“Calling Cards,” a Novel Approach for Identifying Transcription Factor-DNA Interactions, Uncovers AML-related HOXA9-NUP98 Fusion Product Targets

Gabriel A. Bien-Willner[1,2], Akiko Takeda[2], David Mayhew[1], Robi Mitra[1], and Nabeel Yaseen[2]
[1]Department of Genetics, and [2]Department of Pathology and Immunology, Washington University in St. Louis

UPDATED ABSTRACT

Background/Significance:
The NUP98-HOXA9 fusion product is an onco-protein that has been described in acute myeloid leukemia (AML). It is leukemogenic in mice; and has been shown to induce cell proliferation and prevent differentiation in human and mouse hematopoietic precursors. The identification of the targets of such oncogenic fusion proteins is critical to the progress of personalized medicine as it may help aid therapy in a genotype-specific manner. However, the targets of this aberrant transcription factor have been fully elucidated. This is in part because current methods of large-scale identification of DNA-protein interactions rely heavily on quality antibodies, and these are not always available or easy to produce. Furthermore, these events may happen at specific time points in the cell cycle, and the timing of these experiments may prohibit the identification of many relevant targets.

Methods:
“Calling Cards” is a novel approach to identify transcription factor binding sites genome-wide with the aid of cellular machinery that is native to the cell: transposable elements and transposases. Briefly, the eukaryotic PiggyBac (PB) transposase is fused to the transcription factor of interest, and introduced into a relevant cell along with a bar-coded transposon. Thus, when the transcription factor binds its target, the fused PB induces the insertion of the barcoded transposon nearby, thus leaving a permanent “calling card” of where it has bound.

The *NUP98-HOX9A* fusion gene was linked to the PB transposase (helper plasmid) and transfected into mouse embryonic stem cells (RW4) along with the selectable and barcoded transposon (donor plasmid) as a pilot experiment. Transfected cells were then selected, and unique sites of integration of the barcoded transposons (i.e., sites of NUP98-HOX9A binding) were identified via next-generation sequencing (Illumina HiSeq). Subsequently, the novel technique was applied to human hematopoietic stem cells.

Results:
Over 2300 unique insertion events were identified in the preliminary mouse stem cell study, a vast majority forming clusters in the 5’UTR of known genes. Two biological replicates, including adding the PB and both the N and C terminus of *NUP98-HOX9A*, yielded similar results. Following these preliminary studies, the experiment was conducted in human hematopoietic stem cells. Over 160,000 unique insertion events were identified. Compared to controls, these were clustered around >700 genomic loci.

Conclusions:
Calling Cards is a novel technique that can readily identify the targets of transcription factors associated with oncogenic translocations. As such, it may be a useful tool to identify unique genes and pathways associated with specific cancer genotypes, and help advance personalized medicine.

INTRODUCTION

BACKGROUND:
Until recently, identifying protein-DNA interactions was very difficult and required a lot of evidence from a variety of sources to identify. These included:

- Patient data (mutations/rearrangements)
- Lots of *in vitro* work
- Animal studies for validation

CURRENT TECHNIQUES
This process became much simpler with novel genomics approaches- first with ChIP-chip and currently with ChIP-seq, where genomic libraries are constructed from DNA-protein interactions captured with chromatin immunoprecipitation and then massively sequenced.

However, this process has a few drawbacks, namely:

- Good antibodies must be available
- Protein-DNA interactions are fixed at a single point in time, and may miss important interactions
- Amplification steps may increase non-specific interactions

NUP98/HOXA9

Nucleoporin (nuclear pore protein) 98 is thought to be involved in protein import into the nucleus. It is also thought to be involved in RNA export, and may play a role in intranuclear trafficking. *NUP98* has been associated with acute myeloid leukemia (AML) as translocations involving this gene has been observed in 15 chromosomal rearrangement-positive AML variants. Of these, 8 have had transcription factors as the translocation partner for *NUP98,* resulting in a chimeric fusion gene.

HOXA9 is a transcription factor critical in development (anteroposterior patterning) and has a complex role during hematopoiesis. HOXA9 overexpression can immortalize myeloid progenitors and can induce AML in mice.

