-513 (2006).

41. Ott, M.G. et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nature Med.* **5**, 401-409 (2006).

42. Gaspar, H.B. et al. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet* **364**, 2181-2187 (2005).

43. Buonamici, S., Chakraborty, S., Senyuk, V. & Nucifora, G. The role of EVI1 in normal and leukemic cells. *Blood Cells Mol Dis.* **31**, 206-212 (2003).

44. Buonamici, S. et al. EVI1 induces myelodysplastic syndrome in mice. *J. Clin. Invest.* **114**, 713-719 (2004).

45. Li, Z. et al. Murine leukemia induced by retroviral gene marking. *Science* **296**, 497 (2002).

46. Mikkers, H. et al. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nature Genet.* **32**, 153-159 (2002).

47. Wu, X., Luke, B.T. & Burgess, S.M. Redefining the common insertion site. *Virol.* **344**, 292-295 (2006).

48. Morgan, R.A. et al. Cancer regression in patients after transfer of genetically engineered lymphocytes. *Science* **314**, 126-129 (2006).

49. Ivics, Z. & Izsvak, Z. Transposable elements for transgenesis and insertional mutagenesis in vertebrates: a contemporary review of experimental strategies. *Meth. Mol. Biol.* **260**, 255-276 (2004).

50. Hackett, P.B., Ekker, S.C., Largaespada, D.A. & McIvor, R.S. *Sleeping Beauty* transposon-mediated gene therapy for prolonged expression. *Adv. Genet.* **54**, 187-229 (2005).

51. Hackett, P.B., Ekker, S.E. & Essner, J.J. Applications of transposable elements in fish for transgenesis and functional genomics. *Fish Development and Genetics (Z. Gong and V. Korzh, eds.) World Scientific, Inc.* **Chapter 16**, 532-580 (2004).

52. Ivics, Z. & Izsvak, Z. Transposons for gene therapy! *Curr. Gene Ther.* **6**, 593-607 (2006).

53. Ivics, Z., Hackett, P.B., Plasterk, R.H. & Izsvak, Z. Molecular reconstruction of *Sleeping Beauty*, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-510 (1997).

54. Dupuy, A.J., Fritz, S. & Largaespada, D.A. Transposition and gene disruption using a mutagenic transposon vector in the male germline of the mouse. *Genesis* **30**, 82-88 (2001).

55. Davidson, A.E. et al. Efficient gene delivery and expression in zebrafish using *Sleeping Beauty*. *Dev. Biol.* **263**, 191-202 (2003).

56. Balciunas, D. et al. Enhancer trapping in zebrafish using the *Sleeping Beauty* transposon. *BMC Genomics* **5**, 62 (2004).

57. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic *Sleeping Beauty* transposon system. *Nature* **436**, 221-226 (2005).

58. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumors using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-276 (2005).

59. Keng, V.W. et al. Region-specific saturation germline mutagenesis in mice using the *Sleeping Beauty* transposon system. *Nature Meth.* **2**, 763-769 (2005).

60. Carlson, C.M. & Largaespada, D.A. Insertional mutagenesis in mice: new perspectives and tools. *Nature Rev. Genet.* **6**, 568-580 (2005).

61. Nakai, H. et al. AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nature Genet.* **34**, 297-302 (2003).

62. Wu, X. & Burgess, S.M. Integration target site selection for retroviruses and transposable elements. *Cell. Mol. Life Sci.* **61**, 2588-2596 (2004).

63. Wu, X., Li, Y., Crise, B. & Burgess, S.M. Transcription start regions in human genome are favored targets for MLV integration. *Science* **300**, 1749-1751 (2003).

64. Bushman, F. et al. Genome-wide analysis of retroviral DNA integration. *Nature Rev. Microbiol.* **3**, 848-858 (2005).

65. Kay, M.A., Glorioso, J.C. & Naldini, L. Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics. *Nature Med.* **7**, 33-40 (2001).

66. Muruve, D.A., Barnes, M.J., Stillman, I.E. & Libermann, T.A. Adenoviral gene therapy leads to rapid induction of multiple chemokines and acute neutrophil-dependent hepatic injury in vivo. *Hum. Gene Ther.* **10**, 965-976 (1999).

67. Graham, A., Walker, R., Baird, P., Hahn, C.N. & Fazakerley, J.K. CNS gene therapy applications of the semliki forest virus 1 vector are limited by neurotoxicity. *Mol. Ther.* **13**, 631-635 (2006).
68. Reeves, L., Smucker, P. & Cornetta, K. Packaging cell line characteristics and optimizing retroviral vector titer: The National Gene Vector Laboratory experience. *Hum. Gene Ther.* **11**, 2093-2103 (2000).
69. Kumar, M., Bradow, B.P. & Zimmerberg, J. Large-scale production of pseudotyped lentiviral vectors using baculovirus GP64. *Hum. Gene Ther.* **14**, 67-77 (2003).
70. Wagner, E., Culmsee, C. & Boeckle, S. Targeting polyplexes: Toward synthetic virus vector systems. *Adv. Genet.* **53**, 333-354 (2005).
71. Putnam, D. Polymers for gene delivery across length scales. *Nature Mat.* **5**, 439-451 (2006).
72. Merdan, T. et al. PEGylation of poly(ethylene imine) affects stability of complexes with plasmid DNA under in vivo conditions in a dose-dependent manner after intravenous injection into mice. *Bioconjugate Chem.* **16**, 785-792 (2006).
73. Liu, F., Song, Y. & Liu, D. Hydrodynamics-based transfection in animals by systemic administration of plasmid DNA. *Gene Therapy* **6**, 1258-1266 (1999).
74. Zhang, G., Budker, V. & Wolff, J.A. High levels of foreign gene expression in hepatocytes after tail vein injections of naked plasmid DNA. *Human Gene Therapy* **10**, 1735-1737 (1999).
75. Suda, T., Gao, X., Stolz, D.B. & Liu, D. Structural impact of hydrodynamic injection on mouse liver. *Gene Ther.* **14**, 129-137 (2007).
76. Yoshino, H., Hashizume, K. & Kobayashi, E. Naked plasmid DNA transfer to the porcine liver using rapid injection with large volume. *Gene Ther.* **13**, 1696-1702 (2006).
77. Herweijer, H. & Wolff, J.A. Gene therapy progress and prospects: Hydrodynamic gene delivery. *Gene Ther.* **14**, 99-107 (2007).
78. Kichler, A. Gene transfer with modified polyethylenimines. *J. Gene Med.* **6**, S3-S10 (2004).
79. Demeneix, B. & Behr, J.P. Polyethylenimine (PEI). *Adv. Genet.* **53**, 217-230 (2005).
80. Neu, M., Fischer, D. & Kissel, T. Recent advances in rational gene transfer vector design based on poly(ethylene imine) and its derivatives. *J. Gene Med.* **7**, 992-1009 (2005).
81. Breunig, M. et al. Gene delivery with low molecular weight linear polyethylenimines. *J. Gene Med.* **7**, 1287-1298 (2005).
82. Huang, X., Wilber, A.C. et al. Stable gene transfer and expression in human primary T-cells by the *Sleeping Beauty* transposon system. *Blood* **107**, 483-491 (2005).
83. Yant, S.R. et al. Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nature Genetics* **25**, 35-41 (2000).
84. Ohlfest, J.R. et al. Phenotypic correction and long-term expression of factor VIII in hemophilic mice by immunotolerization and nonviral gene transfer using the *Sleeping Beauty* transposon system. *Blood* **105**, 2691-2698 (2005).
85. Baus, J., Liu, L., Heggestad, A.D., Sanz, S. & Fletcher, B.S. Correction of murine hemophilia a by hematopoietic stem cell gene therapy. *Mol. Ther.* **12**, 1034-1042 (2005).
86. Liu, L., Mah, C. & Fletcher, B.S. Sustained FVIII expression and phenotypic correction of hemophilia A in neonatal mice. *Mol. Ther.* **13**, 1006-1015 (2006).
87. Montini, E.P. et al. In vivo correction of murine tyrosinemia type I by DNA-mediated transposition. *Mol. Therap.* **6**, 759-769 (2002).

88. Balciunas, D. et al. Harnessing an efficient large cargo-capacity transposon for vertebrate gene transfer applications. *PLoS Genet.* **4**(2006).

89. Ortiz, S. et al. Sustainable correction of junctional epidermollysis bullosa via transposon-mediated nonviral gene transfer. *Gene Ther.* **10**, 1099-1104 (2003).

90. Ohlfest, J.R., Lobitz, P.D., Perkinson, S.G. & Largaespada, D.A. Integration and long-term expression in xenografted human glioblastoma cells using a plasmid-based transposon system. *Mol. Ther.* **10**, 260-268 (2004).

91. Ohlfest, J.R. et al. Combinatorial anti-angiogenic gene therapy by nonviral gene transfer using the *Sleeping Beauty* transposon causes tumor regression and improves survival in mice bearing intracranial human glioblastoma. *Mol. Ther.* **12**, 778-788 (2005).

92. Chen, Z.T., Kren, B.T., Wong, P.Y.P., Low, W.C. & Steer, C.J. *Sleeping Beauty-mediated down-regulation of huntingtin expression by RNA interference. . Biochem. Biophys. Res. Comm.* **329**, 646-652 (2005).

93. Liu, H., Liu, L., Fletcher, B.S. & Visner, G.A. *Sleeping Beauty*-based gene therapy with indoleamine 2,3-dioxygenase inhibits lung allograft fibrosis. *FASEB J.* **20**, 2384-2386 (2006).

94. Aronovich, E.L. et al. *Sleeping Beauty* transposon-mediated gene therapy in the murine models of mucopolysaccharidoses (MPS) Type I and MPS Type VII. *(submitted)* (2006).

95. Miskey, C., Izsvak, Z., Plasterk, R.H.A. & Ivics, Z. The Frog Prince: a reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells. *Nucl. Acids Res.* **31**, 6873-6881 (2003).

96. Ding, S. et al. Efficient transposition of the *piggyBac* (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-483 (2005).

97. Rio, D.C., Barnes, G., Laski, F.A., Rine, J. & Rubin, G.M. Evidence for *Drosophila* P element transposase activity in mammalian cells and yeast. *J. Mol. Biol.* **200**, 411-415 (1988).

98. Berry, C., Hannenhalli, S., Leipzig, J. & Bushman, F.D. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comp. Biol.* **2**, e157 (2006).

99. Ciuffi, A. et al. Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Mol. Ther.* **13**, 366-373 (2006).

100. Ciuffi, A., Diamond, T.L., Hwang, Y., Marshall, H.M. & Bushman, F.D. Modulating target site selection during human immunodeficiency virus DNA integration in vitro with an engineered tethering factor. *Hum. Gene Ther.* **17**, 960-967 (2006).

101. Lewinski, M.K. et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog. 2* **2**, e60 (2006).

102. Bushman, F.D. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc. Natl. Acad. Sci. USA* **91**, 9233-9237 (1994).

103. Bushman, F.D. & Miller, M.D. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J. Virol.* **71**, 458-464 (1997).

104. Zhu, Y., Dai, J., Fuerst, P.G. & Voytas, D.F. Controlling integration specificity of a yeast retrotransposon. *Proc. Natl. Acad. Sci. USA* **100**, 5891-5895 (2003).

105. Yant, S.R. et al. High-resolution genome-wide mapping of transposon integration in mammals. *Mol. Cell. Biol.* **25**, 2085-2094 (2005).

106. Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919-925. (2006).

107. Mitchell, R.S. et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS* **2**, 1127-1136 (2004).

108. Laufs, S. et al. Insertion of retroviral vectors in NOD/SCID repopulating human peripheral blood progenitor cells occurs preferentially in the vicinity of transcription start regions and in introns. *Mol. Ther.* **10**, 874-881 (2004).

109. De Palma, M. et al. Promoter trapping reveals significant differences in integration site selection between MLV and HIV vectors in primary hematopoietic cells. *Blood* **105**, 2307-2315 (2005).

110. Holman, A.G. & Coffin, J.M. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc. Natl. Acad. Sci. USA* **102**, 6103-6107 (2005).

111. Schroder, A.R.W. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-529 (2002).

112. Nakai, H. et al. Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. *J. Virol.* **79**, 3606-3614 (2005).

113. Lewinski, M.K. et al. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol.* **79**, 6610-6619 (2005).

114. Narezkina, A. et al. Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78**, 11656-11663 (2004).

115. Geurts, A.M. et al. Gene transfer into genomes of human cells by the *Sleeping Beauty* transposon system. *Mol. Therap.* **8**, 108-117 (2003).

116. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-778. (2005).

117. Pryciak, P.M., Muller, H.P. & Varmus, H.E. Simian virus 40 minichromosomes as targets for retroviral integration in vivo. *Proc Natl Acad Sci U S A* **89**, 9237-41 (1992).

118. Groth, A.C., Olivares, E.C., Thyagarajan, B. & Calos, M.P. A phage integrase directs efficient site-specific integration in human cells. *Proc. Nat. Acad. Sci. USA* **97**, 5995-6000 (2000).

119. Thyagarajan, B., Olivares, E.C., Hollis, R.P., Ginsburg, D.S. & Calos, M.P. Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol. Cell. Biol.* **21**, 3926-3934 (2001).

120. Olivares, E.C. et al. Site-specific genomic integration produces therapeutic Factor IX levels in mice. *Nature Biotech.* **20**, 1124-1128 (2002).

121. Chen, L. & Woo, S.L.C. Complete and persistent phenotypic correction of phenylketonuria in mice by site-specific genome integration of murine phenylalanine hydroxylase cDNA. *Proc. Natl. Acad. Sci. USA* **102**, 15581-15586 (2005).

122. Kuduvalli, P.N., Mitra, R. & Craig, N.L. Site-specific Tn7 transposition into the human genome. *Nucl. Acids Res.* **33**, 857-863 (2005).

123. Vigdal, T.J., Kaufman, C.D., Izsvak, Z., Voytas, D.F. & Ivics, Z. Common physical properties of DNA affecting target site selection of *Sleeping Beauty* and other Tc1/mariner transposable elements. *J. Mol. Biol.* **323**, 411-452 (2002).

124. Carlson, C.M. et al. Transposon mutagenesis of the mouse germline. *Genetics* **165**, 243-256 (2003).

125. Wu, X., Li, Y., Crise, B., Burgess, S.M. & Munroe, D.J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79**, 5211-5214. (2005).

126. Grandgennet, D.P. Symmetrical recognition of cellular DNA target sequences during retroviral integration. *Proc. Nat. Acad. Sci. USA* **102**, 5903-5904 (2005).

127. Pryciak, P.M., Sil, A. & Varmus, H.E. Retroviral integration into minichromosomes in vitro. *EMBO J.* **11**, 291-303 (1992).

128. Muller, H.P. & Varmus, H.E. DNA-bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**, 4704-4714 (1994).

129. Liao, G.C., Rehm, E.J. & Rubin, G.M. Insertion site preferences of the P transposable element in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **97**, 3347-51 (2000).

130. Liu, G. et al. Target-site preference for *Sleeping Beauty* transposons. *J. Mol. Biol.* **346**, 161-173 (2005).

131. Geurts, A.M. et al. DNA structural patterns influence integration site preferences for mobile elements. *Nucl. Acids Res.* **34**, 2803-2811 (2006).

132. Posey, J.E., Pytlos, M.J., Sinden, R.R. & Roth, D.B. Target DNA structure plays a critical role in RAG transposition. *PLoS Biol.* **4**, e350 (2006).

133. Gorin, A.A., Zhurkin, V.B. & Olson, W.K. *B*-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.* **247**, 34-48 (1995).

134. Ivanov, V.I. et al. CRP-DNA complexes: inducing the *A*-like form in the binding sites with an extended central spacer. *J. Mol. Biol.* **245**, 228-240 (1995).

135. Lu, X.J., Shakked, Z. & Olson, W.K. *A*-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300**, 819-840 (2000).

136. Olson, W.K. et al. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* **313**, 229-237. (2001).

137. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. & Zhurkin, V.B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Nat. Acad. Sci. USA* **95**, 11163-11168. (1998).

138. Olson, W.K. & Zhurkin, V.B. Modeling DNA deformations. *Curr. Opin. Struct. Biol.* **10**, 286-297. (2000).

139. Brukner, I., Sanchez, R., Suck, D. & Pongor, S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* **14**, 1812-1818 (1995).

140. Mandal, P.K., Rawal, K., Ramaswamy, R., Bhattacharya, A. & Bhattacharya, S. Identification of insertion hot spots for non-LTR retrotransposons: computational and biochemical application to *Entamoeba histolytica. Nucl. Acids Res.* **34**, 5752-5763 (2006).

141. Nesbit, C.E., Tersak, J.M. & Prochownik, E.V. MYC oncogenes and human neoplastic disease. *Oncogene* **18**, 3004-3006 (1999).

142. Maragathavally, K.J., Kaminski, J.M. & Coates, C.J. Chimeric *Mos1* and *piggyBac* transposases result in site-directed integration. *FASEB J.* **20**, 1880-1882 (2006).

143. Yant, S.R., Huang, Y. & Kay, M.A. Fusion proteins consisting of the *Sleeping Beauty* transposase and the polydactyl zinc finger protein hE2C direct transposon integration into a unique human chromosomal sequence. *Mol. Ther.* **11, Suppl. 1**, S424 (2005).

144. CPMP. Insertional mutagenesis and oncogenesis: update from non-clinical and clinical studies. Gene Therapy Expert Group of the Committee for Proprietary Medical Products (CPMP). *J. Gene Med.* **6**, 127-129 (2004).
145. Levine, B.L. et al. Gene transfer in humans using a conditionally replicating lentiviral vector. *Proc. Natl. Acad. Sci. USA* **103**, 17372-17377 (2006).
146. Buchholz, C.J. & Cichutek, K. Is it going to be SIN? *J. Gene Med.* **8**, 1274-1276 (2006).
147. Chalberg, T.W. et al. Integration specificity of phage phiC31 integrase in the human genome *J. Mol. Biol.* **357**, 28-48 (2006).
148. Liu, J., Jeppesen, I., Nielsen, K. & Jensen, T.G. phiC31 integrase induces chromosomal aberrations in primary human fibroblasts. *Gene Ther.* **13**, 1188-1190 (2006).
149. Ehrhardt, A., Engler, J.A., Xu, H. & Kay, M.A. Molecular analysis of chromosomal rearrangements in mammalian cells after phiC31-mediated integration. *Human Gene Ther.* **17**, 1077-1094 (2006).
150. Collier, L.S. & Largaespada, D.A. Hopping around the tumor genome: transposons for cancer gene discovery. *Cancer Res.* **65**, 9607-9610 (2005).

Table 6.1. Properties of non-viral integrating vectors proposed for gene therapy

VECTOR		**PROPERTIES**	
SYSTEM[a]	**Activity[b]**	**Target Preferences[c]**	**Positive / Negative Attributes[d]**
SB	Standard*	TA sites, random	highly tested / cargo capacity decreases efficiency
φC31	lower	pseudo-*att* sites	highly tested / induces chromosomal mutations and rearrangements
PB	same	TTAA sites (genes)	too new to evaluate / targets transcription units
Tol2	higher	unknown	cured tyrosinemia type 1 in mice / may target genes, too new to evaluate
φBT1 l	lower	pseudo-*att* sites	cured PKU in mice / too new to evaluate
FP	same	TA sites	too new to evaluate

a) Systems: SB, *Sleeping Beauty*, PB, *piggyBac*, FP, *Frog Prince*

b) Activity: Relative to SB in HeLa cells or other cells where SB has been tested

c) Target sites for phage integrases φC31 and φBT1 are not found in mammalian genomes; sequences with similarities to the phage attachment sites (*att* sites) are targets, but they vary with cell type.

d) Evaluation with respect to gene therapy. Only SB and □C31 have been extensively tested, the others are too new to know positive and negative attributes. PKU = phenylketonuria.

Figure 6.1: Potential genetic consequences of integration of transgenic cassettes into chromatin. An expression cassette (orange box) in a viral or non-viral vector (represented by purple inverted arrowheads that indicate either inverted or direct terminal repeats) can integrate into four classes of chromatin. 1) Integration into heterochromatin will most likely result in the suppression of expression of the transgene and essentially no genetic consequences to the host. 2) Integration into intergenic regions of euchromatin is the most desirable outcome – the transgenic cassette is expressed leading to a gain-of function (GOF) in the host cell. 3) Integration into a transcriptional regulatory region can have several outcomes including expression (GOF) of the transgenic cassette, potentially modified by neighboring enhancer and silencer elements in the region. Regulatory elements in the transgenic cassette may either enhance expression of the neighboring gene (GOF for Gene X) or in rare cases block expression of an active gene. 4) Integration of the vector into a transcriptional unit may allow expression of the transgene but block expression of the host gene leading to a phenotypic loss of function (LOF). Integration within some genes can also lead to a dominant gain-of-function (DGF) and/or production of a dominant-negative form (DNF) of the original Gene X. A further discussion of effects of insertional mutagenesis can be found in refs [60,150].

Fig 6.1

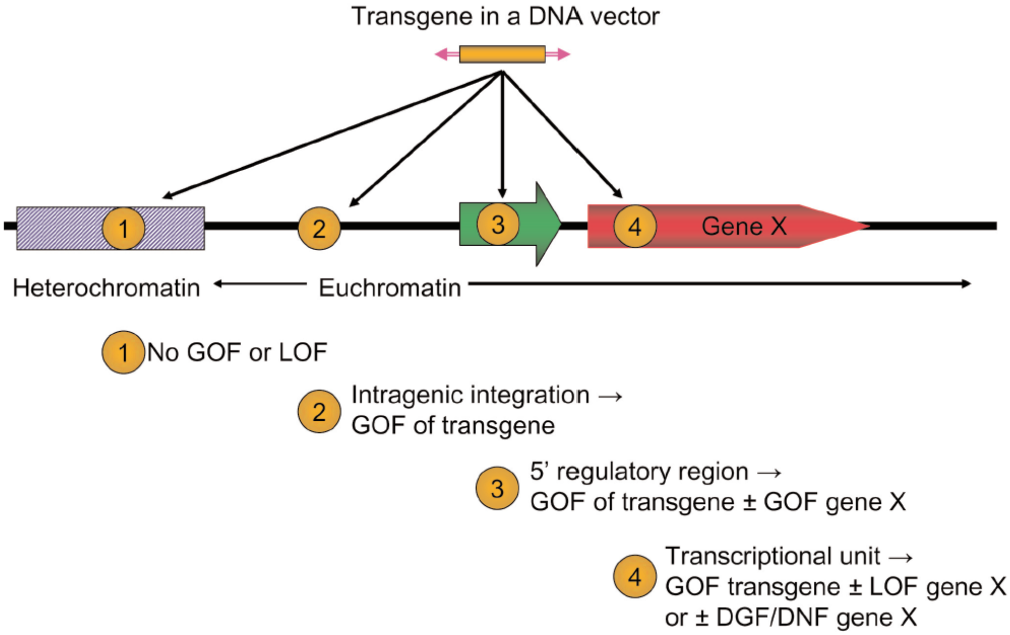

Transgene in a DNA vector
1
2
3
4
Gene X
Heterochromatin
Euchromatin
1 No GOF or LOF
2 Intragenic integration →
GOF of transgene
3 5' regulatory region →
GOF of transgene ± GOF gene X
4 Transcriptional unit →
GOF transgene ± LOF gene X
or ± DGF/DNF gene X

Figure 6.2: Deviations of DNA structure from the average B-form DNA that play a role modeling three-dimensional structures of specific DNA sequences. The figure illustrates physical parameters of B-form DNA structure that are altered in preferred sites for integration of insertional vectors. **A)** B-form DNA. **B)** A-DNA. Interactions between neighboring nucleotides govern the variable energy needed to convert from B- to A-DNA. The propensity of a sequence of B-form DNA to adopt the A-form is referred to as A-philicity [134]. **C)** Parameters of base-pair orientation affected by protein-DNA binding. *Twist* (horizontal looping arrow) refers to the rotation of base pairs around a central axis (heavy vertical black line); the average rotation between two base pairs is 36^0. *Tilt* (dotted lines) refers to the inclination of the base pairs with respect to the central axis; the average tilt is 0^o between basepairs, which are normally parallel in B-form DNA. *Rise* (vertical double arrowhead) is the distance between adjacent base pairs; the normal spacing is slightly more than 3.3Å, but can be more than 3.4Å at preferred target sites. *Slide* (horizontal double-arrowhead) refers to the shifting of the axis of a base pair out of alignment with the central axis. *Roll* (vertical looping arrow) refers to rotation of the nucleotide plane around a horizontal axis. A given base pair may be distorted in more than one of these parameters. V_{step} analysis is a method of examining these, and other physical parameters such as *shift*, in terms of a single number that derives from the transition from one base pair to another [130,137]. **D)** DNA bendability.