NUP98-HOXA9 FUSION

- Has been shown to prevent maturation and increase proliferation in cell culture
- When introduced in stem cells, found in all leukocytes except B cells
- When introduced into mice, induces chronic and acute leukemia (Kroon, EMBO 2001)
- These mice get MPD that evolved into AML
- Is thought interact with *Meis1* for pathogenic effect
- Microarray experiments (Takeda, Cancer Res 2006) show that there is a difference in gene expression profiles between cells transfected with the fusion product compared to controls
- Thus this fusion likely acts by altering gene expression
- No antibodies for this fusion exist, no binding sites are described
- Would be interesting to map direct binding sites for this fusion, and see if we can identify specific targets that lead to leukemogenesis

CALLING CARDS

BIOLOGY
Much of the human genome is made up of repeat elements, such as LINES and SINEs (Alu repeats). These are actually transposons- segments of DNA that “jump” from one place to another in the genome. They can cause novel gene functions that drive evolution or deleterious mutations. Proteins called transposases can “cut and paste” transposons randomly in the genome. One mammalian active transposase called PiggyBac (PB) is the focus of this study. PB transposase specifically recognizes PB inverted terminal repeats (ITRs) that flank the transposon; it binds to these sequences and catalyzes excision of the transposon. PB then integrates at TTAA sites.

CONCEPT
The novel concept is to use the transposable machinery described above to mark desired protein-DNA interactions by fusing a transcription factor of interest to PB, and barcode the transposon it directs. As the transcription factor binds its target, the linked transposase would recruit the coded transposon to integrate into the genome at that spot- a mark that would be left after the DNA-protein interaction was dissolved. Thus, the transcription factor would leave a “calling card” of where it had been. The efficiency of transposition is somewhere between 1-10% of events.

METHOD

CONSTRUCTS
Two plasmids were created. The first (here called the “Donor”) contains the PB transposon with ten different barcodes. This ensures that if transposition occurs at the same locus, at least 10 unique insertion events will be observed. Also included are selectable markers (Td-tomato, neomycin), sequencing primers, and restriction enzyme sites for library preparation. The second plasmid (“Helper”) contains a fusion of the PB transposase to the transcription factor of interest (NUP98-HOXA9) by an 18AA linker sequence. This product is driven by a strong CMV promoter.

CELL LINES
Two cell lines are used in this experiment. First, mouse embryonic stem cells (RW4) are used because of their rapid growth and potential to express a variety of genes since they are not yet differentiated. Secondly, human hematopoietic stem cells were obtained and maintained in a restrictive media that prevents differentiation.

EXPERIMENT
Cells were transfected with a mixture of both plasmids. Success of the transfection was observed by red fluorescence of transfected cells. Controls included the same donor plasmid along with a helper plasmid that contained PB transopsase only. Cells then underwent selection (G418 in the RW4 cell line, no selection in the hematopoietic stem cells) and were allowed to grow for 4 days (RW4 cells) or 1 day (human cells). They were then harvested and DNA was extracted.
Extracted DNA was used to create a genomic library of the insertion sites (see Fig. 1). DNA was digested with specific enzymes that would cut at the transposon insertion site and nearby DNA. The fragments were then circularized, and a nested PCR was done to amplify all insertion sites. The resulting DNA fragments of 200-1000 bp were submitted for Solexa sequencing with the Illumina HiSeq 2000, and 100-150 million paired-end reads were conducted. These were then aligned with the reference mouse and human genomes.