Fig 6.2

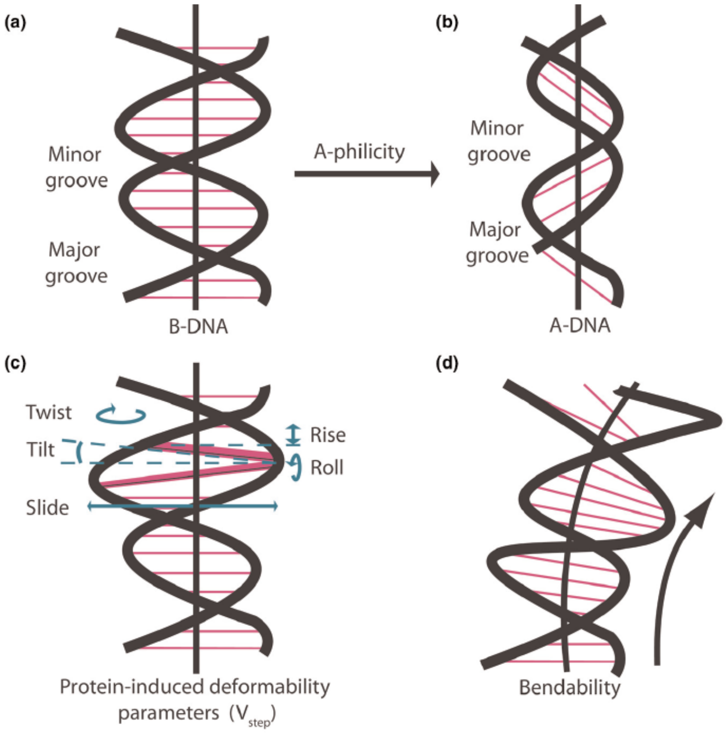

(a)
Minor groove
Major groove
B-DNA
A-philicity
(b)
Minor groove
Major groove
A-DNA
(c)
Twist
Tilt
Slide
Rise
Roll
Protein-induced deformability parameters (V_{step})
(d)
Bendability

Figure 6.3: Approaches to identification of DNA structural characteristics governing insertion site preferences for Tol2 and SB transposons. A) Averaging of all available insertion sites smoothes trends observed in individual plots. Plot of $\boldsymbol{V_{step}}$ profiles of 18 20-basepair Tol2 insertions (left, from ref [88]) compared to 18 randomly generated sequences (right). Averages are shown by thick black lines. Although individual Tol2 profiles appear jagged, peaks are not position-specific, thus, the plot of the average of 36 sites reveals only one small, distinct peak. Individual random sequences also appear jagged, however, an average of over 9,000 random sequences is a flat line. **B)** Analyses of Tol2-insertion site A-philicity profiles, compared to 18 random sequences. Trends are similar to V_{step} patterns. **C)** Plot of trinucleotide bendability for Tol2 and random sites, indicating only small common trends compared to random sequence. Random sequences in A-C acquired from a 10Mb portion of human chromosome 1p. **D)** Bendability plots for *Sleeping Beauty* insertion sites (from [105]). The average trinucleotide bendability at each position of 12-base insertion sites is shown for 574 insertions (red), as well as a subset of 189 insertions classified as "preferred" based on $\boldsymbol{V_{step}}$ profiles (dark blue). Random insertions are shown in light blue. This plot shows how identification of "preferred" sites can be useful to distinguish structural patterns for common insertion sites; preferred sites (based common patterns of protein-induced deformability with recurrently-hit sites) show an overall increase in a separate parameter, DNA bendability, when "basal" sites are removed.

Fig 6.3

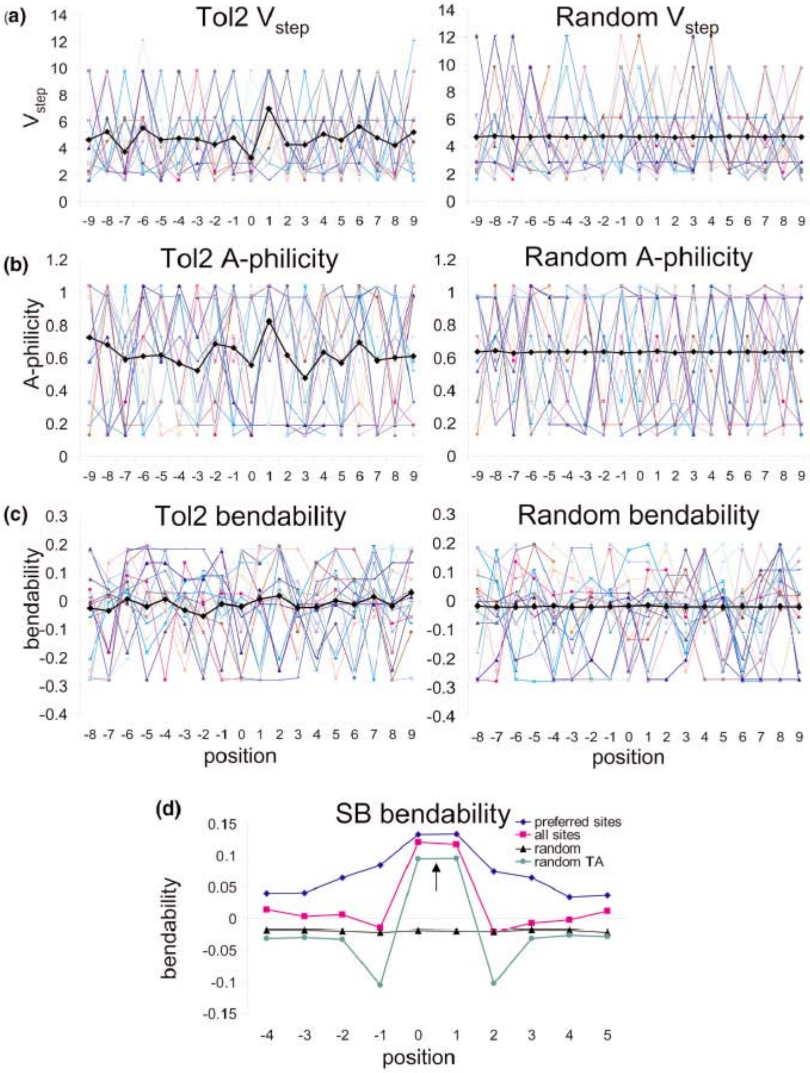
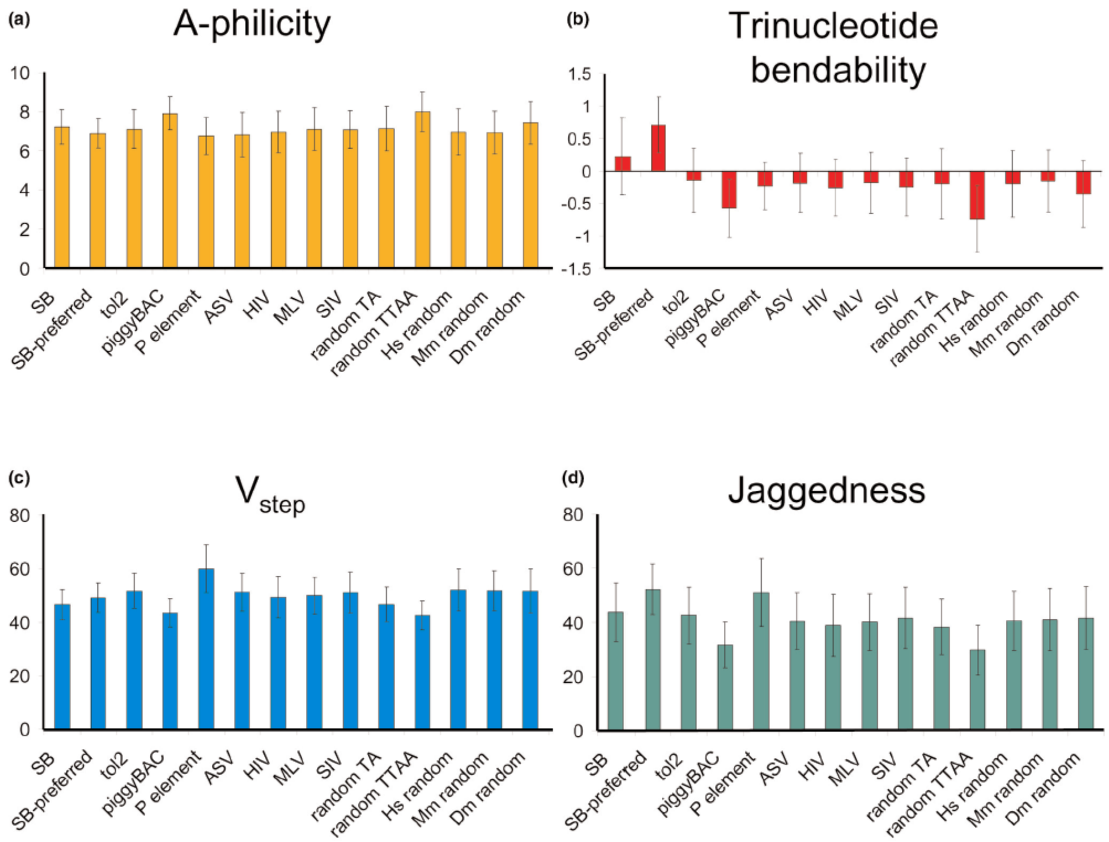

(a)
Tol2 V_{step}
Random V_{step}
V_{step}
14 12 10 8 6 4 2 0
-9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9
(b)
Tol2 A-philicity
Random A-philicity
A-philicity
1.2 1 0.8 0.6 0.4 0.2 0
-9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9
(c)
Tol2 bendability
Random bendability
bendability
0.3 0.2 0.1 0 -0.1 -0.2 -0.3 -0.4
-8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9
position
(d)
SB bendability
preferred sites
all sites
random
random TA
bendability
0.15 0.1 0.05 0 -0.05 -0.1 -0.15
-4 -3 -2 -1 0 1 2 3 4 5
position

Figure 6.4: Variability of DNA structural characteristics between insertion sites for various vectors. All A-philicity **(A)**, trinucleotide bendability **(B),** and ***V_{step}*** values **(C)** were summed across 12 nucleotides and averaged for all sites of each vector class. **D)** "Jaggedness" was measured by taking the absolute value of differences between adjacent ***V_{step}*** values, summed and average as in panels A-C. Error bars represent standard deviations. SB: 574 Sleeping Beauty integrations into human cells identified by Yant et al. [105]. SB-preferred: subset of 189 sites from the Yant dataset classified as "preferred" by ProTIS [115]. Tol2: 63 Tol2 integrations [88]. *PiggyBac*: 297 *piggyBac* insertions deposited into Genbank by Exelexis containing a single TTAA sequence flanked by 10 bases on each side. P-element: 920 P-element insertion sites mapped by Liao, et al [129]. ASV: 357 ASLV insertions into 293T-TVA cells. HIV: 334 HIV integrations into SubT1 cells. MLV: 695 MLV integrations into HeLa cells. SIV: 148 SIV integrations into CEMx164 cells. All P-element, ASV, HIV, MLV and SIV sequences kindly provided by Dr. Xioalin Wu. All sites were compared with three sets of over 9,000 randomly selected 12-mers from 10 Mb sections of human chromosome 1 (Hs), mouse chromosome 4 (Mm), and Drosophila chromosome 3L (Dm), and 10,000 randomly selected TA and TTAA sites from human chromosome 1.

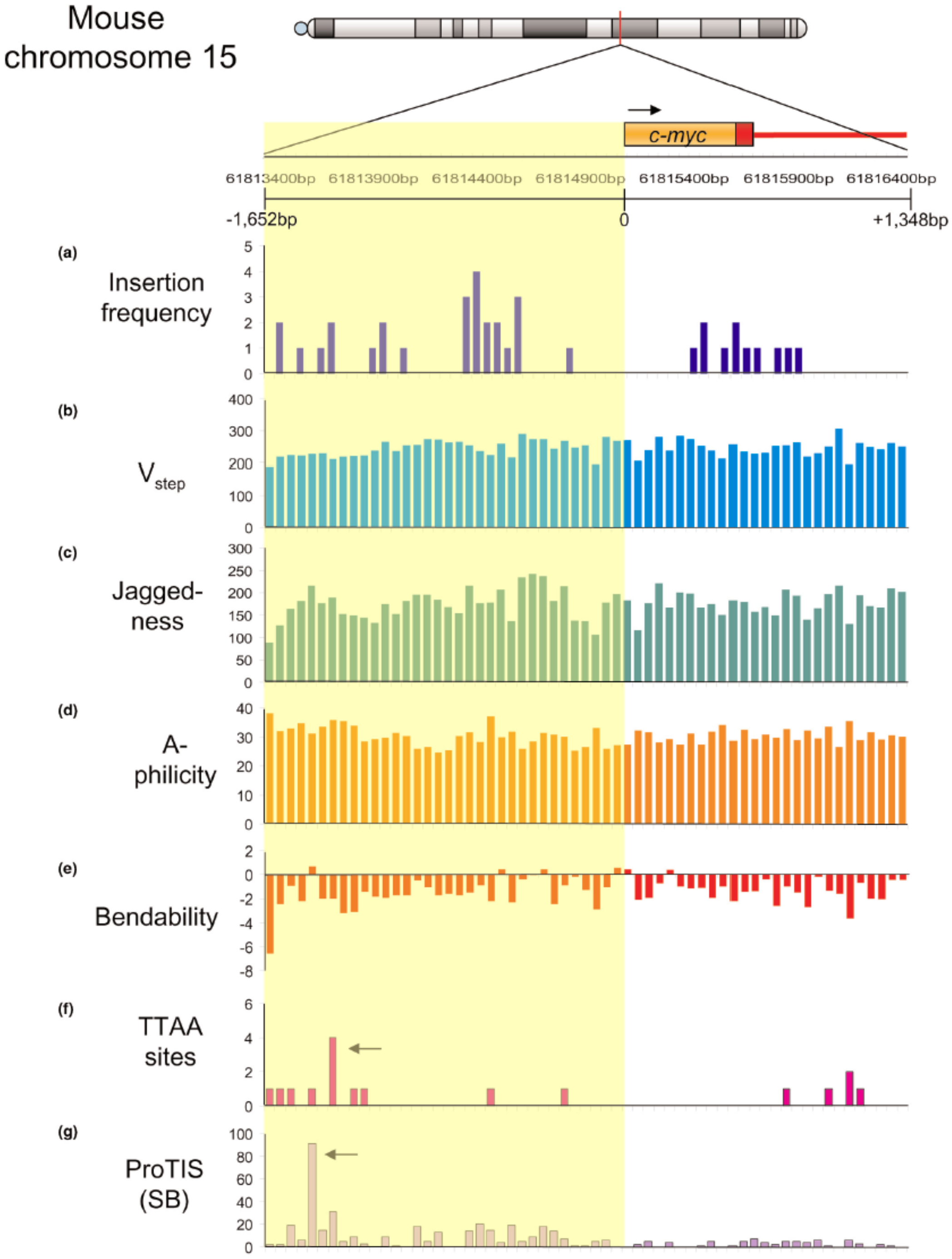
Fig 6.4



(a)
A-philicity
10
8
6
4
2
0
SB
SB-preferred
tol2
piggyBAC
P element
ASV
HIV
MLV
SIV
random TA
random TTAA
Hs random
Mm random
Dm random
(b)
Trinucleotide bendability
1.5
1
0.5
0
-0.5
-1
-1.5
SB
SB-preferred
tol2
piggyBAC
P element
ASV
HIV
MLV
SIV
random TA
random TTAA
Hs random
Mm random
Dm random
(c)
V_{step}
80
60
40
20
0
SB
SB-preferred
tol2
piggyBAC
P element
ASV
HIV
MLV
SIV
random TA
random TTAA
Hs random
Mm random
Dm random
(d)
Jaggedness
80
60
40
20
0
SB
SB-preferred
tol2
piggyBAC
P element
ASV
HIV
MLV
SIV
random TA
random TTAA
Hs random
Mm random
Dm random

Figure 6.5: SB insertions across the mouse *Braf* gene. 30 SB insertions deposited in the RTCGD were mapped across the entire *Braf* transcript and 10kb upstream (NCBI 36 build, note that *Braf* is transcribed right-to-left). Most oncogenic insertions occurred in introns 11 and 12 (formerly annotated as intron 9). ProTIS profiling across the entire gene reveals predicted hotspots for SB integration **(A)**, however, most actual integrations were found in a relatively low scoring-region corresponding to introns 11 and 12 **(B)**. A blowup of this local 4.9-kb region demonstrates that ProTIS scores **(C)** closely match patterns of actual transposon integration **(D)**.

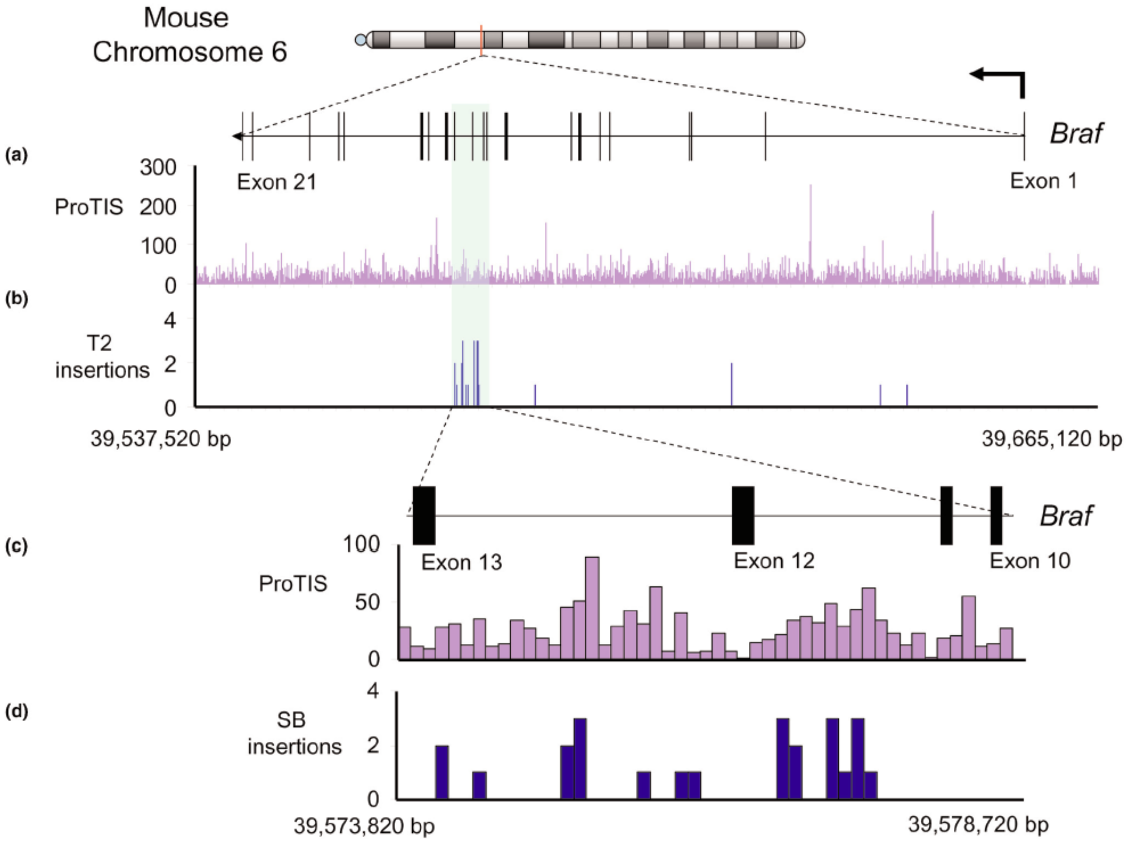
Fig 6.5



Mouse chromosome 15
c-myc
61813400bp
61813900bp
61814400bp
61814900bp
61815400bp
61815900bp
61816400bp
-1,652bp
0
+1,348bp
(a)
Insertion frequency
5
4
3
2
1
0
(b)
V_{step}
400
300
200
100
0
(c)
Jagged-ness
300
250
200
150
100
50
0
(d)
A-philicity
40
30
20
10
0
(e)
Bendability
2
0
-2
-4
-6
-8
(f)
TTAA sites
6
4
2
0
(g)
ProTIS (SB)
100
80
60
40
20
0

Figure 6.6: Insertion prediction for transposon vectors surrounding the *c-myc* locus on mouse chromosome 15. A 3-kb sequence from the mouse *c-myc* locus (from 61813400 to 61816400 bp) harboring 37 retroviral insertions submitted to the RTCGB (http://rtcgd.ncifcrf.gov/) is shown. The first exon and intron of *c-myc* are shown in orange; the upstream promoter sequence is shaded in yellow. **A)** Retrovirus insertion frequency per 50-bp segments. **B-G)** DNA structural characteristics at 50-bp resolution. **B)** Total $\boldsymbol{V_{step}}$ for each bin across the region. **C)** Total $\boldsymbol{V_{step}}$ jaggedness. **D)** Total A-philicity values. **E)** Total trinucleotide bendability. **F)** Number of TTAA sequences per 50-bp bin, representing the total number of possible *piggyBac* insertion sites. Notably, many regions harboring oncogene-selected retroviral insertions have few or no TTAA sequences, suggesting the likelihood of a *piggyBac* insertion causing an oncogenic event may be lower than that for retroviruses. Arrow represents a potential "hotspot" for integration, over 1 kb upstream of exon 1. **G)** *ProTIS* prediction shows a similar, low-incidence of preferred SB integration sites. Arrow indicates predicted hotspot for integration over 1kb upstream of exon 1, and slightly upstream of the TTAA hotspot.

Fig 6.6


Mouse
Chromosome 6
Braf
(a)
Exon 21
Exon 1
300
ProTIS 200
100
0
(b)
4
T2
insertions 2
0
39,537,520 bp
39,665,120 bp
Braf
(c)
100
Exon 13
Exon 12
Exon 10
ProTIS
50
0
(d)
4
SB
insertions 2
0
39,573,820 bp
39,578,720 bp

Chapter 7: Whole-body *Sleeping Beauty* mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality

Source: The following chapter was published as a manuscript in Cancer Research in 2009 (PMID 19843846)

Contributions: I performed the array CGH (sample preparation, hybridization, data acquisition, analysis and interpretation) leading to one of the major conclusions of the paper: that Sleeping Beauty does not cause tumors through large-scale genomic instability, but that the system can induce rearrangements around the donor concatemer. The data are presented in **Figures 7.4** and **7.8**, and I wrote the relevant sections of the text. The other authors contributed to the other aspects of the paper. The entire manuscript is included to provide context for my data, as well as to provide additional background for later chapters.