OVERVIEW

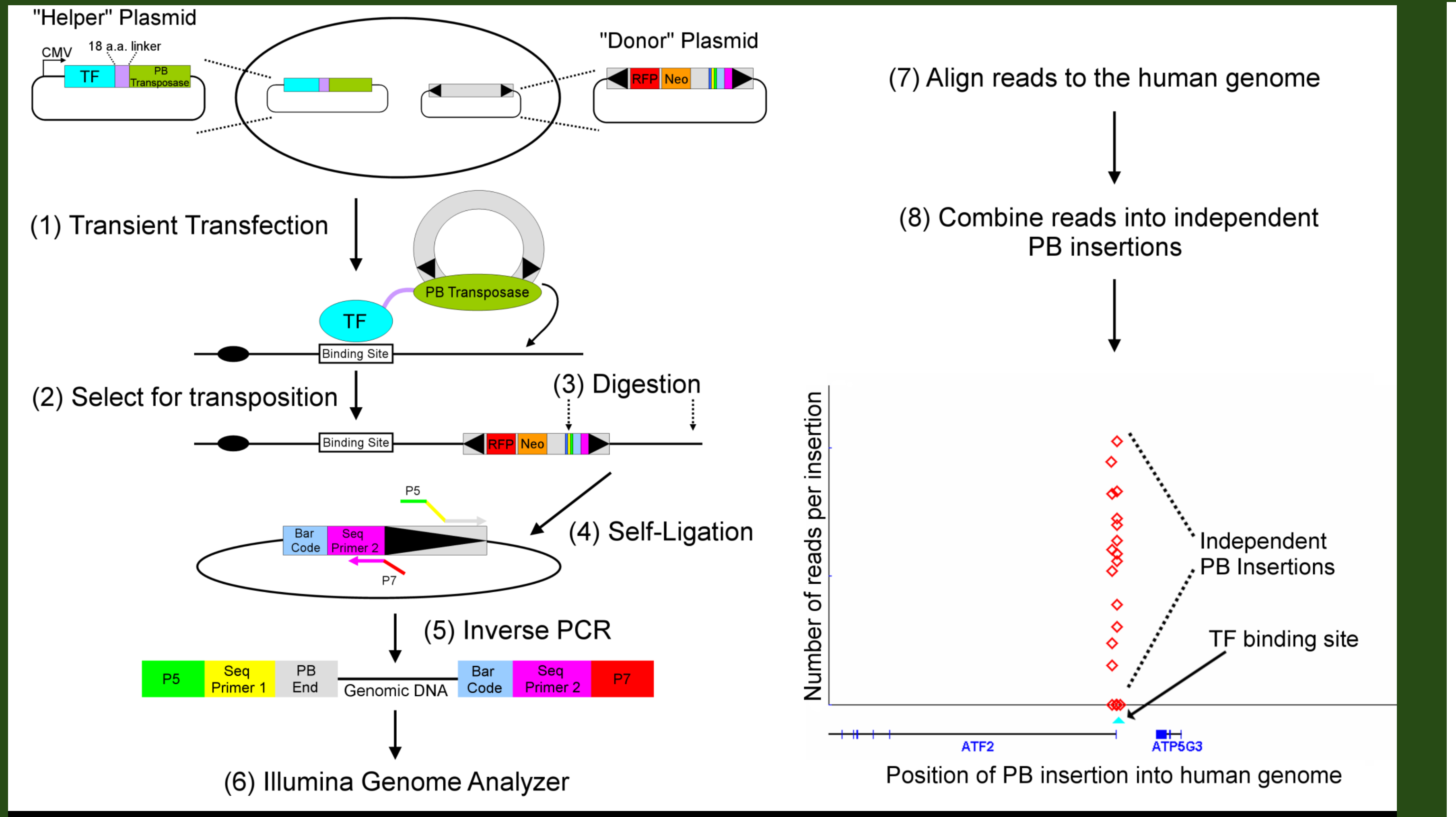


Fig.1. Illustration of the Callingcards assay. (1) Helper and Donor plasmids are transiently transfected into the cell of choice. (2) cells undergo selection for the transfected cells to maximize yield. This can be chemical selection or cell sorting. Cells are allowed to grow, then the DNA is harvested. (3) DNA is digested with enzymes that cut the inserted transposon and nearby DNA. (4) These fragments then undergo autoligation, and an inverse PCR step (5) amplifies the insertions. These are then submitted for Solexa sequencing (6) and aligned to the genome (7). (8) Finally, analysis is based not on total reads, but independent insertion events.

RESULTS

RW4 (mouse ES cell) experiment:

The helper plasmid was constructed with the NUP98-HOXA9 insert both at the N and C terminus of the PB transposase (called “PB-fusion” and “Fusion-PB”. A lipofectamine-based transfection was performed. After DNA harvesting, library preparation, Solexa sequencing, and analysis, the PB-fusion construct yielded 1425 unique genomic insertions and fusion-PB 909 unique insertions.
Together, there were more than 100 clusters of insertions (more than 3 independent binding events/site), and 27 sites with more than 4 independent insertions. Of these:

- 67% were observed in both study sets
- 44% are known to be expressed in leukemia cell lines

A sample of such genes includes:

- *Mll3*
- *Asxl1*
- *Bag3*

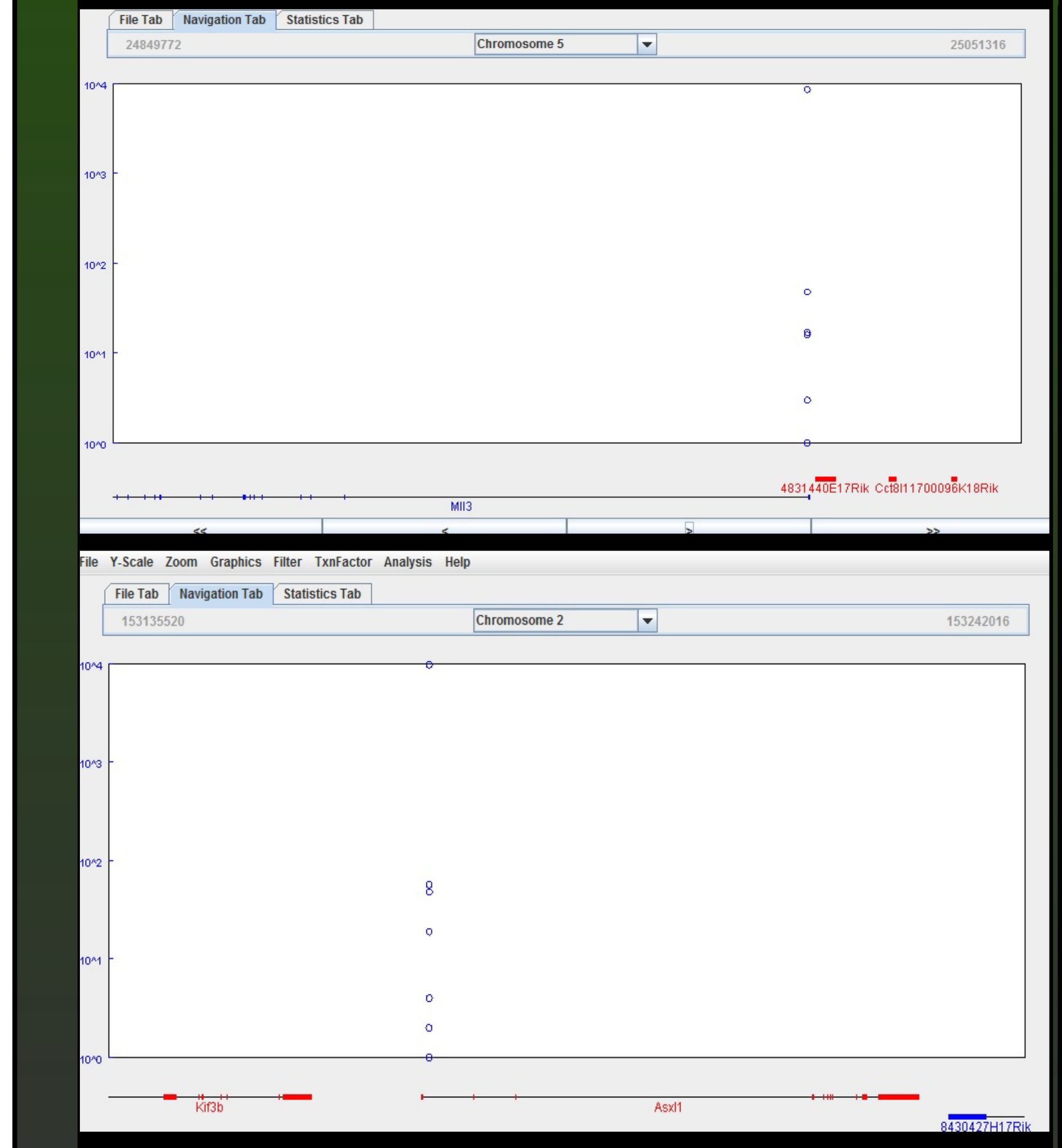


Fig.2. A sample of insertion events mapped back to the mouse genome. The upper panel displays Insertions that map to the 5’ end of the *Mll3* gene. The lower panel shows similar findings in *Asxl1*.