Whole-body *Sleeping Beauty* mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality

Lara S. Collier[1,2*], David J. Adams[3], Christopher S. Hackett[4], Laura E. Bendzick[1],

Keiko Akagi[5,6], Michael N. Davies[1], Miechaleen D. Diers[1], Fausto J. Rodriguez[7], Aaron M. Bender[7], Christina Tieu[7], Ilze Matise[8], Adam J. Dupuy[9], Neal G. Copeland[10], Nancy A. Jenkins[10], J. Graeme Hodgson[4,11], William A. Weiss[4], Robert B. Jenkins[7] &

David A. Largaespada[1*]

[1]Department of Genetics, Cell Biology and Development; Masonic Cancer Center; University of Minnesota; Minneapolis, MN. [2]Present address: School of Pharmacy; University of Wisconsin-Madison; Madison, WI. [3]Wellcome Trust Sanger Institute; Hinxton; Cambs; UK. [4]University of California, San Francisco, CA. [5]Mouse Cancer Genetics Program; National Cancer Institute at Frederick; Frederick, MD. [6] Present address: The Ohio State University Comprehensive Cancer Center; Columbus, OH. [7] Division of Experimental Pathology; Mayo Clinic; Rochester, MN [8]Masonic Cancer Center Histopathology Core; University of Minnesota; Minneapolis, MN. [9]Department of Anatomy and Cell Biology; University of Iowa; Iowa City, IA. [10]Institute of Molecular and Cell Biology, Singapore [11]Present address:

Funding: K01CA122183 and an American Cancer Society pre-doctoral fellowship (LSC), MN Dept of Employment and Economic Development SPAP-05-0013-P-FY06 (DAL and RBJ), R01CA113636-01A1 (DAL), R01NS055750 (WAW), Cancer Research-UK and the Wellcome Trust (DJA).

Authors for correspondence: Dr. Lara Collier, lcollier@wisc.edu,

Ph:608-890-2149, Fax:608-262-5345 or Dr. David Largaespada, larga002@umn.edu, Ph:612-626-4979, Fax:612-626-6140

Conflicts of interest: In the past, SB technology was exclusively licensed to Discovery Genomics Inc. (DGI), which is co-founded by DAL. DAL has an equity interest in and is an unpaid scientific advisor to DGI. DGI is pursuing the use of SB for human gene therapy. Neither DGI money nor personnel were involved in the work reported here. The University of Minnesota has a pending patent on the process of using transposons such as SB for cancer gene discovery. DAL, LSC, AJD, NAJ and NGC are named inventors.

Running head: Whole-body *Sleeping Beauty* mutagenesis phenotypes

Key words: Transposon, leukemia, glioma

Abstract

The *Sleeping Beauty* (SB) transposon system has been used as a somatic mutagen to identify candidate cancer genes. In previous studies, efficient leukemia/lymphoma formation on an otherwise wild-type genetic background occurred in mice undergoing whole-body mobilization of transposons, but was accompanied by high levels of embryonic lethality. To explore the utility of SB for large-scale cancer gene discovery projects, we have generated mice that carry combinations of different transposon and transposase transgenes. We have identified a transposon/transposase combination that promotes highly penetrant leukemia/lymphoma formation on an otherwise wild-type genetic background, yet does not cause embryonic lethality. Infiltrating gliomas also occurred at lower penetrance in these mice. SB-induced tumors do not harbor large numbers of chromosomal amplifications or deletions, indicating that transposon mobilization likely promotes tumor formation by insertional mutagenesis of cancer genes, and not by promoting wide-scale genomic instability. Cloning of transposon insertions from lymphomas/leukemias identified common insertion sites at known and candidate novel cancer genes. These data indicate that a high mutagenesis rate can be achieved using SB without high levels of embryonic lethality or genomic instability. Furthermore, the SB system can be used to identify new genes involved in lymphomagenesis/leukemiogenesis.

Introduction

Forward somatic cell genetic screens in model organisms are a powerful approach for the identification and validation of tumor suppressor genes (tsgs) and oncogenes relevant in human cancer [1-3]. Insertional mutagens such as retroviruses and transposable elements are frequently used for this purpose because the mutagen itself serves as a molecular tag, allowing rapid identification of mutagenized genomic loci. Candidate cancer genes are identified by finding regions of the genome that are insertionally mutated in multiple independent tumors, so-called common insertion sites (CISs).

The SB transposon system has been used as such an insertional mutagen. The SB system is bipartite; consisting of the mobilized piece of DNA, the transposon, and the enzyme that catalyzes the transposition reaction, the transposase [4]. Different combinations of SB transposon and transposase transgenics have been used for whole-body somatic cell genetic screens *in vivo* [5,6]. For these studies, different lines of mice harboring multiple copies of the T2/onc transposon in a head-to-tail arrangement in a chromosomally resident concatomer were utilized. Lines harboring about 25 copies of T2/onc in the donor concatomer were designated as low-copy lines [5] while lines harboring greater than 140 copies of T2/onc were designated as high-copy lines [6]. Two SB transposase transgenic lines were used to mobilize T2/onc throughout the soma. One transgenic was engineered with the SB11 version of the transposase "knocked" into the *Rosa26* locus (*Rosa26*–SB11) [6] while one transgenic expresses the SB10 version of the transposase under the control of the CAGGS promoter [7] (CAGGS-SB10) [5]. Mobilizing T2/onc from low-copy lines by CAGGS-SB10 could not generate tumors on an otherwise wild-type genetic background, yet did accelerate sarcoma formation in mice deficient for the tsg *p19Arf* [5]. T2/onc mobilization from high-copy lines by *Rosa26*-SB11 on an otherwise wild-type genetic background resulted in high levels of embryonic lethality which limited the number of transposon;transposase doubly transgenic mice that could be generated [6]. All mice surviving to birth eventually succumbed to tumors, primarily lymphocytic lymphoma/leukemia, by 120 days. Medulloblastoma and other hyperplasias/neoplasias were also observed at low penetrance. Cloning insertions from 15 lymphoma/leukemias and one medulloblastoma identified 33 CISs at known and candidate cancer genes, only a few of which had been previously identified in retroviral screens for lymphoma/leukemia genes [6].

The SB system is two-component (consisting of both transposons and transposase), so the possibility exists to modify each component individually to determine the effects on tumorigenesis. To this end, we crossed a T2/onc high-copy line to CAGGS-SB10 and two T2/onc low-copy lines to *Rosa26*-SB11. We have discovered that a rate of mutagenesis sufficient for promoting highly penetrant tumor formation yet insufficient for causing embryonic lethality can be achieved with the SB system. Leukemias/lymphomas predominate the tumor spectrum in mice undergoing whole-body transposon mutagenesis. Gliomas also occur with reduced penetrance, indicating that this tumor type can be modeled using SB mutagenesis. Furthermore, wide-spread genomic instability is not observed in SB-induced or accelerated tumors, suggesting that transposon insertional mutagenesis and not genomic instability drives tumorigenesis in these models. Transposon insertion sites from SB-induced leukemias/lymphomas identify CISs at both known and candidate novel cancer genes, suggesting that the SB system can reveal a different spectrum of cancer loci than retroviruses.

Materials and Methods

Mice: Mouse work was performed under University of Minnesota IACUC guidelines. All strains have been described [5,6,8]. At necropsy, tissues were snap frozen for DNA preparation and formalin fixed/paraffin embedded for pathological analysis at the Masonic Cancer Center Histopathology Core and the Mayo Clinic Tissue and Cell Molecular Analysis Shared Resource. Kaplan-Meier survival analysis was performed using Prism software.

Genotyping: Transposase transgenics were PCR genotyped using the following primers: 5'GGACAACAAAGTCAAGGTAT3' and 5'TAACTTGGGTCAAACGTTTC3'. T2/onc mice were genotyped as described [9].

Flow cytometry: Cell staining and flow cytometry techniques were as described [10,11]. Antibodies used were CY5-conjugated anti-CD4, APC-conjugated anti-CD8, FITC-conjugated anti-B220, PE-conjugated anti-TCRβ, PE-conjugated Gr1 and FITC-conjugated Mac1 (BD Biosciences, San Jose, CA). Data were analyzed using FlowJo software.

Linker-mediated PCR: For tumor DNA, linker-mediated PCR was performed as described [12]. PCR products from tumors were shotgun cloned into pCR4-Topo (Invitrogen, Carlsbad, CA). For each PCR, 96 bacterial colonies were robot picked, prepped and sequenced on the ABI 3730 platform. For the dataset from tail DNA, sequencing was performed on the 454 platform as described [9].

Insertion mapping and CIS analysis: Mapping of insertion sites to NCBI 36 build of the mouse genome was performed as described [6,13]. CIS analysis was based on published methods [14,15]. Because of the possibility for transposons to local hop after a prior mobilization, insertions from the same animal were not allowed to solely define a CIS. Insertion data is deposited in the RTCGD [13].

Array CGH: Tumor DNA samples (1 μg each) were labeled with Cy-3-dUTP and control DNA samples from muscle or spleen tissue (from the same animal when possible and from littermates in all other cases) were labeled with Cy-5 d-UTP essentially as described [16], with the omission of the *Dpn* II digest. Samples were combined and mixed with mouse Cot-1 DNA and hybridized to 1344-element BAC arrays [17] as described. Array images were captured using a CCD camera, and automated spot identification and statistical analysis was carried out using custom software [18] as described [16].

IHC: IHC for transposase was performed using the M.O.M. kit (Vector Laboratories, Burlingame, CA). Anti-transposase antibody (R&D Systems, Minneapolis, MN) was used at 1 μg/ml. Immunostain was developed using the ABC Vectastain peroxidase system (Vector Laboratories, Burlingame, CA), and sections were counterstained with hematoxylin. IHC was performed using anti-GFAP antibodies (Dako, Denmark, polyclonal, dilution 1:4000) and synaptophysin (ICN, Costa Mesa, CA, clone SY38, dilution 1:40) and showed a similar pattern in all gliomas examined. Primary antibody incubation was performed for 30 minutes, followed by 20 minutes in the Envision+Dual Link detection system on a Dako autostainer.

Results and Discussion

Combining CAGGS-SB10 with high-copy T2/onc does not result in tumor formation due to limited transposase expression

To determine if mobilization of T2/onc from high-copy lines by CAGGS-SB10 is sufficient to induce tumors, mice doubly transgenic for a T2/onc high-copy concatomer located on chromosome 4 [6] and the CAGGS-SB10 transgene [8] were generated. No evidence of embryonic lethality was observed (data not shown). CAGGS-SB10 only controls (n=11) and T2/onc high-copy;CAGGS-SB10 experimental mice (n=9) were aged and monitored for tumor formation for 18 months. No statistical difference in survival was observed (P=.6848, Log rank test) (**Figure 7.5**), indicating that the mutagenesis rate achieved by mobilizing T2/onc from a high-copy line by CAGGS-SB10 is insufficient for tumor formation on an otherwise wild-type genetic background.

To investigate if transposase expression levels influence mutagenesis rates, immunohistochemistry (IHC) was performed to detect transposase in CAGGS-SB10 and *Rosa26*-SB11 mice. Normal adult tissues were examined, as was a sarcoma from a *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy mouse [5]. Although transposase was detected in the sarcoma, it was absent from most normal somatic tissues in CAGGS-SB10 mice (**Figures 7.1** and **7.6**). When expression was detected, it occurred in a highly variegated pattern (liver in **Figure 7.1**; kidney in **Figure 7.6**). In contrast, transposase was robustly expressed in the majority of cell types in *Rosa26*-SB11 mice (**Figure 7.1** and **Figure 7.6**). The low level and variegated expression in CAGGS-SB10 mice is potentially due to epigenetic silencing that is often observed in standard transgenics.

The presence of transposase in a *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc sarcoma indicates that transposase is expressed in these mice in an appropriate cell type to promote sarcomagenesis. It could be hypothesized that T2/onc high-copy;CAGGS-SB10 mice could have developed sarcomas on an otherwise wild-type genetic background due to the availability of many T2/onc copies for mutagenesis. However, tumor formation was not observed. In murine models, *p19Arf* is known to play a role in oncogene-induced senescence [19-21]. Therefore, in sarcoma-initiating cells in *p19Arf*$^{+/+}$ mice, T2/onc mutagenesis of cancer genes could promote Arf-mediated senescence, providing a block to tumor formation. This experiment suggests that performing SB-screens in tumor-predisposed genetic backgrounds may be necessary for robust tumor formation in certain tissue types.

Combining T2/onc low-copy lines with *Rosa26*-SB11 does not cause embryonic lethality but promotes tumor formation in otherwise wild-type mice

To determine if mobilization from low-copy lines is sufficient for tumor formation, two T2/onc low-copy lines (lines 68 and 76 [5]) were crossed to *Rosa26*-SB11. Chi square analysis of the resulting progeny (**Table 7.1**) revealed no evidence for non-Mendelian inheritance of the transgenes (p= 0.4153, 3 degrees of freedom). A large cohort of doubly transgenic mice and single transgenic littermate controls were therefore generated. T2/onc low-copy;*Rosa26*-SB11 mice became moribund with an average latency of 187 days while controls had normal lifespans (**Figure 7.2A**). Separate analysis of each T2/onc low-copy line revealed that T2/onc low-copy line 68;*Rosa26*-SB11 mice develop disease much more rapidly than T2/onc low-copy line 76;*Rosa26*-SB11 mice (**Figure 7.7**). However, the tumor spectrum was the same for both lines. At necropsy, 89% (97 of 109) of analyzed doubly transgenic mice had signs of hematopoietic disease including splenomegaly, lymphadenopathy and/or an enlarged thymus. Twenty-seven mice with hematopoietic involvement were analyzed by veterinary pathologists at the Masonic Cancer Center Comparative Pathology core. Nineteen mice were diagnosed with lymphocytic lymphoma/leukemia (**Figure 7.2B**), three with hematopoietic neoplasia of undetermined lineage, four with hematopoietic hyperplasia and one with myeloid leukemia. Flow cytometry analysis for cell surface markers on nineteen tumors verified that the majority of leukemias/lymphomas arising in these mice are phenotypically T-cell lymphocytic disease (**Table S1** accompanying the online version of this manuscript).

Several mice presented with neurological symptoms at morbidity. Medulloblastomas, a tumor of the cerebellum, occurred at low penetrance in T2/onc high-copy;*Rosa26*-SB11 doubly transgenic mice [6]. To determine if medulloblastomas also occur in T2/onc low-copy;*Rosa26*-SB11 mice, 82 brains were extensively sectioned for pathology. Fourteen brain tumors were discovered. One tumor was a sarcoma growing on the surface of the brain while one was a glioma of undetermined origin (data not shown). Histopathological analysis determined that the remaining tumors were high-grade astrocytomas (**Figure 7.3A, B**). Pseudopalisading necrosis was present in three cases that were therefore classified as glioblastomas (**Figure 7.3C**). IHC for Glial fibrillary acidic protein (GFAP) and synaptophysin were performed on a sub-set of tumors to confirm the diagnosis (**Figure 7.3D**). Immunoreactivity for GFAP was noted at least focally in all tumors examined, supporting the diagnosis of astrocytomas. No gliomas were detected in 28 aged-matched transposon or transposase only mice sacrificed for analysis, indicating that T2/onc mobilization induces high-grade astrocytomas. Molecularly defined hyperproliferative lesions were also found in the prostates of moribund T2/onc low-copy;*Rosa26*-SB11 mice [22]; but no additional overt tumor types were commonly observed despite the fact that transposase is expressed in most cell types. The aggressive nature of the leukemias/lymphomas and gliomas in these mice limits animal survival, and therefore likely prevents the ability of transposon mobilization by *Rosa26*-SB11 to model more slowly developing tumor types.

In contrast to T2/onc high-copy lines, no embryonic lethality was observed in T2/onc low-copy;*Rosa26*-SB11 mice. Embryonic lethality was proposed to potentially result from unrepaired DNA damage after transposition [6]. The lower number of mobilizing transposons in T2/onc low-copy;*Rosa26*-SB11 mice could result in fewer double strand breaks for cellular machinery to repair. Another explanation for embryonic lethality in T2/onc high-copy;*Rosa26*-SB11 double transgenics could be the generation of concatomer-associated rearrangements which can accompany SB germline transposition [23]. High transposition rates associated with high-copy concatomers could increase the severity of these rearrangements. Whatever the cause of embryonic lethality, the Mendelian inheritance of *Rosa26*-SB11 with T2/onc low-copy transgenes indicates that a mutagenic rate sufficient to promote tumor formation but insufficient to interfere with normal development can be achieved.

Brain tumors have been found in mice in which T2/onc is mobilized by *Rosa26*-SB11. In high-copy lines, the tumors were medulloblastomas [6] while in low-copy lines the tumors were infiltrating gliomas. Differences in tumor latency could contribute to differences in brain tumor type. Medulloblastomas are predominantly found in children, and it is hypothesized that they arise from granule cell precursor cells. It is hypothesized that most granule cell precursors have completed proliferation and differentiation by adulthood, and therefore fewer potential medulloblastoma-initiating cells exist in adults. The high mutagenesis rate in T2/onc high-copy;*Rosa26*-SB11 mice could allow enough mutations in cancer genes to occur prior to terminal differentiation. Conversely, in T2/onc low-copy;*Rosa26*-SB11 mice, the mutagenesis rate may be too low to promote tumor formation prior to terminal differentiation of these cells.

This slower mutagenesis rate could still promote mutagenesis in the longer-lived glial precursor cell.

Somatic mobilization of transposons does not cause substantial genomic instability

Transposition of DNA transposons involves double strand break formation and repair [24]. It is possible that somatic mobilization of SB transposons could cause tumor formation by promoting genomic stability due to illegitimate repair of these breaks. To investigate this possibility, BAC-based array comparative genomic hybridization (CGH) [17] was used to look at genomic copy number changes in six T2/onc low-copy;*Rosa26*-SB11 lymphomas/leukemias and three sarcomas from *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy mice [5] using non-tumor DNA as a reference sample (**Figure 7.4**). Two spontaneously arising sarcomas in *p19Arf*$^{-/-}$ mice served to demonstrate the ability of the BAC array platform to detect copy number changes in tumors (**Figure 7.8**). Whole chromosome gains or losses were rarely detected in tumors with mobilizing transposons (for example, the gain of chromosomes 14 and 15 in 76Rosa521 lymph node). Deletions or amplifications defined by one or two adjacent probes were occasionally detected. However, for line 76 leukemias/lymphomas, two of three displayed evidence of amplification or deletion on chromosome 1 at probes flanking 166Mb, the approximate location of the T2/onc concatomer in line 76 [5] (**Figure 7.4B**). The exact chromosomal location of the line 68 concatomer has not been determined, but FISH and local hopping patterns have placed it at approximately 45Mb on chr. 15. One (68Rosa467 spleen) of 3 leukemias/lymphomas from T2/onc line 68 and all three sarcomas from *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy mice line 68 showed evidence for amplification or deletion on chromosome 15.

Previously, chromosomal rearrangements flanking a SB transposon donor concatomer on chromosome 11 have been detected as a result of transposition in the germline and in normal splenocytes [23]. The array CGH reported here indicates that this phenomenon is not limited to the concatomer on chromosome 11, and that they can occur in SB-induced tumors. The data suggest that SB-induced tumorigenesis does not promote genome-wide instability, but does frequently cause amplifications and/or deletions flanking the donor concatomer that could contribute to tumor formation if the donor concatomer happens to reside near a tsg or oncogene. As other DNA elements are known to cause genomic rearrangements [25], it will be important to determine if additional transposons proposed to be used as somatic mutagens including *piggyBac* [26], *Tol2* [27] and *Minos* [28] also promote deletions or amplifications flanking the donor locus.

T2/onc local hops in somatic cells

Cancer gene identification using insertional mutagens relies on performing CIS analysis to identify chromosomal regions where transposons have inserted in tumors at a rate greater than that expected by random chance. *In vivo*, SB is known to have a preference for re-inserting at loci linked to the starting integration site, a phenomenon termed local hopping [29], which complicates CIS analysis. Previously, an unselected insertion set (n=490) obtained from T2/onc high-copy;*Rosa26*-SB11 embryos was used to examine local hopping rates from chromosomal concatomers in somatic cells *in vivo* [6]. Although only 6-11% of insertions were found in the 25-megabase region surrounding donor concatomers, 38.6% of insertions were found on the same chromosome as the donor concatomer. However, it is unclear if transposon copy number or the chromosomal location of a concatomer influences local hopping rates.

To determine how local hopping from low-copy concatomers may influence analysis of SB transposon tumor insertion sites, a dataset of unselected SB transposon insertion sites from tail biopsy DNA isolated from 14-21 day old T2/onc low-copy;*Rosa26*-SB11 mice was generated [9,30]. As no leukemias have been observed in mice this young, the insertions cloned from this material likely represent the SB transposon insertion site profile under un-selected conditions. 16,411 unique SB insertion sites were cloned from 88 doubly transgenic mice using linker-mediated PCR and 454 pyrosequencing methods (**Table S2**). Of these, 22.8% were located in the 25 Mb surrounding the donor concatomer, indicating that transposons in low-copy concatomers local hop in somatic cells. Furthermore, 49.3% of insertions were located on the same chromosome as the donor concatomer (8095 of 16411). The percentage of insertions on the donor chromosome contrasts with T2/onc high-copy line embryo insertions in which 38.6% of insertions reside on the same chromosome as the donor concatomer. This could potentially be explained by differences in local hopping rates from different copy number concatomers. Nevertheless, both of these datasets indicate that in somatic cells, SB transposons have a local hopping interval that encompasses the whole chromosome on which the concatomer resides.

CIS analysis was performed on this control dataset after removal of insertions mapping to the donor chromosomes (n= 8316 insertions) using criteria described by Mikkers *et al* [14] at an expected fraction (Efr) of 0.001 (2 insertions in .325 kb, 3 insertions in 14.75 kb and 4 insertions in 62 kb), which is predicted by Monte Carlo simulations to result in approximately 25 CISs being identified by random chance alone. 43 CISs were identified in the control dataset. Clustering criteria using an Efr of 0.005 was also applied (2 insertions in 1.625 kb, 3 in 33.75 kb and 4 in 109.75 kb) which is predicted to result in a total of 124.5 CISs identified based on random chance alone. Using this criteria, 134 CISs were identified (of which 43 also met the criteria used above for an Efr of 0.001) (**Table S3**).

More CISs were identified in this control dataset than would be predicted by Monte Carlo simulations. This observation could actually be due to random chance, as Monte Carlo simulations predictions of false positive clustering rates are based on averages for an infinite number of experiments. Therefore, 50% of the time a random dataset of 8316 insertions is generated, the number of CISs identified at an Efr of .005 would be ≥ 124.5 and 50% of the time the number of CISs identified would be ≤ 124.5. Alternatively, SB is known to have some insertion preferences for specific sequences or DNA conformations [31,32], and preferred SB insertion sites may not be randomly distributed through the genome. Five of the CISs in the control dataset are also CISs in leukemias (see below), supporting the hypothesis that there are genomic "hot spots" for SB integration. Interestingly, CISs are found in the control dataset on proximal regions of chrs. 5, 11, 12, 13 and 18; indicating the possibility that SB transposons have affinity for inserting into centromeric regions. The generation and analysis of additional datasets under unselected conditions will help refine the statistical methods used for CIS analysis in SB-induced tumors.

T2/onc identifies candidate genes involved in lymphoma/leukemia formation not identified by MuLV

To determine if T2/onc identifies new lymphoma/leukemia cancer genes, 2296 independent T2/onc insertion sites were cloned from 59 lymphomas/leukemias from 58 T2/onc low-copy;*Rosa26*-SB11 animals (**Table S4**). Local hopping was observed as 13.1% of insertions from line 68 occurred within the 25 Mb surrounding the donor concatomer, 18.6% within the 40 Mb surrounding the donor concatomer and 33.6% on the entire donor chromosome. For line 76 the local hopping percentages were similar at 11.2%, 15.6% and 31.5%, respectively.

The local hoping rate in T2/onc low-copy;*Rosa26*-SB11 tumors is intermediate between those reported for sarcomas from *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy mice (23% of insertions within the 40Mb surrounding the donor concatomer [5]) and that reported for T2/onc high-copy;*Rosa26*-SB11 leukemias/lymphomas ("little local hopping" [6]). The chromosomal location and environment of the concatomers could influence local hopping rates. Transposase levels could also influence local hopping rates as differences were observed between *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy sarcomas and *Rosa26*-SB11;T2/onc low-copy leukemias/lymphomas. The lower local hopping rates in *Rosa26*-SB11;T2/onc low-copy leukemias/lymphomas compared to weaning tail biopsies from the same cohort of mice could be due to increased time for transposons to remobilize in tumors from older mice compared to tissue from young mice.

CIS analysis was performed on insertions cloned from 59 *Rosa26*-SB11;T2/onc low-copy induced leukemias/lymphomas after removal of insertions residing on the donor concatomer chromosome (n=1547 insertions) to identify candidate leukemia/lymphoma genes. Analysis was performed using criteria described by Mikkers *et al* [14] at an expected fraction (Efr) of 0.001 (2 insertions in 1.95 kb, 3 insertions in 88.5 kb and 4 insertions in 371.5 kb), which is predicted to result in 4.5 CISs being identified by random chance alone. This resulted in the identification of 28 CISs in the leukemia dataset. Clustering criteria of insertions using an Efr of 0.005 was also applied (2 insertions in 9.75kb, 3 in 202kb and 4 in 658.5 kb) which is predicted to result in a total of 22.5 CISs identified based on random chance alone. Using this criteria a total of 49 CISs were identified (of which 28 also met the criteria used above for an Efr of 0.001). Five of these CISs were also CISs in the unselected dataset, indicating that they likely do not tag a locus important for cancer formation. Removal of these resulted in a final total of 44 CISs in leukemias/lymphomas (**Table 7.2**).

Of the 44 CISs identified, 15 are CISs in the RTCGD of retroviral screens for lymphoma/leukemia genes or in a recent report analyzing MuLV insertions from over 500 tumors [13,33] (**Table 7.2**). Therefore, the majority of CISs identified by SB have not been previously identified in retroviral screens. Notably, in lymphomas/leukemias resulting from mobilization from T2/onc low-copy lines, a CIS is found in *Pten*. Although an important tsg in the hematopoietic system and other cancers [34,35], *Pten* has not been previously identified as a CIS in retroviral screens [36]. This supports the hypothesis that SB can identify cancer genes that are not readily tagged by retroviruses, including tsgs. Only seven CISs were common between the leukemia/lymphoma dataset described here and the dataset from the 15 leukemias/lymphomas generated using high-copy lines [6], indicating that cloning insertions from a larger cohort of tumors increases CIS identification power.

In summary, by combining T2/onc low-copy lines with *Rosa26*-SB11 we have achieved whole-body mobilization rates that are sufficient to promote penetrant tumorigenesis without the complication of embryonic lethality or genomic instability. Although lymphomas/leukemias predominant the tumor spectrum, whole-body mobilization of T2/onc can also promote glioma formation including glioblastoma, a tumor type in humans with an extremely poor prognosis. In lymphomas/leukemias, T2/onc tags both known and candidate novel cancer genes. Recent reports have demonstrated the ability of T2/onc mobilization by tissue-specific transposase expression to generate liver tumors and intestinal tumors useful for candidate cancer gene discovery [9,30]. In these models, true carcinoma/adenocarcinoma on a wild-type genetic background occurred with long latency and incomplete penetrance, indicating that additional improvements to increase mutagenesis rates are still needed for the SB system. Our data indicate that such mutagenesis rates can be obtained without undesired consequences such as lethality or genome-wide instability and that further development of the SB system is warranted for cancer gene discovery in a wider range of cell types.

Acknowledgements

We thank Erin Riley for technical assistance, Paul Marker and Michael Taylor for critical reading of the manuscript and the members of the Center for Genome Engineering for many helpful discussions.

References

1. Uren, A.G., Kool, J., Berns, A. & van Lohuizen, M. Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24**, 7656-72 (2005).
2. Collier, L.S. & Largaespada, D.A. Transforming science: cancer gene identification. *Curr Opin Genet Dev* **16**, 23-9 (2006).
3. Callahan, R. & Smith, G.H. MMTV-induced mammary tumorigenesis: gene discovery, progression to malignancy and cellular pathways. *Oncogene* **19**, 992-1001 (2000).
4. Ivics, Z., Hackett, P.B., Plasterk, R.H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-10 (1997).
5. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6 (2005).
6. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6 (2005).
7. Okabe, M., Ikawa, M., Kominami, K., Nakanishi, T. & Nishimune, Y. 'Green mice' as a source of ubiquitous green cells. *FEBS Lett* **407**, 313-9 (1997).
8. Dupuy, A.J., Fritz, S. & Largaespada, D.A. Transposition and gene disruption in the male germline of the mouse. *Genesis* **30**, 82-8 (2001).
9. Starr, T.K. et al. A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323**, 1747-50 (2009).
10. Kim, W.I., Matise, I., Diers, M.D. & Largaespada, D.A. RAS oncogene suppression induces apoptosis followed by more differentiated and less myelosuppressive disease upon relapse of acute myeloid leukemia. *Blood* **113**, 1086-96 (2009).
11. Kim, W.I., Wiesner, S.M. & Largaespada, D.A. Vav promoter-tTA conditional transgene expression system for hematopoietic cells drives high level expression in developing B and T cells. *Exp Hematol* **35**, 1231-9 (2007).
12. Largaespada, D.A. & Collier, L.S. Transposon-mediated mutagenesis in somatic cells: identification of transposon-genomic DNA junctions. *Methods Mol Biol* **435**, 95-108 (2008).
13. Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A. & Copeland, N.G. RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res* **32**, D523-7 (2004).
14. Mikkers, H. et al. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nat Genet* **32**, 153-9 (2002).
15. Johansson, F.K. et al. Identification of candidate cancer-causing genes in mouse brain tumors by retroviral tagging. *Proc Natl Acad Sci U S A* **101**, 11334-7 (2004).
16. Hackett, C.S. et al. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-73 (2003).
17. Hodgson, J.G. et al. Copy number aberrations in mouse breast tumors reveal loci and genes important in tumorigenic receptor tyrosine kinase signaling. *Cancer Res* **65**, 9695-704 (2005).

18. Jain, A.N. et al. Fully automatic quantification of microarray image data. *Genome Res* **12**, 325-32 (2002).
19. de Stanchina, E. et al. E1A signaling to p53 involves the p19(ARF) tumor suppressor. *Genes Dev* **12**, 2434-42 (1998).
20. Zindy, F. et al. Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis and immortalization. *Genes Dev* **12**, 2424-33 (1998).
21. Palmero, I., Pantoja, C. & Serrano, M. p19ARF links the tumour suppressor p53 to Ras. *Nature* **395**, 125-6 (1998).
22. Rahrmann, E.P. et al. Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping Beauty transposon-based somatic mutagenesis screen. *Cancer Res* **69**, 4388-97 (2009).
23. Geurts, A.M. et al. Gene Mutations and Genomic Rearrangements in the Mouse as a Result of Transposon Mobilization from Chromosomal Concatemers. *PLoS Genet* **2**(2006).
24. van Luenen, H.G., Colloms, S.D. & Plasterk, R.H. The mechanism of transposition of Tc3 in C. elegans. *Cell* **79**, 293-301 (1994).
25. Gray, Y.H. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* **16**, 461-8 (2000).
26. Ding, S. et al. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-83 (2005).
27. Balciunas, D. et al. Harnessing a High Cargo-Capacity Transposon for Genetic Applications in Vertebrates. *PLoS Genetics* **In press.**(2006).
28. Zagoraiou, L. et al. In vivo transposition of Minos, a Drosophila mobile element, in mammalian tissues. *Proc Natl Acad Sci U S A* **98**, 11474-8 (2001).
29. Luo, G., Ivics, Z., Izsvak, Z. & Bradley, A. Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* **95**, 10769-73 (1998).
30. Keng, V.W. et al. A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-74 (2009).
31. Liu, G. et al. Target-site preferences of Sleeping Beauty transposons. *J Mol Biol* **346**, 161-73 (2005).
32. Geurts, A.M. et al. Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Res* **34**, 2803-11 (2006).
33. Uren, A.G. et al. Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133**, 727-41 (2008).
34. Chow, L.M. & Baker, S.J. PTEN function in normal and neoplastic growth. *Cancer Lett* **241**, 184-96 (2006).
35. Maser, R.S. et al. Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447**, 966-71 (2007).
36. http://rtcgd.ncifcrf.gov.

Table 7.1: No evidence for non-Mendelian transgene inheritance in the T2/onc low-copy;*Rosa*26-SB11 cross.

genotype (T2/onc,*Rosa*26-SB11):	+,-	-,+	+,+	-,-
number of mice:	123	138	112	129

T2/onc low-copy heterozygous mice were crossed to *Rosa*26-SB11 heterozygous mice and the resulting progeny were genotyped for each transgene. The four possible genotypes and number of mice observed for each genotype are shown.

Table 7.2: CISs identified in leukemias/lymphomas from T2/onc;*Rosa*26-SB11 mice.

CIS name	Chr	CIS address start	CIS address end	# of tumors	# of independent insertions	other genes in CIS interval
Myb *	10	20824216	20921673	4	4	
Ube2d1 #	10	70669435	70673276	2	2	
Stab2 #	10	86412419	86505987	3	3	
Ikzf1 *	11	11407716	11749020	14	18	4930415F15Rik, 4930512M02Rik, Q3UW08, Fignl1, Ddc
Csf2 #	11	54031828	54569276	4	4	ENSMUSG00000060068, Il3, Acsl6, 4930404A10Rik, Fnip1, Rapgef6, Cdc42se2
Cox10 #	11	63785103	63794563	2	2	
Rab5c *#	11	100370989	100543223	3	3	Ttc25, Cnp1, Dnajc7, Nkiras2, A930006D11Rik, D11Lgp2e, Gcn5l2, Hspb9
Dusp22	13	30572119	30813368	4	4	Irf4
Ibrdc2 #	13	47050885	47464428	4	5	Tpmt, Aof1, Dek
H2afy *#	13	56100124	56645908	4	4	BC027057, Neurog1, Cxcl14, Q8CDW6, AU042651, Il9, Fbxl21, Lect2, Q3U1K8, Tgfbi
Mef2c *	13	83767562	84048492	5	5	
Cenpk	13	105370547	105371252	2	2	
Zmiz1 *	14	24269629	24414687	5	5	
Heg1	16	33678449	33679145	2	2	
Btla #	16	44873922	45525600	4	4	Cd200r3, Ccdc80, Q3UVS9, Slc35a5, Atg3, EG547267, Cd200, Gm609, ENSMUST00000060550, Slc9a10
Erg *	16	95072653	95570852	20	10	Kcnj6, Kcnj15
Heatr5b #	17	78655675	78761809	3	3	ENSMUST00000059920, 2310002B06Rik, ENSMUST00000043373, Eif2ak2
Mbd2 *	18	70309981	70807502	7	6	2310002L13Rik, ENSMUST00000096551, 4930503L19Rik, Poli, ENSMUST00000031200
Pten	19	32611775	32885582	7	12	Papss2, Atad1, B430203M17Rik
Notch1 *	2	26281337	26321310	21	20	
Zbtb34	2	33060035	33919549	5	6	Angptl2, Ralgps1, Lmx1b, C130021I20Rik, ENSMUST00000100174, 2610528K11Rik
Rasgrp1 *	2	117030934	117031956	2	2	
Gm414	3	70203770	70205270	2	2	
Ppp3ca #	3	136626657	136865395	2	6	
Bach2 *#	4	32416902	33046612	4	6	XR_001707.1, Q8BQ29, ENSMUST00000093133, Gja10, Casp8ap2, Mdn1
Ptpn12 #	5	20125026	21032439	5	5	Magi2, Q3U0Y7, Q8CEH7, Phtf2, Tmem60, Rsbn1l, ENSMUST00000053060, EG626903, A530088I07Rik, 4930528G09Rik, Fgl2, AI847670, Fbxl13
Kit/Kdr *	5	75834997	76133546	4	5	
AB041803	6	31101079	31233559	4	4	
Wnk1 #	6	119698409	120263825	4	4	Erc1, 3110021A11Rik, ENSMUST00000036010, Rad52, EG406236, mmu-mir-706, Ninj2, B4galnt3, ENSMUSG00000053059
Etv6	6	134104562	134557768	4	4	Bcl2l14, Lrp6, Q8BPW4
Akt2	7	27305138	27309525	3	3	
Klf13	7	63514893	63864073	4	4	Otud7a, ENSMUST00000003521
Mctp2 #	7	72254322	72263587	2	2	
Eed	7	89832779	89844617	2	3	
EG209380 #	7	105978304	106615569	4	4	Gvin1, Q922V0, Q9D303, ENSMUST00000071162, Gm1966, Olfr693-701
Zfp629	7	127271619	127383373	4	4	Fbs1, Q8C4A9, D030022P06Rik, Phkg2, Gm166, Rnf40
Dcun1d5 #	9	7186494	7196156	2	2	
Naalad2 #	9	18088306	18093482	2	2	
Fli1 *	9	31723619	32229349	6	5	Grit, Kcnj5, Kcnj1
BC033915	9	45938617	45993423	3	4	Apoa1, Tcea1, Apoc3, Efhc1
4833427G06Rik	9	50485627	50898988	4	4	Dixdc1, 2310030G06Rik, Cryab, Hspb2, 1110032A03Rik, D630004A14Rik, Alg9, Ppp2r1b, Snf1lk2, Layn, Btg4, mmu-mir-34b,c
Tcf12 * #	9	71786484	71892068	3	3	
Eras * #	X	7019492	7220920	3	3	Otud5, Pim2, Slc35a2, Pqbp1, Timm17b, ENSMUST00000085330, Q3UUQ2, Pcsk1n, Hdac6, Gata1, 2010001H14Rik, EG632013, Suv39h1
Tbl1x #	X	73774303	74215530	4	4	EG628893, Prkx, Pbsn

The name of the CIS is presented along with the chromosome (chr), the base pair of the first insertion defining the CIS, the base pair of the last insertion defining the CIS, the number of tumors defining the CIS, the number of insertions in independent TA dinucleotides, additional Ensembl annotated genes found within the bounds of the CIS. CISs previously identified in MuLV mutagenesis studies [13,33] are marked with an * and those defined by an expected fraction (Efr) of .005 are marked with a #.

Note: Supplementary tables are not included, but accompany and are available with the online version of the published manuscript.

Supplementary Table 1: A summary of immunophenotypes of leukemias/lymphomas from *Rosa26*-SB11;T2/onc low-copy mice. The mouse number, tissue type as well as the surface markers present on the tumor cells are presented (CD4 and CD8 for T cells, B220 for B cells). In some tumors, a secondary population expressing different markers was also present. Tumor cells from 68Rosa39 spleen were positive for B220 and therefore were classified as B cell disease. The remaining tumors display markers of mature (CD4 or CD8 single positive) or immature (CD4/8 double positive) T cells. One tumor (76Rosa267 lymph node) was negative for CD4, CD8 and B220 expression, but was found to express the T cell receptor β chain (TCRB) and was therefore classified as immature T cell disease.

Supplementary Table 2: The chromosomal locations of unselected T2/onc insertions. T2/onc insertions were cloned from weaning tail clips of *Rosa26*-SB11;T2/onc low-copy mice. The chromosome (chr) and basepair of each insertion is presented, followed by the mouse symbol of the closest gene (within 100kb, N/A= not applicable), the gene accession number (gene_ac) of the closest gene, the location of the insertion in relation to the closest gene, the distance of the insertion from the closest gene (kb=kilobase, CDS=coding sequence), the direction of the MSCVLTR in T2/onc in relation to the direction of transcription of the closest gene (dir=direction, inv=inverse) and the number of times the PCR product for the particular insertion was sequenced on the 454 platform (freq=frequency).

Supplementary Table 3: CISs identified in weaning tail clips from T2/onc low-copy;*Rosa*26-SB11 mice. The name of the CIS is presented along with the chromosome (chr= chromosome, unordered=unordered contig), the base pair of the first insertion defining the CIS, the base pair of the last insertion defining the CIS and the number of independent insertions in independent TA dinucleotides. CISs defined by an expected fraction (Efr) of .005 are marked with a #.

Supplementary Table 4: The chromosomal locations of T2/onc insertions from leukemias/lymphomas. T2/onc insertions were cloned from leukemias (sp=spleen, th=thymus, ln=lymph node, ma=mass) of *Rosa26*-SB11;T2/onc low-copy mice. The chromosome (chr) and basepair of each insertion is presented, followed by the mouse symbol of the closest gene (within 100kb, N/A= not applicable), the gene accession number (gene_ac) of the closest gene, the location of the insertion in relation to the closest gene, the distance of the insertion from the closest gene (kb=kilobase, CDS=coding sequence), the direction of the MSCVLTR in T2/onc in relation to the direction of transcription of the closest gene (dir=direction, inv=inverse) and the number of times the PCR product for the particular insertion was sequenced in 96 well plate format (freq=frequency).

Figure 7.1: IHC reveals transposase expression in transgenic mice. Transposase is poorly expressed in somatic tissues in CAGGS-SB10 mice (liver shown), while a sarcoma from a *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy mouse contains many transposase expressing cells. Most cells in *Rosa26*-SB11 mice stain positive for transposase (liver shown). Brown indicates antibody staining, nuclei are counter-stained blue. A liver from a non-transposase transgenic mouse demonstrates antibody specificity. Scale bar=50 microns.

Fig 7.1

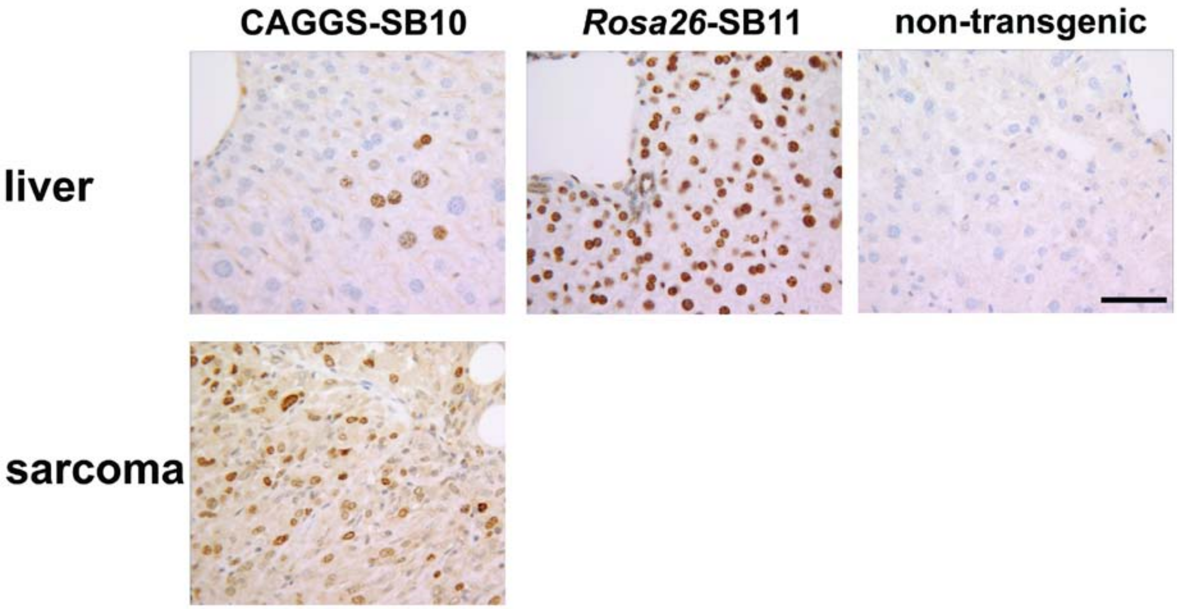


Figure 7.2: *Rosa26*-SB11; T2/onc low-copy mice are tumor prone. A) Kaplan-Meier survival curve for *Rosa26*-SB11;T2/onc low-copy (SB;T2, triangles), *Rosa26*-SB11 (SB, squares), and T2/onc low-copy (T2, circles) mice. *Rosa26*-SB11;T2/onc low-copy mice become moribund more rapidly than controls (p<.001, Logrank test). B) Hematoxylin and Eosin (H&E) stained example of a lymphocytic leukemia/lymphoma from a *Rosa26*-SB11;T2/onc low-copy mouse.

Fig 7.2

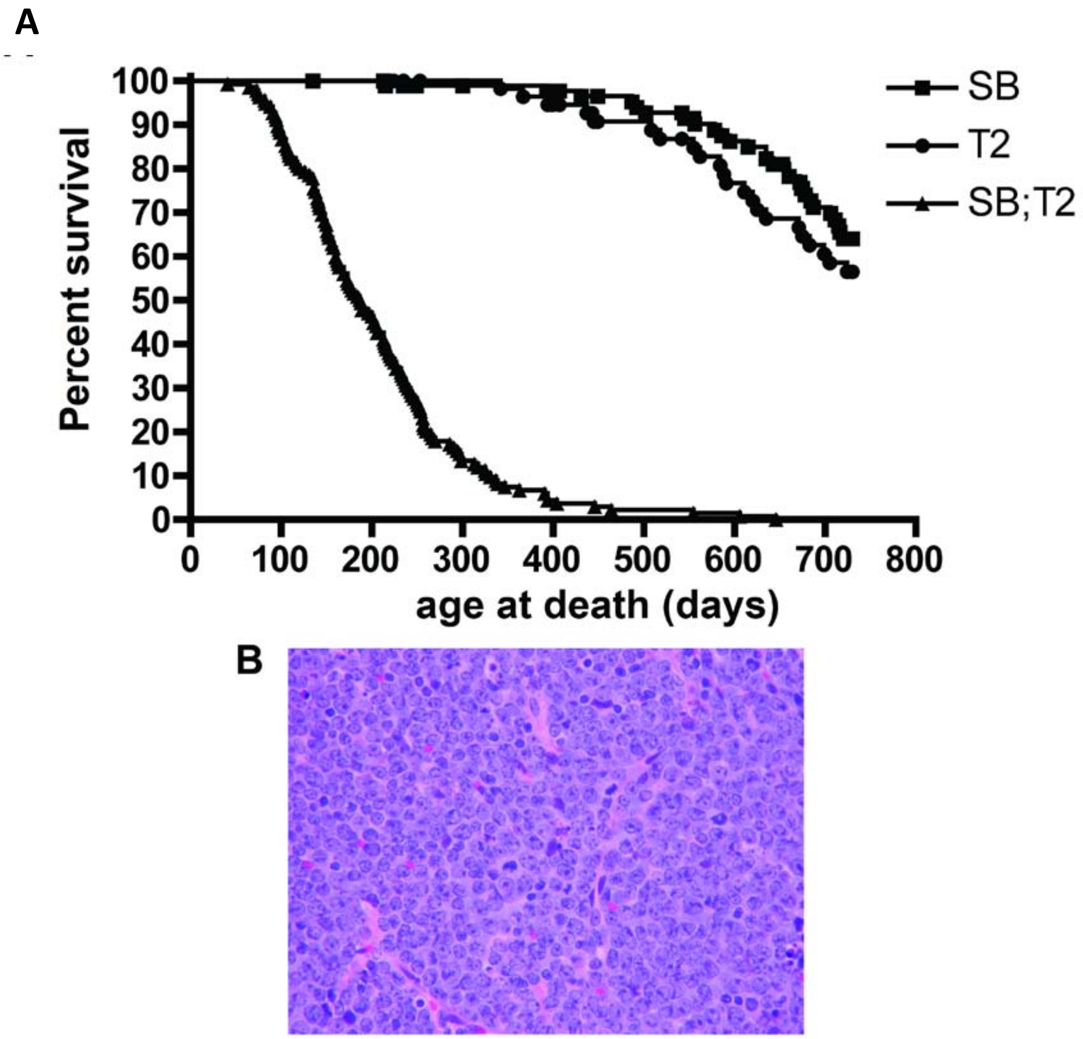

A
Percent survival
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
700
800
age at death (days)
SB
T2
SB;T2
B

Figure 7.3: Gliomas are present in *Rosa26*-SB11; T2/onc low-copy mice. A) Gliomas sometimes involved essentially an entire hemisphere (H&E). B) Neoplastic cells were characterized by round to oval nuclei with indistinct nucleoli (H&E, x400). C) Pseudopallisading necrosis was evident in a subset of cases which were therefore classified as glioblastoma (H&E, x400). D) GFAP immunostain highlighted numerous reactive astrocytes in areas of tumor cell infiltration (x200), but also occasionally labeled small round to oval cells lacking conspicuous cell processes consistent with tumor cells (inset, arrows)(x600).