Human Hematopoietic stem cells

The “PB-Fusion” construct was transfected along with the donor plasmid via electroporation. Protocols were optimized to ensure a maximum number of unique insertion events. Chemical selection was not necessary. 5M viable stem cells were used at the start of the transfection process for both the PB-Fusion construct and controls. The resulting sequences were mapped back to the human genome.

- Control plasmid yielded 126,139 unique insertion events
- PB-Fusion calling cards yielded 163,049 unique insertions

Analysis first created clusters of insertion sites. These were then compared to controls, and were removed from further analysis if a similar cluster was also present in controls.

Final results:

- 724 clusters of insertion sites identified, these were comprised of:
- 575 Ensembl genes/coding regions, and
- 348 named genes

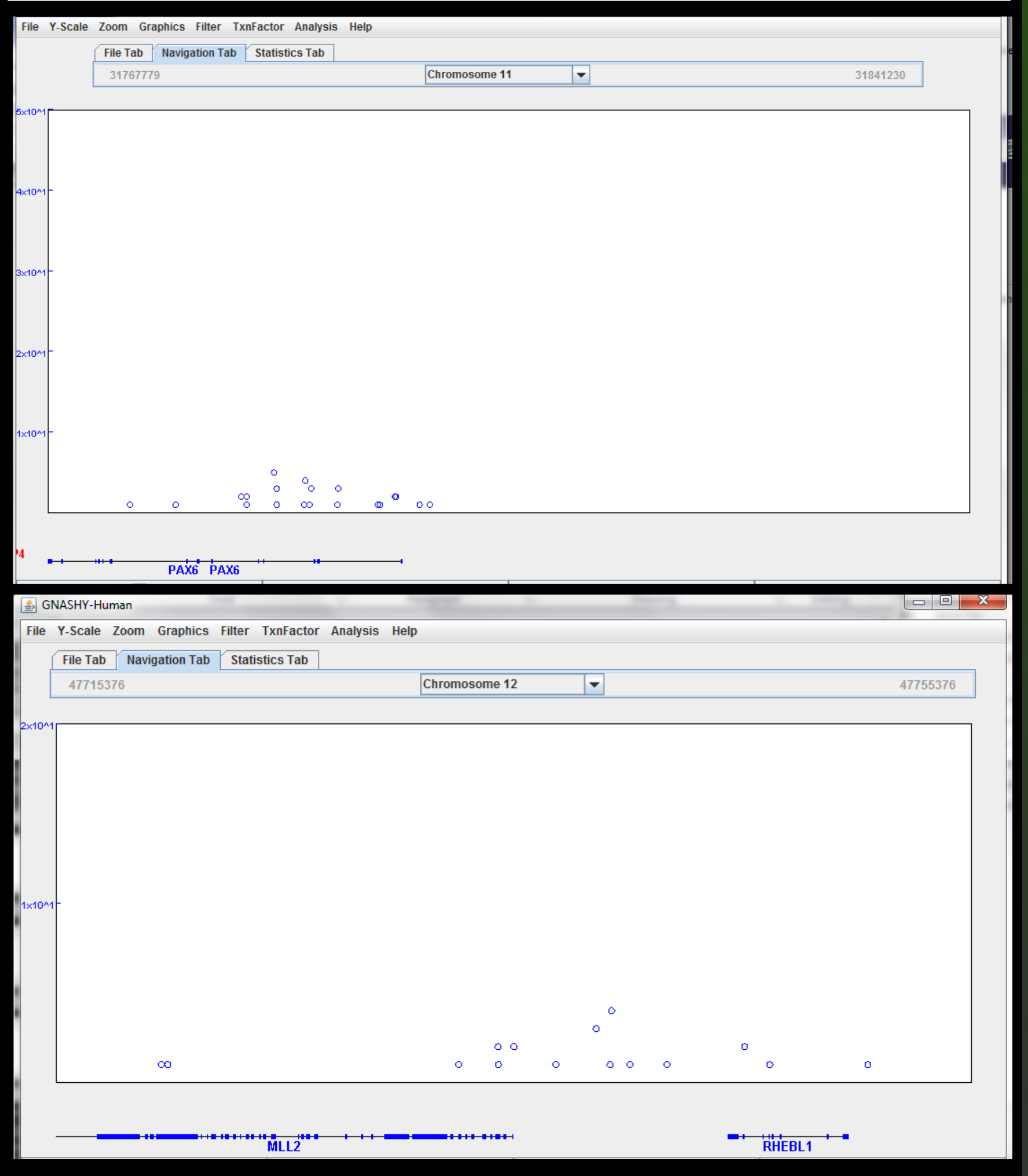
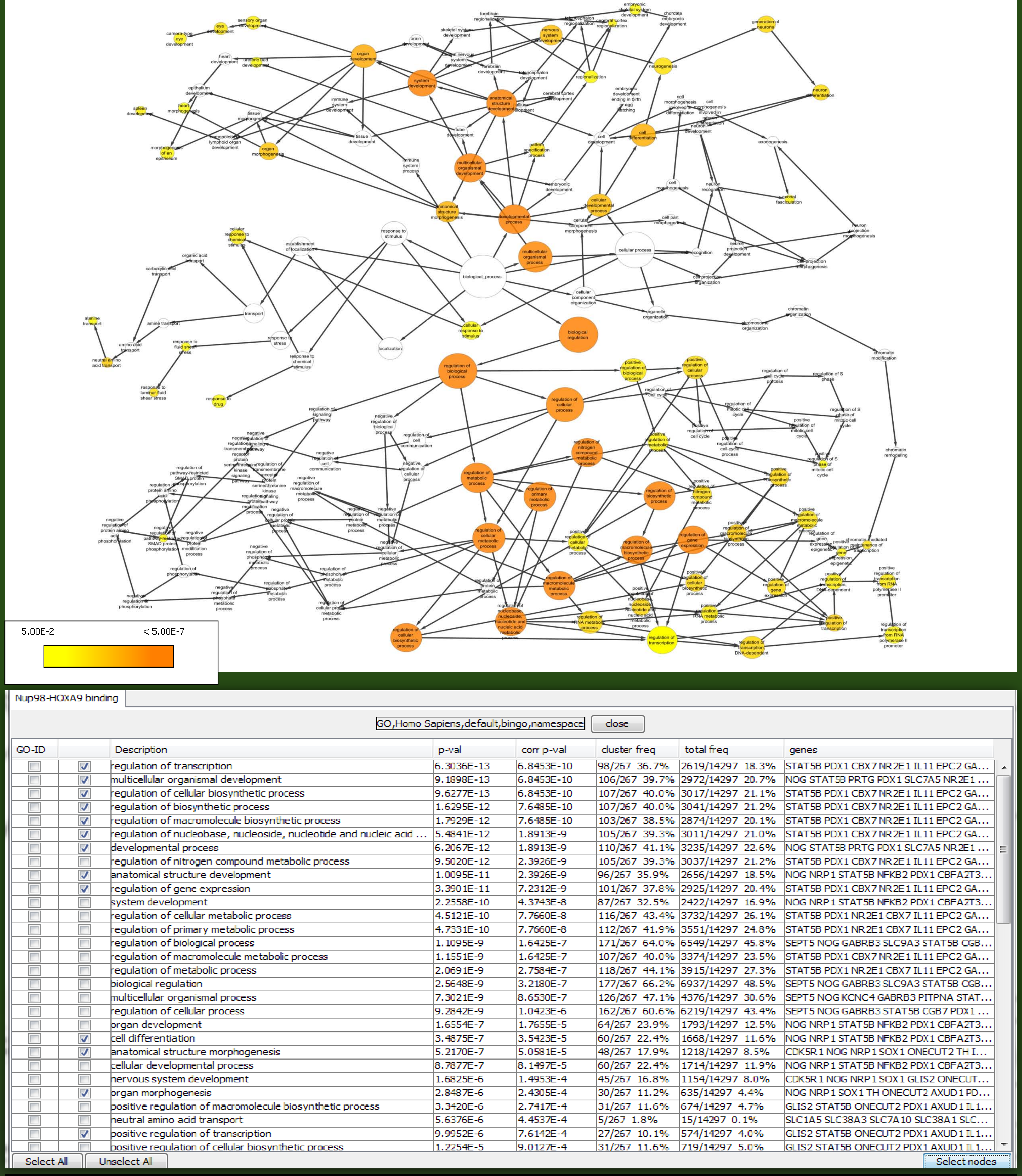


Fig.3. Examples of “calling cards” left by NUP98-HOXA9 in the human genome. Both genes pictured are known to be involved in leukemogenesis.