Fig 7.3

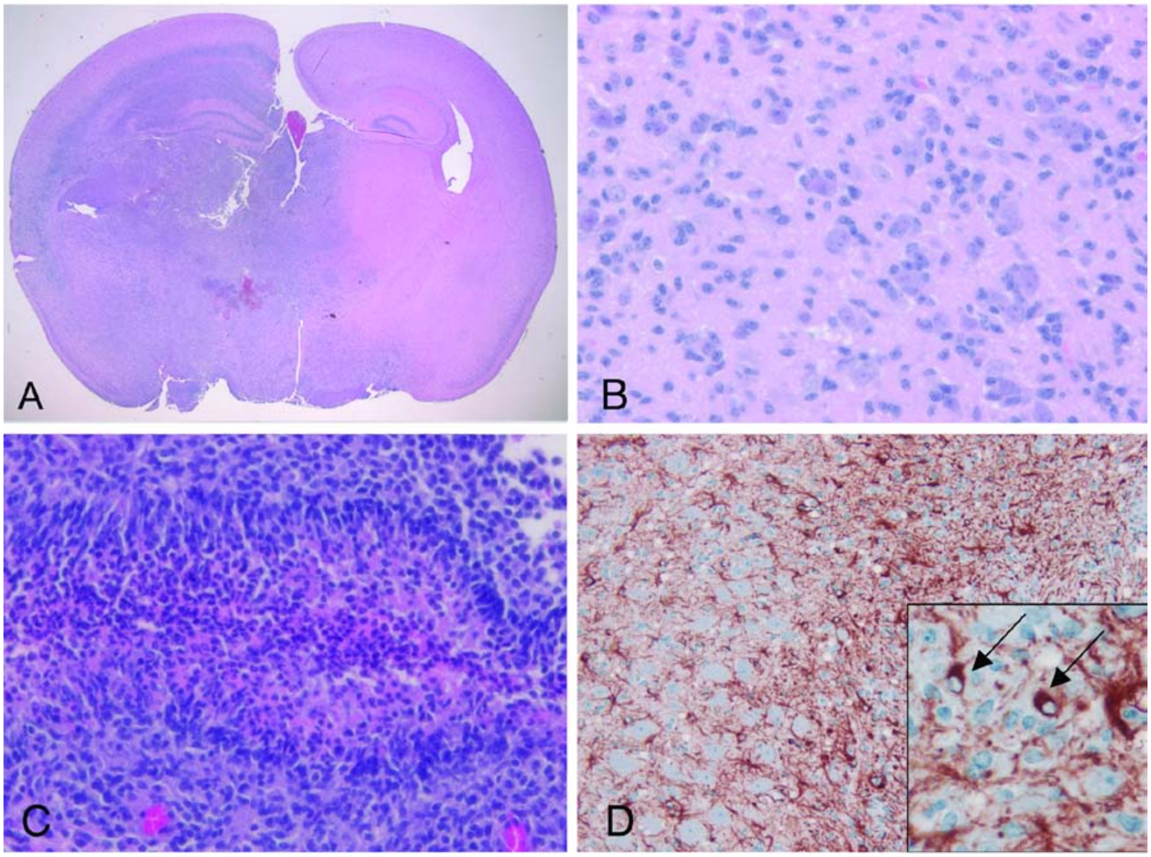


Figure 7.4: *Rosa26*-SB11;T2/onc low-copy leukemias do not show genome-wide chromosomal amplifications and deletions. Array CGH profiles from six leukemias/lymphomas from *Rosa26*-SB11;T2/onc low-copy mice (76Rosa and 68Rosa) and three sarcomas from *p19Arf*$^{-/-}$;CAGGS-SB10;T2/onc low-copy mice (p1968Caggs). Each row represents one tumor. A) Genome-wide $\log_2$ ratios. Dotted lines represent 3 standard deviations from the central mean of all clones genome wide, indicating cutoffs for gains and losses, respectively. Tumors from 76Rosa167, 76Rosa517, 68Rosa467, p1968Caggs8, p1968Caggs24 and p1968Caggs98 show localized rearrangements in the region surrounding the transposon concatomer on chromosome 1 and 15 (shaded areas). The apparent gain of the X chromosome in the tumor from 68Rosa467 is due to DNA from normal male spleen being used as reference DNA for a tumor from a female littermate mouse. B) Profiles of clones on the donor concatomer chromosome for the tumors profiled in (A).

Fig 7.4

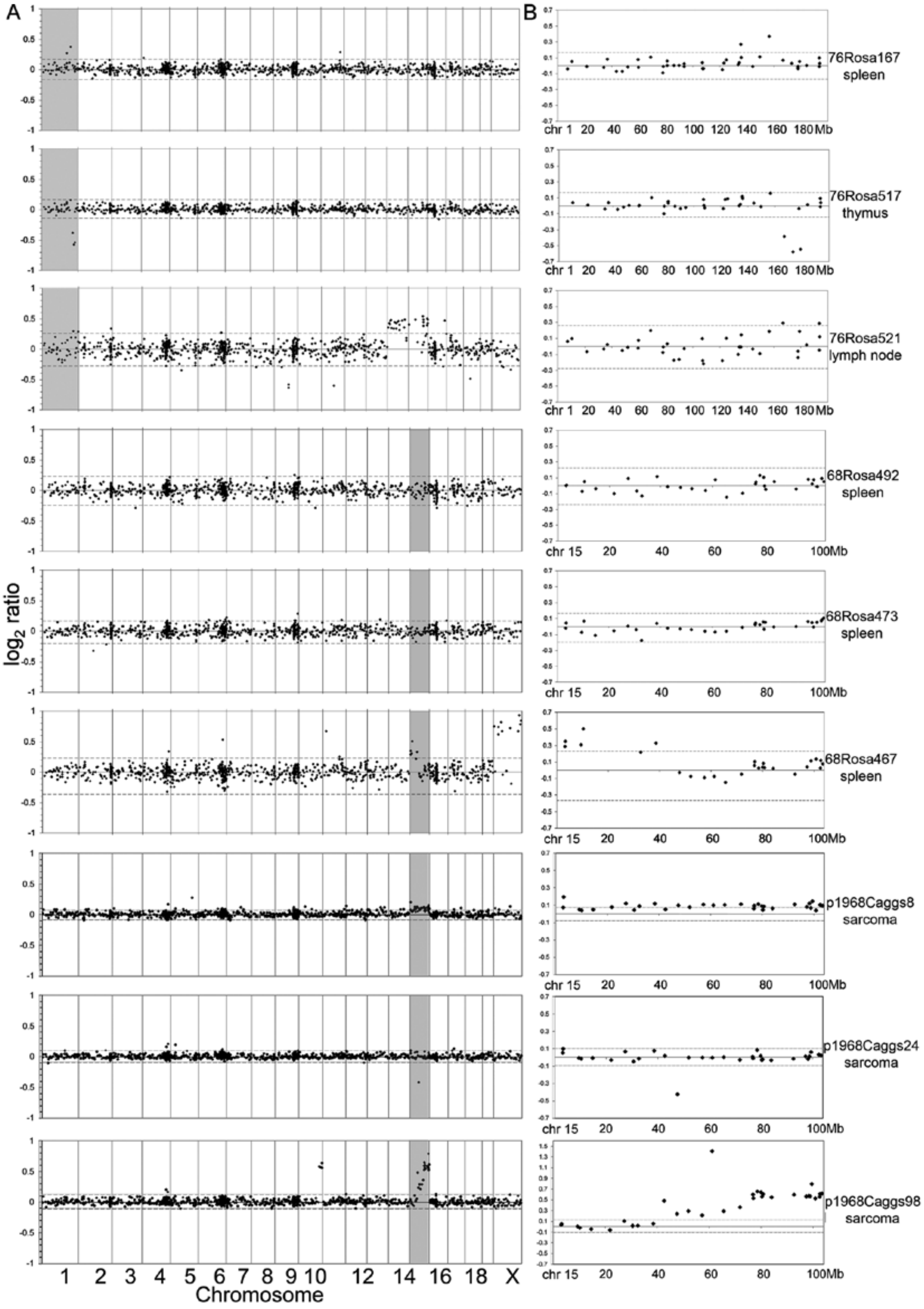

A
B
log2 ratio
1
0.5
0
-0.5
-1
Chromosome
1
2
3
4
5
6
7
8
9
10
12
14
16
18
X
76Rosa167
spleen
76Rosa517
thymus
76Rosa521
lymph node
68Rosa492
spleen
68Rosa473
spleen
68Rosa467
spleen
p1968Caggs8
sarcoma
p1968Caggs24
sarcoma
p1968Caggs98
sarcoma
chr 1 20 40 60 80 100 120 140 160 180 Mb
chr 15 20 40 60 80 100Mb
0.7
0.5
0.3
0.1
-0.1
-0.3
-0.5
-0.7
1.5
1.3
1.1
0.9

Figure 7.5: CAGGS-SB10;T2/onc high-copy mice have no increase in morbidity/mortality compared to control mice. A Kaplan-Meier survival curve for control CAGGS-SB10+ (triangles; T2-SB+) and experimental CAGGS-SB10+;T2/onc high-copy+ (squares; T2+SB+). No mice presented with outward signs of tumor formation by 18 months, however a limited number of both control and experimental mice were found dead of unknown causes during the observation period. No statistically significant difference in survival was observed.

Fig 7.5

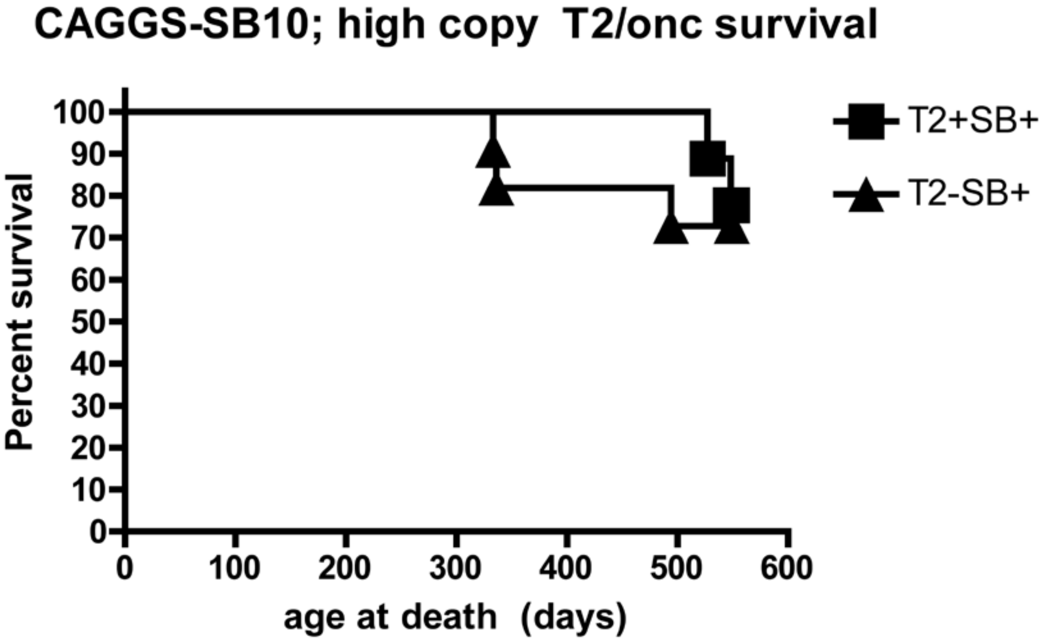

CAGGS-SB10; high copy T2/onc survival
Percent survival
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
age at death (days)
T2+SB+
T2-SB+

Figure 7.6: IHC reveals transposase expression in adult CAGGS-SB10 and *Rosa26*-SB11 mice.

Tissue types are arranged in rows, while genotypes and antibody combinations are in columns. anti-SB= incubated with anti-transposase antibody. Sections stained with secondary antibody only serve to control for staining that results from reaction of the anti-mouse IG secondary antibody used with mouse tissue. Testes are at 200x magnification, all other tissues are at 400x magnification. For testes scale bar=100 microns, all other tissues scale bar=50 microns. Brown indicates antibody staining, nuclei are counter-stained blue. Most cells in *Rosa26*-SB11 mice stain positive for transposase, while transposase expressing cells outside of the germline are rare in CAGGS-SB10 mice. Arrow points to an example of a rare transposase positive cell in the kidney of a CAGGS-SB10 mouse. Some variegated transposase expression is also detected in CAGGS-SB10 livers (same panel as in Figure 1). However, transposase expression is detected in a sarcoma from a p19*Arf*$^{-/-}$;CAGGS-SB10;T2/onc mouse (same panel as in Figure 1).

Fig 7.6

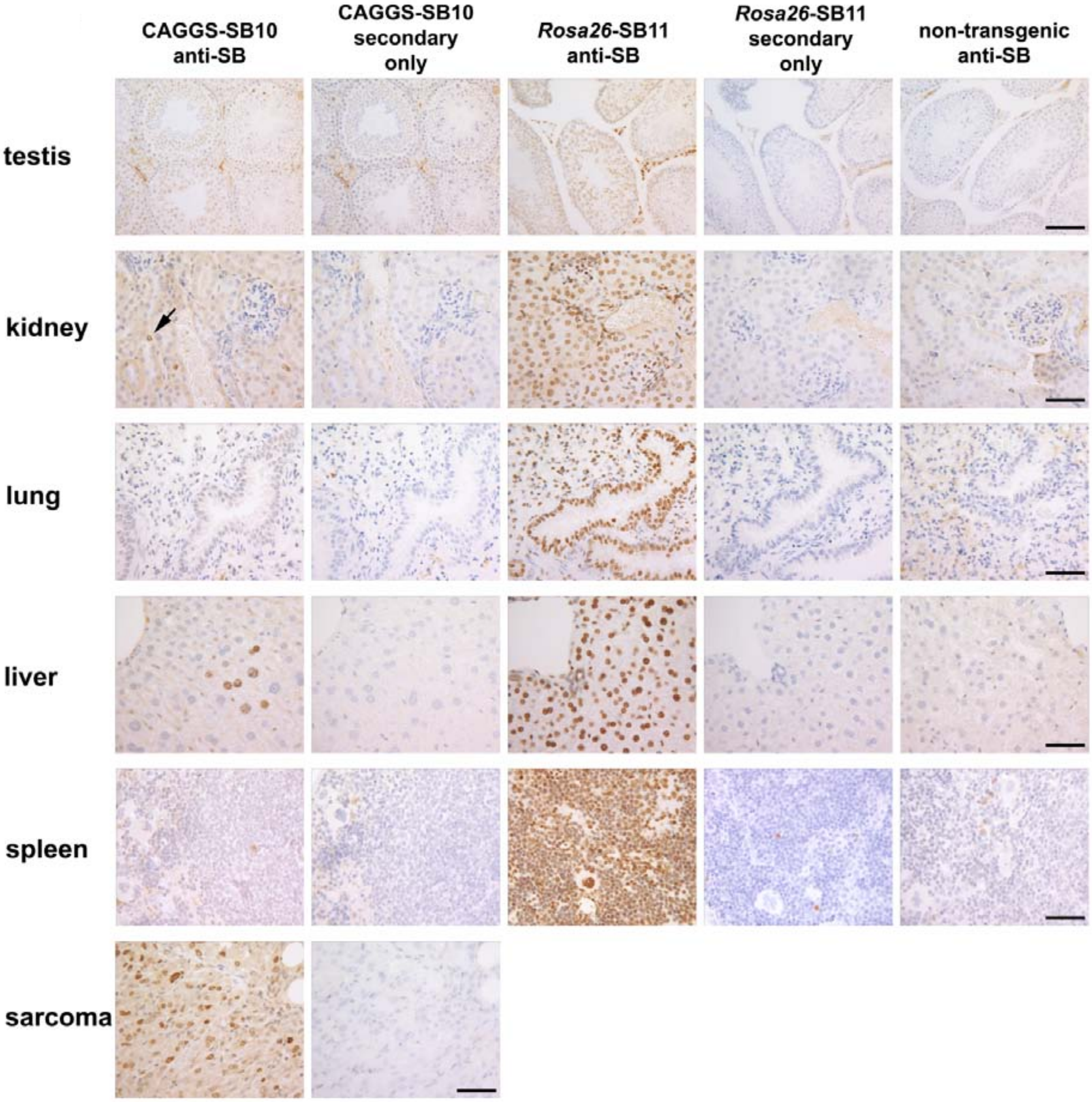

CAGGS-SB10 anti-SB
CAGGS-SB10 secondary only
Rosa26-SB11 anti-SB
Rosa26-SB11 secondary only
non-transgenic anti-SB
testis
kidney
lung
liver
spleen
sarcoma

Figure 7.7: *Rosa26*-SB11;T2/onc low-copy line 68 mice develop tumors more rapidly than *Rosa26*-SB11;T2/onc low-copy line 76 mice

A Kaplan-Meier curve demonstrating that *Rosa26*-SB11;T2/onc low-copy line 68 mice (circles; 68) become moribund more rapidly than *Rosa26*-SB11;T2/onc low-copy line 76 mice (triangles; 76) (p<.001, Logrank test). The average age of morbidity was 146 and 234 days, respectively.

Fig 7.7

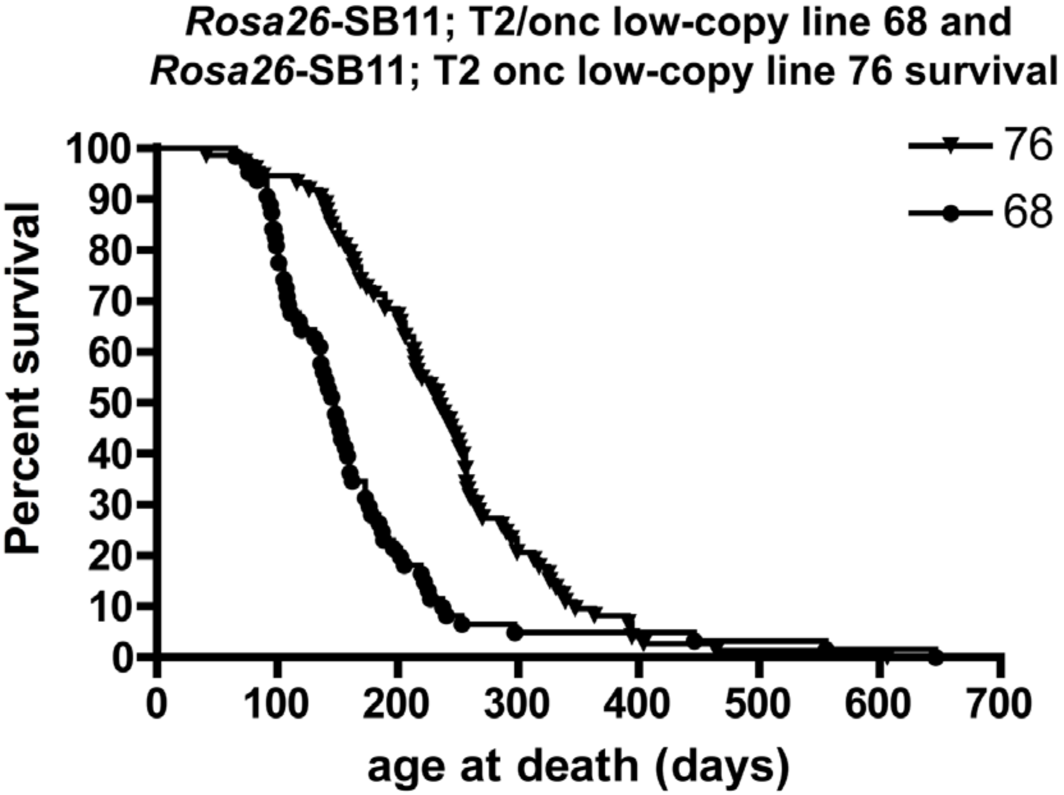
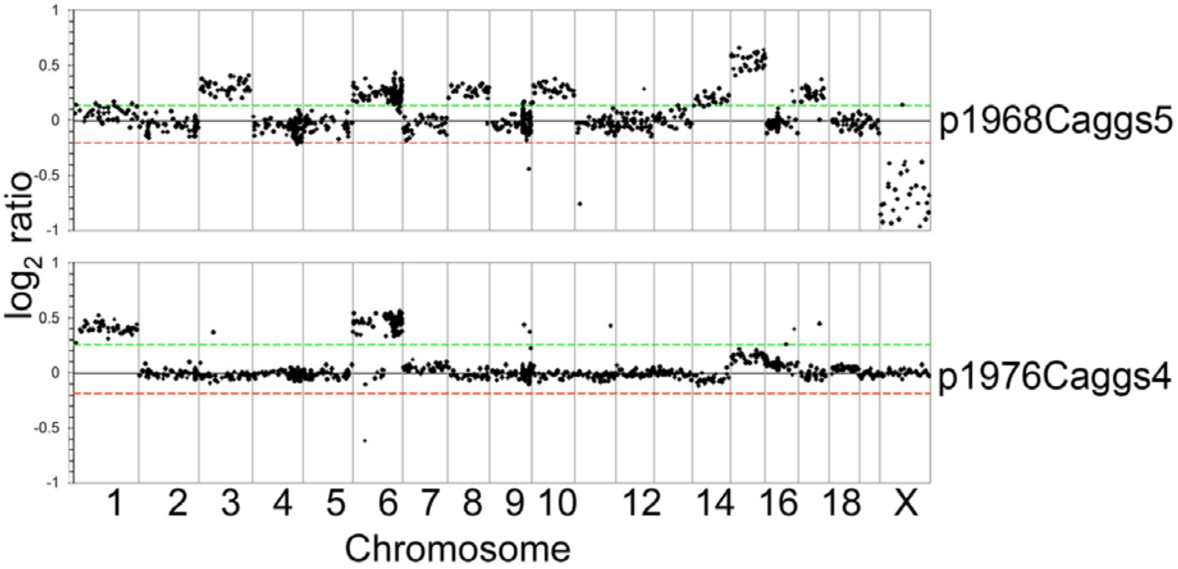

Rosa26-SB11; T2/onc low-copy line 68 and
Rosa26-SB11; T2 onc low-copy line 76 survival
Percent survival
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
700
age at death (days)
76
68

Figure 7.8: The BAC array CGH platform can detect copy number changes in tumors. Array CGH profiles for two sarcomas generated on the *p19Arf*$^{-/-}$ genetic background. Both tumors are from mice that harbor transposons but not transposase, therefore the tumors are spontaneously arising and do not contain mobilizing transposons. Each row represents one tumor and genome-wide $\log_2$ ratios are plotted. Green and red dotted lines represent 3 standard deviations from the central mean of all clones genome wide, indicating cutoffs for gains and losses, respectively.

Fig 7.8



1
0.5
0
-0.5
-1
log2 ratio
p1968Caggs5
p1976Caggs4
1
2
3
4
5
6
7
8
9
10
12
14
16
18
X
Chromosome

Chapter 8: Development of a Sleeping Beauty insertional mutagenesis system for neuroblastoma

Source: The following chapter contains unpublished data.

Contributions: I performed all of the experiments presented. WC Gustafson, Kim Nguyen, and Slava Yakovenko assisted in maintaining mouse populations. Joanna Phillips provided histological classification of tumor samples. The UCSF neuropathology core assisted with the histology and IHC. Nigel Killeen, Adam Dupuy, and David Largaespada provided expert advice. William Weiss supervised the project.

Development of a Sleeping Beauty insertional mutagenesis system for neuroblastoma

Christopher S. Hackett, William C. Gustafson, Joanna Phillips, Kim Nguyen, Slava Yakovenko, Nigel Killeen, Adam Dupuy, David A. Largaespada, and William A. Weiss.

Abstract

The molecular pathways driving neuroblastoma remain poorly characterized. As such, knowledge of the disease and overall clinical outcomes could benefit from novel genetic approaches to identify genes and pathways involved in tumorigenesis. Here we have applied a new forward genetic screening technology, Sleeping Beauty (SB) insertional mutagenesis, to a mouse model for neuroblastoma driven by a TH-MCYN transgene. We show that existing constructs used to drive SB transposase expression do not express in the peripheral sympathetic nervous system (the tissue of origin for neuroblastoma). We have developed a novel transgenic mouse, TH-SB11, that shows robust SB expression in the adrenal medulla. When crossed to the TH-MYCN transgene and a low-copy mutagenic transposon (T2/Onc), this system was capable of producing tumors on an otherwise fully resistant genetic background. This system provides a novel tool to discover new genes and therapeutic targets in neuroblastoma.

Introduction

As discussed extensively in the introduction to Part I, neuroblastoma is a disease for which the molecular abnormalities underlying growth are poorly understood. Genetics provides a powerful tool to discover the key molecules involved in cancer initiation and progression, and to lay the foundation for subsequent biochemical analysis and therapeutic development. We have utilized a mouse model for high-risk neuroblastoma driven by MYCN expression targeted to the sympathetic nervous system[1] to identify both genomic aberrations in tumors[2] and candidate genes underlying tumor susceptibility (Part I). We next sought to conduct an unbiased forward genetic screen to identify novel genomic lesions contributing to tumor development.

Since neuroblastoma is an embryonal tumor that originates from migrating peripheral neural crest tissues that goes on to differentiate into the diffuse, mostly post-mitotic peripheral sympathetic nervous system, targeting the tissue of origin using an integrating retroviral mutagenesis vector is difficult. Because of this, the Sleeping Beauty transposon-based insertional mutagenesis system, in which all components are delivered as transgenes in the germline, is an ideal system for forward somatic mutagenesis in neuroblastoma. We first attempted to utilize the existing mouse constructs for Sleeping Beauty insertional mutagenesis in neuroblastoma. When these tools proved inadequate, we investigated transposase expression in neural-crest derived tissue and showed that the presumably ubiquitously active *Rosa26* locus did not drive transposase expression in the tissue of origin for neuroblastoma. We then constructed a novel transgenic mouse, TH-SB11, to drive transposase expression in the peripheral sympathetic neural-crest derived tissue. We show robust transposase expression in the adrenal medullas from two founder lines, and observe tumors in an otherwise resistant genetic background in TH-SB11 lines showing low to moderate levels of transposase expression.

Results:

Existing Sleeping Beauty transposase expressing mice do not initiate or accelerate neuroblastoma

To test whether mobilized Sleeping Beauty transposase could overcome the strain-specific resistance of the FVB/N genetic background in the TH-MYCN model, we crossed the TH-MYCN mice (fully resistant to neuroblastomas) to T2/Onc2 transgenics, and doubly-transgenic pups were crossed to CAGGS-SB10 mice (all constructs were on an FVB/N background). In our initial cohort, all 4 triply-transgenic mice succumbed to aggressive neuroblastomas. However, the tumors did not demonstrate transposase expression by either IHC or Western Blot (data not shown). Mapping of insertion sites did not reveal any common insertion sites in tumors from the 4 animals. A subsequent expanded cohort of over 20 triply-transgenic mice, with an equivalent control cohort, did not produce any tumors. We concluded that the tumors observed in the original litter of mice were not accelerated by mobilized transposons, but were due either to subtle strain background differences in mice acquired from outside collaborators, or from some epigenetic artifact of the rederivation process required to import the mice; in both cases, the propagation of the mice in our colony would have erased these effects.

We next crossed mice carrying the TH-MYCN transgene on an FVBN background with mice carrying a Cre transgene knocked into the endogenous tyrosine-hydroxylase locus (TH-Cre[3]). We crossed mice carrying both TH-MYCN and TH-Cre to mice doubly homozygous for the conditional Rosa-lsl-SB11 construct[4] and the T2/Onc2 transposon concatemer[5]. No tumors arose in mice carrying only TH-Cre and the transposase/transposon constructs (in a cohort of 20 mice), suggesting that the system was not sufficient to drive tumorigenesis in the absence of a predisposing oncogene. While 8 tumors were detected in the cohort of mice carrying TH-Cre, TH-MYCN, and the transposase/transposon constructs, tumor onset was not accelerated, and the number of tumors was not significantly higher than what would be expected on a mixed, mostly C57/B6 background.

The Rosa-SB11 mouse does not show Sleeping Beauty Transposase expression in peripheral neural crest derived tissue

After failing to accelerate tumorigenesis using the CAGGS-SB10 transposase, it was established that the construct was essentially restricted to muscle tissue in adult mice (Lara Collier, personal communication), providing an explanation for the lack of tumor acceleration and transposase expression in tumors. To determine whether a similar lack of transposase expression in sympathetic peripheral nervous system tissue was responsible for the failure of the *Rosa26* –based transposase constructs, we next analyzed transposase expression in these mice. The *Rosa26* locus was identified by gene trapping as a position on mouse chromosome 6 that showed near-ubiquitous expression of gene-trap β-galactosidase expression in all tissues analyzed[6]. The *Rosa26* locus has become a popular target for genomic knock-in experiments using ES cells due to both the chromatin accessibility (making ES cell targeting highly efficient) and the presumed ubiquitous expression of inserted genes from the endogenous promoter. Sleeping Beauty transposase was knocked into the *Rosa26* locus for both constitutive[5] and conditional[4] transposase expression. The constitutive expressing construct shows transposase expression in all major tissues analyzed; however, this analysis did not include the adrenal gland or any peripheral nerve ganglia. To assess whether the transposase was expressed in peripheral neural crest derived sympathetic nervous system tissue, we isolated adrenal glands from RosaSB11 mice and doubly-transgenic Rosa-lsl-SB11 and TH-Cre mice and assayed transposase expression using IHC. The adrenal gland consists of two layers: the outer adrenal cortex, and the inner adrenal medulla (**Figure 8.1A, B**). The adrenal medulla is neural-crest derived sympathetic nervous tissue that, unlike the adrenal cortex, is a tissue of origin for neuroblastoma, as well as pheochromocytoma. While nuclear SB transposase expression was detected in the adrenal cortex of the RosaSB11 mice, no expression was detected in the adrenal medulla (**Figure 8.1C, E)**. Consistent with this, no expression was detected in either layer of the adrenal glands in mice carrying the conditional SB11 allele and the TH-Cre knock-in (**Figure 8.1D, 8.1F**). We conclude that SB transposase is not expressed in the adrenal medulla when driven by the endogenous *Rosa26* promoter.

TH-SB11 transgenic mice display strong transposase expression in the adrenal medulla

To drive high-level transposase expression in the peripheral sympathetic nervous system, we made transgenic mice in which SB11 transposase expression was driven by the same tyrosine hydroxylase fragment used to drive MYCN expression in the TH-MYCN model system[1] (**Figure 8.2A)**. Transgenic mice were made on an FVB/N background. We identified 11 transgenic founders by PCR. These lines were expanded, and adrenal glands were isolated from second- and third-generation progeny and screened using IHC for SB transposase. As shown in **Figure 8.2B-I**, expression levels were highly variable across the different founder lines, with lines A and L (not shown) showing the highest expression levels, lines D and E showing intermediate levels, and the other lines showing low or no expression of the transposase. In strains showing moderate to strong expression, expression was also detected in spleen, but not in any other major organs.

TH-SB11 mice develop primitive tumors after a long latency

Though transgenic line TH-SB11-E showed only moderate SB expression in the adrenal medulla, and line TH-SB11-G showed low levels of transposase expression, we aged mice carrying the transposase construct, the low-copy T2/Onc (low copy) transposon concatemer, and the TH-MYCN transgene on a neuroblastoma-resistant FVB/N genetic background. Two out of three triply-transgenic mice carrying the TH-SB11-E construct developed large palpable masses at 5 months and 8.5 months of age (the third mouse in the cohort was found dead with no detectable tumor at 6.5 months of age). No tumors were detected in control mice carrying only two of the three constructs (N=10, aged to 12 months). Tumors were attached to spleen, and pathological analysis showed them to be primitive small round blue cell tumors with rare, scattered mitotic figures, with large areas of necrosis and extramedullary hematopoiesis (**Figure 8.3A, B**). One mouse carrying TH-SB11-G developed large abdominal masses at 12 months of age, but the pathology of this mass appeared non-malignant (not shown).

Discussion

The ability to conduct forward genetic screens to identify genes driving neuroblastoma would be of great benefit to the field, as the molecular characterization of this tumor is relatively underdeveloped. However, the diffuse nature of the sympathetic peripheral nervous system (the tissue of origin for neuroblastoma), and the lack of retroviruses with specific tropism for sympathetic nervous tissue, complicate this strategy. The Sleeping Beauty system provides a means to perform these screens using transposon elements propagated through the germline that circumvent the delivery issue, but introduce the requirement for strong, tissue-specific transposase expression.

We have established that the existing SB11 transposase mouse lines do not drive detectable transposase expression in the adrenal medulla, a large peripheral sympathetic tissue that is a frequent origin of neuroblastomas. We then generated several lines of transgenic mice expressing SB11 under the control of the tyrosine hydroxylase promoter to drive expression in the sympathetic nervous system. Several of these mice displayed strong specific SB expression in the adrenal medulla and in spleen, but not in any other major organs. When mice were mated to TH-MYCN and T2/Onc on a fully-resistant FVB/N background (see Chapter 2), two mice developed small round blue cell tumors next to the spleen, indicating that the system was promoting tumorigenesis, and providing proof-of-principle that the system is capable of generating neuroblastoma-like tumors. An expanded cohort of mice will validate the system as a means to drive tumors in this model, as well as to generate a pool of tumors from which to identify common transposon insertion sites and candidate oncogenes and tumor suppressors. Additionally, cohorts of mice carrying TH-SB11 and a high-copy T2/Onc2, with and without the TH-MYCN predisposing oncogene, are also being generated. While these mice will not be on a fully resistant background, mobilized transposons in mice carrying the TH-MYCN construct may accelerate tumorigenesis and help to identify genes contributing to tumor development. Similarly, tumors arising in mice not carrying TH-MYCN may provide a means to identify novel pathways in the disease.

Methods

Mice: T2/Onc mice were obtained from David Largaespada (University of Minnesota). T2/Onc2, Rosa-SB11, and Rosa-lsl-SB11 mice were obtained from Adam Dupuy, Neal Copeland, and Nancy Jenkins (NCI, Frederick, MD). All mice were maintained according to the standards of the UCSF Animal Care and Use Committee.

Generation of TH-SB11 mice: The EcoRI fragment containing the tyrosine hydroxylase promoter and rabbit beta-globin intron were excised from the TH-MYCN targeting construct[1] and inserted into the EcoRI site of pGEM-7Zf+ MCS (Promega, Madison, WI). SB11 transposase was cut from pCMV-SB11 (a gift from P. Hackett, University of Minnesota) using EagI and SalI, blunted, and ligated into the SmaI site of the pGEM-7Zf+ MCS. The construct was separated from the vector using NsiI, which cut once inside the TH promoter (generating the same TH segment used to make TH-MYCN) and once in the pGEM-7Zf+ MCS downstream of the EcoRI and SmaI sites. DNA was injected into fertilized FVBN eggs by the UCSF transgenic core, yielding 13 transgenic founders as identified by PCR for SB11 transposase. Founders were mated to FVBN mates, and adrenal glands were isolated from transgenic progeny and screened by IHC for SB11 transposase (R&D Systems, Minneapolis, MN) using established protocols[7].

References

1. Weiss, W.A., Aldape, K., Mohapatra, G., Feuerstein, B.G. & Bishop, J.M. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *EMBO J* **16**, 2985-95 (1997).
2. Hackett, C.S. et al. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-73 (2003).
3. Lindeberg, J. et al. Transgenic expression of Cre recombinase from the tyrosine hydroxylase locus. *Genesis* **40**, 67-73 (2004).
4. Keng, V.W. et al. A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-74 (2009).
5. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6 (2005).
6. Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev* **5**, 1513-23 (1991).
7. Collier, L.S. et al. Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality. *Cancer Res* **69**, 8429-37 (2009).

Figure 8.1: Rosa-SB11 mice do not express transposase in the adrenal medulla. A. H&E stain of a mammary gland showing distinct layers: an outer adrenal cortex, and inner, less eosinophilic, sympathetic adrenal medulla. **B.** SB11 IHC on a wild-type B6 adrenal gland, showing no nuclear staining for SB11 transposase. **C.** SB11 IHC on a Rosa-SB11 adrenal gland, showing staining in the adrenal cortex, but not adrenal medulla. **D.** SB11 IHC on a Rosa-lsl-SB11 mouse crossed with a TH-Cre construct, showing no staining in either layer. **E & F:** Higher magnifications of the samples in **C** and **D**, respectively.

Fig 8.1

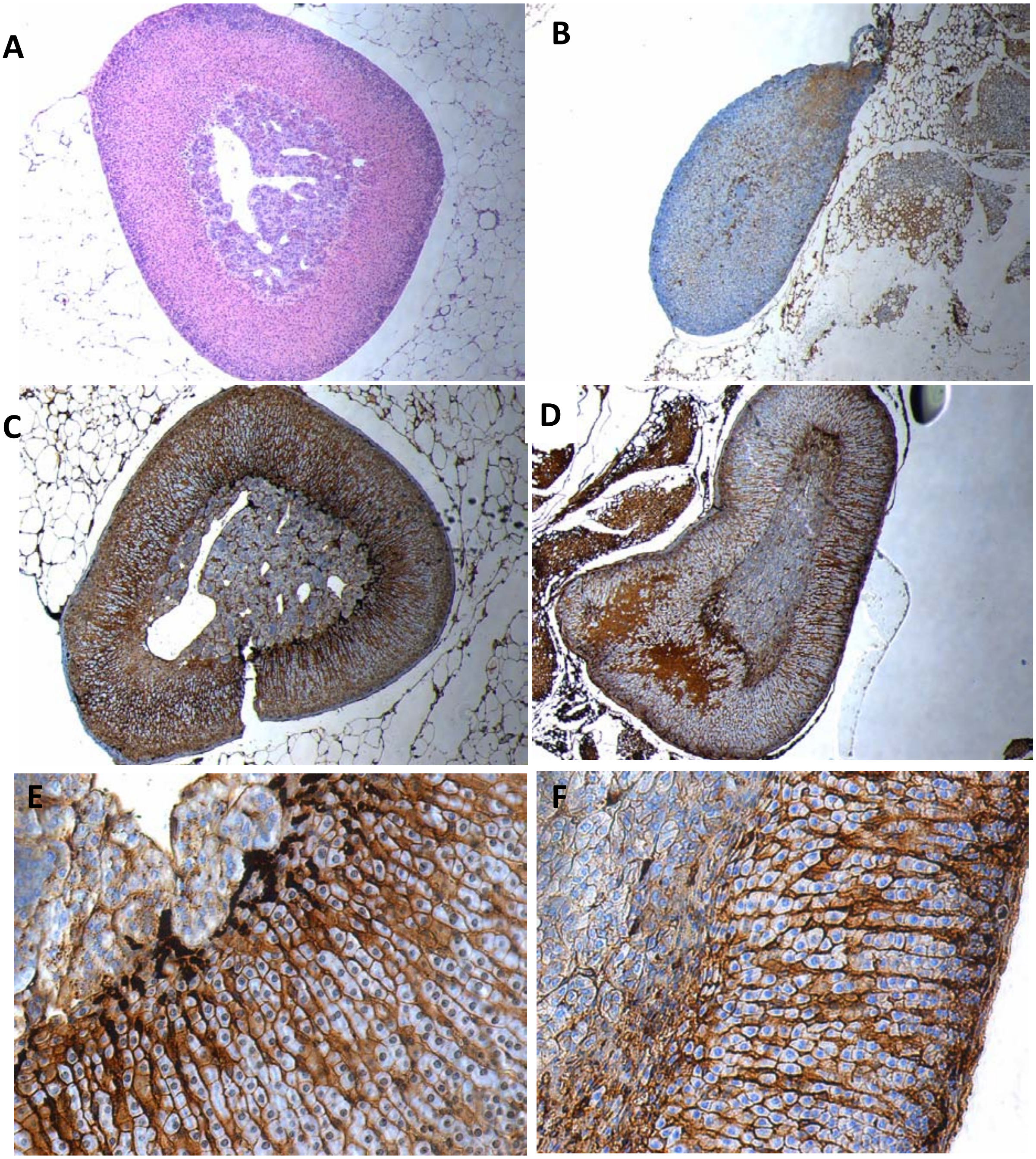


Figure 8.2: IHC showing SB11 expression in TH-SB11 adrenal glands. A. Structure of the TH-SB11 construct. B-G: individual TH-SB11 lines **B:** Line TH-SB11A **C:** Line TH-SB11C **D:** Line TH-SB11D **E:** Line TH-SB11E **F:** Line TH-SB11F. **G:** Line TH-SB11G **H.** Higher magnification of TH-SB11A. **I.** Higher magnification of TH-SB11E.

Fig 8.2

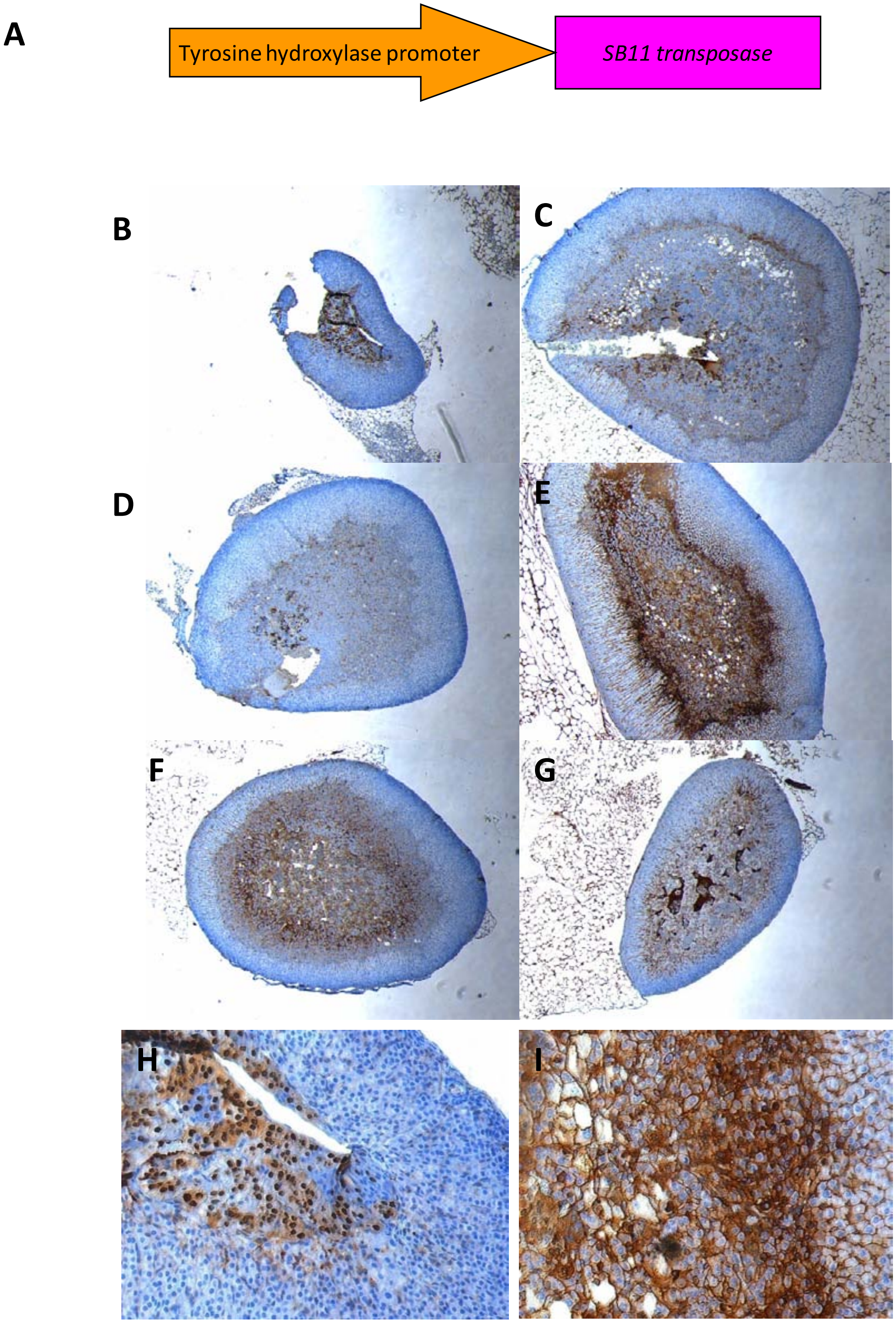
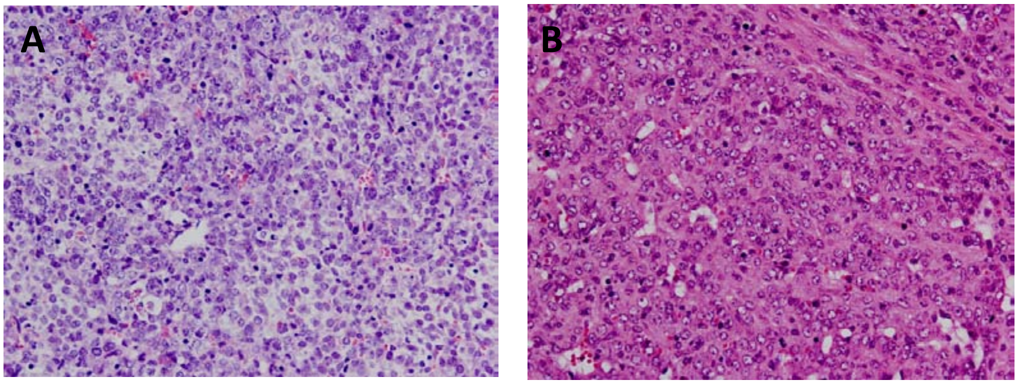

A
Tyrosine hydroxylase promoter
SB11 transposase
B
C
D
E
F
G
H
I

Figure 8.3: H&E stains of TH-SB11 tumors. A. Tumor isolated from a mouse at 8.5 months of age. **B.** Tumor isolated from a mouse at 5 months of age.

Fig 8.3



Chapter 9: Sleeping Beauty insertional mutagenesis in breast cancer

Source: The following chapter contains unpublished data.

Contributions: I performed all of the experiments described. I received technical assistance and supervision from Pengfei Lu, Joanna Phillips, Aditi Sharma, and Peter Dijkgraaf. Slava Yakovenko and Kim Nguyen assisted in the maintenance of mouse stocks. David A. Largaespada and Adam J. Dupuy provided expert advice and unpublished mouse stocks. Zena Werb provided a particularly significant amount of expert advice. William A. Weiss supervised the project.

Sleeping Beauty insertional mutagenesis in breast cancer

Christopher S. Hackett, Pengfei Lu, Joanna Phillips, Aditi Sharma, Peter Dijkgraaf, Slava Yakovenko, Kim Nguyen, David A. Largaespada, Adam J. Dupuy, Zena Werb, and William A. Weiss.

Abstract

Breast cancer is a disease that afflicts hundreds of thousands of women in the US each year. While recently-developed targeted therapeutics have had a significant impact on survival in the disease, the number and utility of these agents remains limited. Here we have attempted to utilize a new forward genetic screening technology, Sleeping Beauty (SB) insertional mutagenesis, to discover novel molecular pathways driving breast tumors in mice. We first show that existing SB transposase constructs are insufficient to drive mammary carcinogenesis in mice, either failing to drive tumors, or causing mortality at a young age, presumably due to leukemia. We next provide proof-of-concept that a mammary epithelial transplant system, combined with lentiviral transduction of SB transposase, may overcome the limitations of the transgenic/knock-in transposase constructs. If true, this system may uncover novel molecular pathways driving breast cancer that were not revealed by previous retroviral insertional mutagenesis screens.

Introduction

Breast cancer is one of the most common malignancies, affecting over 200,000 women in the United States per year (www.cancer.org). Genomic and pathological characterization of the disease has broken the disease into molecularly-distinct subgroups, which differ in the degree to which the genetic lesions underlying tumor development are understood. Though hundreds of genes have been studied for their role in the progression of breast cancer, the translation of this knowledge to the clinic has been limited; currently, molecular diagnosis and targeted therapeutics are focused on three genes: the HER2 growth factor receptor, and the estrogen and progesterone steroid receptors. Targeted therapeutics against these agents (depending on molecular screening at diagnosis), in combination with surgery, general chemotherapy, and radiation remain the standard of care for the disease. Survival in some subgroups has improved dramatically with targeted agents such as Herceptin (a HER2 inhibitor), and aromatase inhibitors targeting estrogen synthesis. However, identification of other genes driving tumorigenesis may present more targets for therapeutic intervention for subgroups currently without an effective targeted therapeutic (for example, the "triple negative" class of tumors lacking HER2 and the steroid receptors), as well as for patients whose tumors present biomarkers for targeted agents but fail to respond to therapy or who relapse after remission.