5.00E-2
< 5.00E-7
regulation of transcription
regulation of cellular biosynthetic process
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
regulation of RNA metabolic process
positive regulation of transcription
regulation of transcription, DNA-dependent
regulation of transcription from RNA polymerase II promoter
Nup98-HOXA9 binding
GO,Homo Sapiens,default,bingo,namespace
close
Select All
Unselect All
Select nodes

GO-ID		Description	p-val	corr p-val	cluster freq	total freq	genes
☐	☑	regulation of transcription	6.3036E-13	6.8453E-10	98/267 36.7%	2619/14297 18.3%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☑	multicellular organismal development	9.1898E-13	6.8453E-10	106/267 39.7%	2972/14297 20.7%	NOG STAT5B PRTG PDX1 SLC7A5 NR2E1 ...
☐	☑	regulation of cellular biosynthetic process	9.6277E-13	6.8453E-10	107/267 40.0%	3017/14297 21.1%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☑	regulation of biosynthetic process	1.6295E-12	7.6485E-10	107/267 40.0%	3041/14297 21.2%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☑	regulation of macromolecule biosynthetic process	1.7929E-12	7.6485E-10	103/267 38.5%	2874/14297 20.1%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☑	regulation of nucleobase, nucleoside, nucleotide and nucleic acid ...	5.4841E-12	1.8913E-9	105/267 39.3%	3011/14297 21.0%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☑	developmental process	6.2067E-12	1.8913E-9	110/267 41.1%	3235/14297 22.6%	NOG STAT5B PRTG PDX1 SLC7A5 NR2E1 ...
☐	☐	regulation of nitrogen compound metabolic process	9.5020E-12	2.3926E-9	105/267 39.3%	3037/14297 21.2%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☑	anatomical structure development	1.0095E-11	2.3926E-9	96/267 35.9%	2656/14297 18.5%	NOG NRP1 STAT5B NFKB2 PDX1 CBFA2T3...
☐	☑	regulation of gene expression	3.3901E-11	7.2312E-9	101/267 37.8%	2925/14297 20.4%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☐	system development	2.2558E-10	4.3743E-8	87/267 32.5%	2422/14297 16.9%	NOG NRP1 STAT5B NFKB2 PDX1 CBFA2T3...
☐	☐	regulation of cellular metabolic process	4.5121E-10	7.7660E-8	116/267 43.4%	3732/14297 26.1%	STAT5B PDX1 NR2E1 CBX7 IL11 EPC2 GA...
☐	☐	regulation of primary metabolic process	4.7331E-10	7.7660E-8	112/267 41.9%	3551/14297 24.8%	STAT5B PDX1 NR2E1 CBX7 IL11 EPC2 GA...
☐	☐	regulation of biological process	1.1095E-9	1.6425E-7	171/267 64.0%	6549/14297 45.8%	SEPT5 NOG GABRB3 SLC9A3 STAT5B CGB...
☐	☐	regulation of macromolecule metabolic process	1.1551E-9	1.6425E-7	107/267 40.0%	3374/14297 23.5%	STAT5B PDX1 CBX7 NR2E1 IL11 EPC2 GA...
☐	☐	regulation of metabolic process	2.0691E-9	2.7584E-7	118/267 44.1%	3915/14297 27.3%	STAT5B PDX1 NR2E1 CBX7 IL11 EPC2 GA...
☐	☐	biological regulation	2.5648E-9	3.2180E-7	177/267 66.2%	6937/14297 48.5%	SEPT5 NOG GABRB3 SLC9A3 STAT5B CGB...
☐	☐	multicellular organismal process	7.3021E-9	8.6530E-7	126/267 47.1%	4376/14297 30.6%	SEPT5 NOG KCNC4 GABRB3 PITPNA STAT...
☐	☐	regulation of cellular process	9.2842E-9	1.0423E-6	162/267 60.6%	6219/14297 43.4%	SEPT5 NOG GABRB3 STAT5B CGB7 PDX1 ...
☐	☐	organ development	1.6554E-7	1.7655E-5	64/267 23.9%	1793/14297 12.5%	NOG NRP1 STAT5B NFKB2 PDX1 CBFA2T3...
☐	☑	cell differentiation	3.4875E-7	3.5423E-5	60/267 22.4%	1668/14297 11.6%	NOG NRP1 STAT5B NFKB2 PDX1 CBFA2T3...
☐	☑	anatomical structure morphogenesis	5.2170E-7	5.0581E-5	48/267 17.9%	1218/14297 8.5%	CDK5R1 NOG NRP1 SOX1 ONECUT2 TH I...
☐	☐	cellular developmental process	8.7877E-7	8.1497E-5	60/267 22.4%	1714/14297 11.9%	NOG NRP1 STAT5B NFKB2 PDX1 CBFA2T3...
☐	☐	nervous system development	1.6825E-6	1.4953E-4	45/267 16.8%	1154/14297 8.0%	CDK5R1 NOG NRP1 SOX1 GLIS2 ONECUT...
☐	☑	organ morphogenesis	2.8487E-6	2.4305E-4	30/267 11.2%	635/14297 4.4%	NOG NRP1 SOX1 TH ONECUT2 AXUD1 PD...
☐	☐	positive regulation of macromolecule biosynthetic process	3.3420E-6	2.7417E-4	31/267 11.6%	674/14297 4.7%	GLIS2 STAT5B ONECUT2 PDX1 AXUD1 IL1...
☐	☐	neutral amino acid transport	5.6376E-6	4.4537E-4	5/267 1.8%	15/14297 0.1%	SLC1A5 SLC38A3 SLC7A10 SLC38A1 SLC...
☐	☑	positive regulation of transcription	9.9952E-6	7.6142E-4	27/267 10.1%	574/14297 4.0%	GLIS2 STAT5B ONECUT2 PDX1 AXUD1 IL1...
☐	☐	positive regulation of cellular biosynthetic process	1.2254E-5	9.0127E-4	31/267 11.6%	719/14297 5.0%	GLIS2 STAT5B ONECUT2 PDX1 AXUD1 IL1...