Although they arise from completely different tissue types, breast cancer shares some molecular hallmarks with neuroblastoma. Namely, 60% of breast tumors show copy number loss or loss of heterozygosity of the chromosome 1p arm[1], while roughly one-third of neuroblastomas harbor this deletion[2], suggesting a potential tumor suppressor in the region common to both diseases. While this lesion is present in a wide variety of solid tumors, it is particularly common in breast cancer, and has a particular association with poor outcome in neuroblastoma. Additionally, somewhere between one quarter to one half of all breast tumors show amplification or overexpression of the c-myc oncogene (with some estimates even higher, depending on the assay), while one-third of neuroblastomas show amplification of the related MYCN oncogene[3]. Mouse models for both of these lesions exist; mice expressing MYCN under the control of the tyrosine hydroxylase promoter (targeting expression to the peripheral sympathetic neural crest) develop neuroblastomas[4]. Similarly, mice expressing *c-myc* in mammary epithelium develop mammary carcinomas, the first mouse model for human cancer utilizing a human oncogene[5]. In both model systems, the tumors are focal and clonal, and show recurrent secondary genetic lesions suggesting the need for additional mutations to drive tumorigenesis. Since the same secondary lesions (for example, 1p deletion in human tumors and the corresponding loss of chromosome 4 in mouse tumors seen in both models[6,7]) may cooperate with both *c-myc* in breast cancer and MYCN in neuroblastoma, parallel identification of these lesions may be of benefit for both treatment of diseases.

As discussed previously, forward genetic insertional mutagenesis screens are a powerful tool to discover novel mutations driving cancer. The mammary gland has been amenable to insertional mutagenesis screens in mice because of the existence of the mouse mammary tumor virus (MMTV), which shows specific tropism for mammary tissue. MMTV has been used to implicate several genes in breast cancer, most notably members of the *wnt* and *FGF* families (reviewed in[8]). However, as discussed previously, integrating vectors are subject to biases towards specific integration sites, and viruses in particular can show a bias towards insertion near specific genes, even in the absence of oncogenic selection[9]. The seminal studies demonstrating these biases did not include a analysis of MMTV, however, a subsequent high-throughput insertional mutagenesis screen using MMTV produced a surprisingly short list of common insertion sites[10]. In this screen of 160 tumors, members of the *Fgf* and *Wnt* families were hit dozens of times, no gene outside of this family harbored an insertion in more than four tumors, and genes known to be able to initiate breast cancer when overexpressed in the mammary gland (e.g. *HER2/neu*, *c-myc*, and *ras*) were not detected by this screen. These observations suggest that MMTV is not capable of revealing all genes associated with breast cancer progression, either due to limitations in the cellular subtype the virus can infect or, more likely, an insertion site bias of the viral integrase.

In this study, we sought to complement the repertoire of common insertion sites identified in breast cancer using MMTV with Sleeping-Beauty-mediated insertional mutagenesis. As described previously, the Sleeping Beauty insertional mutagenesis system works via a two part mechanism: 1) a transposon (T2/Onc, T2/Onc2, or T2/Onc3), a DNA sequence capable of activating or deactivating surrounding host genes, and 2) a transposase protein (SB10 or SB11) which excises the T2/Onc transposon and re-inserts it into a random location in the host genome. Cells harboring insertions conferring tumorigenic characteristics are selected and clonally expanded, allowing for identification of the insertion site and nearby genes.

We first attempted to utilize the existing SB transgenic mice to drive mutagenesis in the mammary gland in combination with overexpression of either the *c-myc* or *HER2/neu* oncogenes. We then observed that the activity of the MMTV LTR/promoter in the hematopoietic compartment made the mice susceptible to leukemias at a young age before they could develop breast cancer, and saw no acceleration of breast cancer in the surviving animals. We next developed a transplant system in which mammary epithelial cells (MECs) from mice carrying dormant mutagenic transposons were transduced with lentiviral transposase and transplanted into recipient mice in which the mammary fat pad had been cleared of epithelial cells. The development of this system facilitates transposon-mediated insertional mutagenesis in the mammary gland without complications from leaky transposase expression, and provides a means to identify novel genes involved in breast cancer development.

Results

CAGGS-SB10 fails to accelerate existing mouse models of breast cancer

We initially used the first-generation SB insertional mutagenesis system[11], consisting of a (presumably) constitutively expressed SB10 transposase under the control of a CAGGS (beta-actin) promoter. Transgenic mice harboring this construct and a transposon transgene did not succumb to tumors without a predisposing tumor suppressor knockout. We combined this system with mice carrying the MMTV-*HER2/neu* (unactivated/wild-type) transgene[12], susceptible to mammary tumors after a long latency. While we were able to successfully map transposon insertions, we did not observe any significant tumor acceleration compared to mice carrying MMTV-*HER2/neu* alone, nor did we see any transposase expression in tumors by Western blot or IHC (data not shown). The Largaespada group subsequently discovered that the CAGGS-SB10 construct was only expressed in muscle tissue, most likely due to a positional effect at the transgene insertion site. Consistent with this, most of the insertion sites we were able to map in tumors were also present in spleens from the corresponding mice, suggesting the transposase was active early in development but dormant during tumor development.

The conditional Rosa-lsl-SB11 construct drives hematopoietic disease but does not significantly accelerate mammary tumorigenesis

We next attempted to increase transposase activity using a Cre-activated Rosa-SB11 knock-in[13] in combination with a high-copy transposon donor locus (T2/Onc2). We attempted to activate this construct with an MMTV-Cre transgenic mouse[14]. However, we observed a high rate of lethality at an early age, presumably from leukemia (though not verified beyond the frequent observation of enlarged spleens), which would be consistent with results from other groups (A. Dupuy, personal communication). This is likely caused by the known activity of , the MMTV-Cre construct in the hematopoietic compartment[15], where the SB system is much more effective at inducing tumors. Among the mice that did not succumb to leukemia, none developed breast cancer, in spite of the MMTV-*HER2/neu* transgene they carried. We speculate this resistance is derived from the mouse strain background; unlike the first generation SB system, which was on an FVB/N background susceptible to MMTV-HER2 induced tumors, the second and third (described below) generations are on a mixed, mostly B6, genetic background that shows less susceptibility to these tumors. In parallel, we combined this system with the MMTV-*c-myc* model of mammary tumorigenesis. Mice in this model showed an exceptionally high rate of lethality within the first two months of life, often accompanied by large spleens, suggesting a predisposition to leukemia. Though two mice in our screen developed solid tumors which were unlikely to have originated in the mammary gland, the high rate of early mortality in these mice prevented us from generating a sufficient collection of mammary tumors for insertion site analysis.

Viral delivery of transposase *in vitro* eliminates hematapoietic disease

In an attempt to overcome strain issues and leukemia incidence, we next turned to an *in vitro*/transplant based strategy. Briefly, in this system, mammary epithelial cells (MECs) are isolated from donor mice, in this case, mice carrying a dormant transposon concatemer (T2/Onc, T2/Onc2, or T2/Onc3). These cells are cultured *in vitro*, and the SB transposase is then introduced with an integrating viral vector. Virally-transduced cells are then transplanted into recipient mice in which the developing endogenous mammary epithelium has been surgically ablated. The transduced/transplanted MECs then repopulate the mammary fat gland, forming a normal branched ductal system, such that the mammary epithelium is genetically modified, but the surrounding tissue is wild-type.

Lentiviruses can infect non-dividing cells and thus can infect the mammary stem cell subpopulation in a MEC culture, providing more efficient outgrowth of transduced cells. Importantly, in this system (unlike a retroviral system), transduced cells do not have to have a growth advantage to repopulate the mammary fat pad[16]. To take advantage of this aspect of lentiviral vectors to provide the greatest chance of generating SB-positive transplanted mammary epithelial cells, we cloned the SB11 transposase into two lentiviral constructs, pEIZ and pEIR[16] to make the constructs pEIZ-SB11 and pEIR-SB11 (**Figure 9.1A**). These constructs express the fluorescent markers (**Figure 9.1B**) and transposase (**Figure 9.1C**) in transduced 293T cells.

We then isolated MECs from mice carrying the high-copy T2/Onc2 transposon, as well as mice carrying the transposon on a *p53+/-* background. MECs were transduced with either control virus or SB11 virus at an MOI of 60. 300,000 cells were transplanted into cleared fat pads of 40 nude mice (20 receiving *p53+/+* cells and 20 receiving *p53+/-* cells) with the right gland receiving SB11-transduced cells and the left gland receiving vector-transduced cells. A control mouse was dissected to establish that MECs were reforming branched epithelial structures; verification of the viability of these cells in comparison with wild-type glands and cleared fat pads is shown in **Figure 9.2**. While the tissue was re-grown, we did not observe any fluorescence in the ducts, suggesting a lack of transposase expression. Consistent with this, no mice in the cohort aged for over 16 months developed tumors. We speculate that this could be due to several technical issues. First, though a very high multiplicity of infection (as determined by viral titering in 293T cells) was used to transduce the MECs prior to transplantation, the transduction efficiency may have been low. Second, expression of SB transposase in vivo may have conferred a selective disadvantage to cells repopulating the mammary fat pad. Third, cells may have been successfully transduced, but the expression levels of transposase and the fluorescent markers may have been below the level of detection, and transposase expression may have been insufficient to drive or accelerate tumorigenesis.

The next phase in this process will involve a newer lentiviral construct developed in the Largaespada group which utilizes the newest transposase, SB100X, which shows 3- to 100-fold increased activity as compared to SB10 in various assays[17]. These constructs also express both luciferase and GFP, allowing us to monitor mammary fat pad repopulation and tumorigenesis without sacrificing the recipient mice. We will test this construct more extensively in MECs cultured over several days in vitro to verify expression of the transposase and reporters prior to transplantation.

Discussion

While insertional mutagenesis has been a powerful tool to uncover genes involved in breast cancer and other malignancies, the results of the high-throughput MMTV insertional mutagenesis screen suggest viral insertional mutagenesis will not uncover a large number of genes relevant to breast cancer. We thus sought to adapt the Sleeping Beauty transposon insertional mutagenesis system to identify genes that caused breast cancer, or cooperated to accelerate tumorigenesis in combination with overexpression of the *HER2* and *c-myc* oncogenes or with hemizygosity of the tumor suppressor *p53*. We have established that a system based on *in vitro* transduction of cells carrying donor transposons with a lentiviral transposase vector can overcome the most significant technical limitation of a purely transgenic approach: the development of leukemia due to leaky transposase expression in the hematopoietic compartment from promoters active in mammary epithelia.

Should this system prove to be effective at inducing tumors, it will provide a powerful tool for further dissection of breast-cancer pathways, as cells from any transgenic or knockout donor mouse can be transduced and transplanted, and one 10-week-old donor mouse can generate enough cells to transplant into 10 recipient mice in a process that takes less than a week. The transposase system should be capable of identifying a wider array of genes than the traditional viral vectors, and the flexibility of the system may be utilized to identify networks of genes (and potential drug targets) that interact with several genes with a known role in breast cancer.

Methods

Mice: Mice harboring the Sleeping Beauty elements were acquired as described in Chapter 8. MMTV-Cre, *p53*$^{+/-}$, and *MMTV-HER2/neu* (wild type) were a kind gift from Zena Werb. MMTV-*c-myc* was acquired through the NCI repository (Frederick, MD). Mice were checked weekly for tumor formation.

Construction of Viral Constructs: SB11 was cloned between the NotI and BamHI sites of HIV-ZsGreen, and blunt-ligated into the SmaI site in HIV-H2B-mRFP[16]. Viruses were packaged with pVSVG envelope in 293T cells, and viral titers were determined by FACs in transduced 293T cells. SB transposase expression was validated in transduced 293T cells by Western blot using an antibody specific for SB transposase (R&D Systems, Minneapolis, MN)

Mammary Epithelial Cell Transplants: Mammary epithelial cells were isolated and transplanted as described[16]. Briefly, mammary glands were isolated from 8-10 week old female mice carrying T2/Onc2 on a p53$^{+/+}$ or p53$^{+/-}$ background, grown in culture for two days, and transduced with lentivirus at an MOI of 60. After 24 hours, 300,000 cells were transplanted into 21 day old recipient nu/nu females in which the developing endogenous mammary epithelium had been surgically ablated in the 4th mammary glands medial to the lymph node. SB and empty-vector transduced cells were implanted in contralateral glands. Mice were tracked for 18 months for tumor formation, with two sample mice sacked at 10 weeks to assess SB and fluorescent marker expression. Dissection and transplant control mammary glands were visualized using a Carmine stain.

References

1. Ingvarsson, S. et al. High incidence of loss of heterozygosity in breast tumors from carriers of the BRCA2 999del5 mutation. *Cancer Res* **58**, 4421-5 (1998).
2. Fong, C.T. et al. Loss of heterozygosity for the short arm of chromosome 1 in human neuroblastomas: correlation with N-myc amplification. *Proc Natl Acad Sci U S A* **86**, 3753-7 (1989).
3. Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E. & Bishop, J.M. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science* **224**, 1121-4 (1984).
4. Weiss, W.A., Aldape, K., Mohapatra, G., Feuerstein, B.G. & Bishop, J.M. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *EMBO J* **16**, 2985-95 (1997).
5. Sinn, E. et al. Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo. *Cell* **49**, 465-75 (1987).
6. Hackett, C.S. et al. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res* **63**, 5266-73 (2003).
7. Weaver, Z.A. et al. A recurring pattern of chromosomal aberrations in mammary gland tumors of MMTV-cmyc transgenic mice. *Genes Chromosomes Cancer* **25**, 251-60 (1999).
8. Jonkers, J. & Berns, A. Retroviral insertional mutagenesis as a strategy to identify cancer genes. *Biochim Biophys Acta* **1287**, 29-57 (1996).
9. Wu, X., Luke, B.T. & Burgess, S.M. Redefining the common insertion site. *Virology* **344**, 292-5 (2006).
10. Theodorou, V. et al. MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. *Nat Genet* **39**, 759-69 (2007).
11. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6 (2005).
12. Guy, C.T. et al. Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc Natl Acad Sci U S A* **89**, 10578-82 (1992).
13. Keng, V.W. et al. A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-74 (2009).
14. Wagner, K.U. et al. Cre-mediated gene deletion in the mammary gland. *Nucleic Acids Res* **25**, 4323-30 (1997).
15. Wagner, K.U. et al. Spatial and temporal expression of the Cre gene under the control of the MMTV-LTR in different lines of transgenic mice. *Transgenic Res* **10**, 545-53 (2001).
16. Welm, B.E., Dijkgraaf, G.J., Bledau, A.S., Welm, A.L. & Werb, Z. Lentiviral transduction of mammary stem cells for analysis of gene function during development and cancer. *Cell Stem Cell* **2**, 90-102 (2008).
17. Mates, L. et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet* **41**, 753-61 (2009).

Figure 9.1: SB lentiviral vectors. A. Structure of EIZ-SB11 and EIR-SB11. **B.** Detection of fluorescence markers in transduced 293T cells. **C.** Western blot validation of SB expression in lentiviral-transduced 293T cells.

Fig 9.1

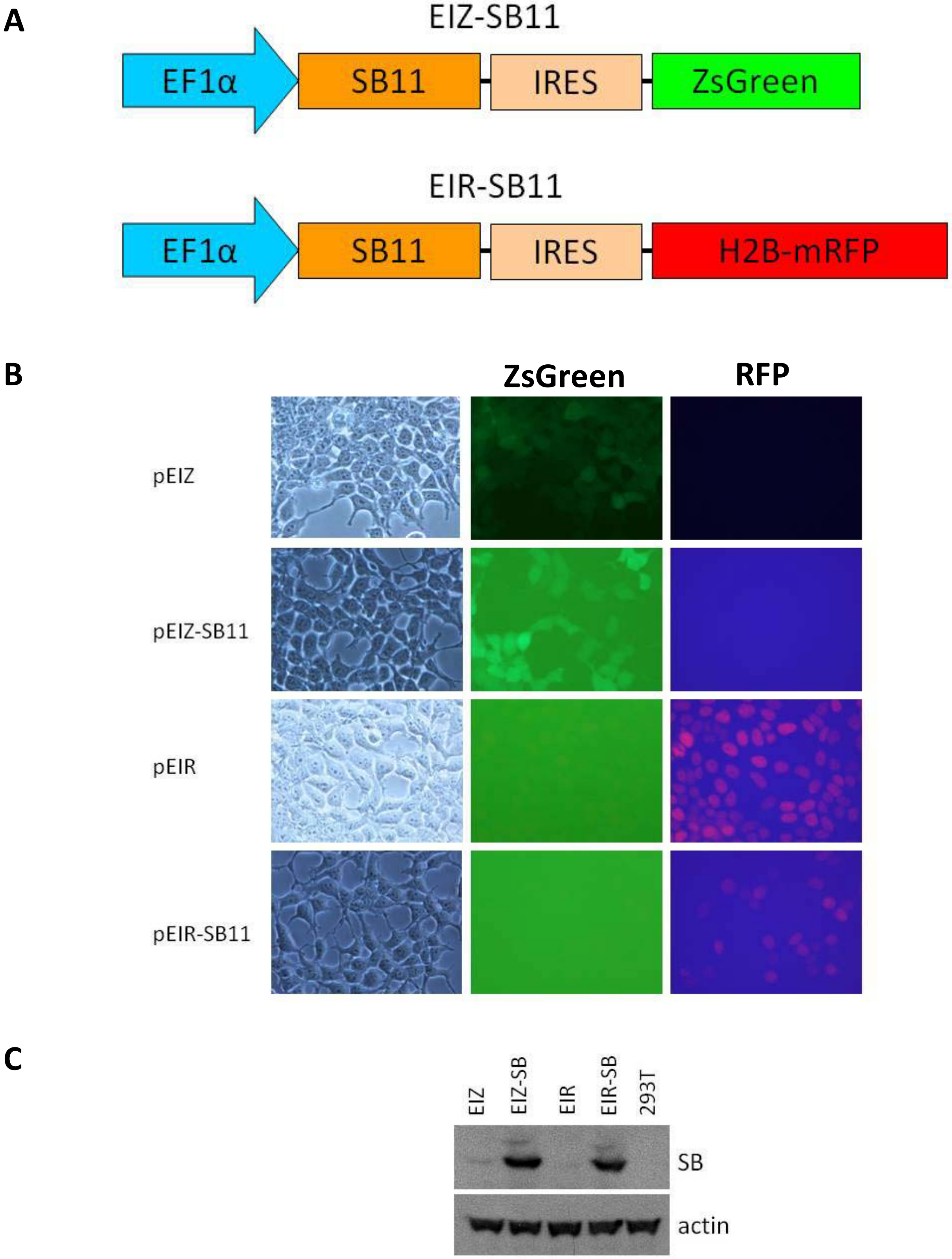
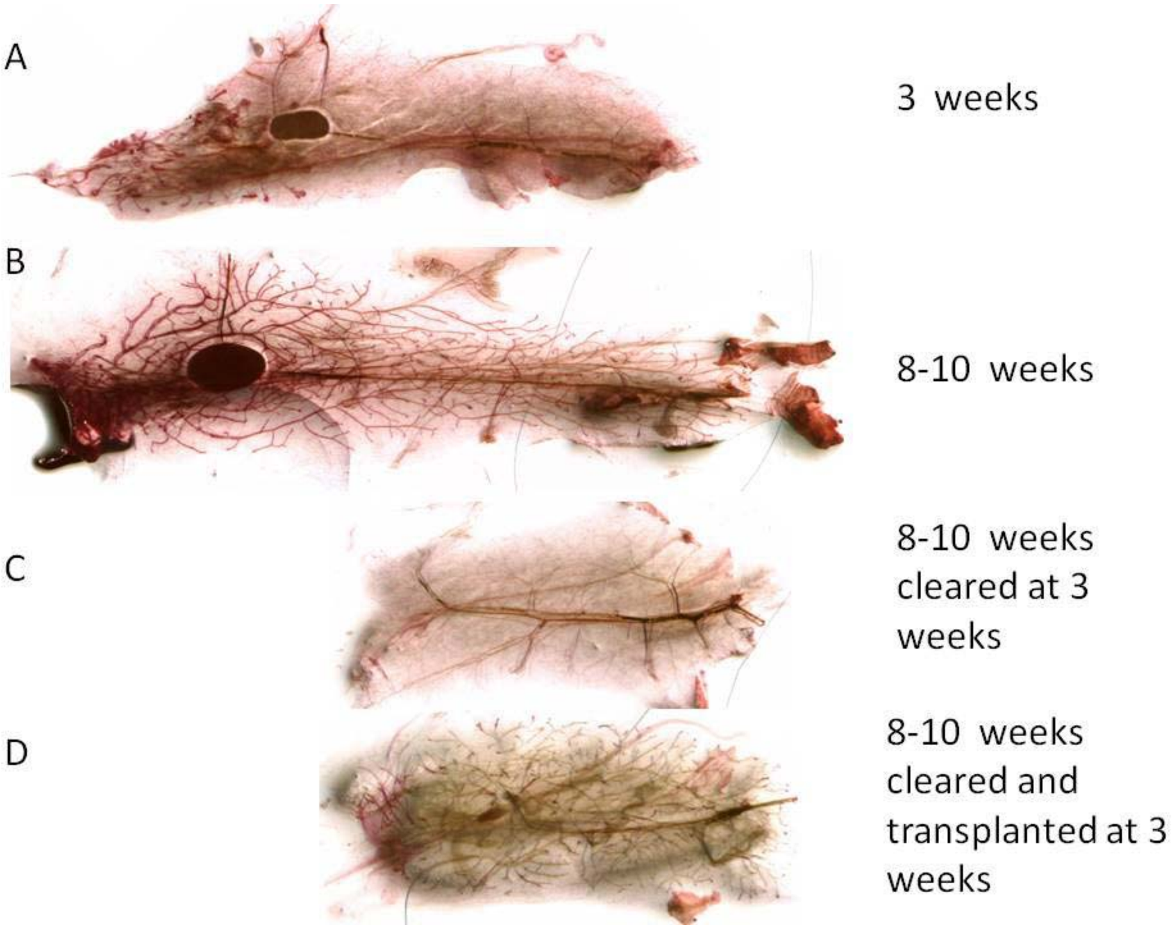

A
EIZ-SB11
EF1α
SB11
IRES
ZsGreen
EIR-SB11
EF1α
SB11
IRES
H2B-mRFP
B
ZsGreen
RFP
pEIZ
pEIZ-SB11
pEIR
pEIR-SB11
C
EIZ
EIZ-SB
EIR
EIR-SB
293T
SB
actin

Figure 9.2: Outgrowth of transplanted mammary epithelial cells. A. Mammary epithelial growth at 3 weeks of age. Branching epithelial structures are medial/distal to the lymph node (left). **B.** Mammary gland structure after 8 weeks of age. **C.** Cleared mammary gland at 10 weeks, showing reduced size, absence of lymph node, and lack of branched epithelium. **D.** Transplanted mammary epithelium. Cells were implanted near the medial edge (right), and show reversed directionality compared to endogenous structures (**B**).

Fig 9.2


A
3 weeks
B
8-10 weeks
C
8-10 weeks cleared at 3 weeks
D
8-10 weeks cleared and transplanted at 3 weeks

Chapter 10: Development of a highly–active conditional Sleeping Beauty transposase mouse for insertional mutagenesis in a diverse array of tissues.

Source: The following contains unpublished data.

Contributions: I performed the experiments presented, with the exception of the Ad-Cre administration in Figure 2, which was done by Anny Shai. Steven Chmura assisted with the screening of knock-in ES cells. Vincent Keng and Adam Dupuy provided unpublished constructs used to make components of the knock in construct. Martin McMahon, David A. Largaespada, and Nigel Killeen provided expert advice, and William A. Weiss supervised the project.