Fig.4. Gene ontology of named genes identified in the calling cards experiment. Of the 348 named genes, 297 were identified in the clusters of biological processes with the software described below. TOP: Gene network created by Cytoscape Desktop and BiNGO ontology software. The circles represent biological processes (nodes) associated with the identified genes found in a statistically significant manner. The size of the circle represents the number of genes within a node from the data set and the color denotes the significance. Lines/arrows between the nodes link biologically related processes. Three main branches of the ontology tree appear: the top includes processes involved with development, the small group to the left includes metabolic processes, and the large group of nodes near the bottom includes processes related to gene transcription/regulation. BOTTOM: List of gene ontology (GO) terms that are overrepresented in the data set compared to a reference set (all human genes). The functional groups (GO terms) are listed in order of statistical significance. p-values are calculated by two statistical test: hypergeometric and binomial. Corrected values attempt to eliminate false discovery rate with Benjamini & Hochberg correction. Cluster frequency describes the proportion of genes from the data set that were within a functional group. Total frequency describes the frequency of total genes in the reference that belong to that group. “genes” includes some of the identified genes in the data set.

CONCLUSION

- The NUP98-HOXA9 fusion protein binds a number of genes/genomic loci in both mouse ES cells and human hematopoietic stem cells
- Experiments in mouse ES cells served as “proof-of-principle” for the Calling Cards technique
- Experiments in human hematopoietic cells identified over 160,000 unique insertion events by the PB-transposase, signifying that the fusion product binds these sites
- These unique insertion events comprised 724 clusters of insertion events, which were within 575 Ensembl genes
- A number of the binding sites are genes known to be involved in leukemogenesis
- Bound genes involved in gene regulation and development are overrepresented in the data set in a significant manner
- “Calling cards” is a novel approach to identifying protein-DNA interactions

- FUTURE DIRECTIONS:
- Validate the most likely candidates in cell culture and mouse experiments