Development of a highly–active conditional Sleeping Beauty transposase mouse for insertional mutagenesis in a diverse array of tissues.

Christopher S. Hackett, Steven Chmura, Vincent Keng, Anny Shai, Adam Dupuy, Martin McMahon, David A. Largaespada, Nigel Killeen, and William A. Weiss

Abstract

The Sleeping Beauty (SB) transposon system has proven a useful tool for insertional mutagenesis for a limited number of cancer types. We hypothesized that generating a system capable of driving higher levels of transposase expression may extend the number of tissue types for which SB insertional mutagenesis could be used to identify novel genes driving cancer. Here we characterize a new genetically engineered mouse in which SB11 transposase has been knocked in to the *Rosa26* locus, under the control of a strong CAGGS promoter upstream of a lox-stop-lox cassette, conferring conditional expression under the control of Cre recombinase. We then validate the construct by demonstrating the activation of transposase in the lung via intranasal administration of Adeno-Cre. This system should facilitate forward mutagenesis screens in tissues in which the endogenous Rosa26 promoter does not drive adequate transposase expression, providing the opportunity to identify novel molecular pathways in several tumor types.

Introduction

The Sleeping Beauty transposase system has proven to be a powerful tool for forward genetic screens in somatic tissues not amenable to retrovirus-mediated insertional mutagenesis [1-6]. However, while several screens have successfully identified genes driving or accelerating tumorigenesis in many tissues, success with this approach has been far from universal. The system has worked well in mouse tumor model systems for the hematopoietic compartment[5], muscle[6], colon[1], liver[2] and the cerebellum ([5], Michael Taylor, personal communication). Results have also been achieved in glioma/astrocytoma[3] and prostate[4], though the efficiency of the system to induce tumors in these models was less than what was seen in the other screens. In contrast, the Sleeping Beauty insertional mutagenesis system has been used in several model systems without success. Though brain tumors have been observed, SB has not consistently induced or accelerated glioblastoma multiforme (GBM), the most malignant primary human brain tumor. The system has also been used without success to date in lung cancer, breast cancer, neuroblastoma, peripheral nerve sheath tumors, among others, and results have not yet been reported in skin and pancreatic tumor models. The several successful screens illustrate the broad potential of this technology. However, the system has not worked in model systems representing, collectively, some of the most common and deadly tumors, as well as several rare tumors for which forward genetic screens could greatly improve the current poor state of genetic and molecular characterization.

The success or failure of the Sleeping Beauty system to initiate or accelerate tumors in different tissues has been painfully hard to predict prior to actually carrying out a long, expensive screen that involves generating and tracking a large cohort of mice, in some cases for far more than a year. Few common themes can be observed from the group of successful endeavors; there seems to be no correlation with the requirement for a predisposing oncogene or a tumor suppressor knockout, or a specific genetic lesion, or a general tissue type (the technology has worked in some epithelial lineages but not others, as well as some neuronal and glial lineages but not others). Thus, no convincing mechanistic explanation exists for why the transposon system would be more efficient in a particular cell type or model system than another. Elucidating this phenomenon, or at the least modifying the system such that it has a broader range of efficacy, is an important step in expanding the utilization of this system to tissues where it is most needed.

One common theme among tissues in which Sleeping Beauty insertional mutagenesis has worked well is that the tissues are usually large organs, and/or have a high rate of cellular turnover. Additionally, the early progression of the system demonstrated that modulating the transposon dosage affected the severity of the resulting tumors; when RosaSB11 transposase mice were combined with T2/Onc2 high-copy transposon concatemers (200-400 copies), all mice that survived to birth rapidly developed a range of tumors[5], but when the RosaSB11 mice were combined with the lower-copy (20-40 copies) T2/Onc donor transposon concatemers, tumor onset was much less rapid[3]. From these two observations, we could build a model in which the frequency of tumor formation or acceleration is a function of the transposase expression levels (and duration of expression throughout life), the donor transposon dosage/copy number, and the number of cells under mutagenesis (itself a product of organ size, cellular turnover, and rates of differentiation past tumor progenitor cell status). From this model, we hypothesize that increasing transposase expression levels would increase mutagenic potential, and may increase the chances of developing or accelerating tumors in tissues that were previously unaffected by Sleeping Beauty mobilization. Modulation of transposase expression has not been tested directly from the existing constructs, as the most common and universal transposase lines are both driven by the endogenous *Rosa26* promoter. However, our IHC analysis showed that *Rosa26*-driven transposase expression varied between tissues (e.g. high levels in liver vs. moderate levels in brain vs. no detectable expression in adrenal medulla), and expression in brain varied greatly between cells. Additionally, anecdotally, transposase driven by tissue-specific promoters targeting expression to the cerebellum were more efficient at inducing tumors than the *Rosa26* constructs (Michael Taylor, unpublished observations). In this study, to test the hypothesis that higher transposase expression levels would confer more efficient mutagenesis, as well as to generate a general tool for Sleeping Beauty insertional mutagenesis in a broader range of tissues, we built a genomic targeting construct to drive Cre-induced SB11 expression from a ubiquitous, highly-active CAGGS promoter knocked into the Rosa26 locus.

Results

Construction of Rosa-CAGGS-LSL-SB11 knock-in mice

We generated a Rosa26-CAGGS-lox-EGFP-stop-lox-SB11 (Rosa-CAGGS-lsl-SB11) targeting construct by modifying multiple constructs used to generate the Rosa-lsl-SB11 mouse[2] by inserting a CAGGS (CMV enhancer, chicken β-actin promoter) element upstream of the LSL construct (**Figure 1A**). This vector was recombineered into a larger Rosa26 targeting plasmid, linearized, and transfected into E16 ES cells. Selected ES cell clones were screened by Southern blot using a probe outside of the targeting construct (**Figure 2B, C**). A Southern blot showing validation of the positive clones is shown in **Figure 2D**. One clone was injected into blastocysts to generate 13 chimeric mice. Chimeric males were mated to 129/SvJ females; one chimera was able to propagate the targeted allele through the germline.

Activation of the transposase in lung via Adeno-Cre

Adeno-Cre and Adeno- βgal was administered to the lung of a 10-week old male RCLSB mouse (with Adeno- βgal administered to a male littermate as a control) using standard protocols. After 1 week, the lungs were perfused, dissected, fixed, and analyzed using IHC. As shown in Figure 2, robust nuclear staining of cells lining the lung was detected in many cells of the mouse given Adeno-Cre, with no cells displaying staining in a control mouse that received Adeno- βgal, suggesting the RCLSB construct is functional and capable of driving high expression levels in the lung.

Discussion

While the Sleeping Beauty transposon system unlocked the potential to discover novel oncogenes and tumor suppressors using somatic insertional mutagenesis in tissues not accessible to mutagenic retroviruses, the system has failed to drive tumorigenesis in several tissue types. We hypothesized that this was a result, at least in part, from low, inconsistent, or absent transposase expression in several tissue types using the most widely-used transposase expression mice. The most commonly used mice harbor a knock in of the SB11 transposase (either constitutively expressed or Cre-activated) into the *Rosa26* locus. While the *Rosa26* locus is a popular site for targeted insertions due to its presumed open chromatin structure and ubiquitous expression in all tissues[7], the actual activity of the *Rosa26* promoter is not particularly high. Additionally, expression patterns are not even across tissues, and even within different cell types of single tissues. In particular, we demonstrated that expression of SB11transposase was absent in the adrenal medulla of Rosa-SB11 mice, suggesting the *Rosa26* promoter was not active in the peripheral sympathetic nervous system. We also observed that staining in the brain was overall less strong than the liver (not shown) and that expression levels varied highly from cell to cell, possibly explaining the difficulty of consistent production of brain tumors using this system.

These observations led us to hypothesize that stronger, ubiquitous (if conditional) transposase expression may yield higher rates of mutagenesis and tumor formation in several tissues. To test this, we modified the conditional RosaSB11 targeting constructs to contain a constitutive, highly-active CAGGS promoter, a fusion of the chicken β actin promoter and CMV enhancer. We validated the activity of this construct in knock-in mice by administering Adeno-Cre to the lung, and observing activation of SB11 transposase using IHC.

If the probability of generating a tumor via transposon-based insertional mutagenesis is a product of cell number (a function of tissue size and cellular turnover), transposon dosage, and transposase expression, this construct should tilt the balance further in favor of tumor formation in tissues where previous attempts to generate tumors have failed. This may facilitate the identification of oncogenes and tumor suppressors in tissues such as the peripheral nervous system, certain compartments of the brain, the lung, the pancreas, and other tissues for which no large-scale insertional mutagenesis screen has been successful.

Methods

Construction of Rosa-CAGGS-lsl-SB11: SB11 transposase was inserted into the SalI site of a Rosa26 targeting vector (a gift from Vincent Keng (University of Minnesota) and Adam Dupuy, (University of Iowa)) modified to contain a CAGGS promoter upstream of the lox-EGFP-stop-lox element to make Rosa-CAGGS-LSL-SB11 (RCLSB). This construct was recombineered into a larger Rosa targeting vector for ES cell targeting.

Generation of RCLSB ES cells and mice: ES cell clones were transfected and selected by the UCSF transgenic core. Genomic DNA samples from 9 ES cell clones were screened by Southern blot using a probe outside of the targeting construct. The probe was generated by TOPO cloning a short sequence amplified from wild-type genomic DNA; the probe was then digested out of the plasmid and radiolabeled with Kleinow fragment polymerase (Prime-It II, Stratagene, Santa Clara, CA) . DNA was digested with ApaI; targeted alleles generated a 13kb band, while the wild-type allele generated an 8kb band. 11 clones generating two bands by Southern blot were expanded and re-screened, with 8 of these showing definitive patterns. ES cells from one clone were injected into blastocysts. Chimeric pups were screened by PCR for SB11 transposase. Nine positive male chimeras were mated to 129/SvJ females. One chimera propagated the targeted allele through the germline.

Administration of Adeno-Cre to lung and detection of SB11 expression: Solutions containing either Adeno Cre or Adeno-βgal (each at 10^8 CFU) were administered intranasally to 10 week old male RCLSB following standard protocols. After 1 week, the lungs were perfused with PBS and mice were euthanized. Lungs were dissected, fixed, and analyzed by IHC for SB11 transposase (R&D Systems, Minneapolis, MN).

References

1. Starr, T.K. et al. A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323**, 1747-50 (2009).
2. Keng, V.W. et al. A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-74 (2009).
3. Collier, L.S. et al. Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality. *Cancer Res* **69**, 8429-37 (2009).
4. Rahrmann, E.P. et al. Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping Beauty transposon-based somatic mutagenesis screen. *Cancer Res* **69**, 4388-97 (2009).
5. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6 (2005).
6. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6 (2005).
7. Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev* **5**, 1513-23 (1991).

Figure 10.1: Generation of RCLSB. A. The RCLSB construct. **B-D.** Southern screening strategy for the knock-in construct. DNA was digested with ApaI and screened using probe us1, hybridizing to the position shown on the maps. **B.** Position of the probe relative to the the wild-type allele. An ApaI digest generates an 8kb fragment. **C.** Due to a novel ApaI site within the knock-in construct, the targeted allele generates a 13kb fragment. **D.** Validation Southern blot showing the two bands in 8 of the selected targeted clones. E14 is an archival wild-type negative control.

Fig 10.1

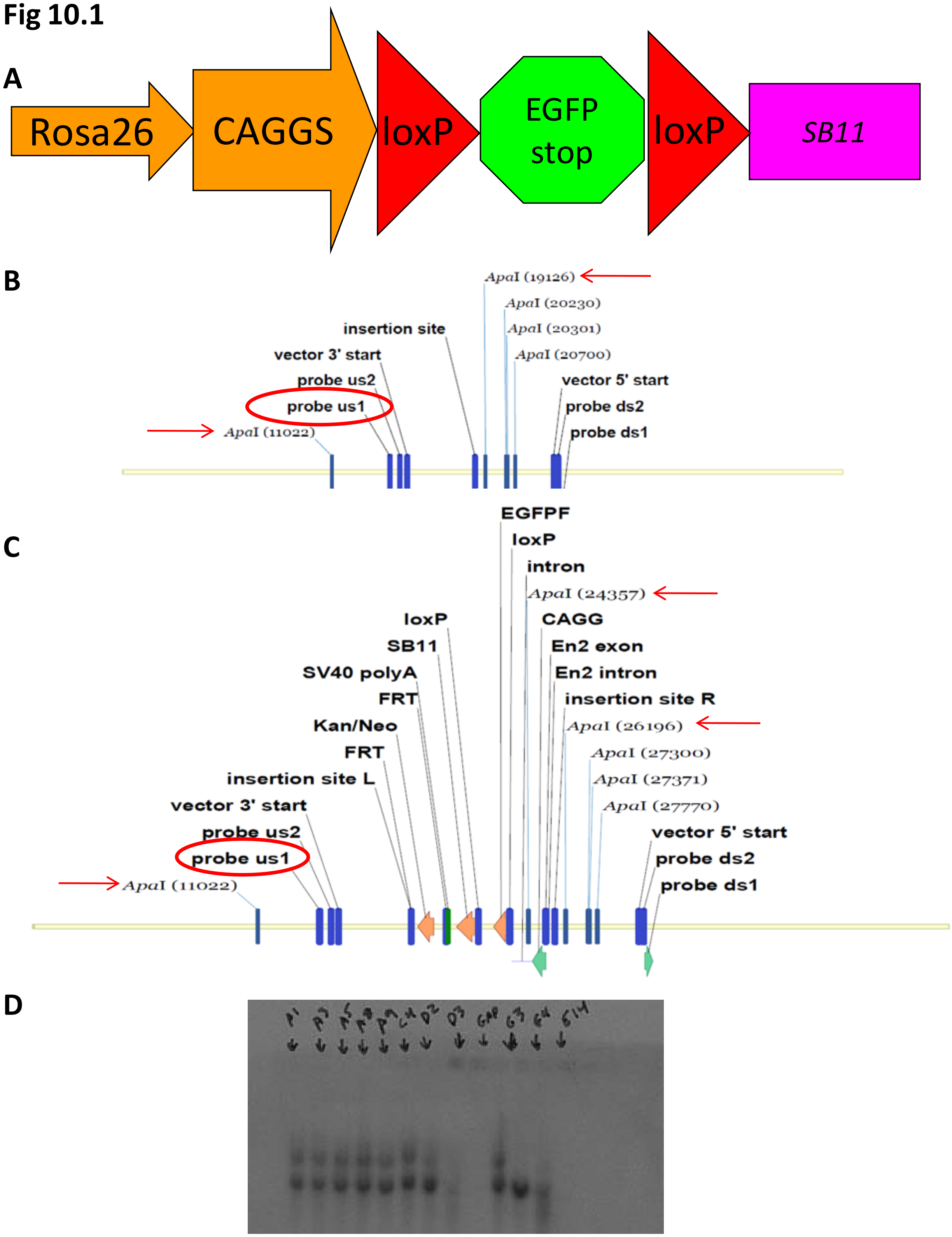
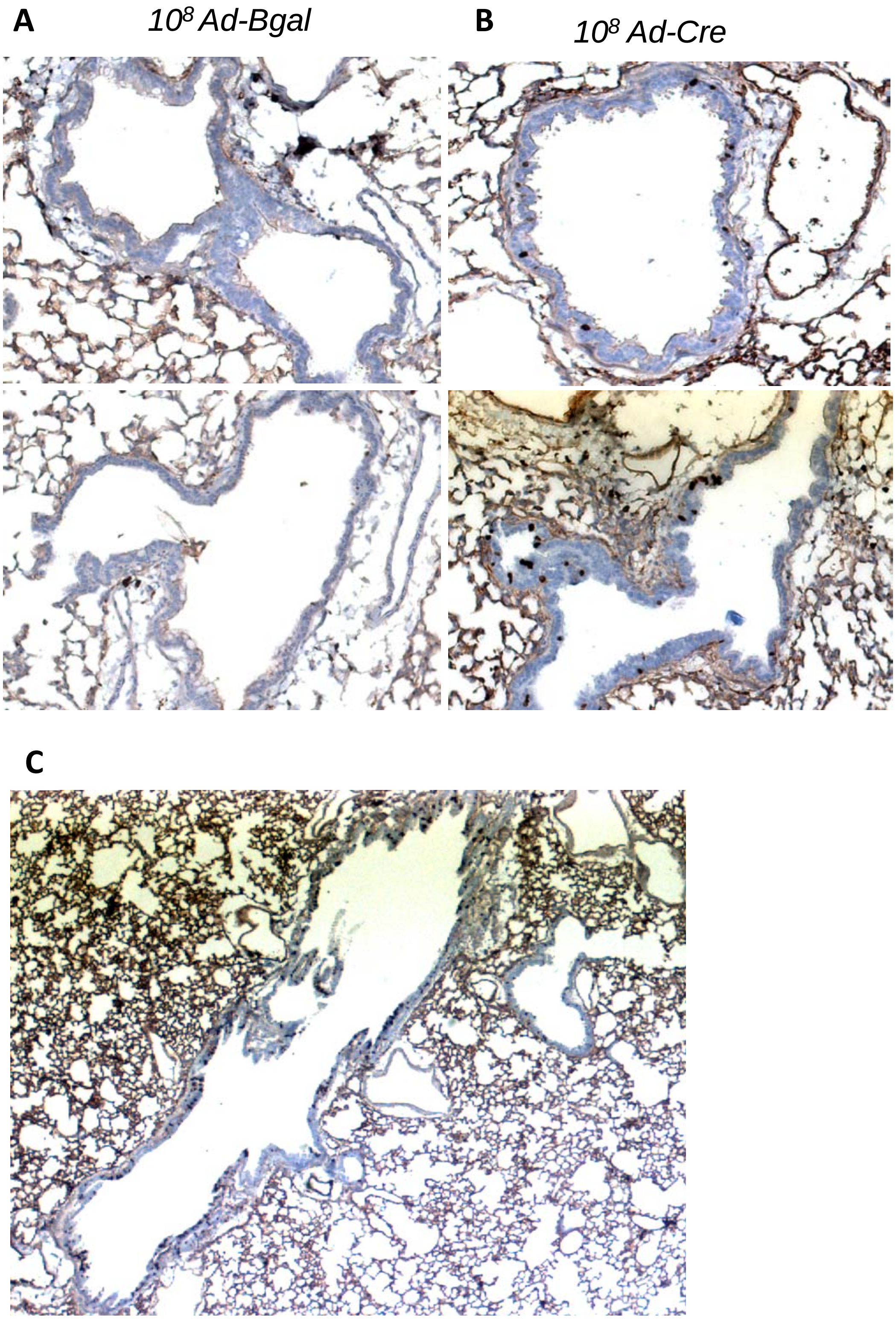

A
Rosa26
CAGGS
loxP
EGFP
stop
loxP
SB11
B
ApaI (19126)
ApaI (20230)
ApaI (20301)
ApaI (20700)
insertion site
vector 3' start
probe us2
probe us1
ApaI (11022)
vector 5' start
probe ds2
probe ds1
C
EGFPF
loxP
intron
ApaI (24357)
CAGG
En2 exon
En2 intron
insertion site R
ApaI (26196)
ApaI (27300)
ApaI (27371)
ApaI (27770)
loxP
SB11
SV40 polyA
FRT
Kan/Neo
FRT
insertion site L
vector 3' start
probe us2
probe us1
ApaI (11022)
vector 5' start
probe ds2
probe ds1
D
A1
A3
A5
A8
A9
C4
D2
D3
GAP
G3
G4
G14

Figure 10.2: Adeno-Cre administration activates SB11 in lung. A. Lung section from an Ad-βgal control showing no nuclear staining for SB11. **B.** Lung section from an Ad-Cre, showing strong nuclear staining for SB11 in multiple cells. **C.** A lower magnification showing a wider field from the lung in (B), demonstrating widespread activation of SB11.

Fig 10.2



Chapter 11: Conclusion

Conclusion

In this collection of work, we have applied complex genetic and genomic approaches to uncover novel pathways involved in neuroblastoma development. We first pursued the observation that in a mouse model for the disease, tumor incidence was dependent on strain background, suggesting that polymorphisms in endogenous genes interacted with the transgene. We demonstrated with a classical linkage analysis that the genetics of this phenomenon was complex. We then performed an expression-QTL analysis of normal tissue (superior cervical ganglia) and tumors to identify a gene, *Arg1*, that interacts with components of the GABA signaling pathway both genetically and biochemically. The convergence of susceptibility genetics, gene expression genetics, and common functional biochemistry implicates this set of genes in tumor development. We have begun to test existing *Arg1* inhibitors *in vitro,* showing that they are capable of impacting proliferation of neuroblastoma cell lines. The progression from an observation of strain-specific tumor susceptibility to a compound ready for preclinical testing represents a rare outcome in the field of modifier genetics.

The microarray data used to identify the *Arg1*-GABA interaction represents one of the most extensive transcriptional characterizations of the peripheral sympathetic nervous system to date, as well as a body of data from a relatively pure neuronal population growing *in vivo*. Additionally, since this dataset came from a large number of genetically heterogeneous mice, subtle genetic perturbations allowed us to explore the regulation of gene expression within the dataset without the need for an arbitrary outside frame of reference. This dataset provided an opportunity for several spinoff projects to explore gene expression patterns and identify putative functional connections between genes in neurons *in vivo*. We used this network to explore interactions with genes contributing to neuroblastoma, diseases of the peripheral nervous system, and neuronal apoptosis. Most interestingly, we identified a large network of genes very closely linked to genes involved in hearing. Several of the genes in this network co-localize in the genome with loci for hereditary hearing loss, providing intriguing new candidate genes from a range of functional groups.

We then returned to neuroblastoma and sought to apply a forward genetics approach to identify genes driving tumorigenesis. We adopted the Sleeping Beauty insertional mutagenesis system, based on a mutagenic vertebrate DNA transposon. As “early adopters”, we became involved in the basic characterization of the system. We first developed a bioinformatics approach to predict the insertion preferences of the vector. We then performed a similar analysis for other vectors used in insertional mutagenesis and gene therapy to illustrate the differences in biases between vectors. This area is important for forward genetics because it assists in determining whether a given insertion is biologically relevant in the context of a particular screen, or whether the insertion is an effect of the vector’s inherent insertion bias. Though not described here, this area is also important for gene therapy, as inadvertent insertional mutagenesis leading to malignancy is among the major concerns in the field and thus an understanding of the characteristics of the vectors in use is critical. We next used array CGH to demonstrate that the Sleeping Beauty system does not promote tumor development primarily through genome-wide genomic instability, in spite of the fact that the transposase creates temporary double-strand breaks during the process of transposition. This result lends further credibility to the model in which insertional activation or deactivation of genes drives tumors.

While contributing to the fundamental characterization of the system, we also attempted to use the Sleeping Beauty system for neuroblastoma and breast cancer. After establishing that the existing tools were insufficient to drive tumors in the peripheral neural crest and the mammary gland, we developed novel approaches. For neuroblastoma, we developed the TH-SB11 transgenic line, which expresses high levels of the SB transposase in the adrenal medulla, and was capable of producing multiple tumors in a small initial cohort of mice. In breast cancer, we established that a mammary epithelial transplant system represents the most promising approach to driving tumors specifically in the mammary gland. We also developed a conditional SB knock in driven by a strong promoter, which may be capable of generating tumor in several tissues where, as in the peripheral nervous system, the conventional tools have failed.

Our overall goal was to identify novel genes and pathways driving neuroblastoma development. Our early genetic results indicated that our system was consistent with many contemporary genomics and genetics results in biomedical sciences: diseases are often driven by strikingly complex interactions of many genes. As demonstrated most notably by recent genome wide association studies, in many cases, there is no “low hanging fruit”; single genes or small groups of genes acting as the primary drivers of disease. Thus, more advanced experimental approaches that take this complexity into consideration will be necessary. We have utilized some of these approaches here, identifying several genes, multiple functional groups, and a potential therapeutic target. It is hoped that this work lays the foundation for the discovery of even more genes and molecular pathways involved in neuroblastoma and other diseases.

Library Release

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

9-2-10

Author Signature Date